TECHNICAL REPORTS
Annexed to the Final Report
of the ESPON 2013 Database Project

❖ Application

- Thematic structuring and variables labeling within the ESPON 2013 Database: An empirical method *(27 p)*
- Text mining methods and visualization tools as means to support the thematic structuring of the ESPON 2013 DB *(35 p)*
- ESPON Database application - Towards a web interface for the ESPON 2013 Database *(15 p)*
- Update of the ESPON Database 2006 into the 2013 version *(27 p)*

❖ Methodological issues

- Towards an approach of time series data issues: from empirical methods to application *(31 p)*
- Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase *(37 p)*
- Spatial analysis for quality control 1 *(81 p)*
- Spatial analysis for quality control 2 (28 p)
- Using downscaled population in local data generation – A country level examination *(16 p)*
- Mapping guide for ESPON Projects and the external community *(33 & 24 p)*

❖ Enlarging the data collection to new scales and cities

- World database – Towards a World Dictionary of units (46 p)
- Analysis of the availability and the quality of data on Western Balkans and Turkey (49 p)
- Local data – First investigations in Romania, Bulgaria, Czech Republic and Slovakia (26 p)
- Local and regional data – Producing innovative indicators (25 p)
- Naming UMZ: A database now operational for urban studies *(39 p)*
- LUZ specifications – Urban Audit 2004 *(72 p)*
- The Functional Urban Areas database *(16 p)*

From the Second Interim Report to the Final Report…
Information concerning the Technical Reports

❖ Obsolete Technical Reports – To be deleted

- Metadata guide - Acquisition and storage of Data and Metadata in ESPON 2013 Database*. Reason: New interface for data and metadata creation*
- ESPON Database application - Towards a web interface for the ESPON 2013 Database. *Reason: New interface for the ESPON Database*

❖ Unchanged Technical Reports

- Thematic structuring and variables labeling within the ESPON 2013 Database: An empirical method
- Spatial analysis for quality control
- Mapping guide for the external community
- World database – Towards a World Dictionary of units
- Local data – First investigations in Romania, Bulgaria, Czech Republic and Slovakia

❖ Updated Technical Reports (with new elements)

- Towards an approach of time series data issues: from empirical methods to application
- Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase
- Naming UMZ: Methods and results
- Using downscaled population in local data generation – A country level examination
- Mapping guide for ESPON Projects
- Analysis of the availability and the quality of data on Western Balkans and Turkey

❖ Additional Technical Reports (brand new!)

- The ESPON Database application[1]
- Text mining methods and visualization tools as means to support the thematic structuring of the ESPON 2013 DB
- Spatial analysis for quality control. Phase 1: The identification of logical input errors and statistical outliers (second draft)
- LUZ specifications (Urban Audit 2004)
- The Functional Urban Areas database
- Update of the ESPON Database 2006 into the 2013 version
- Local and regional data – Producing innovative indicators

[1] This Technical Report will be finished until the end of February 2011, when all the new functionalities of the ESPON Database application will be definitively implemented.

TECHNICAL REPORT

ESPON

Thematic structuring and variables labelling within the ESPON 2013 DB: An empirical method

MAIN RESULTS

- International database classifications used as source of information for structuring the ESPON 2013 DB
- We employ a visual grouping technique to identify a first set of themes
- Data allocation into themes and sub-themes
- Development of an harmonised scheme (TtOYS) to code ESPON indicators
- Future work should validate the usability of this method and enhance sub-themes definition
- Text mining tools will support keywords extraction from ESPON interim reports, naming conventions improvements and optimized coding

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

32 PAGES

LIST OF AUTHORS

Nuno Madeira, University of Luxembourg
Geoffrey Caruso, University of Luxembourg

Contact

E-mail: nuno.madeira@uni.lu; Tel. +352 46 66 44 9691
E-mail: geoffrey.caruso@uni.lu; Tel. +352 46 66 44 6625

TABLE OF CONTENT

Introduction

The ESPON 2013 DB aims to improve the access to regional and spatial information. This process has been initiated by the previous ESPON Programme in order to increase the number of indicators and indices that may positively support analysis of spatial structures and trends across European cities and regions (see, for instance, ESPON project 4.1.3).

The goal of this technical report is to determine a short-term solution to structure the ESPON 2013 DB by themes and sub-themes. This report complements the technical report entitled "Towards an ESPON thesaurus? Some preliminary considerations for the thematic structuring of the ESPON database" that seeks to derive themes and subthemes from a corpus of words and concepts mentioned in various EU reports. In the current report we argue that database classifications, nomenclatures and taxonomies developed by other organisations should also be considered when structuring the ESPON 2013 DB. The reason is quite straightforward. That is, many of those databases have established common themes that often aggregate similar data.

By focusing on the main themes of each database we use the information to analyse the similarities of the classifications. Additionally, we employ matrix visualisation techniques to assist us in looking at the data and therefore make the description more comprehensible.

The results will be then used to further progress on the user interface prototype and hopefully constitute a robust basis for improving the performance of text mining methods (see previous technical report). Arguably, it is worth mentioning that methods employed in this report will only take into account statistical and geographical sources used to develop indicators by applied research projects under Priority 1 and 2 of the ESPON 2013 Programme. In other words, only indicators delivered up to date will be considered in this analysis (see Appendix 1).

As a second step, we propose to link each indicator to a theme and sub-theme. Eventually, this process will facilitate harmonisation of codes – variable names – defined by the other ESPON projects in an uncoordinated manner. This is significant because it would offer some consistency to the entire database and assist other research projects when naming indicators, indices and other measures used by ESPON to evaluate territorial trends, structures and policy impacts in Europe.

1 Research background and methodology

1.1 Research background

As a first approach we assembled a list of first-level themes defined by organisations on which current ESPON projects have obtained raw data, namely UNEP, EEA, EUROSTAT, OECD, UNESCO, WDI, and ILO[1]. This is meaningful because most of these databases have provided and will continue to provide raw data both in terms of environmental and socio-economic matters to develop ESPON indicators and indices. With this regard, each word or expression used as a theme has been listed, evaluated in terms of similarity, and ultimately aggregated into similar themes. However, we must point out that the aggregation of words into thematic clusters has been purely inductive and based on the semantic value of each theme. For detailed information, please see Annex 2 and 3 to this report.

The dataset consists of 85 words or expressions taken from the seven database classifications. Each database classification has in average 18 first-level themes. Both UNEP and WORLD BANK share the largest classification with 26 themes whereas UNESCO has structured its database with only 6 themes.

A prior step in this analysis is data preparation. The input data matrix is described by a binary (presence/absence) relationship model as shown in Table 1. That is, all values range between convergent (1) and divergent (0). Table 1 lists some of the words (rows) and database classifications (columns) employed in this analysis. If we take the first example, we would be able to understand that '*Tourism'* is considered as a first-level theme by UNEP and EEA while other databases do not devote the same attention to such topic. On the other hand, '*Unemployment'* has only been labelled as a first-level theme by ILO. This is reasonable due to the purposes of each database.

	UNEP	EEA	EUROSTAT	OECD	UNESCO	ILO	WPI
(…)							
Tourism	1	1	0	0	0	0	0
Trade	1	0	1	1	0	0	1
Transport	1	1	1	0	0	0	0
Unemployment	0	0	0	0	0	1	0
(…)							

Table 1: *Short example of data input for analysis*

In order to understand the structure generated by this binary matrix some graphical techniques have been applied to determine clusters, identify blocks within the matrix and increase visual perception of commonly used themes. Following the well-known methods developed by Bertin (1967), we explore the concept of matrix visualization and cluster analysis offered by generalized association plots, or GAP

[1] For more detailed information on each database classification, please visit the following Internet sites: UNESCO (http://stats.uis.unesco.org); ILO (http://laborsta.ilo.org); EUROSTAT (http://epp.eurostat.ec.europa.eu), OECD (http://www.oecd.org/statsportal), EEA (http://themes.eea.europa.eu), UNEP (http://geodata.grid.unep.ch), WDI (http://ddp-ext.worldbank.org).

(Chen, 2002; Wu et al., 2008). This open source tool can be understood as recordable matrix to communicate data structures and patterns. Basically it offers the possibility to visualise raw data and display tabular quantities and relationships by means of colour-based representation. The output of such experiment is displayed in a rather natural, inductive perspective but sufficiently helpful to identify proximities between subjects and variables.

The proximity measure one can employ to relate objects in such an experiment depends, on the data type (i.e. binary, nominal, ordinal, etc). Within this context, the choice of proximity measure has an effect on the association patterns which directly influences the visual representation of the interaction structure (Wu et al., 2008). GAP offers some specific measurements for asymmetric information. As our matrix corresponds to a binary data type (presence or absence of a theme in a given classification) we have applied Jaccard's coefficient.

1.2 Methodology

The choice of each database derives from the fact that ESPON evidence is strongly based on raw data provided by those above mentioned institutions. As a consequence, it seems appropriate to consider each database classification and validate by means of generalized association plots the degree of similarity and dissimilarity. The usefulness of such approach is to harmonise words or expressions used by some of the most prominent statistical databases and, therefore, enabling policy-makers, practitioners, and researchers in the field to adopt a common language of understanding.

The matrix visualization is illustrated by a series of images that explore correlation between themes (subjects) and databases (variables). In order to capture potential differences among those databases we decided to include the ESPON 2006 DB structure of first-level themes and identify specific features that could validate or refute our cluster analysis (see Appendix 2 and 3). To this end, each exercise is illustrated by two matrices as an attempt to reveal possible changes. Clearly, some patterns can be discerned from those matrices. Next, we will explore and understand the structure embedded in each data matrix and determine a hierarchy of themes that could support the thematic structuring of the ESPON 2013 DB.

As a first step, we added to our correlation matrix the classification defined by the previous ESPON database (ESPON, 2005) and applied the same methodology. Surprisingly, some of the results indicated a weak correlation between ESPON 2006 DB and other classifications. Even though, EUROSTAT has the strongest similarity whereas UNEP and UNESCO reveal less significant correlation coefficients. Somehow this reflects how crucial it would be for ESPON to be in accordance with main data providers.

In order to demonstrate the existing similarities between different classifications we employed a simple correlation analysis. The different goals defined for each institution's database led to low correlation values. However, some interesting results emerged from this exercise. For instance, it is clear from the correlation matrix that EUROSTAT and OECD share the strongest correlation value (0.50). One reason that could be claimed to justify the degree of resemblance between these two classifications is the nature of the content. Indeed, the fact that EUROSTAT

and OECD collect and disseminate similar data for similar audiences has produced an impact on the classification of both databases.

The opposite scenario, i.e. weak correlation values, is rather frequent and little interpretation can be discerned. Still some explicit assumptions must be stressed due to its degree of clearness, particularly among environmental databases. More precisely, the fact that those sources are committed to cover specific issues such as *environmental hazards*, *marine and coastal areas*, or *air pollution* (see Appendix 2 and 3) that often require detailed data also intensifies the number of discrepancies in most of the themes or categories adopted by each organisation. Preliminary results of this analysis are illustrated by Table 2.2.

	UNEP	EEA	EUROSTAT	OECD	UNESCO	ILO	WPI	ESPON 2006
UNEP	1.00							
EEA	0.22	1.00						
EUROSTAT	0.12	0.27	1.00					
OECD	-0.09	-0.12	0.50	1.00				
UNESCO	-0.10	-0.09	0.14	0.22	1.00			
ILO	-0.20	-0.11	-0.07	-0.10	-0.11	1.00		
WPI	-0.12	-0.15	0.22	0.24	-0.02	-0.07	1.00	
ESPON 2006	0.10	0.18	0.32	0.20	0.07	0.13	0.18	1.00

Table 2: *Correlation matrix of the database classifications employed in this experiment*

Interesting enough in this analysis is the fact that environmental databases tend to be more detailed when compared with socio-economic databases. To a certain extent, this ensures a high level of accuracy and promotes its utility for large audiences. However, there is an enormous discrepancy on the content provided by each environmental classification. On the contrary, both EUROSTAT and OECD have defined a broad list of categories to search and retrieve socio-economic data. As a consequence, semantic similarities are higher and the degree of resemblance between those two entities is much stronger.

Despite the purpose and content of each database it is obvious that organisations do not give much importance to labelling harmonisation. Given the role of the ESPON 2013 Programme for policy advice and development, the ESPON 2013 DB project constitutes a major opportunity to demonstrate the advantages in establishing a harmonised thematic structuring that could rely on classifications defined by the main data providers but also taking into consideration the INSPIRE initiative for the creation of an European Spatial Data Infrastructure.

2 Matrix visualisation techniques for cluster analysis

As explained before this approach will be part of a short-term solution that wishes to integrate text mining methods to extract major themes and sub-themes from a large corpus of qualitative and unstructured data (e.g. ESPON and other related EU reports). We assume therefore that such methods have the capacity to define standards that can lead to improved harmonisation and coherence of spatial concepts and eventually organise knowledge for information retrieval by end users. Next we discuss the results obtained by GAP to determine clusters and identify blocks.

The figures presented in Table 3 demonstrate a clustering of words. We decided to sort data by the GAP ranking that includes the ESPON 2006 DB. The ranking has no absolute meaning but the relative position of words is useful to interpret. GAP ranking is actually the result of a permutation of words so that words that share a similar pattern of presence/absence within the different classifications are positioned in neighbouring rows. (We used the single linkage algorithm to obtain the blocky structure of rows from the permutation) . Figures 3 and 4 to this report display some of the techniques to help identify blocks. Despite its value in terms of matrix visualisation we will give a primary focus on Figures 1 and 2 (for details, see Annexes).

2.1 Discussion of preliminary results

Our initial assumption is that GAP offers very helpful features to interpret data matrix association, patterns and ultimately behaviours. This helped to identify some of the key ideas underlying matrix visualisation needs, namely in terms of adopting a practical solution to display matrices. In fact, the main advantage of such tool corresponds to what Wilkinson & Friendly (2009) designated by *cluster heat maps*. The expression itself is very fortune because it gives the idea of clusters by shading association. That is, data matrices structured by similarity and/or dissimilarity to facilitate analysis and interpretation.

In this section we report our results using GAP (Wu et al., 2008). That is, considering database classifications to illustrate by means of correlation matrices relevant patterns that could easily be interpreted and communicated. More precisely, we propose to find relatively homogeneous clusters of themes. In order to enrich our analysis the number of citations by theme will also be taken into account. Then, we discuss the results from this experiment to propose a first set of themes. Ultimately, the results are compared and clusters are interpreted with respect to the indicators collected up to date for the ESPON 2013 DB.

The preliminary results have provided substantial information to comprehend our data collection. According to Figure 1 (see Appendix) it became clear that certain themes are very representative to the different databases while others are less

visible. For instance, if we consider the bottom right hand corner of Figure 1 we observe that the correlation of certain themes (subjects) is very strong among the different databases (variables) employed. Themes such as *Agriculture*, *Population, Transport* or *Energy* are exceptionally transversal and consequently among the most-cited categories established by certain database classifications. This is significant and somehow justifies the need for adopting such themes within the ESPON 2013 DB.

Figure 2 (see Appendix enclosed to this report) does not include any reference to the ESPON 2006 structure. This was intentional as explained above. Indeed, after computing data the association matrix has slightly changed its appearance. With this regard, some themes have gained more visibility while others expressed a reverse tendency (both results are displayed on Table 3). However, it should be highlighted that the primary group of four themes identified in the previous matrix has been kept very alike (i.e. *Energy*, *Transport*, *Population*, and *Agricultur*e). Similarly, we have identified a less prominent group of themes, mostly clustered on environmental issues, but totally disconnected from the above mentioned cluster. Themes such as *Tourism*, *Land Use*, *Climate*, *Resources* or *Health* loose their importance if not included in the same matrix as ESPON 2006 DB.

Surprisingly enough in this experiment is the fact in both matrices the number of citations is fairly similar, respectively 25% and 28.6% (see Table 3). Two themes, however, react in a different way and demonstrate common behaviours. Both *Tourism* and *Land Use* assume different raking positions when GAP is employed and somehow the percentage of citations reflects that situation. This is extremely relevant because it justifies the ranking of each theme. Ultimately, it confirms that *Tourism* and *Land Use*, two themes credited to the previous ESPON database, are not so important when considering the entire group of words or expressions used in this experiment. An opposite dynamic is observed with *Trade* and *Environment* (3). Both themes are cited as much as those observed in the first cluster but apparently emerge too disconnected from the structure if the ESPON 2006 DB is considered. Despite this situation, it is clear that such themes should be aggregated to the first set of themes for the ESPON 2013 DB. Besides, it would compensate some of the environmental-oriented themes identified previously (i.e. *Water, Climate, Consumption*, *Resources*).

Table 3: *GAP ranking of words or expressions used as a theme*

Themes	GAP Ranking (including ESPON 2006)	GAP Ranking (excluding ESPON 2006)	Number of citations, including ESPON 2006 (%)	Number of citations, excluding ESPON 2006 (%)	Groups
Agriculture	1	1	62.5	57.1	
Population	2	2	75.0	71.4	
Transport	3	5	50.0	42.9	
Energy	4	6	50.0	42.9	(1)
Tourism	5	17	37.5	28.6	
Land use	6	19	37.5	28.6	
Climate	7	14	25.0	28.6	
Water	8	13	25.0	28.6	
Urban	9	15	25.0	28.6	
Consumption	10	16	25.0	28.6	
Resources	11	18	25.0	28.6	
Health	12	20	25.0	28.6	(2)
Trade	13	4	50.0	57.1	
Environment	14	3	62.5	71.4	(3)
Finance	15	11	37.5	42.9	
Development	16	22	37.5	28.6	
Social	17	10	50.0	42.9	
Regional	18	26	25.0	28.6	
Science	19	12	37.5	42.9	
Technology	20	9	62.5	57.1	
Fisheries	21	8	37.5	42.9	
Industry	22	7	37.5	42.9	
Communication	23	21	37.5	28.6	
Infrastructure	24	25	37.5	28.6	
Economy	25	24	25.0	28.6	
Education	26	23	37.5	42.9	(4)
Air	27	27	12.5	14.3	
Biodiversity	28	28	12.5	14.3	
Chemicals	29	29	12.5	14.3	
Coastals	30	31	12.5	14.3	
Waste	31	30	12.5	14.3	
Soil	32	32	12.5	14.3	
Seas	33	33	12.5	14.3	
Scenarios	34	34	12.5	14.3	
Pollution	35	35	12.5	14.3	
Noise	36	36	12.5	14.3	
Welfare	37	60	12.5	14.3	
Demography	38	61	12.5	14.3	
Taxation	39	62	12.5	14.3	
Services	40	63	12.5	14.3	
Productivity	41	64	12.5	14.3	
Patents	42	65	12.5	14.3	
Market regulation	43	66	12.5	14.3	
Globalisation	44	68	12.5	14.3	
Information	45	67	12.5	14.3	
Boundaries	46	49	12.5	14.3	
Vegetation	47	50	12.5	14.3	
Elevation	48	52	12.5	14.3	
Threatened (species)	49	51	12.5	14.3	
Slopes	50	53	12.5	14.3	
Fertilizer	51	57	12.5	14.3	
Food (supply)	52	59	12.5	14.3	
Pesticides	53	54	12.5	14.3	
Marine	54	55	12.5	14.3	
Land cover	55	56	12.5	14.3	
Hazards	56	58	12.5	14.3	
Employment	57	44	25.0	14.3	
Labour	58	48	37.5	28.6	
Household	59	39	37.5	28.6	
Wages	60	40	12.5	14.3	
Consumer price (indices)	61	42	12.5	14.3	
Unemployment	62	41	12.5	14.3	
Strikes & lockouts	63	43	12.5	14.3	
Occupational (injuries)	64	45	12.5	14.3	
International labour migration	65	46	12.5	14.3	
Hours of work	66	47	12.5	14.3	
Wealth	67	-	12.5	-	
Spatial typologies	68	-	12.5	-	
Research	69	-	12.5	-	
Public sector	70	-	12.5	-	
Culture	71	37	12.5	14.3	
Literacy	72	38	12.5	14.3	
Balance of payments	73	69	12.5	14.3	
Exchange rates & prices	74	70	12.5	14.3	
External debt	75	72	12.5	14.3	
Governance	76	73	12.5	14.3	(5)

The results summarized by Table 3 reveal as well other groups of themes that may require further attention. The main feature of the fourth group is related with the predominant focus on socio-economic issues. Themes such as *Finance*, *Development*, *Science*, *Infrastructure* or *Education* assume greater importance within this cluster. On one hand, this is essentially due to the ranking defined by GAP when grouping themes that intersect both OECD and EUROSTAT database classifications. On the other, it justifies the fact that most of these themes are linked to economic, social and development-oriented data. Nevertheless, it is also clear from Table 3.1 that an independent subgroup emerges within this primary group of themes. Indeed, it seems that the choice of computing a correlation matrix without including the ESPON 2006 DB structure has produced some significant impacts on the permutation result, especially on the position of *Technology*, *Fisheries* and *Industry*. Our interpretation is that those themes are strongly linked with the classification adopted by EEA and the motivation for this behaviour seems to stem from the fact that ESPON has not been considered in one of those occasions.

From this point onwards the structure is much more balanced both in terms of ranking and number of citations. This means that little interpretation can be discerned if the ESPON 2006 DB classification is employed by one of the correlation matrices. Next, we argue that those less prominent themes should be included or grouped within bigger groups since most of them are often related to a specific theme. This process has been developed in a rather inductive way and merely based on the semantic value or weight attributed to each theme. That is, the meaning of a given word (or expression) will define its value or weight when compared with themes and therefore determine the level of closeness.

As stated above, this section justifies the choice of aggregating some themes that otherwise would be completely disconnected from our analysis. Consequently, we should stress that this experiment has to a considerable extent been influenced by the level of semantic closeness to other major themes previously identified. Against this background, it seems obvious that an important set of less prominent terms (or expressions) should be treated as environmental-oriented issues. A strong argument to support this view is related to the fact that most of those themes derive from environmental database classifications such as EEA or UNEP. Thus, it is not surprising that our aggregation method considered domains on *Biodiversity*, *Waste*, *Elevation*, or *Slopes* as traditional environmental issues. The same applies to socio-economic issues largely labelled as integrative components. For instance, we noticed that *Taxation*, *Market Regulation*, *Employment*, *Labour*, or *Wages* can be understood as basic socio-economic themes that characterize the diversity of data published by OECD or ILO on their respective portals.

For those terms (or expressions) where uncertainties arise we adopted a more pragmatic solution. Themes like *Globalisation*, *Governance*, or *Welfare* which may be interpreted as very general concepts with meanings that often gravitate between different subjects, we decided to analyse what type of data was being labelled as such. Indeed, we noticed that such themes have not been equally considered by the database classifications employed in this experiment. Somehow, this explains the singularity and different purposes attached to each database classification.

2.2 Towards a first set of themes

The thematic structure of the ESPON 2013 DB should not be seen as a normative approach, but rather as a descriptive one. However, the choice of themes itself is very crucial for the success of the ESPON 2013 Programme because it offers the possibility to support policy development which can and will be used by different target groups (e.g. policy makers, researchers, academics, or practitioners) who wish to promote policy documents, technical reports, or academic studies. Moreover, data publically available for retrival on the ESPON 2013 DB can be used as a source for developing trends and scenarios.

This has significant gains for policy development on European spatial planning but most likely is subject of criticism. Indeed, one could ask if this theme or that were emphasized more, or if an attempt was made to add one theme or another. We believe that our preliminary results should be seen as images of the future or as elements that correspond to the needs of a particular moment. We listed below a first set of themes to help end users to understand the structure we propose for the ESPON 2013 DB. Taking into consideration the methodology applied in this experiment, we label the themes as follows:

01. Agriculture & Fisheries

02. Demography

03. Transport

04. Energy & Environment

05. Land Use

06. Social Affairs

07. Economy

99. No-Cross-Thematic Data

99.01 Integrative indices, typologies and scenarios

99.99 Geographical objects

Table 4: *Preliminary first-level thematic structure for the ESPON 2013 DB*

This list aggregates themes used by the main data providers. Occasionally, the meaning of the word derives from similar terms or expressions. This was the case for *Social Affairs* that often recalls societal-related issues that have great effects on many members of those societies and, for that reason, considered to be problems (e.g. poverty, unemployment) or matters that need further improvement (e.g. healthcare, education). We also added a group to cover cross-thematic and non-thematic data. A first subset then includes variables that mix themes on purpose (e.g. integrative indicators, complex typologies, scenarios), or for non thematic data such as base maps. The second subset refers to base maps (administrative units) and other geographical objects (e.g. grids, cities, networks) or spatial delineations (e.g. morphological zones, functional areas).

Those themes that have not been mentioned in this list should be considered as less interesting for the moment, although this assumption should not be taken as granted. Besides, it is not feasible to address all the relevant political, environmental or social matters. Nevertheless, we can still consider different

approaches to conjecture about the degree to which different topics will develop and gain more or less visibility. For instance, we argue that our on-going experiments with text mining tools have the capacity to identify key words on documents and reports that both employ and communicate ESPON evidence. We assume that such approach would contribute to a comprehensive thematic structuring of the ESPON 2013 DB (see previous technical report). For the moment, it is not obvious that this analysis will introduce new themes or sub-themes within the predefined structure. The emphasis on a particular theme also depends on other variables such as data deliveries (i.e. indicators, indices, typologies), demand from users and potential users, or even EU policy agenda. Whether this occurs or not, many other themes and sub-themes are likely to be added to the ESPON 2013 DB.

2.3 Allocation of data into themes and inductive definition of sub-themes

Most likely the demand from end users of the ESPON 2013 DB will be characterised by immediate, easy and practical access to data. A properly constructed classification is therefore the key to meet this request. The next step in this analysis is to allocate data into themes previously defined. For this purpose, we will consider data from of the ESPON 2006 Programme and data delivered up to date for the ESPON 2013 Programme, i.e. 30 October 2009. During the course of this analysis we also suggest a potential second theme that could improve classification and data retrieval. If some doubts subsist in our evaluation we propose other words to describe data. Hopefully this rather inductive process will rationalize the ability to restrict a search when seeking specific information and allow end users to achieve greater level of precision and recall.

The definition of sub-themes is intended to be data-driven and occasionally some of the less prominent terms (or expressions) that came out from our experiment will be used to complement the thematic structuring. Similarly, we propose to further explore the potentialities of text mining methods to extract key words (see previous technical report). For the moment, we will make use of sub-themes defined by the previous database (with some exceptions). This should be seen as a temporary solution to overcome some of the difficulties that arose during the course of this analysis.

Details of these allocation processes into themes and sub-themes are summarized in Appendix 4 to this report for the variables delivered up to date within the ESPON 2013 DB (i.e. 30 October 2009).

3 Naming conventions and coding scheme

Naming indicators is an important component of indicator development. Therefore research teams should strive to be objective and consistent. Taking into consideration the updated list of indicators (see Appendix 1) we noticed a wide variation of naming conventions that differs according to the criterion defined by each research team. Indeed, some teams have chosen very descriptive names that precisely define what is being measured while others have chosen to use more simplistic names that capture the essence of what is being measured. The latest list of ESPON indicators discloses, however, some similar indicators. These indicators have been examined in order to reduce redundancy and other potential overlaps. The example of *unemployment* is very illustrative. We noticed that data on unemployment was labelled in three different ways by TIPTAP, ESPON 2013 DB, and TeDi projects, respectively as: *Unemployment*, *Unemployed persons*, and *Number of unemployed persons, total*. Consequently, the benefit of developing consistent definitions for commonly used terms would allow us to harmonise the naming conventions and avoid the duplication of indicators.

Given that this is a very difficult matter to resolve, mainly because we are dealing with textual information. Moreover, naming conventions should not be seen as a theme to replace metadata. Thus it requires additional efforts, such as the development of a glossary or handbook to assist in clarifying terms that could potentially be used to label ESPON indicators. For the moment, this is not being considered as an option to keep some consistency. However, similar exercises could be conducted in the future to overcome this difficulty.

	DEMIFER	*TIP TAP*	*TeDi*	*ESPON DB*
Total Population	POP	-	D-NS_1a	pop_t
Unemployment	-	PIM_E2_DEF	E-NS_4a	unemp
Active Population	-	-	E-NS_1a	activ

Table 5: *Examples of indicators with arbitrary naming convention*

Another problem that emerged alongside to this experiment deals with coding systems. As it can be observed in Table 4.1 some of the applied research projects under Priority 1 and 2 of the ESPON 2013 Programme have defined their own rationale to label indicators. Despite the usefulness of such exercises, the truth is that research teams are increasing the degree of ambiguity when apply different methods to label the same indicators. This is often the case among well-popularized indicators. Again, the example of unemployment is very illustrative because it has already been tagged in three different ways and the labelling method differs from research project to research project.

Within the ESPON DN project this situation is becoming increasing problematic to further progress on user interface prototype. Indeed, if no harmonisation is employed the capacity to deduce information from the codes becomes rather difficult. To a certain extent, coding conventions are not used to express the content of data but rather an attempt to homogenise codes for indicators, indices and other measures. However, some information needs to be provided and most importantly it needs to be arranged in a consistent way to avoid such problems in the future. We therefore propose a set of guidelines that could positively contribute to harmonise the coding system of indicators that have been, and will be delivered by the different consortiums involved in ESPON.

4 TtOYS coding scheme to label indicators

In this section we introduce the **TtOYS** coding scheme to label ESPON indicators. TtOYS is an abbreviation for **T**heme, sub-**t**heme, **O**pen field, **Y**ear, and **S**pace. It serves the purpose of assembling relevant information about data to code ESPON indicators using a minimum number of characters (Table 6).

Theme		Sub-theme		Open field								Year				Space	
#	#	#	#	A	B	C	d	e	f	_	_	_	_	#	#	X	X

Table 6 *TtOYS structure to code variables*

The coding scheme for each indicator consists of five fields and can be fulfilled with up to eighteen characters. It recommends two characters for *theme*, *sub-theme*, and *space (i.e. type of geographical object)*. Conversely, both the year and the *open field* are much more flexible. It ranges between two up to four digits for the *year* (to allow description of a period of time) and from six characters to a maximum of eight for the open field. The restriction to eight characters for the open field corresponds to the limitation of a very widely used table format (DBF IV) within GIS data. To improve harmonization, we further propose that letters and numbers should be written in specific order and text displayed as either upper or lower case as it is proposed on this scheme and within our examples.

Results of the coding scheme applied to the current 2013 database can be found in Appendix 6. Exceptionally, we used a non-proportional font (i.e. `Courrier New`) to avoid change of width between codes though they have equivalent number of digits (18 including underscores when the 4 free digits are left empty).

We now further explain the content and rationale of each of the five fields and then provide indicative examples.

Themes and sub-themes (**Tt**)

The list of codes for themes and sub-themes provides already much of the information that is needed to catalogue each indicator. Details of this approach have been explained in the previous section to this report as well as in Appendix 4. The pairs of digits representing themes and sub-themes are simply indicated in the first four characters of the code.

Open field (**O**)

Beyond themes and sub-themes, it is necessary to give further details on the information that is being measured. We think it is impossible to fully harmonize this field given the chosen width restriction (8 digits) and the variety of indicators. Also some flexibility should be allowed. Nevertheless we propose a harmonization process and suggest three lists of abbreviations based on the current state of the database (see table 6). The first two lists will certainly be adapted with time as the database is enriched. They relate to subjects and to some adjectives and names

widely used when labelling indicators (e.g. total, gender, index, shares, change, etc….). The third list should preferably remain fixed since it corresponds to measurement scales as recognised in the geographical/statistical literature.

The process for structuring the field is then the following:

(i) Start with 3 upper case letters best identifying the subject. Where possible, pick up those 3 letters in the provided list of subjects (see table 7),

(ii) Refine the subject using 1,2 or 3 lower case characters (at your convenience, no list provided)

(iii) Complement the code with lower case characters using the proposed list of abbreviations for common adjectives and types (2 or 3 characters), and/or preferably the list of abbreviations for measurement scales (3 characters)

As explained earlier, the process is constrained since it should lead to a maximum of 8 characters, or only 6 characters when two dates are to be indicated (to reflect indicators of change between two periods) – see below.

We wish that users stick to the proposed structure of the open field and to the lists provided. Nevertheless, the first two lists are not exhaustive but based on the current state of the database. Moreover, we are well aware that in some cases adaptations will be needed particularly to obtain more degrees of freedom when facing rather complex but similar indicators. The structure is thus a guide but we think cannot be mandatory.

In table 7, the first list referring to subjects (see first column) is rather straightforward, usually referring to the first three letters of a word, or other letters representing at best the core subject tackled by a variable.

The second list refers to widely used labelling and abbreviations of variables. A typical example in demographic data is *gender*, abbreviated with *f* or *m*. Other abbreviations refer to commonly used terms to describe a variable such as *index*, *rate of*…, *relative*, *change*, etc. They sometimes directly relate to the nature of data (particularly terms such as *volumes, absolute, relative, rate*) though it is only loosely related to the measure itself. In most cases we would advise the authors to give preference to a strict description of the level of measurement.

We believe that the open field should contain an unambiguous description of the *level of measurement* (expression coined by Stevens, 1946) since the data needs to be described as accurately as possible. Levels of measurement are particularly important in order to allow the user to draw a direct relationship between data classification and the cartographic representation of data, as well as to capture its use within mathematical operations. Together with metadata it is an important feature that could lead in the future to an automatic ("intelligent") data management system.

Open field lists					
Subjects		*Common abbreviations*		*Levels of measurements*	
Name	*Code*	*Name*	*Code*	*Name*	*Code*
Accessibility	ACC	Absolute	abs	**Nominal**	**nom**
Active	ACT	Relative	rel	Nominal unique	nou
Births	BTH	Standardized	std	Nominal dichotomous	nod
CO2	CO2	Rate	rt	Nominal categorical	noc
Construction	CON	Index	ix	Nominal graded membership	nog
Congestion	COS	Share	sh	**Ordinal**	**ord**
Deaths	DTH	Change	ch	Complete ordinal	oru
Economic(s)	ECO	Average	av	Classed ordinal	orc
Employment	EMP	Male	m	**Interval**	**int**
Environment	ENV	Female	f	**Ratio**	**rto**
Firm(s)	FIR	Total	t	Extensive ratio	rte
Fisheries	FIS	…		Count ratio	rtc
Farm(s)	FRM			Derived ratio	rtd
Fertility	FRT			Density ratio	rde
Gas	GAS			Cyclic ratio	rty
GDP	GDP			Constrained ratio	rtp
Growth	GWH				
Landscape	LAN				
Life expectancy	LIF				
Land Use	LUS				
Manufacturing	MAN				
Migration	MIG				
Mining	MIN				
Market	MKT				
Natural	NAT				
Productivity	PDT				
Population	POP				
Regional	REG				
Retail	RET				
Safety	SFT				
Tourism	TOU				
Transport	TRA				
Traffic	TRF				
Unemployment	UMP				
…					

Table 7: *Non-exhausted list of abbreviations for the open field*
Note: First two columns derive from the current database of indicators and should be seen as merely indicative examples based on terms or expressions used. The third list (greyed) is derived from (Forrest, 1999) and is meant to be fixed.

In the literature, four levels of measurement are commonly distinguished following the proposals made by Stevens (1946) on the theory of scales and measurements. In ascending order of precision: *nominal*, *ordinal*, *interval*, and *ratio* data. Statistical analysis and most of the spatial analysis handbooks refer to those four scales (see e.g. (Haining, 2003)).

At the nominal level of measurement, the numbers are used to classify the data (e.g. land use). On the contrary, the ordinal scale illustrates some ordered or ranked relationship between categories (e.g. income category). Despite the fact that both levels correspond to categorical data the major difference between them lies on the hierarchical, non-sequential relationship

The interval and the ratio scales are quantitative data (numerical measures). The interval level has equal units of measurement, thus making it possible to interpret not only the order of scale but also the distance between them. Nevertheless, the zero point of an interval scale is arbitrary and is not a true zero. The ratio scale of measurement has a fixed origin or zero point. This is the most advanced scale. Most of common statistical methods of analysis however require only interval level of measurement.

Geographers, particularly those involved in GIS and cartography (e.g. Forrest, 1999, (Chrisman, 2002),(Slocum et al., 2005)) argue that those scales should be

refined when dealing with geographic data. We therefore choose to follow the naming and definition of measurement scales proposed by Forrest, 1999. Chrisman (2002, p.31-33) and Slocum et al., 2005, p.60-61) also discuss the rationale for these subdivisions. Nominal data are divided up into four types: *unique* (no duplicated value), *dichotomous* (binary data), *categorical* and c*ategories with graded membership*. Ordinal data are subdivided into *complete* and *classed ordinal* depending on whether all values are unique or not. Finally Ratio data are subdivided into 6 subtypes. The first two are often referred to as *volumes* or *absolute numbers* in cartographic literature and mapped with proportional symbols: *extensive ratio* (where additive properties apply) and *count* (number of something). Then follows those ratios that are mapped using choropleths and often referred to as *relative* data: *derived ratios* (resulting from the division of any quantity by another), *density ratio* (the denominator is a geographical surface) and *constrained ratio* (values bounded between 0 and 1, representing proportions or probabilities). The last subtype, less in use within the territorial agenda field, is the *cyclic ratio* (e.g. angles). A short description and examples for each of those measurement levels is provided in Appendix 5. Corresponding abbreviations are also displayed in Table 7.

Years and Space (YS)

Finally, regarding Years and Space we propose a rather pragmatic approach to code ESPON indicators. As stated above, **TtOYS** uses a minimum number of characters. This is important to keep the structure as simple as possible. Therefore these last two categories of the coding scheme will include a code for the year(s) of reference and description of the different geographical objects (e.g. NUTS, LAU, UMZ). The general idea is to follow a structure that appears to be most intuitive. Table 8 illustrates how these categories would like by following the coding scheme. As it can be seen, the code for each Year is associated with the last two digits whereas Space suggests a combination of words to specify the name of the geographical object and, if needed, numbers to identify the level unit. In order to understand changes over a period of time on Year code we kept the same rationale but added two more digits which can be placed on the available positions used by Open field (see Table 7).

***Y**ear*		***S**pace*	
Name	*Code*	*Name*	*Acronym*
1995	95	NUTS0	N0
2000	00	NUTS1	N1
2005	05	NUTS2	N2
2010	10	NUTS3	N3
1995-2000	9500	NUTS Mix	NX
2002-2010	0210	LAU3	L1
		LAU2	L2
		LAU3	L3
		LAU Mix	LX
		UMZ	MZ
		GRID	GR
		NETS	NT
		Other Mix	MX

Table 8: *Descriptive example of codes defined for Years and Space (YS).*

Example

As an example, we applied the TtOYS coding scheme on two different indicators (i.e. Migratory Population Change, 2001-2005; Potential Accessibility by Air [absolute level], 2006) to demonstrate the usefulness of our approach. We tested those examples using five types of features that comprise both rigid and flexible divisions. Table 9 illustrates the result that derived from our method in order to capture, as much as possible, the content of each indicator. First we added the code defined for each theme and sub-theme. Secondly, we determined the gender, subject, and level of measurement and, lastly, we included the year of reference and the geographical object (i.e. space). The second example, however, does not respect the rigidness of the previous one, mostly because it reflects a change over a period of time. The procedure is applied on approximately 140 ESPON indicators delivered up to date (see Appendix 5).

(a)

Theme		Sub-theme		Open field*								Year				Space	
0	2	0	2	M	I	G	t	v	o	l	_	0	1	0	5	N	2

* Subject, level of measurement and other common abbreviations.

(b)

Theme		Sub-theme		Open field*								Year				Space	
0	3	0	3	A	I	R	a	b	s	_	_	_	_	0	6	N	3

* Subject, level of measurement and other common abbreviations.

Table 9: *Two different examples of indicators coded with TtOYS. The first (a) reflects the example of "Migratory population change, 2001-2005", and (b) "Potential accessibility by air [absolute level], 2006".*

The result is rather helpful and easy to comprehend. Besides it constitutes an attempt to harmonize coding conventions. However, additional improvements are needed to further increase the quality of this proposal. At this point, is not possible to foresee or describe many of the indicators that will come out from the current and future applied research projects. This will require the involvement of the ESPON research community through a continuous, dynamic process.

Conclusions and future work

In this technical report, we have proposed a pragmatic solution for the thematic structuring for the ESPON 2013 DB. We assumed that international database classifications constitute an important resource of information for ESPON. To a certain extent, this helped to shape the structure of the previous database and certainly will influence the current developments. Therefore we applied a visual grouping technique (GAP) to illustrate, by means of correlation matrices, homogeneous clusters of themes. The results of our experiment constitute the basis to derive **a first set of themes** and eventually facilitate data allocation. The process itself was often time-consuming mostly because the list of indicators contained similar data labelled with different names and codes. As a consequence, we propose a harmonized coding system, **TtOYS**, to capture some of the main features that differentiate each indicator.

The present work points, however, to considerable future work, of both empirical and conceptual nature. At the empirical level, it is clear that we need to refine our understanding of what is being measured to better allocate each indicator to a specific theme and sub-theme. The quality of metadata is of course crucial in that regard. Perhaps more fundamentally, there are some open questions at the conceptual level. Primarily, the future work should validate the usability of this method. Secondly, it should better understand of what kind of knowledge is being labelled as such to ease data allocation, naming conventions and ultimately optimize codification. That is, extract commonly used terms or expressions from qualitative and unstructured data (e.g. ESPON Interim Reports, EU Policy Notes) to improve ESPON 2013 DB thematic structure and eventually offer some consistency on how to name indicators. As a consequence, some of the difficulties that emerged in this technical report should be further investigated by means of text mining tools.

References

Chen, C.-H. (2002). Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. *Statistica Sinica*, 12(1), 7-29.

Chrisman, N. (2002). *Exploring Geographic Information Systems*. New York: John Wiley & Sons.

ESPON (2005). *Integrated Tools for European Spatial Development. ESPON 3.1 Project.* Luxembourg: ESPON Coordination Unit, pp. 141-174.

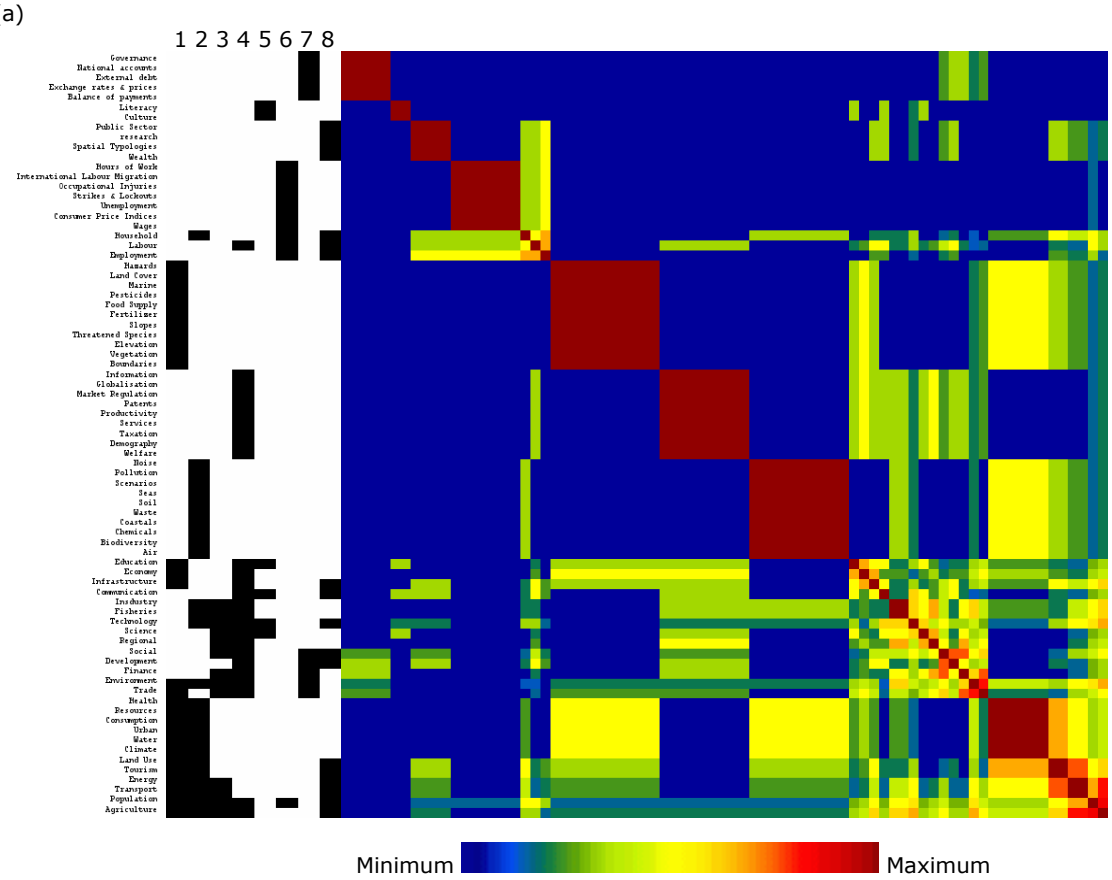
Forrest, D. (1999). Geographic Information: Its Nature, Classification, and Cartographic Representation. *Cartographica*, 36(2), 31-53.

Haining, R. (2003). *Spatial Data Analysis. Theory and Practice*. Cambridge: Cambridge University Press.

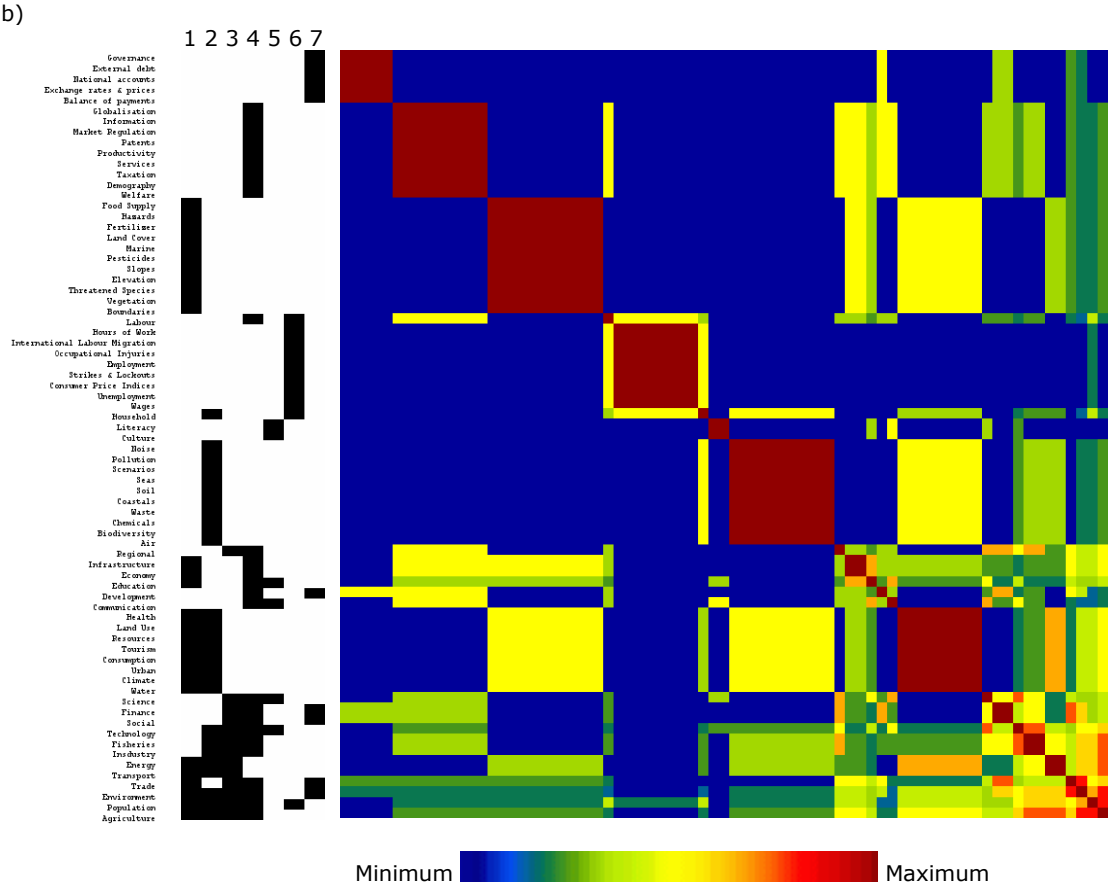
Slocum, T., et al. (2005). *Thematic Cartography and Geographic Visualization*. New Jersey: Pearson Prentice Hall.

Wu, H.-M., et al. (2008). GAP: A graphical environment for matrix visualization and cluster analysis. *Computational Statistics & Analysis*, in press.

(a)



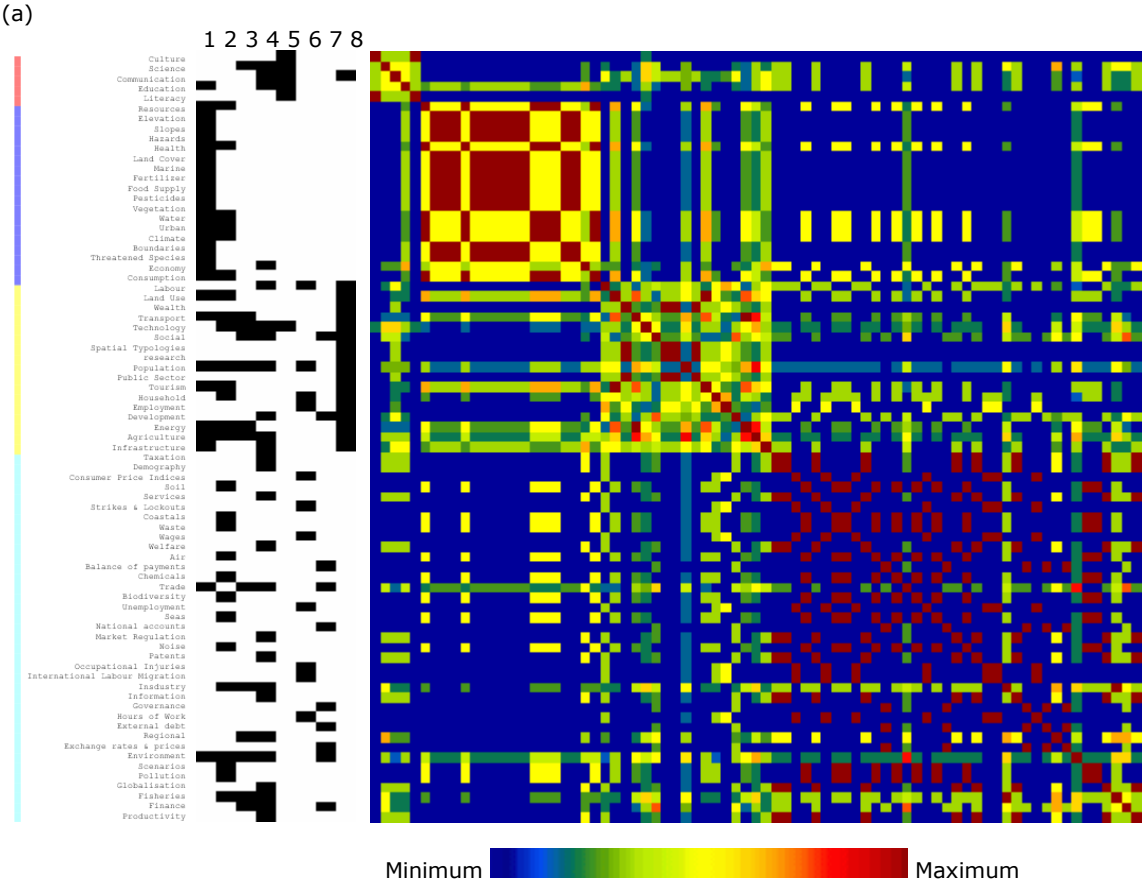
(b)



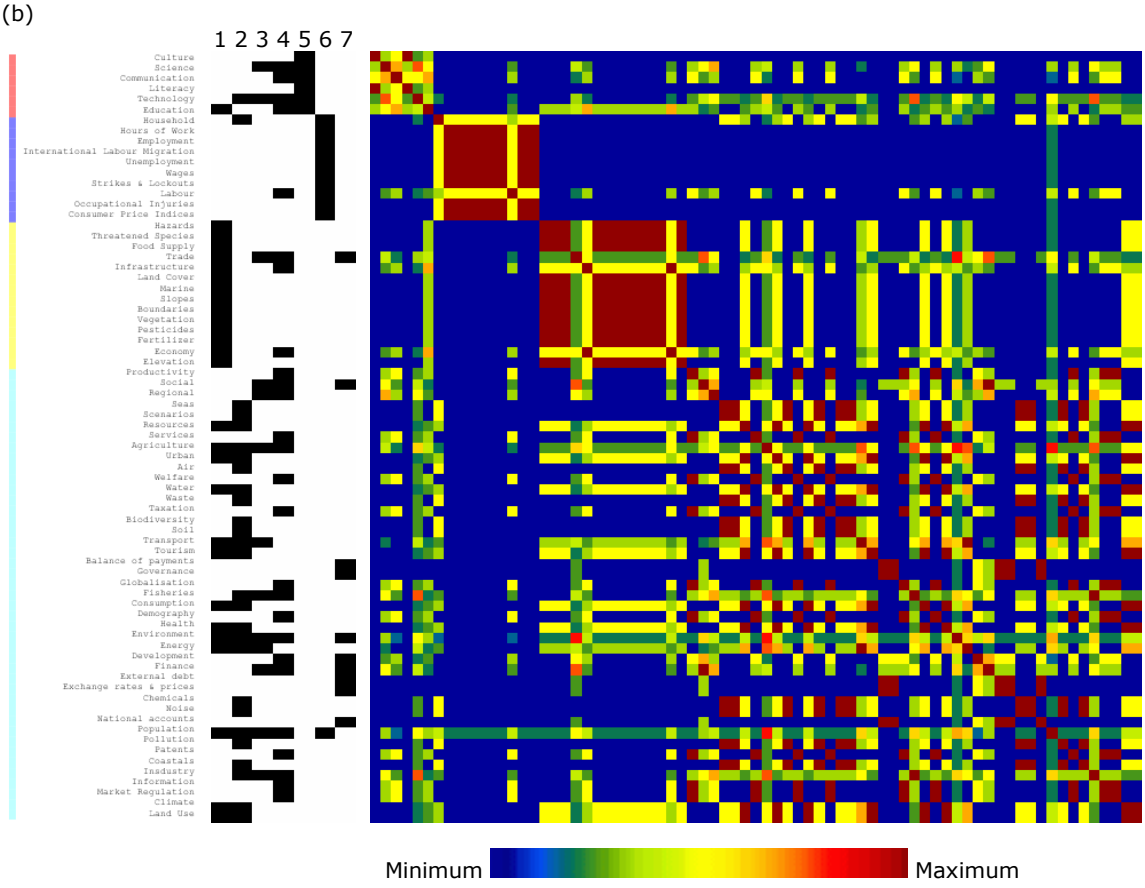
Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

Figure 1
Matrix visualisation in GAP environment of nomenclatures (subjects) used by statistical databases (variables). The first sorted matrix (a) includes ESPON 2006 structure whereas the second one (b) ignores it.

(a)



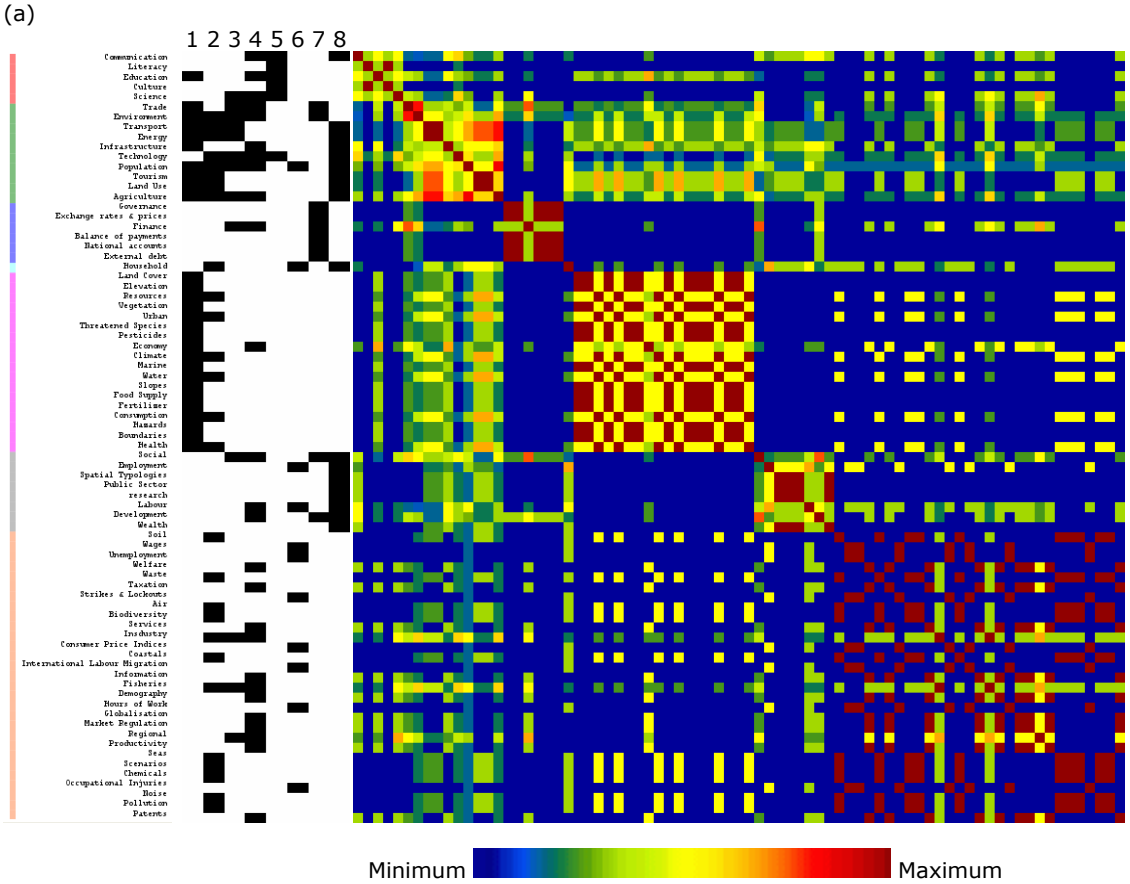
(b)



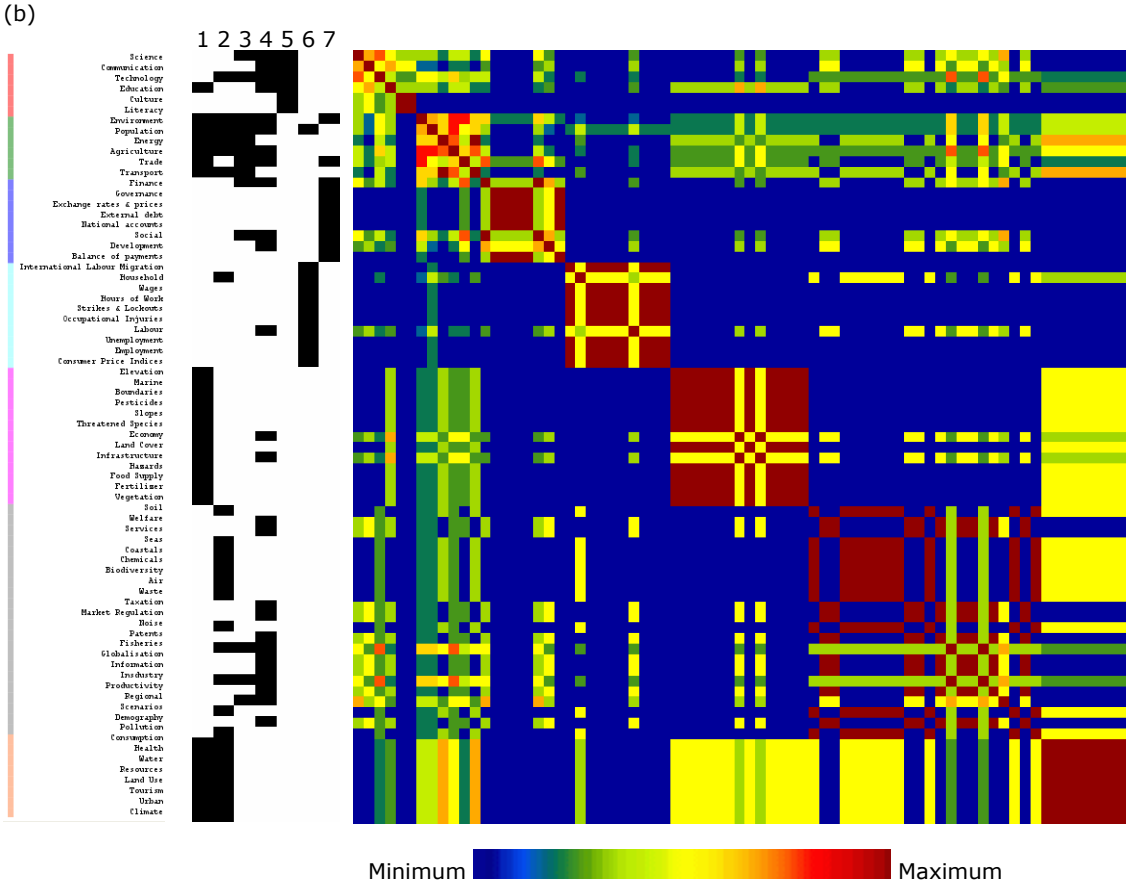
Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

Figure 2
Cluster analysis based on Jaccard's coefficient in GAP environment. The first sorted matrix (a) includes ESPON 2006 structure whereas the second one (b) ignores it.

(a)



(b)



Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

Figure 3
Cluster analysis based on Jaccard's coefficient in GAP environment. The first sorted matrix (a) includes ESPON 2006 structure whereas the second one (b) ignores it.

Appendices

Appendix 1: Description of ESPON indicators delivered up to date (i.e. 30 October 2009)

Report/Project	Indicator
T. Observation #1	Total population
T. Observation #1	Total Population change
T. Observation #1	Migratory population change
T. Observation #1	Core Indicator 1: Annual population growth rate
T. Observation #1	Core Indicator 2: Annual net migration development
T. Observation #1	Core Indicator 3: Annual natural population
T. Observation #1	Core Indicator 4: Annual natural population
T. Observation #1	Multimodal potential accessibility, absolute level
T. Observation #2	Multimodal potential accessibility, standardised
T. Observation #2	Multimodal potential accessibility, change
T. Observation #2	Multimodal potential accessibility, relative change
T. Observation #2	Multimodal potential accessibility, absolute change
T. Observation #2	Potential accessibility by air, absolute level
T. Observation #2	Potential accessibility by air, standardised
T. Observation #2	Potential accessibility by air, change of standardised
T. Observation #2	Potential accessibility by air, relative change
T. Observation #2	Potential accessibility by air, absolute change
T. Observation #2	Potential accessibility road, standardised
T. Observation #2	Potential accessibility road, relative change
T. Observation #2	Potential accessibility road, absolute change
T. Observation #2	Potential accessibility road, index change
T. Observation #2	Potential accessibility rail, 2006, EU27 = 100
T. Observation #2	Potential accessibility rail, relative change
T. Observation #2	Potential accessibility rail, absolute change
T. Observation #2	Potential accessibility road, index change
DEMIFER	Total population
DEMIFER	Population aged 20-39 years
DEMIFER	Population aged 20-64 years
DEMIFER	Population aged 65 years and over
DEMIFER	Population aged 75 years and over
DEMIFER	Annual average population change
DEMIFER	Annual average population change, 20-39 years
DEMIFER	Annual average population change, 20-39 years
DEMIFER	Share of 20-39 years
DEMIFER	Share of population aged 65 years and over
DEMIFER	Average share of population aged 65 years and over
DEMIFER	Life expectancy at birth
DEMIFER	Natural population change
DEMIFER	Net migration change
DEMIFER	Annual average natural population change
DEMIFER	Annual average migration population change
DEMIFER	Total fertility rate
DEMIFER	Internal net migration between the NUTS2 regions
DEMIFER	Basic typology of the demographic status 2005
ESPON2013DB	Unemployed persons
ESPON2013DB	Active population
ESPON2013DB	Total population
ESPON2013DB	Age pyramid by 5 years age-group
ESPON2013DB	GDP in euros
ESPON2013DB	GDP in PPS
TIPTAP	Productivity of inland transport infrastructure
TIPTAP	Productivity of airports
TIPTAP	Congestion costs
TIPTAP	Traffic freight passing through
TIPTAP	CO2 emissions by road traffic
TIPTAP	Safety of roads
TIPTAP	Market opportunities
TIPTAP	Landscape fragmentation
TIPTAP	Exposure to external visitors
TIPTAP	Regional integration
TIPTAP	Economic growth
TIPTAP	Unemployment
TIPTAP	Tourism diversification
TIPTAP	Environmental quality
TIPTAP	Community viability
TIPTAP	CO2 emissions
TIPTAP	Risk of soil erosion
TIPTAP	Landscape diversity
TIPTAP	Community identity
TIPTAP	Heritage products
TeDi	Land use
TeDi	Number of farm holders by age (24-75+)
TeDi	Number of farm holdings
TeDi	Number of persons working in the agricultural sector
TeDi	Number of passengers at airport
TeDi	Freights handled by airports
TeDi	Number of passengers at maritime ports
TeDi	Freights handled by maritime ports
TeDi	Total population, males
TeDi	Total population, females
TeDi	Population by age group
TeDi	Number of births
TeDi	Number of deaths

Report/Project	Indicator
TeDi	Number of in-migrants
TeDi	Number of out-migrants
TeDi	Number of persons born abroad
TeDi	Number of unemployed persons, total
TeDi	Number of unemployed persons, males
TeDi	Number of unemployed persons, females
TeDi	Active population, total
TeDi	Active population, males
TeDi	Active population, females
TeDi	Number of employed persons by economic branch
TeDi	Unemployed persons by age
TeDi	Long term unemployment
TeDi	Part-time unemployment
TeDi	Number of companies created and closed
TeDi	Number of employees by size of the company
TeDi	Number of persons by level of education

Appendix 2: Database classifications ordered by first-level theme

UNESCO
1 Education
2 Science & Technology
3 Culture & Communication
4 Literacy

ILO
1 Economically Active Population
2 Employment
3 Unemployment
4 Hours of Work
5 Wages
6 Labour Cost
7 Consumer Price Indices
8 Occupational Injuries
9 Strikes and Lockouts
10 Household Income and Expenditure
11 International Labour Migration

EUROSTAT
1 General and Regional Statistics
2 Economy and Finance
3 Population and Social Conditions
4 Industry, Trade and Fisheries
5 External Trade
6 Transport
7 Environment and Energy
8 Science and Technology

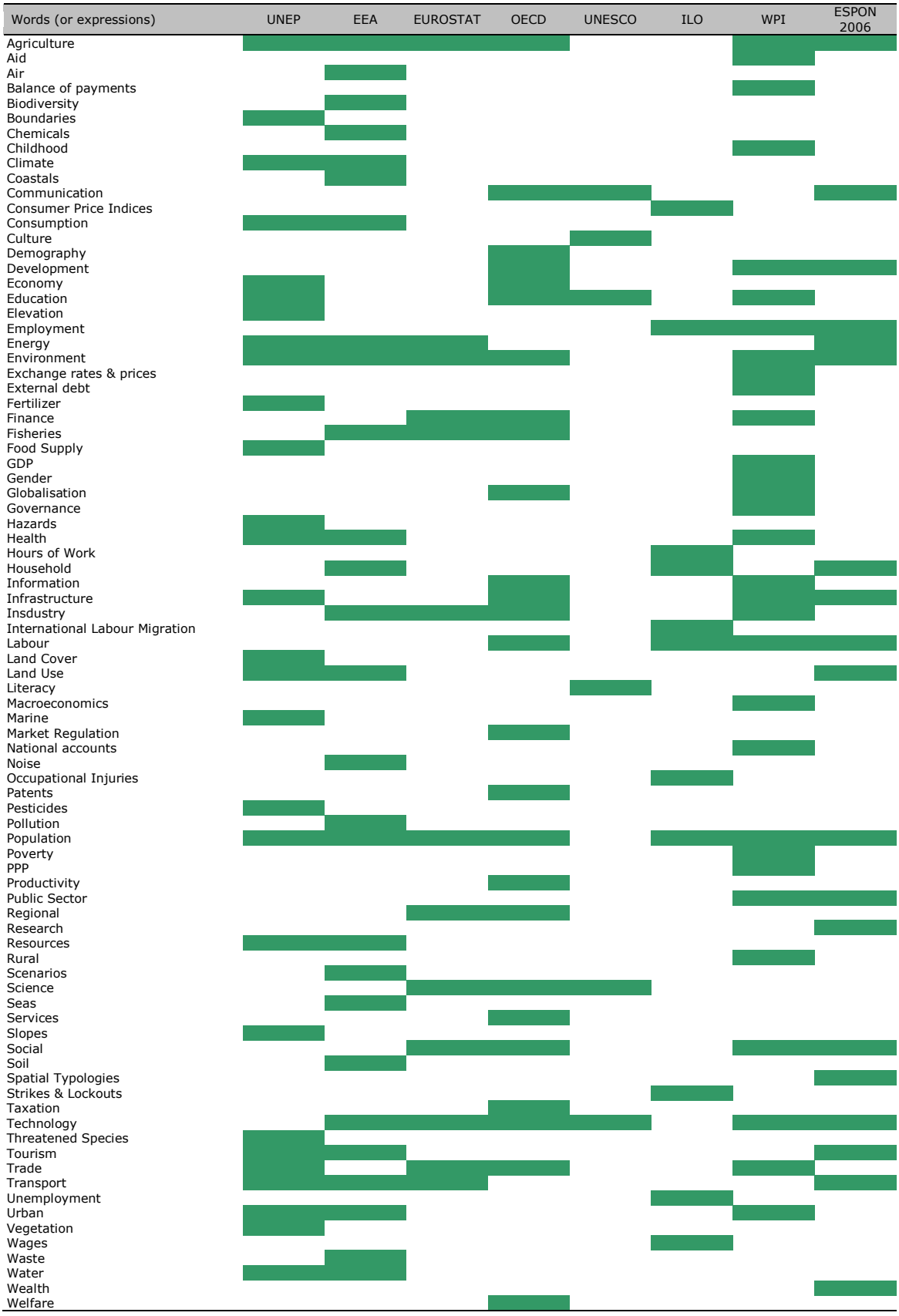
OECD
1 General Statistics
2 Agriculture and Fisheries
3 Demography and Population
4 Development
5 Economic Projections
6 Education and Training
7 Environment
8 Finance
9 Globalisation
10 Health
11 Industry and Services
12 Information and Communication Technology
13 International Trade and Balance of Payments
14 Labour
15 Monthly Economic Indicators
16 National Accounts
17 Prices and Purchasing Power Parities
18 Productivity
19 Public Sector, Taxation and Market Regulation
20 Regional Statistics
21 Science, Technology and Patents
22 Social and Welfare Statistics

EEA
1 Agriculture
2 Air
3 Biodiversity Change
4 Chemicals
5 Climate Change
6 Coastals and Seas
7 Energy
8 Environmental Scenarios
9 Fisheries
10 Households
11 Human Health
12 Industry
13 Natural Resources
14 Noise
15 Policy Analysis
16 Population and Economy
17 Regions
18 Soil
19 Tourism
20 Transport
21 Urban Environment
22 Waste
23 Water

UNEP
1 Agricultural Production
2 Boundaries
3 Climate
4 Economy
5 Education
6 Elevation and Slopes
7 Emissions of GHG and ODS
8 Energy Consumption and Production
9 Environmental Hazards
10 Fertilizer & Pesticides
11 Food Supply & Caloric Intake
12 Health
13 Infrastructure
14 Land Use
15 Marine and Coastal Areas
16 Population
17 Private Consumption
18 Protected Areas and Environmental Protection
19 Technological Hazards
20 Total and Threatened Species
21 Tourism
22 Trade Balances
23 Transport
24 Urbanisation
25 Vegetation and Land Cover
26 Water Consumption and resources

WORLD BANK
1 Agriculture
2 Aid
3 Childhood Development
4 Debt
5 Education
6 Environment
7 Finance
8 Gross Domestic Production
9 Gender
10 Globalisation
11 Governance
12 Health
13 Information Technology
14 Infrastructure
15 Industry
16 Labour & Employment
17 Macroeconomics & Growth
18 Population
19 Poverty
20 Purchasing Power Parity
21 Private Sector
22 Public Sector
23 Rural Development
24 Social Development
25 Trade
26 Urban Development

Appendix 3: Words (or expressions) used as first-level theme in some of the most prominent database classifications for ESPON



Appendix 4: Preliminary thematic structure for the ESPON 2013 DB.

Note: Main themes derived from our experiment with GAP. As a merely indicative example, we used sub-themes from the ESPON 2006 Database (ESPON, 2005).

01. Agriculture & Fisheries

01.01 Land Use

01.02 Farmer Structure

01.03 Employment

01.04 Livestock

01.05 Production

02. Demography (including Household, Population, …)

02.01 Population Structure

02.01 Population Movement

03. Transport (including Accessibility, Communication , Infrastructure, …)

03.01 Transport Infrastructure

03.02 Passengers and Goods Transport

03.03 Accessibility

03.04 Impacts of Transport Policies

04. Energy & Environment (including Climate, Consumption, Hazards, Pollution, Resources, …)

04.01 Natural Hazards

*04.02 Environmental quality**

05. Land Use (including Land Cover, …)

05.01 Land Use

06. Social Affairs *(including Culture, Education, Health, Literacy, …)*

06.01 Education

06.02 Poverty

07. Economy (including Employment, Finance, Industry, Labour, Technology, Trade, Tourism, R&D …)

07.01 Employment

07.02 Unemployment

07.03 Income and Consumption

07.04 Finances and Expenditures

07.05 Tourism

99. Non-/Cross-Thematic Data

*99.01 Integrative indices, typologies and scenarios**

*99.99 Geographical objects**

* Sub-theme not present in ESPON DB 2006 structure

Appendix 5: Details of levels of measurement

	Steven's scales (1946)	Forrest's extended levels (1999)	Required information and nature of data	Examples	Abbreviation
Qualitative	**Nom**inal				**nom**
		unique	all different (no duplication)	country names, NUTS identificators	**nou**
		dichotomous	membership (presence/absence)	coastal areas, regions benefitting from convergence	**nod**
		categorical	categories	land use type, main religion	**noc**
		graded membership	categories plus degree of membership	soil type with percentage conformance	**nog**
Ranking but no quantity	**Ord**inal				**ord**
		complete ordinal	**u**nique ordering	any ranking of regions or cities without ex-aequo	**oru**
		classed ordinal	categories plus ordering	densely/intermediate/weakly populated regions	**orc**
Quantitative	**Int**erval		measure plus arbitrary zero point	temperatures in degrees	**int**
	Rat**io**				**rto**
		extensive ratio	measure (additive rules apply)	GDP, CO2 emissions, acces time, transported tons	**rte**
		count	measure (with unit =1, i.e. not half a person)	population, number of births, firms, migration volumes	**rtc**
		derived ratio	measure (quantity divided by quantity)	GDP per inhabitant, labour productivity, cars per household	**rtd**
		density ratio	measure (quantity divided by area)	population density, firms density	**rde**
		c**y**clic ratio	measure plus length of cycle	angles, slopes orientation	**rty**
		constrained ratio	**p**robability or **p**roportion, range [0,1]	unemployment rate, share of youngs,	**rtp**

Adapted: Forrest (1999)

Appendix 6: Applying TtOYS code on indicators delivered by the current ESPON 2013 projects.

Note: (*) 2003 amendment version; (**) 2006 amendment version.

Indicator	Theme	Sub-theme	Year(s)	Geographical Object	Original Code	T (2)	t (2)	O (6-8)	Y (2-4)	S (2)	TtOYS Code
Total population	02.Demography	01.Population Structure	1995	NUTS2/3**	Pop_t_1995	0 2	0 1	POPtrtc _ _	9 5	N 2	0201POPtrtc__95N2
Total population	02.Demography	01.Population Structure	1999	NUTS2/3**	pop_t_1999	0 2	0 1	POPtrtc _ _	9 9	N 2	0201POPtrtc__99N2
Total population change	02.Demography	02.Population Movement	1995-1999	NUTS2/3**	Pop_ch_95_99	0 2	0 2	POPtrtc _	9 5	N 2	0202POPtrtc_9599N2
Natural population change	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	Nat_ch_96_99	0 2	0 2	NATtrtc _	9 6	N 2	0202NATtrtc_9699N2
Migratory population change	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	Mig_ch_96_99	0 2	0 2	MIGtrtc _	9 6	N 2	0202MIGtrtc_9699N2
Core Indicator 1: Annual growth rate	02.Demography	02.Population Movement	1995-1999	NUTS2/3**	CI1_TOT_95_99	0 2	0 2	GWTrtp_ _	9 5	N 2	0202GWTrtp__9599N2
Core Indicator 2: Annual net migration development	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	CI2_MIG_96_99	0 2	0 2	MIGnet_ _	9 6	N 2	0202MIGnet__9699N2
Core Indicator 3: Annual natural population development	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	CI3_NAT_96_99	0 2	0 2	NATpop_ _	9 6	N 2	0202NATpop__9699N2
Core Indicator 4: Population development by components	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	CI4_TYPO_96_99	0 2	0 2	POPdev_ _	9 6	N 2	0202POPdev__9600N2
Total population	02.Demography	01.Population Structure	2000	NUTS2/3**	Pop_t_2000	0 2	0 1	POPtrtc _ _	0 0	N 2	0201POPtrtc__00N2
Total population	02.Demography	01.Population Structure	2000	NUTS2/3**	pop_t_2000	0 2	0 1	POPtrtc _ _	0 0	N 2	0201POPtrtc__00N2
Sum of births	02.Demography	01.Population Structure	2001-2005	NUTS2/3**	Birth_01_05	0 2	0 1	BTHtrtc _	0 1	N 2	0201BTHtrtc_0105N2
Sum if death	02.Demography	01.Population Structure	2001-2005	NUTS2/3**	Death_01_05	0 2	0 1	DTHtrtc _	0 1	N 2	0201DTHtrtc_0105N2
Total population change	02.Demography	02.Population Movement	2001-2005	NUTS2/3**	Pop_ch_00_05	0 2	0 2	POPtrtc _	0 1	N 2	0202POPtrtc_0105N2
Natural population change	02.Demography	02.Population Movement	2001-2005	NUTS2/3**	Nat_ch_01_05	0 2	0 2	NATtrtc _	0 1	N 2	0202NATtrtc_0105N2
Migratory population change	02.Demography	02.Population Movement	2001-2005	NUTS2/3**	Mig_ch_01_05	0 2	0 2	MIGtrtc _	0 1	N 2	0202MIGtrtc_0105N2
Core Indicator 1: Annual growth rate	02.Demography	02.Population Movement	2000-2005	NUTS2/3**	CI1_TOT_00_05	0 2	0 2	GWTrtp_ _	0 0	N 2	0202GWTrtp__0005N2
Core Indicator 2: Annual net migration development	02.Demography	02.Population Movement	2001-2005	NUTS2/3**	CI2_MIG_01_05	0 2	0 2	MIGnet_ _	0 1	N 2	0202MIGnet__0105N2
Core Indicator 3: Annual natural population development	02.Demography	02.Population Movement	2001-2005	NUTS2/3**	CI3_NAT_01_05	0 2	0 2	NATpop_ _	0 1	N 2	0202NATpop__0105N2
Core Indicator 4: Population development by components	02.Demography	02.Population Movement	2001-2005	NUTS2/3**	CI4_TYPO_01_05	0 2	0 2	POPdev_ _	0 1	N 3	0202POPdev__0105N3
Multimodal potential accessibility, absolute level	03.Transport	03.Accessibility	2001,2006	NUTS3**	MM	0 3	0 3	ACCmmab s	_ 6	N 3	0303ACCmmabs__06N3
Multimodal potential accessibility, standardised	03.Transport	03.Accessibility	2001,2006	NUTS3**	MM_i	0 3	0 3	ACCmmst d	_ 6	N 3	0303ACCmmstd__06N3
Multimodal potential accessibility, change of standardised	03.Transport	03.Accessibility	2001-2006	NUTS3**	MM_i_ch	0 3	0 3	ACCmmst d	0 1	N 3	0303ACCmmstd0106N3
Multimodal potential accessibility, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	MM_r	0 3	0 3	ACCmmre l	0 1	N 3	0303ACCmmrel0106N3
Multimodal potential accessibility, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	MM_a	0 3	0 3	ACCmmab s	0 1	N 3	0303ACCmmabs0106N3
Potential accessibility by air, absolute level	03.Transport	03.Accessibility	2001,2006	NUTS3**	Air	0 3	0 3	ACCaiab s	_ 6	N 3	0303ACCaiabs__06N3
Potential accessibility by air, standardised	03.Transport	03.Accessibility	2001,2006	NUTS3**	Air_i	0 3	0 3	ACCaist d	_ 6	N 3	0303ACCaistd__06N3
Potential accessibility by air, change of standardised	03.Transport	03.Accessibility	2001-2006	NUTS3**	Air_i_ch	0 3	0 3	ACCaist d	0 1	N 3	0303ACCaistd0106N3
Potential accessibility by air, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	Air_r	0 3	0 3	ACCaire l	0 1	N 3	0303ACCairel0106N3
Potential accessibility by air, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	Air_a	0 3	0 3	ACCaiab s	0 1	N 3	0303ACCaiabs0106N3
Total Population	02.Demography	01.Population Structure	2001,2006	NUTS3**	Pop_t	0 2	0 1	POPtrtc _ _	0 6	N 3	0201POPtrtc__06N3
Potential accessibility road, standardised	03.Transport	03.Accessibility	2006	NUTS3**	ROAD_Index	0 3	0 3	ACCrdst d	_ 6	N 3	0303ACCrdstd__06N3
Potential accessibility road, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	ROAD_Relative change	0 3	0 3	ACCrdst d	0 1	N 3	0303ACCrdstd0106N3
Potential accessibility road, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	ROAD_Absolut echange	0 3	0 3	ACCrdab s	0 1	N 3	0303ACCrdabs0106N3
Potential accessibility road, index change	03.Transport	03.Accessibility	2001-2006	NUTS3**	ROAD_Indexch ange	0 3	0 3	ACCrdix _	0 1	N 3	0303ACCrdix_0106N3
Potential accessibility rail, standardised	03.Transport	03.Accessibility	2006	NUTS3**	RAIL_Index	0 3	0 3	ACCrlst d	_ 6	N 3	0303ACCrlstd__06N3
Potential accessibility rail, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	RAIL_Relativec hange	0 3	0 3	ACCrlre l	0 1	N 3	0303ACCrlrel0106N3
Potential accessibility rail, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	RAIL_Absolute change	0 3	0 3	ACCrlab s	0 1	N 3	0303ACCrlabs0106N3
Potential accessibility rail, index change	03.Transport	03.Accessibility	2001-2006	NUTS3**	RAIL_Indexcha nge	0 3	0 3	ACCrlix _	0 1	N 2	0303ACCrlix_0106N2
Total population	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP	0 2	0 1	POPtrtc _ _	0 0	N 2	0201POPtrtc__00N0
Population aged 20-39 years	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP2039	0 2	0 1	POP2039 t	_ 0	N 0	0201POP2039t__00N0
Population aged 20-64 years	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP2064	0 2	0 1	POP2064 t	_ 0	N 0	0201POP2064t__00N0
Population aged 65 years and over	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP65+	0 2	0 1	POP65t_ _	_ 0	N 0	0201POP65t__00N0
Population aged 75 years and over	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP75+	0 2	0 1	POP75t_ _	_ 0	N 0	0201POP75t__00N0
Annual average population change	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	POP_ch	0 2	0 2	POPartc _	_ 0	N 0	0202POPartc__00N0
Annual average population change, 20-39 years	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	POP2039_ch	0 2	0 2	POP2064 ach	0 0	N 0	0202POP2064ach00N0
Annual average population change, 20-64 years	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	POP2064_ch	0 2	0 2	POP2064 ach	0 0	N 0	0202POP2064ach00N0
Annual average population change, 65 years and over	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	POP65+_ch	0 2	0 2	POP65ac h	_ 0	N 0	0202POP65ach__00N0
Annual average population change, 75 years and over	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	POP75+_ch	0 2	0 2	POP75ac h	_ 0	N 0	0202POP75ach__00N0
Share of 20-39 years	02.Demography	01.Population	2005	NUTS0/1/	POP2039_sh	0 0	0 0	POP2039	_ 0	N	0201POP2039s

Indicator	Theme	Sub-theme	Year(s)	Geographical Object	Original Code	T (2)	t (2)	O (6-8)	Y (2-4)	S (2)	TtOYS Code
		Structure		2**		2	1	sh	5	0	h_05N0
Share of population aged 65 years and over	02.Demography	01.Population Structure	2005	NUTS0/1/2**	POP65+_sh	0 2 1	0	POP65sh _	0 5	N 0	0201POP65sh_ __05N0
Average share of population aged 65 years and over	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP65+_ash	0 2 1	0	POP65as _ h	0 7	N 0	0201POP65ash __07N0
Total population	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP	0 2 1	0	POPtrtc _ _	0 7	N 0	0201POPtrtc_ __07N0
Life expectancy at birth	02.Demography	01.Population Structure	2002-2004	NUTS0/1/2**	E0	0 2 1	0	LIFtrtc _ _	0 2	N 0	0201LIFtrtc_ __02N0
Natural population change	02.Demography	02.Population Movement	2000-2006	NUTS0/1/2**	NAT_CH	0 2 2	0	NATtrtc _ _	0 0	N 0	0202NATtrtc_ __00N0
Net migration change	02.Demography	02.Population Movement	2000-2006	NUTS0/1/2**	MIG_CH	0 2 2	0	MIGtrtc 0 _	0 6	N 0	0202MIGtrtc_ 0006N0
Annual average natural population change	02.Demography	02.Population Movement	2000-2006	NUTS0/1/2**	trend_nat	0 2 2	0	NATartc _ _	0 6	N 0	0202NATartc_ __06N0
Annual average net migration rate	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	trend_mig	0 2 2	0	MIGartp _ _	0 7	N 0	0202MIGartp_ __07N0
Annual average population change per 1000 inhabitants	02.Demography	02.Population Movement	2000-2006	NUTS0/1/2**	tot_ch	0 2 2	0	POPartc _ _	0 6	N 0	0202POPartc_ __06N0
Total fertility rate	02.Demography	01.Population Structure	2005	NUTS0/1/2**	TFR	0 2 1	0	FRTtrtp _ _	0 5	N 0	0201FRTtrtp_ __05N0
Internal net migration between the NUTS2 regions	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	NIMIGR	0 2 2	0	MIGintr tc	0 0	N 0	0202MIGintrtc_00N0
Basic typology of the demographic status 2005	02.Demography	01.Population Structure	2005	NUTS0/1/2**	ST_TYPO	0 2 1	0	DEMtyp_ _	0 5	N 0	0201DEMtyp__ __05N0
Age pyramid by 5 years age-group	02.Demography	01.Population Structure	2005	NUTS0/1/2**	A-NS_5	0 2 1	0	DEMpyr_ _	0 5	N 0	0201DEMpyr__ __05N0
Number unemployed persons, total	07.Economy	02.Unemployment	2000-2007	NUTS0/1/2/3**	unemp	0 7 2	0	UMPtrtc _ _	0 0	N 0	0702UMPtrtc_ __00N0
Active population, total	07.Economy	01.Employment	2000-2007	NUTS0/1/2/3**	activ	0 7 1	0	ACTtrtc _ _	0 0	N 0	0701ACTtrtc_ __00N0
Total population, total	02.Demography	01.Population Structure	2000-2006	NUTS0/1/2/3**	pop_t	0 2 1	0	POPtrtc _ _	0 0	N 0	0201POPtrtc_ __00N0
GDP in euros	07.Economy	03.Income and Consumption	2000-2006	NUTS0/1/2/3**	gdp_eur	0 7 3	0	GDPeurr te	0 0	N 0	0703GDPeurrte_00N0
GDP in PPS	07.Economy	03.Income and Consumption	2000-2006	NUTS0/1/2/3**	gdp_pps	0 7 3	0	GDPppsr te	0 0	N 0	0703GDPppsrte_00N0
Productivity of inland transport infrastructure	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_E1_PROD	0 3 1	0	PDTinrt d	_ 0	N 3	0301PDTinrtd __30N3
Productivity of airports	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_E2_Prod_air	0 3 1	0	PDTairt d	_ 0	N 3	0301PDTairtd __30N3
Economic growth (€/inhabitant)	07.Economy	03.Income and Consumption	2005,2030	NUTS3*	PIM_E3_GDP_CAP	0 7 3	0	ECOgwhi x	_ 0	N 3	0703ECOgwhix __30N3
Congestion costs	03.Transport	03.Accessibility	2005,2030	NUTS3*	PIM_E4_Cong_level	0 3 3	0	COScgti x	_ 0	N 3	0303COScgtix __30N3
Traffic freight passing through	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_Q1_Fre_PT	0 3 1	0	TRFfghi x	_ 0	N 3	0301TRFfghix __30N3
CO2 emissions by road traffic	04.Energy & Environment	02.Environmental Quality	2005,2030	NUTS3*	PIM_Q2_CO2/km2	0 4 2	0	CO2rdix _	_ 0	N 3	0402CO2rdix_ __30N3
Safety of roads	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_Q3_Traffic_seg	0 3 1	0	SFTrdix _	_ 0	N 3	0301SFTrdix_ __30N3
Market opportunities	07.Economy	03.Income and Consumption	2005,2030	NUTS3*	PIM_Q4_GDP_3h	0 7 3	0	MKToppi x	_ 0	N 3	0703MKToppix __30N3
Landscape fragmentation	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_I1_HCI_dens	0 3 1	0	LANfrgi x	_ 0	N 3	0301LANfrgix __30N3
Exposure to external visitors	03.Transport	02.Passangers and G. Transport	2005,2030	NUTS3*	PIM_I2_Ext_no_3h	0 3 2	0	EXTvsti x	_ 0	N 3	0302EXTvstix __30N3
Regional integration	03.Transport	03.Accessibility	2005,2030	NUTS3*	PIM_I3_Road_N2_n	0 3 3	0	REGinti x	_ 0	N 3	0303REGintix __30N3
Economic growth (Modulation/Total GDP)	07.Economy	03.Income and Consumption	2000-2002	NUTS2**	PIM_E1_DEF	0 7 3	0	GWHecoi x	_ 2	N 2	0703GWHecoix __02N2
Unemployment	07.Economy	02.Unemployment	2004	NUTS2**		0 7 2	0	UMPtrtc _ _	0 4	N 2	0702UMPtrtc_ __04N2
Tourism diversification	07.Economy	05.Tourism	2004	NUTS2**	PIM_E3_DEF	0 7 5	0	TOUtrtp _ _	0 4	N 2	0705TOUtrtp_ __04N2
Environmental quality	04.Energy & Environment	02.Environmental Quality	/	NUTS2**	PIM_Q1_DEF	0 4 2	0	ENVquai x	_ _	N 2	0402ENVquaix ____N2
Community viability	99.Non-/Cross-thematic data	01.Integrative Indices, Typologies	/	NUTS2**	PIM_Q2_DEF	0 9 1	0	VIAcomi x	_ _	N 2	9901VIAcomix ____N2
CO2 emissions	04.Energy & Environment	02.Environmental Quality	/	NUTS2**		0 4 2	0	CO2rte_ _	_ _	N 2	0402CO2rte__ ____N2
Risk of soil erosion	04.Energy & Environment	01.Natural hazards	2004	NUTS2**	PIM_Q4_DEF	0 4 1	0	ERОrski x	_ 4	N 2	0401ERОrskix __04N2
Landscape diversity	04.Energy & Environment	02.Environmental Quality	/	NUTS2**	PIM_I1	0 4 2	0	LANdivi x	_ _	N 2	0402LANdivix ____N2
Community identity	99.Non-/Cross-thematic data	01.Integrative Indices, Typologies	/	NUTS2**	PIM_I2	0 9 1	0	IDTcomi x	_ _	N 2	9901IDTcomix ____N2
Heritage products	99.Non-/Cross-thematic data	01.Integrative Indices, Typologies	/	NUTS2**	PIM_I3_DEF	0 9 1	0	HRTprdi x	_ _	N 2	9901HRTprdix ____N2
Land use	05.Land Use	01.Land Use	1978-2008	LAU2**	A-NS_7	0 5 1	0	LUStnoc _ _	0 8	L 2	0501LUStnoc_ __08L2
Number of farm holders by age (24-75+)	07.Economy	02.Unemployment	2003-2007	LAU2**		0 1 2	0	FRM2475 t	_ 7	L 2	0102FRM2475t __07L2
Number of farm holdings	07.Economy	02.Unemployment	1991-2007	LAU2**	A-NS_1	0 1 2	0	FRMtrtc _ _	0 7	L 2	0102FRMtrtc_ __07L2
Number of persons working in the agricultural sector	07.Economy	01.Employment	2003-2007	LAU2**	A_NS_3a	0 7 1	0	AGRtrtc _ _	0 7	L 2	0701AGRtrtc_ __07L2
Number of persons working in forestry and logging	07.Economy	01.Employment	2003-2007	LAU2**	A-NS_3b	0 7 1	0	FRTtrtc _ _	0 7	L 2	0701FRTtrtc_ __07L2
Number of persons working in fishing and aquaculture sector	07.Economy	01.Employment	2003-2007	LAU2**	A-NS_3c	0 7 1	0	FIStrtc _ _	0 7	L 2	0701FIStrtc_ __07L2
Number of passengers at airport	03.Transport	02.Passangers and G. Transport	2006-2007	LAU2**	I-NS_2	0 3 2	0	PSGairr te	0 7	L 2	0302PSGairrte_07L2
Freights handled by airports	03.Transport	01.Transport Infrastructure	2006-2008	LAU2**	I-NS_3	0 3 1	0	FRGairr te	0 8	L 2	0301FRGairrte_08L2
Number of passengers at maritime ports	03.Transport	02.Passangers and G. Transport	2006-2009	LAU2**	I-NS_4	0 3 2	0	PSGmrtr te	0 9	L 2	0302PSGmrtrte_09L2
Freights handled by maritime ports	03.Transport	01.Transport Infrastructure	2006-2010	LAU2**	I-NS_3a	0 3 1	0	FRGmrtr te	1 0	L 2	0301FRGmrtrte_10L2
Total population	02.Demography	01.Population Structure	1981-2007	LAU2**	D-NS_1a	0 2 1	0	POPtrtc _ _	0 7	L 2	0201POPtrtc_ __07L2
Total population, males	02.Demography	01.Population Structure	1981-2008	LAU2**	D-NS-1c	0 2 1	0	POPmrtc _ _	0 8	L 2	0201POPmrtc_ __08L2
Total population, females	02.Demography	01.Population Structure	1981-2009	LAU2**	D-NS_1b	0 2 1	0	POPfrtc _ _	0 9	L 2	0201POPfrtc_ __09L2
Population by age group	02.Demography	01.Population Structure	1990-2007	LAU2**	D-NS_2	0 2 1	0	POPaget _ _	0 9	L 2	0201POPaget_ __90L2
Number of live births per year	02.Demography	01.Population Structure	1981-2008	LAU2**	D-NS_3_x	0 2 1	0	BTHtrtc _ _	0 8	L 2	0201BTHtrtc_ __08L2
Number of deaths per year	02.Demography	01.Population Structure	1981-2009	LAU2**	D-NS_3_y	0 2 1	0	DTHtrtc _ _	0 9	L 2	0201DTHtrtc_ __09L2
Number of out migrants	02.Demography	02.Population Movement	1981-2010	LAU2**	D-NS_4_a	0 2 2	0	MIGotrt c	_ 1	L 2	0202MIGotrtc __10L2
Number of in migrants	02.Demography	02.Population Movement	1981-2011	LAU2**	D-NS_4_b	0 2 2	0	MIGitrt c	_ 1	L 2	0202MIGitrtc __11L2

Indicator	Theme	Sub-theme	Year(s)	Geographical Object	Original Code	T (2)	t (2)	O (6-8)	Y (2-4)	S (2)	TtOYS Code
Net migration	02.Demography	02.Population Movement	2000-2007	LAU2**	D-NS_4_c	0 2	0 2	MIGntrt c	_ _	0 7	0202MIGntrtc 07L2
Number of persons born abroad	02.Demography	01.Population Structure	2000-2007	LAU2**	D-NS_5	0 2	0 1	POPbrnr tc	_	0 7	0201POPbrnrt c_07L2
Total number of unemployed persons	07.Economy	02.Unemployment	2007	LAU2	E-NS_4a	0 7	0 2	UMPtrtc _	_	0 7	0702UMPtrtc_ 07L2
Number of unemployed persons, female	07.Economy	02.Unemployment	2007	LAU2	E-NS_4b	0 7	0 2	UMPfrtc _	_	0 7	0702UMPfrtc_ 07L2
Number of unemployed persons, male	07.Economy	02.Unemployment	2007	LAU2	E-NS_4c	0 7	0 2	UMPmrtc _	_	0 7	0702UMPmrtc_ 07L2
Active population, total	07.Economy	01.Employment	2007	LAU2	E-NS_1a	0 7	0 1	ACTtrtc _	_	0 7	0701ACTtrtc_ 07L2
Active population, males	07.Economy	01.Employment	2007	LAU2	E-NS_1c	0 7	0 1	ACTmrtc _	_	0 7	0701ACTmrtc_ 07L2
Active population, females	07.Economy	01.Employment	2007	LAU2	E-NS_1b	0 7	0 1	ACTfrtc _	_	0 7	0701ACTfrtc_ 07L2
Total persons working in agriculture, hunting	07.Economy	01.Employment	2005	LAU2	E-NS_2a	0 7	0 1	AGRtrtc _	_	0 5	0701AGRtrtc_ 05L2
Total persons working in fishing	07.Economy	01.Employment	2005	LAU2	E-NS_2b	0 7	0 1	FIStrtc _	_	0 5	0701FIStrtc_ 05L2
Total persons working in mining and quarrying	07.Economy	01.Employment	2005	LAU2	E-NS_2c	0 7	0 1	MINtrtc _	_	0 5	0701MINtrtc_ 05L2
Total persons working in manufacturing	07.Economy	01.Employment	2005	LAU2	E-NS_2d	0 7	0 1	MANtrtc _	_	0 5	0701MANtrtc_ 05L2
Total persons working in electricity, gas and water supply	07.Economy	01.Employment	2005	LAU2	E-NS_2e	0 7	0 1	GAStrtc _	_	0 5	0701GAStrtc_ 05L2
Total persons working in construction	07.Economy	01.Employment	2005	LAU2	E-NS_2f	0 7	0 1	CONtrtc _	_	0 5	0701CONtrtc_ 05L2
Total persons working in wholesale and retail	07.Economy	01.Employment	2005	LAU2	E-NS_2g	0 7	0 1	RETtrtc _	_	0 5	0701RETtrtc_ 05L2
Total persons working in hotels and restaurants	07.Economy	01.Employment	2005	LAU2	E-NS_2h	0 7	0 1	HOTtrtc _	_	0 5	0701HOTtrtc_ 05L2
Total persons working in transport, storage	07.Economy	01.Employment	2005	LAU2	E-NS_2i	0 7	0 1	TRAtrtc _	_	0 5	0701TRAtrtc_ 05L2
Total persons working in financial intermediation	07.Economy	01.Employment	2005	LAU2	E-NS_2j	0 7	0 1	FINtrtc _	_	0 5	0701FINtrtc_ 05L2
Total persons working in real estate, renting and business	07.Economy	01.Employment	2005	LAU2	E-NS_2k	0 7	0 1	REStrtc _	_	0 5	0701REStrtc_ 05L2
Total persons working in public administration and defence	07.Economy	01.Employment	2005	LAU2	E-NS_2l	0 7	0 1	PADtrtc _	_	0 5	0701PADtrtc_ 05L2
Total persons working in education	07.Economy	01.Employment	2005	LAU2	E-NS_2m	0 7	0 1	EDUtrtc _	_	0 5	0701EDUtrtc_ 05L2
Total persons working in health and social work	07.Economy	01.Employment	2005	LAU2	E-NS_2n	0 7	0 1	HEAtrtc _	_	0 5	0701HEAtrtc_ 05L2
Total persons working in other community activities	07.Economy	01.Employment	2005	LAU2	E-NS_2o	0 7	0 1	COMtrtc _	_	0 5	0701COMtrtc_ 05L2
Total persons working in activities of households	07.Economy	01.Employment	2005	LAU2	E-NS_2p	0 7	0 1	HOUtrtc _	_	0 5	0701HOUtrtc_ 05L2
Total persons working in extra-territorial organizations	07.Economy	01.Employment	2005	LAU2	E-NS_2q	0 7	0 1	ORGtrtc _	_	0 5	0701ORGtrtc_ 05L2
Number of unemployed persons by age	07.Economy	02.Unemployment	2007	LAU2	E-NS_5	0 7	0 2	UMPtrtc _	_	0 7	0702UMPtrtc_ 07L2
Long-term unemployment	07.Economy	02.Unemployment	2007	LAU2	I-NS_1	0 7	0 2	UMPlngr tc	_	0 7	0702UMPlngrt c_07L2
Part-time unemployment	07.Economy	02.Unemployment	2007	LAU2	E-NS_7	0 7	0 2	UMPprtr tc	_	0 7	0702UMPprtrt c_07L2
Number of employees by size of the company	07.Economy	01.Employment	2007	LAU2	E-NS_8	0 7	0 1	EMPcpnr tc	_	0 7	0701EMPcpnrt c_07L2
Number of persons with secondary education degree	06.Social Affairs	01.Education	2007	LAU2	E-NS_10a	0 6	0 1	EDUscdr tc	_	0 7	0601EDUscdrt c_07L2
Number of persons with tertiary education degree	06.Social Affairs	01.Education	2007	LAU2	E-NS_10b	0 6	0 1	EDUtrtr tc	_	0 7	0601EDUtrtrt c_07L2
Number of students of higher education institutions	06.Social Affairs	01.Education	2007	LAU2	E-NS_11	0 6	0 1	EDUhghr tc	_	0 7	0601EDUhghrt c_07L2
Number of companies created	07.Economy	01.Employment	2007	LAU2	E-NS_12	0 7	0 1	FIRopnr tc	_	0 7	0701FIRopnrt c_07L2
Number of companies closed	07.Economy	01.Employment	2007	LAU2	E-NS_13	0 7	0 1	FIRclsr tc	_	0 7	0701FIRclsrt c_07L2

TECHNICAL REPORT

ESPON

Text mining methods and visualization tools as means to support the thematic structuring of the ESPON 2013 DB

MAIN RESULTS

- Text mining methods are employed to derive sub-themes
- ESPON evidence is used as textual data to extract potential words
- Two distinct parameters of data co-occurrence are employed to understand relational similarities
- We evaluate the explanatory power of words with an algorithm that weights the importance of each word in large corpora of textual data
- Mapping techniques of MDS are used to depict similarities and ease interpretation
- Our results suggest the ESPON 2013 DB should be structured in 7+1 themes and 29 sub-themes

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Table of contents

Introduction

In our last technical report we assumed that database structures adopted by international organisations constitute an important source of information to extract first-level themes. For this purpose, we applied a visual grouping technique to illustrate, by means of correlation matrices, homogenous clusters of words. This approach suggested that the ESPON 2013 DB should be structured in 7+1 themes meaning that we would also add a theme to cover both cross-thematic and non-thematic data.

The rationale defined for sub-themes is slightly different. In this case, we use a large collection of textual data to extract potential keywords to label sub-themes. More concretely, this means that for each of the seven themes that emerged from our experiment with database structures we employ text mining tools.

The goal of text mining is to find patterns across textual data and, therefore, derive new information. Such methods enable users to identify keywords that, inductively, create thematic overviews of text collections. Against this background, we argue that text mining methods may positively support the thematic structuring of the ESPON 2013 DB. It is accepted that ESPON introduced new vocabulary of spatial concepts which somehow influenced the terminology adopted by EU institutions. We make use of this evidence to extract keywords from qualitative and unstructured data, in particular ESPON reports and texts delivered by EU institutions that use or make reference to ESPON results. This approach is further enriched by applying mapping techniques of multidimensional scaling.

Methods

The goal of our investigation is to identify keywords on textual data according to their co-occurrence and use that information to conveniently structure the ESPON 2013 DB. As explained in the introductory part of this report, our contribution will only focus on keywords that can be used as sub-themes. For this purpose, we will employ the findings of the previous technical report where we propose a list of 7+1 first-level themes. Primarily, it is important to identify documents that potentially address each of the themes proposed in the last report and, secondly, ensure that we integrate ESPON reports with evidence-based knowledge on European territorial potentials and dynamics.

The most challenging task before applying text mining tools is data preparation (Berry, 2004; Weiss et al. 2005). Due to the fact that textual data is unstructured and often arranged inconveniently it is necessary to follow certain procedures to ensure some consistency to the overall process. The first step is obviously to collect data. In our case this represents any document, study or policy note addressing ESPON evidence and results. For this purpose, we have initially identified 27 final project reports delivered by the ESPON 2006 Programme. Our desk research expanded then to reports delivered by the current ESPON Programme and documents published by other sources that offer a wide range of perspectives to ESPON knowledge (e.g. European Parliament, European Commission). In total, we have collected 53 documents (see Annexes 1-7). Altogether, these documents constitute a large textual database that needs to be structured as efficiently as possible before applying any methodological approach.

Similarly, we have to bear in mind that textual data is a complex conjunction of words and phrases that frequently need to be considered as a whole. There is a quite huge amount of dependency that should not be ignored. Moreover, it is also important to overcome word and semantic ambiguities that may adversely influence our analysis. To this end, the usability offered by WordStat is quite straightforward and no additional expertise is needed (Lewis, 1999; Davi et al., 2005). The pre-processing of our text collection took into account some of the features offered by this text mining module, particularly with regard to stop-word lists and lower case conversion.

One of the most interesting features provided by this software is the compilation of non information-bearing words that basically exclude terms without any predictive capability, such as articles, pronouns or prepositions. These words are often characterised as noise

data and hardly add new information. Besides, it is also possible to add more words to this dictionary of stop-words and improve the accuracy of the corpus for analysis. It should also be mentioned that WordStat merely records the number of times a word appears within a text regardless the content of a sentence or paragraph. After computing data there is a wide variety of ways in which the result can be displayed. The most basic output offered by this application is the word frequency distribution. This knowledge will constitute the basis to explain our results.

As a first step, we apply a pre-defined list on non-information bearing words with no semantic value. Next, we make use of word lemmatization to reduce inflectional form of words to a common root word and ultimately exclude words based upon a frequency criterion. With this regard, we suggest that words below 100 occurrences should be ignored from our analysis. This option exposed a significant number of words that allowed us to further analyse the knowledge structure in text collections.

However, several authors state that words with high frequency distribution do not offer a robust and solid basis for analysis. With this regard, Luhn (cited by Blanchard, 2007: 309) says that 'mid-frequency words in the distribution have both high significance and high discriminating power'. This means that words above an upper cut-off and below a lower cut-off should be removed and, as a consequence, more effort should be added to those that have a mid-frequency (see Figure 1). We took into consideration these aspects and defined a threshold from the estimated densities in order to make emerge other words with explanatory power. The threshold for words with high explanatory power is based on the difference between term frequencies. In other words, the highest gap determines the cut-off.

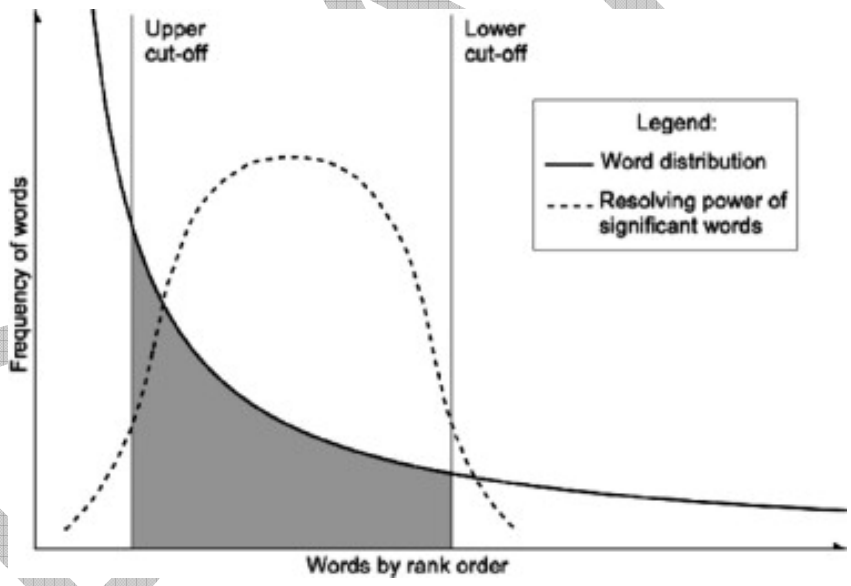


Figure 1: Illustrative plot of word distribution. The power of significant words in higher between the two cut-offs (in grey). Source: Luhn (cited in Blanchard, 2007: 310).

In order to further evaluate the explanatory power of terms in large corpora of textual data, we decided to use TF*IDF (i.e. Term Frequency * Inverse Document Frequency). Basically, this scoring algorithm is a weight that measures the importance of each term in a collection of documents. As a result, the importance increases proportionally to the number of times a term appears in a specific document but is counterbalanced by the frequency of the term in the same corpus or text collection. To a certain extent, TF*IDF algorithm computes the relevance of a document with respect to a particular term. Here we limited our analysis to the 100 most significant terms.

In addition, other features have been applied. Our initial assumption also questioned the ability to use such tools to determine co-occurrence analysis on data derived from a large textual corpus. This function is available on WordStat and it can be used in a variety of ways. For this experiment we have considered two distinct parameters of data co-occurrence – paragraph and group of 5 words. The output file is square matrix where the rows and columns refer to the words and the value of each cell corresponds to the number of co-occurrences. However, it is important to point out that the methods described in this part of the report are solely applied for each of the seven themes that came out from our analysis with database structures. That is, we selected a set of documents for each theme and, based on these documents, prepared data for empirical analysis.

Ultimately, we employ mapping techniques to depict similarities. According to Yang et al (2008), one of the crucial needs in text mining is the ability to visualize relationships between words extracted from large text collections. In order to simply the analysis of data co-occurrence for words with both explanatory power and non-explanatory power we apply a mapping technique of multidimensional scaling to construct and view maps of keywords – VOSviewer (Eck & Waltman, 2007; Eck & Waltman, 2008; Eck et al, 2010). The comparison of these two approaches is also subject to discussion.

Results

In this section we describe the results obtained for the seven themes that came out from our analysis with database structures. In addition, we review the rationale defined for the theme on both cross-thematic and non-thematic data.

For each theme and sub-theme we provide an ordinal code of, respectively, two and four digits to simplify membership. In addition, we coded the last theme with the number 99 so that more themes can be added to the ESPON 2013 DB. Apart from the theme on cross-thematic and non-thematic data, the interpretation of the other themes solely refers to the annexes present in this report.

Agriculture & Fisheries

The maps illustrated in Figure 1a and 1b correspond to the density view of data co-occurrence in a multidimensional scaling. For data on ‘Agriculture & Fisheries’, we selected 7 texts that combine ESPON evidence and results on issues related to the above mentioned topic (see Annex 1).

The comparison of these maps suggests that words in Figure 1a are less concentrated that Figure 1b. Somehow this can be explained by the dimension of each parameter for data co-occurrence. Despite the relevance of this aspect what is interesting here is that, in both cases, several densities can be depicted. The main densities aggregate more or less the same number of words and therefore similarities are very alike. This is quite evident for inflectional variants of words, such as ‘territorial’, ‘develop’, and ‘impact’, or ‘structur’, ‘farm’, and ‘agricultur’.

As stated in the methodology part of this report, the threshold for words with high explanatory power is based on the difference between word frequencies. As it can be seen in Figure 1b, such approach increases the importance of other words, such as ‘environment’, ‘farm’ and ‘product’. Despite the significance of this adjustment to identify an explanatory power on words with mid-frequency distribution, the exclusion of anchor words produces and effect that, somehow, suggests the concentration of many words in just one individual density. This is particularly clear on both parameters.

The ambiguity of the results observed in the previous maps is now reduced to clear and visible densities of words. In addition, it also indicates that terms with high frequency distribution may also discloser similar TF*IDF importance. That seems to be the case for

'territori' and 'impact'[1]. If we look to the results in more detail, it also becomes evident that rural development and the Common Agricultural Policy (CAP) receive similar attention.

The number of sub-themes as well as the labelling to describe the content of each sub-theme is therefore based on textual data co-occurrence by means of visualisation techniques in VOSviewer. The sub-themes related to 'Agricultural & Fisheries' looks as follows:

01 AGRICULTURE & FISHERIES

0101 Farm structure (e.g. farm type, income from farming, organic farming)

0102 Livestock (e.g. livestock output)

0103 Aquaculture and sea fisheries (e.g. aquaculture resources in coastal areas)

0104 Forestry (e.g. production, consumption, and import/export products)

0105 Rural development (e.g. rural employment, rural access to services)

The first sub-theme is designed to incorporate data on farm type, size of farms, income from farming and other related components, such as organic farming. The decision to set out this sub-theme derives from the relatedness between 'farm', 'agricultur', and 'structur' (see Figure 1a and 1b). The third sub-theme is meant to integrate data on 'aquaculture and sea fisheries' and its presence is justified by the high TF*IDF distribution observed for the word 'aquacultur'. The same applies for 'forestry' and 'rural development'.

Demography

The capacity of visualizing and analysing data co-occurrence matrices with VOSviewer reduces the complexity inherent to the content of texts. The collection of documents in the filed of 'Demography' (see Annex 2) indicates a strong emphasis on subjects associated with the following root words: 'popul', 'migrat', and 'labour'. The co-occurrence of such words, regardless the parameter used in this analysis, is quite evident on Figure 2a. Hence, one could assume that the presence of such words is highly important for analysis and eventually contribute to the definition of sub-themes. This assumption is not completely accurate and, as explained previously, some adjustments should be added by withdrawing these words from the corpus and therefore identify other potential words with explanatory power. If we look to both maps in Figure 2b, we recognize inflectional forms of specific words strongly related to the theme on 'demography'. This is the case for 'fertile', 'develop', 'scenario', 'trend', and 'rural'.

But what would happen to our dataset if we just consider words that emerge frequently in a document but do not have the same frequency in the remainder of the corpus? In other words, what would be the results of our analysis if we just keep a maximum of 100 words based on TF*IDF distribution? With this regard, some of the major words identified previously can still be seen in Figures 2c and 2d. Despite the usefulness of this approach, Figure 2c does not improve the distribution observed in Figure 2d. By excluding words that demonstrate high TF*IDF distribution, the density view in both maps is rather similar. Somehow this feature suggests that our corpus is not sensitive to such boundaries and words with high frequency distribution tend to be distributed more homogenously across documents (e.g. 'labour', immigrate', 'depopula').

[1] Of course, documents chosen to integrate the corpus for analysis might have influenced the results. But, in any case, this can be explained by the strong emphasis given to ESPON projects on tools and methodologies to evaluate the effects of agricultural policies on the EU territory.

Having said that, it is important to interpret the results and derive sub-themes related to this specific theme. Some of the terms are implicitly associated with migration (e.g. immigration, migration replacement) or natural change in population (i.e. fertility, mortality, life expectancy) while others demonstrate close links to demographic ratios, trends, and scenarios in rural, urban and peripheral areas. Surprisingly enough we observe a constant correlation between 'labour' and 'forc'. The fact that these two inflectional forms of words disclose a strong correlation is normal, what is not normal is to see these words isolated from other words equally relevant, such as 'ageing' and 'decline'. The topic itself has been on the EU agenda for some years now and many policy measures have been implemented to address the issue of ageing labour force and ageing in general.

Despite the feasibility of text mining techniques in analysing documents related to specific themes or subjects, we would expect to find other words with high explanatory power. This is the case for 'household', 'health', or 'dependency'. Nevertheless, we believe the information collected for analysis is solid and robust to extract sub-themes. The structure of second-level themes related to 'Demography' looks as follows:

02 DEMOGRAPHY

0201 Population structure (e.g. age distribution by group and gender)

0202 Natural changes (e.g. fertility, mortality, life expectancy)

0203 Households (e.g. number of households)

0204 Migrations (e.g. migration replacement, high-skilled labour migration)

0205 Territorial impacts (e.g. population by territorial typologies)

The analysis performed on documents related to 'demography' sets out five distinct sub-themes. These sub-themes try to cover the most significant topics present in the corpus for our initial experiments. It is worth mentioning, however, that this proposal only takes into consideration textual data delivered up to now on ESPON evidence and results. In the future, more documents should be added to the corpus in the field of 'Demography' to eventually extract new terms and label other potential sub-themes.

Transport

Transport-related issues have always represented a major issue for analysis and discussion within the ESPON Programme. Our corpus features 6 documents that explore subjects related to the transport sector in Europe (see Annex 3). The major topics under discussion stem from the need to monitor the EU transport policies, such as mobility, accessibility, sustainable transport policies, impacts on the environment, competitiveness of the economy, or leverage effects on EU territorial development. With this regard, a series of applied research projects commissioned by the ESPON 2006 Programme dedicated a large attention to the Trans-European Transport Network (TEN-T) Initiative. Its importance for competitiveness and growth has been recognized as one of the most fundamental initiatives to set out a proper EU transport policy. The findings of these studies enriched many policy discussions on the developments of the TEN-T with new data on transport networks and traffic flows, including performance indicators, typologies, and scenario-based projections.

The mapping perspective of data co-occurrence in Figure 3a varies according to the parameter. To a large extent, it is correct to say that we observe the same densities. For instance, 'transport', 'access', 'impact', or 'network' confirm their importance. But when we exclude these terms from our dataset the result conveys new densities or clusters of words. This is the case for 'infrastructur', 'develop', 'model', 'scenario', 'road', and 'rail'. If

we now focus our analysis on terms with high TF*IDF distribution, we see that some of the terms identified previously become less visible or, inclusively, unobserved. Due to the indexing rationale behind TF*IDF measure, words like 'flow', 'model', 'ten', or 'gdp' gain more visibility. To some extent, this situation increases the robustness of the dataset by focusing on terms with high levels of explanatory power. As it can be seen in Figure 3c and 3d, the best terms to label sub-themes correspond to 'model', 'scenario', and 'ten'.

Clearly, one of the best densities that can be identified in both Figure 3c and 3d correspond to a group of words that include 'air', 'rail', 'maritim', 'traffic', and 'flow'. Somehow this shows a strong focus of our corpus on transportation systems. Similarly, it seems that 'scenario', 'model', and 'forecast' suggest a distinct density. In this case, one could say that the impact of EU transport policy compromises medium and long-term scenarios. Also important, but less merged with other words, is 'ten'. Here the word corresponds to Trans-European Transport Network (TEN-T) and most of the maps depict data without making any sort of similarity linkage to other potential words; the only exception is data co-occurrence maps based on TF*IDF distribution.

In sum, we believe our analysis discloses some hidden knowledge on a dataset of text information related to 'transport' and, by doing so, facilitates the decision-making process by reducing uncertainly and doubt on words to label sub-themes. Against this background, we suggest that the above-mentioned theme of the ESPON 2013 DB should integrate the following sub-themes:

03 TRANSPORT

0301 Accessibility (e.g. performance indicators, multimodal accessibility)

0302 Flows (e.g. vehicles, passengers, goods, freight)

0303 Infrastructure (e.g. transportation systems, railways, airports, harbours)

To a certain extent, the proposal of sub-themes to allocate data in the field of 'Transport' is similar to the one suggested in the previous ESPON Database. Nevertheless, we believe that such proposal integrates a comprehensive structure of transport information, ranging from data on flows of vehicles, passengers and goods to infrastructures, safety, and investments in the transport sector. It considers as well data that could be delivered on sustainable development, modal split, and environmental impact indicators, including the contribution of each mode of transport, used alone and in combination with others.

Energy & Environment

We assume that energy and environment are complementary and, in many ways, essential for sustainable development. Several policy documents delivered by EU institutions state that sustainable development corresponds to the improvement of citizen's quality of life while reducing the use of natural resources and pressures on the environment (CEC, 2001; EEA, 2002). However, the quality of life is enhanced by costly energy services. The main question, according to these institutions, is how to make use of available energy resources without preventing the needs of future generations. In order to meet the right balance it is necessary to consider other aspects, such as climate change, loss of biodiversity, or ozone layer depletion.

The empirical comparison of data co-occurrence for the two parameters (i.e. paragraph and group of 5 words) is far too similar. The densities observed in both maps only give prominence to 'energy' (Figure 4a). This result was expected, in part, due to the high presence of inflectional forms of words related to this specific word. However, when we add this word to an exclusion list the picture obtained is rather different as other densities emerge. This seems to be the case for 'climat' and 'chang', 'environment' and 'sustain', or 'urban', 'transport' and 'demand'. It is also interesting to note the presence

of a density related to industry, biomass and fossil fuels. In fact, many authors state that replacing fossil fuels by sustainable-produced biomass is seen has a safe method to reduce CO_2 emissions to the atmosphere and therefore the impact on the environment (Gustavsson, 1995; Forsberg, 2002).

Let us now consider the results of data co-occurrence based on TF*IDF distribution in somewhat more detail. The density view of both maps is meaningful because it shows the presence of some terms identified previously and related to nuclear, fossil, and renewable energy sources (e.g. 'oil', 'coal', 'wind', 'solar', 'nuclear', 'thermal'). The difference is that now we just consider terms with high explanatory power and try to depict the information using distinct parameters of linguistic discourse. Here, it is visible the presence of terms that express concern about the subject in analysis, such as 'sensit' or 'vulnerab', or even terms that call for adaption, such as 'adapt'. This is even more evident for densities that combine 'household', 'gdp', and 'employ' to illustrate some of the possible effects emanating from climate change. In a way, the combination of these terms is understandable, especially if we consider that low income households tend to live in areas with low GDP growth, high unemployment rates and therefore more likely to be affected by climate change, and have far less ability to move or make the necessary adjustments to their living conditions.

After conducting the co-occurrence analysis in our corpus, we propose three sub-themes to structure data delivered by TPGs in the field of 'Energy & Environment'. The sub-theme structure and inheritance is the following:

04 ENERGY & ENVIRONMENT

0401 Energy and resources (e.g. renewable, nuclear, and fossil energies)

0402 Environmental facets of climate change (e.g. GHG emissions, air pollution)

0403 Vulnerability impacts (e.g. households, GDP, employment, quality of life)

The first sub-theme is intended to include data on energy and resources, including renewable, nuclear, and fossil energies. The second sub-theme should incorporate distinct features or elements that actively contribute to climate change (e.g. air and soil pollution, biodiversity loss, water management, greenhouse gas emissions). Finally, the third sub-theme is designed to integrate indicators and indices aimed at evaluate the sensitivity and vulnerability impacts on households, employment, quality of life, industry, among other topics.

Land Use

Land use refers how the earth's surface is used, including the location, type and design of human development. As a result, land use patterns have diverse economic, social and environmental impacts. In the previous ESPON Database, 'Land Use' is defined as a first-level theme but its inheritance is somehow vague in the description[2].

The collection of documents used for analysis is not very substantial. Nevertheless, we have managed to gather some documents based on research activities conducted by EU institutions or commissioned to universities and research institutions on behalf of the funding entity (see Annex 5). The visualization of similarities of terms extracted from such context provided a better understanding of data co-occurrence. In Figure 5a we can see the pattern of similarities between terms with high frequency distribution for documents related to 'Land Use'. Clearly, 'urban', 'model', 'land', and 'area' are among those. However, the knowledge obtained from the exclusion of these terms conveys

[2] Two sub-themes have been defined to structure data in the field of 'Land Use'. These are: '111 Natural resources' and 'Land use' (ESPON, 2005).

other relationships. This seems to be the case for ‘chang’, ‘impact’, ‘environment’ and ‘develop’. The same applies for ‘agricultur’, ‘produc’, and ‘cropland’. Less visible, but still important, is relationship between ‘transport’, ’sprawl’, and ‘energi’. The term ‘scenario’, itself very important in land use discussions, appears completely isolated from the other main densities. Overall, the maps presented in Figure 5b disclose relevant information for analysis.

Among the visible interactions established by our corpus in the field of ‘Land Use’, it is possible to identify a sequence of terms that correspond to changes in land-use for both rural and urban settings. This facet seems to be evident for terms like ‘chang’, ‘impact’, ‘rural’ and ‘urban’. In fact, most of the impacts related to land-use have an effect in rural and urban contexts. This also holds true for socio and economic factors. In addition to these, the term ‘scenario’ also discloses some significance. The ability to forecast land-use scenarios is essential to better understand dynamic processes which are determined by a range of driving forces, including demographic, socio-economic, and environmental change.

If we consider the strong emphasis given to these terms and its similarity, one could assume that the focus of our corpus is oriented to land-use changes, impacts and scenarios. As a consequence, we propose the following structure of sub-themes for ‘Land Use’ data:

05 LAND USE

0501 Land use and land cover types (e.g. CORINE Land Cover, GMES)

0502 Urban land use attributes and changes (e.g. LUZ, Urban Atlas)

0503 Rural land use attributes and changes (e.g. Natura 2000)

The first sub-theme has not been defined with text mining tools. However, its purpose is to integrate data related to CORINE Land Cover (CLC), Natura 2000, and the Urban Atlas Initiative. The other two sub-themes derive from a qualitative description of term similarity maps within a corpus of documents in the field of ‘Land Use’ and are meant to integrate data on changes, including indicators and indices on land use changes and impacts.

Social Affairs

The theme related to ‘Social Affairs’ is meant to cover data on social, economic and cultural issues with an emphasis on employment, labour market, income, living conditions, and poverty. Our collection of documents is based on ESPON evidence and results and, alternately, findings from other sources used by TPGs while undertaking research in this sort of topics. In total, we collected 12 documents from different sources, ranging from studies, policy notes and technical reports (see Annex 5).

Unfortunately, in this case, the density maps of data co-occurrence do not add any relevant information. This means that the maps on both parameters are dominated by terms with little explanatory power and therefore terms with the highest frequency distribution should be added to an exclusion list. This seems to be the example for ‘polici’ and ‘social’ (see Figure 6a). However, if we remove these terms from our analysis the density map will depict other similarities offering a better understanding of how inflectional forms of words are associated with each other. Here, a special mentions should be made for ‘labour’ and ‘employ’, econom’ and ‘develop’, and ‘indic’ and ‘cultur’ (see Figure 6b).

The capacity of visualising similarities in a multidimensional scaling dramatically increases with the TF*IDF scoring algorithm. As it can be seen in Figure 6c and 6d, the

knowledge structure is relatively easy to comprehend. Somehow, these maps suggest that TF*IDF measure reinforces the importance of terms less visible within the corpus. Besides, it clearly differentiates the major similarities. For instance, 'cultur' and 'heritag' reveal a distinct similarity. The same applies for 'labour' and 'job', 'household' and 'health', or 'incom' and 'famili'. Equally relevant is the presence of 'poverti'. Most of these similarities underline the rationale behind the theme in the field of 'Social Affairs' and, as a consequence, facilitate the definition of sub-themes.

The thematic structure designed for the ESPON 2006 Database suggests that data on employment and labour market should be disconnected from social exclusion (e.g. poverty) (ESPON, 2007). However, the results of this experiment indicate the opposite meaning that both should integrated and, if possible, include data on similar issues, such as living conditions or health systems.

06 SOCIAL AFFAIRS

0601 Education (e.g. training, lifelong learning)

0602 Labour market (e.g. labour force, labour costs, economic inactivity, earnings)

0603 Living conditions (e.g. poverty, social exclusion, health systems)

0604 Culture (e.g. socio-cultural activities, cultural consumption)

The sub-theme on 'education is designed to integrate data on training and lifelong learning while 'labour market' is more focused on economic inactivity, average earnings, and productivity. Within the same structure we suggest a second-level theme related to 'living conditions'. Ideally, this sub-theme will serve the purpose of integrating data on poverty, social exclusion as well as other types of living conditions, including data on health systems. Finally, our experiment with text mining also suggests a strong emphasis on issues related with culture and heritage. With this respect, we propose a sub-theme to allocate data on socio-cultural activities, cultural consumption and participation.

Economy

The economic analysis of EU strengths and weaknesses is of great importance to understand the policy designed by the Lisbon Strategy. According to the European Council its goal was to make the EU 'the most competitive and dynamic knowledge-based economy in the world capable of sustainable economy growth with more and better jobs and greater social cohesion' (CEC, 2000).

Despite joint efforts to achieve these goals only a small number of actions could have been fully implemented. One of the recent drawbacks to justify the moderate success of this initiative is the serious economic crisis that hit Europe and its citizens. As a response to such event, the European Commission has launched the 'Europe 2020 Strategy' in order to re-adapt the Lisbon Strategy to new challenges (CEC, 2010).

Access to accurate data is therefore of crucial importance to comprehend, for instance, the role of R&D, innovation, and patents to boost competitiveness in Europe. With this regard, the ESPON 2013 Database has the opportunity to structure both statistical and geographical data related to these topics.

The analysis undertaken to extract terms and label sub-themes in the field of 'Economy' is based on a dataset of 10 documents that combine ESPON evidence and results (see Annex 7). The visualisation of knowledge structure created by these documents in a VOSviewer environment is meant to capture the similarity degree of terms with both explanatory power and non-explanatory power. As is can be seen in Figure 7a and 7b, the keyword co-occurrence of inflectional words is quite obvious for 'develop' and 'econom'. The same applies for 'innov' and 'research', or 'fund' and 'structur'.

Surprisingly enough, the exclusion of terms with high frequency distribution (i.e. 'develop' and 'econom') does not demonstrate the expected impact on the density view of data co-occurrence and, therefore, its structure neither changes nor generate unseen similarities. In this sense, the major subgroups identified previously can still be seen in the VOSviewer maps of Figure 7b. It is particularly interesting to observe the strong correlation between 'servic', 'industri', and 'capit', the continuous emphasis on 'knowledge', 'innov', and research', or the association established between 'territorial' and 'structur'.

The application of the TD*IDF measure also plays a decisive role in this analysis. The VOS maps presented in Figure 7c depict the knowledge structure of data co-occurrence based on TF*IDF distribution and, Figure 7d, illustrates what would happen if we added the most significant TF*IDF terms to an exclusion list. In general, our results do not add more information than what we have so far exposed. Therefore, the relationship established by terms with high TF*IDF distribution did not extract unknown and potential information. The terms identified by the preceding experiment confirm this evidence. Hence, we propose the following sub-themes in the field of 'Economy':

07 ECONOMY

0701 Aggregated accounts (e.g. GDP, purchasing power parities, balance of payments)

0702 Employment (e.g. employment, unemployment, long-term unemployment)

0703 Production and costs per sector (e.g. production of manufactured goods)

0704 Research and innovation (e.g. R&D expenditure, ICT research, patents)

The knowledge structure of our corpus suggests the presence of four sub-themes. The first sub-theme is the less visible in our maps. Still, we decided to introduce it so that TPGs involved in ESPON projects may allocate data on GDP and its main components. The other three sub-themes emerged more implicitly due to the high similarity or relatedness of terms and its clear organisation in clusters. Therefore, the sub-theme on 'employment' should allocate data on employment and unemployment. The sub-theme related to 'production and costs per sector' is meant to incorporate data on production of manufactured goods. Finally, for 'research and innovation' TPGs are encouraged to integrate date on R&D expenditure, ICT research, patents, and public investments.

Non/Cross-Thematic Data

This theme is meant to cover both cross-thematic and non-thematic data. A first sub-theme should include variables that mix themes (e.g. integrative indicators; complex typologies; trends and impacts on both CAP and TEN-T; scenario-based projections on urban development; environmental, social, and economic concerns associated with lad-use changes). The second subset refers to base maps (i.e. administrative units) and other geographical objects (e.g. grids, cities, networks) or spatial delineations (e.g. morphological zones, functional areas). However, this proposal is open to more sub-themes. Its current structure looks as follows:

99 NON/CROSS-THEMATIC DATA

9901 Integrative indices, indicators and scenarios (e.g. typologies, scenarios)

9902 Geographical objects (e.g. administrative units, grids, networks)

Conclusion

This report complements our two-step approach to structure the ESPON 2013 DB by themes and sub-themes. For first-level themes, we argue that database structures adopted by international organisations constitute an important source of information. This has been further explained in our last technical report.

Here, we investigate the definition of sub-themes. For this purpose, we believe the demand from the ESPON 2013 DB end-users will correspond to immediate, easy and practical access to data. In order to meet this request, we explore the potentialities offered by text mining tools. This approach widely used to find patterns across textual data generates a multi-faceted overview of topics in large text collections.

For this purpose, we investigate the potentialities of such tools to classify sub-themes. Our collection of texts in based on ESPON evidence and, for each of the first-level themes, we demonstrate that is possible to shed some light on ways to further progress in this field. This assumption has greatly beneficiated from mapping techniques of multidimensional scaling to ease the interpretation of relational similarities that emerged from data co-occurrence with both explanatory power and non-explanatory power. The identification of such patterns on data co-occurrence suggests that the ESPON 2013 DB should be structured in 29 sub-themes unveiling the inheritance of 7+1 themes (see Annex 8).

References

Berry, M. [ed.] (2004). Survey of text mining. Clustering, classification, and retrieval. Springer: New York.

Blanchard, A. (2007) Understanding and customizing stopword lists for enhanced patent mapping. World Patent Information, 29 (1), 308-316.

CEC (2001) Environment 2010: Our future, our choice. Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions. Commission of the European Communities: Brussels.

CEC (2010) Europe 2020. A European Strategy for Smart, Sustainable and Inclusive Growth. Commission of the European Communities.

Davi, A.; Haughton, D.; Nasr, N. et al (2005). A review of two text mining packages: SAS TextMining and WordStat. *American Statistician*, 59(1), 89-103.

Eck, N. & L. Waltman (2007). VOS: a new method for visualizing similarities between objects. In H.-J. Lenz and R. Decker [ed.]. Advances in Data analysis: Proceedings of the 30th Annual conference of the German Classification Society. Studies in Classification, Data Analysis, and Knowledge Organization. Springer: Heidelberg, 299-306.

Eck, N. & L. Waltman (2009). A computer program for bibliometric mapping. In B. Larsen and J. Leta [eds.]. Proceedings of the XII International Conference on Scientometrics and Informetrics, 886-897.

Eck, N. et al (2010) A comparison of two techniques for bibliographic mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, 61(12), 2405-2416.

EEA (2002) Energy and environment in the European Union. Environmental Issue Report No 31. European Environment Agency: Copenhagen.

ESPON (2005). Integrated Tools for European Spatial Development. ESPON 3.1 Project. Luxembourg: ESPON Coordination Unit, pp. 141-174.

European Council (2000) Presidency conclusions. Lisbon European Council, 23-24 March 2000. European Council: Brussels (available for download on www.consilium.europa.eu).

Forsberg, C. (2008) Sustainability by combining nuclear, fossil, and renewable energy sources. *Progress in Nuclear Energy*, 1-9.

Gustavsson, L. et al (1995) Reducing CO_s emissions by substituting biomass for fossil fuels. Energy 20(11), 1097-1113.

Lewis, B. (1999). Simstat with Wordstat: A Comprehensive Statistical Package with a Content Analysis Module. *Field Methods*, 11(2), 166-179.

Weiss, S.; Indurkhya, N.; Zhang, T. and F. Damerau (2005). Text Mining. Predictive Methods for Analysing Unstructured Information. Springer: New York.

Yang, Y.; Akers, L.; Klose, T. and Yang, C. (2008) Text mining and visualization tools. Impressions of emerging capabilities. *World Patent Information*, 30 (1), 280-293.

Annex 1

Reports taken into account for text mining purposes on 'Agriculture & Fisheries'

Code	Report
report#01	European Parliament (2007) Regional Dependency on Fisheries. European Parliament: Brussels.
report#02	European Parliament (2007) Reflection on the possibilities for the future development of the CAP. European Parliament: Brussels.
report#03	ESPON (2005) The territorial impact of CAP and Rural Development Policy. ESPON Project 2.1.3 Final Report. Arkleton Institute: Aberdeen.
report#04	ESPON (2005) Territorial Impacts of European Fisheries Policy. ESPON Project 2.1.5 Final Report. NIBR: Oslo.
report#05	ESPON (2010) European Development Opportunities for Rural Areas. EDORA Draft Final Report. UHI Millenium Institute: Inverness.
report#06	ESPON (2010) Territorial Impact for Transport and Agricultural Policies. TIP TAP Final Report A/B. Politecnico di Milano: Milan.
report#07	ESPON (2010) Territorial Impact for Transport and Agricultural Policies. TIP TAP Final Report C. Politecnico di Milano: Milan.

Annex 2

Reports taken into account for text mining purposes on 'Demography'

Code	Report
report#08	European Parliament (1999) Regional development in less-densely populated regions in the EU. European Parliament: Brussels.
report#09	ESPON (2005) The spatial effects of demographic trends and migration. ESPON Project 1.1.4 Final Report. ITPS: Stockholm.
report#10	ESPON (2010) Demographic and migratory flows affecting European regions and cities. DEMIFER Draft Final Report. NIDI: The Hague.
report#11	CEC (2006) The Demographic Future of Europe: From Challenge to Opportunity. Commission of the European Communities: Brussels.
report#12	CCE (2007) Europe's Demographic Future: Facts and Figures. Commission of the European Communities: Brussels.

Annex 3

Reports taken into account for text mining purposes on 'Transport'

Code	Report
report#13	European Parliament (2006) The Impact of Trans-European Networks on Cohesion and Employment. European Parliament: Brussels.
report#14	ESPON (2004) Transport services and networks: Territorial trends and basic supply on Infrastructure for Territorial Cohesion. ESPON Project 1.2.1 Final Report. University of Tours: Tours.
report#15	ESPON (2005) Territorial Impact of EU Transport and TEN policies. ESPON Project 2.1.1 Final Report. Spiekermann & Wegener: Dortmund.
report#16	ESPON (2009) Territorial Impact package for Transport and Agricultural Policies. TIPTAP Draft Final Report. Politecnico de Milano: Milan.
report#17	ESPON (2007) Update of selected potential accessibility indicators. Final report. Spiekermann & Wegener: Dortmund.
report#18	CEC (2007) Trans-European Networks: Towards an integrated approach. Commission of the European Communities: Brussels.

Annex 4

Reports taken into account for text mining purposes on 'Energy & Environment'

Code	Report
report#19	ESPON (2005) Territorial Trends and Policy Impacts in the field of EU Environmental Policy. ESPON 2.4.1 Final Report. Geological Survey of Finland: Helsinki.
report#20	ESPON (2005) Territorial Trends of Energy Services and Networks and Territorial Impact of EU Energy Policy. ESPON 2.1.4 Final Report. CEEETA: Lisbon.
report#21	ESPON (2010) Climate Change and Territorial Effects on Regions and Local Economies. ESPON CLIMATE Interim Report. TU Dortmund: Dortmund.
report#22	ESPON (2010) Regions at Risk of Energy Poverty. ReRisk Draft Final Report. Inasmet-Tecnalia: Donostia/San Sebastian.
report#23	European Parliament (2007) Using Sustainable and Renewable Energies in the context of Structural Policy 2007-2013. European Parliament: Brussels.
report#24	European Parliament (2006) Energy and Structural and Cohesion Policies. European Parliament: Brussels.
report#25	European Parliament (1998) Sustainable Development: A Key Principle for European Regional Development. European Parliament: Brussels.
report#26	European Parliament (2003) The Enlargement Process of the EU: Consequences in the Field of Environment. European Parliament: Brussels.
report#27	European Parliament (2008) The Challenge of Climate Change for Structural and Cohesion Policies. European Parliament: Brussels.

Annex 5

Reports taken into account for text mining purposes on 'Land Use'

Code	Report
report#28	EEA (2006) Urban sprawl in Europe. The ignored challenge. EEA Report No 10/2006. European Environment Agency: Copenhagen.
report#29	EEA (2007) Land-use scenarios for Europe: qualitative and quantitative analysis on a European scale. EEA Technical Report No 9/2007. European Environment Agency: Copenhagen.
report#30	EEA (2009) Ensuring quality of life in Europe's cities and towns. Tackling the environmental challenges driven by European and global change. EEA Report No 5/2009. European Environmental Agency: Copenhagen.
report#31	DG Environment (2008) Modelling of EU Land-use choices and environmental impacts. Scoping study. BIO Intelligence Services: Ivry-sur-Seine.

Annex 6

Reports taken into account for text mining purposes on 'Social Affairs'

Code	Report
report#32	ESPON (2006) The role and spatial effects of cultural heritage and identity. ESPON Project 1.3.3 Final Report. Universita' degli Studi Ca' Foscari: Venice.
report#33	ESPON (2006) Territorial dimension of the Lisbon-Gothenburg strategy. ESPON Project 3.3 Final Report. Universita' degli Studi 'Tor Vergata' di Roma. CEIS: Rome.
report#34	European Parliament (2005) Adaption of Cohesion Policy to the enlarged Europe and the Lisbon and Gothenburg objectives. European Parliament: Brussels.
report#35	ESPON (2006) Preparatory study on social aspects of EU territorial development. ESPON Project 1.4.2 Final report. OIR: Vienna.
report#36	European Parliament (2007) Impact of Accession on the Labour Markets of the New Member States. European Parliament: Brussels.
report#37	European Parliament (2007) The role of minimum income for social inclusion in the European Union. European Parliament: Brussels.
report#38	European Parliament (2009) Indicators of Job Quality in the European Union. European Parliament: Brussels.
report#39	European Parliament (2010a) The link between job creation, innovation, education and training: An assessment of policies pursued at EU level. European Parliament: Brussels.
report#40	European Parliament (2010b) Structural and Cohesion Policies following the Treaty of Lisbon. European Parliament: Brussels.
report#41	European Parliament (2010c) Social Policy Agenda. Directorate-General for Internal Policies, European Parliament: Brussels.
report#42	European Parliament (2010d) Mobility and Integration of People with Disabilities into the Labour Market. European Parliament: Brussels.
report#43	European Parliament (2010e) EU Cooperation in the field of Social Inclusion. European parliament: Brussels.

Annex 7

Reports taken into account for text mining purposes on 'Economy'

Code	Report
report#44	ESPON (2005) The Territorial Impact of EU Research and Development Policies. ESPON Project 2.1.2 Final Report. ECOTEC: Birmingham.
report#45	ESPON (2006) Territorial dimension of the Lisbon-Gothenburg strategy. ESPON Project 3.3 Final Report. Universita' degli Studi 'Tor Vergata' di Roma. CEIS: Rome.
report#46	ESPON (2007) Identification of spatially relevant aspects of information society. ESPON Project 1.2.3 Final Report. EUROREG: Warsaw.
report#47	ESPON (2006) Territorial impacts of EU economic policies and location of economic activities. ESPON Project 3.4.2 Final Report. IGEAT: Brussels.
report#48	CEC (2002) First progress report on economic and social cohesion. Commission of the European Communities: Brussels.
report#49	CEC (2003) Second progress report on economic and social cohesion. Commission of the European Communities: Brussels.
report#50	CEC (2005) Third progress report on cohesion: towards a new partnership for growth, jobs and cohesion. Commission of the European Communities: Brussels.
report#51	CEC (2006) The growth and jobs strategy and the reform of European cohesion policy. Fourth progress report on cohesion. Commission of the European Communities: Brussels.
report#52	CEC (2008) Fifth progress report on economic and social cohesion. Growing regions, growing Europe. Commission of the European Communities: Brussels.
report#53	CEC (2009) Sixth progress report on economic and social cohesion. Commission of the European Communities: Brussels.

Annex 8

Overview of the ESPON 2013 DB thematic structure by themes and sub-themes

01 AGRICULTURE & FISHERIES
- 0101 Farm structure (e.g. farm type, size of farms, income from farming, organic farming)
- 0102 Livestock (e.g. livestock output)
- 0103 Aquaculture and sea fisheries (e.g. aquaculture resources in coastal and marine areas)
- 0104 Forestry (e.g. production, consumption, import/export products)
- 0105 Rural development (e.g. rural development, rural access to services)

02 DEMOGRAPHY
- 0201 Population structure (e.g. age distribution by group and gender)
- 0202 Natural changes (e.g. fertility, mortality, life expectancy)
- 0203 Households (e.g. number and sizes of households)
- 0204 Migrations (e.g. immigration, migration replacement, high-skilled labour migration)
- 0205 Territorial impacts (e.g. population by territorial typologies)

03 TRANSPORT
- 0301 Accessibility (e.g. performance indicators, multimodal accessibility)
- 0302 Flows (e.g. vehicles, passengers, goods, freight)
- 0303 Infrastructures (e.g. transportation systems, railways, airports, harbours)

04 ENERGY & ENVIRONMENT
- 0401 Energy and resources (e.g. renewable, nuclear, and fossil energies)
- 0402 Climate change (e.g. GHG emissions, air pollution)
- 0403 Vulnerability impacts (e.g. households, GDP, employment, quality of life)

05 LAND USE
- 0501 Land use and land cover types (e.g. CORINE Land Cover, GMES)
- 0502 Urban land use attributes and changes (e.g. LUZ, Urban Atlas)
- 0503 Rural land use attributes and changes (e.g. Natura 2000)

06 SOCIAL AFFAIRS
- 0601 Education (e.g. training, lifelong learning)
- 0602 Labour market (e.g. labour force, labour costs, economic inactivity, earnings)
- 0603 Living conditions (e.g. poverty, social exclusion, health systems)
- 0604 Culture (e.g. socio-cultural activities, cultural consumption)

07 ECONOMY
- 0701 Aggregated accounts (e.g. GDP, balance of payments)
- 0702 Employment (e.g. employment, unemployment, long-term unemployment)
- 0703 Production and costs per sector (e.g. production of manufactured goods)
- 0704 Research and innovation (e.g. R&D expenditure, ICT research, patents, investments)

99 NON/CROSS-THEMATIC DATA
- 9901 Integrative indices, indicators and scenarios (e.g. typologies, scenarios)
- 9902 Geographical objects (e.g. administrative units, grids, networks)

Figure 1a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)

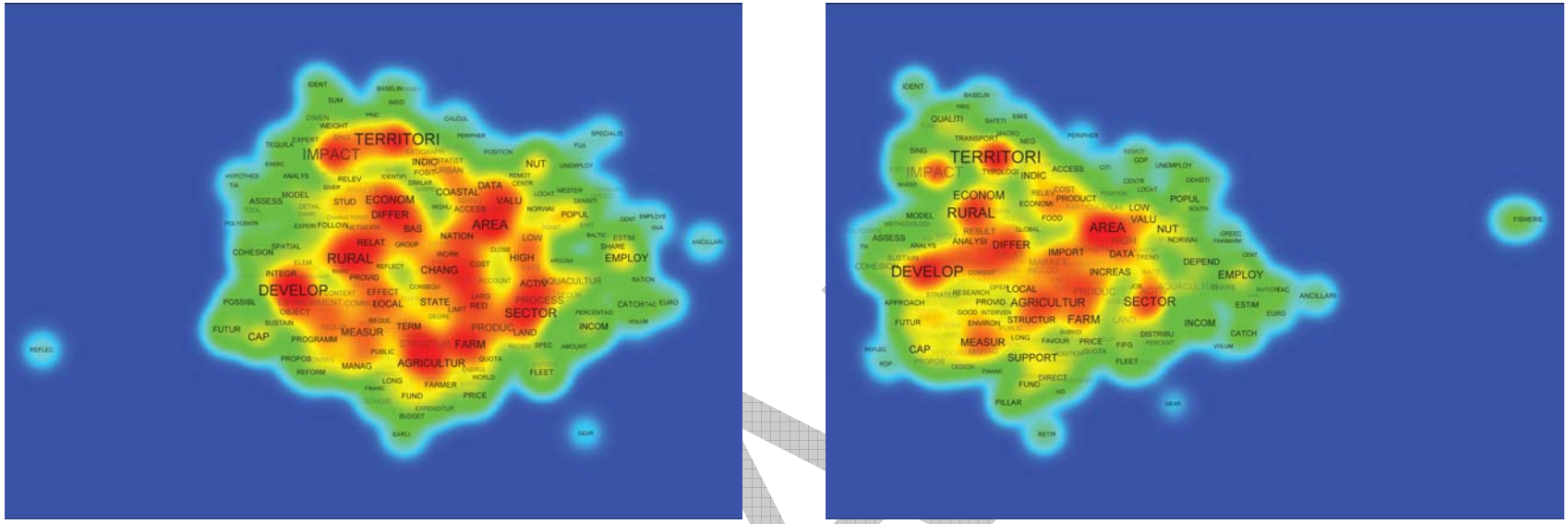
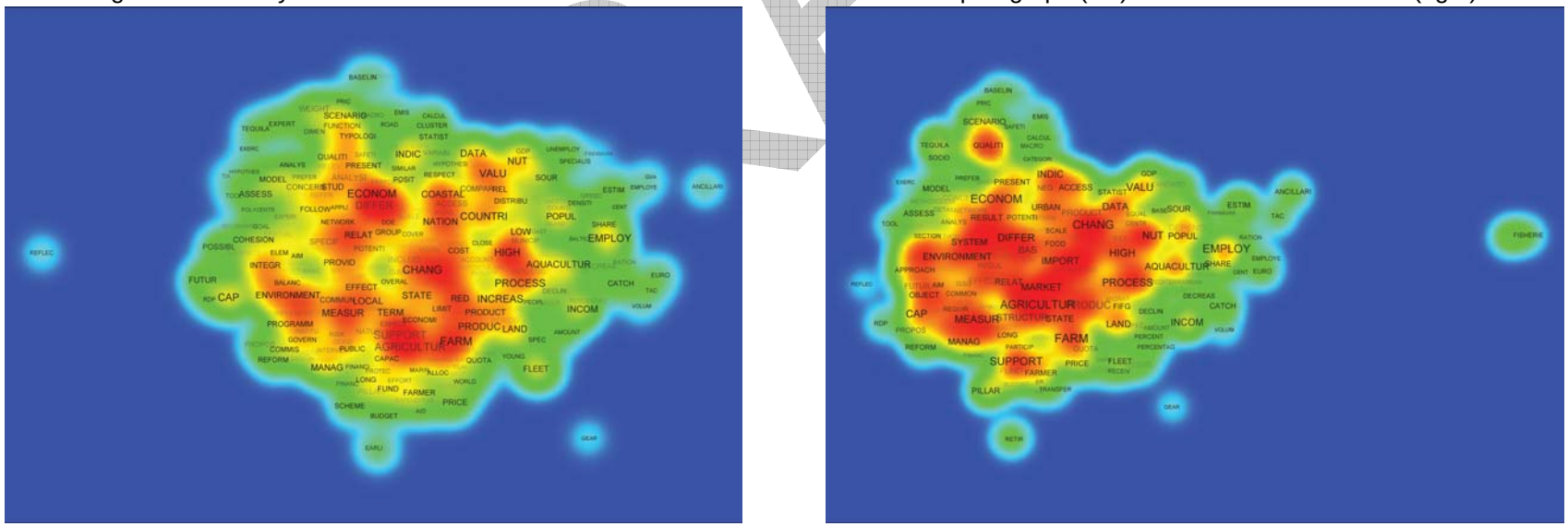


Figure 1b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘territori’, ‘develop’, ‘impact’, ‘area’, ‘rural’, and sector’, present in (1a), were added to a exclusion list in (1b). TF: Term Frequency.

Figure 1c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a window of 5 words (right)

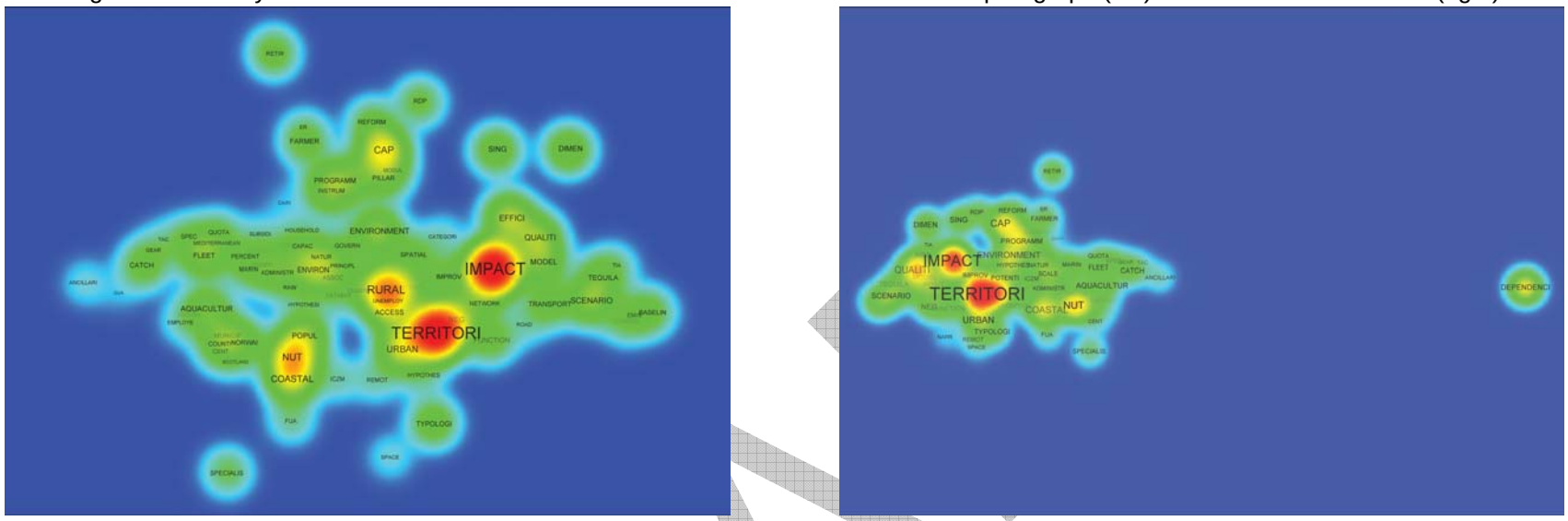
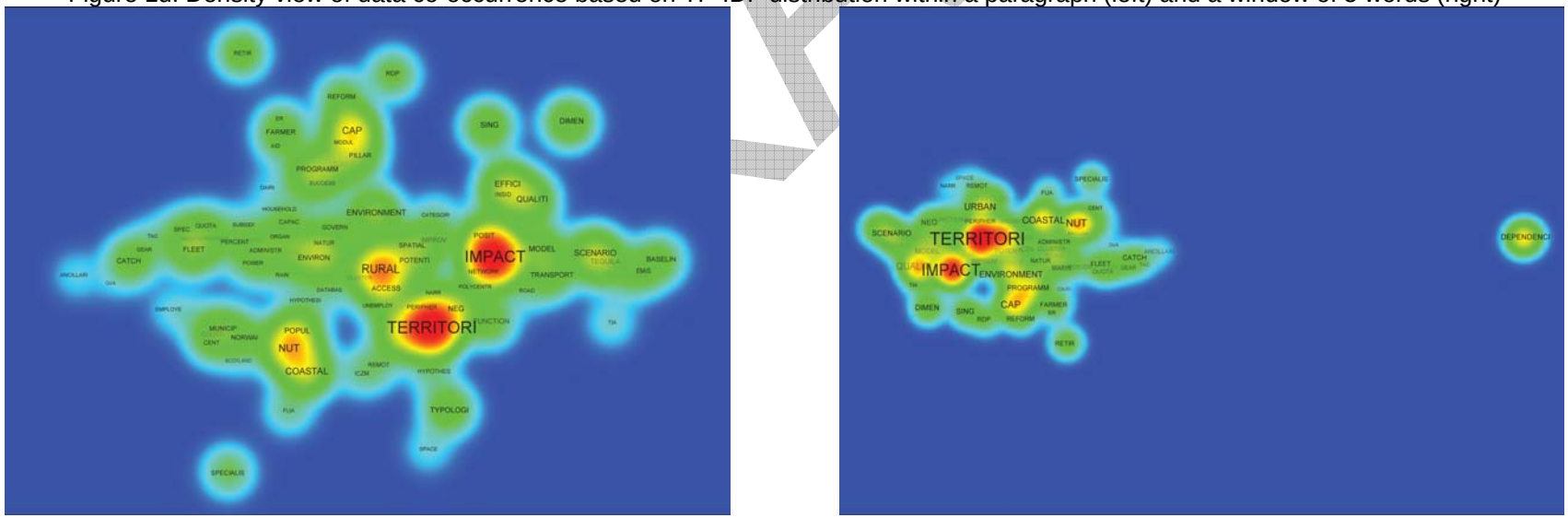


Figure 1d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a window of 5 words (right)



Note: The word ‘acquacultur’, present in (2c), was added to a exclusion list in (2d)

Figure 2a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

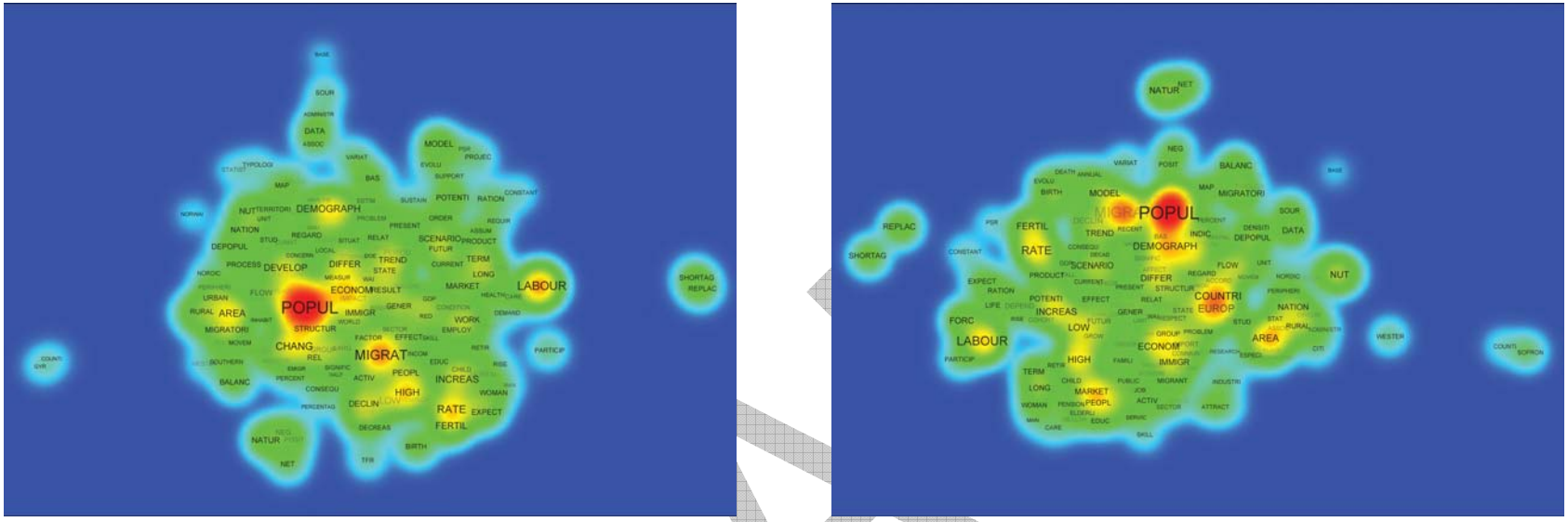
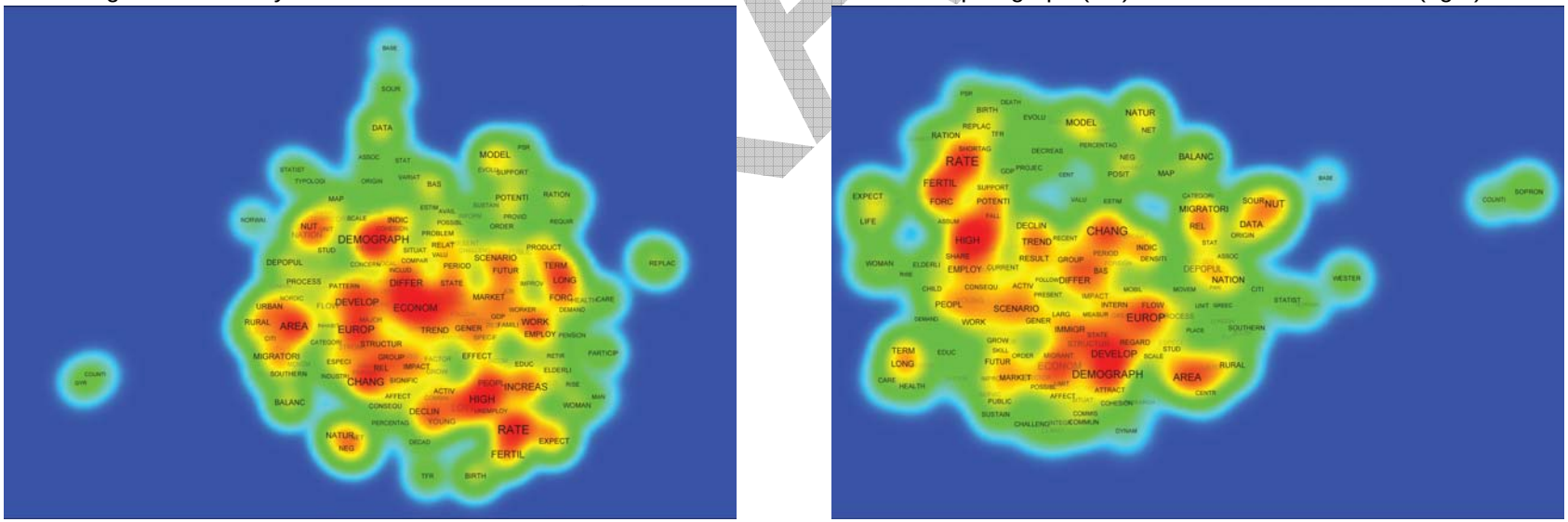


Figure 2b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘popul’, ‘migrat’, ‘countri’, ‘labour’, present in (2a), were added to a exclusion list in (2b). TF: Term Frequency.

Figure 2c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

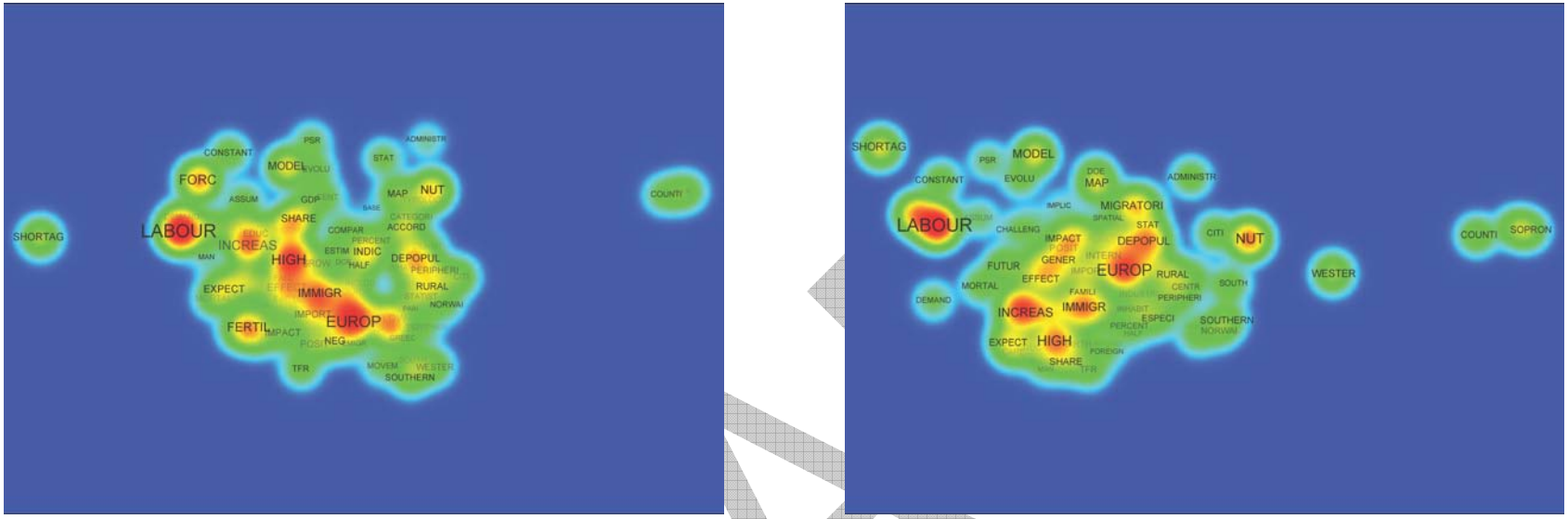
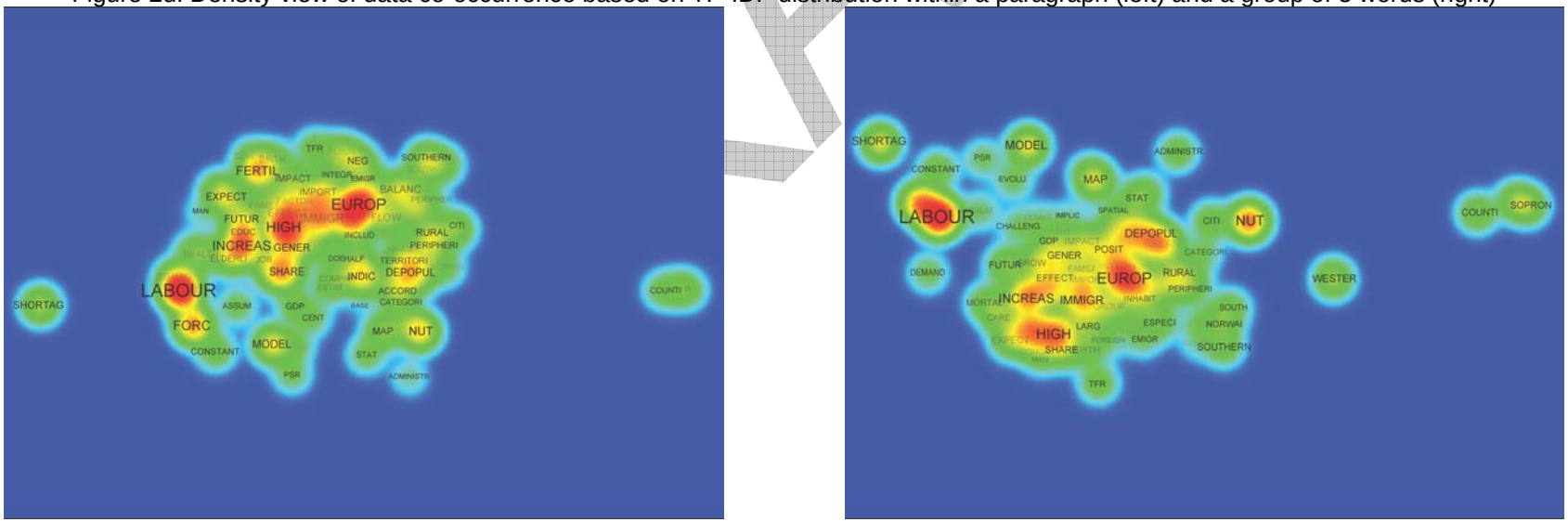


Figure 2d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘migratori’, present in (2c), was added to a exclusion list in (2d)

Figure 3a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

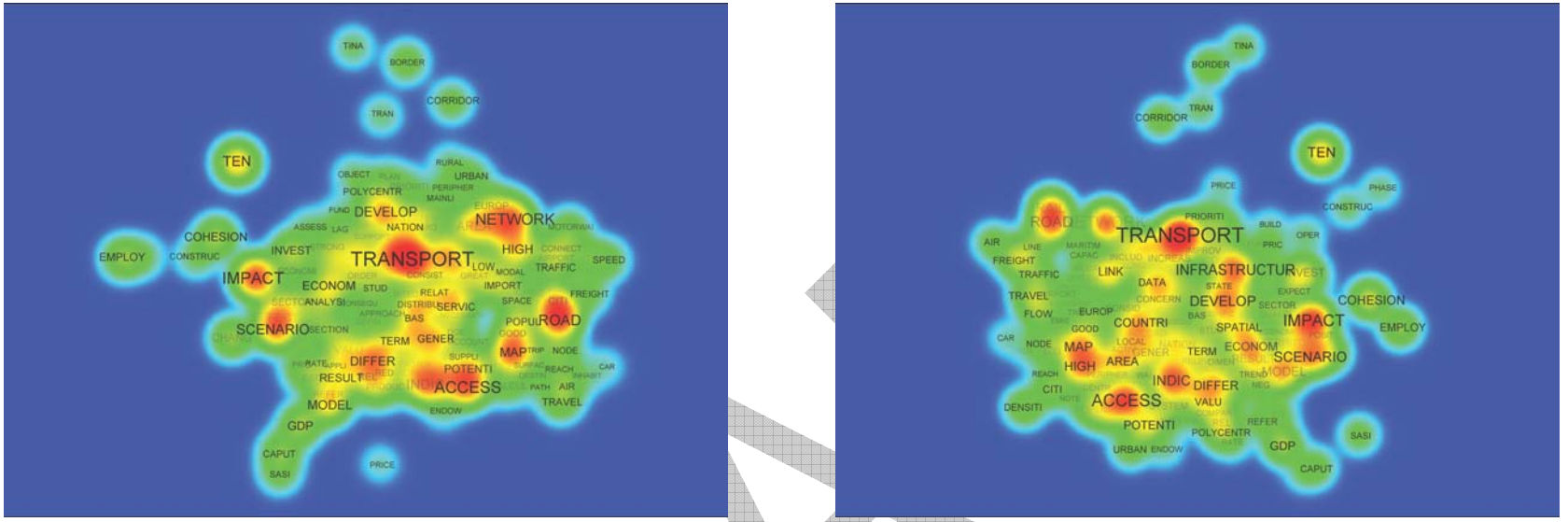
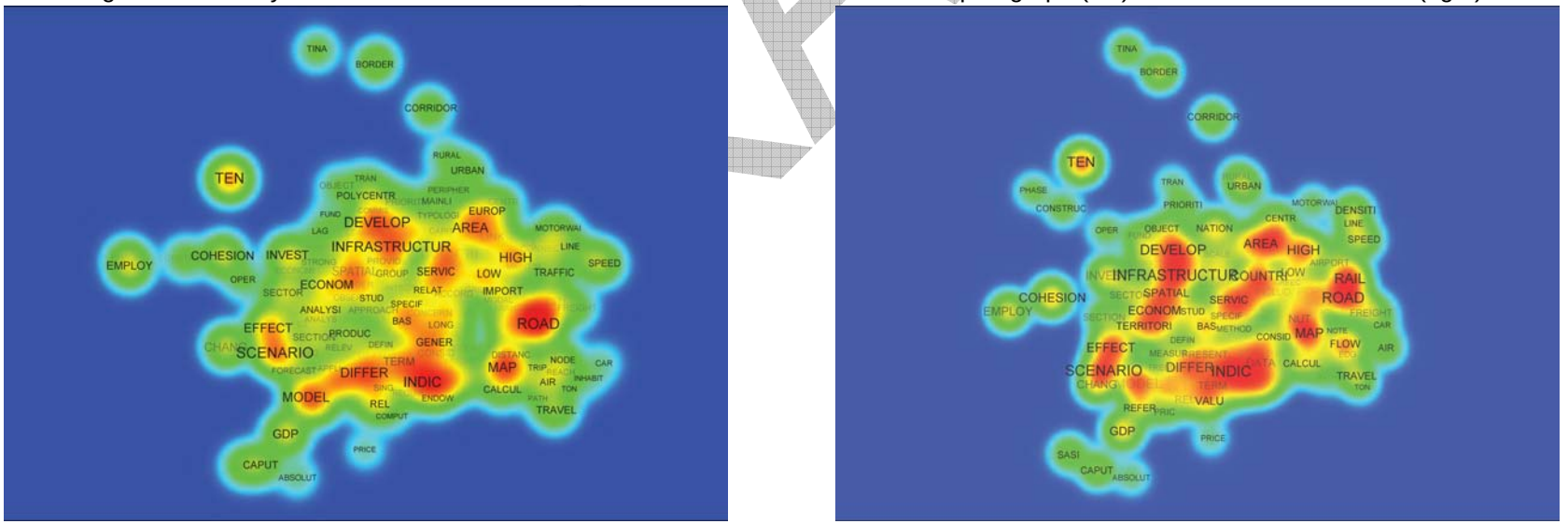


Figure 3b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘transport’, ‘access’ ‘network’, ‘impact’, present in (3a), were added to a exclusion list in (3b). TF: term Frequency.

Figure 3c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

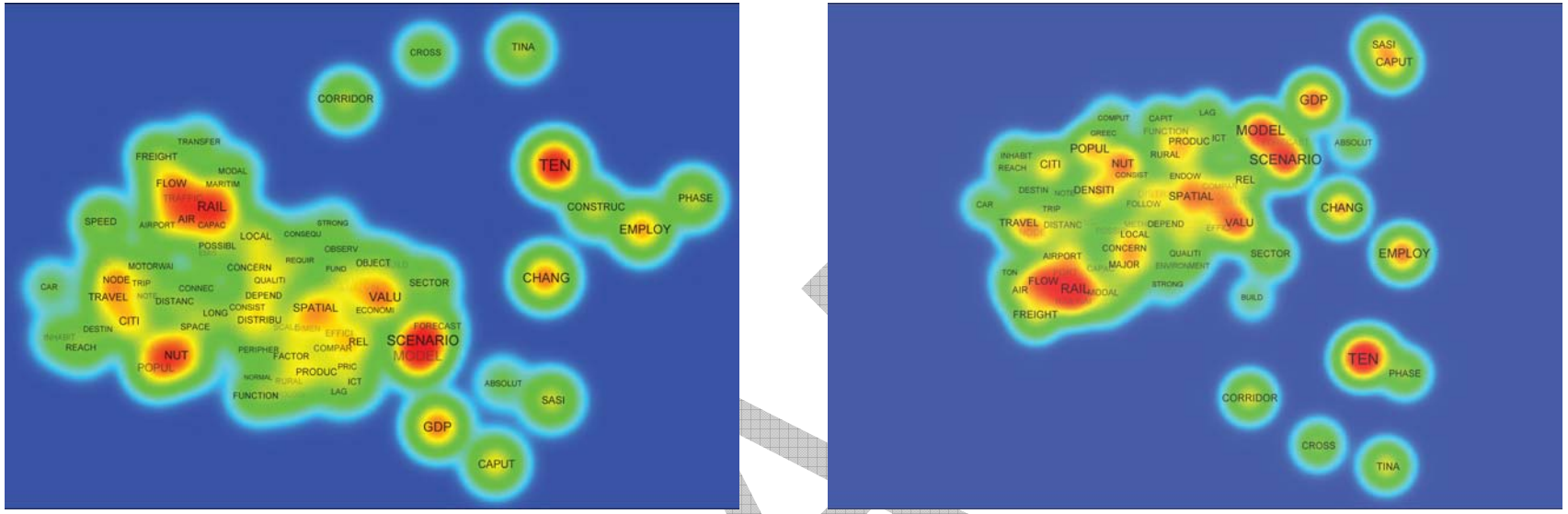
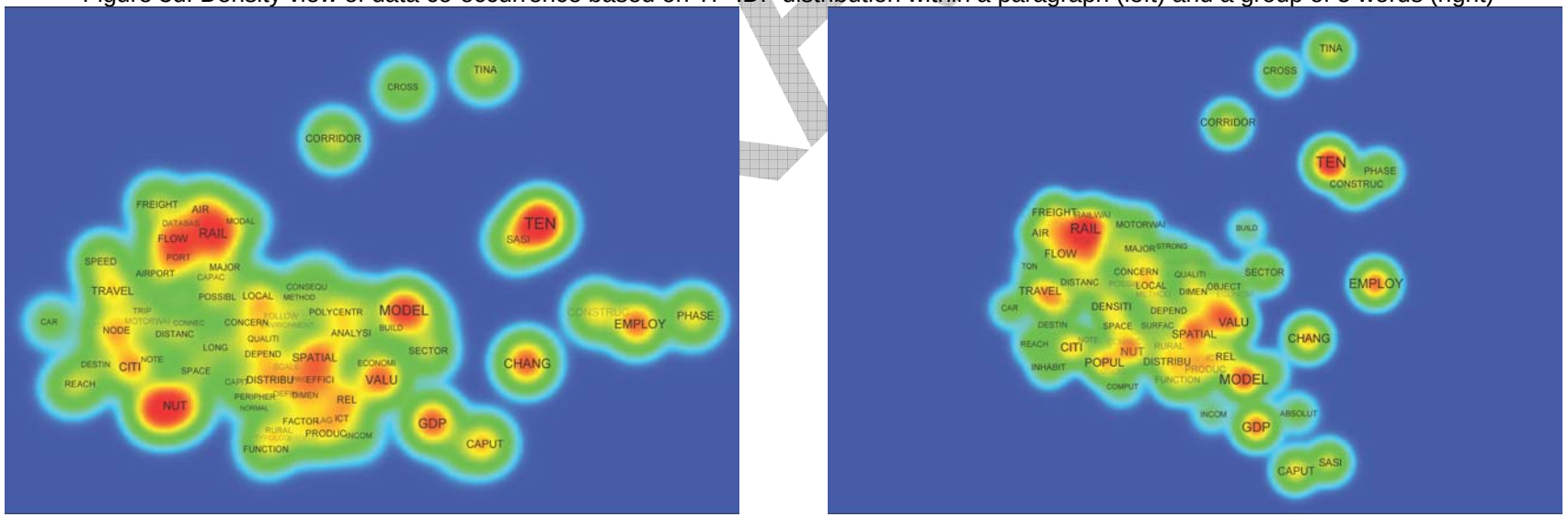


Figure 3d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘scenario’, present in (3c), was added to a exclusion list in (3d)

Figure 4a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

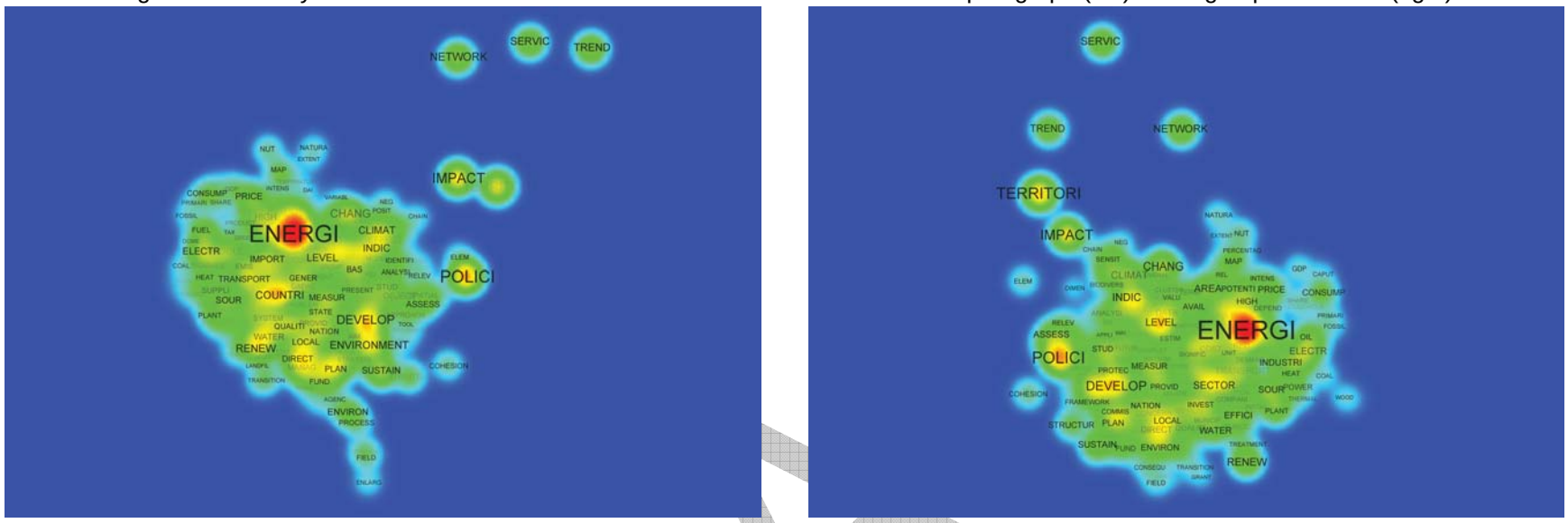
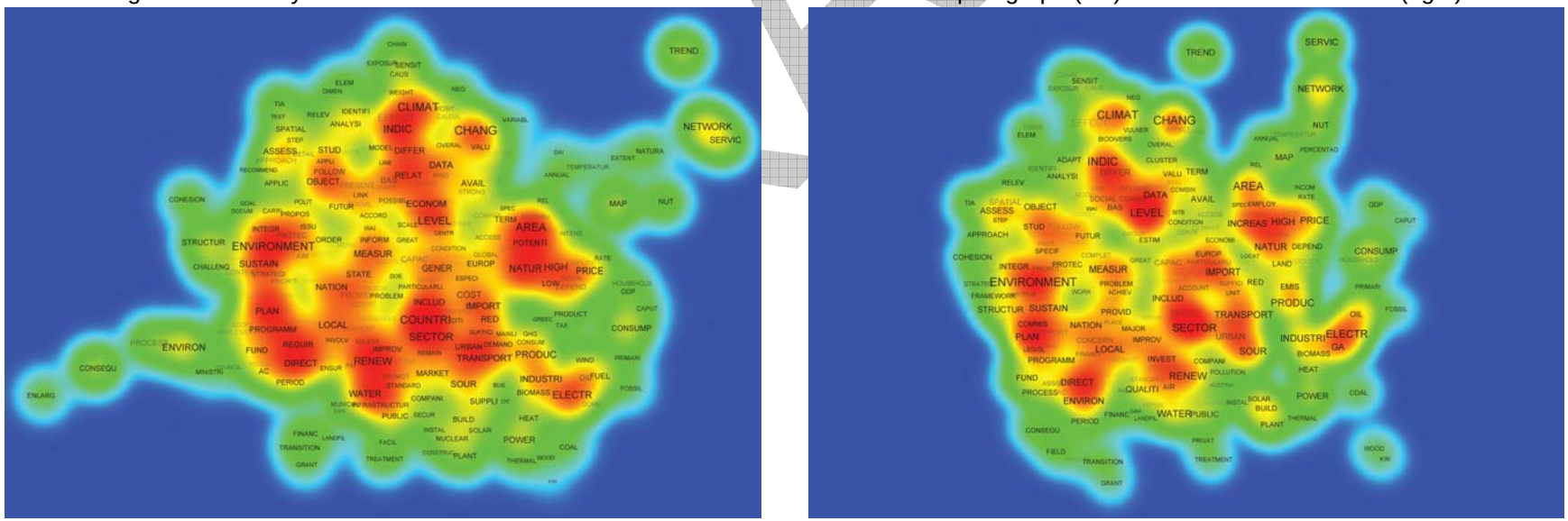


Figure 4b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘energi’, ‘polici’, ‘territori’, ‘impact’, and ‘develop’, present in (4a), were added to a exclusion list in (4c). TF: Term Frequency.

Figure 4c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

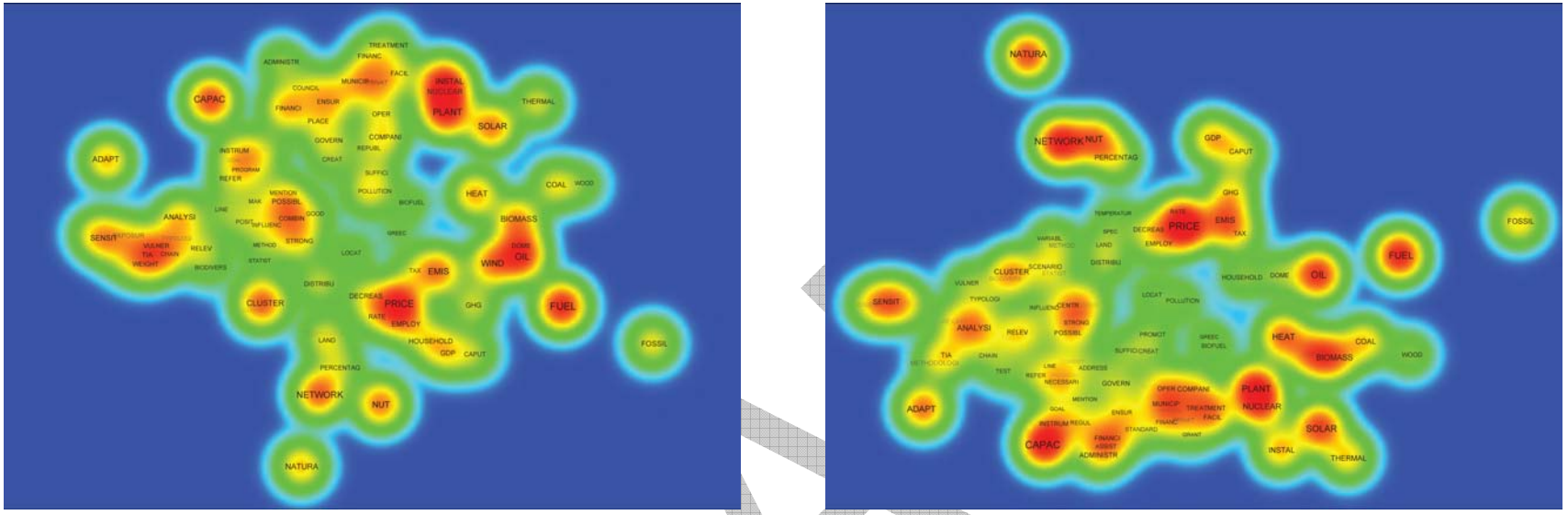
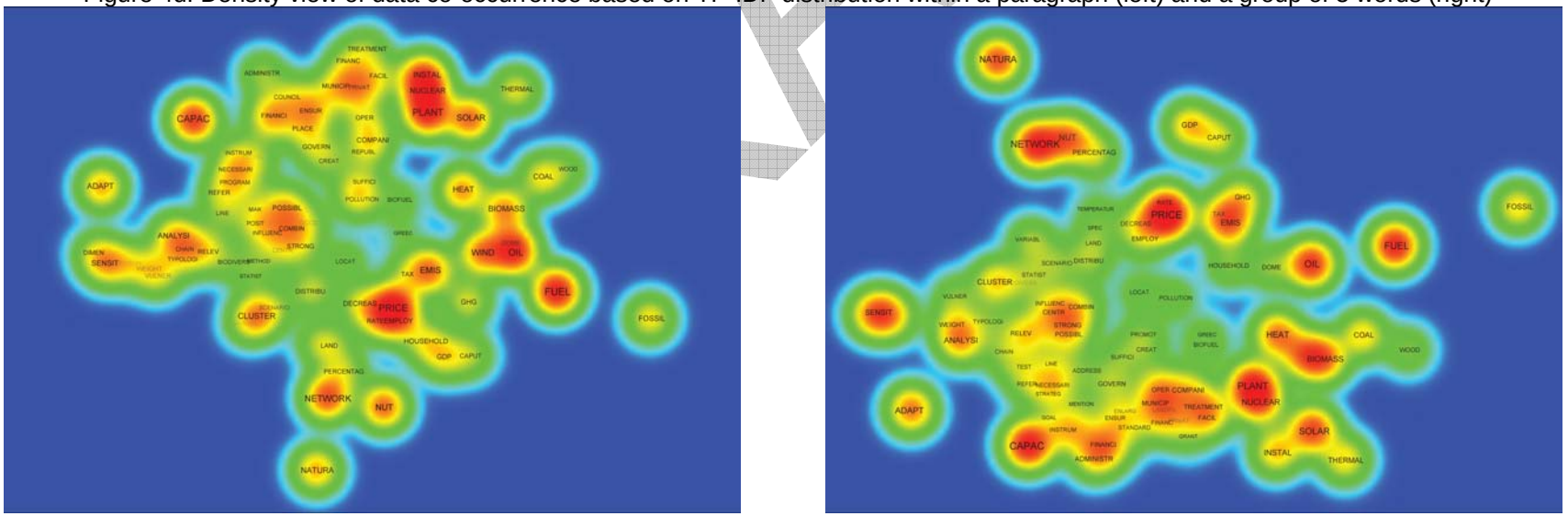


Figure 4d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘tia’, present in (4c), was added to a exclusion list in (4d)

Figure 5a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)

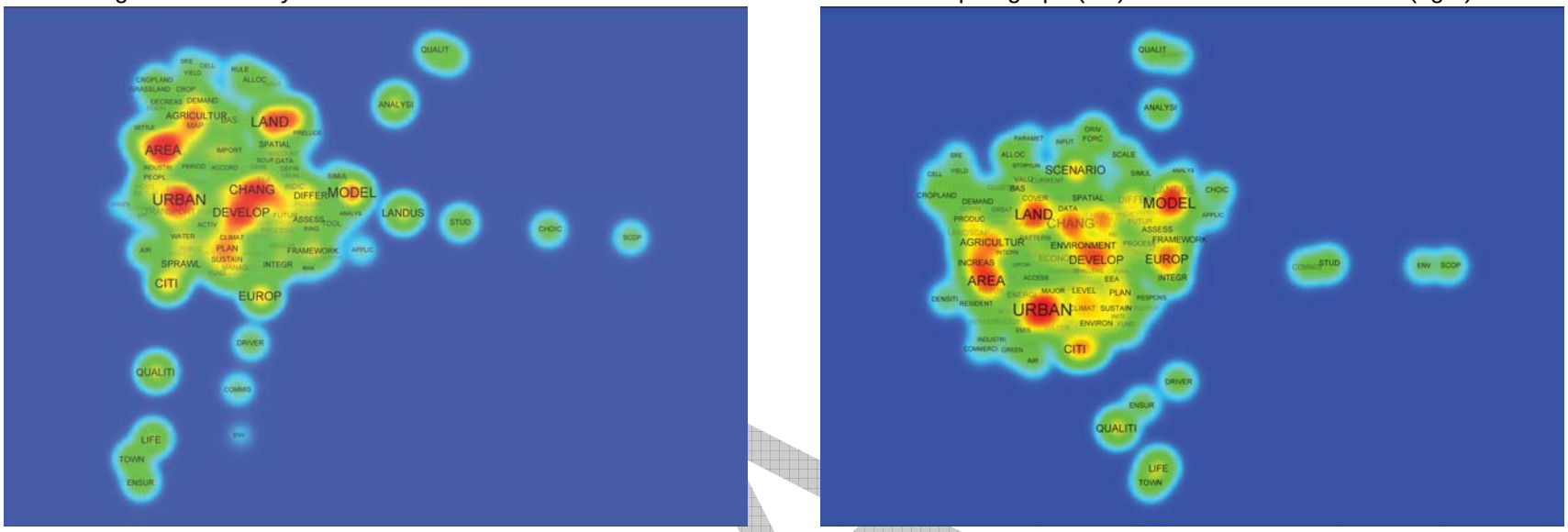
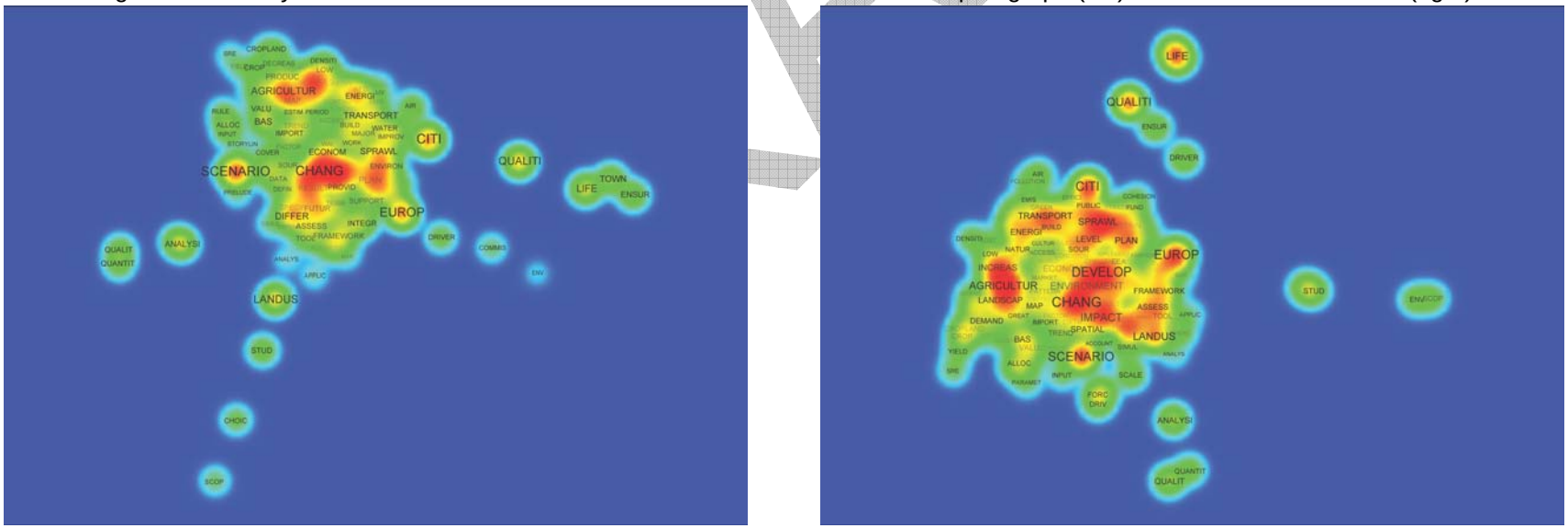


Figure 5b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘urban’, ‘model’, ‘land’, and ‘area’, present in (5a), were added to a exclusion list in (5b)

Figure 5c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

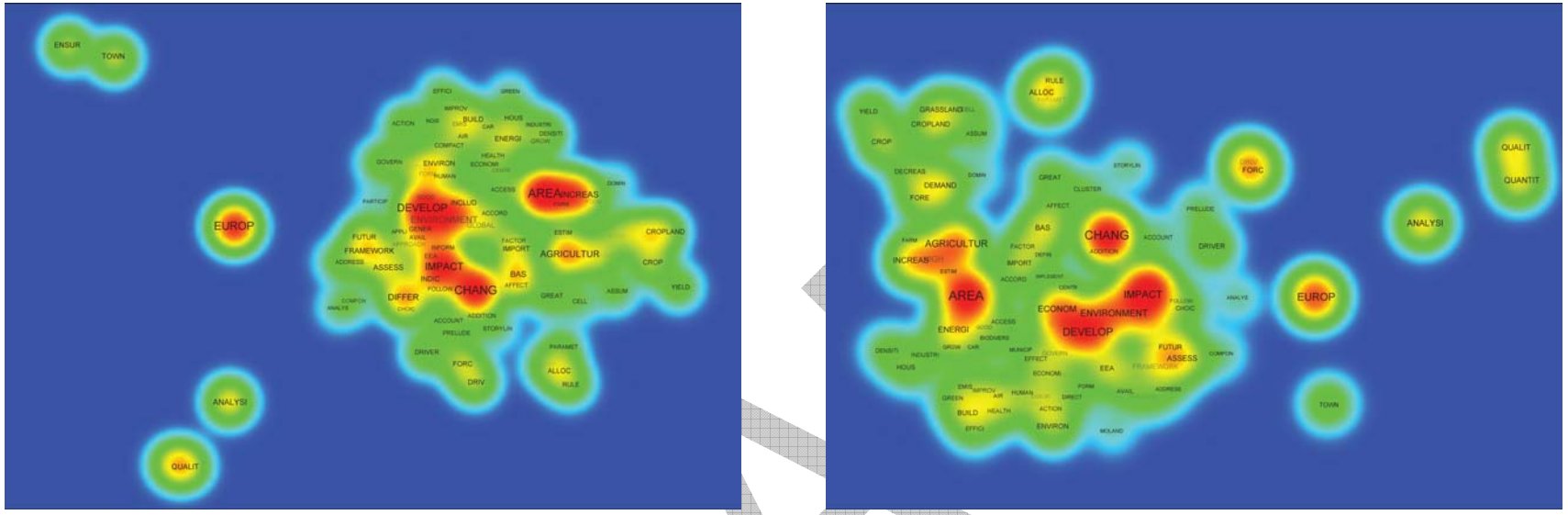
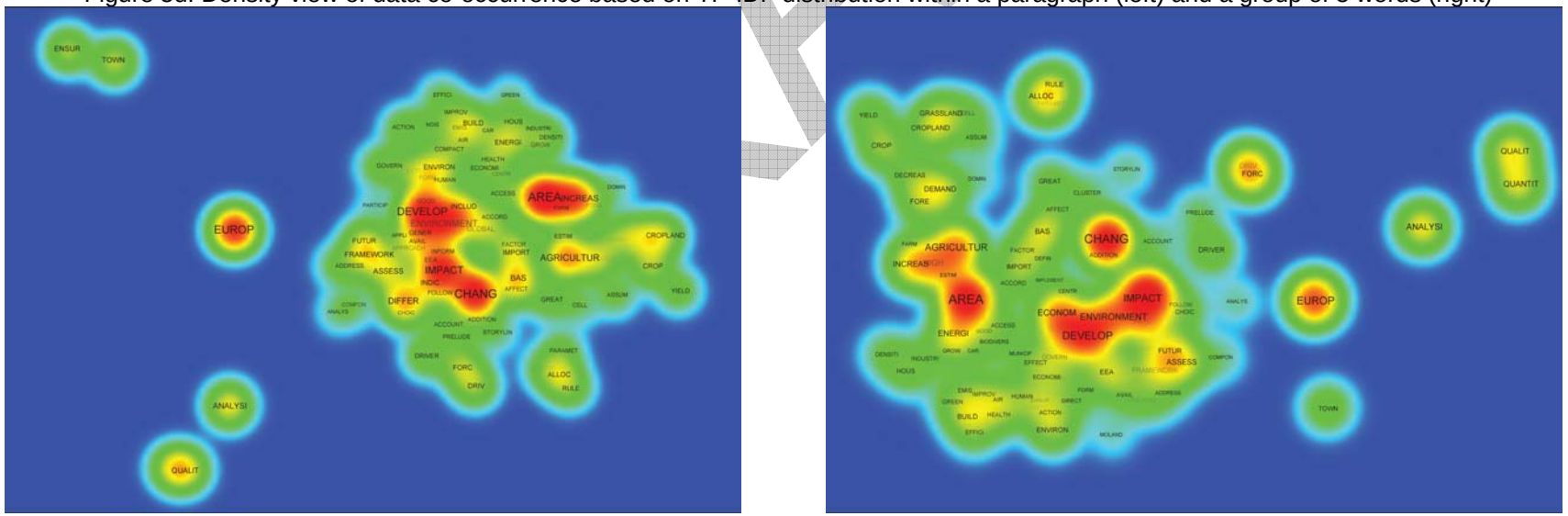


Figure 5d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The word ‘scenario’, present in (5c), was added to a exclusion list in (5d)

Figure 6a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

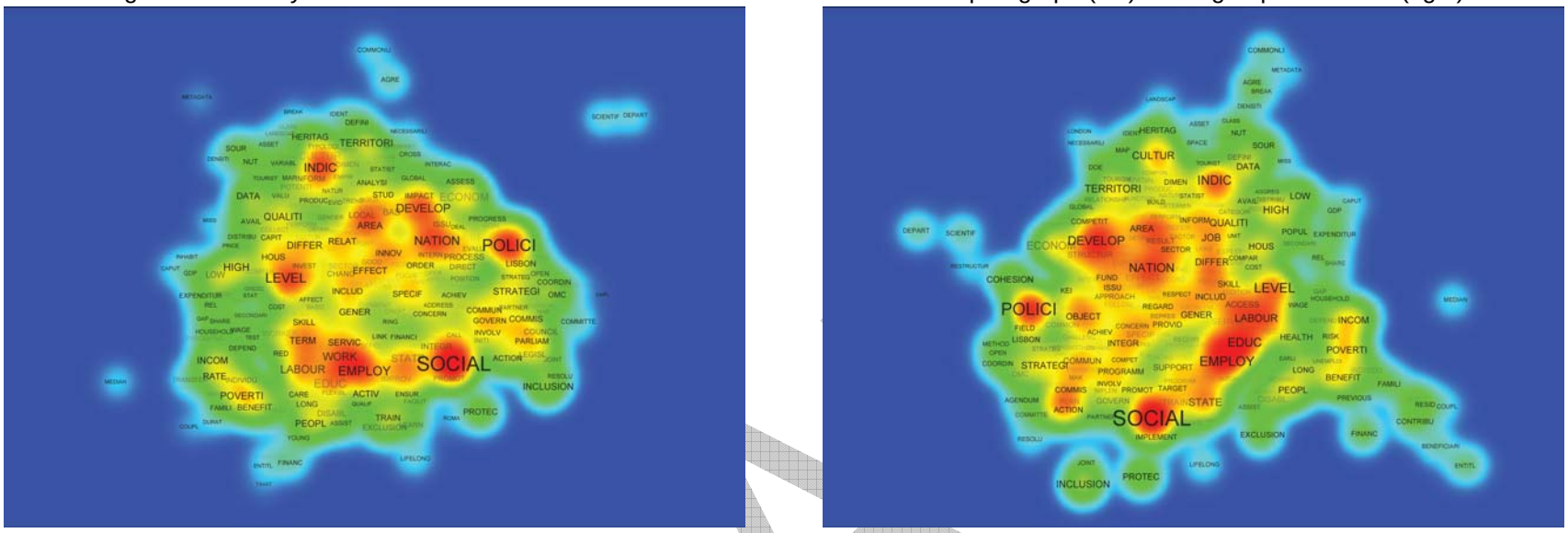
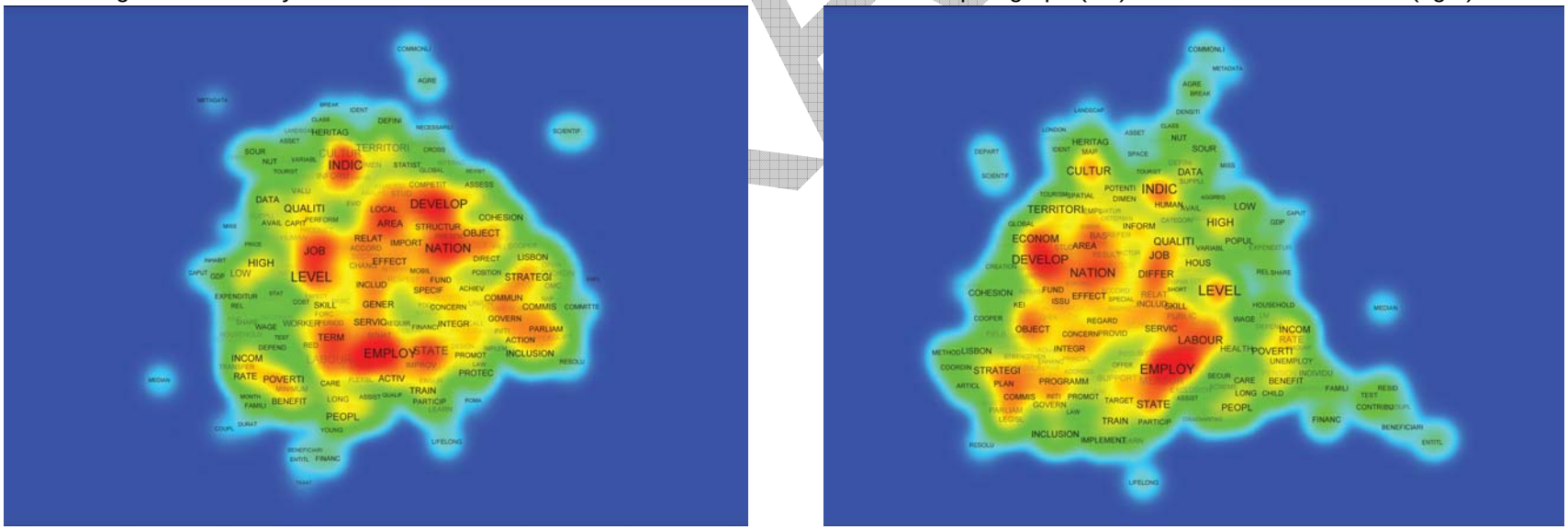


Figure 6b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘socio’, polici’, present in (6a), were added to a exclusion list in (6b). TF: Term Frequency.

Figure 6c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

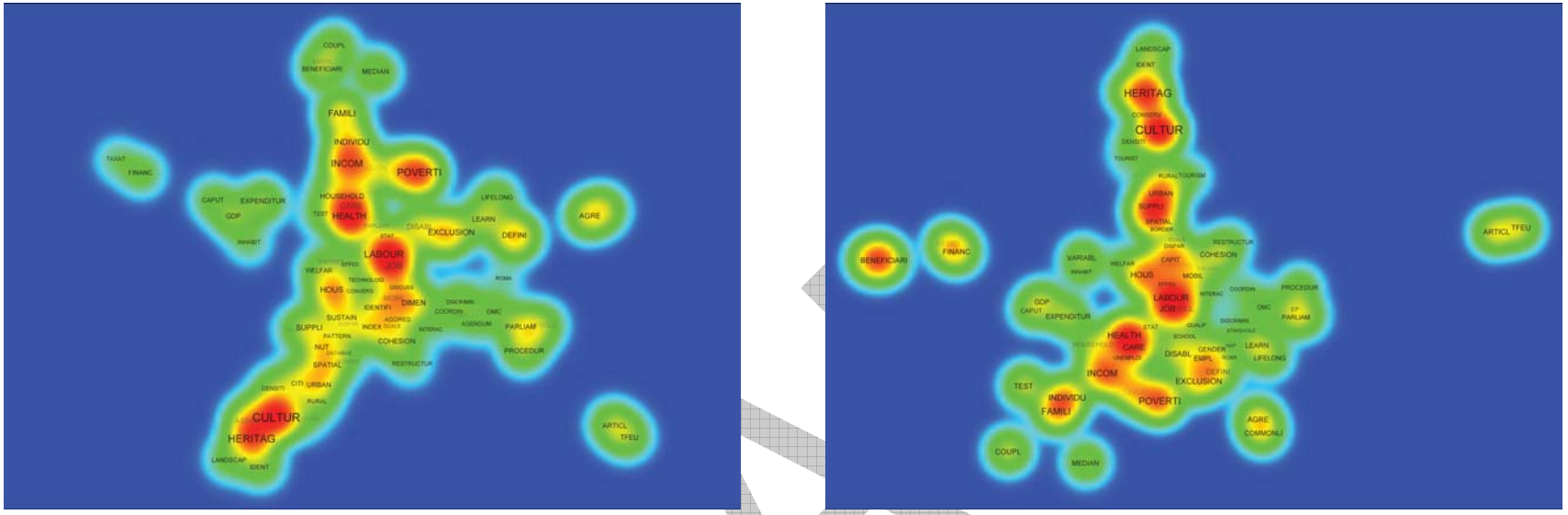
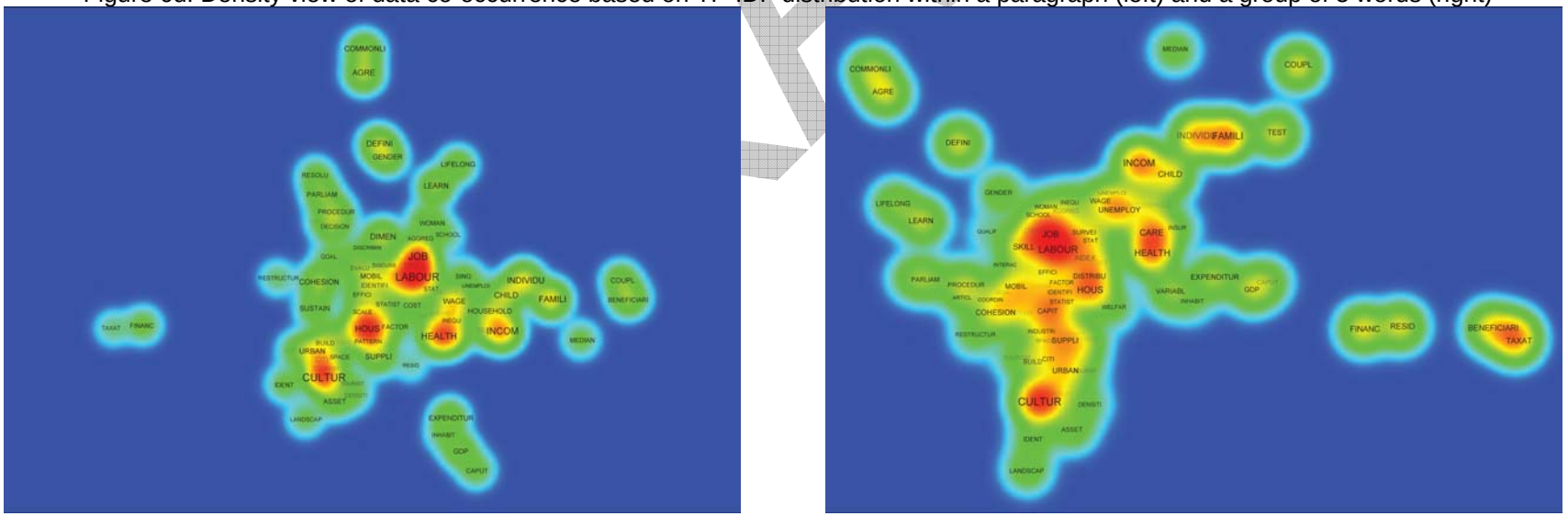


Figure 6d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)



Note: The words ‘heritag’, ‘empl’, ‘spatial’, pension’, ‘disabl’, and ‘poverti’, present in (6c), was added to a exclusion list in (6d)

Figure 7a: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a group of 5 words (right)

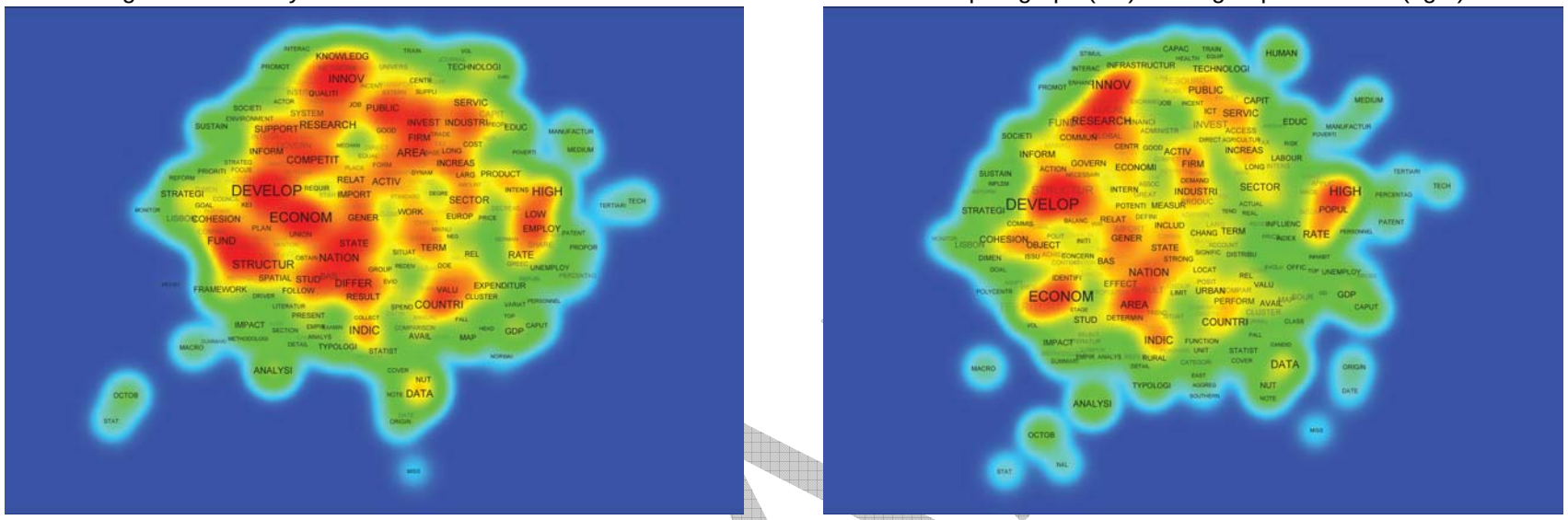
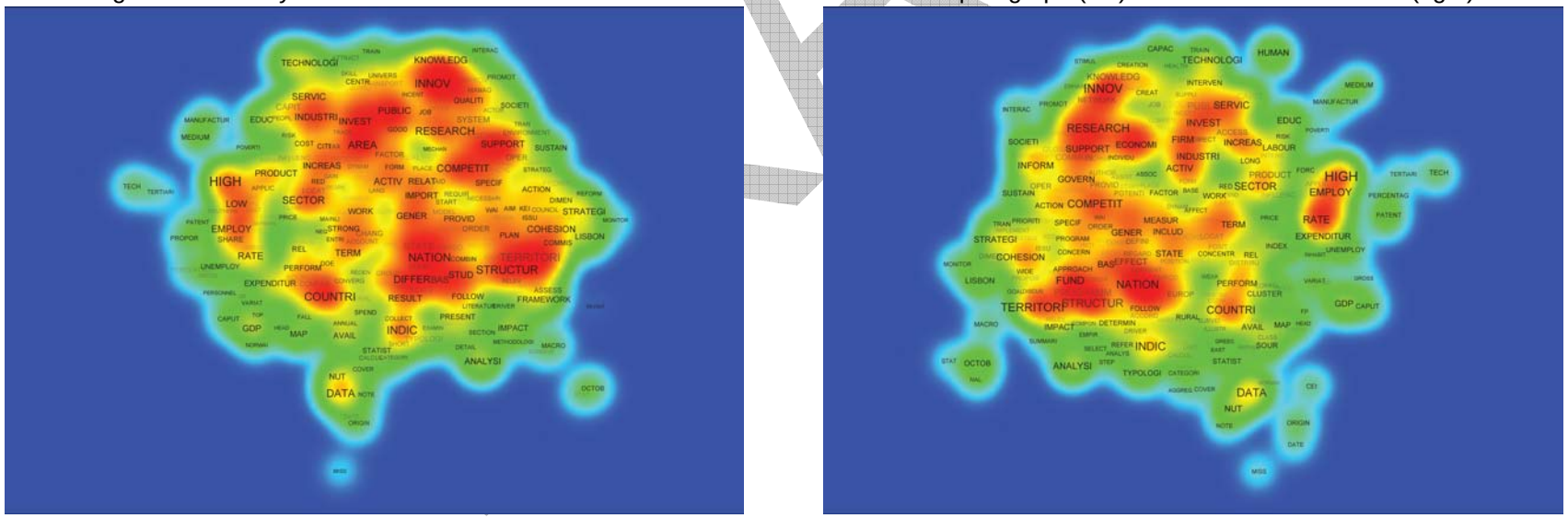


Figure 7b: Density view of data co-occurrence based on TF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘develop, ‘econom’, present in (7a), were added to a exclusion list in (7b). TF: Term Frquency.

Figure 7c: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a group of 5 words (right)

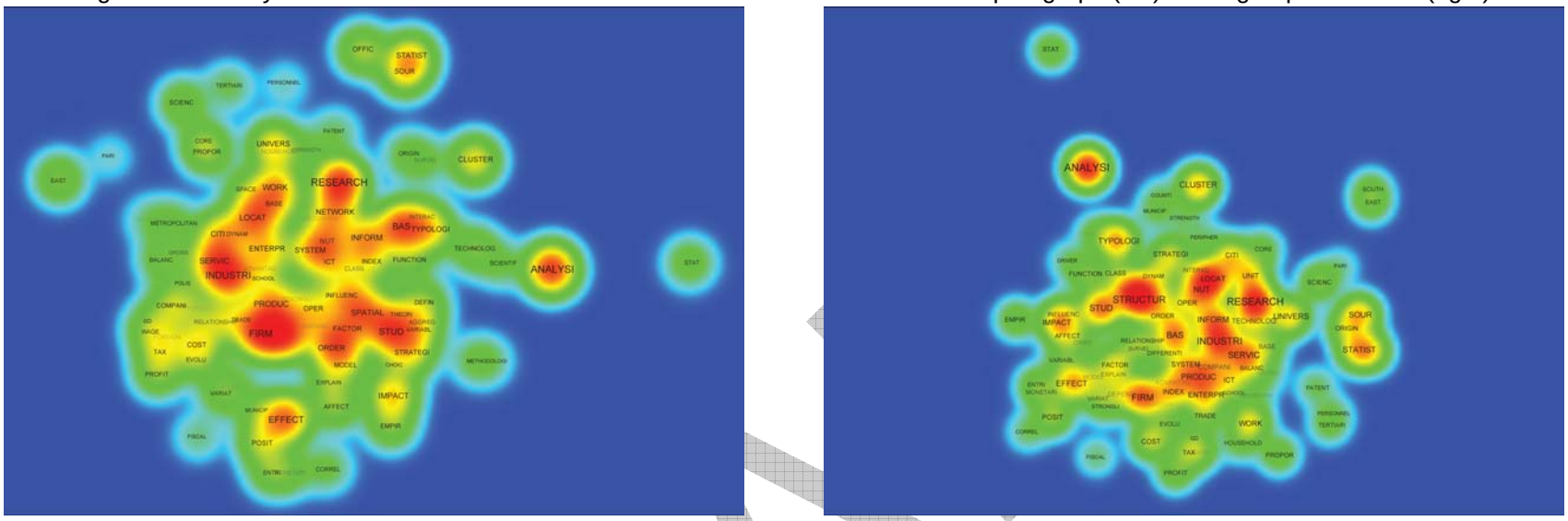
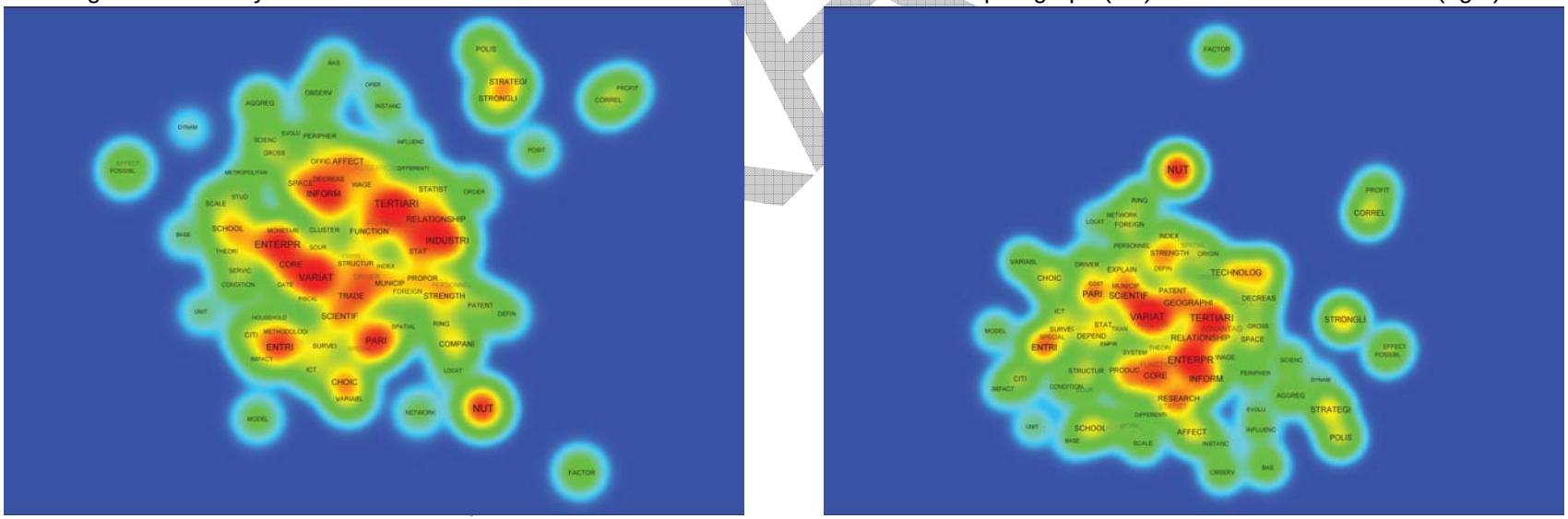


Figure 7d: Density view of data co-occurrence based on TF*IDF distribution within a paragraph (left) and a window of 5 words (right)



Note: The words ‘typologi’, ‘firm’, ‘tax’, ‘counti’, and ‘gdp’, present in (7c), were added to a exclusion list in (7d)

TECHNICAL REPORT

ESPON

The ESPON Database Application

CONTENTS

Thee ESPON 2013 Database Application is a complex information system dedicated to the management of statistical data about the European territory, spanning over a long period of time. The overall architecture relies on two databases: One is used for storing ontology data. The other, called the ESPON Database, is meant to be queried by end-users. The latter only is made accessible to users through Web interfaces that each correspond to the four main functionalities offered by the ESPON 2013 Database Application:

- Registration
- Administration
- Upload both data and metadata
- Query and retrieval such data and metadata.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

20 PAGES

LIST OF AUTHORS

Bogdan Moisuc, LIG Steamer
Jérôme Gensel, LIG Steamer
Anton Telechev, LIG Steamer
Benoit Le Rubrus, LIG Steamer

Contact

moisuc@imag.fr

jerome.gensel@imag.fr

anton.telechev@imag.fr

benoit.le-rubrus@imag.fr

.

LIG Grenoble

Bâtiment ENSIMAG D

681 rue de la Passerelle

38400 Saint Martin d'Hères

(+33) 4 76 82 72 11

TABLE OF CONTENTS

Introduction

This technical report describes the main screens and features of the web application interface version 1.0, also known as the ESPON 2013 Database web extraction tool.

The first section of the document gives a walkthrought of the application by following a a typical user session, from the authentication to the download of query results.

The second section of this document contains a description of thje databases upon which this application is based: the ontology base and the ESPON database.

1 The Web Interface

The ESPON DB Web download interface is an on-line application designed to offer fast browsing and searching capabilities over the ESPON DB. The Web download interface implements several innovative elements that guaranties scalable performance to accommodate the fast growing size of the ESPON DB :

- The use of a server-side application cache system allows the application to avoid querying the database for all browsing tasks excepting the advanced search. This insures fast data searching, whatever the database size.
- The use of an XML exchange format for the answer to queries allows decreasing the size of the data transfers between server and client.
- The use of AJAX techniques (Asynchronous JavaScript and XML) allows further decreasing the size of the traffic between the client (Web browser) and the server (ESPON Web site), by transferring only the parts of query that have changed (in XML) and redisplaying them accordingly on the client (using JavaScript). This allows for load balancing between client and server, as the task of building the presentation from the XML file is performed on the client.
- The dropdown lists used in the interface have been developped as new components in order to match the ESPON look&feel requirements.

The underlying subsections contain a description of the different pages that can be displayed by the application.

1.1 Login page

When typing the URL of the ESPON 2013 Database web application in his/her browser address bar, any user is first invited to choose between both following types of login:

- *anonymous* login;
- *registered account* login.

Figure 1: *The login page of the ESPON DB Web Application*

The differences between these two logins essentially consist in the features the user will be offered by the application. Though differences will be described as one goes along this section, roughly speaking, a registered user will be allowed to access the whole set of pages that can be delivered by the application; an anonymous user will be allowed to search and download results from only a subset of available data. Besides the choice for the kind of session, Figure 1 shows that the login page displays a link in order to ask for new login or password.

The login page of the application allows users to enter the application either in an anonymous way (restricted rights) or by typing their login and password, when they are registered.

On the login page there a link redirected the user towards a registration form. The same form can be accessed by clicking on the menu item "Register", which is only visible to unregistered users. Upon clicking on one of this links, the user arrives on the registration form page. This page is shown on figure 2.

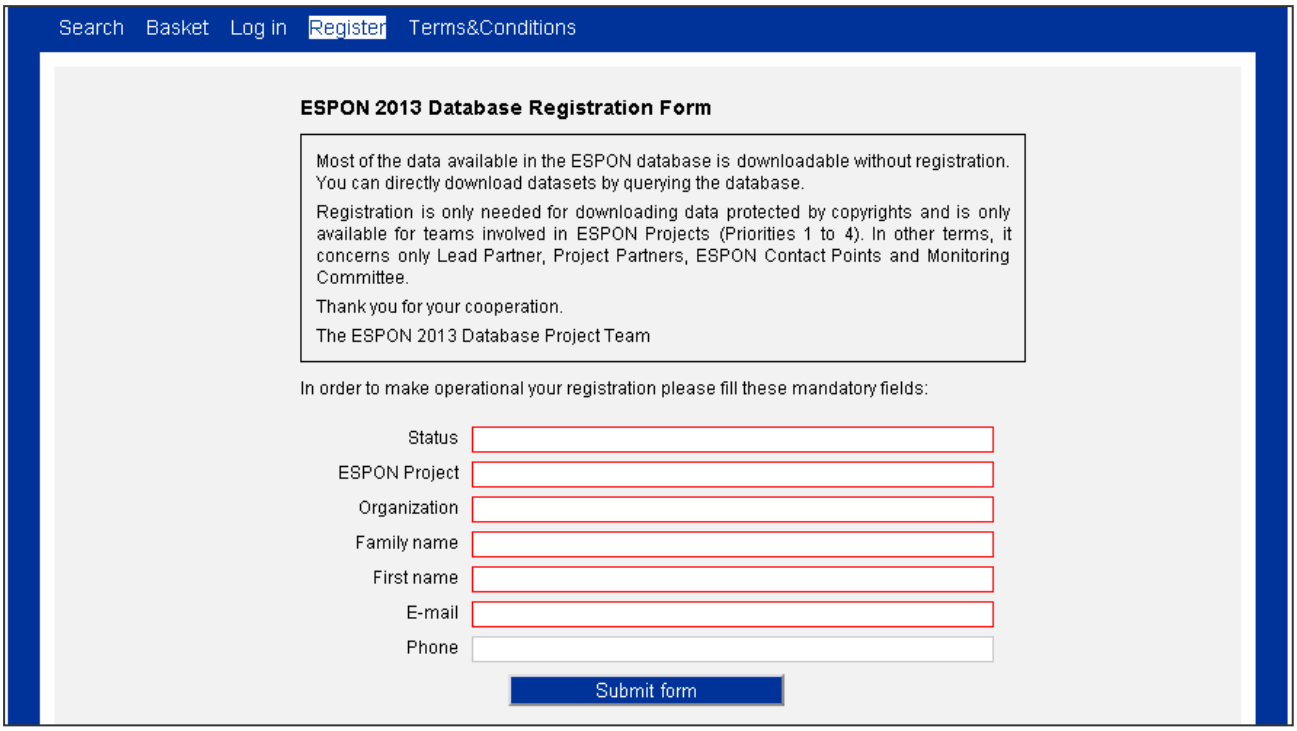


Figure 2: *The registration page of the ESPON DB Web Application*

The registration form contains a small number of fields required for establishing the identity of the user (first name and family name and organization) their contact coordinates (telephone number and email address) and their relationship to ESPON (the ESPON Project which they are members of and their function within this project).,

Depending on the kind of session, the authentication process redirects the user either on the search form page (anonymous session) or on a custom home page (for registered users only). Moreover, although the layout of pages is identical for both types of session, the displayed menu bar on the header of pages is different. Figure 3 shows the menu bar for an anonymous session, figure 4 shows the menu bar for a registered user. The main difference between both menus is the availability of the upload page for a registered user, this feature is described in section 3.

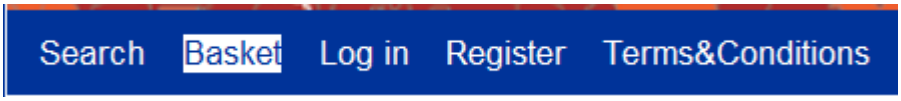


Figure 3 Menu bar for an anonymous session

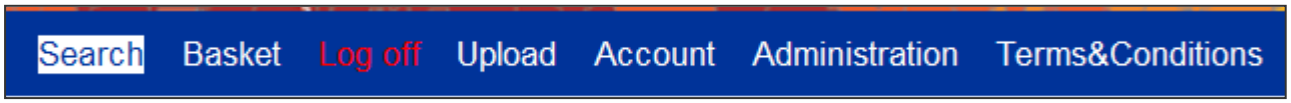


Figure 4 Menu bar for a registered user

For a registered user, the authentication drives him to the ESPON Database Web Interface Home Page. Figure 5 shows that this home page first displays the start date of the current session, then a list of hypertext links to ad-hoc

pages shows the available features of the web application:

- the home page;
- a page to update one's account;
- the search page;
- the basket page;
- the upload page;
- a link to quit the session;
- the on-line help.

Note that except the link to the “Update password” page, the links on the home page are also available from the menu bar.

Following sections aim at describing each of these pages. Common features to anonymous and registered users are described first.

1.2 Search page

Both following modes of research can currently be performed:

- basic mode;
- advanced mode.

As shown on figures 6 and 7, the user may switch from one mode to the other mode by clicking a link under the “Find” button. A “Basket” icon is also displayed near the form in order to easily visualize the number of saved results. Please consult section 3 for more details about the basket.

The basic mode currently contains mandatory fields for which the user must select at least one value before launching a query:

- a spatial dimension: at least one country must be selected in the study area group;
- a thematic dimension: at least one indicator must be selected.

Figure 5: *The search page*

The advanced mode provides optional search criteria. Thus, the user may select:

- one or several NUTS levels;
- one NUTS revision;
- a temporal dimension: one or several covered periods by the dataset, one or several dates of publication;
- a project dimension.

The study area group of criteria provides two selection boxes whose values respectively include:

- the groups of countries (for example, EU15);
- the list of countries that are concerned by available records in database.

Selected values in these selection boxes are dynamically bound: for instance, selecting one group of countries will automatically select the list of countries that belong to this group in the "Country" select box.

As previously mentioned, the query is performed only if the user has selected at least one country and one indicator. If this requirement is forgotten, clicking the "Find" button will return a warning message. Else, the application performs the query and the results are displayed in the initial form, as shown on figure 8.

The screenshot above shows that the query returned two results. Each result is displayed on one line. Considering this table of results, the displayed information for each result is respectively made of following columns:

1. First column (no header title): the icon is displayed if the result concerns data with a copyright status. This case implies both behaviours:

    - if the user is anonymous, the checkbox in the second column is disabled: though he/she can see this result and its metadata, he/she is not allowed to download it;
    - if the user has logged in under a registered account, the checkbox is enabled. Nevertheless, he/she is warned about about the confidential status of data for this result.

2. Second column (no header title): the checkbox allows to select the current result in order to perform two possible actions:

    - adding this result to the basket;
    - immediately downloading this result.

3. The basket is roughly speaking a temporary area where the user can save multiple search results while he/she performs various queries. This basket functionality is further described in section 3.
4. Third column (no header title): the name of the dataset which is concerned by this result item. Clicking this name displays further details for the different NUTS levels which are included in this dataset, if any.
5. Data type column: the current version of the application only manages NUTS, future versions may also include WUTS, GRID, UMZ data types for example.
6. Version: as only NUTS data type is currently managed, this field displays the NUTS revision.
7. Covered period: this field shows the temporal coverage of the current result.
8. Study area: this field shows the spatial coverage of the current result.
9. Last update: this field shows when this dataset was last updated.
10. Source: this field shows the provider of this result.
11. Completeness: the percentage of the global completeness of the current result is represented by a coloured bar. Long blue bar: high rate; short blue bar: low rate.

Metadata: on clicking the icon, the application opens a popup window where the user can consult further details about the metadata of this current result. Figure 9 shows that four different types of information are available in this metadata page:

- identification of the dataset;
- the list of indicators for this dataset;
- the lineage of this dataset;
- the map of Europe showing the rate of completeness of data for each country which is concerned by the dataset. Figure 9 describes this feature.

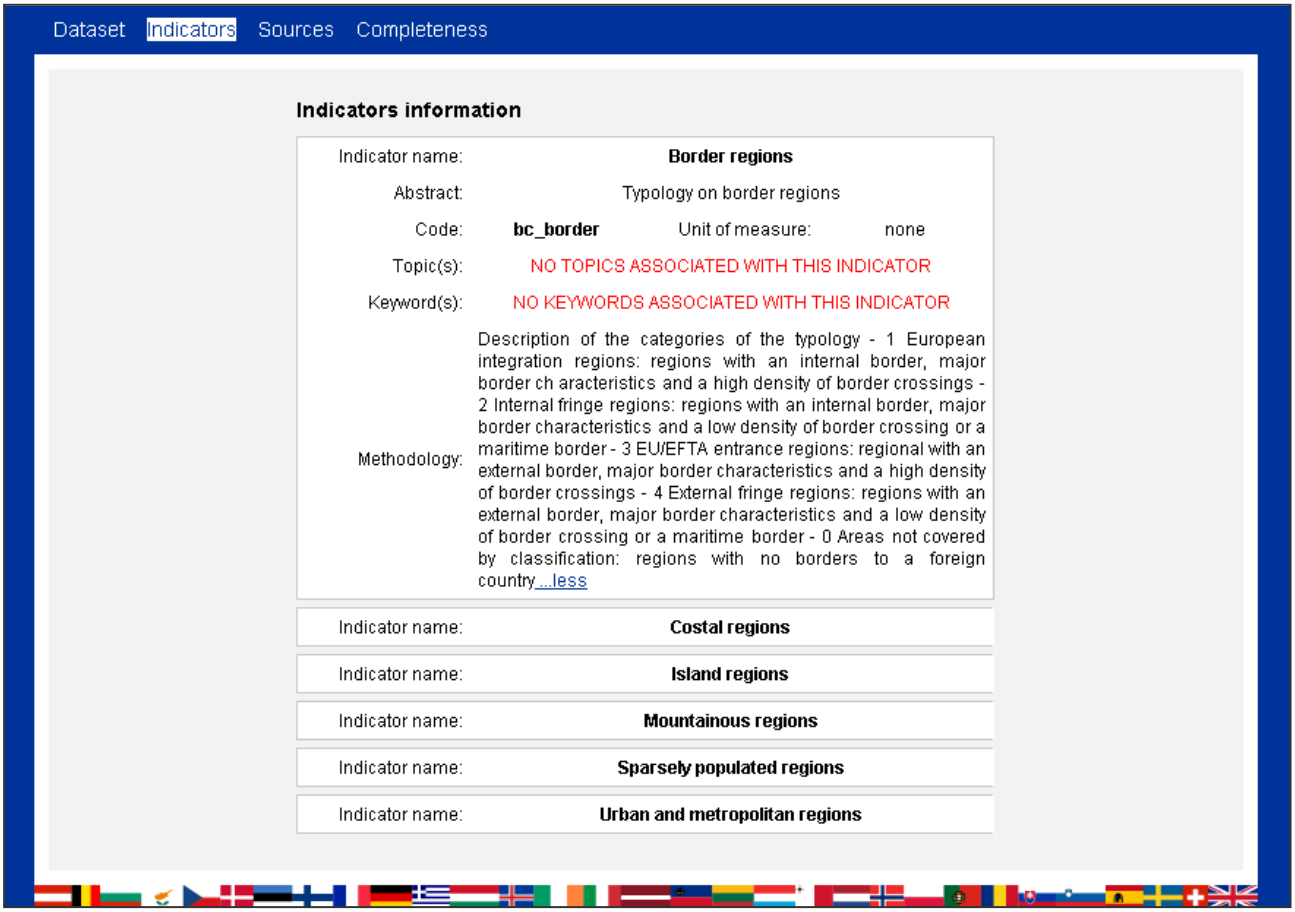


Figure 6: *The detailed metadata page*

Finally, from this search result table, the user can perform the two following actions on selected items (by checking checkboxes on each line):

- adding selected results to his/her basket;
- downloading selected items.

The "basket" feature allows to temporary save search result items. Thus, the user keeps concentrated on his/her queries, the basket allows him/her to drill his/her research process.
Nevertheless, the basket is not a mandatory gateway to download results: if he/she found the expected results, he/she may directly use the "Direct download" functionality. This "direct download" can be considered as a short-cut to the basket service, which is further described in the following section.

1.3 Basket page

Figure 10 shows the basket page: except the colour, the displayed information is quite similar to the search results table on the search page: indeed, the basket is filled of search result items. Once the user completed his/her research, he/she can now refine the selection of search results items that

he/she wants to download.

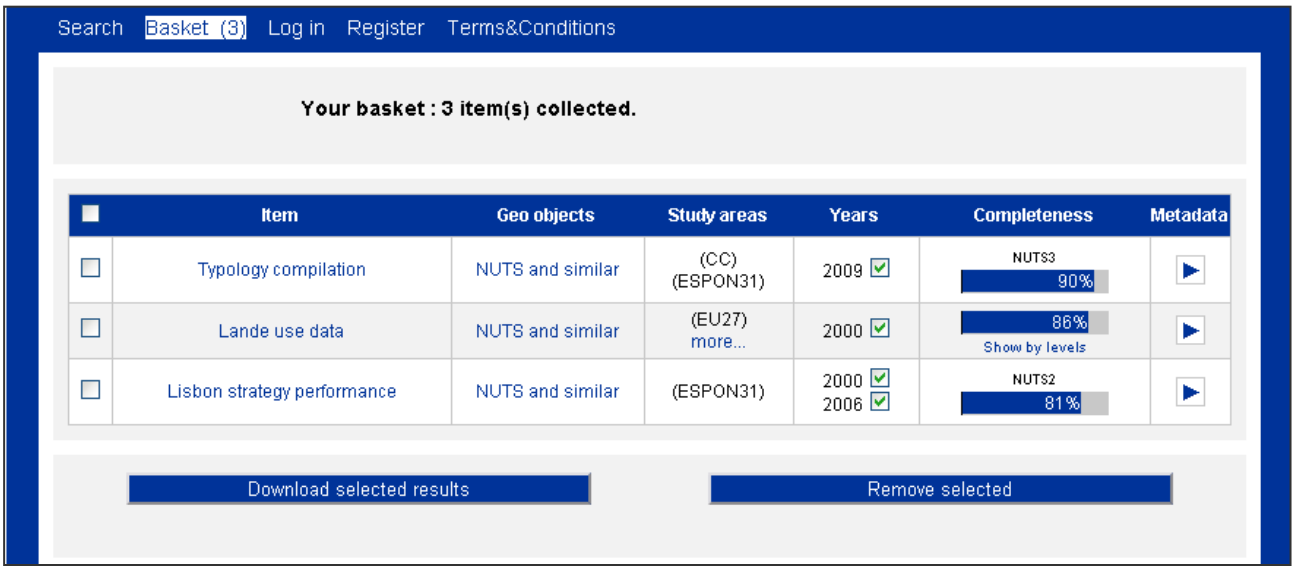


Figure 7: *The basket page*

Through the checkboxes on the left side of each line, the user is invited to select items on which he/she can perform both following actions:

- a deletion;
- a download.

The deletion consists in removing the selected items from the basket. Note that in such a case, the user is asked to confirm before processing.

The download action triggers the following set of tasks on the server-side:
a Microsoft Excel file is build for each selected item which is expected to be downloaded. This spreadsheet is composed of at least four sheets:

1. the first sheet is untitled "Dataset", it provides general information about the current dataset (name, contact, etc.);
2. the second sheet is untitled "Indicators", it provides metadata information about included indicators in this dataset: a description, the unit, the methodology, etc.;
3. the third sheet is untitled "Lineage", it provides metadata information about the lineage of the dataset: validity start, methodology, etc.
4. the fourth sheet displays the values of the dataset.

If the user selected several items to be downloaded, a zip archive is build, gathering the build xls files (one per item) by the previous step.
The application finally returns the build file to the user as an attachment (an xls file or a zip archive, depending on the number of selected items).

On the client-side, the user is finally invited by his/her browser to open or to save the build file (see figure 11) on his/her disk.
In the case of one selected item, the proposed filename for the downloadable

file will be **ESPON_Database_SearchResult.xls**. In the case of several selected items to be downloaded, the default filename for the proposed downloadable file will be **ESPON_Database_SearchResults.zip**.
Once retrieved on his/her disk, the content of the zip archive can be extracted with a standard tool, the included files are simply named **ESPON_Database_SearchResult_1.xls**, **ESPON_Database_SearchResult_2.xls**, etc (as many files as there are selected items).

Caution: note that the basket is unfilled when the user logs out. This is the reason why the user is asked to confirm his/her wish when he/she clicks the "Log out" button on the menu bar.

1.4 Upload page (registered users only)

The upload page aims at contributing to fill the database. Indeed, one key principle of the ESPON 2013 Programme Strategy is to "*improve the European knowledge base on territorial development and cohesion, including data, indicators, typologies, models and maps".* Thus, the upload page is a useful step in this strategy, as it provides the opportunity for registered users to transfer data files, metadata files and other documents from their disks to the ESPON server.

Figure 12 shows the threefold available input fields to provide this service:

- the data input field: clicking the "Browse" button, the user is invited to select a Microsoft Excel file on his/her disk. Caution: as a data file, an xls file is currently required.
- the metadata input field: clicking the associated "Browse" button, the user can select either an xml file or an xls file on his/her disk. Caution: only xls or xml files formats are accepted.
- the additional (optional) documents input field: any further electronic documents may be transferred to the server thanks to this field. Caution: only a zip archive file is currently supported here, even for a single additional document. Please compress and gather your add(s).

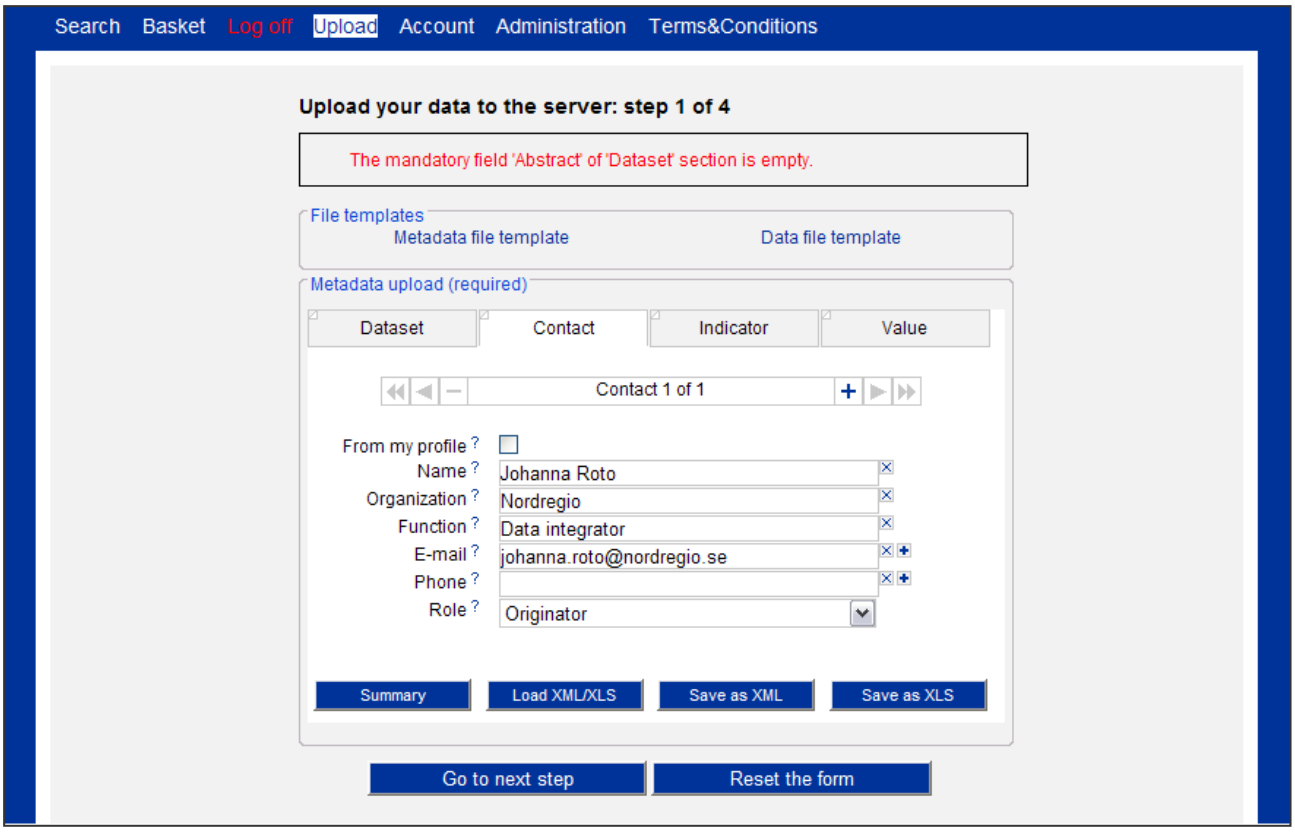


Figure 8: *The upload page, step 1*

Both data and metadata files whose expected formats were previously described are required to successfully achieve an upload.

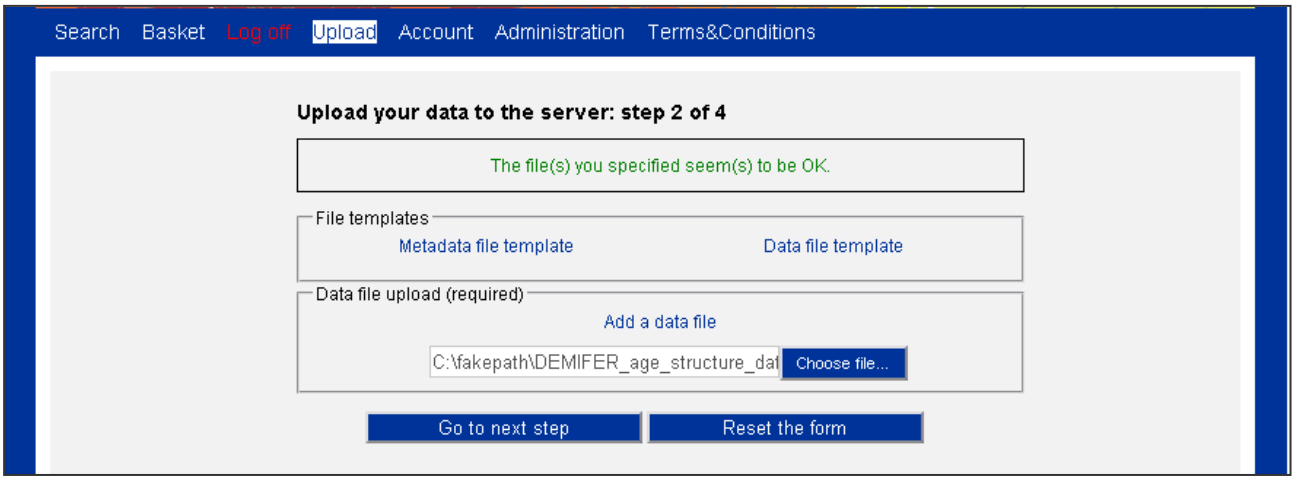


Figure 9: *The upload page, step 2*

Note that the uploaded files are obviously not directly integrated to the ESPON Database, this upload step has to be followed by several processes: verification, harmonization, etc.

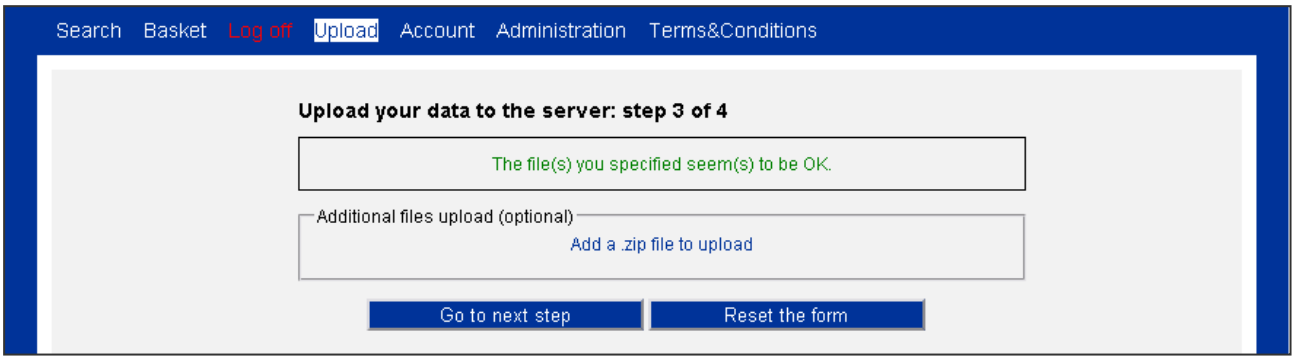


Figure 10: *The upload page, step 3*

In the case when the user does not own a metadata file yet, this upload page provides a link to an ESPON specific metadata editor (via the icon). Please consult the technical report untitled **Technical_Report_metadata.doc** for further information about this customized Geonetwork Opensource tool.

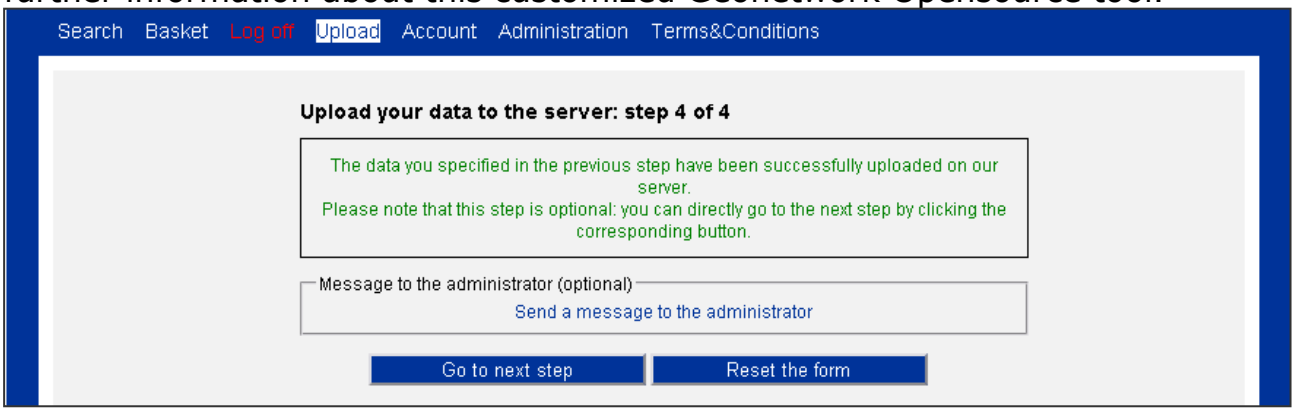


Figure 11: *The upload page, step 4*

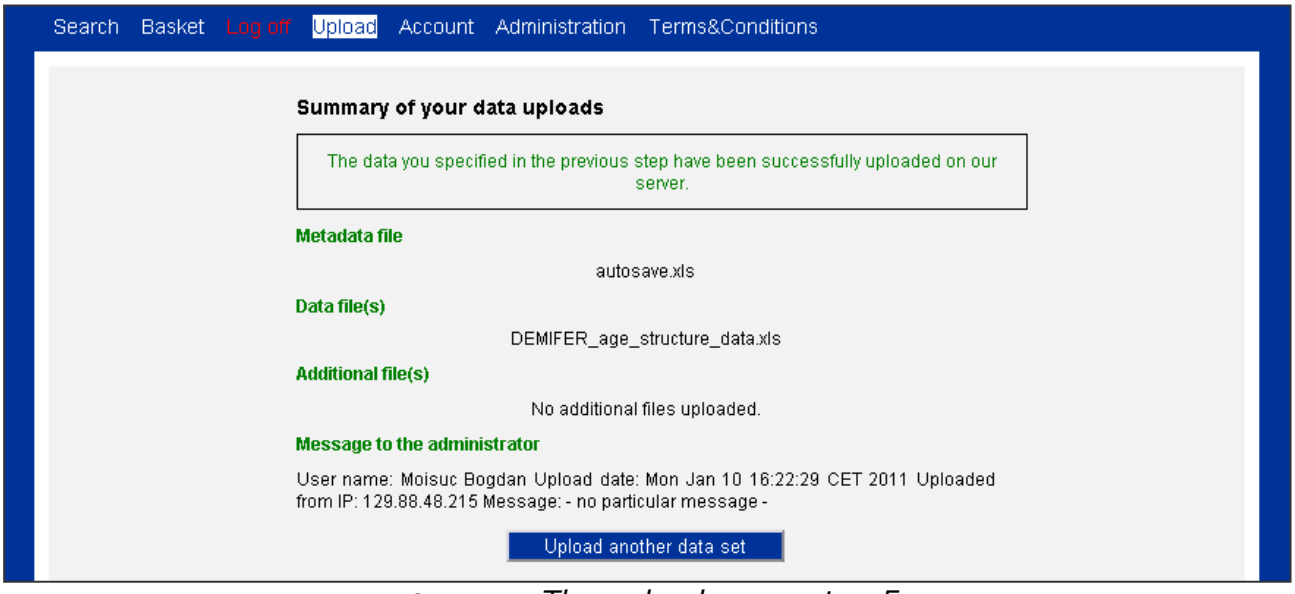


Figure 12: *The upload page, step 5*

1.5 Profile page (registered users only)

The profile page allows users to update and complete their personal information and to change their password for accessing the application. In order to update any profile element, the users are required to enter their password.

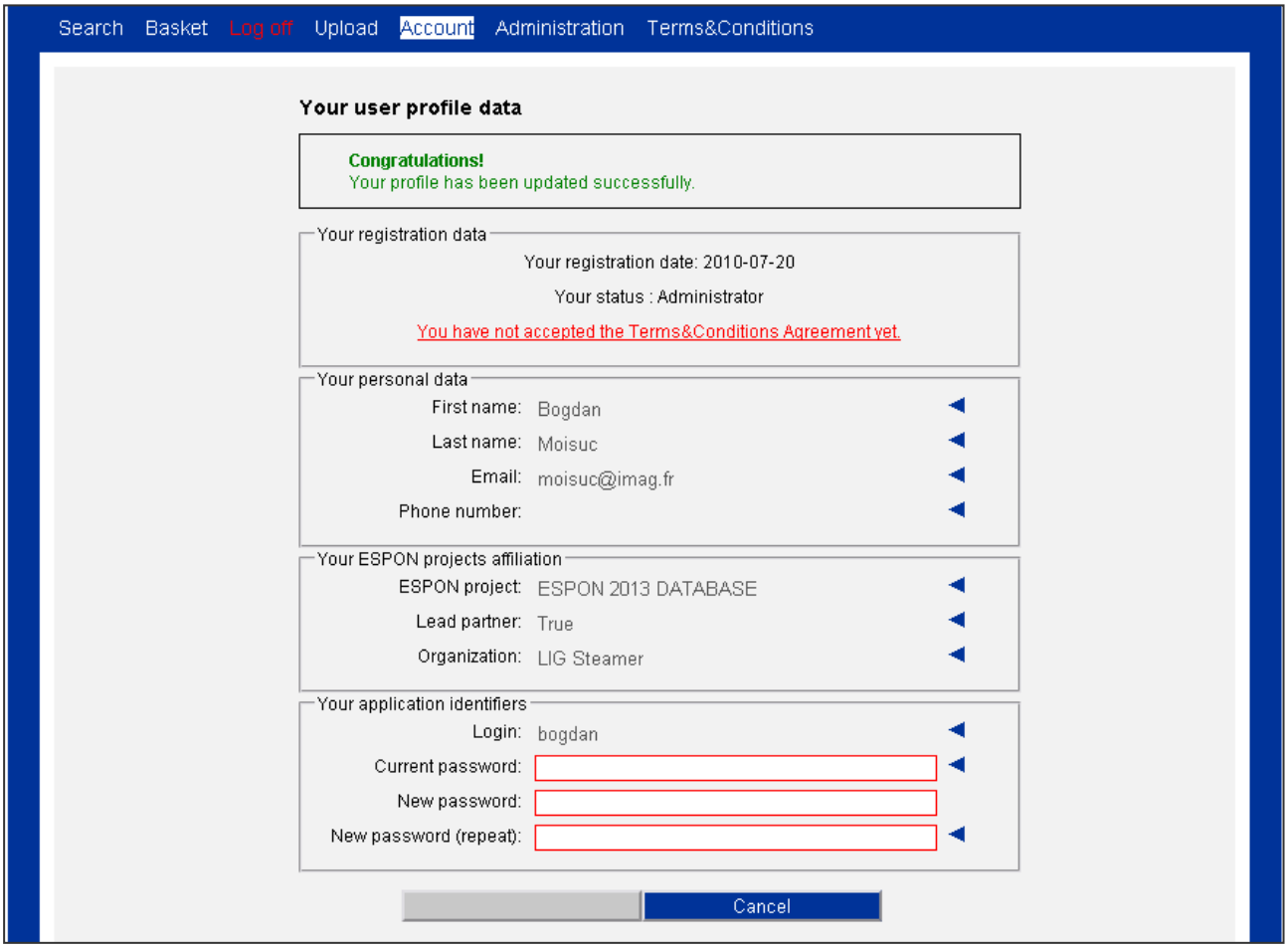


Figure 13: *The profile page*

2 The Databases

THIS PART WILL BE DEVELOPED FOR FEBRUARY 2011

2.1 The ontology database

2.2 The Espon database

TECHNICAL REPORT

ESPON

UPDATE OF ESPON DATABASE 2006 INTO THE 2013 VERSION

MAIN RESULTS

- Two ESPON Databases can be considered within the previous programme. An internal, used by ESPON projects and an external available on ESPON Website which synthesizes and presents as a clear way ESPON 2001-2006 results.
- The update of the ESPON Database available on Web appears more logical.
- Thanks to the new data and metadata model developed in the ESPON DB Project, the level of description of each indicator is improved.
- Expected integration in the ESPON 2013 DB: December 2010

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

27 PAGES

LIST OF AUTHORS

Ronan Ysebaert, UMS RIATE, Paris, France

Bogdan Moisuc, LIG, Grenoble, France

Contact

(1) ronan.ysebaert@ums-riate.fr
(2) Bogdan.Moisuc@imag.fr

(1) Tel. + 33 1 57 27 65 32

(2) Tel. + 33 4 76 82 72 25

TABLE OF CONTENT

Introduction

One of the main objective of challenge 1 – collection of basic regional data – is to " *be able in a short delay to connect the new information elaborated by ESPON 2013 Program with former datasets elaborated by ESPON 2006 Program*" (First Interim Report of ESPON 2013 Database Project). It is a clear fact that ensuring the continuity between the different ESPON databases is a crucial bullet point for the development of the ESPON Program. In other words, it raises the question of the integration and implementation of indicators produced by ESPON 2006 Program into the new version of the database.

However this work is not obvious at all. Some questions have to be answered before making any update: What information is contented in the former ESPON Database? Is all information interesting? If not, which indicators have to be integrated? How making possible the concordance between the two different metadata template? … and so on.

The aim of this paper is to try to establish some milestones in this way. In a first section, we will make a global overview of the previous ESPON Database. Considering this expertise, we will propose a strategy for the update of indicators. Finally, we will discuss about the degree of precision of the update considering the possibilities which exist with the new metadata template.

1 Overview of the ESPON 2006 Database

In concrete terms, ESPON 2006 Database is developed in two parts: an "internal" database, dedicated to ESPON Projects during the ESPON Program and an "external" one where people can download datasets with a limited number of constraints.

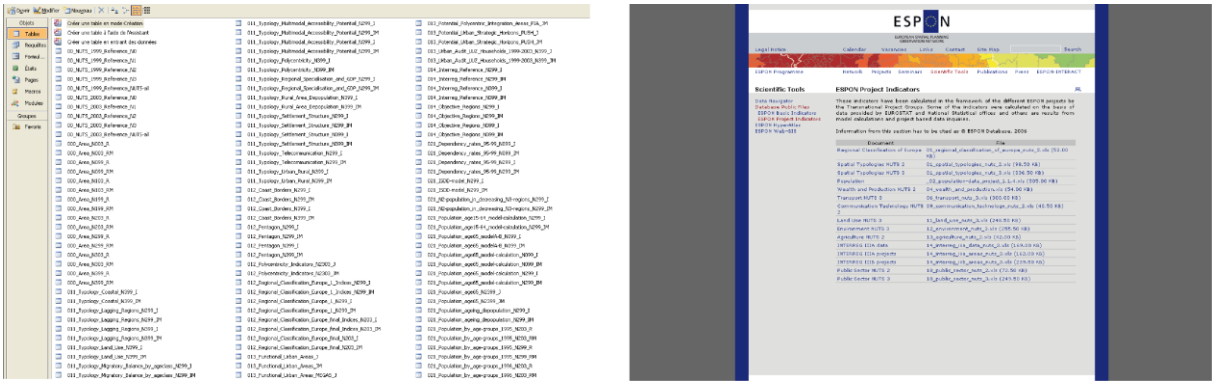


Figure 1 – The "Internal" (on left) and "external" (on right) ESPON 2006 Databases

1.1 The "internal" ESPON 2006 Database

UMS RIATE analysed the content of the internal ESPON 2006 Database to have a global overview of the information available within ESPON in June 2006[1]. The geographical coverage of the ESPON 2006 Database contains the ESPON area during the period of the program, that is to say the current ESPON area minus Iceland and Liechtenstein.

In June 2006, ESPON internal Database contains more than 4350 indicators[2] grouped into 361 tables structured **by themes** (figure 2). Regarding the structuring of the database, population and employment/labour market themes are over represented with 1036 and 1458 indicators. Reversely, some themes contain few indicators, such as tourism or communication (less than 20 indicators)

[1] It depicts the situation in June 2006. Some indicators has been added since this period (last version = April 2007). However, we can consider that these modifications do not change considerably the structure of this internal database.

[2] We consider that an indicator is a column of data in a table. For instance, if GDP in euros is described for NUTS0, NUTS1, NUTS2 and NUTS3 we consider that there is 4 indicators.

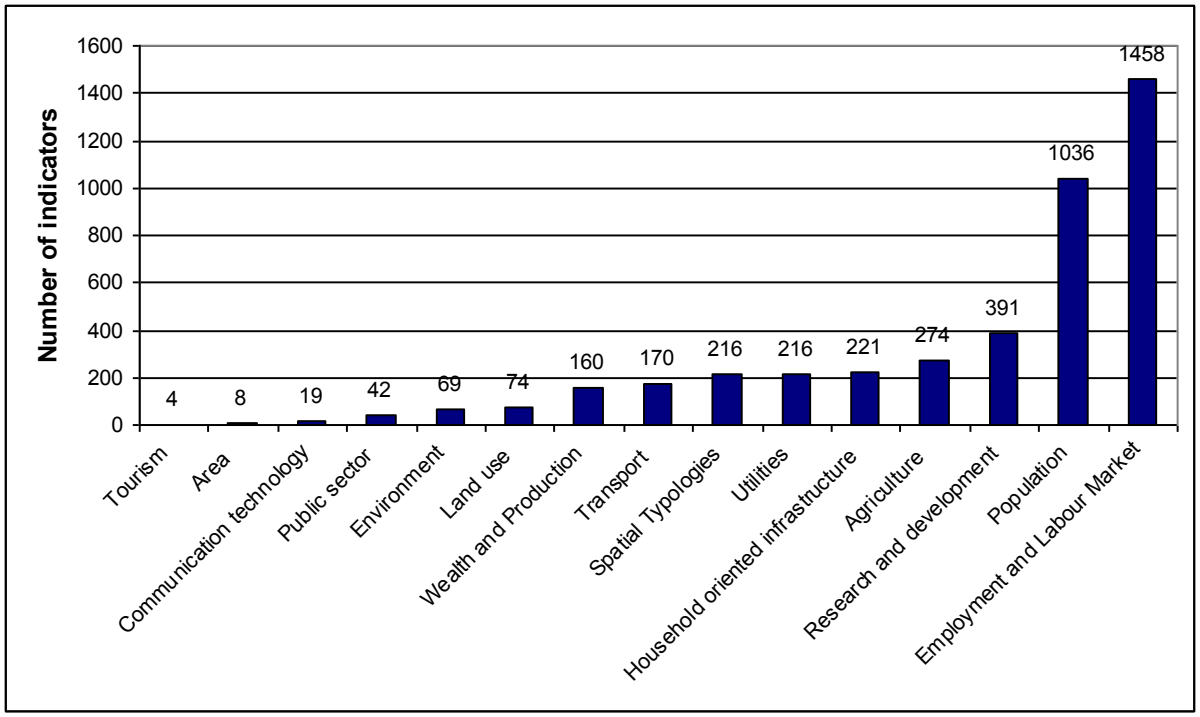


Figure 2 – Number of indicators by themes into the internal ESPON 2006 Database

The analysis of the quantity of information by **type of NUTS** (figure 3) reveals that most of the indicators have been collected in NUTS 2 (more than 3000 indicators, 70 % of the total) and in the NUTS 1999 version (3200 indicators, 74 % of the total.

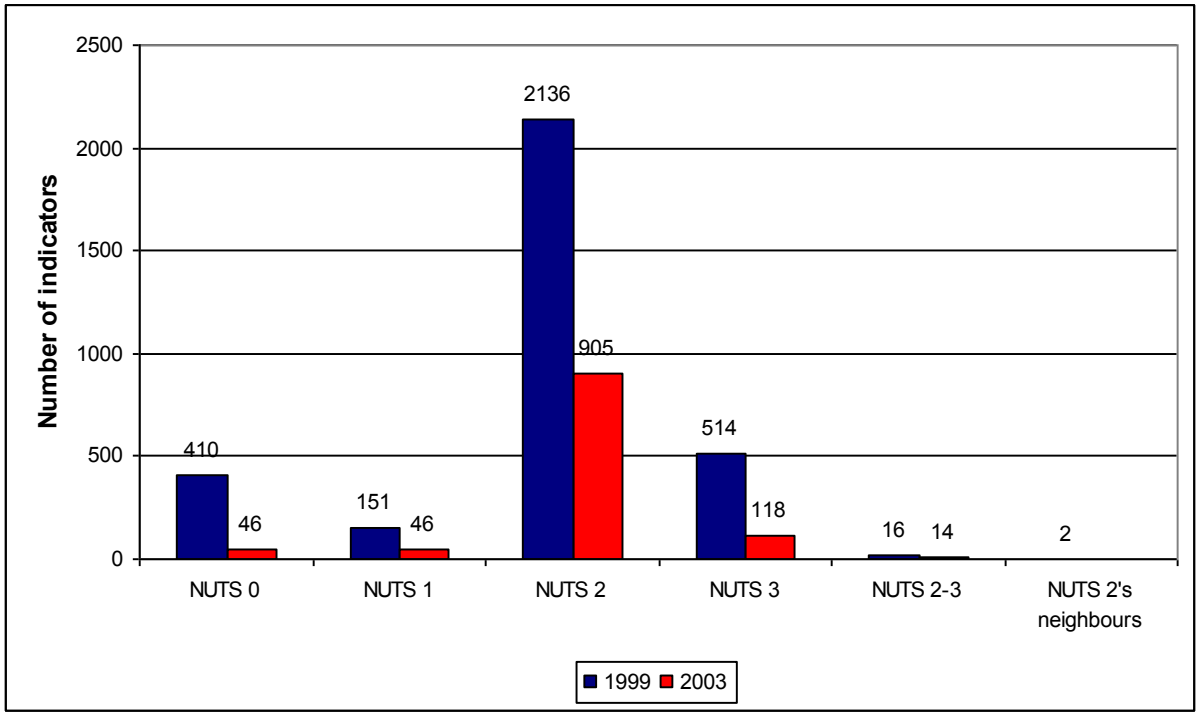


Figure 3 – Number of indicators by type of NUTS into the internal ESPON 2006 Database

Most of the indicators contained in the ESPON 2006 Database have a **period of reference** comprised from 1999 to 2003, 60 % of the total (figure 4).The Database proposed also some indicators described by an evolution. It is possible to have data from 1974 but the degree of completeness is small.

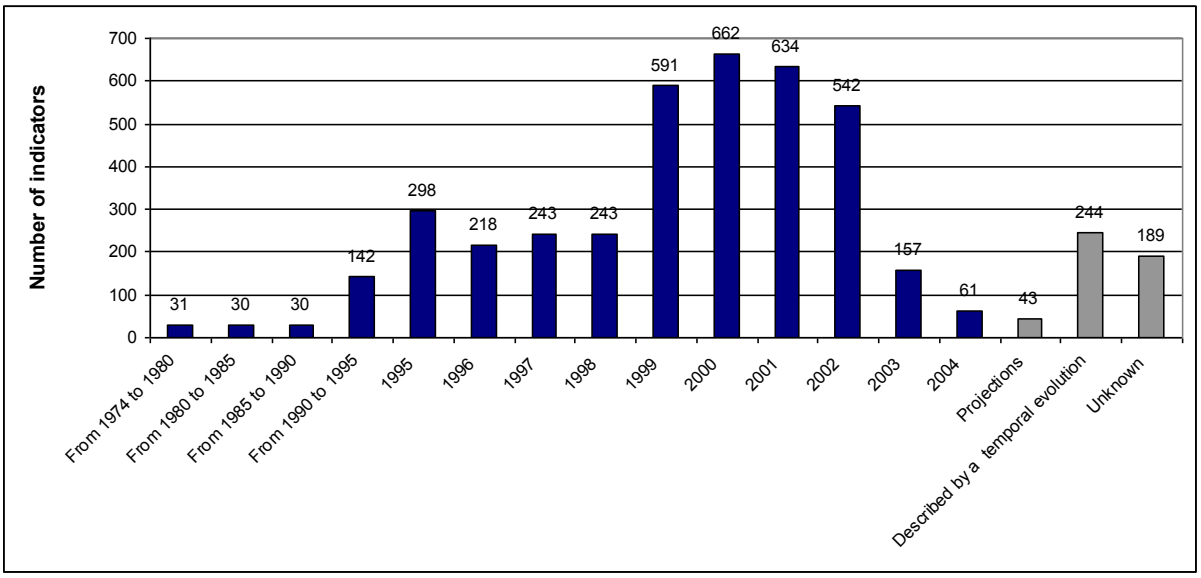


Figure 4 – Period of reference of indicators into the internal ESPON 2006 Database

Considering the **origin of the data**, 78 % of the indicators have a data source coming from Eurostat (Eurostat Regio mainly). The other indicators come from ESPON projects directly (typologies, calculation of innovative indicators) and other data centers, like the European Environmental Agency.

The degree of completeness of values in the different indicators of the database is very heterogeneous (figure 5). The main problem raised by researchers during the last ESPON program considering the database is that it was impossible to have an overview of the quality of information before downloading the dataset. Indeed, most of the indicators are described by a very poor completeness. Around 50 % of the indicators available can be characterised by too important missing values – no data, very poor or poor quality – to be used in the all ESPON area.
However, some indicators present a good degree of completeness. But it represents a minor part of the database (around 40 %) It is particularly the case for indicators described in NUTS0. In NUTS2 and NUTS3 the thematic field where this quality is the most important is wealth and production and population. In NUTS 2 or NUTS 3, most of the datasets are described by missing values in French overseas territories, Canaries, Madeira or Ceuta and Mellila.

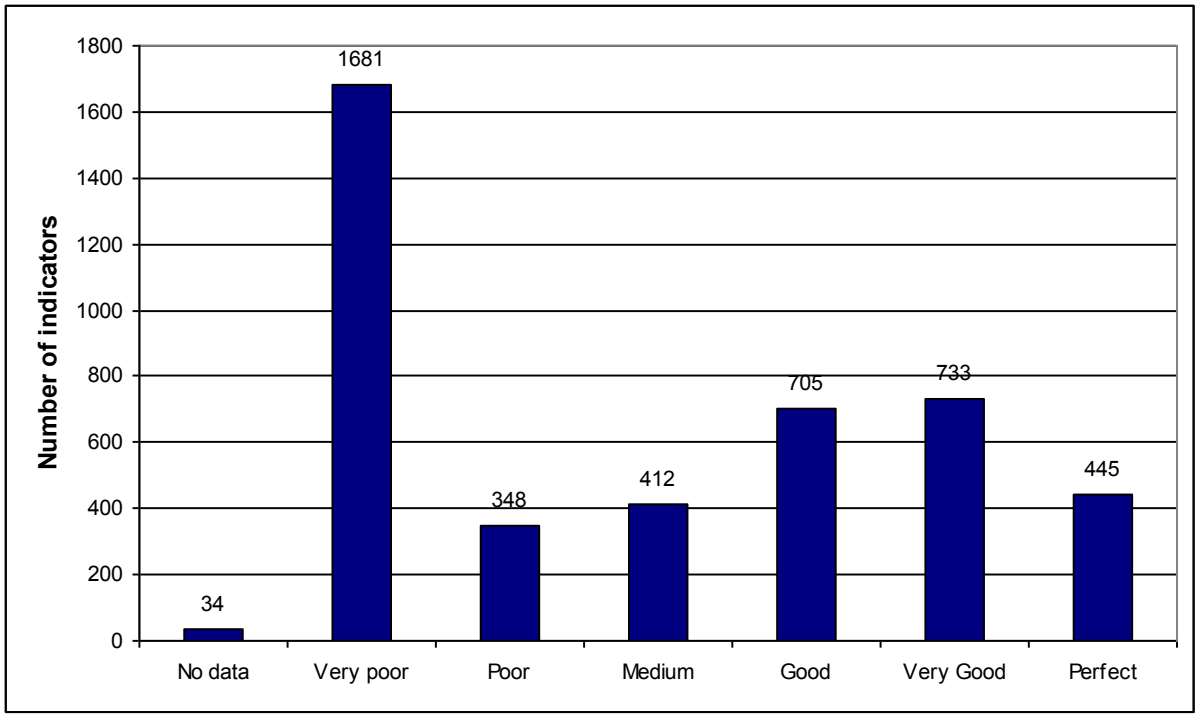


Figure 5 – Degree of completeness[3] of indicators into the internal ESPON 2006 Database

1.2 The "external" ESPON 2006 Database

The so called "external" ESPON 2006 Database is available on ESPON website[4] and is divided in two parts: "Basic indicators" and "project indicators. Basically the main difference between these two fields is that "Basic indicators" coming mainly from Eurostat (excepted for Switzerland and Norway where data come from National Statistical Offices) and has being made in form by ESPON community; on the contrary "project indicators" can be considered as an ESPON Projects production (new indicators, synthetic typologies and so on.

When considering the content of the database by **thematic field** (figure 6), we can notice that population concentrate the most part of the information (80 indicators coming from ESPON, 21 coming from Eurostat). But ESPON makes available a lot of synthetic typologies (47) which summarizes the scientific support of ESPON program to European statistical information. It shows also the synthesis work that has been done during the previous ESPON program.

[3] This classification is a result of a typology: "Perfect" means that 100 % of the data are available; "very good" means that there is a missing values for 1 country; "good" for 2-3 countries; medium for 4-5 countries; "poor" for 6-7 countries and "very poor" for more than 8 countries. The aim of this typology is not to be objective at all (specific cases appear regularly) but to give a global overview of the quality of information available on ESPON 2006 Database.

[4] *http://www.espon.eu/mmp/online/website/content/tools/832/index_EN.html*

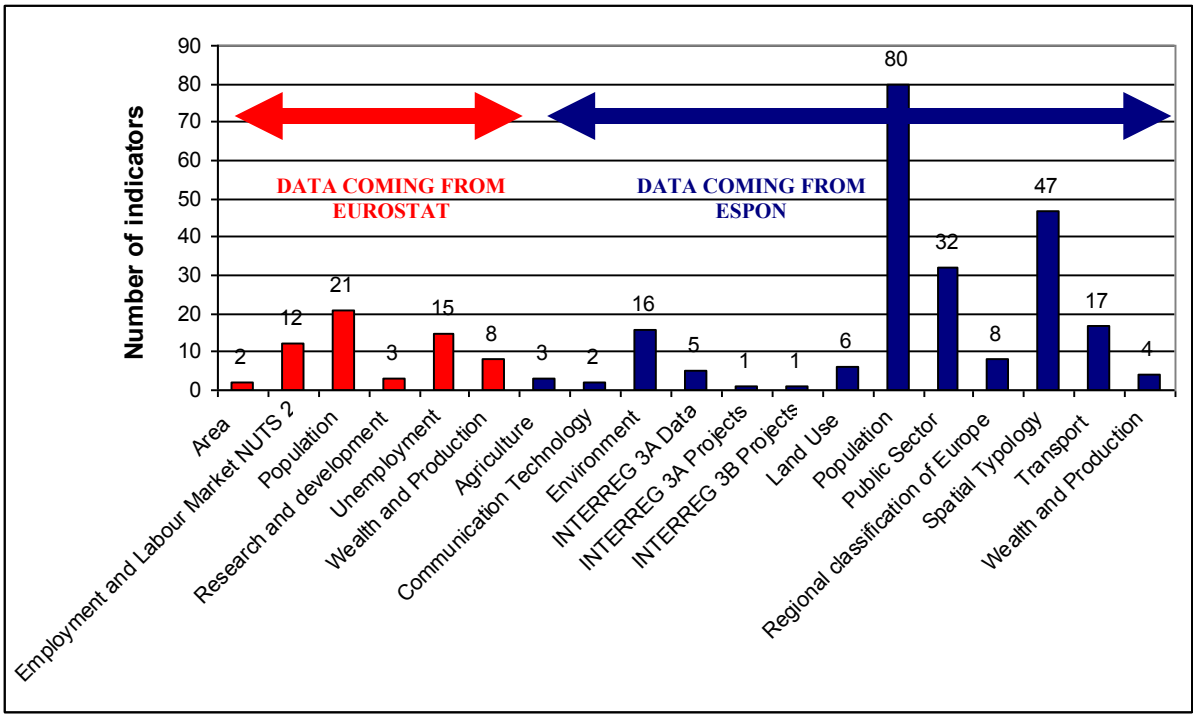


Figure 6 – Number of indicators by themes into the external ESPON 2006 Database

The geographical scope of ESPON 2006 Database available on line is the **regional one**. It contains data for NUTS 2 or NUTS 3 level (or a mix of NUTS 1-2-3). Most indicators are described in NUTS 1999 version and the main geographical level of analysis is NUTS 2.

	NUTS VERSION	
	NUTS 1999	NUTS 2003
NUTS2	126	33
NUTS3	89	23
Mix	12	0

Table 1 – Number of indicators by type of NUTS into the external ESPON 2006 Database

The degree of completeness of this database is great since most of the indicators are above 90 % (figure 7).

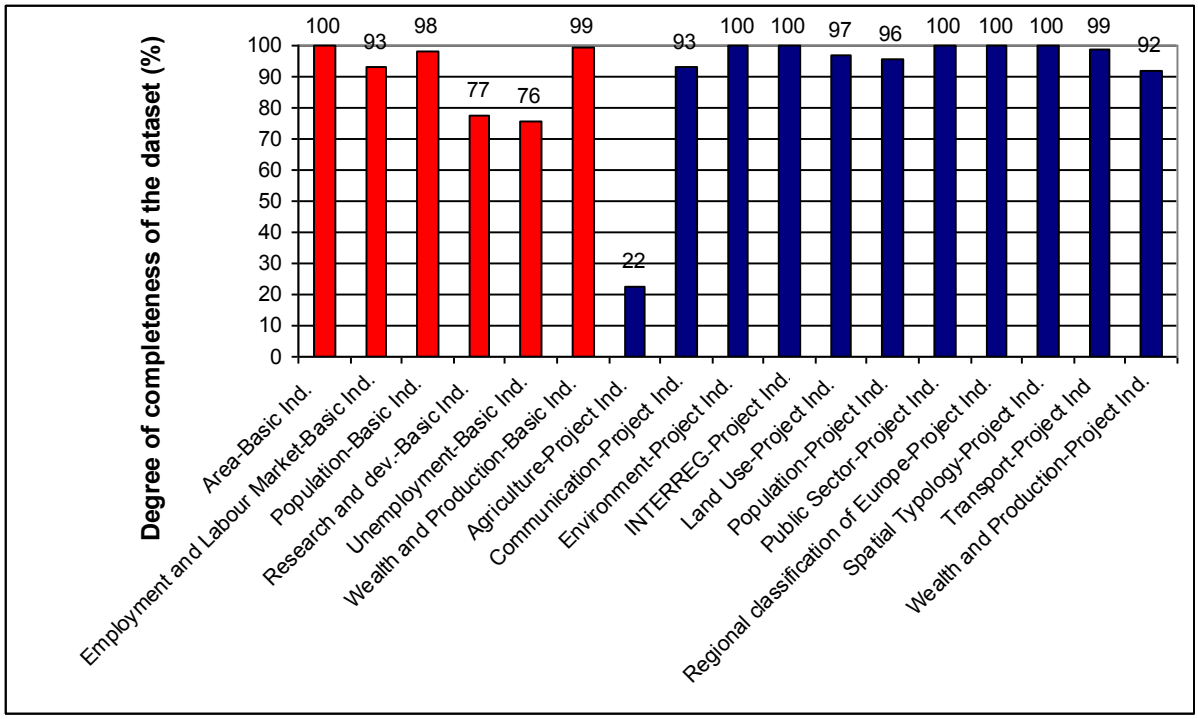


Figure 7 – Degree of completeness of the datasets of the external ESPON 2006 database by themes

2 What strategy for updating the ESPON 2006 Database?

2.1 Trying to integrate the external ESPON 2006 Database (available on the ESPON Website)

ESPON 2013 Database project proposes to focus the work of integration of indicators from the previous program into the external ESPON 2006 Database. Regarding to the part 1 of the paper, different reasons explain this choice:

- The **degree of completeness** of the datasets is significantly higher in the external than in the internal one. It does not make sense to update indicators which are characterised by 80 % of missing values.

- The fact that indicators from ESPON projects have been made available to the external world shows that the reflection needeen for choosing the **most pertinent** indicators produced by the program has already been done.

- It is impossible to integrate in a short term and in a good way more than **4 000 indicators** (internal ESPON Database) and continue to integrate data in a new database structure. The risk is high to make confusions.

2.2 Integrate both Eurostat indicators and ESPON indicators

As shown in the previous section, the origin of the data of ESPON 2006 Database is heterogeneous. However, data come mainly from Eurostat and ESPON projects. ESPON 2013 Database project proposes to integrate both indicators developed by ESPON Projects (called "ESPON Project indicators" in ESPON website) and collected from Eurostat website (called "basic indicators" in ESPON website).

The integration of selected indicators from ESPON 2006 Projects (ESPON Project indicators) does not raise any questions for us: they are the result of innovative methodology using knowledge in statistics and spatial analysis (spatial typologies, regional classification of Europe, estimations of missing values). This data has consequently have to be integrated in the ESPON Database in that form.

However, the question is more open for data coming from Eurostat. We have decided to store this information taking into account that it is not possible to obtain historical data on the Eurostat website. Eurostat proposes only data in the current NUTS version (e.g. NUTS 2006 version). It is possible to download this historical data by using the NewCronos Database (the historical database from Eurostat), which is not obvious for basic end-users (it requires using sql queries…). But it is important to keep in mind that indicators proposed by Eurostat are regularly revised. As an example we have compared the difference of values between two sources for a same indicator (total population in 2003): ESPON 2006 Database (origin of data: Eurostat, 2005) and Eurostat in 2009 (figure 8). If the data are the same for most of the NUTS2, 25 % of the values of total population in 2003 have been re-estimated (France, Slovakia, United Kingdom) since 2005. Indeed, considering that Eurostat revised regularly its basic indicators (population, employment, production), it is better to update these indicators by using directly official sources (national statistical offices, Eurostat). It implies that some values stored in the ESPON 2013 Database and coming from Eurostat are not necessary the same as the ones that are downloadable at present.

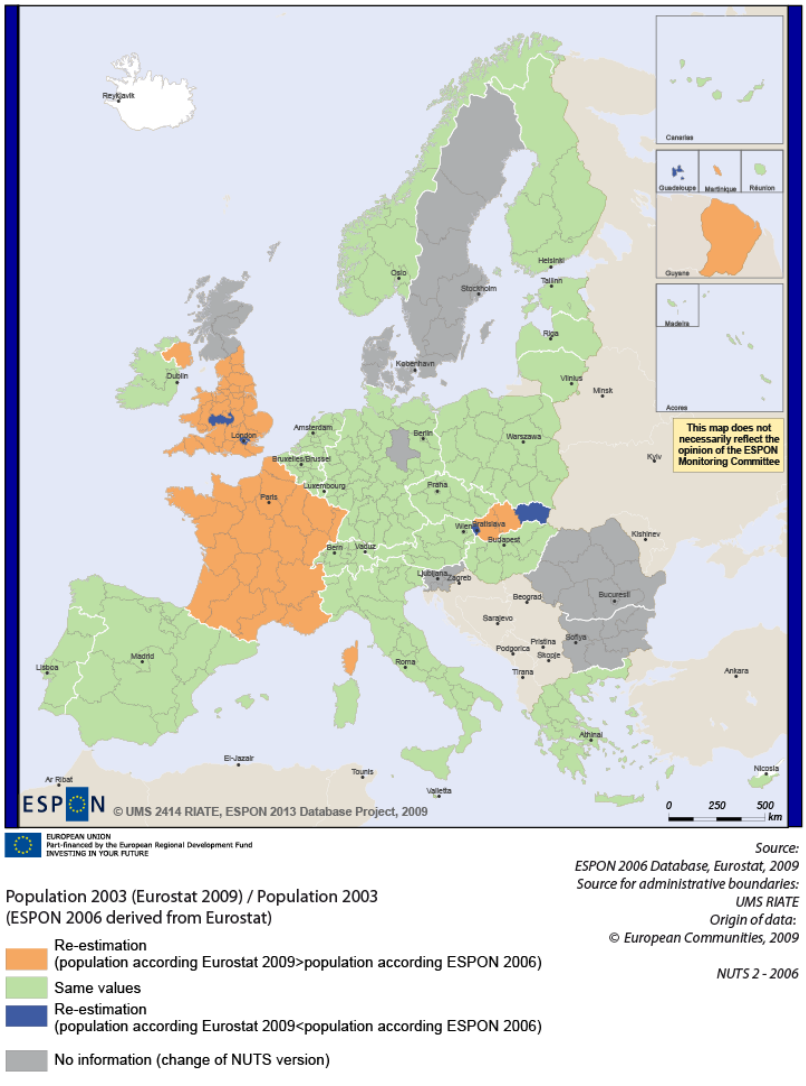


Figure 8 – NUTS 2 where population in 2003 has been re-estimated according to Eurostat

2.3 Description of datasets who will integrate the ESPON DB 2013

More concretely, the ESPON 2013 Database Project proposes to integrated indicators coming from the ESPON 2006 External Database. As a whole, this concerns 15 datasets (including 198 indicators) coming from ESPON Project indicators and 10 datasets (including 54 indicators) coming from ESPON Basic indicators (figure 9[5])

ESPON Project Indicators

These indicators have been calculated in the framework of the different ESPON projects by the Transnational Project Groups. Some of the indicators were calculated on the basis of data provided by EUROSTAT and National Statistical offices and others are results from model calculations and project based data inquiries.

Information from this section has to be cited as © ESPON Database, 2006

Name	Type	Size
Regional Classification of Europe	XLS	52,00 KB
Spatial Typologies NUTS 2	XLS	98,50 KB
Spatial Typologies NUTS 3	XLS	336,50 KB
Population	XLS	505,00 KB
Wealth and Production NUTS 2	XLS	54,00 KB
Transport NUTS 3	XLS	300,00 KB
Communication Technology NUTS 2	XLS	40,50 KB
Land Use NUTS 3	XLS	248,50 KB
Environment NUTS 3	XLS	255,50 KB
Agriculture NUTS 2	XLS	42,00 KB
INTERREG IIIA data	XLS	169,00 KB
INTERREG IIIA projects	XLS	162,00 KB
INTERREG IIIB projects	XLS	239,50 KB
Public Sector NUTS 2	XLS	72,50 KB
Public Sector NUTS 3	XLS	249,50 KB

ESPON Basic indicators

The tables presented in this section include basic information on population, employment, unemployment and economic output. This information is mainly based on regional statistics from EUROSTAT. It covers the regions of the EU25 and Candidate Countries and in some cases it has been complemented with additional data from National Statistics Offices (NSO) in order to cover the entire ESPON territory (EU 25 + Romania and Bulgaria + Switzerland and Norway).

Information from this section has to be cited as © ESPON Database, 2006

Name	Type	Size
Area NUTS 2	XLS	37,50 KB
Area NUTS 3	XLS	129,50 KB
Population NUTS 2	XLS	112,50 KB
Population NUTS 3	XLS	219,50 KB
Employment and Labour Market NUTS 2	XLS	88,00 KB
Unemployment NUTS 2	XLS	52,00 KB
Unemployment NUTS 3	XLS	189,00 KB
Wealth and Production NUTS 2	XLS	51,50 KB
Wealth and Production NUTS 3	XLS	188,00 KB
Research and Development NUTS 2	XLS	44,50 KB

Figure 9 – Targeted datasets to update in the ESPON 2013 Database
http://www.espon.eu/main/Menu_ScientificTools/ESPON2006Tools/DatabasePublicFiles/projectindicators.html

[5] Listing and description of all indicators contented in these datasets in annexes.

3. Integration of the ESPON 2006 Database in the 2013 DB architecture - a real added value.

The integration of ESPON 2006 indicators to the ESPON 2013 Database infrastructure implies to adapt the data and metadata model to the new requirements. In some cases it implies (3.1) to add information to the metadata of the ESPON 2006 Database, which is not fully compliant with the ISO-19115 norm. In other cases, thanks to information available in the different ESPON Reports and metadata mentioned in the datasets downloadable on ESPON Website, it is possible to improve the quality of metadata (3.2). Finally, the metadata and data models proposed by the ESPON 2013 Database Project makes possible a better structuring of the information contained in the metadata of ESPON 2006 indicators (3.3)

3.1. Elements added by the ESPON 2013 Database Project to the metadata of ESPON 2006 indicators

In relation with the data and metadata models proposed by the ESPON 2013 Project, some elements were not described in the datasets available on ESPON Website. This has been corrected for the following elements:

- **Metadata point of contact (figure 10).** The metadata template requires a contact person. Considering that it is complicated to find valid contact 4 years after the end of these projects, we have decided to put ESPON Coordination Unit as a metadata point of contact.

Metadata point of contact	
name	ESPON
email	info@espon.eu
organization	ESPON Coordination Unit
function	Project launcher
role	Data collector

Figure 10 – Adding information concerning point of contact and upload date

- **Date of the scope (figure 11):** Defining the precise date of upload allows defining when the data have been published. Thanks to this it is possible to see immediately that it corresponds to the first ESPON Programme. Considering that this information is not available in the dataset we have chosen to indicate "May 2006" as an upload date.

It corresponds more or less to the end of scientific activities in the first ESPON Program. This date can be changed if necessary.

- **Data provider (figure 11):** for each datasets uploaded, we have considered "ESPON 2006 Database" as a first level data provider and the real data provider (ESPON Projects, Eurostat etc.) as a second level data provider, in the "estimation methodology" field. As an example the data provider of a given scope could be "ESPON 2006 Database", and in the estimation methodology field it could be specified that this ESPON 2006 DB data came primarily from Eurostat (regional database) or another data source (National Statistical Offices of Switzerland for instance).

label	1	
lineage		
	provider	ESPON 2006 Database
	date	05/2006
	URL	
	methodolo	Data comes from Eurostat (Regional database) but are described in an outdated NUTS version. Consequently they are not downloadable in Eurostat website.
	methodology URI	
reliability		
	estimatior	FAUX
	quality	high
constraints		
	public dat	VRAI
	public met	VRAI
	copyright	© ESPON Database, 2006
label	2	
lineage		
	provider	ESPON 2006 Database
	date	05/2006
	URL	
	methodolo	Data comes from the National Statistical Office of Switzerland (Office fédéral de la Statistique)
	methodology URI	
reliability		
	estimatior	FAUX
	quality	high
constraints		
	public dat	VRAI
	public met	VRAI
	copyright	© ESPON Database, 2006

Figure 11 – Definition of date and provider of ESPON 2006 datasets scope

3.2. Improvements of the metadata and data information

Considering the different fields that have to be filed in the metadata and data templates, it allows to envisage new possibilities for using the ESPON 2006 indicators

- **Precise the methodology (figure 12).** For some indicators, as typologies, the methodology description has been improved. Thanks to the information contented in the final reports of concerned projects, we have added substantial elements in the methodology field.

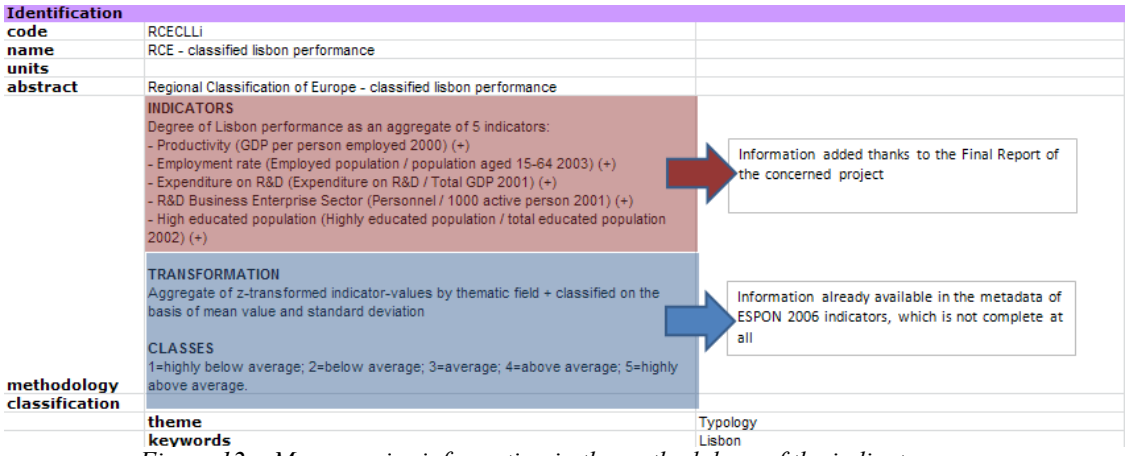


Figure 12 – More precise information in the methodology of the indicator

- **Precise the data source (figure 13).** Thanks to the new data and metadata models, it is possible to integrate the URL of the ESPON Report linked to the data calculation, which is really a value added for valorise ESPON results.

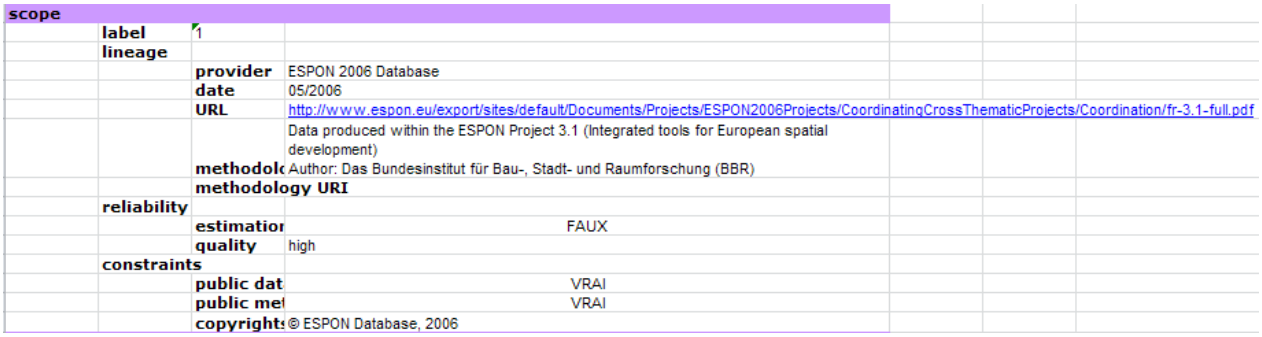


Figure 13 – Identification of the source

3.3. New metadata structure for existing information

The metadata integrated in the ESPON 2006 Database template are done in "free text" fields. Thanks to the new metadata model, it is possible to better structure these fields (figure 14):

1. For each territorial unit contained in the dataset, the NUTS level and the NUTS version has been identified.

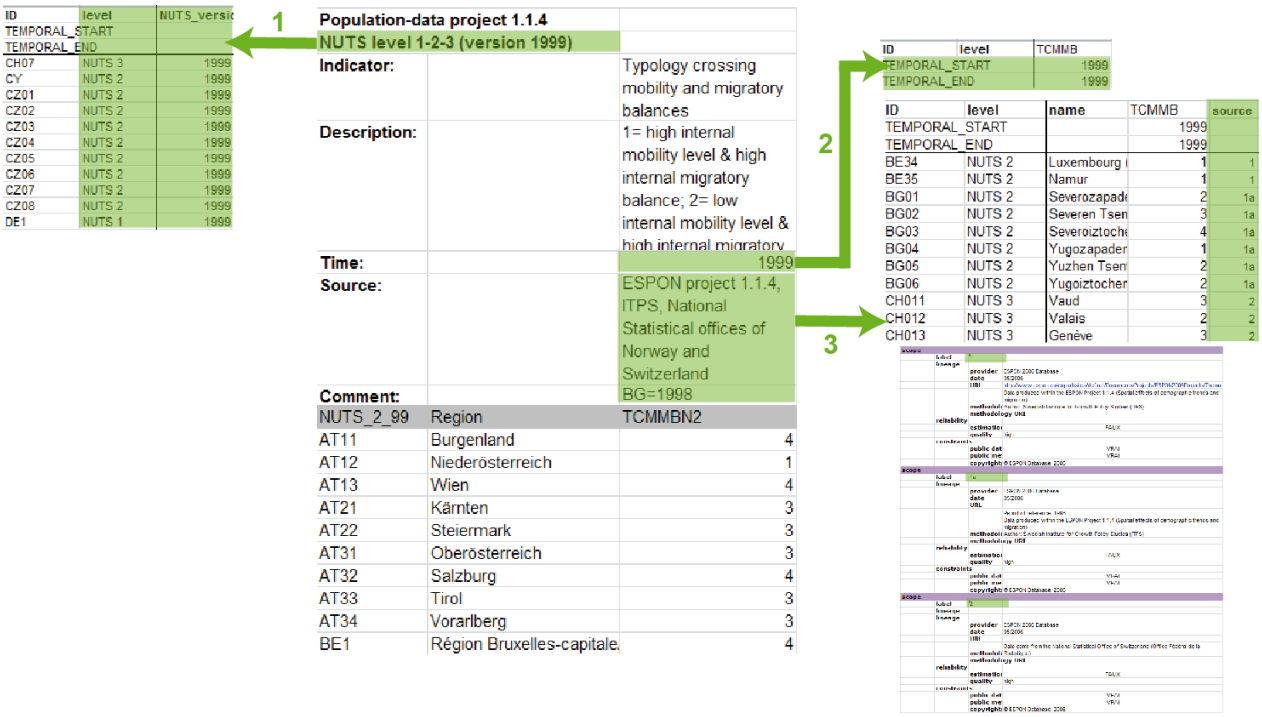


Figure 14 – Integration of metadata in the 2013 database structure (theoretical example)

2. The temporal validity of the indicator has been indicated in dedicated field
3. All sources and comment have been better integrated (column source which is linked to the metadata sheet where values are precisely described).

Consequently, it is now possible for the end-user to better query the ESPON 2006 database:

- What data are the available in the NUTS 1999 delineation?
- What are the available indicators with a temporal extent starting in 1999 and finishing in 2000?
- What is the information available coming from ESPON 1.1.4 Project?

These basic questions, typically adapted to the content of the ESPON 2006 Database will be possible in a near future.

Last but not least, this work makes possible to integrate the results from ESPON 2006 and ESPON 2013 Programs in the same database architecture.

APPENDIX 1 - Updated indicators from the ESPON 2006 Database – Data from TPG's (ESPON Project Indicators)

DATASET	NAME OF INDICATOR	Geog. Object	Temporal start/end	SOURCE (main)
01_regional_classification_of_europe_nuts_2_data	RCE - classified economy	NUTS2 (version 2003)	2002/2002	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified lisbon performance	NUTS2 (version 2003)	2002/2002	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified labour market	NUTS2 (version 2003)	2003/2003	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified demography	NUTS2 (version 2003)	2002/2002	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified naturalness	NUTS2 (version 2003)	2000/2000	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified natural hazards	NUTS2 (version 2003)	2002/2002	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified technological hazards	NUTS2 (version 2003)	2002/2002	ESPON 2.4.2
01_regional_classification_of_europe_nuts_2_data	RCE - classified accessibility	NUTS2 (version 2003)	2001/2001	ESPON 2.4.2
01_spatial_typologies_nuts_2	Settlement Structure Typology	NUTS2 (version 1999)	1999/1999	ESPON project 3.1
01_spatial_typologies_nuts_2	A typology of levels of household telecommunications uptake	NUTS2 (version 1999)	2002/2002	ESPON project 1.2.2
01_spatial_typologies_nuts_2	A typology of estimated levels of business telecommunications access and uptake	NUTS2 (version 1999)	2002/2002	ESPON project 1.2.2
01_spatial_typologies_nuts_2	A typology comparing levels of household and business telecommunications uptake	NUTS2 (version 1999)	2002/2002	ESPON project 1.2.2
01_spatial_typologies_nuts_2	An overall typology of combined household and business telecommunications development	NUTS2 (version 1999)	2002/2002	ESPON project 1.2.2
01_spatial_typologies_nuts_2	Typology Multimodal Accessibility Potential	NUTS2 (version 1999)	2001/2001	ESPON project 2.1.1
01_spatial_typologies_nuts_2	Typology of lagging regions	NUTS2 (version 1999)	2000/2000	ESPON project 2.1.1
01_spatial_typologies_nuts_2	Typologies of regional specialisation and GDP per capita 2001	NUTS2 (version 1999)	2001/2001	ESPON project 1.1.3
01_spatial_typologies_nuts_2	Typologies of regional specialisation and GDP per capita 1995-2001	NUTS2 (version 1999)	1995/2001	ESPON project 1.1.3
01_spatial_typologies_nuts_2	Coast	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_2	Border	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_2	Pentagon EU 15	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_2	Pentagon EU 27 plus 2	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_2	Part of Interreg	NUTS2	2000/2000	ESPON

	North-Sea Programme	(version 1999)		project 3.1
01_spatial_typologies_nuts_2	Part of Interreg CADSES Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Atlantic-Area Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "Non continental and overseas cooperation areas"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "Northern-Peripherie"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Alpine-Space Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "Archimedes"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "Baltic Sea"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "Medoc-Area"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "South-West-Europe"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Part of Interreg Programme "North-West-Europe"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	"Objective 1" regions= regions situated within objective 1 regions	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_2	Objective 2 regions includes regions containing at least one Objective 2 region (partly)	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Typology Settlement Structure	NUTS2 (version 1999)	1999/1999	ESPON project 3.1
01_spatial_typologies_nuts_3	Typology Multimodal Accessibility Potential	NUTS2 (version 1999)	2001/2001	ESPON project 2.1.1
01_spatial_typologies_nuts_3	Typologie of lagging regions	NUTS2 (version 1999)	2000/2000	ESPON project 2.1.1
01_spatial_typologies_nuts_3	Urban-rural typology	NUTS2 (version 1999)	1999/1999	ESPON project 1.1.2
01_spatial_typologies_nuts_3	Coast	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_3	Border	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_3	Pentagon EU 15	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_3	Pentagon EU 27 plus 2	NUTS2 (version 1999)	2003/2003	ESPON project 2.1.1
01_spatial_typologies_nuts_3	Part of Interreg North-Sea Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg CADSES Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Atlantic-Area Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Programme "Non continental and overseas cooperation	NUTS2 (version 1999)	2000/2000	ESPON project 3.1

	areas"			
01_spatial_typologies_nuts_3	Part of Interreg Programme "Northern-Peripherie"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Alpine-Space Programme	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Programme "Archimedes"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Programme "Baltic Sea"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Programme "Medoc-Area"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Programme "South-West-Europe"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Part of Interreg Programme "North-West-Europe"	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	"Objective 1" regions= regions situated within objective 1 regions	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
01_spatial_typologies_nuts_3	Objective 2 regions includes regions containing at least one Objective 2 region (partly)	NUTS2 (version 1999)	2000/2000	ESPON project 3.1
04_wealth_and_production	GDP per capita in PPS	NUTS2 (version 1999)	1999/1999	ESPON project 3.1
04_wealth_and_production	GDP per capita in PPS- Global deviation (EU15)	NUTS2 (version 1999)	1999/1999	ESPON project 3.1
04_wealth_and_production	GDP per capita in PPS- Medium deviation	NUTS2 (version 1999)	1999/1999	ESPON project 3.1
04_wealth_and_production	GDP per capita in PPS- Local deviation	NUTS2 (version 1999)	1999/1999	ESPON project 3.1
02_population-data_project_1.1.4_N2	Typology of migratory balances by age classes	NUTS2 (version 1999)	1995/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Dependency rate	NUTS2 (version 1999)	1995/1995.1999/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Share of NUTS 2 average population living in NUTS 3 regions with population decline	NUTS2 (version 1999)	1995/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Share of NUTS 2 area comprising NUTS 3 regions with population decline	NUTS2 (version 1999)	1995/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population with 65 and more years (%) (Model A)	NUTS2 (version 1999)	2000/2000. 2025/2025, 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population with 65 and more years (%) (Model B0)	NUTS2 (version 1999)	2000/2000. 2025/2025, 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population with 65 and more years (%) (Model B1)	NUTS2 (version 1999)	2000/2000. 2025/2025, 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population with 65 and more years (%) (Model B2)	NUTS2 (version 1999)	2000/2000.2025/2025, 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population with 65 and more years (%) (Model B3)	NUTS2 (version 1999)	2000/2000. 2025/2025, 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Average score on indirect "ageing"/ "depopulating" indicators	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Average score on indirect "ageing"/ "depopulating"	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4

	indicators, Grouped (quartiles)			
02_population-data_project_1.1.4_N2	National Total Fertility Rates	NUTS2 (version 1999)	1999/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Ageing Population (4 groups) 65+/Tot.	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Ageing "Labour Force" (4 groups) 55-64/20-64	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	"Labour Force" Replacement (4 groups) 10-19/55-64	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Post-Active Dependency (4 groups) 65+/20-64	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Aged People vs. Youth (4 groups) 65+/15-24	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Share of children (4 groups) 0-14/Tot.pop	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Changes in Natural Growth Potential (4 groups)20-29 years in 2020 (born 1991-2000)/20-29 years in 2000 (born 1971-1980)	NUTS2 (version 1999)	2000/2020	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Ageing Population (indexes) 65+/Tot.	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Ageing "Labour Force" (indexes) 55-64/20-64	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	"Labour Force" Replace-ment (indexes) 10-19/55-64	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Post-Active Dependency (indexes) 65+/20-64	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Aged People vs. Youth (indexes) 65+/15-24	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Share of children (indexes) 0-14/Tot.pop	NUTS2 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Changes in Natural Growth Potential (indexes)20-29 years in 2020 (born 1991-2000)/20-29 years in 2000 (born 1971-1980)	NUTS2 (version 1999)	2000/2020	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Total fertility rate	NUTS2 (version 1999)	1990/1990; 1995/1995; 1999/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	External immigration	NUTS2 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Migratory balance by regions between	NUTS2 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Absolute migratory balance	NUTS2 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Migratory balance 17.5 to 27.5 years old	NUTS2 (version 1999)	1995/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Migratory balance 32.5 to 42.5 years old	NUTS2 (version 1999)	1995/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Migratory balance 52.5 to 67.5 years old	NUTS2 (version 1999)	1995/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Synthetic cartography of migratory balances	NUTS2 (version 1999)	1995/2000	ESPON project 1.1.4

	for the main age classes			
02_population-data_project_1.1.4_N2	Variation of the population (%) (Model A)	NUTS2 (version 1999)	2000/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Variation of the population (%) (Model B0)	NUTS2 (version 1999)	2000/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Variation of the population (%) (Model B2)	NUTS2 (version 1999)	2000/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Variation of the population (%) (Model B3)	NUTS2 (version 1999)	2000/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population between 15 and 64 years (%) (Model A)	NUTS2 (version 1999)	2000/2000; 2025/2025; 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population between 15 and 64 years (%) (Model B1)	NUTS2 (version 1999)	2000/2000;2025/2025 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population between 15 and 64 years (%) (Model B1)	NUTS2 (version 1999)	2000/2000; 2025/2025; 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population between 15 and 64 years (%) (Model B2)	NUTS2 (version 1999)	2000/2000; 2025/2025; 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N2	Population between 15 and 64 years (%) (Model B3)	NUTS2 (version 1999)	2000/2000; 2025/2025; 2050/2050	ESPON project 1.1.4
02_population-data_project_1.1.4_N3	Type of rural area	NUTS3 (version 1999)	2000/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N3	Relative depopulation, quartiles	NUTS3 (version 1999)	1990/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N3	Population change	NUTS3 (version 1999)	1990/2000; 1990/1995; 1995/2000	ESPON project 1.1.4
02_population-data_project_1.1.4_N23	Share (%) of population in the ages 65+	NUTS 2-3 (version 1999)	1990/1990; 1995/1995; 1999/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N123	Typology crossing mobility and migratory balances	NUTS 1-2-3 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N123	Internal migratory balance	NUTS 1-2-3 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N123	Total migratory balance	NUTS 1-2-3 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N123	External migratory balance	NUTS 1-2-3 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N123	Internal mobility by region	NUTS 1-2-3 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N123	Mobility by region relative to national mobility	NUTS 1-2-3 (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N23a	Total population development	NUTS 2-3a (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N23a	Natural population development	NUTS 2-3a (version 1999)	1996/1999	ESPON project 1.1.4
02_population-data_project_1.1.4_N23a	Net migration	NUTS 2-3a (version 1999)	1996/1999	ESPON project 1.1.4
ESPON_2006_06_transport_nuts_3	Number of commercial airports	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Number of commercial seaports	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Length of road network (km)	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Length of railway network, km	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1

ESPON_2006_06_transport_nuts_3	Traffic in commercial airports	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Connectivity to commercial airports by car of the capital or centroid representative of the NUTS3	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Connectivity to commercial seaports by car of the capital or centroid representative of the NUTS3	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Time to the nearest motorway access, by car of the capital or centroid representative of the NUTS3	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Daily population accessible by car	NUTS3 (version 1999)	1999/1999	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Daily market accessible by car in terms of GDP	NUTS3 (version 1999)	2000/2000	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Potential accessibility air, ESPON space = 100	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Potential accessibility rail, ESPON space = 100	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Potential accessibility road, ESPON space = 100	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Potential accessibility multimodal, ESPON space = 100	NUTS3 (version 1999)	2001/2001	ESPON project 1.2.1
ESPON_2006_06_transport_nuts_3	Accessibility time to market by road,	NUTS3 (version 1999)	1997/1997	ESPON Project 2.1.1
ESPON_2006_06_transport_nuts_3	Accessibility time to market by rail,	NUTS3 (version 1999)	1997/1997	ESPON Project 2.1.1
ESPON_2006_06_transport_nuts_3	Accessibility time to market by rail and road,	NUTS3 (version 1999)	1997/1997	ESPON Project 2.1.1
ESPON_2006_09_communication_technology_nuts_2	Share of Internet users	NUTS2 (version 1999)	2002/2002	ESPON Project 1.2.2
ESPON_2006_09_communication_technology_nuts_2	Share of firms with own website	NUTS2 (version 1999)	2002/2002	ESPON Project 1.2.2
ESPON_2006_11_land_use_nuts_3	Share of artificial surfaces	NUTS3 (version 1999)	1996/1996	Espon project 1.1.2
ESPON_2006_11_land_use_nuts_3	Share of artificial surfaces per 1000 inh.	NUTS3 (version 1999)	1996/1996	Espon project 1.1.2
ESPON_2006_11_land_use_nuts_3	Share of artificial surfaces per 100 million GDP PPS	NUTS3 (version 1999)	1996/1996	Espon project 1.1.2
ESPON_2006_11_land_use_nuts_3	Share of urban fabric	NUTS3 (version 1999)	1996/1996	ESPON Project 3.1
ESPON_2006_11_land_use_nuts_3	Share of arable land	NUTS3 (version 1999)	1996/1996	ESPON Project 3.1
ESPON_2006_11_land_use_nuts_3	Share of permanent crops	NUTS3 (version 1999)	1996/1996	ESPON Project 3.1
ESPON_2006_12_environment_nuts_3	Occurrence of snow avalanches	NUTS3 (version 1999)	2004/2004	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Large scale droughts in Europe	NUTS3 (version 1999)	1904/1995	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Regional earthquake hazard potential	NUTS3 (version	1998/1998	ESPON Project 1.3.1

		1999)		
ESPON_2006_12_environment_nuts_3	Extreme temperatures	NUTS3 (version 1999)	1961/1990	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Regional flood hazard potential	NUTS3 (version 1999)	1987/2002	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Forest fire hazard	NUTS3 (version 1999)	1997/2003	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Occurrence of landslides	NUTS3 (version 1999)	2004/2004	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Occurrence of storm surges	NUTS3 (version 1999)	2004/2004	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Occurrence of tsunami runups and tsunami potential areas in Europe	NUTS3 (version 1999)	-1628/2003	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Volcanic eruptions during the last 10 000 years	NUTS3 (version 1999)	-1628/2003	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Approximate probability of having winter storms and for tropical storms probable maximum intensity	NUTS3 (version 1999)	2004/2004	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Air traffics hazard potential	NUTS3 (version 1999)	1996/2003	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Chemical plants hazard potential	NUTS3 (version 1999)	2001/2004	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Potential risk of radioactive contamination on NUTS3 regions	NUTS3 (version 1999)	2003/2003	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Classification of Oil-SUM values	NUTS3 (version 1999)	2002/2002	ESPON Project 1.3.1
ESPON_2006_12_environment_nuts_3	Sum of all weighted hazard values	NUTS3 (version 1999)	2004/2004	ESPON Project 1.3.1
ESPON_2006_13_agriculture_nuts_2	Share of UAA which is arable	NUTS2 (version 1999)	2001/2001	ESPON Project 2.1.3
ESPON_2006_13_agriculture_nuts_2	Share of UAA that is fallow	NUTS2 (version 1999)	2001/2001	ESPON Project 2.1.3
ESPON_2006_13_agriculture_nuts_2	Share of farm holders aged <65	NUTS2 (version 1999)	1997/1997	ESPON Project 2.1.3
ESPON_2006_14_interreg_iiia_areas_nuts_3	Table of belonging to the 64 INTERREG IIIA Programs	NUTS3 (version 2003)	2000/2000	ESPON
ESPON_2006_14_interreg_iiia_data_nuts_3	Typology of borders in NUTS3 regions participating in INTERREG IIIA Programmes	NUTS3 (version 2003)	2006/2006	ESPON INTERACT/KTH
ESPON_2006_14_interreg_iiia_data_nuts_3	Intensity of projects per INTERREG IIIA Programme	NUTS3 (version 2003)	2006/2006	ESPON INTERACT/KTH
ESPON_2006_14_interreg_iiia_data_nuts_3	Geographic type of land border of NUTS3 INTERREG IIIA programme areas	NUTS3 (version 2003)	2006/2006	ESPON INTERACT/KTH
ESPON_2006_14_interreg_iiia_data_nuts_3	Density of border crossings in INTERREG IIIA areas	NUTS3 (version 2003)	2006/2006	ESPON INTERACT/KTH
ESPON_2006_14_interreg_iiia_data_nuts_3	Economic disparities per programme	NUTS3 (version 2003)	2006/2006	ESPON INTERACT/KTH
ESPON_2006_14_interreg_iiib_areas_nuts_3	Table of belonging to the 14 INTERREG IIIB Programs	NUTS3 (version 1999)	2006/2006	ESPON Project 3.1
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-	NUTS2&NUTS3 (version	1998/2000	ESPON Project 2.2.2

	Accession-Aid (PHARE, PHARE CBC, ISPA) addressing capital-supply-potential	1999)		
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing environmental quality	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing geographical position	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing potential of innovation	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing institutional conditions	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing labour market potenital	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing regional market potential	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Percentage of regional Pre-Accession-Aid (PHARE, PHARE CBC, ISPA) addressing urbanisation&localisation advantages	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Total Pre-Accession-Aid spending (PHARE, PHARE CBC, ISPA)	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	Average Annual Pre-Accession-Aid spending (PHARE, PHARE CBC, ISPA)	NUTS2&NUTS3 (version 1999)	1998/2000	ESPON Project 2.2.2
ESPON_2006_18_public_sector_nuts_2_3_data	All Structural and Cohesion Fund expenditure	NUTS2&NUTS3 (version 1999)	1994/1999	ESPON Project 2.2.1
ESPON_2006_18_public_sector_nuts_2_3_data	Structural Fund expenditure (R)	NUTS2&NUTS3 (version 1999)	1994/1999	ESPON Project 2.2.1
ESPON_2006_18_public_sector_nuts_2_3_data	Structural Fund expenditure (S)	NUTS2&NUTS3 (version 1999)	1994/1999	ESPON Project 2.2.1
ESPON_2006_18_public_sector_nuts_2_3_data	Structural Fund expenditure (A)	NUTS2&NUTS3 (version 1999)	1994/1999	ESPON Project 2.2.1
ESPON_2006_18_public_sector_nuts_2_3_data	Cohesion Fund expenditure (T)	NUTS2&NUTS3 (version 1999)	1994/1999	ESPON Project 2.2.1
ESPON_2006_18_public_sector_nuts_2_3_data	Cohesion Fund expenditure (E)	NUTS2&NUTS3 (version 1999)	1994/1999	ESPON Project 2.2.1

APPENDIX 2 - Updated indicators from the ESPON 2006 Database – Data from Eurostat (ESPON Basic Indicators)

DATASET	NAME OF INDICATOR	Geog. Object	Temporal start/end	SOURCE (main)
00_area_nuts_2-3_data	Total land area	NUTS2, NUTS3 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Population total	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Population density	NUTS2 (version 2003)	2002/2002	Eurostat
02_population_nuts_2	Share of female population	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of male population	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of population < 14 years	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of population > 65 years	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of high aged population (> 75 years)	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of female population < 14 years	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of female population > 65 years	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of female high aged population (> 75 years)	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of male population < 14 years	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of male population > 65 years	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Share of male high aged population (> 75 years)	NUTS2 (version 2003)	2003/2003	Eurostat
02_population_nuts_2	Development of total population in %	NUTS2 (version 2003)	1995/2003	Eurostat
02_population_nuts_2	Development of female population in %	NUTS2 (version 2003)	1995/2003	Eurostat
02_population_nuts_2	Development of male population in %	NUTS2 (version 2003)	1995/2003	Eurostat
02_population_nuts_3	Average Population	NUTS3 (version 2003)	2003/2003	Eurostat
02_population_nuts_3	Average male Population, share in %	NUTS3 (version 2003)	2003/2003	Eurostat
02_population_nuts_3	Average female Population, share in %	NUTS3 (version 2003)	2003/2003	Eurostat
02_population_nuts_3	Population density	NUTS3 (version 2003)	2002/2002	Eurostat
02_population_nuts_3	Development average population	NUTS3 (version 2003)	1995/2003	Eurostat
03_employment_and_labourmarket_nuts_2	Active population total 2001	NUTS2 (version 1999)	2001/2001	Eurostat
03_employment_and_labourmarket_nuts_2	Share of active population < 25 years	NUTS2 (version	2001/2001	Eurostat

		1999)		
03_employment_and_labourmarket_nuts_2	Persons employed	NUTS2 (version 1999)	2001/2001; 2002/2002	Eurostat
03_employment_and_labourmarket_nuts_2	Share of persons employed male	NUTS2 (version 1999)	2001/2001	Eurostat
03_employment_and_labourmarket_nuts_2	Share of persons employed female	NUTS2 (version 1999)	2001/2001	Eurostat
03_employment_and_labourmarket_nuts_2	Share of persons employed in Agriculture in % of total	NUTS2 (version 1999)	2001/2001	Eurostat
03_employment_and_labourmarket_nuts_2	Share of persons employed in Industry in % of total	NUTS2 (version 1999)	2001/2001	Eurostat
03_employment_and_labourmarket_nuts_2	Share of persons employed in Services in % of total	NUTS2 (version 1999)	2001/2001	Eurostat
03_employment_and_labourmarket_nuts_2	Share of employed persons, national, < 25 years, in % of total	NUTS2 (version 1999)	2002/2002	Eurostat
03_employment_and_labourmarket_nuts_2	Share of employed persons, national, < 25 years, in % of total, >65 years	NUTS2 (version 1999)	2003/2003	Eurostat
03_unemployment_nuts_2	Unemployment rate total	NUTS2 (version 2003)	2004/2004	Eurostat
03_unemployment_nuts_2	Unemployment rate female	NUTS2 (version 2003)	2004/2004	Eurostat
03_unemployment_nuts_2	Unemployment rate male	NUTS2 (version 2003)	2004/2004	Eurostat
03_unemployment_nuts_2	Unemployment rate of persons < 25 years	NUTS2 (version 2003)	2004/2004	Eurostat
03_unemployment_nuts_2	Development of unemployment rate	NUTS2 (version 2003)	1999/2004	Eurostat
03_unemployment_nuts_2	Development of unemployment rate, female	NUTS2 (version 2003)	1999/2004	Eurostat
03_unemployment_nuts_2	Development of unemployment rate, male	NUTS2 (version 2003)	1999/2004	Eurostat
03_unemployment_nuts_3	Unemployment rate total	NUTS3 (version 1999)	2001/2001	Eurostat
03_unemployment_nuts_3	Unemployment rate female	NUTS3 (version 1999)	2001/2001	Eurostat
03_unemployment_nuts_3	Unemployment rate male	NUTS3 (version 1999)	2001/2001	Eurostat
03_unemployment_nuts_3	Unemployment rate of persons < 25 years	NUTS3 (version 1999)	2001/2001	Eurostat
03_unemployment_nuts_3	Development of unemployment rate	NUTS3 (version 1999)	1998/2001	Eurostat
03_unemployment_nuts_3	Development of unemployment rate, female	NUTS3 (version 1999)	1998/2001	Eurostat
03_unemployment_nuts_3	Development of unemployment rate, male	NUTS3 (version 1999)	1998/2001	Eurostat
03_unemployment_nuts_3	Development of unemployment rate, young pop	NUTS3 (version 1999)	1998/2001	Eurostat
wealth_and_production_nuts_2-3_data	GDP in Purchasing Power Parities per inhabitant 2002	NUTS2, NUTS3 (version 2003)	2002/2002	ESPON project 3.2
wealth_and_production_nuts_2-3_data	GDP in Euro per inhabitant 2002	NUTS2, NUTS3 (version 2003)	2002/2002	ESPON project 3.2
wealth_and_production_nuts_2-3_data	Development of GDP in	NUTS2,	1998/2002	ESPON project

	Purchasing Power Parities per inhabitant 1998-2002	NUTS3 (version 2003)		3.2
wealth_and_production_nuts_2-3_data	Development of GDP in Euro per inhabitant 1998-2002	NUTS2, NUTS3 (version 2003)	1998/2002	ESPON project 3.2
07_research_and_development_nuts_2	Patent applications to the EPO per persons employed	NUTS2 (version 2003)	2002/2002	Eurostat
07_research_and_development_nuts_2	Total intramural R&D expenditure	NUTS2 (version 2003)	2002/2002	Eurostat
07_research_and_development_nuts_2	FuE Business Enterprise Sector, personnell	NUTS2 (version 2003)	2003/2003	Eurostat

TECHNICAL REPORT

ESPON

Towards an approach of time series data issue: from empirical methods to applications

CONTENTS

The problem of time series data is the lack of data a territorial unit either because the territorial unit has changed in the course of time or because data are simply missing.

Three complementary methods have been investigated: data collection in historical regional databases, NUTS changes modeling and missing values estimation.

The building of a dictionary of changes is a fundamental step to build harmonized spatiotemporal database. It consists to describing and modeling of territorial changes.

This theoretical framework was tested through applications.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

31 PAGES

LIST OF AUTHORS

Ben Rebah Maher, UMS 2414 RIATE

Plumejeaud Christine, LIG

Ysebaert Ronan, UMS 2414 RIATE

Peeters Didier, IGEAT, Université Libre de Bruxelles

Contact

maher.benrebah@ums-riate.fr
christine.plumejeaud@imag.fr
ronan.ysebaert@ums-riate.fr
dpeeter1@ulb.ac.be

UMS 2414 RIATE
Tel. (+ 33) 1 57 27 65 35

LIG Grenoble
Tel. (+33) 4 76 82 72 11

Université Libre de Bruxelles
Tel. (+32) 2 650 50 77

TABLE OF CONTENT

Introduction

ESPON DB 2013 project aims to improve the access to time series data. The issue of time series is a recurring necessity for ESPON projects and several European institutions primarily DG REGIO and ESUROSTAT. In spite of it importance, this process has not very well initiated by the previous ESPON DB project (2006)1.

This brief technical report is for the purpose of providing a background to time series challenge of ESPON DB project during the timeframe December 2008 to December 2009. In this first period of the project, activities done are mainly conceptual. Some applications and operational results have been developed since November 2009 (figure 25).

The issue of time series data could be fundamentally assimilated to the lack of data for a territorial unit either because the territorial unit has changed in the course of time or because data are simply missing. Difficulties to build time series data can be related firstly to the lack of achieved databases EUROSTAT, as the principal provider of European statistics, does not archive its database versions. It keeps just the last version of database. Secondly, information about historical changes of NUTS is often missing or uncertain.

Time series approach can be organized in two main steps. Firstly, collection and exploration of historical data bases (NewCronos from EUROSTAT, cohesion reports from DG-REGIO…) was undertaken. This works aims to provide a review of continuous time-data series could be built form this data bases. Additionally, we have explored NUTS changes between 1995 and 2006. The dictionary of NUTS changes is the main result of this exploration. It allows a review of territorial changes (codes, names and geometries). But the most contribution of the dictionary of changes is the identification of the genealogy (lineage) of NUTS which is very useful for the harmonization of time-series data.

The result of this first step will be used to build continuous time-series data. The conceptual model and the implementation of the computing (automation) of the process is in progress.

[1] http://www.espon.eu/mmp/online/website/content/tools/832/index_EN.html

MAUP (First Interim Report 15/06/2006)

1.7 In what sense is the MAUP a problem? (p. 15)

[…]

The most important problem is about international and historical comparisons: do the elementary spatial units which are used for the analysis have the same meaning in two different countries? At two different time periods? It is not easy to determine if a difference in the results is due to a difference in the processes which are underlying the observed phenomena, or simply to a difference in the meaning of the spatial entities that are used for the observation.

http://www.espon.eu/mmp/online/website/content/projects/261/431/index_EN.html

Data Navigator 2 (Final Report, Part 1 – Handbook for data collection, January 2007)

4.3.4.1 Temporal integration

(p. 110)

[…] **Identificators or the geometries of the NUTS change strongly during the period**. These changes introduce very big difficulties in the survey of variables in the time. It doesn't exist any simple ties between two dates.

(p. 120)

Changes of geometry and changes of units identification, don't permit to get directly evolutions of population basing on initial data, as the shows following example:

We estimate an evolution for a middle time (1990-2000) and represent it for different geographical grids (NUTS 23 1988 and NUTS 23 1999) whereas data of population initial are the similar, calculations of evolution (1990-2000) defer very strongly from a geographical grid to the other.

http://www.espon.eu/mmp/online/website/content/tools/127/index_EN.html

Europe in the World (Final Interim Report, December 2007)

6.5. Conclusion (vol. 1 p 242)

The synthesis of the regional insertion of the ESPON region into the world economy and the typology of gateway cities that we have elaborated in this final section of the report cannot be considered as definitive results as their elaboration was based on a limited number of criteria. **Better results could be obtained in the future if, for example, international trades statistics can be obtain for the regional level or if coherent time series could be analyzed concerning the evolution of air traffic linking European cities to the rest of the World.** The current set of results does however uncover some important findings in accordance with the objectives of the ESDP.

http://www.espon.eu/mmp/online/website/content/projects/260/720/index_EN.html

1 Territorial changing information sources

Data availability and data quality are crucial for the understanding and the formalization of Nuts genealogy. Our attempt to harmonize NUTS versions is the result of a meticulous combination of several sources provided by European and national institutions.

1.1 Legal source: Official journal of the European Union

The Official Journal is the legal source. It constitutes the legal framework of regulation of NUTS since 2003. The regulation EC n° 1059/2003 defines the NUTS and states the conditions of their modifications. This information is very useful to understand and formalize the changes of NUTS. This founder juridical text is amended and updated when new countries joined the European Union (figure1).

REGULATIONS

REGULATION (EC) No 176/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 20 February 2008

amending Regulation (EC) No 1059/2003 on the establishment of a common classification of territorial units for statistics (NUTS) by reason of the accession of Bulgaria and Romania to the European Union

Annex I to Regulation (EC) No 1059/2003 is amended as follows:

1. the following table is inserted between BE — BELGIQUE/BELGIË and CZ — ČESKÁ REPUBLIKA:

'БЪЛГАРИЯ

CODE	NUTS 1	NUTS 2	NUTS 3
BG			
BG3	СЕВЕРНА И ЮГОИЗТОЧНА БЪЛГАРИЯ		
BG31		Северозападен	
BG311			Видин
BG312			Монтана
BG313			Враца
BG314			Плевен
BG315			Ловеч

Figure 1: *Example of regulations amendment following the accession of Czech Republic and Romania*

1.2 Eurostat

Eurostat is the most valuable source. Many kinds of documents are produced by Eurostat2 allowing with NUTS changes, among which the most interesting is the description of changes occurring between each version. However, this description does not usually define types of changes. It is also, sometimes very imprecise, in the case of complex territorial modification, like the Danish territorial modification in 2006 which is described as follow: “Following an extensive regional reform in Denmark, where new administrative regions were created, Denmark will be divided into NUTS level 2 regions. The previous NUTS 3 regions do not generally correspond to the new NUTS level 2 regions. […] “The previous 15 administrative regions have been abolished and in their place, 11 new non administrative regions have been created by combining municipalities. Only two NUTS 3 level 3 regions remain intact”.

Concerning the update and the of EUROSTAT database, EUROSTAT does not archive its database versions. First of all, it keeps just the last version of database. Secondly, information about historical changes of NUTS is often missing or uncertain.

Besides the data available on the Eurostat internet portal, we obtained a CDRom with the Windows-only New Cronos application (figure 2), i. e. the Eurostat archives. This CDRom was unsuitable for the needs of the project because of its web interface designed exclusively for data consultation and not for data exportation. The data were also stored on the CDRom in a specific file format unknown from us which led us to spend time on finding technical workarounds to finally extract and store these in a format we could handle.

The data appeared to be organised in 271 tables and 16 categories. We made an inventory of their content in order to have an idea of their completeness, so to say the time span covered and the territories covered. The nuts are from 1999, and all European countries that are currently EU members are represented, but this of course depending on the type of data, the nuts level, the years considered, and logically the completeness of these archive decreases with older data.

To provide here an exhaustive list of the content, even in a synthesised way, would be a non-sense because of the number of variables and parameters. The data currently available on the Eurostat internet portal and the data included in these archives are partially the same, except that they do not use the same nuts reference system.

These data will be included in the Database system but will depend on the time series conversion tool to mix them with the current Eurostat data. Reversely, since they refer to an older nuts genealogy (1999) they might be useful in the next step of the time series harmonisation challenge, but probably as a validation mean, to be compared with the values our tool will compute for the 1999 nuts references.

[2] http://ec.europa.eu/eurostat/ramon/nuts/splash_regions.html

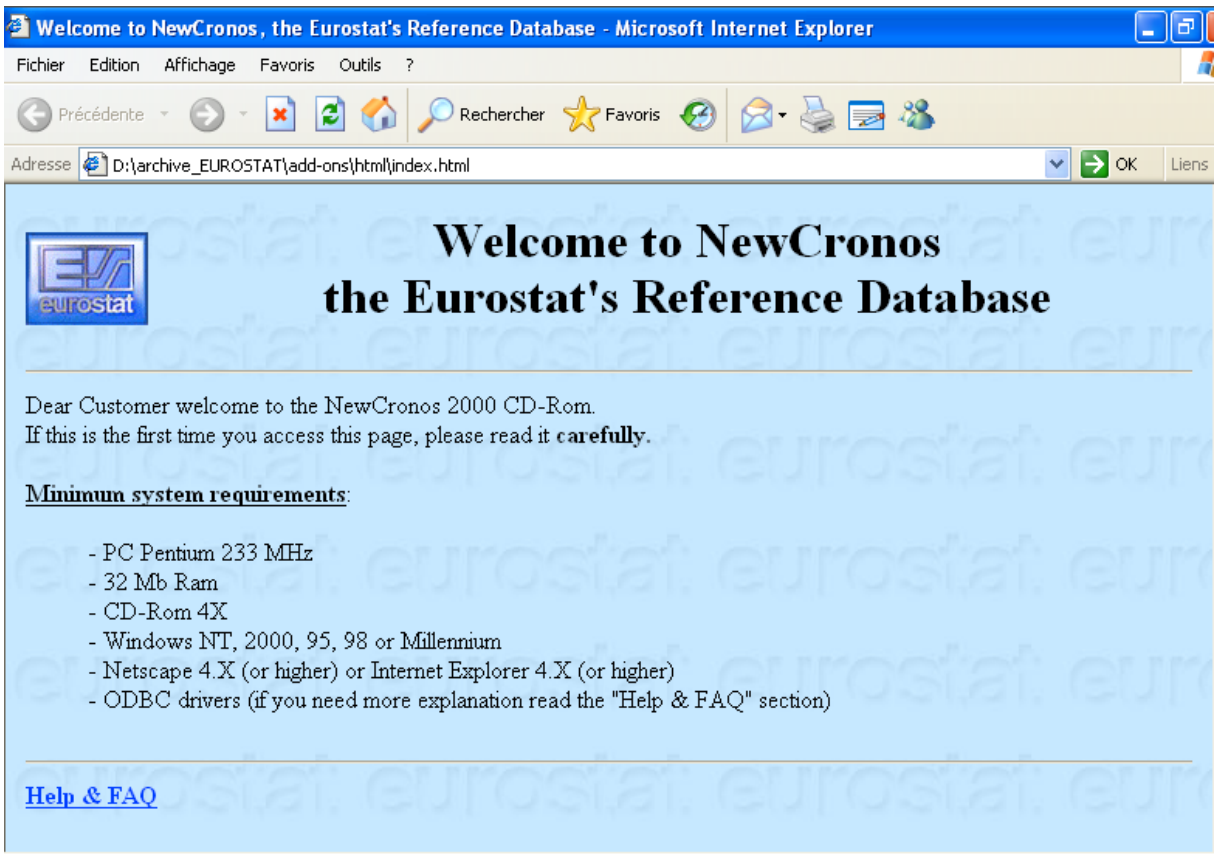


Figure 2: *NewCronos Web interface*

1.3 National Statistical Institutes

The European national statistical offices provide historical databases of national administrative boundaries. These sources are very useful for the understanding of local changes (national) which may affect geometry or structure of NUTS levels. National boundaries historical database is also very essential in the case of accessing new countries (EU15, EU25, and EU27) because EUROSTAT databases do not provide long term information about the historical administrative boundaries of these New Members.

Although a relatively high numbers of countries describe the changes of their administrative boundaries, the attempt does not construct a real temporal database. The most frequent method is to list changes as events (juridical rules) without relations between administrative units states. Historical “communes” database in France, done by the INSEE, describe changes of the local level from 1930 until today. However, this database is not easily workable because of describing textual change
(Figure 3)

ID MUNICIPALITY	DATE OF THE CHANGE	DESCRIPTION
51 261	14/12/1930	Fresnes-lès-Reims become Fresne-lès-Reims.
67 161	21/02/1948	Gottenhausen become Gottenhouse.
12 307	03/12/1952	Curan is created thanks to some areas of Salles-Curan.
51 136	01/03/2006	Châtillon-sur-Marne is separated from Cuisles.
21 551	01/01/2009	Saint-Germain-Source-Seine is merged to Blessey, which become Source-Seine

Figure 3: *Extract of the "Historical French communes' database" done by INSEE*

Danish National statistical Institute has built population time series data based in the current Nuts version (figure3). Earlier 2003 territorial nuts changes data sets are also available.

NUTS VERSION	CODE	NAME	1979	1980	...	2004	2005	2006	...	2009
NUTS 2003	DK001	Copenhagen	505974	498850	...	501664	502362	501158	...	No data
NUTS 2003	DK001	Frederiksberg	88835	88287	...	91721	91886	91855	...	No data
NUTS 2003	DK002	Copenhagen County	629928	627245	...	618407	618237	618529	...	No data
NUTS 2003	DK003	Frederiksborg County	325855	329141	...	373688	375705	378686	...	No data
NUTS 2003	DK004	Roskilde County	199672	202017	...	237089	239049	241523	...	No data
NUTS 2003	DK005	West Zealand County	275985	277833	...	302479	304761	307207	...	No data
NUTS 2003	DK006	Storstrøm County	259445	260081	...	261884	262144	262781	...	No data
NUTS 2003	DK007	Bornholm (excl. Christiansø)	47605	47780	...	43673	43347	43245	...	No data
NUTS 2003	DK008	Funen County	451727	452965	...	475082	476580	478347	...	No data
NUTS 2003	DK009	South Jutland County	248985	249949	...	252936	252980	252433	...	No data
NUTS 2003	DK00A	Ribe County	211492	212624	...	224595	224454	224261	...	No data
NUTS 2003	DK00B	Vejle County	323418	325774	...	355691	358055	360921	...	No data
NUTS 2003	DK00C	Ringkøbing County	261028	262751	...	274830	274574	275065	...	No data
NUTS 2003	DK00D	Århus County	571702	573916	...	653472	657671	661370	...	No data
NUTS 2003	DK00E	Viborg County	230536	231517	...	234659	234434	234896	...	No data
NUTS 2003	DK00F	North Jutland County	479349	481335	...	495669	495068	495090	...	No data
NUTS 2006	DK011	Province København by	No data	No data	...	No data	646986	645875	...	667228
NUTS 2006	DK012	Province Københavns omegn	No data	No data	...	No data	504634	504317	...	508183
NUTS 2006	DK013	Province Nordsjælland	No data	No data	...	No data	436570	440036	...	444215
NUTS 2006	DK014	Province Bornholm	No data	No data	...	No data	43445	43337	...	42659
NUTS 2006	DK021	Province Østsjælland	No data	No data	...	No data	228712	231150	...	233605
NUTS 2006	DK022	Province Vest- og sydsjælland	No data	No data	...	No data	577242	580361	...	587647
NUTS 2006	DK031	Province Fyn	No data	No data	...	No data	476580	478347	...	484346
NUTS 2006	DK032	Province Sydjylland	No data	No data	...	No data	707171	707504	...	715321
NUTS 2006	DK042	Province Østjylland	No data	No data	...	No data	792934	798671	...	820558
NUTS 2006	DK041	Province Vestjylland	No data	No data	...	No data	419853	421054	...	427174
NUTS 2006	DK050	Province Nordjylland	No data	No data	...	No data	577278	576807	...	580515

Figure 4: *Population 1979-2009 in the NUTS3 of Denmark (both NUTS 2003 and 2006 version, according to the Danish National Statistic Institute)*

Italian National statistical Institute proposes another example of time series handling. It provides information related to national territorial changes and its correspondence with regional (European) level (figure 5).

However, the Italian example may be considered as the best attempt because it provides much information to describe the change of administrative units: type of change, juridical texts and relation between versions of unit (genealogy). It allows also to analyse the effect of national administrative boundaries change on the NUTS geometry and hierarchy.

Cartografia: confini amministrativi e dei sistemi locali del lavoro

Censimento 2001, 31 dicembre 2008 e 1 gennaio 2010

L'Istat fornisce la **versione generalizzata dei confini amministrativi** (Regioni, Province e Comuni) **e dei sistemi locali del lavoro**. Gli strati informativi sono costituiti da tre livelli gerarchici a copertura nazionale per i limiti di regione, provincia e comune.

I dati sono in formato shapefile; tale formato dati è stato reso pubblico già da parecchi anni ed utilizzato per lo scambio di dati in ambito GIS (Geografic Information System). I dati cartografici forniti sono nel sistema di riferimento ED_1950_UTM zona 32; il dettaglio tecnico della proiezione è riportato nel file apposito, associato a ciascun file geografico.

La scala dei dati non è certificabile uniformemente dall'Istat, in quanto le basi di acquisizione utilizzate provengono da fonti e scale differenti, che variano dal 1:5.000 in ambito urbano fino 1:25.000 in ambito extraurbano. I dati sono stati inoltre generalizzati e semplificati nelle forme geometriche, per renderne disponibile una versione da utilizzare agevolmente, per la creazione di cartografia simbolica o di riferimento a livello nazionale.

I file geografici di regioni, province e comuni, già pubblicati alla data del Censimento del 2001, sono stati aggiornati comprendendo le variazioni (territoriali e di nome) intercorse tra la data del Censimento 2001 e il 31 dicembre 2008 e successivamente al 1 gennaio 2010. Sono stati quindi acquisiti i codici e le denominazioni delle tre nuove province (Monza e della Brianza, Fermo e Barletta-Andria-Trani) e le nuove codifiche dei comuni ad esse appartenenti. Inoltre sono state acquisite anche altre variazioni comunali (si veda la Struttura dei dati). Per una più approfondita descrizione delle variazioni amministrative e territoriali intervenute successivamente alla data del Censimento del 2001 si può consultare la pagina web con i codici dei comuni, delle province e delle regioni.

Sono inoltre forniti i confini dei Sistemi locali del lavoro e delle NUTS2 (Nomenclature of territorial units for statistics), che rappresenta l'articolazione ufficiale europea del territorio di livello 2 (Regioni e le Province autonome per l'Italia) finalizzata alla produzione di statistiche.

descrizione **dati**
- **Struttura dei dati**

confini amministrativi

Censimento 2001
- **Regioni**
- **Province**
- **Comuni**

31 dicembre 2008
- **Regioni**
- **Province**
- **Comuni**

1 gennaio 2010
- **Regioni**
- **Province**
- **Comuni**

altra **cartografia**
- **Sistemi locali del lavoro**
  Censimento 2001
- **NUTS2**
  2008

per informazioni

Informazioni territoriali e sistema informativo geografico
tel. 06 4673.4861
email int@istat.it

Figure 5: *Web Interface of Italian National statistical Institute*

Furthermore the National Statistical Institutes, other government departments such as the interior ministers could publish documents related to national territorial changes. The Danish ministry of the interior and social affairs has published guide paper to understand the local boundaries reforms, which have consequences on the definition of NUTS3 and NUTS2 units (figure 6). This kind of document is not usually available.

Figure 6: *Explanation of the Danish Reform on Territorial units*

1.4 Conclusion of the first part

Building NUTS temporal database is very complex for the following main reasons:

- NUTS changes vary greatly from country to country due to their different structures;
- The available data are very heterogeneous;
- Data quality varies largely;
- Lack of good practice and experiences of handling territorial boundaries.

Based on a compilation of several sources and methods, we will propose formalization adapted to the specificity of NUTS.

2 Nuts changes knowledge: from elementary to systemic approach of territorial changes

The benchmarking of sources and experiences has showed the complexity of NUTS territorial changes. Following Swianczny, (2000) who stated that: “In order to create a truly time integrative GIS, the focus has to change from spatial to temporal and from analyzing changes between events to the analysis of the change itself”, we propose an appropriate approach to formalize the Nuts changes. This approach will be based on an explicit description of changes.

Based on the characteristics of NUTS, determined by regulations, we can distinguish the following changes: name, geometry, code and hierarchical level. These changes can occur at the same time because territorial changes are very complex. The changes analysis may be presented from tow angles: elementary approach and systemic approach.

2.1 Elementary changes

Elementary approach consists in describing the change of territorial units one undependably of the others (figure 7).

- Change of name: two cases can be distinguished. If the unit in question belongs to two levels (it is at the same time a NUTS 2 and a NUTS 3) the change of name can concern either one or the two levels.

  1999: BE31 Brabant Wallon
  2003: B310 Arrondissement Nivelles

- Change of code: it may result from different types of territorial modifications: political decision, territorial reorganisation. In the first case, we can list many changes of NUTS 2 level in 2003. In the second case, we point the code changing of Italian NUTS 2 and NUTS 3 units since 2003 due to regional reorganisation of NUTS 1.
- Change of geometry: It is the most complicated change type. Generally, the deformation of a spatial unit can be done in three different ways: the loss of area, the gain of area, or the redistribution of boundaries even while keeping the same area value. Most of the time, there is no relation between the different versions of the NUTS, like in Poland (figure 9). This kind of situation makes the harmonisation very difficult to implement.
- Change of hierarchy: it consists to the change of the territorial reference or of belonging. As shown by the figure 7, Italian Nuts 2 level have changed their Nuts 1 units of belonging because of territorial reforms of nuts 1 level in 2003.

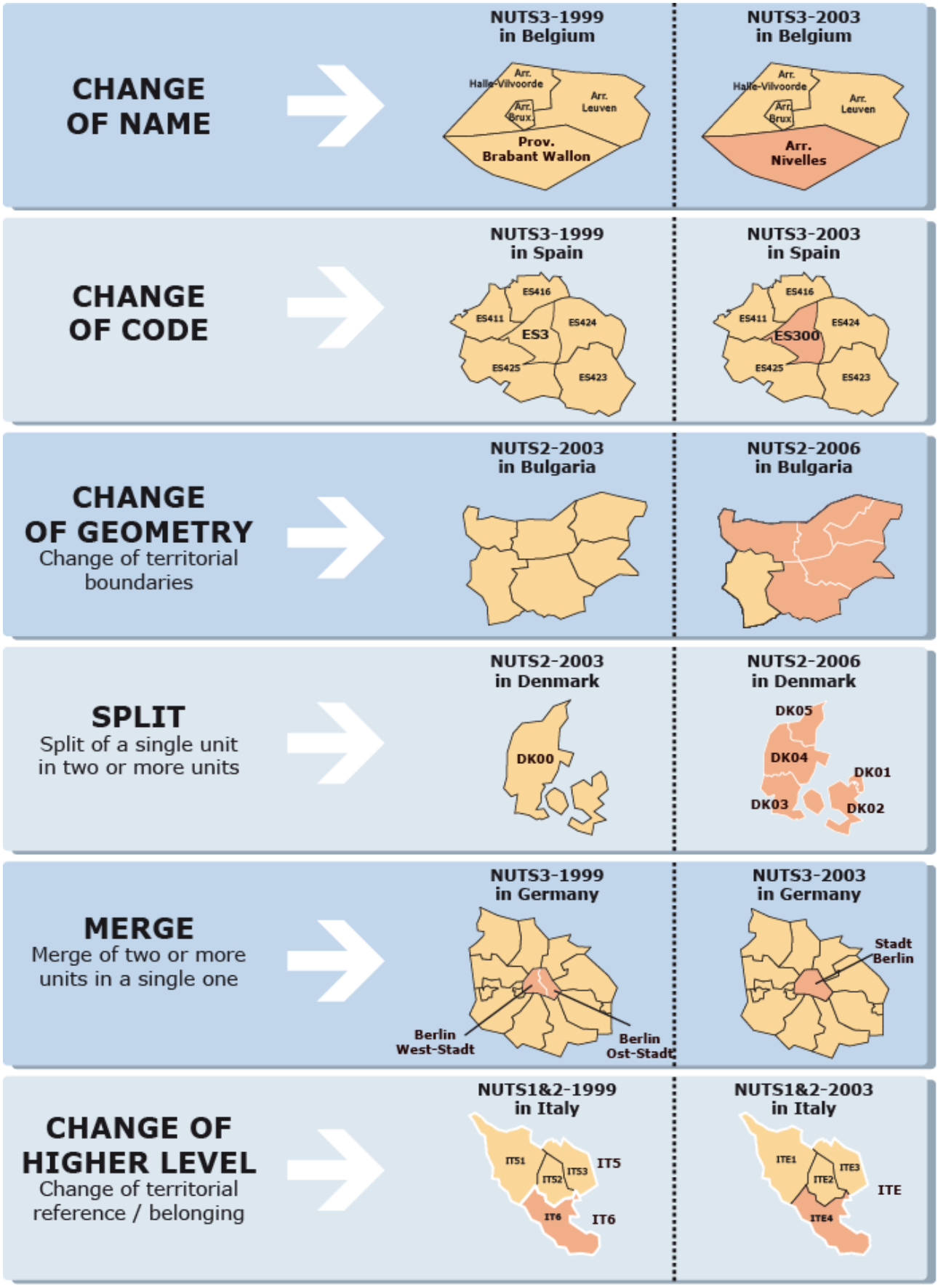


Figure 7: *Examples of elementary Nuts changes*

CODE 1999	CODE 2003	Name 1999	Name 2003	Status
ES3	ES3	Comunidad de Madrid	Comunidad de Madrid	No change
ES3	ES30	Comunidad de Madrid	Comunidad de Madrid	Changed
ES3	ES300	Comunidad de Madrid	Comunidad de Madrid	Changed
IT1	ITC	Nord Ovest	Nord Ovest	Changed
IT11	ITC1	Piemonte	Piemonte	Changed
IT111	ITC11	Torino	Torino	Changed
IT112	ITC12	Vercelli	Vercelli	Changed
IT113	ITC13	Biella	Biella	Changed
IT114	ITC14	Verbano-Cusio-Ossola	Verbano-Cusio-Ossola	Changed
IT115	ITC15	Novara	Novara	Changed
IT116	ITC16	Cuneo	Cuneo	Changed
IT117	ITC17	Asti	Asti	Changed
IT118	ITC18	Alessandria	Alessandria	Changed

Figure 8: *Examples of changing of unit's code in Italy and Spain from 1999 to 2003*

Considering the characteristics of Nuts units which are mentioned, limiting the investigation to basic (elementary) changes does not allow to reconstruct genealogy of Nuts. To achieve this, we have considered nuts structure as a system and we have focused on the relationship between changes.

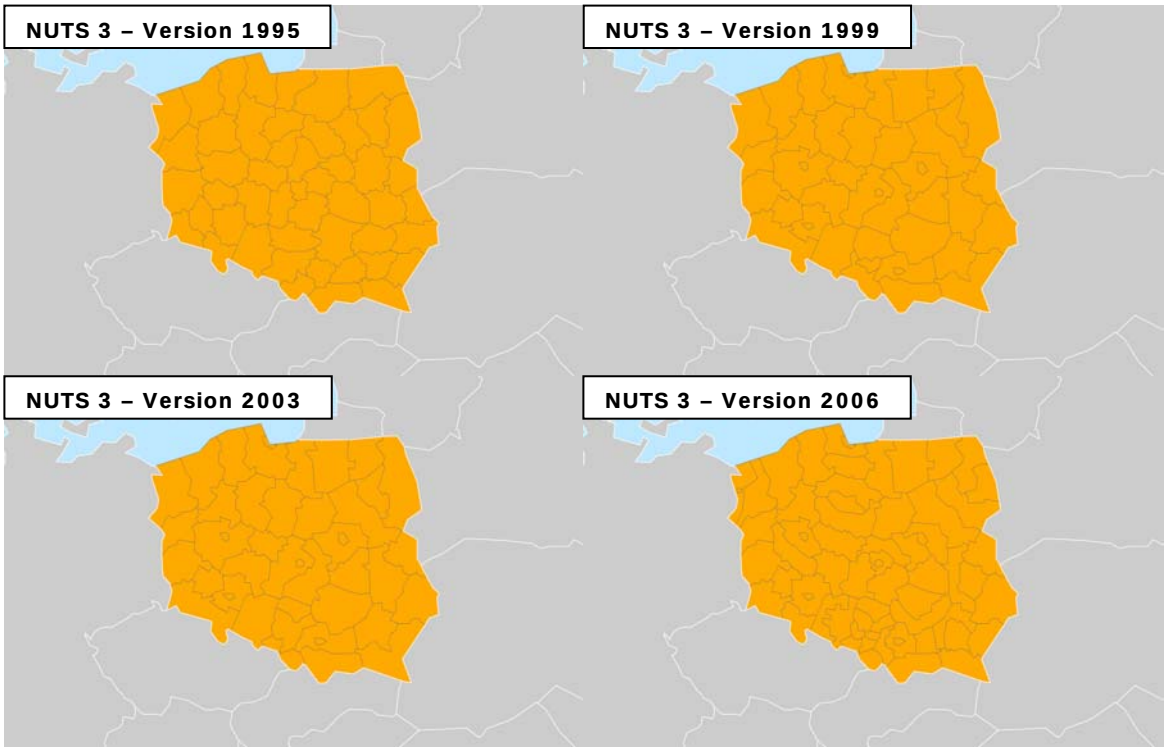


Figure9: *Complex geometry change in Polish nuts 3 units*

2.2 Systemic approach of NUTS changes

Because the formalization of Nuts changes is complex and has to take into account several parameters (type of changes, temporal period and scalar dimension), we propose a cubic model which emphasizes the relationships between these parameters. This means that the result of territorial modifications depends on the type of changes (name, code, and geometry), the period of time and the territorial level. Thus, we used the concept of systemic approach.

The systemic conception emphasizes the relationships between the changes:

- A change affecting a unit may have implications on the other units.
- A change happening on a given level may have implications on the other levels.

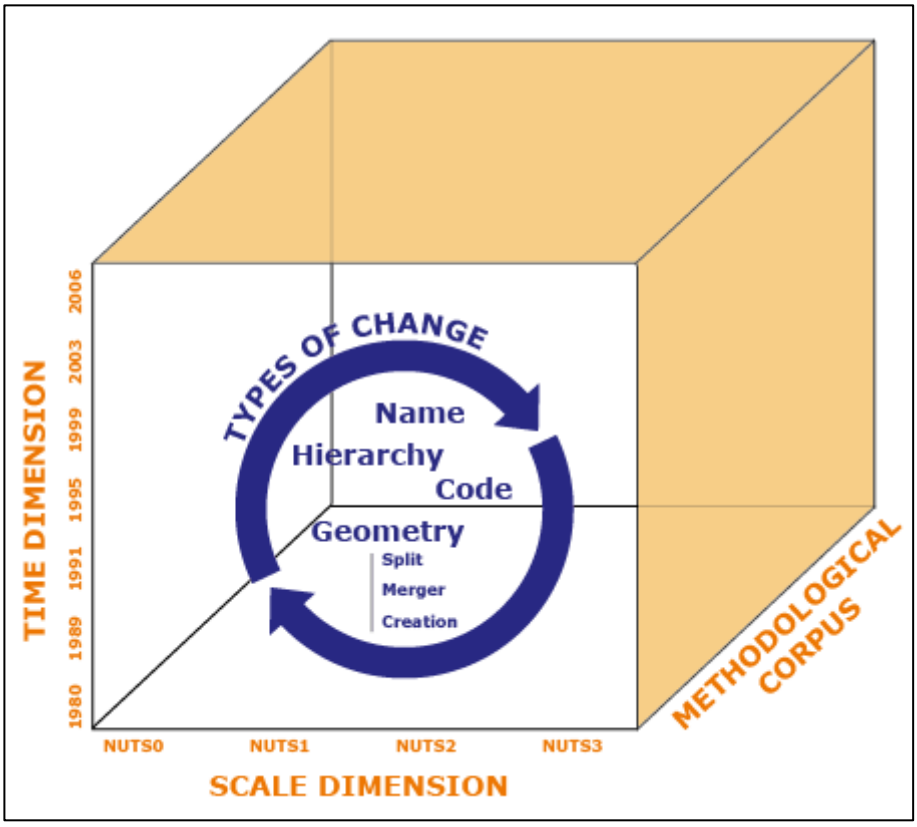


***Figure 10**: Cube structure of NUTS formalisation*

We demonstrate our approach through the analysis of the example of Italian Nuts between 1995 and 2006 (figure 10):

Concerning the temporal dimension, tow orders can be distinguished:

- The period of time determines the degree of discontinuity of the data sets. Indeed, the extension of the period increases the discontinuity because of the complexity of changes that may have occurred. In the case of Italian Nuts, if we consider the whole period (1995-2006), we can see a big discontinuity in the data sets. However, the data set will be complete between 2003 and 2006.

- The building of time series data could be considered in either a prospective or retrospective territorial approach. The prospective view consists in transposing old data sets onto a recent version of Nuts (data 1995 onto Nuts 2006 for example). However, the retrospective view consists in transposing recent data sets onto old Nuts versions (data 2006 to Nuts 1995). Each approach requires a different methodology. For example, 2003 version data should be disaggregated to be integrated in Italian Nuts 1 level 1999 version. However, the 1995 version data should be aggregated.
- As for the Scalar dimension, it is linked to the hierarchical structure of Nuts (Nuts 1 level is subdivided into Nuts 2 level which is in turn subdivided into Nuts 3 level). In fact, the changes which occur in higher levels (1 and 2) have various consequences on lower territorial levels. As it was shown by the figure 7, the territorial reform of Italian Nuts 1 level in 2003, consisting in merging and changing codes of units, has caused a change of codes of Nuts 2 and Nuts 3 units. Moreover, reforms of higher Nuts levels (Nuts 1 and Nuts 2) could have more complex implications on lower levels. The creation of 5 new Nuts 2 units in Denmark in 2003, by splitting DK00, has caused very complex territorial reorganization on Nuts 3 level units (Figure 7).

Regarding Relationships between changes, the change of geometry is a determining factor in the time series data building process. On the whole, three types of unit spatial changes can be identified: the loss of area, the gain of area and deformation (which means territorial boundaries redistribution without loss of area). Based on these primary types of changes, we have developed a conceptual corpus to describe further types of changes (dictionary of changes). The dictionary of changes aims to answer the following questions: what happened? How did it happen? And what were the results?
For example, the Danish territorial reforms in 2003 could be described as follows:

Nuts 1 level: there are no changes

Nuts 2 level:
- The Split of DK00 (change of geometry)
- Official disappearance of DK00
- Creation of 5 new Nuts 2 units: DK01, DK02, DK03, DK04 and DK05

Nuts 3 level:
- Change of code which means change of belonging to a superior unit (hierarchy): Funy DK008 (2003) and DK031 (2006), Bornholm DK007 (2003) and DK014 (2006)
- Complex changes of geometry for the rest of units which have caused the disappearance of 12 units and the creation of 10 new units

3 Building historical database of territorial changes: from conceptual approach to operational solutions

This process of formalization of Nuts changes is not an end in itself. Aiming to build continuous time series data, its first objective is to understand how spatial units have been changed. Information describing Nuts changes represent a very useful metadata. In this final section of this technical report, we will examine how these results could be presented to the users. A first application made on cohesion reports will also be presented.

The results of this exploration may be presented in different ways depending on the users' needs. The examples that we present illustrate the progress of the complexity of the issue of nuts changes formalization. Location of change, identification of change and genealogy (lineage) of spatial units are the most important information that could be allowed to the users at this stage of work.

3.1 Improved Snapshot model (presentation)

One of the simplest spatio-temporal data models is the snapshot model (Armstrong 1988). Temporal information was incorporated into this spatial data model by time-stamping layers. Every layer shows the states of NUTS at different times without explicit relation between versions. Changes and genealogy cannot be depicted. This prevents harmonized database to be built.

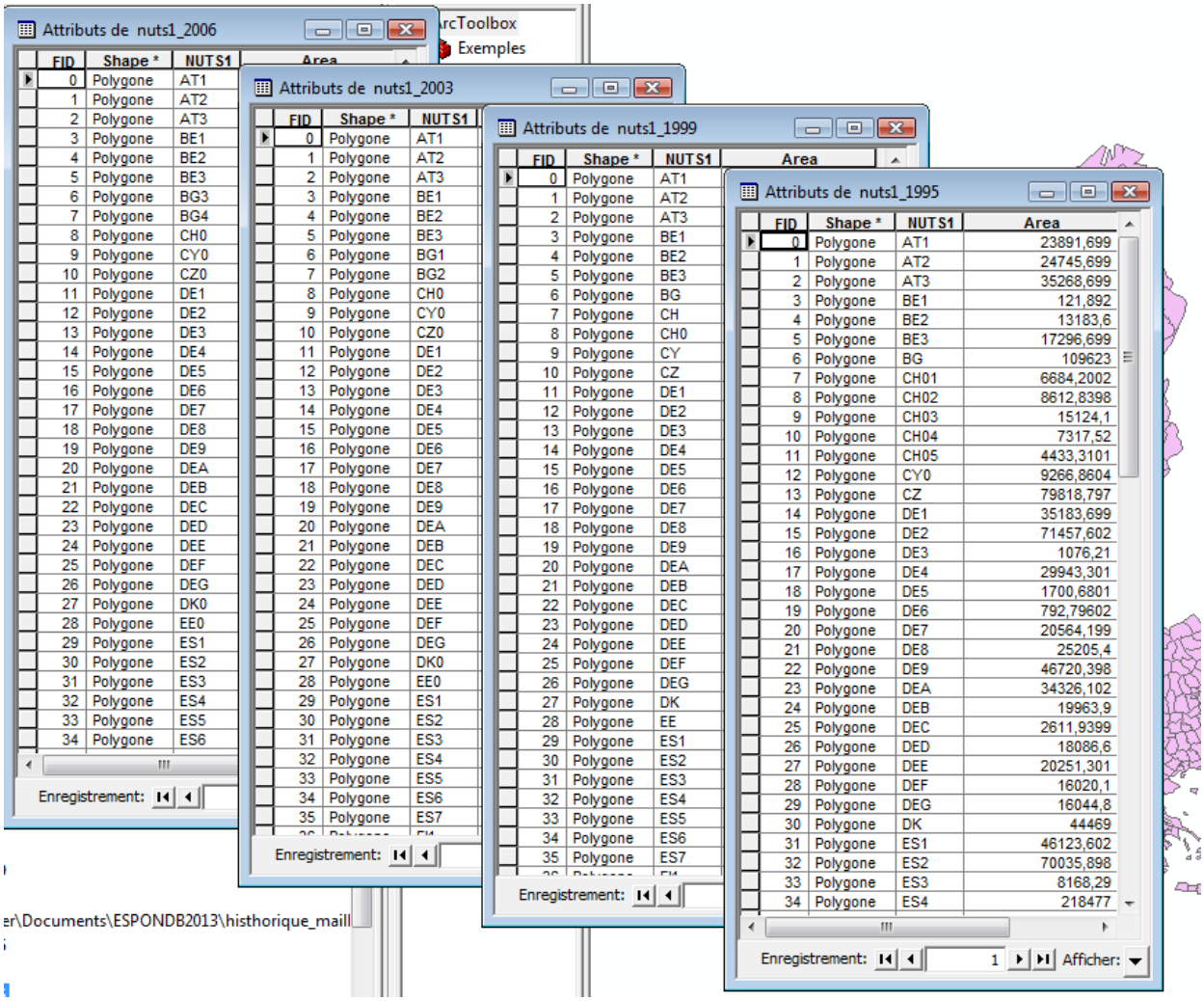


Figure 11: *Snapshot of NUTS versions between 1995 and 2006 (example of Nuts 1 level)*

This method could be improved by location of changes occurred. A binary code was attributed to the state of Nuts: 0 for stability and 1 in case of change. This simple coding can identify potential discontinuities in the time series. The figure 11 shows an extract of tables covering all Nuts levels.

Code 2006	NUTS0	NUTS level 1	NUTS level 2	NUTS level 3	Change	Change since 2003
DK0		DANMARK			same	0
DK01			Hovedstaden		changed	1
DK011				Byen København	changed	1
DK012				Københavns omegn	changed	1
DK013				Nordsjælland	changed	1
DK014				Bornholm	changed	1
DK02			Sjælland		changed	1
DK021				Østsjælland	changed	1
DK022				Vest- og Sydsjælland	changed	1
DK03			Syddanmark		changed	1
DK031				Fyn	changed	1
DK032				Sydjylland	changed	1
DK04			Midtjylland		changed	1
DK041				Vestjylland	changed	1
DK042				Østjylland	changed	1
DK05			Nordjylland		changed	1
DK050				Nordjylland	changed	1

Figure 12: *Extract of table of location of changes: Danish Nuts between 2006 and 2003*

The figure 12 shows more developed stage of change describing process. Identifying kinds of change and the consequences, even in scalar dimension and relationship between changes, were added to the location of change.

- The column Check: change/no change
- The column Change: identify the initial change
- The column Life: indicates if units exists (E) or did not exist (D: deleted)
- The column Hierarchy : change of geometry (0/1)
- The column Geometry: specifies the change of geometry of the unit

Semantic describing changes (code) should be improved. A table of Metadata will be established when this semantic description will be consolidated.

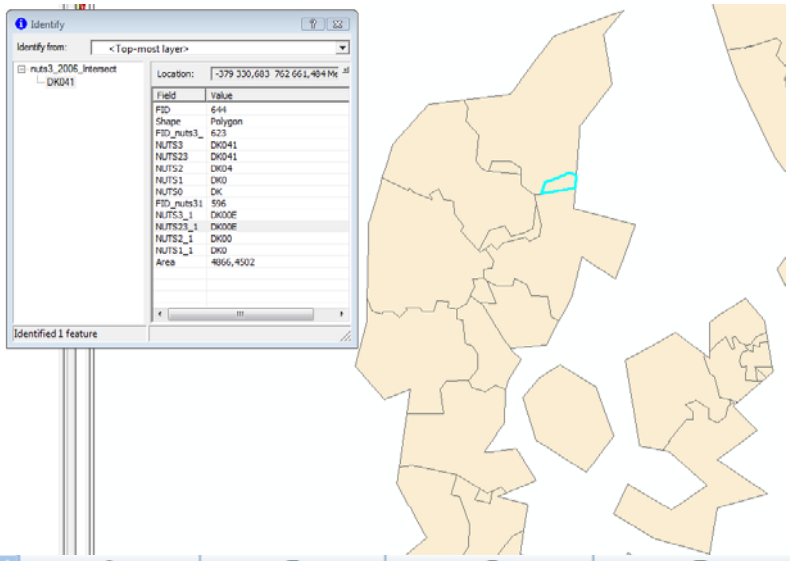
Code 2003	Code 2006	Country	NUTS level 1	NUTS level 2	NUTS level 3	CHECK	CHANGE	LIFE	HIERARCHY	GEOMETRY
	DK01			Hovedstaden		changed	GEOM	N	0	GEOM+
	DK011				Byen København	changed	GEOM	N	0	GEOM+
	DK012				Københavns omegn	changed	GEOM	N	0	GEOM+
	DK013				Nordsjælland	changed	GEOM	N	0	GEOM+
DK007	DK014				Bornholm	changed	GEOM	N	0	GEOM+
	DK02			Sjælland		changed	GEOM	N	0	GEOM+
	DK021				Østsjælland	changed	GEOM	N	0	GEOM+
	DK022				Vest- og Sydsjælland	changed	GEOM	N	0	GEOM+
	DK03			Syddanmark		changed	GEOM	N	0	GEOM+
DK008	DK031				Fyn	changed	GEOM	N	0	GEOM+
	DK032				Sydjylland	changed	GEOM	N	0	GEOM+
	DK04			Midtjylland		changed	GEOM	N	0	GEOM+
	DK041				Vestjylland	changed	GEOM	N	0	GEOM+
	DK042				Østjylland	changed	GEOM	N	0	GEOM+
	DK05			Nordjylland		changed	GEOM	N	0	GEOM+
	DK050				Nordjylland	changed	GEOM	N	0	GEOM+
DK00				*Danmark*		changed	GEOM	D	0	0
DK001					*København og Frederiksberg kommuner*	changed	GEOM	D	0	0
DK002					*Københavns amt*	changed	GEOM	D	0	0
DK003					*Frederiksborg amt*	changed	GEOM	D	0	0
DK004					*Roskilde amt*	changed	GEOM	D	0	0
DK005					*Vestsjællands amt*	changed	GEOM	D	0	0
DK006					*Storstrøms amt*	changed	GEOM	D	0	0
DK009					*Sønderjyllands amt*	changed	GEOM	D	0	0
DK00A					*Ribe amt*	changed	GEOM	D	0	0
DK00B					*Vejle amt*	changed	GEOM	D	0	0
DK00C					*Ringkøbing amt*	changed	GEOM	D	0	0
DK00D					*Århus amt*	changed	GEOM	D	0	0
DK00E					*Viborg amt*	changed	GEOM	D	0	0
DK00F					*Nordjyllands amt*	changed	GEOM	D	0	0

Figure 13: *Extract of table of identification of changes: Danish Nuts between 2006 and 2003*

3.2 The space-time composite model: reconstructing genealogy of Nuts versions

This model was proposed by Langran (1992). It consists to decompose NUTS, geometries through time by intersecting different versions (1995-1999-2003-2006). Small spatio-temporal entities (polygons) are created as the results of this intersection. It represents the lowest common denominator.

As it is shown by the figure 13 the intersection of geometries of Danish NUTS 3 2003 and 2006 versions results 22 polygons. The belonging to NUTS is defined as a temporal attribute. For example, the polygon n°11 (table and selected in the maps) belong to DK00E unit in 2003 and to DK041 unit in 2006. In fact, genealogy of units can be deducted and may be represented by a graph.



	A	B	C	D
1	Polygons	NUTS3-2003	NUTS3-2006	Area
2	p1	DK001	DK011	50,8357
3	p2	DK002	DK011	32,7178
4	p3	DK002	DK012	321,644
5	p4	DK002	DK013	71,7531
6	p5	DK003	DK013	774,849
7	p6	DK004	DK021	863,461
8	p7	DK005	DK022	3171,58
9	p8	DK006	DK022	4059,64
10	p9	DK007	DK014	614,29
11	p10	DK00C	DK041	5227,71
12	p11	DK00E	DK041	4866,45
13	p12	DK00B	DK041	183,677
14	p13	DK00D	DK042	4056,62
15	p14	DK00E	DK042	335,029
16	p15	DK00B	DK042	1459,76
17	p16	DK00D	DK050	186,306
18	p17	DK00E	DK050	147,482
19	p18	DK00F	DK050	6069,42
20	p19	DK008	DK031	2659,36
21	p20	DK00A	DK032	4063,59
22	p21	DK009	DK032	3891,81
23	p22	DK00B	DK032	1351,82

Figure 14: *Space time composite model*

We have improved this cartographic method by quantifying lineage of Nuts. Information collected through the exploration of sources has allowed us to establish a relationship between spatial units. We tried to quantify this genealogy by calculating the proportion of area transferred in case of change. The proportion of the population transferred will also be tested. This method is very useful for estimating missing data because of territorial changes

The figure 14 illustrates the table built. It does not yet cover all the ESPON area because of lack of accurate data especially for the new member countries.

NAME	code 2006	% Geom	code 2003	% Geom	code 1999	% Geom	code 1995	NAME
Denmark	0	0	DK00	100	DK00	100	DK00	Danmark
	DK01	4,2	DK00	4,2	DK00	4,2	DK00	Hovedstaden
	DK02	18,2	DK00	18,2	DK00	18,2	DK00	Sjælland
	DK03	26,9	DK00	26,9	DK00	26,9	DK00	Syddanmark
	DK04	36,3	DK00	36,3	DK00	36,3	DK00	Midtjylland
	DK05	14,4	DK00	14,4	DK00	14,4	DK00	Nordjylland

Figure 15: *Extract of table of genealogy of Nuts: Danish Nuts2 level between 2006 and 1995*

Another simple approach is to tag every object (NUTS) with a pair of timestamps, one for the time of creation and one for the time of cessation. Current objects have their cessation time given by a special value “NOW”, “CURRENT”, or “NULL”.

To conclude this section, we emphasize that these applications were developed in the following time periods: 2006-2003; 2006-1999; 2006-1995, 2003-1999; 2003-1995 and 1999-1995.

3.3 Towards a NUTS changes mapping

A cartographic display of the results of formalization would be very useful for the better understanding of Nuts territorial changes. However the visualization of changes is as complex as the formalization and it depends on all parameters prsented in the previous section. We propose the following thoughts.

3.3.1 A general view

The map of static units distinguishes static units (no change observed) from changed units as shown by the figure 15. Yellow units are static. However, white units have been changed since 1995.

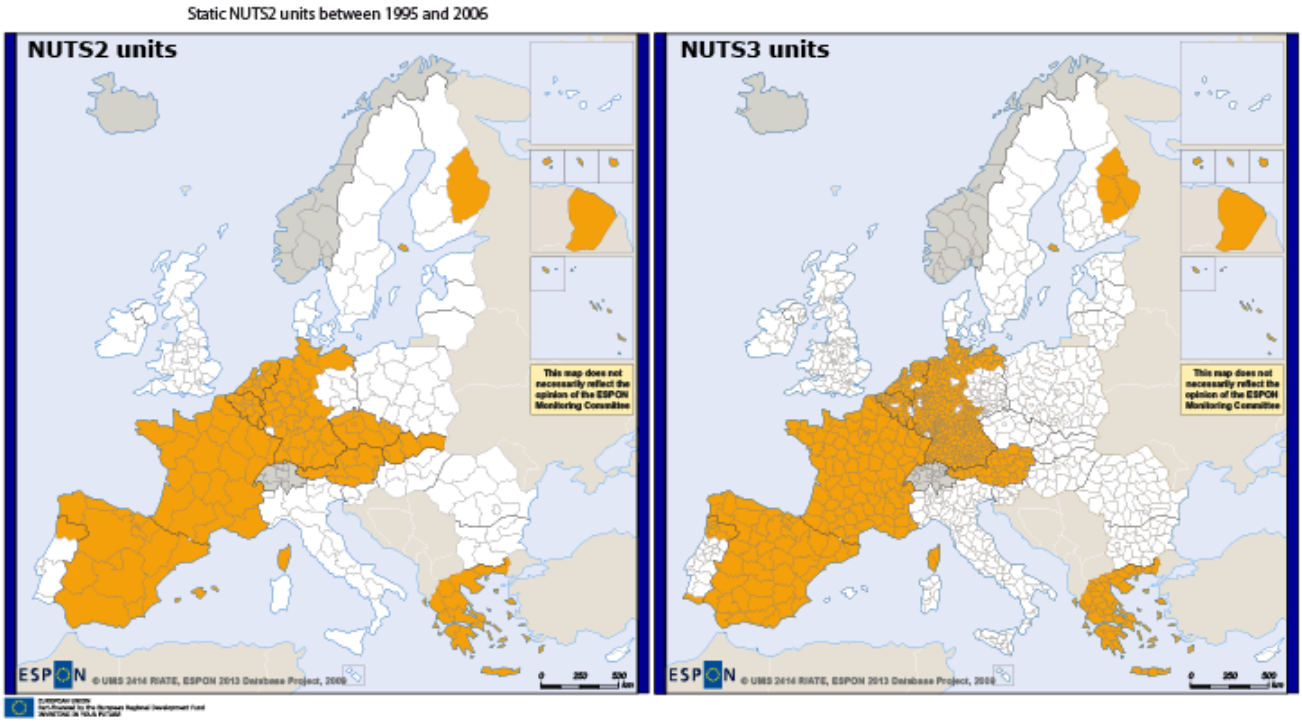


***Figure 16*:** Static NUTS 2 and Nuts 3 units between 1995 and 2006 (all criteria).

3.3.2 Typology of changes

These maps give more information about changes. It aims to emphasize on the diversity of changes. The most important kinds of changes are: name, code, geometry and hierarchy.

3.3.3 Maps by times intervals

The interval of time (period) is a very important parameter for building time series database. Naturally, the enlargement of the covered period makes the building process more difficult. Mapping changes by period of time could help users in their building time series process.

3.3.4 Examples of lineage (genealogy) visualization

The attempt is to propose a graphic visualization of the units' genealogy. We selected the most demonstrative examples in order to emphasize in the complexity of Nuts changes. These patterns are in an early stage implementation and should be improved.

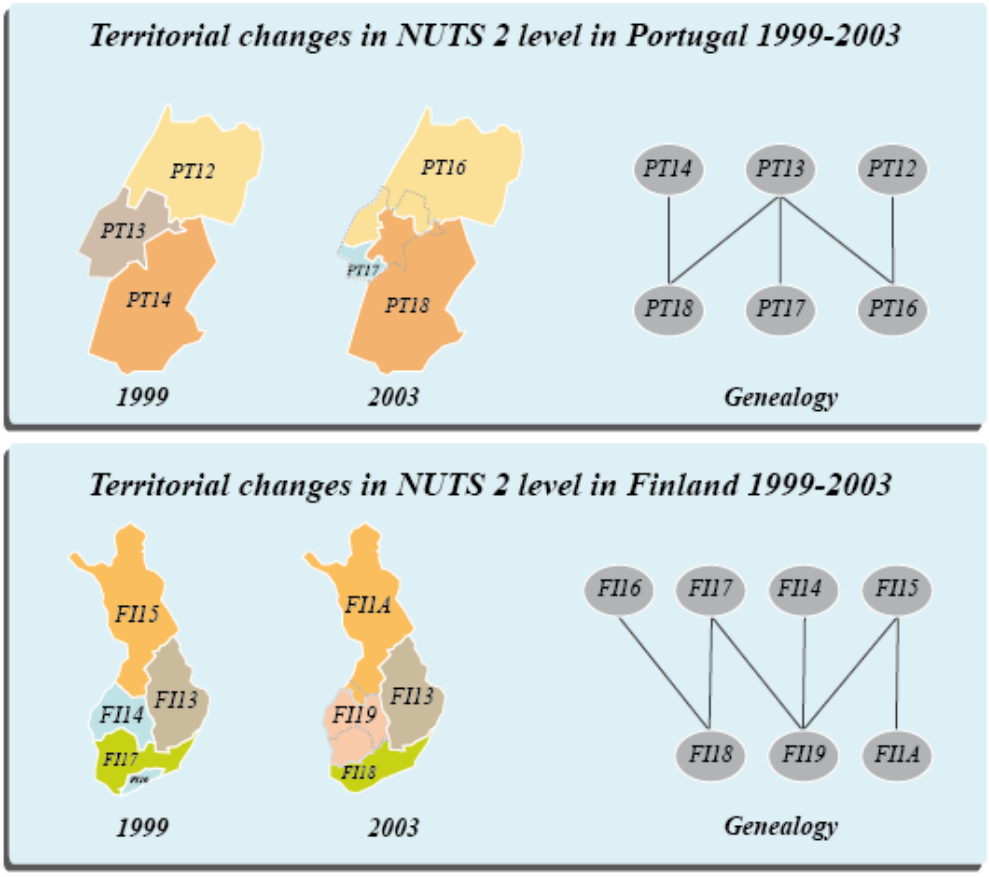


Figure 17: examples of NUTS genealogy visualization

3.4 Tables structure of NUTS changes

Table structuring of NUTS changes is a very challenging topic. The aim of this topic is to provide a very friendly using data about territorial changes. The complexity of NUTS changes requires simplifying the information to be useful and comprehensive by a large public dealing with time series data implementation.

We propose a several ways of structuring change tables. Theses tables are complementary and propose different visions of territorial changes that meet to needs of many users in terms of spatiotemporal databases.

3.4.1 Changes location tables

It is a binary description of NUTS situation between each version: 1995-1999, 1999-2003 and 2003-2006 (more time intervals combinations are possible). The aim of this table is to locate changes regarding all analysis criteria (name, code, geometry and belonging to a higher level).

Code 2006	NUTS0	NUTS level 1	NUTS level 2	NUTS level 3	Change since 2003
DK	DANMARK				0
DK0		DANMARK			0
DK01			Hovedstaden		1
DK011				Byen København	1
DK012				Københavns omegn	1
DK013				Nordsjælland	1
DK014				Bornholm	1
DK02			Sjælland		1
DK021				Østsjælland	1

Figure 18: extract of change location table (Danish Nuts between 2003 and 2006)

3.4.2 Changes typology tables

These tables give more detailed information about changes. The identification of nuts changes is very helpful for understanding territorial changes. It could contribute to the database harmonization. Regarding to nuts changes formalization presented in the previous section (2.2), we identified four kinds of change: name, code, geometry and belonging to a higher level.

Code 2003	Code 1999	NUTS level 1	NUTS level 2	NUTS level 3	OP	CHANGE	LIFE	HIERARCHY	GEOMETRY
DE3	DE3	BERLIN			0	0	E	0	0
DE30	DE3		Berlin		CODE	CODE	E	1	0
0	DE300			Berlin	MERGER	NO	D	0	NO
DE301	0			Berlin-West, Stadt	MERGER	NO	N	0	GEOM+
DE302	0			Berlin-Ost, Stadt	MERGER	NO	N	0	GEOM+
DE4	DE4	BRANDENBURG			SPLIT	H	D	1	NO
0	DE4		Brandenburg		SPLIT	NO	D	0	NO
DE41	0			Brandenburg - Nordost	SPLIT	NO	N	0	GEOM+
DE42	0			Brandenburg - Südwest	SPLIT	NO	N	0	GEOM+
DE411	DE403			Frankfurt (Oder), Kreisfreie Stadt	SPLIT	CODE	E	2	0
DE412	DE405			Barnim	SPLIT	CODE	E	2	0
DE413	DE409			Märkisch-Oderland	SPLIT	CODE	E	2	0
DE414	DE40A			Oberhavel	SPLIT	CODE	E	2	0
DE415	DE40C			Oder-Spree	SPLIT	CODE	E	2	0
DE416	DE40D			Ostprignitz-Ruppin	SPLIT	CODE	E	2	0
DE417	DE40F			Prignitz	SPLIT	CODE	E	2	0
DE418	DE40I			Uckermark	SPLIT	CODE	E	2	0

Figure 19: extract of change typology table (German Nuts between 2003 and 1999)

3.4.3 "Units life" tables

The aim of these tables is to provide a general view of the official existence of units in each nuts version. The use of legal definition of statistical units (Official journal of the European Union) avoided the several interpretations of the notion of "life". We regroup in one table all units from 1995 to 2006 and we indicate each version, by binary description, if it exists or it was deleted.

Code 2003	Code 1999	NUTS level 1	NUTS level 2	NUTS level 3	OP	LIFE
DE3	DE3	BERLIN			0	1
DE30	DE3		Berlin		CODE	1
0	DE300			Berlin	MERGER	0
DE301	0			Berlin-West, Stadt	MERGER	1
DE302	0			Berlin-Ost, Stadt	MERGER	1
DE4	DE4	BRANDENBURG			SPLIT	0
0	DE4		Brandenburg		SPLIT	0
DE41	0			Brandenburg - Nordost	SPLIT	1
DE42	0			Brandenburg - Südwest	SPLIT	1
DE411	DE403			Frankfurt (Oder), Kreisfreie Stadt	SPLIT	1
DE412	DE405			Barnim	SPLIT	1
DE413	DE409			Märkisch-Oderland	SPLIT	1
DE414	DE40A			Oberhavel	SPLIT	1
DE415	DE40C			Oder-Spree	SPLIT	1
DE416	DE40D			Ostprignitz-Ruppin	SPLIT	1
DE417	DE40F			Prignitz	SPLIT	1
DE418	DE40I			Uckermark	SPLIT	1

Figure 20: extract of "Units life" table (German Nuts between 2003 and 1999)

3.4.4 "genealogy units" tables

It is the most complicated table and the most innovative. It provides a conversion keys between nuts versions in order to harmonize temporal databases. To emphasize on the lineage of units, we have defined tow possible kinds of proportionalities: area and population. The proportionality of area transferred is very useful for the conversion of environmental indicators (figure 15). However the proportionality of population is useful for social economic indicators conversion.

4 Example of time series data building process: the European regional cohesion indicators

This section presents a harmonization use-case based on the formalization previously developed. It emphasizes on the contribution of time series data building and harmonization to the thematic analysis. The analysis of the European regional cohesion spatio-temporal evolution offers a very interesting example of methodological demonstration of the time-series building process. The analysis focused on data attached to the three latest reports on territorial cohesion, which include all methodological difficulties such as missing values and mixture of NUTS versions.

4.1 Data collection and temporal series maps

Based on the results of the exploration of nuts changes, the first step of the time series data building process is to collect data and to identify the latest version of nuts. The aim of this work is to produce thematic maps based on the 2003 and 2006 Nuts version (figure 16 and figure 17). These data sets were delivered at ESPON Coordination Unit in November (ESPON DB Update) and were presented at ESPON seminar in Malmö (2-4 December 2009).

id	level	name	NUTS_VER	source
TEMPORAL_START			1999	
TEMPORAL_END			2006	
BE	NUTS0	België/Belgique	2006	1
BE10	NUTS2	Région de Bruxelles-Capitale / Brussels Hoofdstedelijk Gewest	2006	1
BE2	NUTS1	Vlaams Gewest	2006	1
BE21	NUTS2	Prov. Antwerpen	2006	1
BE22	NUTS2	Prov. Limburg (BE)	2006	1
BE23	NUTS2	Prov. Oost-Vlaanderen	2006	1
BE24	NUTS2	Prov. Vlaams-Brabant	2006	1
BE25	NUTS2	Prov. West-Vlaanderen	2006	1
BE3	NUTS1	Région Wallonne	2006	1
BE31	NUTS2	Prov. Brabant Wallon	2006	1
BE32	NUTS2	Prov. Hainaut	2006	1
BE33	NUTS2	Prov. Liège	2006	1
BE34	NUTS2	Prov. Luxembourg (BE)	2006	1
BE35	NUTS2	Prov. Namur	2006	1
BG	NUTS0	Bulgaria	2006	1

Figure 21: Nuts version identification of spatial units used by the 4th Cohesion report

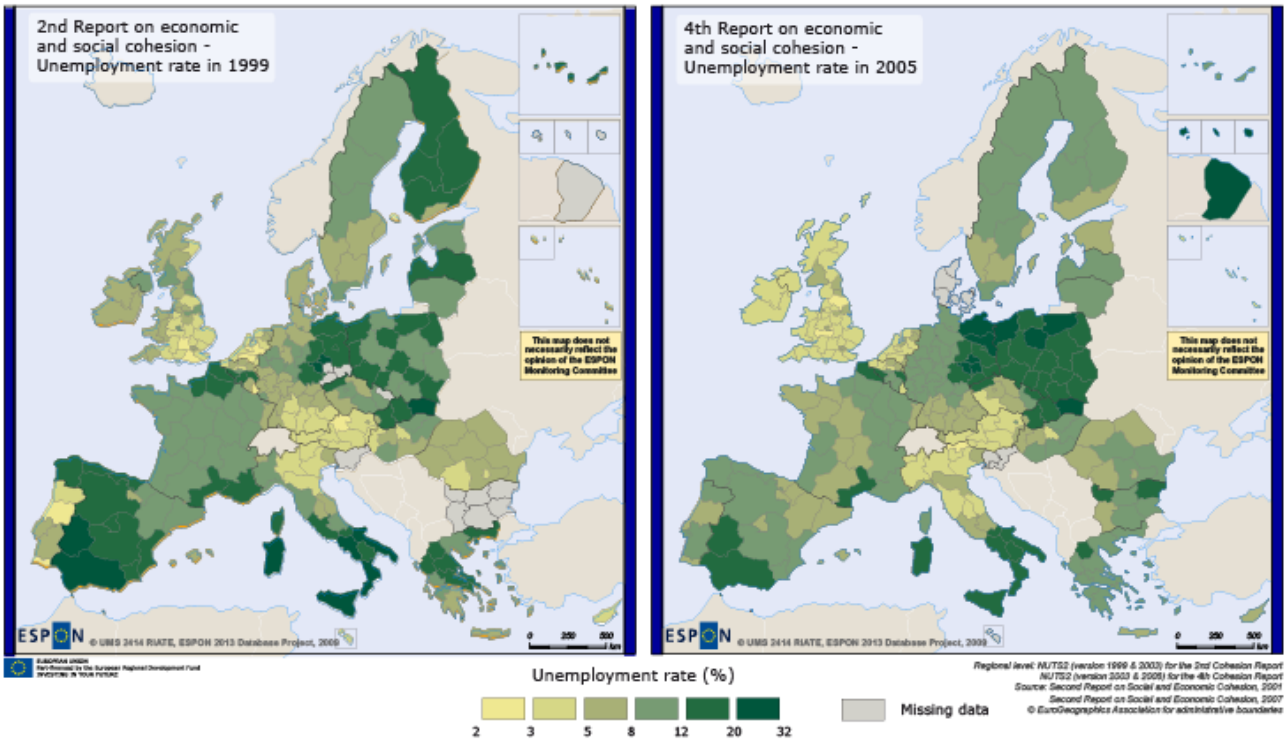


Figure 22: Mapping the Unemployment rate in 1999 and 2005

This task aims to provide data users and policy makers with maps allowing the comparison between two situations of unemployment in the European regions. Mapping is an important added value.

4.2 Building a harmonized temporal data base for the European regional cohesion analysis

The harmonization of time series data is a fundamental step for a more advanced temporal analysis and mapping. The built of time-series data can be done by two complementary approaches: The ESTI framework for estimating missing values is one of them1 and the modelling of NUTS changes.

The methodology consisted on these main steps:

- Based on the dictionary of changes of territorial units, we defined the versions of NUTS. The 2003 NUTS version was considered as a reference. The data were transposed into this version.
- Missing data were estimated using the ESTI framework methods
- Selected indicators cover the following sectors: economy, demography, unemployment and GDP per capita
- We proceed first to a hierarchical cluster on the second report data and we allocate the results to third and fourth reports data

These steps will be developed in the next sections.

4.2.1 Cohesion reports datasets exploration: missing values and lack of compatibility between NUTS versions

The exploration of data at NUTS 2 level shows the lack of compatibility of NUTS between datasets. Indeed, the collected data is a mixture of nuts versions from 1995 to 2006. If we look at the NUTS 2 level, we can observe many kinds of change. We illustrate some of them by the following tables.

ID (CR2)	name	ID (CR3)	name
IT11	PIEMONTE	ITC1	Piemonte
IT12	VALLE D'AOSTA	ITC2	Valle d'Aosta/Vallée d'Aoste
IT13	LIGURIA	ITC3	Liguria
IT20	LOMBARDIA	ITC4	Lombardia
IT31	TRENTINO-ALTO ADIGE	IT31	*Trentino-Alto Adige*
		ITD1	Provincia Autonoma Bolzano/Bozen
		ITD2	Provincia Autonoma Trento
IT32	VENETO	ITD3	Veneto
IT33	FRIULI-VENEZIA GIULIA	ITD4	Friuli-Venezia Giulia
IT40	EMILIA-ROMAGNA	ITD5	Emilia-Romagna
IT51	TOSCANA	ITE1	Toscana
IT52	UMBRIA	ITE2	Umbria
IT53	MARCHE	ITE3	Marche
IT60	LAZIO	ITE4	Lazio
IT71	ABRUZZO	ITF1	Abruzzo
IT72	MOLISE	ITF2	Molise
IT80	CAMPANIA	ITF3	Campania
IT91	PUGLIA	ITF4	Puglia
IT92	BASILICATA	ITF5	Basilicata
IT93	CALABRIA	ITF6	Calabria
ITA	SICILIA	ITG1	Sicilia
ITB	SARDEGNA	ITG2	Sardegna

Figure 23: The lack of compatibility between Italians NUTS2 units between 1999 and 2003.

ID (CR2)	name	ID (CR3)	name
FI14	VÄLI-SUOMI	FI18	Etelä-Suomi
FI15	POHJOIS-SUOMI	FI19	Länsi-Suomi
FI16	UUSIMAA (SUURALUE)	FI1A	Pohjois-Suomi
FI17	ETELÄ-SUOMI		

Figure 20: A complicated territorial change of NUTS2 level in Finland

Concerning the missing values, we can distinguish between cases caused by changes of NUTS and cases of non-available data. Depending of these types we use the adequate framework of estimation.

id	level	name	NUTS_VER	GDP_head_av	Emp_agri	Emp_ind	Emp_serv	Unemp	Unemp_LT	Pop_tot
BG31	NUTS2	Severozapaden	2006	25,6				12,1		983
BG32	NUTS2	Severen tsentralen	2006	26,4				12,1		963
BG33	NUTS2	Severoiztochen	2006	29,3				12,1		1004
BG34	NUTS2	Yugoiztochen	2006	29,9				8,3		1143
BG42	NUTS2	Yuzhen tsentralen	2006	25,6				11,0		1576
CY0	NUTS2	KYPROS	2006	78,9	10,2	23,2	66,6	3,1		741
DE41	NUTS2	Brandenburg - Nor	2006	68,4						1177
DE42	NUTS2	Brandenburg - Süd	2006	77,8						1419
DED1	NUTS2	CHEMNITZ	2006	60,5					41,4	1690
DED2	NUTS2	DRESDEN	2006	71,1					44,7	1762
DED3	NUTS2	LEIPZIG	2006	76,5					42,2	1112
ES63	NUTS2	Ciudad Autónoma	2006							75
ES63	NUTS2	Ciudad Autónoma d	2006	90,4				19,7	51,5	71
ES64	NUTS2	Ciudad Autónoma	2006							67
ES64	NUTS2	Ciudad Autónoma d	2006	87,9				13,9	36,7	67
FI16	NUTS2	UUSIMAA (SUUR	1999	130,6	1,06	22,34	76,27	7	28,7	1336
FI17	NUTS2	ETELÄ-SUOMI	1999	91,4	6,32	32,94	60,48	12,3	24,5	1815
FI18	NUTS2	Etelä-Suomi	2006	133,3						2537
FI19	NUTS2	Länsi-Suomi	2006	99,9						1319
FI1A	NUTS2	Pohjois-Suomi	2006	97,7						628

Missing value in the 4 cohesion report
Missing value in the 3 cohesion report
Missing value in the 2 cohesion report

Figure 24: An extract of the summary table of missing values of the 2, 3 and 4 cohesion reports.

4.2.2 NUTS harmonisation process

After the identification of NUTS version (Figure 17), the second step of database harmonisation consists to define a reference version of NUTS depending of the complexity of territorial changes. In this reference we will transpose all indicators. Theoretically, three formalizations are possible:

- Prospective approach consists to consider the oldest version of nuts as the spatial reference.
- Retrospective approach consists to consider the most recent version as a spatial reference.
- Intermediary approach uses the intermediary version as a reference: the lowest common denominator

Based on the dictionary of changes (figure 15), we considered the 2003 NUTS2 version as a reference because it allows building the most complete temporal database. Selected indicators were transposed in this NUTS version.

We selected 13 indicators covering economic sectors, Unemployment, demography, education and GDP.

The following figure illustrates the structure of the built temporal database.

Codes_Unit	Indicator 1	.....	Indicator 13
AT1-CR2			
AT11-CR2			
AT12-CR2			
AT13-CR2			
AT2-CR2			
AT1-CR3			
AT11-CR3			
AT12-CR3			
AT13-CR3			
AT2-CR3			
AT1-CR4			
AT11-CR4			
AT12-CR4			
AT13-CR4			
AT2-CR4			

Figure 25: Theoretical structure of the complete temporal database based on 2, 3 and 4 cohesion reports.

CR	id	NUTS_VER	GDP_head_av	Emp_agri	Emp_ind	Emp_serv	Unemp	Unemp_LT	Pop_tot
CR2	BE10	2006	170,3	0,21	13,42	86,38	14	62,0	952
CR2	BE21	2006	137,5	2,15	28,07	69,79	6,5	53,2	1637
CR2	BE22	2006	109,4	2,09	34,52	63,39	7	53,7	782
CR3	BE10	238,5	0,1	13,1	86,9	14,5	55,1	17,8	65,4
CR3	BE21	135,9	1,2	29,7	69,2	5,5	44,0	17,2	65,9
CR3	BE22	98,7	1,6	32,9	65,5	5,3	32,5	17,4	68,8
CR4	BE10	248,3	0,2	11,1	88,7	16,3	56,4	18,2	66,1
CR4	BE21	144,5	1,7	28,4	69,9	6,2	44,0	16,8	65,7
CR4	BE22	101,5	1,8	32,2	66,0	7,1	44,2	16,5	68,3
CR4	BE23	111,0	2,0	28,4	69,6	4,9	37,1	16,4	65,8

Figure 26: An extract of the complete temporal database based on 2, 3 and 4 cohesion reports.

4.3 Spatio-temporal analysis of the European regional cohesion

This section shows that it is possible to analyze the evolution of spatial patterns of regional inequalities with changing territorial units, through the example of a cross analysis of statistical annexes of 2nd, 3rd and 4th Cohesion Reports. The built temporal dataset allows implementing a temporal hierarchical cluster analysis.

The territorial cohesion of the European regions is organised along four main groups. Each group can be either central or peripheral (Figure 27):

- *North-western European type*:

    - Central: it is characterized mainly by: a very high education level, a very high rate of the service sector employments and a high rate of elderly people.
    - Peripheral: it is characterised by a relative decrease in the level of education and service sector employments.
    - Ultra-peripheral: it is characterised by a very high rate of young population. There is also an important decrease in education and service sector employments levels.
- *Southern European type*:
    - Central: it is characterized mainly by a high rate of elderly people, an important rate of industry sector employment and a low education level.
    - Peripheral: it is characterised by a relative decrease in the rate of elderly people and an increase in the rate of the agriculture sector employment.
- *Central European type:*
    - Central: it is characterized mainly by an important rate of adult population, a high education rate and the primacy of service sector employments rate.
    - Peripheral: it is characterised by the decrease the rate of high education level and an increase in the rate of the industry sector employment.
- Eastern European type:
    - Central: it is characterised by an important rate of the industry and the agriculture sectors employment and an important rate of adult population.
    - Peripheral: it is characterised by the primacy of the rate of the agriculture sectors employment and the increase of the rate of young population rate.

This organization is very stable through time. Indeed, no major changes happened.

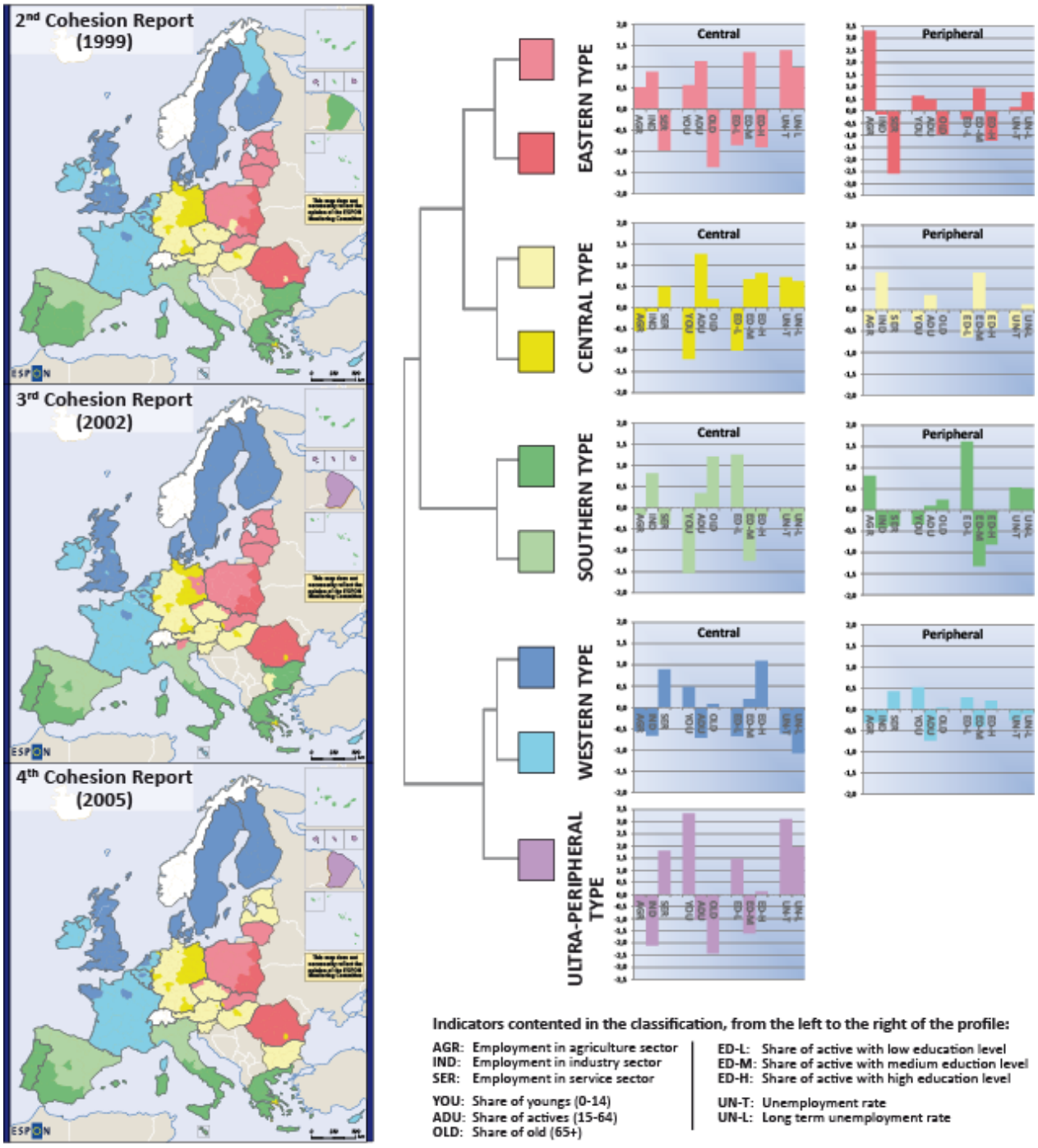


Figure 27: An extract of the complete temporal database based on 2, 3 and 4 cohesion reports.

Conclusion

The contribution of the change analysis based-approach to the topic of time-series is very important. The theoretical formalization and modelling allow the building of a complete temporal database (figure 24). They provide a key of conversion between NUTS versions. This formalization is not a normative approach, but rather it requires an improvement.
The improvement perspectives could be oriented in the following directions:

- The enlargement of the temporal period to the beginning of 1980. The archived New-Cronos database of Eurostat is a very relevant case study.
- The integration of a various geographic objects like cities and grids could have a very potential contribution to the time series building process. It could also affine and improve the thematic analysis
- The automation of the dictionary of changes and the estimation methods

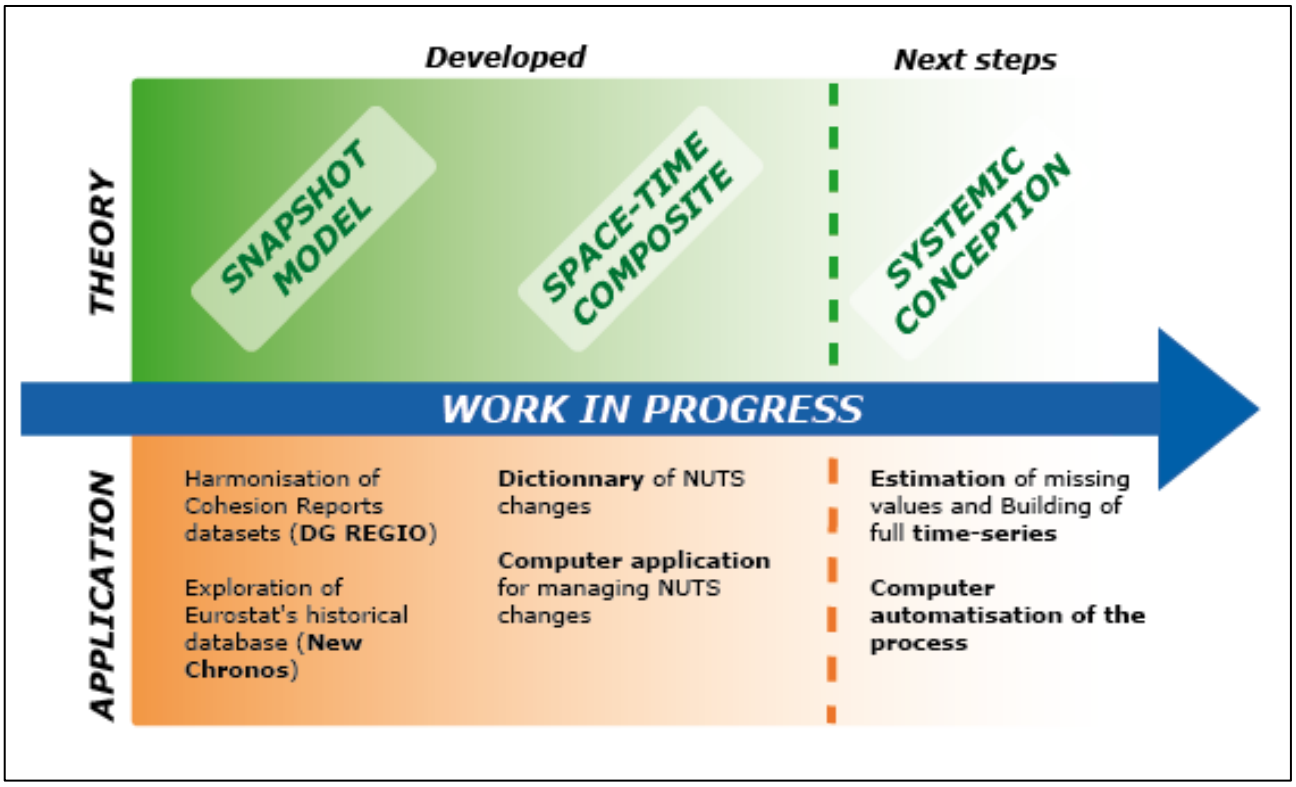


Figure 28: Synthesis of the work of the "time-series" group.

TECHNICAL REPORT

ESPON

Disaggregation of socioeconomic data into a regular grid and combination with other types of data

CONTENT

- Introduction and methodology description. This section describes the background details and methodological solution adopted with regard to the disaggregation of data.
- Integration of socio-economic and environmental information. This part presents the OLAP Cube solution implemented in order to combine data from several data sources and formats, mainly socioeconomic statistics together with continuous data, such as land cover.
- Future steps. Taking into account the results from the challenge along the ESPON DB projects this section drafts the following steps that could be developed as a continuation.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

37 PAGES (incl. annexes)

LIST OF AUTHORS

Roger Milego, Universitat Autònoma de Barcelona, UAB 08193 Bellaterra

Maria José Ramos, Universitat Autònoma de Barcelona , UAB 08193 Bellaterra

Contact

roger.milego@uab.cat
mariajose.ramos@uab.cat

tel. + 34 93 581 35 46

TABLE OF CONTENT

1 Introduction

The ESPON 2013 DB project has been structured in several challenges in order to fulfill its objectives. The challenge to which this technical report refers to is the challenge number 5: "Combining socio-economic data measured for administrative zoning (NUTS level) and environmental data defined on a regular grid (like Corine Land cover)". The UAB (Universitat Autònoma de Barcelona) is the responsible partner with regard to this challenge.

Most of the socioeconomic variables or indicators are typically given by administrative unit, i.e. NUTS regions, whereas the environmental data is usually not following those boundaries, but given by natural units or regular grid cells.

The aim of this challenge is to define a suitable methodology for integrating and making comparable data coming from statistical sources (e.g. EUROSTAT) and measured by administrative unit, together with environmental data stored by natural unit or regular grid structure (e.g. Corine Land Cover).

The ESPON 2006 program developed some indicators in which the environmental data was transposed to NUTS division by means of GIS tools, in order to make them comparable to socioeconomic data. The results from this integration strategy, not always convincing, make clear the necessity of implementing a new integration process based on grid methods as it is said in the tender of the Espon 2013 Database project and in the Modifiable Areal Unit Problem study.

ESPON 2013 Database challenges

According to the E.S.T.I (space, Source, Time, Indicator) framework presented in the "Handbook for data collection" (ESPON 3.2, Final Report, Annex) four main objectives were identified in the tender of the ESPON 2013 DB Programme.

This challenge is included in the second key-question**: "Combination of heterogeneous sources- balancing Eurostat data"**, that emphasizes the need of the integration between different types and sources of data. The harmonization of the database in a fixed spatial division (NUTS3) solution that was chosen by many ESPON 2006 projects presented some doubts and not always convincing results. It is in this scenario that **a new integration methodology based on the reverse operation** is mentioned, **a projection of socio-economic information into units elaborated for the monitoring of natural resources.**

Modifiable Areal Unit Problem (ESPON 3.4.3)

The MAUP study, in its chapter 4 "*Exploration of gridding methods",* highlights the integration of heterogeneous databases as one of the most promising application of gridding methods for ESPON.

Two potential fields of applications are distinguished for gridding methods:

- **Time harmonisation of changing territorial units.**

"The use of grid help to build an harmonised territorial framework where all changing territorial divisions are harmonised and can further be used for the analysis of time variation" MAUP study (ESPON 3.4.3)

- **Thematic harmonisation and combination of heterogeneous spatial sources.**

The ESPON 2006 integration strategy, called "Eurostat oriented" by MAUP study, based on transferring all the information that it is not delivered on the basis of administrative units (NUTS 2 or NUTS 3) toward administrative units, is questioned, and the use of a new strategy is proposed.

"Information of good quality (as CLC) is therefore transformed into information of bad quality when projected in spatial units which are not adapted" MAUP study (ESPON 3.4.3).

" "Eurostat oriented" strategy could be replaced by another strategy that could be called the "EEA oriented" where all data would be transformed into grid and integrated on this basis" MAUP study (ESPON 3.4.3).

Methodology Proposal

In the First Interim Report of the project (Feb 2009), a methodology proposal was made on the basis of existing applications made by other institutions, such as:

- ***"A Donwnscaled Population Density Map of the EU from Commune Data and Land Cover Information"*** by Javier Gallego, JRC-ISPRA**.**

  A combination of commune population data with Corine Land Cover to produce an EU-wide grid with 1 ha resolution of downscaled population density[1].

- **G-Econ Research project of the University of Yale to develop a geophysically based data set on economic activity.**

  Estimation of gross output at a 1-degree longitude by 1-degree latitude resolution at a global scale for virtually all terrestrial grid cells based on spatial rescaling settled on **proportional allocation[2]**.

- **FARO-EU (Foresight Analysis of Rural areas Of Europe)**

  The project is aimed to analyse Rural Development in Europe by analysing patterns and trends of a selection of territorial indicators specific for rural areas within a Spatial Regional Reference Framework[3]**.**

---

[1] http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/S14P3_JAVIER_GALLEGO_DOWNSCALED_POPULATION_DENSITY.pdf

[2] "*New Metrics for Environmental Economics: Gridded Economic Dats*" by William D. Nordhaus. http://www.oecd.org/dataoecd/44/7/37117455.pdf

[3] www.faro-eu.org

- ***"Transforming Population Data for Interdisciplinary Usages: From census to grid"*** by Deborah Balk & Greg Yetman from Columbia University.

  Creation of the Gridded Population of the World (GPW) data base implementing **a proportional allocation** of population from administrative units to grid cells [4].

The objectives established in the tender of the ESPON 2013 DB, the MAUP study results and recommendations, the bibliography research on existing methodologies and our experience at the UAB, as European Topic Centre on Land Use and Spatial Information, supporting the EEA in monitoring the land use/land cover change in Europe and analyzing the environmental consequences; lead us to the conclusion that the best way to downscale socioeconomic data and make them comparable with other kind of data, is **using a regular grid structure**, in which each cell takes a figure of the indicator or variable.

It was also concluded that depending on the nature of each variable, a different integration method should be applied. In other words, the way of calculating the actual figure for each grid cell might differ between different types of data, according to their definition.

The European Reference Grid

The EEA recommends the use of EEA reference grids for projection ETRS89-LAEA 52N 10E. The recommendation is based on proposal at the 1st European Workshop on Reference Grids[5].

The 1st Workshop on European Reference Grids was organized by the Joint Research Centre of the European Commission following a request of the EEA and the request of the INSPIRE Implementing Strategies Working Group that recommended the adaptation of a Europe-wide reference grid to facilitate the management and analyses of spatial information. The interest of the creation of a common coordinate reference system and a common equal-area grid to represent EU and Pan-Europe was also expressed by the National Statistical Institutes.

Taking into account this recommendations and our UAB/ETC-LUSI experience under some EEA projects such as LEAC (Land and Ecosystem Accounting), we proposed to disaggregate socioeconomic data into the **1 km European Reference Grid**[6], as it is the way in which valuable data for the ESPON projects, such as the Corine Land Cover changes, are stored as well.

Testing process

After making all these decisions, we have carried out several tests with different data using different integration methods, trying to achieve useful results, on the one hand,

[4] http://sedac.ciesin.columbia.edu/gpw-v2/GPWdocumentation.pdf
[5] http://eusoils.jrc.ec.europa.eu/projects/alpsis/Docs/ref_grid_sh_proc_draft.pdf
[6] http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=760

but aiming at preparing the basis for the automation of the processes and integration with environmental data in an OLAP (On-Line Analytical Processing) cube.

Whenever an ESPON project would like to make the comparison of any kind of socioeconomic data together with land cover or environmental data not measured by administrative unit, the ESPON 2013 DB project, through the challenge 5 outcomes, will be ready to facilitate this task and provide them with the expected results or methodological tools to be applied.

The goal of the results presented in this report is to highlight the high potential of the three methodologies proposed. At this point the layers distributed by 1km grid resulting from applying the different methods are not available; their future distribution will be based on the building of OLAP's cubes using the most updated data.

Objectives

This technical report is aimed at explaining the details of the methodology and tests undertaken regarding the challenge 5 of the ESPON 2013 DB and it has, in particular, the following objectives:

- Review and summarise the background of the challenge 5.
- Give a detailed description of the methodology to be applied in order to downscale socioeconomic data.
- Describe the different disaggregation methods.
- Explain the tests undertaken and the results that have been achieved so far.
- Make some conclusions about all the processes undertaken so far.
- Define the next steps to be carried out.

2 Methodology description

This section describes the details of the methodology proposed to face the challenge of combining data measured using different reporting units.

According to what it has been explained in the introduction, and after reviewing several studies and taking into account our experience at the UAB (ETC-LUSI) and the EEA, we propose to integrate socioeconomic data in the 1 km European Reference Grid, because, besides some other reasons, this unit is used to summarise land cover data and other types of environmental data processed at the EEA.

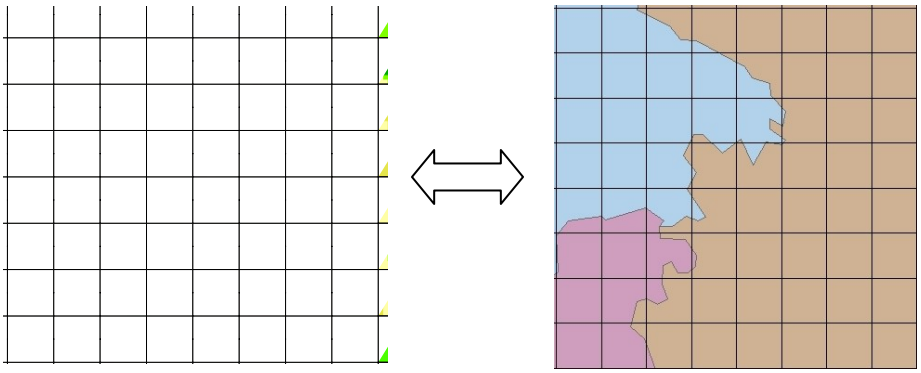


Figure 1. *The 1 km European Reference Grid will hold both environmental and socioeconomic information*.

Therefore, the first step to be carried out should be the **intersection** between the 1 km European Reference Grid and the administrative units by which the indicator is given. This is done by a physical overlay of both layers in vector format, by means of the ArcGIS tool Intersect. This tool creates a new layer holding both the geometries and the attributes of the source layers.

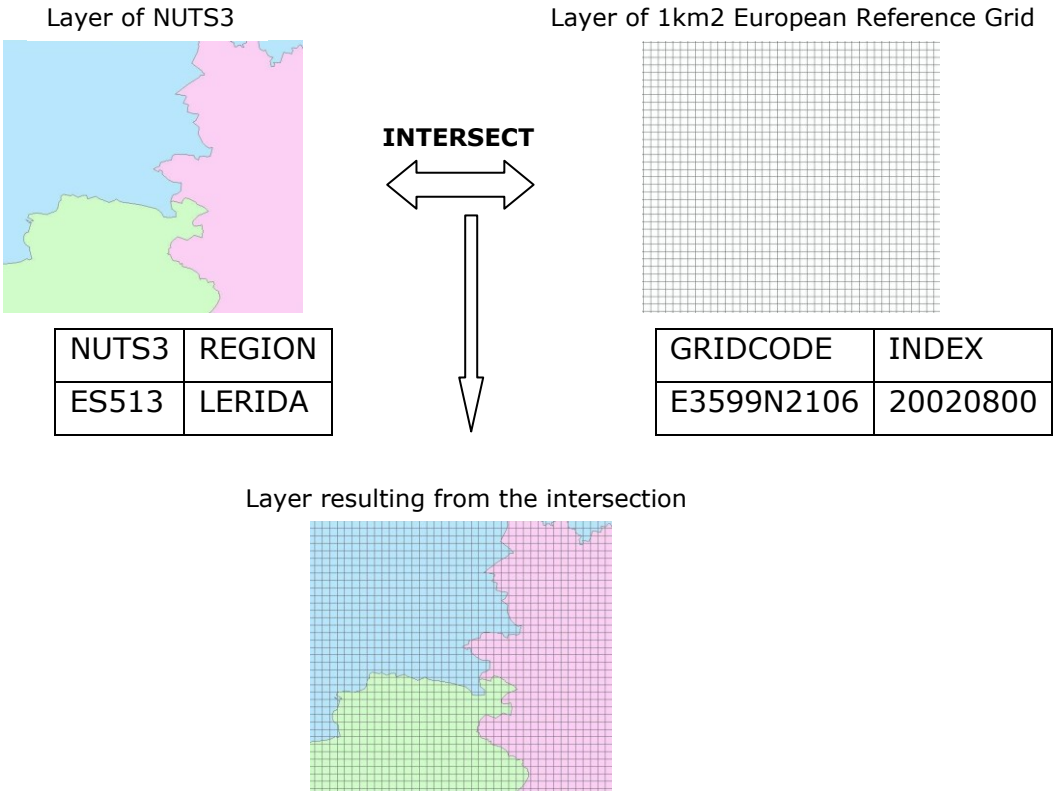
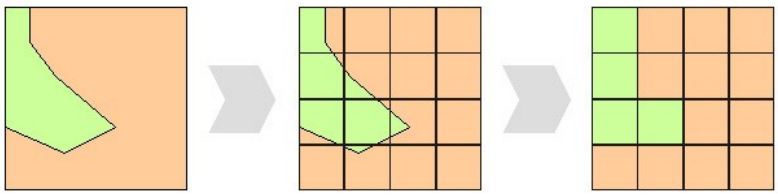


***Figure 2**. Example of the intersection tool.*

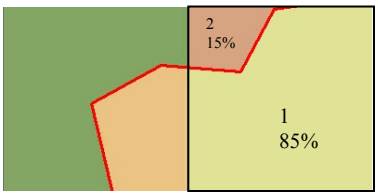
Once the intersection has been computed, a way of computing a single figure by each reference grid cell should be defined. It has been stated that depending on the nature of each indicator or variable, a different kind of integration procedure should be defined. In this regard, we have defined and tested with different data the following three integration methods:

2.1 Integration methods

Maximum area criteria: the cell takes the value of the unit which covers most of the cell area. It should be a good option for uncountable variables.



Proportional calculation: the cell takes a calculated value depending on the values of the units falling inside and their share within the cell. This method seems very appropriate for countable variables.



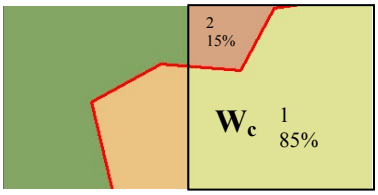
Cell value = Σ (V_i * $Share_i$)

Where: V_i = Value of unit i

$Share_i$ = Share of unit i within the cell

In the example: V_1 * 0.85 + V_2 * 0.15

Proportional and weighted calculation: the cell takes also a proportionally calculated value, but this value is weighted for each cell, according to an external variable (e.g. population). This method can be applied to improve the territorial distribution of a socioeconomic indicator. For instance, a GDP indicator can be redistributed by 1 km grid and weighted by the population figures of each cell (coming from the 1 km population density dataset produced by JRC).



Cell value = W_c Σ (V_i * $Share_i$)

Where: V_i = Value of unit i

$Share_i$ = Share of unit i within the cell

W_c = weight assigned to cell c

In the example: W_c * (V_1 * 0.85 + V_2 * 0.15)

Depending on each type of indicator or variable to be integrated within the reference grid, a different type of integration should be decided and tested. Besides the method finally chosen to integrate, it is important to highlight that indicator figures given by area unit, e.g. by square kilometre, should be converted considering that each cell has a total area of 1 km^2.

Whenever it is possible, the third method has been used and should be used to disaggregate the data, as it gives an added value to the source data, providing more interesting results when different data are put together on a cell-by-cell basis.

Annex 1 shows the test results for the three integration methods.

3 Integration of socio-economic and environmental information

Once the variable has been distributed by 1 km cell, it can be compared to other data on a cell-by-cell basis, and it can be integrated into an **OLAP (Online Analytical Processing) cube**.

The OLAP technology[7] use a multidimensional data model, allowing complex analytical and ad-hoc queries with a rapid execution time. This solution facilitates the integrated analysis of several types of geographic and statistical data to users with or without GIS knowledge.

The next schema presents the general process to make possible the integration within an OLAP cube

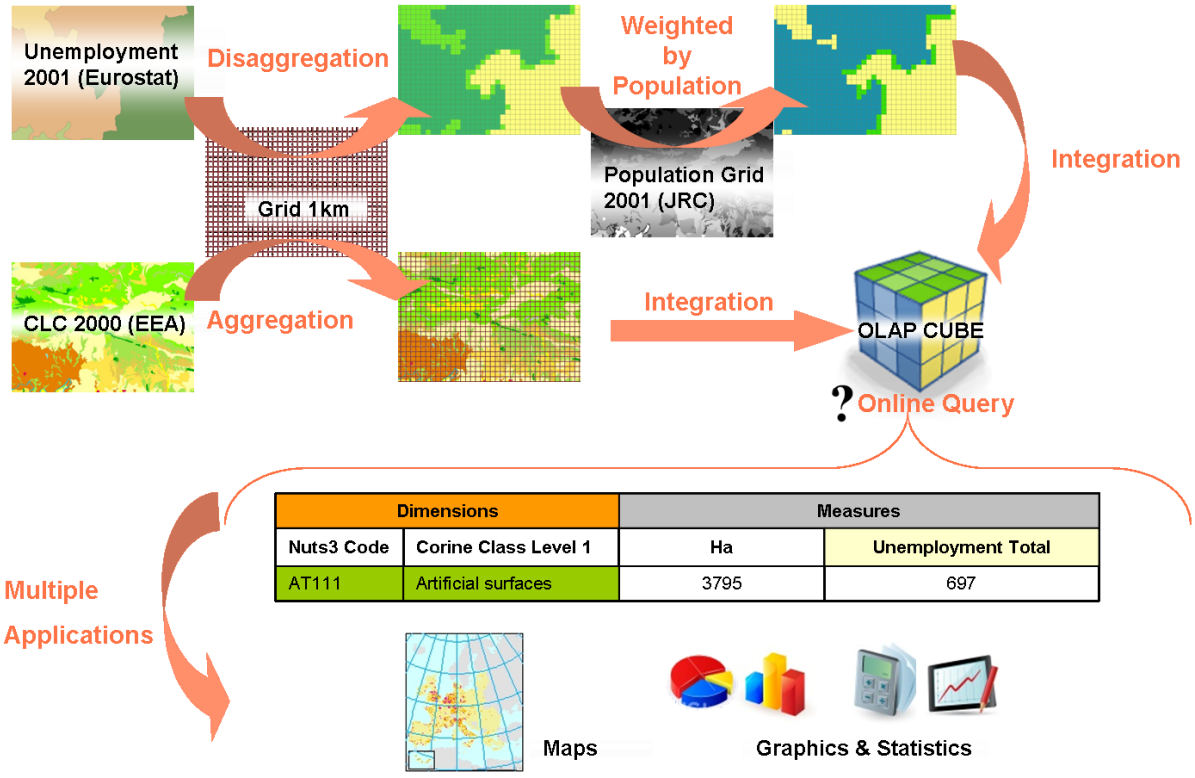


Figure 13. *Simplified schema of data integration in an OLAP Cube*

In the case of ESPON OLAP cube, it consists on the **ESPON socio-economic** variables as numerical attributes or measures that will be aggregated using a set of dimensions. The **dimensions** or topics of interest for the user are generally represented by different types: **spatial dimensions,** usually represented by

[7]Some OLAP information resources:
1- http://en.wikipedia.org/wiki/Online_analytical_processing
2- http://www.cs.sfu.ca/CC/459/han/papers/chaudhuri97.pdf
3- http://es.wikipedia.org/wiki/Cubo_OLAP

administrative units for Europe (NUTS), a number of **thematic dimensions**: land use data or dominant land cover types; and a third type, which is the **temporal dimension** that shows the difference between two years time.

In this way, we are able to ask to the ESPON-OLAP cube some questions taking into account the socioeconomic variables or indicators. For example, having the GDP and CLC changes in the ESPON OLAP cube, we could analyse which land cover flows occur by different GDP ranges, and, in the end, get the results on a NUTS3, NUTS2 or country (NUTS0) basis.

The OLAP Cube delivered together with the ESPON 2013 DB Final Report is the version 3.0, and holds the following dimensions and measures:

- Socioeconomic variables
  - GDP 2003
  - GDP 2006
  - Active population 2003
  - Active population 2006
  - Unemployment 2003
  - Unemployment 2006
- Land cover
  - Corine Land Cover 1990
  - Corine Land Cover 2000
  - Corine Land Cover 2006
- Land cover changes
  - Land Cover Flows 1990-2000
  - Land Cover Flows 1990-2006
  - Land Cover Flows 2000-2006
- Measures
  - Population density 2001 (inhab/km^2)
  - Area (ha)
- Geographical dimensions
  - Elevation Breakdown
  - Biogeographic Regions
  - Large Urban Zones and City Names
  - Massifs
  - NUTS 2006
  - NUTS 2003
  - River Basin Districts UE

The ESPON OLAP Cube can be accessed either online or offline. Online connection has not been implemented yet in the framework of the ESPON DB. So far, a CUB file has been provided in order to use the ESPON OLAP Cube offline. **Annex 2** explains how to use this file and make queries to the Cube by means of MS-Excel.

4 Future steps

There are several improvements and future implementations we have in mind to make this methodological approach and derived tools grow up in the future. We list below some of them:

a) Methodological improvements: Some aspects that have to be deeply analysed are:

   - Treatment of administrative units with no data values.
   - Differences between geographical extents, for example between Nuts 3 2006 layer and Corine Land Cover.
   - Improvement of the disaggregation performance in terms of time and manageability of the final layer.
   - Calculation of deviations with respect to the source data.

b) Extent increase: It is foreseen to increase the number of neighbours, countries and areas to process. The possible candidates will comprise Maghreb and part of Saharan Africa together with a portion of the CIS[8].

c) Follow-up of the EFGS: It might be very useful to keep following up the outcomes of the European Forum for Geostatistics[9] in order to contrast the proposed method, and, eventually, improve it.

d) Questionnaire: The preparation of a questionnaire could be proposed for the ESPON projects to fill in, in order to get feedback about which types of data they would like to be converted into the grid format and, eventually, included within the ESPON OLAP Cubes.

e) Integration with the ESPON 2013 Database: It should be deeply studied the most suitable solution to integrate grid data in the ESPON 2013 Database and to make the ESPON OLAP Cube available to users through the database interface. The main idea is to facilitate the comparison and analysis of socio-economic data with environmental data that usually not follows an administrative distribution.

f) Visualisation tool: A suitable web visualisation tool could be adapted and integrated into the ESPON database framework allowing the visualisation of grid and socio-economic ESPON OLAP data and, eventually, the other ESPON database data. Two specific environments could be defined depending on the specific needs of the user: one more simple, designed for a basic user just for visualisation purposes, and a more advanced one, being able to show comprehensive results like interactive maps, graphs or tables.

g) Different ESPON Cubes: Several ESPON OLAP Cubes could be built up in order to integrate the gridded data and make them ready to be queried and analysed together with other types of data, like land cover, natural units, protected areas, and so on, for different scales (i.e. Urban Cube).

[8] Commonwealth of Independent States
[9] http://www.efgs.ssb.no/

5 Conclusions

After having defined a methodology in order to be able to put together and analyse socioeconomic and environmental data, and having made several tests using different datasets, we can make now some conclusions about the main outcomes of the work done and the things we have learned so far:

- Disaggregating socioeconomic data by a regular grid is the best solution in order to downscale such information reported by administrative areas.
- The 1 km European Reference Grid is a good option to undertake the disaggregation because:
    - It has an European coverage
    - It follows Inspire specifications
    - It is used for several institutions as the reference grid
    - Its resolution is optimal in order not to lose data precision
- For uncountable data (non-numeric values), the best aggregation method is the "maximum area criterion".
- For countable data, the best method is the proportional one, which calculates the final value according to the area share of each of the values.
- Whenever it is possible, it is better to weight the final figures when using a proportional method, e.g. by population.
- The "proportional and weighted" aggregation method is the one that gives better results, plus some added value to the downscaling.
- The different methods are independent from the source data format and can be applied to vector and raster format.
- In order to achieve good results following this methodology it is important to use data sources which follow the same spatial and temporal specifications (extent, spatial resolution, temporal resolution…).
- This methodology allows the integration of socio-economic in an OLAP cube, which facilitates the comparison and analysis of such data together with land cover data, for example.

To sum up, it can be added that any kind of socioeconomic data can be processed using the methodology proposed and tested, in order to have them downscaled and stored by 1 km grid, facilitating their comparison with many other data not reported by administrative units.

Our next steps will be aimed at improving the performance of the methodologies proposed but also to analyse the introduction of some changes based on our results and new projects at European context.

6 References

- ***Litterature***

Arévalo J., *Land and Ecosystem Accounting. Technical Procedure, Internal Report v.2,* 2009, ETC-LUSI, European Environmental Agency.

Chaudhuri S., Dayal U., *An overview of Data Warehousing and OLAP Technology,* Simon Fraser University Canada (SFU.CA).

Deichmann U., Balk D., Yetman G., 2001, *Transforming Population Data for Interdisciplinary Usages: From census to grid,* NASA Socioeconomic Data and Applications Center (SEDAC).

Gallego J., *A Downscaled Population Density Map of the EU from Commune Data and Land Cover Information,* JRC-Ispra.

Gallego J., *Downscaling population density in the European Union with a land cover map and a point survey,* JRC-Ispra.

Gallego J., *Population density grid of EU-27+, version 4. Summary of the downscaling method,* JRC-Ispra.

Malinowski E., Zimányi E., 2009, *Advanced Data Warehouse Design- From Conventional to Spatial and Temporal Applications*, Springer.

Short Proceedings of the 1st European Workshop on Reference Grids, Ispra, 27-19 October 2003, JRC- Institute for Environmental and Sustainability, Ispra

William D. Nordhaus, 2006, *New Metrics for Environmental Economics: Gridded Economic Data,* Yale University

- ***Websites***

Espon programme: http://www.espon.eu/

- **Future Orientation for Cities (FOCI). ESPON programme**

http://www.espon.eu/mmp/online/website/content/programme/1455/2233/2236/2239/index_EN.html

- **The modifiable areas unit problem (MAUP). ESPON Scientific Support Project 3.4.3**

http://www.espon.eu/mmp/online/website/content/projects/261/431/index_EN.html

European Environment Agency (EEA). European Commission: http://www.eea.europa.eu/

- **Population density disaggregated with Corine land cover 2000.**

http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-1

- **EEA reference grids**

http://www.eea.europa.eu/data-and-maps/data/eea-reference-grids

European forum for Geostatistics 2010. http://www.efgs.ssb.no/

Eurostat, European Commission:
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

Foresight Analysis of Rural areas of Europe (Faro-eu.org) http://www.faro-eu.org

Geographically based Economic data (G-Econ) project, University of Yale
http://gecon.yale.edu/

Joint Research Centre (JRC). European Comission:
http://ec.europa.eu/dgs/jrc/index.cfm

ANNEX 1. Testing the methodology

In order to test the methodology proposed and the different methods of data integration into the reference grid, we have chosen some socioeconomic variables or other data not being measured by grid cell, but stored using administrative units or other kind of irregular delimitation. Thus, we have chosen variables such as GDP, unemployment or urban dominance. The next sections show the steps undertaken and results achieved for any kind of test done.

Testing the maximum area criterion: Urban dominance (2000) based on the Urban Morphological Zones.

The maximum area criterion is suitable for uncountable data or data given by ranges. In particular, this approach can be applied to better differentiate the urban/non urban character of the land as has been highlihted by the team of **ESPON FOCI**[10] project. The advantage of taking percentage of Urban Morphological Zones over pure land cover classess is that UMZ indicates if certain area of urban fabric is part of a greater entity (equivalent to a city). Then, the urban dominance could be used to characterise certain areas and to relate them to other descriptors.

a) Source data:

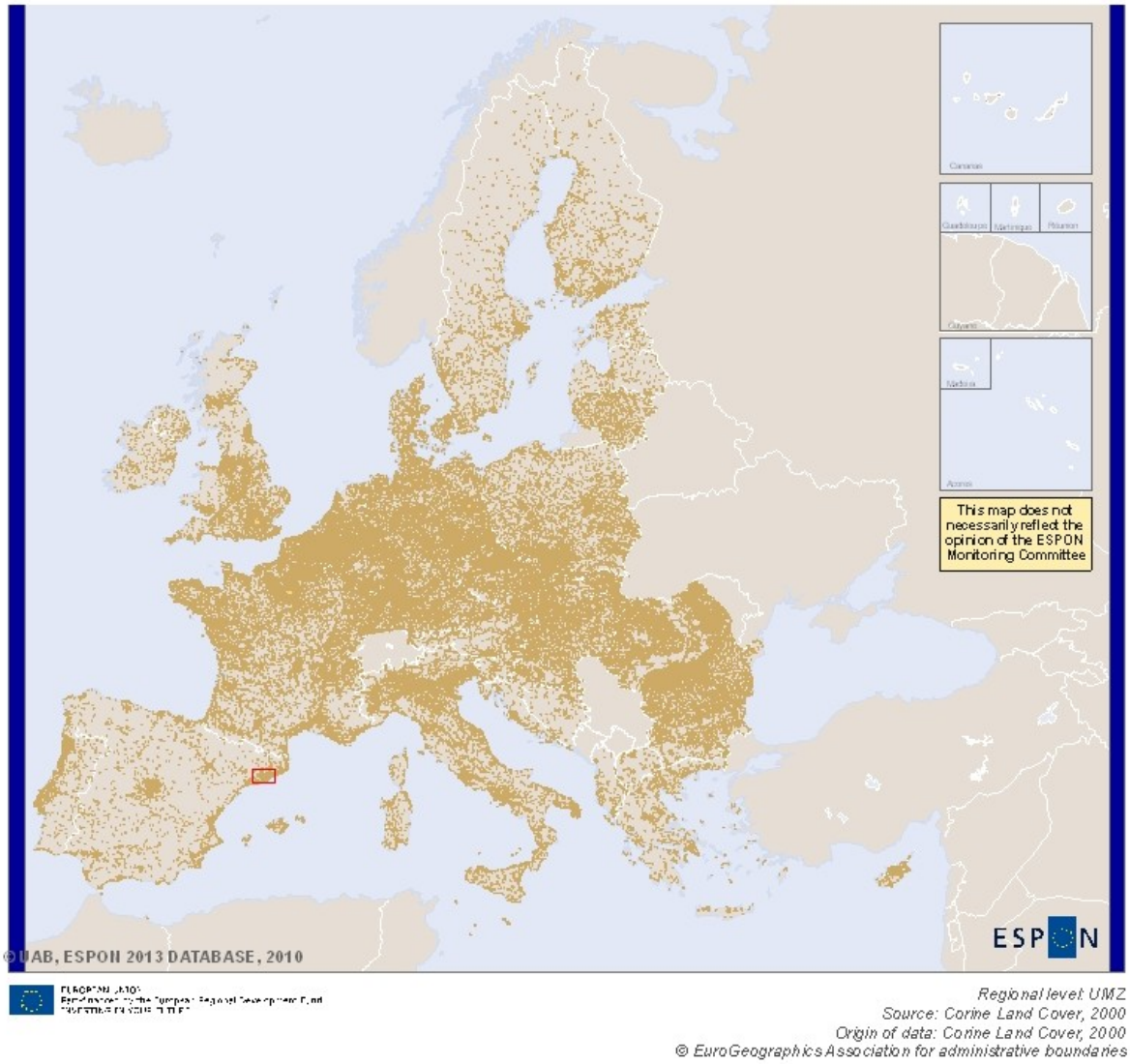
In order to make this test we have used the Urban Morphological Zones 2000 (Map1. Urban Morphological Zones 2000), an EEA's dataset which is a delimitation of urban areas according to a functional definition. The UMZ2000 come from a reclassification of different land cover classes of CLC2000 following different criteria and, therefore, they are not following administrative boundaries but an artificial boundary created by addition of land cover polygons.

For further details about UMZ, please visit:

http://dataservice.eea.europa.eu/download.asp?id=17335&filetype=.pdf

[10] http://www.espon.eu/mmp/online/website/content/programme/1455/2233/2236/2239/index_EN.html

Map 1 . Urban Morphological Zones 2000



b) Process steps:

As it has been stated in the methodology section, the first step has been the intersection between the Reference Grid and the UMZ2000. In this way, we are able to calculate which share of grid cell is occupied by an UMZ.

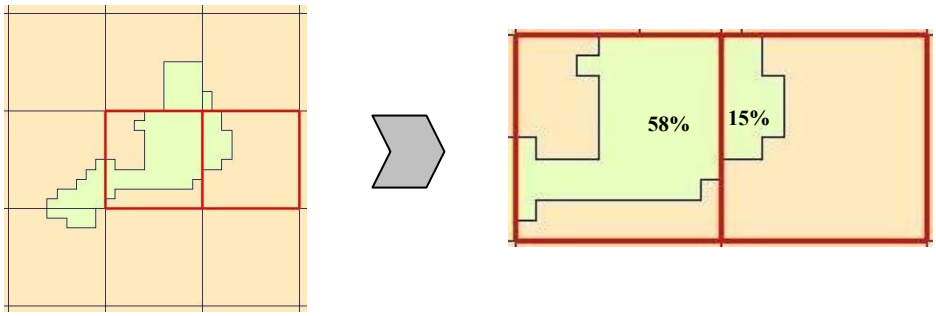


Figure 3. *Intersection of UMZ2000 and the Reference Grid*.

If the cell has more than its half (50%) covered by an UMZ polygon, we define it as an urban-dominant cell, whereas if it has less than 50% of UMZ inside, it is defined as a non-urban cell.

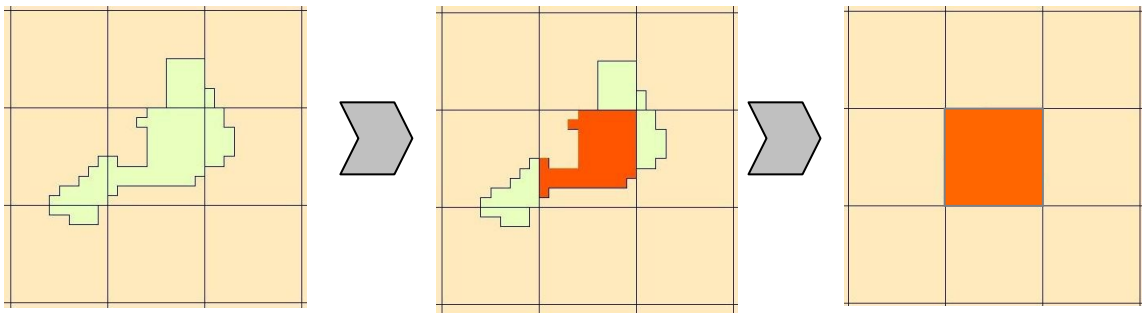
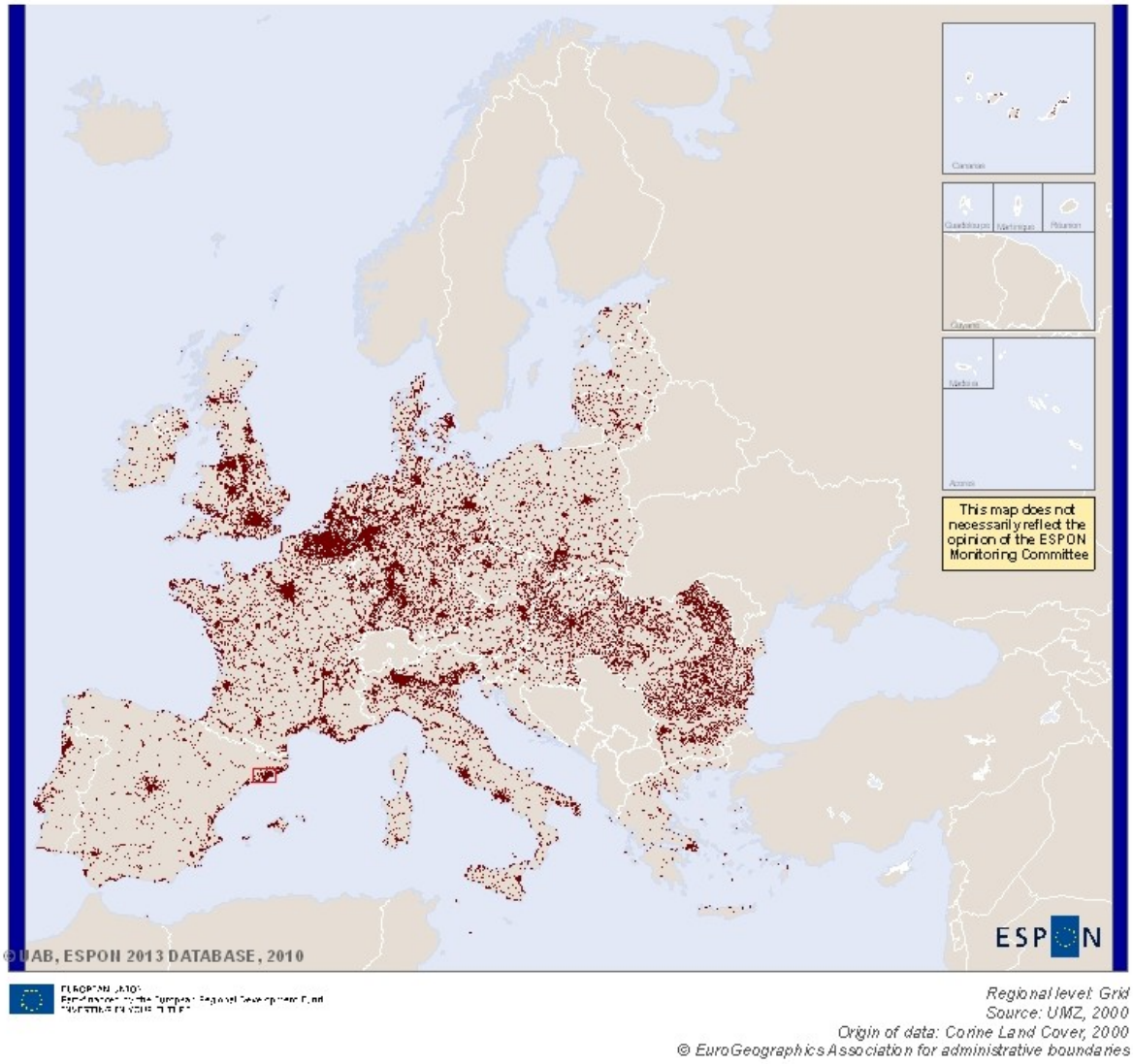


Figure 4. *Urban dominance definition process*.

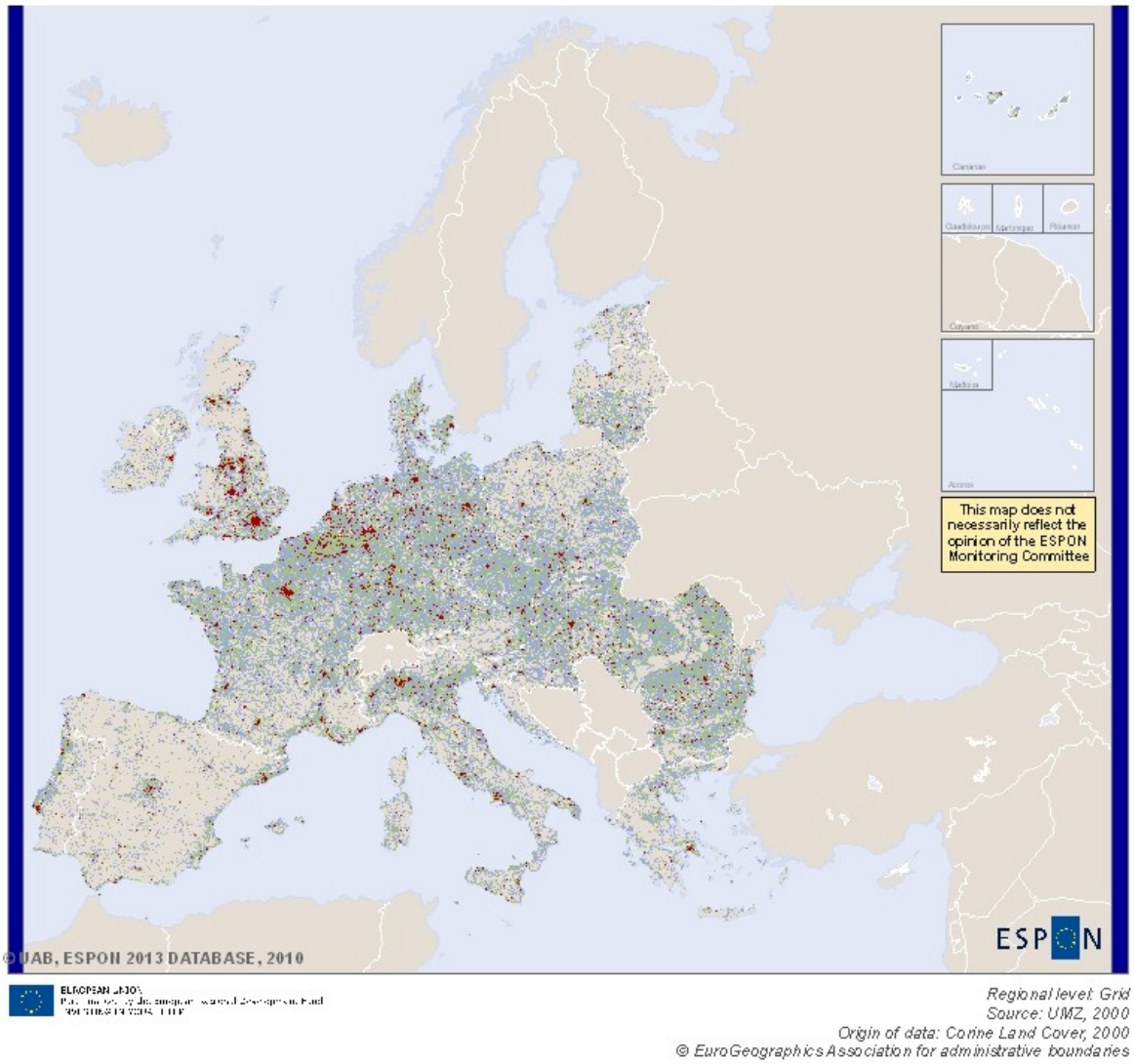
Finally, this kind of map can be elaborated, where we can see the urban dominance in Europe by 1 km grid cell being able to identify quicker the main points of urban surfaces in Europe:

Map 2 . Urban Dominance layout



What is interesting is that now we are able to analyse land cover changes and flows by urban or non-urban cells, or any other indicator. Furthermore, we can create an urban classification by percentage ranges, e.g. 0-20%, 20-60%, 60-100% (low urban dominance, mid urban dominance, high urban dominance):

Map 3 . Urban Classificatiion by density ranges



Conclusion: the maximum area criteria is useful for non-numeric values or in case we are interested in a discrete classification of grid cells, either by a thematic attribute or a value range. Moreover, the **ESPON FOCI** project found interesting this kind of approach applied to UMZ because it can be used as a criteria to define urban dominance and integrate it in the urban-intherland analysis.

Testing the "Proportional calculation" method: Unemployment rate total (2001)

a) Source data:

As for the proportional calculation method, we have chosen an indicator of unemployment: the unemployment rate total (2001) from Eurostat. It represents unemployed people as a percentage of the economically active population, and it is measured by NUTS3 region.

Although the most suitable methodology to be applied on this indicator is the third one, proportional and weighted calculation, we have selected the unemployment rate total as an example of the possible results that can be obtained by applying the second methodology.

	A	B	C	D	E	F	G	H	I	J
1	**Employment and Labour Market**									
2	**NUTS level 3 (version 1999)**									
3	**Subtheme:**		Unemployment	Unemployment	Unemployment	Unemployment	Unemployment	Unemployment	Unemployment	Unemployment
4	**Indicator:**		Unemployment rate total	Unemployment rate female	Unemployment rate male	Unemployment rate under 25 years	Development of unemployment rate 1998-2001	Development of unemployment rate, female, 1998-2001	Development of unemployment rate, male, 1998-2001	Development of unemployment rate, <25 years, 1998-2001
5	**Description:**		in %	in %	in %	in %	in percantage points	in percantage points	in percantage points	in percantage points
6	**Time:**		2001	2001	2001	2001	1998-2001	1998-2001	1998-2001	1998-2001
7	**Source:**		Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices
8	**Comment:**									
9	NUTS_3_99	Region	UNRT01N3	UNRF01N3	UNRM01N3	UNRU2501N3	UNRT98N3	UNRF98N3	UNRM98N3	UNRU2598N3
10	AT111	Mittelburgenland	3,1	4,8	2,2	3,4	-1,0	-1,9	-0,5	-0,1
11	AT112	Nordburgenland	2,3	2,8	2,0	4,0	-0,8	-1,2	-0,5	-0,6
12	AT113	Südburgenland	4,3	5,2	3,5	7,6	-0,9	-1,4	-0,5	0,0
13	AT121	Mostviertel-Eisenwurzen	2,1	2,7	1,5	3,5	-0,6	-0,8	-0,6	-0,5
14	AT122	Niederösterreich-Süd	3,4	3,8	3,0	5,2	-0,9	-1,2	-0,8	-0,1
15	AT123	Sankt Pölten	3,3	3,7	2,9	6,0	-0,8	-1,5	-0,4	0,7
16	AT124	Waldviertel	3,0	3,7	2,4	4,7	-0,9	-1,8	-0,3	-0,1
17	AT125	Weinviertel	2,9	3,5	2,4	4,1	-0,6	-1,1	-0,3	-0,5
18	AT126	Wiener Umland/Nordteil	2,3	2,7	2,1	3,6	-0,7	-1,1	-0,4	0,0
19	AT127	Wiener Umland/Südteil	2,9	3,3	2,8	4,3	-1,0	-1,2	-0,8	-0,2
20	AT13	Wien	4,9	4,9	4,9	7,2	-1,7	-1,8	-1,7	-1,9
21	AT211	Klagenfurt-Villach	4,1	5,2	3,2	8,1	-1,1	-1,3	-1,0	-1,2
22	AT212	Oberkärnten	5,6	8,5	3,6	9,4	-1,4	-2,0	-0,9	-2,4
23	AT213	Unterkärnten	3,7	5,1	2,6	5,7	-1,2	-1,7	-0,9	-2,0

Figure 5. *Unemployment source data.*

b) Process steps:

We start joining the figures of the indicator on unemployment with the layer of NUTS3 using the unique identifier of the NUTS3 regions, this operation allows to have a geometry representation of the information.

In order to have a single dataset holding the NUTS3 and the Reference Grid geometries, including the attribute information, we carried out an overlay process.

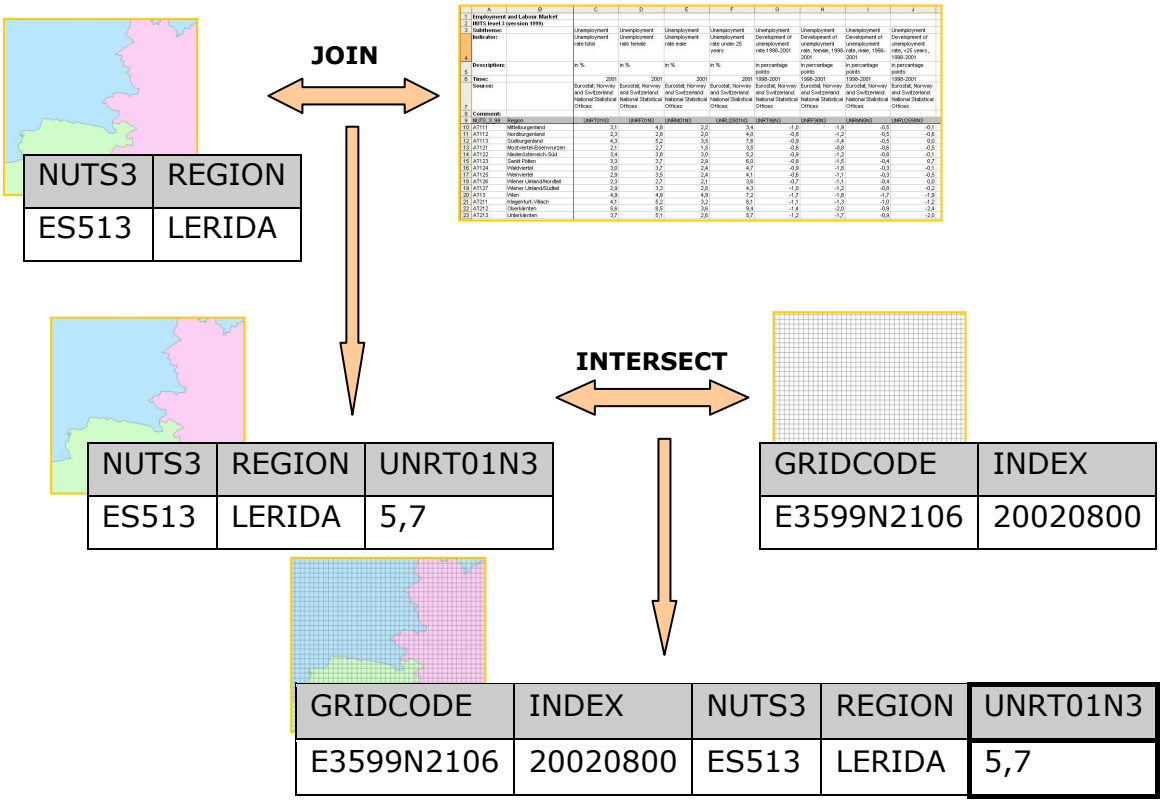
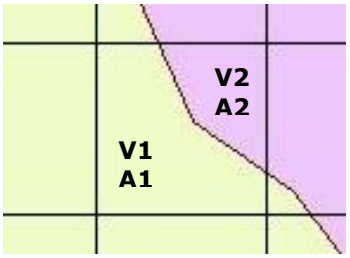


Figure 6. *Creation process schema for unemployment downscaling.*

After that, we should recalculate a single value for each grid cell, based on the NUTS3 value or values that fall inside. If a cell contains different NUTS3 values, the final value will be calculated depending on the share of surface of each NUTS3 region within the cell. For example, if a cell has 0.2 out of 1 covered by one NUTS3 region and 0.8 covered by a second NUTS3, the figures should be calculated accordingly (multiplying the first figure by 0.2 and the second one by 0.8).



V1= 8,8 % Unemployment people
A1 = 0,6 km2

V2 = 8,6 % Unemployment people
A2 = 0,4 km2

Final value of the cell:

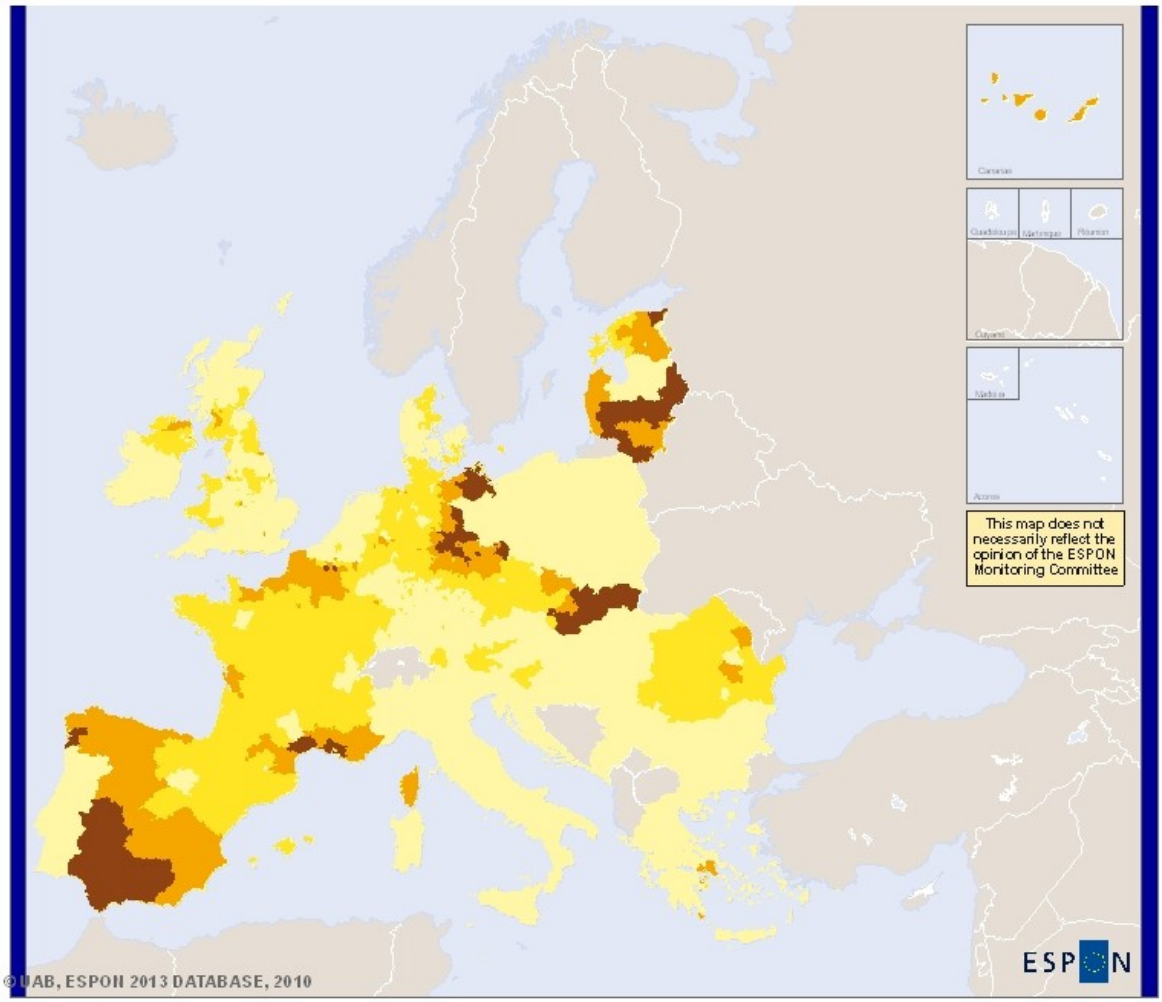
V1= 8,8 * 0,6 = 5,28
V2 = 8,6 * 0,4 = 3,44

Vfinal = 5,28 + 3,44 = 8,72 % Unemployment people

c) Results:

The results put on a map have a look like this:

Map 4 . Distribution of unemployment rate total by grid



Regional level: Grid
Source: Unemployment rate total (EUROSTAT), 2001
Origin of data: Unemployment rate total (EUROSTAT), 2001
© EuroGeographics Association for administrative boundaries

Urban Classification

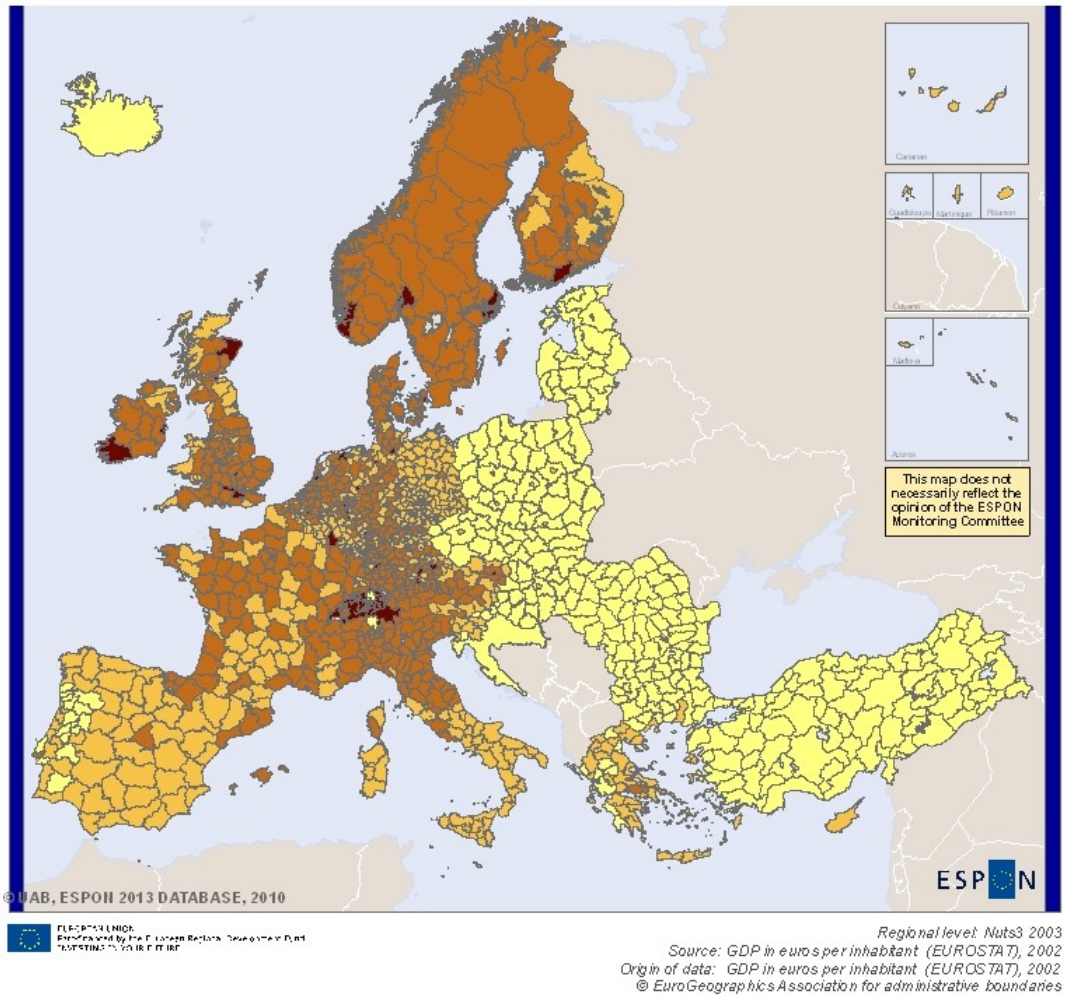
- < 5 %
- 5 - 10
- 10 - 15
- >15 %

Although it seems a map where NUTS3 have been coloured as such, the data are stored by 1 km grid cell and, therefore, they can be compared with other data stored under the same grid coding, such as land cover data.

Testing the "Proportional and weighted calculation" method: GDP – Wealth and Production (2002)

a) Source data: In order to test the third aggregation method, we have chosen an economic variable, the GDP in euro per inhabitant 2002 (Eurostat) (Map 5. GDP €/inhab. 2002), and decided to weight its values by the population living in each 1 km grid cell. In this way, the GDP value is downscaled in a more realistic manner. As for population, we have used the JRC's population density grid dataset[11] for the year 2001. In this grid, population data for communes is remapped based on Corine Land Cover classes and a quite complex algorithm[12].

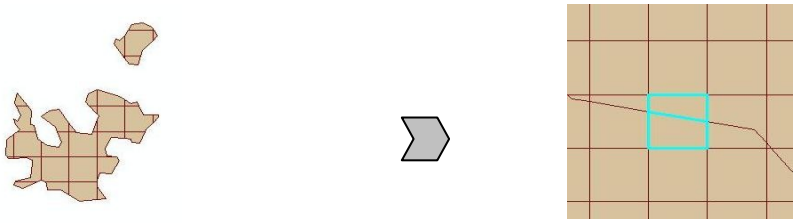
Map 5 . GDP €/inhab. 2002 distributed by Nuts3 2003



[11] http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_RESEARCH/PGE_RESEARCH_NTTS/S14P3%20-%20JAVIER%20GALLEGO%20-%20%20DOWNSCALED%20POPULATION%20DENSITY.PDF

12 http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-1

b) Process steps: In this case, the GDP is measured by NUTS3 regions. Therefore, the first step, as in the previous two cases, is overlaying the layer in which the data is given with the 1 km Reference Grid. After that, a single figure should be calculated for each 1 km grid cell, depending on the values coming from the NUTS3 regions overlaying it. If more than one value is shared by a grid cell, the final figure is calculated proportionally with regard to the area that each value occupies within the cell.



GRID CODE	AREA	GDP	GDP * AREA
GRIDCODE1	A1	GDP1	A1 * GDP1
GRIDCODE1	A2	GDP2	A2 * GDP2

GRID CODE	GDP
GRIDCODE1	A1 * GDP1 + A2*GDP2

Finally, the value of GDP per capita that has been calculated by each grid cell is multiplied by the population figure in that cell, giving a final GDP value which reflects not only the richness of the region but also the distribution of that richness amongst the inhabitants. The next figure presents the general schema followed to calculate the indicator.

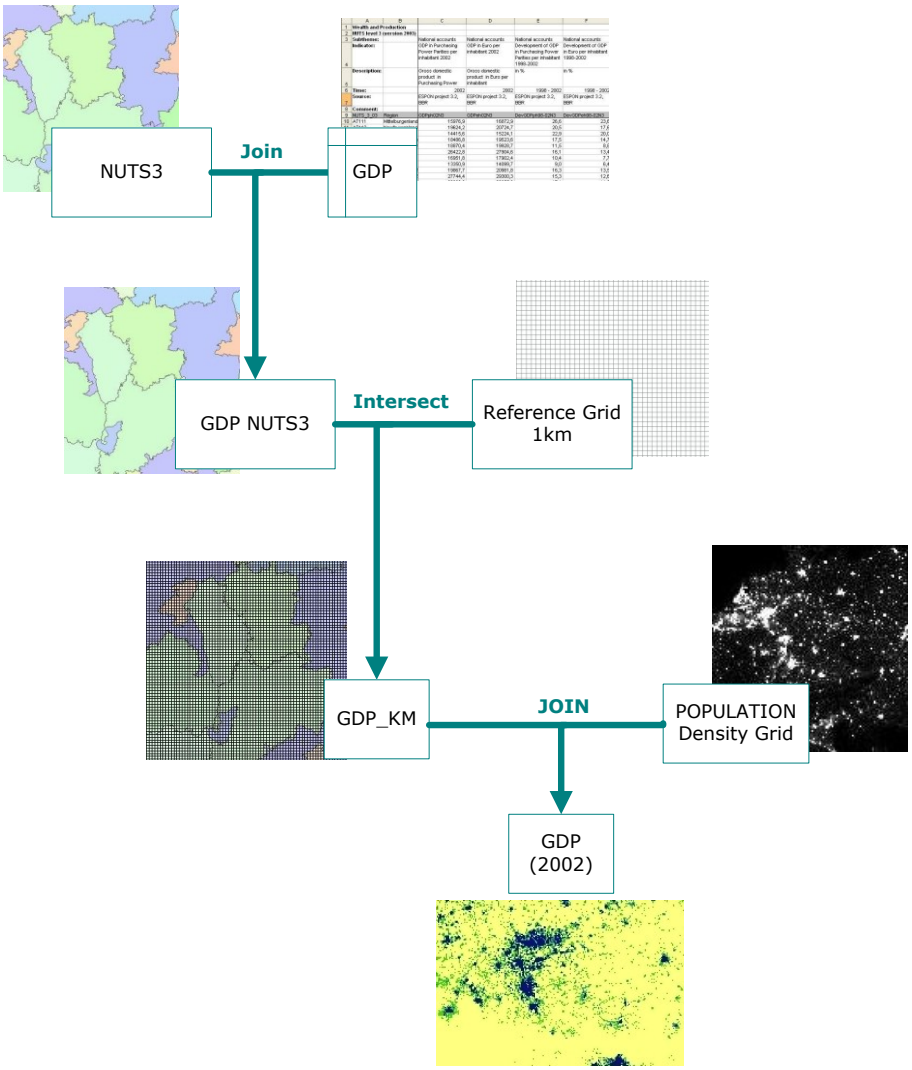


Figure 7. *Creation process schema for GDP downscaling.*

Exceptions: To obtain the final GDP (2002) weighted by the population it has been necessary to use the population density grid but also the population 2003 distributed by NUTS3 (Eurostat). The reason is that the population density grid doesn't cover all the extension of the layer NUTS3 , this mainly happened in islands zones like Canary islands. In these regions, to be able to calculate the GDP total, the weighted process has been made with the information of the population from Eurostat.

	A	B	C	D	E	F	G
1	**Population**						
2	**NUTS level 3 (version 2003)**						
3	**Subtheme:**		Population structure	Population structure	Population structure	Population structure	Population structure
4	**Indicator:**		Average Population 2003	Average male Population, share in %, 2003	Average female Population, share in %, 2003	Population density 2002	Development average population 1995-2003 in %
5	**Description:**						
6	**Time:**		2003	2003	2003	2002	2003
7	**Source:**		Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices	Eurostat; Norway and Switzerland: National Statistical Offices
8	**Comment:**		UKM + UKN = 2002	UKM + UKN = 2002	UKM + UKN = 2002		UKM + UKN = 2002
9	NUTS_3_03	Region	AvgPopN303	AvgmPopN303	AvgfPopN303	PopdensN302	DavgPop9503N3
25	AT222	Liezen	81.800	48,4	51,5	24,9	1,1
26	AT223	Östliche Obersteiermark	173.700	48,4	51,6	57,4	-10,6
27	AT224	Oststeiermark	268.400	49,3	50,7	77,3	4,8
28	AT225	West- und Südsteiermark	190.700	49,1	50,9	84,5	3,0
29	AT226	Westliche Obersteiermark	108.200	49,0	50,9	36,2	-4,0
30	AT311	Innviertel	273.200	49,2	50,9	96,8	1,9
31	AT312	Linz-Wels	531.300	48,5	51,5	302,4	0,6
32	AT313	Mühlviertel	202.700	50,0	50,0	75,3	4,9
33	AT314	Steyr-Kirchdorf	152.600	49,0	51,0	69,3	1,1
34	AT315	Traunviertel	227.100	48,7	51,3	89,7	2,9
35	AT321	Lungau	21.300	49,3	50,7	21,5	-1,8

Figure 8. *Eurostat's 2003 population source data.*

The next figure highlights in blue and red the main zones not covered by the population density grid and where it has been necessary to use the population information provided by Eurostat.

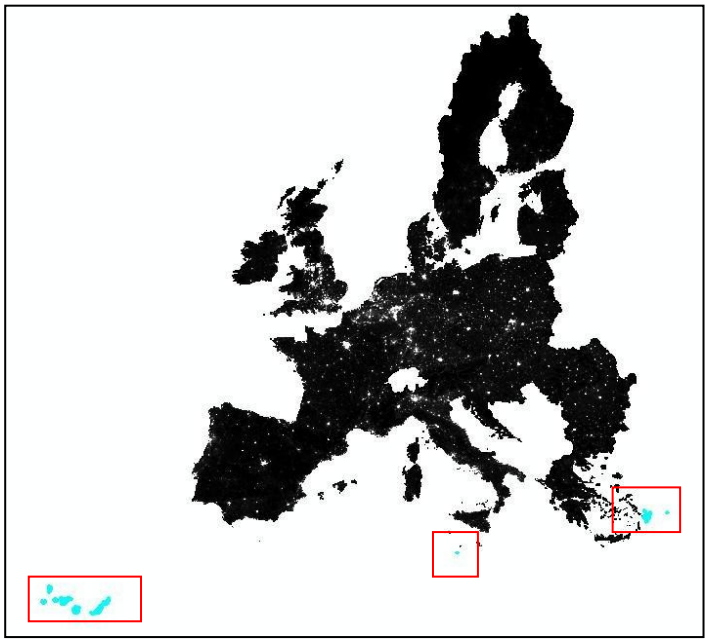
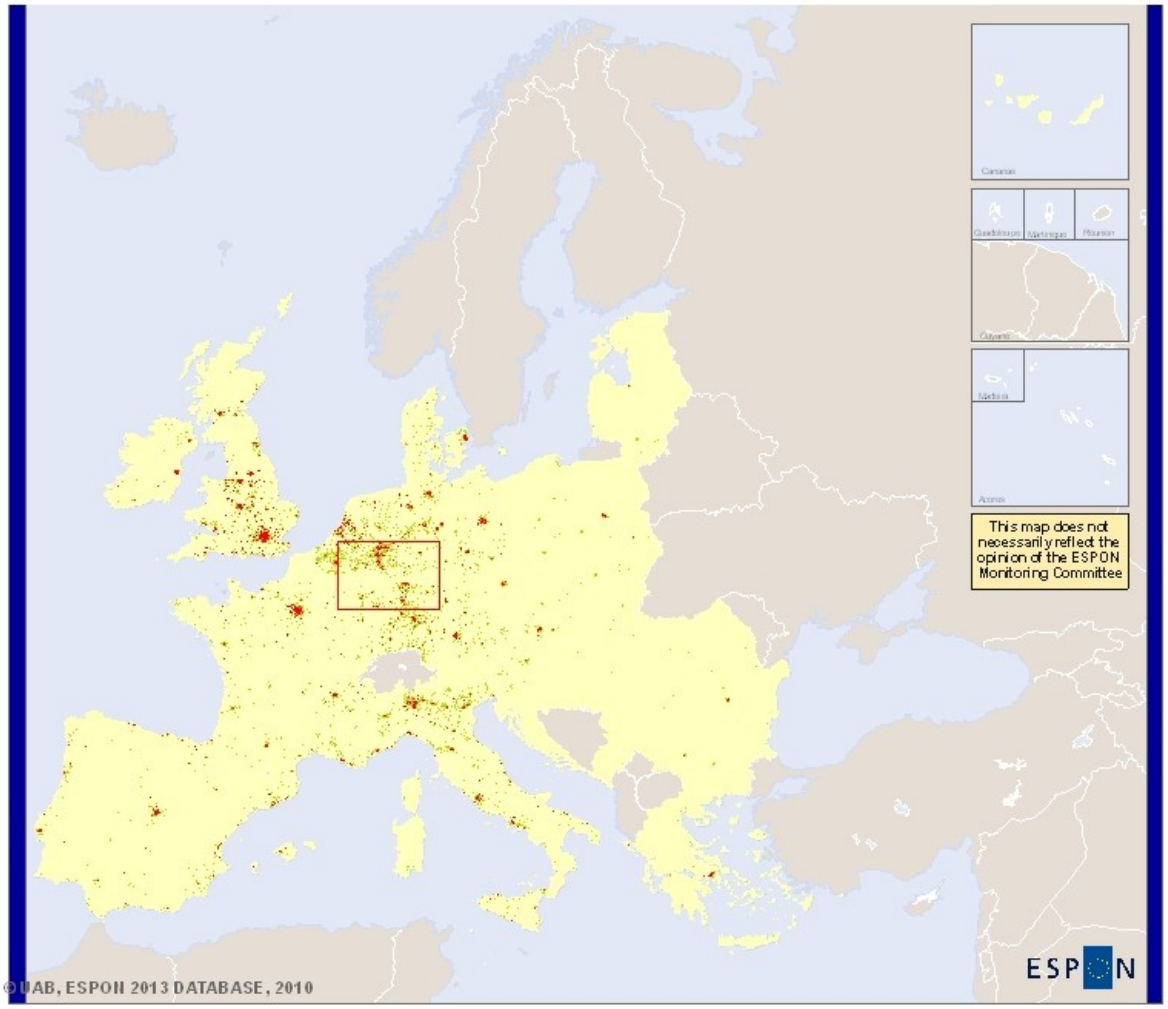


Figure 9. *Location of areas out of population grid's scope.*

Results: when we put the results on a map, we have the following layouts:

Map 6 . Distribution of GDP in Euro 2002 by grid

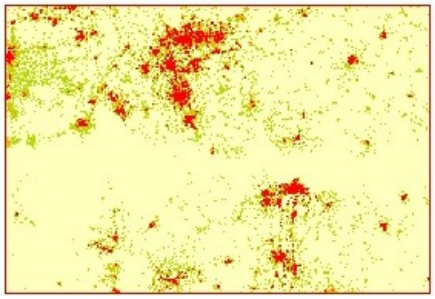


Regional level: Grid
Source: GDP in euros per inhabitant (EUROSTAT), 2002
Origin of data: GDP in euros per inhabitant (EUROSTAT), 2002
© EuroGeographics Association for administrative boundaries

Wealth and Production

GDP in Euro 2002

- < 10 million euros
- 10 - 30
- 30- 50
- > 50 million euros



As it is obvious according to the method, the GDP is concentrated in the biggest urban areas, where most of the people are living and somehow higher in the grid cells belonging to the richest regions in Europe. Consequently, this method of redistributing and weighting data by grid cells is useful to be somehow independent of the administrative (arbitrary) divisions. This case is highlighted for example in the south-west of Ireland and in the north of Italy.

a) South-west of Ireland:

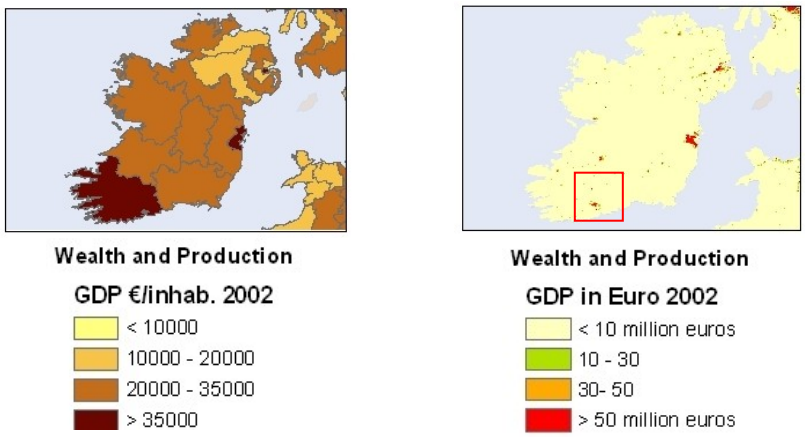


Figure 10. *Original GDP €/inhab distributed by Nuts3 2003 vs. GDP in Euro distributed by grid at the South-west of Ireland.*

In this case the Nuts3 region (IE025) is very big, but the richness is concentrated mainly around the Cork city (a small dot at the mapped scale, highlighted with a red square).

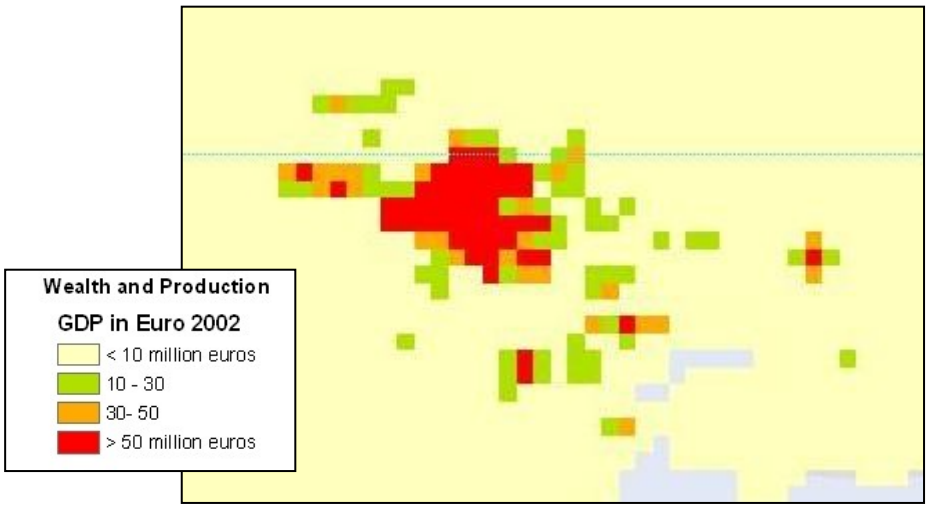


Figure 11. *Zoom in on the Cork City at the south-west of Ireland*

b) North of Italy:

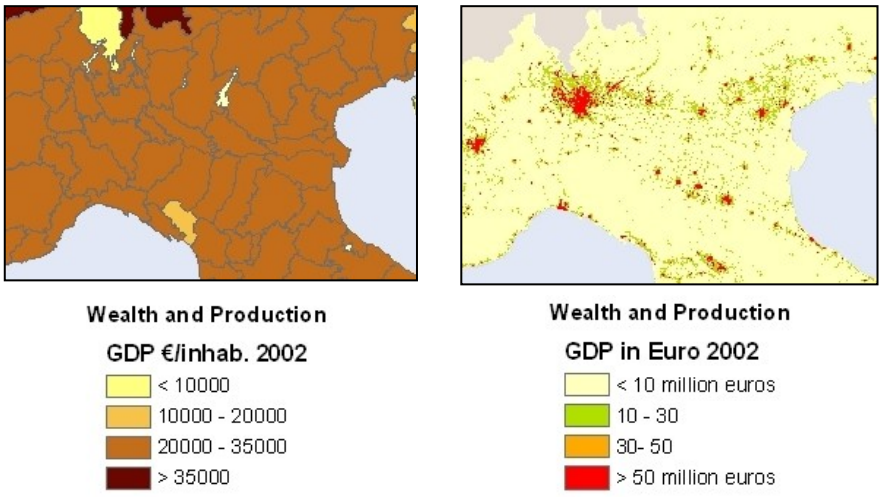


***Figure 12**. Original GDP €/inhab distributed by Nuts3 2003 vs. GDP in Euro distributed by grid at the North of Italy.*

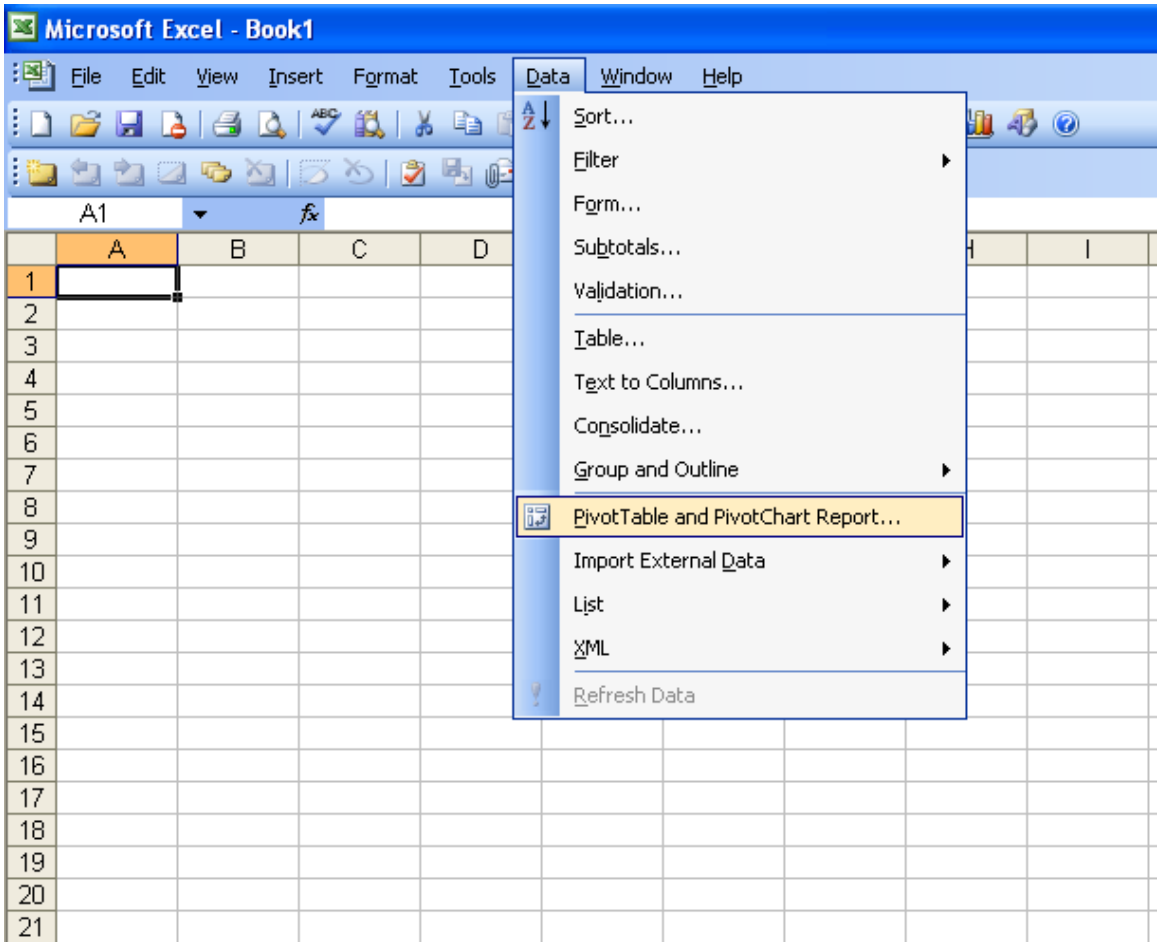
In the case of the north of Italy, the regions are much smaller and have quite a high GDP/inhab. When theses values are weighted by population and distributed by grid cells they better show the concentration of richness in the big cities, like Milano in this case.

ANNEX 2. ESPON OLAP Cube User Manual

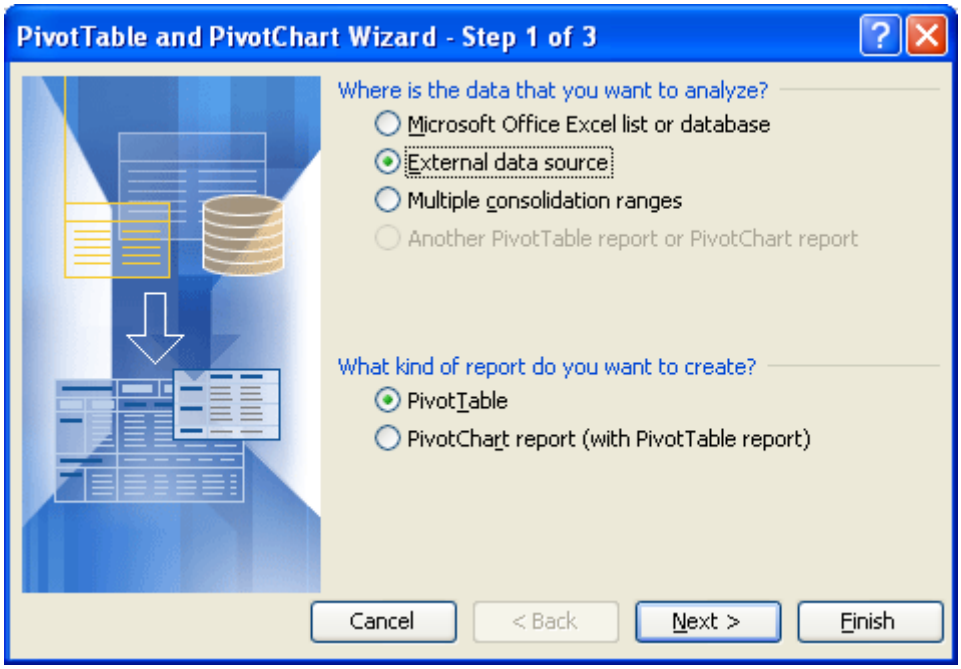
An OLAP Cube can be queried **online** and **offline**. So far, the online connection has not been implemented. In order to use the cube, a single file **.CUB**, which works offline, has been provided.

The .CUB file can be connected to and queried from Microsoft Excel with a few steps detailed in the following pages:

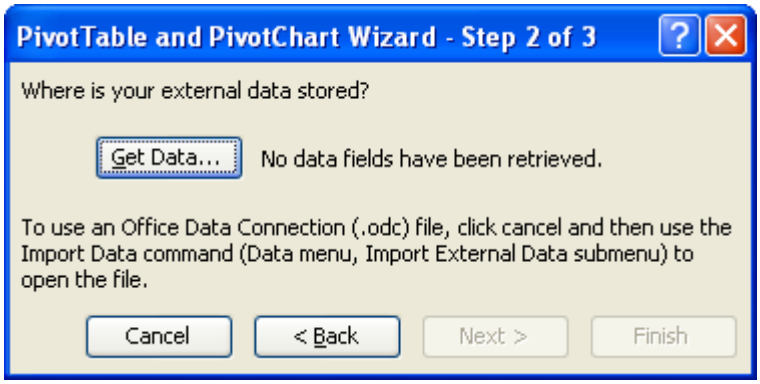
- Select "Pivot Table…" in the Data Menu



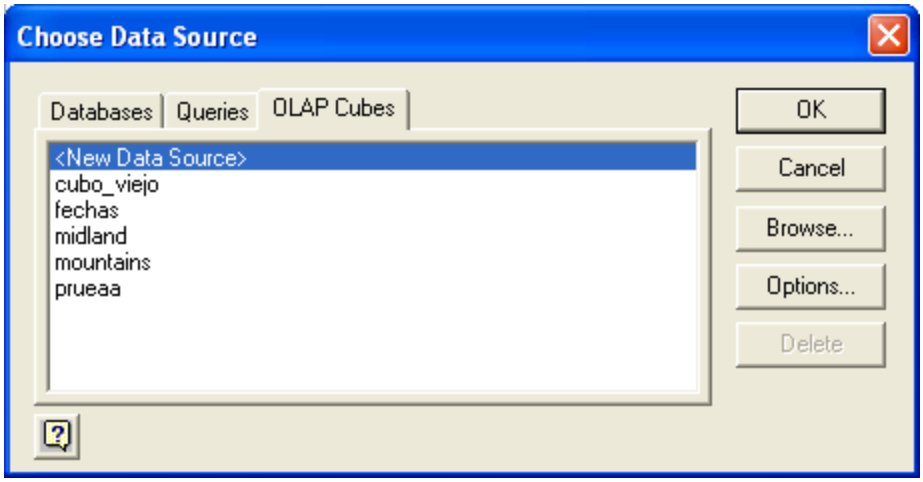
- Select “External data source” and Pivot Table as report type



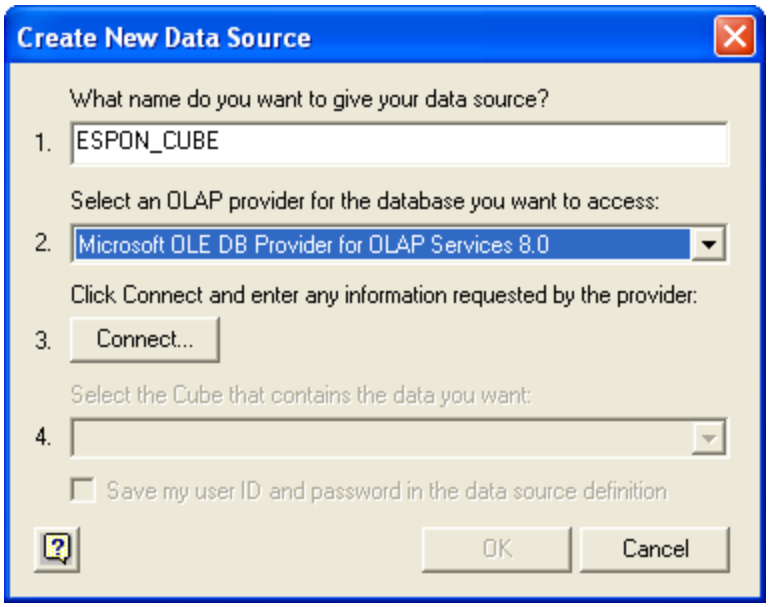
- Click on “Get Data…”



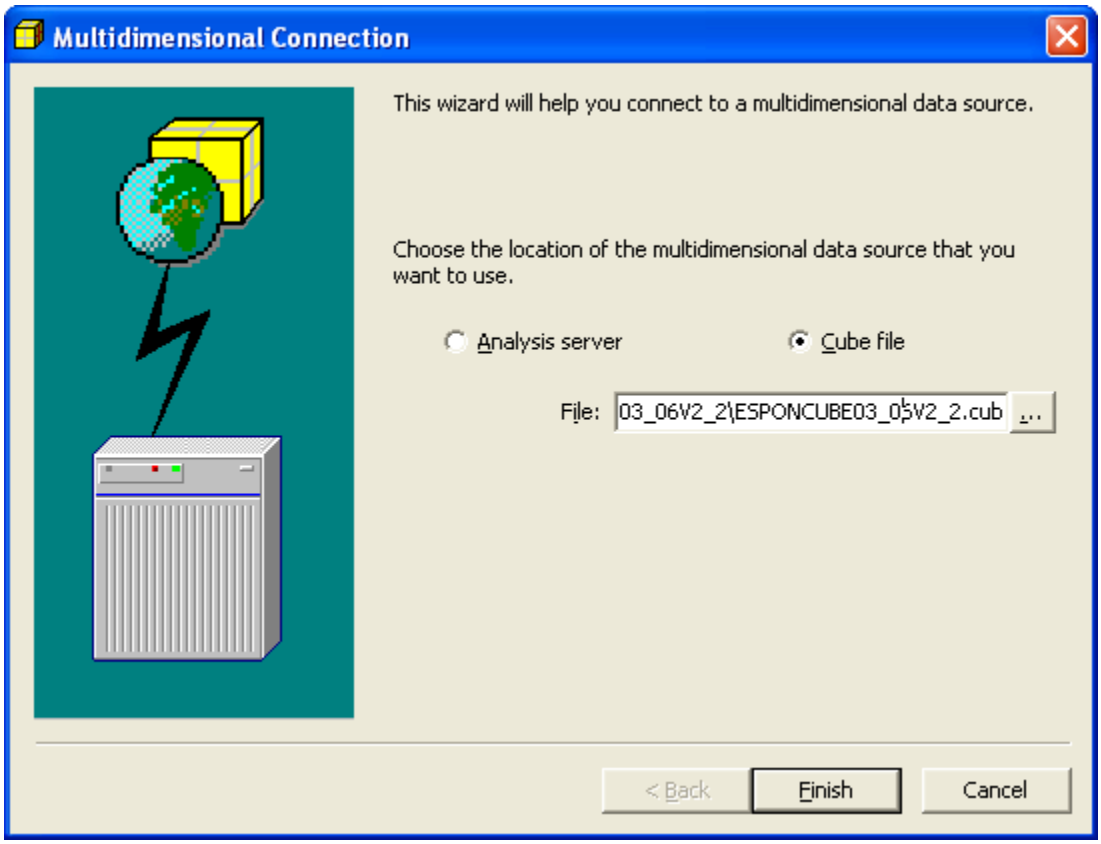
- Go to “OLAP Cubes” Tab. Choose <New Data Source>. Click OK



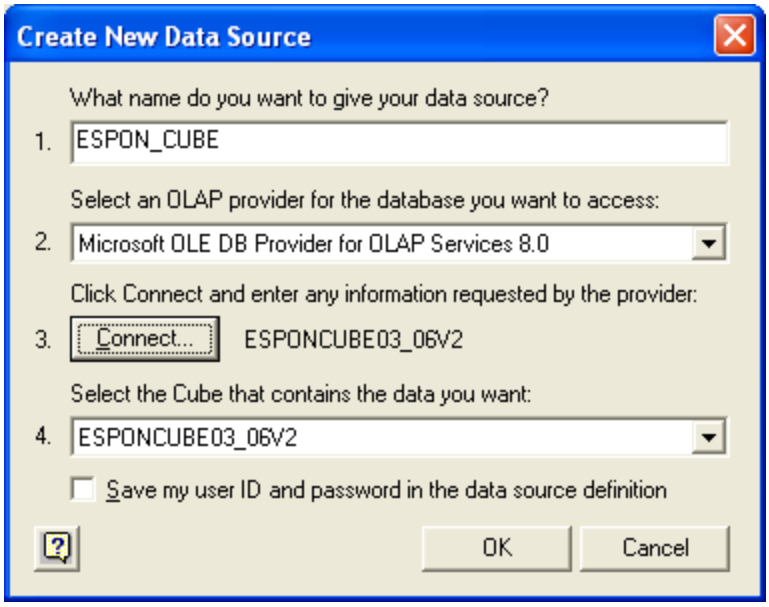
- Write down a name for your connection. Choose "MS OLE DB Provider for OLAP Services 8.0 (note: this component should be installed in order to connect to an OLAP Cube). Click Connect… button.



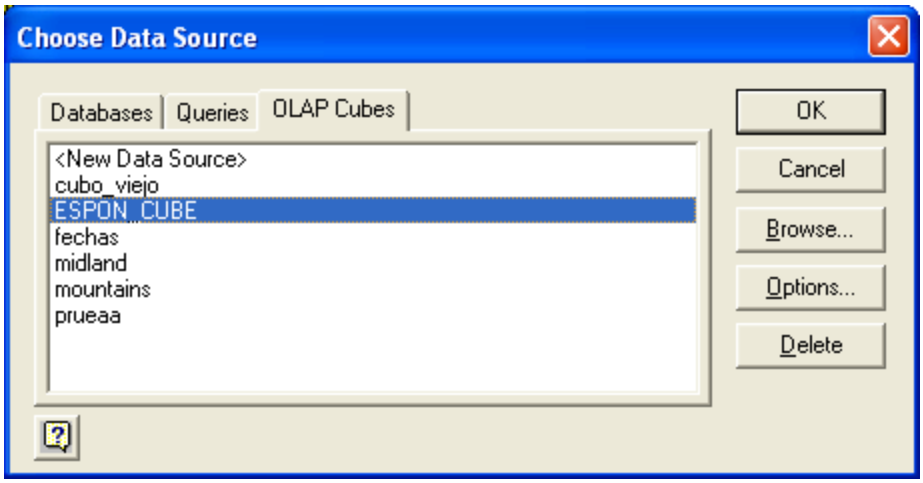
- Choose Cube file, and browse and choose the .cub file in your computer. Click Finish.



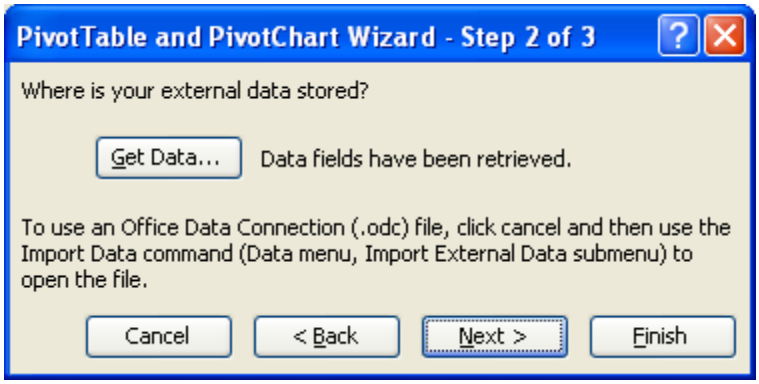
- Click OK



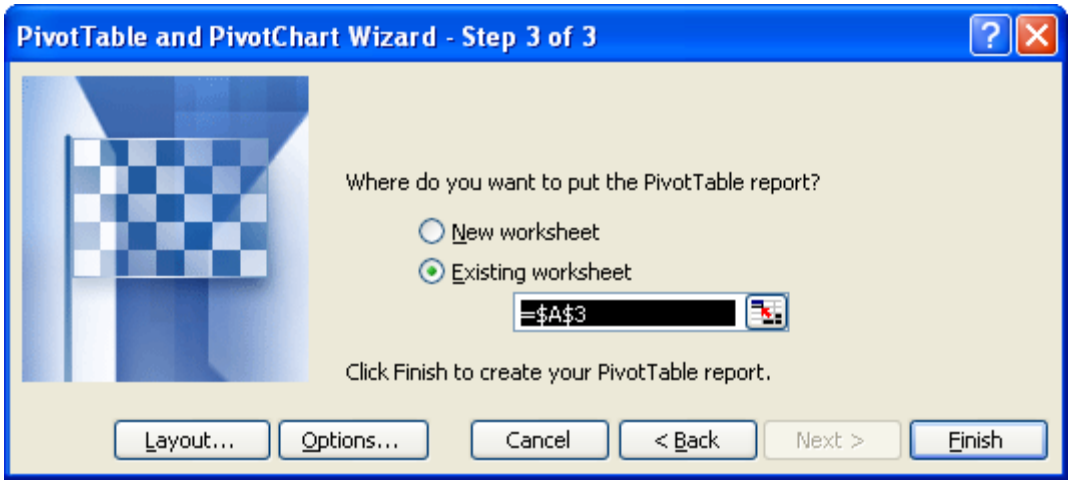
- Choose the connection just created. Click OK



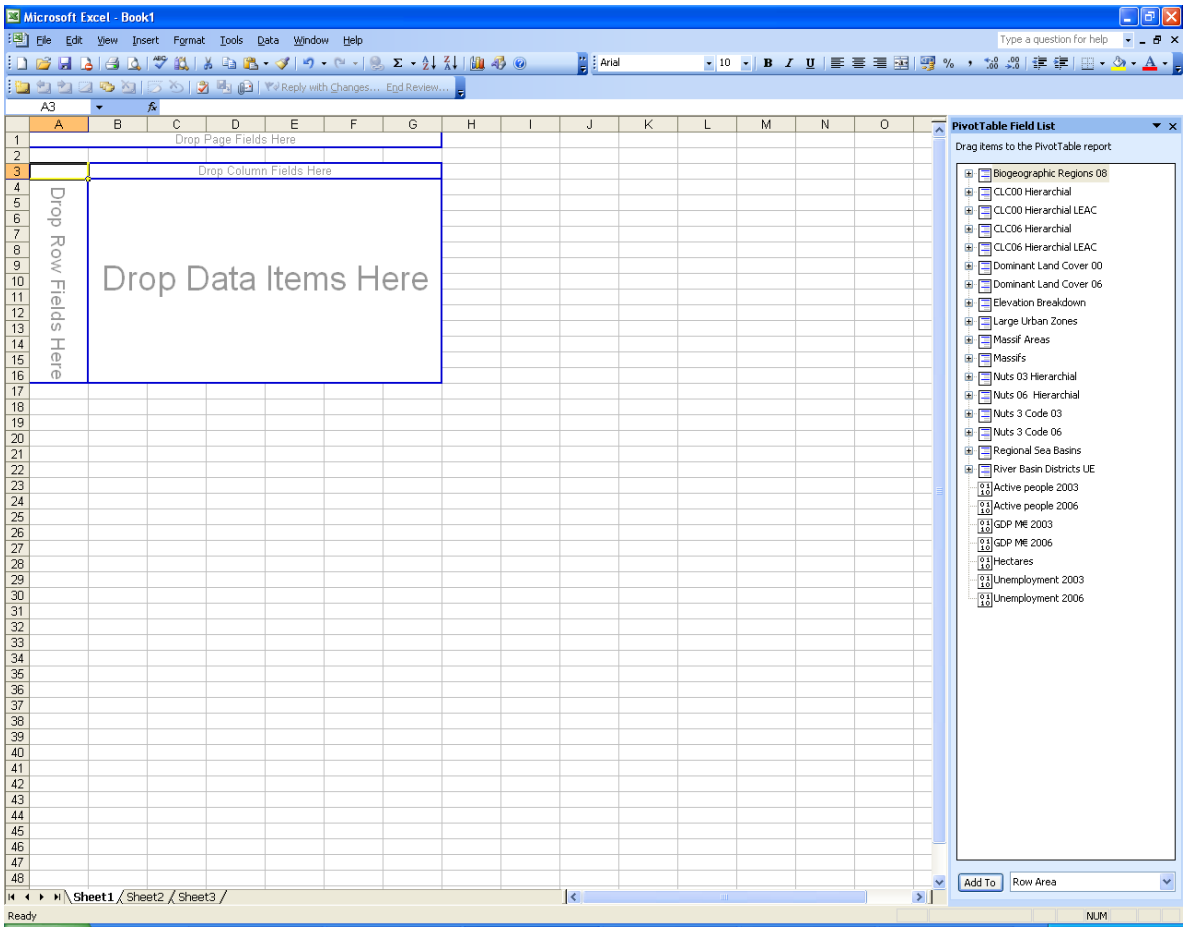
- Click "Next >"



- Choose either a new or existing worksheet. Click Finish.

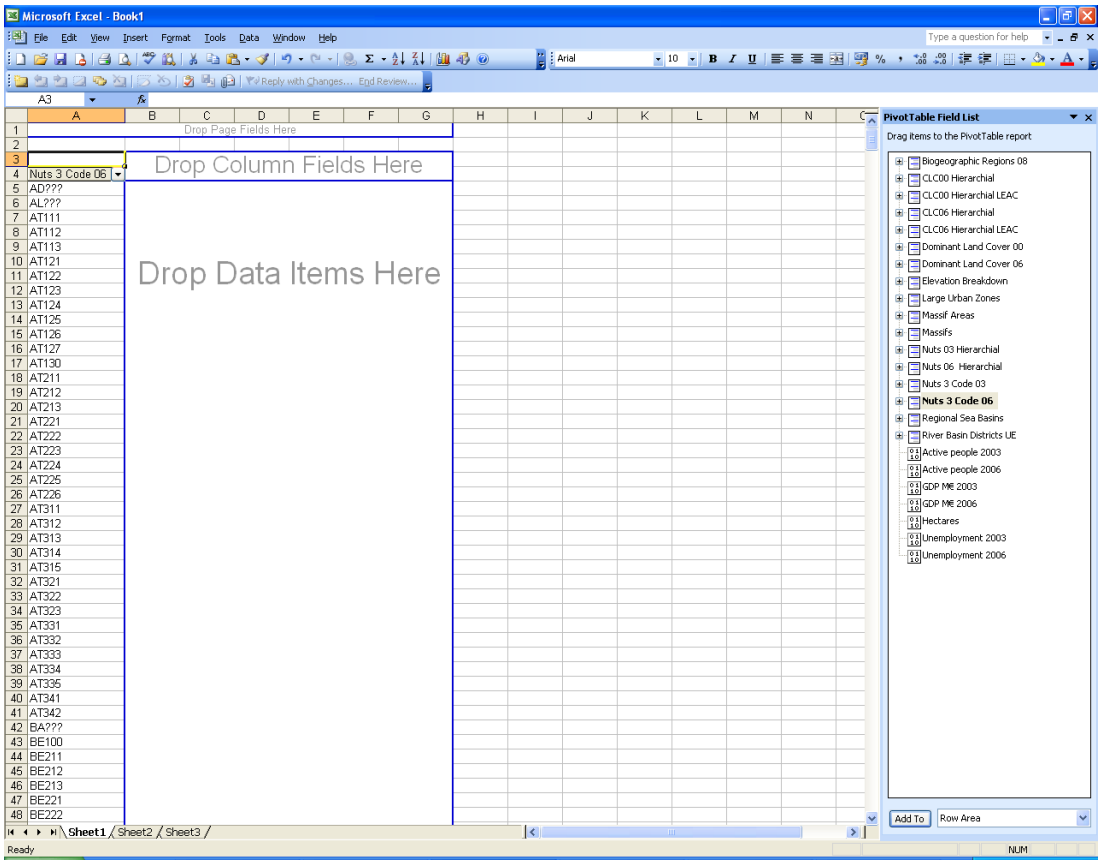


- The Cube is ready!

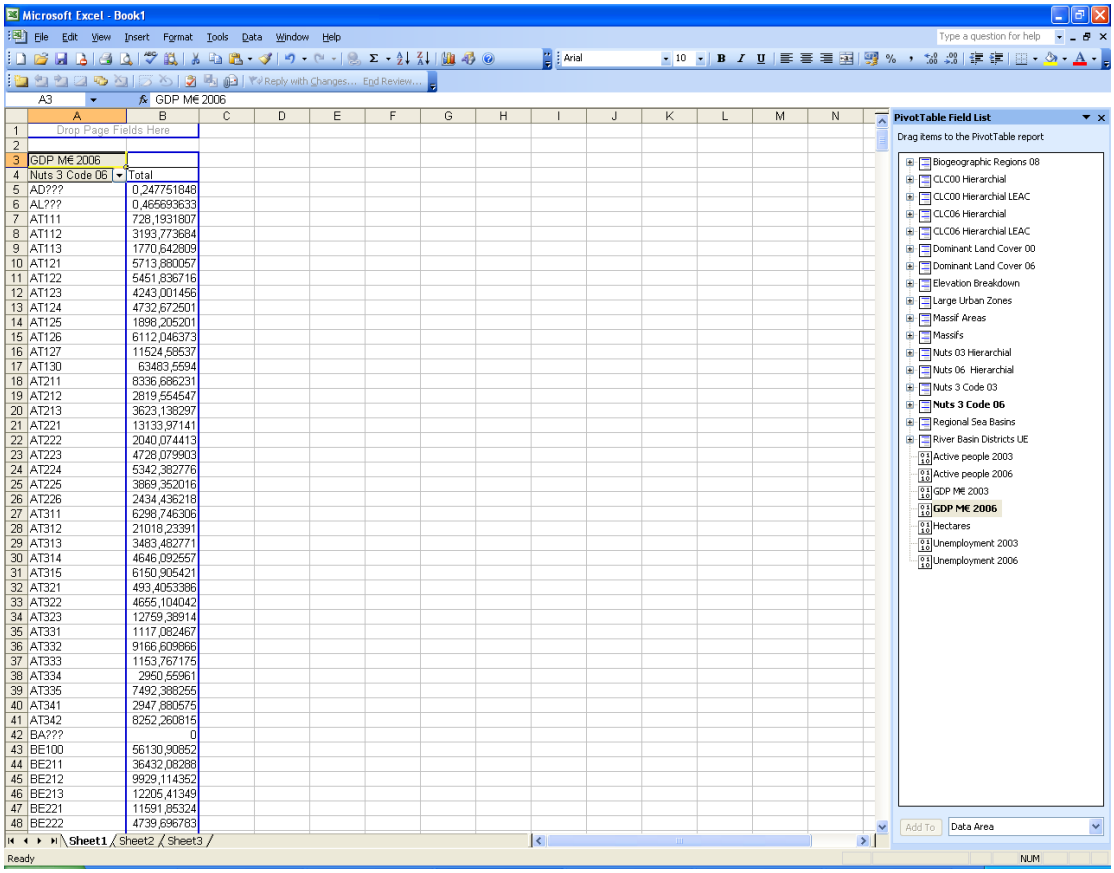


How to query the Cube

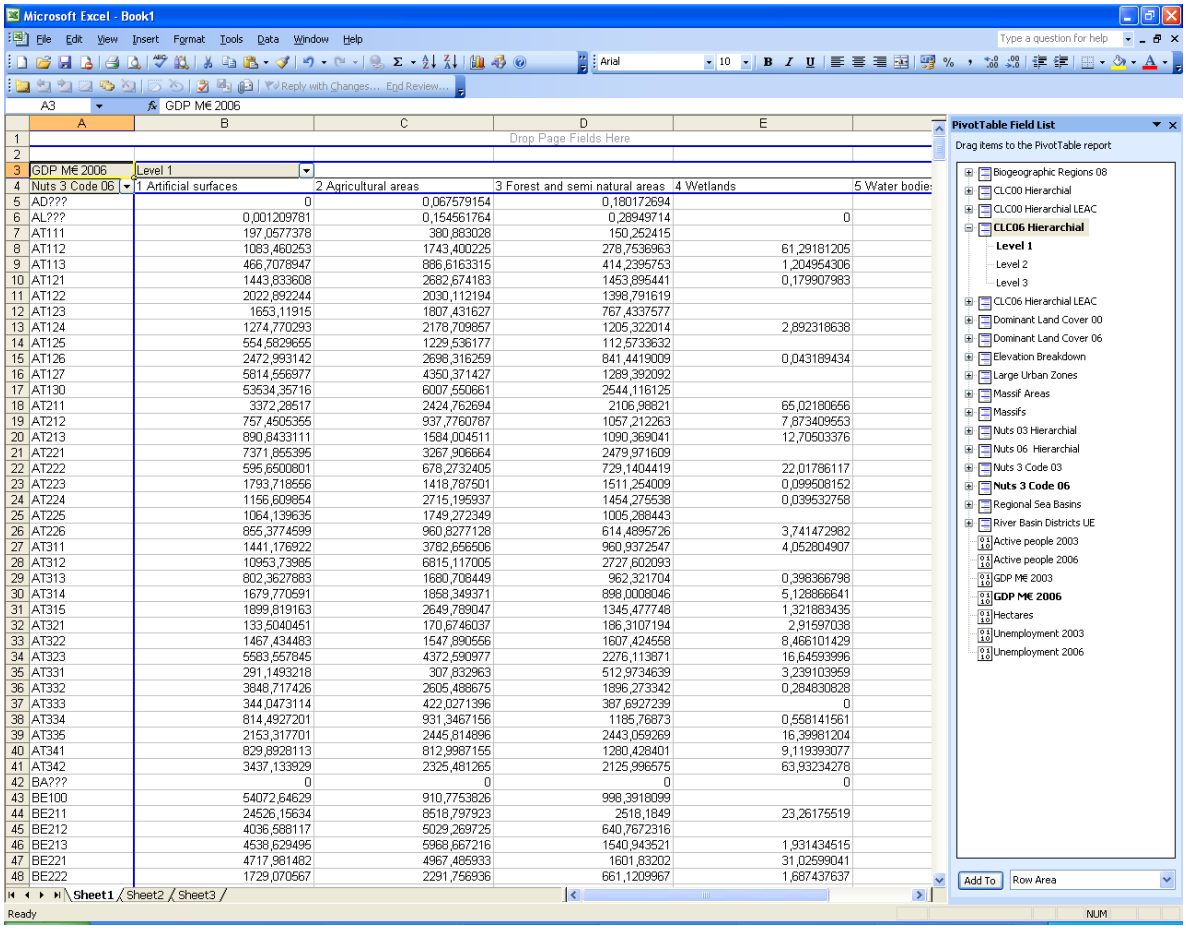
- Drag & Drop dimensions (e.g. Nuts 3 Code 06) in the Row Area



- Drag & Drop measures (e.g. GDP M€ 06) in the Data Area



- Drag &Drop other measures (e.g. CLC06 hierarchial Level 1) in the Column Area



- Play around with the ESPON OLAP Cube to build new queries.
- You can also build Pivot charts.
- If you are using MS Excel 2007, please check:

http://office.microsoft.com/en-gb/excel/CH062528071033.aspx

- Documentation about OLAP and OLAP Cubes:

http://en.wikipedia.org/wiki/Online_analytical_processing

http://en.wikipedia.org/wiki/OLAP_cube

TECHNICAL REPORT

ESPON

SPATIAL ANALYSIS FOR QUALITY CONTROL

Phase 1: The identification of logical input errors and statistical outliers

MAIN RESULTS

- Exceptional values can arise from logical input errors and true outlying data.
- The accurate identification of an exceptional value is important as input errors should be treated differently to true outlying data.
- Input errors can usually be identified mathematically or sometimes, statistically. Outliers are identified statistically.
- Techniques to statistically identify outliers are presented using worked examples that have been coded with R open source software.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

81 PAGES

LIST OF AUTHORS

Paul Harris, National Centre for Geocomputation (NCG), National University of Ireland (Maynooth)

Martin Charlton, National Centre for Geocomputation (NCG), National University of Ireland (Maynooth)

Contact

Paul.Harris@nuim.ie

martin.charlton@nuim.ie

tel. + 353-(0)1-7086208

TABLE OF CONTENT

Introduction

The ESPON 2013 Database should be as free from errors as possible. It follows from this that detecting errors is an important activity in both data entry and data checking. This technical report is to examine how mathematical, statistical and spatial analysis tools can be applied to the ESPON 2013 Database in order to find 'logical input errors' and 'statistical outliers'. In both cases, 'exceptional values' can arise but it is not always clear if such values relate to input errors or true values that are statistically-outlying. In this respect, reliably determining the nature of an exceptional value is important, especially as input errors should be treated differently to statistical outliers. For example, input errors are usually corrected or removed, whilst suspected outliers are usually flagged for further scrutiny.

The outcome of this report is a targeted review of existing outlier-detection tools in the field of statistics, data mining and spatial analysis, and an examination how they can assist in the detection of errors/outliers in the ESPON 2013 Database for improved quality control. This methodological review has a clear focus on spatial analysis with respect to outlier-detection; and is complemented by worked examples on an ESPON-type data set, where chosen techniques are demonstrated. Worked examples are coded using open-source software so that the applied techniques are easily transferable. The list of techniques that are applied should not be considered as exhaustive, but form a cross-section of useful techniques which are appropriate for ESPON 2013 Database.

A related aim of this report is to examine the effects of the Modifiable Areal Unit Problem (MAUP) with respect to error/outlier identification. This follows previous research by NCG for the ESPON 2006 project on this topic (ESPON 2006).

1 Exceptional values: types and identification

1.1 Logical input errors

Logical input errors can arise for a number of reasons. For example, the wrong NUTS1 code could be specified; incorrect data values could be input; data could be repeated exactly but assigned to different variables; data could be displaced within or between columns; data could be swapped within or between columns. In general, the identification of an input error will follow some logical, mathematical approach. For example, if a land use class could only take a positive integer value from 1 to 9 say, then an input error of say, -2, 4.5 or 10 would be easily identified.

An input error may also be identified statistically. For example, if the number 27 is inadvertently entered as 72 for a region's unemployment rate, the value 72 may lie in the extreme tail of this variable's distribution and as such, is statistically-outlying. A difficulty here would be to distinguish between an input error of 72 and a true value of 72.

In this respect, when dealing with errors/outliers, most input errors can be either be corrected or removed, whilst most outliers should be flagged as: (i) suspected outliers and (ii) potential (undetected) input errors. Flagged observations would then require further scrutiny, which should ascertain whether the observation should be: (a) replaced; (b) removed; or if specifically an outlier, (c) retained or possibly down-weighted in some way (so as to provide some robust model fit or statistic of the data).

1.2 Aspatial statistical outliers: identification in univariate to multivariate data sets

A simple, graphical tool for the detection of outliers in univariate data sets is the boxplot (e.g. Frigge et al. 1989). Central to the creation of the boxplot is the inter-quartile range (Q3-Q1) around the median value Q2. Commonly, at the upper end of the distribution, the *inner fence* is defined as the value given by Q2+1.5(Q3-Q1) and the *outer fence* as the value given by Q2+3(Q3-Q1); and there are corresponding values for the lower end of the distribution. Observations whose values lie between the inner and outer fences are usually referred to as *outside* and those whose values lie beyond the outer fence are usually referred to as *far out*. In either case, such observations can be flagged as outlying, however most attention should be placed on observations that lie beyond the outer fence. In this report, we not only demonstrate the use of the standard boxplot but also an adjusted boxplot for skewed distributions (Hubert and Vandervieren 2008). For bivariate data sets, a simple extension of the boxplot, the bagplot (Rousseeuw et al. 1999) can be constructed.

[1] NUTS stands for "nomenclature of territorial units for statistics".

To detect outliers in multivariate data sets, we first demonstrate a technique where outliers are observations that have a *large* squared Mahalanobis Distance (MD^2), where the MD itself is estimated in a robust manner (Filzmoser et al. 2005). MDs are used as they take into account the covariance matrix from which the shape and size of the multivariate data set can be quantified. In this outlier detection technique, robust MD^2 values are related to some pre-determined (upper) quantile of a chi-square distribution (e.g. the 97.5th percentile), where *large* robust MD^2 values lie above this pre-determined threshold. Furthermore, to address subjectivity in choosing the threshold, the technique automatically adjusts the pre-determined threshold (downwards or upwards) via simulation reflecting specific properties of the sample data. The technique (called here RMD2-AQ-outlier) is applied incorporating useful graphical displays of suspected outliers.

We also demonstrate two further multivariate techniques that each use principal component analysis (PCA) to reduce the dimensions of the multivariate data set, where in the resultant transformed space, outliers may be more readily observable. Of the many PCA-based techniques for outlier detection that have been proposed (e.g. see Rousseeuw et al. 2006; Daszykowski et al. 2007; Filzmoser et al. 2008), we demonstrate: (a) the 'sign' approach of Locantore et al. (1999) (call this technique, PCA-outlier-1) and (b) the 'PCOut' approach of Filzmoser et al. (2008) (call this technique, PCA-outlier-2). Both techniques are computationally fast and thus suited to large, high dimensional data sets (see the comparisons given in Filzmoser et al. 2008).

1.3 Spatial statistical outliers: identification in univariate data sets

Commonly outlier detection techniques ignore any spatial element to the data. Data not observed as an outlier when an *aspatial* technique is used, may nevertheless be a *spatial* outlier. Therefore it is important to consider spatial aspects if false negatives (i.e. outliers undetected by an aspatial technique) are to be avoided. In this respect, we demonstrate a technique of Hawkins (1980) to detect spatial outliers in univariate data sets[2]. This technique has much in common with the more recent techniques of Lui et al (2001); Kou et al. (2005).

For this technique, all observations $z(\mathbf{x}_i)$ are suspected a priori as spatial outliers, where $z(\mathbf{x}_i)$ is a spatial outlier if

$$\left(\sqrt{N}\left(z(\mathbf{x}_i) - m_l\right)\right) \Big/ \left(\sqrt{N + \bar{s}_l^2}\right) \geq \chi_{\ldots it-} \qquad (1)$$

Here, $i = ,\ldots,n$; $\mathbf{x}$ is spatial location; N is the number of neighbouring values of $z(\mathbf{x}_i)$; m_l is the local mean; $\bar{s}_l^2$ is the average variance for equivalently sized neighbourhoods across the sample area (i.e. the average local variance) and $\chi_{\ldots it-}$ is

[2] We only present a technique to identify spatial outliers in a univariate sense. Extensions to bivariate and multivariate spatial data sets are not considered here. However our current research in this area concerns the development of geographically weighted PCA techniques with respect to outlier identification (see Charlton et al. 2010), which should allow the identification of multivariate spatial outliers in the ESPON database.

a critical value of the chi-squared distribution for 1 degree of freedom. As there is no objective function for cross-validation, then neighbourhood definitions (for the local mean and variances) are chosen subjectively for this test statistic. In this report, the local mean and variances are found using a geographically weighted approach (see sections 2.4 and 2.5), with 95%, 99% and 99.9% critical levels chosen as appropriate cut-offs.

1.4 The use of statistical models and residual data in outlier identification

In a statistical analysis, it is common to identify outliers via large (positive or negative) prediction errors (or residuals) from some predictive model fit. Observations that are poorly predicted produce large residuals when compared with the actual data, and are therefore deemed as outlying. The key drawback to this approach is the need to specify a model in the first place, where different models may produce different outlying observations. However if several prediction models are applied, then it is reasonable to expect that the most influential outlying observations should be repeatedly identified.

In this respect, we first identify outliers (in a univariate sense) simply using the key component of expression 1, where a spatial outlier relates to a large (absolute) value of the error $z_i - m_l$. Here our prediction model is simply the one chosen to find the local mean m_l, which in this case is some simple spatial predictor using geographical weights (which we shall call the local mean predictor, LM). The widely-used inverse distance weighting model would be one example of such an LM model.

Furthermore, we also identify outliers (via residual data) using univariate and multivariate regressions in both aspatial and spatial forms. In particular we apply: (a) standard multiple linear regression (MLR) models, (b) attribute-space local regression (LR) models (see Loader 2004) and (c) geographic-space local regression models (Fotheringham et al. 2002) (i.e. geographically weighted regression, GWR). Here LR accounts for nonstationarity and nonlinearity in attribute-space, whilst GWR accounts for nonstationary and nonlinearity in geographic-space. Both LR and GWR are nonparametric in design. The conventional MLR model assumes stationarity and linearity in both attribute- and geographic-space; and is parametric in design. Consequently, each of the three regression forms will identify outliers (or possibly groups of outliers, see section 2.5) according to their particular specification (or set of modelling assumptions).

The investigation of residual data plays a central role in the formulation of a robust regression model, where the influence of outlying data on the regression fit is reduced (e.g. see Faraway 2004, p98-106; Cruz Ortiz et al. 2006). MLR, LR (see Loader 2004) and GWR (Fotheringham et al. 2002, p73-82; Harris et al. 2010) all have robust forms. Commonly, a robust regression will identify outliers as observations with large standardised (or studentised) residuals via a leave-one-out approach. However, in this report we only identify outliers simply, via the raw residuals and without the benefit of a leave-one-out fit.

1.5 The identification of spatial clusters

A group of observations identified as outliers may actually be spatially clustered with a substantive reason for their 'unusualness' (i.e. false positives are to be avoided as well). In this respect, it is worthwhile applying techniques that identify local (or regional) changes in the spatial process according to some key moment or relationship[3].

Furthermore, seemingly significant clusters can be sometimes be attributable to only a few (influential and outlying) observations; so although the local techniques described below are not specifically designed to identify spatial outliers, they sometimes do so. Indeed, a corresponding robust form of the given local technique would out of necessity identify spatial outliers in order to reduce their influence.

Thus in the first instance, local summary univariate and bivariate statistics are calculated and investigated. In particular, we assess changes in the mean, standard deviation and correlation across space, where these (spatial) moments are all found in a geographically weighted form (Fotheringham et al. 2002)[4]. For the multivariate case, GWR can be applied, which complements a local correlation analysis when investigating relationship-change across space.

From a spatial autocorrelation viewpoint, a local version of Moran's I (Anselin 1995) is used. Positive spatial autocorrelation exists when neighbouring spatial units tend to have similar values of a variable; whilst negative spatial autocorrelation exists when they do not. Local Moran's I is only used to investigate univariate data, but the statistic could be adapted to investigate cross-autocorrelation in bivariate and multivariate data sets.

1.6 Summary: MAUP, temporal outliers and data imputation

We have presented a typology of techniques where variables are analysed singly or in combination; and aspatially or spatially. Underlying all of these techniques is the spatial structure of the reporting units, where results can be influenced not only by the level of spatial aggregation used but also by the spatial configuration of the reporting units (i.e. a MAUP; e.g. see Wong 1996). In this report we demonstrate the consequences of the MAUP for outlier identification via a worked example, where outlier-detection techniques are applied at different NUTS levels (NUTS level 3 through to NUTS level 0).

We have not addressed the identification of temporal (or by extension, spatio-temporal) outliers. This is not an oversight, as ESPON time series data is not expected to be of a sufficient length for an outlier detection technique to be reliably

[3] Brunsdon and Charlton (2010) assess the effectiveness of multiple hypothesis testing for detecting clusters of geographical anomalies. These tests would complement the techniques demonstrated from this section of the report.

[4] Robust forms of geographically weighted summary statistics (GWSS) can be found in Brunsdon et al. (2002) and in Harris and Brunsdon (2010).

applied. Instead it should suffice that the aspatial/spatial detection methods demonstrated here can be repeated at different time intervals. The consequences of the reporting units changing over time (i.e. another MAUP) are addressed elsewhere in ESPON 2013 database project.

As already discussed, once an input error has been identified the observation can either be corrected or removed (i.e. replaced with the missing value notation, NA[5]). On the other hand, suspected outliers (which may be an input error) can (after some additional scrutiny) be: (a) replaced; (b) removed (i.e. replaced by NA); or if indeed an outlier, (c) retained or possibly down-weighted in some way. This entails that some form of imputation or prediction of missing valued data will be required, and here the chosen regression models of section 2.4 may be of value.

1.7 Further reading

This report provides a brief overview to subject of error or outlier identification with respect to the task of identifying outliers in the ESPON 2013 Database. There is an extensive literature on outlier detection, where the following reading list may be useful.

- An evaluation of aspatial techniques to detect input errors and true outliers (here known as data editing), together with imputation techniques, for large scale survey data can be found in Charlton (2004). This and related articles arose from the EUREDIT project[6]. Related articles include: an outlier identification technique for multivariate data by Béguin and Hulliger (2004); a robust regression technique for data edits by Chambers et al. (2004); and a classification and regression tree technique for data edits by Petrakos et al. (2004).
- An aspatial Bayesian technique that both edits and imputes data in a multivariate context can be found in Ghosh-Dastidar and Schafer (2003).
- Reviews of aspatial outlier identification techniques from univariate to multivariate data sets can be found in Reimann et al. (2005); Rousseeuw et al. (2006); Daszykowski et al. (2007); Morgenthaler (2007).
- Further aspatial outlier identification techniques for multivariate data sets can found in Hoo et al. (2002); Jackson and Chen (2004), where the former article also imputes data.
- Imputation (aspatial) techniques can be found in Plaia and Bondi (2006); Vanden Branden and Verboven (2009), where the former article focuses on time series data.
- Alternative spatial outlier identification techniques can be found in D'Alimonte and Cornford (2007); Ainsworth and Dean (2008); Meiklit et al. (2009).

[5] NA is the missing data indicator used in the R statistical computing package (see section 4).

[6] See http://www.cs.york.ac.uk/euredit/. The project website was still active as of 1/12/09.

2 Data for worked examples

In the worked examples, NUTS3 level data are used. Here 1351 values (with two missing) for the variable 'evolution of gross domestic product (GDP) from the years 2000 to 2005' at NUTS2006 divisions are related to sixteen contextual variables at NUTS1999 divisions (with a maximum of 1329 values for each contextual variable). As the NUTS2006 spatial units can differ to the NUTS1999 spatial units, this combining of data results in at least 438 (1351 minus 913) missing values for each contextual variable (i.e. NUTS2006 and NUTS1999 divisions have 913 reporting units in common). Thus in summary, NUTS3 level data using the NUTS2006 divisions are the spatial units that are retained.

2.1 The full data set

The 'evolution of GDP' variable is named EVOGDP_2000_2005_2006, where the first two numbers (2000 and 2005) relate to the collection time (i.e. year) of the data and the last number (2006) relates to the NUTS division or version. Similar naming conventions were used for all other variables. EVOGDP_2000_2005_2006 is itself calculated from four stock variables which are presented in Table 1, together with the formula for calculating EVOGDP_2000_2005_2006.

The sixteen contextual variables are presented in Tables 2 to 6. These variables were selected from the basic and project indicator files posted on the ESPON website[7]. Contextual data include: two spatial typology variables, one unemployment variable, six land use variables, one natural hazards variable and six regional policy variables. In total, the full data set consists of twenty-three variables (plus the coordinates/centroids of each region).

Observe that as variables were collected over different time periods (from 1996 to 2005) this data set is purely used to demonstrate the outlier identification techniques of section 2 via the worked examples in section 4. It is essentially a fabricated data set and as such, all analytical results need to be interpreted with this in mind.

However the contextual variables were selected in expectation that if all variables were relatable (i.e. collected over the same period), then this particular set of contextual variables may help explain variation in EVOGDP_2000_2005_2006 (see sections 2.4, 4.5 and 4.6).

[7] See http://www.espon.eu/mmp/online/website/content/tools/832/850/588_EN.html and http://www.espon.eu/mmp/online/website/content/tools/832/873/605_EN.html

Variable type	Variable name	Indicator	Year	Unit
STOCK (1)	GDP_2000_2006	Gross Domestic Product	2000	Million Euros
STOCK (2)	GDP_2005_2006	Gross Domestic Product	2005	Million Euros
STOCK (3)	POP_T_2000_2006	Total population (annual average)	2000	Thousands inhabit.
STOCK (4)	POP_T_2005_2006	Total population (annual average)	2005	Thousands inhabit.
RATIO (5)	GDP_POP_2000_2006	GDP per inhabit. = [illegible] × 000	2000	Euros
RATIO (6)	GDP_POP_2005_2006	GDP per inhabit. = [illegible] × 000	2005	Euros
RATIO	EVOGDP_2000_2005_2006	Evolution of GDP = [illegible] × 00	2000-2005	Percentage

***Table 1**: Description of the EVOGDP_2000_2005_2006 variable*

Theme	**Spatial typology**	**Spatial typology**
Indicator	Typology Settlement Structure (nine basic types defined by population density and situation regarding centres) – 1: city core region; 2: very densely populated; 3: densely populated; 4: rural region; 5: city core region; 6: densely populated region; 7: rural region; 8: more densely populated region; 9: less densely populated region	Urban-rural typology (six basic types) – 1: High urban influence, high human intervention; 2: High urban influence, medium human intervention; 3: High urban influence, low human intervention; 4: low urban influence, high human intervention; 5: Low urban influence, medium human intervention; 6: Low urban influence, low human intervention
Original variable name	Settyp99N3	URTypN3
New variable name	SPAT_TYPE_1_1999_1999	SPAT_TYPE_2_1999_1999
Min. possible	1	1
Max. possible	9	6
Unit or variable type	CLASS	CLASS

***Table 2**: Descriptions of spatial typology contextual variables*

Theme	**Unemployment**	**Land use**	**Land use**	**Land use**
Indicator	Unemployment rate	Share of artificial surfaces	Artificial surfaces per 1000 inhabitants	Artificial surfaces per GDP
Original variable name	UNRT01N3	ArSu96N3	ArSc96N3	ArSg96N3
New variable name	UNEMP_R_2001_1999	LU_AS_1_1996_1999	LU_AS_2_1996_1999	LU_AS_3_1996_1999
Min. possible	0	0	0	0
Max. possible	100	100	100	100
Unit or variable type	PERCENTAGE	PERCENTAGE	PERCENTAGE	PERCENTAGE

***Table 3**: Descriptions of unemployment and three land use contextual variables*

Theme	**Land use**	**Land use**	**Land use**	**Environment - Hazards**
Indicator	Share of urban fabric	Share of arable land	Share of permanent crops	Sum of all weighted hazard values
Original variable name	UFL296N3	ALL296N3	PCL296N3	smwh04
New variable name	LU_UF_1996_1999	LU_AR_1996_1999	LU_PC_1996_1999	NAT_HAZ_2004_1999
Min. possible	0	0	0	10
Max. possible	100	100	100	INFINITY
Unit or variable type	PERCENTAGE	PERCENTAGE	PERCENTAGE	INTEGER

***Table 4**: Descriptions of three land use and an environmental hazards contextual variable*

Theme	Regional policy	Regional policy	Regional policy
Indicator	All Structural & Cohesion Fund expenditure	Structural Fund expenditure related to Regional Development & Productive Infrastructure	Structural Fund expenditure related to Social Integration & Human Resources
Original variable name	SFT99N3	SFR99N3	SFS99N3
New variable name	SF_CF_1999_1999	SF_R_1999_1999	SF_S_1999_1999
Min. possible	0	0	0
Max. possible	INFINITY	INFINITY	INFINITY
Unit or variable type	REAL NUMBER	REAL NUMBER	REAL NUMBER

***Table 5**: Descriptions of three regional policy contextual variables*

Theme	Regional policy	Regional policy	Regional policy
Indicator	Structural Fund expenditure related to Agriculture, Rural Development & Fishery	Cohesion Fund expenditure related to Transport	Cohesion Fund expenditure related to Environment
Original variable name	SFA99N3	SFCT99N3	SFCE99N3
New variable name	SF_A_1999_1999	CF_T_1999_1999	CF_E_1999_1999
Min. possible	0	0	0
Max. possible	INFINITY	INFINITY	INFINITY
Unit or variable type	REAL NUMBER	REAL NUMBER	REAL NUMBER

***Table 6**: Descriptions of three regional policy contextual variables*

2.2 Data subsets and analytical objectives

Subsets of the full data set are analysed in two basic forms: (a) in their original state and (b) in a state where some commonly encountered logical input errors are deliberately introduced. Here Tables 7 and 8 summarise how variable subsets of the full data set are used in each of six worked examples presented in section 4.

Worked example	1	2	3
Variables investigated	NUTS3 code, GDP_2000_2006, GDP_2005_2006, POP_T_2000_2006 & POP_T_2005_2006	EVOGDP_2000_2005_2006 (plus the coordinate data)	Some subset of EVOGDP_2000_2005_2006, its 16 contextual variables & the coordinate data
Introduced input errors?	Yes	No	No
Identification type: logical or statistical or both	Both	Statistical	Statistical
Identification type: univariate or multivariate or both	Univariate	Univariate	Multivariate
Identification type: aspatial or spatial or both	Aspatial	Both	Aspatial
Key statistical identification techniques applied	Boxplots	Boxplots; Hawkins test; residuals from LM, MLR, LR & GWR fits	Bagplots; Robust MD^2 analysis (RMD2-AQ-outlier); & PCA for outliers (PCA-outlier-1 & PCA-outlier-2)
Analysis objective	Identify logical input errors so that EVOGDP_2000_2005_2006 can be investigated for statistical outliers	Identify statistical outliers	Identify statistical outliers

***Table 7**: Data subsets and analytical objectives for worked examples 1 to 3*

Worked example	4	5	6
Variables investigated	EVOGDP_2000_2005_2006 in relation to some subset of its 16 contextual variables and the coordinate data	EVOGDP_2000_2005_2006 in relation to some subset of its 16 contextual variables and the coordinate data	EVOGDP_2000_2005_2006 (plus the coordinate data)
Introduced input errors?	No	No	No
Identification type: logical or statistical or both	Statistical	Statistical	Statistical
Identification type: univariate or multivariate or both	Multivariate	Both	Univariate
Identification type: aspatial or spatial or both	Both	Spatial	Both
Key statistical identification techniques applied	Residuals from MLR, LR & GWR fits	Data exploration with GWSS, GWR & local Moran's I	Boxplots; Hawkins test; residuals from LM, MLR, LR & GWR fits
Analysis objective	Identify statistical outliers	Identify statistical clusters	Investigate the consequences of MAUP with respect to outlier identification

***Table 8**: Data subsets and analytical objectives for worked examples 4 to 6*

It is envisaged that when identifying exceptional values in an ESPON database data set, a first pass should identify input errors using both mathematical and statistical techniques. That is the first pass screens the data. Identified input errors should then be corrected (and as such, a revised data set can be *assumed* input error-free) before a second pass is undertaken that only uses the statistical techniques to identify outlying observations. It is essential that two passes are conducted otherwise the detection of true outliers will be compromised by input errors.

Thus from Tables 7 and 8, worked example 1 relates to a first pass (for input errors) before its corresponding second pass (for outliers), which is (effectively) worked example 2. For worked examples 2, 3, 4, 5 and 6, it should be assumed that this data has already been screened for input errors. Observe that worked example 6 is the same as worked example 2, but applied at different NUTS levels (i.e. spatial scales) to investigate the effects of MAUP with respect to outlier identification.

2.3 A data subset with deliberate logical input errors

We now present a list of logical input error-types that have been introduced to: (a) the NUTS3 codes; and (b) the variables GDP_2000_2006, GDP_2005_2006, POP_T_2000_2006 and POP_T_2005_2006 (i.e. only those variables used in the calculation of EVOGDP_2000_2005_2006). This list of input errors is given with appropriate solutions (i.e. for worked example 1).

This list is not exhaustive, and should grow as different input error-types become apparent (i.e. at this stage, we are not expected to foresee all input error possibilities). The spatial location of the input errors is depicted in Fig. 1. Consequences of input errors for the correct calculation of EVOGDP_2000_2005_2006 are depicted in Fig. 2.

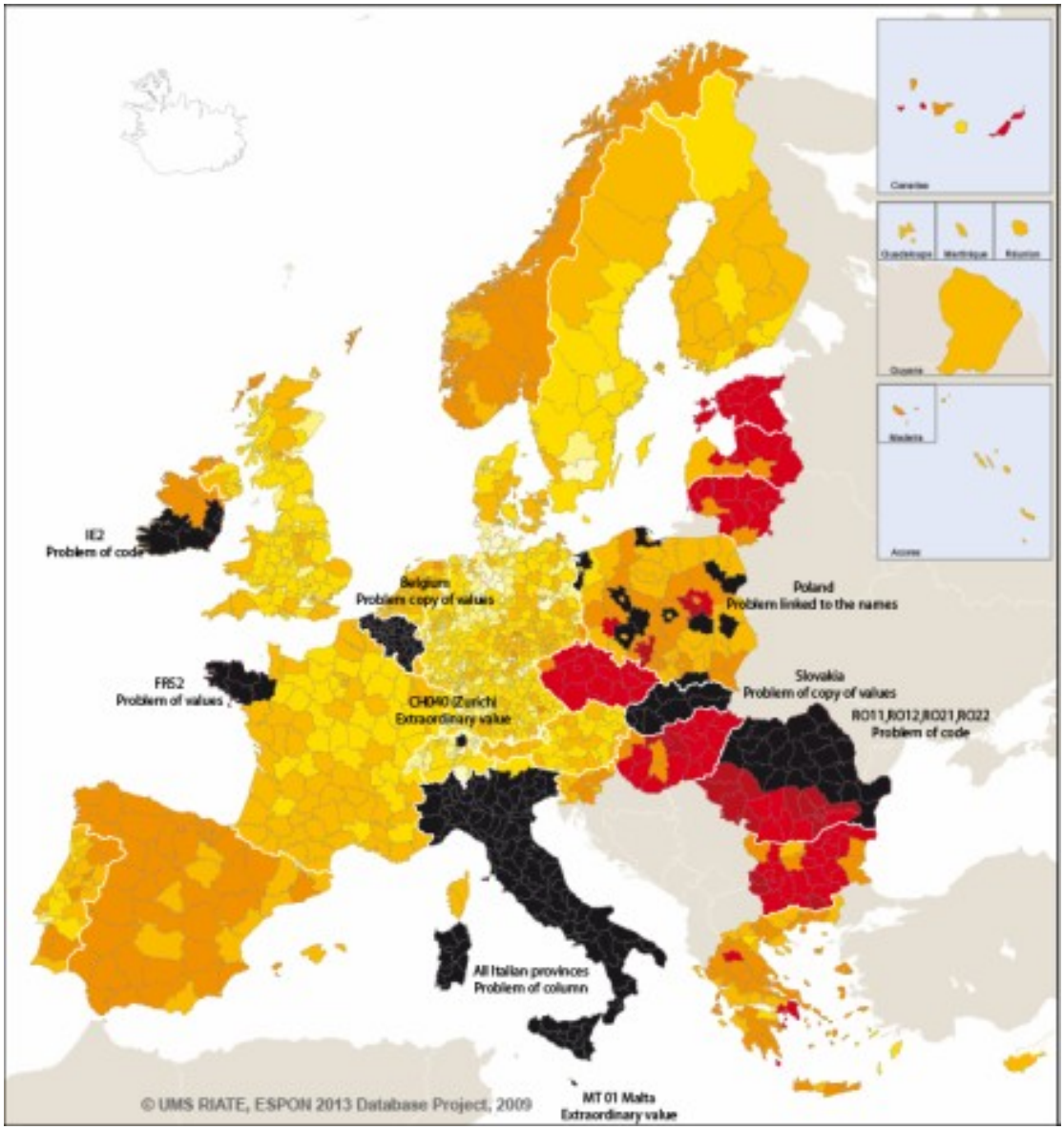


***Figure 1**: Location of input errors (in black) overlaid on the true EVOGDP_2000_2005_2006 data (see Fig. 2)*

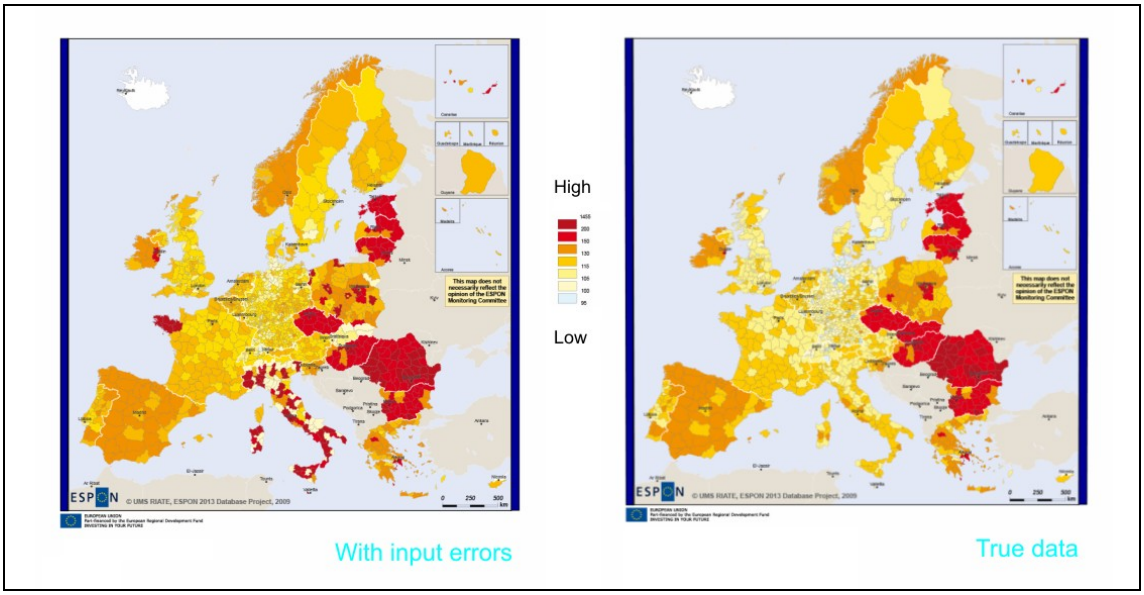


***Figure 2**: Maps of EVOGDP_2000_2005_2006 with and without input errors*

Problems with NUTS code (29 input errors)

Input error-type 1: For Ireland, 5 wrong codes have been input at NUTS3 level. In the NUTS hierarchy, this does not imply changes at NUTS2 level (see Fig. 3a). Solution: codes can be checked by a simple relationship to the correct NUTS name and code pairs.

Input error-type 2: For Romania, 24 wrong codes have been input at NUTS3 level. In the hierarchy, this does imply changes at NUTS2 level (see Fig. 3a). Solution: codes can be checked by a simple relationship to the correct NUTS name and code pairs.

Problems with values (6 input errors)

Input error-type 3: For Zürich (NUTS3 - CH040), the total population in 2005 (POP_T_2005_2006) has been multiplied by -1 (see Fig. 3b). This value is impossible for this variable and as such, should be easily identified.

Input error-type 4: For Brittany (NUTS2 - FR52), the total population in 2005 (POP_T_2005_2006) has been divided by 10 at the NUTS3 level (all 4 of them, see Fig. 3b). These values are possible, but should be easily identified by a simple subtraction of both population variables (POP_T_2005_2006 minus POP_T_2000_2006) and looking for unusually large (negative) differences. Large (negative) differences could be identified as statistically outlying (which upon further scrutiny would indicate *potential* input errors).

Input error-type 5: For Malta (NUTS3 - MT001), the total GDP in 2005 (GDP_2005_2006) has been incorrectly entered as 9999. 9999 is sometimes used to denote a missing value (see Fig. 3b). This value is possible, but should be identified when all *potential* missing values (e.g. values of -99, -999, -9999, 99, 999, 9999, etc.) are identified for further scrutiny.

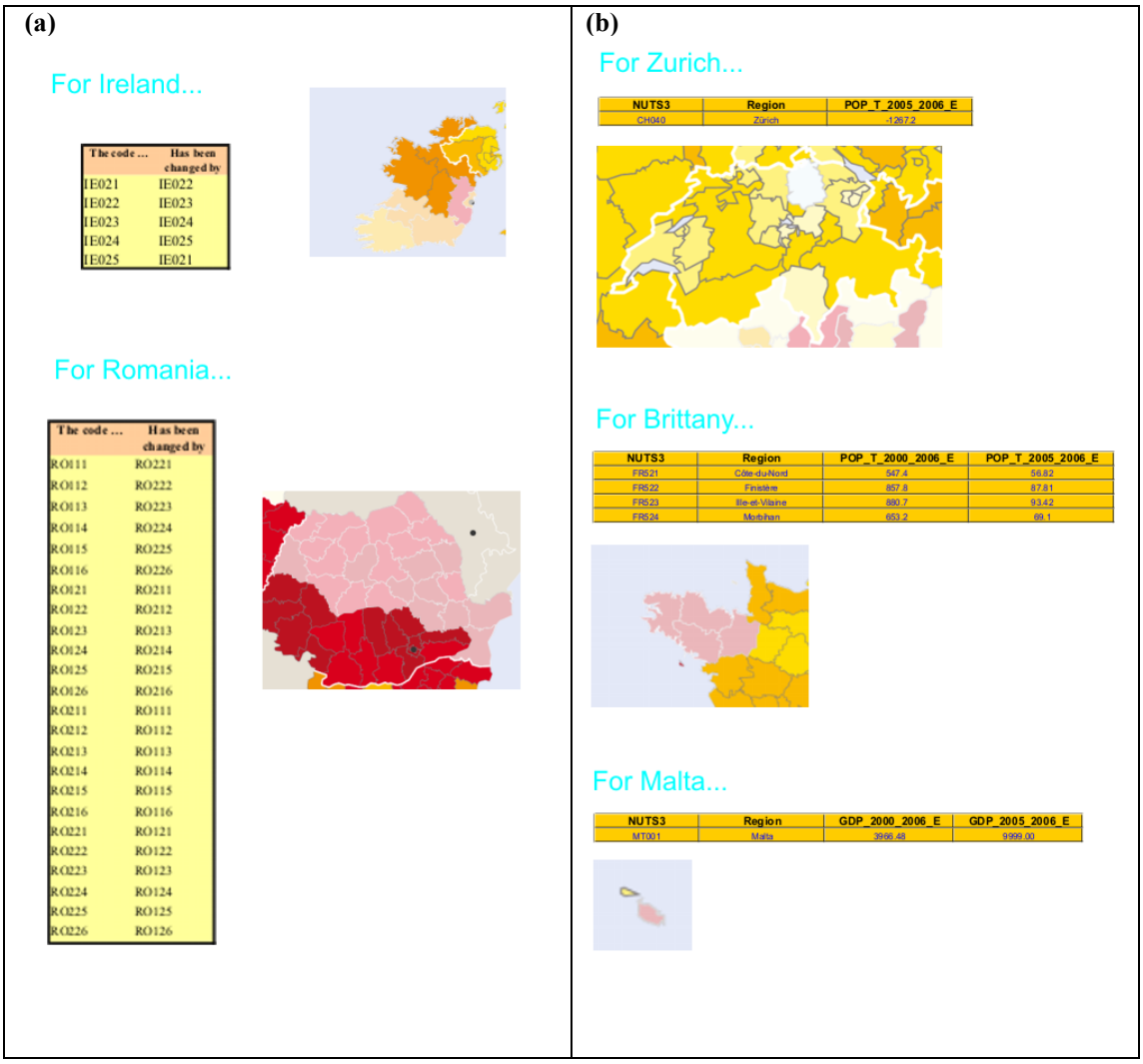


Figure 3 (a): *incorrect NUTS code entries (b) incorrect value entries*

Problems with copied or repeated data (52 input errors)

Input error-type 6: For Belgium (44 entries), the total population in 2000 (POP_T_2000_2006) is repeated exactly for the total population in 2005 (POP_T_2005_2006) (all at NUTS3 level, see Fig. 4). These values are possible, but should be easily identified by a simple subtraction of the two population variables and looking for (exact) zero values. It can be assumed that equal populations for the two years are highly unlikely. Also observe that differences of zero are unlikely to be statistically outlying. A difficulty here would be to decide whether the values for POP_T_2000_2006 or the values for POP_T_2005_2006 were inputted incorrectly. This would require further scrutiny.

Input error-type 6: For Slovakia (8 entries), the total GDP in 2000 (GDP_2000_2006) is repeated exactly for the total GPD in 2005 (GDP_2005_2006) (all at NUTS3 level, see Fig. 4). These values are possible, but again should be easily identified by a simple subtraction of the two GDP variables and looking for zero values. As with the population data, it can be assumed that equal GDP data for the two years is highly unlikely. Again, it is difficult to know whether the values for GDP_2000_2006 or the values for GDP_2005_2006 were inputted incorrectly.

NUTS3	Region	GDP_2000_2006	GDP_2005_2006	POP_T_2000_2006	POP_T_2005_2006
BE100	Arr, de Bruxelles-Capitale/Arr	47934.56178	57922.33354	961.9	961.9
BE211	Arr, Antwerpen	29639.0728	35421.28545	931.8	931.8
BE212	Arr, Mechelen	7394.05749	9666.820472	306	306
BE213	Arr, Turnhout	10297.66717	12190.7848	406.9	406.9
↓					
BE345	Arr, Virton	788.351089	772.513648	48.5	48.5
BE351	Arr, Dinant	1421.734879	1590.899159	100	100
BE352	Arr, Namur	5400.430206	6678.291105	283.3	283.3
BE353	Arr, Philippeville	732.603405	865.381684	61.6	61.6
SK010	Bratislavský kraj	5466.133514	5466.133514	617.2	602.4
SK021	Trnavský kraj	2366.623155	2366.623155	551.3	553.7
SK022	Trencianský kraj	2344.433595	2344.433595	609.1	600.9
SK023	Nitrianský kraj	2533.37388	2533.37388	715.3	708.9
SK031	Zilinský kraj	2316.321555	2316.321555	693.4	694.4
SK032	Banskobystrický kraj	2246.515253	2246.515253	662.3	657.7
SK041	Presovský kraj	1953.076092	1953.076092	786	797.7
SK042	Kosický kraj	2808.992955	2808.992955	766.2	771.2

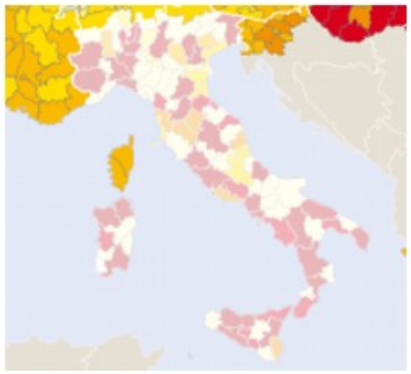
Figure 4*: Problems of copied or repeated data*

Shift in data values (up one or down one line in its data column) (107 input errors)

Input error-type 7: For Italy (107 entries), the total population in 2005 (POP_T_2005_2006) has been shifted up by one line (all at NUTS3 level, see Fig. 5). These values are possible, but most values (not all) should be statistically identified as *potential* input errors. Again a subtraction of the two population variables should for most cases, produce unusually large positive or unusually large negative values which should give rise to suspicion.

Observe that this error-type has created an extra missing value (NUTS3 - ITG2C) and one value has effectively been lost (NUTS3 - ITC11).

For Italy...

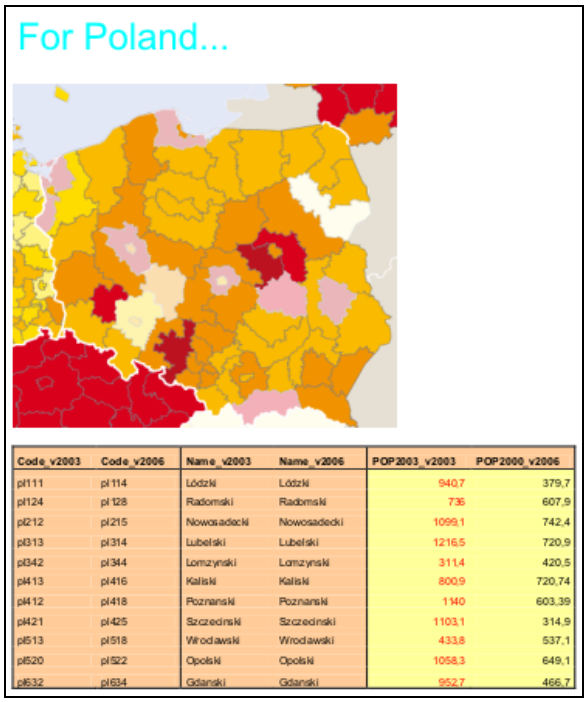


NUTS3	Region	GDP_2000_2006	GDP_2005_2006	POP_T_2000_2006	POP_T_2005_2006
ITC11	Torino	52029.49531	61725.00907	2172.6	177.2
ITC12	Vercelli	3730.69389	4584.692785	177.7	187.9
ITC13	Biella	4385.872603	4731.489232	187.7	161.6
ITC14	Verbano-Cusio-Oss	3046.651604	3481.234948	159.3	354.5
ITC15	Novara	8224.855319	9667.118664	341.1	570.9
ITC16	Cuneo	12881.37037	16439.24116	554.7	213.8
ITC17	Asti	4136.673486	4924.755622	208.2	430.2
ITC18	Alessandria	9045.646615	10968.53334	421	123.4
↓					
ITG25	Sassari	4965.776727	5752.202514	323.6	162.7
ITG26	Nuoro	2363.182555	2779.217366	165.1	551.9
ITG27	Cagliari	10414.23012	12089.44068	542.9	168.6
ITG28	Oristano	2079.185795	2726.430351	169	144.7
ITG29	Olbia-Tempio	2580.152981	3355.462608	137.6	58.1
ITG2A	Ogliastra	684.8490646	902.704739	58.7	104.2
ITG2B	Medio Campidano	1086.953999	1282.341192	106.5	131.6
ITG2C	Carbonia-Iglesias	1666.656429	1759.374031	133.3	

Figure 5: Problems of shifted data

Problems with NUTS codes and names (11 input errors)

Input error-type 8: For some regions of Poland (11 entries), the total population in 2000 (POP_T_2000_2006) has been estimated by the total population in 2003 using NUTS2003 divisions (i.e. POP_T_2003_2003). Such estimations are fine provided the NUTS3 codes are used to relate the regions and not the region names. In this case, the region names have been erroneously used (see Fig. 6). Here the region names have not changed but the geometries for the regions have (and thus the sensible use of different NUTS codes for such instances). The resultant (erroneous) population values are all possible, and in this case, may not be easily identified. They may be identified by a subtraction of POP_T_2005_2006 from POP_T_2000_2006 provided the subtraction happens to result in large outlying differences.



Code_v2003	Code_v2006	Name_v2003	Name_v2006	POP2003_v2003	POP2000_v2006
pl111	pl114	Łódzki	Łódzki	940,7	379,7
pl124	pl128	Radomski	Radomski	736	607,9
pl212	pl215	Nowosadecki	Nowosadecki	1099,1	742,4
pl313	pl314	Lubelski	Lubelski	1216,5	720,9
pl342	pl344	Lomzynski	Lomzynski	311,4	420,5
pl413	pl416	Kaliski	Kaliski	800,9	720,74
pl412	pl418	Poznanski	Poznanski	1140	603,39
pl421	pl425	Szczecinski	Szczecinski	1103,1	314,9
pl513	pl518	Wrodawski	Wrodawski	433,8	537,1
pl520	pl522	Opolski	Opolski	1058,3	649,1
pl632	pl634	Gdanski	Gdanski	952,7	466,7

***Figure 6**: Problems with NUTS codes and names.*

3 Worked examples: commented R scripts and results

3.1 The R statistical environment

All worked examples are coded in the R statistical computing environment (Ihaka and Gentleman 1996), which is open source[8]. In particular we use version 2.9.0 of the base system. For each worked example, only contributed packages are used (i.e. can be downloaded from the R website) except for a useful R mapping package, GISTools (Brunsdon pers. comm.), which is currently under development and will be made available on the R website shortly. The (unsupported) version of GISTools used here (version 0.5-4) is posted on ESPON 2013 database extranet, together with all other relevant materials that are needed to repeat each worked example. The GISTools package is not essential for outlier detection and maps could be constructed using other R packages or outside of R in a GIS.

3.2 Worked example 1: univariate & residual analyses for input errors & outliers

The R script for worked example 1 is given in Appendix 1. The results are summarised in Fig. 7, where the rates of false negatives, false positives and overall misclassification with respect to input error identification were found to be 13.2%, 2.0% and 3.7% respectively. These rates are promisingly low, where their existence, is in part, a reflection of the automated nature of the identification procedures undertaken. Rates should tend to zero upon further (manual) scrutiny of the input errors that only have the *potential* to be so. For example, it would be expected that the rate of false negatives would reduce upon a manual scrutiny of the data in Italy, where the shift in data values (input error-type 7) should be quickly identified.

Observe also that there are many instances of false positives in Spain. This may reflect (unknown) input errors that were already present in this data set (i.e. before our deliberate introduction of input errors) or may reflect true (but unusual) data values (i.e. outliers). Either way, the corresponding data should be scrutinised and checked.

[8] The R website is http://www.r-project.org/

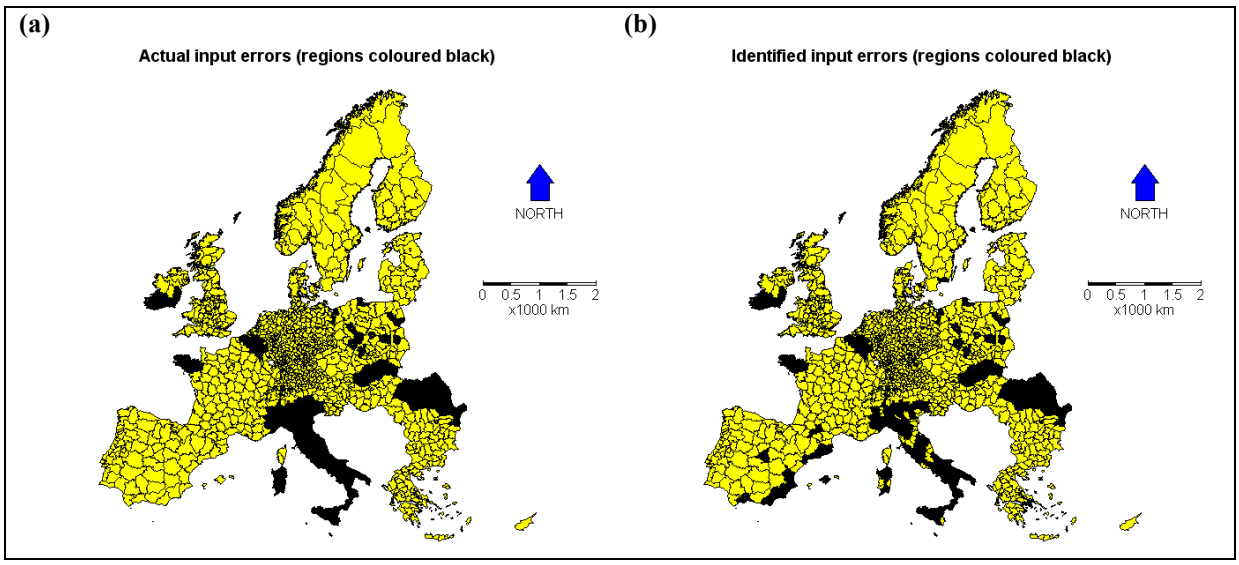


***Figure 7**: Location of (a) true (deliberate) input errors versus (b) identified input errors. Rates of false negatives, false positives and overall misclassification are 13.2%, 2.0% and 3.7% respectively*

3.3 Worked example 2: univariate & residual analyses for outliers

The R script for worked example 2 is given in Appendix 2. The results are summarised in Fig. 8, where the spatial distribution of EVOGDP_2000_2005_2006 is compared with the spatial distribution of (suspected) outliers for this variable. In total, seven indicators were used to gauge whether or not an observation is outlying: (1) standard boxplot statistics; (2) adjusted boxplot statistics; (3) Hawkins' test for spatial outliers; and (4 to 7) large (absolute and raw) residuals from LM/MLR/LR/GWR fits (each calibrated with the coordinate data). These indicators are summarised in Fig. 8b, where a strong case for an outlier relates to an observation that has positive results for all seven outlier identification analyses. It appears that the most outlying EVOGDP_2000_2005_2006 observations lie in the south-east of the ESPON region.

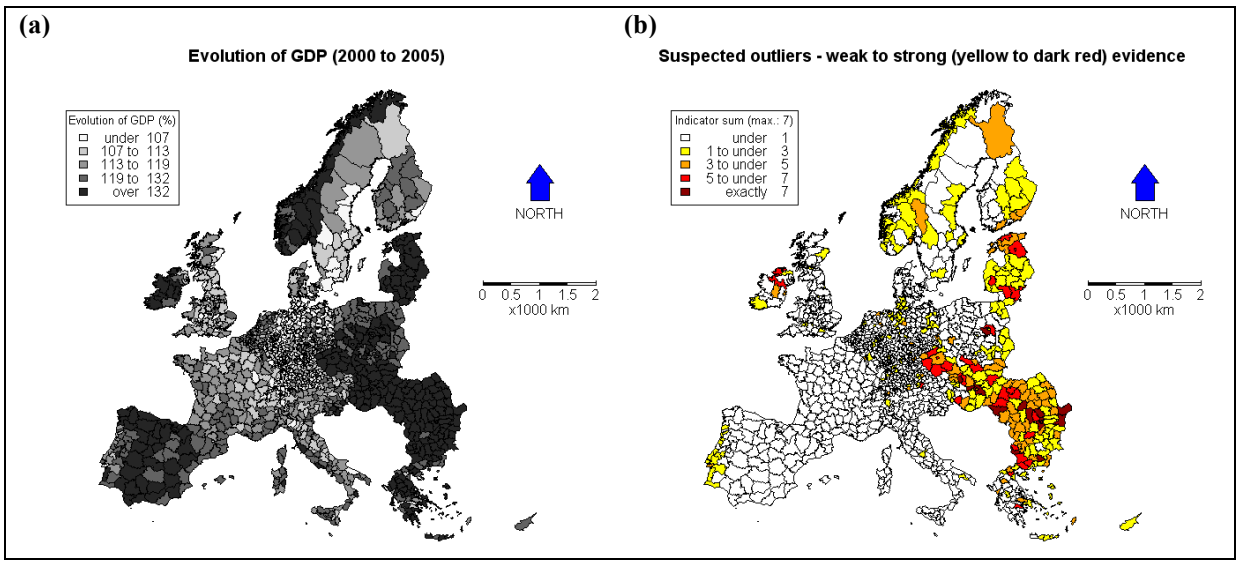


***Figure 8**: Spatial distribution of (a) EVOGDP_2000_2005_2006 and (b) suspected outliers for EVOGDP_2000_2005_2006 (seven univariate indicators)*

3.4 Worked example 3: multivariate analyses for outliers

The R script for worked example 3 is given in Appendix 3. The results are summarised in Fig. 9, where only a much reduced data set of 731 regions could be used in this set of analyses (a consequence of a considerable amount of missing data). In Fig. 9, potential outliers are found according: (a) a bagplot of EVOGDP_2000_2005_2006 with UNEMP_R_2001_1999; (b) the technique, RMD2-AQ-outlier; (c) the technique, PCA-outlier-1; and (d) the technique, PCA-outlier-2. The bagplot identifies outliers in a bivariate-sense whilst the other three techniques identify outliers in a multivariate-sense (in this case, with respect to outlying or unusual relationships amongst EVOGDP_2000_2005_2006, UNEMP_R_2001_1999 NAT_HAZ_2004_1999 and SF_CF_1999_1999). Clearly, the bivariate approach is missing key information. The three multivariate approaches give broadly similar results, but where the relatively large number of potential outliers (as many as 201 observations using RMD2-AQ-outlier) suggests multiple (statistical) populations rather than one population with many outlying observations. This would require further investigation.

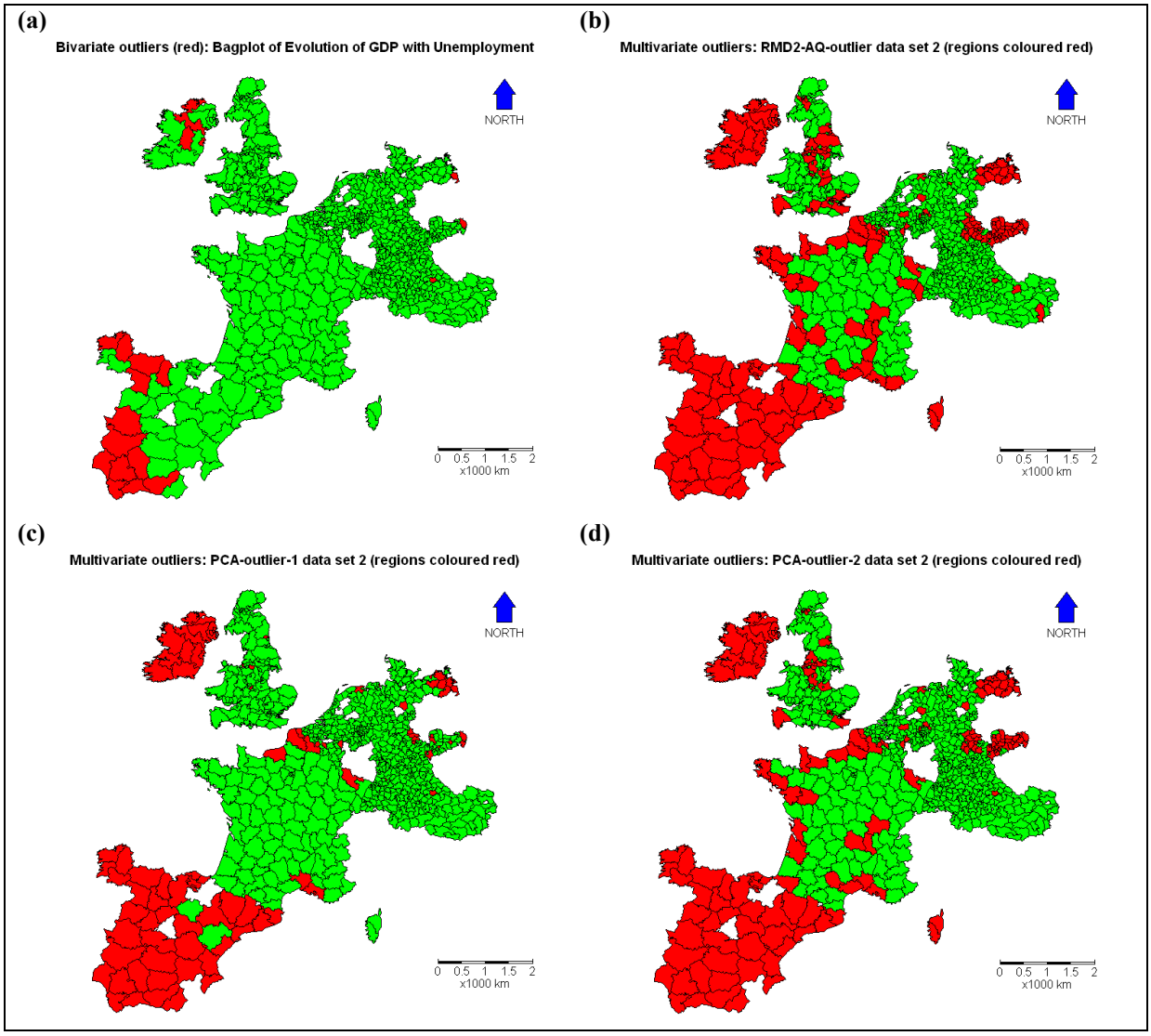


***Figure 9**: Spatial distribution of suspected (bivariate and multivariate) outliers according to (a) a bagplot of EVOGDP_2000_2005_2006 with UNEMP_R_2001_1999; (b) the RMD2-AQ-outlier technique; (c) the PCA-outlier-1 technique; and (d) the PCA-outlier-2 technique*

3.5 Worked example 4: multivariate residual analyses for outliers

The R script for worked example 4 is given in Appendix 4. The results are summarised in Fig. 10, where the spatial distribution of EVOGDP_2000_2005_2006 is compared with the spatial distribution of (suspected) outliers for this variable. Again, only a much reduced data set of 731 regions could be used for this multivariate analysis. In total, three indicators were used to gauge whether or not an observation is outlying: (1) large (absolute and raw) residuals from an MLR fit; (2) large (absolute and raw) residuals from an LR fit; and (3) large (absolute and raw) residuals from a GWR fit. All three models were calibrated using the coordinates, SF_CF_1999_1999 and SPAT_TYPE_2_1999_1999 as independent contextual data. These indicators are summarised in Fig. 10b, where a strong case for an outlier relates to an observation that has positive results for all three outlier identification analyses.

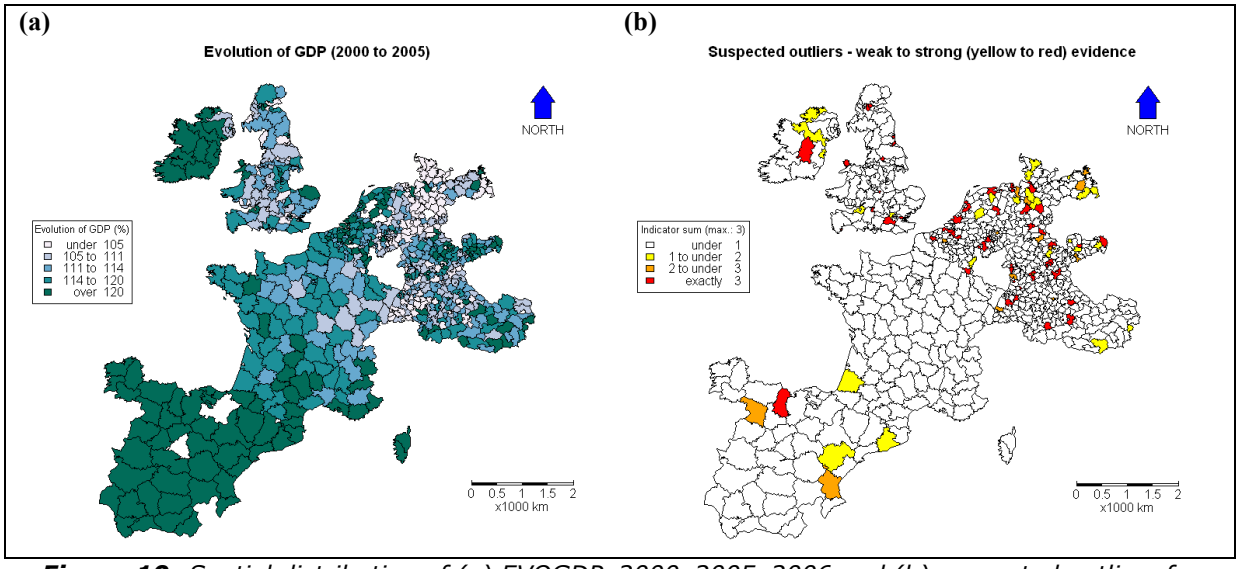


***Figure 10**: Spatial distribution of (a) EVOGDP_2000_2005_2006 and (b) suspected outliers for EVOGDP_2000_2005_2006 (three multivariate indicators).*

3.6 Worked example 5: identification of spatial clusters

The R script for worked example 5 is given in Appendix 5. The results are summarised in Fig. 11, where the aim is to identify 'unusual' clusters in EVOGDP_2000_2005_2006 with respect to (a) its local variability (using GW standard deviations); (b) its local relationship to SF_CF_1999_1999 (via a GW correlation analysis); (c) its local relationship to class 2 of SPAT_TYPE_2_1999_1999 (via a GWR analysis); and (d) its local spatial autocorrelation (via local Moran's I statistic). Again, only a much reduced data set of 731 regions could be used for this combined

univariate and multivariate analysis. Observe that the shown local relationships for EVOGDP_2000_2005_2006 are examples, as different local relationship can be investigated.

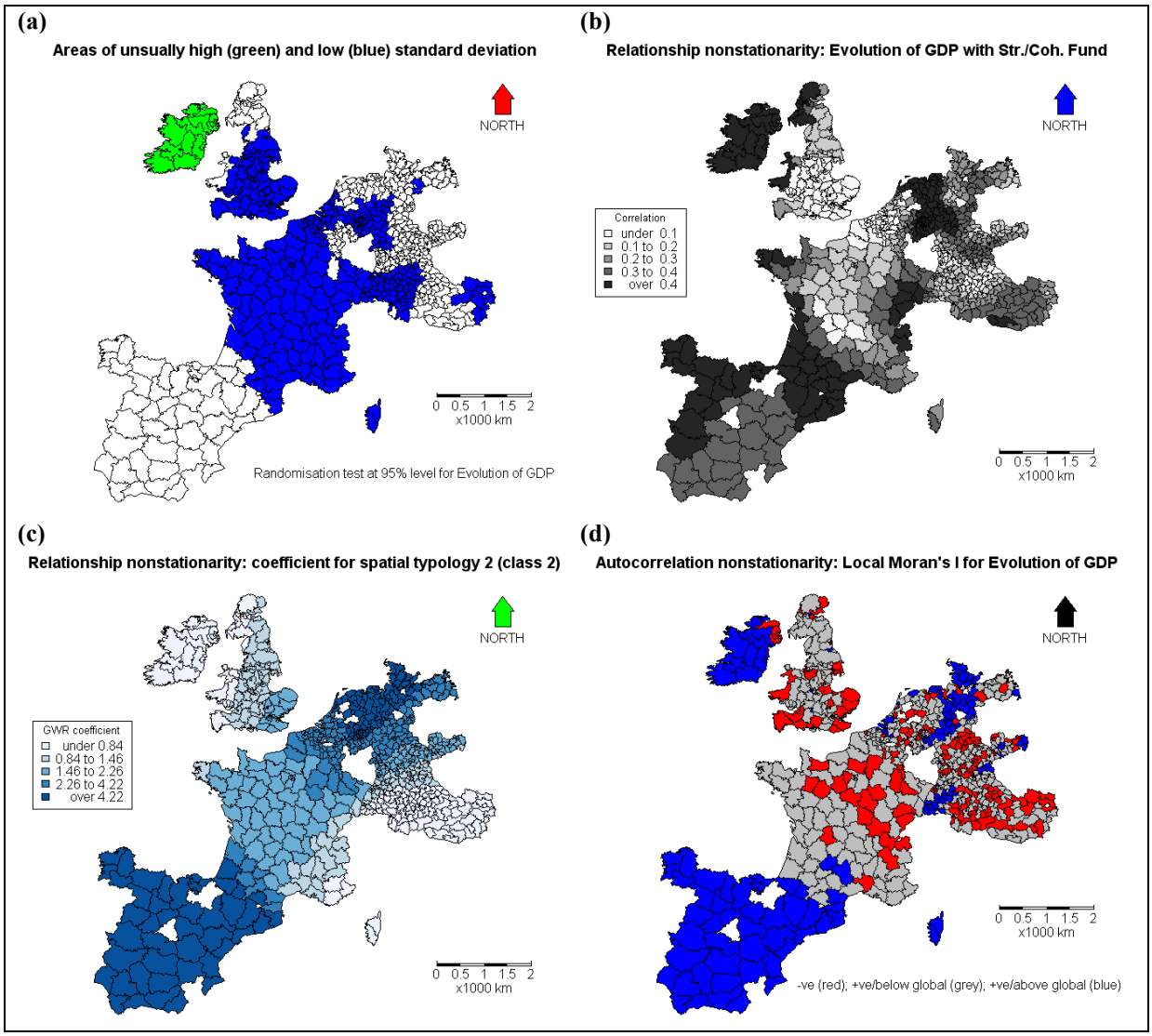


Figure 11*: Identification of 'unusual' clusters in EVOGDP_2000_2005_2006 with respect to (a) its local variability (using GW standard deviations); (b) its local relationship to SF_CF_1999_1999 (via a GW correlation analysis); (c) its local relationship to class 2 of SPAT_TYPE_2_1999_1999 (via a GWR analysis); and (d) its local spatial autocorrelation (via local Moran's I statistic)*

Briefly and focusing on EVOGDP_2000_2005_2006 for Ireland and Northern Ireland only; Fig. 11a indicates that these regions tend to have unusually high levels of variation in EVOGDP_2000_2005_2006; Fig. 11b suggests that these regions tend to have an unusually strong relationship between EVOGDP_2000_2005_2006 and SF_CF_1999_1999; Fig. 11c suggests that these regions tend to have an unusually weak relationship between EVOGDP_2000_2005_2006 and class 2 of SPAT_TYPE_2_1999_1999; and Fig. 11d suggests that some regions of Northern Ireland can have an unusual negative spatial autocorrelation for EVOGDP_2000_2005_2006 (i.e. neighbouring values of EVOGDP_2000_2005_2006 tend to be dissimilar).

3.7 Worked example 6: some consequences of MAUP

The R script for worked example 6 is given in Appendix 6. The results are summarised in Figs. 12 and 13, where the spatial distribution of EVOGDP_2000_2005_2006 and the spatial distribution of (suspected) outliers for EVOGDP_2000_2005_2006 are shown at four different NUTS levels (i.e. 3, 2, 1 and 0), respectively. At each NUTS level, the same seven indicators are used to gauge whether or not an observation is outlying (as in worked example 2).

Scatterplots and correlations are given in Fig. 14, where the "Strongest indication of an outlier for any constituent NUTS level 3 region" is related to the "Indication of an outlier in a corresponding aggregated NUTS level 2/1/0 region". If the effects of MAUP on outlier identification are minimal, then a strong relationship (and correlation) would be expected.

From Figs. 12 to 14, we can observe that:

- A NUTS level 3 region that is an outlier does not imply that the NUTS level 2/1/0 region that contains it will also be an outlier.
- Several adjacent NUTS level 3 regions that are outliers, which belong to two or more adjacent NUTS level 2 regions, do not imply that those NUTS level 2 regions will be outliers (and so forth down the NUTS levels).
- A NUTS level 0/1/2 region that is an outlier is likely to contain one or more NUTS level 3 regions that are outliers.
- Evidence of an outlier weakens as the level of spatial aggregation increases.

In summary, it is recommended that outliers should be identified at the smallest spatial scale.

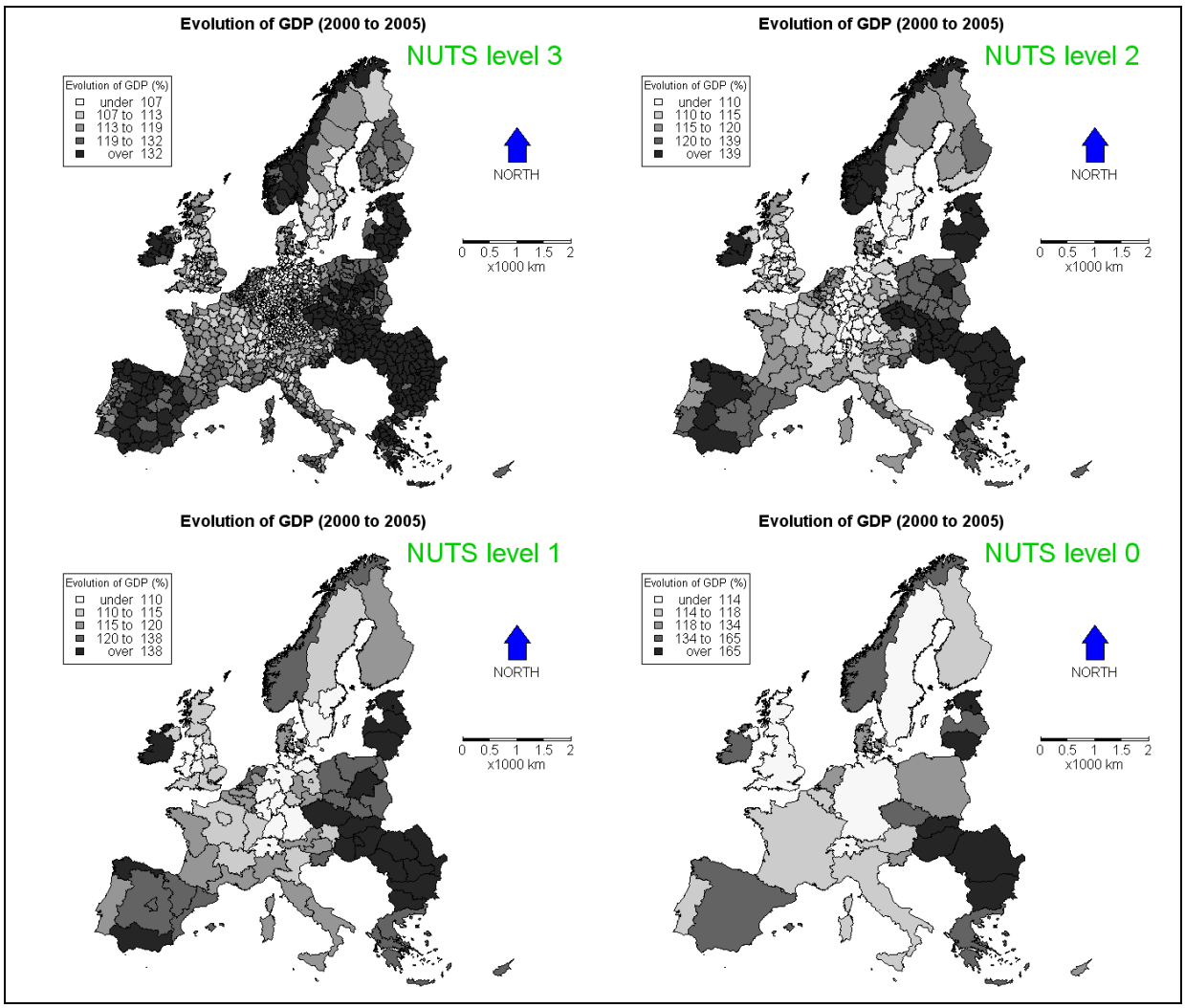


***Figure 12**: Spatial distribution of EVOGDP_2000_2005_2006 at four different NUTS levels (3, 2, 1 and 0)*

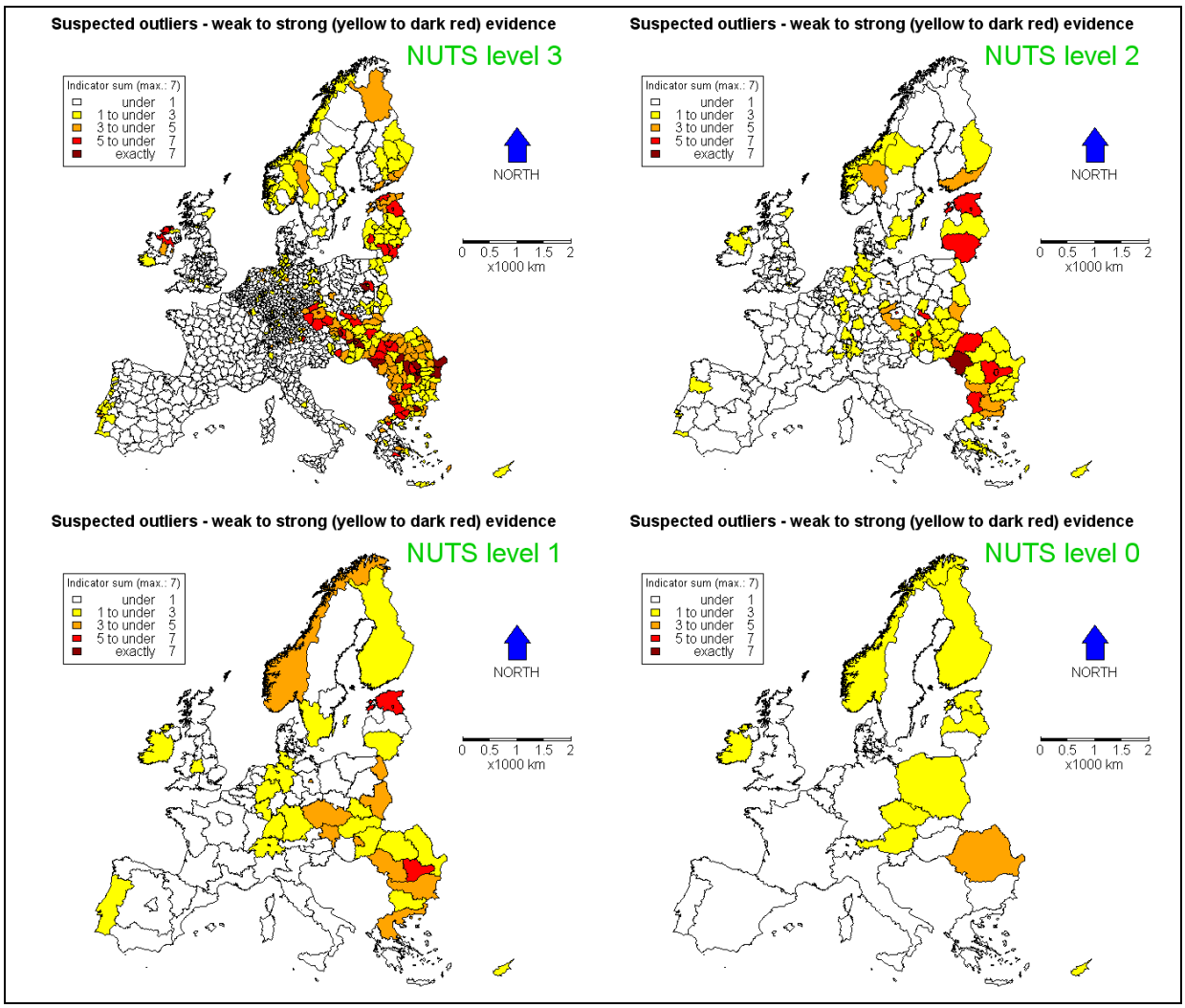


***Figure 13**: Spatial distribution of suspected outliers for EVOGDP_2000_2005_2006 (via seven univariate indicators), where outliers are identified at four NUTS levels (3, 2, 1 and 0)*

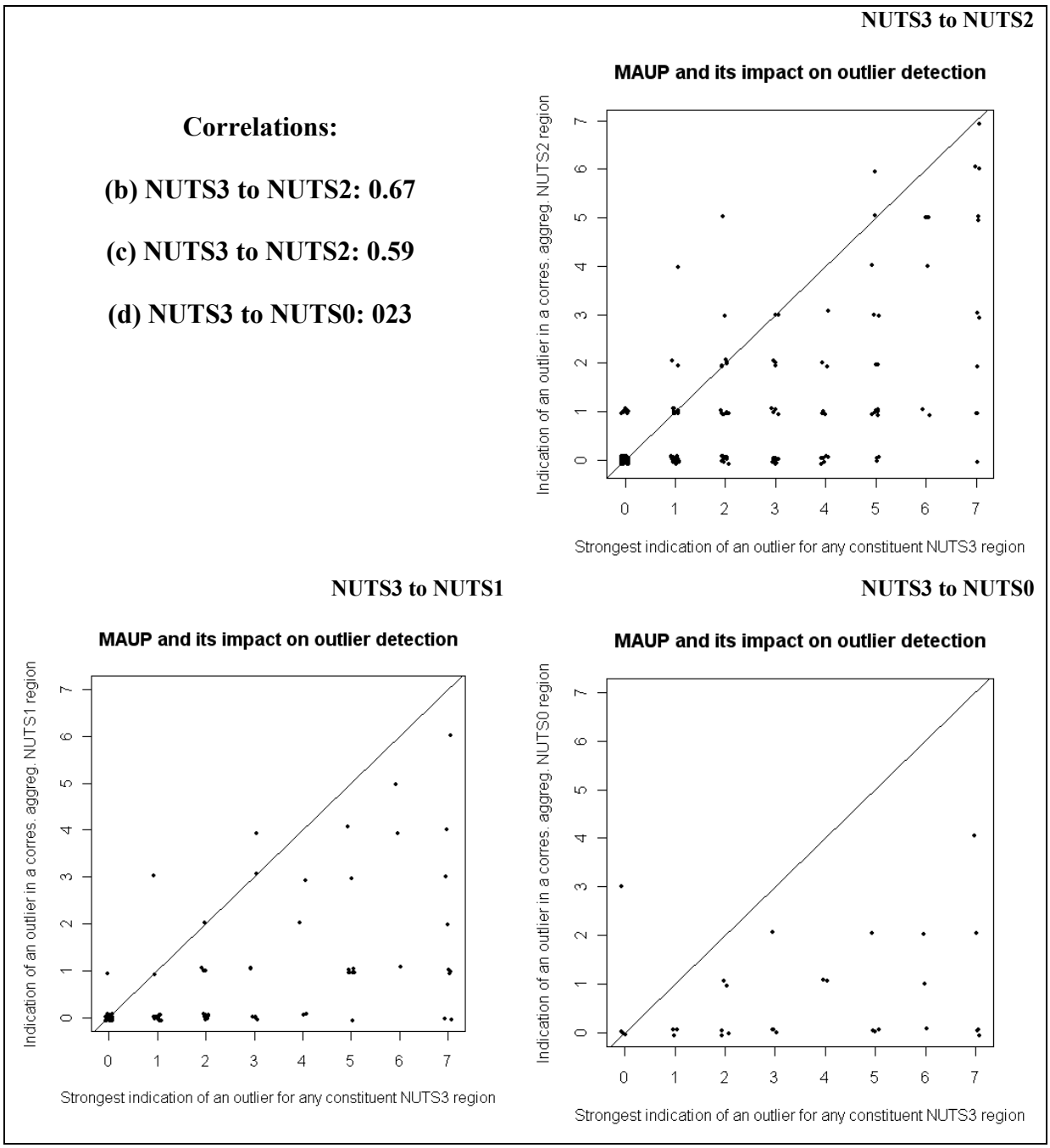


***Figure 14**: Scatterplots and correlations, where the "Strongest indication of an outlier for any constituent NUTS level 3 region" is related to the "Indication of an outlier in a corresponding aggregated NUTS level 2/1/0 region". Scatterplots are jittered to aide interpretation*

4 Discussion and further developments

This technical report provides an introduction to the detection of logical input errors and statistical outliers (i.e. exceptional values) for data sets of the ESPON 2013 Database. Some important aspatial and spatial techniques have been introduced and demonstrated within the R statistical computing environment.

The field of robust statistics and outlier detection is extremely large and diverse, and as such can not be comprehensively reviewed within the terms of reference of this report. However, outlier detection techniques applicable (or designed for) *spatial* data sets are not as developed as those for *aspatial* applications.

In this respect, our current research is focused on this specific area of model development. Here robust versions of geographically weighted summary statistics (GWSS), geographically weighted regression (GWR) and geographically weighted principal component analysis (GWPCA) are to the fore, as they allow the detection of outliers in both univariate and multivariate spatial data sets.

Our expected deliveries for the final report of this phase of the ESPON project will be firmly based on the analytical techniques described and applied here. However we will now hone these procedures using a concrete, real-life data set rather than the fabricated data set used here. This new data set will no doubt present some new analytical challenges that have not been considered. This should enhance the detection methodology, which may need to include the addition of further techniques.

For the final report, we also aim to introduce a selection of the robust geographically weighted techniques that we are currently working on. An improved version of Hawkins' spatial outlier test is also under development, as is a robust version of the local Moran's I statistic (with respect to outlier identification). Here it is envisaged that our relatively advanced robust spatial methods should not be fully presented in the final report of this first phase of the ESPON project, but instead left for the next phase of the ESPON project (i.e. for the 2011 to 2013 stage), when the development of these robust spatial methods has properly matured. Work in this next phase should also include the packaging of the R code for these robust spatial methods, so that techniques are fully portable, transferable and openly documented.

Acknowledgements

Thanks are due to Ronan Ysebaert and Claude Grasland at UMS RIATE for their work on providing us with the (univariate) test data set and the deliberate introduction of a range of challenging logical input errors to this data set.

References

Ainsworth LM, Dean CB (2008), *Detection of local and global outliers in mapping studies*. Environmetrics 19, 21-37.

Anselin L. (1995) *Local indicators of spatial association*. Geographical Analysis 27, 93 -115.

Béguin C, Hulliger B (2004) *Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations*. Journal of the Royal Statistical Society, Series A 167(2), 275-294.

Brunsdon C, Fotheringham AS, Charlton ME (2002) *Geographically weighted summary statistics - a framework for localised exploratory data analysis*. Computers, Environment and Urban Systems 26, 501-524.

Brunsdon C, Charlton ME (2010) *An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection*. Submitted to Environment and Planning B

Chambers R, Hentges A, Zhao X (2004) *Robust automatic methods for outlier and error detection*. Journal of the Royal Statistical Society, Series A 167(2), 323-339.

Charlton ME, Brunsdon C, Demšar U, Harris P, Fotheringham AS (2010) Principal *component analysis: from global to local*. In preparation.

Charlton S (2004) *Evaluating automatic edit and imputation methods, and the EUREDIT Project*. Journal of the Royal Statistical Society, Series A 167(2), 199-207.

Cruz Ortiz M, Sarabia LA, Herrero A (2006) *Robust regression techniques: A useful alternative for the detection of outlier data in chemical analysis*. Talanta 70, 499-512.

D'Alimonte D, Cornford D (2007) *Outlier detection with partial information: application to emergency mapping*. Stochastic Environmental Research and Risk Assessment 22, 613-620.

Daszykowski M, Kaczmarek K, Vander Heyden Y, Walczak B (2007) *Robust statistics in data analysis – a review Basic concepts*. Chemometrics and Intelligent Laboratory Systems 85, 203-219.

ESPON (2006) 3.4.3 *The modifiable areas unit problem* – Final Report http://www.espon.eu/mmp/online/website/content/projects/261/431/file_4970/

Faraway J (2004) *Linear models with R. Chapman & Hall/CRC*, Boca Raton/FL

Filzmoser P, Garrett R, Reimann C (2005) *Multivariate outlier detection in exploration geochemistry*. Computers & Geosciences 31, 579-587.

Filzmoser P, Maronna R, Werner M (2008) *Outlier identification in high dimensions*. Computational Statistics and Data Analysis 52, 1694-1711.

Fotheringham AS, Brunsdon C, Charlton ME (2002) *Geographically Weighted Regression - the analysis of spatially varying relationships*. Wiley, Chichester.
Frigge M, Hoaglin DC, Iglewicz B (1989) *Some implementations of the Boxplot*. The American Statistician 43, 50–54.

Ghosh-Dastidar B, Schafer JL (2003) *Multiple edit/multiple imputation for multivariate continuous data*. Journal of the American Statistical Association 98(464), 807-817.

Harris P, Brunsdon C (2010) *Exploring spatial variation and spatial relationships in a freshwater acidification critical load data set for Great Britain using geographically weighted summary statistics*. Computers & Geosciences 36, 54-70.

Harris P, Fotheringham AS, Juggins S (2010) *Robust Geographically Weighed Regression: A Technique for Quantifying Spatial Relationships Between Freshwater Acidification Critical Loads and Catchment Attributes*. To appear in the Annals of the Association of American Geographers.

Hawkins RM (1980) *Identification of Outliers*. Chapman & Hall, London.

Hoo KA, Tvarlapati KJ, Piovoso MJ, Hajare R (2002) *A method of robust multivariate outlier replacement*. Computers and Chemical Engineering 26, 17-39.

Hubert M, Vandervieren E (2008) *An adjusted boxplot for skewed distributions*. Computational Statistics and Data Analysis 52, 5186-5201.

Ihaka R, Gentleman R (1996) *R: A language for data analysis and graphics*. Journal of Computational and Graphical Statistics 5, 299-314.

Jackson DA, Chen Y (2004) *Robust principal component analysis and outlier detection with ecological data*. Environmetrics 15, 129-139.

Kou Y, Lu C-T, Chen D (2006) *Spatial Weighted Outlier Detection*. In proceedings of the 2006 SIAM International Conference on Data Mining No. 614 2006.

Liu H, Jezek K, O'Kelly M (2001) *Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS*. International Journal of Geographical Information Science 15(8), 721-741.

Loader C (2004) *Smoothing: Local Regression Techniques*. In Gentle J, Härdle W, Mori Y (eds) Handbook of Computational Statistics. Springer-Verlag, Heidelberg.

Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K (1999) *Robust principal components for functional data*. Test 8, 1–73.

Meklit T, Van Meirvenne M, Verstraete S, Bonroy J, Tack F (2009) *Combining marginal and spatial outliers identification to optimize the mapping of the regional geochemical baseline concentration of soil heavy metals*. Geoderma 148, 413-420.

Morgenthaler S (2007) *A survey of robust statistics*. Statistical Methods & Applications 15, 271-293.

Petrakos G, Conversano C, Farmakis G, Mola F, Siciliano R, Stavropoulos P (2004) *New ways of specifying data edits*. Journal of the Royal Statistical Society, Series A 167(2), 249-274.

Plaia A, Bondi A (2006) *Single imputation method of missing values in environmental pollution data sets*. Atmospheric Environment 40, 7316-7330.

Reimann C, Filzmoser P, Garrett R (2005) *Background and threshold: critical comparison of methods of determination*. Science of the Total Environment 346, 1-16.

Rousseeuw PJ, Ruts I, Tukey JW (1999) *The Bagplot: A Bivariate Boxplot.* The American Statistician 53, 382–387.

Rousseeuw PJ, Debruyne M, Engelen S, Hubert M (2006) *Robust and outlier detection in chemometrics*. Critical Reviews in Analytical Chemistry 36, 221-242.

Vanden Branden K, Verboven S (2009) *Robust data imputation*. Computational Biology and Chemistry 33, 7-13.

Wong D (1996) *Aggregation effects in geo-referenced data*. In Arlinghaus SL (ed) Practical Handbook of Spatial Statistics. CRC Press, Boca Raton, FL.

Appendices

Appendix 1 – R script for worked example 1

```
# 1. Preamble ########################################################################


# Worked example 1 - for technical report - challenge 10 - ESPON 2013 database
# NCG - P. Harris & M. Charlton
# 7/2/10


# Objective - to identify input errors in:
# "NUTS_2006" (the NUTS3 code)
# "GDP_2000_2006"
# "GDP_2005_2006"
# "POP_T_2000_2006"
# "POP_T_2005_2006"


# Methods: univariate - aspatial
# Mixture of logical & statistical methods
# Statistical methods:
# 1. Standard boxplots only


# R packages needed.....
# 1. GISTools (version 0.5-4) - depends on 2 to 11...
# 2. foreign (version 0.8-30)
# 3. gpclib (version 1.4-3)
# 4. maptools (version 0.7-16)
# 5. Matrix (version 0.999375-18)
# 6. RColorBrewer (version 1.0-2)
# 7. sp (version 0.9-28)
# 8. spam (version 0.15-2)
# 9. spdep (version 0.4-29)
# 10. spgwr (version 0.6-2)
# 11. tripack (version 1.2-11)

# Base R system version 2.9.0
# N.B. Some of the above packages may still depend on other R packages - download these from R website...


# Relevant data files (see data & ArcGIS directories):

# Excel files...
# 1. ESPON_DATA_NCG_CHALLANGE_10_original.xls
# 2. ESPON_DATA_NCG_CHALLANGE_10_subsets.xls
```

```
# Text files...
# 3. Worked example 1 true codes & new ID.txt

# ArcGIS files...
# 4. Worked_example_1a.shp - ArcGIS shapefile of the data...

# Only files 3 and 4 are needed in this worked example...



# The variables - some with deliberate input-errors...

# The following 5 variables are all suspected (i.e. in this case, known) to have input errors...
# "NUTS3_2006_E",
# "GDP_2000_2006_E",
# "GDP_2005_2006_E",
# "POP_T_2000_2006_E",
# "POP_T_2005_2006_E"

# These 3 variables are calculated from above so will be effected by an input error...
# "GDP_POP_2000_2006_E",
# "GDP_POP_2005_2006_E",
# "EVOGDP_2000_2005_2006_E"

# Remaining variables - all known to have no input errors...
# "NUTS3","NUTS23","NUTS2","NUTS1","NUTS0" - different NUTS levels
# "Error_type" - type of input error according to technical report (a number between 1 and 8)
# "New_ID" - relates to the regions name only & is purely numeric
# "Region_2006_E" - name of region (NB this does not have any errors)
# "NUTS3" - repeated (a consequence of an ArcGIS operation)
# "X","Y" - centroids of regions



# NOTE that this example dataset has been reduced to
# 1329 values from an original 1351 values (i.e. 22 values removed).
# See readme in excel files on worked example data.

# Note here that 13 of the 22 values removed relate to regions that are highly
# spatially disjoint from mainland Europe (i.e. parts of Portugal, France, and Spain
# such as the Azores, Canaries etc.). Before inclusion into the analyses, we need
# to decide on an appropriate distance metric for these regions.









# 2. Importing data as a ArcGIS shapefile & using GISTools to do some maps... ##

require(GISTools)
#help(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
data1 <- readShapePoly("Worked_example_1a.shp",
proj4string=CRS("+proj=Lambert_Azimuthal_Equal_Area+datum=D_ETRS_1989+ellps=GCS_ETRS_1989"))
colnames(data1@data)

# Renaming each variable - as they have been truncated in ArcGIS...
```

```
colnames(data1@data) <- c("NUTS3","NUTS23","NUTS2","NUTS1","NUTS0",
"Error_type","New_ID","NUTS3_2006_E","Region_2006_E",
"GDP_2000_2006_E","GDP_2005_2006_E","POP_T_2000_2006_E","POP_T_2005_2006_E",
"GDP_POP_2000_2006_E","GDP_POP_2005_2006_E","EVOGDP_2000_2005_2006_E","NUTS3","X","Y")

# Size of data set and adding an order ID...
n <- length(data1@data[,1])
Order_ID <- seq(1,n)
data1@data <- cbind(data1@data, Order_ID)
attach(data1@data)

# Creating a shading scheme and plotting a choropleth map...
shades.1 = auto.shading(GDP_2000_2006_E,5, cols=brewer.pal(5,'Greens'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,GDP_2000_2006_E,shades.1)
title("GDP_2000_2006_E: with input errors")
choro.legend(1300000,400000,shades.1,fmt="%4.0f",title='GDP',cex=0.8)
map.scale(1800000,-950000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1000000,80000, col="blue")

# Creating a shading scheme and plotting a choropleth map...
shades.2 = auto.shading(GDP_2005_2006_E,5, cols=brewer.pal(5,'Greens'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,GDP_2005_2006_E,shades.2)
title("GDP_2005_2006_E: with input errors")
choro.legend(1200000,400000,shades.2,fmt="%4.0f",title='GDP',cex=0.8)
map.scale(1800000,-950000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1000000,80000, col="blue")

# Creating a shading scheme and plotting a choropleth map...
shades.3 = auto.shading(POP_T_2000_2006_E,5, cols=brewer.pal(5,'Greens'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,POP_T_2000_2006_E,shades.3)
title("POP_T_2000_2006_E: with input errors")
choro.legend(1400000,400000,shades.3,fmt="%4.0f",title='POP.',cex=0.8)
map.scale(1800000,-950000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1000000,80000, col="blue")

# Creating a shading scheme and plotting a choropleth map...
shades.4 = auto.shading(POP_T_2005_2006_E,5, cols=brewer.pal(5,'Greens'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,POP_T_2005_2006_E,shades.4)
title("POP_T_2005_2006_E: with input errors")
choro.legend(1400000,400000,shades.4,fmt="%4.0f",title='POP.',cex=0.8)
map.scale(1800000,-950000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1000000,80000, col="blue")









# 3. Input error-types 1 and 2 - wrong NUTS code ################################

# This is one approach to deal with these input error-types...

# Order our data by the New_ID (ie a numeric ID of the NUTS region name)...
```

```
data2 <- data1@data[order(New_ID),]
attach(data2)

# Read in a data set where NUTS codes and names (again given as new_ID) are known to be correct...
data3 <- read.table("Worked example 1 true codes & new ID.txt", header=T)
colnames(data3)
attach(data3)

# Scan for input errors in the NUTS code...
# i.e. relate the "New_ID and NUTS3_2006" variables in datasets, data2 and data3...
data4 <- cbind(data2[,7],data2[,8],data3)
#fix(data4) # data spreadsheet

# Or better still - automatically identify input errors as follows...
x <- match(data4[,2], data4[,4]) # matches the New_ID values and assigns matches by position in data set
y <- seq(1,n) # sequence of numbers from 1 to the size of data set (same as Order_ID)
z <- y-x # should be a data set of zeros if all NUTS codes are inputted correctly
sort(-1*(abs(z))) # in this case 29 NUTS codes are inputted incorrectly...

# Updating input error information in one file - using our ordered data set...
indicator.1 <-ifelse(z==0, 0, 1)
data1.update.1 <- cbind(data2, indicator.1)
data1.update.1 <- as.data.frame(data1.update.1)
attach(data1.update.1)
#fix(data1.update.1)

# Re-order our data back to its original state...
data1@data <- data1.update.1[order(data1.update.1[,20]),] # note using data1.update.1[,20] not Order_ID
attach(data1@data)

# A choropleth map of input errors ...
shades.5 = shading(c(0,1,2),c("blue","white","red")) # this actually gives: white - no errors & red - errors
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,21],shades.5) # use data1@data[,21] not indicator.1
title("Input error-types 1 & 2 (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Assessing the identification procedure
# Comparing "Error_types 1 & 2" with "indicator.1"
sum(indicator.1)
Assessment <- cbind(data1@data[,6], data1@data[,21])
Assessment.1 <- Assessment[order(-Assessment[,1]),]
#fix(Assessment.1)

# All 29 input errors correctly identified in Ireland & Romania...
# No false positives...

# COMMENT - THIS TYPE OF INPUT-ERROR IS PROBABLY BETTER DETECTED OUTSIDE OF R - i.e.
IN A DATABASE




# 4. Input error-type 3 - impossible values ######################################

# This is one approach to deal with this input error-type...
```

```
# Checks for impossible values (in this case, impossible values for positive continuous data,
POP_T_2005_2006_E)

# POP_T_2005_2006_E in the ordered dataset
imp.val <- data2[,13]

# Explore the data...
summary(imp.val) # summary statistics
sort(imp.val) # ordered data
X11(width=5.3,height=5.7)
boxplot(imp.val, main="Input error-type 3", pch=19, cex=0.5) # boxplot

# Define minimum and maximums
Min_pop <- 0
Max_pop <- 10000 # This upper-limit is chosen by judgement

# Identifying & updating input error information in one file - using our ordered data set...
indicator.2 <-ifelse(Min_pop < imp.val & imp.val < Max_pop, 0, 1)
data1.update.2 <- cbind(data1.update.1,indicator.2)
data1.update.2 <- as.data.frame(data1.update.2)
attach(data1.update.2)
#fix(data1.update.2)

# Again re-order our data back to its original state...
data1@data <- data1.update.2[order(data1.update.2[,20]),]
attach(data1@data)

# A choropleth map of input errors ...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,22],shades.5)
title("Input error-type 3 (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Assessing the identification procedure
# Comparing "Error_type 3" with "indicator.2"
sum(indicator.2)
Assessment <- cbind(data1@data[,6], data1@data[,22])
Assessment.2 <- Assessment[order(-Assessment[,1]),]
#fix(Assessment.2)

# 1 input error correctly identified in Zurich...
# No false positives...









# 5. Input error-type 5 - potential missing value ################################

# This is one approach to deal with this input error-type...

# Investigate all entries of -99, -999, -9999, 99, 999, 9999 as potential missing values...

# In this case do this for GDP_2005_2006_E

# GDP_2005_2006_E in the ordered dataset
miss.val <- data2[,11]
```

```
# Identifying & updating potential input error information in one file - using our ordered data set...
indicator.3 <-ifelse(miss.val!=abs(99) & miss.val!=abs(999) & miss.val!=abs(9999), 0, 1)
data1.update.3 <- cbind(data1.update.2,indicator.3)
data1.update.3 <- as.data.frame(data1.update.3)
attach(data1.update.3)
#fix(data1.update.3)

# Again re-order our data back to its original state...
data1@data <- data1.update.3[order(data1.update.3[,20]),]
attach(data1@data)

# A choropleth map of potential input errors...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,23],shades.5)
title("Potential input error-type 5 (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Assessing the identification procedure
# Comparing "Error_type 5" with "indicator.3"
sum(indicator.3)
Assessment <- cbind(data1@data[,6], data1@data[,23])
Assessment.3 <- Assessment[order(-Assessment[,1]),]
#fix(Assessment.3)

# 1 input error correctly identified in Malta...
# No false positives...




# 6. Input error-type 4,6,7 and 8 - all (relatively) unexpected values #########

# This is one approach to deal with these input error-types...

# Checking POP_T_2000_2006_E with POP_T_2005_2006_E for unusual data...
# This time four input error-types (4, 6, 7 and 8) can be investigated together...
# From section 3, impossible input error-types have already been identified for POP_T_2005_2006_E
# (i.e. we do not need to account for this error-type)
# but further input error-types can be identified if we relate/compare POP_T_2000_2006_E with
POP_T_2005_2006_E

# Intuitively, these data pairs should be broadly similar (but not exactly the same, i.e. error-type 6)
# Interest lies in the data pairs that are very different (i.e. differences are statistically outlying)
# or are identical (i.e. error-type 6 - copied or repeated data)...

# Again naming the relevant variables in the ordered dataset
x1 <- data2[,12] # POP_T_2000_2006_E
y1 <- data2[,13] # POP_T_2005_2006_E

# Exploring the data with a scatterplot (data should broadly lie on the 45 degree line)...
X11(width=5.3,height=5.7)
plot(x1,y1, main="Potential input error-types 4,6,7 or 8", pch=19, cex=0.5) # scatterplot
abline(0,1) # the 45 degree line

# Difference data...
# POP_T_2005_2006_E minus POP_T_2000_2006_E
z1 <- (y1-x1) # actual differences
```

```
#z1 <- abs(y1-x1) # absolute differences

# Exploring the difference data...
summary(z1) # summary statistics
sort(z1) # ordered data
X11(width=5.3,height=5.7)
hist(z1, main="Potential input error-types 4,6,7 or 8") # histogram
X11(width=5.3,height=5.7)
boxplot(z1, main="Potential input error-types 4,6,7 or 8", pch=19, cex=0.5) # boxplot

# Boxplot statistics...
# Change 'coef' accordingly...
# Default 'coef' is 1.5...
# The higher the 'coef' value the stricter the limits/cut-offs & vice versa...
bp <- boxplot.stats(z1, coef=6)
bp$stats
bp$stats[1] # the lower limit/cut-off - i.e. differences below are deemed outlying...
bp$stats[5] # the upper limit/cut-off - i.e. differences above are deemed outlying...
bp$conf
sort(bp$out)
length(bp$out) # number of potential outliers/errors....
# help(boxplot.stats) # for boxplot details...

# Identifying & updating potential input error information in one file - using our ordered data set...
indicator.4 <-ifelse(z1!=0 & z1>bp$stats[1]& z1<bp$stats[5], 0, 1) # i.e. identical or outlying differences...
data1.update.4 <- cbind(data1.update.3,indicator.4,z1) # note - including the difference data
data1.update.4 <- as.data.frame(data1.update.4)
attach(data1.update.4)
#fix(data1.update.4)

# Again re-order our data back to its original state...
data1@data <- data1.update.4[order(data1.update.4[,20]),]
attach(data1@data)

# A choropleth map of potential input errors...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,24],shades.5)
title("Potential input error-types 4,6,7 or 8 (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Observe that this methodology has also identified the impossible value for POP_T_2005_2006_E
# compare...
# indicator.2
# with
# indicators.4
# comparison.1 <- cbind(indicator.2, indicator.4) # see row 119

# We can now investigate these potential outliers more closely...

# Comparing "Error_types 4, 6, 7 & 8" with "indicator.4" and with the "difference data"
sum(indicator.4)
Assessment <- cbind(data1@data[,6], data1@data[,24], data1@data[,25])
Assessment.4 <- Assessment[order(-Assessment[,1]),]
#fix(Assessment.4)

# Results...

# Input error-type 3 - 1 out of 1 input error is correctly re-identified in Zurich...
# Input error-type 4 - 4 out of 4 input errors are correctly identified in Brittany...
# Input error-type 6 - 44 out of 44 input errors are correctly identified in Belgium...
# Input error-type 7 - 81 out of 107 input errors are correctly identified in Italy...
```

```
# i.e. 26 False negatives
# Input error-type 8 - 10 out of 11 input errors are correctly identified in Poland...
# i.e. 1 False negative

# False positives...
# For input error-types 4, 7 or 8 -
# 16 out of 1162
# i.e. 16 unusally large increases/decrceases in population are actually true...

# False positives...
# For input error-type 6 -
# 6 out of 1162
# i.e. the population remained exactly the same in 6 regions...









# 7. Input error-type 6 only - repeated or copied data ##########################

# This is for GDP_2000_2006_E with GDP_2005_2006_E - but only for repeated data
# These data pairs should be exactly the same

# Again using the relevant variables in the ordered dataset
x2 <- data2[,10] #GDP_2000_2006_E
y2 <- data2[,11] #GDP_2005_2006_E

# Difference data...
z2 <- abs(y2-x2) # absolute differences
sort(z2) # ordered absolute data

# Identifying & updating potential input error information in one file - using our ordered data set...
indicator.5 <-ifelse(z2!=0, 0, 1) # i.e. identical differences...
data1.update.5 <- cbind(data1.update.4,indicator.5,z2) # note - including the difference data
data1.update.5 <- as.data.frame(data1.update.5)
attach(data1.update.5)
#fix(data1.update.5)

# Again re-order our data back to its original state...
data1@data <- data1.update.5[order(data1.update.5[,20]),]
attach(data1@data)

# A choropleth map of potential input errors...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,26],shades.5)
title("Potential input error-type 6 (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Assessing the identification procedure
# Comparing "Error_type 6" with "indicator.5" and with the "difference data"
sum(indicator.5)
Assessment <- cbind(data1@data[,6], data1@data[,26], data1@data[,27])
Assessment.5 <- Assessment[order(-Assessment[,1]),]
#fix(Assessment.5)

# 8 out of 8 input errors correctly identified in Slovakia...
# No false positives...
```

```
# 8.  All input error-types together ##############################################

# Put all indicator data together...
indicator.6 <- indicator.1+indicator.2+indicator.3+indicator.4+indicator.5

data1.update.6 <- cbind(data1.update.5,indicator.6)
data1.update.6 <- as.data.frame(data1.update.6)
attach(data1.update.6)
#fix(data1.update.6)

# Again re-order our data back to its original state...
data1@data <- data1.update.6[order(data1.update.6[,20]),]
attach(data1@data)

# A choropleth map of all identified input errors...
shades.6 = shading(c(0,1,3),c("blue","yellow","black")) # yellow - no errors & black - errors
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,28],shades.6)
title("Identified input errors (regions coloured black)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# A choropleth map of actual input errors...
shades.6 = shading(c(0,1,9),c("blue","yellow","black")) # yellow - no errors & black - errors
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,6],shades.6)
title("Actual input errors (regions coloured black)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# See Figure 7 in Technical report for above maps...

# Missclassification rates...

n # Data set size
tn.ie <- 205 # Total number of deliberate (known) input errors

tn.f <- 27 # Total number of false negatives
tn.p <- 22 # Total number of false positives

# Rate of false negatives
(tn.f/tn.ie)*100

# Rate of false positives
(tn.p/(n-tn.ie))*100

# Overall missclassification rate
((tn.f+tn.p)/n)*100
```

Appendix 2 – R script for worked example 2

```
# 1. Preamble ###########################################################################


# Worked example 2 - for technical report - challenge 10 - ESPON 2013 database
# NCG - P. Harris & M. Charlton
# 8/2/10


# Objective - to identify statistical outliers in:
# "EVOGDP_2000_2005_2006"


# Methods: univariate - aspatial & spatial
# Only statistical methods:
# 1. Standard and Adjusted boxplots,
# 2. Hawkins' test (includes the use of GWSS -
#            geographically weighted summary statistics - GW means and variances),
# 3. LM (local mean, i.e. a GW mean)
# 4. MLR (multiple linear regression),
# 5. LR (local regression) &
# 6. GWR (geographically weighted regression)


# R packages needed.....
# 1. GISTools (version 0.5-4) - depends on 2 to 11...
# 2. foreign (version 0.8-30)
# 3. gpclib (version 1.4-3)
# 4. maptools (version 0.7-16)
# 5. Matrix (version 0.999375-18)
# 6. RColorBrewer (version 1.0-2)
# 7. sp (version 0.9-28)
# 8. spam (version 0.15-2)
# 9. spdep (version 0.4-29)
# 10. spgwr (version 0.6-2) - for GWSS & GWR
# 11. tripack (version 1.2-11)
# 12. moments (version 0.11) - for skewness
# 13. robustbase (version 0.4-5) - for adjusted boxplots
# 14. locfit (version 1.5-4)- for LR

# Base R system version 2.9.0
# N.B. Some of the above packages may still depend on other R packages - download these from R website...


# Relevant data files (see data & ArcGIS directories):

# Excel files...
# 1. ESPON_DATA_NCG_CHALLANGE_10_original.xls
# 2. ESPON_DATA_NCG_CHALLANGE_10_subsets.xls

# ArcGIS files...
# 3. Worked_example_2a.shp - ArcGIS shapefile of the data...
```

```
# The 11 variables...

# "NUTS3","NUTS23","NUTS2","NUTS1","NUTS0" - 5 different NUTS levels
# "New_ID" - relates to the regions name only & is purely numeric
# "NUTS3_2006" - the 2006 NUTS3 version
# "Region_2006" - name of 2006 NUTS3 version
# "X","Y" - centroids of regions
# "EVOGDP_2000_2005_2006" - the variable of interest



# NOTE - this example dataset has been reduced to 1329 values
# from an original 1351 values (i.e. 22 values removed)
# see readme in excel files on worked example data.

# NOTE - this dataset is NOT one corrected for input errors from worked example 1.
# It is just the corresponding dataset without the introduction of deliberate input errors.









# 2. Importing data as a ArcGIS shapefile & using GISTools to do a map... ######

require(GISTools)
#help(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
data1 <- readShapePoly("Worked_example_2a.shp",
proj4string=CRS("+proj=Lambert_Azimuthal_Equal_Area+datum=D_ETRS_1989+ellps=GCS_ETRS_1989"))
colnames(data1@data)

# renaming each variable - as they have been truncated in ArcGIS...
colnames(data1@data) <- c("NUTS3","NUTS23","NUTS2","NUTS1","NUTS0",
"New_ID","NUTS3_2006","Region_2006",
"X","Y","EVOGDP_2000_2005_2006")

# Size of data set and adding an order ID...
n <- length(data1@data[,1])
Order_ID <- seq(1,n)
data1@data <- cbind(data1@data, Order_ID)
attach(data1@data)

# Creating a shading scheme and plotting a choropleth map of EVOGDP_2000_2005_2006...
shades.1 = auto.shading(EVOGDP_2000_2005_2006,5, cols=brewer.pal(5,'Greys'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,EVOGDP_2000_2005_2006,shades.1)
title("Evolution of GDP (2000 to 2005)")
choro.legend(-2400000,2200000,shades.1,fmt="%4.0f",title='Evolution of GDP (%)',cex=0.8)
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")
#text(1500000,2400000, "NUTS level 3", cex=2, col=3)
```

```
# 3. Boxplots ###########################################################################

# Let EVOGDP_2000_2005_2006 be z1...
z1 <- EVOGDP_2000_2005_2006

# Exploring this data...
summary(z1) # summary statistics
sort(z1) # ordered data

# Histogram
X11(width=5.3,height=5.7)
hist(z1, main="Histogram: Evolution of GDP (2000 to 2005)",xlab="Evolution of GDP")

# Standard boxplot with defaults
X11(width=5.3,height=5.7)
boxplot(z1, main="Std. boxplot: Evolution of GDP (2000 to 2005)", pch=19, cex=0.5)

# Standard Boxplot statistics...
# Change 'coef' accordingly...
# Default 'coef' is 1.5...
# The higher the 'coef' value the stricter the limits/cut-offs & vice versa...
bp <- boxplot.stats(z1, coef=1.5)
bp$stats
bp$stats[1] # the lower limit/cut-off - i.e. values below are deemed outlying...
bp$stats[5] # the upper limit/cut-off - i.e. values above are deemed outlying...
bp$conf
sort(bp$out)
length(bp$out) # number of potential outliers...
# help(boxplot.stats) # for details...

# Identifying & updating outlier information in one file
indicator.1 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1) # i.e. suspected outliers...
data1@data <- cbind(data1@data, indicator.1)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of standard boxplot outliers
shades.2 = shading(c(0,1,2),c("blue","white","red")) # i.e. white - no & red - yes
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,13],shades.2)
title("Std. boxplot outliers (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Need moments package to assess skewness (before adjusted boxplots)
require(moments)
# Ignore warning message...
skewness(z1) # skew is fairly high and positive

# Package for adjusted boxplots...
require(robustbase)

# Adjusted boxplot with defaults
X11(width=5.3,height=5.7)
adjbox(z1, main="Adj. boxplot: Evolution of GDP (2000 to 2005)", pch=19, cex=0.5)
```

```
# Adjusted Boxplot statistics...
# Change 'coef' accordingly...
# Default 'coef' is 1.5...
# The higher the 'coef' value the stricter the limits/cut-offs & vice versa...
abp <- adjboxStats(z1, coef=1.5)
abp$stats
abp$stats[1] # the lower limit/cut-off - i.e. values below are deemed outlying...
abp$stats[5] # the upper limit/cut-off - i.e. values above are deemed outlying...
abp$conf
sort(abp$out)
length(abp$out) # number of potential outliers...
#help(adjboxStats) # for details...

# Identifying & updating outlier information in one file
indicator.2 <-ifelse(z1>abp$stats[1]& z1<abp$stats[5], 0, 1) # i.e. suspected outliers...
data1@data <- cbind(data1@data, indicator.2)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of adjusted boxplot outliers
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,14],shades.2)
title("Adj. boxplot outliers (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")




# 4. GW summary statistics and Hawkins' Spatial Outlier Test ###################

# First need to find GW means (i.e. LMs) and GW variances for Hawkin's test...
# In this case, using the gw.cov function in spgwr to find the GW means/variances...

# Note 1. - we could define our own weighting scheme to use with the gw.cov function.
# For example, an IDW-type scheme.
# But in this case, the default bi-square weighting scheme is used.

# Note 2. - we could find an optimal bandwidth (i.e. the optimal number of nearby data)
# for a GW mean using 'leave-one-out' cross-validation.
# But in this case, a user-specified bandwidth is defined as the nearest 10% of data.
# It is not so easy to find an optimal bandwidth for a GW variance
# and as such, is commonly chosen subjectively.

# Note 3. - Hawkins' test should ideally use GW means/variances that have been
# calculated without the observation at each calibration/observation location.
# However, this oversight is not expected to advesely affect results.

# Future work can investigate the above issues...

# To re-cap...
colnames(data.1)

# Defining coordinates....
coordinates(data.1) <- c("X", "Y")
```

```
# GW summary statistics at observation locations (i.e. region centroids)...
# Calculated using 10% of nearby EVOGDP_2000_2005_2006 data.
bwd.1 <- 0.1
gwss <- gw.cov(data.1, vars=11, adapt=bwd.1)
#help(gw.cov) # for details...
names(gwss$SDF) # The GW summary statistics calculated...

# GW means and variances...
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2

# Hawkins' Test for Spatial Outliers...
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth

Hawk.test <- (Hawk.N*(EVOGDP_2000_2005_2006-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
summary(Hawk.test)

# Critical values of the chi-squared distribution
chi_10 <- 2.70554
chi_5 <- 3.84146
chi_2.5 <- 5.02389
chi_1 <- 6.63490
chi_0.5 <- 7.87944
chi_0.01 <- 10.828

# Updating outlier information in one file
indicator.3 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...
data1@data <- cbind(data1@data, Hawk.test, indicator.3)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of spatial outliers ...
shades.3 = shading(c(chi_5,chi_1,chi_0.01),c("white","yellow","orange","red"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,15],shades.3)
title("Spatial outliers: at 5/1/0.01 % (yellow/orange/red) critical levels")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")




# 5. Residual analysis with LM, MLR, LR and GWR models ##########################



# LM...
# Using GW.mean from before...
GW.mean.r <- EVOGDP_2000_2005_2006-GW.mean # Actual minus prediction
summary(GW.mean.r)

# Identifying & updating outlier information in one file
cut.off.1 <- quantile(GW.mean.r, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
```

```
indicator.4 <-ifelse(GW.mean.r>=cut.off.1[2] & GW.mean.r<=cut.off.1[20], 0, 1)
data1@data <- cbind(data1@data, GW.mean.r, indicator.4)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for LM...
shades.4 = shading(c(cut.off.1[2],cut.off.1[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,17],shades.4)
title("Raw resids. from LM: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# MLR...
# First- & second-order polynomial fits of the coordinate data...
mlr.1 <- lm(EVOGDP_2000_2005_2006 ~ X+Y)
mlr.2 <- lm(EVOGDP_2000_2005_2006 ~ X+Y+I(X^2)+I(Y^2)+I(X*Y))
summary(mlr.1)
summary(mlr.2)

# Choosing a second-order MLR fit...

# Using raw residuals as in LM fit...
raw.resids.mlr <- EVOGDP_2000_2005_2006-mlr.2$fitted # Actual minus prediction
summary(raw.resids.mlr)

# Identifying & updating outlier information in one file
cut.off.2 <- quantile(raw.resids.mlr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
indicator.5 <-ifelse(raw.resids.mlr>=cut.off.2[2] & raw.resids.mlr<=cut.off.2[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.mlr, indicator.5)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for MLR...
shades.5 = shading(c(cut.off.2[2],cut.off.2[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,19],shades.5)
title("Raw resids. from MLR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# LR...
# With coordinate data as explanatory variables (i.e. first-order polynomial).

# Using locfit...
require(locfit)
# Ignore warning message...

# Finding the bandwidth for a non-robust LR (i.e. not a lowess fit)
# using generalised cross-validation (GCV) approach.
summary(gcvplot(EVOGDP_2000_2005_2006~X+Y,data=data.1, scale=F,alpha=seq(0.005,0.01,by=0.001),
deg=1,kern="tricube",lfproc=locfit.raw))
```

```
# Choosing a LR fit with bandwidth chosen from above...
bwd.2 <- 0.008
lr <- locfit(EVOGDP_2000_2005_2006~X+Y,data=data.1, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)

# Raw residuals...
lr.p <- fitted.locfit(lr)
raw.resids.lr <- EVOGDP_2000_2005_2006-lr.p # Actual minus prediction
summary(raw.resids.lr)

# Identifying & updating outlier information in one file
cut.off.3 <- quantile(raw.resids.lr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
indicator.6 <-ifelse(raw.resids.lr>=cut.off.3[2] & raw.resids.lr<=cut.off.3[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.lr, indicator.6)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for LR...
shades.6 = shading(c(cut.off.3[2],cut.off.3[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,21],shades.6)
title("Raw resids. from LR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# GWR...
# With coordinate data as explanatory variables (i.e. first-order polynomial).
# Using spgwr...

# Defining the coordinates...
coords.1<-cbind(data.1[,9],data.1[,10])

# Finding the bandwidth for GWR using Akaike Information Criterion (AIC) approach.
#gwr.aic.bwd <-gwr.sel(EVOGDP_2000_2005_2006~X+Y,data=data.1,coords=coords.1,adapt=TRUE,
#gweight=gwr.bisquare, method="aic")
#gwr.aic.bwd[1] # the optimum bandwidth

# Or finding the bandwidth for GWR using cross-validation approach.
#gwr.cv.bwd <-gwr.sel(EVOGDP_2000_2005_2006~X+Y,data=data.1,coords=coords.1,adapt=TRUE,
#gweight=gwr.bisquare, method="cv")
#gwr.cv.bwd[1] # the optimum bandwidth

# Above optimisation can take a long time...
# So choosing a GWR fit with user-specified bandwidth of 0.03...
bwd.3 <- 0.03
gwr.p <-gwr(EVOGDP_2000_2005_2006~X+Y,data=data.1,coords=coords.1,adapt=bwd.3,
gweight=gwr.bisquare,predictions=T)
#gwr.p$SDF

# GWR raw residuals...
raw.resids.gwr <- EVOGDP_2000_2005_2006-gwr.p$SDF$pred
summary(raw.resids.gwr)

# Identifying & updating outlier information in one file
cut.off.4 <- quantile(raw.resids.gwr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
indicator.7 <-ifelse(raw.resids.gwr>=cut.off.4[2] & raw.resids.gwr<=cut.off.4[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.gwr, indicator.7)
data1@data <- as.data.frame(data1@data)
```

```
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for GWR...
shades.7 = shading(c(cut.off.4[2],cut.off.4[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,23],shades.7)
title("Raw resids. from GWR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# NB - Future work should explore the identification of outliers using
# standardiased residuals & corresponding robust regression models...







# 6.  All identified outliers together ##########################################

# Put all indicator data together...
indicator.8 <- indicator.1+indicator.2+indicator.3+indicator.4+indicator.5+indicator.6+indicator.7
summary(indicator.8)
# Histogram
X11(width=5.3,height=5.7)
hist(indicator.8,br=c(0,1,2,3,4,5,6,7))

# Thus a strong case for an outlier relates to an observation
# that has a indicator.8 value of 7...

data1@data <- cbind(data1@data, indicator.8)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)
#write.table(data.1,"Outliers_NUTS_level3.txt", col.names=T,row.names=F)

# A choropleth map of suspected outliers...
shades.7 = shading(c(1,3,5,7),c("white","yellow","orange","red","dark red"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,25],shades.7)
title("Suspected outliers - weak to strong (yellow to dark red) evidence")
choro.legend(-2400000,2200000,shades.7,
over="exactly", between="to under",
fmt="%4.0f",title='Indicator sum (max.: 7)',cex=0.8)
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")
#text(1500000,2400000, "NUTS level 3", cex=2, col=3)
```

Appendix 3 – R script for worked example 3

```
# 1. Preamble ################################################################################



# Worked example 3 - for technical report - challenge 10 - ESPON 2013 database
# NCG - P. Harris & M. Charlton
# 7/2/10



# Objective - to identify statistical outliers in some subset of this data:
# "X"
# "Y"
# "EVOGDP_2000_2005_2006"
# "SPAT_TYPE_1_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# "UNEMP_R_2001_1999"
# "LU_AS_1_1996_1999"
# "LU_AS_2_1996_1999"
# "LU_AS_3_1996_1999"
# "LU_UF_1996_1999"
# "LU_AR_1996_1999"
# "LU_PC_1996_1999"
# "NAT_HAZ_2004_1999"
# "SF_CF_1999_1999"
# "SF_R_1999_1999"
# "SF_S_1999_1999"
# "SF_A_1999_1999"
# "CF_T_1999_1999"
# "CF_E_1999_1999"



# Methods: multivariate - aspatial only
# Only statistical methods:
# 1. Bagplots
# 2. Robust MD-squared analysis (RMD2-AQ-outlier)
# 3. Two techniques based on PCA for outlier detection (PCA-outlier-1 & PCA-outlier-2)



# R packages needed.....
# 1. GISTools (version 0.5-4) - depends on 2 to 11...
# 2. foreign (version 0.8-30)
# 3. gpclib (version 1.4-3)
# 4. maptools (version 0.7-16)
# 5. Matrix (version 0.999375-18)
# 6. RColorBrewer (version 1.0-2)
# 7. sp (version 0.9-28)
# 8. spam (version 0.15-2)
# 9. spdep (version 0.4-29)
# 10. spgwr (version 0.6-2)
# 11. tripack (version 1.2-11)
# 12. aplpack (version 1.2-1) - for bagplots
# 13. robustbase (version 0.4-5) - required for mvoutlier package
# 14. mvoutlier (version 1.4) - for robust MD-squared analysis and PCA outlier detection
```

```
# Base R system version 2.9.0
# N.B. Some of the above packages may still depend on other R packages - download these from R website...



# Relevant data files (see data & ArcGIS directories):

# Excel files...
# 1. ESPON_DATA_NCG_CHALLANGE_10_original.xls
# 2. ESPON_DATA_NCG_CHALLANGE_10_subsets.xls

# ArcGIS files...
# 3. Worked_example_345a_reduced.shp - ArcGIS shapefile of the data...



# The 27 variables...

# "NUTS3","NUTS23","NUTS2","NUTS1","NUTS0" - 5 different NUTS levels
# "New_ID" - relates to the regions name only & is purely numeric
# "NUTS3_2006" - the 2006 NUTS3 version
# "Region_2006" - name of 2006 NUTS3 version
# "X","Y" - centroids of regions
# "EVOGDP_2000_2005_2006" - Evolution of GDP
# and 16 likely contextual variables of "EVOGDP_2000_2005_2006" ...
# "SPAT_TYPE_1_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# "UNEMP_R_2001_1999"
# "LU_AS_1_1996_1999"
# "LU_AS_2_1996_1999"
# "LU_AS_3_1996_1999"
# "LU_UF_1996_1999"
# "LU_AR_1996_1999"
# "LU_PC_1996_1999"
# "NAT_HAZ_2004_1999"
# "SF_CF_1999_1999"
# "SF_R_1999_1999"
# "SF_S_1999_1999"
# "SF_A_1999_1999"
# "CF_T_1999_1999"
# "CF_E_1999_1999"



# NOTE - Methods demonstrated in this worked example do not require
# a relationship between "EVOGDP_2000_2005_2006" and its likely
# contextual data - see worked examples 4 and 5 for this.



# NOTE - This example data set has been reduced to 731 values
# from an original 1351 values
# see readme in excel files on worked example data.









# 2. Importing data as a ArcGIS shapefile & using GISTools to do some maps #####
```

```
require(GISTools)
#help(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
data1 <- readShapePoly("Worked_example_345a_reduced.shp",
proj4string=CRS("+proj=Lambert_Azimuthal_Equal_Area+datum=D_ETRS_1989+ellps=GCS_ETRS_1989"))
colnames(data1@data)

# renaming each variable - as they have been truncated in ArcGIS...
colnames(data1@data) <- c("NUTS3","NUTS23","NUTS2","NUTS1","NUTS0",
"New_ID","NUTS3_2006","Region_2006",
"X","Y","EVOGDP_2000_2005_2006",
"SPAT_TYPE_1_1999_1999","SPAT_TYPE_2_1999_1999",
"UNEMP_R_2001_1999",
"LU_AS_1_1996_1999","LU_AS_2_1996_1999","LU_AS_3_1996_1999",
"LU_UF_1996_1999","LU_AR_1996_1999","LU_PC_1996_1999",
"NAT_HAZ_2004_1999",
"SF_CF_1999_1999",
"SF_R_1999_1999","SF_S_1999_1999","SF_A_1999_1999",
"CF_T_1999_1999","CF_E_1999_1999")

# Size of data set and adding an order ID...
n <- length(data1@data[,1])
Order_ID <- seq(1,n)
data1@data <- cbind(data1@data, Order_ID)
attach(data1@data)

# Coordinate data only...
coords <- cbind(data1@data[,9],data1@data[,10])

# Example multivariate data set one...
Mult.data.1 <- cbind(EVOGDP_2000_2005_2006,SPAT_TYPE_1_1999_1999,SPAT_TYPE_2_1999_1999,
UNEMP_R_2001_1999,LU_AS_1_1996_1999,LU_AS_2_1996_1999,LU_AS_3_1996_1999,LU_UF_1996_1999,
LU_AR_1996_1999,LU_PC_1996_1999,NAT_HAZ_2004_1999,SF_CF_1999_1999,SF_R_1999_1999,
SF_S_1999_1999,SF_A_1999_1999,CF_T_1999_1999,CF_E_1999_1999)
Mult.data.1 <- as.data.frame(Mult.data.1)
attach(Mult.data.1)

# Example multivariate data set two...
Mult.data.2 <- cbind(EVOGDP_2000_2005_2006,UNEMP_R_2001_1999,
NAT_HAZ_2004_1999,SF_CF_1999_1999)
Mult.data.2 <- as.data.frame(Mult.data.2)
attach(Mult.data.2)

# Example multivariate data set three (data set two with coordinates)...
Mult.data.3 <- cbind(X,Y,EVOGDP_2000_2005_2006,UNEMP_R_2001_1999,
NAT_HAZ_2004_1999,SF_CF_1999_1999)
Mult.data.3 <- as.data.frame(Mult.data.3)
attach(Mult.data.3)

# Creating a shading scheme and plotting a choropleth map of EVOGDP_2000_2005_2006...
shades.1 = auto.shading(EVOGDP_2000_2005_2006,5, cols=brewer.pal(5,'PuBuGn'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,EVOGDP_2000_2005_2006,shades.1)
title("Evolution of GDP (2000 to 2005)")
choro.legend(-2300000,250000,shades.1,fmt="%4.0f",title='Evolution of GDP (%)',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")

# Creating a shading scheme and plotting a choropleth map of UNEMP_R_2001_1999...
```

```
shades.2 = auto.shading(UNEMP_R_2001_1999,5, cols=brewer.pal(5,'Greys'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,UNEMP_R_2001_1999,shades.2)
title("Unemployment rate")
choro.legend(-2300000,250000,shades.2,fmt="%4.1f",title='Unemployment (%)',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")

# Creating a shading scheme and plotting a choropleth map of NAT_HAZ_2004_1999...
shades.3 = auto.shading(NAT_HAZ_2004_1999,5, cols=brewer.pal(5,'Greens'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,NAT_HAZ_2004_1999,shades.3)
title("Natural hazards")
choro.legend(-2300000,250000,shades.3,fmt="%4.0f",title='Indication',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")

# Creating a shading scheme and plotting a choropleth map of SF_CF_1999_1999...
shades.4 = auto.shading(SF_CF_1999_1999,5, cols=brewer.pal(5,'Reds'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,SF_CF_1999_1999,shades.4)
title("All Structural & Cohesion Fund expenditure")
choro.legend(-2350000,250000,shades.4,fmt="%4.0f",title='Str. & Coh. Fund',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")









# 3. Bagplots ###########################################################################

# Package for bagplots...
require(aplpack)

# For example, exploring these 2 variables...
z1 <- EVOGDP_2000_2005_2006
z2 <- UNEMP_R_2001_1999

# Summary statistics...
summary(z1)
summary(z2)

# Univariate boxplots...
X11(width=5.3,height=5.7)
boxplot(z1, main="Evolution of GDP", pch=19, cex=0.5)
X11(width=5.3,height=5.7)
boxplot(z2, main="Unemployment", pch=19, cex=0.5)

# The bagplot function
#help(bagplot)

# Example...
X11(width=5.3,height=5.7)
bagp.1 <- bagplot(z1,z2,xlab="Evolution of GDP",ylab="Unemployment",
main="Example bagplot: outliers in red (outside of bag)",cex=0.6)
bivariate.outliers.1 <- bagp.1$pxy.outlier
```

```
length(bivariate.outliers.1[,1])
bivariate.not.outliers.1 <-rbind(bagp.1$pxy.bag,bagp.1$pxy.outer)
length(bivariate.not.outliers.1[,1])

# Some data manipulations for mapping...
# Note can also use library(sqldf) to match datasets...
bivariate.outliers.1x <- merge(data1@data, bivariate.outliers.1,
by.x=c("EVOGDP_2000_2005_2006","UNEMP_R_2001_1999"), by.y=c("x","y"))
indicator.1 <-c(rep(1,length(bivariate.outliers.1x[,1])))
bivariate.outliers.1x <- cbind(bivariate.outliers.1x, indicator.1)

bivariate.not.outliers.1x <- merge(data1@data, bivariate.not.outliers.1,
by.x=c("EVOGDP_2000_2005_2006","UNEMP_R_2001_1999"), by.y=c("x","y"))
indicator.1 <-c(rep(0,length(bivariate.not.outliers.1x[,1])))
bivariate.not.outliers.1x <- cbind(bivariate.not.outliers.1x, indicator.1)

xx1 <- rbind(bivariate.not.outliers.1x, bivariate.outliers.1x)
data1@data <- xx1[order(xx1[,28]),] # get data in correct order with Order_ID
attach(data1@data)

# A choropleth map...
shades.5 = shading(c(0,1,2),c("white","green","red")) # i.e. green - no & red - yes
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,29],shades.5)
title("Bivariate outliers (red): Bagplot of Evolution of GDP with Unemployment")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")




# 4. Robust MD-squared analysis (RMD2-AQ-outlier) ##############################

# Following the paper of Filzmoser et al. (2005)...

# Load the necessary package...
require(mvoutlier)

# Note - multivariate data set one and similar data subsets can give rise to some technical problems
# with this technique, as it is designed for continuous multivariate normal data with outlying observations,
# whereas we have data sets that include categorical data.
# In this respect, we only explore example multivariate data sets two & three, which only have continuous
variables.
# This is not considered a problem as outliers are expected to be more hidden in continuous variables.
# Non-normality of continuous data may however still cause problems.
# Similar comments apply to the PCA methods of section 5...

# Therefore using example multivariate dataset two...

# The key function/plot for this identification technique is...
#help(aq.plot)
#X11(width=12,height=8)
#aq.plot(Mult.data.2)
#help(aq.plot)

# However, slightly adapting the aq.plot function to suit our needs...
aq.plot.1 <- function (x, delta = qchisq(0.975, df = ncol(x)), quan = 1/2,
    alpha = 0.025)
```

```
{
    if (is.vector(x) == TRUE || ncol(x) == 1) {
        stop("x must be at least two-dimensional")
    }
    covr <- covMcd(x, alpha = quan)
    dist <- mahalanobis(x, center = covr$center, cov = covr$cov)
    s <- sort(dist, index = TRUE)
    z <- x
    if (ncol(x) > 2) {
        p <- princomp(x, covmat = covr)
        z <- p$scores[, 1:2]
        sdprop <- (p$sd[1] + p$sd[2])/sum(p$sd)
        cat("Projection to the first and second robust principal components.\n")
        cat("Proportion of total variation (explained variance): ")
        cat(sdprop)
        cat("\n")
    }
    par(mfrow = c(2, 2), mai = c(0.8, 0.6, 0.2, 0.2), mgp = c(2.4,
        1, 0))
    plot(z, col = 3, type = "n",
        main="(A) Data (by ID) projected on the first two RPCs",
        xlab = "First Robust Principal Component (RPC)", ylab = "Second Robust Principal Component (RPC)")
    text(z, dimnames(as.data.frame(z))[[1]], col = 3, cex = 0.8)
    plot(s$x, (1:length(dist))/length(dist), col = 3,
        main = paste("(B) Outlier detection: above ",
        100 * (1 - alpha), "%  & adj. quantiles", sep = ""),
        xlab = "Ordered squared robust Mahalanobis distances",
        ylab = "Cumulative probability", type = "n")
    text(s$x, (1:length(dist))/length(dist), as.character(s$ix),
        col = 3, cex = 0.8)
    t <- seq(0, max(dist), by = 0.01)
    lines(t, pchisq(t, df = ncol(x)), col = 6)
    abline(v = delta, col = 5)
    xarw <- arw(x, covr$center, covr$cov, alpha = alpha)
    # note - arw() is the adaptive reweighted estimator for multivariate location and scatter...
    abline(v = xarw$cn, col = 4)
    legend(11000, 0.3, c("Chi-squared dist. func.", paste(100 * (1 - alpha), "% quantile", sep = ""),
    "Adjusted quantile"), col = c(6,5,4), lty = c(1,1,1), bty="n")
    plot(z, col = 3, type = "n", main = paste("(C) Outliers (in red) based on (user-specified) ",
        100 * (1 - alpha), "% quantile", sep = ""),
        xlab = "First RPC", ylab = "Second RPC")
    for (i in 1:nrow(x)) {
        if (dist[i] >= delta)
            text(z[i, 1], z[i, 2], dimnames(as.data.frame(x))[[1]][i],
                col = 2, cex = 0.8)
        if (dist[i] < delta)
            text(z[i, 1], z[i, 2], dimnames(as.data.frame(x))[[1]][i],
                col = 3, cex = 0.8)
    }
    plot(z, col = 3, type = "n", main = "(D) Outliers (in red) based on adjusted quantile",
        xlab = "First RPC", ylab = "Second RPC")
    for (i in 1:nrow(x)) {
        if (dist[i] >= xarw$cn)
            text(z[i, 1], z[i, 2], dimnames(as.data.frame(x))[[1]][i],
                col = 2, cex = 0.8)
        if (dist[i] < xarw$cn)
            text(z[i, 1], z[i, 2], dimnames(as.data.frame(x))[[1]][i],
                col = 3, cex = 0.8)
    }
    o <- (sqrt(dist) > min(sqrt(xarw$cn), sqrt(qchisq(0.975,
        dim(x)[2]))))
    l <- list(outliers = o)
    l
```

```
}

# Thus our take on the adjusted quantile plot...
X11(width=12,height=8)
mult.out.d2.m1 <- aq.plot.1(Mult.data.2)

# Identifying & updating outlier information in one file
indicator.2 <-ifelse(mult.out.d2.m1$outliers==F, 0, 1) # i.e. suspected outliers...
data1@data <- cbind(data1@data, indicator.2)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of the multivariate outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,30],shades.5)
title("Multivariate outliers: RMD2-AQ-outlier data set 2 (regions coloured red)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")

# Univariate presentation of the same multivariate outliers...
# (i.e multivariate outliers in red as in the above choropleth map)
X11(width=12,height=8)
uni.plot(Mult.data.2)

# NB - see Filzmoser 2005 paper & mvoutlier reference manual for more options
# on the visualisation of multivariate outliers...




# And using example multivariate dataset three...

# The adjusted quantile plot...
X11(width=12,height=8)
mult.out.d3.m1 <- aq.plot.1(Mult.data.3)

# Identifying & updating outlier information in one file
indicator.3 <-ifelse(mult.out.d3.m1$outliers==F, 0, 1) # i.e. suspected outliers...
data1@data <- cbind(data1@data, indicator.3)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of the multivariate outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,31],shades.5)
title("Multivariate outliers: RMD2-AQ-outlier data set 3 (regions coloured red)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")

# Univariate presentation of the same multivariate outliers...
# (i.e multivariate outliers in red as in the above choropleth map)
X11(width=16,height=8)
uni.plot(Mult.data.3)
```

```
# 5. PCA for outlier detection ###################################################

# Following the paper of Filzmoser et al. (2008)...

# Again using the mvoutlier package
# And using only Mult.data.2 data set for simplicity...

# Sign Method for Outlier Identification in High Dimensions...
# i.e. PCA-outlier-1
# Simple version (sign1) & sophisticated (sign2) versions are possible...
# Using the simple version...
mult.out.d2.m2 <- sign1(Mult.data.2)

# Identifying & updating outlier information in one file
indicator.4 <- ifelse(mult.out.d2.m2$wfinal01==1, 0, 1) # i.e. suspected outliers are the wrong way around in this
case...
data1@data <- cbind(data1@data, indicator.4)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of the multivariate outliers
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,32],shades.5)
title("Multivariate outliers: PCA-outlier-1 data set 2 (regions coloured red)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")



# PCOut Method for Outlier Identification in High Dimensions
# i.e. PCA-outlier-2
mult.out.d2.m3 <- pcout(Mult.data.2)

# Identifying & updating outlier information in one file
indicator.5 <- ifelse(mult.out.d2.m3$wfinal01==1, 0, 1) # i.e. suspected outliers are the wrong way around in this
case...
data1@data <- cbind(data1@data, indicator.5)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of the multivariate outliers
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,33],shades.5)
title("Multivariate outliers: PCA-outlier-2 data set 2 (regions coloured red)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")
```

```
# 6. Summary ##########################################################################

# Number of potential outliers

# Bagplot...
sort(-data1@data[,29])

#RMD2-AQ-outlier with multivariate data set 2
sort(-indicator.2)

#PCA-outlier-1 with multivariate data set 2
sort(-indicator.4)

#PCA-outlier-2 with multivariate data set 2
sort(-indicator.5)
```

Appendix 4 – R script for worked example 4

```
# 1. Preamble ##########################################################################



# Worked example 4 - for technical report - challenge 10 - ESPON 2013 database
# NCG - P. Harris & M. Charlton
# 6/2/10



# Objective - to identify statistical outliers in "EVOGDP_2000_2005_2006"
# in relation to some subset of the following explanatory/contextual data:
# "X"
# "Y"
# "SPAT_TYPE_1_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# "UNEMP_R_2001_1999"
# "LU_AS_1_1996_1999"
# "LU_AS_2_1996_1999"
# "LU_AS_3_1996_1999"
# "LU_UF_1996_1999"
# "LU_AR_1996_1999"
# "LU_PC_1996_1999"
# "NAT_HAZ_2004_1999"
# "SF_CF_1999_1999"
# "SF_R_1999_1999"
# "SF_S_1999_1999"
# "SF_A_1999_1999"
# "CF_T_1999_1999"
# "CF_E_1999_1999"



# Methods: multivariate - aspatial & spatial
```

```
# Only statistical methods:
# 1. MLR (multiple linear regression),
# 2. LR (local regression) &
# 3. GWR (geographically weighted regression)



# R packages needed.....
# 1. GISTools (version 0.5-4) - depends on 2 to 11...
# 2. foreign (version 0.8-30)
# 3. gpclib (version 1.4-3)
# 4. maptools (version 0.7-16)
# 5. Matrix (version 0.999375-18)
# 6. RColorBrewer (version 1.0-2)
# 7. sp (version 0.9-28)
# 8. spam (version 0.15-2)
# 9. spdep (version 0.4-29)
# 10. spgwr (version 0.6-2) - for GWSS & GWR
# 11. tripack (version 1.2-11)
# 12. car (version 1.2-12) - for MLR
# 13. locfit (version 1.5-4)- for LR

# Base R system version 2.9.0
# N.B. Some of the above packages may still depend on other R packages - download these from R website...



# Relevant data files (see data & ArcGIS directories):

# Excel files...
# 1. ESPON_DATA_NCG_CHALLANGE_10_original.xls
# 2. ESPON_DATA_NCG_CHALLANGE_10_subsets.xls

# ArcGIS files...
# 3. Worked_example_345a_reduced.shp - ArcGIS shapefile of the data...



# The 27 variables...

# "NUTS3","NUTS23","NUTS2","NUTS1","NUTS0" - 5 different NUTS levels
# "New_ID" - relates to the regions name only & is purely numeric
# "NUTS3_2006" - the 2006 NUTS3 version
# "Region_2006" - name of 2006 NUTS3 version
# "X","Y" - centroids of regions
# "EVOGDP_2000_2005_2006" - Evolution of GDP
# and 16 likely contextual variables of "EVOGDP_2000_2005_2006" ...
# "SPAT_TYPE_1_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# "UNEMP_R_2001_1999"
# "LU_AS_1_1996_1999"
# "LU_AS_2_1996_1999"
# "LU_AS_3_1996_1999"
# "LU_UF_1996_1999"
# "LU_AR_1996_1999"
# "LU_PC_1996_1999"
# "NAT_HAZ_2004_1999"
# "SF_CF_1999_1999"
# "SF_R_1999_1999"
# "SF_S_1999_1999"
# "SF_A_1999_1999"
# "CF_T_1999_1999"
# "CF_E_1999_1999"
```

```
# NOTE - This example data set has been reduced to 731 values from an original 1351 values
# see readme in excel files on worked example data.




# 2. Importing data as a ArcGIS shapefile & using GISTools to do some maps #####

require(GISTools)
#help(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
data1 <- readShapePoly("Worked_example_345a_reduced.shp",
proj4string=CRS("+proj=Lambert_Azimuthal_Equal_Area+datum=D_ETRS_1989+ellps=GCS_ETRS_1989"))
colnames(data1@data)

# renaming each variable - as they have been truncated in ArcGIS...
colnames(data1@data) <- c("NUTS3","NUTS23","NUTS2","NUTS1","NUTS0",
"New_ID","NUTS3_2006","Region_2006",
"X","Y","EVOGDP_2000_2005_2006",
"SPAT_TYPE_1_1999_1999","SPAT_TYPE_2_1999_1999",
"UNEMP_R_2001_1999",
"LU_AS_1_1996_1999","LU_AS_2_1996_1999","LU_AS_3_1996_1999",
"LU_UF_1996_1999","LU_AR_1996_1999","LU_PC_1996_1999",
"NAT_HAZ_2004_1999",
"SF_CF_1999_1999",
"SF_R_1999_1999","SF_S_1999_1999","SF_A_1999_1999",
"CF_T_1999_1999","CF_E_1999_1999")

# Size of data set and adding an order ID...
n <- length(data1@data[,1])
Order_ID <- seq(1,n)
data1@data <- cbind(data1@data, Order_ID)
attach(data1@data)

# Coordinate data only...
coords <- cbind(data1@data[,9],data1@data[,10])

# Creating a shading scheme and plotting a choropleth map of EVOGDP_2000_2005_2006...
shades.1 = auto.shading(EVOGDP_2000_2005_2006,5, cols=brewer.pal(5,'PuBuGn'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,EVOGDP_2000_2005_2006,shades.1)
title("Evolution of GDP (2000 to 2005)")
choro.legend(-2300000,250000,shades.1,fmt="%4.0f",title='Evolution of GDP (%)',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")
```

```
# 3. Exploratory analyses for EVOGDP_2000_2005_2006 relationships ##############

data.1 <- cbind(EVOGDP_2000_2005_2006,X,Y)
data.2 <- cbind(EVOGDP_2000_2005_2006,UNEMP_R_2001_1999,NAT_HAZ_2004_1999)
data.3 <-
cbind(EVOGDP_2000_2005_2006,LU_AS_1_1996_1999,LU_AS_2_1996_1999,LU_AS_3_1996_1999)
data.4 <- cbind(EVOGDP_2000_2005_2006,LU_UF_1996_1999,LU_AR_1996_1999,LU_PC_1996_1999)
data.5 <- cbind(EVOGDP_2000_2005_2006,SF_R_1999_1999,SF_S_1999_1999,SF_A_1999_1999)
data.6 <- cbind(EVOGDP_2000_2005_2006,SF_CF_1999_1999,CF_T_1999_1999,CF_E_1999_1999)

cor(data.1,use="pairwise.complete.obs")
cor(data.2,use="pairwise.complete.obs")
cor(data.3,use="pairwise.complete.obs")
cor(data.4,use="pairwise.complete.obs")
cor(data.5,use="pairwise.complete.obs")
cor(data.6,use="pairwise.complete.obs")

X11(width=6,height=6)
pairs(data.1)
X11(width=6,height=6)
pairs(data.2)
X11(width=6,height=6)
pairs(data.3)
X11(width=6,height=6)
pairs(data.4)
X11(width=6,height=6)
pairs(data.5)
X11(width=6,height=6)
pairs(data.6)

X11(width=6,height=4)
boxplot(EVOGDP_2000_2005_2006~SPAT_TYPE_1_1999_1999,xlab="SPAT_TYPE_1_1999_1999",
ylab="EVOGDP_2000_2005_2006",cex=0.5, main="Evolution of GDP with Spatial typology 1")

X11(width=6,height=4)
boxplot(EVOGDP_2000_2005_2006~SPAT_TYPE_2_1999_1999,xlab="SPAT_TYPE_2_1999_1999",
ylab="EVOGDP_2000_2005_2006",cex=0.5, main="Evolution of GDP with Spatial typology 2")

# Exploratory investigations suggests that
# "X"
# "Y"
# "SF_CF_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# have moderate relationships with EVOGDP_2000_2005_2006

# Coding for a categorical variable in a regression model using factor()...
SPAT_TYPE_2_1999_1999.f <- factor(SPAT_TYPE_2_1999_1999)

# For basic MLR analysis...
require(car)

# Full MLR model
mlr.1 <- lm(EVOGDP_2000_2005_2006 ~ X+Y+SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f)
summary(mlr.1)
vif(mlr.1) # Variance inflation factor (for collinearity)
AIC(mlr.1) # note R gives n*AIC

# AIC stepwise MLR model
mlr.2 <- step(mlr.1)
summary(mlr.2)
vif(mlr.2)
AIC(mlr.2)
```

```
# Results suggest that mlr.1 model is OK...

# We now assume (for section 4.) that the same explanatory variables
# are also important locally with LR and GWR...



# We can also investigate GW correlations using the spgwr function gw.cov

data.1 <- data1@data
coordinates(data.1) <- c("X", "Y")

# GW summary statistics at observation locations (i.e. region centroids)...
# Calculated using 10% of nearby data.
bwd.1 <- 0.1
gwss <- gw.cov(data.1, vars=c(11,22), adapt=bwd.1, cor = TRUE)
names(gwss$SDF) # The GW summary statistics calculated...

# GW correlations...
GW.corr <- gwss$SDF$cor.EVOGDP_2000_2005_2006.SF_CF_1999_1999.
summary(GW.corr) # some evidence of relationship nonstationarity...

# Updating information in one file
data1@data <- cbind(data1@data, GW.corr)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of GW correlations ...
shades.2 = auto.shading(GW.corr,5, cols=brewer.pal(5,'Greys'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,29],shades.2)
title("Relationship nonstationarity: Evolution of GDP with Str./Coh. Fund")
choro.legend(-2300000,250000,shades.2,fmt="%4.1f",title='Correlation',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")









# 4. Residual analysis with MLR, LR and GWR models ##############################


# Using raw residuals from mlr-1 fit...
raw.resids.mlr <- EVOGDP_2000_2005_2006-mlr.1$fitted # Actual minus prediction
summary(raw.resids.mlr)

# Identifying & updating outlier information in one file
cut.off.1 <- quantile(raw.resids.mlr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
indicator.1 <-ifelse(raw.resids.mlr>=cut.off.1[2] & raw.resids.mlr<=cut.off.1[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.mlr, indicator.1)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for MLR...
```

```
shades.3 = shading(c(cut.off.1[2],cut.off.1[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,30],shades.3)
title("Raw resids. from MLR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")



# LR using locfit...
require(locfit)
# Ignore warning message...

# Finding the bandwidth for a non-robust LR (i.e. not a lowess fit)
# using generalised cross-validation (GCV) approach.
summary(gcvplot(EVOGDP_2000_2005_2006~X+Y+SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1, scale=F,alpha=seq(0.1,1,by=0.1),
deg=1,kern="tricube",lfproc=locfit.raw))

# Choosing a LR fit with bandwidth chosen from above...
bwd.2 <- 0.7
lr <- locfit(EVOGDP_2000_2005_2006~X+Y+SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1, scale=F, alpha=bwd.2,deg=1,kern="tricube",lfproc=locfit.raw)

# Raw residuals...
lr.p <- fitted.locfit(lr)
raw.resids.lr <- EVOGDP_2000_2005_2006-lr.p # Actual minus prediction
summary(raw.resids.lr)

# Identifying & updating outlier information in one file
cut.off.2 <- quantile(raw.resids.lr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
indicator.2 <-ifelse(raw.resids.lr>=cut.off.2[2] & raw.resids.lr<=cut.off.2[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.lr, indicator.2)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for LR...
shades.4 = shading(c(cut.off.2[2],cut.off.2[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,32],shades.4)
title("Raw resids. from LR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")



# GWR using spgwr...
# Defining the coordinates...
coords.1<-cbind(data.1[,9],data.1[,10])

# Finding the bandwidth for GWR using Akaike Information Criterion (AIC) approach.
#gwr.aic.bwd <-gwr.sel(EVOGDP_2000_2005_2006~X+Y+SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
#data=data.1,coords=coords.1,adapt=TRUE,
#gweight=gwr.bisquare, method="aic")
#gwr.aic.bwd[1] # the optimum bandwidth

# Or finding the bandwidth for GWR using cross-validation (CV) approach.
gwr.cv.bwd <-gwr.sel(EVOGDP_2000_2005_2006~X+Y+SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1,coords=coords.1,adapt=TRUE,
```

```
gweight=gwr.bisquare, method="cv")
gwr.cv.bwd[1] # the optimum bandwidth

# Using CV bandwidth...
bwd.3 <- gwr.cv.bwd[1]
gwr.p <-gwr(EVOGDP_2000_2005_2006~X+Y+SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1,coords=coords.1,adapt=bwd.3,gweight=gwr.bisquare,predictions=T)
#gwr.p$SDF

# GWR raw residuals...
raw.resids.gwr <- EVOGDP_2000_2005_2006-gwr.p$SDF$pred
summary(raw.resids.gwr)

# Identifying & updating outlier information in one file
cut.off.3 <- quantile(raw.resids.gwr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails - alter accordingly...
indicator.3 <-ifelse(raw.resids.gwr>=cut.off.3[2] & raw.resids.gwr<=cut.off.3[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.gwr, indicator.3)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for GWR...
shades.5 = shading(c(cut.off.3[2],cut.off.3[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,34],shades.5)
title("Raw resids. from GWR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")



# NB - Future work should explore the identification of outliers using
# standardiased residuals & corresponding robust regression models...









# 5.  All identified outliers together ##########################################

# Put all indicator data together...
indicator.4 <- indicator.1+indicator.2+indicator.3
summary(indicator.4)
# Histogram
X11(width=5.3,height=5.7)
hist(indicator.4,br=c(0,1,2,3))

# Thus a strong case for an outlier relates to an observation
# that has a indicator.4 value of 3...

data1@data <- cbind(data1@data, indicator.4)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of suspected outliers...
shades.6 = shading(c(1,2,3),c("white","yellow","orange","red"))
```

```
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,36],shades.6)
title("Suspected outliers - weak to strong (yellow to red) evidence")
choro.legend(-2300000,250000,shades.6,over="exactly", between="to under",
fmt="%4.0f",title='Indicator sum (max.: 3)',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")
```

Appendix 5 – R script for worked example 5

```
# 1. Preamble ##########################################################################



# Worked example 5 - for technical report - challenge 10 - ESPON 2013 database
# NCG - P. Harris & M. Charlton
# 7/2/10



# Objective - to identify statistical clusters in "EVOGDP_2000_2005_2006" with
# respect to key spatial moments...
# GW means/standard deviations; Relationships - GW correlations/regressions;
# Spatial autocorrelation (Moran's I)
# For relationships -
# EVOGDP_2000_2005_2006 is related to some subset of the following explanatory/contextual data:
# "X"
# "Y"
# "SPAT_TYPE_1_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# "UNEMP_R_2001_1999"
# "LU_AS_1_1996_1999"
# "LU_AS_2_1996_1999"
# "LU_AS_3_1996_1999"
# "LU_UF_1996_1999"
# "LU_AR_1996_1999"
# "LU_PC_1996_1999"
# "NAT_HAZ_2004_1999"
# "SF_CF_1999_1999"
# "SF_R_1999_1999"
# "SF_S_1999_1999"
# "SF_A_1999_1999"
# "CF_T_1999_1999"
# "CF_E_1999_1999"



# Methods: univariate & multivariate - all spatial
# Only statistical methods:
# 1. GWSS (geographically weighted summary statistics)
# 2. MLR (multiple linear regression) & GWR (geographically weighted regression)
```

```
# 3. Global and local Moran's I



# R packages needed.....
# 1. GISTools (version 0.5-4) - depends on 2 to 11...
# 2. foreign (version 0.8-30)
# 3. gpclib (version 1.4-3)
# 4. maptools (version 0.7-16)
# 5. Matrix (version 0.999375-18)
# 6. RColorBrewer (version 1.0-2)
# 7. sp (version 0.9-28)
# 8. spam (version 0.15-2)
# 9. spdep (version 0.4-29) - for global and local Moran's I
# 10. spgwr (version 0.6-2) - for GWSS and GWR
# 11. tripack (version 1.2-11)
# 12. car (version 1.2-12) - for MLR

# Base R system version 2.9.0
# N.B. Some of the above packages may still depend on other R packages - download these from R website...



# Relevant data files (see data & ArcGIS directories):

# Excel files...
# 1. ESPON_DATA_NCG_CHALLANGE_10_original.xls
# 2. ESPON_DATA_NCG_CHALLANGE_10_subsets.xls

# ArcGIS files...
# 3. Worked_example_345a_reduced.shp - ArcGIS shapefile of the data...



# The 27 variables...

# "NUTS3","NUTS23","NUTS2","NUTS1","NUTS0" - 5 different NUTS levels
# "New_ID" - relates to the regions name only & is purely numeric
# "NUTS3_2006" - the 2006 NUTS3 version
# "Region_2006" - name of 2006 NUTS3 version
# "X","Y" - centroids of regions
# "EVOGDP_2000_2005_2006" - Evolution of GDP
# and 16 likely contextual variables of "EVOGDP_2000_2005_2006" ...
# "SPAT_TYPE_1_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# "UNEMP_R_2001_1999"
# "LU_AS_1_1996_1999"
# "LU_AS_2_1996_1999"
# "LU_AS_3_1996_1999"
# "LU_UF_1996_1999"
# "LU_AR_1996_1999"
# "LU_PC_1996_1999"
# "NAT_HAZ_2004_1999"
# "SF_CF_1999_1999"
# "SF_R_1999_1999"
# "SF_S_1999_1999"
# "SF_A_1999_1999"
# "CF_T_1999_1999"
# "CF_E_1999_1999"



# NOTE - This example data set has been reduced to 731 values from an original 1351 values
```

```
# see readme in excel files on worked example data.




# 2. Importing data as a ArcGIS shapefile & using GISTools to do some maps #####

require(GISTools)
#help(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
data1 <- readShapePoly("Worked_example_345a_reduced.shp",
proj4string=CRS("+proj=Lambert_Azimuthal_Equal_Area+datum=D_ETRS_1989+ellps=GCS_ETRS_1989"))
colnames(data1@data)

# renaming each variable - as they have been truncated in ArcGIS...
colnames(data1@data) <- c("NUTS3","NUTS23","NUTS2","NUTS1","NUTS0",
"New_ID","NUTS3_2006","Region_2006",
"X","Y","EVOGDP_2000_2005_2006",
"SPAT_TYPE_1_1999_1999","SPAT_TYPE_2_1999_1999",
"UNEMP_R_2001_1999",
"LU_AS_1_1996_1999","LU_AS_2_1996_1999","LU_AS_3_1996_1999",
"LU_UF_1996_1999","LU_AR_1996_1999","LU_PC_1996_1999",
"NAT_HAZ_2004_1999",
"SF_CF_1999_1999",
"SF_R_1999_1999","SF_S_1999_1999","SF_A_1999_1999",
"CF_T_1999_1999","CF_E_1999_1999")

# Size of data set and adding an order ID...
n <- length(data1@data[,1])
Order_ID <- seq(1,n)
data1@data <- cbind(data1@data, Order_ID)
attach(data1@data)

# Coordinate data only...
coords <- cbind(data1@data[,9],data1@data[,10])

# Creating a shading scheme and plotting a choropleth map of EVOGDP_2000_2005_2006...
shades.1 = auto.shading(EVOGDP_2000_2005_2006,5, cols=brewer.pal(5,'YlOrBr'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,EVOGDP_2000_2005_2006,shades.1)
title("Evolution of GDP (2000 to 2005)")
choro.legend(-2300000,250000,shades.1,fmt="%4.0f",title='Evolution of GDP (%)',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")







# 3. Exploratory analyses for EVOGDP_2000_2005_2006 spatial moments ############

# Global summary statistics...
summary(EVOGDP_2000_2005_2006)
```

```
# standard deviation
sd <- (var(EVOGDP_2000_2005_2006))^0.5
sd



# As in worked example 4 we can investigate GW summary statistics
# using the spgwr function gw.cov

data.1 <- data1@data
coordinates(data.1) <- c("X", "Y")

# GW summary statistics at observation locations (i.e. region centroids)...
# Calculated using 10% of nearby data.
bwd.1 <- 0.1
gwss <- gw.cov(data.1, vars=c(11,22), adapt=bwd.1, cor = TRUE)
names(gwss$SDF) # The GW summary statistics calculated...



# GW means...
GW.mean <- gwss$SDF$mean.EVOGDP_2000_2005_2006
summary(GW.mean) # some evidence of mean nonstationarity...

# Updating information in one file
data1@data <- cbind(data1@data, GW.mean)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of GW means ...
shades.2 = auto.shading(GW.mean,5, cols=brewer.pal(5,'YlOrBr'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,29],shades.2)
title("Mean nonstationarity: Evolution of GDP")
choro.legend(-2300000,250000,shades.2,fmt="%4.1f",title='Mean',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")



# GW standard deviations...
GW.sd <- gwss$SDF$sd.EVOGDP_2000_2005_2006
summary(GW.sd) # some evidence of SD nonstationarity...

# Updating information in one file
data1@data <- cbind(data1@data, GW.sd)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of GW SDs ...
shades.3 = auto.shading(GW.sd,5, cols=brewer.pal(5,'RdPu'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,30],shades.3)
title("Standard Deviation nonstationarity: Evolution of GDP")
choro.legend(-2300000,250000,shades.3,fmt="%4.1f",title='SD',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")
```

```
# For all GW summary statistics randomisation tests can be used to identfy
# locations of unsually different local statistics -
# see Fotheringham et al. (2002); Harris and Brunsdon (2010)

# As an example: randomisation test for the standard deviation...
n.r1 <- 99  # Number of randomisations (the more the better)
out.x <- matrix(nrow=n,ncol=n.r1)
for(i in 1:n.r1)
{
rand.dat <- sample(data.1[,11])
data.2 <- cbind(data1@data, rand.dat)
data.2 <- as.data.frame(data.2)
attach(data.2)
coordinates(data.2) <- c("X", "Y")
gwss.rand <- gw.cov(data.2, vars=c(31), adapt=bwd.1)
out.x[,i] <- gwss.rand$SDF$sd.V1
}
# combining the randomisation results with the actual result...
out.x1 <- cbind(GW.sd, out.x)
out.x2 <- t(apply(out.x1,1,rank))
Random.sd <- out.x2[,1]

# Updating information in one file
data1@data <- cbind(data1@data, Random.sd)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of randomisation test result ...
shades.4 = shading(c(2.5,97.5),c("blue","white","green")) # test at 95% level...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,31],shades.4)
title("Areas of unsually high (green) and low (blue) standard deviation")
map.scale(100000,-750000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="red")
text(-320000,-1150000, "Randomisation test at 95% level for Evolution of GDP")







# 4. Exploratory analyses for EVOGDP_2000_2005_2006 w.r.t global relationships #

# From exploratory investigations of worked example 4:
# "SF_CF_1999_1999"
# "SPAT_TYPE_2_1999_1999"
# have moderate relationships with EVOGDP_2000_2005_2006
# i.e.

# Correlation & scatterplot...
cor(EVOGDP_2000_2005_2006,SF_CF_1999_1999,use="pairwise.complete.obs")
X11(width=6,height=6)
plot(EVOGDP_2000_2005_2006,SF_CF_1999_1999, main="Evolution of GDP with Str/Coh Fund",
pch=19, cex=0.5)
```

```
# Boxplot for categorical variable...
X11(width=6,height=4)
boxplot(EVOGDP_2000_2005_2006~SPAT_TYPE_2_1999_1999,xlab="SPAT_TYPE_2_1999_1999",
ylab="EVOGDP_2000_2005_2006",cex=0.5, main="Evolution of GDP with Spatial typology 2")

# Coding for a categorical variable in a regression model using factor()...
SPAT_TYPE_2_1999_1999.f <- factor(SPAT_TYPE_2_1999_1999)

# For useful basic MLR analysis...
require(car)

# Full MLR model
mlr.1 <- lm(EVOGDP_2000_2005_2006 ~ SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f)
summary(mlr.1)
vif(mlr.1) # Variance inflation factor (for collinearity)
AIC(mlr.1) # note R gives n*AIC

# AIC stepwise MLR model
mlr.2 <- step(mlr.1)
summary(mlr.2)
vif(mlr.2)
AIC(mlr.2)

# Results suggest that mlr.1 model is OK...

# We also assume that the same explanatory variables
# are also important locally with GWR...




# 5. Exploratory analyses for EVOGDP_2000_2005_2006 w.r.t local relationships ##

# We can also investigate GW correlations from the spgwr function gw.cov output in section 3.

# GW correlations...
GW.corr <- gwss$SDF$cor.EVOGDP_2000_2005_2006.SF_CF_1999_1999.
summary(GW.corr) # some evidence of relationship nonstationarity...

# Updating information in one file
data1@data <- cbind(data1@data, GW.corr)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of GW correlations ...
shades.5 = auto.shading(GW.corr,5, cols=brewer.pal(5,'Greys'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,32],shades.5)
title("Relationship nonstationarity: Evolution of GDP with Str./Coh. Fund")
choro.legend(-2300000,250000,shades.5,fmt="%4.1f",title='Correlation',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")



# Now using GWR to explore local relationships (using spgwr)...
```

```
# Defining the coordinates...
coords.1<-cbind(data.1[,9],data.1[,10])

# Finding the bandwidth for GWR using cross-validation (CV) approach.
gwr.cv.bwd <-gwr.sel(EVOGDP_2000_2005_2006~SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1,coords=coords.1,adapt=TRUE,
gweight=gwr.bisquare, method="cv")
gwr.cv.bwd[1] # the optimum bandwidth

# GWR using CV bandwidth...
bwd.2 <- gwr.cv.bwd[1]
gwr.1 <-gwr(EVOGDP_2000_2005_2006~SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1,coords=coords.1,adapt=bwd.2,gweight=gwr.bisquare)
#gwr.1$SDF

# As an example, only investigating coefficients (or parameters) for SPAT_TYPE_2_1999_1999 class 2
gwr.coeff.1 <- gwr.1$SDF$SPAT_TYPE_2_1999_1999.f2

# Updating information in one file
data1@data <- cbind(data1@data, gwr.coeff.1)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of this particular part of the GWR output ...
shades.6 = auto.shading(gwr.coeff.1, 5, cols=brewer.pal(5,'Blues'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,33],shades.6)
title("Relationship nonstationarity: coefficient for spatial typology 2 (class 2)")
choro.legend(-2300000,250000,shades.6,fmt="%1.2f",title='GWR coefficient',cex=0.8)
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="green")



# Randomisation test -
# To gauge whether or not any observed variation in the GWR coefficients is unusual...

# First get the actual variance of each local coefficient estimate...
output.x1 <- as(gwr.1$SDF, "data.frame")[,2:8]
output.1 <- as.vector(apply(output.x1,2,var))

# Now find the same variances based on 99 randomisations of the data...
output.2 <- matrix(nrow=n.r1,ncol=7) # i.e. 7 columns (intercept plus number of indep. variables)
for(i in 1:n.r1)
{
print(i) # counter
c1<-t(coords.1)
c1 <- as.data.frame(c1)
c1.s <-sample(c1,length(data.1[,1]))
c1.s <- as.data.frame(c1.s)
c1.s <- t(c1.s)
coords.2 <- as.matrix(c1.s)
gwr.2 <-gwr(EVOGDP_2000_2005_2006~SF_CF_1999_1999+SPAT_TYPE_2_1999_1999.f,
data=data.1,coords=coords.2,adapt=bwd.2,gweight=gwr.bisquare)
output.x2 <- as(gwr.2$SDF, "data.frame")[,2:8]
output.2[i,] <- as.vector(apply(output.x2,2,var))
}

# p-values for each coefficient estimate
```

```
output.3 <- rbind(output.1,output.2)
r.1 <- rank(output.3[,1])
r.11 <- ((n.r1+2)-r.1[1])/(n.r1+1)
r.2 <- rank(output.3[,2])
r.22 <- ((n.r1+2)-r.2[1])/(n.r1+1)
r.3 <- rank(output.3[,3])
r.33 <- ((n.r1+2)-r.3[1])/(n.r1+1)
r.4 <- rank(output.3[,4])
r.44 <- ((n.r1+2)-r.4[1])/(n.r1+1)
r.5 <- rank(output.3[,5])
r.55 <- ((n.r1+2)-r.5[1])/(n.r1+1)
r.6 <- rank(output.3[,6])
r.66 <- ((n.r1+2)-r.6[1])/(n.r1+1)
r.7 <- rank(output.3[,7])
r.77 <- ((n.r1+2)-r.7[1])/(n.r1+1)

# Thus in this case, all Monte Carlo tests are based on 99 randomisations of the data.
# The larger the p-value,the more support is given to the null hypothesis
# of a stationary regression coefficient estimate.
rand.test.1 <- cbind(r.11,r.22,r.33,r.44,r.55,r.66,r.77)
rand.test.1









# 6. Global and local autocorrelation ##############################################

# Global Moran's I
# Local Moran's I (a Local Indicator of Spatial Association LISA)

require(spdep) # for global and local Moran's I

# Firstly, two different examples to define spatial distances or spatial topology.
# A measure of distance is needed to calculate local and global Moran's I statistics.

data1.labs = poly.labels(data1)

# 1. Queen's case spatial topology (from chess)
data1.nb1 = poly2nb(data1)
X11(width=8,height=7)
par(mar=c(0,0,2,0))
plot(data1,col='grey')
plot(data1.nb1,coordinates(data1.labs),col='red',add=TRUE)
title("Queen's case spatial topology")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")

# 2. Alternatively, base topology on nearness of polygons rather than contiguity.
# Here we define two polygons as neighbours if they are within some distance d of one another.
# Thus letting d = 100000m, for example...
data1.nb2 = dnearneigh(poly.labels(data1),0,100000)
X11(width=8,height=7)
par(mar=c(0,0,2,0))
plot(data1,col='grey')
plot(data1.nb2,coordinates(data1.labs),col='red',add=TRUE)
title("Regions whose centroids are within 100km of each other")
map.scale(100000,-1050000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="blue")
```

```
# Tests for global Moran's I statistic using 'Queens case topology' contiguity matrix
data1.lw = nb2listw(data1.nb1,zero.policy=T)
attach(data.frame(data1))

# Global Moran's I:
moran(EVOGDP_2000_2005_2006, data1.lw, n, Szero(data1.lw), zero.policy=T)

# The assumptions underlying the following test are sensitive to the form
# of the graph of neighbour relationships and other factors.
# Results may be checked against those of moran.mc permutations.
moran.test(EVOGDP_2000_2005_2006,data1.lw, zero.policy=T)

# A permutation test for Moran's I statistic calculated by using nsim random permutations of x for
# the given spatial weighting scheme, to establish the rank of the observed statistic in relation to the
# nsim simulated values.
moran.mc(EVOGDP_2000_2005_2006,data1.lw, zero.policy=T,nsim=10000)



# Local Moran's I also using 'Queens case topology' contiguity matrix
Local.moran <- localmoran(EVOGDP_2000_2005_2006,data1.lw, zero.policy=T)

# Summary of local Moran's I
summary(Local.moran)

# Updating information in one file
data1@data <- cbind(data1@data, Local.moran[,1])
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of local Moran's I for EVOGDP_2000_2005_2006...
shades.7 = shading(c(0,0.572),c("red","grey","blue")) # shading relates to global value...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,34],shades.7)
title("Autocorrelation nonstationarity: Local Moran's I for Evolution of GDP")
map.scale(100000,-750000,500000,"x1000 km",4,0.5)
north.arrow(200000,750000,40000, col="black")
text(-360000,-1150000, "-ve (red); +ve/below global (grey); +ve/above global (blue)")
```

Appendix 6 – R script for worked example 6

```
# 1. Preamble ##################################################################



# Worked example 6 - for technical report - challenge 10 - ESPON 2013 database
```

```
# NCG - P. Harris & M. Charlton
# 8/2/10



# Objective - to identify statistical outliers in:
# "EVOGDP_2000_2005_2006" at three different NUTS levels (2, 1 & 0)
# - i.e. investigate the effects of MAUP on outlier identification
# Results are compared with those of worked example 2 for NUTS3.



# Methods: univariate - aspatial & spatial (i.e. the same as that presented in worked example 2)
# Only statistical methods:
# 1. Standard and Adjusted boxplots,
# 2. Hawkins' test (includes the use of GWSS -
#            geographically weighted summary statistics - GW means and variances),
# 3. LM (local mean, i.e. a GW mean)
# 4. MLR (multiple linear regression),
# 5. LR (local regression) &
# 6. GWR (geographically weighted regression)



# R packages needed.....
# 1. GISTools (version 0.5-4) - depends on 2 to 11...
# 2. foreign (version 0.8-30)
# 3. gpclib (version 1.4-3)
# 4. maptools (version 0.7-16)
# 5. Matrix (version 0.999375-18)
# 6. RColorBrewer (version 1.0-2)
# 7. sp (version 0.9-28)
# 8. spam (version 0.15-2)
# 9. spdep (version 0.4-29)
# 10. spgwr (version 0.6-2) - for GWSS & GWR
# 11. tripack (version 1.2-11)
# 12. moments (version 0.11) - for skewness
# 13. robustbase (version 0.4-5) - for adjusted boxplots
# 14. locfit (version 1.5-4)- for LR

# Base R system version 2.9.0
# N.B. Some of the above packages may still depend on other R packages -
# download these from R website...



# Relevant data files (see data & ArcGIS directories):

# Excel files...
# 1. ESPON_DATA_NCG_CHALLANGE_10_original.xls
# 2. ESPON_DATA_NCG_CHALLANGE_10_subsets.xls

# ArcGIS files...
# 3. Worked_example_6c_Dissolve_nuts2a.shp - ArcGIS shapefile of the NUTS2 data (278 values)
# 4. Worked_example_6c_Dissolve_nuts1a.shp - ArcGIS shapefile of the NUTS1 data (95 values)
# 5. Worked_example_6c_Dissolve_nuts0a.shp - ArcGIS shapefile of the NUTS0 data (30 values)



# The 10 variables...

# "NUTSx", - the NUTS level - 2, 1 or 0
# "NUTS3_outlier_mean" - mean of outlier indicator.8 from worked example 2
```

```
#                    when going from NUTS 3 to larger scale
# "NUTS3_outlier_max" - maximum of outlier indicator.8 from worked example 2
#                    when going from NUTS 3 to larger scale
# "GDP_2000_2006" - mean of NUTS3 data when going from NUTS 3 to larger scale
# "GDP_2005_2006" - mean of NUTS3 data when going from NUTS 3 to larger scale
# "POP_T_2000_2006" - mean of NUTS3 data when going from NUTS 3 to larger scale
# "POP_T_2005_2006" - mean of NUTS3 data when going from NUTS 3 to larger scale
# "X","Y" - centroids of NUTS regions
# "EVOGDP_2000_2005_2006" - the variable of interest re-calculated from
#                     the relevant variables above




# Change the following script in six places to go through each NUTS level...









# 2. Importing data as a ArcGIS shapefile & using GISTools to do a map... ######

require(GISTools)
#help(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
#data1 <- readShapePoly("Worked_example_6c_Dissolve_nuts2a.shp",
data1 <- readShapePoly("Worked_example_6c_Dissolve_nuts1a.shp",
#data1 <- readShapePoly("Worked_example_6c_Dissolve_nuts0a.shp",
proj4string=CRS("+proj=Lambert_Azimuthal_Equal_Area+datum=D_ETRS_1989+ellps=GCS_ETRS_1989"))
colnames(data1@data)

# Renaming each variable - as they have been altered by ArcGIS commands...
colnames(data1@data) <- c("NUTS2", "NUTS3_outlier_mean", "NUTS3_outlier_max",
"GDP_2000_2006", "GDP_2005_2006", "POP_T_2000_2006", "POP_T_2005_2006",
"X","Y")

# Size of data set and adding an order ID...
n <- length(data1@data[,1])
Order_ID <- seq(1,n)
data1@data <- cbind(data1@data, Order_ID)
attach(data1@data)

# Calculating the new EVOGDP_2000_2005_2006 values...
EVOGDP_2000_2005_2006 <-
((data1@data[,5]/data1@data[,7])*1000)/((data1@data[,4]/data1@data[,6])*1000)*100
data1@data <- cbind(data1@data, EVOGDP_2000_2005_2006)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Creating a shading scheme and plotting a choropleth map of EVOGDP_2000_2005_2006...
shades.1 = auto.shading(EVOGDP_2000_2005_2006,5, cols=brewer.pal(5,'Greys'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,EVOGDP_2000_2005_2006,shades.1)
title("Evolution of GDP (2000 to 2005)")
choro.legend(-2400000,2200000,shades.1,fmt="%4.0f",title='Evolution of GDP (%)',cex=0.8)
```

```
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")
#text(1500000,2400000, "NUTS level 2", cex=2, col=3)
text(1500000,2400000, "NUTS level 1", cex=2, col=3)
#text(1500000,2400000, "NUTS level 0", cex=2, col=3)









# 3. Boxplots ########################################################################

# Let EVOGDP_2000_2005_2006 be z1...
z1 <- EVOGDP_2000_2005_2006

# Exploring this data...
summary(z1) # summary statistics
sort(z1) # ordered data

# Histogram
X11(width=5.3,height=5.7)
hist(z1, main="Histogram: Evolution of GDP (2000 to 2005)",xlab="Evolution of GDP")

# Standard boxplot with defaults
X11(width=5.3,height=5.7)
boxplot(z1, main="Std. boxplot: Evolution of GDP (2000 to 2005)", pch=19, cex=0.5)

# Standard Boxplot statistics...
# Change 'coef' accordingly...
# Default 'coef' is 1.5...
# The higher the 'coef' value the stricter the limits/cut-offs & vice versa...
bp <- boxplot.stats(z1, coef=1.5)
bp$stats
bp$stats[1] # the lower limit/cut-off - i.e. values below are deemed outlying...
bp$stats[5] # the upper limit/cut-off - i.e. values above are deemed outlying...
bp$conf
sort(bp$out)
length(bp$out) # number of potential outliers...
# help(boxplot.stats) # for details...

# Identifying & updating outlier information in one file
indicator.1 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1) # i.e. suspected outliers...
data1@data <- cbind(data1@data, indicator.1)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of standard boxplot outliers
shades.2 = shading(c(0,1,2),c("blue","white","red")) # i.e. white - no & red - yes
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,12],shades.2)
title("Std. boxplot outliers (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")

# Need moments package to assess skewness (before adjusted boxplots)
require(moments)
# Ignore warning message...
```

```
skewness(z1)

# Package for adjusted boxplots...
require(robustbase)

# Adjusted boxplot with defaults
X11(width=5.3,height=5.7)
adjbox(z1, main="Adj. boxplot: Evolution of GDP (2000 to 2005)", pch=19, cex=0.5)

# Adjusted Boxplot statistics...
# Change 'coef' accordingly...
# Default 'coef' is 1.5...
# The higher the 'coef' value the stricter the limits/cut-offs & vice versa...
abp <- adjboxStats(z1, coef=1.5)
abp$stats
abp$stats[1] # the lower limit/cut-off - i.e. values below are deemed outlying...
abp$stats[5] # the upper limit/cut-off - i.e. values above are deemed outlying...
abp$conf
sort(abp$out)
length(abp$out) # number of potential outliers...
#help(adjboxStats) # for details...

# Identifying & updating outlier information in one file
indicator.2 <-ifelse(z1>abp$stats[1]& z1<abp$stats[5], 0, 1) # i.e. suspected outliers...
data1@data <- cbind(data1@data, indicator.2)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of adjusted boxplot outliers
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,13],shades.2)
title("Adj. boxplot outliers (regions coloured red)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")




# 4. GW summary statistics and Hawkins' Spatial Outlier Test ####################

# First need to find GW means (i.e. LMs) and GW variances for Hawkin's test...
# In this case, using the gw.cov function in spgwr to find the GW means/variances...

# To re-cap...
colnames(data.1)

# Defining coordinates....
coordinates(data.1) <- c("X", "Y")

# GW summary statistics at observation locations (i.e. region centroids)...
# Calculated using 10% of nearby EVOGDP_2000_2005_2006 data.
bwd.1 <- 0.1
gwss <- gw.cov(data.1, vars=11, adapt=bwd.1)
#help(gw.cov) # for details...
names(gwss$SDF) # The GW summary statistics calculated...
```

```
# GW means and variances...
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2

# Hawkins' Test for Spatial Outliers...
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth

Hawk.test <- (Hawk.N*(EVOGDP_2000_2005_2006-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
summary(Hawk.test)

# Critical values of the chi-squared distribution
chi_10 <- 2.70554
chi_5 <- 3.84146
chi_2.5 <- 5.02389
chi_1 <- 6.63490
chi_0.5 <- 7.87944
chi_0.01 <- 10.828

# Updating outlier information in one file
indicator.3 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...
data1@data <- cbind(data1@data, Hawk.test, indicator.3)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of spatial outliers ...
shades.3 = shading(c(chi_5,chi_1,chi_0.01),c("white","yellow","orange","red"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,14],shades.3)
title("Spatial outliers: at 5/1/0.01 % (yellow/orange/red) critical levels")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")









# 5. Residual analysis with LM, MLR, LR and GWR models ##########################



# LM...
# Using GW.mean from before...
GW.mean.r <- EVOGDP_2000_2005_2006-GW.mean # Actual minus prediction
summary(GW.mean.r)

# Identifying & updating outlier information in one file
cut.off.1 <- quantile(GW.mean.r, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails
indicator.4 <-ifelse(GW.mean.r>=cut.off.1[2] & GW.mean.r<=cut.off.1[20], 0, 1)
data1@data <- cbind(data1@data, GW.mean.r, indicator.4)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for LM...
```

```
shades.4 = shading(c(cut.off.1[2],cut.off.1[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,16],shades.4)
title("Raw resids. from LM: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# MLR...
# First- & second-order polynomial fits of the coordinate data...
mlr.1 <- lm(EVOGDP_2000_2005_2006 ~ X+Y)
mlr.2 <- lm(EVOGDP_2000_2005_2006 ~ X+Y+I(X^2)+I(Y^2)+I(X*Y))
summary(mlr.1)
summary(mlr.2)

# Choosing a second-order MLR fit...

# Using raw residuals as in LM fit...
raw.resids.mlr <- EVOGDP_2000_2005_2006-mlr.2$fitted # Actual minus prediction
summary(raw.resids.mlr)

# Identifying & updating outlier information in one file
cut.off.2 <- quantile(raw.resids.mlr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails
indicator.5 <-ifelse(raw.resids.mlr>=cut.off.2[2] & raw.resids.mlr<=cut.off.2[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.mlr, indicator.5)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for MLR...
shades.5 = shading(c(cut.off.2[2],cut.off.2[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,18],shades.5)
title("Raw resids. from MLR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# LR...
# With coordinate data as explanatory variables (i.e. first-order polynomial).

# Using locfit...
require(locfit)
# Ignore warning message...

# Finding the bandwidth for a non-robust LR (i.e. not a lowess fit)
# using generalised cross-validation (GCV) approach.
summary(gcvplot(EVOGDP_2000_2005_2006~X+Y,data=data.1, scale=F,
#alpha=seq(0.02,0.1,by=0.01), # for NUTS2
alpha=seq(0.1,0.2,by=0.01), # for NUTS1
#alpha=seq(0.2,1,by=0.1), # for NUTS0
deg=1,kern="tricube",lfproc=locfit.raw))

# Choosing a LR fit with bandwidth chosen from above...
#bwd.2 <- 0.03 # for NUTS2
bwd.2 <- 0.2 # for NUTS1
#bwd.2 <- 0.6 # for NUTS0
lr <- locfit(EVOGDP_2000_2005_2006~X+Y,data=data.1, scale=F, alpha=bwd.2,
```

```
deg=1,kern="tricube",lfproc=locfit.raw)

# Raw residuals...
lr.p <- fitted.locfit(lr)
raw.resids.lr <- EVOGDP_2000_2005_2006-lr.p # Actual minus prediction
summary(raw.resids.lr)

# Identifying & updating outlier information in one file
cut.off.3 <- quantile(raw.resids.lr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails
indicator.6 <-ifelse(raw.resids.lr>=cut.off.3[2] & raw.resids.lr<=cut.off.3[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.lr, indicator.6)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for LR...
shades.6 = shading(c(cut.off.3[2],cut.off.3[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,20],shades.6)
title("Raw resids. from LR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")



# GWR...
# With coordinate data as explanatory variables (i.e. first-order polynomial).
# Using spgwr...

# Defining the coordinates...
coords.1<-cbind(data.1[,8],data.1[,9])

# Finding the bandwidth for GWR using Akaike Information Criterion (AIC) approach.
#gwr.aic.bwd <-gwr.sel(EVOGDP_2000_2005_2006~X+Y,data=data.1,coords=coords.1,adapt=TRUE,
#gweight=gwr.bisquare, method="aic")
#gwr.aic.bwd[1] # the optimum bandwidth

# Or finding the bandwidth for GWR using cross-validation approach.
gwr.cv.bwd <-gwr.sel(EVOGDP_2000_2005_2006~X+Y,data=data.1,coords=coords.1,adapt=TRUE,
gweight=gwr.bisquare, method="cv")
gwr.cv.bwd[1] # the optimum bandwidth

bwd.3 <- gwr.cv.bwd[1]
gwr.p <-gwr(EVOGDP_2000_2005_2006~X+Y,data=data.1,coords=coords.1,adapt=bwd.3,
gweight=gwr.bisquare,predictions=T)
#gwr.p$SDF

# GWR raw residuals...
raw.resids.gwr <- EVOGDP_2000_2005_2006-gwr.p$SDF$pred
summary(raw.resids.gwr)

# Identifying & updating outlier information in one file
cut.off.4 <- quantile(raw.resids.gwr, probs = seq(0, 1, 0.05), na.rm=T) # for 5% tails
indicator.7 <-ifelse(raw.resids.gwr>=cut.off.4[2] & raw.resids.gwr<=cut.off.4[20], 0, 1)
data1@data <- cbind(data1@data, raw.resids.gwr, indicator.7)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# Raw residual map for GWR...
```

```
shades.7 = shading(c(cut.off.4[2],cut.off.4[20]),c("red","white","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,22],shades.7)
title("Raw resids. from GWR: High/-ve(red) & High/+ve(black) (5% tails)")
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")









# 6.  All identified outliers together ##########################################

# Put all indicator data together...
indicator.8 <- indicator.1+indicator.2+indicator.3+indicator.4+indicator.5+indicator.6+indicator.7
summary(indicator.8)
# Histogram
X11(width=5.3,height=5.7)
hist(indicator.8,br=c(0,1,2,3,4,5,6,7))

# Thus a strong case for an outlier relates to an observation
# that has a indicator.8 value of 7...

data1@data <- cbind(data1@data, indicator.8)
data1@data <- as.data.frame(data1@data)
attach(data1@data)
data.1 <- data1@data
#fix(data.1)

# A choropleth map of suspected outliers...
shades.7 = shading(c(1,3,5,7),c("white","yellow","orange","red","dark red"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,data1@data[,24],shades.7)
title("Suspected outliers - weak to strong (yellow to dark red) evidence")
choro.legend(-2400000,2200000,shades.7,
over="exactly", between="to under",
fmt="%4.0f",title='Indicator sum (max.: 7)',cex=0.8)
map.scale(1800000,650000,1000000,"x1000 km",4,0.5)
north.arrow(1800000,1400000,80000, col="blue")
#text(1500000,2400000, "NUTS level 2", cex=2, col=3)
text(1500000,2400000, "NUTS level 1", cex=2, col=3)
#text(1500000,2400000, "NUTS level 0", cex=2, col=3)









# 7. The effects of MAUP ########################################################

# This relationship would be expected to weaken from NUTS2 to NUTS0
X11(width=5.3,height=5.7)
plot(jitter(NUTS3_outlier_max,factor=0.4), jitter(indicator.8,factor=0.4),
main="MAUP and its impact on outlier detection",
xlab="Strongest indication of an outlier for any constituent NUTS3 region",
#ylab="Indication of an outlier in a corres. aggreg. NUTS2 region", ylim=c(0,7),
```

```
ylab="Indication of an outlier in a corres. aggreg. NUTS1 region", ylim=c(0,7),
#ylab="Indication of an outlier in a corres. aggreg. NUTS0 region", ylim=c(0,7),
pch=19, cex=0.5)
abline(0,1)
cor(NUTS3_outlier_max, indicator.8)
```

TECHNICAL REPORT

ESPON

Spatial analysis for quality control

Phase 1:
The identification of logical input errors and statistical outliers

Draft: 2

MAIN RESULTS

- Exceptional values can arise from logical input errors and true outlying data.
- The accurate identification of an exceptional value is important as input errors should be treated differently to true outlying data.
- Input errors can usually be identified mathematically or sometimes, statistically. Outliers are identified statistically.
- A weighted evidence approach to statistically identify outliers is presented using a worked example that has been coded with R open source software.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

28 PAGES

Authors

National Centre for Geocomputation (NCG), National University of Ireland (Maynooth)

Paul Harris

Martin Charlton

Stewart Fotheringham

1 Introduction and aims

The ESPON 2013 Database should be as free from errors as possible. It follows from this that detecting errors is an important activity in both data entry and data checking. This technical report (which accompanies the final report) aims to clarify the results in the Second Interim Report (SIR) on how mathematical, statistical and spatial analysis tools can be applied to the ESPON 2013 Database in order to find 'logical input errors' (stage 1 detection) and 'statistical outliers' (stage 2 detection).

We first respond to queries or criticisms raised by the ESPON management on the SIR. Secondly (and in some respects, a response to criticisms), we propose a novel approach to the detection of statistical outliers (stage 2 detection), which we have developed since the SIR. As demonstration of this 'weight of evidence' approach, it is applied to a real (space-time) ESPON data set. The case study is presented with an R script so that the 'weight of evidence' approach is reproducible.

2 Responses to queries or criticisms

We shall simply respond to each point made in turn.

"Please also deliver the R-Scripts as a file." **Response** - This was already done, and the R script for this technical report has been similarly delivered.

"Chapter 4, however, lacks a comparison between the various techniques. Which techniques have shown the best results for these worked examples and are therefore the most promising?" **Response** - For statistical outliers, this is a very difficult task to do with any objectivity or certainty. Outliers in the ESPON database can arise for many different reasons and as such, require different approaches for their detection. Furthermore, the calibration of an outlier detection technique commonly involves many judged, user-specified model decisions, where in the end, it is common to refer to an unusual observation as simply *potentially* outlying (i.e. statistical detection cannot usually say that an observation is outlying with certainty). For these reasons and more, we have adopted a 'weight of evidence' approach to outlier detection, using many of the SIR techniques. The approach is still in its infancy and future work should consider various refinements (see section 3 for details). For the detection of logical input errors (stage 1 detection), we recommend that their methods of detection are simply updated as they arise. It is impossible to foresee the infinite number of ways that a logical input error can occur. We have provided a start to this issue in the SIR and here, the metadata for each variable should provide a helpful basis for this detection.

"For ESPON it is also important to know what exactly the benefit is for ESPON of the research done on improving these techniques and procedures. The benefit in general is clear: the detection of outliers and errors can be done easier, faster and better. However, the report does not show (yet) what has been detected in the datasets that are/will be incorporated in the ESPON 2013 Database. The usefulness of this research for ESPON should be made more clear." **Response** - We demonstrate our 'weight of evidence' approach to a real case study dataset in section 3, where observations are flagged as outlying together with the nature of their 'outlyingness' (i.e. spatial, temporal, etc.). The approach can easily be embedded within the database architecture. Many of the detection techniques from the SIR have already been incorporated in this respect, some of which are key components of the more coherent 'weight of evidence' approach.

"Furthermore, detecting is one step, solving the problem is the next. Are the outliers and errors detected real errors? Can they be solved? And if so, how are they solved. It would be

good if this could be shown in the next version of the report." **Response** - Observations can only be flagged as *potentially* outlying, but using a 'weight of evidence' approach allows some observations to have a higher *potential* in this respect, than others. Observations that are flagged as outlying from many detection techniques should be viewed with the strongest suspicion. The decision of what to do with an outlier (e.g. remove, replace, etc.) should ultimately reside with some expert on the given data set (or its provider). Mechanisms can be put in place within the database architecture to do this in an efficient and effective manner.

3 Case study: the 'weight of evidence' approach

In this case study, we apply our 'weight of evidence' approach to detect outliers in an ESPON space-time, *univariate* data set (where an extension to *multivariate* case is simple and direct). This approach applies nine representative and complementary statistical outlier detection techniques, where observations are flagged as outlying according the outcome of each technique. This then builds up a 'weight of evidence' for the likelihood of a given observation being outlying (i.e. statistical evidence is strongest for an observation that is flagged as outlying for all nine techniques).

According to our chosen techniques, we can indicate the nature of the outliers. Outliers can have any combination of aspatial, spatial, temporal or relational characteristics (see Fig. 1). Furthermore, this methodology can be specified such that a different weighting is attached to each technique. For example, we may want to attach more weight to observations that are outlying in a temporal sense. The detection techniques are all non-robust in form. However robust forms are preferred, as outliers can compromise the calibration of a given technique prior to its use as a method of detection. Extensions to robust forms, for each technique, should be the first set of refinements that are considered with this approach.

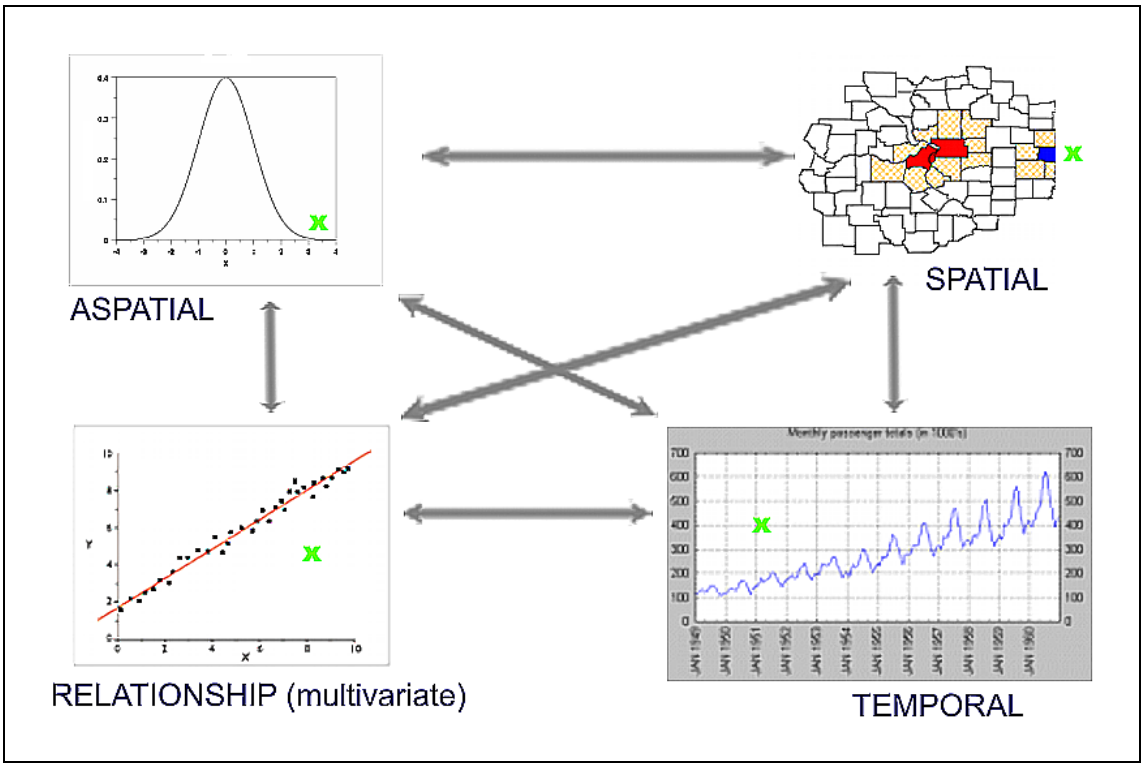


Figure 1 Types of outliers detected.

The nine techniques are: (i) boxplot statistics (for aspatial, univariate outliers); (ii) Hawkins' statistic (for spatial, univariate outliers); (iii) time series statistics (for temporal, univariate outliers); (iv) regression (for outliers from an aspatial, linear relationship); (v) locally weighted regression (for outliers from an aspatial, nonlinear relationship); (vi) geographically weighted regression (for outliers from a spatial, nonlinear relationship); (vii) principal component analysis (PCA, for outliers from a 'model-free', aspatial, linear relationship); (viii) locally weighted PCA (for outliers from a 'model-free', aspatial, nonlinear relationship); and (ix) geographically weighted PCA (for outliers from a 'model-free', spatial, nonlinear relationship). Here the PCA methods do not require domains-specific knowledge (i.e. how to choose covariates in a regression) and as such, are termed 'model-free'.

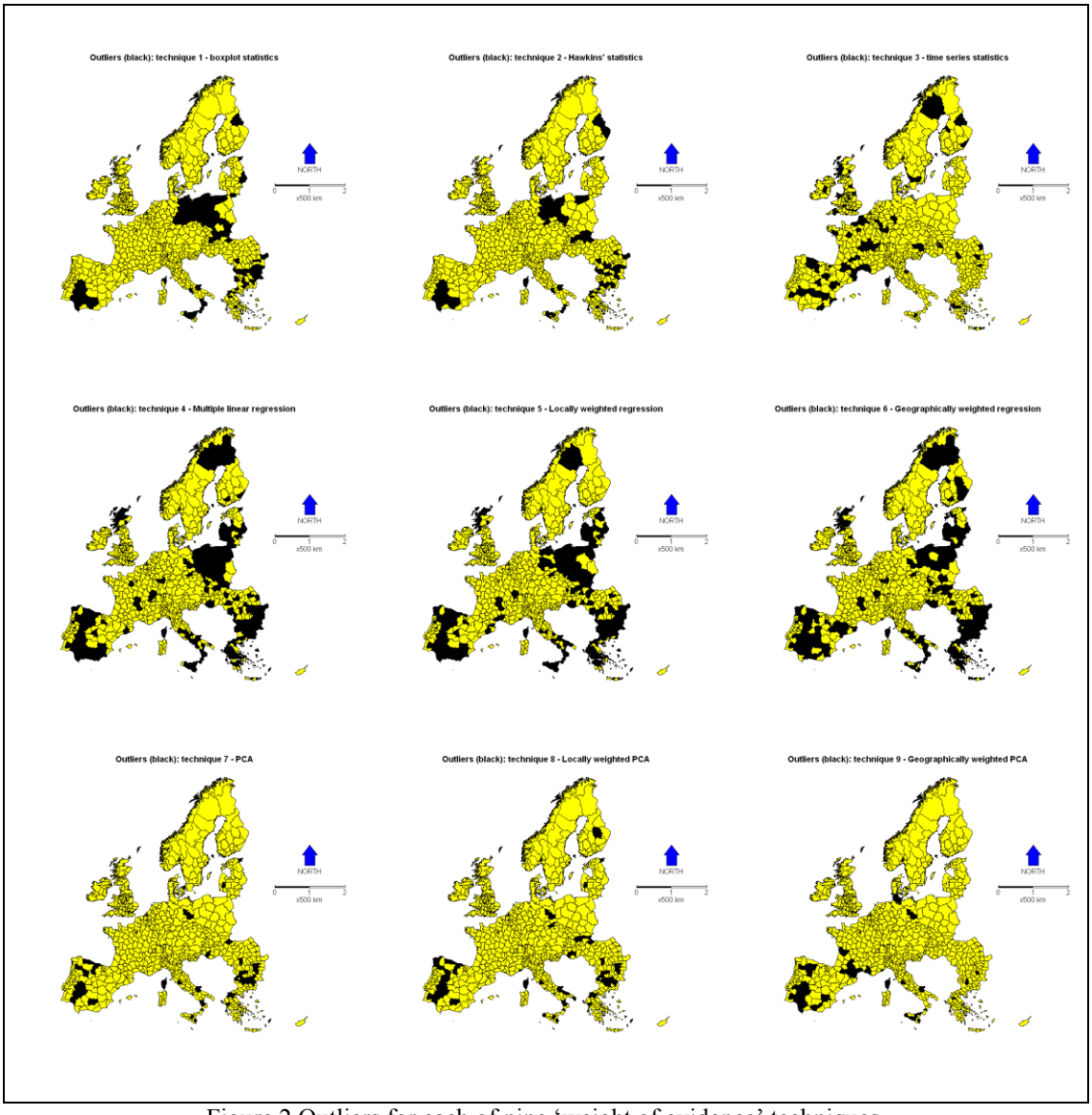


Figure 2 Outliers for each of nine 'weight of evidence' techniques.

We demonstrate our 'weight of evidence' outlier detection approach to ESPON unemployment rate data covering 709 NUTS23 regions for eight consecutive years 2000 to 2007. Here we show that outlier detection can not only be used as a data cleaning or screening exercise, but can also be used to uncover interesting or unusual relationships in the

process that has not been considered before. Observe also that the regression methods can also be used for the prediction of missing data.

Full details of the outlier detection approach are given in the associated R script (see Appendix). Fig. 2 presents the results (maps) for each of the nine detection techniques and Fig.3 presents the summary of the results (i.e. the final 'weight of evidence' map). Thus as examples, there is strong evidence of at least one outlying unemployment rate (from the eight years) in a NUTS23 region in SW Spain and a NUTS23 region in N Corsica. Conversely, there is little evidence of at least one outlying unemployment rate (from the eight years) in a NUTS23 region in SW Ireland and all NUTS23 regions in Norway.

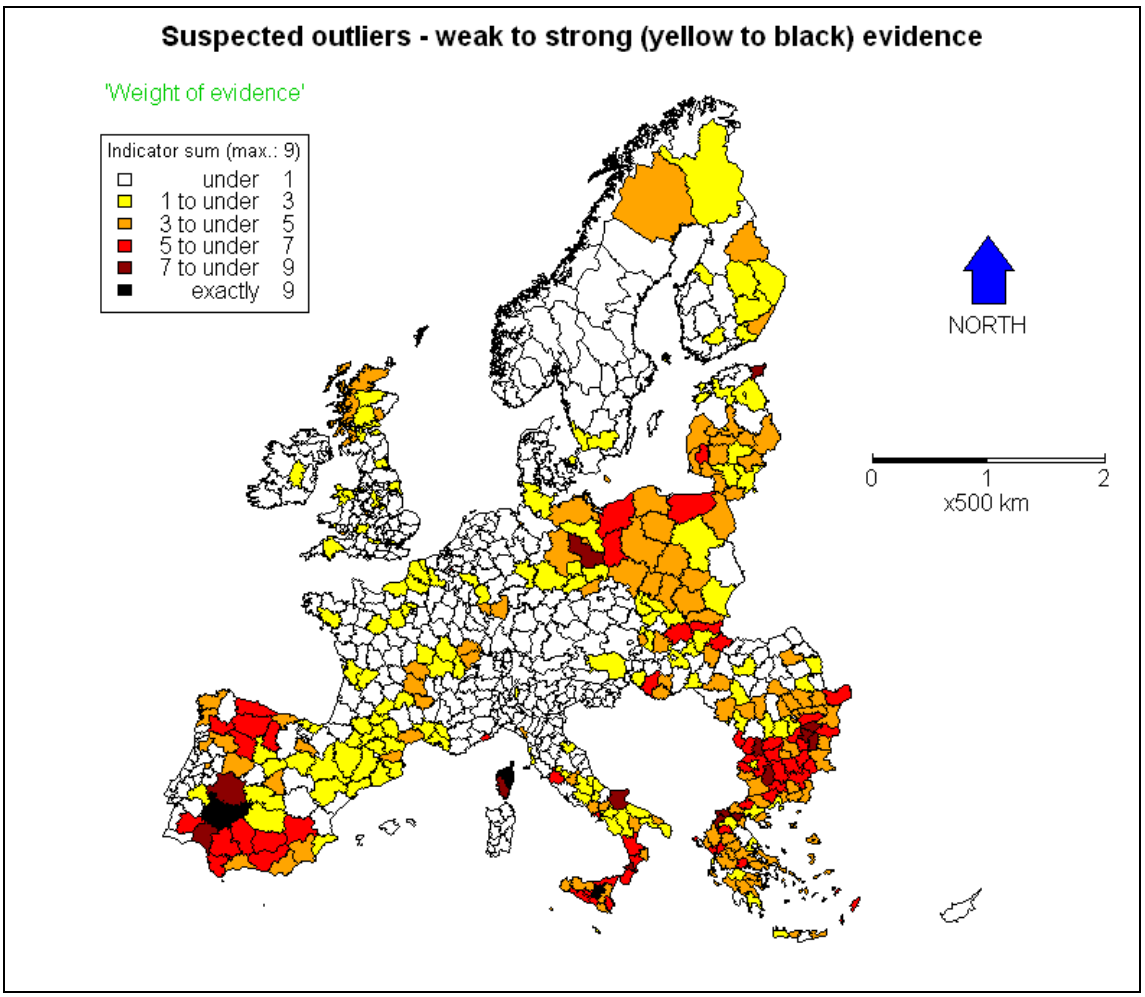


Figure 3 Final 'weight of evidence' map.

Reading

Fotheringham AS, Brunsdon C, Charlton ME (2002) Geographically Weighted Regression - the analysis of spatially varying relationships. Wiley, Chichester.

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5, 299-314.

Joliffe IT (2002) Principal Components Analysis, 2nd edition, New York: Springer-Verlag

Loader C (2004) Smoothing: Local Regression Techniques. In Gentle J, Härdle W, Mori Y (eds) Handbook of Computational Statistics. Springer-Verlag, Heidelberg.

Appendix: Worked example of the 'weight of evidence' approach

```
#    This program is free software: you can redistribute it and/or modify
#    it under the terms of the GNU General Public License as published by
#    the Free Software Foundation, either version 3 of the License, or
#    (at your option) any later version.

#    This program is distributed in the hope that it will be useful,
#    but WITHOUT ANY WARRANTY; without even the implied warranty of
#    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#    GNU General Public License for more details.

#    You should have received a copy of the GNU General Public License
#    along with this program.  If not, see <http://www.gnu.org/licenses/>.


# Challenge 10 - ESPON 2013 database

# P. Harris (NCG), M. Charlton (NCG) & C. Brunsdon (University of Leicester, UK)

# 24/10/10

# 'Weight of evidence' approach for detecting statistical outliers.

# There is no single 'best' outlier detection technique, so our aim is to:
# a. Apply a representative selection of outlier detection techniques.
# b. Flag an observation if it is a likely outlier according to each technique.
# c. Build up a weight of evidence for the likelihood of a given observation being statistically outlying.
# d. Suggest what type of outlier it is likely to be
#        - aspatial, spatial, temporal, relationship, some mixture…
# e. Consult an expert on the data to decide on the appropriate course of action.



# In this case, we employ 9 complementary and representative outlier detection techniques:

# 1. Boxplots
# 2. Hawkins' spatial statistic
# 3. Time series statistics
# 4. MLR residuals
# 5. LWR residuals
# 6. GWR residuals
# 7. PCA 'residuals'
# 8. LWPCA 'residuals'
# 9. GWPCA 'residuals'

# Techniques 1,2,4,5,6 and 7 were all investigated in the SIR.
```

```
# This script is for the weight of evidence approach in its most basic form.
# All outlier detection techniques are non-robust in design.
# Further work should replace each technique with a robust version
# to give a robust weight of evidence approach.

# Currently robust forms are readily available in R for:

# 1. Boxplots - robust form - adjusted boxplot - robustbase package
# 2. Hawkins' spatial statistic - robust form needs development/coding
# 3. Time series statistics - robust form needs development/coding
# 4. MLR residuals - robust form - many choices - e.g. robustbase package
# 5. LWR residuals  - robust form - many choices - locfit package
# 6. GWR residuals - robust form needs development/coding
# 7. PCA 'residuals' - robust form - many choices - e.g. mvoutlier package
# 8. LWPCA 'residuals' - robust form needs development/coding
# 9. GWPCA 'residuals' - robust form needs development/coding



# This particular script is also for the univariate space-time case but
# techniques are directly extendable to the multivariate space-time case.

# In 7 cases - outliers are chosen according to standard boxplot statistics.
# We could have specified adjusted boxplots statistics here instead
# but this can be part of any robust weight of evidence approach (future work, see above).

# Furthermore, in all cases - statistical tests could be used instead of boxplots.
# As it stands - this is only done in 2 cases.



# This weight of evidence approach can be performance tested using data that have been infected with outliers.
# However results can be difficult to interpret since:
# 1. It is difficult to guarentee that the infections actually produce outliers.
# 2. The data may already contain outliers.

# Further work should explore the use of simulated data sets
# or more attention should be paid to the theoretical properties of each outlier detection technique.



# Currently - this basic weight of evidence approach gives equal weights to each of the 9 detection techniques.
# There is no reason why these weights cannot vary...

# For example, these weights may work better:
# 1. Boxplots - 1
# 2. Hawkins - 1
# 3. Time series - 1
```

```
# 4. MLR - 1/3
# 5. LWR - 1/3
# 6. GWR - 1/3
# 7. PCA - 1/3
# 8. LWPCA - 1/3
# 9. GWPCA - 1/3



# The case study data set is unemployment rate ( = unemployment population/active population)
# at the NUTS 23 level covering 8 years: 2000-2007.
# Thus there are 8 variables at each of 790 regions = 6320 observations in total.

# We assume that the data has been checked for logical input errors (i.e. stage 1 detection is done).

# For brevity, we also assume that we only need at least one of eight time-specific unemployment rates
# in a region to be outlying.
# However in all techniques, execpt PCA/LWPCA/GWPCA, we actually identify outliers by year.


# An earlier version of output from this basic script (with clean and infected data) can be found at:
# http://www.espon.eu/main/Menu_Events/Menu_OpenSeminars/openseminar100609After.html
# This is the presenatation given at the Madrid open seminar 2010.


# Observe that outlier detection can reveal the most interesting results
# i.e. - those not expected.

# Observe also that the regression outlier detection methods can be used for the prediction of missing data too.


# R packages needed.....
# 1. GISTools (version 0.5-4) and all its dependencies (see SIR)
# 2. locfit (version 1.5-4)- for LWR



##############################################################################
# A. Functions needed for PCA, LWPCA & GWPCA ##################################
##############################################################################


# Author: Chris Brunsdon (plus some edits by Paul Harris)

# Details can be found in forthcoming articles

# Sub-functions....

wpca <- function(x,wt,...) {
        local.center <- function(x,wt)
```

```
                sweep(x,2,colSums(sweep(x,1,wt,'*'))/sum(wt))
        svd(sweep(local.center(x,wt),1,sqrt(wt),'*'),...)}

bw.by.nn <- function(loc,nn,eloc=loc) {
        result <- numeric(nrow(eloc))
        int.part <- floor(nn)
        frac.part <- nn - int.part
        if (frac.part == 0) {
                for (i in 1:nrow(eloc)) {
                        d.sqr <- rowSums(sweep(loc,2,eloc[i,])**2)
                        result[i] <- sort(d.sqr,partial=nn)[nn]}}
        else {
                for (i in 1:nrow(eloc)) {
                        d.sqr <- rowSums(sweep(loc,2,eloc[i,])**2)
                        d.tmp <- sort(d.sqr,partial=c(nn,nn+1))[c(nn,nn+1)]
                        result[i] <- d.tmp[1]*(1-frac.part) + d.tmp[2]*frac.part}}
        sqrt(result)}

bw.by.nn.1 <- function(x,nn,ex=x) {
        result <- numeric(nrow(ex))
        int.part <- floor(nn)
        frac.part <- nn - int.part
        if (frac.part == 0) {
                for (i in 1:nrow(ex)) {
                        d.sqr <- rowSums(sweep(x,2,ex[i,])**2)
                        result[i] <- sort(d.sqr,partial=nn)[nn]}}
        else {
                for (i in 1:nrow(ex)) {
                        d.sqr <- rowSums(sweep(x,2,ex[i,])**2)
                        d.tmp <- sort(d.sqr,partial=c(nn,nn+1))[c(nn,nn+1)]
                        result[i] <- d.tmp[1]*(1-frac.part) + d.tmp[2]*frac.part}}
        sqrt(result)}



# Basic functions....

lwpca <- function(x,bw,k=2,ex=x,pcafun=wpca,...) {
        n <- nrow(ex)
        m <- ncol(x)
        w <- array(0,c(n,m,k))
        d <- matrix(0,n,k)
        bw <- bw*bw
        for (i in 1:n) {
                wt <- rowSums(sweep(x,2,ex[i,])**2)/bw
                use <- wt < 1
                if (sum(use) < 5)
                        stop(paste('You will need a larger bandwidth: location ',i,
                        ' has only ',sum(use),'neigbours'))
                wt <- (1 - wt[use])**2
```

```
                temp <- pcafun(x[use,],wt,nu=0,nv=k)
                w[i,,] <- temp$v
                d[i,] <- temp$d[1:k]}
        if (!is.null(rownames(x))) dimnames(w)[[1]] <- rownames(x)
        if (!is.null(colnames(x))) dimnames(w)[[2]] <- colnames(x)
        dimnames(w)[[3]] <- paste("PC",1:k,sep='')
        list(loadings=w,var=d,bw=sqrt(bw))}

gwpca <- function(x,loc,bw,k=2,eloc=loc,pcafun=wpca,...) {
        n <- nrow(eloc)
        m <- ncol(x)
        w <- array(0,c(n,m,k))
        d <- matrix(0,n,k)
        bw <- bw*bw
        for (i in 1:n) {
                wt <- rowSums(sweep(loc,2,eloc[i,])**2)/bw
                use <- wt < 1
                if (sum(use) < 5)
                        stop(paste('You will need a larger bandwidth: location ',i,
                        ' has only ',sum(use),'neigbours'))
                wt <- (1 - wt[use])**2
                temp <- pcafun(x[use,],wt,nu=0,nv=k)
                w[i,,] <- temp$v
                d[i,] <- temp$d[1:k]}
        if (!is.null(rownames(x))) dimnames(w)[[1]] <- rownames(x)
        if (!is.null(colnames(x))) dimnames(w)[[2]] <- colnames(x)
        dimnames(w)[[3]] <- paste("PC",1:k,sep='')
        list(loadings=w,var=d,bw=sqrt(bw))}



# Automatic bandwidth functions...

lwpca.autobw <- function(x,bw.interval,k=2,ex=x,verbose=FALSE,...) {
        bw <- optimise(function(z) lwpca.cv(x,z,k=k,...),bw.interval)
        if (verbose) {
                cat("'Optimal' bandwidth = ",bw$minimum)
                cat(" Score is ",bw$objective)}
        lwpca(x,bw$minimum,k=k,ex=ex,...)}

lwpca.autobw.by.nn <- function(x,nn.interval,k=2,ex=x,verbose=FALSE,...) {
        bw <- optimise(function(z) lwpca.cv(x,bw.by.nn.1(x,z),k=k,...),nn.interval)
        if (verbose) {
                cat("'Optimal' nn = ",bw$minimum)
                cat(" Score is ",bw$objective)}
        lwpca(x,bw.by.nn(x,bw$minimum),k=k,ex=ex,...)}

gwpca.autobw <- function(x,loc,bw.interval,k=2,eloc=loc,verbose=FALSE,...) {
        bw <- optimise(function(z) gwpca.cv(x,loc,z,k=k,...),bw.interval)
        if (verbose) {
```

```
                cat("'Optimal' bandwidth = ",bw$minimum)
                cat(" Score is ",bw$objective)}
        gwpca(x,loc,bw$minimum,k=k,eloc=eloc,...)}

gwpca.autobw.by.nn <- function(x,loc,nn.interval,k=2,eloc=loc,verbose=FALSE,...) {
        bw <- optimise(function(z) gwpca.cv(x,loc,bw.by.nn(loc,z),k=k,...),nn.interval)
        if (verbose) {
                cat("'Optimal' nn = ",bw$minimum)
                cat(" Score is ",bw$objective)}
        gwpca(x,loc,bw.by.nn(loc,bw$minimum),k=k,eloc=eloc,...)}



# Cross-validation functions...

lwpca.cv <- function(x,bw,k=2,...) {
        n <- nrow(x)
        m <- ncol(x)
        w <- array(0,c(n,m,k))
        bw <- bw*bw
        score <- 0
        for (i in 1:n) {
                wt <- rowSums(sweep(x,2,x[i,])**2)/bw
                wt[i] <- Inf
                use <- wt < 1
                wt <- (1 - wt[use])**2
                v <- wpca(x[use,],wt,nu=0,nv=k)$v
                v <- v %*% t(v)
                score <- score + sum((x[i,] - x[i,] %*% v))**2}
        score}

gwpca.cv <- function(x,loc,bw,k=2,...) {
        n <- nrow(loc)
        m <- ncol(x)
        w <- array(0,c(n,m,k))
        bw <- bw*bw
        score <- 0
        for (i in 1:n) {
                wt <- rowSums(sweep(loc,2,loc[i,])**2)/bw
                wt[i] <- Inf
                use <- wt < 1
                wt <- (1 - wt[use])**2
                v <- wpca(x[use,],wt,nu=0,nv=k)$v
                v <- v %*% t(v)
                score <- score + sum((x[i,] - x[i,] %*% v))**2}
        score}



# Outlier/residual functions...
```

```
lwpca.cv.contrib <- function(x,bw,k=2,...) {
        n <- nrow(x)
        m <- ncol(x)
        w <- array(0,c(n,m,k))
        bw <- bw*bw
        score <- numeric(n)
        for (i in 1:n) {
                wt <- rowSums(sweep(x,2,x[i,])**2)/bw
                wt[i] <- Inf
                use <- wt < 1
                wt <- (1 - wt[use])**2
                v <- wpca(x[use,],wt,nu=0,nv=k)$v
                v <- v %*% t(v)
                score[i] <- sum((x[i,] - x[i,] %*% v))**2}
        score}

gwpca.cv.contrib <- function(x,loc,bw,k=2,...) {
        n <- nrow(loc)
        m <- ncol(x)
        w <- array(0,c(n,m,k))
        bw <- bw*bw
        score <- numeric(n)
        for (i in 1:n) {
                wt <- rowSums(sweep(loc,2,loc[i,])**2)/bw
                wt[i] <- Inf
                use <- wt < 1
                wt <- (1 - wt[use])**2
                v <- wpca(x[use,],wt,nu=0,nv=k)$v
                v <- v %*% t(v)
                score[i] <- sum((x[i,] - x[i,] %*% v))**2}
        score}



################################################################################
# B. Data input and exploratory data analysis ##################################
################################################################################



# Read in a data
data0 <- read.table("Worked example data.csv", ",", header=T)
colnames(data0) # Note the first 8 columns relate to an experiment with infected data - ignore this
attach(data0)



# Variables of interest are:
# UNEMP.R1.2000, UNEMP.R1.2001, UNEMP.R1.2002, UNEMP.R1.2003, UNEMP.R1.2004, UNEMP.R1.2005,
# UNEMP.R1.2006 and UNEMP.R1.2007
```

```
# Distributions...
X11(width=12,height=7)
par(mfrow=c(2,4))
hist(UNEMP.R1.2000, freq=F)
hist(UNEMP.R1.2001, freq=F)
hist(UNEMP.R1.2002, freq=F)
hist(UNEMP.R1.2003, freq=F)
hist(UNEMP.R1.2004, freq=F)
hist(UNEMP.R1.2005, freq=F)
hist(UNEMP.R1.2006, freq=F)
hist(UNEMP.R1.2007, freq=F)

X11(width=12,height=7)
par(mfrow=c(2,4))
hist(log(UNEMP.R1.2000), freq=F)
hist(log(UNEMP.R1.2001), freq=F)
hist(log(UNEMP.R1.2002), freq=F)
hist(log(UNEMP.R1.2003), freq=F)
hist(log(UNEMP.R1.2004), freq=F)
hist(log(UNEMP.R1.2005), freq=F)
hist(log(UNEMP.R1.2006), freq=F)
hist(log(UNEMP.R1.2007), freq=F)



# Correlations...
cor(data0[9:16])
X11(width=18,height=12)
pairs(data0[9:16],cex=0.5)



# Importing data as a ArcGIS shapefile & using GISTools to do some maps
require(GISTools)
# Ignore all warnings - this code is under development...

# Read in the shapefile...
data1 <- readShapePoly("FinalTS23.shp",
proj4string=CRS("+proj=laea+lat_0=50.0+lon_0=15.0+x_0=0+y_0=0"))
colnames(data1@data)

# Renaming each variable - as they have been truncated in ArcGIS...
colnames(data1@data) <- c("NUTS23x","NUTS23Ax",
"ACTIVE_2000x","ACTIVE_2001x","ACTIVE_2002x","ACTIVE_2003x","ACTIVE_2004x","ACTIVE_2005x","ACTIVE_2006x",
"ACTIVE_2007x",
"UNEMP_2000x","UNEMP_2001x","UNEMP_2002x","UNEMP_2003x","UNEMP_2004x","UNEMP_2005x","UNEMP_2006x",
"UNEMP_2007x",
"Xx","Yx")
```

```
# Checking the two data sets match up OK...
length(data0[,1])
length(data1@data[,1])
data.x <- cbind(data1@data$NUTS23x, NUTS23, data1@data$Xx, X, data1@data$Yx, Y)

# Coordinates....
Nuts23Coords <- as.matrix(cbind(X/1000, Y/1000))
colnames(Nuts23Coords) <- c("Easting","Northing")

# Maximum distances of sampled area in main directions...
max(X/1000)-min(X/1000)
max(Y/1000)-min(Y/1000)

# Renaming data...
data1@data <- data0
attach(data1@data)

# Create another data matrix for the analysis
#Data.1.scaled <- scale(as.matrix(data1@data[9:16])) # scaled
Data.1.scaled <- as.matrix(data1@data[9:16]) # not scaled
rownames(Data.1.scaled) <- data1@data[,17]
n1 <- length(Data.1.scaled[,1])

# An example choropleth map...
shades.e = auto.shading(UNEMP.R1.2000,5, cols=brewer.pal(5,'PuBuGn'))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,UNEMP.R1.2000,shades.e)
title("Unemployment rate: year 2000")
choro.legend(1300000,400000,shades.e,fmt="%4.2f",title='Rate',cex=0.8)
map.scale(1800000,-650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")



################################################################################
# 1. Standard boxplot outliers ################################################
################################################################################


# Need to do this for each unemployment variable (i.e. each year) ...

z1 <- data0[,9]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.1 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1) # i.e. suspected outliers...

z1 <- data0[,10]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.2 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)
```

```
z1 <- data0[,11]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.3 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)

z1 <- data0[,12]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.4 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)

z1 <- data0[,13]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.5 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)

z1 <- data0[,14]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.6 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)

z1 <- data0[,15]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.7 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)

z1 <- data0[,16]
bp <- boxplot.stats(z1, coef=1.5)
indicator.1.8 <-ifelse(z1>bp$stats[1]& z1<bp$stats[5], 0, 1)

# This indictor is one value out of eight is an outlier at a given location
indicator.1 <- indicator.1.1+indicator.1.2+indicator.1.3+indicator.1.4+indicator.1.5+indicator.1.6+indicator.1.7+indicator.1.8


# A choropleth map of the boxplot outliers...
shades.1 = shading(c(0,1,9),c("white","yellow","black")) # i.e. yellow - no & black - yes
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.1,shades.1)
title("Outliers (black): technique 1 - boxplot statistics")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")



################################################################################
# 2. Outliers via Hawkins' statistic - using chi-squared ######################
################################################################################



# GW means and GW variances for Hawkin's statistic found
# using the gw.cov function in spgwr with the default bi-square weighting scheme.

# Bandwdth is user-specifed...
```

```
bwd.1 <- 0.1 # 10% of nearby data

# Coordinates....
coordinates(data0) <- c("X", "Y") # note do not need to divide by a 1000

# Critical values of the chi-squared distribution
chi_10 <- 2.70554
chi_5 <- 3.84146
chi_2.5 <- 5.02389
chi_1 <- 6.63490
chi_0.5 <- 7.87944
chi_0.01 <- 10.828

# Hawkins test for all 8 unemloyment variables (or years)

gwss <- gw.cov(data0, vars=9, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2000-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.1 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...

gwss <- gw.cov(data0, vars=10, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2001-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.2 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...

gwss <- gw.cov(data0, vars=11, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2002-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.3 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...

gwss <- gw.cov(data0, vars=12, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2003-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.4 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...
```

```
gwss <- gw.cov(data0, vars=13, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2004-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.5 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...

gwss <- gw.cov(data0, vars=14, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2005-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.6 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...

gwss <- gw.cov(data0, vars=15, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2006-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.7 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...

gwss <- gw.cov(data0, vars=16, adapt=bwd.1)
GW.mean <- gwss$SDF$mean.V1
GW.variance <- (gwss$SDF$sd.V1)^2
Hawk.N <- bwd.1*length(X) # number of neighbouring data
Hawk.lm <- GW.mean # the local mean at observation points
Hawk.alv <- mean(GW.variance) # the average local variance with same bandwidth
Hawk.test <- (Hawk.N*(UNEMP.R1.2007-Hawk.lm)^2)/((Hawk.N+1)*Hawk.alv)  # test statistic
indicator.2.8 <-ifelse(Hawk.test <=chi_5, 0, 1) # change critical level accordingly...


# This indictor is one value out of eight is an outlier at a given location
indicator.2 <- indicator.2.1+indicator.2.2+indicator.2.3+indicator.2.4+indicator.2.5+indicator.2.6+indicator.2.7+indicator.2.8


# A choropleth map of Hawkins' test results...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.2,shades.1)
title("Outliers (black): technique 2 - Hawkins' statistics")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")
```

```
################################################################################
# 3. Time series outliers - using chi-squared ##################################
################################################################################



# Simplified version of Claude Grasland's time series outlier identification
# In this case for one, not multiple spatial units...

Data.1.ts <- data1@data[9:16]
row.means <- apply(Data.1.ts,1,mean)
row.vars <- apply(Data.1.ts,1,var)
row.sds <- row.vars^0.5
op.1 <- sweep(Data.1.ts,1,row.means)
resids.ts <- (op.1)^2/row.vars
# Note the use of variances - thus we do a chi-square test here instead of looking at boxplot statistics
#sort(resids.3)
quant.ts <- quantile(resids.ts, probs = seq(0, 1, 0.01))
threshold.ts <- quant.ts[100] # in this case the 99%tile

# Assume Gaussianity and do a Chi-sqaured test...
threshold.tsx <- 3.84146 # chi-square test at 95% ...
indicator.tsx <-ifelse(resids.ts>0& resids.ts<threshold.tsx, 0, 1) # i.e. suspected outliers...
indicator.3 <- apply(indicator.tsx,1,sum)

# A choropleth map of the time series outliers
shades.2 = shading(c(0,1,2),c("white","yellow","black")) # i.e. yellow - no & black - yes
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.3,shades.2)
title("Outliers (black): technique 3 - time series statistics")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")



################################################################################
# 4. Outliers via MLR  - via boxplots ##########################################
################################################################################



# In all regressions, we use unemployment from the two closest years to explain unemployment for a given year of interest.

# Note using Boxplot statistics for residuals

# Need to do this eight times...
# one for each unemployment variable & changing the independent variables accordingly

mlr <- lm(UNEMP.R1.2000 ~ UNEMP.R1.2001+UNEMP.R1.2002)
```

```
raw.resids.mlr <- UNEMP.R1.2000-mlr$fitted # Actual minus prediction (raw residuals)
# Identifying & updating outlier information in one file using Boxplot statistics for residuals
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.1 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2001 ~ UNEMP.R1.2000+UNEMP.R1.2002)
raw.resids.mlr <- UNEMP.R1.2001-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.2 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2002 ~ UNEMP.R1.2001+UNEMP.R1.2003)
raw.resids.mlr <- UNEMP.R1.2002-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.3 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2003 ~ UNEMP.R1.2002+UNEMP.R1.2004)
raw.resids.mlr <- UNEMP.R1.2003-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.4 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2004 ~ UNEMP.R1.2003+UNEMP.R1.2005)
raw.resids.mlr <- UNEMP.R1.2004-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.5 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2005 ~ UNEMP.R1.2004+UNEMP.R1.2006)
raw.resids.mlr <- UNEMP.R1.2005-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.6 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2006 ~ UNEMP.R1.2005+UNEMP.R1.2007)
raw.resids.mlr <- UNEMP.R1.2006-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.7 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

mlr <- lm(UNEMP.R1.2007 ~ UNEMP.R1.2005+UNEMP.R1.2006)
raw.resids.mlr <- UNEMP.R1.2007-mlr$fitted
bp <- boxplot.stats(raw.resids.mlr, coef=1.5)
indicator.4.8 <-ifelse(raw.resids.mlr>bp$stats[1]& raw.resids.mlr<bp$stats[5], 0, 1)

# This indictor is one value out of eight is an outlier at a given location
indicator.4 <- indicator.4.1+indicator.4.2+indicator.4.3+indicator.4.4+indicator.4.5+indicator.4.6+indicator.4.7+indicator.4.8

# A choropleth map of MLR outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.4,shades.1)
title("Outliers (black): technique 4 - Multiple linear regression")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")
```

```
##############################################################################
# 5. Outliers via LWR  - via boxplots ########################################
##############################################################################



# Using locfit...
require(locfit)
# Ignore warning message...

# In this case - we only use the closest year to the dependent variable, as too many variables can create problems...
# Automatic bandwidth selection also was problematic in some cases...

# Automatic bandwidth for a non-robust LR (i.e. not a lowess fit) using generalised cross-validation approach.
#summary(gcvplot(UNEMP.R1.2000 ~ UNEMP.R1.2001+UNEMP.R1.2002,data=data0, scale=F,alpha=seq(0.1,1,by=0.1),
#deg=1,kern="tricube",lfproc=locfit.raw))

# As such - the same user-specified bandwidth is used for all eight fits
bwd.2 <- 0.4

# The eight LWR model fits...

lwr <- locfit(UNEMP.R1.2000 ~ UNEMP.R1.2001,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2000-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.1 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2001 ~ UNEMP.R1.2002,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2001-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.2 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2002 ~ UNEMP.R1.2003,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2002-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.3 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2003 ~ UNEMP.R1.2004,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2003-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
```

```
indicator.5.4 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2004 ~ UNEMP.R1.2005,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2004-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.5 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2005 ~ UNEMP.R1.2006,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2005-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.6 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2006 ~ UNEMP.R1.2007,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2006-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.7 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)

lwr <- locfit(UNEMP.R1.2007 ~ UNEMP.R1.2006,data=data0, scale=F, alpha=bwd.2,
deg=1,kern="tricube",lfproc=locfit.raw)
lwr.p <- fitted.locfit(lwr)
raw.resids.lr <- UNEMP.R1.2007-lwr.p # Actual minus prediction
bp <- boxplot.stats(raw.resids.lr, coef=1.5)
indicator.5.8 <-ifelse(raw.resids.lr>bp$stats[1]& raw.resids.lr<bp$stats[5], 0, 1)


# This indictor is one value out of eight is an outlier at a given location
indicator.5 <- indicator.5.1+indicator.5.2+indicator.5.3+indicator.5.4+indicator.5.5+indicator.5.6+indicator.5.7+indicator.5.8

# A choropleth map of the LWR outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.5,shades.1)
title("Outliers (black): technique 5 - Locally weighted regression")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")


################################################################################
# 6. Outliers via GWR  - via boxplots ##########################################
################################################################################

# Using spgwr...

# As with MLR - we again use unemployment from the two closest years to explain unemployment for a given year of interest.
```

```
# In this case - all eight GWR fits have an optimally found bandwidth (by cross-validation)

# The eight GWR model fits...

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2000 ~ UNEMP.R1.2001+UNEMP.R1.2002,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2000 ~ UNEMP.R1.2001+UNEMP.R1.2002,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2000-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.1 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2001 ~ UNEMP.R1.2000+UNEMP.R1.2002,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2001 ~ UNEMP.R1.2000+UNEMP.R1.2002,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2001-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.2 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2002 ~ UNEMP.R1.2001+UNEMP.R1.2003,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2002 ~ UNEMP.R1.2001+UNEMP.R1.2003,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2002-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.3 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2003 ~ UNEMP.R1.2002+UNEMP.R1.2004,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2003 ~ UNEMP.R1.2002+UNEMP.R1.2004,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2003-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.4 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2004 ~ UNEMP.R1.2003+UNEMP.R1.2005,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2004 ~ UNEMP.R1.2003+UNEMP.R1.2005,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2004-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.5 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2005 ~ UNEMP.R1.2004+UNEMP.R1.2006,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2005 ~ UNEMP.R1.2004+UNEMP.R1.2006,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2005-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
```

```
indicator.6.6 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2006 ~ UNEMP.R1.2005+UNEMP.R1.2007,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2006 ~ UNEMP.R1.2005+UNEMP.R1.2007,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2006-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.7 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)

gwr.cv.bwd <-gwr.sel(UNEMP.R1.2007 ~ UNEMP.R1.2005+UNEMP.R1.2006,
data=data0,adapt=TRUE,gweight=gwr.bisquare, method="cv")
gwr.p <- gwr(UNEMP.R1.2007 ~ UNEMP.R1.2005+UNEMP.R1.2006,data=data0,adapt=gwr.cv.bwd[1],
gweight=gwr.bisquare,predictions=T)
raw.resids.gwr <- UNEMP.R1.2007-gwr.p$SDF$pred
bp <- boxplot.stats(raw.resids.gwr, coef=1.5)
indicator.6.8 <-ifelse(raw.resids.gwr>bp$stats[1]& raw.resids.gwr<bp$stats[5], 0, 1)


# This indictor is one value out of eight is an outlier at a given location
indicator.6 <- indicator.6.1+indicator.6.2+indicator.6.3+indicator.6.4+indicator.6.5+indicator.6.6+indicator.6.7+indicator.6.8

# A choropleth map of GWR outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.6,shades.1)
title("Outliers (black): technique 6 - Geographically weighted regression")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")



################################################################################
# 7. PCA outliers  - via boxplots ##############################################
################################################################################



# For PCA, LWPCA and GWPCA we need to define k the number of components to 'keep'....
# In this case 2 or 3 seems ideal - can check this by cross-validation...
# This part of the PCA/LWPCA/GWPCA approach to outlier detection needs further work...

# Lets choose k=3...
comp.keep <- 3

# Global PCA result using the gwpca function with a very large fixed bandwidth...
pca.result <- gwpca(Data.1.scaled,Nuts23Coords,bw=100000000000000000,k=comp.keep,verbose=TRUE)

# Identify some outliers
resids.pca <- sqrt(gwpca.cv.contrib(Data.1.scaled, Nuts23Coords, pca.result$bw,k=comp.keep))
```

```
#sort(resids.pca)
quant.pca <- quantile(resids.pca, probs = seq(0, 1, 0.01))
threshold.pca <- quant.pca[96] # in this case the 95%tile

# Standard Boxplot statistics for resids.pca
bp <- boxplot.stats(resids.pca, coef=1.5)
bp$stats
bp$stats[1]
bp$stats[5]
bp$conf
sort(bp$out)
length(bp$out)

# Identifying & updating outlier information in one file
indicator.7 <-ifelse(resids.pca>bp$stats[1]& resids.pca<bp$stats[5], 0, 1) # i.e. suspected outliers...

# A choropleth map of the PCA outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.7,shades.2)
title("Outliers (black): technique 7 - PCA")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")



################################################################################
# 8. LWPCA outliers  - via boxplots ############################################
################################################################################



# Automatically finding the optimal adaptive bandwidth by cross-validation...
lwpca.result <- lwpca.autobw.by.nn(Data.1.scaled,c(100,790),k=comp.keep,verbose=TRUE)

# Some output
#lwpca.result$bw
#lwpca.result$var
#lwpca.result$loadings

# Identify some outliers
resids.lwpca <- sqrt(lwpca.cv.contrib(Data.1.scaled, lwpca.result$bw,k=comp.keep))
quant.lwpca <- quantile(resids.lwpca, probs = seq(0, 1, 0.01))
threshold.lwpca <- quant.lwpca[96]

# Standard Boxplot statistics for resids.lwpca
bp <- boxplot.stats(resids.lwpca, coef=1.5)
bp$stats
bp$stats[1]
bp$stats[5]
bp$conf
```

```
sort(bp$out)
length(bp$out)

# Identifying & updating outlier information in one file
indicator.8 <-ifelse(resids.lwpca>bp$stats[1]& resids.lwpca<bp$stats[5], 0, 1) # i.e. suspected outliers...

# A choropleth map of the LWPCA outliers...
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.8,shades.2)
title("Outliers (black): technique 8 - Locally weighted PCA")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")



################################################################################
# 9. GWPCA outliers - via boxplots ############################################
################################################################################



# Automatically finding the optimal adaptive bandwidth by cross-validation...
gwpca.result <- gwpca.autobw.by.nn(Data.1.scaled,Nuts23Coords,c(100,790),k=comp.keep,verbose=TRUE)

# Some output
#gwpca.result$bw
#gwpca.result$var
#gwpca.result$loadings

# Identify some outliers
resids.gwpca <- sqrt(gwpca.cv.contrib(Data.1.scaled, Nuts23Coords, gwpca.result$bw,k=comp.keep))
quant.gwpca <- quantile(resids.gwpca, probs = seq(0, 1, 0.01))
threshold.gwpca <- quant.gwpca[96]

# Standard Boxplot statistics for resids.gwpca
bp <- boxplot.stats(resids.gwpca, coef=1.5)
bp$stats
bp$stats[1]
bp$stats[5]
bp$conf
sort(bp$out)
length(bp$out)

# Identifying & updating outlier information in one file
indicator.9 <-ifelse(resids.gwpca>bp$stats[1]& resids.gwpca<bp$stats[5], 0, 1) # i.e. suspected outliers...



# A choropleth map of the GWPCA outliers...
shades.4 = shading(c(0,1,2),c("white","yellow","black")) # i.e. yellow - no & black - yes
```

```
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.9,shades.2)
title("Outliers (black): technique 9 - Geographically weighted PCA")
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1000000,80000, col="blue")


################################################################################
# C. Linking all data together for weight of evidence map ######################
################################################################################



data.out <- cbind(indicator.1, indicator.2, indicator.3,indicator.4, indicator.5, indicator.6,
indicator.7, indicator.8, indicator.9, data1@data)

# To a text file
#write.table(data.out,"WoE outliers 1.txt", col.names=T,row.names=F)

data.out.1 <- cbind(X, Y, Order_ID, indicator.1, indicator.2, indicator.3,
indicator.4, indicator.5, indicator.6,indicator.7, indicator.8, indicator.9)

# To a text file
#write.table(data.out.1,"WoE outliers indicators 1.txt", col.names=T,row.names=F)

max(indicator.1)
max(indicator.2)
max(indicator.3)
max(indicator.4)
max(indicator.5)
max(indicator.6)
max(indicator.7)
max(indicator.8)
max(indicator.9)

indicator.1a <- indicator.1 >= 1
indicator.2a <- indicator.2 >= 1
indicator.4a <- indicator.4 >= 1
indicator.5a <- indicator.5 >= 1
indicator.6a <- indicator.6 >= 1

max(indicator.1a)
max(indicator.2a)
max(indicator.3)
max(indicator.4a)
max(indicator.5a)
max(indicator.6a)
max(indicator.7)
max(indicator.8)
```

```
max(indicator.9)


# Put all indicator data together...
indicator.10 <- indicator.1a+indicator.2a+indicator.3+indicator.4a+indicator.5a+indicator.6a+indicator.7+indicator.8+indicator.9

summary(indicator.10)
# Histogram
X11(width=5.3,height=5.7)
hist(indicator.10,br=c(0,1,2,3,4,5,6,7,8,9), main="Distribution of 'weight of evidence'",
xlab="Indicator sum")

# A choropleth map of suspected outliers...
shades.7 = shading(c(1,3,5,7,9),c("white","yellow","orange","red","dark red","black"))
X11(width=8,height=7)
par(mar=c(0,0,2,0))
choropleth(data1,indicator.10,shades.7)
title("Suspected outliers - weak to strong (yellow to black) evidence")
choro.legend(-2400000,2200000,shades.7,
over="exactly", between="to under",
fmt="%4.0f",title='Indicator sum (max.: 9)',cex=0.8)
map.scale(1800000,650000,1100000,"x500 km",2,1)
north.arrow(1800000,1400000,80000, col="blue")
text(-1900000,2400000, "'Weight of evidence'", cex=1, col=3)
```

TECHNICAL REPORT

ESPON

USING DOWNSCALED POPULATION IN LOCAL DATA GENERATION

A COUNTRY-LEVEL EXAMINATION

CONTENT

- Research Context and Approach. This part outlines the background to and methodology of the examination of downscaled population data.
- A Country-Level Examination. This part presents the results of a Swedish examination of downscaled population, focusing on 1) population estimates in varying local settings, and 2) the estimation of overall population for UMZs of different sizes.
- Summary and Discussion. This part points out that while there are obvious limitations to downscaled population data, it is a quite reasonable tool for certain purposes. In particular, fairly good estimations of UMZ population can be obtained.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

16 PAGES

LIST OF AUTHORS

Magnus Strömgren, Dept. of Social and Economic Geography, Umeå University

Einar Holm, Dept. of Social and Economic Geography, Umeå University

Contact

magnus.stromgren@geography.umu.se
einar.holm@geography.umu.se

tel. + 46 90 786 52 58

TABLE OF CONTENTS

Introduction

In the ESPON db context, there is a need to utilize or present population data with a high degree of spatial resolution. For instance, in the disaggregation of socioeconomic data to grid level, detailed local population data is required for a proper downscaling of certain variables. Similarly, in reporting population figures for geographical subdivisions such as Urban Morphological Zones (UMZs), NUTS or even Local Administrative Units (LAU), level 2 population figures won't suffice. The approach that has been taken is to make use of downscaled population data—a population grid produced by the Joint Research Centre (JRC). This dataset, "Population density disaggregated with CORINE land cover 2000", distributes LAU, level 2 population data to a grid, mainly using CORINE land cover data.

However, there are a limited number of tests of the suitability of using the population grid for different purposes, as well as of its reliability in different settings. This technical report presents the results of a country-level examination of the population grid, using Swedish register population data.

1 Research Context and Approach

In addition to exploring the role of survey data, an important ESPON db activity for the Department of Social and Economic Geography, Umeå University is to carry out comparisons between Swedish data and data with EU coverage. The department has access to Swedish register data, which not only covers the entire population of Sweden for a substantial time period, but also has a high degree of spatial resolution. This resource makes possible a broad range of exploratory studies and evaluations.

This technical report presents the results of a country-level examination of downscaled population for the EU. In the study, Swedish register population data is used to examine the JRC population grid "Population density disaggregated with CORINE land cover 2000". The population grid—which allocates LAU, level 2 2001 census population data to 100 m^2 squares, mainly using CORINE land cover data—is an important tool in the ESPON db project. First, it is part of the workflow to disaggregate socioeconomic data into a grid structure. This is presented in more detail in the technical report "Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase". Second, it is utilized in order to assign population to Urban Morphological Zones (UMZs).

However, the suitability of using the population grid for different purposes, as well as its reliability in different settings, has not been subject to much scrutiny. Still, some validations of the population grid have been performed. For instance, a comparison with Austrian reference data at the km^2 level showed an overall reduction by 50 percent in the disagreement with reference data, when compared to a non-weighted distribution of the population.[1] Against this background, it is not without interest to examine how the population grid compares to Swedish register data.

1.1 Methodology

Since the population grid departs from population figures for LAU, level 2—in the Swedish case, municipalities—grid population summarized at that level can be expected to largely correspond to register data. However, at the local level, it may be more or less reliable. Similarly, the performance of the grid in estimating population figures for other geographical subdivisions (e.g., UMZs) is unclear. Taking into account how the population grid is employed in the ESPON db project, the examination focuses on 1) population estimates in varying local settings, and 2) the estimation of overall population for UMZs of different sizes.

The basis for the **first test**, concerning grid population estimates in varying local settings, is a calculation of residuals. This is carried out by comparing an aggregation of the population grid to square kilometers with corresponding Swedish register data. Absolute residuals are then summarized at the municipality level (n=290). In addition to looking at the results per municipality, results are also categorized by "municipality groups"—a classification of Sweden's municipalities in nine different groups, created in 2005 by the Swedish Association of Local Authorities and Regions. The municipality group classification aims at defining homogenous regions, which share similar characteristics in terms of for instance population size, commuting patterns and

[1] Gallego J., Downscaling population density in the European Union with a land cover map and a point survey, JRC-Ispra.

employment profile. There are nine municipality groups, presented in Figure 1 and Table 2. This first test considers not only the absolute residual sum, but also the residual sum in relation to municipality area (expressed in km^2) and population size. When relating residual sum to population size, initial figures are multiplied by 50 in order to get a more gini-style estimation of the overall discrepancy between grid and register data.

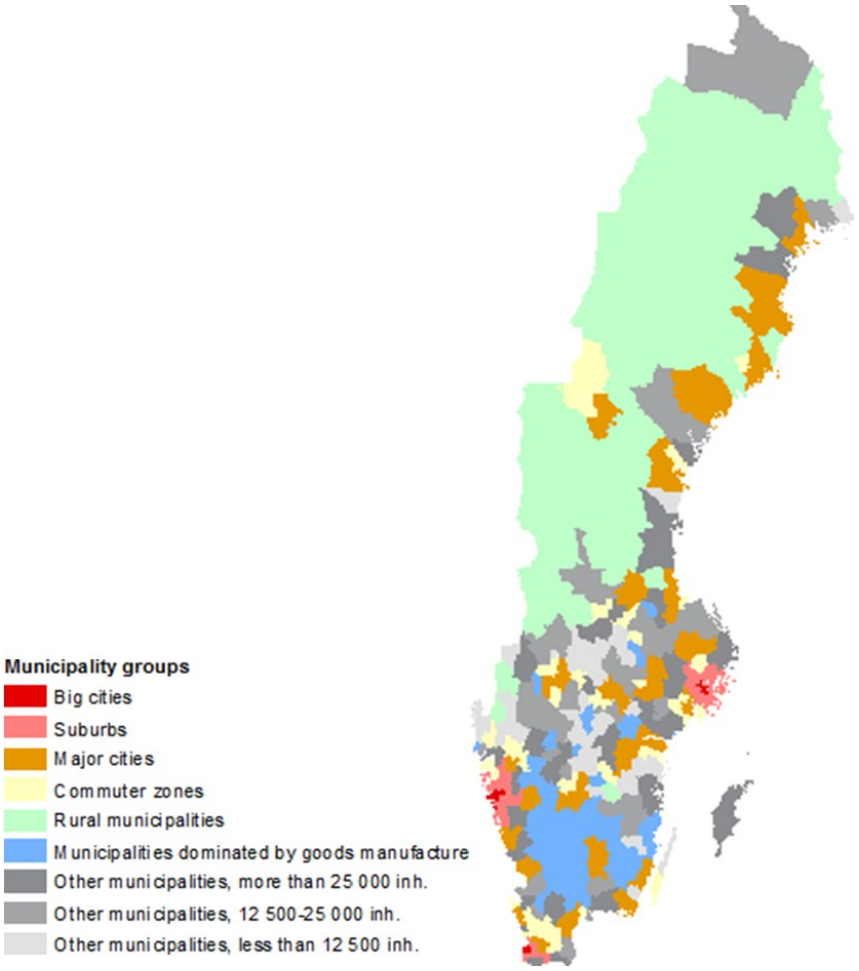


Figure 1: *Municipality group map*

ID	Category	Number of municipalities
1	Big cities	3
2	Suburbs	38
3	Major cities	27
4	Commuter zones	41
5	Rural municipalities	39
6	Municipalities dominated by goods manufacture	40
7	Other municipalities, more than 25,000 inhabitants	34
8	Other municipalities, 12,500–25,000 inhabitants	37
9	Other municipalities, less than 12,500 inhabitants	31

Table 1*: Municipality groups: IDs and frequencies*

The **second test** concerns using the population grid to estimate the population of Urban Morphological Zones (UMZs)—a delimitation of urban areas with EU coverage. In this test, overall population figures for each UMZ are calculated using both the original population grid and register data, which then are used for calculation of per-UMZ residuals. Thus, in contrast to the first test—which is based on the sum of absolute square residuals—this test focuses is the overall predictive capabilities of the grid, when it comes to UMZs of different sizes. Grid residuals within each UMZ have also been produced, primarily for purposes of trying to clarify patterns of over- and underestimation.

2 A Country-Level Examination

2.1 Population Estimates in Varying Local Settings

The first test of the population grid concerns population estimates in varying local settings. Clearly, the way discrepancies between grid and register data is associated to the local context depends on whether absolute residuals are just summarized or related to area or population. In Figure 2, the ten municipalities with highest (red) and lowest (green) absolute residuals are displayed, using three different measures: residual sum (left) as well as residual sum related to municipality area (middle) and population size (right). Municipalities with a comparatively large population (e.g., Malmö) tend to fare quite bad regarding residual sum and residual sum related to area, but pretty good when residual sum is related to population size. For municipalities with a comparatively small population (e.g., several municipalities in the inland of Northern Sweden), the situation is the opposite.

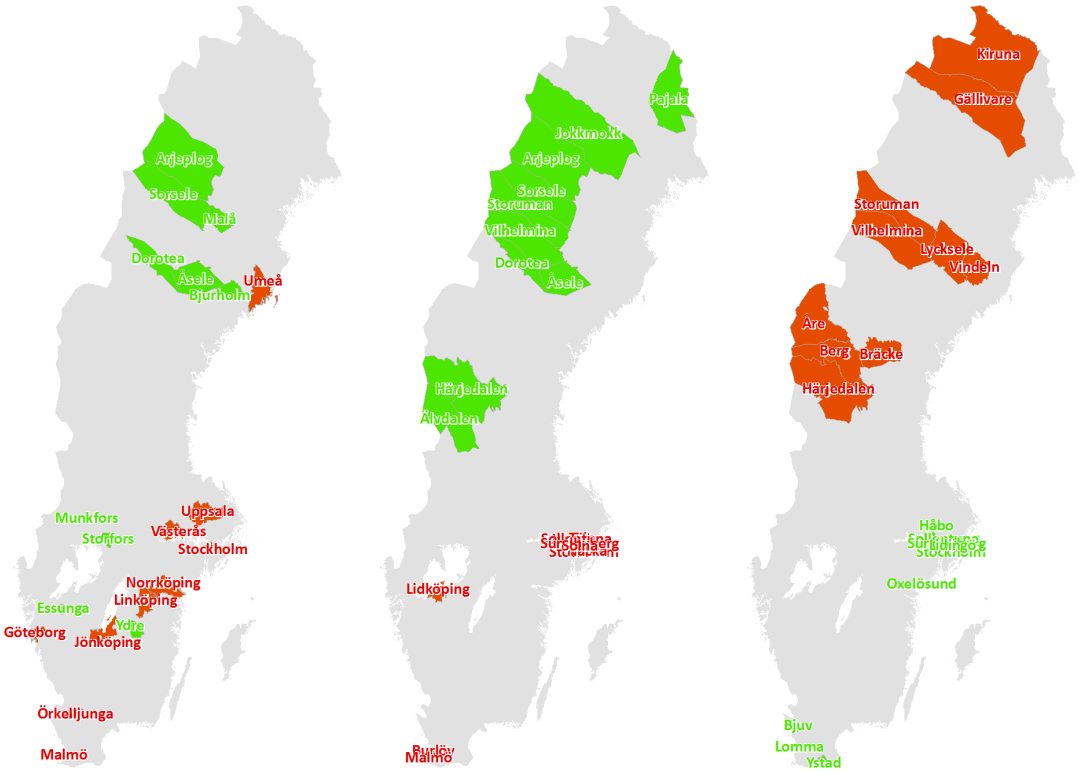


***Figure 2**: The ten municipalities with highest (red) and lowest (green) absolute residuals summarized (left) and in relation to area (middle) and population size (right)*

In Figure 3, all 290 municipalities are displayed in a scatterplot, with population size on the x-axis and absolute residuals by population size on the y-axis. The scale on the x-axis is logarithmical. There is a clear relationship between the two dimensions. As municipality population size increases, residual sum relative population size tends to decrease. However, this overall relationship is not without exceptions. In particular, for municipalities with a population of about 10,000 inhabitants, there are considerable variations in the level of overall discrepancy between grid and register data.

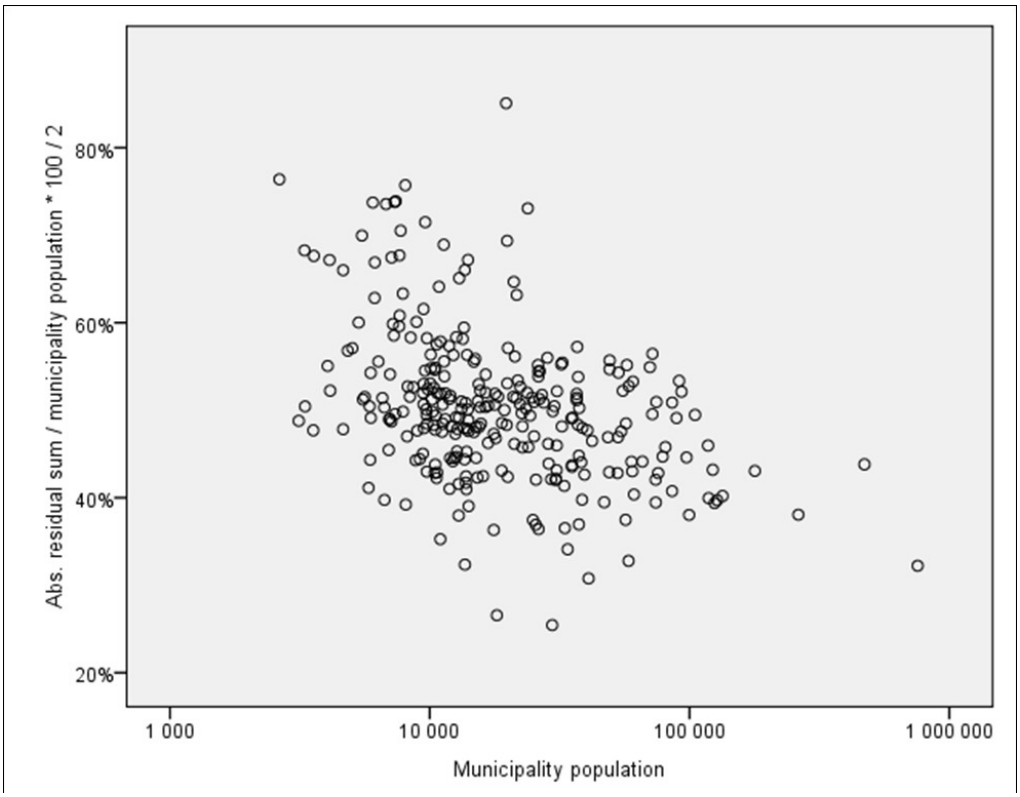


***Figure 3**: Municipality population size compared to absolute residual sum in relation to municipality population size*

2.1.1 Results by Municipality Group

In order to get a better understanding of how the three residual measures are related to the local context, municipalities are categorized by the municipality group to which they belong (see Figure 1 and Table 1). By the use of boxplots to graphically present the results, variations within and between these varying kinds of local settings becomes apparent. The first boxplot (Figure 4) displays absolute residual sum. The by far highest median error can be found in group 1, "big cities". Municipality groups 3 ("major cities") and 7 ("other municipalities, more than 25,000 inhabitants") also exhibit comparatively large median errors (cf. Figure 2, left). It should be noted that the scale on the y-axis is logarithmical.

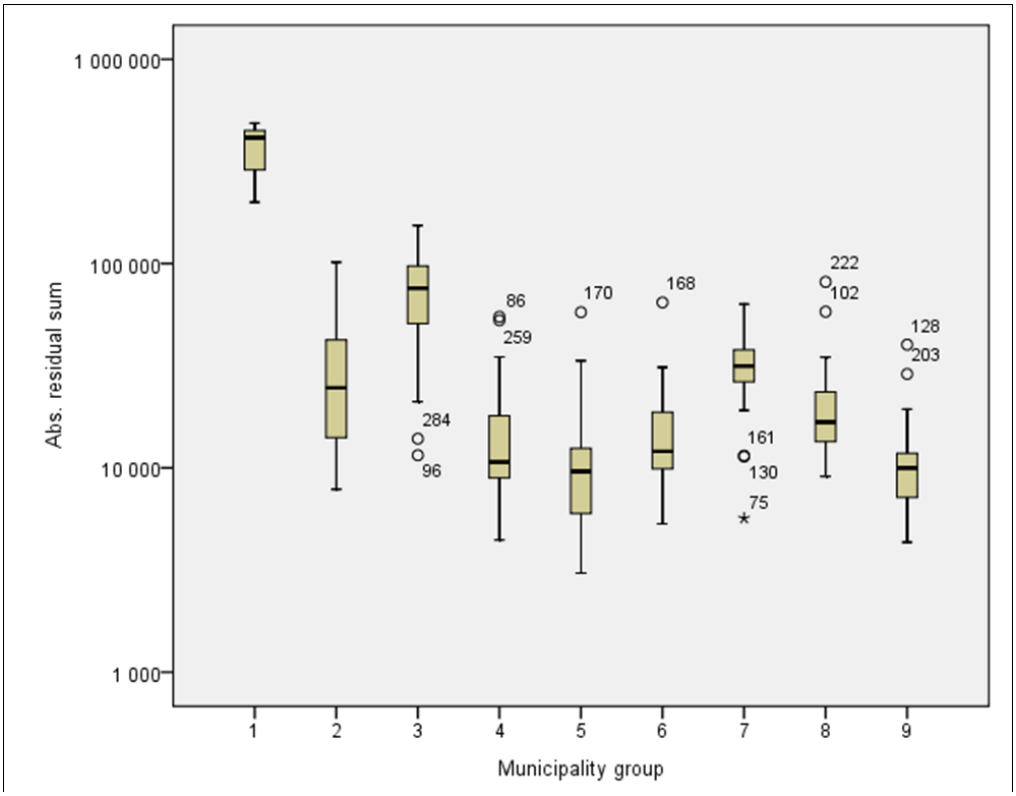


***Figure 4**: Absolute municipality residual sum subdivided by municipality group*

In relation to area (Figure 5), a somewhat similar pattern of differences between municipality groups emerges. The big cities category (1) exhibits the largest median error; rural municipalities (5) clearly the smallest (cf. Figure 2, middle). When it comes to municipality group 2, which represents suburban municipalities, there is a substantial internal variation. It should be noted that the scale is the y-axis is logarithmical. For the gini-style measure of residual sum related to municipality size, the pattern is quite different (Figure 6). Rural municipalities exhibit the largest median error; big cities by far the smallest (cf. Figure 2, right). There are substantial variations within not only suburban, but also rural municipalities.

Like the results presented in Figure 2 and Figure 3, the municipality group comparison indicates that there is a relationship between the three residual measures and population size. In addition, it reveals substantial variations within certain municipality groups. In the case of suburban municipalities, it is easy to see why such diversity may arise. The settlement structure in suburban areas varies considerably, ranging

from spacious residential area to crowded housing estates. Presumably, the mix of suburban housing in certain municipalities produces a population distribution more in line with the figures of the population grid.

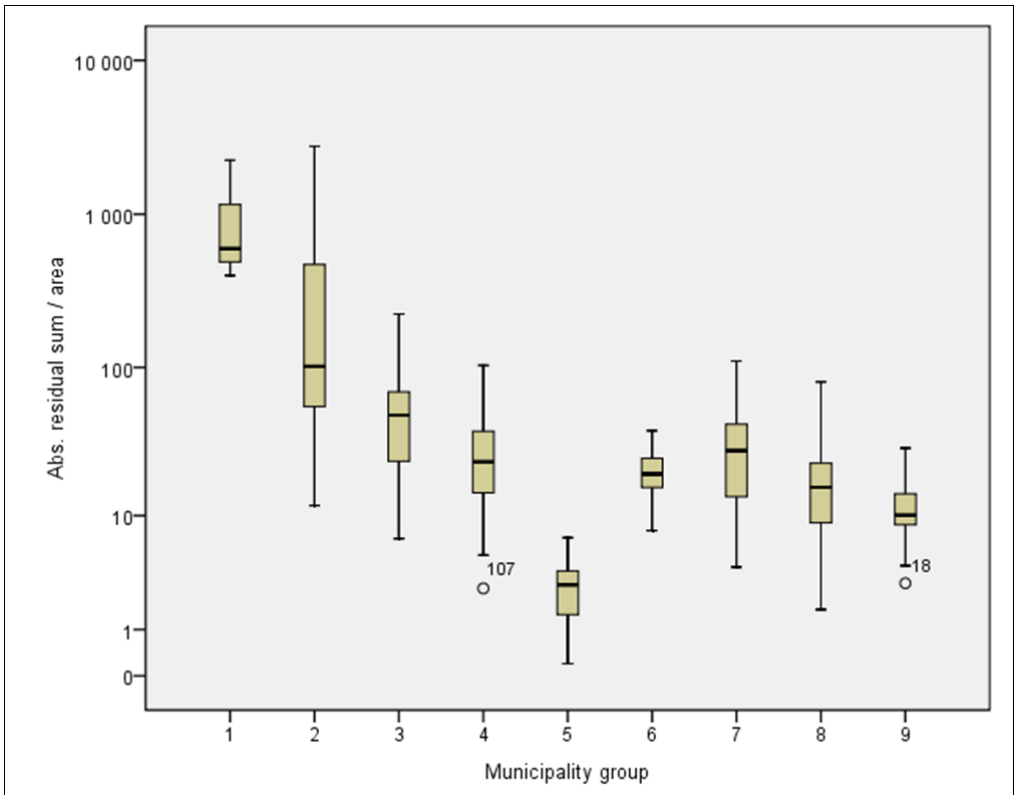


Figure 5 *: Absolute municipality residual sum in relation to area, subdivided by municipality group*

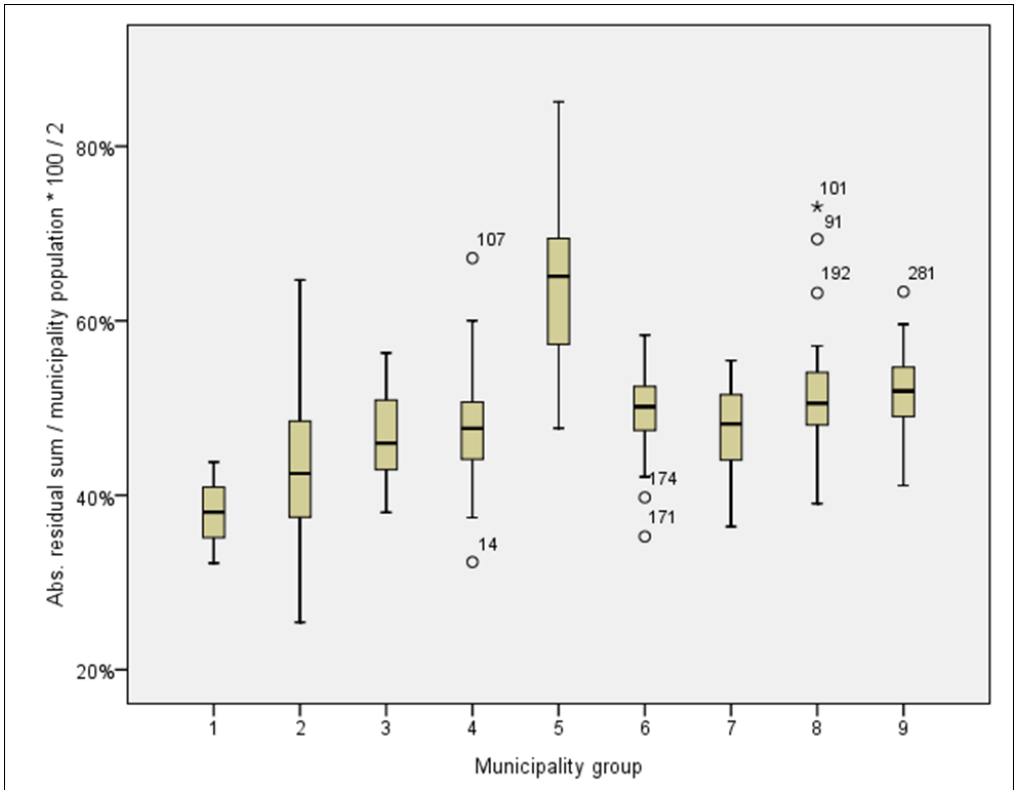


Figure 6*: Absolute municipality residual sum in relation to population size, subdivided by municipality group*

2.1.2 The Residual Map

In the interpretation of the results, the underlying km^2 residual map may yield some clues. Figure 7 shows a residual map for Sweden as a whole; Figure 8 two close-ups of the residuals, representing a part of Southern (left) and Northern (right) Sweden. In these maps, red color means that the grid population is larger than the register population. Conversely, blue color indicates that the grid population is smaller than the register population. It should be noted that these maps only show squares that are inhabited in register data. As can be seen in Figure 7 and—even more clearly—Figure 8, there is a tendency for the grid to underestimate the population size of inhabited squares in the inland of Northern Sweden. Primarily, this is due to the assignment of population figures to many actually uninhabited squares. Naturally, this is a phenomenon that is likely to be more pronounced in sparsely populated areas, such as the rural municipalities in the inland of Northern Sweden (cf. Figure 1).

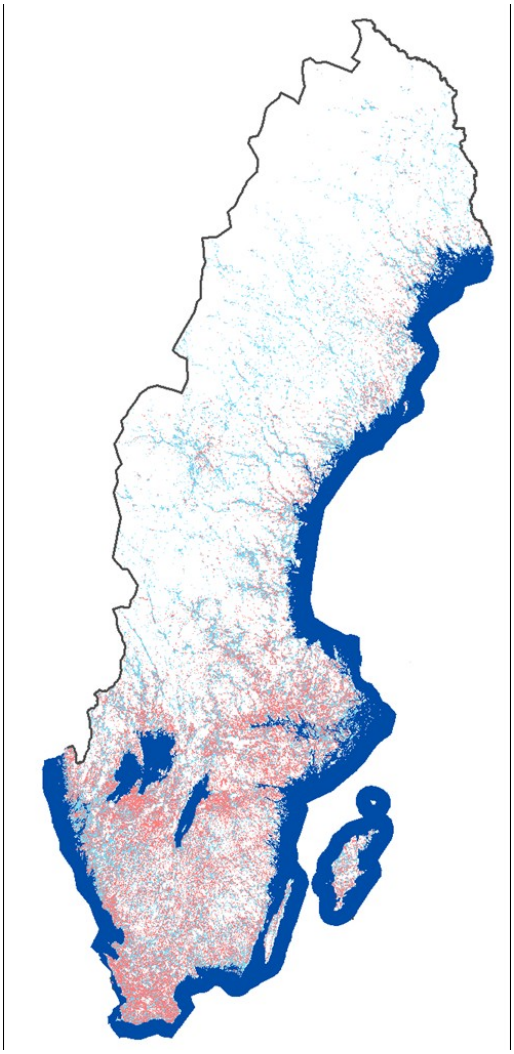


***Figure 7**: Absolute km^2 square residuals*

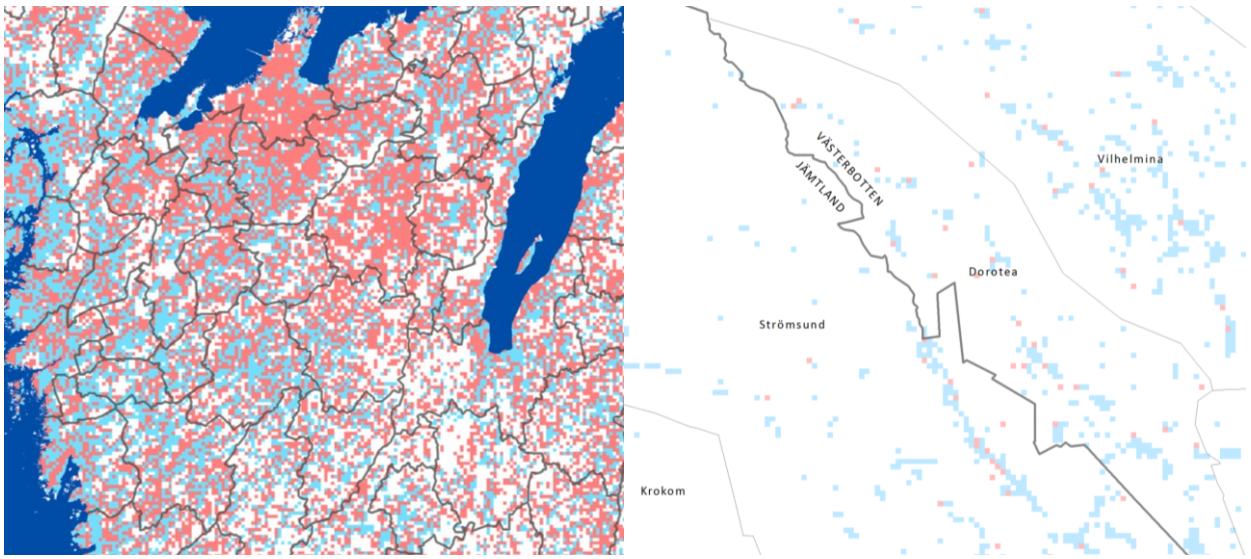


Figure 8: *Close-up of km^2 residuals in Southern (left) and Northern (right) Sweden*

2.2 Estimations of UMZ Population

The second test of the population grid, which concerns estimations of overall UMZ population, reveals an intriguing pattern of varying degrees over- and underestimation depending on UMZ size. Figure 9 displays, in scatterplot form, UMZ register population size on the x-axis, and overall residuals (register population – grid population) on the y-axis. For both the x- and the y-axis, the scale is logarithmical. In addition, the observations are binned: the larger the dots, the more UMZs are located in and around that point in the scatterplot. All in all, the number of over- and underestimated UMZs are about equal. Two separate clusters—one of which exhibits a linear trend—are clearly evident. First, for most UMZs with about 1,000 inhabitants and more according to register data, population is underestimated, and the underestimation increases with UMZ size. Second, the population of many UMZs with a register population below 1,000 inhabitants is—more or less—overestimated. In the boxplot that makes up Figure 10, this phenomenon is evident by the large range and many outliers in the UMZ size classes "200-999" and "1000-9,999", respectively. Generally, the amount of over- and underestimation is quite modest, especially when viewed in the light of actual population size.

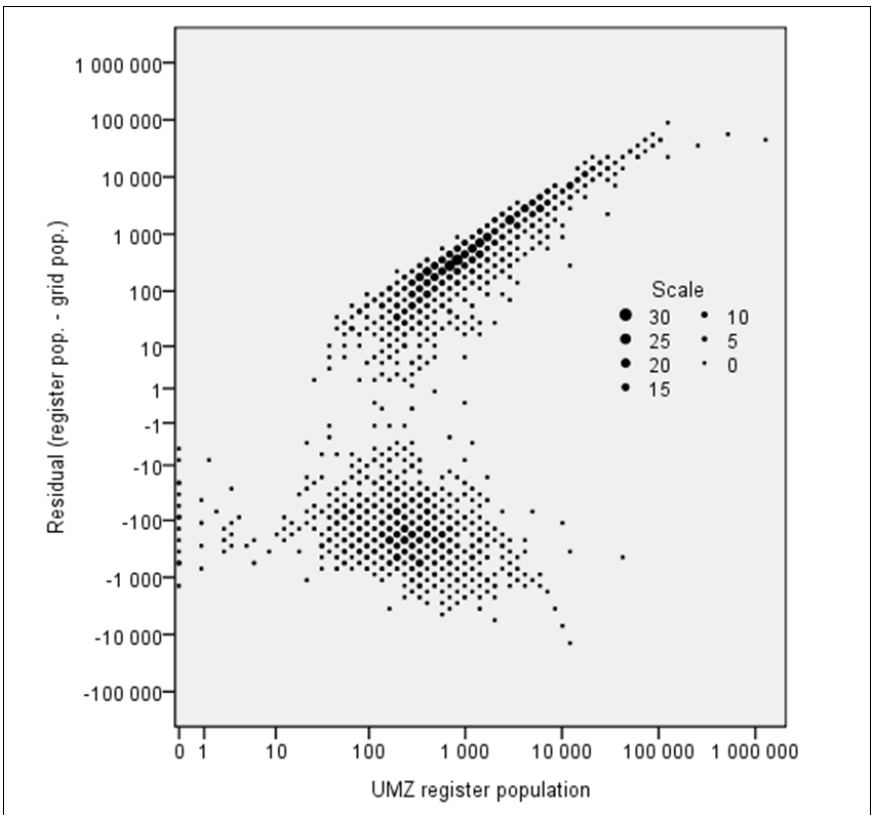


***Figure 9**: UMZ residuals by UMZ register population*

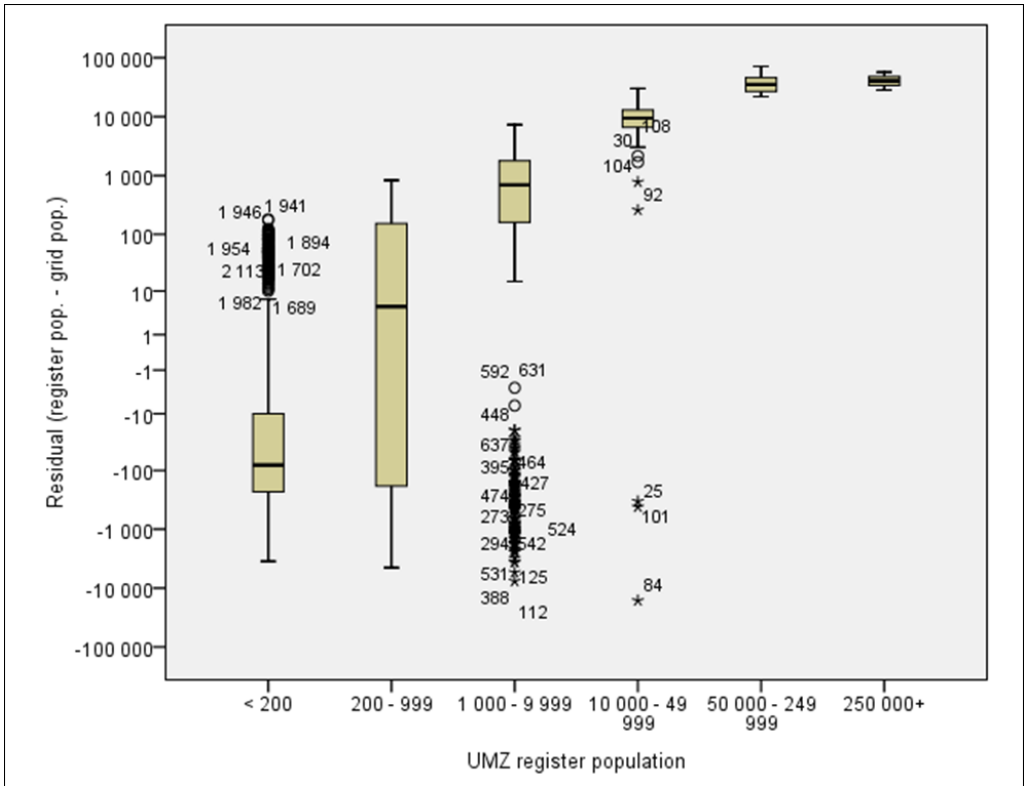


***Figure 10**: UMZ residuals by UMZ register population classes*

When it comes to the overestimation of many small UMZs, a possible explanation could be that the population grid overestimates areas with many buildings but small resident population. Spatial agglomerations of second homes, which are quite common in Sweden, are obvious examples of this kind of area. In Table 2, the UMZ

boundaries are related to on the one hand the Swedish definition of urban localities (*tätort*)—basically, any agglomeration of 200 or more inhabitants—and on the other hand a delimitation of "other concentrated settlement". For UMZs that intersect any urban locality, 40 percent are overestimated. By contrast, for UMZs intersecting either only "other concentrated settlement" or neither category, the overestimation figure rises to about 80 percent. In practice, the "other concentrated settlement" category is largely made up by second home areas. Clearly, then, this finding lends some support to the notion of second home areas being responsible for the cluster of overestimated UMZs. Still, 80 percent of UMZs actually overlap urban localities, and a substantial proportion of those UMZs are also overestimated. In other words, the pattern of overestimation may also be a question of UMZ size.

Relation to Swedish delimitations	% of UMZs	% of UMZs overestimated
UMZ intersects neither urban locality (*tätort*) nor "other concentrated settlement"	10	82
UMZ intersects only "other concentrated settlement"	10	79
UMZ intersects urban locality (*tätort*)	80	40

***Table 2**: UMZs in relation to Swedish definitions of settlements*

Concerning the general and increasing underestimation of large UMZs, there is harder to find an explanation for the phenomenon. Figure 11 shows grid residuals for UMZs in the Stockholm area, including residuals for Stockholm UMZ—the largest UMZ in Sweden in terms of population size, and also among the most underestimated using the population grid. In this map, red and—in particular—dark red color means that the grid population is larger than the register population. Conversely, the two shades of blue indicate that the grid population is more or less smaller than the register population. In the city center, there is—not surprisingly—a clear tendency for the grid to overestimate the population, while suburban areas generally exhibit a mixed pattern of over- and estimation. While this overall residual pattern is likely to occur in many other larger UMZs, it gives no obvious clue as to the reasons for the overall trend towards increased population underestimation with increased UMZ size.

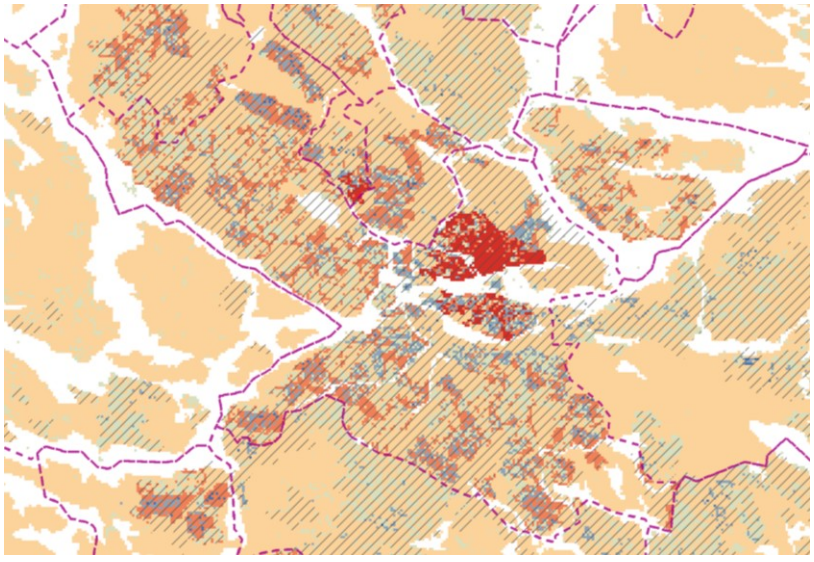


***Figure 11**: Close-up of 100 meter^2 residuals in UMZs in the Stockholm area*

3 Summary and Discussion

In the ESPON db context, there is a need for population data with a high degree of spatial resolution. For instance, when it comes to disaggregating certain socioeconomic variables, or reporting population figures for geographical subdivisions such as UMZs, available data with good spatial coverage (e.g. for NUTS 2 or LAU, level 2 regions) have obvious limitations. Therefore, downscaled population data—specifically, a JRC population grid covering the entire EU—has been employed for these purposes. Against this background, the dataset has been subject to a country-level examination using Swedish register population data. Taking into account how the population grid is employed in the ESPON db project, the examination focuses on 1) population estimates in varying local settings, and 2) the estimation of overall population for UMZs of different sizes.

The first test summarizes absolute local residuals for Sweden's 290 municipalities (i.e., LAU, level 2 subdivisions), using three different measures: residual sum as well as residual sum related to municipality area and population size. Results are also presented categorized by municipality group—a classification of municipalities in nine different groups according to their characteristics. The results indicate that there is a relationship between municipality population size on the one hand, and the three residual measures on the other. Municipalities with a large population are associated with low discrepancies between grid and register data when residual sum is related to population size, but high discrepancies in terms of residual sum and residual sum related to area. For small municipalities, such as many rural municipalities in Northern Sweden, the situation is the opposite. A map of the actual local residuals reveals that there is a tendency for the population grid to underestimate the population size of inhabited squares in such settings. Primarily, this is due to the assignment of population figures to many actually uninhabited squares. In an EU perspective, this is likely to be less of an issue. The municipality group comparison reveals substantial variations within certain municipality groups, particularly regarding the category representing suburban municipalities. Presumably, this reflects the considerable diversity in settlement structure and population distribution that exist in suburban areas.

The second test concerns using the population grid to estimate the population of Urban Morphological Zones (UMZs). When overall residuals are related to UMZ population size, the about equal number of over- and underestimated UMZs form two separate clusters. For large UMZs the number of inhabitants tends to be underestimated, and the underestimation increases with UMZ size, while the population of many small UMZs is—more or less—overestimated. A plausible explanation for the latter phenomenon is that the population grid overestimates areas with many buildings but small resident population, such as second home areas. Generally, the amount of overall over- and underestimation is quite modest, especially when actual UMZ population size is taken into account.

In this country-level examination of downscaled population data, the discrepancies between downscaled and register data varies depending on local setting, and is also highly influenced by the way residuals are expressed. In any case, it is hardly a stretch to conclude that local grid population estimates often are quite unreliable. Still, using the population grid to downscale socioeconomic data is a likely to enhance to quality of data, and—and least for now—no better alternative exists. In the estimation of overall UMZ population size, the population grid works quite well—at least in the Swedish context. Consequently, while there are obvious limitations to downscaled

population data, the JRC population grid is a quite reasonable tool for the enhancement of certain ESPON datasets.

TECHNICAL REPORT

ESPON

MAPPING GUIDE

CARTOGRAPHY FOR ESPON PROJECTS

CONTENT

- Maps and ESPON 2013. This part presents the content of the map-kit tool ("European, local and global"). All the elements that have to be necessarily represented on each map are described.
- Enhancing information. This part explain how symbolize ESPON 2013 data with the good rules of graphic semiology.
- Maps are tool for communication. This part insists on the fact that a map has necessarily to deliver a clear message.

ESPON 2013 DATABASE

January 2011

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

33 PAGES

LIST OF AUTHORS

Christine Zanin, University Paris 7, UMS 2414 RIATE

Nicolas Lambert, UMS 2414 RIATE

Ronan Ysebaert, UMS 2414 RIATE

Contact

christine.zanin@univ-paris-diderot.fr
nicolas.lambert@ums-riate.fr
ronan.ysebaert@ums-riate.fr

tel. + 33 1 57 27 65 32

TABLE OF CONTENT

Introduction

Maps are a great way of displaying statistical data. It allows summarizing a complex and important information into clear and compact presentation. They can bring a great help in spotting patterns within data.

Maps are accessible for many reasons. People understand maps (at least, think they do). People like maps because they attract attention and brighten up presentation. Nevertheless, and in a scientific versus, the interest of the representation of geographical information on maps can be summarized in three main points1.

The localisation is the most elementary subject related to geographic information. It allows answering to question "Where can we find this phenomenon?" The precision of the localisation depends on the quality of this kind of information such as statistical databases, statistical yearbook and so on. Locate a geographical object has generally a sense only if it is possible to compare it to other one "Why this object is located here and not there?". Answers can be read off directly from the map without any other help.

The comparison: Geographical objects analysis makes a concrete sense when it is possible to compare them. "What is the situation of this region as compare to the other one?"; "Can we observe geographical pattern, such as discontinuities, concentration?" Maps are useful tools for interpreting and pointing out specific geographical patterns, which are impossible to catch with an only statistical analysis.

Planning: Since the relations between European territories are very intensive, territorial planning on a special location must interfere with other territories and have to.

Despite many interests to use maps within ESPON, these kinds of documents have also their limits. Maps always generalise and simplify information. Mapping is more than just rendering; it also getting to know the phenomenon which is to be mapped. That's why mapping is not an easy action. Deliver the right message must remain the first objective of map design and mapping allows you to orchestrate the elements of the map to best convey its message to its audience. Thus, the design of maps is mainly concerned with making choices: the choice of mapping method (proportional symbol or choropleth map, isoline or grid map or even a cartogram), the choice of the aggregation level on which information as to be depicted, the choice on the level of statistic areas and the type of data (absolute or relative representation), the choice of graphic variables (such as differences in size, value, grain, colour, direction and shape) to be used. These choices are fundamental's one, they influence people's conception and visualisation of space.

This technical report is not a formal cartography book but allows everyone to understand easily how to produce an effective and operational map in the ESPON 2013 program. The report is organized in 3 parts: (i) Maps and ESPON 2013 (description and explanation of map-kit tool); (ii) Enhancing information (mapping methods and graphic semiology); (iii) Maps and communication (map is to deliver a simple and clear message).

[1] Béguin M., Pumain D., 2003, *La représentation des données géographiques – statistique et cartographie,* Armand Colin, 192p.

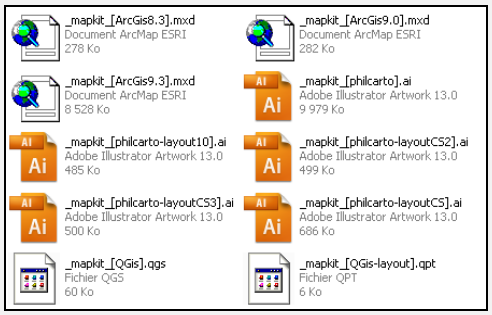
1 Maps and ESPON 2013

Description of the Map-Kit tool.

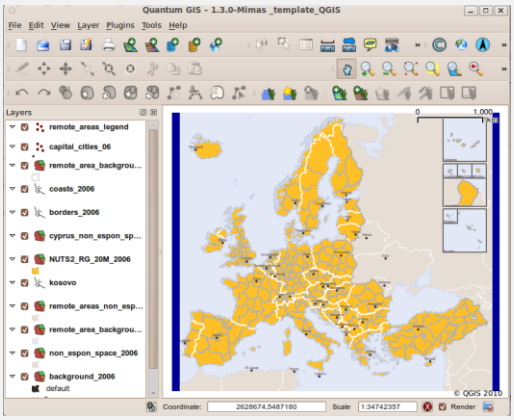
1.1 The "European" Map-Kit

To ensure the harmonisation of all maps produced by ESPON projects, an ESPON Map-Kit tool is operational. Among others, this tool contains geometries. Theses geometries are an extraction of EBM of Eurogeographics with a scale of 1:20 Million downloaded on the Eurostat website (GISCO). To finalise the cartographical template, other elements are also available (e.g. Coast lines, North part of Cyprus, The delineation of the Kosovo, remote territories, capital cities). Compatible with the EPSON 2013 database, all these elements are included in an ARCGIS mxd document, which is an easy way to make harmonized maps. But, the map kit is also available in an open source format. Quantum GIS (QGIS) is a user friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo). It runs on Linux, Unix, Mac OSX, and Windows and supports numerous vector, raster, and database formats and functionalities. It is possible to download the application on the following URL: *http://www.qgis.org/en.html*. It is also possible to use a third application to make thematic maps: Philcarto. It is not a GIS application, is it a free tool dedicated to thematic mapping and spatial analysis. The application and the documentation is downloadable on this URL: *http://philcarto.free.fr/Inscriptions.html*

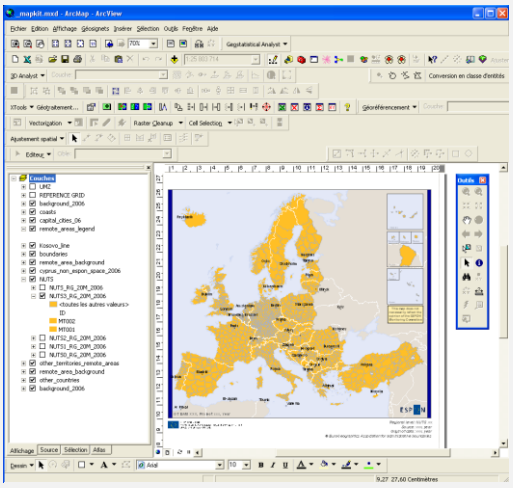
3 TOOLS FOR THE SAME MAPKIT



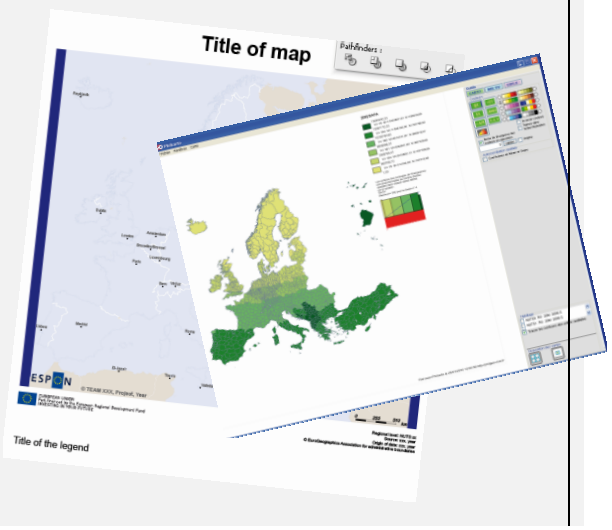
QGIS screenshot

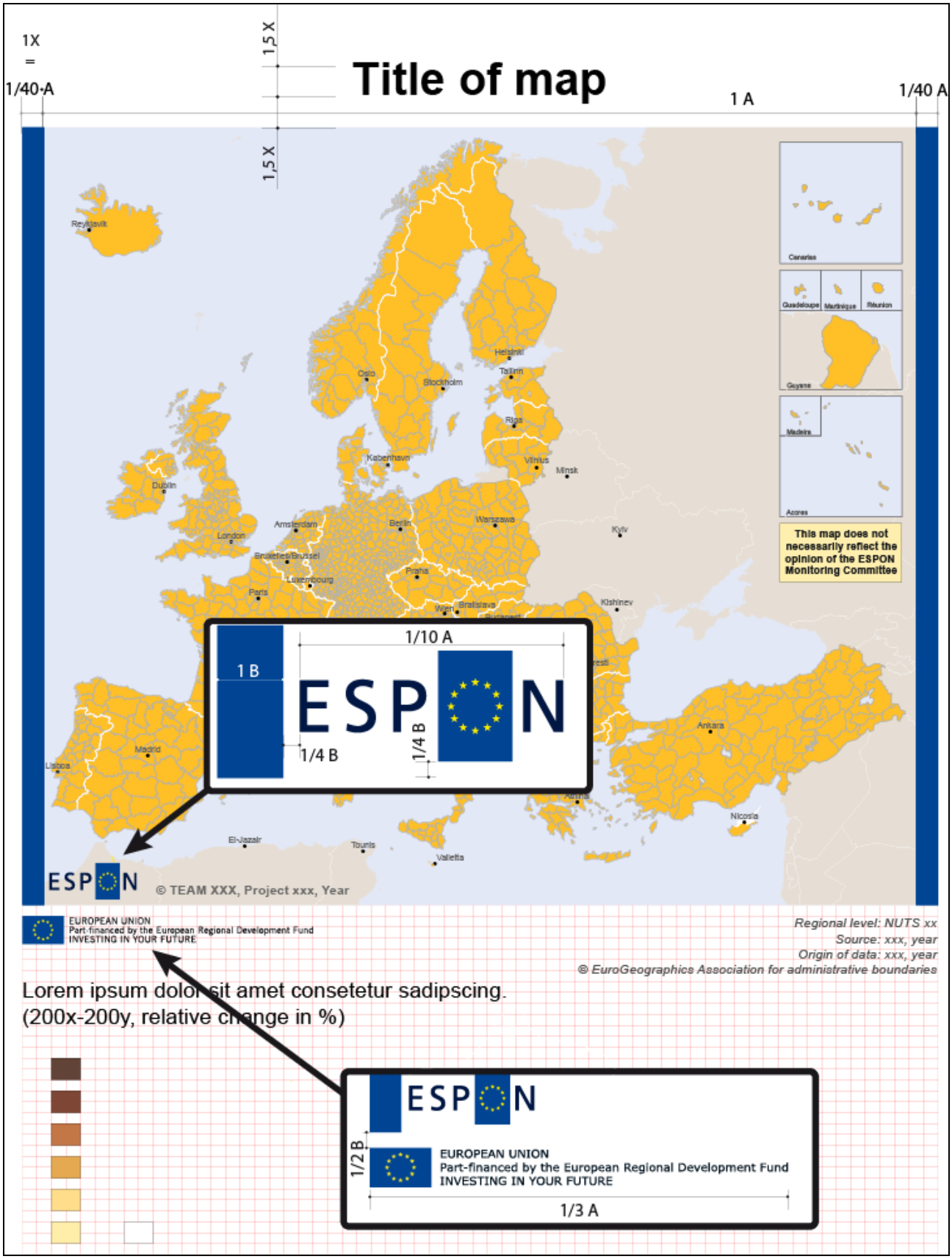


ArcGIS screenshot



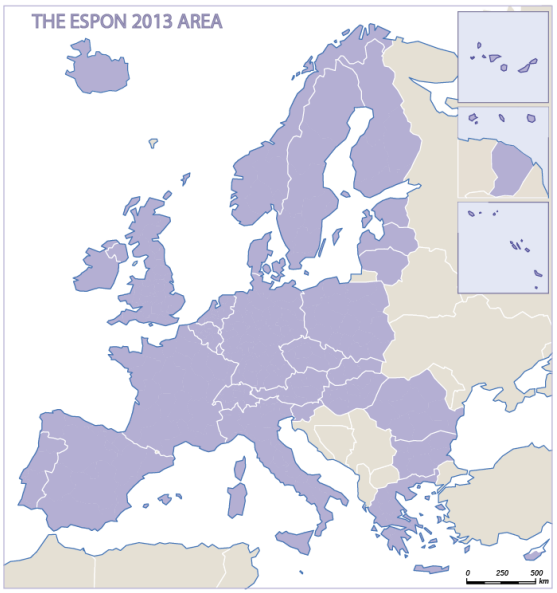
PHILCARTO screenshot





Layout of an ESPON map

1.1.1 ESPON area: 31 countries



The ESPON area defined in the current program is composed by all the Member States of European Union (27 countries) plus Switzerland, Norway, Iceland and Liechtenstein = 31 countries.

1.1.2 Candidate countries and western Balkans

More than the ESPON area, the map-kit includes the candidate countries and the Western Balkans.

For Candidate Countries (Croatia, FYROM, Turkey), the NUTS system already exists. But, concerning the western Balkans, a system named "SIMILAR NUTS" has been created (Albania, Kosovo, Montenegro, Bosnia-Herzegovina, Serbia).

Concerning the rules of cartography, drawing of borders, for some countries, must follow precise rules for political reason. In general, ESPON follows the rules established by European Commission. When these rules do not exist at EU level (for

example because of lack of consensus) the rules of UN are used as reference. According to these considerations, we have to use always the reference to the UN resolution when referring to Kosovo, i.e. under UNSCR 1244/99. On the map, the borders of Kosovo are thinner (0.20 pt) than the other boundaries (0.30 pt) and the name of the city of Pristina is not written.

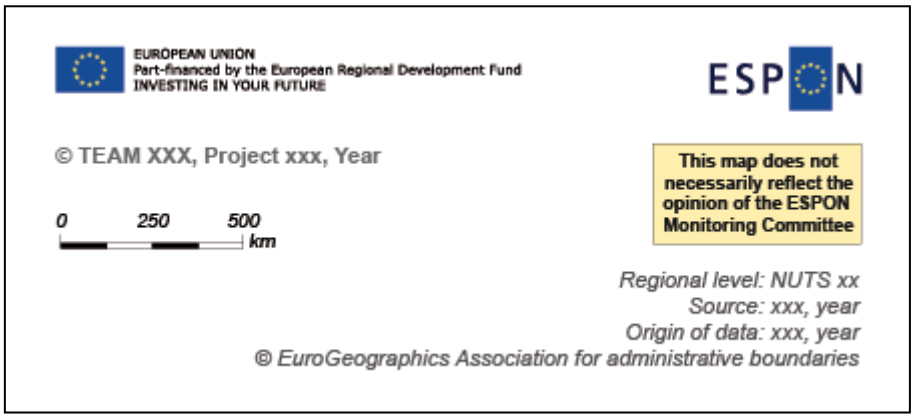
1.1.3 Projection and Ellipsoid

The projection of the ESPON MAP KIT is now based on the ETRS-LAEA system: ETRS89 Lambert Azimuthal Equal Area Coordinate Reference System. This projection is the standard in Europe for pan-European statistical mapping at all scales. In particular, this projection is used by the European Environment Agency.

Parameters: latitude of origin 52° N, longitude of origin 10° E, false northing 3 210 000.0 m, false easting 4 321 000.0 m.

EPSG code: 3035

1.1.4 Logos, disclaimer, layout, etc.



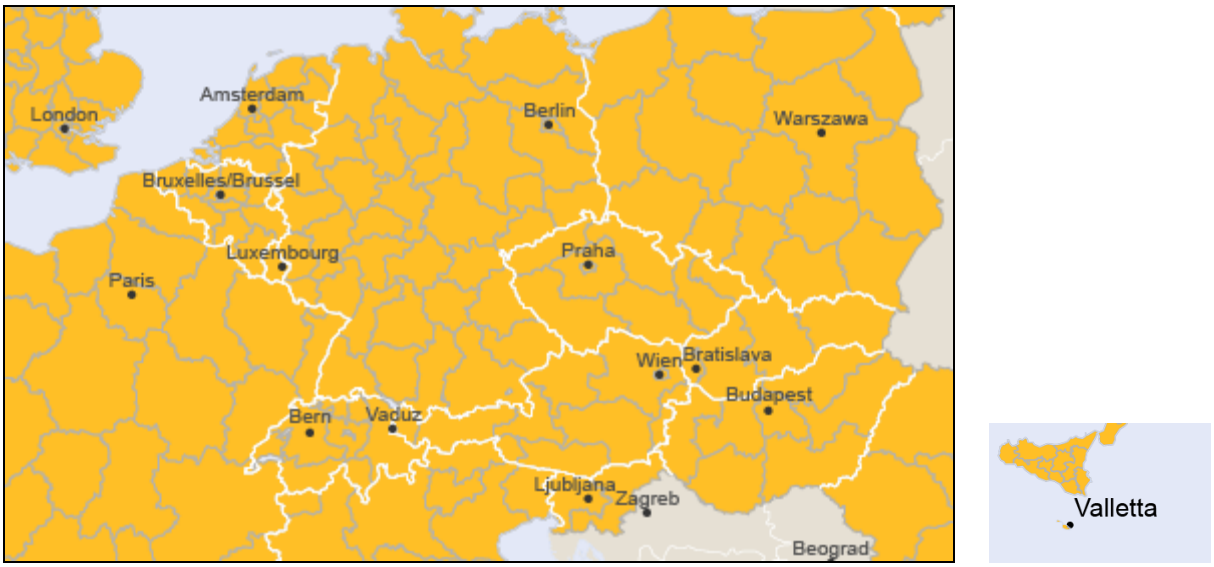
Some elements absolutely have to appear on the map layout:

- The EC publicity requirements, with the following sentence: "EUROPEAN UNION; Part-financed by the European Regional Development Fund; INVESTING IN YOUR FUTURE".
- The ESPON logo.
- The MC disclaimer: "This map does not necessarily reflect the opinion of the ESPON Monitoring Committee"
- Team, Project, Date.
- The Regional level and the NUTS version (e.g. NUTS3 2006).
- Data sources (e.g. ESPON 2013 DATABSE)
- Origin of data (e.g., European communities, June 2009)
- Eurogeographics copyright: © Eurogeographics Association for administrative boundaries.
- The Scale.

1.1.5 Capital cities

45 capital cities have to be written on the map.

Vilnius, Minsk, Dublin, Berlin, Amsterdam, Warszawa, London, Bruxelles/Brussel, Kyiv, Praha, Paris, Wien, Budapest, Bern, Beograd, Bucuresti, Sofiya, Tirana, Madrid, Ankara, Helsinki, Zagreb, Nicosia, Luxembourg, Bratislava, Tallinn, Sarajevo, Skopje, Athinai, Kishinev, Kobenhavn, Lisboa, Oslo, Reykjavik, Riga, Roma, Stockholm, Valletta, Ljubljana, El-Jazair, Tounis, Ar Ribat, Podgorica, Vaduz, Ankara



The localisation of each capital city is shown by a black bullet point. Except for Malta, the name of the city is always above the bullet point. For a better visibility, Valetta is written slightly on the right of the bullet point.

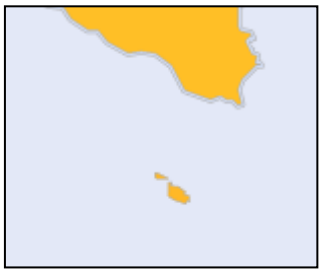
1.1.6 Remote territories

Remote territories of France (Martinique, Guadeloupe, Guyane française, Réunion), Spain (Canarias) and Portugal (Acores, Madeira) are territories members of the European Union. They have to be represented on maps even when data are not available.

1.1.7 Cyprus

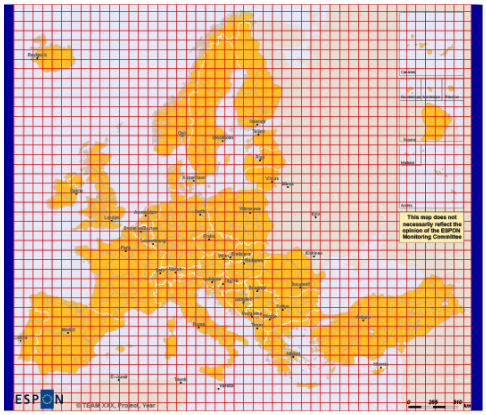
Cyprus is represented in two different colours. The North area appears in white as "no data".

1.1.8 Coast of Malta



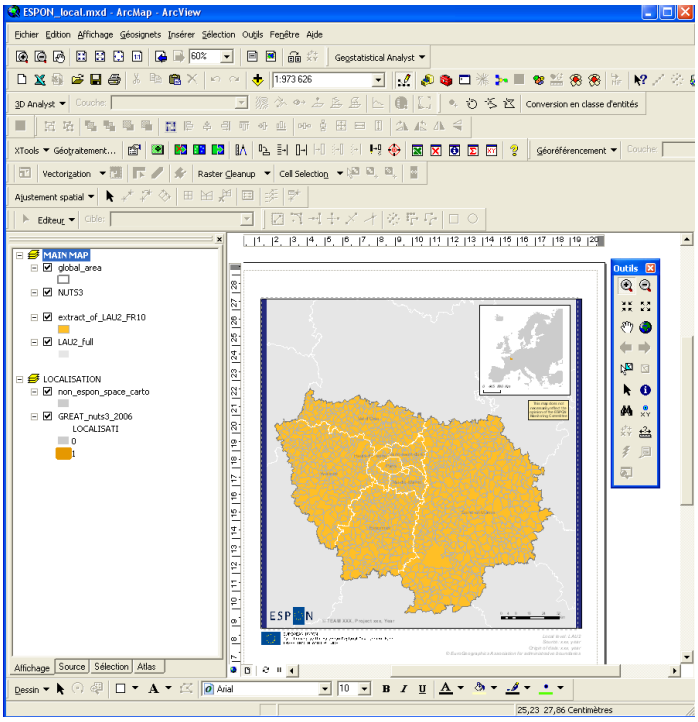
To ensure the visibility of Malta islands on the maps, we do not use light blue coast-line that could be reduce, on the map, the size of this country. Moreover, the line of the Malta polygon is drawn as thinner as possible.

1.1.9 Mapping on reference grid



The fact that the projection of the MAP KIT is the same as the projection used by EEA (EPSG 3035) ensure the compatibility between EEA reference grids and the ESPON template defined. Theses grids are included in the Map Kit. As a consequence, it is possible to use them in the same European Map Kit as previously.

1.2 The "Local" Map-Kit

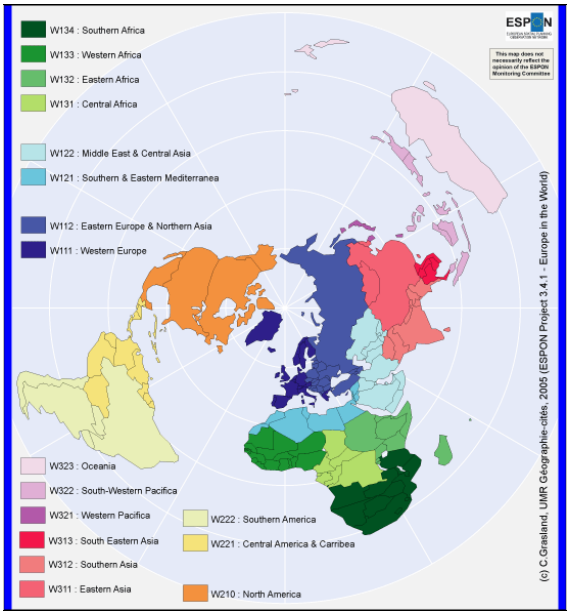


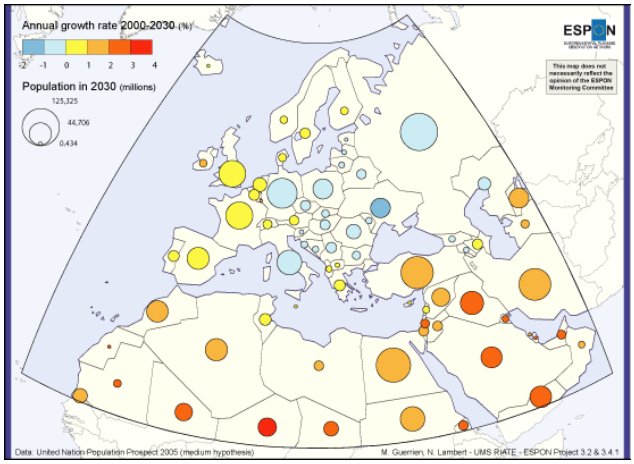
For Priority 2 projects, a "Local" map-kit has been created. As the same of the "European" Map-kit, some elements must figure on map: disclaimer, sources, logos, scale…

This template consists in 2 parts: the main map and the localisation map. The main map is an extract of the LAU2 base map. To be useful, this map have to be represented with the good local or national projection (e.g. for France: LAMBERT93). The map of localisation indicates where exactly the study area is located in Europe. This element is positioned on the top-right of the template.

1.3 The "Global" Map-Kit *(ESPON 2006 version)*

Developed during the first period of the program by the project ESPON 3.4.1 "Europe in the World", this specific map-kit is also available. It will probably be improved by teams in charge of the project "Continental territorial structures and flows (Globalisation)".





1.4 The "Zoom out" Map-Kit

In order to allow comparisons between Europe and other countries of the World at regional level, a specific map-kit has been created.

Take China, for example, its template must be divided in two parts:
The main map shows the case study at regional level and its neighbourhood (Mongolia, Russia, India etc. in grey), shown in an adapted projection (Asia Lambert Conformal Conic in this case).
The localisation map on the top right of the template indicates the location of the study area in the World. The localisation map is based on orthographic projection.
For any other part of the World you must adapt the template to your case study (geographical coverage, adapted projection…)

Classically, the map has to content the ESPON elements (disclaimer, layout, data source, scales…).

2 Enhancing information

2.1 Differentiation of data type

Many possibilities exist to show data on map. Choosing relevant representation is not an obvious task and has to be considered seriously. Indeed, choosing the wrong type of map can completely misrepresent the data. It is important to keep in mind that **the choice in cartography is always dependant on the type of data**. It is possible to identify four main types of data:

1. **Qualitative data**
2. **Quantitative data with absolute values**
3. **Quantitative data with ratios values**
2. **Ordinal (or ranked) data**

For each type of data it is possible to relate it to a **geographical reference: points, lines or areas.**

There are many possibilities to show correctly data on maps. The aim of this paper is not to present all types of correct visualisation, but an extract of the most usual and efficient ones.

2.1.1 Qualitative data

A data is qualitative when its value is a nominal one with qualitative differences: components do not allow establishing range relations between them.

For example, considering the different geographical references:

Points: location universities by type (university, polytechnics…) – ***Figure 1***

Lines: communication network without hierarchy (ferry connections, main roads) – ***Figure 2***

Areas: results from typology (rural area, urban area…) – ***Figure 3***

Qualitative data have to be shown such a manner that do not suggest rank either quantity. Two possibilities: use **geometric symbols** or **differential colour** in order to **differentiate** the different elements of the map.

With points (figure 1) the most efficient is to show information by colour or geometric symbols. It is important to use a limited quantity of symbols or colours to make the map understandable.

For lines or areas, differential colours should be used (figure 2 and 3).

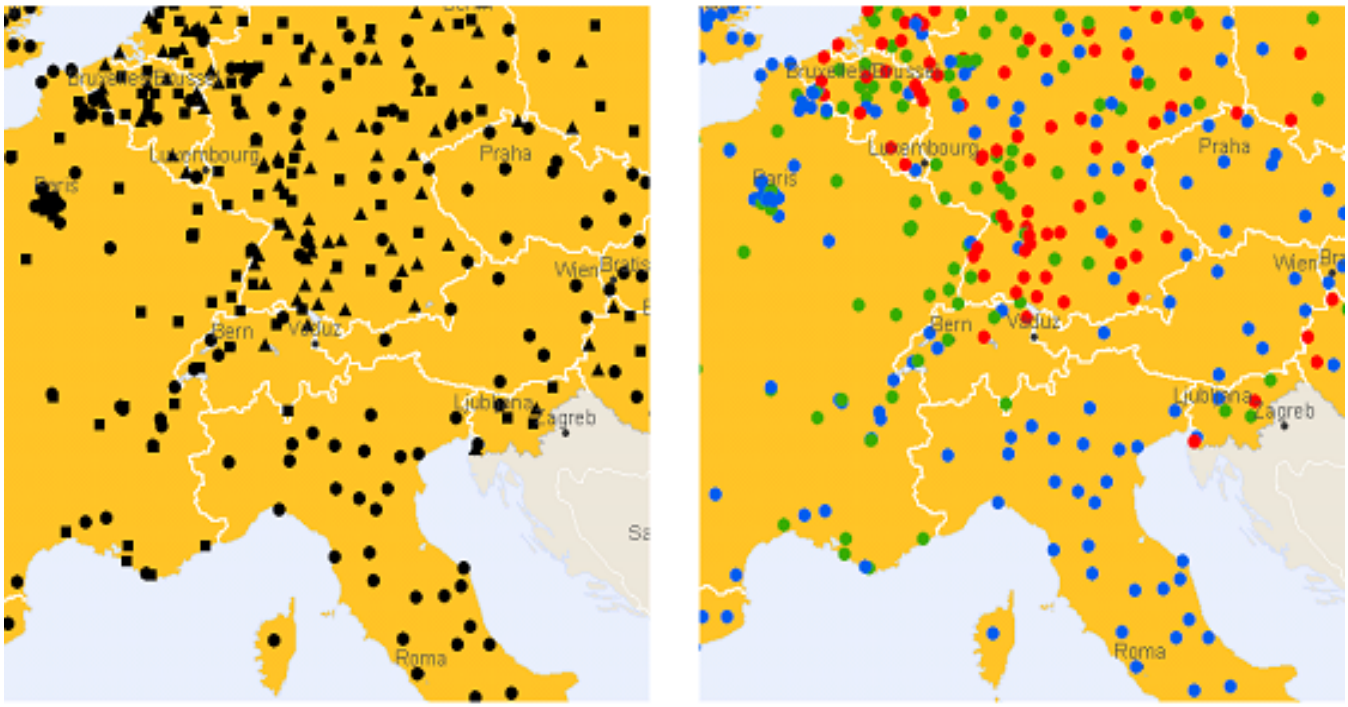


Figure 1 *– Universities by types – Two possibilities*
Good map = points + symbols or points + colours

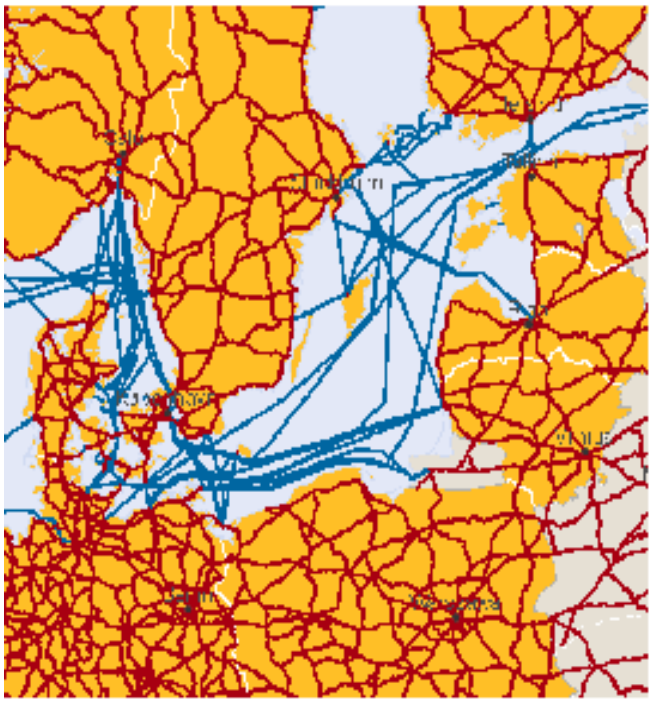


Figure 2 *– Mains roads and ferry connections*
Good map = line + colours

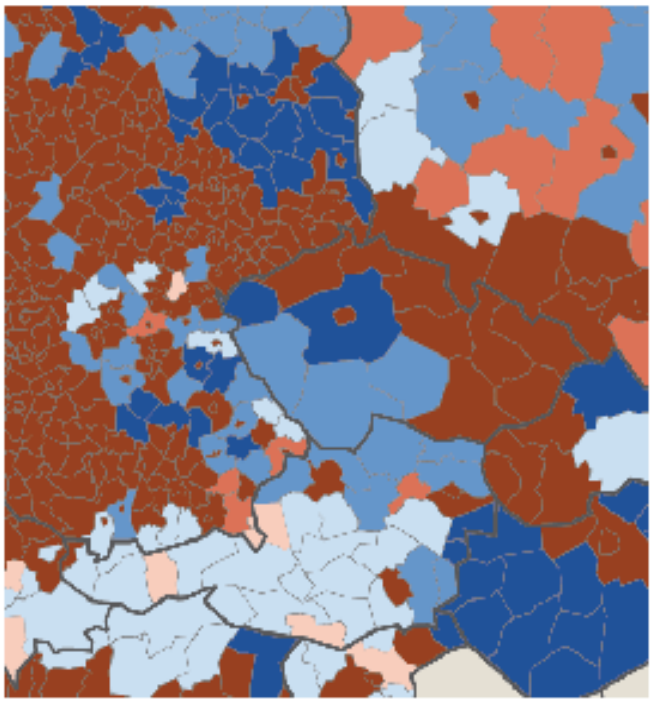


Figure 3 *– Results from Urban-Rural typology*
Good map = areas + colours

2.1.2 Quantitative data with absolute values

Quantitative data with absolute values means concrete **quantity**; the sum of the different values can be calculated and has a real sense. For example, population, GDP, CO2 emissions are absolute quantitative data if we consider the number of inhabitants, number of euros or tons of gas emissions.

For example, considering the different geographical references:

Points: Cities of Europe (number of inhabitants)

Lines: Containers flows across the world (millions tons) – ***Figure 4***

Areas: Population of NUTS 3 – ***Figure 5***

Whatever the type of geographical objects (points, line, areas), the cartography of quantitative data with absolute values has to **respect the quantity** and differences of proportionality. For points or areas objects, the most common representation is to use maps with area **proportional circles**. The circled area is proportional to the size of the data value.

The map showing data in line format (***figure 4***) has to use lines of different width. The width of the line is proportional to the data value.

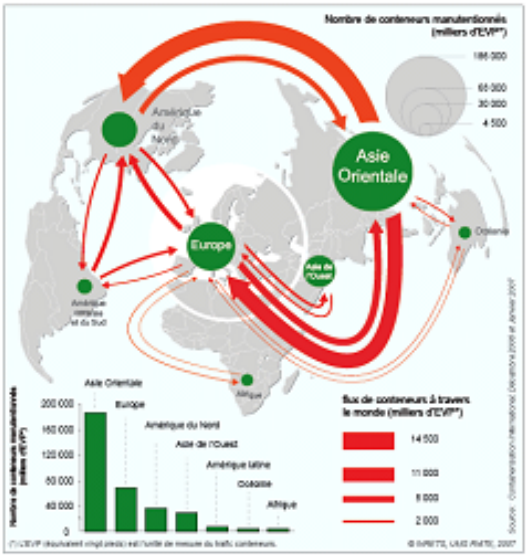


Figure 4 *- Containers flows across the World*
Good map = line + variation of line size

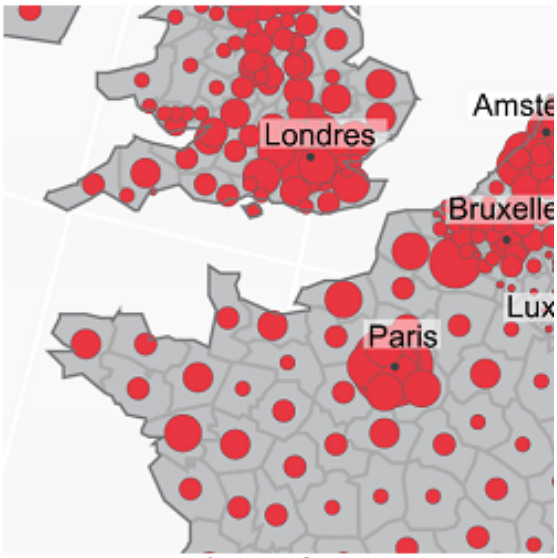


Figure 5 *– Population of NUTS 3 in Europe*
Good map = dot + proportional variation of size

2.1.3 Quantitative data with interval or ratio values

The ratio values are calculated and expressed a series of ratios or proportional values, such as percentage, per km, per inhabitant. This kind of data is the most common.

For example, considering the different geographical references:

Points: Cities of France (cinema attendance index) ***Figure 6***

Lines: GDP per inhabitants discontinuities (relative difference between two territories) – ***Figure 7***

Areas: Abstention, European elections 2009, in Ile-de-France municipalities – ***Figure 8***

For ratios values, the most relevant representation is a choropleth map where density is linked to the class of the data value for each area. The efficiency of the map depends on the range between the leased dense (lightest) area and the densest (darkest) area. When correctly applied, percentage or densities that are twice as high are represented by a grey value that is twice as dark.

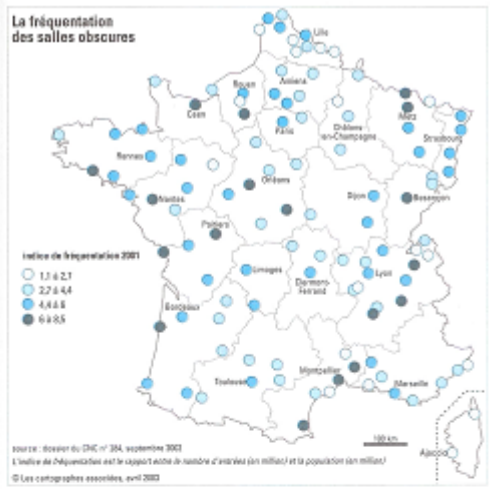


Figure 6: Cinema attendance Index in French main cities

Good map = dots + variation of colours

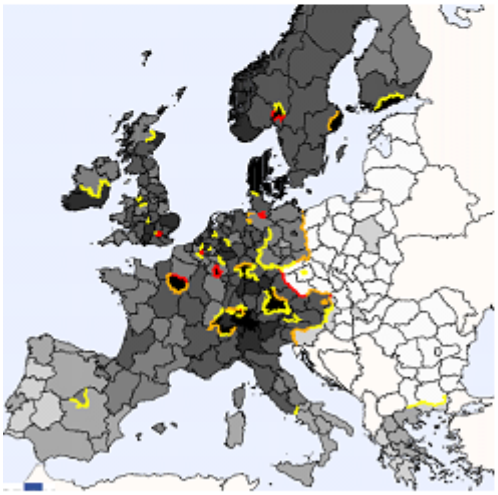


Figure 7: GDP per inhabitants discontinuities

Good map = lines + variation of colours

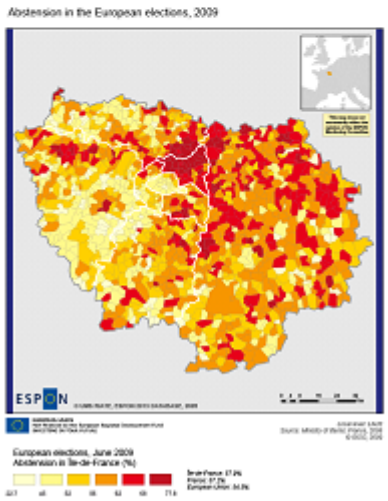


Figure 8: Abstention European votes in Île-de-France

Good map = area + variation of colours

2.1.4 Ordinal or ranked data

Ordinal data are categorical data where there is a logical ordering to the categories. A good example is the Likert scale that you see on many surveys: 1=strongly disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=strongly agree. Another example could be found with modalities like first, second, third etc., or small, medium and high.

For example, considering the different geographical references:

Points: Typology of Functional Urban Areas – MEGA, national FUA, regional FUE

Figure 9

Lines: Road hierarchy – ***Figure 10***

Areas: Degree of policentricity – ***Figure 11***

The representation of these data is based on the expression of natural modalities order. Considering the different geographical references (point, line or area) you can only use 2 graphics variables: grey value or the intensity of a colour. They allow denoting differences in intensity of a phenomenon and expressing order between geographical areas, points or lines. Because differences in grey value or in intensity of colour are used, a hierarchy or order between ordinal modalities can be perceived.

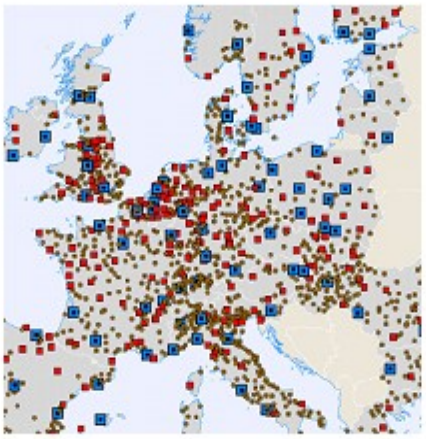


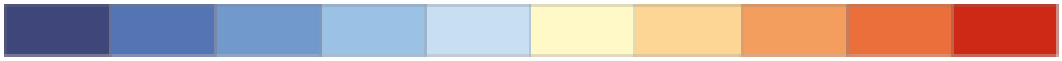
Figure 9: *Typology of Functional Urban Areas*
Good map: points + variation of colour (or size)

Figure 10: *Road hierarchy in Europe*
Good map: lines + variation of colour

Figure 11: *Degree of policentricity in Europe*
Good map: areas + variation of colour

2.2 When using two variations of colour?

It is sometimes necessary to show a phenomenon by a variation of two colours fundamentally different:



This kind of representation is very useful since it allows making more differentiation between the classes of the map. However, it is possible to use these oppositions of colours only if the **break has an objective sense** in the dataset, for instance:

- Opposition between negative and positive values (decrease and increase of population between two periods)
- Values above/under the average value or median value of the dataset (level of accessibility above or under the EU27 average)
- Values above/under a value which have a concrete reality (unemployment rate under/above the threshold of 10 %).

Opposition of variation of two colours should be used only for quantitative data with ratio values and ranked data.

To ensure the **harmonisation** of all maps produced by ESPON projects, it is important that also the use of colours is being guided in the case of opposite colours. In general, it is advised not to combine red and green in one map in order to serve the colour-blind people. Other general rules do not exist. The choice of opposite colors is very subjective and cultural. However, it is quite confusing if two different ESPON maps are published where red has a positive meaning in one of the maps and a negative meaning in the other map. Therefore, in the case of ESPON maps with opposite colours, it is decided to have the following principle as guideline:

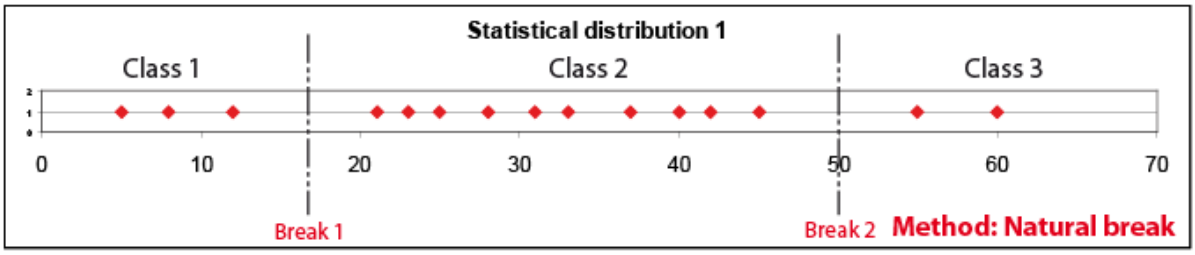
When combining red (warm colours) and blue (cold colours), **red is 'not good'/'negative'** and **blue is 'good'/'positive'"**

2.3 Choice of data ranges

Nevertheless, this kind of representation introduces always a **loss of information** since it transforms a complex statistical distribution into a limited number of classes. Information becomes more generalised and simplified. The accuracy of original values is lost, but **this operation is needed in order to present a synthetic overview of the dataset**. Indeed, a good class division will focus on what is the main content of the dataset, and minimise the loss of accuracy by generalisation. Further below you will find five different classes dividing methods ranging data values. Of course it is also possible to combine different methods, in particular when there are an important number of records. This step before mapping is needed for quantitative values with ratios only.

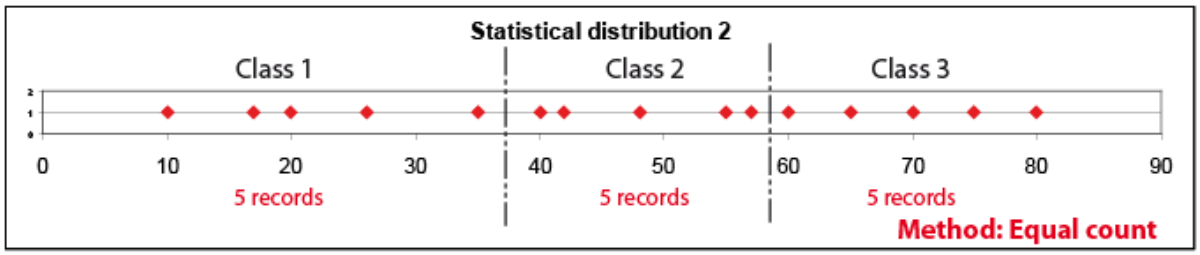
2.3.1 Natural Break

This method sets the breakpoint to "natural points" in the dataset. The strength of this method is that it increases the information content. T**his method is suited when important breaks** describe the dataset.



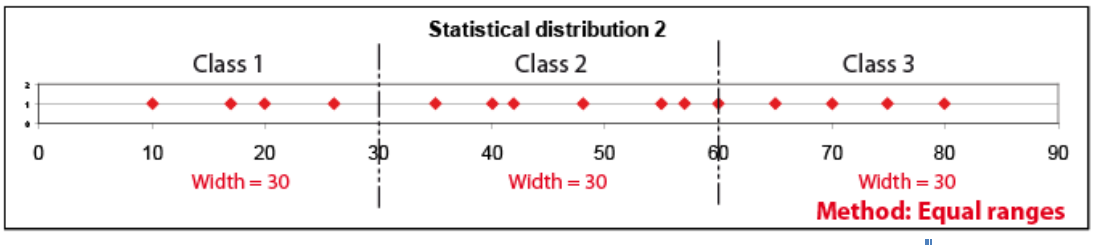
2.3.2 Equal Count or quantile

Equal range contains approximately the same number of records. With 5 classes, each contains 20 % of the total number of the data values. **This method is suited for comparing one dataset with datasets from other themes**. If the data deviate from a linear distribution, the absolute class width will show large variations. Equal count methodology does not take into account exceptional values in the distribution.



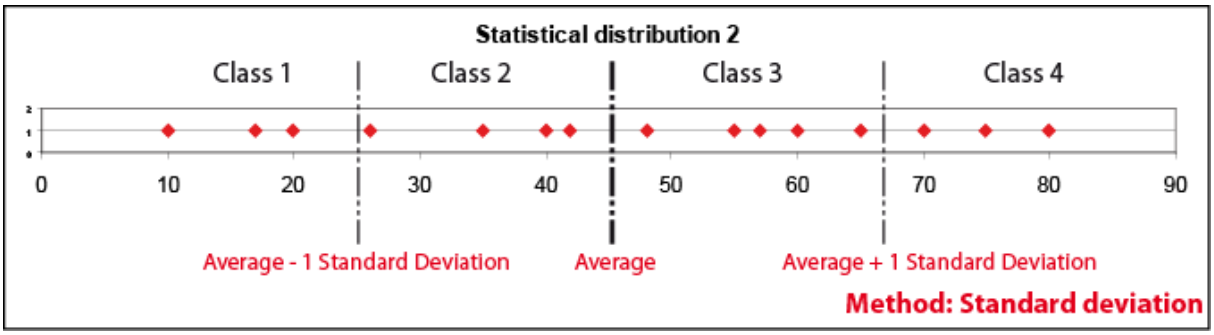
2.3.3 Equal Ranges

The difference between the top and bottom values in each range is the same. This means that we can use values like 0-20; 20-40 etc. or calculate the width of the dataset, and divide by the number of classes wanted. In this case the lowest class will start with the lowest value; the width between the classes will be the same, and the top of the highest value in the dataset. **This method is suited for datasets with a smooth linear distribution.** If the method is used on dataset that are not linear distributed, you will have some classes with many values and others with few or no values.



2.3.4 Standard Deviation (Jenks method)

The class borders are calculated from the mean value and the standard deviation. Standard deviation is a way to describe statistical dispersion. The width of the class is equal to the standard dispersion (or an half depending on the number of classes expected). **This method is suited for normal distributed datasets only**.



2.3.5 Geometric progression

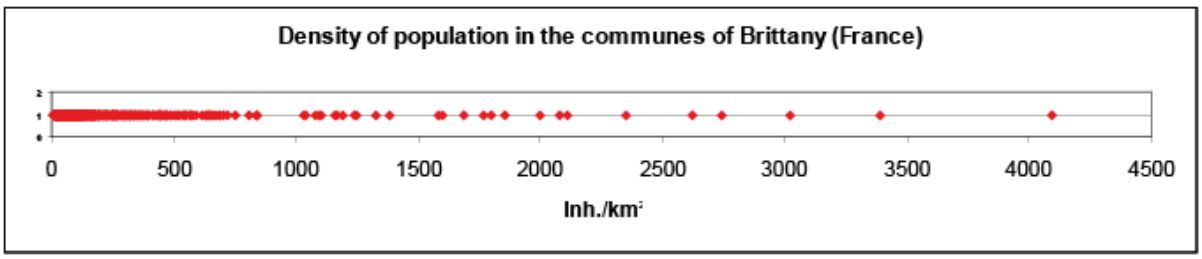
The widths of the class follow a geometric progression. To calculate the width of the different class, it is necessarily to estimate the geometric ratio, such as:

log R = (log10 Max – log10 Min) / number of classes wanted

$R = 10^{\log r}$

Width of the Classes = (min, min x R); (min x R; min x R x R) and so on.

This method is suited for uneven distribution and particularly distribution described by a lot of low values and few high values, such as density of population distribution.



From the example of Brittany, the data ranges, following the geometric progression, should be in 6 classes:

Class	Class boundaries	Number of communes
1	[9;25[	128
2	[25;70[	626
3	[70;190[	343
4	[190;525[	117
5	[525;1470[	39
6	[1470;4100[	15

Whatever the method chosen for ranging the distribution, it is important to use smooth values for the break, in order to understand and memorize easier the sense of the map, e.g. use 30 instead of 29,77; 1500 instead of 1508 etc.

Figure 12 shows the importance of the choice of data range on the visualisation of phenomena.

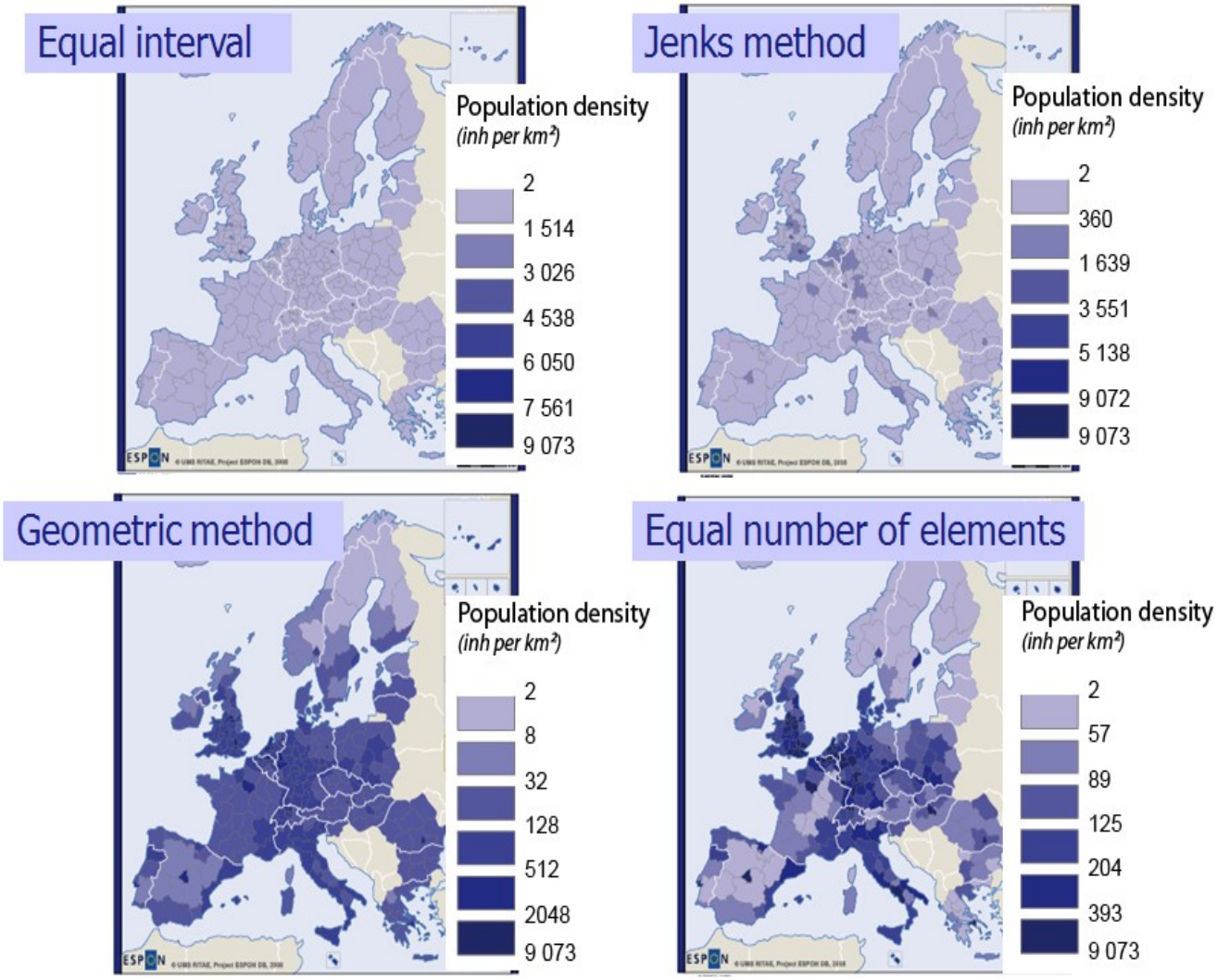


Figure 12*: Result and efficiency are dependent upon the data classification method*

3 Maps are tool for communication

As we explain in the introduction of this technical report: "Maps are perhaps as fundamental to society as language and the written word. They are the preeminent means of recording and communicating information about the location and spatial characteristics of the natural world and of society and culture[2] ".

Maps are produced all over the world and used by people as different as scientists, researchers, scholars, governments or businesses. These maps are most of the time statistical ones connected with the environment, the economy, the politics, the society etc.

The biggest strength of these maps is to allow an effective and relevant communication of the information. However, cartography is a special type of visual communication that does require some preliminary learning: a special purpose language for describing spatial relationships. "The analogy with language also helps explain why training in principles of effective cartography is so important--it allows us to communicate more effectively. Without knowledge of some of these basic principles, the beginning cartographer is likely to be misunderstood or cause confusion[2]".

Of course, cartographers must pay special attention to coordinate systems, map projections, and issues of scale and direction but that's not the first issue of map as a tool for communication. Maps are symbolic abstractions and representations. **The first question when mapping is related to know how to simplify, generalize, represent and symbolize the relationships being represented with graphics symbols**. In other words, what is a good map?

If a design is always more effective than a long speech, the measure of a good map is how well it conveys the right information to its readers and how well it communicates with its audience. This raises a series of questions that must be addresses at the start of a map conception: What is the motive, intent, or goal of the map? Who will read the map? Where will the map be used? What data is available for the composition of the map?

Beyond aesthetic characteristics, the communication also passes by a complete and effective layout: some elements must appear within the base map and the thematic representation, a complete legend, explicit title and source, a precise date of data or even a scale.

[2] *Kenneth E. Foote and Shannon Crum, The Geographer's Craft Project, Department of Geography, The University of Colorado at Boulder*

From data to map, 7 fundamental goals need to be identified to realize a good map:

1. Identify the goal of the map;
2. Identify the audience of the map and where it will be used;
3. Identify the information to be communicated;
4. Identify the geographical reference (point, line or area?);
3. Choose the base map (map projection and scale);
4. Choose the visual variable (symbolic graphic language);
5. Choose layout and identify all the elements to be added.

When these different elements are not correctly taking into account, the map will be characterised probably by some mistakes and misunderstandings.

3.1 Bad choices in term of representation of the data

Most of the problems of visualisation and map design are generally linked to **bad choices in term of representation of the data** (cf. part 2 of the technical report). When comparing ***figures 13 and 14***, which represent the same information, e.g. a typology showing age structure and total population in the municipalities from Brittany (France), it is quite clear that the second map is really clearer than the first one. Two main reasons can explain it (***figure 13***):

- Absolute values (e.g. total population in 2000) don't have to be shown by variation of intensity of black (hachure). This kind of representation does not respect the ratio of proportionality of the indicator, which is fundamental and needed information. Using hachure is also a visual mistake; the map is not readable at all and the representation is not the most efficient. These data have to be shown by proportional symbols, circles for instance.
- This typology, derived from age structure cannot be considered as a qualitative data, since there is an implicit order when considering the progression in term of age. In concrete terms, showing each class by a different colour is not the best solution. To show correctly this data it is important to think about the goal of the map. Here, it is important to represent the municipalities described by high share of young, active and old people. As a consequence, it is important to differentiate these information (3 colours) and also to make possible the analyse of the graduation of the phenomenon (high/medium shares), e.g. using variation of intensity of these 3 colours.

The solution proposed in **figure 14** try to correct these different elements. The most adapted solution for the representation of these data is to combine circles and colours in order to make the map as clear as possible. On top of that, it allows nuancing the interpretation of the map, e.g. Brittany is a region where ageing is important, but it concerns specific small and rural cities.

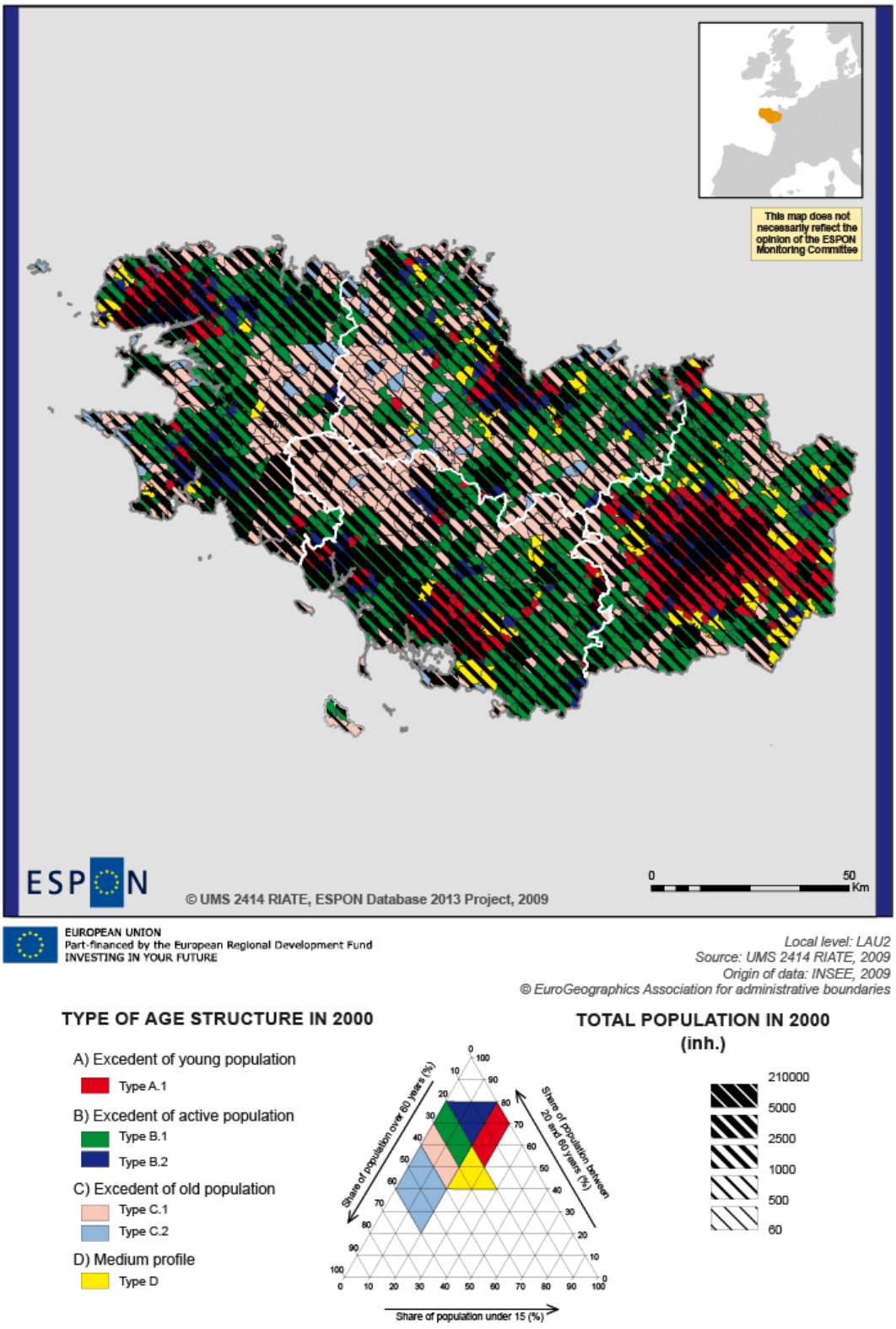


Figure 13**: Population and age structure in Brittany (France) – **with semiologic problems

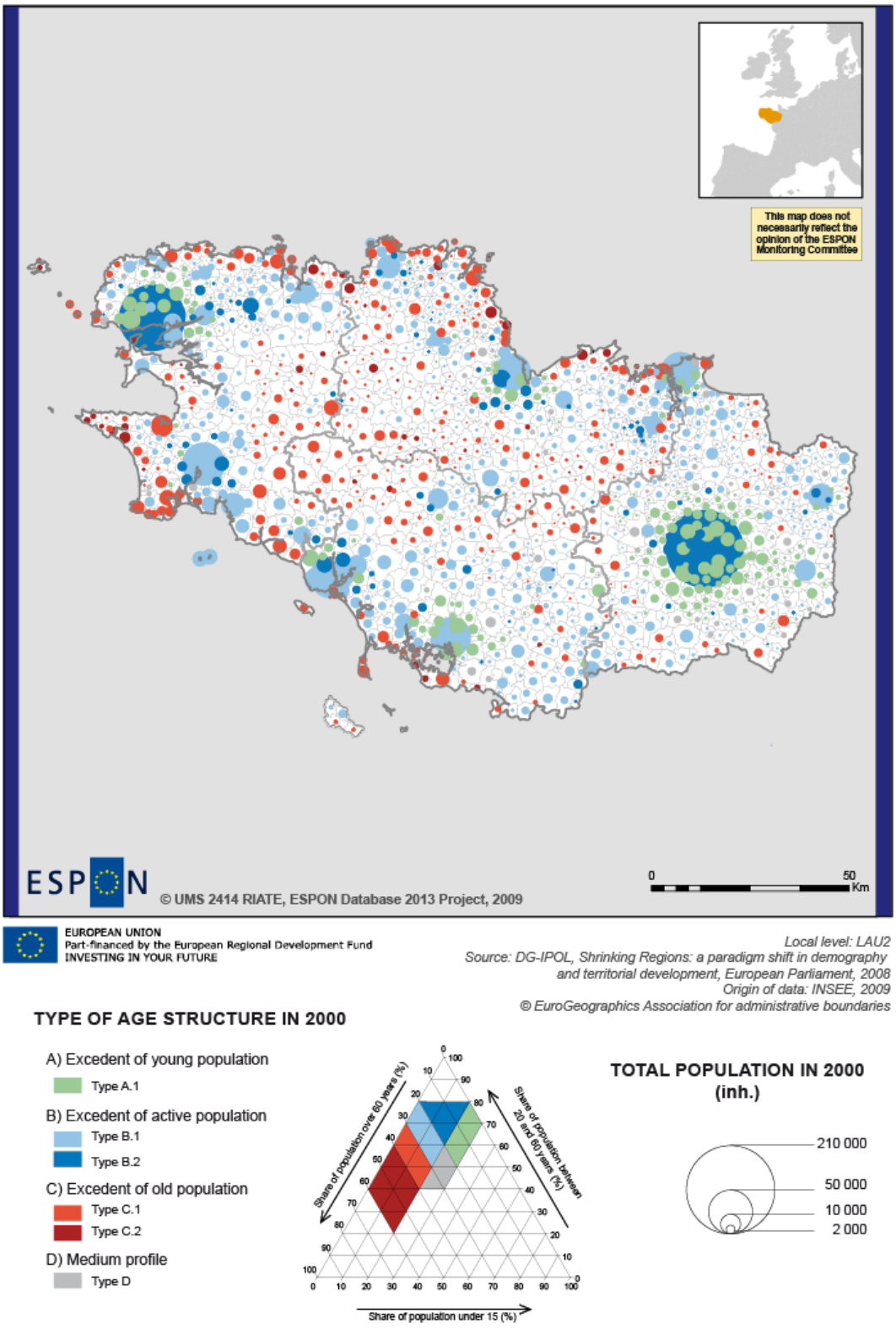


Figure 14**: Population and age structure in Brittany (France) – **without semiologic problems

3.2 Improving the efficiency of the map

Other problem which appears regularly is the degree of complexity of the map. The aim of the maps is to be synthetic. When representing too much information, the eye cannot distinguish the different elements of the map. This kind of figure can be solved by thinking to the design of the map: where is the best location for legend? How using with the most efficiency the place available?

The figures 15 and 16 show the same information, e.g. a typology of population development by components during the period 1995-2004 in EU27; this data is crossed with expected population evolution in 2030.

Figure 15 proposes solution which is correct in term of graphic semiology: ordinal data are shown by variation of colour (green/red) and shrinking/non shrinking regions (qualitative data) are represented by the opposition of hachure and no hachure. However, the combination of these two visual variables makes the map hard to interpret and the message become not so clear!

When there is too much information it becomes difficult to be able to synthesise the message of the map. That is why in some cases it is more efficient to split information in two maps instead of concentrating all the elements in a single one. This has been done on figure 16, where the map located on left of the document shows the regions described by an expected growth of population; and the map on the right shows the regions where a demographic decrease is planned. This template allows immediately to observe that during the period 1995-2005 most of the 'shrinking regions' have witnessed a downturn linked to both natural change and a negative migratory balance.

There is never an optimal solution

Whatever the examples proposed and demonstrated, it is important to keep in mind that there is never a single solution to show information on maps. In fact, each person has his own perception when interpreting graphic documents or pictures. **Map is always a compromise**. But during the creation of the map, is fundamental to try to make the map as understandable as possible. In concrete terms, it is not an obvious task and it is kindly recommended to make different attempts and share the results with other colleagues before saying "OK, my map is ready for the report"!

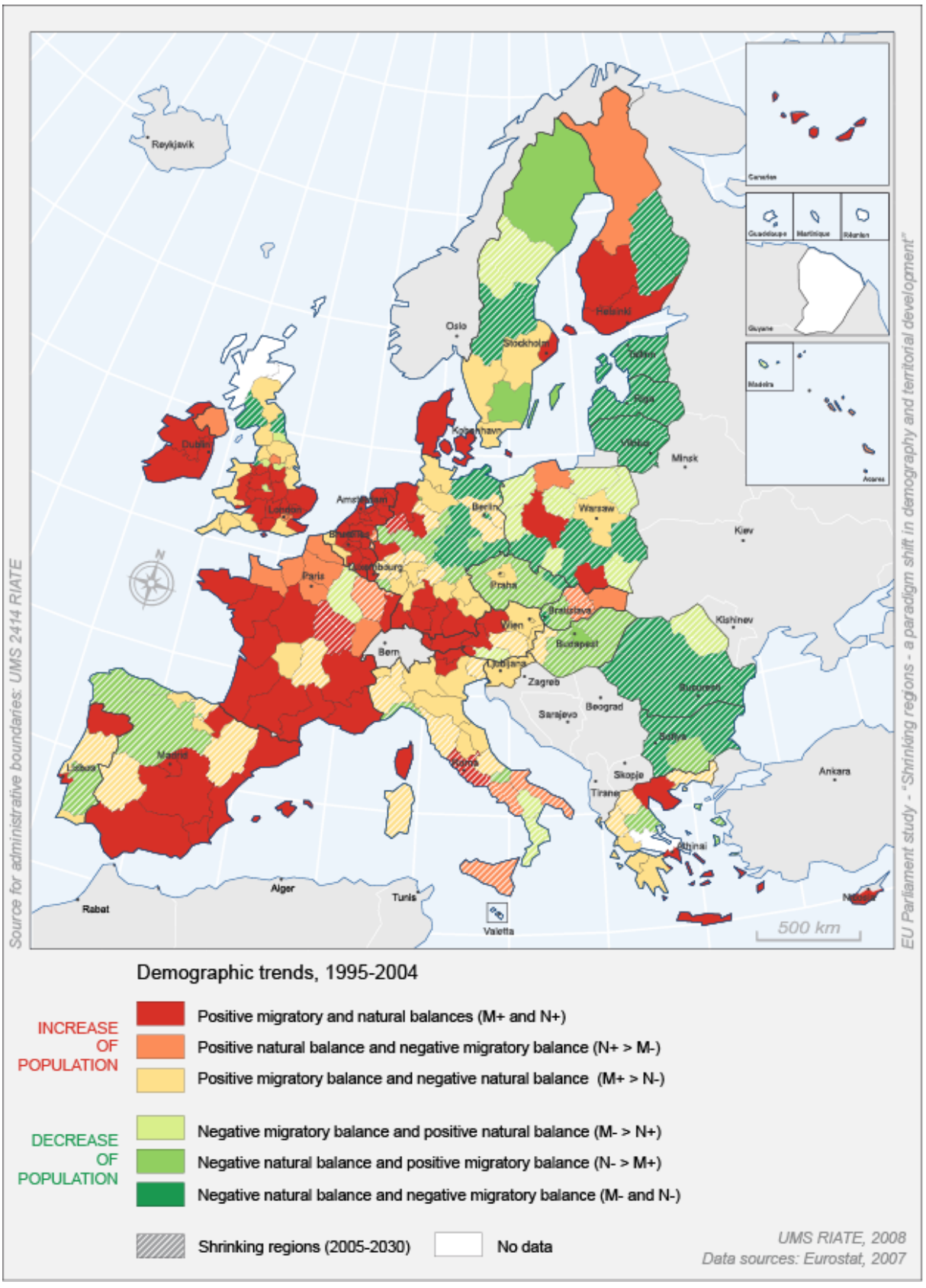


Figure 15**: Typology of regional growth patterns – **Possibility 1

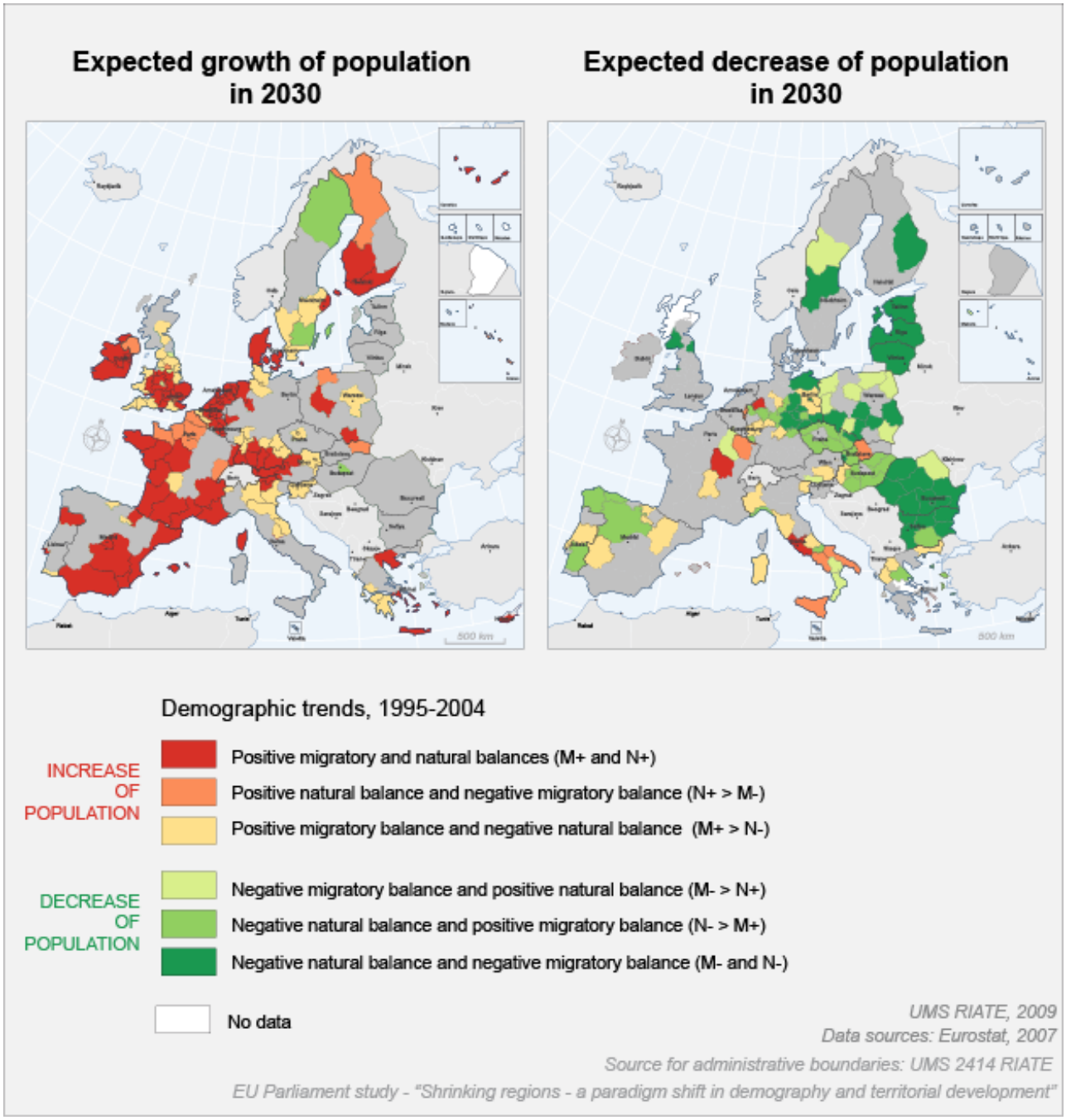


Figure 15**: Typology of regional growth patterns – **Possibility 2

ANNEXES

These annexes allow you to choose some efficient graphic variables to communicate differences in size, order or quality.

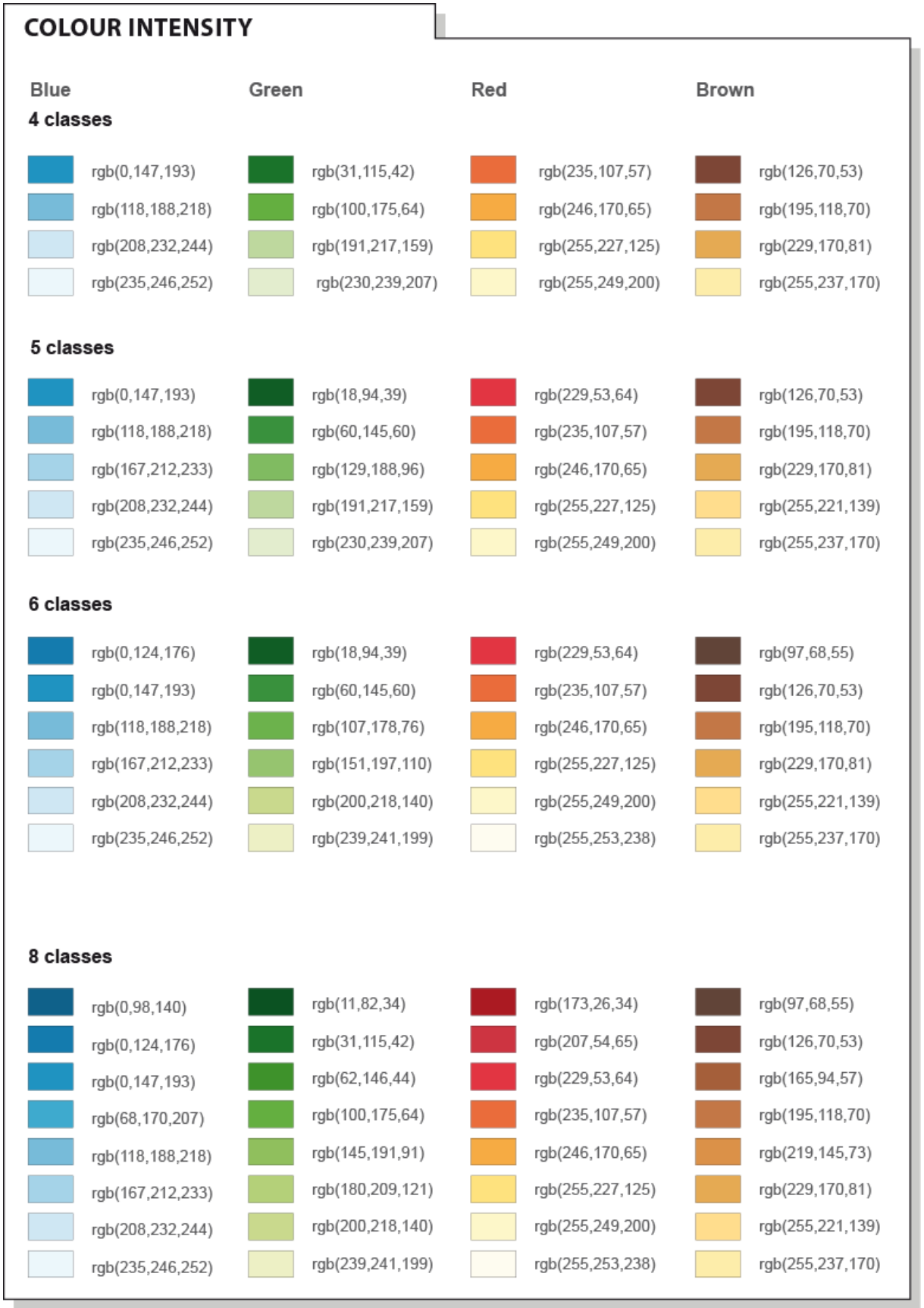
ANNEXE 1 - Relation of graphical variables to perceptual characteristics

Graphical variable	Type of data			
	nominal	ordinal	Interval/ratio	quantity
Size		x	x	x
Grey or colour value		x	x	
Grain/texture		x	x	
Colour hue	x			
Orientation	x			
Shape	x			

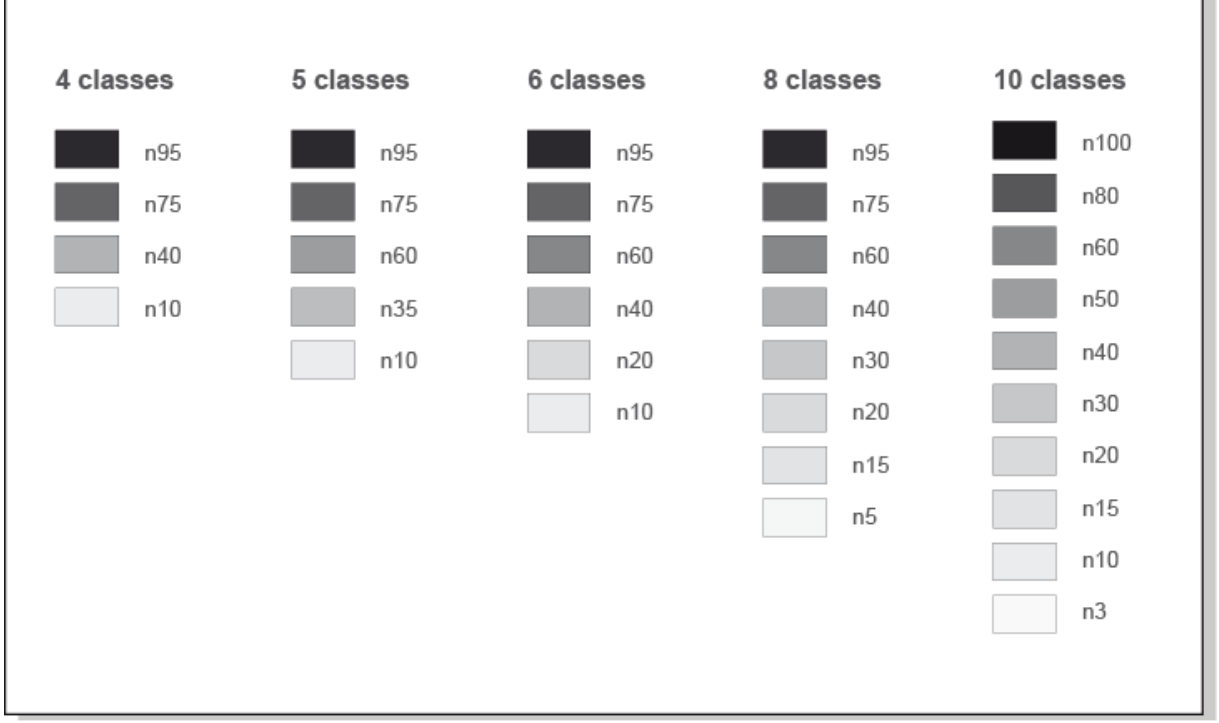
ANNEXE 2 - Numbers of categories that can be perceived at a glance

Graphical variable	Point	Line	Area
Size	4	4	5
Grey or colour value	3	4	5
Grain/texture	2	4	5
Colour hue	7	7	8
Orientation	4	2	4
Shape	3	3	3

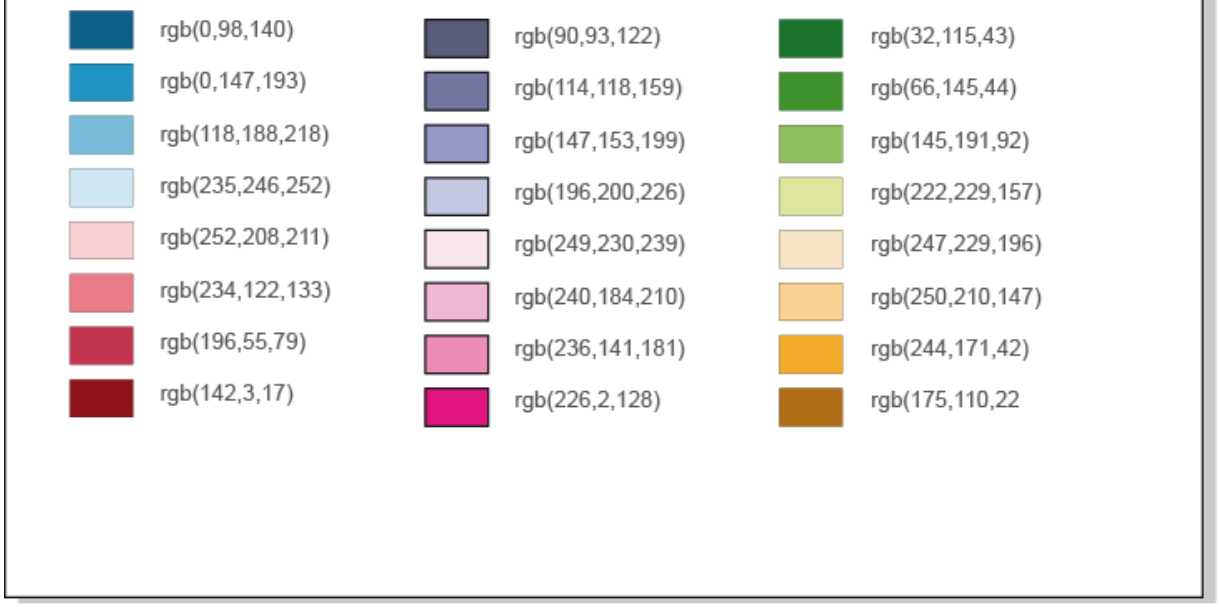
ANNEXE 3: Differences in value or lightness



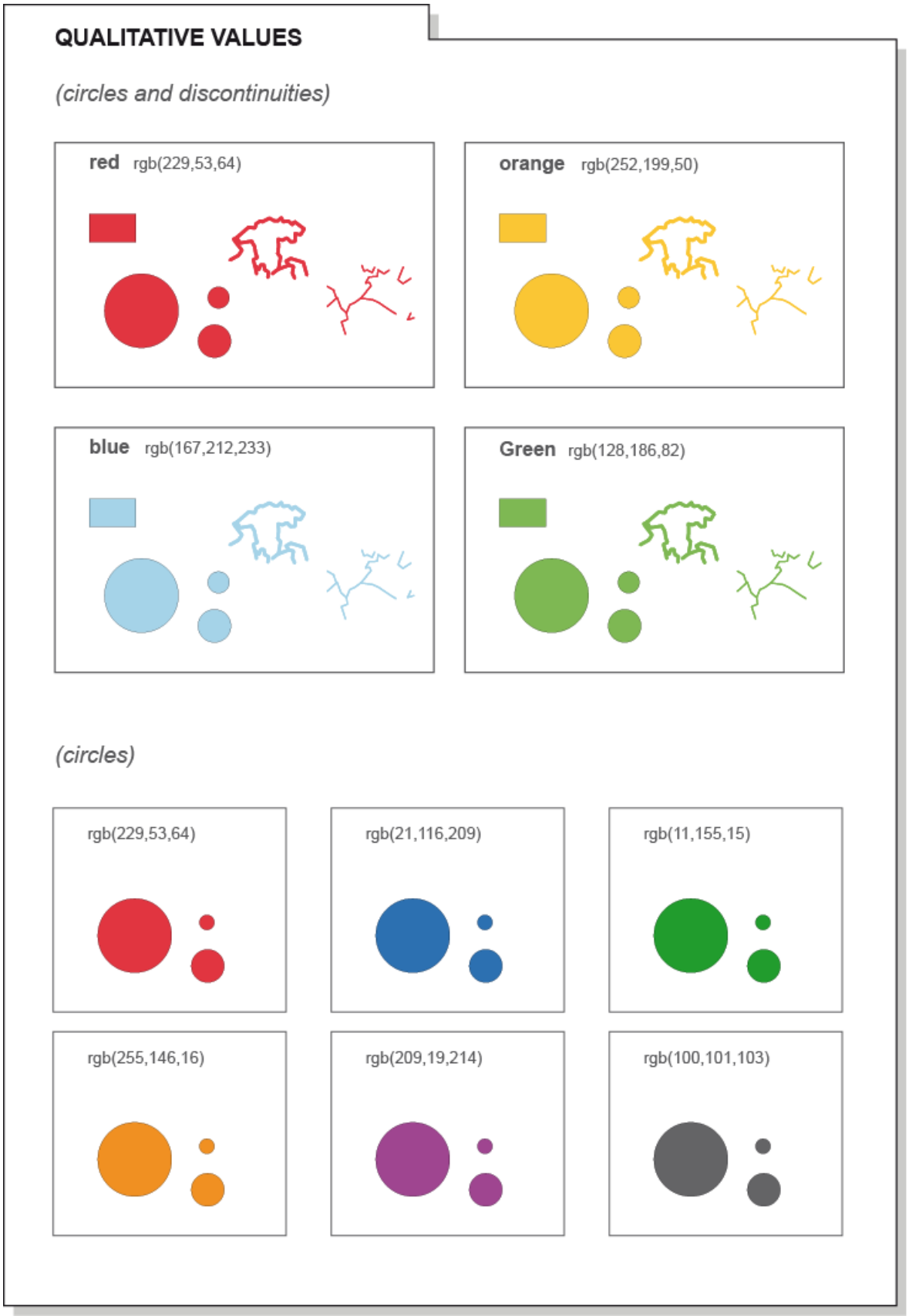
GREY VALUE



OPPOSITE COLOURS



ANNEXE 4: Colours for differences typology or qualitative value



References

- ***Litterature***

Béguin M., Pumain D., 2003, *La représentation des données géographiques – statistique et cartographie,* Armand Colin.

Bertin J., 1967, *Sémiologie graphique*, Gauthiers-Villars.

Cambrezy L., de Maximy R. (Ed.), 1995, La cartographie en débat, représenter ou convaincre, Editions Kathala et Orstom, Paris

Harris R. L., 1996, Information graphics, a comprehensive illustrated reference, visual tools for analysing, managing and communicating, Management Graphics ed., USA

Harley, J. B., 1988, Maps, knowledge and power. In COSGROVE, D. (Ed.) The Iconography of Landscape.Cambridge, MA, Cambridge University Press.

Kraak M.-J., Ormeling F., 2003, *Cartography, Visualization of Geospatial Data,* 2nd edition, Pearson Education, Prentice Hall.

Kraak, M.-J., 1998, Exploratory cartography, map as tools for discovery, *ITC Journal* (1), pp.46-54

MacEachren A.M., 1994, *Some truth with maps: a primer on design and symbolization*, Association of American Geographers, Washington DC.

Monmonnier M., 1996, *How to lie with maps*, University of Chicago Press.

Robinson A.H., Morrison J.L., Muehrcke P.C., 1995, *Elements of cartography*, New York, J.Willey & Sons.

Wilkinson L., 1999, The grammar of graphics. New York, Springer.

Wood, D., 1992, The Power of Maps. New York, The Guildford Press.

Wood C. H., Keller C. P., 1996, Cartographic design: theoretical and practical perspectives, Wiley, USA

Zanin C., Trémélo M-L, 2003, *Savoir faire une carte: Aide à la conception et à la réalisation d'une carte thématique univariée,* Belin.

- ***Websites***

***Colorbrewer* 2.0** is an online tool designed to help people select good color schemes for maps and other graphics: http://colorbrewer2.org/

Philcarto is a free tool for cartography, available on the net: http://philcarto.free.fr/

Quantum GIS is an Open Source Geographic Information System. It runs on Linux, Unix, Mac OSX, and Windows and supports numerous vector, raster, and database formats and functionalities: http://www.qgis.org/

TECHNICAL REPORT

ESPON

MAPPING GUIDE

CARTOGRAPHY IN ESPON 2013

CONTENT

- Enhancing information. This part explain how symbolize ESPON 2013 data with the good rules of graphic semiology.
- Maps are tool for communication. This part insists on the fact that a map has necessarily to deliver a clear message.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

24 PAGES

LIST OF AUTHORS

Christine Zanin, University Paris 7, UMS 2414 RIATE

Nicolas Lambert, UMS 2414 RIATE

Ronan Ysebaert, UMS 2414 RIATE

Contact

christine.zanin@univ-paris-diderot.fr
nicolas.lambert@ums-riate.fr
ronan.ysebaert@ums-riate.fr

tel. + 33 1 57 27 65 32

TABLE OF CONTENT

Introduction

Maps are a great way of displaying statistical data. It allows summarizing a complex and important information into clear and compact presentation. They can bring a great help in spotting patterns within data.

Maps are accessible for many reasons. People understand maps (at least, think they do). People like maps because they attract attention and brighten up presentation. Nevertheless, and in a scientific versus, the interest of the representation of geographical information on maps can be summarized in three main points1.

The localisation is the most elementary subject related to geographic information. It allows answering to question "Where can we find this phenomenon?" The precision of the localisation depends on the quality of this kind of information such as statistical databases, statistical yearbook and so on. Locate a geographical object has generally a sense only if it is possible to compare it to other one "Why this object is located here and not there?". Answers can be read off directly from the map without any other help.

The comparison: Geographical objects analysis makes a concrete sense when it is possible to compare them. "What is the situation of this region as compare to the other one?"; "Can we observe geographical pattern, such as discontinuities, concentration?" Maps are useful tools for interpreting and pointing out specific geographical patterns, which are impossible to catch with an only statistical analysis.

Planning: Since the relations between European territories are very intensive, territorial planning on a special location must interfere with other territories and have to.

Despite many interests to use maps within ESPON, these kinds of documents have also their limits. Maps always generalise and simplify information. Mapping is more than just rendering; it also getting to know the phenomenon which is to be mapped. That's why mapping is not an easy action. Deliver the right message must remain the first objective of map design and mapping allows you to orchestrate the elements of the map to best convey its message to its audience. Thus, the design of maps is mainly concerned with making choices: the choice of mapping method (proportional symbol or choropleth map, isoline or grid map or even a cartogram), the choice of the aggregation level on which information as to be depicted, the choice on the level of statistic areas and the type of data (absolute or relative representation), the choice of graphic variables (such as differences in size, value, grain, colour, direction and shape) to be used. These choices are fundamental's one, they influence people's conception and visualisation of space.

This technical report is not a formal cartography book but allows everyone to understand easily how to produce an effective and operational map in the ESPON 2013 program. The report is organized in two parts: (i) Enhancing information (mapping methods and graphic semiology); (ii) Maps and communication (map is to deliver a simple and clear message).

[1] Béguin M., Pumain D., 2003, *La représentation des données géographiques – statistique et cartographie,* Armand Colin, 192p.

1 Enhancing information

1.1 Differentiation of data type

Many possibilities exist to show data on map. Choosing relevant representation is not an obvious task and has to be considered seriously. Indeed, choosing the wrong type of map can completely misrepresent the data. It is important to keep in mind that **the choice in cartography is always dependant on the type of data**. It is possible to identify four main types of data:

1. **Qualitative data**
2. **Quantitative data with absolute values**
3. **Quantitative data with ratios values**
2. **Ordinal (or ranked) data**

For each type of data it is possible to relate it to a **geographical reference: points, lines or areas.**

There are many possibilities to show correctly data on maps. The aim of this paper is not to present all types of correct visualisation, but an extract of the most usual and efficient ones.

1.1.1 Qualitative data

A data is qualitative when its value is a nominal one with qualitative differences: components do not allow establishing range relations between them.

For example, considering the different geographical references:

Points: location universities by type (university, polytechnics…) – ***Figure 1***

Lines: communication network without hierarchy (ferry connections, main roads) – ***Figure 2***

Areas: results from typology (rural area, urban area…) – ***Figure 3***

Qualitative data have to be shown such a manner that do not suggest rank either quantity. Two possibilities: use **geometric symbols** or **differential colour** in order to **differentiate** the different elements of the map.

With points (figure 1) the most efficient is to show information by colour or geometric symbols. It is important to use a limited quantity of symbols or colours to make the map understandable.

For lines or areas, differential colours should be used (figure 2 and 3).

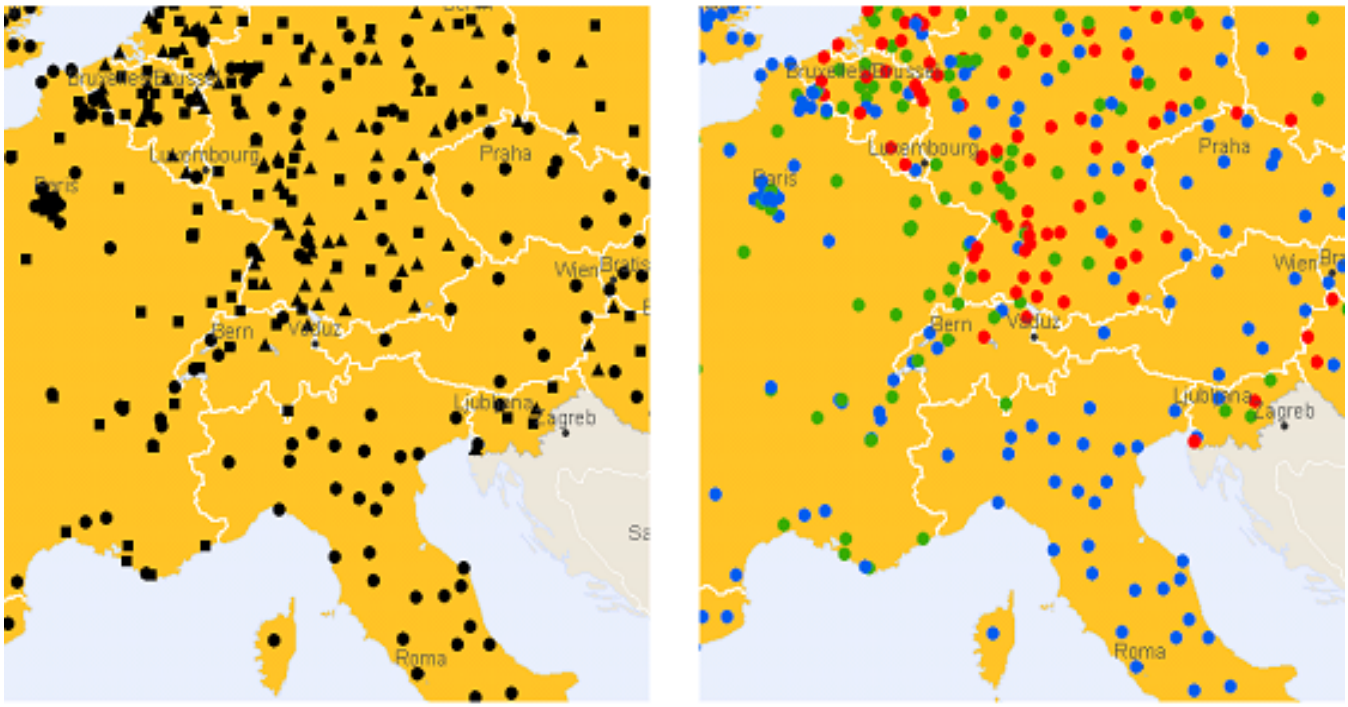


Figure 1 *– Universities by types – Two possibilities*
Good map = points + symbols or points + colours

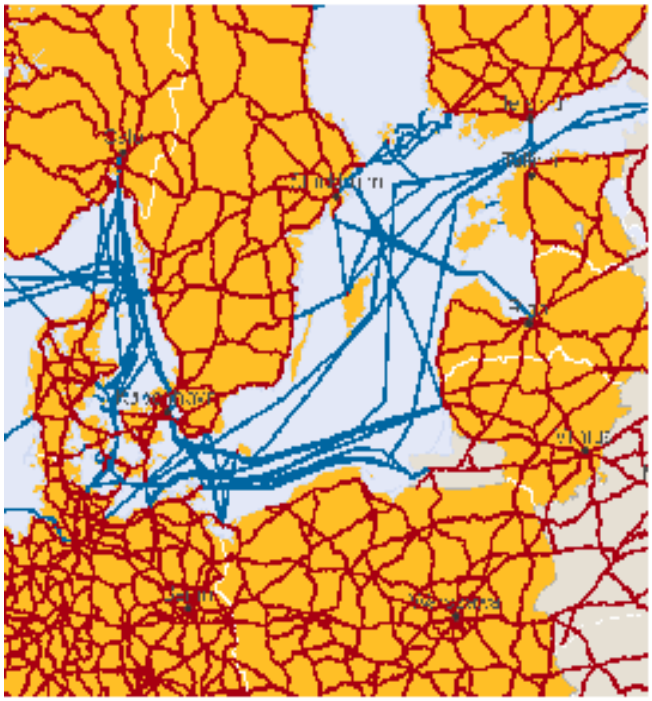


Figure 2 *– Mains roads and ferry connections*
Good map = line + colours

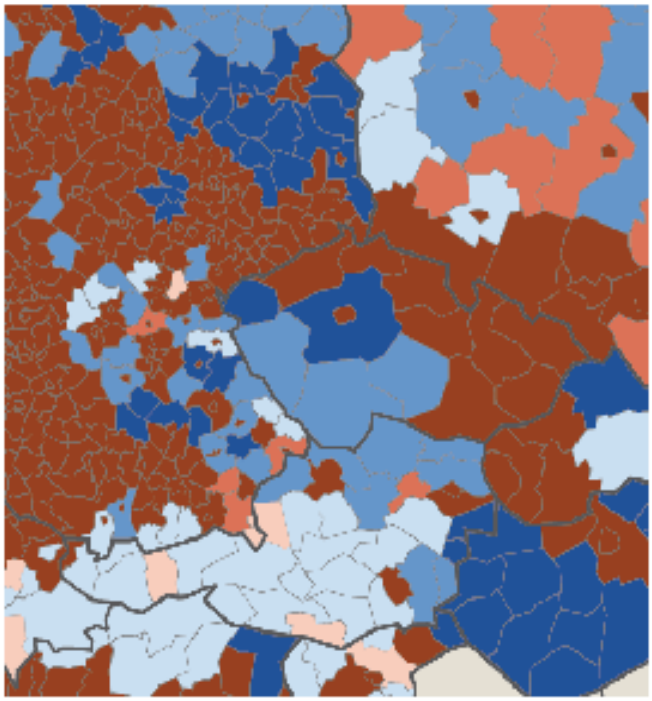


Figure 3 *– Results from Urban-Rural typology*
Good map = areas + colours

1.1.2 Quantitative data with absolute values

Quantitative data with absolute values means concrete **quantity**; the sum of the different values can be calculated and has a real sense. For example, population, GDP, CO2 emissions are absolute quantitative data if we consider the number of inhabitants, number of euros or tons of gas emissions.

For example, considering the different geographical references:

Points: Cities of Europe (number of inhabitants)

Lines: Containers flows across the world (millions tons) – ***Figure 4***

Areas: Population of NUTS 3 – ***Figure 5***

Whatever the type of geographical objects (points, line, areas), the cartography of quantitative data with absolute values has to **respect the quantity** and differences of proportionality. For points or areas objects, the most common representation is to use maps with area **proportional circles**. The circled area is proportional to the size of the data value.

The map showing data in line format (***figure 4***) has to use lines of different width. The width of the line is proportional to the data value.

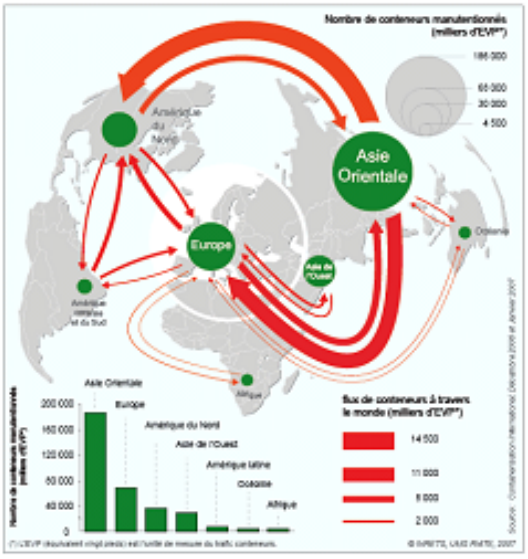


Figure 4 *- Containers flows across the World*
Good map = line + variation of line size

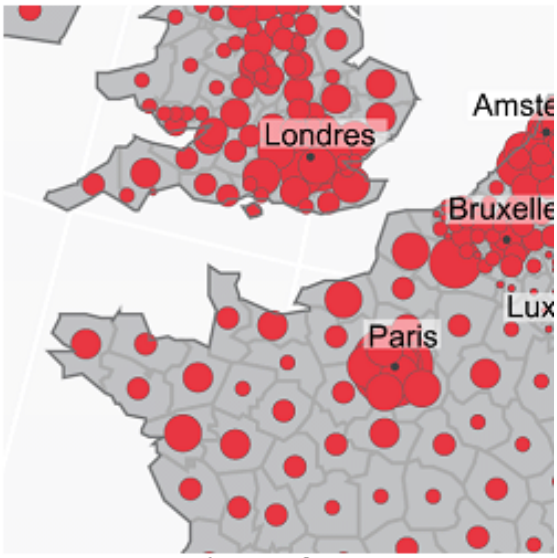


Figure 5 *– Population of NUTS 3 in Europe*
Good map = dot + proportional variation of size

1.1.3 Quantitative data with interval or ratio values

The ratio values are calculated and expressed a series of ratios or proportional values, such as percentage, per km, per inhabitant. This kind of data is the most common.

For example, considering the different geographical references:

Points: Cities of France (cinema attendance index) ***Figure 6***

Lines: GDP per inhabitants discontinuities (relative difference between two territories) – ***Figure 7***

Areas: Abstention, European elections 2009, in Ile-de-France municipalities – ***Figure 8***

For ratios values, the most relevant representation is a choropleth map where density is linked to the class of the data value for each area. The efficiency of the map depends on the range between the leased dense (lightest) area and the densest (darkest) area. When correctly applied, percentage or densities that are twice as high are represented by a grey value that is twice as dark.

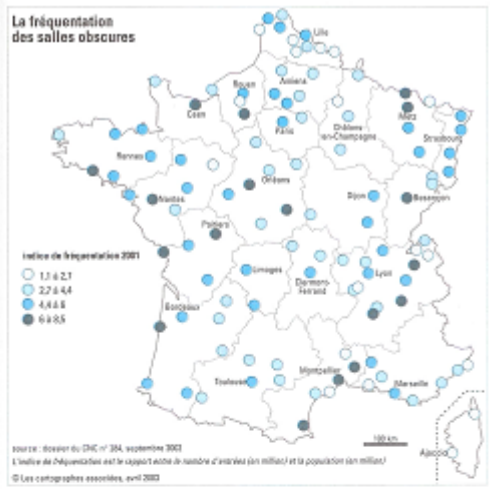


Figure 6: Cinema attendance Index in French main cities

Good map = dots + variation of colours

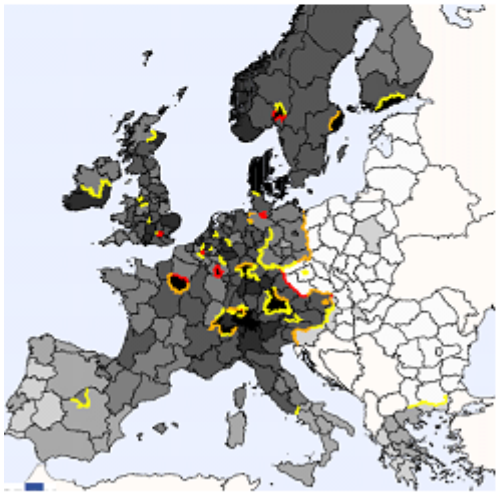


Figure 7: GDP per inhabitants discontinuities

Good map = lines + variation of colours

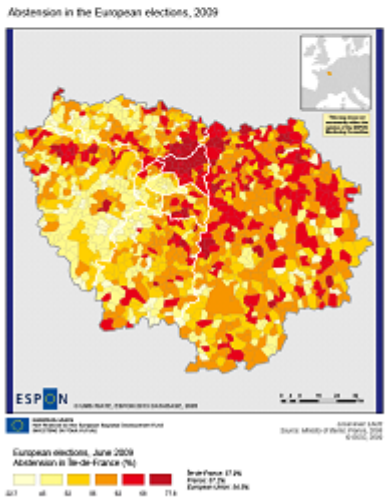


Figure 8: Abstention European votes in Île-de-France

Good map = area + variation of colours

1.1.4 Ordinal or ranked data

Ordinal data are categorical data where there is a logical ordering to the categories. A good example is the Likert scale that you see on many surveys: 1=strongly disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=strongly agree. Another example could be found with modalities like first, second, third etc., or small, medium and high.

For example, considering the different geographical references:

Points: Typology of Functional Urban Areas – MEGA, national FUA, regional FUE

Figure 9

Lines: Road hierarchy – ***Figure 10***

Areas: Degree of policentricity – ***Figure 11***

The representation of these data is based on the expression of natural modalities order. Considering the different geographical references (point, line or area) you can only use 2 graphics variables: grey value or the intensity of a colour. They allow denoting differences in intensity of a phenomenon and expressing order between geographical areas, points or lines. Because differences in grey value or in intensity of colour are used, a hierarchy or order between ordinal modalities can be perceived.

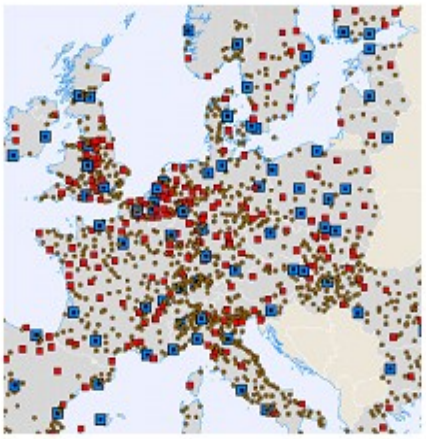


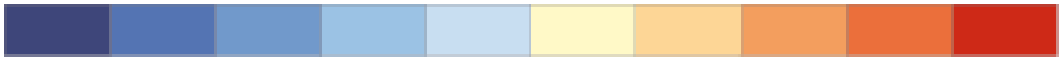
Figure 9: *Typology of Functional Urban Areas*
Good map: points + variation of colour (or size)

Figure 10: *Road hierarchy in Europe*
Good map: lines + variation of colour

Figure 11: *Degree of policentricity in Europe*
Good map: areas + variation of colour

1.2 When using two variations of colour?

It is sometimes necessary to show a phenomenon by a variation of two colours fundamentally different:



This kind of representation is very useful since it allows making more differentiation between the classes of the map. However, it is possible to use these oppositions of colours only if the **break has an objective sense** in the dataset, for instance:

- Opposition between negative and positive values (decrease and increase of population between two periods)
- Values above/under the average value or median value of the dataset (level of accessibility above or under the EU27 average)
- Values above/under a value which have a concrete reality (unemployment rate under/above the threshold of 10 %).

Opposition of variation of two colours should be used only for quantitative data with ratio values and ranked data.

To ensure the **harmonisation** of all maps produced by ESPON projects, it is important that also the use of colours is being guided in the case of opposite colours. In general, it is advised not to combine red and green in one map in order to serve the colour-blind people. Other general rules do not exist. The choice of opposite colors is very subjective and cultural. However, it is quite confusing if two different ESPON maps are published where red has a positive meaning in one of the maps and a negative meaning in the other map. Therefore, in the case of ESPON maps with opposite colours, it is decided to have the following principle as guideline:

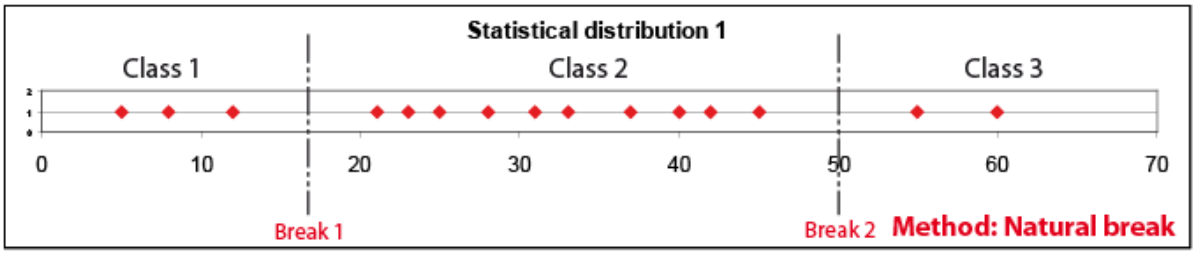
When combining red (warm colours) and blue (cold colours), **red is 'not good'/''negative'** and **blue is 'good'/''positive'"**

1.3 Choice of data ranges

Nevertheless, this kind of representation introduces always a **loss of information** since it transforms a complex statistical distribution into a limited number of classes. Information becomes more generalised and simplified. The accuracy of original values is lost, but **this operation is needed in order to present a synthetic overview of the dataset**. Indeed, a good class division will focus on what is the main content of the dataset, and minimise the loss of accuracy by generalisation. Further below you will find five different classes dividing methods ranging data values. Of course it is also possible to combine different methods, in particular when there are an important number of records. This step before mapping is needed for quantitative values with ratios only.

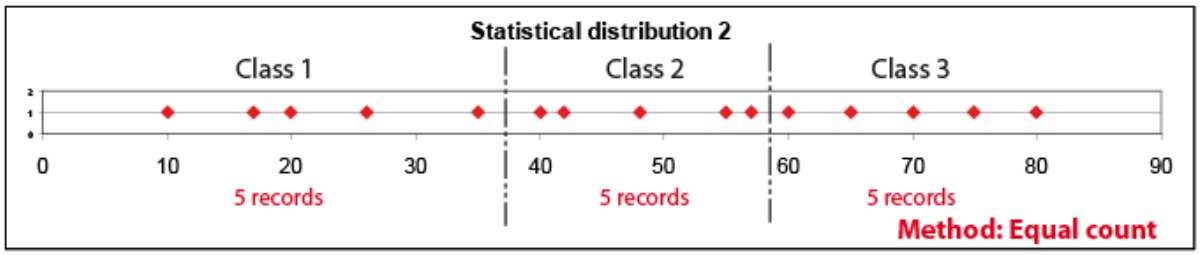
1.3.1 Natural Break

This method sets the breakpoint to "natural points" in the dataset. The strength of this method is that it increases the information content. T**his method is suited when important breaks** describe the dataset.



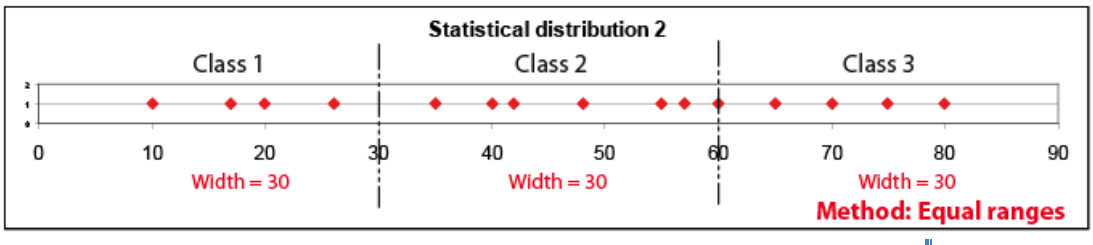
1.3.2 Equal Count or quantile

Equal range contains approximately the same number of records. With 5 classes, each contains 20 % of the total number of the data values. **This method is suited for comparing one dataset with datasets from other themes**. If the data deviate from a linear distribution, the absolute class width will show large variations. Equal count methodology does not take into account exceptional values in the distribution.



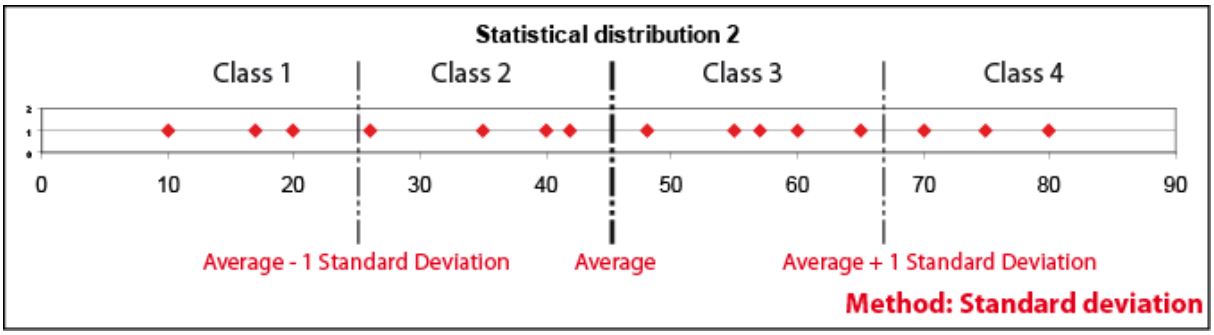
1.3.3 Equal Ranges

The difference between the top and bottom values in each range is the same. This means that we can use values like 0-20; 20-40 etc. or calculate the width of the dataset, and divide by the number of classes wanted. In this case the lowest class will start with the lowest value; the width between the classes will be the same, and the top of the highest value in the dataset. **This method is suited for datasets with a smooth linear distribution.** If the method is used on dataset that are not linear distributed, you will have some classes with many values and others with few or no values.



1.3.4 Standard Deviation (Jenks method)

The class borders are calculated from the mean value and the standard deviation. Standard deviation is a way to describe statistical dispersion. The width of the class is equal to the standard dispersion (or an half depending on the number of classes expected). **This method is suited for normal distributed datasets only**.



1.3.5 Geometric progression

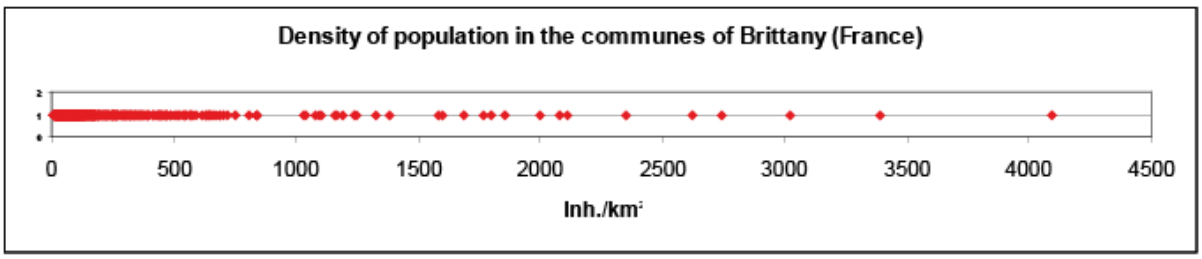
The widths of the class follow a geometric progression. To calculate the width of the different class, it is necessarily to estimate the geometric ratio, such as:

log R = (log10 Max – log10 Min) / number of classes wanted

$R = 10^{\log r}$

Width of the Classes = (min, min x R); (min x R; min x R x R) and so on.

This method is suited for uneven distribution and particularly distribution described by a lot of low values and few high values, such as density of population distribution.



From the example of Brittany, the data ranges, following the geometric progression, should be in 6 classes:

Class	Class boundaries	Number of communes
1	[9;25[	128
2	[25;70[	626
3	[70;190[	343
4	[190;525[	117
5	[525;1470[	39
6	[1470;4100[	15

Whatever the method chosen for ranging the distribution, it is important to use smooth values for the break, in order to understand and memorize easier the sense of the map, e.g. use 30 instead of 29,77; 1500 instead of 1508 etc.

Figure 12 shows the importance of the choice of data range on the visualisation of phenomena.

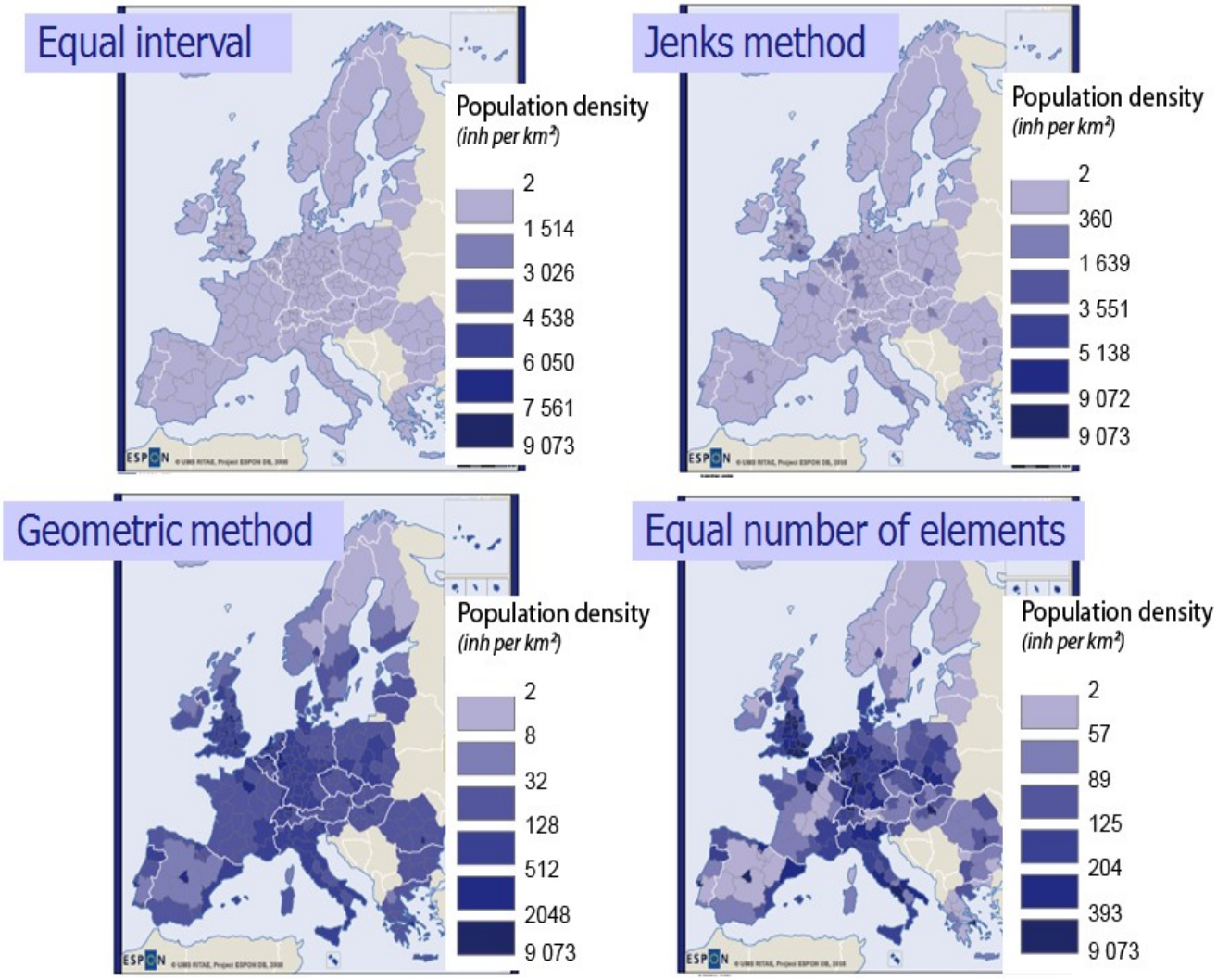


Figure 12*: Result and efficiency are dependent upon the data classification method*

2 Maps are tool for communication

As we explain in the introduction of this technical report: "Maps are perhaps as fundamental to society as language and the written word. They are the preeminent means of recording and communicating information about the location and spatial characteristics of the natural world and of society and culture[2] ".

Maps are produced all over the world and used by people as different as scientists, researchers, scholars, governments or businesses. These maps are most of the time statistical ones connected with the environment, the economy, the politics, the society etc.

The biggest strength of these maps is to allow an effective and relevant communication of the information. However, cartography is a special type of visual communication that does require some preliminary learning: a special purpose language for describing spatial relationships. "The analogy with language also helps explain why training in principles of effective cartography is so important--it allows us to communicate more effectively. Without knowledge of some of these basic principles, the beginning cartographer is likely to be misunderstood or cause confusion[2]".

Of course, cartographers must pay special attention to coordinate systems, map projections, and issues of scale and direction but that's not the first issue of map as a tool for communication. Maps are symbolic abstractions and representations. **The first question when mapping is related to know how to simplify, generalize, represent and symbolize the relationships being represented with graphics symbols**. In other words, what is a good map?

If a design is always more effective than a long speech, the measure of a good map is how well it conveys the right information to its readers and how well it communicates with its audience. This raises a series of questions that must be addresses at the start of a map conception: What is the motive, intent, or goal of the map? Who will read the map? Where will the map be used? What data is available for the composition of the map?

Beyond aesthetic characteristics, the communication also passes by a complete and effective layout: some elements must appear within the base map and the thematic representation, a complete legend, explicit title and source, a precise date of data or even a scale.

[2] *Kenneth E. Foote and Shannon Crum, The Geographer's Craft Project, Department of Geography, The University of Colorado at Boulder*

From data to map, 7 fundamental goals need to be identified to realize a good map:

1. Identify the goal of the map;
2. Identify the audience of the map and where it will be used;
3. Identify the information to be communicated;
4. Identify the geographical reference (point, line or area?);
3. Choose the base map (map projection and scale);
4. Choose the visual variable (symbolic graphic language);
5. Choose layout and identify all the elements to be added.

When these different elements are not correctly taking into account, the map will be characterised probably by some mistakes and misunderstandings.

2.1 Bad choices in term of representation of the data

Most of the problems of visualisation and map design are generally linked to **bad choices in term of representation of the data** (cf. part 2 of the technical report). When comparing ***figures 13 and 14***, which represent the same information, e.g. a typology showing age structure and total population in the municipalities from Brittany (France), it is quite clear that the second map is really clearer than the first one. Two main reasons can explain it (***figure 13***):

- Absolute values (e.g. total population in 2000) don't have to be shown by variation of intensity of black (hachure). This kind of representation does not respect the ratio of proportionality of the indicator, which is fundamental and needed information. Using hachure is also a visual mistake; the map is not readable at all and the representation is not the most efficient. These data have to be shown by proportional symbols, circles for instance.
- This typology, derived from age structure cannot be considered as a qualitative data, since there is an implicit order when considering the progression in term of age. In concrete terms, showing each class by a different colour is not the best solution. To show correctly this data it is important to think about the goal of the map. Here, it is important to represent the municipalities described by high share of young, active and old people. As a consequence, it is important to differentiate these information (3 colours) and also to make possible the analyse of the graduation of the phenomenon (high/medium shares), e.g. using variation of intensity of these 3 colours.

The solution proposed in **figure 14** try to correct these different elements. The most adapted solution for the representation of these data is to combine circles and colours in order to make the map as clear as possible. On top of that, it allows nuancing the interpretation of the map, e.g. Brittany is a region where ageing is important, but it concerns specific small and rural cities.

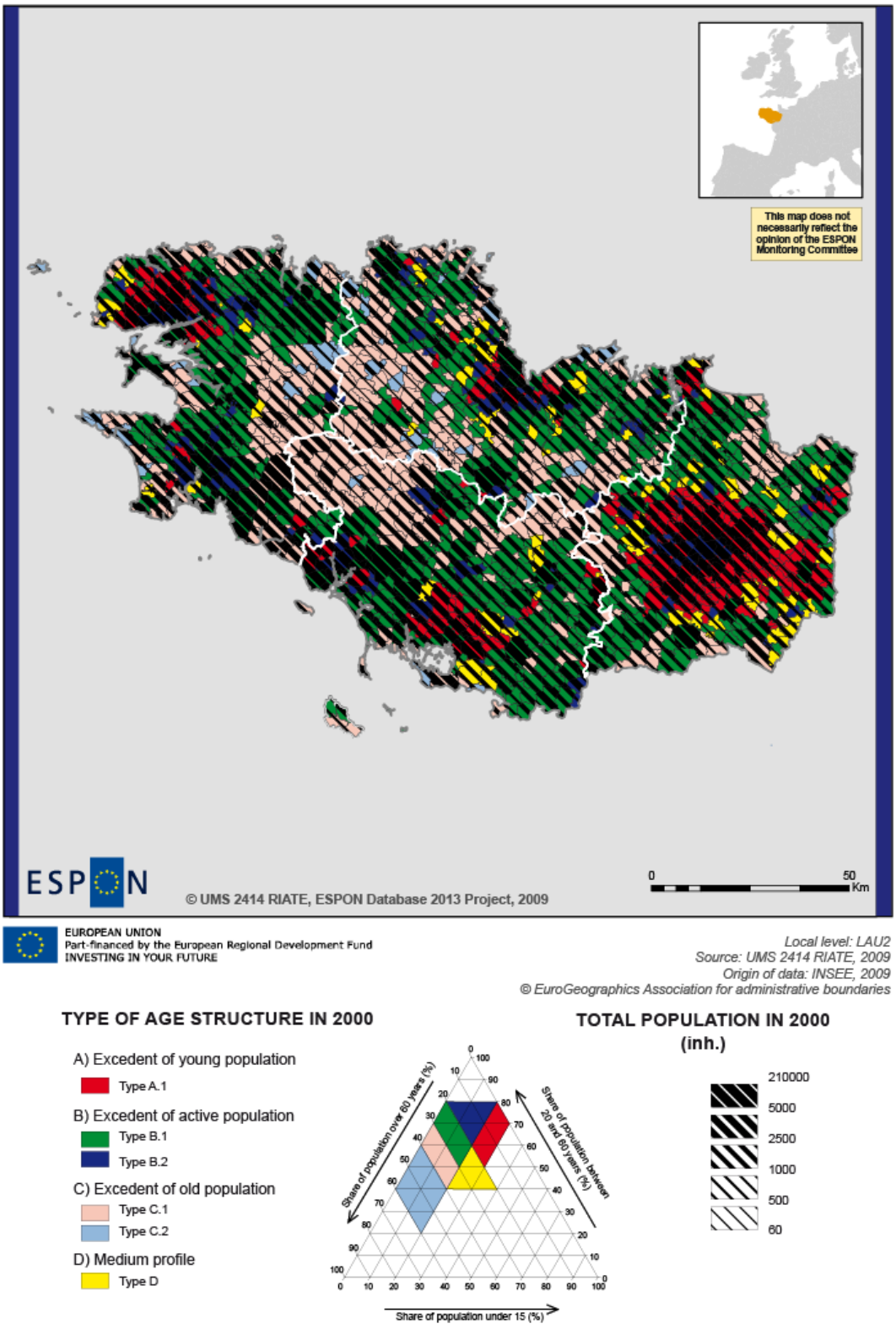


Figure 13**: Population and age structure in Brittany (France) – **with semiologic problems

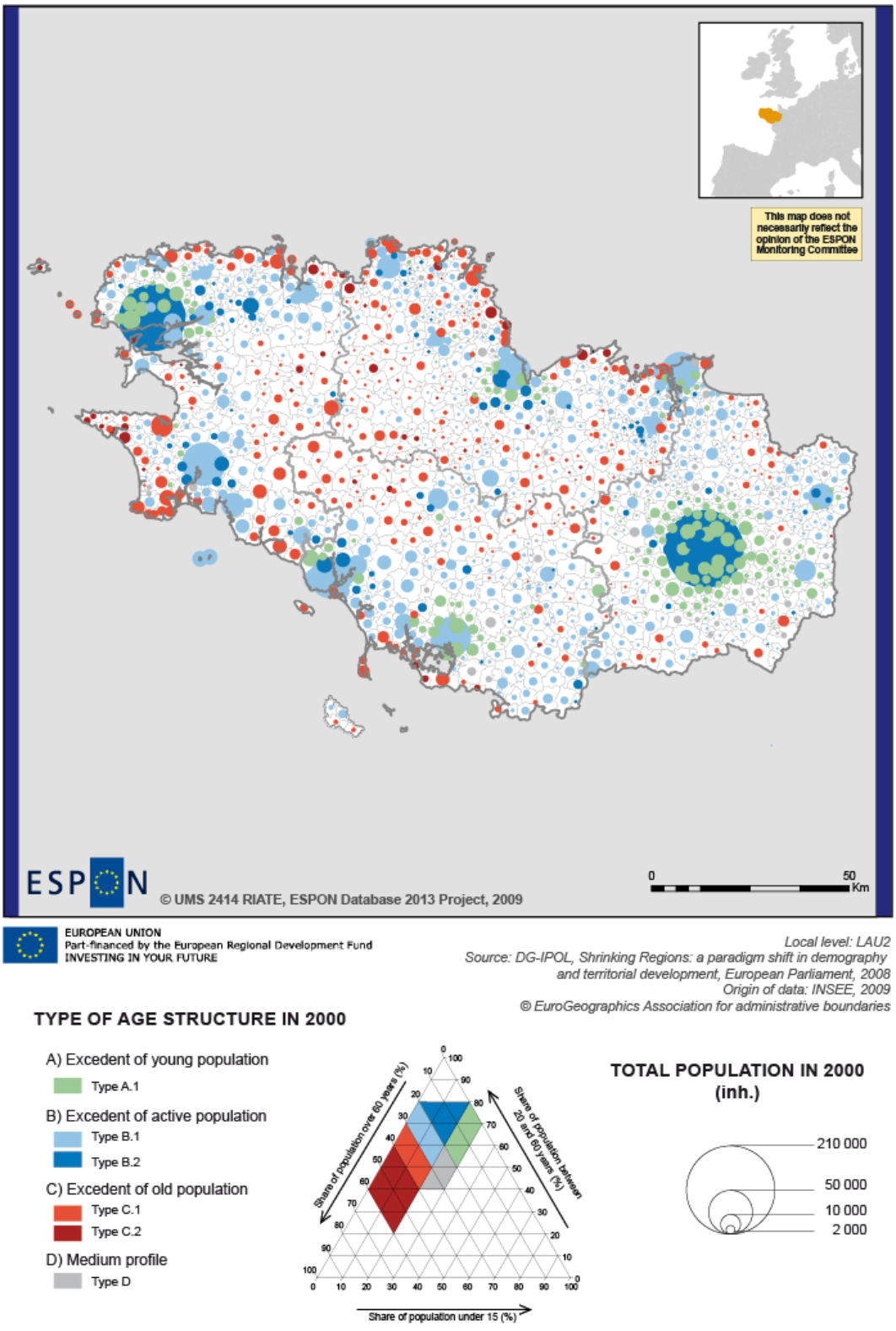


Figure 14**: Population and age structure in Brittany (France) – **without semiologic problems

2.2 Improving the efficiency of the map

Other problem which appears regularly is the degree of complexity of the map. The aim of the maps is to be synthetic. When representing too much information, the eye cannot distinguish the different elements of the map. This kind of figure can be solved by thinking to the design of the map: where is the best location for legend? How using with the most efficiency the place available?

The figures 15 and 16 show the same information, e.g. a typology of population development by components during the period 1995-2004 in EU27; this data is crossed with expected population evolution in 2030.

Figure 15 proposes solution which is correct in term of graphic semiology: ordinal data are shown by variation of colour (green/red) and shrinking/non shrinking regions (qualitative data) are represented by the opposition of hachure and no hachure. However, the combination of these two visual variables makes the map hard to interpret and the message become not so clear!

When there is too much information it becomes difficult to be able to synthesise the message of the map. That is why in some cases it is more efficient to split information in two maps instead of concentrating all the elements in a single one. This has been done on figure 16, where the map located on left of the document shows the regions described by an expected growth of population; and the map on the right shows the regions where a demographic decrease is planned. This template allows immediately to observe that during the period 1995-2005 most of the 'shrinking regions' have witnessed a downturn linked to both natural change and a negative migratory balance.

There is never an optimal solution

Whatever the examples proposed and demonstrated, it is important to keep in mind that there is never a single solution to show information on maps. In fact, each person has his own perception when interpreting graphic documents or pictures. **Map is always a compromise**. But during the creation of the map, is fundamental to try to make the map as understandable as possible. In concrete terms, it is not an obvious task and it is kindly recommended to make different attempts and share the results with other colleagues before saying "OK, my map is ready for the report"!

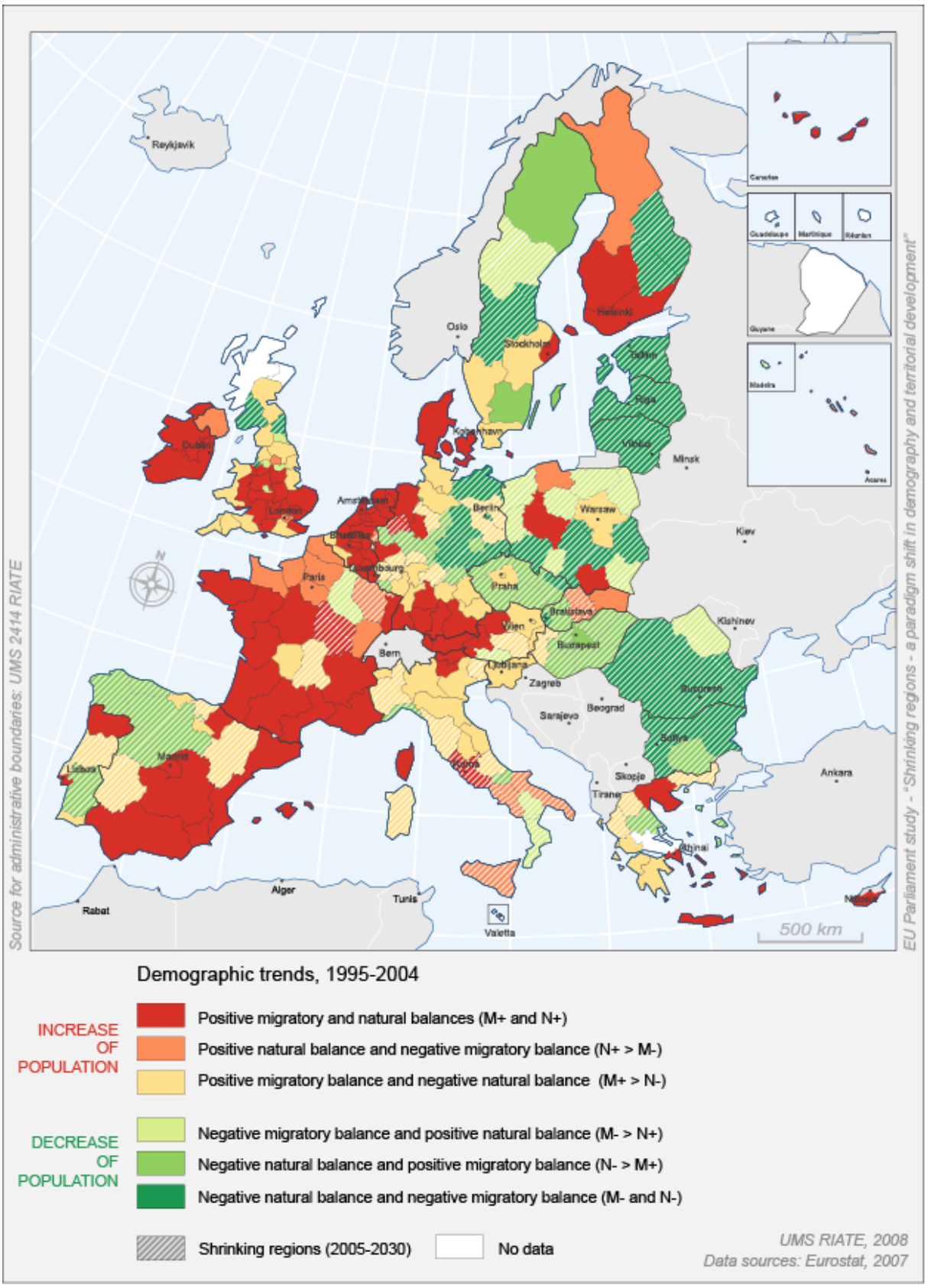


Figure 15**: Typology of regional growth patterns – **Possibility 1

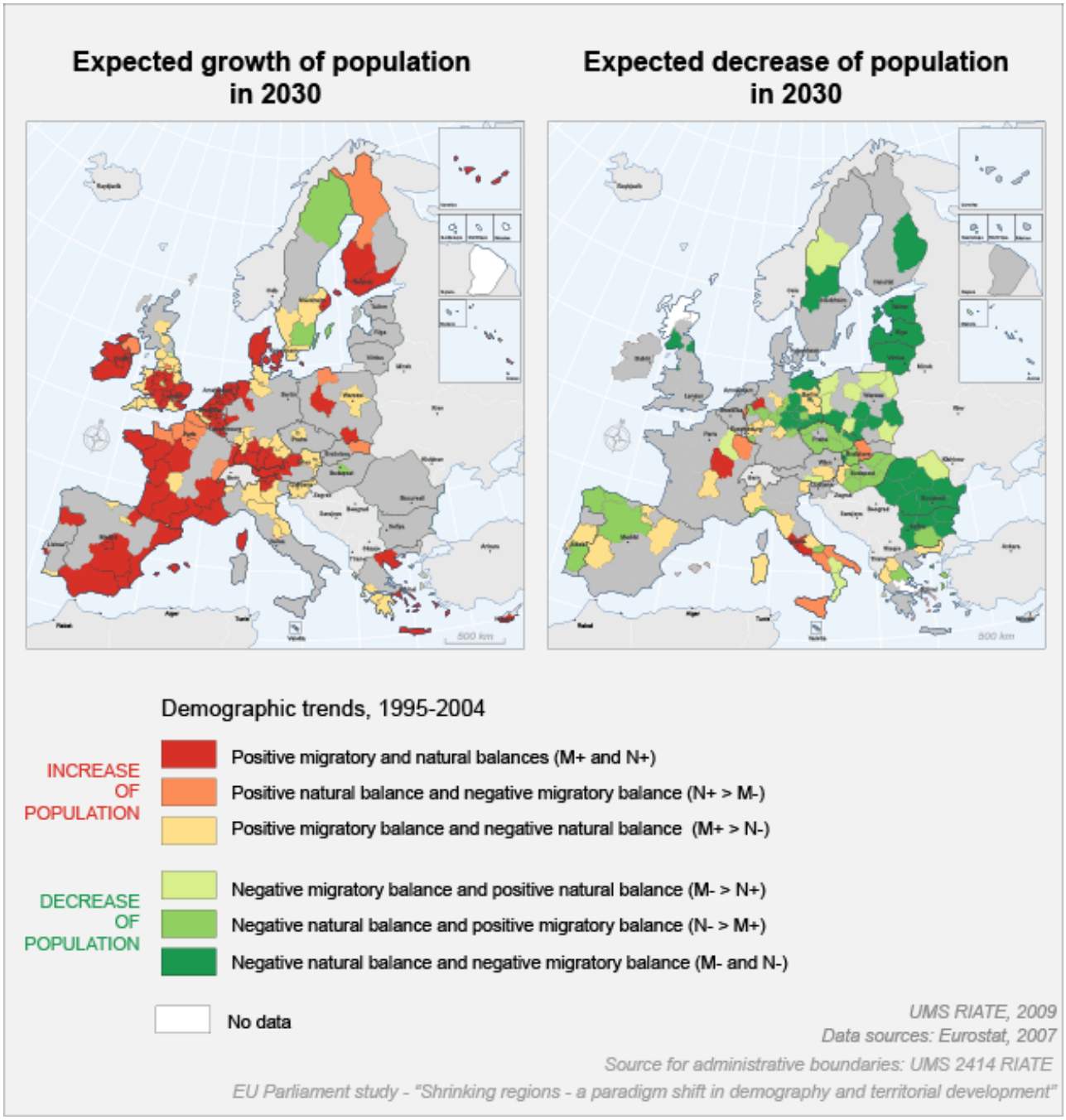


Figure 15**: Typology of regional growth patterns – **Possibility 2

ANNEXES

These annexes allow you to choose some efficient graphic variables to communicate differences in size, order or quality.

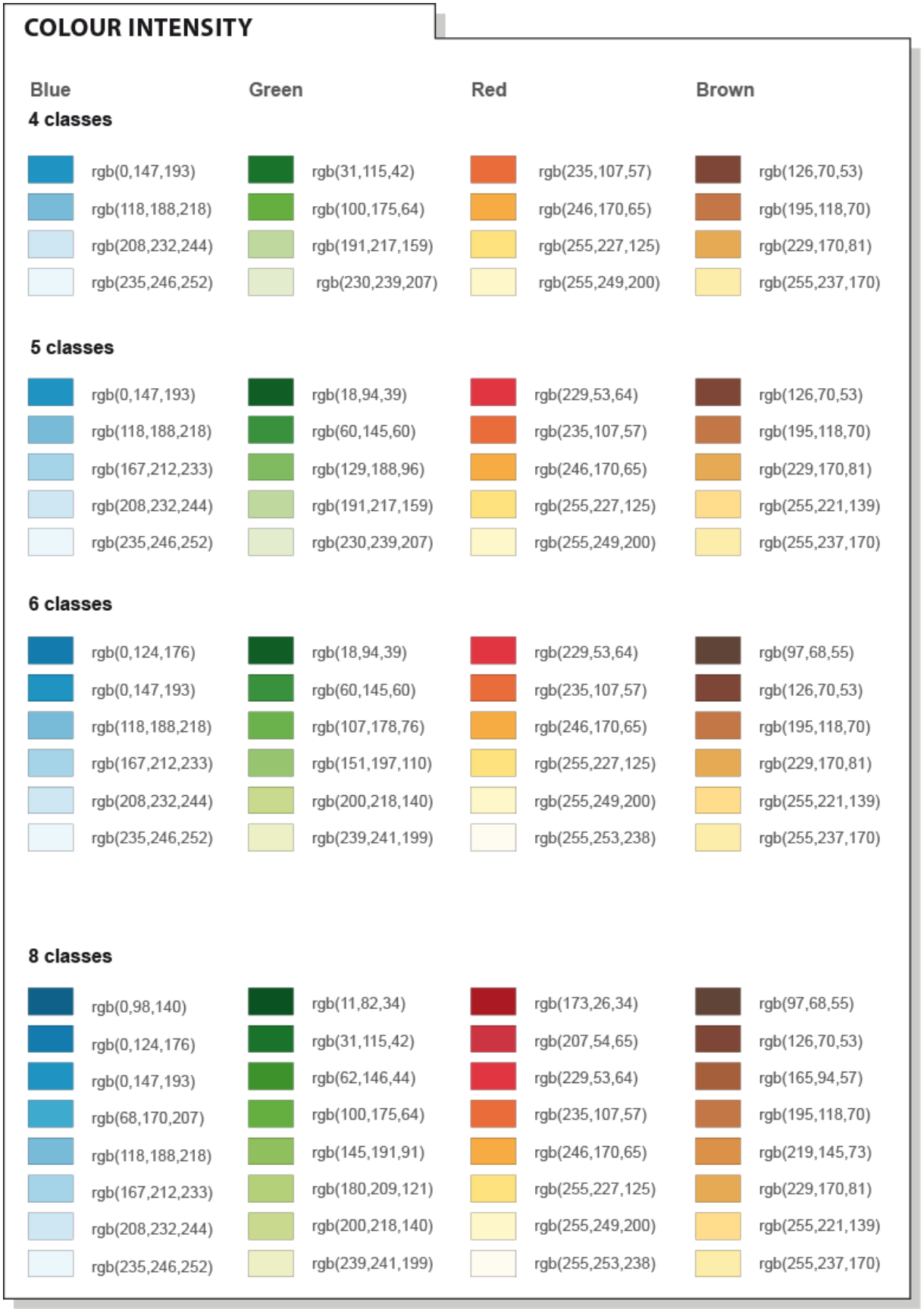
ANNEXE 1 - Relation of graphical variables to perceptual characteristics

Graphical variable	Type of data			
	nominal	ordinal	Interval/ratio	quantity
Size		x	x	x
Grey or colour value		x	x	
Grain/texture		x	x	
Colour hue	x			
Orientation	x			
Shape	x			

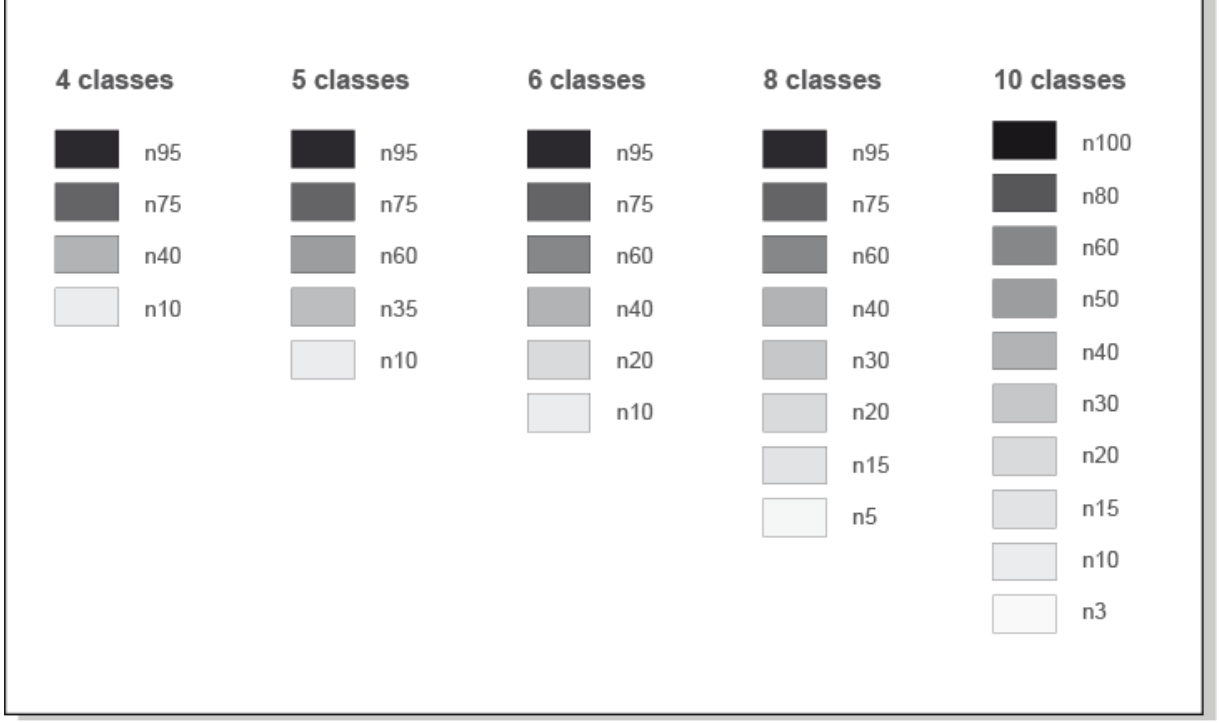
ANNEXE 2 - Numbers of categories that can be perceived at a glance

Graphical variable	Point	Line	Area
Size	4	4	5
Grey or colour value	3	4	5
Grain/texture	2	4	5
Colour hue	7	7	8
Orientation	4	2	4
Shape	3	3	3

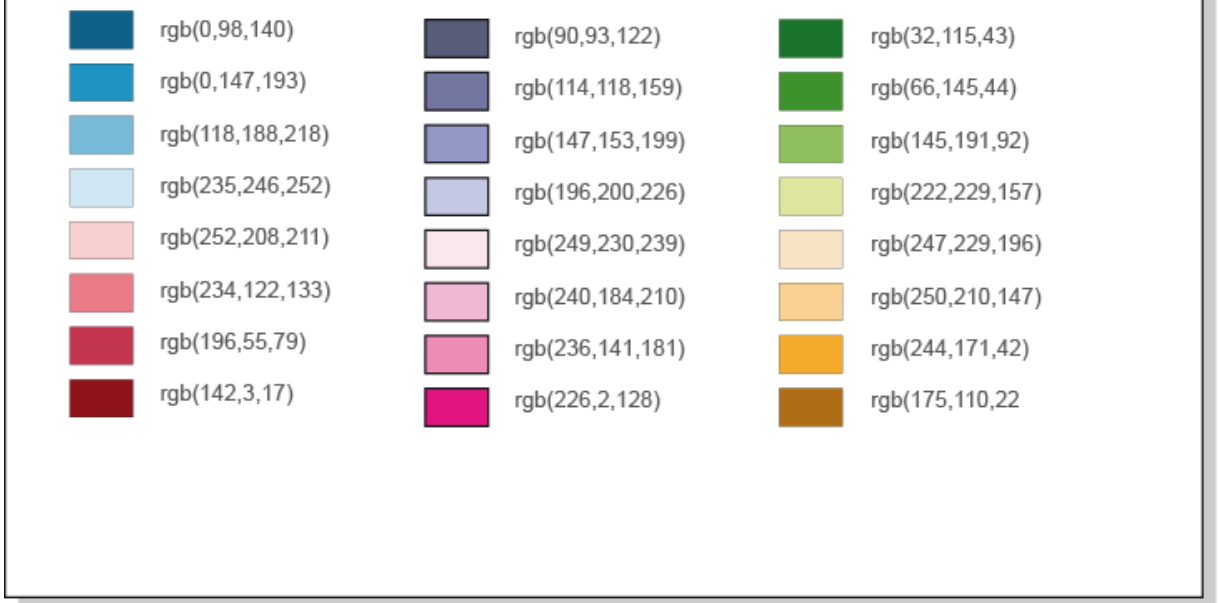
ANNEXE 3: Differences in value or lightness



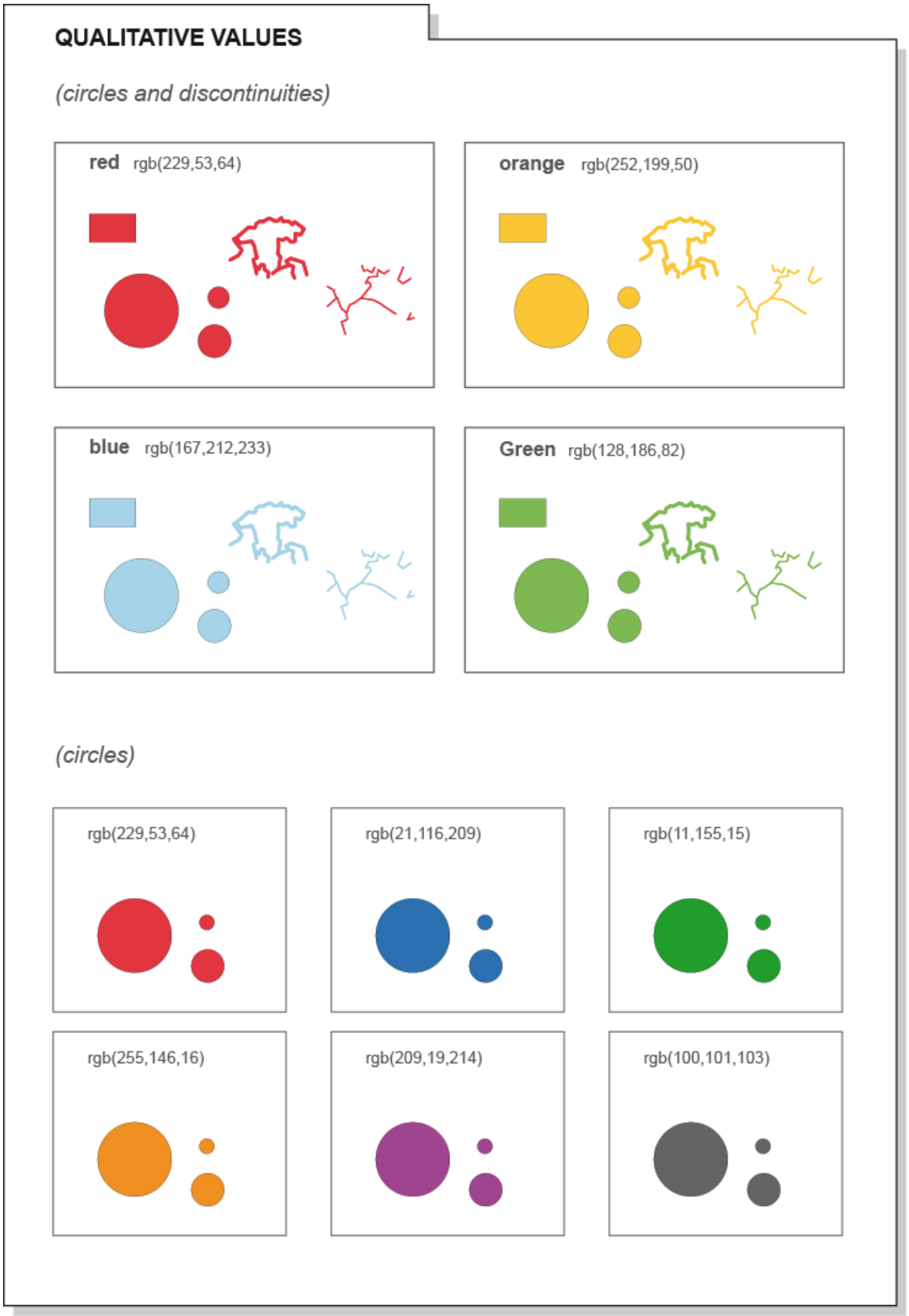
GREY VALUE



OPPOSITE COLOURS



ANNEXE 4: Colours for differences typology or qualitative value



References

- ***Litterature***

Béguin M., Pumain D., 2003, *La représentation des données géographiques – statistique et cartographie,* Armand Colin.

Bertin J., 1967, *Sémiologie graphique*, Gauthiers-Villars.

Cambrezy L., de Maximy R. (Ed.), 1995, La cartographie en débat, représenter ou convaincre, Editions Kathala et Orstom, Paris

Harris R. L., 1996, Information graphics, a comprehensive illustrated reference, visual tools for analysing, managing and communicating, Management Graphics ed., USA

Harley, J. B., 1988, Maps, knowledge and power. In COSGROVE, D. (Ed.) The Iconography of Landscape.Cambridge, MA, Cambridge University Press.

Kraak M.-J., Ormeling F., 2003, *Cartography, Visualization of Geospatial Data,* 2nd edition, Pearson Education, Prentice Hall.

Kraak, M.-J., 1998, Exploratory cartography, map as tools for discovery, *ITC Journal* (1), pp.46-54

MacEachren A.M., 1994, *Some truth with maps: a primer on design and symbolization*, Association of American Geographers, Washington DC.

Monmonnier M., 1996, *How to lie with maps*, University of Chicago Press.

Robinson A.H., Morrison J.L., Muehrcke P.C., 1995, *Elements of cartography*, New York, J.Willey & Sons.

Wilkinson L., 1999, The grammar of graphics. New York, Springer.

Wood, D., 1992, The Power of Maps. New York, The Guildford Press.

Wood C. H., Keller C. P., 1996, Cartographic design: theoretical and practical perspectives, Wiley, USA

Zanin C., Trémélo M-L, 2003, *Savoir faire une carte: Aide à la conception et à la réalisation d'une carte thématique univariée,* Belin.

- ***Websites***

***Colorbrewer* 2.0** is an online tool designed to help people select good color schemes for maps and other graphics: http://colorbrewer2.org/

Philcarto is a free tool for cartography, available on the net: http://philcarto.free.fr/

Quantum GIS is an Open Source Geographic Information System. It runs on Linux, Unix, Mac OSX, and Windows and supports numerous vector, raster, and database formats and functionalities: http://www.qgis.org/

0 TECHNICAL REPORT

ESPON

WORLD DATABASE

Towards a World Dictionary of units

CONTENT

- Description of the provisional WORLD ESPON 2013 DATABASE
- Overview and description of a sample of world databases
- Comparison between world databases and Eurostat databases: preliminary results

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

46 PAGES

LIST OF AUTHORS

Hy Dao, UNEP-GRID Genève

Andrea de Bono, UNEP-GRID Genève

Claude Grasland, UMS 2414 RIATE

Nicolas Lambert, UMS 2414 RIATE

Contact

hy.dao@grid.unep.ch
debono@grid.unep.ch
grasland@parisgeo.cnrs.fr
nicolas.lambert@ums-riate.fr

tel. + 41 22 917 82 40 (UNEP-GRID)
tel. + 33 1 57 27 65 32 (UMS RIATE)

TABLE OF CONTENT

Introduction

The first obvious aim of this challenge is to provide data for ESPON projects working at global scale, like the new projects on "Globalisation" launched in February 2010. But another important objective is to complete some discontinuous time series at NUTS2 or NUTS3 levels by means of disagregation of time series available at State level. But in order to make such a work, it is necessary to define a listing of indicators available, to define a The work done by UMS RIATE and expert team UNEP on this challenge is summarised in the draft technical report "ESPON World database". The following Technical Report presents the work in progress in February 2010. It will be improved until the end of the project.

The first section describes the provisional ESPON 2013 World Database by defining "indicators of reference" and introducing the notion of "units of reference".

Secondly, we have considered the official list of countries from main international "thematic" providers. In fact, the definitions of "what is a country" for each provider do not correspond in several cases. The second section shows concretely this fact.

The section 3 focuses on the linking of World data with Eurostat Regional data. Our goal has been to design a methodological tool (named "Gap Tracker") for explaining the differences between global databases and Eurostat data. This Technical Report shows the first results of the testing phase which will be developed in the next steps of the ESPON Database Project.

1 Description of the provisional WORLD ESPON 2013 DATABASE

This is a preliminary version of the World Espon 2013 database:

The number of indicators is limited and will be improved

The list of "countries" and regions for the "world database ESPON 2013" is under process, as well as the elaboration of their unique ID

The standard output format for the exchange with the main Espon 2013 Database is not yet implemented

The database can be subdivided into two main components:

The Indicators (section 1.1) including the data sensu stricto with global extent, mainly from International organizations and data provided by Eurostat, which cover the European region.

Units (section 1.2) including country subdivisions and the regional/thematic aggregations used by global providers and Eurostat

1.1 Indicators

We propose to assemble our collection starting and testing methodologies on four maingroups of variables: population, Gross Domestic Product (GDP), carbon dioxide

Emissions and land use, that will include in a second stage all the subcategories needed by the ESPON database. This version of database includes data for population and CO2 emissions. A complete list of indicators actually included (end of June 2009) in the database can be found in the annexes

Fields description

category (text) indicator full name ex. population sex ratio

provider_code (text or integer) original country code from data provider (normally ISO or UN)

source (integer) data source code; linked with table source

1950,… 2050 (double) value per year

Data sources

Population: United Nations/Population Division with the World Population Prospects (WPP2008)

http://www.un.org/esa/population/unpop.htm

Emissions: UNFCCC data reported by countries (Annex I parties), and

http://unfccc.int/ghg_data/items/3800.php

CDIAC where data are calculated from energy statistics from UN yearbook

http://cdiac.ornl.gov/trends/emis/overview.html

Table names

Tables wpp2008_stocks, sources WPP2008 and world_co2 sources, CDIAC and UNFCCC

eu2009_co2, et eu2009_pop_stocks (Europe) sources Eurostat 02/09

Notes

All data at this moment do not include any external manipulation. It will be the case when we will work with World Bank World Development Indicators (WDI) that normally display several gaps in time series. Population data from WPP from 2009 to 2050 are projections calculated on the base of "Medium Fertility Variant".

1.2 Units

Under this category we include the countries/territories subdivisions and regional/thematic aggregations supplied by our main data providers. They will be described in detail in the next chapter: "Overview and description of a sample of world databases".

The main_table represents the basic territorial units including 251 countries/territories and their ISO alpha 3 and 2 codes. In order to assign a unique ID for each territorial unit, we added some ISO codes where data were missing (see field notes)

Tables un_main, wdi_main, geo_main include countries/territories with their regional and or thematic aggregations respectively from UN, World Development Indicators and GEO, with their original country codes and Iso3 as primary key.

Table europe_main follows the same structure.

Tables undp_hdi_main et undp_hdi_cat include the last figures per country of Human Development Index (HDI UNDP).

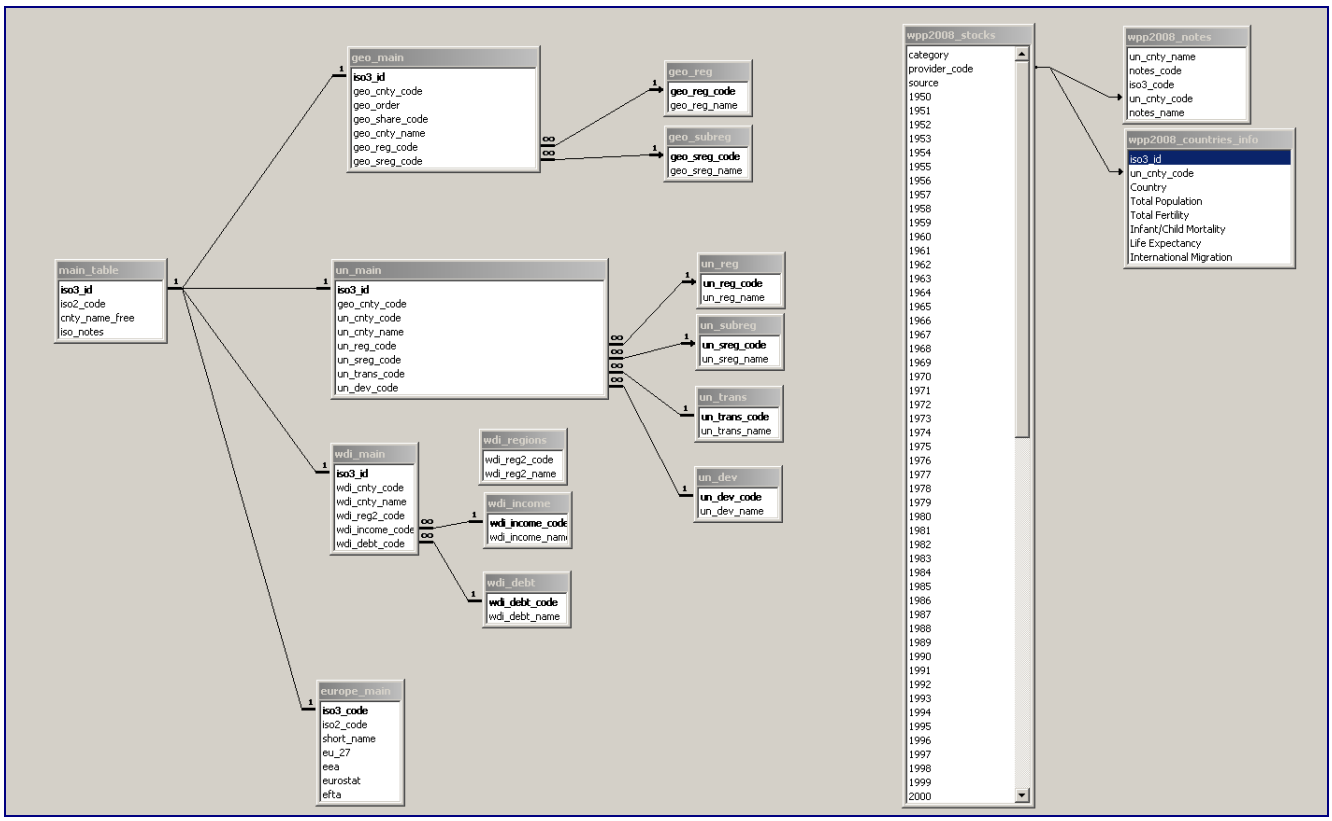


Figure 1 : countries/territories subdivisions and regional/thematic aggregations supplied by our main data providers

2 Overview and description of a sample of World databases

A lot of World databases exist. However, they do not describle the units contented within it in a same way. A first work consists to identify the structure of reference of each of them. The complete description of geographical units is presented in annexe 2.

2.1 CHELEM DATABASE

CHELEM is an economic long term database constructed by the CEPII (1960 to present). The aim of this database is to constitute a coherent view of the world economy. This database is composed of 3 sub-databases: International trade (1) GDP (2) and balance of payments (3).

CHELEM is based on a specific geographic classification with two kinds of partition. The partition in 96 zones (available from 1993 onwards) gives the maximal detail for trade. The partition in 82 zones doesn't detail the countries resulting from former Yugoslavia, USSR and Czechoslovakia.

Partition in 82	Partition in 96
	

Figure 2 : CHELEM Database subdivisions

2.2 ESPON 2006 EUROPE IN THE WORLD DATABASE: The WUTS System

Realised in the first ESPON program, this world database in based on a precise list of 168 states which represent a minimum of 1/10 000 of the population, GDP or area of the World. This list of 168 states provides a clear basis for data collection in an harmonised way, all states being identified by a specific code (WUTS CODE).

The **WUTS** (World Unified Territorial System) is a harmonised hierarchical system of World division which is directly inspired from the **NUTS** (Nomenclature of Territorial Units for Statistics) created by Eurostat more than 25 years ago in order to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union. The WUTS is compesed by 5 hierarchical levels, from the level of States (WUTS5) to the level of the World (WUTS0).

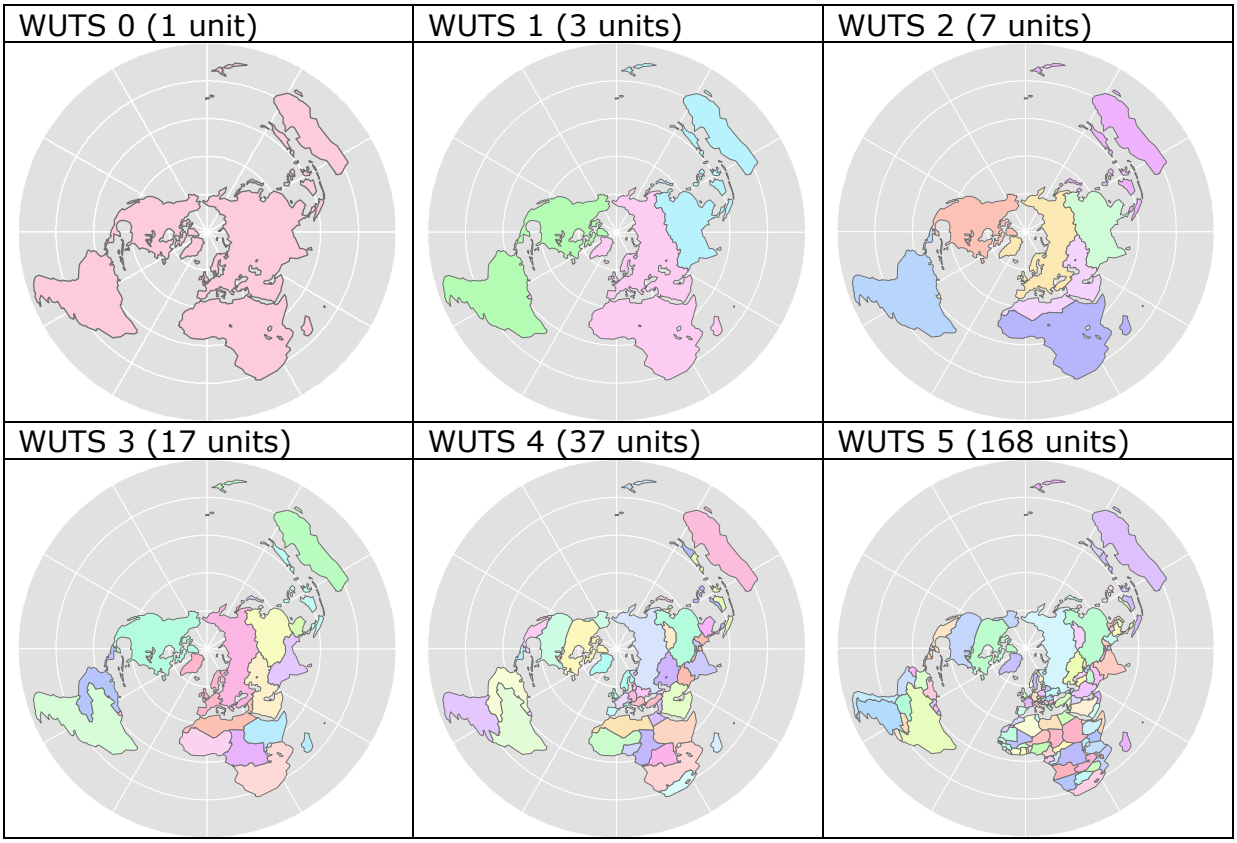


Figure 3 : Europe in the World subdivisions

2.3 UN Standard countries or area and geographical regions

The list of countries or areas includes those countries or areas for which the Statistics Division of the United Nations Secretariat compiles statistical data. The names of countries or areas refer to their short form used in day-to-day operations of the United Nations and not necessarily to their official name as used in formal documents.

The geographical regions and groupings of countries and areas are not comprehensive but only a selection, which are or may be used in the compilation of statistics. In order to ensure consistency in statistics and for convenience, each country or area is shown in one region only. The macro geographical regions are arranged to the extent possible according to continents. Within these groupings more detailed component geographical regions are shown.

The group of least developed countries (LDCs), as defined by the United Nations, comprises 49 countries, of which 33 are in Africa, 10 in Asia, 1 in Latin America and the Caribbean, and 5 in Oceania.

Note that there is no established convention for the designation of "developed" and "developing" countries or areas in the United Nations system: the designations "developed" and developing" are intended for statistical convenience and do not necessarily express a judgement about the stage reached by a particular country or area in the development process.

Criteria for identification of LDCs and Landlocked developing countries can be found at http://www.unohrlls.org/en/ldc/related/59/

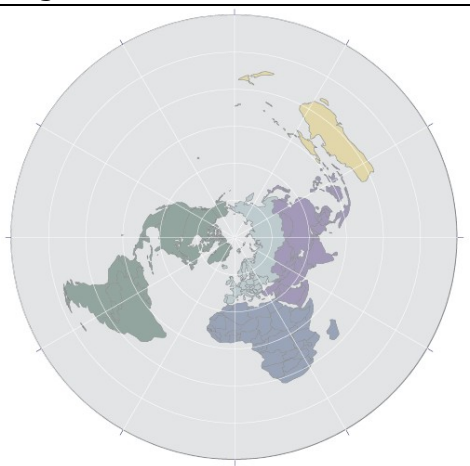
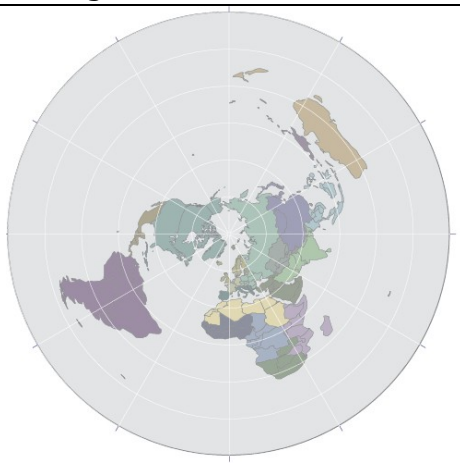
UN regional Breakdown	UN sub-regional Breakdown
	
UN Developing countries	UN Landlocked developing countries
	

Figure 4 : United Nations aggregations

2.4 World Bank: The World Development Indicators (WDI)

The World Development Indicators (WDI) 2009 is the statistical benchmark that helps measure the progress of development.
The 2009 WDI includes more than 800 indicators organized in 6 sections: World View, People, Environment, Economy, States and Markets, and Global Links.
Data are shown for all World Bank member countries (185), and all other economies with populations of more than 30,000 (209 total)

For operational and analytical purposes, the World Bank's main criterion for classifying economies is gross national income (GNI) per capita. Based on its GNI per capita, every economy is classified as low income, middle income (subdivided into lower middle and upper middle), or high income. Other analytical groups based on geographic regions are also used.

Geographic region: Classifications and data reported for geographic regions are for low-income and middle-income economies only. Low-income and middle-income economies are sometimes referred to as developing economies. The use of the term is convenient; it is not intended to imply that all economies in the group are experiencing similar development or that other economies have reached a preferred or final stage of development. Classification by income does not necessarily reflect development status.

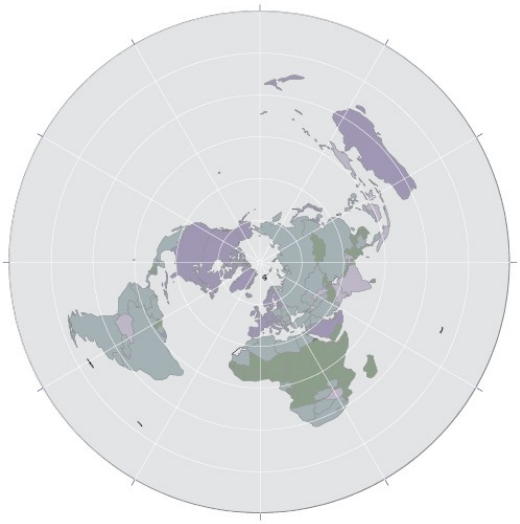
WDI aggregations by Income	WDI aggregations by Debt
	

Figure 5 : WDI aggregations

2.5 The Global Environment Outlook (GEO) Data Portal

The GEO Data Portal gives access to a broad collection of harmonized environmental and socio-economic data sets from authoritative sources at global, regional (7), sub-

regional (23) and national (237) levels. There is no established convention for the designation of regional and sub-regional groups. Geographical aggregations are arranged to the extent possible according to continents. Some inconsistencies exists: for example French Guyana is incorporated in the South America in regional aggregations, but in a political point of view belongs to Europe.
In the other way Israel could be included by its geographic position in West Asia but it is comprised de facto to the Western Europe group.

GEO regional Breakdown (7)	GEO sub-regional Breakdown (23)
	

Figure 5 : GEO aggregations

3 Comparison between world databases and Eurostat databases: preliminary results

3.1 Methodology

Based on the results of ESPON 2006 Program, we propose to examine in a systematic way, how to combine datasets at world/neighbourhood levels (where basic territorial units are the states) and datasets at European/Regional levels (where basic territorial units are NUTS2 or NUTS3 units).
Our focus in this chapter consists to explain the differences in the indicators values, for the same geographical unit, between the global and European databases. This methodological tool is called the "gap tracker tool".

- **Europe in the ESPON database (EIE)**
  Provider: mainly Eurostat
  Coverage: Eurostat countries

- **Europe in the World database (EIW)**

  Provider: International Organizations (eg UN, FAO)
  Coverage: global but the check is only between countries matching with Eurostat coverage

In order to increase compatibility between EIE and EIW datasets, we setup a "process" of systematic analyse of their differences. The steps are descript as follows

Verification Phase

This phase mainly consists to check if the two sources are compatible in terms of:

- Definition of a country (ex.: Cyprus includes both Greek and Turkish administration part or not?)
- Date of update (ex.: EIE once two years, EIW once six months)
- Period considered for the collect of indicator (ex.: Census date)
- Method to collect the indicator: measure or estimation
- Definition of indicators: like unemployment
- Measure units
- Methods of estimation for missing data (ex.: extrapolations, interpolations..)

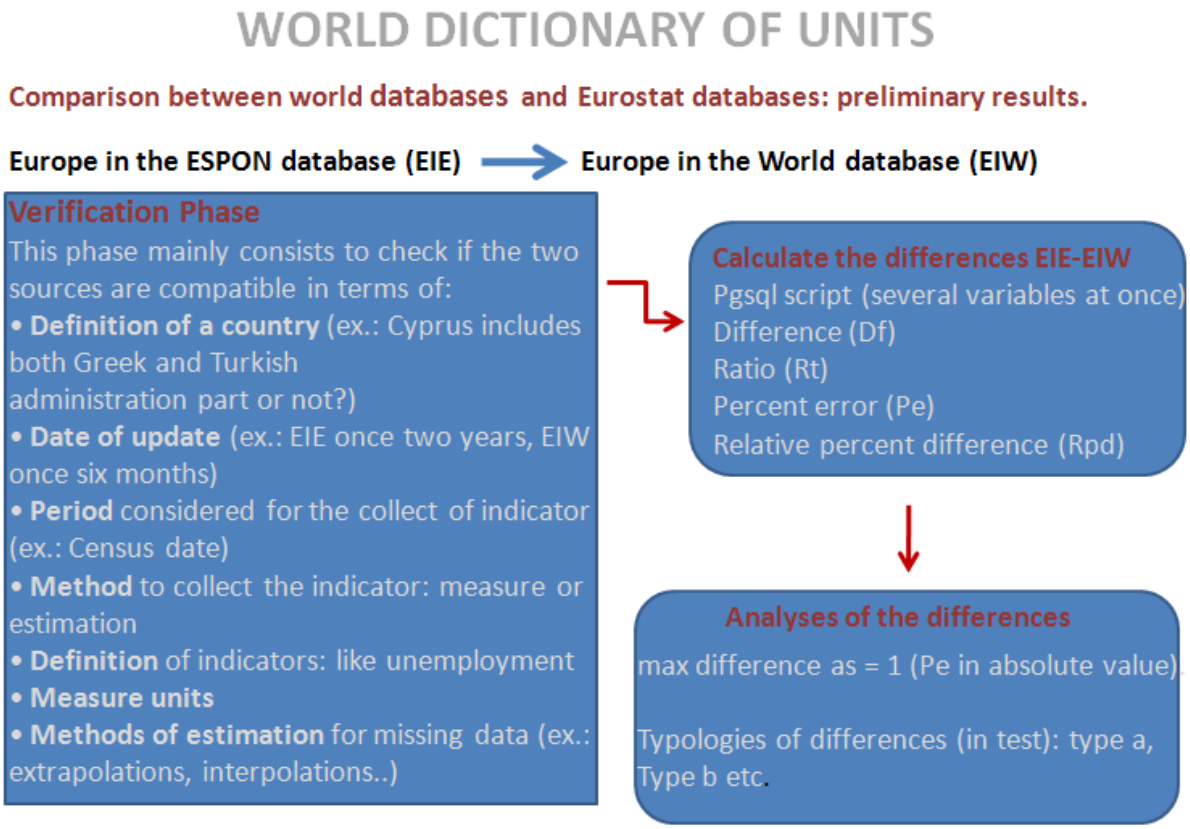


Figure 6 : Summary of the methodology used in the testing phase

Calculate the differences EIE-EIW

A script in Pgsql helps us to calculate the difference between EIE and EIW with several data sets at the same time. For the moment we use four types of simple formulas, but more complicate algorithms can be easily added.

Difference (Df) = EIW – EIE
Ratio (Rt) = EIW / EIE
Percent error (Pe) = ((EIW - EIE) / EIE) * 100
Relative percent difference (Rpd) = ((EIW – EIE) / ((EIW + EIE) / 2)) *100

Analyses of the differences

Once fixed the acceptable max difference as = 1 (Pe in absolute value), we can approach the problem under different perspectives: by country, indicator, group of indicators, year…

However, the threshold = 1 can be debated.

Typologies of differences (in test)

The idea is to subdivide the Pe in several typologies in order to better characterize the analyses of difference. We introduce three concepts to illustrate the typologies:

Magnitude is the numerical value of the difference. It is generally referred to the Percent error (Pe). In a qualitative way a Pe comprise from 1 to 3 % is considered as "moderate"…

Range is the value max minus value min of the magnitude in a time series.

Variability is the measure of the variation of the magnitude over the considered years covered by the indicator. It can be of several Random constant, linear, exponential, composite…

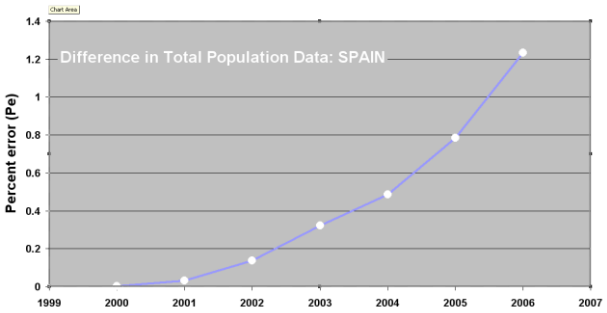
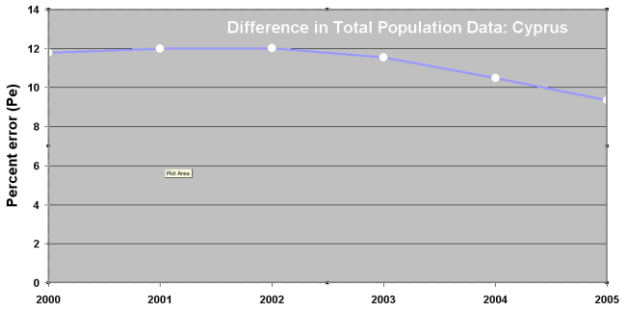
Example of exponential variability	Example of constant variability
	

Figure 7 : Different types of variability

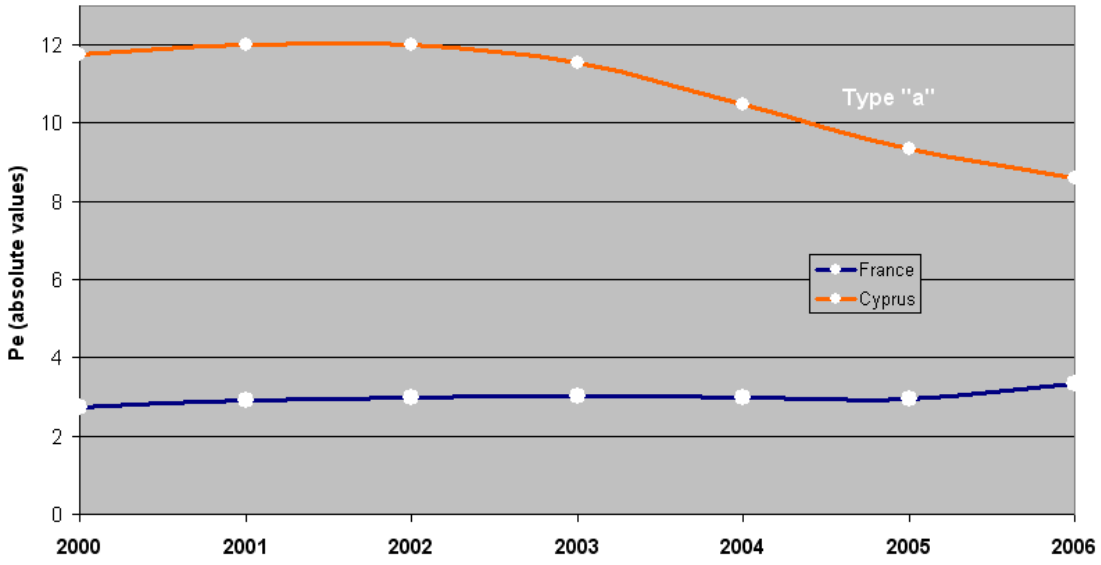
The typologies of difference include:

- Type "a" magnitude moderate to medium, variability constant across all the period of collect: cause probable difference of definition for countries (eg.: population for Cyprus and France)

- Type "b" magnitude moderate, distribution variable, some years without errors. Situation that can have several origins: mixed sources data, interpolations/extrapolations from EIW or /and EIE

- Type "c" magnitude moderate to elevate, variability constant to slightly random across all the period of collect: difference in the indicator definition and as consequence in the collect methods (measure or calculations)
- Type "d"…

Example of Type "a": in this case the origin of discrepancy is the different definition of countries (nb.: Pe in absolute values)



Example of type **"b":** Something happens in 2000-2001 (census differences? Interpolations…)

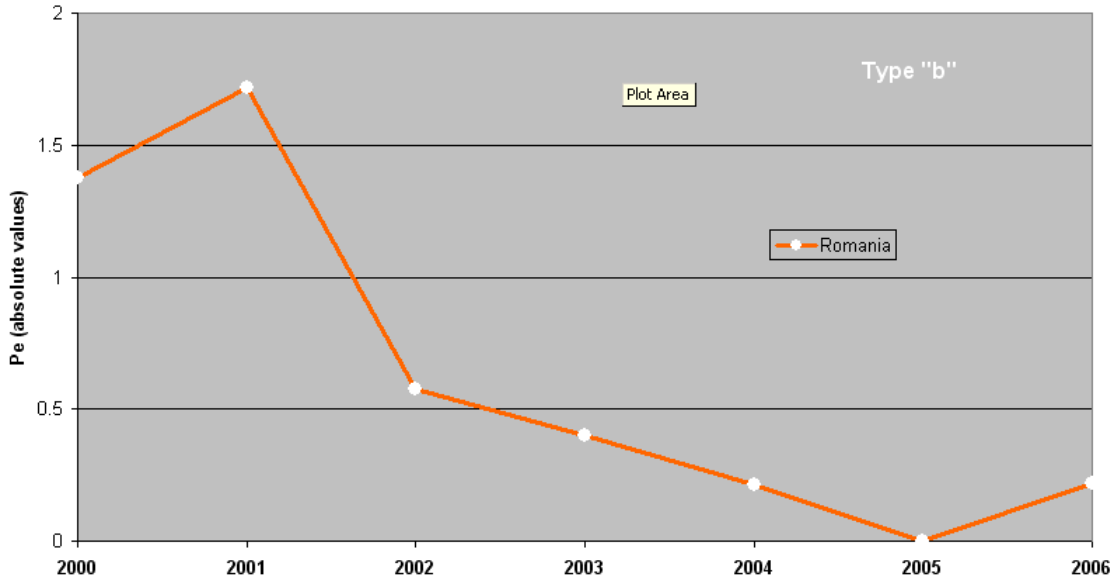


Figure 8 : Towards a typology of differences

3.2 Preliminary results

We analyzed stocks data from the last World Population Prospect (WPP08) including male, female, and both sexes population by age groups.
Although, the majority of Eurostat data on demography is provided by the UN Population Division some differences between the two sets of data exists, especially for population by sex and age groups.

3.2.1 Total population data

Comparisons from "total population both sexes" indicator give very moderates values of Pe, comprise below the threshold of 1, for almost all countries.
Cyprus and France show a distinctive difference of type "a" caused by the different definition of the country:
Data for Cyprus refer only to the areas of Cyprus controlled by the Government of the Republic of Cyprus" in the Eurostat database and France includes the overseas departments (DOM).
Apart France & Cyprus, only Liechtenstein, Romania, Bulgaria, Spain and Malta display values outside the reported in figure xy.
Data for Spain shows variability with exponential trend: Pe increase during time. This could be caused by secondary readjustment of values?.

Countries where value of Pe exceeds 1. Indicator: total population both sexes

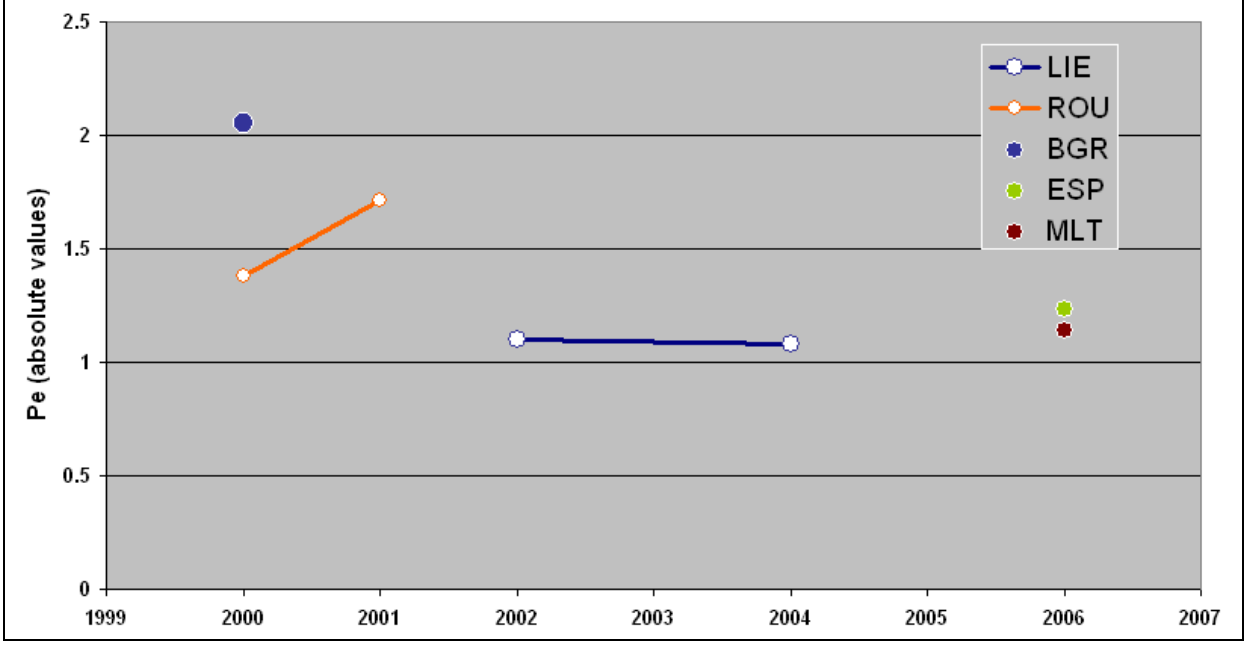


Figure 9 : Identification of countries where significant differences has been identified

3.2.2 Population by sex and age groups

This group of indicators shows significant differences between EIW and EIE data sets. Figure 10 shows the number of countries where Pe > 1 (per age/sex class).

Number of countries where Pe is greater than 1 (year 2005)

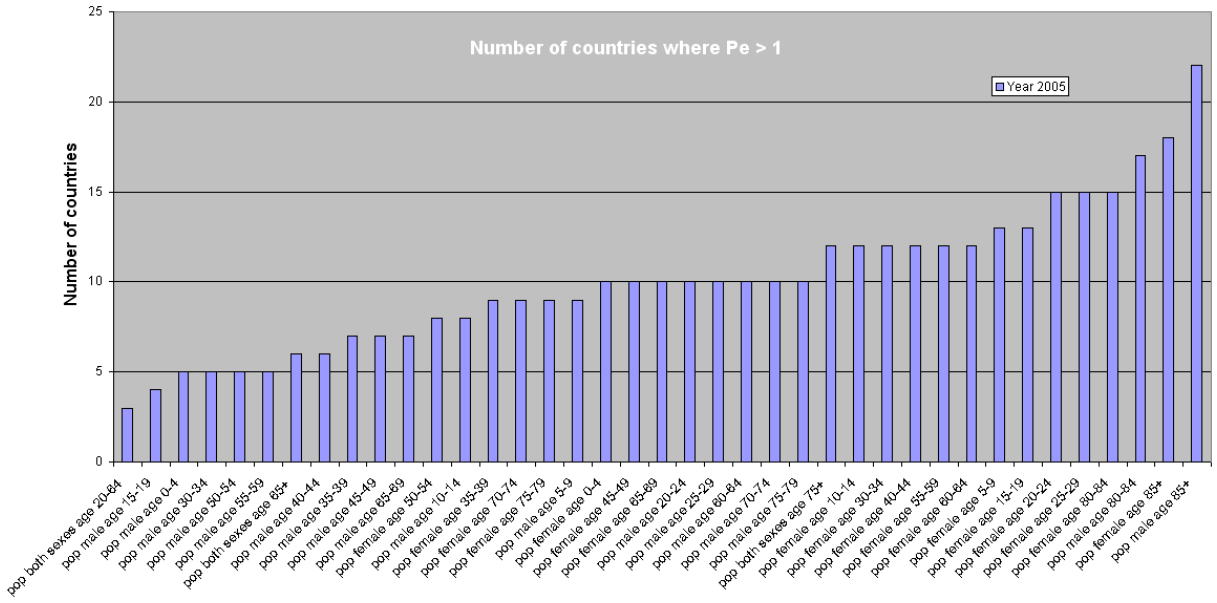


Figure 10 : Identification of age-classes where significant differences has been identified

The following tables illustrate some of these inconsistencies per country and per group of variables.

Differences absolute vs relative per age groups fo Belgium (year 2005)

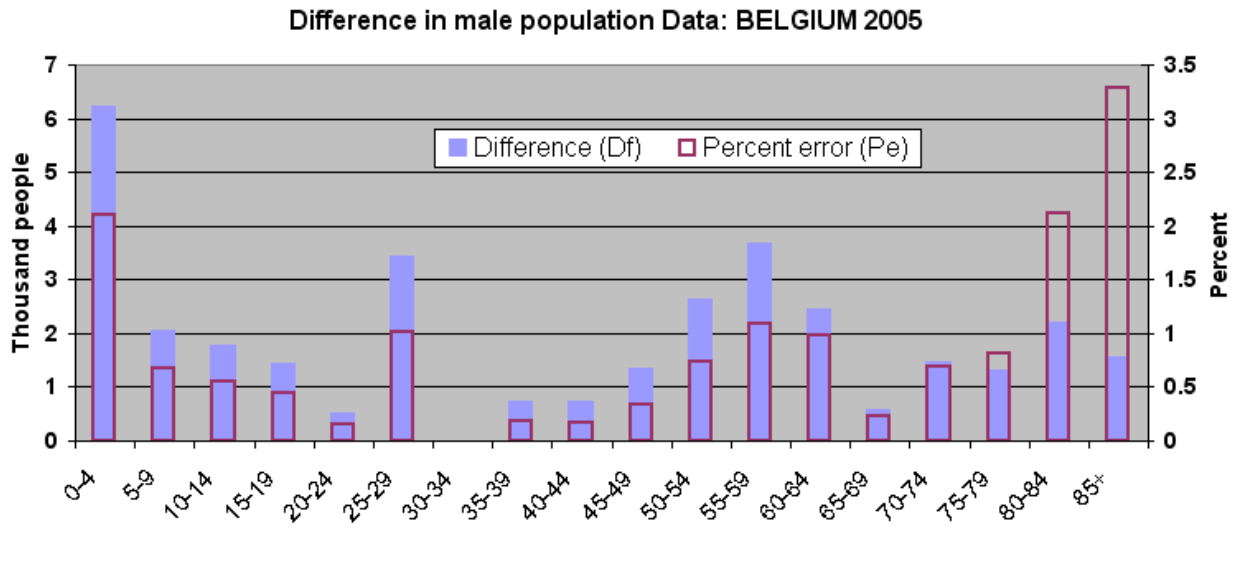
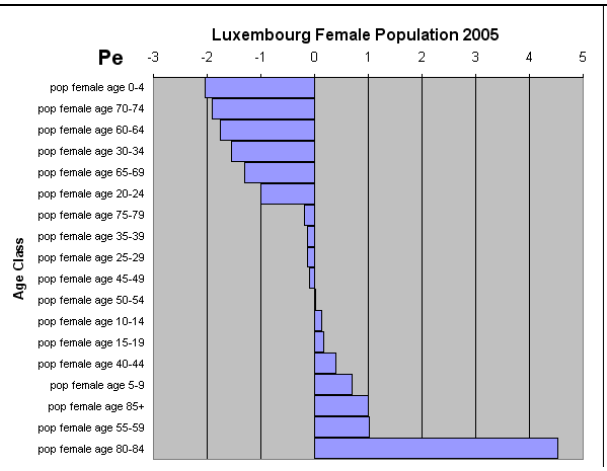


Figure 11 : Identification of age-classes where significant differences has been identified in Belgium

Per age groups for Luxembourg (year 2005)



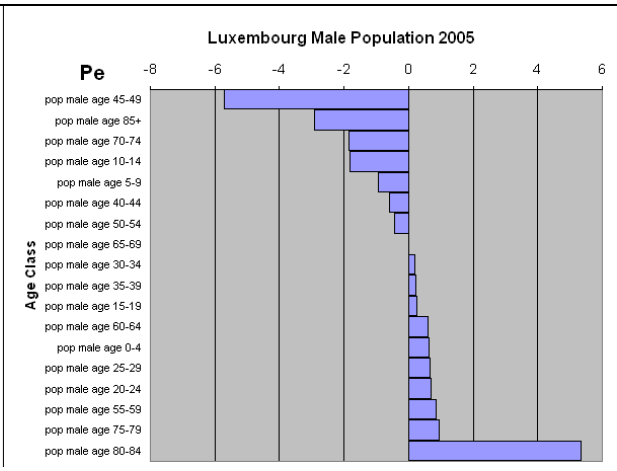


Figure 12 : Identification of age-classes where significant differences has been identified in Luxembourg

…Something happens for population over 80 in England?

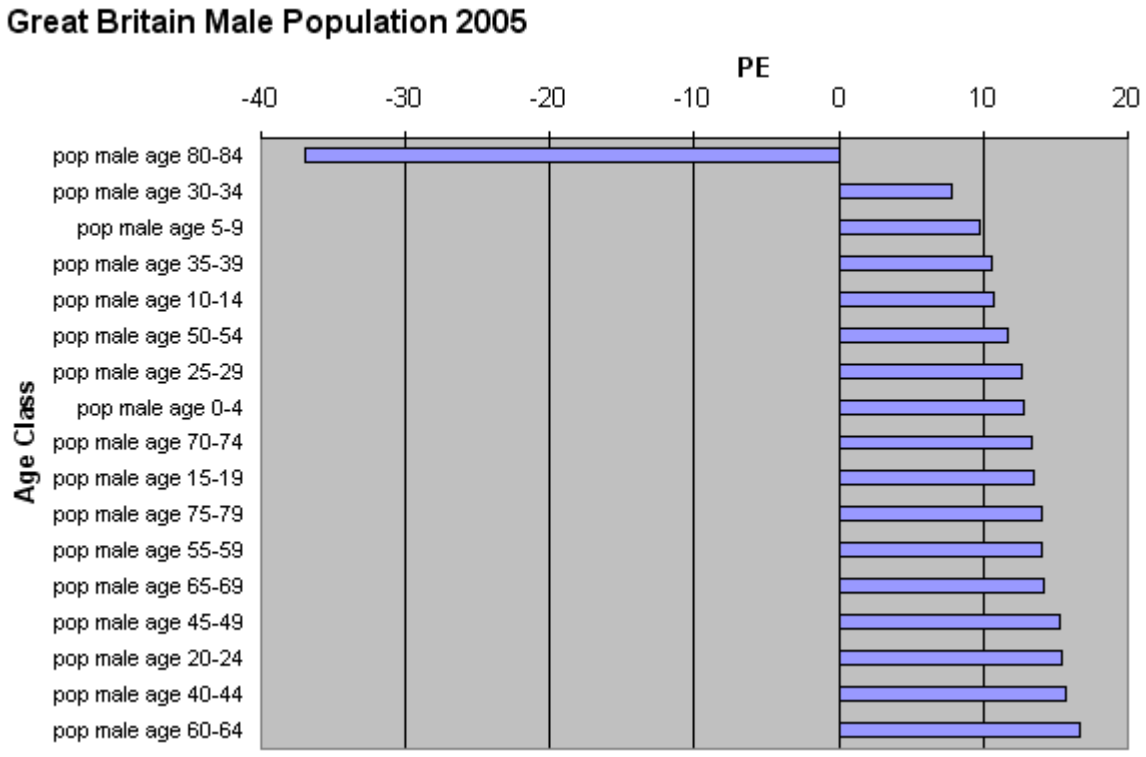


Figure 13 : Identification of age-classes where significant differences has been identified in England

We analyzed 1420 records: all countries show at least a value (but mainly a set of values) with a Pe greater than 1.
85 records have a Pe bigger than 10. There are no apparent relationships between errors in class group and or sex.
One of cause of this inconsistence can be searched in the differences between definitions of variables and, may be, in the interpolations. Work is still in progress…

4 Work in progress

February to June 2010

- Final choices concerning the "World Dictionary of units"
- Finalisation of the "Gap tracker" tool

July to December 2010

- ESPON World Database version 2.0 (global statistical data + codes of spatial units + links to ESPON DB geometries)
- Final version of TECHNICAL REPORT "ESPON World database (I): World Dictionary of units"

Annex 1 - List of EIW (including global coverage) indicators

1 - Data from WPP 2008 population stocks table wpp2008_stocks

indicator	temporal extent
pop female age 80+	1950-2050
pop male age 80+	1950-2050
pop female age 75-79	1950-2050
pop male age 75-79	1950-2050
pop male age 70-74	1950-2050
pop female age 70-74	1950-2050
pop male age 65-69	1950-2050
pop female age 65-69	1950-2050
pop female age 60-64	1950-2050
pop male age 60-64	1950-2050
pop female age 55-59	1950-2050
pop male age 55-59	1950-2050
pop female age 50-54	1950-2050
pop male age 50-54	1950-2050
pop female age 45-49	1950-2050
pop male age 45-49	1950-2050
pop female age 40-44	1950-2050
pop male age 40-44	1950-2050
pop female age 35-39	1950-2050
pop male age 35-39	1950-2050
pop female age 30-34	1950-2050
pop male age 30-34	1950-2050
pop female age 25-29	1950-2050
pop male age 25-29	1950-2050
pop female age 20-24	1950-2050
pop male age 20-24	1950-2050
pop female age 15-19	1950-2050
pop male age 15-19	1950-2050
pop female age 10-14	1950-2050
pop male age 10-14	1950-2050
pop female age 5-9	1950-2050
pop male age 5-9	1950-2050
pop female age 0-4	1950-2050
pop male age 0-4	1950-2050
pop female age all	1950-2050
pop male age all	1950-2050
pop both sexes age all	1950-2050
pop both sexes age 0-14	1950, 1955, ..., 2050
pop both sexes age 0-17	1950, 1955, ..., 2050
pop both sexes age 0-19	1950, 1955, ..., 2050
pop both sexes age 0-24	1950, 1955, ..., 2050
pop both sexes age 15+	1950, 1955, ..., 2050
pop both sexes age 15-17	1950, 1955, ..., 2050
pop both sexes age 15-24	1950, 1955, ..., 2050

pop both sexes age 15-49	1950, 1955, ..., 2050
pop both sexes age 15-59	1950, 1955, ..., 2050
pop both sexes age 15-64	1950, 1955, ..., 2050
pop both sexes age 18+	1950, 1955, ..., 2050
pop both sexes age 18-23	1950, 1955, ..., 2050
pop both sexes age 20+	1950, 1955, ..., 2050
pop both sexes age 20-64	1950, 1955, ..., 2050
pop both sexes age 20-69	1950, 1955, ..., 2050
pop both sexes age 25+	1950, 1955, ..., 2050
pop both sexes age 25-69	1950, 1955, ..., 2050
pop both sexes age 50+	1950, 1955, ..., 2050
pop both sexes age 5-14	1950, 1955, ..., 2050
pop both sexes age 60+	1950, 1955, ..., 2050
pop both sexes age 65+	1950, 1955, ..., 2050
pop both sexes age 70+	1950, 1955, ..., 2050
pop both sexes age 75+	1950, 1955, ..., 2050
pop both sexes age 85+	1950, 1955, ..., 2050
pop both sexes age 90+	1950, 1955, ..., 2050
pop female age 0-14	1950, 1955, ..., 2050
pop female age 0-17	1950, 1955, ..., 2050
pop female age 0-19	1950, 1955, ..., 2050
pop female age 0-24	1950, 1955, ..., 2050
pop female age 100+	1950, 1955, ..., 2050
pop female age 12-14	1950, 1955, ..., 2050
pop female age 15+	1950, 1955, ..., 2050
pop female age 15-17	1950, 1955, ..., 2050
pop female age 15-24	1950, 1955, ..., 2050
pop female age 15-49	1950, 1955, ..., 2050
pop female age 15-59	1950, 1955, ..., 2050
pop female age 15-64	1950, 1955, ..., 2050
pop female age 18+	1950, 1955, ..., 2050
pop female age 18-23	1950, 1955, ..., 2050
pop female age 20+	1950, 1955, ..., 2050
pop female age 20-64	1950, 1955, ..., 2050
pop female age 20-69	1950, 1955, ..., 2050
pop female age 25+	1950, 1955, ..., 2050
pop female age 25-69	1950, 1955, ..., 2050
pop female age 50+	1950, 1955, ..., 2050
pop female age 5-14	1950, 1955, ..., 2050
pop female age 60+	1950, 1955, ..., 2050
pop female age 6-11	1950, 1955, ..., 2050
pop female age 65+	1950, 1955, ..., 2050
pop female age 70+	1950, 1955, ..., 2050
pop female age 75+	1950, 1955, ..., 2050
pop female age 80-84	1950, 1955, ..., 2050
pop female age 85+	1950, 1955, ..., 2050
pop female age 85-89	1950, 1955, ..., 2050
pop female age 90+	1950, 1955, ..., 2050
pop female age 90-94	1950, 1955, ..., 2050
pop female age 95-99	1950, 1955, ..., 2050
pop male age 0-14	1950, 1955, ..., 2050
pop male age 0-17	1950, 1955, ..., 2050
pop male age 0-19	1950, 1955, ..., 2050

pop male age 0-24	1950, 1955, ..., 2050
pop male age 100+	1950, 1955, ..., 2050
pop male age 12-14	1950, 1955, ..., 2050
pop male age 15+	1950, 1955, ..., 2050
pop male age 15-17	1950, 1955, ..., 2050
pop male age 15-24	1950, 1955, ..., 2050
pop male age 15-49	1950, 1955, ..., 2050
pop male age 15-59	1950, 1955, ..., 2050
pop male age 15-64	1950, 1955, ..., 2050
pop male age 18+	1950, 1955, ..., 2050
pop male age 18-23	1950, 1955, ..., 2050
pop male age 20+	1950, 1955, ..., 2050
pop male age 20-64	1950, 1955, ..., 2050
pop male age 20-69	1950, 1955, ..., 2050
pop male age 25+	1950, 1955, ..., 2050
pop male age 25-69	1950, 1955, ..., 2050
pop male age 50+	1950, 1955, ..., 2050
pop male age 5-14	1950, 1955, ..., 2050
pop male age 60+	1950, 1955, ..., 2050
pop male age 6-11	1950, 1955, ..., 2050
pop male age 65+	1950, 1955, ..., 2050
pop male age 70+	1950, 1955, ..., 2050
pop male age 75+	1950, 1955, ..., 2050
pop male age 80-84	1950, 1955, ..., 2050
pop male age 85+	1950, 1955, ..., 2050
pop male age 85-89	1950, 1955, ..., 2050
pop male age 90+	1950, 1955, ..., 2050
pop male age 90-94	1950, 1955, ..., 2050
pop male age 95-99	1950, 1955, ..., 2050
pop median age	1950, 1955, ..., 2050
pop sex ratio age 0-14	1950, 1955, ..., 2050
pop sex ratio age 0-17	1950, 1955, ..., 2050
pop sex ratio age 0-19	1950, 1955, ..., 2050
pop sex ratio age 0-24	1950, 1955, ..., 2050
pop sex ratio age 0-4	1950, 1955, ..., 2050
pop sex ratio age 12-14	1950, 1955, ..., 2050
pop sex ratio age 15+	1950, 1955, ..., 2050
pop sex ratio age 15-17	1950, 1955, ..., 2050
pop sex ratio age 15-24	1950, 1955, ..., 2050
pop sex ratio age 15-49	1950, 1955, ..., 2050
pop sex ratio age 15-59	1950, 1955, ..., 2050
pop sex ratio age 15-64	1950, 1955, ..., 2050
pop sex ratio age 18+	1950, 1955, ..., 2050
pop sex ratio age 18-23	1950, 1955, ..., 2050
pop sex ratio age 20+	1950, 1955, ..., 2050
pop sex ratio age 20-64	1950, 1955, ..., 2050
pop sex ratio age 20-69	1950, 1955, ..., 2050
pop sex ratio age 25+	1950, 1955, ..., 2050
pop sex ratio age 25-69	1950, 1955, ..., 2050
pop sex ratio age 50+	1950, 1955, ..., 2050
pop sex ratio age 5-14	1950, 1955, ..., 2050
pop sex ratio age 60+	1950, 1955, ..., 2050
pop sex ratio age 6-11	1950, 1955, ..., 2050

pop sex ratio age 65+	1950, 1955, ..., 2050
pop sex ratio age 70+	1950, 1955, ..., 2050
pop sex ratio age 75+	1950, 1955, ..., 2050
pop sex ratio age 80+	1950, 1955, ..., 2050
pop sex ratio age 85+	1950, 1955, ..., 2050
pop sex ratio age 90+	1950, 1955, ..., 2050
pop sex ratio age all	1950, 1955, ..., 2050

2 - Data from CO2 Emissions (UNFCCC 2009 and CDIAC 2008) table co2c_cdiac_unfcc

Annex 2.1 - DESCRIPTION OF GEOGRAPHICAL UNITS (PARTITION IN 96) FROM CHELEM

United States	*United States of America (including Puerto Rico and US Virgin Islands in TRADE, US Samoa, Guam, US Virgin Islands and Puerto Rico in BOP)*
Canada	*Canada*
France	*France, Monaco (including French overseas departments in TRADE, and French overseas departments and territories in BOP)*
BLEU	*Belgium, Luxembourg*
Germany	*Germany (including East Germany since 1991)*
Italy	*Italy (including San Marino and the Holy See)*
Netherlands	*Netherlands*
United Kingdom	*United Kingdom of Great Britain and Northern Ireland*
Ireland	*Ireland*
Denmark	*Denmark*
Finland	*Finland*
Norway	*Norway (including Svalbard and Jan Mayen)*
Sweden	*Sweden*
Iceland	*Iceland (and Faroe Islands in TRADE)*
Austria	*Austria*
Switzerland	*Switzerland (including Liechtenstein in TRADE)*
Spain	*Spain*
Greece	*Greece*
Portugal	*Portugal*
Turkey	*Turkey*
Israel	*Israel*
Serbia and Montenegro	*Federal Republic of Yugoslavia (including Macedonia in TRADE in 1992)*
Bosnia and Herzegovina	*Bosnia and Herzegovina*
Croatia	*Croatia*
Macedonia, Republic of	*Macedonia, Republic of*
Slovenia	*Slovenia*
Others in south Europe	*Andorra (in TRADE only), Cyprus, Gibraltar, Malta, West Bank and Gaza (in GDP and BOP only)*
Japan	*Japan*
Australia	*Australia*
New Zealand	*New Zealand*
Southafrican Union	*Botswana, Lesotho, Namibia, South Africa, Swaziland*
Venezuela	*Venezuela*
Ecuador	*Ecuador*
Mexico	*Mexico*
Brazil	*Brazil*
Argentina	*Argentina*
Chile	*Chile*
Colombia	*Colombia*
Peru	*Peru*
Bolivia	*Bolivia*
Paraguay	*Paraguay*
Uruguay	*Uruguay*
Others in America	*Anguilla (in BOP and TRADE), Antigua and Barbuda, Aruba, Bahamas, Barbados, Belize, Bermuda, Costa Rica, Cuba, Dominica, Dominican Republic, El Salvador, French Guiana (in GDP only), Grenada, Guadeloupe (in GDP only), Guatemala, Guyana, Haiti, Honduras, Jamaica, Martinique (in GDP only), Montserrat (in BOP and TRADE), Netherland Antilles, Nicaragua, Panama, Puerto Rico (in GDP only), Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, US Virgin Islands (in GDP only), and all others in America (in TRADE only)*

Algeria	*Algeria*
Morocco	*Morocco (including Western Sahara in BOP)*
Tunisia	*Tunisia*
Egypt	*Egypt*
Libyan Arab Jamahiriya	*Libyan Arab Jamahiriya*
Saudi Arabia	*Saudi Arabia*
Gulf nes	*Bahrein, Iran, Iraq, Kuwait, Oman, Qatar, United Arab Emirates*
Middle East, no OPEC	*Jordan, Lebanon, Syria, Yemen*
Nigeria	*Nigeria*
Gabon	*Gabon*
Cameroon	*Cameroon*
Cote d'Ivoire	*Cote d'Ivoire*
Kenya	*Kenya*
Africa (others)	*Congo, Ghana, Mauritius, Reunion (in GDP only), Seychelles, Western Sahara (in GDP and TRADE), Zimbabwe, and all others in Africa (in TRADE only)*
African LDCs	*Angola, Benin, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Cote d'Ivoire, Democratic Republic of Congo (formerly Zaire), Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gambia, Guinea, Guinea-Bissau, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Niger, Rwanda, Sao Tome and Principe, Senegal, Sierra Leone, Somalia, Sudan, Tanzania, Togo, Uganda, Zambia*
Indonesia	*Indonesia*
India	*India*
South Korea	*Republic of Korea*
Hong Kong	*Hong Kong Special Administrative Region of China*
Singapore	*Singapore*
Taiwan	*Taiwan*
Malaysia	*Malaysia*
Philippines	*Philippines*
Thailand	*Thailand*
Pakistan	*Pakistan*
Brunei Darussalam	*Brunei Darussalam*
Bangladesh	*Bangladesh*
Sri Lanka	*Sri Lanka*
East Asia nes, others	*Fiji, French Polynesia (in GDP and TRADE), Guam (in GDP and TRADE), Macao, Mongolia, New Caledonia (in GDP and TRADE), North Korea, Pacific Islands (in GDP and TRADE), Papua New Guinea, Tonga, US Samoa (in GDP and TRADE), Vanuatu, Western Samoa, and all others in Asia and Oceania (in TRADE only)*
East Asian LDCs	*Afghanistan, Bhutan, Kiribati, Maldives, Myanmar, Nepal, Solomon Islands, Vanuatu, Western Samoa*
Russian Federation	*Russian Federation*
Ukraine	*Ukraine*
Belarus	*Belarus*
Kazakhstan	*Kazakhstan*
Kyrgyzstan	*Kyrgyzstan*
Caucasus	*Armenia, Azerbaijan, Georgia*
Other CIS	*Moldova, Tajikistan, Turkmenistan, Uzbekistan*
Estonia	*Estonia*
Latvia	*Latvia*
Lithuania	*Lithuania*
Bulgaria	*Bulgaria*
Czech Republic	*Czech Republic*
Slovakia	*Slovakia*
Hungary	*Hungary*
Poland	*Poland*
Romania	*Romania*
Former German Democratic Rep.	*Former German Democratic Republic (up to 1990)*
Albania	*Albania*

China, People's Rep.	*The People's Republic of China: Mainland*
Viet Nam	*Viet Nam*
Cambodia, Lao PDR	*Cambodia, Lao PDR*
Miscellaneous	*Not elsewhere specified (international organizations in BOP)*
World	***Total-of-the-33-Areas***

Annex 2.2. - DESCRIPTION OF GEOGRAPHICAL UNITS (168 UNITS) FROM ESPON 2006 PROGRAM (EUROPE IN THE WORLD)

WUTS5_Names	*Note*
Afghanistan	
Angola	
Albania	
United Arab Emirates	
Argentina	
Armenia	
Australia	
Austria	
Azerbaijan	
Burundi	
Belgium	
Benin	
Burkina Faso	
Bangladesh	
Bulgaria	
Bahrain	
Bahamas	
Bosnia and Herzegovina	
Belarus	
Belize	
Bolivia	
Brazil	
Bhutan	
Botswana	
Central African Republic	
Canada	
Switzerland	
Chile	
China	*(China main land + Macao + Hong-Kong)*
Côte d'Ivoire	
Cameroon	
Congo, Dem. Rep. of the	
Congo	
Colombia	
Costa Rica	
Cuba	
Cyprus	
Czech Republic	
Germany	
Djibouti	
Denmark	
Dominican Republic	
Algeria	
Ecuador	
Egypt	
Eritrea	
West Sahara	
Spain	

Estonia	
Ethiopia	
Finland	
Fiji	
France	*France (Mainland) + Guadeloupe + Martinique + Guyane + Réunion*
Gabon	
United Kingdom	
Georgia	
Ghana	
Guinea	
Gambia	
Guinea-Bissau	
Equatorial Guinea	
Greece	
Greenland	
Guatemala	
Guyana	
Honduras	
Croatia	
Haiti	
Hungary	
Indonesia	
India	
Ireland	
Iran, Islamic Rep. of	
Iraq	
Iceland	
Israel	*Israel (without Occupied Palestinian Territories)*
Italy	
Jamaica	
Jordan	
Japan	
Kazakhstan	
Kenya	
Kyrgyzstan	
Cambodia	
Korea, Rep. of	
Kuwait	
Lao People's Dem. Rep.	
Lebanon	
Liberia	
Libyan Arab Jamahiriya	
Sri Lanka	
Lesotho	
Lithuania	
Luxembourg	
Latvia	
Morocco	*Morocco (without Western Sahara)*
Moldova, Rep. of	
Madagascar	
Mexico	
Macedonia, TFYR	
Mali	
Malta	
Myanmar	
Mongolia	
Mozambique	

Mauritania
Mauritius
Malawi
Malaysia
Namibia
Niger
Nigeria
Nicaragua
Netherlands
Norway
Nepal
New Zealand
Oman
Pakistan
Panama
Peru
Philippines
Papua New Guinea
Poland
Puerto Rico
North Korea
Portugal
Paraguay
Qatar
Romania
Russian Federation
Rwanda
Saudi Arabia
Serbia/Montenegro
Sudan
Senegal
Singapore
Sierra Leone
El Salvador
Somalia
Suriname
Slovakia
Slovenia
Sweden
Swaziland
Syrian Arab Republic
Chad
Togo
Thailand
Tajikistan
Turkmenistan
Trinidad and Tobago
Tunisia
Turkey
Taiwan
Tanzania, U. Rep. of
Uganda
Ukraine
Uruguay
United States
Uzbekistan
Venezuela

Viet Nam
Occupied Palestinian Territories
Yemen
South Africa
Zambia
Zimbabwe

Annex 2.3 - DESCRIPTION OF GEOGRAPHICAL UNITS from GEO

un_cnty_name	*notes_name*
Aruba	
Afghanistan	
Angola	
Anguilla	
Albania	
Andorra	
Netherlands Antilles	
United Arab Emirates	
Argentina	
Armenia	
American Samoa	
Antarctic	
Antigua and Barbuda	
Australia	
Austria	
Azerbaijan	
Burundi	
Belgium	
Benin	
Burkina Faso	
Bangladesh	
Bulgaria	
Bahrain	
Bahamas	
Bosnia and Herzegovina	
Belarus	
Belize	
Bermuda	
Bolivia	
Brazil	
Barbados	
Brunei Darussalam	
Bhutan	
Botswana	
Central African Republic	
Canada	
Cocos (Keeling) Islands	
Switzerland	
Chile	
China	*Including Macau, Hong Kong and Taiwan*
Cote d'Ivoire	
Cameroon	
Democratic Republic of the Congo	
Congo	
Cook Islands	
Colombia	

Comoros
Cape Verde
Costa Rica
Cuba
Christmas Island
Cayman Islands
Cyprus
Czech Republic
Germany
Djibouti
Dominica
Denmark
Dominican Republic
Algeria
Ecuador
Egypt
Eritrea
Western Sahara
Spain
Estonia
Ethiopia
Finland
Fiji
Falkland Islands (Malvinas)
France
Faroe Islands
Micronesia (Federated States of)
Gabon
United Kingdom of Great Britain and Northern Ireland
Georgia
Guernsey
Ghana
Gibraltar
Guinea
Guadeloupe
Gambia
Guinea-Bissau
Equatorial Guinea
Greece
Grenada
Greenland
Guatemala
French Guiana
Guam
Guyana
Honduras
Croatia
Haiti
Hungary
Indonesia
Isle of Man
India
Ireland

Iran (Islamic Republic of)
Iraq
Iceland
Israel
Italy
Jamaica
Jersey
Jordan
Japan
Johnston Atoll
Kazakhstan
Kenya
Kyrgyzstan
Cambodia
Kiribati
Saint Kitts and Nevis
Republic of Korea
Kuwait
Lao People's Democratic Republic
Lebanon
Liberia
Libyan Arab Jamahiriya
Saint Lucia
Liechtenstein
Sri Lanka
Lesotho
Lithuania
Luxembourg
Latvia
Morocco
Monaco
Moldova, Republic of
Madagascar
Maldives
Mexico
Marshall Islands
Midway Islands
The former Yugoslav Republic of Macedonia
Mali
Malta
Myanmar
Montenegro
Mongolia
Northern Mariana Islands
Mozambique
Mauritania
Montserrat
Martinique
Mauritius
Malawi
Malaysia
Mayotte
Namibia

New Caledonia	
Niger	
Norfolk Island	
Nigeria	
Nicaragua	
Niue	
Netherlands	
Norway	
Nepal	
Nauru	
New Zealand	
Oman	
Pakistan	
Panama	
Pitcairn Island	
Peru	
Philippines	
Palau	
Papua New Guinea	
Poland	
Puerto Rico	
Democratic People's Republic of Korea	
Portugal	
Paraguay	
Occupied Palestinian Territory	*Including West Bank and Gaza*
French Polynesia	
Qatar	
Reunion	
Romania	
Russian Federation	
Rwanda	
Saudi Arabia	
Sudan	
Senegal	
Singapore	
Saint Helena	
Svalbard and Jan Mayen Islands	
Solomon Islands	
Sierra Leone	
El Salvador	
San Marino	
Somalia	
Saint Pierre and Miquelon	
Serbia	*Including Kosovo*
Sao Tome and Principe	
Suriname	
Slovakia	
Slovenia	
Sweden	
Swaziland	
Seychelles	
Syrian Arab Republic	
Turks and Caicos Islands	

Chad
Togo
Thailand
Tajikistan
Tokelau
Turkmenistan
Timor-Leste
Tonga
Trinidad and Tobago
Tunisia
Turkey
Tuvalu
United Republic of Tanzania
Uganda
Ukraine
Uruguay
United States of America
Uzbekistan
Holy See
Saint Vincent and the Grenadines
Venezuela
British Virgin Islands
United States Virgin Islands
Viet Nam
Vanuatu
Wake Island
Wallis and Futuna
Samoa
Yemen
South Africa
Zambia
Zimbabwe

Annex 2.4 - DESCRIPTION OF GEOGRAPHICAL UNITS from WDI

wdi_cnty_name	*notes_name*
Afghanistan	
Albania	
Algeria	
American Samoa	
Andorra	
Angola	
Antigua and Barbuda	
Argentina	
Armenia	
Aruba	
Australia	
Austria	
Azerbaijan	
Bahamas, The	
Bahrain	
Bangladesh	
Barbados	
Belarus	
Belgium	
Belize	
Benin	
Bermuda	
Bhutan	
Bolivia	
Bosnia and Herzegovina	
Botswana	
Brazil	
Brunei Darussalam	
Bulgaria	
Burkina Faso	
Burundi	
Cambodia	
Cameroon	
Canada	
Cape Verde	
Cayman Islands	
Central African Republic	
Chad	
Channel Islands	
Chile	
China	*Unless otherwise noted, data for China do not include data for Hong Kong, Macau, or Taiwan*
Colombia	
Comoros	
Congo, Dem. Rep.	
Congo, Rep.	
Costa Rica	

Cote d'Ivoire	
Croatia	
Cuba	
Cyprus	*Data related to GDP, exclude Turkish-controlled area*
Czech Republic	
Denmark	
Djibouti	
Dominica	
Dominican Republic	
Ecuador	
Egypt, Arab Rep.	
El Salvador	
Equatorial Guinea	
Eritrea	
Estonia	
Ethiopia	
Faeroe Islands	
Fiji	
Finland	
France	*Data related to GDP, include French Guiana, Guadelupe, Martinique and Réunion*
French Guiana	
French Polynesia	
Gabon	
Gambia, The	
Georgia	
Germany	
Ghana	
Greece	
Greenland	
Grenada	
Guadeloupe	
Guam	
Guatemala	
Guinea	
Guinea-Bissau	
Guyana	
Haiti	
Honduras	
Hong Kong, China	
Hungary	
Iceland	
India	
Indonesia	
Iran, Islamic Rep.	
Iraq	
Ireland	
Isle of Man	
Israel	
Italy	
Jamaica	
Japan	
Jordan	
Kazakhstan	

Kenya	
Kiribati	
Korea, Dem. Rep.	
Korea, Rep.	
Kuwait	
Kyrgyz Republic	
Lao PDR	
Latvia	
Lebanon	
Lesotho	
Liberia	
Libya	
Liechtenstein	
Lithuania	
Luxembourg	
Macao, China	
Macedonia, FYR	
Madagascar	
Malawi	
Malaysia	
Maldives	
Mali	
Malta	
Marshall Islands	
Martinique	
Mauritania	
Mauritius	
Mayotte	
Mexico	
Micronesia, Fed. Sts.	
Moldova	*Data related to GDP, exclude Transnistria*
Monaco	
Mongolia	
Montenegro	
Morocco	
Mozambique	
Myanmar	
Namibia	
Nauru	
Nepal	
Netherlands	
Netherlands Antilles	
New Caledonia	
New Zealand	
Nicaragua	
Niger	
Nigeria	
Northern Mariana Islands	
Norway	
Oman	
Pakistan	
Palau	
Panama	

Papua New Guinea	
Paraguay	
Peru	
Philippines	
Poland	
Portugal	
Puerto Rico	
Qatar	
Reunion	
Romania	
Russian Federation	
Rwanda	
Samoa	
San Marino	
Sao Tome and Principe	
Saudi Arabia	
Senegal	
Serbia	*Where available, data from Serbia and Montenegro are shown separately: However some indicators for Serbia prior to 2006 include data for Montenegro*
Seychelles	
Sierra Leone	
Singapore	
Slovak Republic	
Slovenia	
Solomon Islands	
Somalia	
South Africa	
Spain	
Sri Lanka	
St. Kitts and Nevis	
St. Lucia	
St. Vincent and the Grenadines	
Sudan	
Suriname	
Swaziland	
Sweden	
Switzerland	
Syrian Arab Republic	
Tajikistan	
Tanzania	*Data related to GDP, cover mainland Tanzania only*
Thailand	
Timor-Leste	
Togo	
Tonga	
Trinidad and Tobago	
Tunisia	
Turkey	
Turkmenistan	
Tuvalu	
Uganda	
Ukraine	
United Arab Emirates	
United Kingdom	
United States	

Uruguay
Uzbekistan
Vanuatu
Venezuela, RB
Vietnam
Virgin Islands (U.S.)
West Bank and Gaza
Yemen, Rep.
Zambia
Zimbabwe

Annex 2.5 - DESCRIPTION OF GEOGRAPHICAL UNITS from UN (WPP08)

un_cnty_name	*notes_name*
Afghanistan	
Aland Islands	
Albania	
Algeria	
American Samoa	
Andorra	
Angola	
Anguilla	
Antigua and Barbuda	
Argentina	
Armenia	
Aruba	
Australia	*Including Christmas Island, Cocos (Keeling) Islands, and Norfolk Island.*
Austria	
Azerbaijan	
Bahamas	
Bahrain	
Bangladesh	
Barbados	
Belarus	
Belgium	
Belize	
Benin	
Bermuda	
Bhutan	
Bolivia	
Bosnia and Herzegovina	
Botswana	
Brazil	
British Virgin Islands	
Brunei Darussalam	
Bulgaria	
Burkina Faso	
Burundi	
Cambodia	
Cameroon	
Canada	
Cape Verde	
Cayman Islands	
Central African Republic	
Chad	
Channel Islands	*Refers to Guernsey, and Jersey.*
Chile	
China	*For statistical purposes, the data for China do not include Hong Kong and Macao, Special Administrative Regions (SAR) of China.*
Colombia	
Comoros	

Congo	
Cook Islands	
Costa Rica	
Cote d'Ivoire	
Croatia	
Cuba	
Cyprus	
Czech Republic	
Democratic People's Republic of Korea	
Democratic Republic of the Congo	
Denmark	
Djibouti	
Dominica	
Dominican Republic	
Ecuador	
Egypt	
El Salvador	
Equatorial Guinea	
Eritrea	
Estonia	
Ethiopia	
Faeroe Islands	
Falkland Islands (Malvinas)	
Fiji	
Finland	*Including Åland Islands.*
France	
French Guiana	
French Polynesia	
Gabon	
Gambia	
Georgia	
Germany	
Ghana	
Gibraltar	
Greece	
Greenland	
Grenada	
Guadeloupe	
Guam	
Guatemala	
Guernsey	
Guinea	
Guinea-Bissau	
Guyana	
Haiti	
Holy See	*Refers to the Vatican City State.*
Honduras	
Hong Kong Special Administrative Region of China	*As of 1 July 1997, Hong Kong became a Special Administrative Region (SAR) of China.*
Hungary	
Iceland	
India	
Indonesia	
Iran, Islamic Republic of	

Iraq	
Ireland	
Isle of Man	
Israel	
Italy	
Jamaica	
Japan	
Jersey	
Jordan	
Kazakhstan	
Kenya	
Kiribati	
Kuwait	
Kyrgyzstan	
Lao People's Democratic Republic	
Latvia	
Lebanon	
Lesotho	
Liberia	
Libyan Arab Jamahiriya	
Liechtenstein	
Lithuania	
Luxembourg	
Macao Special Administrative Region of China	*As of 20 December 1999, Macao became a Special Administrative Region (SAR) of China.*
Madagascar	
Malawi	
Malaysia	
Maldives	
Mali	
Malta	
Marshall Islands	
Martinique	
Mauritania	
Mauritius	*Including Agalega, Rodrigues, and Saint Brandon.*
Mayotte	
Mexico	
Micronesia, Federated States of	
Monaco	
Mongolia	
Montenegro	
Montserrat	
Morocco	
Mozambique	
Myanmar	
Namibia	
Nauru	
Nepal	
Netherlands	
Netherlands Antilles	
New Caledonia	
New Zealand	
Nicaragua	
Niger	

Nigeria	
Niue	
Norfolk Island	
Northern Mariana Islands	
Norway	*Including Svalbard and Jan Mayen Islands.*
Occupied Palestinian Territory	*Including West Bank and Gaza*
Oman	
Pakistan	
Palau	
Panama	
Papua New Guinea	
Paraguay	
Peru	
Philippines	
Pitcairn	
Poland	
Portugal	
Puerto Rico	
Qatar	
Republic of Korea	
Republic of Moldova	
Réunion	
Romania	
Russian Federation	
Rwanda	
Saint Helena	*Including Ascension, and Tristan da Cunha.*
Saint Kitts and Nevis	
Saint Lucia	
Saint Pierre and Miquelon	
Saint Vincent and the Grenadines	
Saint-Barthélemy	
Saint-Martin (French part)	
Samoa	
San Marino	
Sao Tome and Principe	
Saudi Arabia	
Senegal	
Serbia	
Seychelles	
Sierra Leone	
Singapore	
Slovakia	
Slovenia	
Solomon Islands	
Somalia	
South Africa	
Spain	
Sri Lanka	
Sudan	
Suriname	
Svalbard and Jan Mayen Islands	
Swaziland	
Sweden	

Switzerland	
Syrian Arab Republic	
Tajikistan	
Thailand	
The former Yugoslav Republic of Macedonia	*The former Yugoslav Republic of Macedonia.*
Timor-Leste	
Togo	
Tokelau	
Tonga	
Trinidad and Tobago	
Tunisia	
Turkey	
Turkmenistan	
Turks and Caicos Islands	
Tuvalu	
Uganda	
Ukraine	
United Arab Emirates	
United Kingdom of Great Britain and Northern Ireland	
United Republic of Tanzania	
United States of America	
United States Virgin Islands	
Uruguay	
Uzbekistan	
Vanuatu	
Venezuela (Bolivarian Republic of)	
Viet Nam	
Wallis and Futuna Islands	
Western Sahara	
Yemen	
Zambia	
Zimbabwe	

References

- ***Websites***

CHELEM Database (harmonised counts on exchanges and the world economy), built by CEPII is known since a lot of years as a precious tool for analysing the global World Economy: http://www.cepii.fr/francgraph/bdd/chelem.htm

Geo Data Portal *i*s the authoritative source for data sets used by UNEP and its partners in the Global Environment Outlook (GEO) report and other integrated environment assessments. Its online database holds more than 500 different variables, as national, subregional, regional and global statistics or as geospatial data sets (maps), covering themes like Freshwater, Population, Forests, Emissions, Climate, Disasters, Health and GDP. Display them on-the-fly as maps, graphs, data tables or download the data in different formats: http://geodata.grid.unep.ch/

United Nations database: http://unstats.un.org/unsd/default.htm

World Development Indicators **Online (*WDI*)** provides direct access to more than 800 development indicators, with time series for 209 countries and 18 country groups from 1960 to 2008, where data are available: http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20398986~menuPK:64133163~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html

TECHNICAL REPORT

ESPON

ANALYSIS OF THE AVAILABILITY AND THE QUALITY OF DATA ON WESTERN BALKANS AND TURKEY

CONTENT

- General assessment. This part discusses the Spatial Administrative Divisions of WB and Turkey and presents the set of data delivered
- Assessment per country. This part presents an assessment of the availability and the quality of data per country of WB and Turkey.
- Conclusions on the data availability at NUTS0 to NUTS3 levels and the inclusion of WB and Turkey data in the ESPON Database.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

49 PAGES

LIST OF AUTHORS

Author of the Report and main researcher:
Minas Angelidis, National Technical University of Athens (NTUA)

Contributions:

Gabriella Karka (in parts of the Report), NTUA

Kostas Santimpantakis (in specific parts of data process), NTUA

Epameinondas Tsigkas (in specific parts of data process), NTUA

Vivian Bazoula (in specific parts of data process), NTUA

Contact

angelidi@central.ntua.gr

tel. + 30 210 7721731

TABLE OF CONTENT

Abbreviations

CC: Candidate Countries
FBiH: Federation of Bosnia and Herzegovina
GDP: Gross Domestic Product
NUTS: Nomenclature of Territorial Units for Statistics
PCC: Potential Candidate Countries
WB: Western Balkans

List of Tables, Maps and Figures

Introduction and methodological remarks

The present Technical Report corresponds to the Challenge 11 of the ESPON 2013 project: "Enlargement to Neighbourhood".

Key findings of our work are:

- Availability and quality of the data on the Candidate Countries (CC) and Potential Candidate Countries (PCC): the Western Balkans countries and Turkey, at NUTS2 level is in general terms satisfactory
- Data availability and quality on CC and PCC at NUTS3 level, which is the most challenging for the needs of ESPON, is almost fully satisfactory for Croatia, FYROM and Turkey which have adopted the NUTS classification while it is satisfactory for a wide number of issues for the rest CC / PCC.

Summary of the methodological issues

In order to ensure a sound comparability of data of the CC and PCC which have not adopted the NUTS classification, we have classified the existing administrative units of these countries at different territorial levels in "similar NUTS" territorial units. We have used for this purpose the criterion of population potential of the EU NUTS classification as well as the overall structure of government in these countries with focus on the power of the respective regional and local authorities and the main features of territorial development in each administrative level per country.

The implementation of this method ensured that the "similar NUTS" divisions correspond almost fully with the respective divisions for the EU countries and could be further used in the definition of "similar NUTS" divisions in the Eastern Neighbouring countries (ENC) and the Southern Mediterranean Neighbouring countries (MNC).

Introduction and discussion of the methodological options

The part of the ESPON Database 2013 project referred to the ***Western Balkans countries and Turkey*** data was the main part of the Challenge 11 of the project: "Enlargement to neighbourhood". It aimed to extend the pool of data on the ESPON countries on the WB and Turkey as well as to ensure that the relevant data be harmonized with the rest of the ESPON Database. Some of the methodological conclusions of this work could be used for the future work on the inclusion of data for Eastern Neighbouring countries (ENC) and the Southern Mediterranean Neighbouring countries (MNC) in the scope of the 2nd part of the ESPON database 2013.

Western Balkans countries and Turkey are ***Candidate Countries (CC)*** or ***Potential Candidate Countries (PCC)***. Specifically:

According to the overall enlargement strategy of the EU document adopted by the Commission on 8.11.2006 (http://ec.europa.eu/enlargement/countries/index_en.htm) ***Croatia*** and ***Turkey*** are ***Candidate Countries***. In December 2005, the European Council granted the ***Former Yugoslav Republic of Macedonia (FYROM)*** the status

of a ***Candidate Country***; The European Council of 16-17.12.2010 agreed to give ***Montenegro*** the status of ***Candidate country***; accession negotiations with these two countries have not started.

Albania, Bosnia and Herzegovina and Serbia including Kosovo (Under UN Security Council Resolution 1244) are ***Potential Candidate Countries***: See in more detail in the 8.11.06 document and the corresponding following documents.

We accomplished during 2009 and 2010, the evaluation of the situation of data available in these countries, following the relevant methodology and preliminary studies elaborated in ESPON 2006. We also assessed how it is possible to establish contacts with the national statistical offices of these countries and ensure a regular dataflow among them and the ESPON 2013 Database Project.

Apart from the discussion of the issues concerning the WB and Turkey data, we refer to the data delivered to the Lead Partner of the project, which have been gradually integrated in the ESPON 2013 database.

We comment here on the WB and Turkey data at NUTS0 to NUTS3 levels.

NUTS and regional / territorial classification

We first had to assess ***the conformity of the WB and Turkey spatial administrative divisions to the EU NUTS classification criteria.***

According to Eurostat, **NUTS – Nomenclature of Territorial Units for Statistics**, 2010[1], the NUTS classification is a hierarchical system for dividing up the economic territory of the EU for the purpose of:

- The collection, development and harmonisation of EU regional statistics:
- Socio-economic analyses of the regions.
    - NUTS 1: major socio-economic regions
    - NUTS 2: basic regions for the application of regional policies
    - NUTS 3: as small regions for specific diagnoses
- Framing of EU regional policies.
    - Regions eligible for aid from the Structural Funds (Objective 1) have been classified at the NUTS 2 level.
    - Areas eligible under the other priority objectives have mainly been classified at the NUTS 3 level.
    - The Cohesion report has so far mainly been prepared at the NUTS 2 level.

Principles and Characteristics of NUTS

- **Principle 1**: The NUTS regulation defines minimum and maximum population thresholds for the size of the NUTS regions:

[1] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature Last update 28.07.2010

NUTS 1	3 million	7 million
NUTS 2	800 000	3 million
NUTS 3	150 000	800 000

Despite the aim of ensuring that regions of comparable size all appear at the same NUTS level, each level still contains regions which differ greatly in terms of population.

- **Principle 2**: NUTS favours administrative divisions (normative criterion)

  For practical reasons the NUTS classification is based on the administrative divisions applied in the Member States that generally comprise two main regional levels. The additional third level is created by aggregating administrative units.

- **Principle 3:** NUTS favours general geographical units

  General geographical units are normally more suitable for any given indicator than geographical units specific to certain fields of activity.

EU NUTS classification uses almost exclusively the population criterion.

In the above explanatory texts of the EU NUTS classification, NUTS2 units are seen (to some extent) as "basic regions for the application of regional policies" while NUTS3 units are approached as "small regions for specific diagnoses".

Regarding NUTS2 units in the case of the Western Balkans countries which have not adopted NUTS regulation, it is obvious that these countries should define "similar to NUTS2" units which are appropriate for the application of the EU regional policies (see, among others, in: Knezevic 2010). However, they would evidently apply a political criterion: the use of existing administrative divisions as they are or the limited readjustment of the latter or the creation of new divisions depend on the political will of these countries. Evidently, they will also take into account a relevant technical criterion: the territorial features of the existing "regions".

From our point of view, we should examine if the existing administrative divisions in these countries which fulfill the population criterion for NUTS2, fulfill also the "application of regional policies" criterion. In order to better clarify this criterion, we should refer, even very concisely, to the concept of "development policy region".

Regions can be defined by physical, "functional", economic, social, cultural, environmental characteristics and so on; each of the many branches of territorial analysis and planning approaches regions according to its general theoretical scope. For the case of "development policy regions" in the PCC of WB most appropriate approach is to examine whether "similar NUTS2" divisions according to the NUTS populations criterion, corresponding to existing administrative divisions, comply with the "regional level" of the «spatial governance" system of each of the examined countries.

Even more, it is useful to see whether the "similar NUTS2" and "similar NUTS3" units which will be defined comply with the overall spatial configuration of the interested countries with focus on the hierarchy and the networking of the urban centres of these countries.

In next we will examine these territorial features mainly for each of the Potential CC; we will also examine this issue, relatively less, for the cases of the CC which have already adopted the EU NUTS classification.

1. General assessment

1.1 The WB and Turkey Spatial Administrative Divisions

Turkey, Croatia and FYROM have already adopted the NUTS classification.

The rest of the WB countries are at the present in the procedure of adopting it. *According to the assessment using the population criterion, in the majority of these last the existing administrative divisions (regions, districts etc) could be associated to the EU NUTS definitions without considerable problems.* In other words, we could create ***"similar NUTS" divisions*** which fulfil satisfactorily the NUTS population criterion - see in Tables 1.1 and 1.2 and in detail in section 2.

According to our methodological approach (see previously), we have examined, as possible, the administrative capacity of the NUTS and "similar NUTS" spatial administrative divisions of the WB and Turkey and, further on, the overall territorial features of these countries. As we will see in next, the results have more or less consolidated the selection of the "similar NUTS" divisions.

Table 1.1: NUTS 1, 2, 3 regions in Croatia, FYROM and Turkey

	NUTS 1	**NUTS 2**	S **NUTS 3**
Croatia	Country	Regiia	Counties
FYROM	Country	Country	Statistical Regions
Turkey	Regions	Sub-regions	Provinces

Table 1.2: "Similar NUTS 1, 2, 3" regions in the CC except from Croatia, FYROM and Turkey

	Similar to NUTS 1	**Similar to NUTS 2**	**Similar to NUTS 3**
Albania	Country	(Country)	12 Prefectures
BeH	Country or: FBiH, RS, Brsko district	FBiH, RS, Brsko district	10 Cantons
Serbia	Central Serbia, Voivodina	(Central Serbia, Voivodina)	21 Districts
Montenegro *	Country	Country	Country
Kosovo* **	Country	Country	(Country)

* See in more detail in the per country assessment

** Under UN Security Council Resolution 1244

Second, we examined the ***availability / quality of existing data (for the ESPON Database needs) at NUTS0 to NUTS3 levels*** in the WB and Turkey (including data allowing us to make diachronic comparisons).

We paid particular attention in finding out if there exist at each of the CC / PCC, at NUTS3 level, at least a number of "basic" data / indicators -from censuses, inventories and surveys ***already done and comparable with those realized in the EU-27 countries.***

1.2 The entire set of data required and the interim deliveries

The entire set of data required

In more detail:

The data required are mainly referred to the following aspects of NUTS3 areas:

(a) Demographic and social:

Population, households, dwellings etc per appropriate categories

(b) Economic aspects, employment: Active population, employment / unemployment, GDP etc

(c) Environmental aspects.

We give in the Annex 2 the Table 1 of the existing data per CC / PCC, per group of themes and per census / survey in which are based.

Usually realised censuses and specific statistical surveys concerning the ESPON Database data indicators:

- Population census, building / dwellings census / inventories

- Labour force survey, household budget survey etc.

Since the situation in the CC / PCC varies considerably from country to country, it was necessary to make an in depth assessment per country using:

(a) Primarily the **Eurostat data** and

(b) **Data provided by the Statistical Offices of the CC** as well as

(c) Data from a wide range of other sources: ESPON 2006 projects, ESTIA-SPOSE programme, other relevant INTERREG programmes, Wikipedia, www.citypopulation.de etc -see in References - Sources.

More specifically, the available data on CC / PCC from Eurostat are presented in the following two Annexes:

- Annex 3: short presentation of the respective Eurostat data on CC per sector (topic).

- Annex 4: full description of the respective data.

According to the assessment we have made, for the majority of categories and countries the above datasets for the ESPON Database exist at the appropriate spatial level: NUTS0, NUTS1,2,3 or "similar NUTS1,2,3".

In addition: all CC / PCC except Turkey and Kosovo (Under UN Security Council Resolution 1244) are included in CORINE Land Cover and other land based EU programs providing useful land use and environmental data.

The issue of administrative divisions' shapefiles

For Croatia, FYROM and Turkey there are NUTS shapefiles provided by Eurogeographics. There are not respective shapefiles for Albania, Bosnia and Herzegovina, Serbia, Montenegro and Kosovo (under UN Security Council Resolution 1244). Our workgroup has found non official shapefiles for these countries downloaded from the Berkeley University website. Then, these shapefiles have been appropriately adjusted by the RIATE workgroup and included in the ESPON template.

The "basic" data delivered to the Lead Partner on 2009

Our work should follow the steps of the entire Database project. Therefore, during the stages of the project in the year 2009***, we were more specifically interested in a first set of "basic" data, delivered to the LP, which are gradually integrated in the Database. This set includes more specifically, the following:***

- Total Population,
- GDP in Euros and GDP in PPS,
- Active Population and Unemployment,
- Total Population by sex and age (for the year 2005)

We provided, in addition, data for:

- Total area,
- Land area and
- Population density.

We have sent to the LP Excel tables in the format defined by the LP data for all CC / PCC, coming *from all available sources* (Eurostat, National Statistical Offices and other sources) at NUTS 0, 1, 2, 3, levels.

More specifically

- For Croatia, FYROM, Turkey most of the data is from Eurostat, but we have, also, added data from other sources.
- For the rest CC / PCC, Eurostat provides data only at NUTS0 level only for a few indicators; therefore, we mainly used data from the National Statistical Offices and other sources.

Especially in the "2009 Deliveries» for the CC PCC, are included:

(a) General tables, which include data for all sectors as well as respective ***metadata***.

(b) "Diffusion Tables" for: Total population, GDP, LFS (Labour Force Survey) data, Age pyramid.

Concerning, more specifically, the "2009 Delivery" data, the situation varies considerably according to the country -see in Table 1.3.

Table 1.3: Data delivered per CC / PCC (years 2000-2006*) and NUTS**

	Total area	**Land area**	**Total pop**	**GDP in Euros and PPS****	**Active population**	**Unem-ploy-ment**	**Total pop. By sex-age 2005****	**Pop. densi-ty**
Albania	NUTS0,1,2,3		NUTS0 NUTS3 (2001& 2004)	NUTS0***				NUTS0, 1,2,3
Bosnia & ze-govina	NUTS0,1,2,3		NUTS0 NUTS1,2,3(2007)	NUTS0***				NUTS0 (2000-2006), NUTS3(2007)
Croatia	NUTS0,1,2,3	NUTS0, 1,2,3	NUTS0,1,2,3 (2001-2007)	NUTS0,1,2,3	NUTS0 (2002-2007) NUTS0,1,2 (2007)	NUTSO,1 (2002-2005 &2007) NUTS2(2007)	NUTS0,1,2 (2005) NUTS3 (2001)	NUTS0, 1,2,3
FYROM	NUTS0,1,2	NUTS0, 1,2,3	NUTS0,1,2,3 (2000-2007)	NUTS0 NUTS1,2,3 (2004-2006)			NUTS0,1,2 (2005)	NUTS0, 1,2,3
Serbia	NUTS0 NUTS3		NUTS0 NUTS1,3(2002)	NUTS0***			NUTS3 (2002)	NUTS0 (2000-2007)
Monte-negro			NUTS0,1,2,3	NUTS0***				NUTS0 (2000-2007)
Kosovo ***			NUTS0,1,2 (2002-2006)					NUTS0, 1,2 (2002-2006)
Turkey	NUTS0,1,2,3	NUTS0, 1,2,3	NUTS0,1,2,3	NUTS0 (2000-2005) NUTS1,2,3 (2000 & 2001)				NUTS0 (2000-2007) NUTS1, 2,3 (2000-2006)

* In some cases: 2007. Total population per sex and age is given only for the year 2005.

** Or "similar NUTS" –according to the case.

*** For Croatia, FYROM and Turkey we used data from the Eurostat table for the EU countries.

For the other CC there is data for GDP taken from the Eurostat table for 'GDP and main aggregates" (only in Euros") *for Candidate Countries* (we compiled these data only in the respective "Diffusion Table"). The respective data of this table for Croatia, FYROM and Turkey does not comply with the data of the previous Table (for EU countries). We will clarify further this point later on.

**** For some CC there are data on distribution per sex and age for other years –often for 2001.

***** Under UN Security Council Resolution 1244.

2. Assessment per country

Albania

***Spatial units' levels*:**

The ***total population of the country*** amounted up to ***3.170.000 inhabitants*** in ***2008*** (Eurostat 2010).

Albania is divided into 12 prefectures (counties, Albanian: official qark/qarku, but often prefecture / prefektura), 37 districts and 351 municipalities.

Concerning the EU regulation for NUTS 3: 150 000 - 800 000 inhabitants; all Albania's prefectures, except two, have 150 000 - 800 000 inhabitants in 2001.

Prefectures have a Council and considerable competences which are gradually extended.

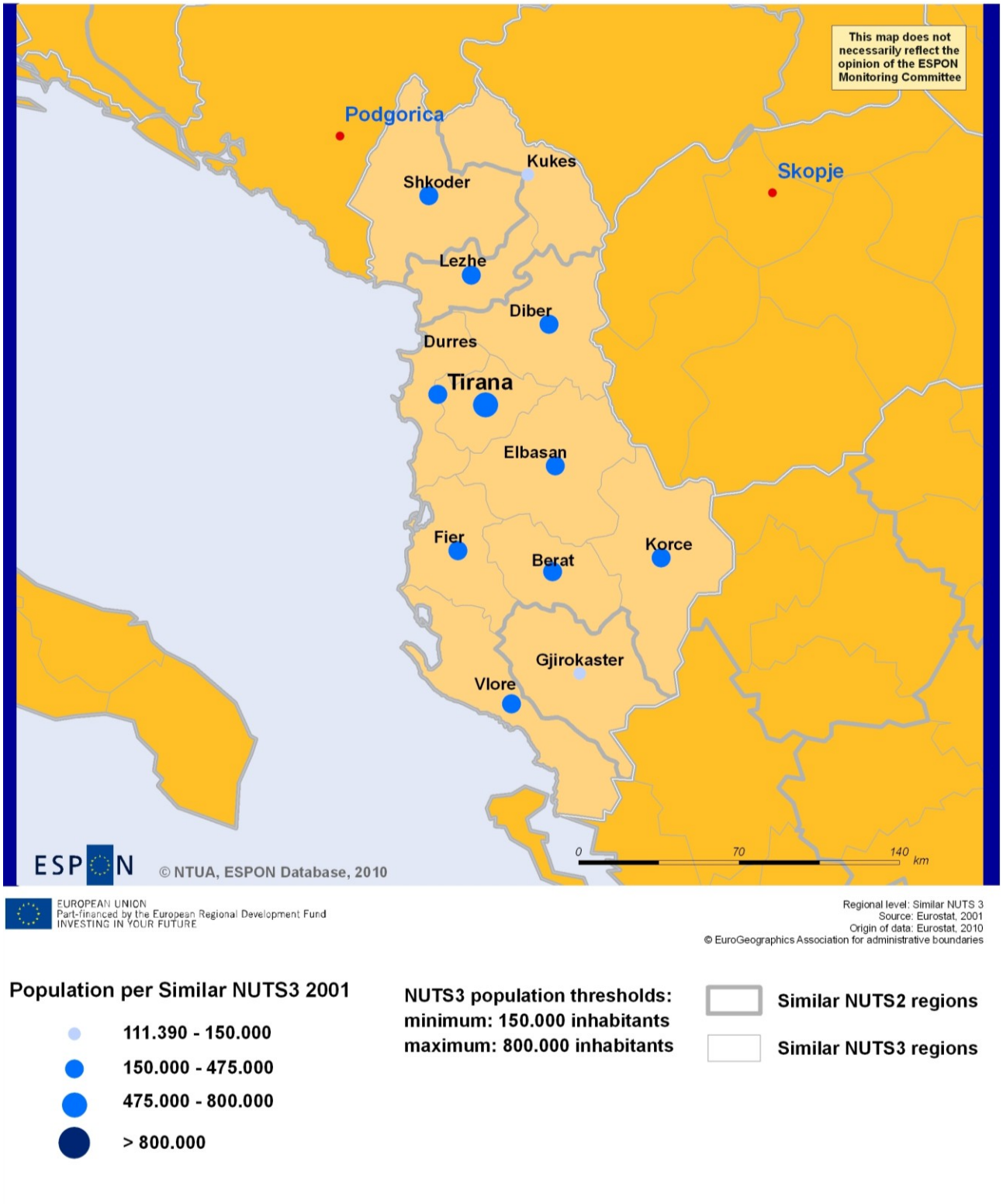
Therefore, ***Albania's prefectures could be assimilated to NUTS3*** – Table AL.1 and Map AL.1.

Table AL.1: Population per prefecture 2001

	Prefecture	Popul. 2001
1	Tiranë	597.899
2	Fier	382.544
3	Elbasan	362.736
4	Shkodër	256.473
5	Durrës	245.179
6	Vlorë	192.982
7	Korçë	265.182
8	Berat	193.020
9	Dibër	189.854
10	Gjirokastër	112.831
11	Kukës	111.393
12	Lezhë	159.182
	Total Alb.	3.069.275

There is not an official territorial division which could be assimilated to ***EU NUTS2 division***. Several scenarios are now examined in the context of the EU – Albania collaboration.

Apart from the capital: Tirana, seven other cities of the country have (roughly) more than 200.000 inhabitants: Fier, Elbasan, Shkoder, Durres, Vlore, Korce and Berat. The eight major urban centres are capitals of prefectures; this is in line with the correspondence of the prefectures to "similar NUTS3" units. However, apart from Tirana there are not other urban centres which obviously provide services of higher level than the prefectural level.



Map AL.1: Albania similar NUTS2 and 3 units, Population per similar NUTS3 2001

Existing data at "similar NUTS3" level (2009)

(1) Official statistical data:

Data at the level of ***prefectures ("counties") / similar NUTS 3****:*

- From the *population censuses of 1989 and 2001:*

(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- From the housing census of 2001

The *Labour Force Survey of 2007* refers to the national level.

Moreover, some research about *population projections 2001-2021, gender perspectives, people and work and living conditions and inequality* exist only for national level or the level of regions (north, centre except from Tirana – Durres, South and Tirana – Durres).

See in the Table 1 in Annex 2.

(2) Data on land uses and environment -from CLC, UMZ.

Data Delivered

The data is only at NUTS0 and NUTS3 levels. We used the data of NUTS0 level for the "similar NUTS1" and "similar NUTS2" regions (total of the country).

- Total area NUTS0 (2000-2006) & NUTS3 (2000-2006)
- Land area No data
- Total population NUTS 0 (2000-2006) & NUTS3 (2001 & 2004)
- GDP (Eur, PPS) See our remark in Table 1.3.
- Active population No data
- Unemployment No data
- Pop by sex and age No data
- Population density NUTS0 2000-2006, NUTS3: 2001 and 2004

Bosnia and Herzegovina

Spatial units' levels:

The ***total population of the country*** amounted up to ***3.843.000 inhab.*** in ***2008*** (Eurostat 2010).

Bosnia and Herzegovina is divided into three entities*: Federation of Bosnia and Herzegovina (FBiH), Republic of Srpska (RS), and Brčko District*, which was established in 2000 out of land from both entities[2].

FBiH is divided in **10 cantons** –*Table BH1. and Map BH.1*- and 79 municipalities; Republic of Srpska has 62 municipalities; City of Brčko is a separate administrative unit - District.

It is difficult to associate the Bosnia and Herzegovina administrative units with corresponding NUTS levels, because the magnitudes of the population of the units belonging to each administrative level are dissimilar.

An additional difficulty relies on the fact that for the RS there is no census or official estimation after 2001; consequently, the estimations occurred by several sources differ significantly among each other. For FBiH there is a very recent (2007) official estimation of the population (from the FBiH's Federal Office of Statistics) that we use in the following.

FBiH (population 2007: 2.328.000), RS (population 2007 estimate: 1.439.700) and Brčko (population 2007 estimate: 68.860) could be assimilated to NUTS1 and / or NUTS2.

EU regulation population criterion for NUTS 3: 150.000-800.000 inhabitants; 6 FBiH's cantons have 227.000-496.000 inh, while 4, have 34.000 - 82.000 inhabitants in 2007. Obviously, according to this criterion, the 4 smaller cantons could be difficultly assimilated to NUTS3 units - See in Table BH.1.

The administrative power / capacity of the cantons is considerable: they have their own cantonal government, which is under the law of the Federation as a whole.

The 10 cantons of FBiH could be assimilated to NUTS3.

Specifically for the ***"similar NUTS2 level"***:

According to the respective population criterion: 800 000 - 3 million inhabitants, both FBiH and RS could be assimilated to NUTS2. Regarding also the administrative power / capacity criterion, these units could be assimilated to NUTS2. However, evidently, in the case of Bosnia and Herzegovina, the political criterion would be taken primarily into account, therefore even Brčko District could be assimilated to NUTS2.

Regarding the ***national urban system***, apart from Sarajevo, the capital (with more than 300.000 inhabitants), which has a really primary role, there are six important regional centres: Banja Luka (in RS: 250.000 inh.) Mostar (FBiH), Tuzla (FBiH), Zenic (FBiH), Bijeljina (RS) and Prijedor (RS); each of these last has 110.000 – 140.000 inhabitants. There are also a number of secondary regional centres (two of them, situated in FBiH, have more than 50.000 inhabitants).

[2] It officially belongs to both, but is governed by neither, and functions under a decentralized system of local government.

Map BH.1: Bosnia and Herzegovina similar NUTS2 and 3 units, Population per similar NUTS3 2001

Table BH.1: Official estimate of the population of FBiH cantons ("similar NUTS3") 2007

	Surface area km2	Population, 2007 [1)]	Population density per km2 2007
Federacija Bosne i Hercegovine	**26.110,5**	**2.328.359**	**89,2**
Unsko-sanski kanton	4.125,0	287.878	69,8
Kanton Posavski	324,6	41.187	126,9
Tuzlanski kanton	2.649,0	496.830	187,6
Zeničko-dobojski kanton	3.343,3	401.796	120,2
Bosanskopodrinjski kanton	504,6	33.662	66,7
Srednjobosanski kanton	3.189,0	256.339	80,4
Hercegovačko-neretvanski kanton	4.401,0	227.473	51,7
Zapadno-hercegovački kanton	1.362,2	82.095	60,3
Kanton Sarajevo	1.276,9	419.030	328,2
Kanton 10	4.934,9	82.069	16,6

The major urban centres are capitals of cantons; this is in line with the correspondence of the cantons to "similar NUTS3" units.

Existing data at "similar NUTS3" level

(1) Official statistical data:

Data at the level of 3 entities: Federation of Bosnia and Herzegovina (FBiH), Republic of Srpska (RS) and Brsko District.

- From the population census of 1991:

(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- From the population official estimate 2008 (for the FBiH): population per sex age etc, active population etc – see above.

- From the Labour Force Survey, carried out in 2007: total active population and its sex distribution, number of employed and unemployed persons, employment per primary, secondary and tertiary sector.

Data on the GDP exist for the FBiH and RS– at entity level.

Data at the level of cantons:

Population 2008 from the population official estimate 2008 –only for the FBiH.

(2) Data on land uses and environment -from CLC, UMZ.

Data Delivered

- Total area No Data
- Land area No Data
- Total population NUTS 0 (2000-2006), similar NUTS1,2,3 (2007)
- GDP (Euros, PPS) See our remark in Table 1.3.
- Active population No data
- Unemployment No data
- Pop by sex and age No data
- Population Density "similar NUTS0" (2007), "similar NUTS3" (2007)

We used as:

Similar NUTS1: the entire country Bosnia and Herzegovina

Similar NUTS2: the Federation of Bosnia and Herzegovina (FBiH), the Republic of Srpska (RS) and the Brčko District

Similar NUTS3: the 10 cantons of FBiH, the Republic of Srpska, and the Brčko District

Croatia

Spatial units' levels

The ***total population of the country*** amounted up to ***4.436.000 inhab.*** in ***2008*** (Eurostat 2010)***.***

- Croatia has already adopted the EU NUTS (1, 2, 3) classification as follows:

NUTS 1: Country (Hrvatska), NUTS 2: Regija (3), NUTS 3: Counties / Jupanija (21). See in Map CR.1.

Only 11 counties had a population ranging between 150.000 and 800.000 inh, in 2001, which are the EU regulation limits for NUTS 3. The 10 remaining counties had a lower population: 54.000-142.000 inhabitants (in 2001).

Croatia includes 3 regions at **NUTS2 level**. It has not established specific administrative structures at NUTS2 level for regional development implementation.

Regarding the **national urban system**, apart from Zagreb, the capital, there are several primary and secondary regional centres; some of these last are located in the touristic coastal zone and / or they are "cities-gates".

In general terms, the capacity of the Croatian urban centres is in line with the divisions in NUTS2 and NUTS3 units.

Existing data at NUTS3 level (2009)

(1) Official statistical data:

*Data at **NUTS3 level**:*

- From the population censuses of 1991 and 2001:

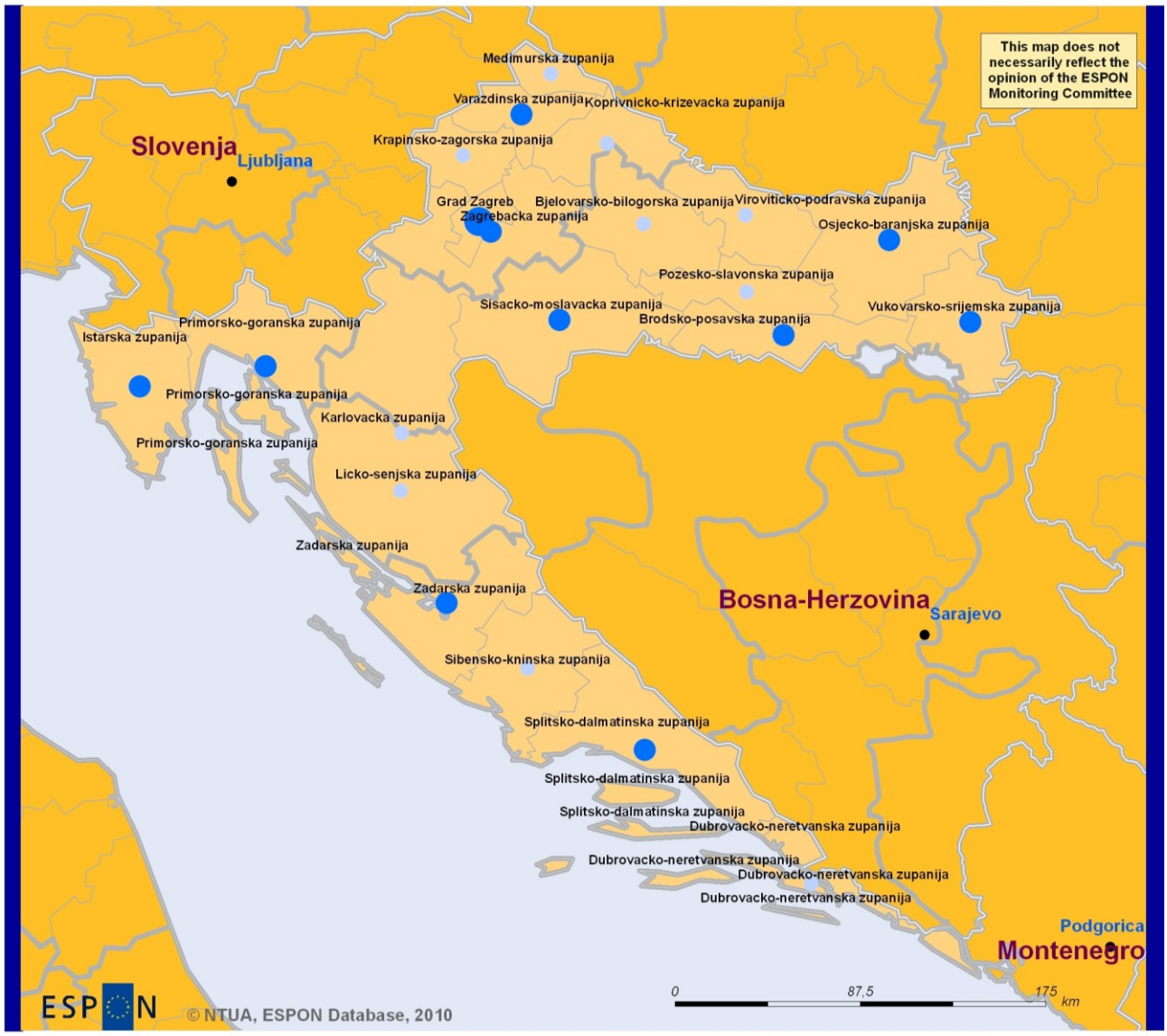
(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- From the population, *households and dwellings* census 2001 (31st March 2001).

Data at National level: Labour force survey -First Quarter of 2008.

(2) Data on land uses and environment -from CLC, UMZ.



Map CR.1: Croatia NUTS2 and 3 units, Population per NUTS3 2001

Table CR.1: Croatian counties (NUTS3) population in 2001

Code	**County of:**	Pop. 2001
hr035	Split-Dalmatia	463.676
hr031	Primorje-Gorskikotar	305.505
hr025	Osijek-Baranja	330.506
hr033	Zadar	162.045
hr024	SlavonskiBrod-Posavina	176.765
hr028	Sisak-Moslavina	185.387
hr034	Šibenik-Knin	112.891
hr027	Karlovac	141.787
hr037	Dubrovnik-Neretva	122.870
hr021	Bjelovar-Bilogora	133.084
hr014	Varaždin	184.769
hr026	Vukovar-Sirmium	204.768
hr023	Požega-Slavonia	85.831
hr015	Koprivnica-Križevci	124.467
hr022	Virovitica-Podravina	93.389
hr016	Međimurje	118.426
hr032	Lika-Senj	53.677
hr013	Krapina-Zagorje	142.432
hr036	Istria	206.344
hro11	Grad Zagreb (City of Za-	779.145
hr012	Zagreb zupan.	309.696
Hr	REP. OF CROATIA	4.127.764

Data Delivered

Data provided mainly by Eurostat

- Total Area data for 2000-2006 and NUTS 0,1,2,3
- Land Area data for 2000-2006 and NUTS 0,1,2,3
- Total population data for 2001-2006 and NUTS 0,1,2,3
- GDP (eur, pps) data for 2000-2006 and NUTS 0,1,2,3
- Active population only for NUTS 0 (2002-2007) and NUTS 0,1,2 (2007)
- Unemployment only for 2007 and NUTS0,1,2
- Pop by sex and age data for 2005 and NUTS 0,1,2, NUTS3 (2001)
- Population Density data for 2001-2006 and NUTS 0,1,2,3 and NUTS0 for 2007

We assimilated NUTS1 level to NUTS0 level.

FYROM

Spatial units' levels

The ***total population of the country*** amounted up to ***2.045.000 inhab.*** in ***2008*** (Eurostat 2010).

FYROM has already adopted the EU classification of spatial units in NUTS; by level:

NUTS 1 and NUTS 2: Country, NUTS 3: Eight (8) Statisticki Regioni / Statistical Regions – See in the ***Map FY.1.***

Table FY.1 Population 2002 of the FYROM regions / NUTS3

Code	*Regions / NUTS3*	*Pop. 2002*
mk008	Skopje	571.040
mk002	Eastern	203.213
mk007	Northeastern	173.814
mk005	Pelagonia	221.019
mk006	Polog	304.125
mk004	Southeastern	171.416
mk001	Vardar	133.248

In August 2004, FYROM was reorganised into 85 municipalities (10 of which comprise Greater Skopje) which could be assimilated to LAU (1) level.

This is reduced from the previous 123 municipalities established in September, 1996. Prior to this, local government was organised into 34 administrative districts (source: Wikipedia).

Seven (7) from the 8 Statistical Regions had a population ranging between 150.000 and 800.000 inh, in 2002 –which are the EU regulation limits for NUTS 3. The one remaining had a lower population: 133.000-inhabitants in 2002 (Table FY.1).

Existing data at NUTS3 level (2009)

(1) Official statistical data:

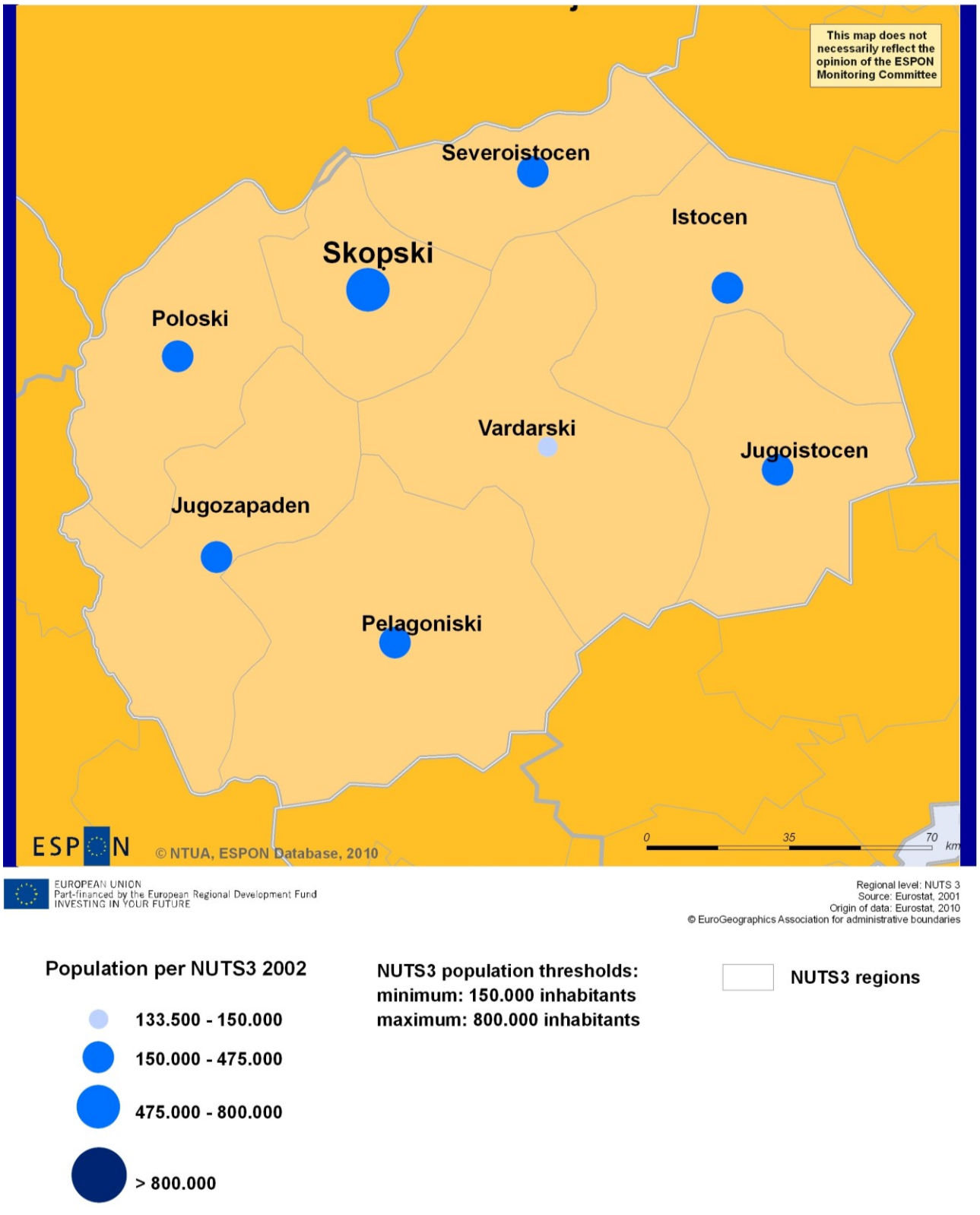
Data at the level of "Statistical Regions" / NUTS 3 (by aggregation of municipalities' data):

- From the population censuses of 1991 and 2002:

(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- From the population, *households and dwellings* census 2002



Map FY.1: FYROM NUTS2 and 3 units, Population per NUTS3 2001

Data at country level (only): from the GDP annual estimations of 2004-2006.

Specific surveys: labour force survey etc.

(2) Data on land uses and environment -from CLC, UMZ.

Data Delivered

Data provided mainly by Eurostat

- Total area data for 2000-2006 and NUTS 0,1,2
- Land area data for 2000-2006 and NUTS 0,1,2,3
- Total population data for 2000-2007 and NUTS 0,1,2,3
- GDP (Euros, PPS) data for NUTS 0 (2000-2006) and NUTS 1,2,3 (2004-2006)
- Active population No data
- Unemployment No data
- Pop by sex and age data for 2005 and NUTS 0,1,2
- Population Density data for 2001-2006 and NUTS 0,1,2,3

See in more detail in the attached "metadata" Excel table.

We assimilated NUTS1 and NUTS2 levels to NUTS0 level.

Serbia

Spatial units' levels

The ***total population of the country*** amounted up to ***7.366.000 inhabitants*** in ***2008*** (Eurostat 2010).

The "similar NUTS3" level

Serbia is divided into two parts: the ***Central Serbia*** and the autonomous province of ***Vojvodina*** and further into **24 districts** (excluding Kosovo) plus the **City of Belgrade**.

The districts and the City of Belgrade are further divided into 157 municipalities – See in **Map SE.1** (the "similar NUTS2" divisions in the Map do not correspond to the recent official division of the Serbia territory in NUTS2 units –see in next)

The territorial organization of the Republic of Serbia is regulated by the Law on Territorial Organization, adopted in the Assembly of Serbia on 29.12.2007. Under the Law, the units of the territorial organization are: municipalities, cities and autonomous provinces.

Districts (okruzi) are regional centres of state authority, but have no assemblies of their own; they present purely administrative divisions, and host various state institutions such as funds, office branches and courts. Districts are not defined by the Law on Territorial Organisation, but are organised under the Government's Enactment of 29 January 1992.

Therefore:
- Districts could be reliably assimilated to NUTS3; 21 from the 25 districts had a population ranging between 150.000 and 800.000 inhabitants, in 2002 (EU regulation limits for NUTS 3). The four remaining had 102.000-147.000 inhab. in 2002.

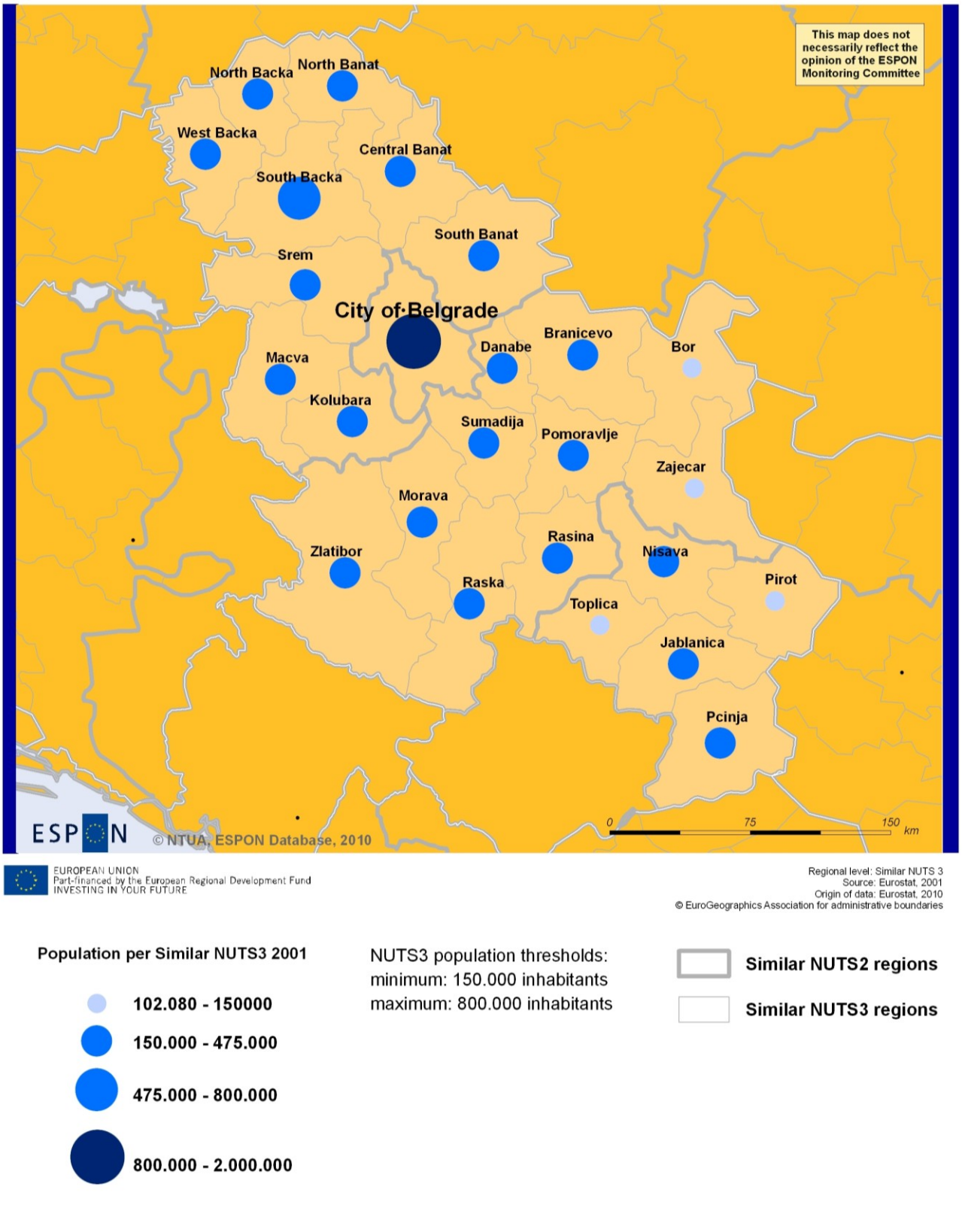
The "similar NUTS2" level

According to the recent available documentation (to be used with caution)

(Wikipedia http://en.wikipedia.org/wiki/Statistical_regions_of_Serbia as of 10.1.2011)

Serbia is divided into **five statistical regions in accordance to NUTS 2**, which are in turn grouped **into two higher NUTS 1 statistical units (North and South**).

In 2009, the Serbian parliament adopted the Law on Equal Territorial Development that formed seven statistical regions on the territory of Serbia. The Law was amended on 7 April 2010, so that the number of regions was reduced to five –see in **Figure SE.2.** The Eastern Serbia region was merged with Southern Serbia and Šumadija was merged with Western Serbia.



Map SE.1: Serbia similar NUTS2* and NUTS3 units, Population per similar NUTS3 2001

* Estimation in 2009, before the reform of 2010 –see in next

According to the above, the statistical regions and their NUTS codes are:

RS: Serbia

- RS1: Serbia - North
  - RS11: Belgrade
  - RS12: Vojvodina
- RS2: Serbia - South
  - RS21: Šumadija and Western Serbia
  - RS22: Southern and Eastern Serbia
  - RS23: Kosovo and Metohija

If we take into account the population criterion, the two Serbian provinces (plus, eventually, the City of Belgrade) could be assimilated to NUTS 2.
As we have mentioned, the statistical NUTS1 and NUTS2 regions created by the government in order to meet the NUTS criteria as well as the requirements of the EU regional policy, do not have actually a considerable administrative power; also, they are not self-governed entities. The political criterion prevailed for their creation.

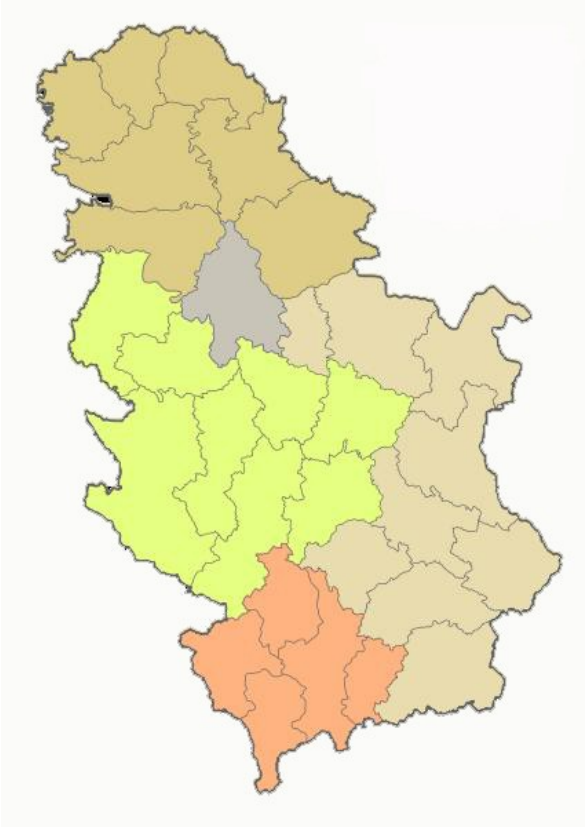


Figure SE.2: Serbia "similar NUTS2" according to the 2010 reform
Source: Wikipedia 2010

The ***national urban system***

Vojvodina has two big cities: Novi Sad and Subotica, two other cities with 50.000-100.000 inhabitants: Zremjamin and Pancevo and 4 cities with 30.000-50.000 inh.

Apart from the capital –Belgrade-, Central Serbia has two other big cities: Nis and

Kragujevac. The rest of the system of cities is balanced with around 10 cities with 50.000- 100.000 inhabitants and a considerable number of cities with 20.000-50.000 inhabitants.

In general terms, the capitals of the districts (corresponding to "similar NUTS3" units) are in most cases enough developed in order to support the development of the respective territorial units. Also, the development of the "new" NUTS2 regions could be supported by respective existing urban centres or networks of urban centres.

Existing data at "similar NUTS3" level (2009)

(1) Official statistical data:

Data at the level of municipalities and districts / similar NUTS 3 (by aggregation of municipalities' data):

-From the population censuses of 1991 and 2002:

(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

There also data on population (distribution per age, sex etc) from a very recent -2006- official estimate.

(2) Data on land uses and environment -from CLC. There are data from CLC2006 but there are not data on UMZ (2009 documentation).

Most of the data concerning censuses of the population and building, specific surveys etc are aggregated and published on the level of *municipalities (LAU1).*

Data Delivered

- Total area NUTS0 and similar to NUTS3 for 2000-2006
- Land area No data
- Total population NUTS 0 (2000-2006) and similar NUTS 3 (2002)
- GDP (Euros, PPS) See our remark in Table 1.3.
- Active population No data
- Unemployment No data
- Pop by sex and age Similar NUTS3 (2002)
- Population Density NUTS0 (2000-2007) and similar NUTS3 (2002)

In 2009, we have assimilated the two parts of Serbia (Central Serbia and Vojvodina) to NUTS1 level as well as to NUTS2 level. This division should be modified according to recent developments on this issue –see previously.

Montenegro

Spatial units' levels

The ***total population of Montenegro*** amounted in ***627.000 inhab.*** in ***2008*** (Eurostat 2010)***,*** therefore ***the total of the country could be assimilated to NUT1, NUTS2 and NUTS3,*** as according to the EU regulation limits for NUTS 3 spatial units their population should range between 150.000 and 800.000 inhabitants and only the municipality of Podgorica had more than 150.000 inhabitants (169.132) in 2003.

We should remind here that Montenegro has been recognised as candidate Country at the end of 2010.

The country is divided in 21 *municipalities* which could be assimilated to *LAU1* level.

Alternatively: only the Municipality of Podgorica (with population over 150.000) could be assimilated to NUTS3.

Existing data at "similar NUTS3" level (2009)

(1) Official statistical data:

Data are available *mainly for the total of the country*

- From the population censuses of 1991 and 2003:

(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- For the education level, the available data exist only in the census of 2003,

(2) Data on land uses and environment -from CLC, UMZ.

Data Delivered

- Total area No data
- Land area No data
- Total population NUTS 0 for 2000-2007
- GDP (Euros, PPS) See our remark in Table 1.3.
- Active population No data
- Unemployment No data
- Pop by sex and age No data
- Population Density NUTS 0 for 2000-2007

Kosovo (Under UN Security Council Resolution 1244)

Spatial units' levels

The ***total population of Kosovo*** amounted up to ***2. 153.000 inhabitants in 2008*** (Eurostat 2010).

Republic of Kosovo is divided in **seven districts** and 30 municipalities.

According to the EU regulation limits for NUTS 3 spatial units their population should range between 150.000 and 800.000 inh; the population of more than the half of the Kosovo districts surpasses 150.000 inhabitants, therefore ***the districts could be (difficultly) assimilated to NUTS 3 units.*** Municipalities could be assimilated to LAU1 level.

Existing data at "similar NUTS3" level (2009)

(1) Official statistical data:

Data at the level of districts / similar NUTS 3:

- From the population census of 1991 (only)

(a) Population: total, distributions: per sex and age group, per education level

(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- No data available on GDP.

Data at national level:

Labour force survey 2002, Labour Market Statistics 2007

(2) There are data from CLC 2006.

Data Delivered

- Total area No data
- Land area No data
- Total population NUTS 0 for 2000-2006
- GDP (Euros, PPS) See our remark in Table 1.3.
- Active population No data
- Unemployment No data
- Pop by sex and age No data
- Population Density NUTS 0 for 2000-2006

Turkey

The ***total population of Turkey*** amounted in ***70.586.000 inhabitants*** in ***2008*** (Eurostat 2010)[3].

Spatial units' levels:

Turkey, which has adopted the EU NUTS/LAU system, has:

- 12 NUTS1 units (Regions, BΦLGELER in Turkish),

- 26 NUTS2 units (Sub-regions, ALT BΦLGELER in Turkish) and

- 81 NUTS3 units (Provinces, ILLER in Turkish) –see in the **Map TU1.**

Seventy eight (78) of these last have a population greater than 50.000 inhabitants (in 2000 and beyond).

Notes on the national urban system

In almost all of the 78 NUTS3 units mentioned above, there is at least one urban region with a population greater than 50.000 inhabitants. The large majority of these urban centres have more than 100.000 inhabitants.

In general terms, the capitals of the Provinces -NUTS3 units- are in almost all the cases enough developed in order to support the development of the respective territorial units. Also, the development of the NUTS2 regions could be supported by respective existing urban centres or networks of urban centres.

We should note that Turkey has already participated in Urban Audit since 2000 with 26 cities[4]. Thus there are enough data for a satisfactory sample of Turkish cities.

Existing data at NUTS3 level (2009)

(1) Official statistical data:

- Data at ***district level***:

From 1990 and 2001 censuses and from the *2007 Population Census which used the Address Based Population Registration System:*

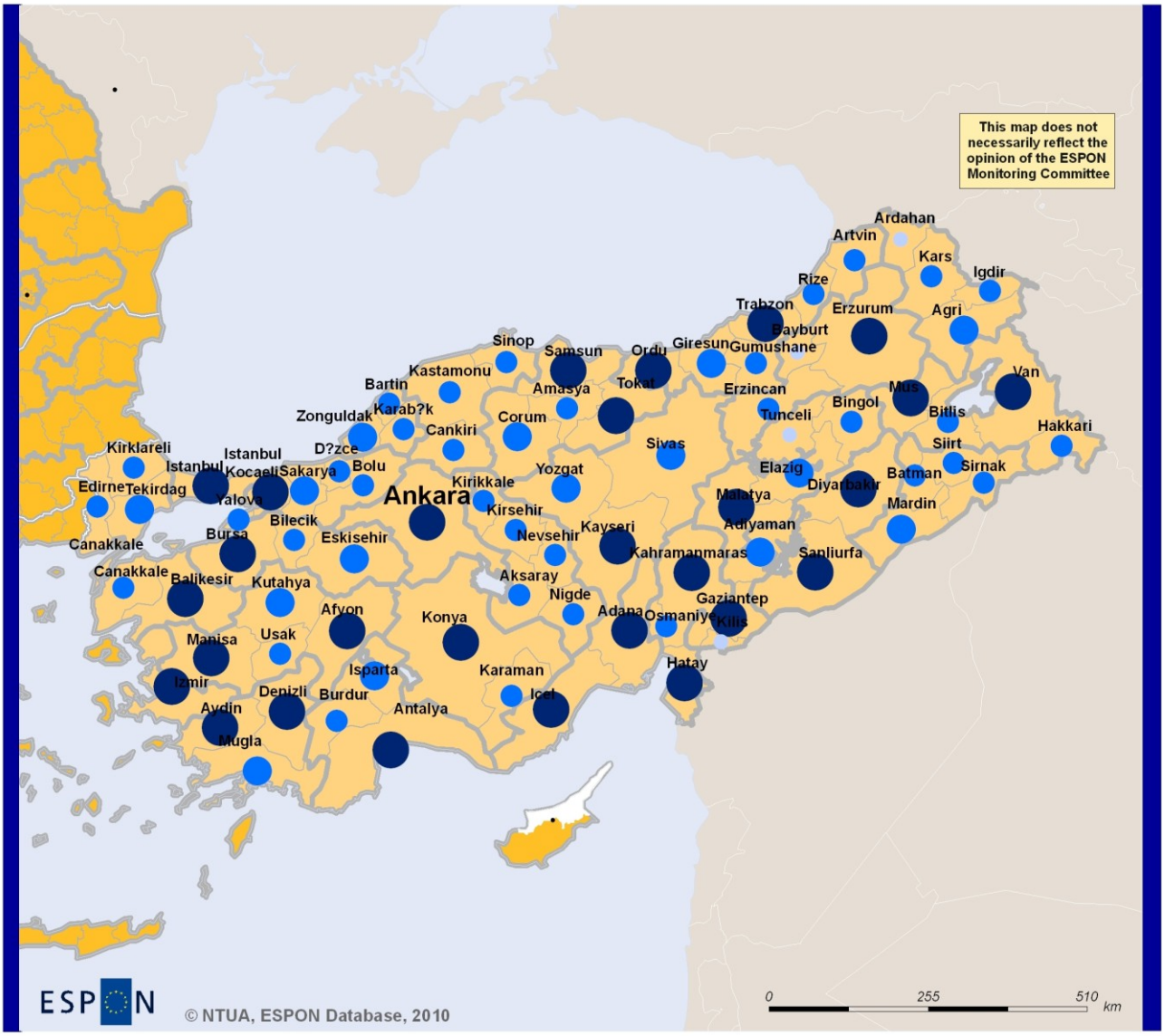
Population by age group and sex, Age dependency ratio, City and village population, Sex ratio, Population density.

- Data from periodic results of households Labour Force Survey for Turkey, Urban and Rural regions (results of 1988 – 1999 terms, results of 2000- October 2007, results of November 2007 and after = Address Based Population Registration System)

(2) There are data from CLC 2006.

[3] According to an official estimate (Address Based Population Registration System) the country population amounted in 70.586.260 inhabitants in December 2007.

[4] Ankara, Adana,, Antaya, Baikesir, Bursa, Denizi, Diyarbakir, Edirne, Erzurum, Gaziantep, Hatay, Istanbul, Izmir, Kars, Kastamonu, Kayseri, Kocaei, Konya, Maatya, Manisa, Nevsehir, Samsun, Siirt, Trabzon, Van and Zongudak.



Map TU.1: Turkey NUTS2 and 3 units, Population per NUTS3 2001

Data Delivered

Data provided mainly by Eurostat

- Total area data for 2000-2006 and NUTS 0,1,2,3
- Land area data for 2000-2006 and NUTS 0,1,2,3
- Total population data for 2000-2006 and NUTS 0,1,2,3
- GDP (eur, pps) data for NUTS0 (2000-2005) and NUTS1,2,3 (2000 & 2001)
- Active population No data
- Unemployment No data
- Pop by sex and age No data
- Population Density data for 2000-2006 and NUTS 0,1,2,3, NUTS0 for 2007

See in more detail in the "metadata" Excel table delivered to the LP.

We used as:

NUTS1: the 12 Regions

NUTS2: the 26 Sub-regions

NUTS3: the 81 Provinces.

3. Conclusions

Compatibility of "similar NUTS" divisions with the EU NUTS classification

Turkey, Croatia and FYROM have already adopted the NUTS classification.

For the rest WB countries, the results of the respective examination per country, using the criteria of the population weight (formal criterion) together with the administrative capacity (informal criterion) -see for the methodology in the Introduction- ensured that the "similar NUTS" divisions used correspond almost fully with the respective divisions for the EU countries; Therefore, the "similar NUTS" could be used for the work on data without considerable problems.

The methodology developed could be further used in the control of the consistency of "similar NUTS" divisions in the Eastern Neighbouring countries (ENC) and the Southern Mediterranean Neighbouring countries (MNC).

Data availability at level NUTS 0

In general, it is very satisfactory for all CC / PCC; most of the data are provided by Eurostat, additional data are provided by the National Statistical Offices (NSO).

Data availability at NUTS2 and 3 levels

- It is in general very satisfactory *for* ***Croatia, FYROM and Turkey****.* Data are fully comparable with the EU ones as these countries have adopted the NUTS classification. Available data from Eurostat cover at NUTS2 level a wide range of topics (see in the Annexes) and at NUTS3 level: demography, economic accounts, tourism, and labour market. Some additional data for specific topics are provided by the NSO of these countries.

- It is less satisfactory for the ***other Western Balkans countries****;* relevant data are provided by the NSO.

In more detail, at "similar NUTS3 level" of these countries:

(a) For demography and labour market, it is good only for some of them while for the rest it is nearly acceptable.

(b) For the rest sections, there are important differences according to the country. Concisely, availability is more satisfactory for Serbia, much less satisfactory for the other PCC.

The closure of the CC / PCC "gap"

Taking into account that necessary reliable data at the appropriate NUTS level or "similar NUTS" level exist for the CC / PCC except Kosovo (under UN Security Council Resolution 1244), all *these countries should remain in the scope of the ESPON Database;* few data for Kosovo could be included at the moment in the Database.

A set of data on WB countries and Turkey is included in the ESPON 2013 Database project's respective Deliveries.

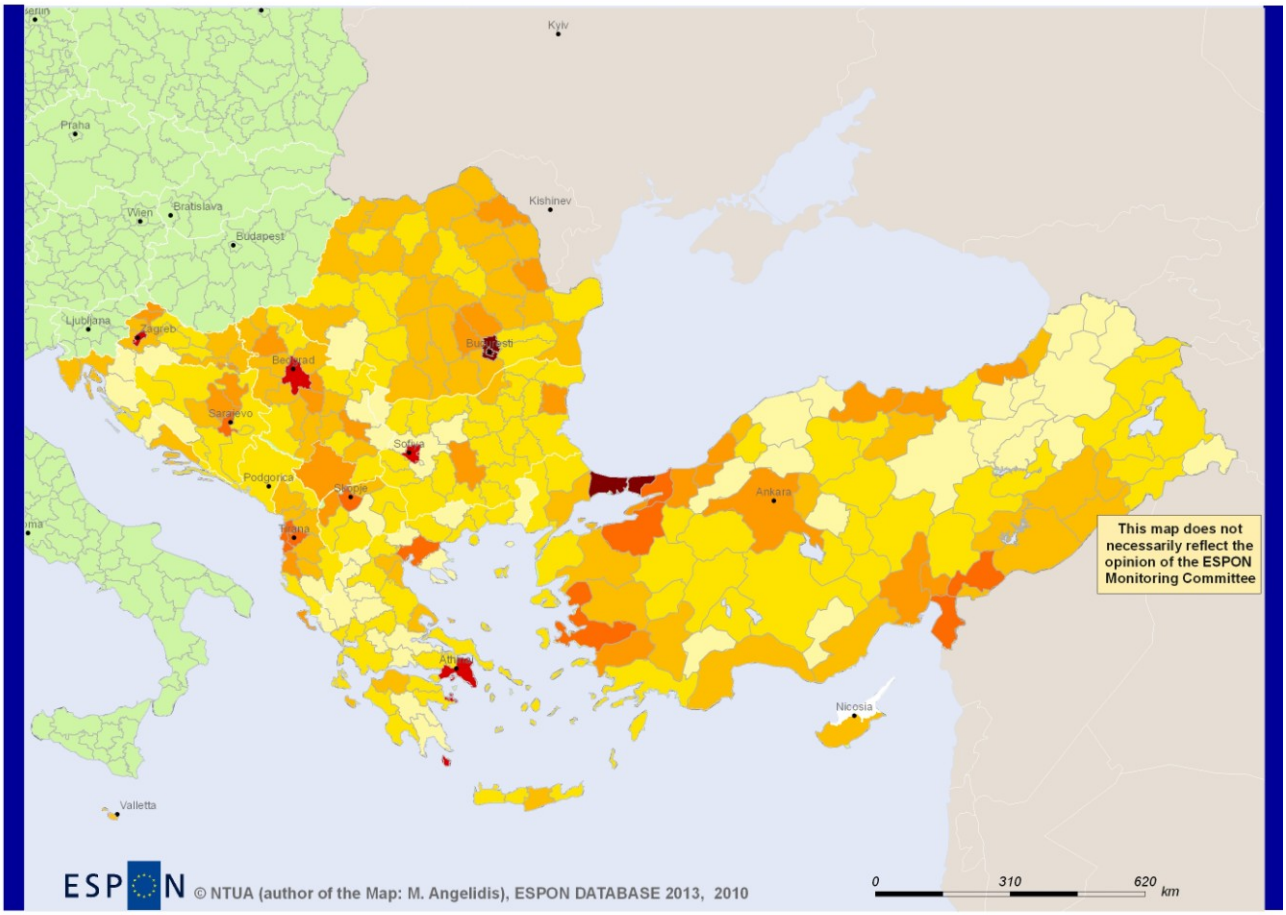
Concluding, the "gap" in the ESPON datasets and Maps corresponding to the WB countries and Turkey (CC / PCC) is closed to a considerable degree –see in the Maps Annex 1 and 2 in Annex 1 –Maps.

This step is very important because it allows the study of the territorial particularities of these countries which should be taken into account in the future Cohesion and Neighbourhood Policies of the EU.

The further development of both formal and informal collaboration of ESPON with Eurostat, DG Regio Official and the Statistical Institutes of the respective countries could ensure a regular bilateral flow of territorial data for these countries.

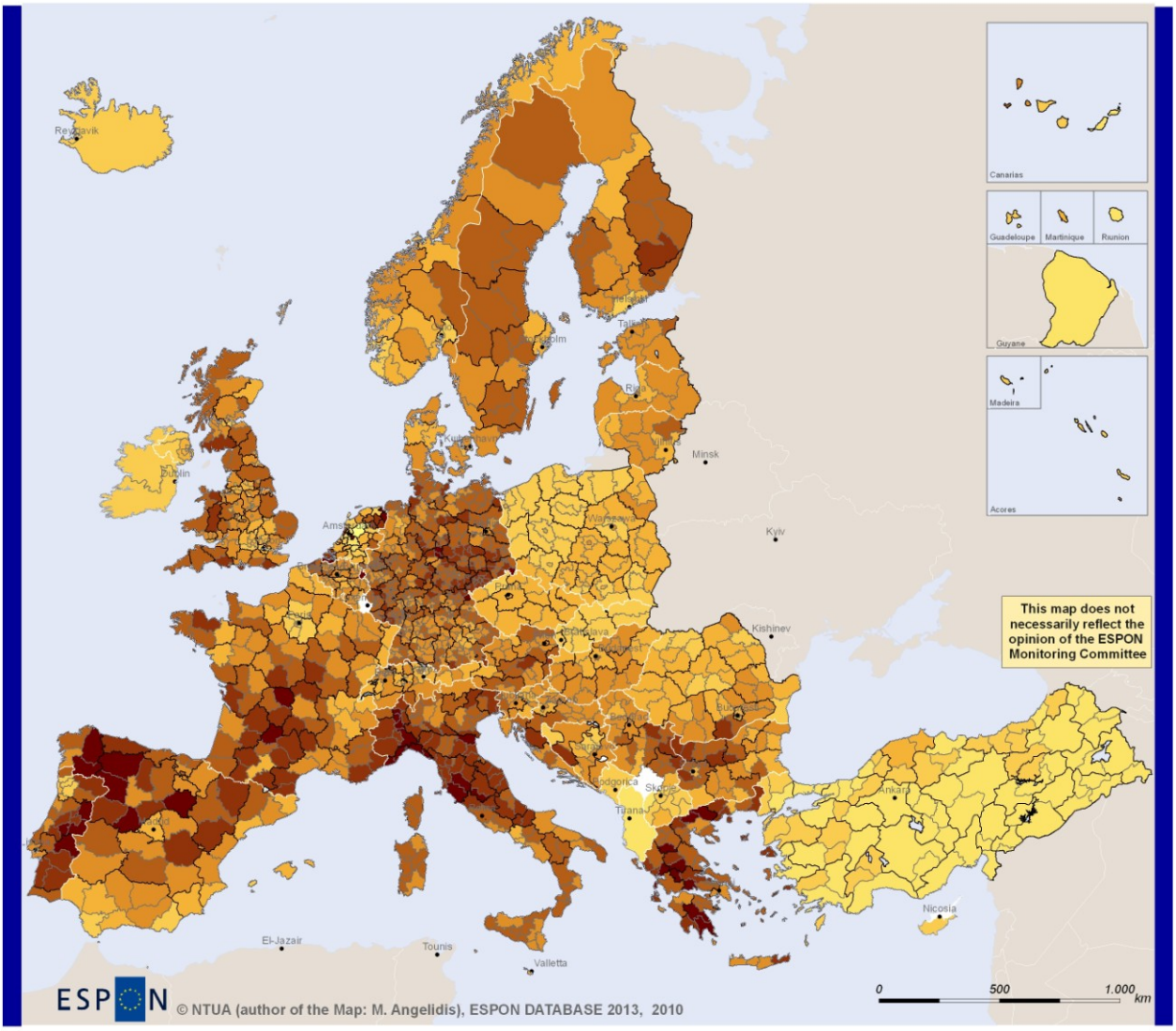
Annex-1 Maps

Map Annex 1: Population density at NUTS3 level in South-eastern Europe: EU and Candidate / Potential Candidate Countries: Western Balkans and Turkey



Map Annex 1: Population density at NUTS3 level in South-eastern Europe: EU and Candidate / Potential Candidate Countries: Western Balkans and Turkey

Map Annex 2: EU and Western Balkans and Turkey* Population 65 years and over Rate % at NUTS3 or similar NUTS3 level 2008



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 3, similar NUTS3
Source: NTUA team for the elaboration of data
Origin of data: Eurostat, National Statist. Organisations of the CC, 2010
© EuroGeographics Association for administrative boundaries

Population 65 years and over
Rate % of the Total population
at NUTS3 or similar NUTS3 level 2008**

- No official data: Kosovo
- 2,8 - 9,3
- 9,5- 13,2
- 13,3 - 15,8
- 15,9 - 18,3
- 18,4 - 21,1
- 21,2 - 24,3
- 24,4 - 30,7

* Candidate or Potential Candidate countries

** Data sources, reference year, NUTS or similar NUTS levels:
Eurostat data for the year 2008 at NUTS3 level (2010)
Except:
- Bosnia & Herzegovina: year 2008, Serbia: year 2007 at "Similar NUTS3" level, estimations of NSO
- Albania and Montenegro: year 2007 at country level, Eurostat data (estimations)
- Germany : Eurostat NUTS3 level, year 2006
- Netherlands: Eurostat NUTS3 level, year 2007
- Ireland and Cyprus: Eurostat NUTS2 level, year 2005

Geometries sources: Eurogeographics administrative boundaries 2006 except: Albania, Serbia, B & H, Montenegro, Kosovo "similar NUTS3": other sources

Map Annex 2: EU and Western Balkans and Turkey Population 65 years and over Rate % at NUTS3 or similar NUTS3 level 2008

Annex 2 – Table 1: Western Balkans and Turkey available (in 2009) territorial data – from all sources

Data at "similar NUTS 3" level in the Western Balkans and Turkey

Data, simple indicators

Data - Indicators / Country	**Albania**	**Bosnia-Herzegov.**	**Croatia**	**FYROM**	**Serbia**	**Montenegro**	**Kosovo** (8)	**Turkey**
	12 Prefectures ("counties") -"similar NUTS3"	**FBiH, RS, and Brsko District (3)**	**Jupanija (21)**	**8 Statisticki Regioni / SR (4a) -"similar NUTS3"**	**Districts (4b) -"similar NUTS3"**	**total of the country**	**seven districts**	**81 ILLER -"similar NUTS3"**
Population census' years 1985 - 2008	1989, 2001	1991	1991, 2001	1991, 2002	1991, 2002 (5)	1991, 2003	1991	1985, 1990, 1997 (6), 2000 (6)
Buildings / dwellings census 1985 - 2008								
Labour force survey 1985 - 2008		2007						
Demographic and social aspects								
Total Population	1989c, 2001c	1991,1995, 2001 -2002 FBiH 2007	1991c,1995, 2001c, 2002 - 2008	1991c, 2002c	1991c, 1995, 2002c, 1998-2005a.e.	1981c,1991c, 2003c	**1981,1991c, 2006, 2007**	1990c, 2000c
Popul. by sex: males, females	2001c	1991c, 2000 - 2003, FBIH 2007	1991,1995,2001 -2008	1991c, 2002c	1991c, 1995, 1998-2007, 2002c	1981c,1991c, 2003c	1991c, 2006, 2007	1990c, 2000c

Population by age group	2001c	1991c, 2000-2003 - FBiH 2007	1991c, 2001c	1991c, 2002c	1991c, 1995, 2002c, 1998-2005a.e.	1991c, 2003c	1991c, 2006, 2007	1990c, 2000c
Population by sex and age group	2001c	1991c	1991c, 2001c	1991c, 2002c	1991c, 1995, 1998-2005, 2002c	1991c, 2003c	1991c	1990c, 2000c
Population per education level	2001c	1991c	1991c, 2001c	1991c, 2002c	1991c, 1995, 1998-2005,	2003-2008, 2003c	1991c	1990c, 2000c
Total number of **house-holds**								
Lone - person households								
Lone - parent households - total / male/ female number								
number of **dwellings**								
Economic aspects, Employment								
Total Active Population	2001c	1991c, 2007 Labour force survey (lfs)	1991c, 2001c	1991c, 2002c	1991c, 1995, 2002c	1991c, 2003c	1991c	1990c, 2000c
Male, Female Active Population	2001c	1991c, 2007 lfs	1991c, 2001c	1991c, 2002c	1991c, 1995, 2002c	1991c, 2003c	1991c	1990c, 2000c
Number of Employed persons	2001c	1991c, 2007 lfs	1991c, 2001c	1991c, 2002c,	1991c, 1995, 1998-2006, 2002c, 2006 (7)	1991c, 2003c, 2004-2007	1991c	1990c, 2000c
Number of unemployed persons	2001c	1991c, 2007 lfs	1991c, 2001c	1991c, 2002c,	1991c, 1995, 1998-2006, 2002c, 2006 (7)	1991c, 2003c, 2004-2007	1991c	2000c, 2004-2007
Employment per primary, secondary, tertiary sector	2001c	1991c, 2007 lfs	1991c, 2001c	1991c, 2002c,	1991c, 1995, 1998-2006, 2002c, 2006 (7)	1991c, 2003c, 2004-2007	1991c	1981-2001

Gross Domestic Product (GDP) (Euros)		FBiH 2005-2007		2004-2006		2000-2004	no data	1990c, 2000c

(1) a.e.=annual estimations
(2) c=census(es)
(3) Federation of Bosnia and Herzegovina (FBiH), Republic of Srpska (RS), and Brsko District
(4a) Existing results are per municipality, we can provide by aggregation results per SR
(4b) Existing results are per municipality, we can provide by aggregation results per Districts
(5) Census not carried out on the territory of Kosovo and Metohia.
(6) Turkey: 1997: Housing census only, 2000: Population census only
(7) Serbia Survey of employed per municipality 2006, we can provide by aggregation results per Districts
(8) Under UN Security Council Resolution 1244

Annex 3 - W. Balkans and Turkey data (2009) from Eurostat / Short presentation

Albania, Bosnia and Herzegovina, FYROM, Serbia, Montenegro, Kosovo[5], Turkey
Data from Eurostat – 2009

(1) (NUTS0, Country level NUTS1)

A) Key indicators on EU policy – Data for all CC – unless a different reference is made:

1) Structural indicators: a) General Economic Background, b) Employment, c) Innovation and Research, d) Economic Reform, e) Social Cohesion, f) Environment (except Kosovo[2])

B) Regional statistics – Data only for Croatia, FYROM and Turkey – unless a different reference is made:

1) Regional science and technology statistics

R&D expenditure and personnel: a) Total R&D personnel by sectors of performance (employment) and region (except FYROM), b) Total intramural R&D expenditure (GERD) by sectors of performance and region (except FYROM)

Human Resources in Science and Technology (HRST) (NUTS level 0, 1 and 2) (except FYROM):

a) Annual data on HRST and sub-groups, b) Annual data on HRST and sub-groups, employed, by sector of economic activity**,** c) Annual data on HRST and sub-groups by age, d) Annual data on HRST and sub-groups by gender

2) Regional labour market statistics

Regional economically active population - LFS series and LFS adjusted series a) Economically active population by sex and age, at NUTS level 1, (1000), b) Economically active population by sex, age and highest level of education attained, at NUTS level 1 (1000), c) Economic activity rates by sex and age, at NUTS level 1 (%), d)Economically active population by sex and age, at NUTS level 1 (1000)

Regional employment - LFS series: a) Average number of usual weekly hours of work in main job (full-time), at NUTS level 1 (hours), b) Employment by professional status, at NUTS level 1 (1000), c) Employment by full-time/part-time and sex, at NUTS level 1 (1000), d) Employment by sex, age and highest level of education attained, at NUTS level 1 (1000), e) Employment rates by sex and age, at NUTS level 1 (%), f) Employment by sex and age, at NUTS level 1 (1000)

Regional unemployment - LFS adjusted series: a) Unemployment rates by sex and age, at NUTS levels 1, 2 and 3 (%), b) Unemployment by sex and age, at NUTS level 1 (1000), c) Long-term unemployment (12 months and more), at NUTS level 1 (1000; %)

Regional socio-demographic labour force statistics - LFS series: a) Life-long learning - participation of adults aged 25-64 in education and training, at NUTS level 1 (1000), b) Population aged 15 and over by sex and age, at NUTS level 1 (1000), c) Population aged 15 and over by sex, age and highest level of education attained, at NUTS level 1 (1000), d) Number of households by degree of urbanisation of residence, at NUTS level 1 (1000)

[5] Under UN Security Council Resolution 1244.

C) Economy and finance – Data for all CC – unless a different reference is made:
1) Main Economic Indicators (except Kosovo[2]), 2) GDP and main aggregates, 3) Annual National Accounts – breakdowns by branches , 4) Annual National Accounts – breakdowns of final consumption expenditure, 5) Government Statistics (except Montenegro), 6) Exchange Rates and Interest Rates, 7) Monetary and other Financial Statistics
8) Prices (except Montenegro), 9) Balance of payments

D) Population and social conditions – Data for all CC – unless a different reference is made:
1) Population Demography, , 2) Education, 3) Labour Market, 4) Living Conditions (except Montenegro, Kosovo[2])

E) Industry, trade and services – Data for all CC – unless a different reference is made:
1) Short-term business Statistics (except Kosovo), 2) Business demography
(except Croatia, FYROM, Turkey, B n H, Montenegro, Kosovo), 3) Information Society Statistics, 4) Tourism (except Kosovo[2])

F) Agriculture, forestry and fisheries – Data for all CC – unless a different reference is made:
1) Agriculture, 2) Forestry Statistics (except Kosovo[2]), 3) Fisheries (except Montenegro, Serbia, Kosovo)

G) External trade – Data for all CC – unless a different reference is made:
1) External Trade, 2) Trading Partners – Flows, 3) Trading Partners – Balance, 4) Trade by Commodity, 5) Terms of trade (except B n H, Montenegro, Kosovo[2])

H) Transport – Data for all CC

I) Environment and Energy – Data for all CC
1) Climate change and waste, 2) Energy

J) Science and technology – Data for all CC – unless a different reference is made: (except Albania, B n H, Kosovo[2])

(2) NUTS2 level

B) Regional statistics – Data only for Croatia, FYROM and Turkey – unless a different reference is made

1) Regions,
2) Regional agriculture statistics: a) Animal populations (December) (except FYROM), b) Areas harvested, yields, production (except FYROM), c) Production of cows' milk on farms (1000 tons) (except FYROM), d) Land (except Croatia and FYROM)
3) Regional demographic statistics
Population and area: a) Population at 1st Jan. by sex and age, from 1980 to 1990, b) Population at 1st January by sex and age from 1990 onwards, c) Average population by sex and age
Population change: a) Births by age of the mother, b) Deaths by sex and age, c) Infant mortality

4) Regional economic accounts

Gross domestic product indicators - ESA95: a) Gross domestic product (GDP) at current market prices at NUTS level 2, b) Real growth rate of regional GDP at market prices - percentage change on previous year, c) Dispersion of regional GDP (%)

Branch accounts - ESA95: a) Compensation of employees at NUTS level 2**,** b) Gross fixed capital formation

Household accounts: a) Income of households**,** b) Secondary distribution of income account of households**,** c) Allocation of primary income account of households

5) Regional science and technology statistics

Human Resources in Science and Technology (HRST): a) Annual data on HRST and subgroups (except FYROM)

Employment in high technology sectors: a) Annual data on employment in technology and knowledge-intensive sectors at the regional level, by gender (except FYROM)

6) Regional tourism statistics a) Nights spent annual data (except Turkey), b) Arrivals - annual data (except Turkey)

7) Regional labour market statistics

Regional economically active population - LFS series and LFS adjusted series: a) Economically active population by sex and age, (1000), b) Economically active population by sex, age and highest level of education attained, (1000), c) Economic activity rates by sex and age, (%), d)Economically active population by sex and age, (1000)

Regional employment - LFS series: a) Average number of usual weekly hours of work in main job (full-time), at NUTS levels 1 and 2 (hours), b) Employment by professional status, (1000), c) Employment by full-time/part-time and sex, (1000), d) Employment by sex, age and highest level of education attained, (1000), e) Employment rates by sex and age, at NUTS levels 1 and 2 (%), f) Employment by sex and age, (1000), g) Employment and commuting among NUTS level 2 regions (1000)

Regional unemployment - LFS adjusted series: a) Unemployment rates by sex and age, (%), b) Unemployment by sex and age, (1000), c) Long-term unemployment (12 months and more), (1000; %)

Regional socio-demographic labour force statistics - LFS series: a) Life-long learning - participation of adults aged 25-64 in education and training, (1000), b) Population aged 15 and over by sex and age, (1000), c) Population aged 15 and over by sex, age and highest level of education attained, (1000), d) Number of households by degree of urbanisation of residence, (1000)

(3) NUTS3 level

B) Regional statistics – Data only for Croatia, FYROM and Turkey – unless a different reference is made

1) Regional demographic statistics

Population and area: a) Population density, b) Population at 1st January by sex and age from 1990 onwards, c) Annual average population by sex, d) Average population by sex and age, e) Area of the regions, f) Population at 1st January by sex and age, from 1980 to 1990

Population change: a) Births and deaths

2) Regional economic accounts

Gross domestic product indicators - ESA95: a) Dispersion of regional GDP at Nuts level 2 and 3 (%), b) Gross domestic product (GDP) at current market prices

Branch accounts - ESA95: a) Employment (in persons), b) Gross value added at basic prices

3) Regional tourism statistics a) Number of establishments, bedrooms and bedplaces - NUTS 3 - annual data (except Turkey)

4) Regional labour market statistics

Regional economically active population - LFS series and LFS adjusted series: a) Economically active population by sex and age, (1000)

Regional unemployment - LFS adjusted series: a) Unemployment rates by sex and age, 3 (%), b) Unemployment by sex and age, (1000)

Annex 4 – Western Balkans and Turkey data (2009) from Eurostat/ Detailed description

Albania, B and H, FYROM, Serbia, Montenegro, Kosovo[6], Turkey Data from Eurostat - 2009

A) Key indicators on EU policy: Structural indicators

1) General Economic Background: Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – level NUTS 0, 1
2) Employment: Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – level NUTS 0, 1
3) Innovation and Research: Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – level NUTS 0, 1
4) Economic Reform: Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – level NUTS 0, 1
5) Social Cohesion: Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – level NUTS 0, 1
6) Environment: Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia (no data for: Kosovo) – level NUTS 0, 1

B) Regional statistics

1) Regions: Croatia, FYROM, Turkey (no data for: Albania, B n H, Montenegro, Serbia, Kosovo) –NUTS level 2
2) Regional agriculture statistics:
a) Animal populations (December) (only Croatia and Turkey) NUTS level 2
b) Areas harvested, yields, production (only Croatia and Turkey) NUTS level 2
c) Land use (only Turkey) NUTS level 2
d) Production of cows' milk on farms (1000 tons) (only Croatia and Turkey) NUTS level 2
3) Regional demographic statistics
A) *Population and area*
a) Population at 1st January by sex and age, from 1980 to 1990 (only Croatia, FYROM and Turkey) NUTS level 2, 3
b) Population at 1st January by sex and age from 1990 onwards (only Croatia, FYROM and Turkey) NUTS level 2, 3
c) Annual average population by sex (only Croatia, FYROM and Turkey) NUTS level 3
d) Average population by sex and age (only Croatia, FYROM and Turkey) NUTS level 2, 3
e) Area of the regions (only Croatia, FYROM and Turkey) NUTS level 3
f) Population density (only Croatia, FYROM and Turkey) NUTS level 3
B) *Population change*
a) Births and deaths (only Croatia, FYROM and Turkey) level NUTS 3
b) Births by age of the mother (only Croatia, FYROM and Turkey) NUTS level 2
c) Deaths by sex and age (only Croatia, FYROM and Turkey) NUTS level 2
d) Infant mortality (only Croatia, FYROM and Turkey) NUTS level 2
4) Regional economic accounts
A) *Gross domestic product indicators - ESA95*
a) Gross domestic product (GDP) at current market prices at NUTS level 2 (only Croatia, FYROM and Turkey)

[6] Under UN Security Council Resolution 1244

b) Gross domestic product (GDP) at current market prices at NUTS level 3 (only Croatia, FYROM and Turkey)
c) Real growth rate of regional GDP at market prices at NUTS level 2 - percentage change on previous year (only Croatia, FYROM and Turkey)
d) Dispersion of regional GDP at Nuts level 2 and 3 (%) (only Croatia, FYROM and Turkey)
B) *Branch accounts - ESA95*
a) Gross fixed capital formation at NUTS level 2 (only Croatia, FYROM and Turkey)
b) Compensation of employees at NUTS level 2 (only Croatia, FYROM and Turkey)
c) Gross value added at basic prices at NUTS level 3 (only Croatia, FYROM and Turkey)
d) Employment (in persons) at NUTS level 3 (only Croatia, FYROM and Turkey)
C) *Household accounts - ESA95*
a) Allocation of primary income account of households at NUTS level 2 (only Croatia, FYROM and Turkey)
b) Secondary distribution of income account of households at NUTS level 2 (only Croatia, FYROM and Turkey)
c) Income of households at NUTS level 2 (only Croatia, FYROM and Turkey)
5) Regional science and technology statistics
A) *R&D expenditure and personnel*
a) Total intramural R&D expenditure (GERD) by sectors of performance and region (only Croatia and Turkey) NUTS level 1
b) Total R&D personnel by sectors of performance (employment) and region (only Croatia and Turkey) NUTS level 1
B) *Human Resources in Science and Technology (HRST)*
a) Annual data on HRST and sub-groups (NUTS level 0, 1 and 2) (only Croatia and Turkey)
b) Annual data on HRST and sub-groups by gender (NUTS level 0 and 1) (only Croatia and Turkey)
c) Annual data on HRST and sub-groups by age (NUTS level 0 and 1) (only Croatia and Turkey)
d) Annual data on HRST and sub-groups, employed, by sector of economic activity (NUTS level 0 and 1) (only Croatia and Turkey)
C) *Employment in high technology sectors (reg_htec)*
a) Annual data on employment in technology and knowledge-intensive sectors at the regional level, by gender (only Croatia and Turkey) NUTS level 1
6) Regional tourism statistics
a) Arrivals - NUTS 2 - annual data (only Croatia and FYROM)
b) Nights spent - NUTS 2 - annual data (only Croatia and FYROM)
c) Number of establishments, bedrooms and bedplaces - NUTS 3 - annual data (only Croatia and FYROM)
7) Regional labour market statistics
A) *Regional economically active population - LFS series and LFS adjusted series*
a) Economically active population by sex and age, at NUTS levels 1, 2 and 3 (1000) (only Croatia, FYROM and Turkey)
b) Economically active population by sex and age, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
c) Economic activity rates by sex and age, at NUTS levels 1 and 2 (%) (only Croatia, FYROM and Turkey)
d) Economically active population by sex, age and highest level of education attained, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
B) *Regional employment - LFS series*
a) Employment by sex and age, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
b) Employment by professional status, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
c) Employment by full-time/part-time and sex, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)

d) Employment by sex, age and highest level of education attained, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
e) Employment and commuting among NUTS level 2 regions (1000) (only Croatia, FYROM and Turkey)
f) Employment rates by sex and age, at NUTS levels 1 and 2 (%) (only Croatia, FYROM and Turkey)
g) Average number of usual weekly hours of work in main job (full-time), at NUTS levels 1 and 2 (hours) (only Croatia, FYROM and Turkey)
C) *Regional unemployment - LFS adjusted series*
a) Unemployment by sex and age, at NUTS levels 1, 2 and 3 (1000) (only Croatia, FYROM and Turkey)
b) Unemployment rates by sex and age, at NUTS levels 1, 2 and 3 (%)(only Croatia, FYROM and Turkey)
c) Long-term unemployment (12 months and more), at NUTS levels 1 and 2 (1000; %) (only Croatia, FYROM and Turkey)
D) *Regional socio-demographic labour force statistics - LFS series*

a) Number of households by degree of urbanisation of residence, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
b) Population aged 15 and over by sex and age, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
c) Population aged 15 and over by sex, age and highest level of education attained, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)
d) Life-long learning - participation of adults aged 25-64 in education and training, at NUTS levels 1 and 2 (1000) (only Croatia, FYROM and Turkey)

C) Economy and finance
1) Main Economic Indicators Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia – NUTS level 1 (no data for: Kosovo)
2) GDP and main aggregates Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
3) Annual National Accounts – breakdowns by branches Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
4) Annual National Accounts – breakdowns of final consumption expenditure
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
5) Government Statistics
Croatia, FYROM, Turkey, Albania, B n H, Serbia, Kosovo – NUTS level 1 (no data for: Montenegro)
6) Exchange Rates and Interest Rates
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
7) Monetary and other Financial Statistics
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
8) Prices
Croatia, FYROM, Turkey, Albania, B n H, Serbia, Kosovo – NUTS level 1 (no data for: Montenegro)
9) Balance of payments
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1

D) Population and social conditions
1) Population Demography
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
2) Education
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
3) Labour Market
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
4) Living Conditions

Croatia, FYROM, Turkey, Albania, B n H, Serbia – NUTS level 1 (no data for: Montenegro, Kosovo)

E) Industry, trade and services
1) Short-term business Statistics
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia – NUTS level 1 (no data for: Kosovo)
2) Business demography
Albania, Serbia – NUTS level 1 (no data for: Croatia, FYROM, Turkey, B n H, Montenegro, Kosovo)
3) Information Society Statistics
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
4) Tourism
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia – NUTS level 1 (no data for: Kosovo)

F) Agriculture, forestry and fisheries
1) Agriculture
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
2) Forestry Statistics
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia – NUTS level 1 (no data for: Kosovo)
3) Fisheries
Croatia, FYROM, Turkey, Albania, B n H – NUTS level 1 (no data for: Montenegro, Serbia, Kosovo)

G) External trade
1) External Trade
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
2) Trading Partners – Flows
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
3) Trading Partners – Balance
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
4) Trade by Commodity
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1
5) Terms of trade
Croatia, FYROM, Turkey, Albania, Serbia – NUTS level 1(no data for: B n H, Montenegro, Kosovo)

H) Transport
Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1

I) Environment and Energy
1) Climate change and waste
Croatia, Turkey, Albania, Montenegro, Serbia – NUTS level 1(no data for: FYROM, B n H, Kosovo)
2) Energy Croatia, FYROM, Turkey, Albania, B n H, Montenegro, Serbia, Kosovo – NUTS level 1

J) Science and technology Croatia, FYROM, Turkey, Montenegro, Serbia – NUTS level 1(no data for: Albania, B n H, Kosovo)

References - Data sources

World and EU Statistical and other data sources

Eurostat, General and Regional Statistics / Non EU countries / *Candidate and potential candidate countries:* Regional data (for Croatia, FYROM and Turkey), other data mainly at national level.

United Nations (UN) / Statistical division (2008), *Several Tables from the UN Statistical Databases: Population and housing censuses: census dates, Population of capital cities and cities of 100,000 and more inhabitants etc.*

Eurostat publications, EC Regulations etc

EC (2003) Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 *on the establishment of a common classification of territorial units for statistics (NUTS)* (Official Journal L 154, 21/06/2003)

EC, *Regulations (EC) No 1888/2005, No 105/2007 and No 176/2008 amending the above Regulation (EC) No 1059/2003*

Eurostat (2010), *NUTS - Nomenclature of territorial units for statistics*, http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction as of 10.1.11.

Eurostat (2010), *Statistical regions outside the EU* http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/statistical_regions_outside_eu as of 10.1.11

Eurostat / Methodologies and working papers - EC (2008), *Statistical regions for the EFTA countries and the Candidate countries,* Office for Official Publications of the EC, ISSN 1977-0375.

Eurostat / Methodologies and working papers - EC (2008), *European Regional and Urban Statistics Reference Guide*, Office for Official Publications of the EC, ISSN 1977-0375

Eurostat / Pocketbooks (2008, 2009, 2010), *Pocketbook on candidate and potential candidate countries*, Office for Official Publications of the EC.

Eurostat Leaflets, *Several leaflets on candidate and potential candidate countries (2008): economic development, population and social conditions etc*, Office for Official Publications of the EC.

Eurostat / Statistical books (2008, 2009, 2010), *Eurostat regional yearbook 2008*, 2009, 2010.

Official Statistical data sources for the CC

Several online publications on economic development, population and social conditions, dwellings, environment etc – see in detail in Chapter 2: assessment per country.

- Albania: Albania Institute of Statistics: http://www.instat.gov.al .
- Bosnia and Herzegovina: Agency for statistics of Bosnia and Herzegovina: http://www.bhas.ba, Federation of Bosnia and Herzegovina Federal office of

Statistics: http://www.fzs.ba and Republika Srpska Institute of Statistics: http://www.rzs.rs.ba

- Croatia: CROSTAT, Republic of Croatia – Central Bureau of Statistics: http://www.dzs.hr/
- FYROM: Republic of Macedonia State Statistical Office: http://www.stat.gov.mk/
- Serbia, Montenegro and Kosovo (Under UN Security Council Resolution 1244): Serbia Republic Statistical office: http://www.statserb.sr.gov.yu/ and Serbia and Montenegro Statistical Office: http://www.szs.sv.gov.yu/
- Montenegro: Statistical Office of the Republic of Montenegro – MONSTAT: http://www.monstat.cg.yu/EngPrva.htm
- Kosovo (Under UN Security Council Resolution 1244): Statistical office of Kosova: www.ks-gov.net/ESK/
- Turkey :Turkey Statistical office: http://www.tuik.gov.tr

Regional and Turkey Urban Audit statistics: http://www.tuik.gov.tr/BolgeselIstatistik/menuAction.do?dil=en

Other documents, research works, publications and data sources

Ambiente Italia Research Institute (2003): *European Common Indicators Final Project Report: Development, Refinement, Management and Evaluation of European Common Indicators Projec*t (ECI), Milano

ESPON Transnational Project Group (TPG), 2005, *project 1.1.3: Enlargement of the European Union and the wider European Perspective as regards its polycentric spatial structure*, KTH, Stockholm.

EC (2007), *Screening report Croatia: Chapter 22 – Regional policy and coordination of structural instruments.*

EC (2006), *EU Enlargement Strategy and Main Challenges 2006 – 2007* http://ec.europa.eu/enlargement/countries/index_en.htm

EU / Regional Policy (2007), *State of European Cities Report: Adding value to the European Urban Audit.*

Rakić B. – Obradović S. (2010), *Absorption problems of the EU development funds in Serbia* http://teme.junis.ni.ac.rs/teme2-2010/teme%202-2010-07.pdf as of 10.1.11.

Eurogeographics, EEA, EIONET, ETCI/LUSI: *Several documents on the EU spatial information system*

European Environment Agency / EEA (2007), Technical report No 17/2007 *CLC2006 technical guidelines,* Office for Official Publications of the EC, ISSN 1725–2237

INSTAT and Swiss Cooperation Office Albania (2010), *Socio-demographic statistics in Albania: selected topics and future developments*, Tirana, 2010.

Knezevic I. (2010), *Absorption Capacity of Serbia for Use of EU Funds: Practical Lessons from Slovakia*, The Pontis Foundation and the Center for Democracy Foundation, Belgrade, July 2010

Milego R. (2007), *Report: Urban Morphological Zones 2000 version F1v0 Definition and procedural steps*, Universitat Autònoma de Barcelona / EEA.

TPG (2004-2006), *ESTIA – SPOSE Programme: First Interim Synthetic Report (WP2) Polycentric Growth Thematic Study Final Report* (WP2.2), *National Re-*

ports of Albania, FYROM, Serbia and Montenegro etc, UEHRI / Panteion University, Athens.

TPG (2006), *ESPON ECPs Transnational Networking activities 097/2005 Data and Indicators of Western Balkans Final Report**, ECP Greece, Athens (*data at NUTS0, 1 and 2 levels).

Wikipedia (2009) www.wikipedia.org, several data on the CC: statistical territorial division, population per regions, prefectures etc, other data –until 2009.

TECHNICAL REPORT

ESPON

Integrating local data

First investigations in Romania, Bulgaria, Czech Republic and Slovakia

CONTENT

- **Exploring and collecting indicators at LAU 2 level is a task that must overcome three problems : the administrative changes, the chronological homogeneity of the datasets and the semantic interpretation of the indicators.**
- **The cumulated experience when working at LAU 1/2 level shows that populating a database becomes a learning by doing process, blocking the construction of a general algorithm.**

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

26 PAGES

LIST OF AUTHORS

Octavian Groza, UAIC, TIGRIS, Iasi, Romania

Alexandru Rusu, UAIC, TIGRIS, Iasi, Romania

Contact

octaviangroza@yahoo.com

tel. + 40 074 55 71 04
+ 40 232 20 14 87

TABLE OF CONTENT

Introduction

The experience accumulated during the previous ESPON contracts proves the necessity of integrating local statistical data (LAU1/LAU2 scale) in order to support more in depth analyses. Such analyses could focus on transnational thematic studies, intra-urban and urban-rural differentiations or trans-scalar approaches. Collecting and the harmonization of this data represent the mission of the TIGRIS team. After an exploring period (identifying the possible data sources, finding the appropriate structure of the database, getting familiar with the geometries or experimenting the exercise of data collection) we started to effectively collect the indicators and build the sample database for two neighboring countries.

In accordance with the proposals set out in the First Interim Report of ESPON 2013 Database Project, Tigris team had to develop a database for two neighboring countries included in the ESPON space. Dealing with this objective involved overcoming a number of problems, most of them being associated with: the harmonization of the spatial geometries, the chronological harmonization and the linguistic barriers. Also, the gap between our initial goal (to exhaustively fill in a database for two neighboring countries) and the outcome (a sample database populated with indicators available online for the Czech Republic and Slovakia at LAU1/2 scale) is mainly due to the large amount of statistical information available on the NSI web sites, that requires additional time for the processing and the integration in a coherent database. To be more explicit, the spatial information and the attribute data needed at LAU 1/2 scales is available not only on the NSI sites [e.g. the population of Slovak municipalities (LAU2) at 31.12.2008], but also from many other sources of information. Thus, building a coherent, comprehensive, comparable and functional database requires additional time and sometimes different collection methods. As a consequence, the completeness of the database was probably the first item that the TIGRIS team quit when starting the effective work.

1 The data sources - specifications

The main source of spatial information (geometries) owned for the moment by the Tigris team is the GISCO geodatabase. Two files were particularly useful: COMM_CENS_RG 2001 and COMM_CENS_2006. The two shape-files provide a base-map at LAU2 scale (polygons and center-points). As there wasn't any comparable base-map for the LAU 1 level in the mentioned database, we were determined to build up a LAU 1 map by merging the LAU2 units, according to the 2001 geometry and integrate some of the collected indicators. Using these maps was essential to our work in order to properly match the information extracted (the statistical indicators) with the available geographic coding system. However, this LAU 1 working map does not guarantee the accuracy of the resulting spatial objects, or its proper correlation with the recent extracted indicators because of the modifications in the administrative organization occurred after 2001.

A second source of information used in our work consists in the official lists with spatial units (LAU1/2) in each country of the European Union. The list being available on the EUROSTAT website1, it's only a matter of proper downloading in order to get an image of the administrative organization of a large part of the ESPON space. Theoretically, these lists are valid for the LAU 1/2 geometry corresponding to 2007. The quasi-chaotic evolution of this geometry at this minimal spatial scale, especially for certain countries (e.g. Romania) makes the official list proposed by EUROSTAT to be regarded with a certain dose of skepticism. Despite limitations associated with chronological inappropriateness, EUROSTAT nomenclature has been extremely useful in building the database at least for two reasons:

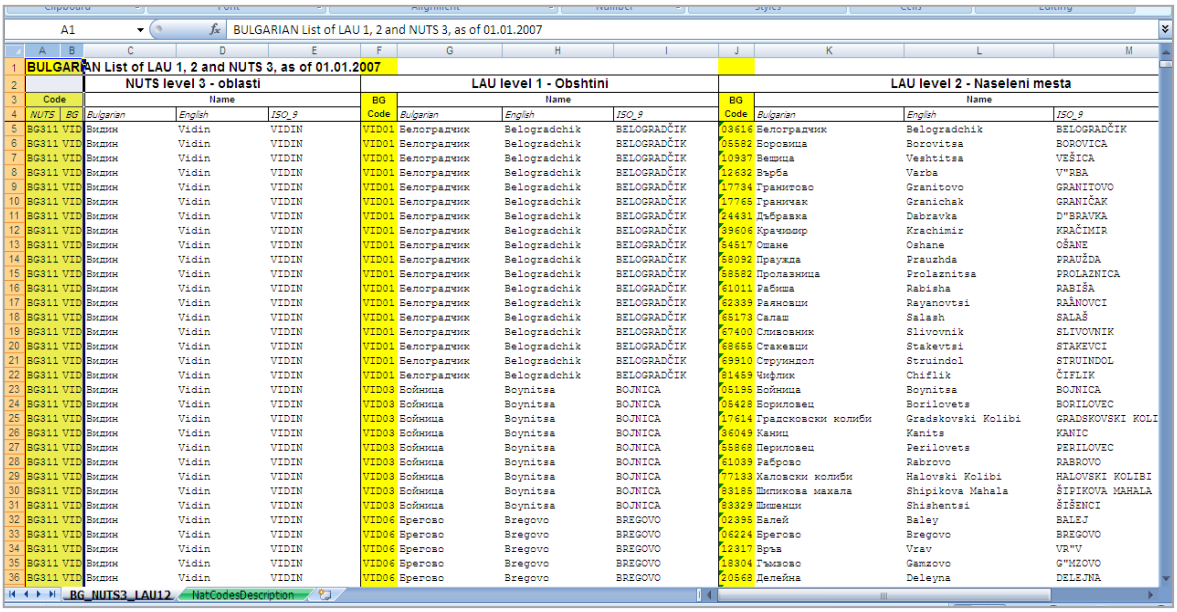
First, this set of lists is one of the few references which allows the appropriate integration of the LAU2 spatial frame in an hierarchically superior administrative levels (LAU2 => LAU1 => NUTS 3 => etc.). For the moment, from the perspective of indentifying the hierarchical spatial units of an LAU, a single file in the database COMM_CENS_2001 in GISCO equals the utility of the EUROSTAT references.

Second, the EUROSTAT classification system includes a useful coding system (national encoding, LAU labels, useful notes and remarks), which somehow permits us to connect the collected information and the indicators with the EUROSTAT references. Some countries, such as Bulgaria, are irrelevant in this respect, the coding system being very sophisticated (the LAU2 national code has its own logic; its construction does not coincide with the coding system used in the GISCO database2, although there are some "filiations" between the two systems).

---

[1] Finding the *bug* in the page permits also the download of information even for Bulgaria; If not, downloading Bulgaria offers Belgian information.

[2] According to the National Code Description included in the EUROSTAT file the "BG [Bulgarian] codes at this level consists of 5 digits. This is not a composite code. The code doesn't contain any information about the belonging of this territorial unit to any upper level of the classification. They are an inheritance from the previous Bulgarian Territorial Classification, created in the `70ies."



BULGARIAN List of LAU 1, 2 and NUTS 3, as of 01.01.2007

	NUTS level 3 - oblasti			LAU level 1 - Obshtini				LAU level 2 - Naseleni mesta			
	Code	Name			BG Code	Name			BG Code	Name	
	NUTS	Bulgarian	English	ISO_9		Bulgarian	English	ISO_9		Bulgarian	English	ISO_9
5	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	03616	Белоградчик	Belogradchik	BELOGRADČIK
6	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	05582	Боровица	Borovitsa	BOROVICA
7	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	10937	Вешица	Veshtitsa	VEŠICA
8	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	12632	Върба	Varba	V"RBA
9	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	17734	Гранитово	Granitovo	GRANITOVO
10	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	17765	Граничак	Granichak	GRANIČAK
11	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	24431	Дъбравка	Dabravka	D"BRAVKA
12	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	39606	Крачимир	Krachimir	KRAČIMIR
13	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	64817	Ошане	Oshane	OŠANE
14	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	68092	Праужда	Prauzhda	PRAUŽDA
15	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	68882	Пролазница	Prolaznitsa	PROLAZNICA
16	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	61011	Рабиша	Rabisha	RABIŠA
17	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	62339	Раяновци	Rayanovtsi	RAÂNOVCI
18	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	65173	Салаш	Salash	SALAŠ
19	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	67400	Сливовник	Slivovnik	SLIVOVNIK
20	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	68655	Стакевци	Stakevtsi	STAKEVCI
21	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	69910	Струиндол	Struindol	STRUINDOL
22	BG311	Видин	Vidin	VIDIN	VID01	Белоградчик	Belogradchik	BELOGRADČIK	61459	Чифлик	Chiflik	ČIFLIK
23	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	05195	Бойница	Boynitsa	BOJNICA
24	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	05428	Бориловец	Borilovets	BORILOVEC
25	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	17614	Градсковски колиби	Gradskovski Kolibi	GRADSKOVSKI KOLI
26	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	36049	Каниц	Kanits	KANIC
27	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	58868	Периловец	Perilovets	PERILOVEC
28	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	61039	Раброво	Rabrovo	RABROVO
29	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	77138	Халовски колиби	Halovski Kolibi	HALOVSKI KOLIBI
30	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	83185	Шипикова махала	Shipikova Mahala	ŠIPIKOVA MAHALA
31	BG311	Видин	Vidin	VIDIN	VID03	Бойница	Boynitsa	BOJNICA	83329	Шишенци	Shishentsi	ŠIŠENCI
32	BG311	Видин	Vidin	VIDIN	VID06	Брегово	Bregovo	BREGOVO	02395	Балей	Baley	BALEJ
33	BG311	Видин	Vidin	VIDIN	VID06	Брегово	Bregovo	BREGOVO	06224	Брегово	Bregovo	BREGOVO
34	BG311	Видин	Vidin	VIDIN	VID06	Брегово	Bregovo	BREGOVO	12317	Връв	Vrav	VR"V
35	BG311	Видин	Vidin	VIDIN	VID06	Брегово	Bregovo	BREGOVO	18304	Гъмзово	Gamzovo	G"MZOVO
36	BG311	Видин	Vidin	VIDIN	VID06	Брегово	Bregovo	BREGOVO	20568	Делейна	Deleyna	DELEJNA

Figure 1: *Bulgarian list of LAU 1/2 spatial units with labels in Bulgarian and English (source : EUROSTAT)*

A third source of information used for the database construction is represented by the NSI websites. Obviously, the information collected from these references is not homogenous/ unequal as presentation system[3] (structuring, organization manner), as time-series included, as semantic relative to the indicators or as spatial dimension.

As a **PRELIMINARY CONCLUSION**: the chronological heterogeneity of our information sources constantly forced our approach to situate itself on some uncertain coordinates, dictated not only by *the lack of accuracy linked with the geometries*, but also by *our direct interference with the inner structures of the files collected*, due to some technical impossibilities related to the spatial variety of the extracted indicators.

[3] The file format used by the NIS sites represents one major drawback during the collection period of indicators. Some NIS (like the Slovakian one) offers free information for LAU 1 spatial units via downloadable software (AXIS), a kind of spreadsheet format which doubles the working time. The Slovak LAU 2 indicators are even more difficult to harvest because they are presented unit by unit, in ***html*** format (probably). The Czech Republic NIS site offers the information in ***.xls*** format, facilitating the collection at LAU2 scale. However, The Czech Republic NIS offers no information at LAU 1 scale.

2 Choosing between 27+4 countries

The selection of the countries included in our analysis was based on several criteria. First, we preferred from the start that the two countries to be located in the eastern part of the ESPON space, starting from the premise that the data collection, due the unequal experience[4] and the numerous readjustments imposed by the transition period could be somewhat more difficult here than in some Western states, which already managed to perfect their statistical systems, thus making it an useful experience and an easy to extrapolate one. In the meanwhile, we had to keep in mind the fact that the main difficulty in the process of extracting statistical indicators (especially in terms of chronological dimension), is linked with the search for an equilibrium between the length of the time series and the number of spatial units involved. That's why we have privileged two medium-size countries, honestly much more suitable for the statistical data collection. In the beginning of our work we have focused on Romania and Bulgaria and the rationale seemed quite logic to us.

First, the Tigris team has some experience in dealing with LAU2 databases for the two countries (e.g. Espace géographique, etc). Moreover, we have already completed a sample database for Romania and Bulgaria using LAU1 and LAU2 indicators, collected in 2007 and 2008 and some of these indicators were already chronologically harmonized. This experience is reinforced by the know-how accumulated during the elaboration of the several versions of the Atlas of Romania (the version available online is basically a LAU2 cartographic tool). All this work already undertaken for the two countries helped us in building a large and quite comprehensive database (several hundreds of indicators only for Romania) for the 2948 or 3175 LAU2 officially designated in this moment. However, this database is relatively old because of the successive administrative "micro-reforms" who multiplied the number of spatial units from 2948 in 2002 to almost 3175 in 2008. Most often, these readjustments in the elementary geometry were produced by the division of LAU2, by administrative redefinition (some rural LAU change status in urban ones) or by the modification of the existing nomenclature.

[4] We had the nice surprise to observe that the Eastern NIS sites are generally comparable with the western ones and sometime extremely innovating in their data layout or in the process of indicator selection.

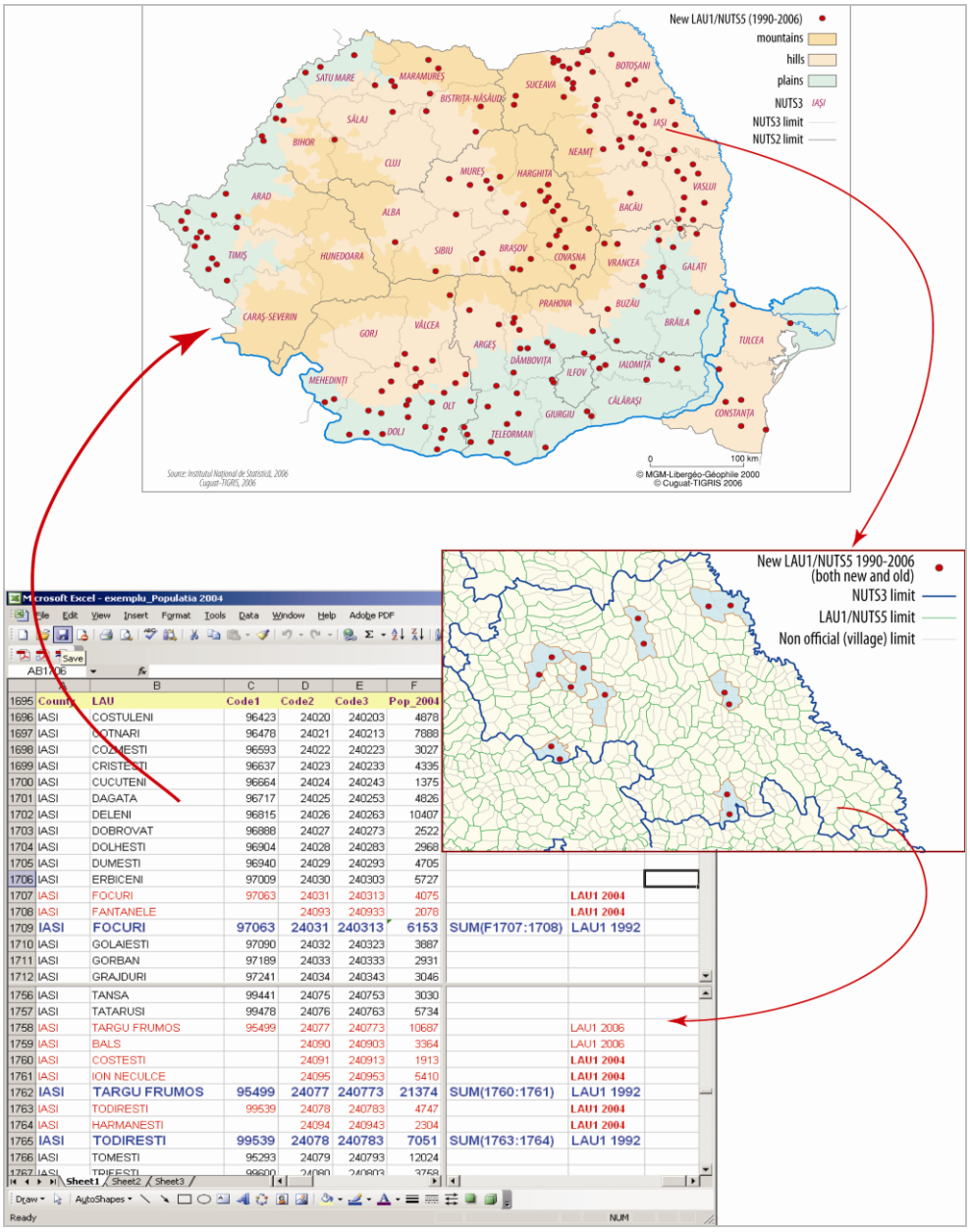


Figure 2: *New LAU2 [5]units in Romania (1990-2006)*

[5] In the titles of the two maps one should read New Lau2/NUTS5

3 Populating the database for Romania and Bulgaria…step by step and inch by inch

At a normal screen resolution, the 4618 Bulgarian LAU 2 spatial units extracted from the GISCO database (COMM_CENS_RG_2001) represent the equivalent of approximately 23 meters of information for only one field in the working file. The 2940 Romanian LAU2 should occupy another approximately 15 meters of information. It might seem somehow anecdotic and irrelevant information but, basically, populating a database means introducing meters of information for every indicator. One could imagine that this process is an automatic one, an easy job for post-modern geographers. Is not quite like that. Populating the database also signifies an endless verification process in order to properly match fields of information extracted from the online sources with the working files to be filled in. This matching issue represented the most time consuming aspect in the working process. However, it was also the simplest intellectual challenge in our approach.

After collecting the data from the GISCO tables and directories we have observed several inadequacies between the list of LAU 1/2 registered in this database and the lists provided by other sources (National Institutes of Statistics, TIGRIS databse, EUROSTAT), both for Romania and Bulgaria. Bulgaria is probably the most interesting challenge in terms of rebuilding the administrative history at minimal spatial scale.

TYPES OF MODIFICATIONS	
observed for the first time	change in the list of composite units
creation	closure
creation by separation (from another populated place)	closure by new administrative-territorial structure
creation by merging	closure by merging
creation by division	closure by division
creation by new administrative-territorial structure	closure by addition
annexation to the country territory	erasion
change by new administrative-territorial structure	closure by loss of territory
change of name	restoration
change of characteristic	restoration by merging
change of administrative centre	restoration by merging
change of administrative territorial belonging	restoration by separation
separation	restoration by division
addition	change of boundaries/structure

Table 1: *Classification of LAU 2 modifications in the administrative geometry (events recorded since 1878)*
Source: NSI Bulgaria, NATIONAL REGISTER OF POPULATED PLACES

Although the Bulgarian Register of Populated Places is extremely generous in terms of information regarding the changes in the administrative geometry of the LAU2, all these references require a systematic approach which is an extremely time-consuming task. For example, 30 units of type 2 LAU were closed by addition after 2001 (and the addition term deserves a definition which was not yet found), another 3 were closed

by merging, 4 villages (LAU2 units) changed their name, one town was restored by separation, one village was restored by merging, 6 villages were created by addition, 6 new units were created by separation and unfortunately this list of modifications is not exhausted. All these territorial metamorphosis have a direct impact on the database that we are supposed to provide.

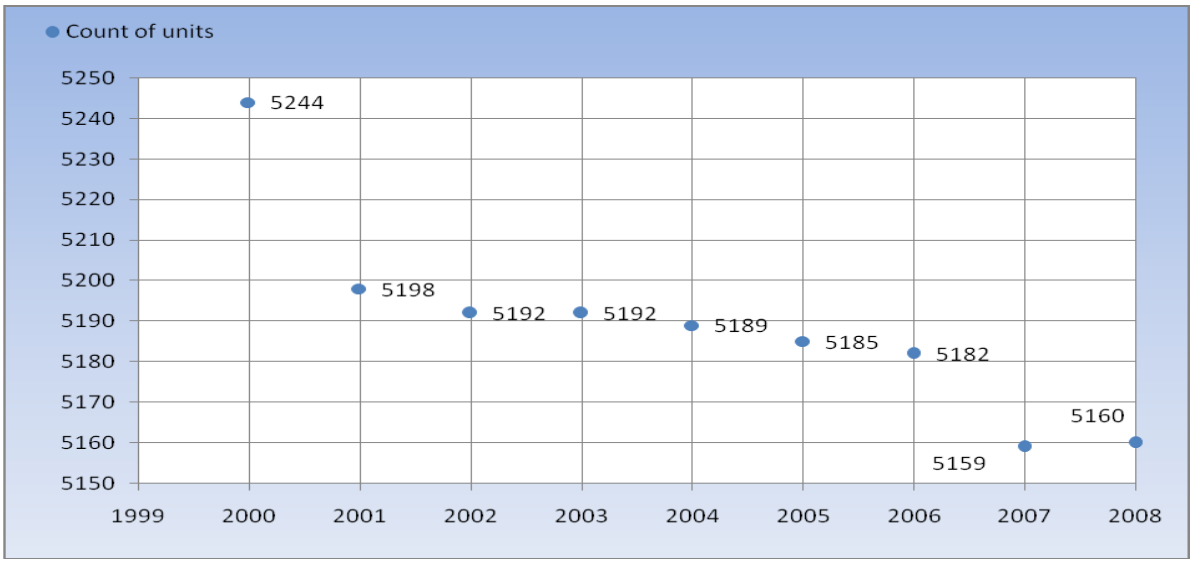


Figure 3: *The evolution of LAU2 in Bulgaria (x axis = time)*

According to this official source, the number of LAU 2 in Bulgaria constantly dropped from more than 5200 spatial units in 2000 to 5160 in 2008. However, a big number of units listed in the National Register of Populated Places was not found either in the GISCO/EUROGEOGRAPHICS reference files (599 LAU2 missing for 2001) or in the official LAU2 list presented by EUROSTAT[6].

Consequently, linking geometry and database tables is impossible for the moment.

In this case, even if we have succeeded to extract one indicator for Bulgaria at this scale of analysis (population for the LAU 2 polygons between 2000 and 2008) in the absence of a proper base-map, the table is unlikely to be useful.

Similar problems have been encountered for Romania. GISCO/EuroGeographics database lists 2940 LAU2 in 2001. Comparing this source with the TIGRIS database or with some official sources (National Institute of Statistics) we found 8 LAU2 missing. If in Bulgaria[7] the trend shows a constant decrease in the number of LAU 2 units, in Romania the situation represents exactly the opposite. TIGRIS had several attempts to rebuild the elementary base-map of Romania but without success due to the fact that new changes in the administrative geometry are occurring almost every month. As a matter of fact, the problem is much more complicated. The decision to create a new Romanian LAU2 (generally by division) is not immediately doubled by a map with the new limits of the new born polygons. Even if we succeed to provide an updated base-map for Romania, we are not quite sure about the accuracy of these polygons.

[6] The two sources offer a different number of spatial units for 2007 (EUROSTAT – 5299 and NSI BULGARIA 5159). Almost 150 spatial units are to be found in list of modifications only for this chronological reference.
[7] According to the Bulgarian National Register of Populated Places.

The following tables synthesize the main steps and problems encountered in the development of the database for the two countries. Despite several attempts, for the moment not every problem is also accompanied by a solution.

STEP	Operation	Source	Done
1	Extracting basemap for Bulgaria (LAU2)	GISCO COMM_CENS_2001_AT	OK
2	Extracting basemap for Romania (LAU2)	GISCO COMM_CENS_2001_AT	OK
3	Merging LAU2 polygons in LAU 1 (only for Bulgaria)	GISCO COMM_CENS_2001_AT	OK
4	Creating basemap with the two countries	GISCO COMM_CENS_2001_AT	OK
5	Extracting indicators from the GISCO database	GISCO COMM_CENS_2001_AT	OK
6	Comparing LAU2 GISCO codes with other coding systems (SIRUTA for Romania and the Bulgarian NSI codes)	GISCO COMM_CENS_2001_AT, TIGRIS database, NSI databases	OK
7	Dealing with the encounterd problems		OK
8	Populating the database with indicators for both countries	GISCO COMM_CENS_2001_AT, TIGRIS database, NSI databases	OK

Table 2: *Operational steps undertaken during the database development process and data sources*

PROBLEMS	Solution
No match between the LAU2 GISCO coding system and the NSI coding system.	Inventing a new coding system.
No LAU1 label in the GISCO database (for Bulgaria) and no match between the LAU1 GISCO coding system and the NSI coding system.	Matching under Excel the labels and the codes
No match between the LAU2 geometry (GISCO) and the indicators extracted from the other sources. (599 LAU2 missing in 2001) No match between the LAU 1 geometry (GISCO) and the indicators extracted from the other sources.	Operation aborted for the moment

Table 2: *Problems encountered in the database development process and solutions developed*

Even if the issues concerning the proper linkage between the base-map and the database should be overcome, it will still be difficult to imagine a solution in order to eliminate the size differences between the LAU2 of the two countries.
The Bulgarian LAU1 has no correspondent in Romania while the Romanian LAU 2 is much bigger than the same units in Bulgaria (Fig. 4). When mapping whatever indicator, this "mass effect" should be considered. We are sure that we will encounter the same problem (linked with the surface difference) at the French-Belgian border.

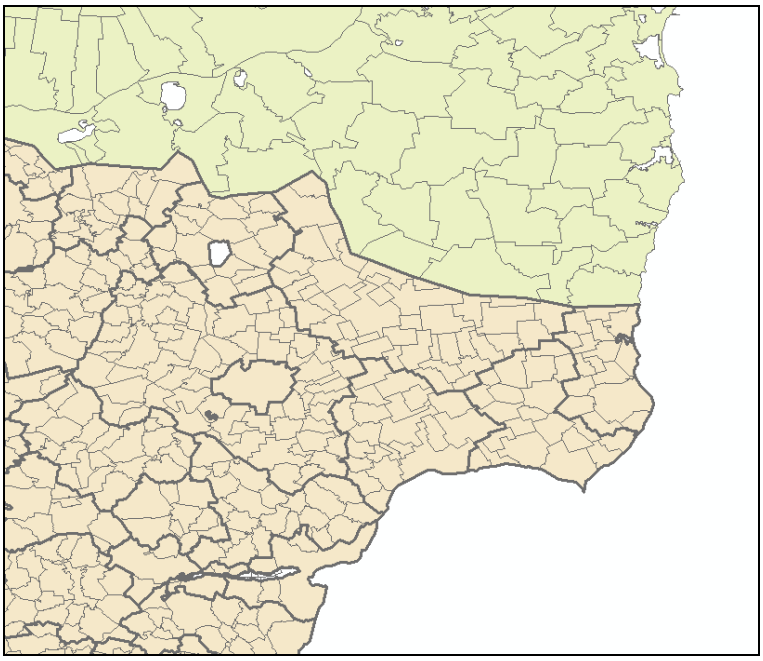


Figure no 4 – LAU2 in Romania and Bulgaria (LAU1)

For all these reasons we have stopped working for the moment at a database for Romania and Bulgaria (technically speaking we are in standby with the history of LAU2 evolution for the both countries), even if we have somehow advanced in this problematic, and as a backup for the technical rapport and for Challenge 4, we have focused on building a database for other two countries (Czech Republic and Slovak Republic).

4 Building a database for the Czech Republic and Slovakia

The choice of the two countries was based on some facilities that have smoothed the collection of information and the matching exercise with the base-map extracted from the GISCO database. First, unlike Romania and Bulgaria, quite a few administrative reforms have altered the administrative geometry of LAU2 and LAU1 during the 2001-2008 period. Such mutations, but not so intense like in the Bulgarian case, are visible in Slovakia. For now, only 8 Slovak LAU2 don't find their correspondent in the GISCO tables which we use to verify the correspondence between the base-map and the database. The collection of the indicators started from the National Institutes of Statistics, in particular the 2001 Census results for the Czech Republic and the Regional database for Slovak Republic.

Despite our intention, we are not able to provide an exhaustive database for the two countries yet. In the case of the Slovak Republic, the information available at LAU 1 exceeds our possibility to collect them just in time. Anyway, a prioritization of the indicators should be considered for a proper extraction, otherwise we might populate the database with interesting but not very useful8 indicators.

Figure 5: *The availability of statistical indicators in the case of the Slovak Republic - Sample view*

(Source: RegDat, The Regional Statistics Database hosted on the Statistical Office of The Slovak Republic website)

[8] As an example, we can download indicators such as the "pension's expenditures in Euro or Slovak currency between 1999 and 2008", for the Slovakian LAU1, but we cannot find the same information (the same indicator) for the Czech Republic. At a smaller scale, for the Slovak LAU2 we may download the earliest recorded mention by historical sources (e.g. Borinka (LAU2) in the District of Malacky (LAU1) was first mentioned in 1273 A.D. An exhaustive collection of the Slovakian indicators should provide even the administrative or economic central places attributes for the Slovak Lau2.

In order to integrate all this information in our data tables we were forced, (especially when collecting the indicators at LAU1 scale for Slovakia) to work with another software (Pc-Axis) allowing the visualization of the chosen variables (Fig. 6 and 7). Just to emphasize the immense data series and the sometimes overwhelming work involved: eight indicators for nine years time-series and 79 spatial units could be regarded as quite a simple case...but not as simple as downloading the agriculture indicators (Fig. 7). On the other hand, the collection of indicators for Czech Republic at LAU1 scale is not simple at all. The site of the Czech Institute of Statistics still uses the term NUTS4 instead of LAU1. Our first researches ignored this aspect. Consequently, we are not able to provide indicators for this type of administrative geometry for this country. Recently, after a routine check of the data sources, we have managed to obtain some LAU1 indicators (some demographic time-series from 1949 to 2007) and these tables will soon be ordered and integrated in the database.

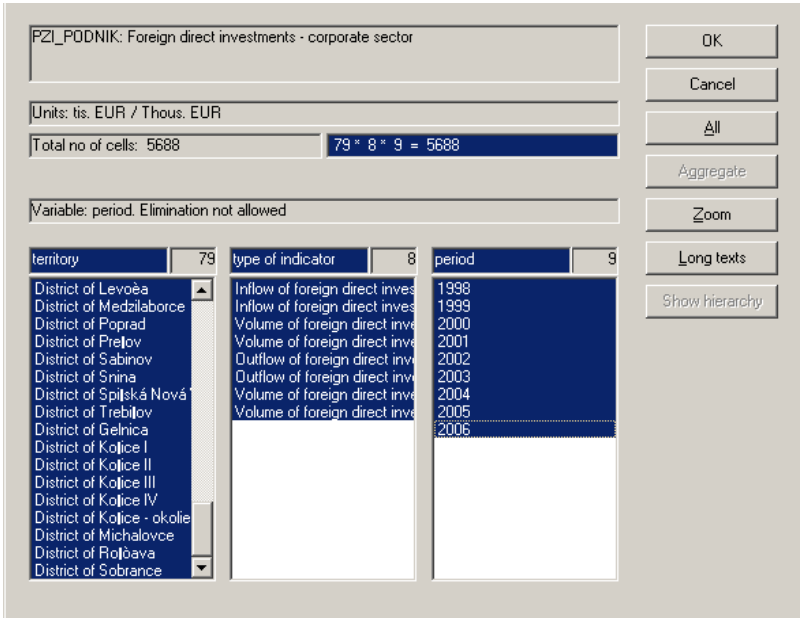


***Figure 6**: Foreign direct investments in Slovakia (LAU1 – 1998-2006, Pc-Axis software view)*

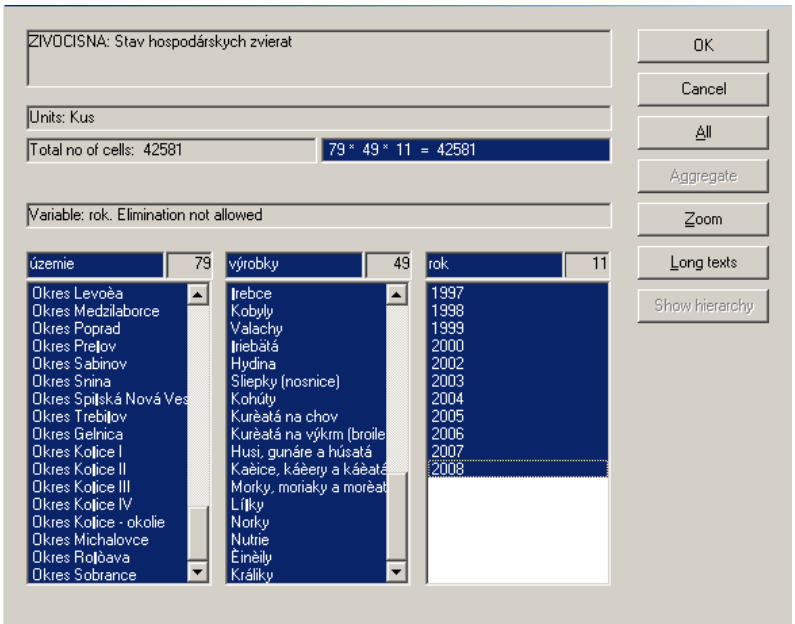
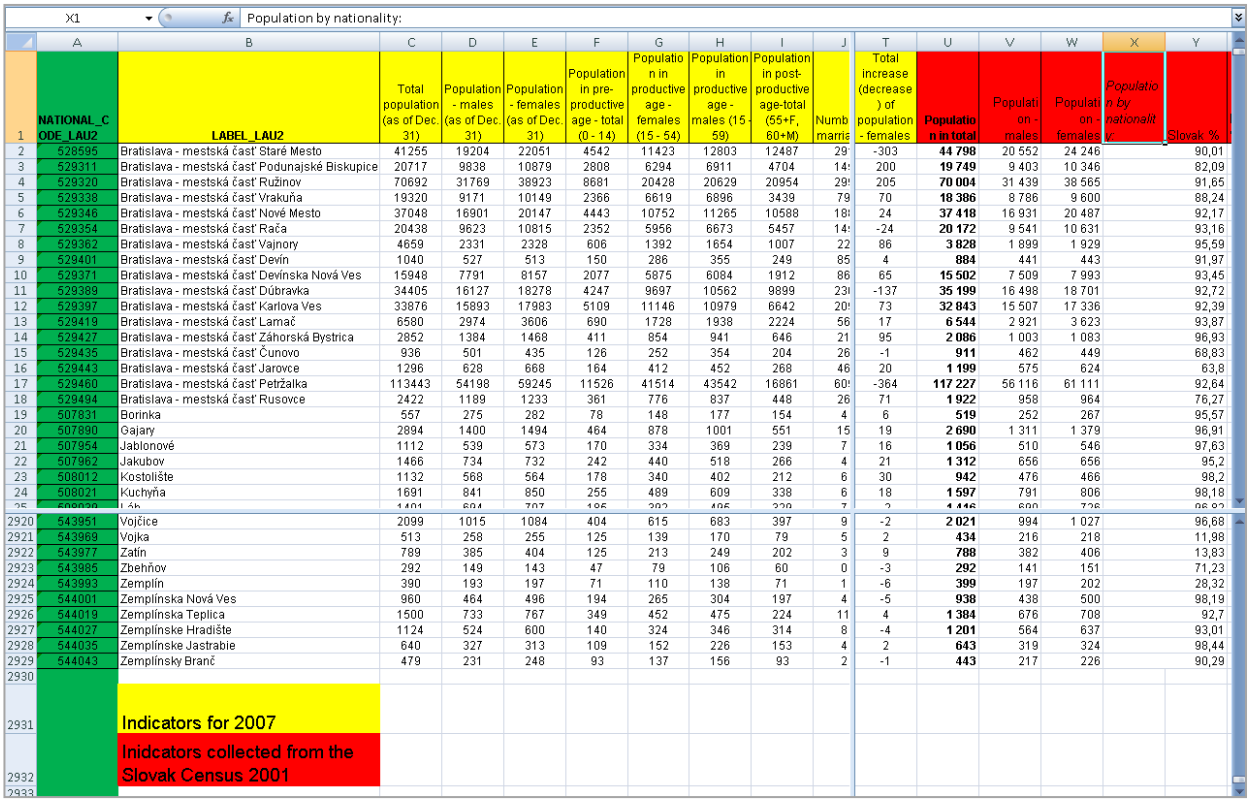


***Figure 7:** Stav hospodárskych zvierat by územie, výrobky and rok*

(**Pure Slovakian… It seems to be a file which presents indicators about the livestock according to the Google translate tool – "Status of livestock by the products and the year"**)

The data collection at LAU2 scale for Slovakia is also uncompleted. We have managed to include in our database 52 indicators, LAU2 by LAU2, after a long copy-paste/import data exercise LAU2 by LAU2 files (2928 multiplied by 2 files copied for each spatial unit). The 52 indicators include different information which we considered relevant at the extraction moment (economic and demographic indicators for 2007 and 2001). Generally, the other variables available for download on the site (Health Services, The Basic Characteristics of the LAU2 or the Environment Indicators) are mainly presented in text format (Boolean opposition of presence/absence). Working on Slovakia LAU2 and LAU1 indicators is a useful exercise, a training routine for the collection of information for Poland and Austria.



	A	B	C	D	E	F	G	H	I	J	T	U	V	W	X	Y
1	NATIONAL_CODE_LAU2	LABEL_LAU2	Total population (as of Dec. 31)	Population - males (as of Dec. 31)	Population - females (as of Dec. 31)	Population in pre-productive age - total (0 - 14)	Population in productive age - females (15 - 54)	Population in productive age - males (15 - 59)	Population in post-productive age-total (55+F, 60+M)	Numb marri	Total increase (decrease) of population - females	Population in total	Populati on - males	Populati on - females	Population by nationalit y:	Slovak %
2	528595	Bratislava - mestská časť Staré Mesto	41255	19204	22051	4542	11423	12803	12487	29	-303	44 798	20 552	24 246		90,01
3	529311	Bratislava - mestská časť Podunajské Biskupice	20717	9838	10879	2808	6294	6911	4704	14	200	19 749	9 403	10 346		82,09
4	529320	Bratislava - mestská časť Ružinov	70692	31769	38923	8681	20428	20629	20954	29	205	70 004	31 439	38 565		91,65
5	529338	Bratislava - mestská časť Vrakuňa	19320	9171	10149	2366	6619	6896	3439	79	70	18 386	8 786	9 600		88,24
6	529346	Bratislava - mestská časť Nové Mesto	37048	16901	20147	4443	10752	11265	10588	18	24	37 418	16 931	20 487		92,17
7	529354	Bratislava - mestská časť Rača	20438	9623	10815	2352	5956	6673	5457	14	-24	20 172	9 541	10 631		93,16
8	529362	Bratislava - mestská časť Vajnory	4659	2331	2328	606	1392	1654	1007	22	86	3 828	1 899	1 929		95,59
9	529401	Bratislava - mestská časť Devín	1040	527	513	150	286	355	249	85	4	884	441	443		91,97
10	529371	Bratislava - mestská časť Devínska Nová Ves	15948	7791	8157	2077	5875	6084	1912	86	65	15 502	7 509	7 993		93,45
11	529389	Bratislava - mestská časť Dúbravka	34405	16127	18278	4247	9697	10562	9899	23	-137	35 199	16 498	18 701		92,72
12	529397	Bratislava - mestská časť Karlova Ves	33876	15893	17983	5109	11146	10979	6642	20	73	32 843	15 507	17 336		92,39
13	529419	Bratislava - mestská časť Lamač	6580	2974	3606	690	1728	1938	2224	56	17	6 544	2 921	3 623		93,87
14	529427	Bratislava - mestská časť Záhorská Bystrica	2852	1384	1468	411	854	941	646	21	95	2 086	1 003	1 083		96,93
15	529435	Bratislava - mestská časť Čunovo	936	501	435	126	252	354	204	26	-1	911	462	449		68,83
16	529443	Bratislava - mestská časť Jarovce	1296	628	668	164	412	452	268	46	20	1 199	575	624		63,8
17	529460	Bratislava - mestská časť Petržalka	113443	54198	59245	11526	41514	43542	16861	60	-364	117 227	56 116	61 111		92,64
18	529494	Bratislava - mestská časť Rusovce	2422	1189	1233	361	776	837	448	26	71	1 922	958	964		76,27
19	507831	Borinka	557	275	282	78	148	177	154	4	6	519	252	267		95,57
20	507890	Gajary	2894	1400	1494	464	878	1001	551	15	19	2 690	1 311	1 379		96,91
21	507954	Jablonové	1112	539	573	170	334	369	239	7	16	1 056	510	546		97,63
22	507962	Jakubov	1466	734	732	242	440	518	266	4	21	1 312	656	656		95,2
23	508012	Kostolište	1132	568	564	178	340	402	212	6	30	942	476	466		98,2
24	508021	Kuchyňa	1691	841	850	255	489	609	338	6	18	1 597	791	806		98,18
2920	543951	Vojčice	2099	1015	1084	404	615	683	397	9	-2	2 021	994	1 027		96,68
2921	543969	Vojka	513	258	255	125	139	170	79	5	2	434	216	218		11,98
2922	543977	Zatín	789	385	404	125	213	249	202	3	9	788	382	406		13,83
2923	543985	Zbehňov	292	149	143	47	79	106	60	0	-3	292	141	151		71,23
2924	543993	Zemplín	390	193	197	71	110	138	71	1	-6	399	197	202		28,32
2925	544001	Zemplínska Nová Ves	960	464	496	194	265	304	197	4	-5	938	438	500		98,19
2926	544019	Zemplínska Teplica	1500	733	767	349	452	475	224	11	4	1 384	676	708		92,7
2927	544027	Zemplínske Hradište	1124	524	600	140	324	346	314	8	-4	1 201	564	637		93,01
2928	544035	Zemplínske Jastrabie	640	327	313	109	152	226	153	4	2	643	319	324		98,44
2929	544043	Zemplínsky Branč	479	231	248	93	137	156	93	2	-1	443	217	226		90,29
2930																
2931		Indicators for 2007														
2932		Inidcators collected from the Slovak Census 2001														
2933																

***Figure 8**: A "working file" for the Slovak LAU 2 database*

The matching process between the NSI tables and the coding system used in the GISCO files for the Slovak LAU2 geometry shows that 8 new LAU2 are to be integrated in the map. These 8 LAU2 present no information recorded from the Slovak Census but they do present some indicators for 2007.

On the other hand, for the Czech Republic we have extracted 149 indicators covering a larger field of domains (from demographics to dwelling stocks and economics, table 4). As the tables and the base-maps extracted from GISCO/EuroGeographics database are chronologically correlated with the Czech Census and because no Czech LAU2 is missing for the moment, populating the database was not as complicated as was the case for Bulgaria.

Types of indicators	
1. Population by age and marital status	8. Commuters to work and schools
2. Population by age	9. Households by type
3. Population by highest educational attainment	10. Housing stock
4. Population by nationality	11. Houses by the floor number and by basic amenities
5. Population by denomination	12. Dwelling stock
6. Population by economic activity	13. Permanently occupied dwellings by legal reason of use and size of dwelling
7. Economically active population by branch of economic activity	

Table 4: *Categories of statistical indicators for the Czech Republic*

A major advantage observed during the population of the database consists in the fact that a most of the data for the Czech Republic and Slovakia comes from the Census conducted in 2001. Thus the main indicators are at least chronologically harmonized. Unfortunately, these indicators are not also semantically linked, except for the ethnic and confessional structure of the population, for the number of dwellings and for some economic variables.

Thus, as a **PRELIMINARY CONCLUSION:** the Tigris team has succeeded in creating two sample databases for 4 countries (Romania, Bulgaria, Czech Republic and Slovakia). The issue we are working to overcome now is that the indicators are not complete or harmonized yet. Several types of problems were identified, some of them having simple and/or no time consuming solutions, while some others might need a supplementary time for a more advanced analysis in order to provide effective solutions and implement them.

Although the focus remained on the construction of a sample database for the two countries, a part of the team has managed to gather indicators at LAU2 scale for three Scandinavian countries (Norway, Sweden and Finland) and for 2 Baltic states (Latvia and Lithuania). These 5 sets of new indicators (generally demographic and social ones) will be processed in the incoming period, in order for them to properly match with a base-map or with other data tables.

As a conclusive summary of our work, the statistical indicators collected and integrated in the database for the Czech Republic and Slovakia are presented in the table in annex. Some of the indicators are constantly repeated[9] (e.g. the LAU2 coding system and their names) in order to facilitate the preliminary extraction, when needed.

[9] This is the case only for the indicators collected for the Czech Republic.

5 Using SIRE database

Another element that TIGRIS team has to deal it is the recovery and transformation of indicators from SIRE database. Having a particular structure (an obsolete coding system and a spatial hierarchical structure starting from NUTS 0 to ex-NUTS 5, in the same field) the integration of information implies acquiring a specific method. The main chronological marks in SIRE are 1981 and 1991. Obviously, not all the countries in the ESPON space are present in the database and one could think that an interesting and complete exploitation of SIRE should be doubled by an investigation of datasets for recently integrated in the EU.

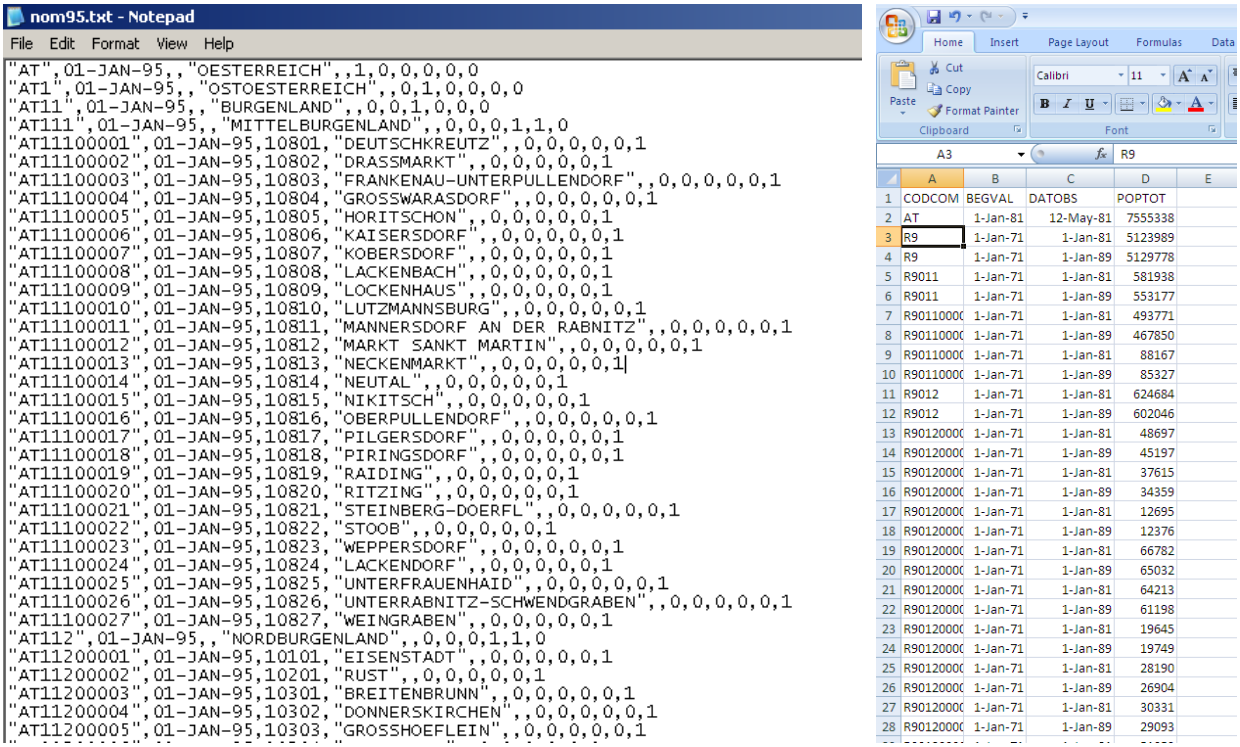


Figure 9: *SIRE database before (on the left – labels and codes) and after (on the right – population in 1991) data basic integration.*

The output of working with SIRE indicators is multiple. It serves for comparison between the coding systems and labels, in order to survey administrative modifications at LAU scale and it's also useful for building some chronologically based indicators between 1991 and 2001, when used in linkage with other databases.

6 Integrating Priority 2 projects

The integration of data obtained in Priority 2 projects represents a priority in TIGRIS work. That's why one of the deliverables was conceived as a container for this kind of information. However, a prioritization of the indicators, based on an analysis of the added value of these new indicators should also be considered as a task. If the information obtained by Priority 2 projects is too recent (2007 or 2008) it may complicate the integration when SOME not spotted administrative changes in geometry occur. A secondary problem could be linked with the eventual cartographic expression of this new information. If two finisterre are to be mapped, a proper projection will highly smooth the visual transmission.

In the next stage, the efforts of the TIGRIS team will be canalized on perfecting the database for the two countries (integrating some recent demographic indicators for Slovakia at LAU2 scale, (re)structuring/refining some data tables at LAU1 scale for the Czech Republic), on sketching a minimum administrative history for Romania and Bulgaria, finalizing the data collection for some other countries in the ESPON space.

One of the issues we are dealing with at the moment is the data validation and the elimination of the possible errors inherently occurring during the data collection and structuring process. Only after we are going to develop a system for data validation, we are going to be able to attach the metadata to our files.

For the moment, our priority still remains that of creating a proper connection between the indicators and the geometry, which could sometimes be problematic (as our experience when working for the Romania and Bulgaria database proved it).

Organizing a working plan in this context seems to depend on variables that are partially controllable by TIGRIS. In the short term our effort will focus on the elaboration of a database with indicators for at least two neighboring countries. For the midterm (December 2099) finalizing a database with indicators at LAU level would be the main priority. In the same time we shall derive a minimal history of LAU1/2 modifications. For February 2010 we had reserved the most time consuming task – recovering SIRE while populating a country by country database with a basic indicator at least.

Conclusion

Gemeinden, Inn, Municipios, Obcine, Comune, Communes, Freguesias, Telepulesek, Ward, is the label for mostly the same geographic reality, the local level of administrative units in some countries of the ESPON space. Exploring them and collecting their basic information is a feasible and necessary task. Dealing with this task means to properly estimate the right balance between the errors in the spatial geometries, the chronological availability, the administrative changes and the sens of words behind the indicators.

The exploration of the available sources of information at LAU 2 scale (NSI, GISCO, SIRE, etc.) shows that building a database for this territorial level should overcome 3 different issues, in order to become a coherent tool. The first issue refers to the chronological heterogeneity of the indicators. Analyzing these indicators country by country, it's quite a luck to find a proper chronological match between them. This problem is underpinned with the second one, the issue of the administrative changes at local level, this last aspect heavily complicating any database populating process. The administrative changes block the construction of a general algorithm (for more than 119 000 LAU2 in ESPON space), especially when intermediate levels of territorial clip are present – the LAU 1 level. Thirdly, the semantic issue of the indicators could also become important. According to country's definition, dwelling or *others* (religion minorities, e.g.) might not have the same sens from Greece to Iceland.

However, despite TIGRIS experience, working at this scale it's learning by doing process, even if doing is pretty fuzzy in this context. The example of the database built for countries such as Slovak Republic, Czech Republic, Bulgaria or Romania shows that another aspect should be taken into account – the relevance of the indicators. The added value of different variables present in the datasets and available for extraction should be prioritize, having in mind the fact that they may largely vary because of the 3 issues already exposed.

When we try to integrate databases such as SIRE in a LAU 2 actual frame we should double the working process by an investigation of the statistical sources available for the '80 and '90 period for some countries recently integrated in the ESPON space. If not, we might obtain a proper image of the past without any link to its future. A comparable problem emerges when we integrate data form the Priority 2 projects. This time, it's not the chronological frame that worries, but the spatial one.

Annex

List of LAU2 indicators for the Czech Republic and Slovakia

INDICATORS	DESCRIPTION	SOURCE AND OBSERVATIONS
Iden	Basemap code	GISCO database
OBJECTID	Inner code used in ARCVIEW	GISCO database
COMM_ID	Basemap code	GISCO database
X	Dummy longitude coordinate	Automatically extracted
Y	Dummy latitude coordinate	Automatically extracted
COMM_NAME	LAU2 label	GISCO database
NAME_ASCI	LAU2 label in ASCII format	GISCO database
NAME_HTML	LAU2 label in HTML format	GISCO database
NAME_SIRE	LAU2 label in SIRE database	GISCO database
TRUE_COMM_	Dummy variable	GISCO database
CNTR_CODE	Country code	GISCO database
AREA_TOTL	Area	GISCO database
AREA_LAND	Area (only null values)	GISCO database
POPL 2001?	Population in 2001 (LAU2)	GISCO database
NSI_CODE	Code used by the National Statistical Institute	GISCO database
LAU2_CODE	LAU2 code (different from the IDEN and COMM_ID)	GISCO database
ADRG_LAU1_	LAU1 hierarchical code	GISCO database
NUTS_CODE	NUTS hierarchical code	GISCO database
DGUR_CODE	Dummy indicator ?	GISCO database
DGUR_AREA_	Area (text values)	GISCO database
DGUR_AREA	Area	GISCO database
POPL_DENS	Population's density in 2001	GISCO database
NATIONAL_CODE_LAU2	Indicator used in the matching process	Automatically extracted (no values only for Czech Republic)
LABEL_LAU2	Indicator used in the matching process	Automatically extracted (no values only for Czech Republic)
Total population (as of Dec. 31)	Total population (as of Dec. 31)	Regional Database (NSI Slovakia) Indicator valid for 2007
Population - males (as of Dec. 31)	Population - males (as of Dec. 31)	Regional Database (NSI Slovakia) Indicator valid for 2007
Population - females (as of Dec. 31)	Population - females (as of Dec. 31)	Regional Database (NSI Slovakia) Indicator valid for 2007
Population in pre-productive age - total (0 - 14)	Population in pre-productive age - total (0 - 14)	Regional Database (NSI Slovakia) Indicator valid for 2007
Population in productive age - females (15 - 54)	Population in productive age - females (15 - 54)	Regional Database (NSI Slovakia) Indicator valid for 2007
Population in productive age - males (15 - 59)	Population in productive age - males (15 - 59)	Regional Database (NSI Slovakia) Indicator valid for 2007
Population in post-productive age-total (55+F, 60+M)	Population in post-productive age-total (55+F, 60+M)	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of marriages	Number of marriages	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of divorces	Number of divorces	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of live births total	Number of live births total	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of live births males	Number of live births males	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of live births females	Number of live births females	Regional Database (NSI Slovakia) Indicator valid for 2007

Number of deaths total	Number of deaths total	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of deaths males	Number of deaths males	Regional Database (NSI Slovakia) Indicator valid for 2007
Number of deaths females	Number of deaths females	Regional Database (NSI Slovakia) Indicator valid for 2007
Total increase (decrease) of population - total	Total increase (decrease) of population - total	Regional Database (NSI Slovakia) Indicator valid for 2007
Total increase (decrease) of population -males	Total increase (decrease) of population -males	Regional Database (NSI Slovakia) Indicator valid for 2007
Total increase (decrease) of population - females	Total increase (decrease) of population - females	Regional Database (NSI Slovakia) Indicator valid for 2007
Population in total	Population in total	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Population - males	Population - males	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Population - females	Population - females	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Population by nationality: Slovak %	Population by nationality: Slovak %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Hungar. %	Hungar. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Gipsy %	Gipsy %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Ruthen. %	Ruthen. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Ukrain. %	Ukrain. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Czech %	Czech %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Morav. %	Morav. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Siles. %	Siles. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
German %	German %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Polish %	Polish %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Population by religions: Roman-Cathol. %	Population by religions: Roman-Cathol. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Evangelic %	Evangelic %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Greek-Cathol. %	Greek-Cathol. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Orthodox %	Orthodox %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Cz.sl. Hussit. %	Cz.sl. Hussit. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
without denom. %	without denom. %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
other %	other %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
not specified %	not specified %	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Economically active persons - total	Economically active persons - total	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Economically active persons - males	Economically active persons - males	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Economically active persons - females	Economically active persons - females	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Employed - total	Employed - total	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)

Employed - males	Employed - males	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Employed - females	Employed - females	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Unemployed - total	Unemployed - total	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Unemployed - males	Unemployed - males	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Unemployed - females	Unemployed - females	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Houses total	Houses total	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Permanent habitational houses total	Permanent habitational houses total	Regional Database (NSI Slovakia) Indicator valid for 2001 (CENSUS)
Iden	Basemap code	GISCO database
OBJECTID	Inner code used in ARCVIEW	GISCO database
COMM_ID	Basemap code	GISCO database
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	GISCO database
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
LAU2	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NAME	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population, total	Population, total	NSI Czech Republic - CENSUS 2002
Economically active, total	Economically active, total	NSI Czech Republic - CENSUS 2003
Agriculture, Forestry, Water economy	Agriculture, Forestry, Water economy	NSI Czech Republic - CENSUS 2004
Industry	Industry	NSI Czech Republic - CENSUS 2005
Construction	Construction	NSI Czech Republic - CENSUS 2006
Wholesale and retail trade, Repair of motor vehicles	Wholesale and retail trade, Repair of motor vehicles	NSI Czech Republic - CENSUS 2007
Transport and Communications	Transport and Communications	NSI Czech Republic - CENSUS 2008
Public administration and Defence; Compulsory social security	Public administration and Defence; Compulsory social security	NSI Czech Republic - CENSUS 2009
Education, Health and social work, Veterinary activities	Education, Health and social work, Veterinary activities	NSI Czech Republic - CENSUS 2010
LAU2code	LAU2 code (identical to the IDEN and COMM_ID)	GISCO database
LAU1	LAU1 code	GISCO database
LAU2	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population, total	Population, total	NSI Czech Republic - CENSUS 2001
Females	Females	NSI Czech Republic - CENSUS 2001
MALESSingle	MALESSingle	NSI Czech Republic - CENSUS 2001
MALESMarried	MALESMarried	NSI Czech Republic - CENSUS 2001
MALESDivorced	MALESDivorced	NSI Czech Republic - CENSUS 2001
MALESWidowed	MALESWidowed	NSI Czech Republic - CENSUS 2001
MALESUnknown	MALESUnknown	NSI Czech Republic - CENSUS 2001
FEMALESSingle	FEMALESSingle	NSI Czech Republic - CENSUS 2001
FEMALESMarried	FEMALESMarried	NSI Czech Republic - CENSUS 2001
FEMALESDivorced	FEMALESDivorced	NSI Czech Republic - CENSUS 2001
FEMALESWidowed	FEMALESWidowed	NSI Czech Republic - CENSUS 2001
FEMALESUnknown	FEMALESUnknown	NSI Czech Republic - CENSUS 2001
LAU2codeOk	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS 4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)

Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population, total	Population, total	NSI Czech Republic - CENSUS 2001
A.G. 0-4	A.G. 0-4	NSI Czech Republic - CENSUS 2001
A.G. 5-14	A.G. 5-14	NSI Czech Republic - CENSUS 2001
A.G. 15-19	A.G. 15-19	NSI Czech Republic - CENSUS 2001
A.G. 20-29	A.G. 20-29	NSI Czech Republic - CENSUS 2001
A.G. 30-39	A.G. 30-39	NSI Czech Republic - CENSUS 2001
A.G. 40-49	A.G. 40-49	NSI Czech Republic - CENSUS 2001
A.G. 50-59	A.G. 50-59	NSI Czech Republic - CENSUS 2001
A.G. 60-64	A.G. 60-64	NSI Czech Republic - CENSUS 2001
A.G. 5-74	A.G. 5-74	NSI Czech Republic - CENSUS 2001
A.G. 75+unknown	A.G. 75+unknown	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population aged 15+	Population aged 15+	NSI Czech Republic - CENSUS 2001
Without education	Without education	NSI Czech Republic - CENSUS 2001
Basic incl. not finished	Basic incl. not finished	NSI Czech Republic - CENSUS 2001
Secondary vocational and technical without GCSE	Secondary vocational and technical without GCSE	NSI Czech Republic - CENSUS 2001
Full secondary general with GCSE	Full secondary general with GCSE	NSI Czech Republic - CENSUS 2001
Higher professional and Extension study	Higher professional and Extension study	NSI Czech Republic - CENSUS 2001
University	University	NSI Czech Republic - CENSUS 2001
Not identified	Not identified	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS 4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population, total	Population, total	NSI Czech Republic - CENSUS 2001
Czech	Czech	NSI Czech Republic - CENSUS 2001
Moravian	Moravian	NSI Czech Republic - CENSUS 2001
Silesian	Silesian	NSI Czech Republic - CENSUS 2001
Slovak	Slovak	NSI Czech Republic - CENSUS 2001
Romany	Romany	NSI Czech Republic - CENSUS 2001
Polish	Polish	NSI Czech Republic - CENSUS 2001
German	German	NSI Czech Republic - CENSUS 2001
Ukrainian	Ukrainian	NSI Czech Republic - CENSUS 2001
Vietnamese	Vietnamese	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population, total	Population, total	NSI Czech Republic - CENSUS 2001
Believers	Believers	NSI Czech Republic - CENSUS 2001
Roman Catholic Church	Roman Catholic Church	NSI Czech Republic - CENSUS 2001
Czechoslovak Hussite Church	Czechoslovak Hussite Church	NSI Czech Republic - CENSUS 2001

Evangelical Church of Czech Brethren	Evangelical Church of Czech Brethren	NSI Czech Republic - CENSUS 2001
Orthodox Church	Orthodox Church	NSI Czech Republic - CENSUS 2001
Jehovah´ Witnesses	Jehovah´ Witnesses	NSI Czech Republic - CENSUS 2001
Undenominational	Undenominational	NSI Czech Republic - CENSUS 2001
Unknown Denomination	Unknown Denomination	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Population, total	Population, total	NSI Czech Republic - CENSUS 2001
Economically active (EA), total	Economically active (EA), total	NSI Czech Republic - CENSUS 2001
employed persons	employed persons	NSI Czech Republic - CENSUS 2001
EA pensioners	EA pensioners	NSI Czech Republic - CENSUS 2001
women on maternity leave	women on maternity leave	NSI Czech Republic - CENSUS 2001
unemployed persons	unemployed persons	NSI Czech Republic - CENSUS 2001
Economically inactive (EI), total	Economically inactive (EI), total	NSI Czech Republic - CENSUS 2001
EI pensioners	EI pensioners	NSI Czech Republic - CENSUS 2001
Pupils,students, apprentices	Pupils,students, apprentices	NSI Czech Republic - CENSUS 2001
Economic activity not identified	Economic activity not identified	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS 4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Commuters to work	Commuters to work	NSI Czech Republic - CENSUS 2001
within municipality	within municipality	NSI Czech Republic - CENSUS 2001
within district	within district	NSI Czech Republic - CENSUS 2001
within region	within region	NSI Czech Republic - CENSUS 2001
into other region	into other region	NSI Czech Republic - CENSUS 2001
Commuters to work daily out of municipality	Commuters to work daily out of municipality	NSI Czech Republic - CENSUS 2001
Pupils commuting to schools daily out of municipality	Pupils commuting to schools daily out of municipality	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Dwelling households, total	Dwelling households, total	NSI Czech Republic - CENSUS 2001
with 1 PV*	with 1 PV*	NSI Czech Republic - CENSUS 2001
with 2+PV*	with 2+PV*	NSI Czech Republic - CENSUS 2001
Private households	Private households	NSI Czech Republic - CENSUS 2001
with 1 census household	with 1 census household	NSI Czech Republic - CENSUS 2001
with 2 and over census household	with 2 and over census household	NSI Czech Republic - CENSUS 2001
Census households (C-H), total	Census households (C-H), total	NSI Czech Republic - CENSUS 2001
Two-parent families	Two-parent families	NSI Czech Republic - CENSUS 2001
with dependent children	with dependent children	NSI Czech Republic - CENSUS 2001
Lone-parent families	Lone-parent families	NSI Czech Republic - CENSUS 2001
with dependent children	with dependent children	NSI Czech Republic - CENSUS 2001
Non-family households	Non-family households	NSI Czech Republic - CENSUS 2001
Households of individuals	Households of individuals	NSI Czech Republic - CENSUS 2001

LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Houses, total	Houses, total	NSI Czech Republic - CENSUS 2001
Permanently occupied houses	Permanently occupied houses	NSI Czech Republic - CENSUS 2001
Family houses	Family houses	NSI Czech Republic - CENSUS 2001
Multi-dwelling houses	Multi-dwelling houses	NSI Czech Republic - CENSUS 2001
Houses by ownershipprivate persons	Houses by ownershipprivate persons	NSI Czech Republic - CENSUS 2001
Houses by ownershipcommunity,state	Houses by ownershipcommunity,state	NSI Czech Republic - CENSUS 2001
Houses by ownershiphousing association	Houses by ownershiphousing association	NSI Czech Republic - CENSUS 2001
Houses builtup to 1919	Houses builtup to 1919	NSI Czech Republic - CENSUS 2001
Houses built1920-1945	Houses built1920-1945	NSI Czech Republic - CENSUS 2001
Houses built1945-1980	Houses built1945-1980	NSI Czech Republic - CENSUS 2001
Houses built1981-2001	Houses built1981-2001	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Houses, total	Houses, total	NSI Czech Republic - CENSUS 2001
by floor (above ground)1-2	by floor (above ground)1-2	NSI Czech Republic - CENSUS 2001
by floor (above ground)3-4	by floor (above ground)3-4	NSI Czech Republic - CENSUS 2001
by floor (above ground)5+	by floor (above ground)5+	NSI Czech Republic - CENSUS 2001
Sewage: connection to the public system	Sewage: connection to the public system	NSI Czech Republic - CENSUS 2001
Water supply system	Water supply system	NSI Czech Republic - CENSUS 2001
Gas supply	Gas supply	NSI Czech Republic - CENSUS 2001
Central heating	Central heating	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Dwellings, total	Dwellings, total	NSI Czech Republic - CENSUS 2001
Permanently occupied dwellings	Permanently occupied dwellings	NSI Czech Republic - CENSUS 2001
Family houses	Family houses	NSI Czech Republic - CENSUS 2001
Multi-dwelling houses	Multi-dwelling houses	NSI Czech Republic - CENSUS 2001
Unoccupied dwellings in permanently occupied houses	Unoccupied dwellings in permanently occupied houses	NSI Czech Republic - CENSUS 2001
Unoccupied dwellings in unoccupied houses	Unoccupied dwellings in unoccupied houses	NSI Czech Republic - CENSUS 2001
occupied temporarily	occupied temporarily	NSI Czech Republic - CENSUS 2001
used for recreation	used for recreation	NSI Czech Republic - CENSUS 2001
LAU2 code	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech	NSI Czech Republic - CENSUS 2001

	Republic	
Permanently occupied dwellings, total	Permanently occupied dwellings, total	NSI Czech Republic - CENSUS 2001
by legal reason of use : in own house	by legal reason of use : in own house	NSI Czech Republic - CENSUS 2001
by legal reason of use : in own dwelling	by legal reason of use : in own dwelling	NSI Czech Republic - CENSUS 2001
by legal reason of use : rented	by legal reason of use : rented	NSI Czech Republic - CENSUS 2001
by legal reason of use : in dwelling of housing association	by legal reason of use : in dwelling of housing association	NSI Czech Republic - CENSUS 2001
1 living room	1 living room	NSI Czech Republic - CENSUS 2001
2 living rooms	2 living rooms	NSI Czech Republic - CENSUS 2001
3 living rooms	3 living rooms	NSI Czech Republic - CENSUS 2001
4 living rooms	4 living rooms	NSI Czech Republic - CENSUS 2001
5+ living rooms	5+ living rooms	NSI Czech Republic - CENSUS 2001
Iden	LAU2 code (identical to the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
OBJECTID	Inner code used in ARCVIEW	GISCO database
COMM_ID	Basemap code	GISCO database
LAU2code	LAU2 code (identical to the IDEN and COMM_ID)	GISCO database
NUTS4	LAU1 hierarchical code in ancient format (NUTS)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality code	LAU2 code (different from the IDEN and COMM_ID)	Automatically extracted (NSI Czech Republic - CENSUS 2001)
Municipality name	Label used by the NSI Czech Republic	NSI Czech Republic - CENSUS 2001
Permanently occupied dwellings, total	Permanently occupied dwellings, total	NSI Czech Republic - CENSUS 2001
dwellings by basic amenities : Gas supply in dwelling	dwellings by basic amenities : Gas supply in dwelling	NSI Czech Republic - CENSUS 2001
dwellings by basic amenities : Water supply in dwelling	dwellings by basic amenities : Water supply in dwelling	NSI Czech Republic - CENSUS 2001
dwellings by basic amenities : Private flush toilet	dwellings by basic amenities : Private flush toilet	NSI Czech Republic - CENSUS 2001
dwellings by basic amenities : Bathroom, shower inside dwelling	dwellings by basic amenities : Bathroom, shower inside dwelling	NSI Czech Republic - CENSUS 2001
dwellings by basic amenities : Central heating	dwellings by basic amenities : Central heating	NSI Czech Republic - CENSUS 2001
dwellings by basic amenities : Single-storey heating	dwellings by basic amenities : Single-storey heating	NSI Czech Republic - CENSUS 2001
Average number of : dwelling persons	Average number of : dwelling persons	NSI Czech Republic - CENSUS 2001
Average number of : persons per living room up to 8 m2	Average number of : persons per living room up to 8 m2	NSI Czech Republic - CENSUS 2001
Average number of: occupied living area per dwelling	Average number of: occupied living area per dwelling	NSI Czech Republic - CENSUS 2001
Average number of : occupied living area per 1 person	Average number of : occupied living area per 1 person	NSI Czech Republic - CENSUS 2001
Average number of : living rooms per dwelling	Average number of : living rooms per dwelling	NSI Czech Republic - CENSUS 2001

References

- ***Litterature***

Korte B. G., 2001, The GIS book 5th edition, Onword Press, New York

Turcanasu G., Rusu A., 2008, Le système des villes en Bulgarie et en Roumanie. Quelles perspectives pour un polycentrisme?, Espace Geographique, no.4/2008

Groza O., 2005, Maillages administratifs officiels et identités territoriales officieuses: les échelons spatiaux de la différenciation identitaire en Roumanie, in V. Rey; T. Saint-Julien - „Territoires d'Europe, la différence en partage", ENS- Editions, Lyon, pp.153-160,

- ***Websites***

http://www.insse.ro/cms/rw/pages/index.en.do, National Institute of Statistics, Romania

http://www.nsi.bg/index_en.htm, National Statistical Institute of Bulgaria

http://px-web.statistics.sk/PXWebSlovak/index_en.htm, Statistical Office of the Slovak Republic

http://www.czso.cz/eng/redakce.nsf/i/home, Czech Statistical Office

http://ec.europa.eu/eurostat/ramon/nuts/lau_en.html, Eurostat – list of LAU

www.territorial-intelligence.eu/caenti/, Coordination action of the European Network of Territorial Intelligence

TECHNICAL REPORT

ESPON

Local and regional data

Producing innovative indicators

MAIN RESULTS

- Local data can be integrated from different sources or it can be produced using the LAU2 as elementary spatial patterns
- Using an appropriate number of LAU2, the data can be mobilized in specific geographical models (the estimated GDP 2006 distribution at local scale)
- The distances calculated at local scale represent an useful analysis tool which can provide new territorial indicators

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

25 PAGES

LIST OF AUTHORS

Octavian Groza, UAIC, CUGUAT-TIGRIS, Iasi, Romania
Alexandru Rusu, UAIC, CUGUAT-TIGRIS, Iasi, Romania

Contact
octaviangroza@yahoo.com

tel. + 40 0232 20 14 88

General plan

1.1 The data harmonization at LAU2 scale is depending on the spatial harmonization of the geometries for the selected countries.
1.2 Working with a large number of LAU2 units involves accepting the extraordinary values.
1.3 Mapping the evolution(s) in connection with some explanatory factors.
1.4 Explaining the demographic decline: some basic hypothesis (the auto-correlation)

2.1 A short introduction in the spatial patterns: points, polygons, lines and networks. The LAU2 as spatial patterns.
2.2 Using the spatial patterns as a base for the indicators construction.
2.3 A case study on Romania and Hungary – the calculation of the LAU2 accessibility.

Data integration and creation of basic indicators (INTERMEDIATE) p. 4

1. The production of some basic indicators, using the **LAU2** geometries and data. Example: calculating the relative evolution of population for selected countries in Eastern Europe. p. 5

2. The production of some basic indicators, using the **LAU2** geometries as a geographic object. Spatial patterns vs. spatial structures. p.9

1.1 How to measure the economic performance at LAU2 scale?
1.2 How to find relevant and harmonized data for this operation? A case study on the local aggregated turnover in 2006 – the Romanian case.
1.3 How to by-pass the lack of harmonization using grid information?
1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.
1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.

2.1 Using the distance as an indicator for territorial coherence in the Eastern Europe.
2.2 The settlement's hierarchy – how many levels function from Prague to Sulina?
2.3 How to map the distance: choroplet vs. "oursins".
2.4 The local fragmentation of the territorial architecture and the hierarchical immobility.

Data integration in geographical models (ADVANCED) p. 13

1.Towards more elaborate indicators and models applied to **LAU2** objects. The potential of interaction as a measure for the local economic performance. p. 14

2.Towards more elaborate indicators and models applied to **LAU2** objects. The settlement's hierarchy and the territorial architecture of the selected countries. p. 24

INTERMEDIATE

1.1 The data harmonization at LAU2 scale is depending on the spatial harmonization of the geometries for the **selected countries**.
1.2 Working with a large number of LAU2 units involves accepting the extraordinary values.
1.3 Mapping the evolution(s) in connection with some explanatory factors.
1.4 Explaining the demographic decline: some basic hypothesis (the auto-correlation)

2.1 A short introduction in the **spatial patterns**: points, polygons, lines and networks. The LAU2 as spatial patterns.
2.2 Using the spatial patterns as a base for the indicators construction.
2.3 A case study on Romania and Hungary – the calculation of the LAU2 accessibility.

Relevant findings

The demographic decline affects area of the size of a medium ESPON country. The process presents spatial homogeneity and has chances to become a trans-scalar and cross-border issue.

The demographic growth is possible, even in a context of turbulent economic performance. It almost concerns only the large metropolitan areas and some regions with specific geographical features.

The elementary spatial patterns (points, polygons, networks) can be mobilized in the production of relevant indicators for the local territories. Basic techniques of spatial analysis can be implemented, even when we deal with a large number of LAU2.

The basic indicators (accessibility in the network) can be complicated (the coefficient of variance for the local **accessibility**) or they can present interest for some local typologies.

1. The production of some basic indicators, using the **LAU2** geometries and data. Example: calculating the relative evolution of population for selected countries in Eastern Europe.

2. The production of some basic indicators, using the **LAU2** geometries as a geographic object. Spatial patterns vs. spatial structures.

Selected countries
There are five countries selected as case study: the Czech Republic, Slovakia, Hungary, Romania and Bulgaria. The number of LAU2 in these 5 states is more than 20 000. The number, the shape of the LAU2 and their variable administrative fragmentation is a good opportunity to test methodologies for data collection and indicators creation.

Spatial patterns
Every LAU2 can be resumed to elementary spatial patterns: an administrative center (point) and an administrative surface (polygon). These LAU2 are connected by different networks. The spatial patterns present specific methods of analysis (e.g.: the weighted barycenter for the points, the number of neighbors for the polygons, the distance calculation in the network etc.). Our option was to integrate all the spatial patterns in one model able to explain how the road network accessibility shapes the territorial structures for some specific countries.

Accessibility
The geographical littérature is abundant in definitions concerning this concept. Some of them are mathematically formalized in order to be applied as spatial models. From an intuitive and superficial point of view, accessibility means the capacity of one place (LAU2) to be reached from other places (LAU2). The output is measured in kilometers, time and costs. An easy way to implement this indicator is to calculate the average distance separating a LAU2 from the others (a basic accessibility indicator).

1.1 The data harmonization at LAU2 scale is depending on the spatial harmonization of the geometries for the selected countries.

Our intention is to describe by a map the demographic evolutions at LAU2 scale for the selected countries, between 2001 and 2006. This intention is based on the presence of harmonized data concerning the number of inhabitants for the two years. If the data is ready to use, the spatial frame (LAU2 geometries) is a major issue. If in 2001 some of the capitals were presented as a single polygon, in 2006 their administrative territory is divided (Budapest more than 20 LAU2, Bratislava more than 15, Bucarest 6 sectors). It isn't a major issue, as a matter of fact the aggregation of data is just a time problem. Other LAU2 are more concerning. Some of them have disappeared during 2001-2006, some others just appeared on the map, without necessary having a connection between the two categories. The logical way to deal with the spatial lack of harmonization is to exclude them from the analysis. This option becomes illogical when the administrative mutations in the local geometry become too important. In that case, it is recommended to apply various techniques that can finally harmonize these spatial dynamics.

1.2 Working with a large number of LAU2 units involves accepting the extraordinary values.

The issue of the extraordinary values is induced by the large number of spatial units with almost insignificant population (in Bulgaria we deal with municipalities that used to have 10 inhabitants in 2001). In this LAU2, every change in the population will transmit "considerable" evolutions when we map relative indicators, such as the variation of population for the mentioned period. The same phenomena can be observed for some LAU2 that are situated in the proximity of large cities, involved in processes of sub-urbanization. As a matter of fact, the extraordinary values are not extra-ordinary at all, if one would look at the local context in which the demographic trends are deployed. Usually, these values are consistent with the regional evolutions that shape the demographic decline or growth, the positive spatial auto-correlation being a general rule. One special case of extraordinary value appears in areas with administrative mutations. If one LAU2 suffered an administrative division/split, it will create a "fake" extraordinary value that reflects the creation of a new LAU2 rather than a demographic evolution.

>>> 1.3 Mapping the evolution(s) in connection with some explanatory factors.

Administrative mutations and extraordinary values

There are many cases of administrative mutations that can influence the mapping process. One of the most common issue is the territorial division, from one LAU2 resulting 2 new spatial units. We cannot correctly estimate the demographic evolution, in this case, for obvious reasons. Sometimes is not the division of LAU that becomes a problem. The administrative union of two LAU also interferes with the data calculation, even for simple indicators like the evolution of population. We can find even more complicated cases where the division is followed by a union, involving massing data or extraordinary values for three LAU. Occurring very often in Romania and Bulgaria, these mutations are reflecting the importance of the local dimension in the political and administrative strategies. We can observe on the map that these changes in the basic geometry are correlated with the economic territorial rhythms. Thus, in some less dynamic regions, the fragmentation indicates the stake of the public administrative finance at local scale, while in the pro-active regions we can interpret it as a trend to concentrate the local financial resources (the proximity of cities, touristic regions, rich industrial and transportation corridors).

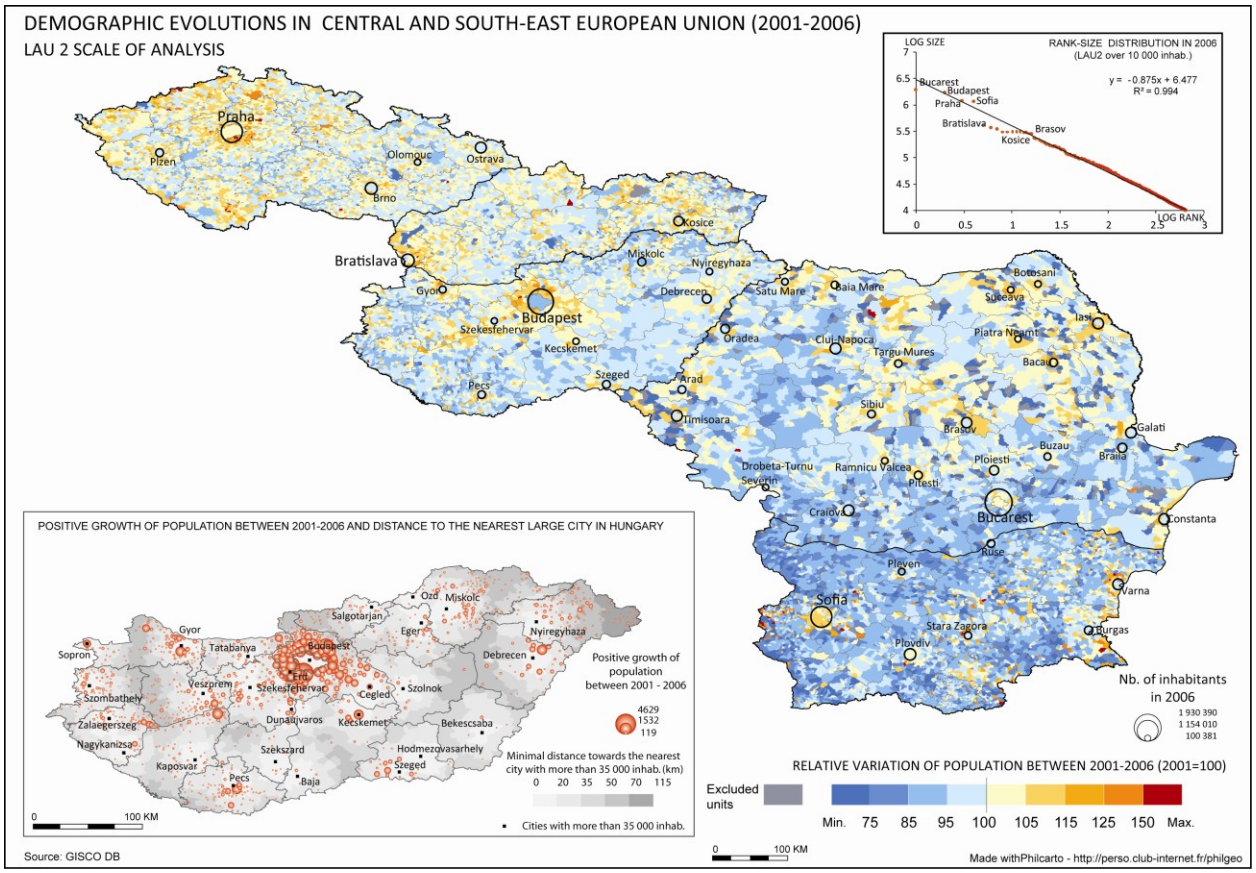


Fig. Trends of demographic evolution in 5 selected countries of the ESPON space (draft map)

1.3 Mapping the evolution(s) in connection with some explanatory factors.

Despite the spatial fragmentation and the lack of coherence regarding the LAU2 geometry of the 5 countries, some general trends can be easily identified. The positive demographic evolution is a function of some qualitative and quantitative transformations that reshape the role of the large cities in the territory: sub-urbanization and metropolitan development. However, defining the large city in the area of the selected countries is not an easy task. The rank size distribution for 2006 indicates that a possible superior limit should be 280 000 inhabitants (an approximation). Many cities below this limit are also involved in the suburbanization process, reflecting their key position in the national urban system and also their economic strength. Sometimes, the demographic growth is controlled by the distance towards the nearest city, such is the case in Hungary – extremely visible for Budapest, but also for Pecs, Szeged or Debrecen. On the other side, some rural regions (with a traditional pro-natalist behavior) like Moldavia (Romania) or Haskovo-Kardzali region (Bulgaria) still conserve positive evolution. Other recent studies emphasized that this natalist behavior is in extinction. A special case is the Central-Transylvania where the final phase of the demographic transition was conjecturally interrupted by some ethnic and confessional local specificity. The demographic decline is a challenging reality for vast rural spaces in Bulgaria, Romania and Hungary, but also regionally present in Czech Republic and Slovakia. Taking into account the surface concerned by this phenomena - the cross-border Danubian region between Romania and Bulgaria or the cross-border region between Romania and Hungary, this decline could represent a policy relevant aspect regarding the demographic evolutions in the ESPON space.

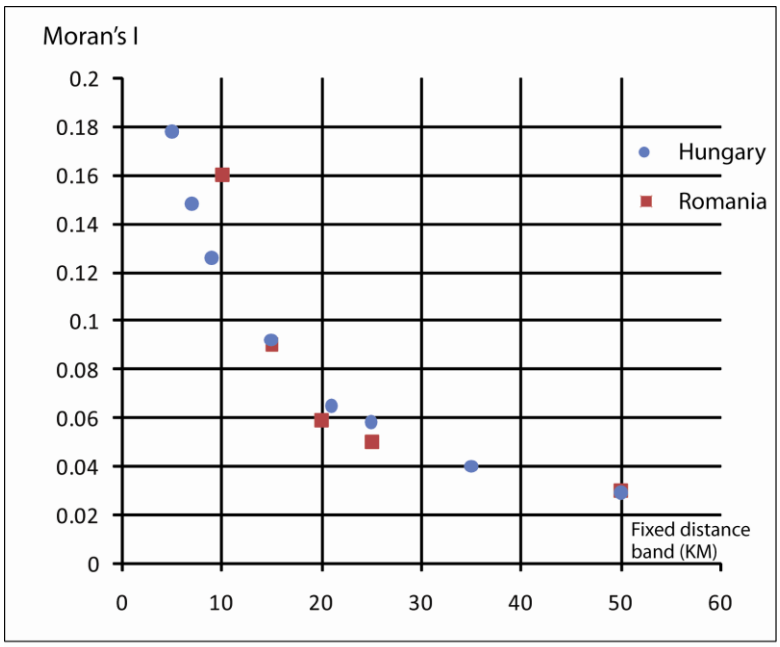
>>> 1.4 Explaining the demographic decline: some basic hypothesis (the auto-correlation)

1.4 Explaining the demographic decline: some basic hypothesis (the auto-correlation)

The distribution of the three dimensions of the demographic evolution (stability, decline and growth) present a specific spatial pattern. As a general rule, the LAU2 characterized by decline seem to have neighbors presenting the same trend. In the proximity of the cities, the LAU2 positive evolution is also included in a locally homogeneous context. The area of relative demographic stability is also subject to similarity with the neighbors (this area is visible in the Romanian Sub-Carpathians). This effect is called spatial auto-correlation and it is largely developed and formalized in the geographic litterature. There are many ways in which we can test its existence (Geary test, Moran's I or the measure of the local dissimilarity). In our case, testing the presence of the spatial auto-correlation in the demographic evolution is a good method to estimate the relationship between the indicator and the local context. Basically, we try to estimate the size of a homogeneous region that is characterized by the same demographic trend, at local scale. That is the sens of the word "explaining" in the title of this fragment. Technically, we have to follow several steps in order to obtain the size of a homogeneous region:

1) Choose a method: testing the spatial auto-correlation using a GIS is a simple task. In the absence of a GIS, there is a spreadsheet method that involves the manipulation of a large table of geographic information.
2) Depending on the method, reflecting on the concept of neighborhood is also useful. At local scale, the administrative contiguity is problematic because we have to take into account the spatial fragmentation (e.g. administrative contiguity is problematic in Bulgaria, exaggerated in Romania, just good in Hungary). A distance bandwidth will be a better option in order to define the neighbors and the proximity effect.
3) Three hypothesis are now available for testing :
   H0: the spatial auto-correlation is null. Any two neighbor LAU are neither similar, neither different to any two LAU that are not neighborhood related.
   H1: the spatial auto-correlation is negative. Two neighbors LAU are less similar than two distanced LAU. The local context is characterized by heterogeneity.
   H2: **the spatial auto-correlation is positive**. Two neighbors LAU are rather similar compared with two distanced LAU. The local context is characterized by homogeneity.
4) If the test confirms the **H2,** by enlarging the neighborhood context we can obtain the size of a locally homogeneous region in relation with the demographic evolution. For more details, see http://grasland.script.univ-paris-diderot.fr/

In our case study, we have made an option for the GIS solution because using a spreadsheet method will involve the manipulation of 20 000 by 20 000 matrix, which is a time consuming strategy, difficult to implement. The test of spatial autocorrelation (Moran's I) confirms the **H2** hypothesis (0.145 in a range of 15 km around each LAU2 in the 5 selected countries, statistically representative for more than 20 000 LAU2) and shows that the local context and the demographic trends are related. In this case, we have proceeded to the development of the point 4 (see above). Despite our effort, we cannot provide the size of the homogeneous regions having the same demographic pattern for all the countries included in our study. Therefore, we have applied the test of spatial autocorrelation at different distance bands only for Hungary and Romania. The general trend of the Morans's I distribution for the two countries shows a decrease of the local context's effect when we extend the definition of the "local" from 5 to 50 km. For Hungary the test was started at 5 km and then incremented by 2 km. In the Romanian case, to avoid data exclusion, the spatial auto-correlation was measured at 10 km and incremented by 5 km, until the 50 km limit was reached. The analysis of the Moran's I distribution pattern shows that the local effects diminish gradually, a clear limit for regions with similar behavior being difficult to estimate.



Hungary and Romania

The general aspect of the Moran's I distribution indicates that the auto-correlation effect decrease almost as a power function of distance in both cases. The low values of the indicator of spatial auto-correlation define an ambiguous relation between the local context and the demographic trends. With more than 3000 spatial units for each country, even these low values are representative. A better illustration of this local effect could be emphasized by using a map of the spatial distribution of the local dissimilarities. Applied for Hungary, this kind of map is an exploration tool (not very sophisticated), showing some of the local sensitivities.

Fig. Distribution of Moran's I as a distance decay function in Hungary and Romania

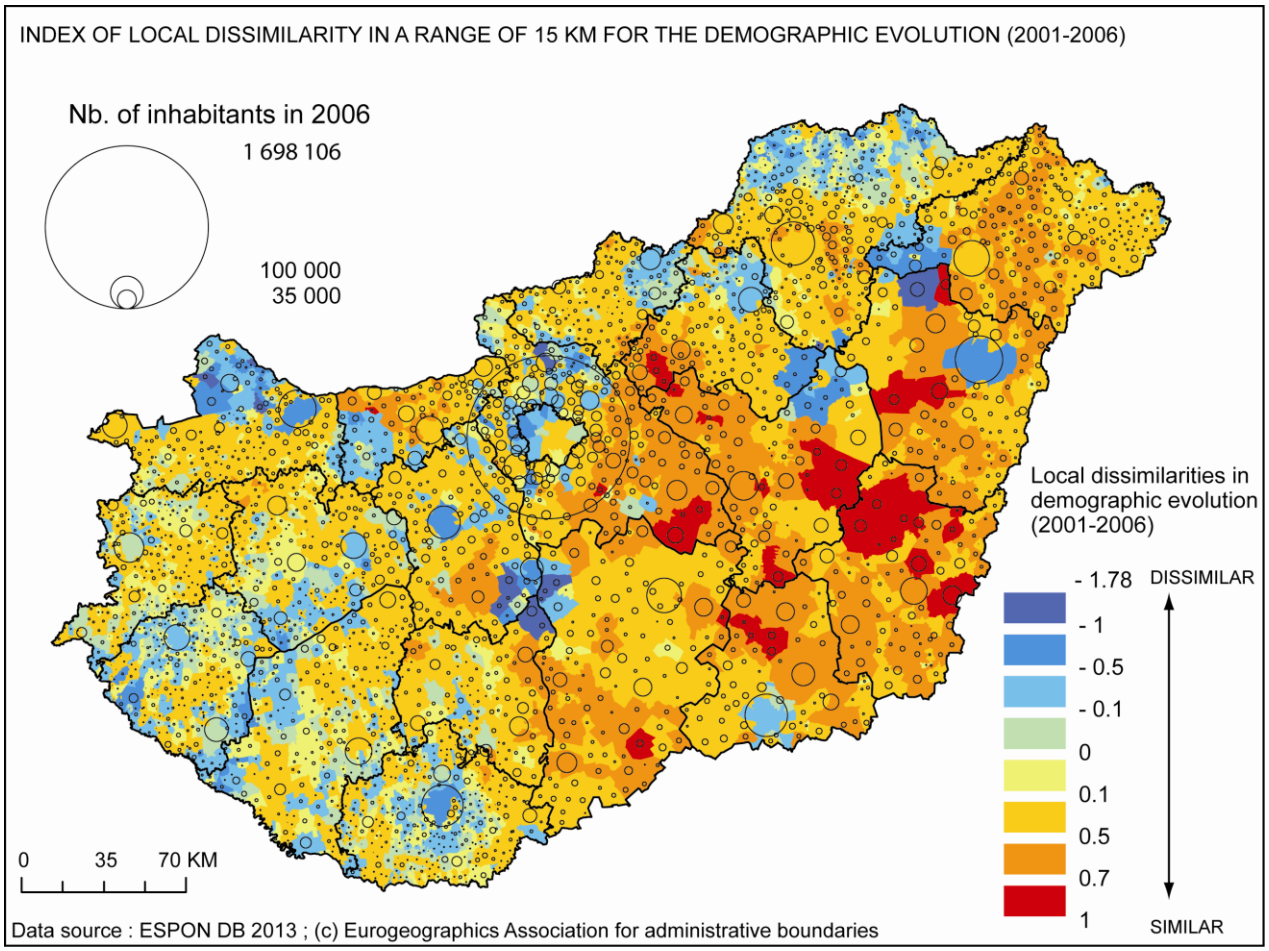


Fig. Differences in the influence of the local context on the demographic evolution of the Hungarian Lau2

Reading a similarity/dissimilarity map is not an easy task. Rather than showing the spatial repartition of a phenomena (demographic dynamics), it shows how the spatial units are acting in relation with their local context (15 km in this case), regardless of the trend – decline, stability, growth. The positive values in the legend indicate local homogeneity, while the negative ones a different behavior compared with the neighborhood. The color's graduation shows the intensity of the homogeneity vs. heterogeneity spatial distribution concerning the demographic evolution in Hungary. In a context of general decline of population, a large part of the Hungarian territory is behaving spatially auto-correlated. As a map description, the heterogeneous areas are situated near the large Hungarian cities and in the proximity of the "triplum confinium" of some of the NUTS 3. The homogeneous zones are occupying the central and the eastern parts of Hungary and this moderate lack of dissimilarities should also be linked to the size of the LAU2 in these regions. This map is the output of a double interrogation in a similarity matrix weighted by distances, a map that involved the manipulation of more than 3000 rows and columns. However, this method (even if time consuming) provides a good tool in order to better interpret the spatial repartition of the demographic dynamics and could be particularly useful in studies concerning the cross-border regions.

2.1 A short introduction in the **spatial patterns**: points, polygons, lines and networks. The LAU2 as spatial patterns.
2.2 Using the spatial patterns as a base for the indicators construction.
2.3 A case study on Romania and Hungary – the calculation of the LAU2 accessibility.

2.1 A short introduction in the spatial patterns: points, polygons, lines and networks. The LAU2 as spatial patterns.

We like it or not, the geographical reality cannot be synthesized in more than 3 elementary geometries: points, lines and surfaces. Each type of spatial pattern (also called sometimes spatial structures, even if arguable concerning the epithet) involves specific methodological approaches. As an example, the point patterns can be analyzed using the weighted centroid technique, the networks by using the graph theory and the surfaces by taking into account the shape of the polygons. Reducing the LAU2 or the LAU2 information to elementary geometrical features allows us to produce some geographic indicators that can be integrated in a local database. The interest of the spatial patterns analysis is not to fill fields of information in a table (it is also a method to increase the inflation of the information), it rather touches the need to intersect or relate indicators in an explanatory process (e.g. the local economic performance as an eventual output of the accessibility).

2.2 Using the spatial patterns as a base for the indicators construction.

Some of the techniques that we can use in order to approach these different spatial patterns will provide only synthetic indicators (the weighted centroid of the population or the standard distance deviation as a function of some central features, such as capitals). Producing specific information for each LAU2 situated in our 5 countries involves a different approach. Using the reticular spatial structures and the LAU2 centroids for Hungary and Romania we have explored the possibility to relate the geographical position of the spatial units to their position in a network. One of the common methods is to calculate distances in both the network and the geographical coordinate system. As a matter of fact, we have tried to confront Euclidean distances between LAU2 to the "real" distances in a network (the road network is sufficiently detailed to allow it). The obtained indicator is not quite an accessibility indicator, it rather functions as a network efficiency measure.

>>>**2.3 A case study on Romania and Hungary – the calculation of the LAU2 accessibility.**

2.3 A case study on Romania and Hungary – the calculation of the LAU2 accessibility.

The Euclidean distances in a point spatial pattern reflect relations that might occur in an isotropic and homogeneous territory. The distances within the network will measure the relations that might occur in a historically planned transportation system. A comparison between the two kinds of distances allows us to map for Hungary and Romania a derived indicator of local accessibility. In both the countries the low values appears in the plain regions or in regions with a fair road network density. When the connectivity in the network is problematic or when we take into account the mountain zones, we observe higher values of the indicator. Sometimes, these high values are linked with the absence of some essential infrastructural features: a bridge across the Danube or other main rivers, sinuous transportation corridors. The deviation to the shortest Euclidean path will have an influence on the time distances and the cost distances, being sometimes an issue in the construction of the local economic performance. It also reflects the high degree of dependence on the mentioned essential infrastructural features and the lack of alternative road segments in the transportation process. In the Romanian case the high and problematic values of the indicator are specific to some remote LAU2 from the Carpathian Mountains and in regions with an intense relief fragmentation. In Hungary we observe a concentration of the high values in the southern part, confirming the spatial discontinuity effect of the Danube. Some LAU2 where not included in the calculation process because their centroid is not in the proximity of the road network (more than 10 km.). It is the case for LAU 2 situated in the Danube Delta (where there is no road network and the access is granted by water) and in very isolated mountain regions. One critical aspect of the two maps is the lack of harmony in the classification of values in the legend. The reason for this discrepancy is the conservation of the classification method – the natural breaks (Jenks). This lack of harmony also shows that low or medium values for Romania do not have the same sens in Hungary, where they can be considered intense deviations to the shortest Euclidean path. We can also speculate and think that in countries with a larger surface, the probability to encounter high deviation of the values is also stronger. There are some steps that we have to take into consideration when working with the road network efficiency indicator:

- verifying the topology of the network (the connectivity algorithm and the presence of bridges) is a long process. In its absence bizarre situations and extraordinary values may appear.
- Choosing an appropriate system of map projection is also needed in order to avoid exaggerated deviations for the Euclidean distances.
- The conversion of the Euclidean distances in km or other distance units should be carefully supervised.

Accessibility at local scale

Some of the ESPON projects already focused on the issues induced by the accessibility, but most of them touched only the NUTS3 frame. When the local dimension was taken into consideration, it was the accessibility to the various networks (roads or rails) that presented interest. One challenging intention would be to confront basic indicators of accessibility (the average distance that separates a LAU2 from all the other) in a double territorial context – the point spatial patterns and the network spatial patterns. The output of this approach reflects the relationship between the so called "natural accessibility" (the role of the natural features in the construction of the local accessibility) and the general characteristics of the road transportation network. The steps needed to complete this kind of analysis start with the construction of the Euclidean accessibility matrix. After that, the extraction of the average distance for every LAU2 is needed. A second problem is to build an origin-destination matrix in a common GIS. By summarization we will obtain the average distances within the network. The indicators will be put into a simple report and the road network efficiency measure is available for the mapping process. This efficiency refers only to the geometric characteristics of the network and it does not take into account the qualitative aspects. If we map only one indicator, such as the average distance that will separate each LAU2 from all the other, a core-peripheries model will appear.

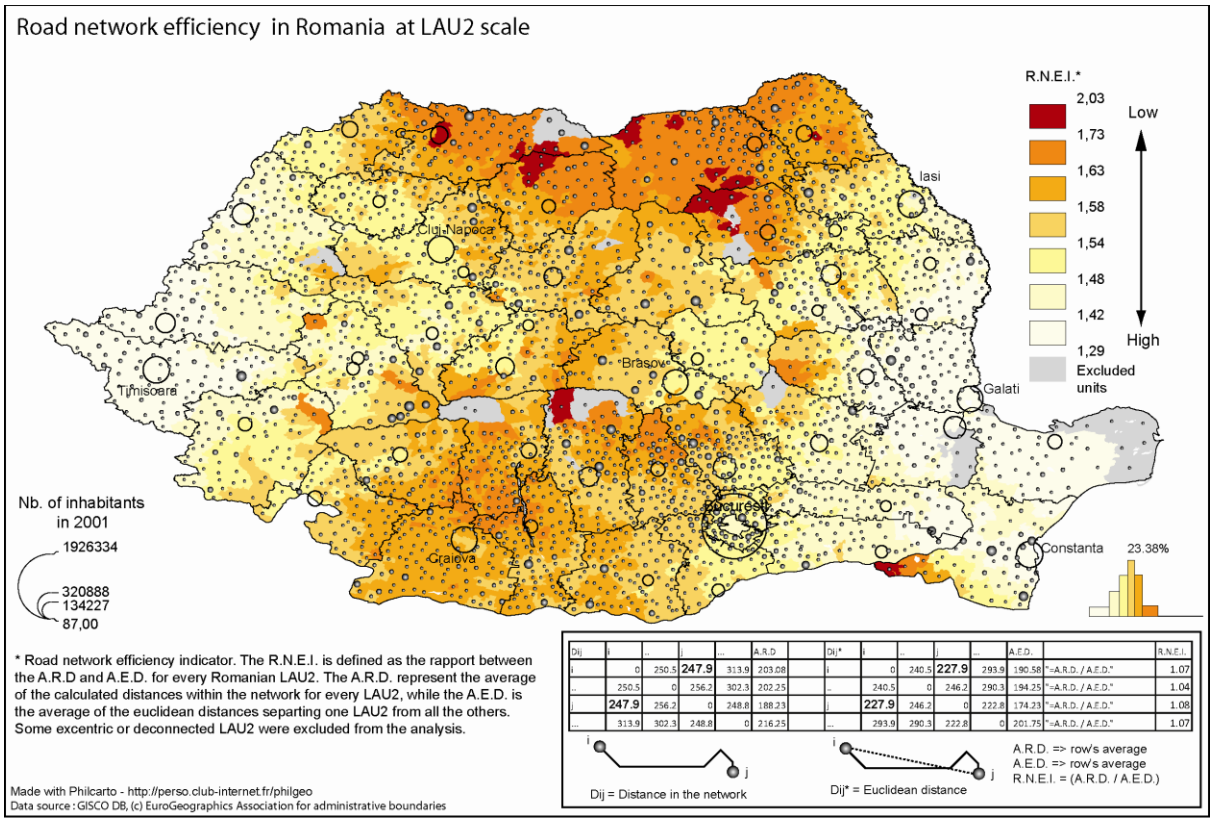


Fig. Comparing Euclidean distances and network distances in Romania

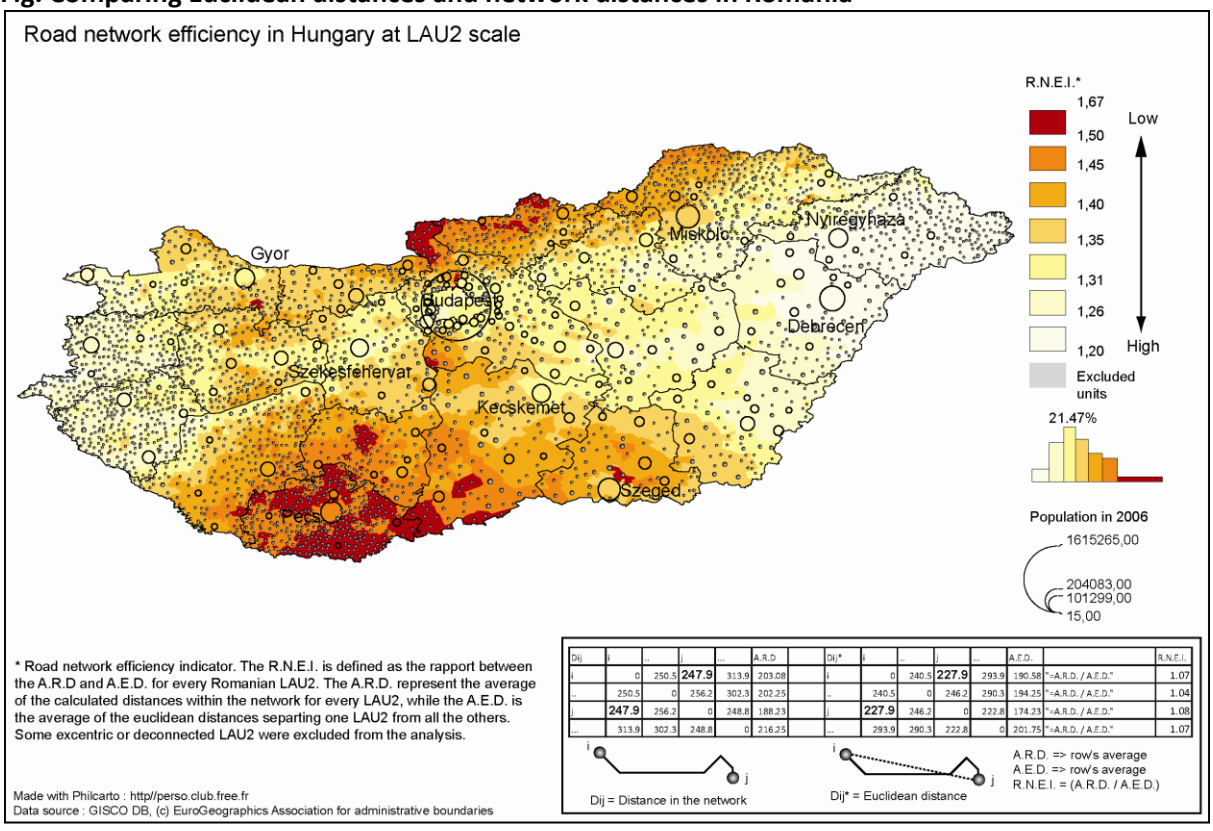


Fig. Comparing Euclidean distances and network distances in Hungary

In the absence of data, indicators like accessibility or deviation to the shortest path could become an interesting set of indicator, a tool in the exploration or description of territories at local scale. In their estimation process we mobilize the elementary spatial patterns (the points, the lines and the surfaces) and we use classical spatial analysis methods. The size of the studied region is an important aspect in the data creation. Despite the efforts, we were not able to work on a space with more than 3000 LAU2 (an approximate value) when calculating the Euclidean and the distances in the network. When we need to investigate the local accessibility for larger regions (two neighbor countries), solutions might appear if we change the working methodology and if we begin to split the table of information in data packages that don't exceed 10 millions cells.

ADVANCED

1.1 How to measure the economic performance at LAU2 scale?
1.2 How to find relevant and harmonized data for this operation? A case study on the local aggregated turnover in 2006 – the Romanian case.
1.3 How to by-pass the lack of harmonization using grid information?
1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.
1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.

2.1 Using the **distance** as an indicator for territorial coherence in the Eastern Europe.
2.2 The settlement's **hierarchy** – how many levels function from Prague to **Sulina**?
2.3 How to map the distance: choroplet vs. "oursins".
2.4 The local fragmentation of the territorial architecture and the hierarchical immobility.

Relevant findings:

The economic performance is a matter of scale and mass (demography and surface).

The past is still present in the future – the key role of the **modern industrial regions**.

Spatial anti-discontinuities: when frontiers are uniting trends of economic performance.

Some of the remote areas are quite well. The regions in difficulty are still to precise.

1.Towards more elaborate indicators and models applied to **LAU2** objects. The potential of interaction as a measure for the local economic performance.

2.Towards more elaborate indicators and models applied to **LAU2** objects. The settlement's hierarchy and the territorial architecture of the selected countries.

Definitions

Distance. In the context of our research, several distance definitions are operational. As a common approach, we will use the Euclidean distance as a weight for the potential of interaction between the spatial units (LAU2). We have also calculated the minimal path linking the LAU2 to the nearest city of more than 50 000 inhabitants in 2006, using the distance in the road network.

Hierarchy. An intuitive concept that shapes the structure of a territory as a decreasing function of the mass (demography, economic power etc.). In our case, the hierarchy is both used as a rank (LAU2 ranks according to the population, starting from 1 to n) and as imbrications of limits between different classes of demographic mass (e.g. LAU2 under 50 000 inhab., LAU2 under 128 000 inhab., more than 128 000 inhab.).

Sulina. An oriental "finisterre" of the EU on the continent, in Romania. City-port of 4983 inhab. in July 2001. Access granted by Danube and sea only.

ADVANCED 1

1.1 How to measure the economic performance at LAU2 scale?

The economic performance is an important dimension in the panel of official indicators defined in the Lisbon strategy. Generally it is measured using the GDP (per capita, per employed or the rhythm of its formation) as a relevant picture of the economic success. However, this success is depending on the scale that we intend to use when we map it. At NUTS level (0,2,3) the spatial pattern in the reparation of the GDP is a core-periphery/peripheries matter of distribution. At local level (LAU), we can assume that this concentration is still visible, if we would have access to data. Possibly, this local concentration would have as actors the capitals, the metropolitan areas and some privileged rural regions. At the opposite, the low or under-average performance might be associated to some **remote areas**, still in transition regions or un-adapted urban networks. This eventual "jeu d'echelle" of the investigation would finally show how the repartition of economic welfare or performance is a subject to the spatial frame in which we try to fit it. For an external neutral reader, these assumptions might sound as the foundations of a hypothesis. Data concerning the GDP at LAU2 level is generally absent for the 31 countries included in the ESPON space. In their relative absence, it is impossible to confirm the existence of the **trans-scalar** spatial processes in the distribution of the economic performance. There are two options in this case: either the use of an alternative measure for the economic performance, either to try to estimate the values for every single LAU2. We have explored the two possibilities, taking into account that the recommended solution should be a dominant strategy for an eventual research group working on the local databases. The first solution is to explore unofficial but reliable data sources that can offer indicators on the economic trends at local scale, knowing that harmonization with other countries will be a difficult task to assume. The second option is to approximate/estimate the values using **spatial analysis** techniques.

>>> 1.2 How to find relevant and harmonized data for this operation?

Remote areas
These territorial units could be considered only one dimension of a larger concept: the territories with specific geographical features (ESPON, 2006).
However, the specificities are not equivalent to some territorial lack of assets, especially when zooming at local scale.

Trans-scalar
A geographical attribute of the spatial distributions that intersect the MAUP (modifiable area unit problem). Basically, one territorial repartition might not have the same pattern, when we observe it at different scales of analysis.

Spatial analysis
According to the modern classifications, the spatial analysis is a method used to observe and measure the spatial structures. In the GIS, especially in the mainstream of the specific jargon, the spatial analysis is a technique used to perform topologic and logic operations between the spatial features: join operations, intersections, updates of geometries etc.

1.1 How to measure the economic performance at LAU2 scale?

1.2 How to find relevant and harmonized data for this operation? A case study on the local aggregated turnover in 2006 – the Romanian case.

1.3 How to by-pass the lack of harmonization using grid information?

1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.

1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.

1.2 How to find relevant and harmonized data for this operation? A case study on the local aggregated turnover in 2006 – the Romanian case.

In Romania, but also in other countries from the ESPON space, accessing data concerning the economic performance at local scale is a challenge. During our networking activities with the official data providers, we have found two significant facts about this type of indicator: 1) the Romanian NSI doesn't necessarily gather it; 2) other official databases, such as the Ministry of Industry, will provide data aggregated by branch at NUTS scale. If GDP is not free and officially available at LAU2, we can use other reliable data providers in order to obtain information about the economic success at local scale. One way to do it is to integrate data about the economic actors that are present in a certain territory. In the Romanian case, we have downloaded and exploited a free product (a database with more than **600 000 firms** and the basic information about their economic behavior: nb. of employees in 2006, the turnover in 2006 and the foreign direct participation). In other ESPON countries, similar products might be available. This free package of data present reliability for several reasons: the data is available for consultation and validation passing by other sources (the Ministry of Finances); the territorial picture fits the expectations and the recent trends. The values concerning the firm's turnover were integrated by matching the name of the settlements with the **SIRUTA codes** (Romanian NSI official code) for more than 12 000 spatial units that compose the 3000 LAU2 from this country. It was a challenging case study because it involved a large amount of data, specific matching algorithms and it also offered a general picture of what is happening under the LAU2 level. Practically we summarized the turnover of all the economic actors that are active in a Romanian LAU2. This approach has advantages and weakness that we will discuss next.

>>> 1.2 A case study on the local aggregated turnover in 2006 – the Romanian case.

Firms

As economic actor, the firm might have different definitions from one state to another. Their number of firms is also a subject to debate, if one will take into account the variety of purposes for which firms appears in an economic system. In the Romanian case, some of the data is an average of the values for the reported year (e.g. the number of employees in 2006). For multinational firms with antennas in the local territory, the data is furnished for the LAU2 of the headquarters and not for the production of sales compartments (e.g. if the social siege is in Bucharest and the production in Iasi, the data is attached to Bucarest).

SIRUTA code

A unique identification code proposed for all the settlements in Romania. The equivalent LAU2 code is called SIRSUP (superior SIRUTA). It is sometimes a problem for the matching operations because it is different from the codes in the basemap database.

1.2 A case study on the local aggregated turnover in 2006 – the Romanian case.

Advantages: a clear picture of the economic performance at local scale, especially when data is smoothed (different methods are possible). In this case, an archipelago-like territorial structure is shaped by a West-East gradient. The data can be used in order to multiply indicators: turnover per employee, local discontinuities or trans-scalar analysis.

Disadvantages: The data is expressed in national currency (1 RON is the approximate equivalent of 0.25 Euro) and a conversion should be made. A good knowledge of the territory should be mobilized in the map interpretation. Some of the values might look as an outlier in the territorial context – the case of Medias and Mioveni, two cities that take profit from the headquarters effect (EON GAS and Renault, two major multinational firms are located in the mentioned LAU2). The same effect could be the cause of an over-representation of Bucarest. The contextualization at an upper scale (Bulgaria and Hungary) is impossible for the moment and an eventual approach should overcome large problems of harmonization (e.g. in the case of Bulgaria, data might be available at LAU1 units).

The map (Fig.)

In 2006, the distribution of the economic performance at LAU2 scale is an aspect of the strong metropolitan concentration. Some of the regions are better situated in this equation: the west of the country or the urban network of Transylvania. Moldavia (East) and some rural peripheral regions in the South are less visible on the map. There are several reasons for this situation: gradients of economic growth, high costs of transportation or a lack of urban economic engines. This situation is also reflected in the distribution of the welfare.

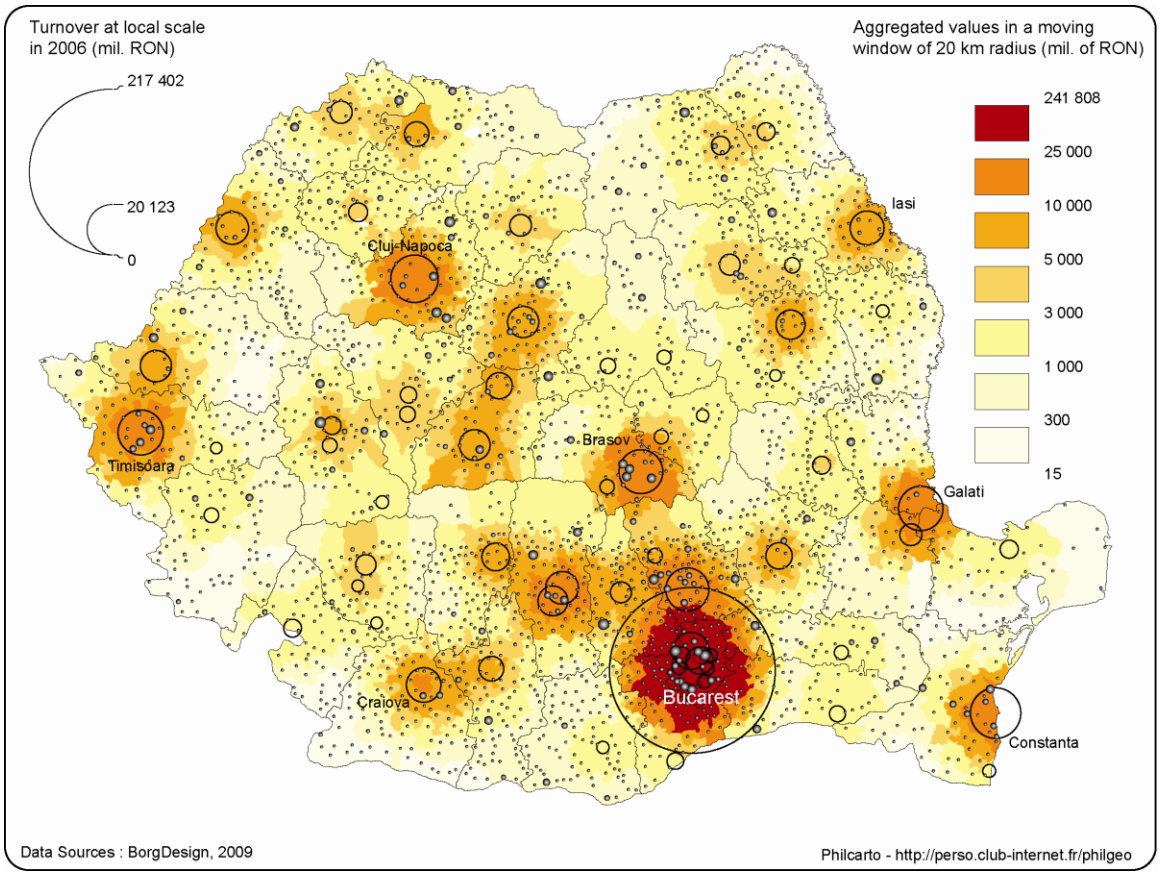


Fig. The distribution of the aggregated turnover in 2006 – Romania

1.1 How to measure the economic performance at LAU2 scale?

1.2 How to find relevant and harmonized data for this operation? A case study on the local aggregated turnover in 2006 – the Romanian case.

1.3 How to by-pass the lack of harmonization using grid information?

1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.

1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.

1.3 How to by-pass the lack of harmonization using grid information?

In ESPON DB 2013 Challenge 5 a specific methodology was developed in order to disaggregate data expressed at NUTS level in a grid (1km). The basic indicators are GDP, unemployment and active population for 2003 and 2006. At the base of the estimation, the methodology used the distribution of population in a 1km grid and the CLC 2006 classification on the built-up area. This deliverable can be used in the analysis of the economic performance at local scale in two ways: mapping the GDP 2006 in the grid or aggregating the data in the LAU2 frame. In the second case, we will obtain an estimation of the distribution of GDP at the scale of the municipalities.

1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.

Despite the number of elements in the grid that have to be manipulated, the technical and methodological operations are not as complicated as one might expect. There are at least two ways to deal with this problem. The first one is to intersect every cell of the grid with the LAU2 geometry. In that case, we will obtain a large number of new polygons that contain information (GDP). Weighting the surface of the new polygons with the values for the requested indicator will approximate the share of GDP in every polygon. However, this method is time and computer resources consuming and it involves the manipulation of many spatial objects. It also has the advantage of the best precision in the estimation (to debate). A second option consists in the by-passing of the surface weight estimation and the use of the centers of every grid cell. As a matter of fact, the 1km grid cell is just a geometrical container of the information that can be reduced to one point. Intersecting the values of the point (GDP) with the LAU2 spatial frame will allow us to estimate (by summarization) the GDP at local scale. From a certain point of view, both methods are useful, it just depends on the context in which we apply them. The first one is suitable for regional research, the second for a large amount of LAU2.

Data aggregation from the 1 km grid. 2nd method

1. Don't expect to finish very soon the aggregation for all the ESPON space. First, choose your region.
2. Create by dissolving the LAU2 limits a mask for the studied territory.
3. Extract by this mask the grid cells that present interest.
4. Create by extraction or selection the LAU2 map of the area. Make sure to have the same time reference as the indicator (e.g. 2006 layer map for GDP 2006 indicator)
5. Create centers of the grid cells. You should obtain a dot/point map. Be sure to join the information to this new spatial frame (GDP or other indicators).
6. Intersect these centers with the LAU2 map of the area. A large file containing both the codes and values of the grid and the codes and values of the LAU2 shall be created.
7. Summarize data using the LAU2 code. A new table shall be created. This table should be saved. Else, go to Step no.1.
8. Join this table to the LAU2 spatial frame and map the result.

1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.

Choosing one method or another, the specific issues of spatial matching cannot be easily overcome. One major problem is to balance the degree of generalization of the grid data representation with the geometry of the LAU2. The spatial frame for 2006 is not always perfectly overlapping the **CLC 2006** grid in which the information (GDP values) is contained. This problem is especially occurring **in the border or coastal zones**, making the estimation difficult. It is also a problem for areas with a high fragmentation of the administrative frame (the municipalities are represented as tiny polygons, intersecting 3 or 4 grid cells). The Czech Republic and Bulgaria are good examples for this last issue. Another spatial matching issue is occurring when we have to take into account the features of the natural environment. Many large LAU2 have lakes on their territory or areas with high elevation. The grid estimation method doesn't exclude them, even when their surface is limited. However, we can assume that even these areas present a contribution to the GDP creation by some economic activities taking place in relation with these zones. Sometimes the grid model penalized them (in regions with touristic vocation), sometimes not (when the natural areas present a considerable human pressure).
A zooming illustration (fig…) of these spatial matching methodological problems is presented in the cross-border region of Dobrogea, situated on the shore of the Black Sea (Romania and Bulgaria). The map indicates that the distribution of the GDP in the grid could also be the subject of a strong territorial auto-correlation effect, induced by the NUTS0 or NUTS3 regions. On other cross-border area (Romania and Hungary or Czech Republic and Slovakia) this effect is less present in the mapping result.

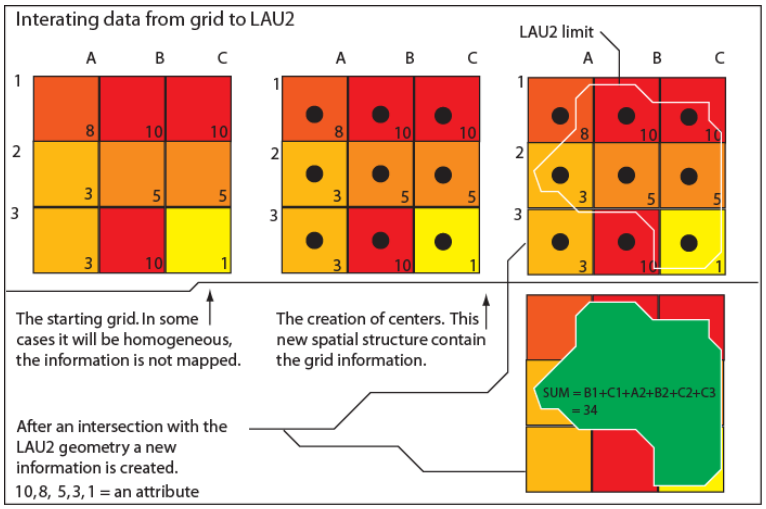


Fig. Graphic representation of grid data integration in the Lau2 geometry

Methodological solutions

In the **border or coastal zones** the data integration from grid to LAU2 is more difficult. For the border we can apply a solution based on the weighted surface of the cell intersected by the frontiers. As long as the shape of the frontier is the result of a dissolve option in a GIS, we have a good degree of shape precision. The calculated values can then be included in the file. For coastal areas a solution based on buffers may be imagined. For every LAU2 that has a limit on the sea we can create a buffer of n km and calculate the values of the indicator. We are now looking for a solution able to optimize and accelerate these two methodological options.

Land use data integration

Using the spatial analysis GIS methods, efforts were made in order to integrate CLC 2003 and 2006 data in the LAU2 geometries. This operation involved the intersection between the CLC 2003 classes and the LAU2 frame. The eventual output of this method would be a LAU2 typology based on the internal land use. Sensible to map projection and map scale, the measurement provides some errors. Taking into account the differences between the CLC 2003 classes and the national land use typologies, these errors are difficult to correct. Due to the large amount of time involved in this operation, the CLC data integration is stoped.

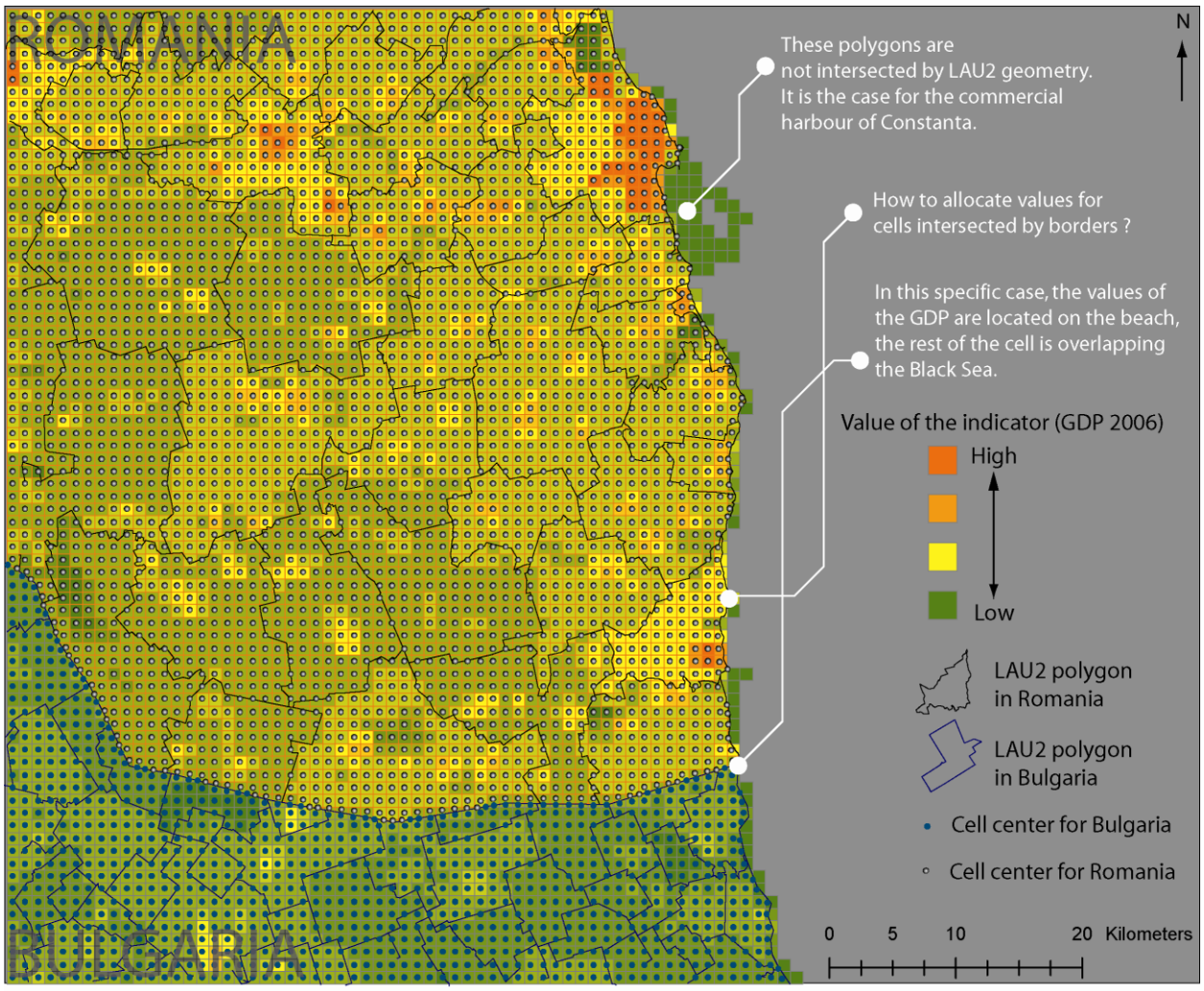


Fig. An illustration of the methodological issues induced by the grid data aggregation at LAU 2 scale

The main question behind this map is how far we can go with the estimation of values at local scale. The harbor of Constanta is serving the city itself, the NUTS 3 of Constanta, Bucarest and Romania. As a matter of fact, it is not an object that should be excluded from a larger spatial context. Its infrastructure is overlapping a nearby LAU2 (Agigea), a municipality almost completely integrated in the metropolitan area of Constanta. As in any port, a constellation of firms are located and they do contribute to the GDP creation. Some of them locate in Constanta, some of them in the LAU2 of Agigea. In the last case, firms are locating in order to take profit from the fiscal advantages of the free-tax zone active in the port. The both categories of economic actors use the port's infrastructure and depend on the metropolitan links created in the recent period. Allocate the GDP from cells to one LAU or another is an attempt to perfectly map the trees and not the forest. As a compromise conclusion, the GDP is created in the area of Constanta and not in the municipalities. This area can be defined as a potential LAU1 region (neither LAU2, neither NUTS3, neither FUA, but relevant for a better integration of the local data) or it can be estimated by quantitative modeling. In Romania and some other ESPON states the LAU1 administrative level of data collection is absent. The only remaining option is to "smooth" the data by a potential of interaction model.

>>> **1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.**

1.1 How to measure the economic performance at LAU2 scale?

1.2 1.2How to find relevant and harmonized data for this operation? A case study on the local aggregated turnover in 2006 – the Romanian case.

1.3 How to by-pass the lack of harmonization using grid information?

1.4 Integrating disaggregated data (GDP at NUTS 3 level expressed in a grid of 1km) on the LAU2 geometry frame – a problem of spatial matching.

1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.

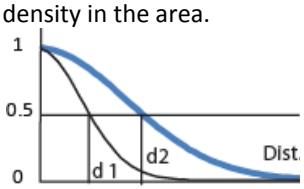
1.5 The calculation of the potential model of interaction for the local estimated GDP. New insights on the model's moving window and the distance decay.

The integration of the grid data in the LAU2 frame was performed for 5 selected countries: Czech Republic, Slovakia, Hungary, Romania and Bulgaria. The operation involved more than 20 000 LAU2 and this considerable amount of spatial units is caused by the high degree of administrative fragmentation in 4 of the 5 states (if we also take into account the western part of Hungary). As we have integrated the 2006 values of the indicators in the 2006 LAU2 geometry, some corrections were necessary because all the capitals (except for Sofia and Prague) and some major Slovakian cities (Kosice) are divided in spatial units with LAU2 administrative competences. In the case of Bucarest, we have 6 sectors. Consequently, the data was summarized according to the most central division of the capitals, usually the sector no.1 or the Staro Mesto (the old center). A second step in the data exploration was to map the result in order to verify if extraordinary values or errors are interfering with the methodology that we used. This step works like an "expert opinion" validation of the collected or integrated data, but it is based on the visual survey and it has obvious limits. The mapping options are basically limited – symbols or choropleth design. In both cases the mass effect will be present on the map and by mass effect we understand the high distance that separates the capitals and the large cities from the rest of the urban or rural entities included in the settlement's hierarchy. A better visualisation will occur if the data is smoothed (GDP in 2006 at local scale, unemployment or active population) and this better visualisation is needed if one will think that we are working with the local data and not with the cities. Smoothing the data also involves a choice to be made between the methods: the average values within a rank 1 neighborhood, the summarized values in a cut-off distance neighborhood or a model based on the potential of interaction of all the LAU2 units. Some of the options take space into account in a discrete way, the last one (the model of spatial potential) in a continuous manner.

>>> 1.5 **The calculation of the potential model of interaction for the local estimated GDP**

Building the model

After the inclusion of the grid data in the LAU2 frame, we have prepared a model of potential of interaction that usually works when the values of the parameters are well approximated. The three variables that we had to take into account are: the interaction function, the distance decay function and the mass (the GDP in 2006 estimated for all the LAU2 in the 5 selected countries). In a classical model the interaction function and the distance decay are based on some constant values. In our model the two variables are receiving multiple values, being weighted with the demographic rank and with de road network density in the area.



d1. Radius of a certain distance for a LAU2

d.2 Radius of a certain distance for a LAU2 with a larger mass.

The black and the blue lines simulate the decrease of the interaction as a distance function.

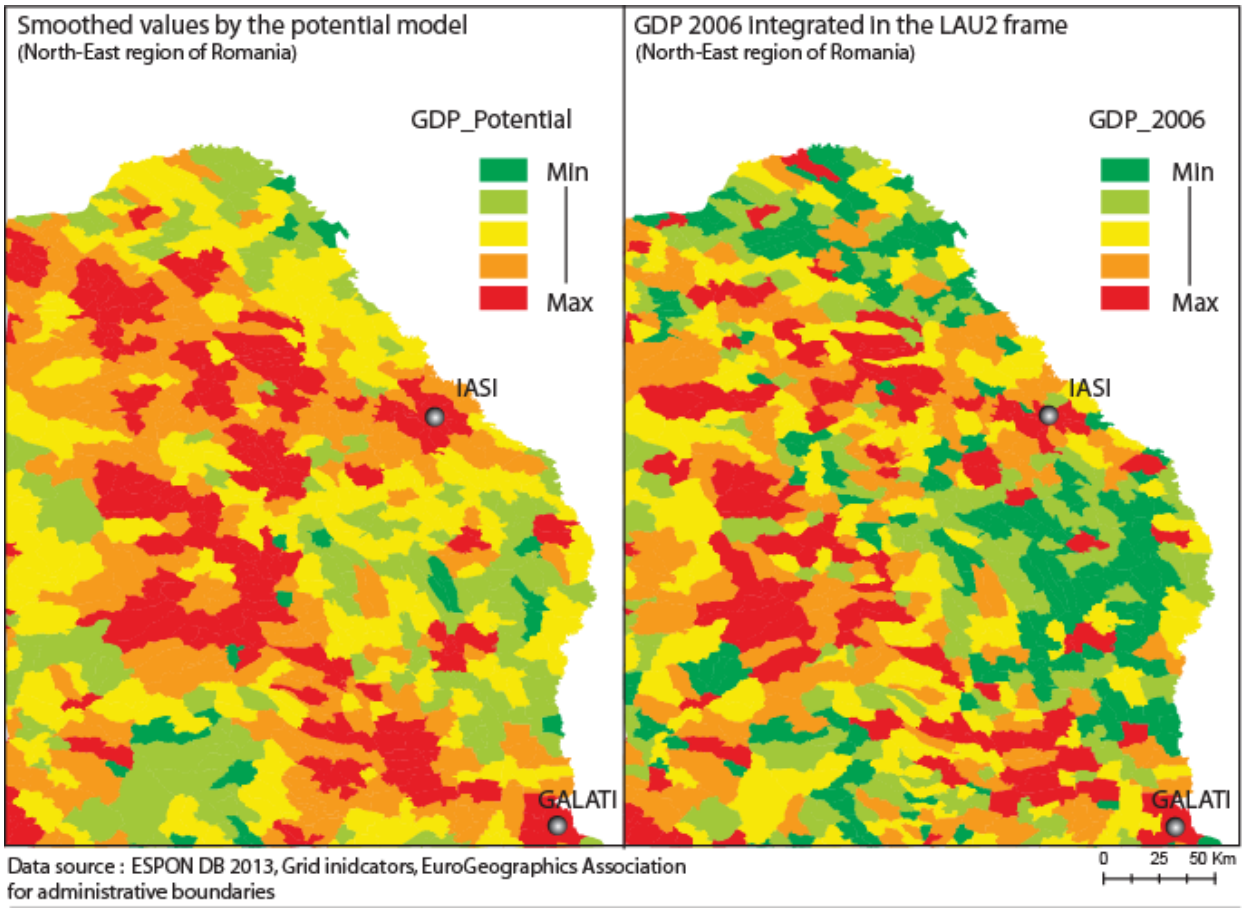


Fig. Smoothed values for GDP_2006 at LAU2 scale vs. the estimated values of GDP_2006

1.5 **The calculation of the potential model of interaction for the local estimated GDP**

In a classical model of potential interaction, we assume that at a "certain distance" from a spatial unit the interaction decreases at 50 % (if we use a Gaussian negative function). We also take into account the friction of the space by using a distance decay function with 2 as exponent (the canonical 2). When we apply the model we can calculate for every single spatial unit how much "interaction" will receive, if our assumptions approximate well the reality. We can better translate this into English by using an example. Let's assume that we want to calculate the potential interactions for Prague. First, we will assume that all the interactions are reduced to half at 20 km. This means that 50 % of the flows are collected in a circle with 20 km radius. The other half, obviously beyond 20 km. We will also take into account the friction induce by space (we will use 2 as value, not too much friction according to the classical models). If one will try to validate this model for Prague using the commuters flows as an empirical validation, he will observe that we were wrong in our assumptions. Prague collects 50 % of its flows from a 45 km radius, while the distance friction is only 1.2. Unfortunately, for the GDP 2006 potential of interaction we cannot validate our assumptions because we don't have the local economical flows, but our model is also supposed to work as a smoothing method for the data. If we reproduce the calculations for Brno, we will observe that the radius is about 25 km and the distance decay has 1.8 as value. One important aspect induced by the empirical "validation" is that we cannot use constant values as parameters.

Our model was built using a variable radius for the interaction function and a variable distance decay parameter. The variability was mathematically induced by using the demographic rank as a weight for the radius. The capitals received 20 km radius (in a Gaussian function) while the other cities values vary between 15 and 1 km, according to their rank. The distance decay was calculated using the road network density (high density = low friction (1.7), low density = high friction (2.3)). The road network density was estimated by intersecting the reticular spatial structure of roads with the LAU2 geometry. After this preliminary "mise en scene" of the parameters, the model was applied for more than 15 000 LAU2. Bulgaria was excluded in the absence of a network reliable file.

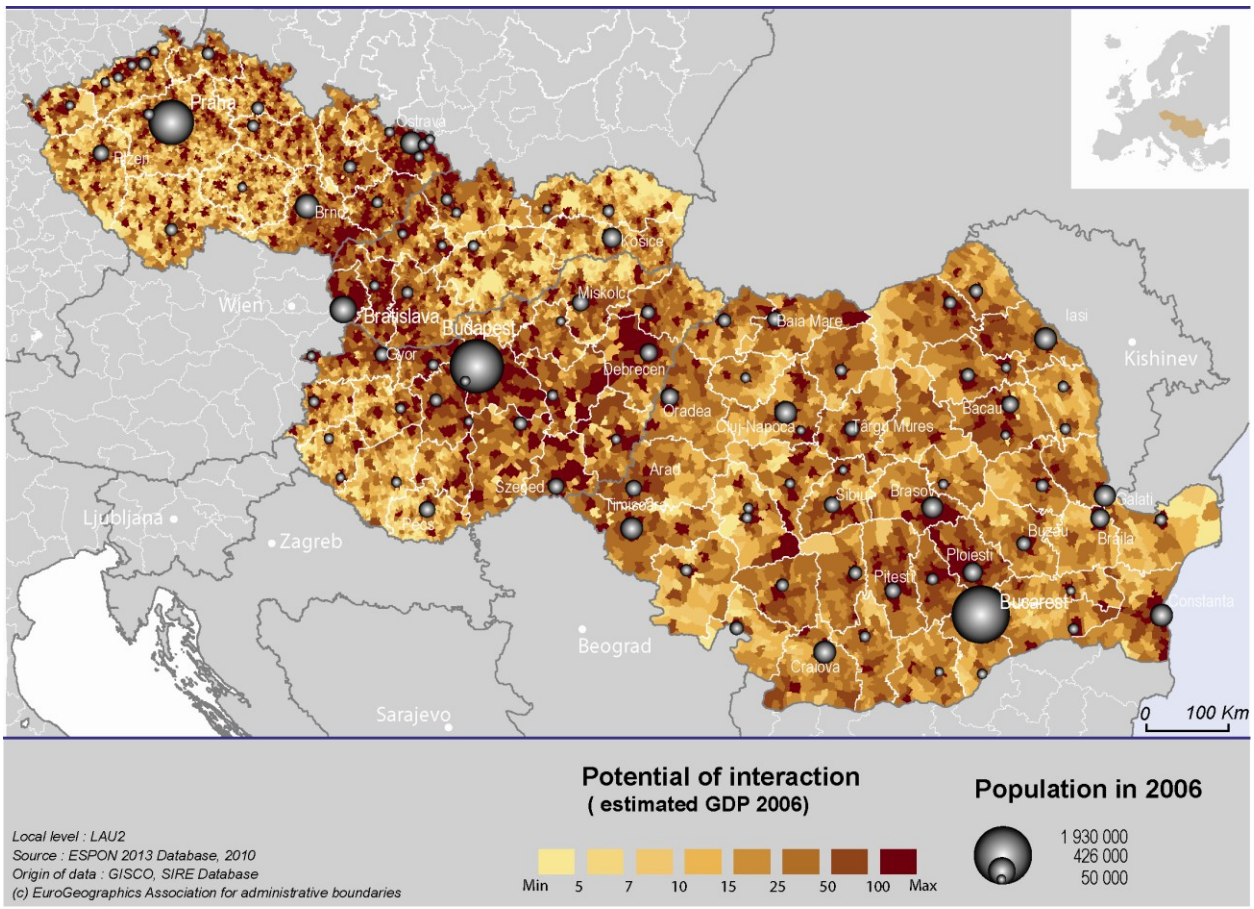


Fig. The potential of interaction for the GDP estimated at local scale in 2006

With this approach the smoothed values present more interest than the basic GDP disaggregated at LAU2 scale. Our intention was to eliminate the noise from the map and leave intact a territorial structure that also allows seeing the local (fig. no). We preferred to use a spreadsheet for the calculation rather than a GIS model, because the methodology can easily be reproduced by interested users. Limited by the hardware, we were forced to implement the model in packages of 15 000 LAU2 x 500 LAU2 (until we finished them all), but other strategies are possible too. A second intervention in the data was necessary, this time in order to allow the mapping process – the values were standardized using the limit of the second hierarchical class (100), letting all the superior values floating to the maximum (Budapest). There are two ways to interpret the cartographic result. A first strategy will try to seek for the outliers and the extraordinary values and the second one involves the mobilization of the spatial structure concept.

Key findings from the map

The economic performance is a matter of scale and mass (demography and surface).

After modeling the distribution of the GDP 2006 at local scale we can observe that the spatial pattern of this repartition is influenced by the size of the LAU2. In Hungary and Romania this phenomena is clearly visible. At the opposite, areas with high administrative fragmentation (in Slovakia or in Czech Republic) seem to be penalized even after the data smoothing. This regularity is less present in the East of the Czech Republic and in the northern region of Bucharest, two zones specialized in industrial activities.

The past is still present in the future – the key role of the modern industrial regions

After the transition period, the industrial regions (some of them old, some of them emerged or modernized during the socialist period) seem to regain a comfortable position in the GDP hierarchy. At local scale, these industrial basins are extremely important if one will take into consideration their impact on employment or in the welfare creation. These regions (e.g. Ostrava in Czech Republic, Gyor in Hungary, Pitesti-Ploiesti in Romania) are there to complete the metropolitan economic nodes and their conjuncture fragility is balanced by resilience, adaptation and integration in the European economy.

Spatial anti-discontinuities: when frontiers are uniting trends of economic performance.

Transforming the frontiers in interfaces that filter the flows of persons, goods and information is a constant trend in the Eastern European countries. In some cases, these frontiers may also work as attractors for the economic activities, such is the case between Romania and Hungary or partially between Slovakia and Hungary. The real discontinuities in the local estimated GDP 2006 distribution seem to be internal, sometimes overlapping old historical limits (Moldavia and Transylvania for Romania). Maybe the new economic paradigm that installed in the transition and pre-adhesion period reactivated these old frontiers, shaping new logics of economic performance compared to the past.

Some of the remote areas are quite well. The regions in difficulty are still precise.

Without being a rule, some of the remote regions and areas with specific geographical features are not marginalized in the distribution of the economic performance (the coastal regions are quite dynamic, some of the mountain areas in the Carpathians present decent values of GDP due to recent turistification and re-industrialization and some of the border regions behave as economic attractors). The regions in difficulty locate (without being a clear regularity) in the “no man’s land” of the metropolitan and urban polarization, making us to assume that the economic performance could be a distance decay function towards the nearest city.

>>> 2.1 Using the distance as an indicator for territorial coherence in the Eastern Europe

Advanced 2

2.1 **Using the distance as an indicator for territorial coherence in the Eastern Europe.**
2.2 **The settlement's hierarchy – how many levels function from Prague to Sulina?**
2.3 How to map the distance: choropleth vs. "oursins".
2.4 The local fragmentation of the territorial architecture and the hierarchical immobility.

2.1 Using the distance as an indicator for territorial coherence in the Eastern Europe

Distance is still an important component that shapes the local space in the selected countries (it filters the flows, it may explain how the economic performance is distributed or how the urban network is functioning). Working with the distances at LAU2 scale is complicated if we take into account the large number of spatial units. However, not all the distances are relevant or interesting in an eventual study. If we want to compare how the estimated GDP at local scale in 2006 is distributed in relation with distance to the nearest city, we will work with a reasonable quantity of elements in the matrix (about 15000 LAU2 x almost 100 spatial units). In this case we can proceed to the calculation of distances using the regional road network. If no GIS instrument is available, the Euclidean distances can also be estimated using the classical model. Both methodologies involve a number of compulsory steps to check, in order to obtain the "oursin" map that we seek for.

2.2 The settlement's hierarchy – how many levels function from Prague to Sulina?

As our intention is to put in relation LAU2 s and cities in the selected countries (Czech Republic, Slovakia, Hungary and Romania) and as the definitions of the cities are heterogeneous, we have assumed that the starting point should be the choice of a demographic cut-off in the urban hierarchy. The analysis of the rank-size distribution for 2006 population shows that 3 hierarchical levels are clearly visible in the region: the capitals, the so called "large cities" (over 128 000 inhab.) and the medium sized cities (over 50 000 inhab.). As an alternative, we could use the distinction between FUA and MEGA, qualitatively complicating the model. These three demographic levels were put into relation with the LAU2 using the distance in the road network as an indicator for some potential and theoretical urban influence area. If there is any association between the economic performance and the distance decay based on cities, the mapping process and the interpretation should show it. We also could assume that this relation will work better in a homogeneous space, like Hungary. That's the reason for focusing our cartographic processing and argumentation on this country. Of course, if we could make the map readable.

>>> 2.3 How to map the distance: choropleth vs. "oursins".

Steps to obtain a link map

1. Every map comes from software. From the point of view of the software, not every sum of lines is a network. Manipulation is needed in order to obtain nodes and links in the network.
2. In the case of an oursin map, not all the points (LAU2, cities, features) shall be involved. The eventual user should think that what he will obtain will be an origin vs. destination asymmetrical matrix.
3. The origin vs. destination matrix will be provided in a vector format which is sometimes difficult to manage.
4. In order to obtain the minimal distance that separates a LAU2 unit from the cities (our case) an interrogation in the matrix is necessary.
5. After the interrogation is applied, a new vector matrix is obtained and it will be saved. Else, go to step no.1.
6. The data obtained from the calculation offers two mapping options: links or choropleth. The links can be used to emphasize the shape of a theoretical zone of urban influence, the choropleth its limits.

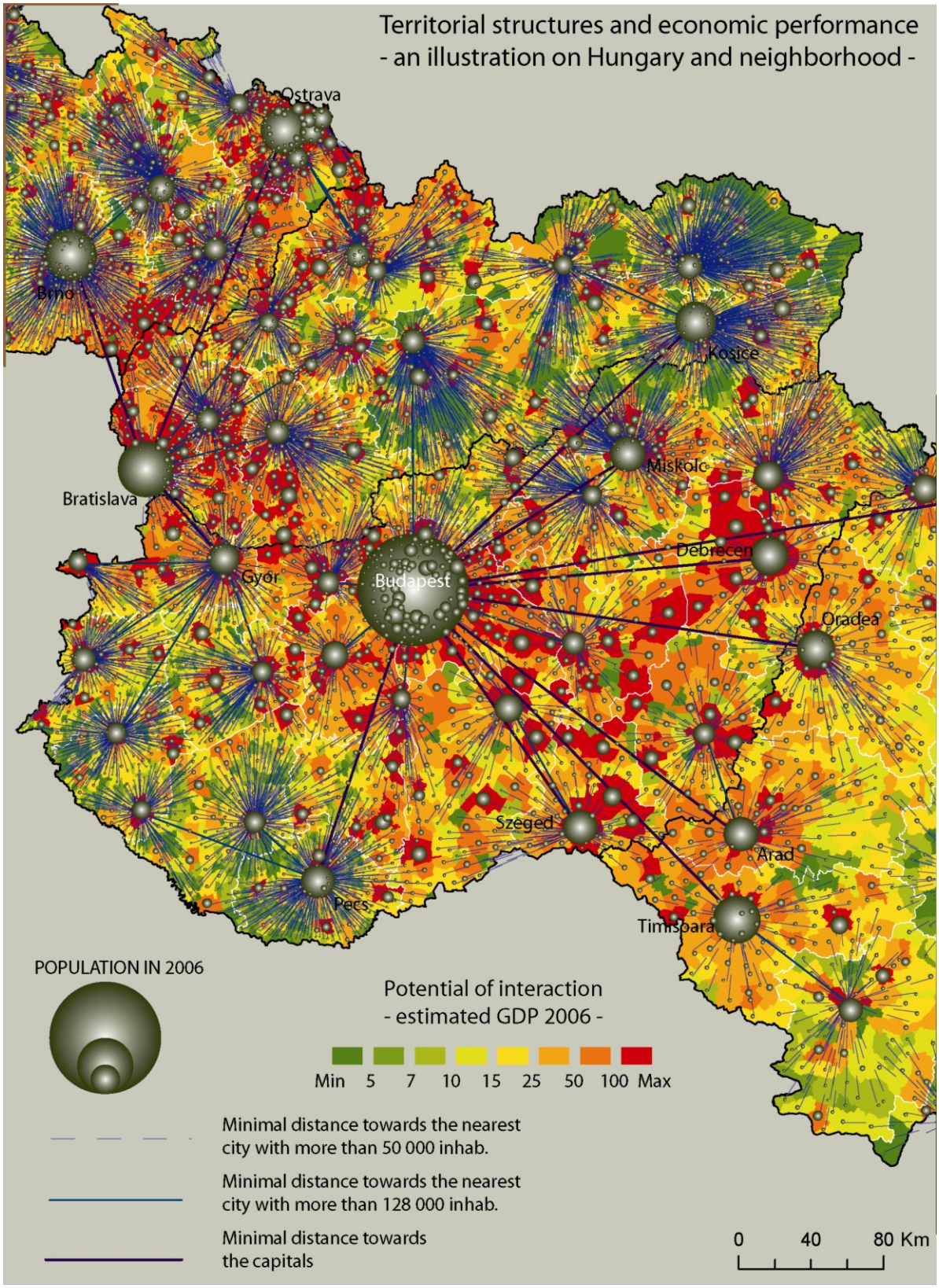


Fig. Distances and economic performance in Hungary and neighborhood

2.3 How to map the distance: choropleth vs. "oursins".

If one will use the minimal distances as an indicator for the territorial or administrative fragmentation at the local scale, the mapping method will be an issue to take into consideration. Using the LAU2 polygons for the cartographic product will produce a "tropical fish" like map. At the opposite, the link map will provide an agglomerated picture of the territorial structures, emphasizing the shapes rather then the limits of the theoretical zones of influence. The main advantage of the second map type (links or "oursin") resides in the opportunities that offer to hierarchically imbricate different distance levels. (e.g. LAU2 vs. nearest city with more than 50 000 inhab., nearest city with more than 50 000 inhab. with the nearest city with more than 128 000 inhab., the nearest city with more than 128 000 inhab. with the capitals). In a choropleth map only one kind of distance could be mapped, excepting the case where we use a cluster analysis. If the LAU2 used as destinations are too dense, there is a risk to make the map unreadable (which is the case for Slovakia or Czech Republic, in our illustration).

2.4 The local fragmentation of the territorial architecture and the hierarchical immobility.

The relation between the economic performance and the distance towards the nearest medium (50 000 inhab.) or large city (128 000 inhab.) seems to obey to a U shaped like function rather than a classical power decay function (based on the observations made on Hungary and neighborhood). This output has two explanations: the christallerian pattern in the city distribution or errors induced by the aggregation method (GDP for 2006 in a grid format to GDP 2006 in LAU2 frame, smoothed by a potential method).
In many local cases, the shape of the theoretical urban influence area is largely overlapping the NUTS 3 limits. It is not the NUTS3 limit that should be put into question, but the limited number of destinations used by the model.
The national borders are irrelevant in the design of the capital's theoretical areas of influence. Some western cities in Romania are closer to Budapest, than Bucharest. Their local hinterland too, inducing some sensible questions regarding the equilibrium between national and trans-national public planning policies.
The context does matter and by context we can imagine the role that would play MEGA such as Vienna in the context of the Eastern and Central Europe. How the potential interaction of the GDP 2006 (locally estimated) would be reshaped, if Vienna were on the map?
In the eventuality of a recalibration of the hierarchical demographic levels (using 27 000 inhab. instead of 50 000), the relation between economic performance and distance towards the nearest city will take (maybe) another form. However, at the top of the hierarchy, the situation will present few changes and Oradea will still be closer to Budapest than to Bucharest. The hierarchical immobility will still function, it is the local that will present interesting dynamics.
The corridors of welfare are complicating the gradients and the core(s)-peripheral spatial patterns present on the map. Some of these corridors reply the major transportation network, some others the linear proximity to economic engines or consecrated MEGA. The disconnections in relation with the metropolitan nodes sometimes give them the attribute of fragility or territorial tunnel effect.

TECHNICAL REPORT

ESPON

Naming UMZ: a data base now operational for urban studies

MAIN RESULTS

- From physical zones to urban settlements. Currently, the UMZ (CLC 2000, EEA) are not described by a name. This attribute is yet essential for creating a semantic link to the territory.
- Automatic algorithm. According to the way an UMZ overlaps reference units (LAU or other), a new method attributes automatically one or several names.
- Validation. A final check is done automatically, by comparing results to other European database names. Some particular cases are corrected by expertise.
- A first thematic insight. An exploration of the main features is proposed in the last part. New results are given concerning the European city size distribution, the general and regional density patterns, and the main characteristics of international UMZ.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

LIST OF AUTHORS

Anne Bretagnolle, University Paris 1, UMR Géographie-cités

Marianne Guérois, University Paris 7, UMR Géographie-cités

Guilhain Averlant, UMR Géographie-cités

Hélène Mathian, C.N.R.S., UMR Géographie-cités

François Delisle, UMR Géographie-cités

Liliane Lizzi, C.N.R.S., UMR Géographie-cités

Timothée Giraud, UMR Géographie-cités, UMS 2414 Riate

Contact

anne.bretagnolle@parisgeo.cnrs.fr

tel. + 33 1 01 40 46 40 00

TABLE OF CONTENT

1 Stakes and matter

Urban Morphological Zones have been created in 2004 by the European Environment Agency. This data base forms a perspective for the future, for three main reasons: it is constructed using highly automated methods, it is regularly updated (two dates are now available for UMZ perimeters, 1990 and 2000 and the 2006 version will be soon available) and it is fully documented.

This database has however not been widely used to date in urban studies, mainly because it is not operational: the objects are simply spots or patches, without names, and hence without semantic links with the territory. They only constitute a set of geometrical objects, and not of geographical objects. The general aim of this Technical Report is to describe the automatic methods and expertises that have been used for naming UMZ and getting them usable for a first exploration of the European urban settlements.

1.1 Presentation of UMZ

UMZ have been created in order to analyze "the extent of urban land-take in Europe, where sprawl happens and how it is shaped" (*EEA activities*, http://www.eea.europa.eu/themes/urban/eea-activities). An UMZ can be described as "a set of urban zones, defined from land cover classes contributing to the urban tissue and function", forming a continuous built-up area (i.e. laying less than 200 m. apart)[1].

Since September 2009, the geographical coverage of the UMZ 2000 database is the following one:

- the 27 countries of the European Union

- 5 countries in the Balkan region (Albania, Bosnia-Herzegovina, Kosovo, Macedonia and Serbia)

- Norway, Lichtenstein and Island[2].

The UMZ dataset can be downloaded freely on EEA website[3].Different attributes are available:

- Identification code (not the same than for UMZ 1990)

- Population (estimated from JRC's Population density grid, see Javier Gallego, *Joint Research Center*)[4].

- Area and perimeter

[1] Urban Morphological Zones 2000 Version F1v0. Definition and procedural steps, Roger Milego, February 2007, http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=995.

[2] CLC2006 should also cover Switzerland.

[3] http://www.eea.europa.eu/data-and-maps/figures/urban-morphological-zones-umz-2000

[4] For further details, see Downscaling population density in the European Union with a land cover map and a point survey, http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2.

1.2 From physical zones to urban settlements?

Going from physical zones to urban settlements is not a trivial operation. For example, overlaying on a GIS the UMZ and Google Earth or LAU 2 names is not sufficient: if we make a zoom on Berlin surroundings, it will be easy to put the "Berlin" name on the UMZ whose centroid is the closest to the historic center, but several difficult questions remain:

- Should other close UMZ inside the same LAU 2 receive also the name "Berlin" or another one (for instance, more local names given by Google Earth)?

- What are the reference units for choosing the right names? In most of the cases, city names fit with LAU 2 names (for instance in France, Germany or Belgium, as the eponym name of city fits with the central municipality). But in some other cases, like in Portugal, Greece or Denmark, city names fit with LAU1 names. And it is even more complicated in United Kingdom or Ireland, where city names don't fit with one only administrative level but with other administrative entities.

- How can we manage the case of polycentric cities, like industrial or littoral conurbations? Should they receive several names, for instance when the population is well distributed among the different cores, or just one name?

- What about large cities cases, which are extending now at the scale of one NUTS 3 rather than LAU 1 or 2? Should we give them the name of the eponymous LAU 2 or the name of the region that currently fits with their spatial coverage?

- How ensuring a quick update of UMZ names, facing the evolution of perimeters (corrections or new dates), the evolution of population density grid (JRC), or the need to apply the methods to smaller objects (in the current data base, names are given only to UMZ larger than 10 000 inhabitants, i.e. less than 50 % of the total number of UMZ)?

The answers given to these different questions are discussed and fully described and illustrated in the following sections of the Technical Report.

1.3 A new version of UMZ data base

Different adjustments have been made to UMZ database in order to facilitate its use by ESPON partners. UMZ larger than 10 000 inhabitants have been considered (a total of 4437 UMZ).

- Updated Population: using automatic methods, we have updated the population of all the UMZ with the last version (v.5) of the Population density grid built in 2007 by *Joint Research Center*[5]. The scale used for this grid is 100x100 meters.

- New indicators: Name(s), Centroid[6], Density (inh./km^2), Country[7], International code (number of countries crossed by the UMZ), International index (% of population not living in the main country).

[5] Gallego J., 2007; Downscaling population density in the European Union with a land cover map and a point survey, http://dataservice.eea.europa.eu/dataservice.

2 Naming methodology

2.1 Automatic algorithms

2.1.1 General presentation

The methodology that has been chosen is largely inspired by the one used by French Census Board (INSEE) to give names to French urban areas (*unités urbaines*)[8]. Rules and criteria have been elaborated to differentiate three types of spatial configurations resulting from the overlap of UMZ data base, Population density grid and the reference units data base (i.e. the data base that has been selected for giving the names, for example LAU 2) (Figure **1**).

In the first situation, the major part of the UMZ population (more than 50%) is located inside one reference unit[9]. The urban settlement extends rather clearly around one morphological centre, and receives one name.

In the second and third situations, no reference unit concentrates more than 50 % of the UMZ population: we retain therefore the unit that has the major contribution as the main one, then we examine the other reference units that largely contribute to the UMZ population. If they represent more than 50% of the main reference unit contribution, we retain them and the UMZ is considered as "UMZ with several cores" (Situation 2). If not, we keep only the main reference unit for naming UMZ. It is then considered as "UMZ with a weak core" (situation 3).

- *Situation 1 : UMZ with a strong core* (it receives one name)
- *Situation 2: UMZ with several cores* (it receives several names)
- *Situation 3: UMZ with a weak core* (it receives one name)

2.1.2 Algorithm steps and illustrations

In order to simplify the presentation of this sub-section, the selected reference source for city names is LAU 2.

---

[6] The centroid is the centre of gravity computed as the average of the coordinates of all the UMZ's vertices.

[7] If the UMZ overlays more than one country, it is associated to the country which includes the largest part of the UMZ population (main country)

[8] « Composition communale des unités urbaines, Population et délimitation 1999, Nomenclatures et codes » ; INSEE, mars 1999.

[9] We have retained, like INSEE, the minimal threshold of 50% inhabitants, which gives rather goods results (see validation section below).

The methodology can be presented as a succession of steps or algorithms. Each step involves automatic calculations.

2.1.2.1 Geometrical and statistical sources

Three different types of objects are overlaid:

- UMZ 2000[10]

- Population density grid from JRC (version 5)[11]

- Local administrative units (LAU 2, EuroBoundaryMap 2006 v2.0 from EuroGeographics, validity: 2006).

2.1.2.2 Computation steps

We compute the population intersecting LAU and UMZ and we observe the maximal value for each LAU related to one UMZ.

Let ***L*** describe the LAU and ***u*** the UMZ. After the intersection, let ***L(u)*** be the part of the LAU ***L*** intersecting the UMZ ***u***, *and* $P_{L(u)}$ be the population of this part, when P_u. is the population of the whole UMZ.

$$P_u = \Sigma \{ P_{L(u)}, L \text{ intersecting } u \}$$

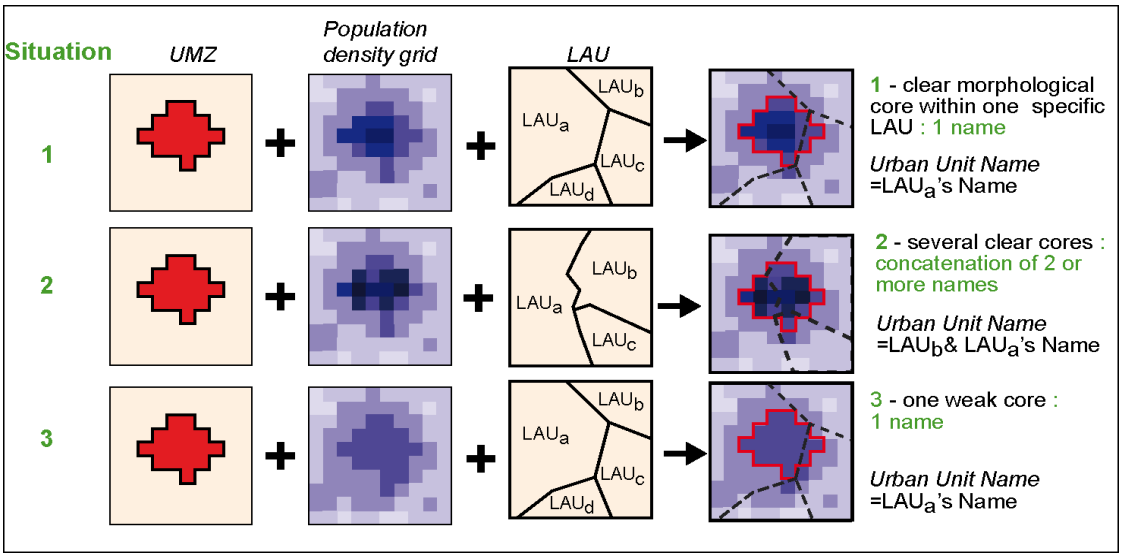
Thus for each UMZ ***u*** , the series of { $P_{L(u)}$, ***L*** intersecting ***u***} is considered, ranked by decreasing order, and let then $L_i(u)$ be the i^{th} part in this ordered series.

For a given UMZ ***u***, three different situations can occur (Figure 1).

[10] Latest version given by the European Topic Center on Land Use and Spatial Information (ETCLUSI) in June 2010, which should be available in the future EEA dataserver. Official distribution: http://www.eea.europa.eu/data-and-maps/figures/urban-morphological-zones-umz-2000.

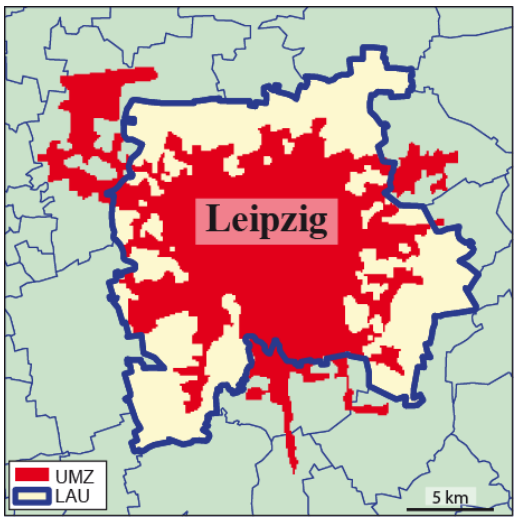
[11] http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2/population-density-grid-geotiff-format.

Figure 1 : Naming methodology (Situation 1, 2 and 3)



SITUATION 1: The largest population of the LAUs intersections is more than 50% of the UMZ's population. We have an UMZ with one strong core, clearly organized around one center. The UMZ is named with the name of this LAU_a (Figure 1). This is the case of Leipzig example (Figure 2).

Figure 2 : Leipzig (Germany), an UMZ with one strong core (Situation 1)



UMZ population: 536 552 inh.
UMZ population in Leipzig LAU 2: 483 285 inh.

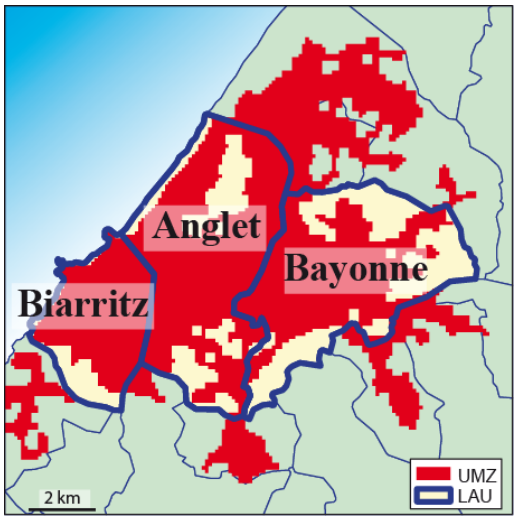
Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

SITUATION 2 and 3: There is not a unique main core as defined above, thus the larger part is retained as the reference, and the other parts are considered successively, in decreasing order of population, as long as their populations exceed 50% of the first part population.

$$Secondary\ units=\{\boldsymbol{L_j(u)} / \boldsymbol{P_{Lj(u)}} >= 0.5 * \boldsymbol{P_{L1(u)}}\}$$

Situation 2 : one or several secondary units' population represent more than 50% of the population of the largest part. We retain the name of the concerned secondary units, and the final name of the UMZ is a compounded name. The order of the names is not alphabetical but follows the decreasing order of population contributions to UMZ. This is the case of Bayonne-Anglet-Biarritz (Figure 3).

Figure 3 : Bayonne-Anglet-Biarritz (France), an UMZ with several cores (Situation 2)

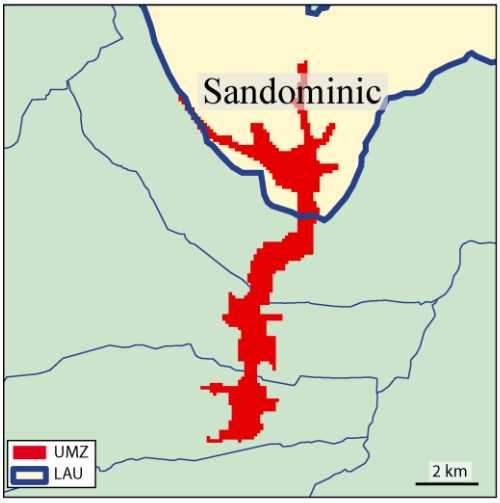


UMZ population: 128 554 inh.
Bayonne LAU 2 population inside UMZ: 39 708inh.
Anglet LAU 2 population inside UMZ: 35 185 inh.
Biarritz LAU 2 population inside UMZ: 30 156 inh.
Other LAU 2 population inside UMZ < 12 000 inh.

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

Situation 3: no secondary unit's population represents more than 50% of the larger part. We retain finally only the name of the main LAU unit, fitting again with a "one core" context (one morphological core, but less strong than in Situation 1) (Figure 4).

Figure 4: Sandominic (Romania), an UMZ with one weak core (situation 3)



UMZ population: 10 678 inh.
Sandominic LAU 2 population inside UMZ: 4 893 (46% of total UMZ inhabitants)
Other LAU 2 population inside UMZ < 2446 inh.

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

The different steps of the algorithm can be summarized by:

IF $P_{L1(u)} >= 0.5 * P_u$ THEN Name(u)=Name($L_{1(u)}$)

ELSE IF *Secondary units*={$L_j(u)$ / $P_{Lj(u)} >= 0.5 * P_{L1(u)}$} $\neq \emptyset$

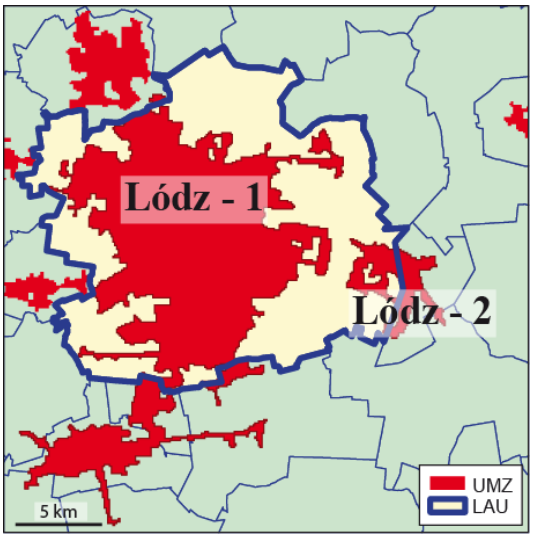
THEN Name(u)=Name($L_{1(u)}$)+{Name($L_{j(u)}$), j / $P_{Lj(u)} >= 0.5 * P_{L1(u)}$}

ELSE Name(u)=Name($L_{1(u)}$)

2.1.2.3 A particular case: different UMZ with identical names

In the large majority of cases, each UMZ could be associated to a unique name (situations 1 and 3) or to a unique combination of names (situation 2). However, in about 10% of the cases, several UMZ share the same administrative unit. It is for instance what happens in the case of the city of Lódz in Poland (Figure 5): the most populated parts of two different UMZ fall into the same LAU2, so that they both receive exactly the same name. In order to maintain the attribution of distinct identifiers for UMZ, we add a number after the name, according to the decreasing size of UMZ populations (Lódz - 1 and Lódz - 2).

Figure 5 : Lódz (Poland), two UMZ with the same name

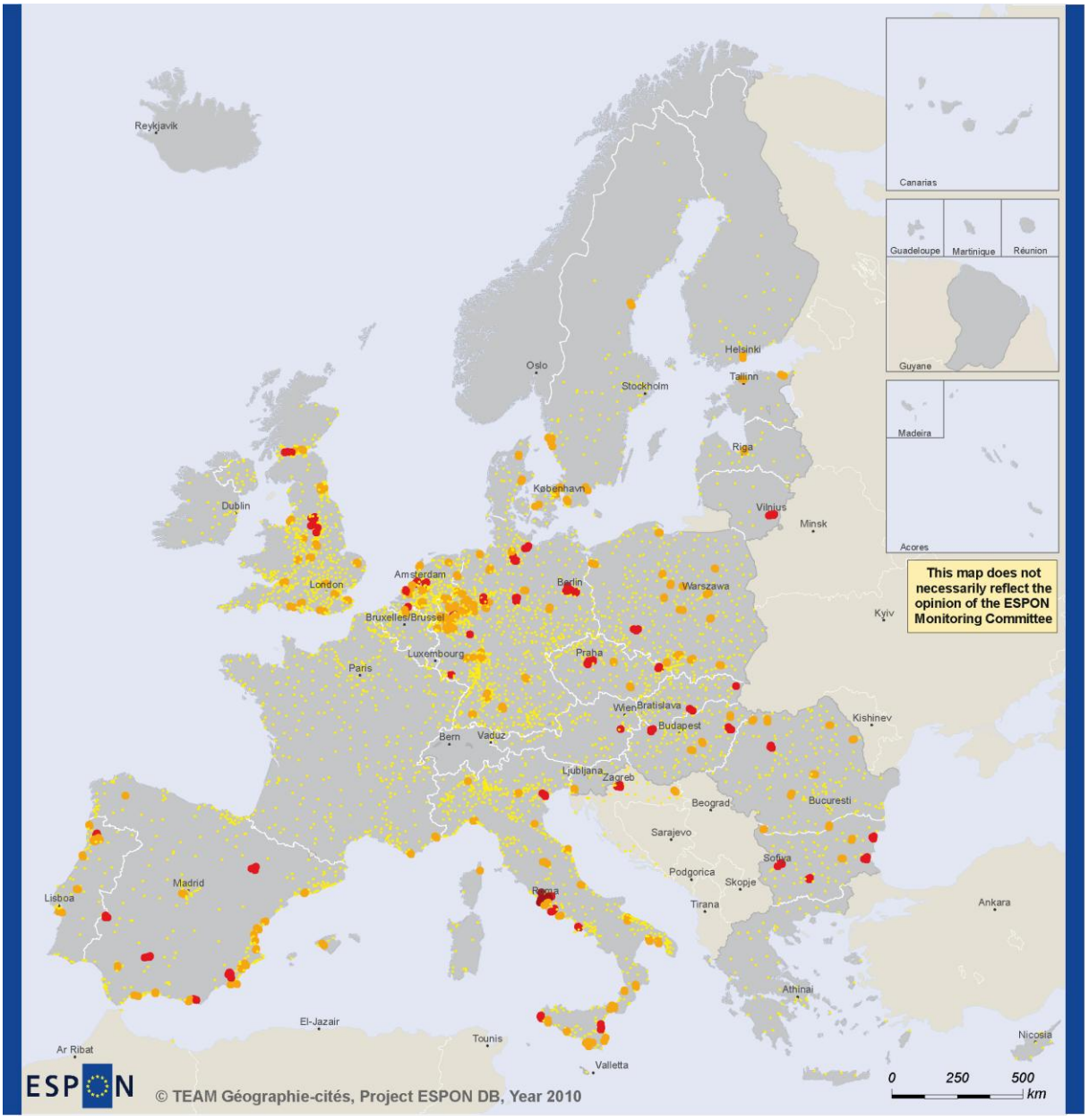


UMZ Lódz - 1: 822545 inh.
UMZ Lódz - 2: 43894 inh.

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

Figure 6 displays the location of those cases all over Europe. They appear to be quite well distributed from one country to another, even if some regions concentrate a large number of cases (like in the Rhine-Ruhr Valley or in the Netherlands) and even if some countries do not host any of them (like in France, where administrative units are particularly small).

Figure 6 : UMZ 2000 with identical names



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

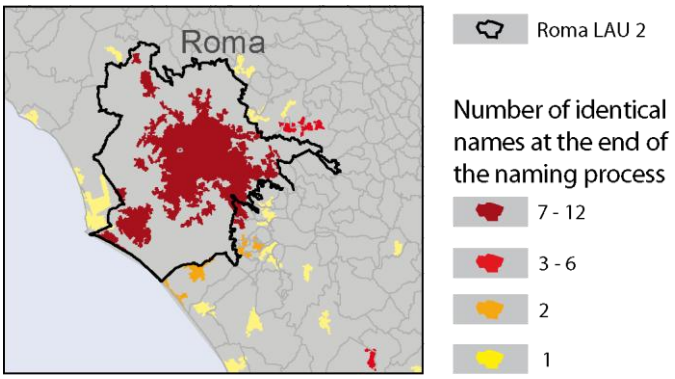
Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: The European Environment Agency (UMZ 2000 V.2), Joint Research Center (Density Grid V.5), LAU2 (2006, V.2) and national sources (see figure 9)
© EuroGeographics Association for administrative boundaries

Number of identical names at the end of the naming process

- 7 - 12
- 3 - 6
- 2
- 1

In most of the cases, only 2 identical names result from the naming process, and in a few cases we obtain 3 to 6 repetitions of the same name. The last class in the map (more than 6 repetitions) is only illustrated by Roma: 12 UMZ share this name! This is due to the very large size of the Roma LAU 2 (Figure 7), where local units have probably been merged into a unique metropolitan level.

Figure 7 : The case of Roma: 12 UMZ sharing the same name



Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

2.2 Automation of processing chain

2.2.1 The need for an automated process

The automation of the rules defined for naming UMZ is necessary for three main reasons:

- The inputs represent a huge mass of data emanating from different files which requires automatic support instead of manual process inside a GIS:
  UMZ: 4 437 UMZ over 10 000 inhabitants
  Euroboundary, LAU2: 106 452 administrative units
  Population Grid: More than 2 billion pixels in the density grid

- The calibration of the naming method supposes to conduct different tests which are useful to choose the right administrative level of reference in some specific cases (see below, 2.3.1).

- The databases contents are constantly evolving and it is essential to be reactive to these changes. Automation allows quick updating with new versions of sources or methods (EEA and JRC; for example, there are at least two different versions of UMZ2000) or new dates (as regards for instance to the integration of the future UMZ 2006, 2010…).

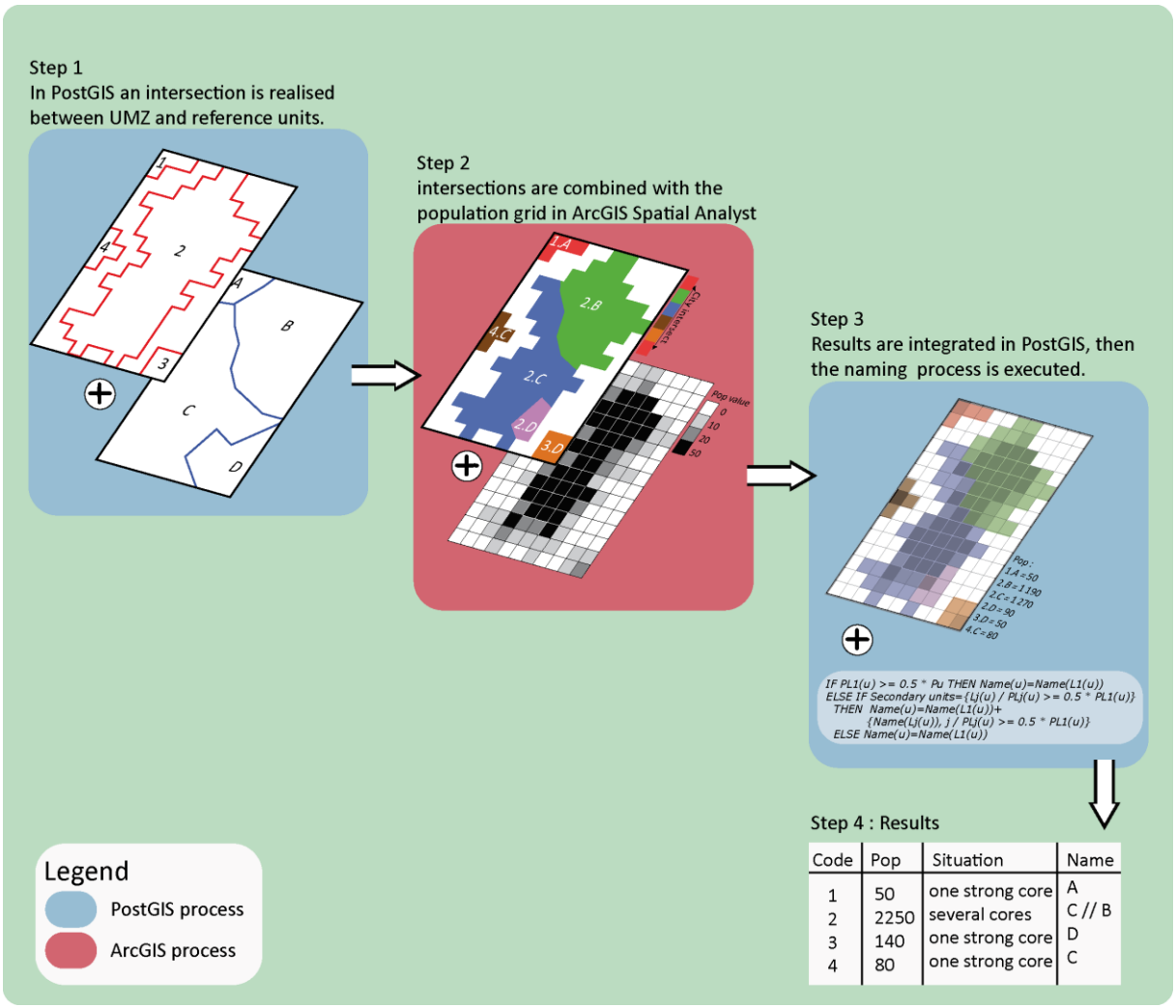
2.2.2 Process description and software solution

From a technical point of view, the automated implementation of the naming algorithm is based on three different steps which have been executed through

PostGIS and ArcGIS softwares, thanks to an integrated processing chain (Figure 8)[12]:

- Step 1: the geometrical intersections between the UMZ and the reference units are created through PostGIS
- Step 2: the population of the resulting intersections is calculated with ArcGIS Spatial Analyst. To a recent date indeed, PostGIS could not allow manipulating any raster data and the program had to use the raster solutions of the spatial analyst add-on of ESRI®. A Python language program using the Geospatial Data Abstraction Library (GDAL) has thus been developed to interface the two softwares.
- Step 3: the population computed is retrieved and integrated into PostGIS, in which the algorithm of naming is implemented

Figure 8 : The different steps of the processing chain



Code	Pop	Situation	Name
1	50	one strong core	A
2	2250	several cores	C // B
3	140	one strong core	D
4	80	one strong core	C

This program can process all the data and all the steps at once, which prevents from errors and duplicates. Eventually the automatic naming for the whole Europe could be realized in about one hour.

[12] Ultimately, this program should only rely on open sources technologies. The statistical processes can be now implemented through a PostgreSQL database with the help of the PostGIS add-on, which allows processing data with a geometry and realizing spatial requests.

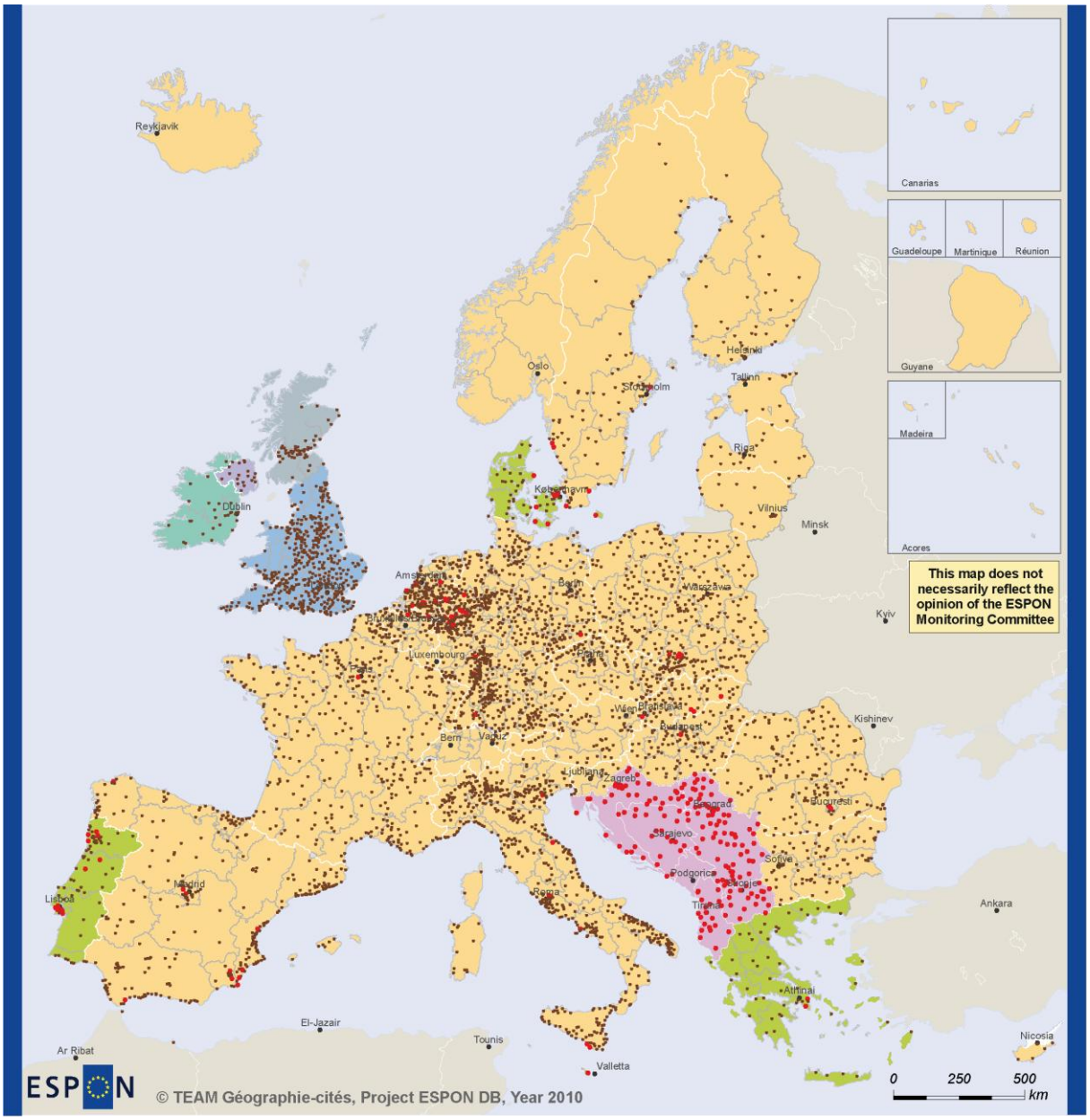
2.3 Expertise and validation

The validation of the naming method results from an expertise based on a comparison with the relevant national data base for city names (LAU 2 in majority, but also LAU1 or national settlement areas). Final results are then systematically matched to other sources (Eurostat, Geopolis) for validation. The last particular cases are checked manually, using Google Earth.

2.3.1 Selecting relevant reference units

The identification of relevant units of reference for choosing UMZ names deals with a critical issue: is there a semantic level more suitable than another for naming towns and cities? At first sight, LAU 2 seems to be the most accurate level and the most usual reference at European scale. This level is relevant in the large majority of cases, but it cannot be used for the whole countries: the correlation between cities usual names and administrative levels names depends indeed on the history of administrative divisions and on the way the status of city was formerly given. It may thus vary in some countries or experiment local variations within some countries. The name "Leipzig" fits for instance with LAU 2 level whereas the name "Dublin" fits with LAU 1 level and the name "Paris" with NUTS 3. An expertise was thus necessary to select the best reference unit for each country (Figure 9).

Figure 9 : Relevant reference units for naming UMZ



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: The European Environment Agency (UMZ 2000 V.2),
Office for National Statistics (England & Wales),
Nothern Ireland Statistics & Research Agency, General Register Office for Scotland.
© EuroGeographics Association for administrative boundaries

City names

- naming from manual process
- naming from automatic

Sources

Homogeneous sources and data grid population

- SIRE (Lau2 level)
- SIRE (Lau1 level)

No data grid population

- www.citypopulation.de

National sources and data grid population

- *Urban areas* (England and Wales)
- *Settlement Development Limits* (Northern Ireland)
- *Census town* (Ireland Republic)
- *Settlements* (Scotland)

2.3.1.1 LAU1 instead of LAU2

In some countries, the relevant administrative level appeared to be rather LAU1 than LAU2. We have used the LAU 1 version of EuroBoundaryMap 2006 v2.0 from EuroGeographics (validity: 2006).

Portugal

In Portugal, the status of a city was formerly given by decree and most of the cities corresponded to LAU 1 capital cities (*capitais de distrito*). This legacy is still present, in the sense that the current names of the LAU 2 have no relation with the names of the cities. We have then chosen LAU 1 (*concelhos-municipios*) for naming UMZ.

Denmark

In Denmark, the LAU 2 level corresponds to a parish level, whose names do not fit with the real names of cities. The most accurate level for naming UMZ is the *Kommuner* level (LAU 1).

Greece

The same issue occurs for Greece where the LAU2 level has no relation with the city name usually used. The LAU 1 level (*Demoi* and *Kointites)* has thus been chosen.

2.3.1.2 NUTS instead of LAU

NUTS level has been selected for some capital cities or other particular cases.

- Paris, Bucharest, and Budapest: the LAU 2 fits with sub-city districts (called "arrondissement" or "sector"), so that NUTS 3 level has been used in the algorithm.

- London: the name "London" is not represented at LAU 2 level (and the algorithm gives a "UMZ with several cores", with several hundred of names) neither at LAU 1 level (28 names obtained). At NUTS 3 level, the names are like "Inner London West" etc., at NUTS 2 level "Inner London" and "Outer London". The best administrative level fitting with the name "London" and with the spatial extent of the UMZ is the NUTS 1.

- Brussels: there is one LAU 2 called Brussels but it is a very little one compared to the present extent of the city, so that the name of the LAU 2 is not retained by the automatic process (the final name of the "UMZ with several cores" would be Antwerpen-Gent). Thus we have chosen the NUTS 3 level ("Arr. de Bruxelles-Capitale / Arr. van Brussel-Hoofdstad"). The definitive name resulting from the algorithm is Brussel-Antwerpen-Gent.

- Valetta (Malta): there is just one administrative level below the national one (a LAU 2 level), and the eponym LAU 2 is too small to emerge from the automatic algorithm in the final name of the UMZ. We have then attributed the name Valetta.

- In Slovakia: Bratislava and Košice are divided in several districts at LAU 2 and LAU 1 levels. The best level for naming is NUTS 3 ("Bratislava region" and

"Košice region") but it is very large compared to the UMZ spatial extents. Here again, we have attributed the names Bratislava and Košice to the UMZ.

2.3.1.3 National settlement areas

In some other countries, neither LAU2 nor LAU1 appeared suitable for naming and national data bases have been used as a reference.

United-Kingdom

In United Kingdom the LAU 2 level does not fit with names given historically to cities. LAU 2 correspond to "electoral wards" (or "parts thereof") and LAU1 to "district/unitary authorities". The city names of the "urban areas" (morphological agglomerations built by the Ordnance Survey) do not necessary fit with LAU2 or LAU1, so that we have used these urban areas as the reference units in the automatic algorithms. Three different data bases have been used:

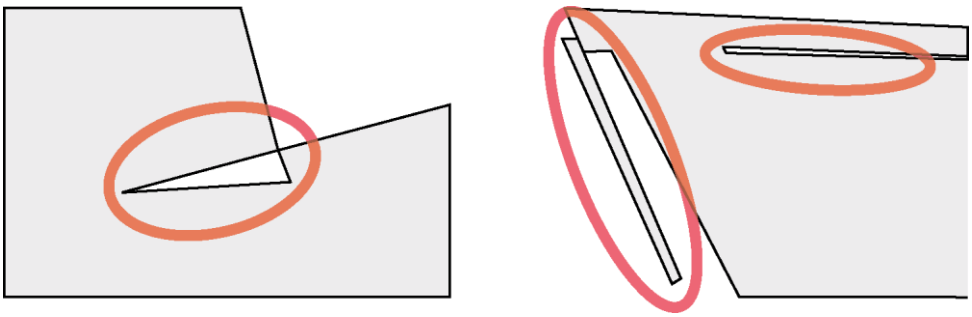
- *Urban area*s of England and Wales (2001)[13]

- *Settlements* of Scotland (2009)[14]

- *Settlements Development Limits* of Northern Ireland (2005)[15]

The vector format versions of these databases have been kindly sent by the National Statistics Office of United Kingdom. They could be used to give names by spatial requests, after correcting a hundred of topological errors that hampered the application of the automatic processes (Figure 10).

Figure 10 : Some examples of topological errors



Ireland

In Ireland, LAU2 corresponds to "electoral districts" and LAU1 do not systematically fit with the city names given to Census Towns by the Central Statistics Office of Ireland (for example when the LAU1 is a county). We have then used the Census Towns of Ireland (2006) data base, sent in vector format by the National Statistics Office of Ireland.

[13] http://www.statistics.gov.uk/
[14] http://www.gro-scotland.gov.uk/
[15] http://www.cso.ie/

2.3.2 Countries without population density grid

For the 149 UMZ larger than 10 000 inhabitants that are located in Balkan countries (Albania, Bosnia-Herzegovina, Kosovo, Macedonia and Serbia), the population has not been attributed by EEA using the Population density grid but using other sources (www.citypopulation.de). Consequently, the automatic algorithms have not been applied to these 149 UMZ and we have used the same source for giving names.

In order to ensure a good comparability in thematic explorations (section 3) these countries have not been included in the analyses.

2.3.3 Validation process

2.3.3.1 Sources

Implementing an automatic process is essential in order to quickly adapt the naming method to new sources, to avoid errors and to establish the process traceability. Yet it is equally important to validate the resulting names by comparing them to other existing urban databases. Two sources were used to check the quality of the method:

- Geopolis database (Moriconi-Ebard, 1994)
- Eurostat compilation of national city names: database "Geographical names: Settlements"[16]

In each of these databases, the cities are only represented by points (centroids) which are associated to a name. The checking method relies on successive steps:

- First, a spatial overlay of the names attributed by the algorithm and of the names associated to Geopolis and Eurostat databases. This comparison is based on a spatial request that retrieves the centroids intersecting UMZ. Specific spatial patterns have to be taken into account (for instance when Eurostat or Geopolis centroids intersect UMZ "holes").
- Secondly, a semantic comparison. The associated names are gathered into common tables and UMZ naming is validated if the names are the same. In order to optimize the matching process, it is necessary to realize textual corrections: lowercases everywhere, same local abbreviations, same spellings and universal translations (differences like Warszawa/Varsovie, Praha/Prague, Aix en Provence/Aix-en-Provence, etc., have been corrected by choosing the name of the referent database –LAU or national settlement areas-). These corrections enabled to identify and correct 1081 mismatches in names.

[16] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/geodata/archives

2.3.3.2 Typology of errors

85 names on 4437 UMZ did not match after those first checks and required case by case semantic modifications. A typology of errors has been proposed in order to make easier the future checks. These errors refer to 4 types of mismatches:

- The name of the reference unit (LAU or national settlement area) does not fit with any urban locality. This case happens generally when the name is related to some topographical features, rather than to the settlement itself (for instance the UMZ named Farum in Denmark was first called Fureso by the algorithm, whereas it is the name of a lake near the city). It happens also sometimes when it refers to a general location (for instance the large LAU named Westland, located in the south of Den Haag, Netherlands, is not suitable for naming the UMZ which corresponds more precisely to Monster locality in Google Earth).

- The name of the reference unit was historically given by an eponymous city that is currently less populated than another city included in this unit. There has been a sort of reversal between historic names and population trends, so that the most important UMZ does not receive the name of the most populated locality. This is for instance the case of the UMZ which is named Pamela (Portugal) according to the algorithm, whereas it should receive the name of the largest city (Pinhal Novo) of this LAU, identified by using Google Earth.

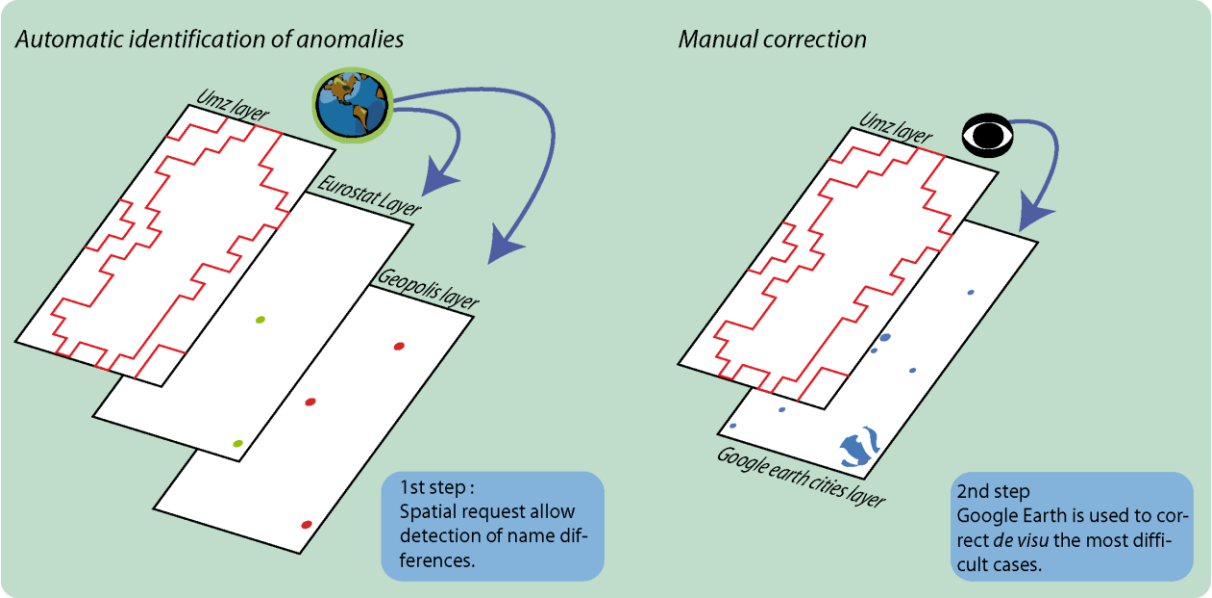
- The manual expertise of UMZ with identical names ("Lodz cases") has revealed another inconsistency: two UMZ included in the same reference unit can receive the same name even if they are very distant from each other. This is for instance the case of the two UMZ named Kristianstad-1 and Kristianstad-2 (Sweden): the UMZ of Kristianstad-2, which is distant from 10.5 km to Kristianstad-1, clearly overlays the locality of Åhus in Google Earth.

- In the peripheral parts of some industrial conurbations (Mannheim, Düsseldorf, Katowice…), a small UMZ located in the periphery of a reference unit takes the name of this unit whereas the most urbanized part of this unit belongs to the conurbation. As this urbanized part counts a relatively few population, the name of the reference unit is not taken into account in the name of the UMZ conurbation, so that this name is finally attributed only to the small UMZ. We have then chosen to give the Google Earth name and not the reference unit name to this small UMZ. An illustration can be given by the case of the locality of Rucheim, in Germany, which is included in the LAU 2 Ludwigshafen, whose larger urbanised part belongs to Mannheim UMZ. As the automatic algorithm gives the name "Ludwigshafen" to the UMZ situated at Rucheim place, we have corrected manually this name.

2.3.3.3 Solutions proposed

Ultimately these remaining mismatches are corrected by referring to the Google Earth database whose names layer is based on Multinet® from TeleAtlas®. The 85 UMZ have been converted to KLM format in order to be overlaid with other Google Earth layers. Names are then corrected *de visu* for the last mismatches (Figure 1).

Figure 11 : Steps of the validation process



2.4 Results : typology of naming situations

A simple count gives a first idea of the results obtained by automatic algorithms coupled with expertise on relevant administrative levels. We have considered UMZ larger than 10 000 inhabitants (4437 objects, including Balkans). The results have been summarized in Table 1:

Table 1 : Naming UMZ through automatic methods

	SITUATION 1 **"UMZ with one strong core"**	**SITUATION 2** **"UMZ with several cores"**	**SITUATION 3** **"UMZ with a weak core"**
Total number	4164	193	80
Percentage	94%	4%	2%

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

The typology presented in Table 1 has been mapped in Figure 12.If we focus first on "situation 2" (several cores), we recognise the industrial conurbations of the Midlands, the French and Belgium basin, the Ruhr basin, Silesia and Galicia regions. We also identify some sea-side conurbations, for example in Portugal, Spain, Italy or France. Another type of "UMZ with several cores" consists in large cities sprawling and connecting other large and close cities, like in Belgium (around Brussels) or in Romania.

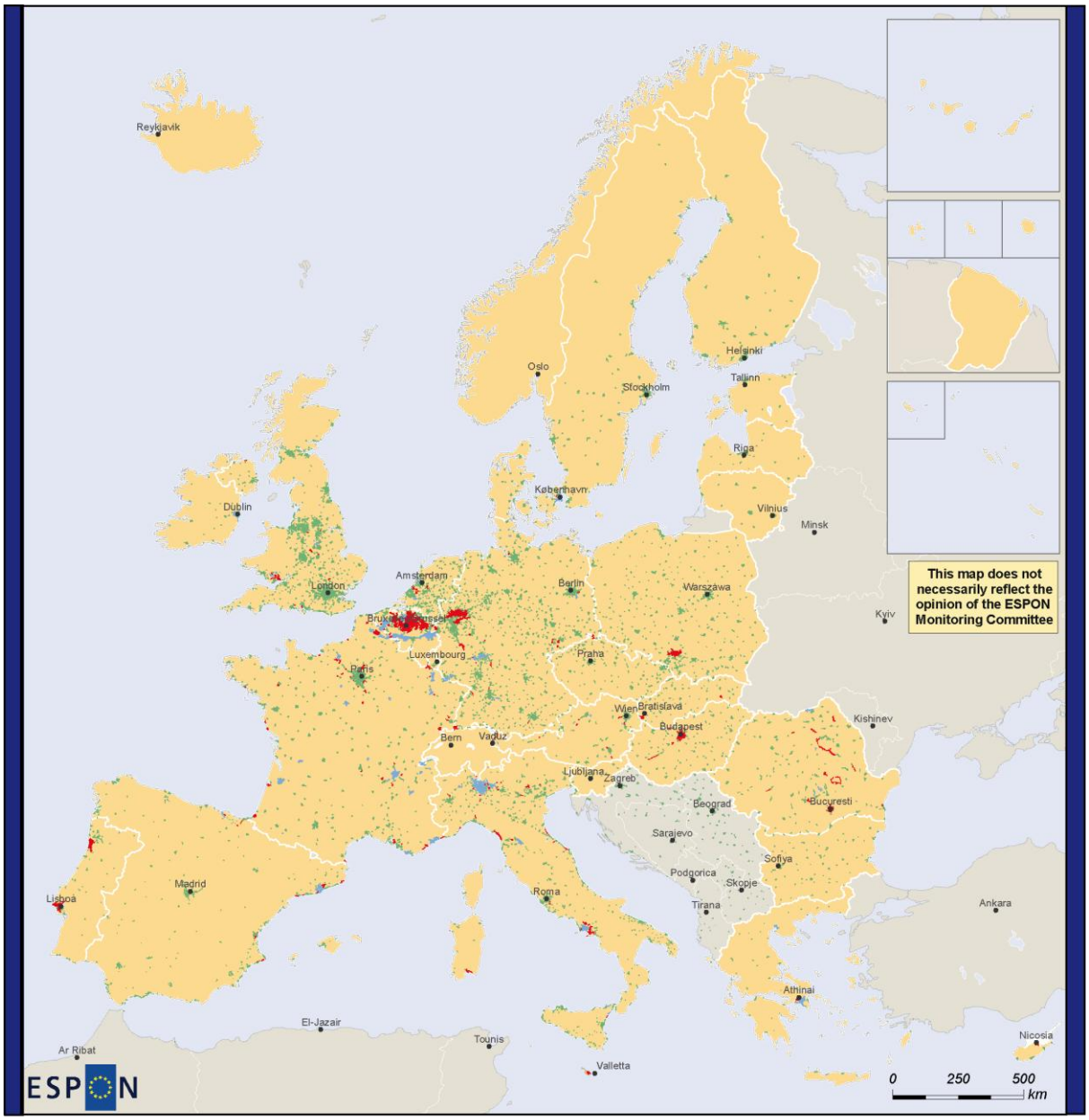
For the "situation 3" (one weak core, less strong than in situation 1), we can notice that locations are mostly the same than for UMZ with several cores (see in Italy, United Kingdom, Belgium, France…).

The "one strong core" cases, which represent the great majority (94% of the UMZ) are spared all around Europe but more represented in Northern Europe

(Sweden, Baltic countries, Denmark), characterised by relatively sparse urban settlements.

Let us notice that it is difficult to give more interpretations: situations 2 and 3 do not necessarily enlighten some "polycentric cities" but may result from the specific local or national average size of the reference units that have been used in the algorithm: we have more chances to obtain a "several cores" situation when this average size is little (like in France), and a real polycentric city could appear as "with one core" if the average size is large (like in Denmark).

Figure 12 : UMZ typology according to naming results



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

UMZ type :

- Situation 1 : UMZ with one strong core
- Situation 2 : UMZ with one weak core
- Situation 3 : UMZ with several cores

Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: The European Environment Agency (UMZ 2000 V.2), Joint Research Center (Density Grid V.5), LAU2 (2006, V.2) and national sources (see figure 9)
© EuroGeographics Association for administrative boundaries

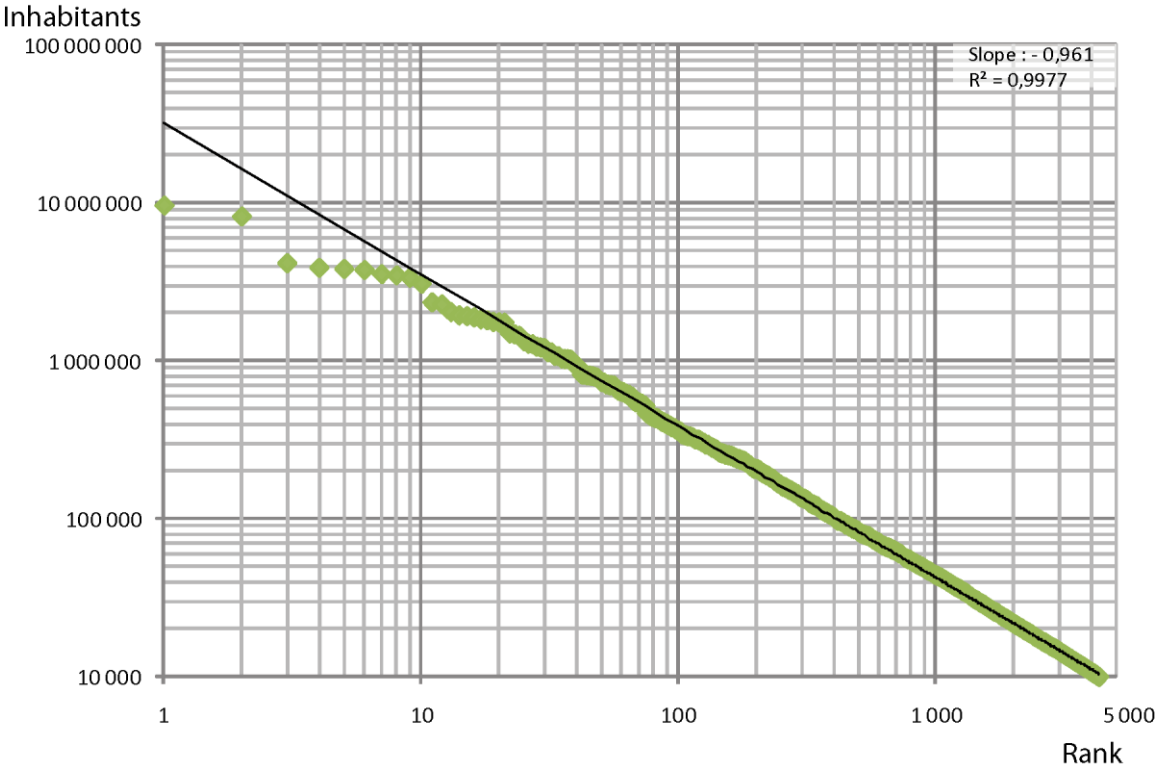
3 A thematic insight into European cities

The UMZ data base is now operational for a deep exploration of the common features and diversity of European urban settlements. Three types of analyses can be presented as a preview: in addition to classical indicators referring to the hierarchical structure of city systems or to the population density of cities, a new indicator has been created to identify and compare international UMZ.

3.1 Urban hierarchy and city-size distribution

The classical rank-size distribution, plotted for the 4437 cities over 10 000 inhabitants (Figure 13), confirms the very high regularity of the hierarchical structure at the European level (the determination coefficient R^2 equals to 0.99). The absolute value of the slope, used as an indicator of city size inequality level, is 0.96, very close to other values computed by European researchers with former databases (for instance, Geopolis data base, 1994). National studies and computation of primacy index should fruitfully complete this overview of urban hierarchy in Europe.

Figure 13 : Pareto-Zipf distribution of city sizes (UMZ 2000 data base)



Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

A closer look at the head of this hierarchy can be proposed through the "top ten" UMZ (Table 2), which are compared here to other urban rankings resulting from

European databases (Morphological Urban Areas from IGEAT[17] and Larger Urban Zones from Urban Audit[18]). Seven of these UMZ are also part of the largest set of MUA and LUZ. Main differences concern conurbations like Bruxelles-Antwerpen-Gent, Liverpool-Manchester and Essen-Dortmund-Duisburg, which are clearly overestimated by UMZ as compared to MUA which are built from similar morphological criteria. Further details about UMZ ranking are given in the table of the 50 first UMZ (see Annex, table 1).

Table 2 : "Top ten" UMZ compared to MUA and LUZ (2000, population in thousand inhabitants)

UMZ			MUA			LUZ		
Rank	**Name**	**Pop.**	**Rank**	**Name**	**Pop.**	**Rank**	**Names**	**Pop.**
1	Paris	9 476	**1**	Paris	9 591	**1**	London	11 917
2	London	8 208	**2**	London	8 265	**2**	Paris	11 089
3	Milano	4 156	**3**	Madrid	4 955	**3**	Madrid	5 805
4	Essen-Dortmund-Duisburg-Bochum	3 891	**4**	Berlin	3 776	**4**	Ruhrgebiet	5 302
5	Madrid	3 843	**5**	Barcelona	3 755	**5**	Berlin	4 971
6	Bruxelles-Antwerpen-Gent	3 790	**6**	Milano	3 698	**6**	Barcelona	4 234
7	Liverpool-Manchester	3 531	**7**	Athinai	3 331	**7**	Athina	4 013
8	Athinai	3 489	**8**	Roma	2 532	**8**	Roma	3 458
9	Berlin	3 435	**9**	Birmingham - Wolverhampton	2 363	**9**	Hamburg	3 135
10	Barcelona	3 106	**10**	Lisboa	2 315	**10**	Milano	3 077

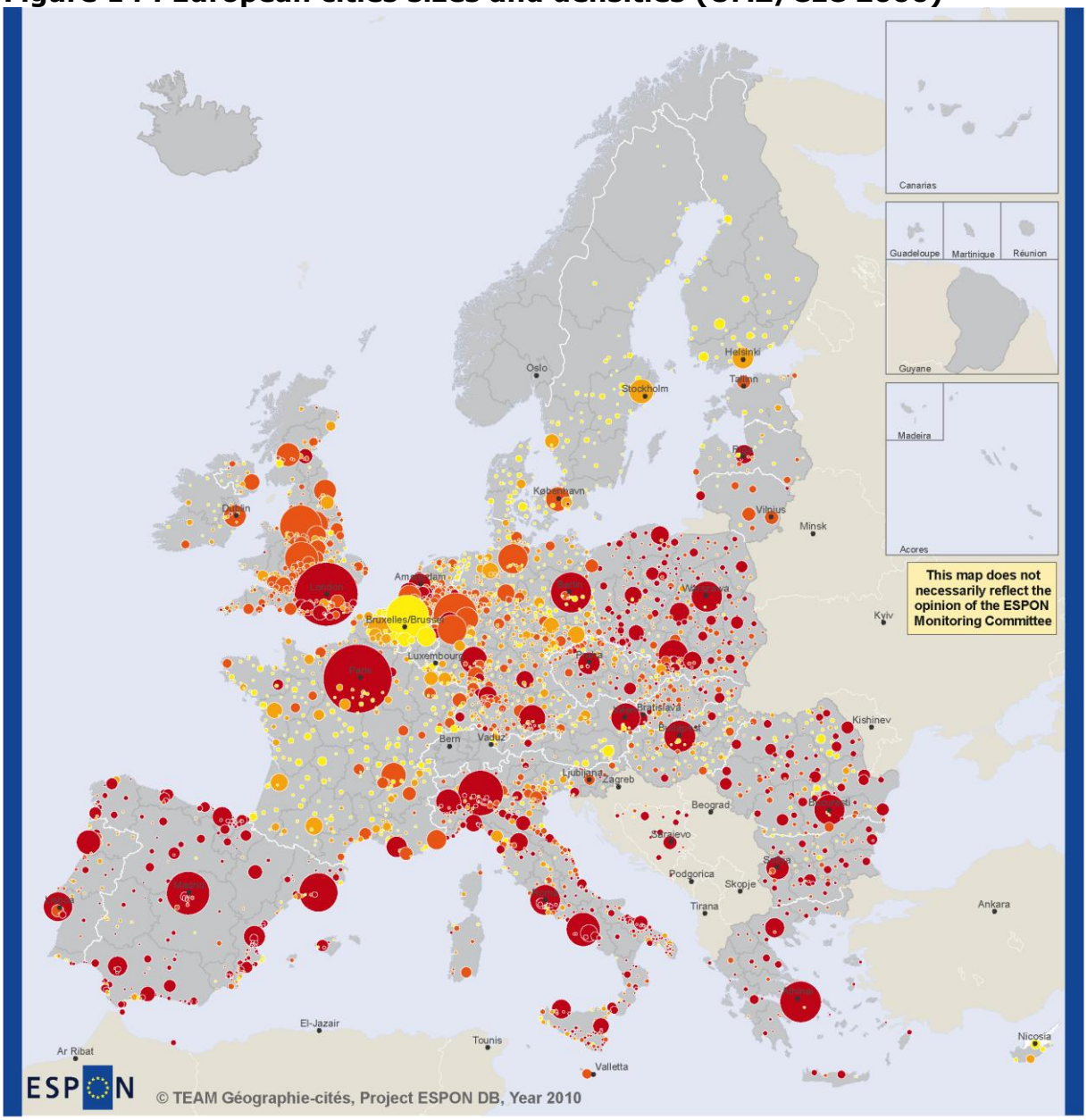
Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center, ESPON 1-4-3, Urban Audit.

3.2 Density patterns

A multiscalar analysis of density levels in Europe gives striking results, with a major North-South gradient (Figure 14): for example, average urban density is lower than 2000 inh./km^2 in Sweden, Denmark, Finland, whereas it reaches 4000 inh./km^2 in Italy and more in Spain or Greece (Table 3).

[17] MUA have been defined in Vandermotten et *alii* 1999 and in ESPON 1-4-3 « Study on urban functions ». IGEAT refers to Institut de Gestion de l'Environnement et d'Aménagement du Territoire, Université Libre de Bruxelles.
[18] http://www.urbanaudit.org/

Figure 14 : European cities sizes and densities (UMZ/CLC 2000)



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: The European Environment Agency (UMZ 2000 V.2), Joint Research Center (Density Grid V.5), LAU2 (2006, V.2) and national sources (see figure 9)
© EuroGeographics Association for administrative boundaries

Number of inhabitants

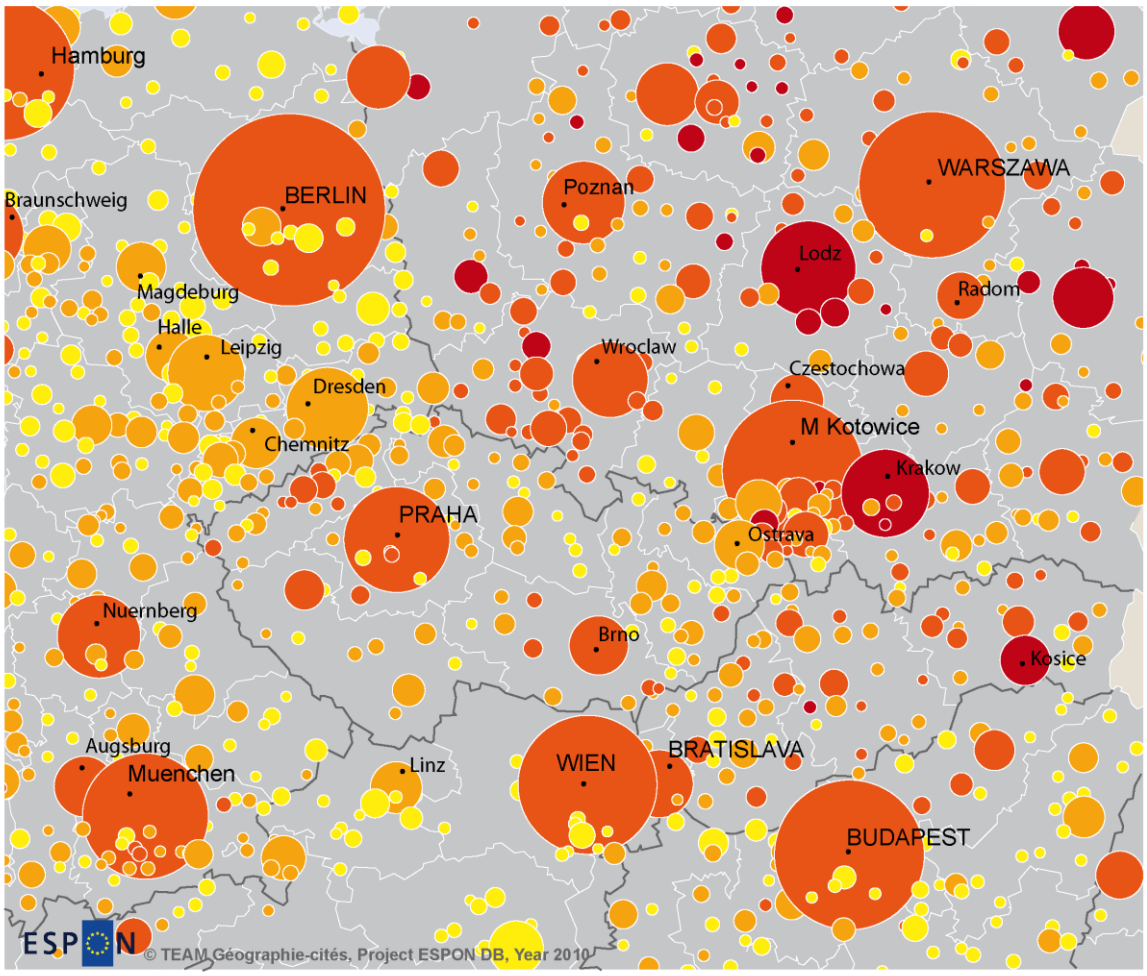
10 000 000
1 000 000
10 000

Density (inhab/km²)

340 - 2 150
2 150 - 2 900
2 900 - 3 800
3 800 - 33 600

Some national specificities appear also very strongly, as revealed by the higher densities of Dutch cities, the strong discontinuities observed for instance at the Franco-Spanish frontier and at the German-Polish border (Figure 14), or as suggested by the high densities of some Eastern countries like Poland (see Annex, Table 2).

Figure 15 : European City sizes and densities (UMZ/CLC 2000)



Furthermore, a strong and regular relationship with city size levels can be enhanced (Table 3): densities exceed 5700 inh./km^2 in cities larger than 2 millions inhabitants, then decrease regularly until 3000 inh./km^2 for cities between 10 000 and 25 000 inhabitants. This higher level of densities in the largest cities can be interpreted as the result of a historical accumulation process and as the expression of a more pronounced centrality and competition for land.

Let us recall that density indicator is of high interest for urban planning issues, for example in environmental topics, especially when it can be coupled with other

transportations indicators. Even if current debates enlighten a lack of consensus between researchers, we can mention for instance the question of the minimal city or sub-district density level necessary for providing efficient public transportation networks, or the one of the possible link between average city density level and pollution gas emissions.

Table 3 : Urban population density per class of population

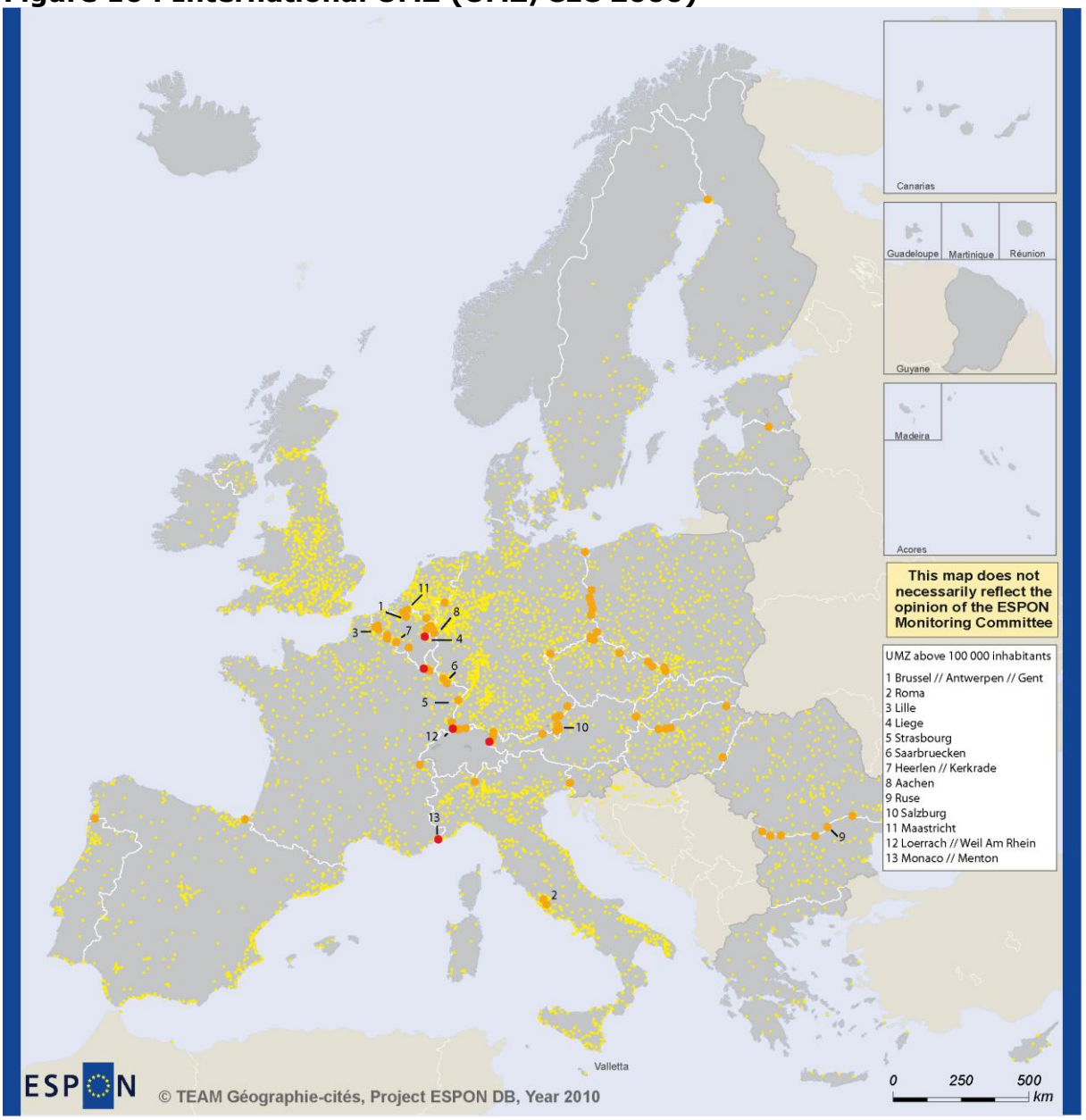
Class of population	Number of UMZ	Density (inh./km²)
1 to 10 millions inh.	39	4787
0,5 to 1 millions inh.	36	4 892
250 to 500 thousands inh.	136	4 235
100 to 250 thousands inh.	203	3 932
50 to 100 thousands inh.	512	3 469
25 to 50 thousands inh.	904	3 214
10 to 25 thousands inh.	2607	3 053

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

3.3 International UMZ

International UMZ can now be identified through a new indicator (international code) that describes the number of countries crossed by each UMZ. The distribution of international UMZ (Figure 16) is a first important result that offers an overview of cross-national cities, independent from institutional or administrative frames. Furthermore, an index of internalization (% of population living in one or more countries different than the main one) has been computed. It allows to qualify in a comparable way to what extent the city is embedded in a multi-national context and completes in a fruitful way the population indicator of these UMZ: for example, the most populated international UMZ is Brussels/Antwerpen/Gent, but it extends in a very small part in Netherlands (international index is only 1%). At the opposite, some UMZ located at the Poland/Germany, Slovakia/Hungary or Austria/Germany frontiers are not very populated but their international index is over 40% (Table 5). Two other tables are presented in Annex, with the most important international UMZ by countries, according to their population (see Annex, Table 3) and according to their international index (see Annex, Table 4).

Figure 16 : International UMZ (UMZ/CLC 2000)



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: The European Environment Agency (UMZ 2000 V.2), Joint Research Center (Density Grid V.5), LAU2 (2006, V.2) and national sources (see figure 9)

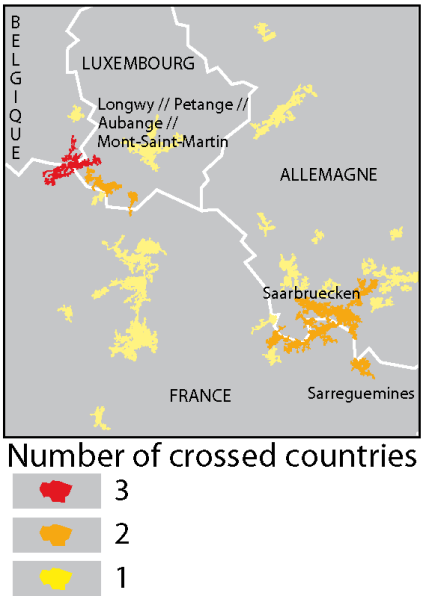
Number of crossed countries

3

2

1

Figure 17 : A zoom on some international UMZ



Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

Table 4 : Top Ten of main populated international UMZ

UMZ Name	Population	Main Country	All countries*	International index**
Brussels-Antwerpen-Gent	3 769 885	BE	BE // NL	0,1
Roma	1 891 236	IT	IT // VA	0,03
Lille	1 335 026	FR	FR // BE	30,1
Liege	760 811	BE	BE // DE // NL	1,03
Strasbourg	435 410	FR	FR // DE	3,4
Saarbruecken	367 294	DE	DE // FR	28,5
Heerlen-Kerkrade	259 447	NL	NL // DE	12,6
Aachen	213 930	DE	DE // NL	3,9
Ruse	201 106	BG	BG // RO	35,1
Salzburg	181 407	AT	AT // DE	10,2

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

* Ranked by decreasing population.
**% of population of UMZ that is not in the main country.

Table 5: Top ten of most internationalised UMZ

UMZ Name	Population	Main country	All countries*	International index**
Comines/Wervik/ Comines-Warneton-Komen-Waasten	33 796	FR	FR // BE	47,2
Longwy- Petange – Aubange-Mont-Saint-Martin	70 371	FR	FR // BE // LU	45
Oberndorf bei Salzburg	10 096	AT	AT // DE	44,1
Braunau am Inn	29 933	AT	AT // DE	42,2
Monaco / Menton	105 193	FR	FR // IT // MC	41,1
Komarno	44 404	SK	SK // HU	40,8
Hamont	16 854	NL	NL // BE	40,7
Cieszyn	53 152	PL	PL // CZ	40,6
Guben	34 413	DE	DE // PL	40,1
Tui	17 801	ES	ES // PT	39,9

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

* Ranked by decreasing population.
** % of population of UMZ that is not in the main country.

4 Conclusion

UMZ present a great potential for future, as they result from the same building methodology and are defined with the same criteria in all the countries. They have been here completed and validated as a European database operational for urban studies: 4437 urban settlements over 10 000 inhabitants are now defined from CLC2000 with harmonized criteria (EEA, last version of Urban Morphological Zone shapes), population (JRC, last version of Population Density Grid), names and metadata. The establishment of an automated process for naming UMZ allows quick updating with new versions of sources or methods (EEA and JRC) or new dates (2006, 2010…). The validation of the method results from an expertise which selects the relevant data base for city names and relies on systematic matches with other sources (Eurostat, Geopolis). This protocol leads to a powerful data base for exploring the features of European cities in 2000 as regards to their settlement characteristics, distribution of city sizes, density patterns or international UMZ configurations.

Further work should improve the operational dimension of UMZ in future:
- Their integration in the ESPON Data Base could be hugely improved by building a zoning correspondence Table with LUZ or other functional database.
- Interoperability with other geo-referenced data bases (urban transport infrastructures, urban mobility, socio-economic LAU data…) opens a wide range of environmental and social studies. Enlargement of urban indicators, towards environmental (grid/raster) and socio-economic data should be realised by using the "OLAP Cube for Urban analysis" developed by UAB[19] and indicators collected at LAU2 level and aggregated in UMZ delineations
- Urban indicators were here computed for the year 2000. A very important challenge lies in the possibility of adding a temporal dimension to these indicators and making them vary in time. Development of a temporal urban data model would enable to follow UMZ urban indicators through time (1990, 2010 and future other dates).

[19] See Technical Report « Social/Environmental data », ESPON Database 2013, written by the Autonomous University of Barcelona.

Annex:

Table 1: UMZ ranking, "Top 50"

Rank	Name	Country	Population 2001	Density (inh./km²)
1	**Paris**	FR	9 656 819	5 386
2	**Greater London Urban Area**	UK	8 221 307	4 861
3	**Milano**	IT	4 164 504	4 166
4	**Essen / Duisburg / Dortmund / Bochum /Gelsenkirchen - 1**	DE	3 892 380	3 674
5	**Madrid - 1**	ES	3 823 031	9 637
6	**Brussel / Antwerpen / Gent**	BE	3 769 885	1 841
7	**Palmas de Gran Canaria, Las**	UK	3 546 819	3 538
8	**Dimos Athinaion**	GR	3 489 768	9 896
9	**Berlin - 1**	DE	3 367 457	4 648
10	**Barcelona**	ES	3 088 470	10 533
11	**Napoli**	IT	2 354 010	9 007
12	**West Midlands Urban Area**	UK	2 286 859	3 480
13	**Budapest**	HU	2 042 024	3 963
14	**M. St. Warszawa - 1**	PL	1 948 024	4 665
15	**Bucaresti - 1**	RO	1 925 741	9 066
16	**Roma - 1**	IT	1 891 236	6 300
17	**Hamburg - 1**	DE	1 838 019	3 549
18	**M. Katowice / M. Sosnowiec / M. Gliwice / M. Zabrze / M. Bytom - 1**	PL	1 810 260	3 845
19	**Koeln - 1**	DE	1 767 659	3 525
20	**Wien**	AT	1 756 034	4 304
21	**Lisboa / Sintra**	PT	1 749 316	5 906
22	**Frankfurt Am Main - 1**	DE	1 493 470	3 971
23	**West Yorkshire Urban Area - 1**	UK	1 473 892	3 541
24	**Muenchen**	DE	1 444 902	4 677
25	**Lille**	FR	1 335 026	2 831
26	**Lyon**	FR	1 287 802	3 092
27	**Torino**	IT	1 278 016	6 640
28	**Stockholm**	SE	1 233 147	2 822
29	**Kobenhavn**	DK	1 218 013	2 986
30	**Porto / Vila Nova De Gaia / Matosinhos / Gondomar**	PT	1 208 098	3 904
31	**Wuppertal / Hagen /**	DE	1 138 180	3 754
32	**Glasgow - 1**	UK	1 135 155	3 290
33	**Sofia - 1**	BG	1 079 088	6 271
34	**Rotterdam - 1**	NL	1 072 014	3 436
35	**Tyneside - 1**	UK	1 037 720	3 463
36	**Dublin**	IE	1 029 106	3 339

Rank	Name	Country	Population 2001	Density (inh./km²)
37	**Amsterdam**	NL	1 028 359	4 270
38	**Praha - 1**	CZ	1 020 584	4 303
39	**Valencia - 1**	ES	967 206	10 874
40	**Helsinki-helsingfors - 1**	FI	917 813	2 218
41	**Marseille**	FR	902 756	5 215
42	**M. Lodz - 1**	PL	822 545	5 232
43	**Bilbao**	ES	819 465	13 869
44	**Nice - 1**	FR	812 330	3 092
45	**Duesseldorf - 1**	DE	809 770	3 899
46	**Dimos Thessalonikis**	GR	804 095	11 600
47	**Sevilla**	ES	797 127	7 984
48	**Palermo**	IT	786 622	7 058
49	**Liege**	BE	760 811	1 401
50	**Gijon**	UK	736 438	3 276

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

Table 2 : Urban population density per country

Country	Number of UMZ	Density (inh./km²)
Albania	19	-
Austria	60	1 986
Bosnia and Herzegovina	21	4 602
Belgium	50	1 449
Bulgaria	83	4 207
Cyprus	5	1 700
Czech Republic	116	2 889
Germany	846	2 799
Denmark	51	1 631
Estonia	16	2 396
Spain	348	5 752
Finland	44	1 135
France	391	2 037
Greece	51	4 493
Croatia	36	-
Hungary	107	2 202
Ireland	26	2 359
North Ireland	1	3 032
Italy	575	3 922
Kosovo	16	-
Liechtenstein	1	1 265
Lithuania	16	3 509
Luxembourg	3	2 532
Latvia	23	2 730
Monaco	1	5 818
Macedonia	20	-
Malta	1	3 297
Netherlands	201	3 398
Poland	327	3 658
Portugal	70	3 545
Romania	159	3 249
Serbia	58	-
Sweden	88	1 316
Slovenia	12	2 943
Slovakia	66	3 151
San Marino	1	4 314
United Kingdom	528	3 114

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

Table 3 : Main international UMZ by country (ranked by population)

Country	Number of international UMZ	Main international UMZ		
		Population	International index*	Name
Austria	6	181 407	16,2	Salzburg
Belgium	4	3 769 885	0,1	Brussel / Antwerpen / Gent
Bulgaria	6	201 106	35,1	Ruse
Czech Republic	5	64 774	0,8	Karvina
Germany	19	367 294	28,5	Saarbruecken
Estonia	1	20 500	39,2	Valga
Spain	2	83 206	15,7	Irun
Finland	1	12 026	36,9	Tornio-Tornea
France	12	1 335 026	30,1	Lille
Hungary	4	26 606	36,9	Esztergom
Italy	4	1 891 236	0,03	Roma
Lithuania	1	10 029	5,8	Mauren / Eschen / Ruggell
Luxembourg	2	62 104	1,3	Esch-Alzette / Differdange
Netherlands	6	259 447	12,6	Heerlen / Kerkrade /
Poland	3	53 152	40,6	Cieszyn
Slovakia	2	44 404	40,8	Komarno

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

*% of population of UMZ that is not in the main country.

Table 4 : Main international UMZ (ranked by International index)

Country	Number of International UMZ	Main international UMZ		
		International index*	Name	Population
Austria	6	44,1	Oberndorf bei Salzburg	10 096
Belgium	4	8,5	Essen (BG)	14 407
Bulgaria	6	35,1	Ruse	201 106
Czech Republic	5	31,1	Nachod	29 071
Germany	19	40,1	Guben	34 413
Estonia	1	31	Valga	20 500
Spain	2	39,8	Tui	17 801
Finland	1	36,9	Tornio-Tornea	12 026
France	12	47,2	Comines / Wervik / Comines-Warneton - Komen-Waasten	33 796
Hungary	4	36,9	Esztergom	27 000
Italy	4	36,7	Gorizia	44 518
Lithuania	1	5,8	Mauren / Eschen / Ruggell	10 029
Luxembourg	2	8,3	Dudelange	18 284
Netherlands	6	40,7	Hamont	16 854
Poland	3	40,6	Cieszyn	53 152
Slovakia	2	40,8	Komarno	44 404

Sources: LAU 2 (EuroBoundaryMap 2006, v2.0) from EuroGeographics, UMZ2000 from European Environment Agency, Population density Grid v.5 from Joint Research Center.

*% of population of UMZ that is not in the main country.

Figures list:

Tables:

Annex Tables:

TECHNICAL REPORT

ESPON

LUZ specifications (Urban Audit 2004)

MAIN RESULTS

- **Stakes:** Larger Urban Zones are based on various national definitions (functional areas, planning regions, local administrative units, etc.). Only a good knowledge of specifications allows identifying bias resulting from the national heterogeneity of these definitions
- **Methodology:** after collecting and expertizing documentation on national specifications, a common "syntax" has been used for categorizing specifications (rules of construction, building blocks, evolution since UA II etc.)
- **Results:** 1) Four general maps enlighten a synthetic typology of definitions, the evolution since UA II, the diversity of thresholds in commuting-based approaches and the specificity of capital city LUZ definitions 2) Thirty country-sheets summary specifications following a common synthax

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

72 PAGES

LIST OF AUTHORS

Anne Bretagnolle, University Paris 1, UMR Géographie-cités

François Delisle, UMR Géographie-cités

Hélène Mathian, C.N.R.S., UMR Géographie-cités

Liliane Lizzi, C.N.R.S., UMR Géographie-cités

Marianne Guérois, University Paris 7, UMR Géographie-cités

Guilhain Averlant, UMR Géographie-cités

Contact

anne.bretagnolle@parisgeo.cnrs.fr

tel. + 33 1 01 40 46 40 00

TABLE OF CONTENT

1 Stakes and matter

1.1 A bottom-up approach for building a European urban data base

Urban Audit[1] (UA) was conducted at the initiative of the Directorate-General for Regional Policy at the European Commission. It aims collecting comparable statistics and indicators for cities, at three different scales (Sub-Districts, City Core and Larger Urban Zones). A large number of indicators (more than one hundred) are sent by countries for each of these levels of definition, and the interest of such variables for urban studies is not to be proved. However, several questions concerning international comparability of indicators results are specifically raised at the larger scale definition, the LUZ one.

Indeed, the work engaged by Urban Audit for collecting and possibly harmonizing LUZ delineations represent a specific approach for trying to build a European set of cities. As opposed to Urban Morphological Zones[2], it is not a top-down approach (starting from identical definition criteria and trying to enrich it by taking into account national diversity) but a bottom-up approach. Countries are required by Urban Audit to choose and send national definitions of LUZ, sometimes changing them when taking into account some recommendations.

1.2 Detecting bias resulting from heterogeneity in LUZ definitions

Before using the large amount of available indicators of Urban Audit, some cautions have to be taken: LUZ definitions are so different from one country to another that an expertise of national specifications must be done in order to identify bias resulting from this heterogeneity.

The first step in our expertise was to collect a huge set of documentation on LUZ specifications (National Report sent at each round of Urban Audit, other information on Statistical Census Boards websites, annex documentations gathered by Urban Audit and kindly sent to us[3]). A second step was dedicated to the construction of a common syntax for describing these specifications, using a common vocabulary and approach in order to understand the national logics in these rules.

---

[1] http://www.urbanaudit.org/

[2] See Technical Report "Naming UMZ: making them more operational for urban studies", ESPON DB 2013.

[3] We want here to thank warmly Teodora Brandmuller, from Urban Audit, without her this work could not have be achieved. She helped us several times to understand some complicated parts of national documentations and she sent us, at different times, a lot of files that fortunately filled some lack in our collect.

This technical report is divided in three parts: in the first one we recall some specificities of Urban Audit and we present the Larger Urban Zones (criteria of selection, evolution, national sources etc.). In the second part we present the methodology used for analyzing LUZ specifications, in particular the general model of country-sheet that has been used for re-writing in common vocabulary and syntax LUZ specification for each of the 30 countries under study. In the third part, results (typologies and maps) are presented and discussed. The 30 country-sheet on LUZ specifications are presented in Annex.

2 Larger Urban Zone presentation

2.1 Urban Audit rounds

Urban Audit, under the coordination of Eurostat, aims gathering comparable data covering most aspects of urban life in European cities and towns. National Urban Audit Coordinators (NUAC), generally represented by National Statistics Offices, are the link between Eurostat and the cities involved. They collect and gather data in their country before transmitting it to Eurostat.

Three different rounds occurred until now (Table 1). A first phase (pilot phase) was launched in 1998, a second one between 2003 for Member States and 2004 for Candidate Countries (UA II 2001), then a third between 2006 and 2007 (UA III 2004). The next round (UA IV 2008) is ongoing and data dissemination is expected for 2011.

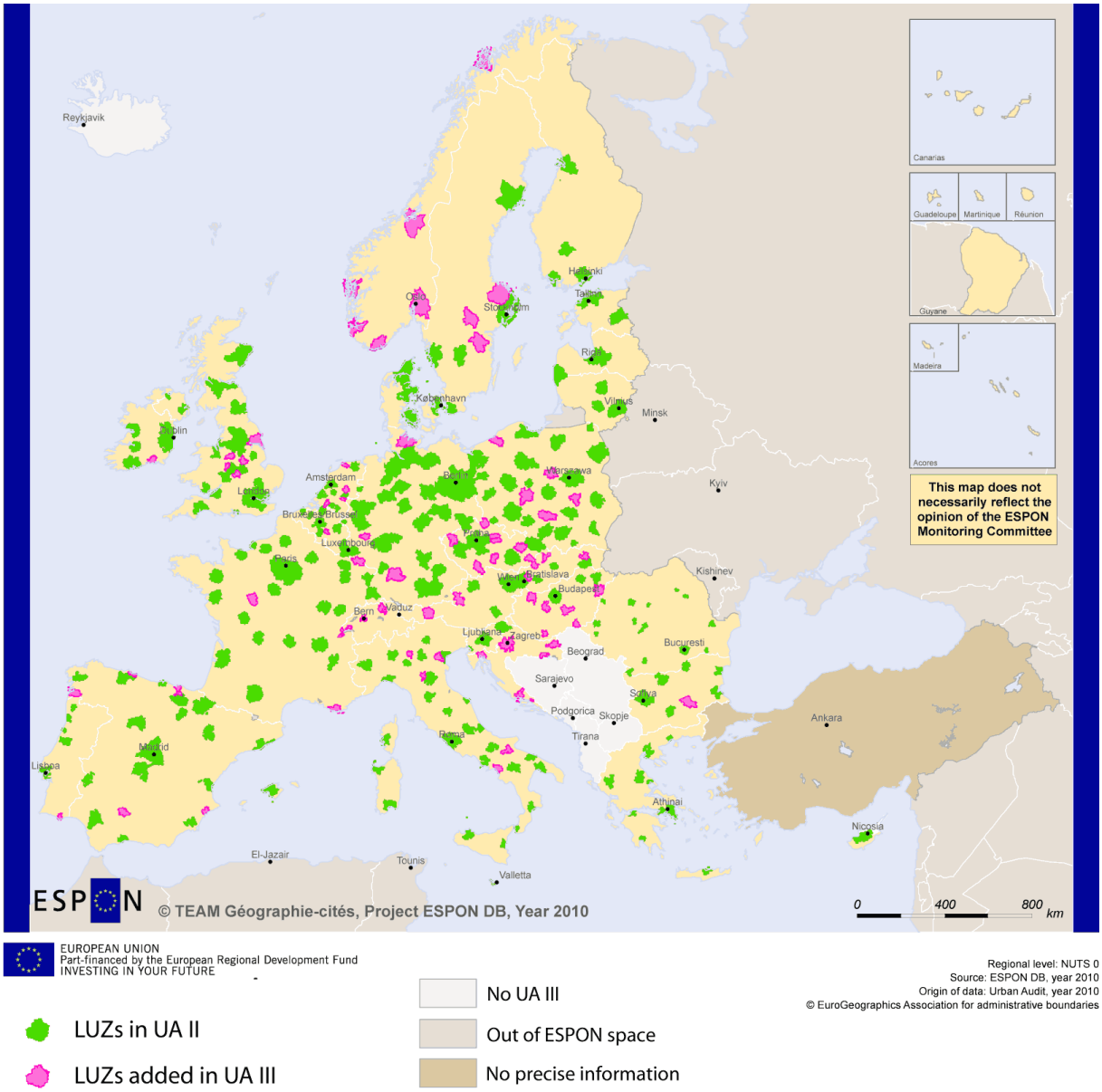
Table 1: Number of cities, countries and indicators involved in the three Urban Audit rounds

	UAPP (1998)	**UA II (2001)**	**UA III (2004)**
Number of cities	58	189 MS +69 CC = **258**	321+46 = **367**
Number of countries	15	15 +12 = **27**	27 MS + Norway + Switzerland + Croatia + Turkey = **31**
Number of indicators	500	336	338
Reference year	1981, 1991, 1996	2001	2004
Launching year	1998	2003-2004	2006-2007

UAPP: Urban Audit Pilot Phase; MS: Member State; CC: Candidate Countries

Some countries which are not included in the UE perimeter have participated to Urban Audit III and we have integrated them in this expertise when finding some data, i.e. Croatia, Norway and Switzerland. For Turkey, we did not found any information concerning LUZ so that we did not consider it here. We have worked on 30 countries in all (Figure 1)

Figure 1: UA II and UA III LUZ



Source: GISCO. Some mismatches between GISCO[4] and National Reports are described in the Country-sheets(Annex) and must be underlined on this map. Some LUZ are missing (the 26 LUZ from Turkey and 5 of the 9 LUZ from Switzerland), some others should be removed (Toulon LUZ in France).

[4] Geographical Information System of the European Commission (http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1).

2.2 LUZ definition by Urban Audit

2.2.1 The three-level approach of the European Cities

Since Urban Audit II, urban entities that have been defined as "cities" may correspond to three representations:

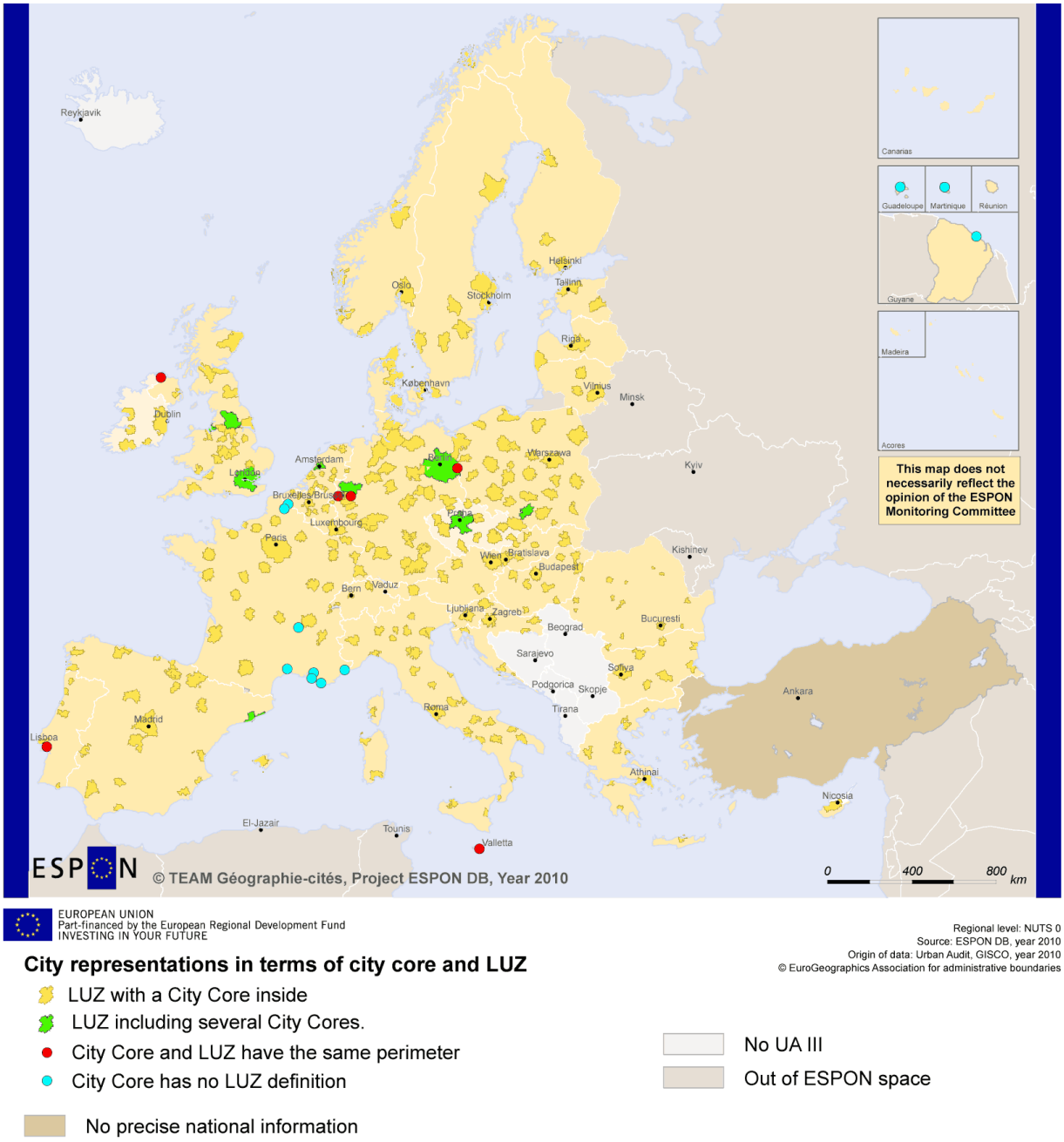
- The ***Core City*** (CC) relates to an administrative approach of cities and towns, and generally fits with the eponymous central LAU2. In order to facilitate international comparisons, some *Kernels* have been created for larger European cities, and consist in an aggregation of LAU2.

- The ***Sub-City District*** (SCD) consists in a subdivision of the city according to population criteria.

- The ***Larger Urban Zone*** (LUZ) is conceived by Urban Audit to approach the functional urban region definition. It must "*allow a comparison between the city and its surroundings. The goal was to have an area from a significant share of the resident commute into the city, a concept known as the 'functional urban region'. To ensure a good data availability, the Urban Audit works with administrative boundaries that approximate the functional urban region*" (http://www.urbanaudit.org/help.aspx).

In most of the cases, Urban Audit data collections concern European Cities simultaneously at the three different scale-levels. However, each Urban Audit "city" hasn't systematically the three representations: some Urban Audit cities have no LUZ but one CC, other have the same perimeter for LUZ and CC; sometimes two CC share the same LUZ…The map presented below (Figure 2) gives a synthetic view of these particular cases in LUZ-CC articulations.

Figure 2: A variety of representations of Urban Audit "cities" (UA III)



2.2.2 Selection criteria of UA cities

Urban Audit cities and the three-scale perimeters for data collection are selected according to specific criteria, that are unusual in the field of urban studies in the sense that classical hierarchical criteria (minimal population threshold) are completed by geographical criteria (spatial dispersion within each country) and administrative criteria (inclusion of national and regional capital) (see Insert 1).

Insert 1: Selection criteria of Urban Audit cities (http://www.urbanaudit.org/help.aspx)

"What is Urban? The Urban Audit aims at a balanced and representative sample of cities in Europe. To obtain such a selection, a few simple rules were followed:

1. Approximately 20% of the national population should be covered by the Urban Audit.

2. All capital cities were included.

3. Where possible, regional capitals were included.

4. Both large (more than 250 000 inhabitants) and medium-sized cities (minimum 50 000 and maximum 250 000 inhabitants) were included.

5. The selected cities should be geographically dispersed within each Member State".

The results give a heterogeneous set of cities, as displayed on the following figure. On the map representing LUZ population (Figure 3), the information seems to be harmonious and close to the one obtained, for example, from the city-size patterns given by Urban Morphological Zone data base (see the Technical Report Naming UMZ, ESPON database 2013).

However, when plotting the distribution of LUZ sizes on a rank-size graph (figure 4), the queue-distribution appears very different from the rest of the city sizes. This is due to the very small number of LUZ under 120 000 inhabitants, under-represented compared to the size of other classes. Some of these small LUZ are indicated on Table 2 (see for example in Germany, Norway, Greece etc.).

Figure 3: Pareto-Zipf distribution of LUZ size (UA III)

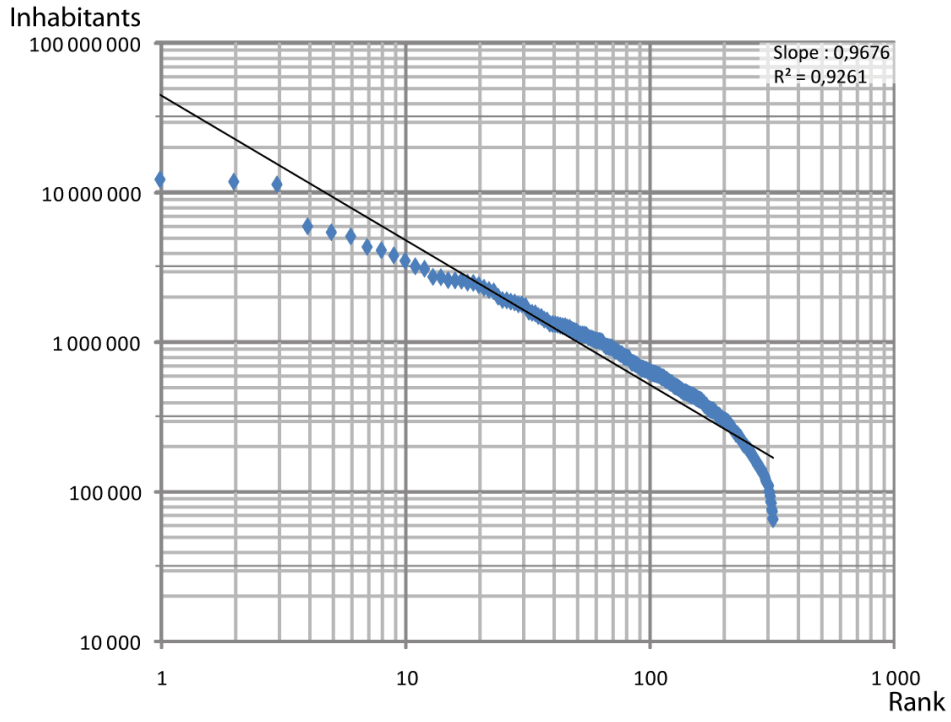
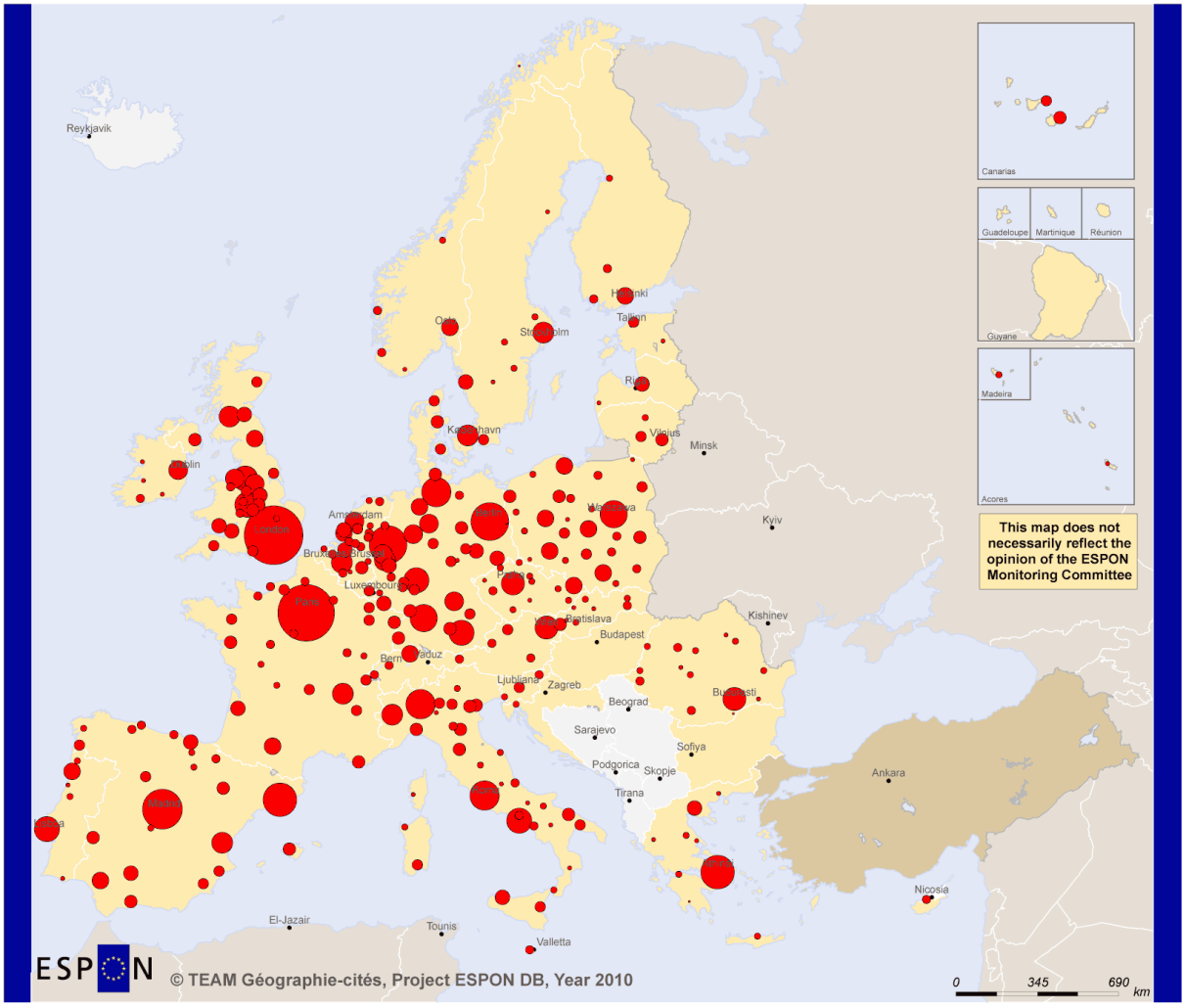


Figure 4: Population of LUZ in 2003-2006 (UA III)



NB: No data for Bulgaria and Hungary (see table 2)
Some LUZ population are missing on Eurostat website[4].

In a more general way, one can observe the fact that the distribution is not well adjusted by the model (the coefficient of determination R^2 is only 0.93) whereas it is not the case for more classical distributions of European city-sizes (see UMZ, in the Technical Report "Naming UMZ" of ESPON database 2013, or Geopolis data base, in François Moriconi-Ebrard, 1993, where the value is over 0.98).

[5] http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?sort=1&file=data%2Furb_vluz.tsv.gz

Table 2: LUZ number and min-max population by country (UA III)

Countries	Nb of LUZ in UA III	Population min and max (2003-2006)
Austria	**5**	**272 677 2 179 769**
Belgium	**7**	**139 024 1 800 663**
Bulgaria	**8**	**No data**
Croatia	**5**	**213 396[6]**
Cyprus	**1**	**289 100**
Czech Republic	**13**	**108 292 1 964 750**
Denmark	**4**	**475 082 1 822 569**
Estonia	**2**	**148 872 521410**
Finland	**4**	**196 096 1 224 107**
France	**24**	**92 633 11 532 409**
Germany	**35**	**65 242 5 302 179**
Greece	**9**	**73 434 4 013 368**
Hungary	**9**	**No data**
Ireland	**5**	**84 489 1 534 426**
Italy	**32**	**99 887 3 419 287**
Latvia	**2**	**131 788 1 003 949**
Lithuania	**3**	**160 656 709 870**
Luxembourg	**1**	**451 600**
Malta	**2**	**370 704**
Netherland	**14**	**158 883 1 443 258**
Norway	**6**	**64 303 1 090 513**
Poland	**27**	**82 539 2 660 406**
Portugal	**9**	**111 782 2 435 837**
Romania	**14**	**72 600 2 140 194**
Slovakia	**8**	**111 419 601 132**
Slovenia	**2**	**319 426 495 101**
Spain	**24**	**167 036 5 804 829**
Sweden	**8**	**139 588 1 860 872**
Switzerland	**4**	**91 437 1 116 089**
United Kingdom	**26**	**253 500 11 917 000**
Total	313	

2.2.3 Documentation (National Reports)

The National Reports consist in a specific document sent by the NUAC (National UA Correspondent) at each UA round, which describes the general context and the specific conditions of the data collection. Even if National Reports are very different from one to another (for instance, the shortest is only 6 pages and the largest 124 pages), one can recognize more or less 4 main fields:

- *Overviews*: a general description of the data collection.

- *Spatial units description*: they are described for each of the three definition levels (City-Core, Sub-City Districts and LUZ). Usually the specifications of the LUZ definition are presented in this part of the Report.

[6] For Croatia, LUZ population is available only for one LUZ on Eurostat website (http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?sort=1&file=data%2Furb_vluz.tsv.gz).

- *Indicators descriptions*: this part may contain lists of data available, meta information, or quality aspects of variables.

- *Conclusion*: some recommendations for improvement or other particular aspects can be mentioned there.

The availability of National Reports by country depends not only on each Urban Audit round but also on particular cases (for example Czech Republic and Ireland didn't send National Report for UA III). A general Table sums up these national differences (Table 3). Every National Report can be uploaded for consultation at the Communication & Information Resource Centre Administrator (Circa) of Eurostat.

Table 3: Availability of National Report per country and per Urban Audit round

Countries	UA II	UA III	Countries	UA II	UA III
Austria	✓	✓	**Ireland**	✓	✘
Belgium	✓	✓	**Italy**	✓	✓
Bulgaria	✓	✓	**Lithuania**	✓	✓
Switzerland		✓	**Luxembourg**	✓	✓
Cyprus	✓	✓	**Latvia**	✓	✓
Czech Republic	✓	✘	**Malta**	✓	✘
Germany	✓	✓	**Netherland**	✓	✓
Denmark	✓	✓	**Norway**		✓
Estonia	✓	✓	**Poland**	✓	✓
Spain	✓	✓	**Portugal**	✓	✓
Finland	✓	✓	**Romania**	✓	✓
France	✓	✓	**Sweden**	✓	✓
Greece	✓	✓	**Slovenia**	✓	✓
Croatia		✓	**Slovakia**	✓	✓
Hungary	✓	✓	**United Kingdom**	✓	✓

3 Expertize methodology

3.1 Documentation data process

3.1.1 Expertize of National Reports

National Reports constitute the main source of information on LUZ specifications. They have been analyzed through different steps:

- Identification of coherence between the different phases of UA (which countries do participate, which cities, which LUZ?)

- Research of regularities in the national descriptions of the process, in order to point out categories that may be systematically extracted (city implications, LUZ evolutions…)

- Identification of the degree of completeness of the information regarding these categories.

- Research of correspondence with the geometric sources of LUZ, provided by GISCO (Geographical Information System of the European Commission[7])

During these processes, different problems have been raised and have made the work highly complex. Some of them are purely formal (for example the National Reports which are not translated in English, a problem that we have naturally solved for France and Belgium but not for Norway). Some others are related to the content (for instance, allusive description of LUZ definition, or lack of explanation concerning LUZ definition) (Insert 2).

[7] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

Insert 2: Main problems encountered in National Report expertise (see country-sheets in Annex for more details)

- National Reports or technical annexes are not written in English (France, Belgium, Norway) ;
- LUZ definitions refer to a national zoning whose construction rules are not specified (for example the *Stadsgewest* of the Netherlands, the *Stobyregioner* of Norway or the *Local Labour Market areas* of Sweden, resulting from a collaboration with Eurostat).
- LUZ building blocks are described in allusive terms (for example the *Amter* of Denmark).
- The National Report does not give any information about LUZ definition (for example in Luxembourg).
- There is no National Report (Ireland, Czech Republic).
- Building Blocks mentioned in National Reports are difficult to compare with the last version of NUTS or LAU available at European scale. Sometimes the version of LAU 2 is not the same, or LAU 1 are not available for the related country.
- Information given in National Report is sometimes in contradiction with information given in GISCO (for example when cities have their City Core inside another LUZ, see Wirral in United Kingdom).

3.1.2 Expertize of other documentation

About one dozen of National Reports contained a fully specified description of LUZ definitions that did not need to be completed by other documentation (Austria, Bulgaria, Croatia, Cyprus, Estonia, Italy, Poland, Portugal, Romania, United Kingdom…).

For the rest of National Reports, we used other sources of information in order to try to fill the lack of information. Different situations occurred:

Situation 1: LUZ is defined as a NUTS or LAU proxy[8].

We have then checked this information, using GISCO (Austria, Cyprus, Estonia, Latvia, Lithuania, Denmark, Slovakia) or search more information, especially when some reports or urban analyses were mentioned in the National Report for justifying the NUTS or LAU proxy (Austria, Latvia, Lithuania, Slovakia).

[8] See Section 4.1.1

Situation 2: LUZ is defined using a national zoning not fully described in the National Report.

We have then tried to find the documents giving the precise methodology on National Statistical Boards websites (for example, the *région urbaine* of Belgium, the *aire urbaine* of France, the *agglomeration* of Switzerland, the *planungsregionnen* of Germany, the *stadsgewest* of Netherlands, the *storbyregioner* of Norway). For Netherlands and Norway, some documents were found but not in English.

Situation 3: LUZ is defined using an allusive reference to a national zoning definition.

Again, National Statistical Boards websites were visited (for example concerning "labour market areas" in Sweden or « suburban communes » in Romania). Another complementary method consisted to find information on national functional areas (Statistical Boards websites, publications) and to compare them to the LUZ perimeters (for example in Sweden).

Situation 4: LUZ is not defined, or in a very imprecise way.

We have then contacted Urban Audit Team (Teodora Brandmuller) and received in most of the cases some additional information (for example Greece, Spain, Finland).

3.2 Construction of the general syntax

Different fields have been chosen for describing, in a common language and syntax, the rules used by each country to construct their LUZ. These fields and their content are described in Insert 3.

Insert 3: Model of the Country-Sheet

NAME OF THE COUNTRY
Summary
◆ Number of cities with a LUZ in UA III and previous UA ■ Changes in LUZ definitions between UA II and UA III ● Description of LUZ delineation rules in UA III
Building blocks
◆ Identification of the building blocks of the LUZ ■ Links between building blocks when LUZ is an aggregation ● LUZ capital specificity (relatively to the other national LUZ)
Particular cases
Particular observations, not concerning LUZ capital
Correspondence with GISCO
Coherence between National Reports and GISCO shape files (number of LUZ, names etc.)
References
Documentation used for building the sheet (National Reports and other sources)

3.3 Unsolved problems

In some cases, it was not possible to achieve the expertise because of a remaining lack in information. In these cases, the country-sheet is not complete but missing information is specifically underlined. The most striking cases are:

- Netherlands and its "Stadsgewesten" (only a rough summary of the definition was found in English)

- Hungary and its LAU1 (statistical entities but without information on aggregation criteria)

- Sweden and its "Local Labour Market Areas" (defined in collaboration with Eurostat but we did not found more information)

- Finland and its LUZ (no information on LAU 2 aggregation criteria).

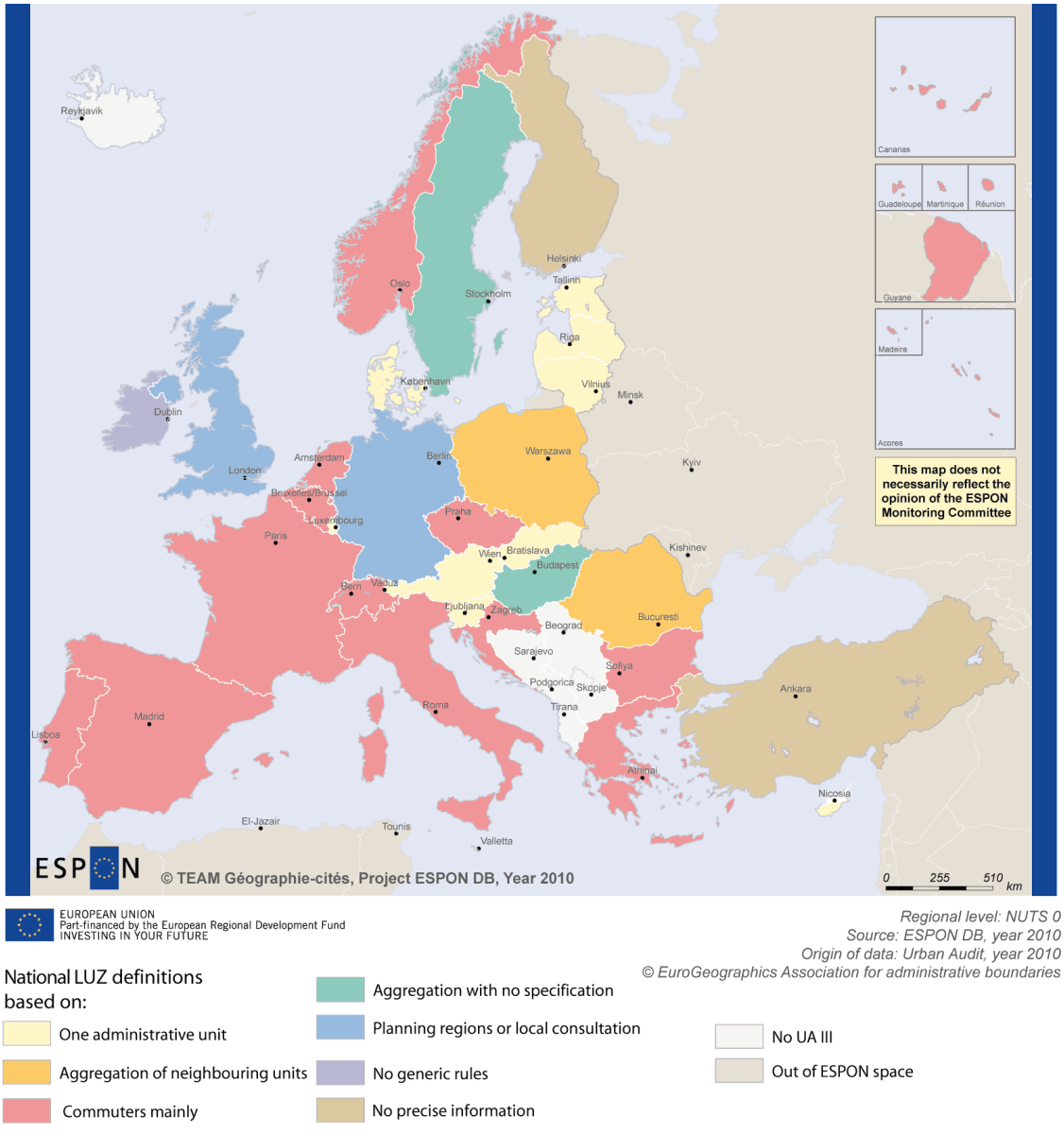
4 Results

After the construction of the 30 country-sheet specifications, different analyses have crossed the results in order to qualify in a synthetic way some national specificities in LUZ delineations.

4.1 Typology of LUZ delineations

A first synthesis of the 30 country-sheets is related to their type of definitions (Figure 5).

Figure 5: Typology of LUZ delineation (UA III)



4.1.1 LUZ as one elementary administrative unit

A first approach characterizes countries that use a sole administrative unit as a LUZ, for example one LAU 1 (Cyprus or Estonia) or one NUTS 3 (Austria or Slovenia). Since these LUZ do not result from an aggregation of LAU, they are not considered as functional in our typology, even if most of the countries mention in their National Report some previous statistical analyses on commuters data that would justify the choice of such or such specific administrative level unit as a proxy (for example Estonia, or Slovakia).

4.1.2 Aggregations of neighbouring units

In two countries, Poland and Romania, LUZ consist in an aggregation of neighboring units, mainly based on distance or contiguity criteria. In Poland, these criteria are completed by hierarchical criteria (the extent of the ring depending on the City-Core size), and in Romania, it is completed by juridical criteria (the selected surrounding LAU 2 must be qualified as "urban" according to a former law, see the country-sheet in Annex).

4.1.3 Aggregations mainly based on commuting data

In about half of European countries, LUZ consist in functional aggregations mainly based on commuting data. However, we have to precise that construction methods are extremely different from one country to another: some take into account only commuters patterns towards a central pole (for example Greece or Croatia), whereas other also consider commuters patterns towards surrounding areas (for example Italy). Sometimes, school commuting are included (see Belgium), population growth (see Switzerland) or transport infrastructures (see Netherlands). In Portugal, the method is first based on LAU 2 commuter levels, then on LAU 1 commuter levels. And we will see in the next sub-section that, even if we examine only commuter thresholds, a great variety of situations appears. Nevertheless, this very rich set of functional approaches is worth studying, as it reflects results of national expertise on "what is a functional urban area" in such or such country.

4.1.4 Aggregation with no specification

In two countries, Sweden and Hungary, LUZ consist in an aggregation of administrative units but the criteria are not specified in the National Reports. In Hungary, for example, the LAU 1 is called a "statistical sub-region" and consists itself in an aggregation of elementary units based on functional criteria (see country-sheet in Annex).

4.1.5 Planning regions or local consultations

In Germany and United Kingdom, LUZ constructions are based on consultations at local or regional levels. LUZ correspond to "Planning regions" in Germany, whereas in United Kingdom "*Office for National Statistics sought the recommendation of relevant Local Authorities and Government Office Regions when constructing the LUZ area for each of the 24 cities under analysis*" (see country-sheets in Annex).

4.1.6 No generic rules

Ireland constitutes a particular case, in the sense that several approaches are used in this country (on a total of 5 LUZ, two of them are based on commuters and two others are based on planning regions) (see country-sheet in Annex).

4.1.7 No precise information

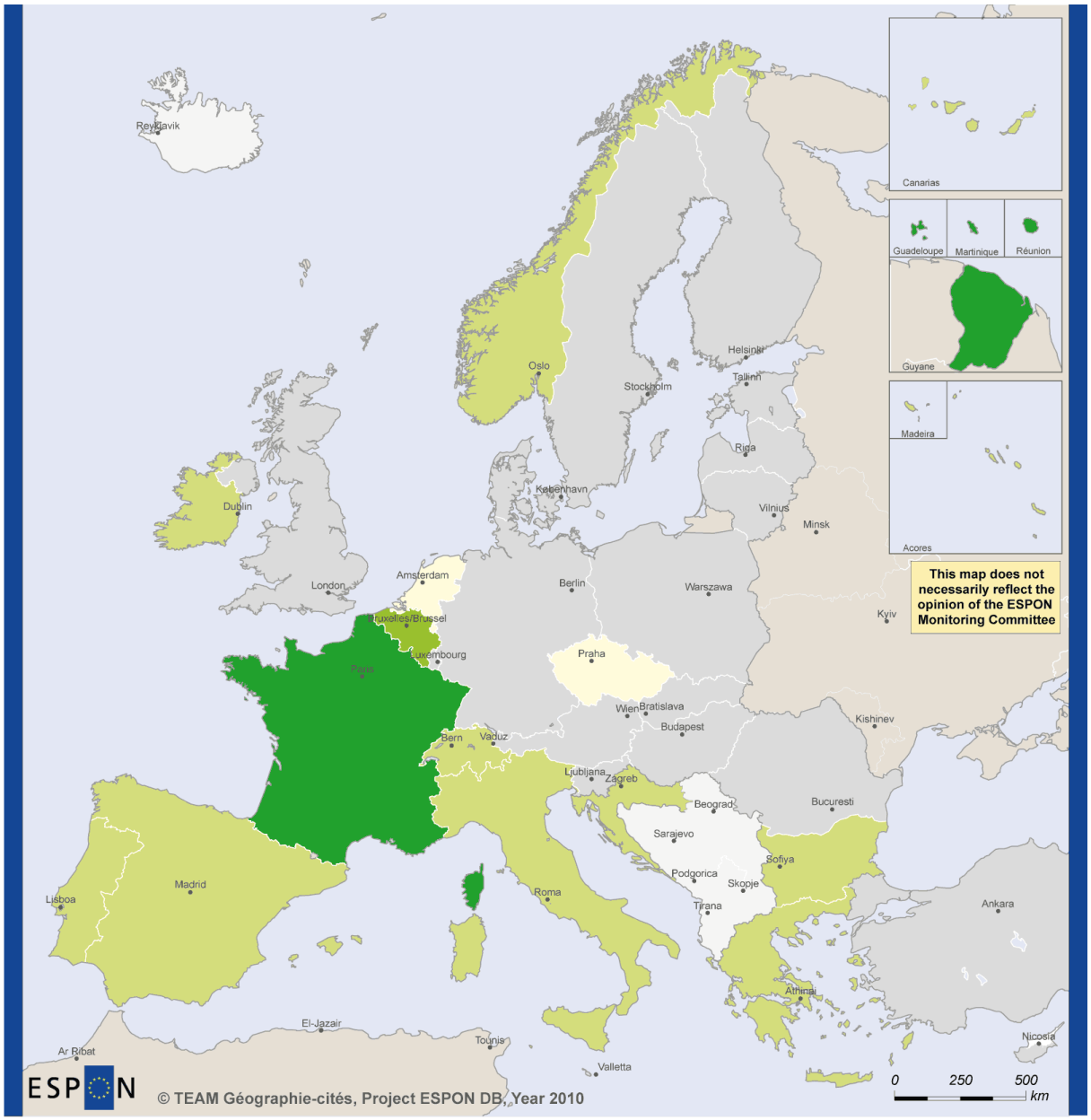
In Finland, the information given in the National Report or other sources collected by us is too incomplete about the way LUZ are constructed (see country-sheet in Annex).

4.2 Diversity of commuting thresholds

When the National Reports or other sources clearly mention the commuter thresholds used in functional definitions of LUZ, it is possible to display these data as values of a quantitative variable (Figure 6). The results seem to be very random and chaotic: countries with similar values are not located in the same part of Europe, and the map does not enlighten a general gradient or center-periphery structure or other macro-structure. Another possible explanation could be represented by the average size of administrative units (variation of thresholds corresponding to variation in LAU surfaces), but we did not find any statistical relations between these two variables. A more specific study should be addressed to the national researches that sustained the choice of these functional criteria in the different countries[9].

[9] In France, for example, this context is given by Thomas Le Jeannic (1996), « Une nouvelle approche territoriale de la ville », INSEE – *Economie et Statistique n°294-295*.

Figure 6: A European variety of LUZ commuting thresholds



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: Urban Audit, year 2010
© EuroGeographics Association for administrative boundaries

Level of commuters in LUZ functional approaches (%)

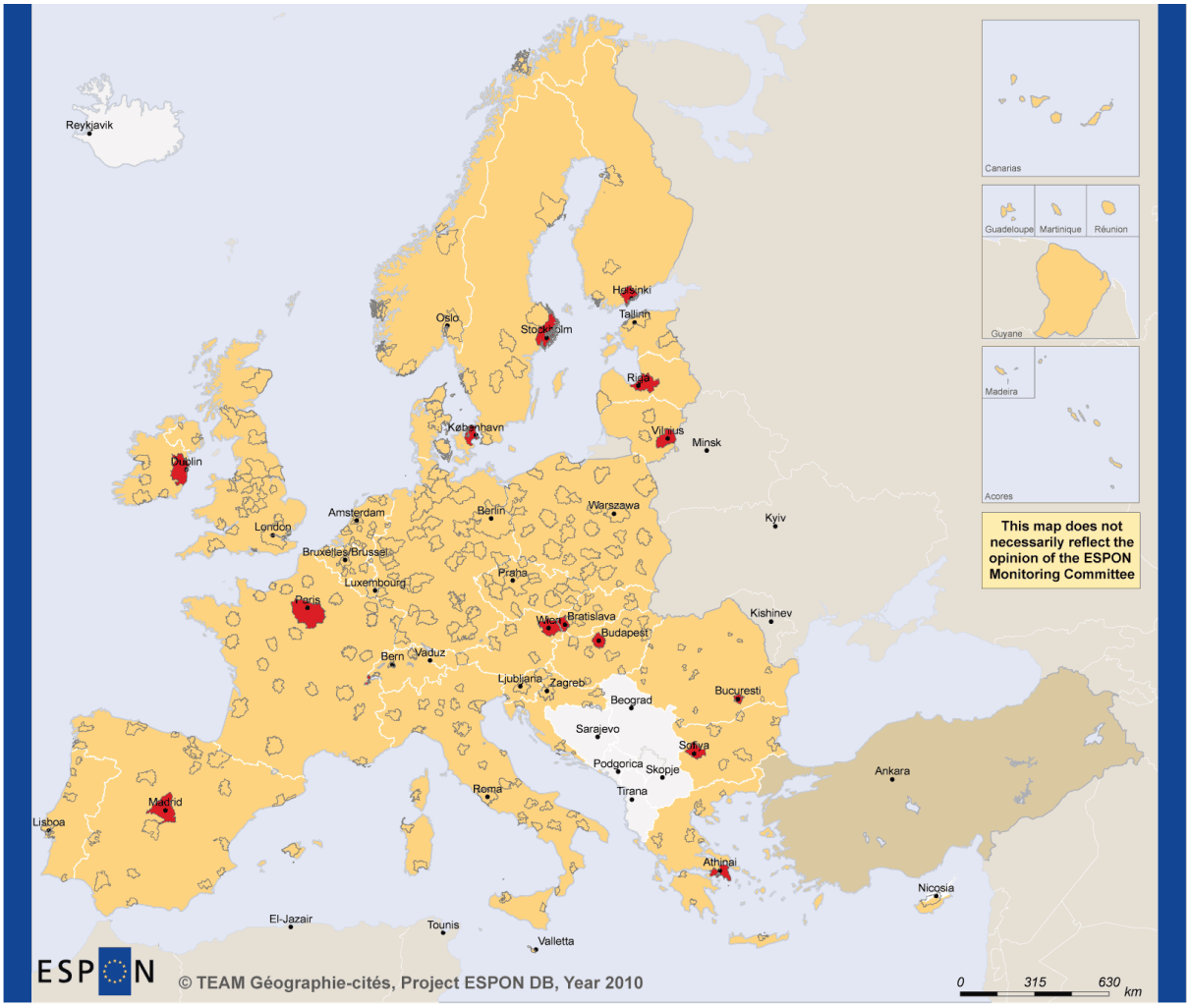
- 10 - 20
- 21 - 30
- ≥31
- no mentioned thresholds
- Other LUZ definition or no information
- No UA III
- Out of ESPON space

4.3 Capital cities as particular cases

In more than half of the European countries, the Capital cities constitute a particular case of the LUZ perimeters (Figure 7). The aim is generally to take into

account a particular influence field of this city (see for example Bucharest), or to fit better to a large administrative unit for questions of data availability (see for example the case of Paris). In most of the cases, NUTS perimeters are used : 4 Capital city LUZ fit with a NUTS 3, one fits with a NUTS 2, 5 are based on an aggregation of NUTS3, and 5 other are based on another type of aggregations (see country-sheets in Annex).

Figure 7: Intra-national heterogeneity in UA III definitions



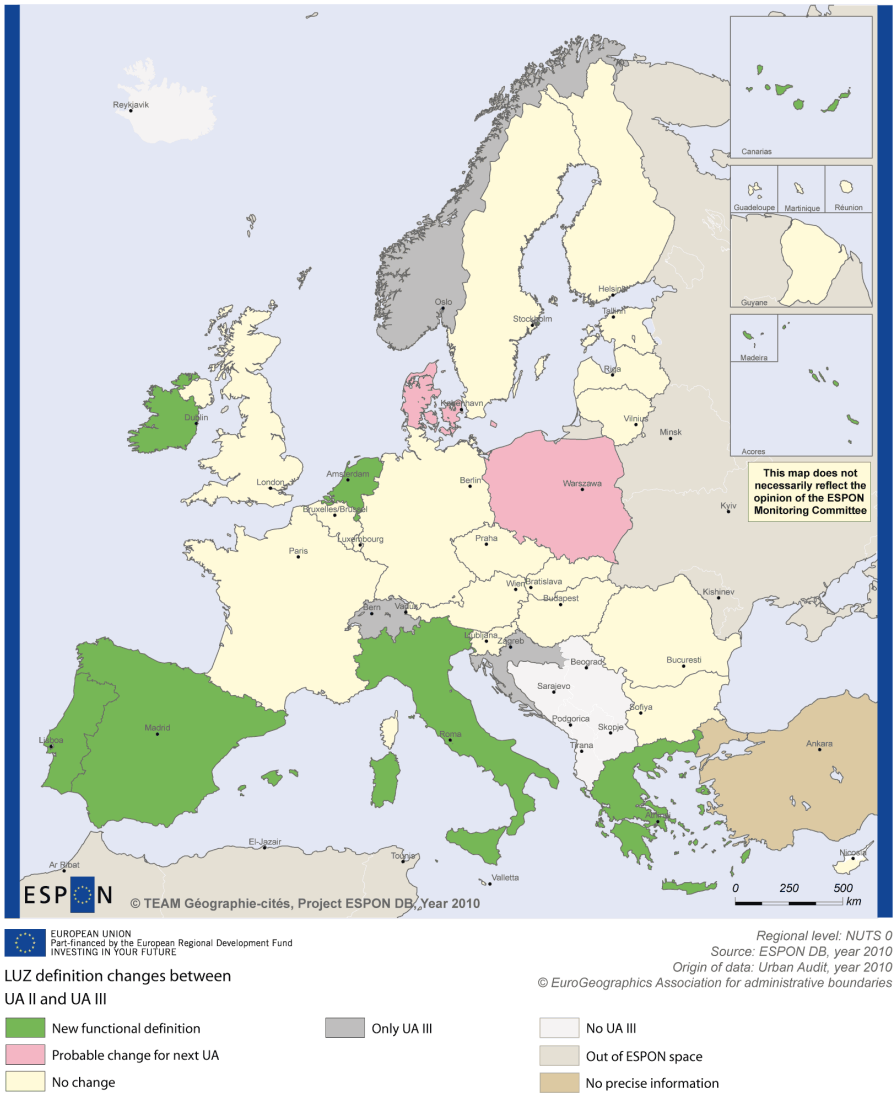
4.4 Towards more functional approaches

One of the fields that have been filled in the country-sheets concerns the evolution of LUZ definition between UA II and UA III. The results enlighten a

clear dynamic towards more functional approaches, largely encouraged by Urban Audit and Eurostat[10].

Between UA II and UA III, six countries have changed to adopt functional definitions (Portugal, Spain, Greece, Italy, Netherlands, Ireland), and two countries have announced in their National Report or annex documentation that they would probably adopt new definition for the next Urban Audit (Poland and Denmark). It is also noteworthy that all the new participant members to UA III, which are not represented on this map because not concerned by evolution since UA II, i.e. Croatia, Norwegian and Switzerland, have all adopted functional definition (mainly based on commuters). We have no information for Turkey.

Figure 8: New LUZ functional definitions



[10] Torbiörn, Carlquist. « The Larger Urban Zones in the Urban Audit data collection. » *Globalisation impact on Regional and Urban Statistics*. Wroclaw, 2006.

5 Conclusion

As presented in this technical report, the work consisting in collecting and analyzing documentation on LUZ specifications was very complex but helped to re-write in a common way the different rules used by each country to define its LUZ. This work is just a step in the whole process, for three different reasons:

- It will have to be updated when the results of the next Urban Audit will be published
- Some explaining factors still have to be explored (for example, concerning the great variety of commuter thresholds in Europe)
- Some country profiles (country sheets) are not yet complete. Other documentation has to be found or to be translated in English.

Anyway, the different maps and the typology allowed to have a global overview and to enlighten different results. The large heterogeneity in the national approaches used to define LUZ engages researchers to be very cautious when interpreting some statistical results. But they also enlighten a very interesting evolution between UA 2001 and UA 2004, towards more functional definitions mainly based on commuters, even if the criteria (commuter thresholds, for instance) are very different from one country to another.

It confirms again that harmonization in definitions must not be only guided by the research of a unique rule and criteria for the whole Europe but must be based firstly on a good knowledge of the regional differences in settlement contexts and secondly on the political and historical ways each country defines cities. These differences are not an obstacle to harmonization when the metadata are fully specified.

Annex: LUZ specifications by country (UA III)

Model of the Country-Sheet

NAME OF THE COUNTRY
Summary
◆ Number of cities with a LUZ in UA III and previous UA ■ Changes in LUZ definitions between UA II and UA III ● Description of LUZ delineation rules in UA III
Building blocks
◆ Identification of the building blocks of the LUZ ■ Links between building blocks when LUZ is an aggregation ● LUZ capital specificity (relatively to the other national LUZ)
Particular cases
Particular observations, not concerning LUZ capital
Correspondence with GISCO
Coherence between National Reports and GISCO shape files (number of LUZ, names etc.)
References
Documentation used for building the sheet (National Reports and other sources)

AUSTRIA

Summary

◇ Austrian cities have been included in Urban Audit since the Pilot Phase (Wien, Graz) **(1; 2)**, then in UA II (Linz) **(2; 3)**, and in UA III (Salzburg, Innsbruck) **(4; 5)**. In total, five cities are concerned.

□ LUZ definitions have not changed between UA II and UA III ("*The spatial units of Vienna and Graz were the same as in the data collection 2001*" **(4)**).

○ Each LUZ corresponds to one NUTS 3 (except for Wien), not only in UA II ("*The proposal for the larger urban zones was to take NUTS 3- regions to get a functional urban region*" **(2)**) but also in UA III ("*For the new cities Salzburg (AT004C) and Innsbruck (AT005C) the Larger Urban Zone is the surrounding NUTS3-Region*" **(4)**). Apparently some commuting data have been used previously, but no details are provided in National Reports ("*For the definition of the larger urban zone of a city we used commuting data*" **(2)**).

Building blocks

◇ NUTS 3, elementary administrative unit (i. e. *Gruppen von PolitischenBezirken*)

□ No aggregation

○ Wien: NUTS 3 aggregation (no details concerning the nature of the links)

Particular cases

Linz: The LUZ "*is not the best solution, because in the NUTS3-region of Linz is another big city called Wels. But Linz and Wels are functionally related, so we decided to take also the NUTS3-region for Linz*" **(2)**.

Correspondence with GISCO

Same number of LUZ[11]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.

2. Schrittwieser, Karin. (Undated). *Urban Audit II Final report for the European Commission*. Wien : Statistik Austria.

3. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.

4. Schrittwieser, Karin. (Undated). *Urban Audit 2004 Austria Final Report for the European Commission*. Wien : Statistik Austria.

5. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.

[11] Files downloaded and checked July 5, 2010

BELGIUM

Summary

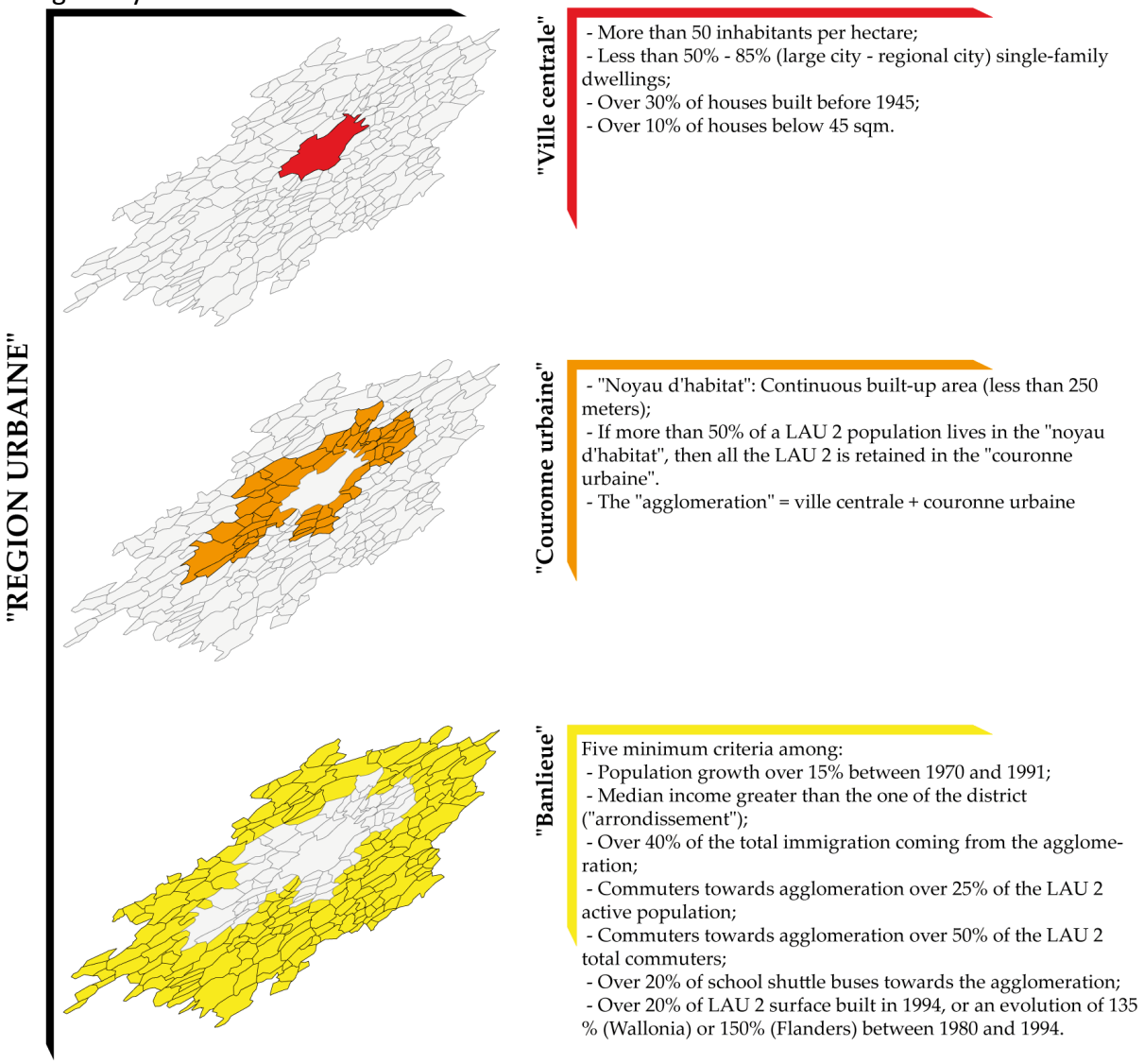
◇ Belgian cities have been included in Urban Audit since the Pilot Phase (Brussels, Anvers) **(1)**, then in UA II (Liège, Charleroi, Gand, Bruges) **(2; 3)**, and in UA III (Namur) **(4)**. In total, seven cities are concerned.

□ LUZ definitions have not changed between UA II and UA III **(5)**.

○ LUZ correspond to the Belgian functional region called « région urbaine » (« *Les « large urban zones » ont été établies à partir de : "Les régions urbaines belges en 1991" »* **(3).** The criteria are functional ([elles] *« reposent sur des critères de fonctionnalité »* **(6*)***. These criteria were formulated in 1996 ***(7)***.

Construction data came from 1991 census **(8)**.

The figure below describes the different steps for building the "régions urbaines", translated in English by us:



Source: ***Van der Haegen, H, Van Hecke, E et G., Juchtmans****.Les régions urbaines belges en 1991. [éd.] Institut national de Statistique. Etudes statistiques. 1996, n°104*

Building blocks
◇ LAU 2, aggregation (i.e. Gemeenten / Communes): “ville centrale + couronne urbaine + banlieue”. □ Links: mainly based on commuters (threshold 25%).
Correspondence with GISCO
Same number of LUZ [12] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1
References
1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes. 2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities. 3. (Undated). *Deuxième rapport sur le projet Urban Audit II*. 4. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities. 5. (Undated). Rapport sur le projet Urban Audit II : Demande de données historiques 1991-1996 et données 2001. 6. Doulou-Ouamba, Marlène. (2008). *Audit Urbain III.* s.l. : Service Publique Fédéral Economie, PME, Classes Moyennes et Energie. 7. Van der Haegen, H, Van Hecke, E et G., Juchtmans. (1996). *Les régions urbaines belges en 1991*. [éd.] Institut national de Statistique. *Etudes statistiques.* n°104. 8. Hermia, J-P. (Undated). *Une nouvelle délimitation spatiale du phénomène périurbain bruxellois*.

[12] Files downloaded and checked June24, 2010

BULGARIA

Summary

◇ Bulgarian cities have been included in UA II (Sofia, Plovdiv, Varna, Burgas, Pleven, Ruse, Vidin), then in UA III (Stara Zagora) **(1; 2; 3; 4)**. In total, eight cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ LUZ are defined with a functional approach. Two steps are followed:
1st step: "*Finding the settlements (LAU2* [i.e. Naseleni Mesta]*) that have more than 20% commuters (out of the employed residents) in both directions to the City Core (incoming and outgoing migration are added)*".
2nd step: "*Aggregating the data for the respective municipality (i. e. LAU 1) and checking again for the criterion*" **(4).** Apparently, this choice is due to lack of data ("*Even though the NUTS5 data is the most suitable for defining the functional urban zones in the case of Bulgaria, due to data availability reasons a combination of NUTS4 units (Obshtini) was chosen*" **(5)**).
Data come from the 2001 Census ("*A particular question in the Census card addresses the location where the individual works/studies and lives*" **(5)**).

Building blocks

◇ LAU 1, aggregation (i.e. *Obshtini*)
□ Links: based on commuting data (threshold 20%).
○ Sofia: Aggregation of one NUTS 3 (i.e. Oblasti) and several LAU 1 ("*The whole NUTS3 level unit (Oblast) was included in the newly formed LUZ, as well as several neighboring NUTS4 territorial units*" **(5)**).

Correspondence with GISCO

Same number of LUZ [13]

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. (Undated). Urban Audit, Phase III - Historic data Interim Report 2006.
3. Eurostat. (2009). *Annuaire régional d'Eurostat 2009.* Luxembourg : Office des publications officielles des Communautés européennes.
4. Tsvetarsky, Serguey. (Undated). *Phare 2005 - Project: Urban Audit IV.* Sofia.
5. Tsetarsky, Serguey et Kotzev, Alexander. (2004). *Eurostat Pilot Phase - Urban Audit II - Phase 1 - Final Report Bulgaria*. Sofia.

[13] Files provided by Urban Audit October 5, 2010

CROATIA

Summary

◇ Croatian cities have been included in UA III (Zagreb, Rijeka, Slavonski Brod, Osijek, Split) **(1)**. In total, five cities are concerned.
□ Croatia did not participate to UA II.
○ LUZ are defined with a functional approach, and the definition was specifically adopted for Urban Audit ("*Larger Urban Zones (LUZ) were created by* CBS [Central Bureau of Statistics] *only for UA needs on the basis of the nearest neighborhood*" **(2)**). Two steps are followed:
1st step: "*Larger urban zones around selected cities have been chosen in the following way: We have investigated for each city which municipalities and towns on the NUTS 5 level have commuting rate over 20% (according to the data from Census 2001)*".
2nd step: "*we have created continuous larger urban zones (no holes or gaps)*" **(2)**.
The construction data come from the 2001 census **(2)**.

Building blocks

◇ LAU 2, aggregation (i.e. *Gradovi i općine*)
□ Links: based on commuting data (threshold 20%)

Particular cases

Split: "*Only one exception was Split LUZ in which municipality of Zagvozd was included even if it has no common border with the other components of the Larger urban zone*" **(2)**.

Correspondence with GISCO

Same number of LUZ [14]

References

1. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
2. Crostat. (2008). *Urban Audit Final Operational Report.* s.l. : Republic of Croatia Central Bureau of statistics.

[14] Files provided by Urban Audit, October 5, 2010

CYPRUS

Summary

◇ Only one city, in UA II and III **(1; 2)** (Lefkosia).
□ LUZ definitions have not changed between UA II and UA III.
○ Lefkosia LUZ corresponds to one LAU 1:
"*The district of Lefkosia was taken as a proxy for the Larger Urban Zone (LUZ) based on data, from the 2001 Population Census, on commuting flows to the core city*" **(1)**: "*in most municipalities and communities in the district of Lefkosia, other than those eight that are considered urban, reside a large percentage of commuters to the urban area of the district (for the purpose of employment): at least 40% (for the majority) when calculated individually*" **(2)**.

Building blocks

◇ LAU 1, elementary administrative unit (i.e. *Eparchie*)
□ No aggregation

Correspondence with GISCO

Same number of LUZ [15]

References

1. (Undated). *Urban Audit 2001 Data Collection Project: Final Report (Reporting country: Cyprus)*.
2. (2008). *Urban Audit 2006/2007 Data Collection Exercise Final Report Member State: Cyprus*.

[15] Files provided by Urban Audit, October 5, 2010.

CZECH REPUBLIC

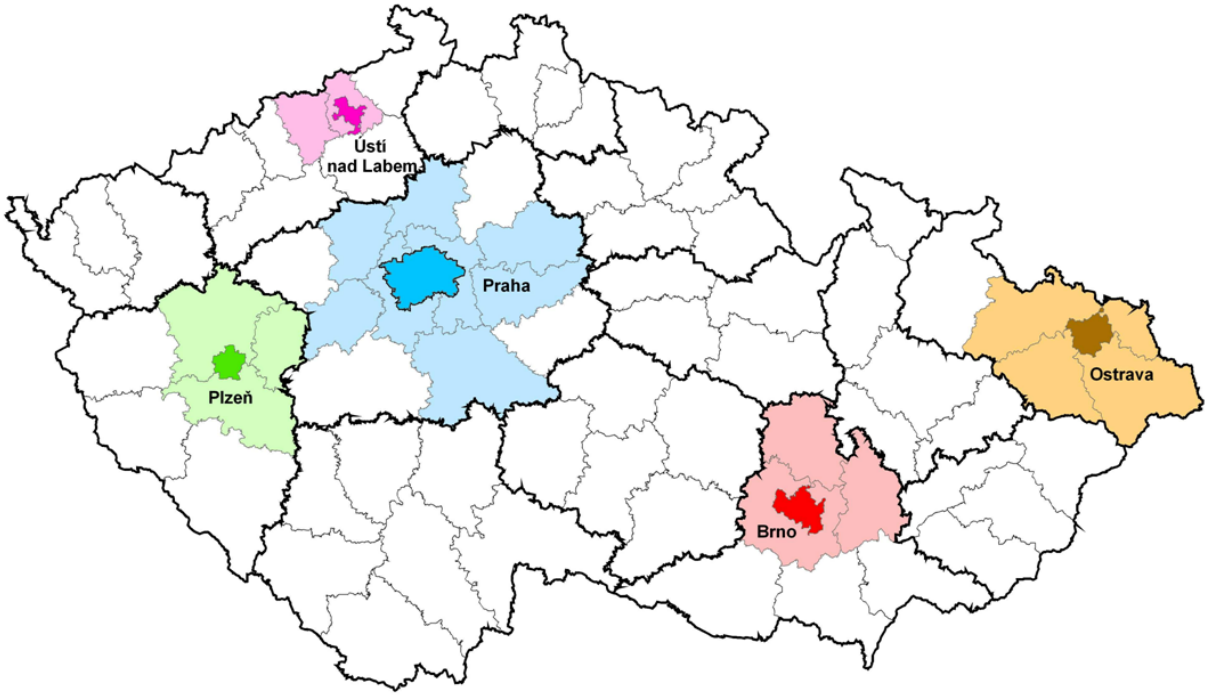
Summary

◇ Czech cities have been included in UA II (Brno, Ostrava, Plzen, Praha, Usti nad Labem) **(1)**, then in UA III (Olomouc, Liberec, Ceske Budejovice, Hradec Kralove, Pardubice, Zlin, Karlovy Vary, Jihlava **(2)**). In total, thirteen cities are concerned.

□ Czech Republic did not transmit National Report for UA III.

○ LUZ are defined in UA II with a functional approach. They correspond to an aggregation of LAU 1 ("*Larger Urban Zones (LUZ) have been defined as groups of LAU 1 units, as the NUTS 3 regions were excessively large*" **(4)**). According to the UA II National Report, LUZ have been built using commuting data from Census 1991, but the methodology is not detailed ("*Delimitation of the adequate functional LUZs was based mainly on the commuting into work and schools from census 1991*" **(5)**). According to the Manual of Data Quality Aspects, the perimeters have been confirmed by Census 2001 ("[LUZ] *have been delimitated as aggregates of LAU-1 units (okresy). (…) Inclusion of individual LAU 1 districts was (…) confirmed by Census 2001, once available*" **(3)**).

The map below gives the delimitation of the five LUZ of UA II. The grey lines correspond to LAU 1 units and the black lines to NUTS 3 **(3)**.



Source: Czech Statistical Office. (2005). *Urban Audit - Czech Republic - Manual of data quality aspects (Phase 1, 2 and historical data).*

Building blocks

◇ LAU 1, aggregation (i.e. *Okrezy*)

□ Links: based on commuting data. Methodology has not been provided in National Report.

Particular cases
- Kladno: the City Core has no LUZ because it is part of Praha LUZ[16].
Correspondence with GISCO
Same number of LUZ [17] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1
References
1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities. 2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities. 3. Czech Statistical Office. (2005). *Urban Audit - Czech Republic - Manual of data quality aspects (phase 1, 2 and historical data)*. 4. Czech Statistical Office. (2005). *Urban Audit - Final Report on Phase 2 - Czech Republic*. Praha. 5. Czech Statistical Office. (2004). *Urban Audit II – Final Report on Phase 1*. Praha.

[16] According to information transmitted by Urban Audit, January 2009.

[17] Files downloaded and checked July 21, 2010

DENMARK

Summary

◇ Danish cities have been included in Urban Audit since the Pilot Phase (Copenhague) **(1)**, then in UA II (Århus, Odense, Aalborg) **(2; 3)**, and in UA III **(4; 5)**. In total, four cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ Each LUZ corresponds to one NUTS 3:
"*The level of the regions* [i.e. Amter] *in which the cities are placed,* [called the larger urban zone]*, (…) is an administrative unit managed by a Council, and is elected every four years*" **(4)**.
These Amters have disappeared on December 31, 2006 but LUZ definitions remained the same for UA III: "*Denmark did not wish to change their LUZ yet. A big regional reform is foreseen for 2007 and in connection with this, new non-administrative NUTS 3 regions will have to replace the old "Amter" or counties*" **(6)**.
Probable changes in LUZ definitions will occur for the next Urban Audit (UA IV).

Building blocks

◇ NUTS 3, elementary administrative unit (i.e. *Amter*).
□ No aggregation
○ Copenhagen: aggregation of several LAU 1 (i.e. *Kommuner*) and several NUTS 3:
"*The capital region has been officially delineated to include the municipality and three counties (Amter):*" **(2)**), i.e, Copenhagen LUZ includes "*(…) the municipality of Copenhagen and Frederiksberg and the regions of Copenhagen, Frederiksberg and Roskilde*" **(4)**.

Correspondence with GISCO

Same number of LUZ [18]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.
2. Danmarks Statistik, Befolkning og Uddannelse. (2004). *Final report concerning agreement N° 2002CE160AT181 about Urban Audit*.
3. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
4. Statistics Denmark, Population. (2007).*Final Report Concerning Agreement N° About Urban Audit 2004 Collection*.
5. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
6. Carlquist, Torbiörn. (2006, August 30). The Larger Urban Zones in the Urban Audit data

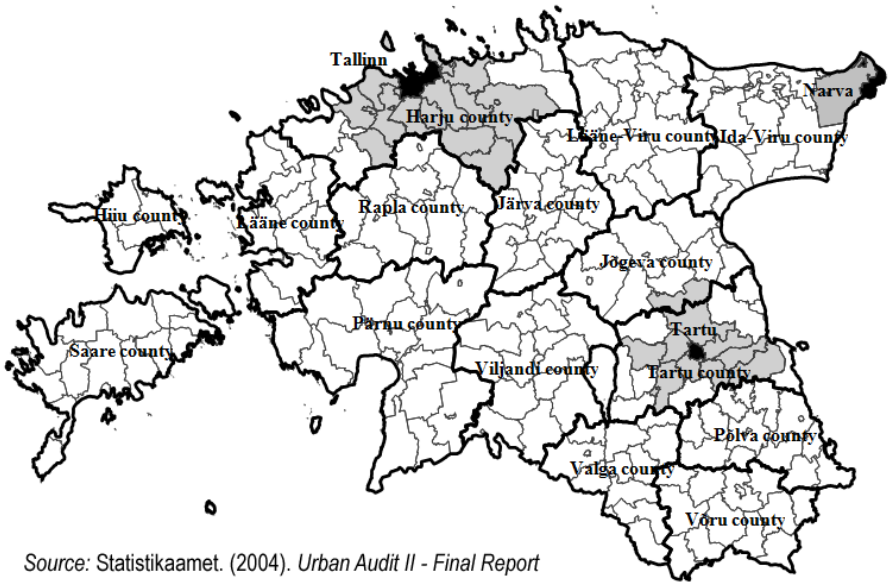
[18] Files downloaded and checked July 5, 2010

collection. *Globalisation Impact on Regional and Urban Statistics*. Wroclaw.
7. Christiansen, Henning. (Undated). *Urban Audit II - State of the Art - Country: Denmark*.

ESTONIA

Summary

◇ Estonian cities have been included in Urban Audit since the second one (Tallin et Tartu **(1)**), and in UA III **(2)**. In total, two cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ Each LUZ corresponds to one LAU 1:
"*In 2002 the working group headed by Professor Jussi S. Jauhiainen of Finland analyzed upon request of the Ministry of Internal Affairs the development potential of Urban Regions of Estonia"* **(3)** . However, it seems that these Urban Regions were not used as LUZ and that LAU 1 were chosen instead. Indeed, a proposal was presented to Eurostat "*to define Harju* [the county that contains Tallinn] *and Tartu counties as LUZs for the above-mentioned cities. The LUZ of Tartu does not exactly overlap with the Urban Region of Tartu specified by Professor Jussi S. Jauhiainen. However, the differences are not very big and considering the administrative concept, both the cities Tallinn and Tartu definitely influence the counties surrounding them*" **(3)**.
The map below gives Urban Regions (grey area) and LAU 1 (bold black line) **(4)**.



Source: Statistikaamet. (2004). *Urban Audit II - Final Report*

Building blocks

◇ LAU1, elementary administrative unit (i.e. *Maakond*)
□ No aggregation

Correspondence with GISCO

Same number of LUZ [19]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

[19] Files downloaded and checked July 21, 2010.

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
3. Statistikaamet. (2004). *Urban Audit II - Final Report*.
4. Statistikaamet. (Undated). *The Report on quality aspects and meta information*.

FINLAND

Summary

◇ Finnish cities have been included in Urban Audit since the Pilot Phase (Helsinki) **(1)**, then in UA II (Tampere, Turku, Oulu) **(2)**, and in UA III **(3)**. In total, four cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ LUZ correspond to an aggregation of LAU 2 and the list of building blocks is given in the UA III National Report **(4)**. However, the rules of aggregation are not described ("*The suitable area divisions for (…) LUZ* (…) were *defined together with Eurostat*" **(5)**).

Building blocks

◇ LAU 2, aggregation (i.e. *Kunnat* / *Kommuner*)
□ Links: No information
○ Helsinki: The LUZ corresponds to "*the functional urban region called the Helsinki Region. This functional region is already used in the NORDSTAT database (Nordic Cities and Regions) and in the Finnish City Indicators database*"[20]

Correspondence with GISCO

Same number of LUZ [21]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.
2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
4. Statistics Finland. (2007). *Eurostat - Urban Audit III - Final Country Report Finland*.
5. Statistics Finland. (2004). *Final Report: Urban Audit II Finland*.

[20] According to information provided by Urban Audit December 3, 2010.
[21] Files downloaded and checked June 24, 2010

FRANCE

Summary

◇ French cities have been included in Urban Audit since the Pilot Phase (Lyon, Toulouse, Strasbourg, Bordeaux, Nantes) **(1)**, then in UA II (Ajaccio, Amiens, Besancon, Caen, Clermont-Ferrand, Dijon, Grenoble, Le Havre, Limoges, Metz, Nancy, Orleans, Paris, Poitiers, Reims, Rennes, Rouen) **(2)**, and in UA III (Tours) **(3)**. In total, thirty-three cities are concerned.

□ LUZ definitions have not changed between UA II and UA III.

○ LUZ are defined with a functional approach. They correspond to the French definition of «*aire urbaine* » (“*Ce sont donc les aires urbaines qui ont été utilisées pour représenter les LUZ de l’audit*” **(4)**). Criteria were formulated in 1996 and the construction data come from the 1999 Census **(7)**.

The figure below describes the different steps for building the “aires urbaines”, translated in English by us:

Source: http://www.insee.fr/fr/regions/auvergne/default.asp?page=themes/donnees_detaillees/aireurbaine/aire-construc.htm

Building blocks
◇ LAU 2, aggregation (i.e. *Commune*) □ Links: based on commuting data (threshold 40%) ○ Paris: NUTS 2 (i.e. *Régions*) ("*Une simplification sur Paris : la LUZ est la région Ile de France. L'aire urbaine n'est guère différente en étendue et les objectifs de l'audit pour la capitale étaient de toutes manières particuliers*" **(4)**).
Particular cases
Pointe-à-Pitre, Fort de France and Cayenne: these cities are not delineated as « *aire urbaine* », so that they don't correspond to a Larger Urban Zone. This is the case for all DOM-TOM cities, as "aires urbaines" have been only defined for Metropolitan France *("Pas de LUZ, là où le concept d' « aire urbaine » n'avait pas été mis en œuvre. Au moment de la collecte, c'était le cas des départements d'outre-mer dont l'étendue territoriale des communes fait qu'il est difficile de reproduire le concept en place sur la métropole."* **(4)**); Saint-Etienne, Marseille, Nice: these cities don't correspond to a LUZ as their City Core is too different from the urban pole of "aire urbaine" called "unite urbaine". Indeed, City Core in France is defined as an EPCI[22] (*"Pas de LUZ, là où les EPCI différaient trop des agglomérations morphologiques. En effet le concept d'aire urbaine est conçu comme la zone d'influence d'un noyau qui est pris a priori comme l'unité urbaine, c'est à dire l'agglomération morphologique. (…) Saint-Etienne, dont l'EPCI est plus étendue que l'agglomération et absorbe même une seconde agglomération ; Marseille, et Nice dont les agglomérations sont chacune partagées sur trois EPCI, dont le rayonnement est donc nettement inférieur à l'aire urbaine »* "**(4)**). Lens: this city does not correspond to a LUZ, due to deviations between City Core and « unité urbaine » ("*Douai et Lens ne forment en fait qu'une seule agglomération, mais dont le contour est significativement différent du seul regroupement des deux EPCI*"**(4)**) Aix-en-Provence, Lille, Montpellier and Toulon: these cities does not correspond to a LUZ ("*The cities of (…) Aix-en-Provence (…) do not have any LUZ*" **(6)**); For Lille and Montpellier, "*LUZ are removed in 2007*", and for Toulon, "*LUZ is removed in 2008*"[23].
Correspondence with GISCO
Different number of LUZ: Toulon's LUZ is not removed in GISCO[24]
References
1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes. 2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities. 3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.

[22]Etablissement Public de Coopération Intercommunale: Political aggregation of LAU 2, especially for elaborating planning projects.

[23]According to information provided by Urban Audit, January 2009.

[24] Files provided by Urban Audit, October 5, 2010.

4. INSEE. (Undated). *Audit urbain 1999-2003, Bilan de collecte*.
5. INSEE. (2010). *Insee - Définitions et méthodes - Aire Urbaine*. Checked June 25, 2010, on Institut National de la Statistique et des Etudes Economiques *:*
http://www.insee.fr/fr/methodes/default.asp?page=definitions/aire-urbaine.htm.
6. INSEE. (Undated). *Audit Urbain 2006, Bilan de collecte*.
7. *Mesurer un univers urbain en expansion.* Julien, Philippe. (2000). Economie et Statistique n°336, pp. 3-33.

GERMANY

Summary

◇ German cities have been included in Urban Audit since the Pilot Phase (Berlin, Hamburg, Munich, Cologne, Frankfurt am Main, Essen, Stuttgart, Leipzig, Dresden) **(1)**, then in UA II (Augsburg, Bielefeld, Bochum, Bonn, Bremen, Darmstadt, Dortmund, Düsseldorf, Erfurt, Frankfurt (Oder), Freiburg-im-Breisgau, Gottingen, Halle-an-der-Saale, Hannover, Karlsruhe, Magdeburg, Mainz, Moers, Monchengladbach, Mulheim-an-der Ruhr, Nurnberg, Regensburg, Schwerin, Trier, Weimar, Wiesbaden, Wuppertal) **(2)**, and in UA III (Kiel, Saarbrucken, Koblenz) **(3; 4)**. In total, forty cities are concerned.

□ LUZ definitions have not changed between UA II and UA III ("*Delineation of the territorial units of the cities of the previous rounds of data collection remained unchanged, i. e. of the 35 cities, the 28 Larger Urban Zones*" **(3)**).

○ Each LUZ corresponds to a « planning region », which is defined through a variety of criteria[25]**(6)**. According to Klaus Trutzel (National Urban Audit Coordinator), "*at this level, the statistical offices of the Länder can provide a particularly wide variety of planning policy data, not all of which are available for individual Kreise (NUTS 3)*" **(5)**. It is also noticed that "*the size of the LUZ is about adequate although the labour market area of the city does not exactly match the LUZ. It might be worth comparing the real commuting areas, delimited by LAU2 units, with the LUZ for better judgments of the adequacy of the LUZ and the information collected for them*"**(3)**.

In 2002/2003, Eurostat checked the correlation between LUZ and functional area (defined with a threshold of 15% and 20%). The results show that LUZ fit quite well with functional areas **(7)**, except for Bielefeld and Schwerin[26]. However, according to Klaus Trutzel, these latter remain unchanged "for comparability needs" **(3)**.

Building blocks

◇ NUTS 3, aggregation (i.e. *Kreis*)

□ Links: variety of criteria used to define "planning regions", not specified in National Reports.

Particular cases

- Essen, Dortmund, Bochum, Mulheim and Moers: One common LUZ, Ruhrgebiet, has been defined for these five cities in the Ruhr area ("*The common LUZ for 5 Urban Audit cities in the Ruhr area was kept*" **(3)**) ;
- Monchengladbach, Wuppertal and Frankfurt (Oder): these cities do not correspond to a LUZ ("*The 3 cities of Monchengladbach, Wuppertal and Frankfurt (Oder) have no LUZ around*

[25] The "planning regions" are defined at the federal level by the BBR (Bundesamtfür Bauwesen und Raumordnung). They are transmitted to the Lander which adapt them taking into account the large-scale projects proposed by local authorities. The number of planning regions is officially 97. However, some of them are subdivided by the Lander into several sub-regions. That's why the number of 115 planning regions can be found sometimes **(6)**.

[26] The comparison was made between the population of the LUZ and the population of the estimated functional areas.

them" **(3)**);
- Potsdam: this city does not correspond to a LUZ because it is part of the LUZ of Berlin[27].
- Stuttgart: this city was removed in UA II but added again in UA III **(5)**

Correspondence with GISCO

Same number of LUZ[28]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.
2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
3. Trutzel, Klaus. (2008). *Urban Audit III (2006): Final Country Report Germany.*
4. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
5. Trutzel, Klaus. (Undated). *URBAN AUDIT - State of the art*.
6. Queva, Christophe. (2007). « Les paradoxes de la Région en Allemagne, entre réseaux et territoires : la région, outil de déterritorialisation ? » *Annales de Géographie* (653).
7. Carlquist, Torbiörn. (2006, August 30). The Larger Urban Zones in the Urban Audit data collection*. Globalisation Impact on Regional and Urban Statistics*. Wroclaw.

[27] According to information transmitted by Urban Audit, January 2009.

[28] Files downloaded and checked July 21, 2010

GREECE

Summary

◇ Greek cities have been included in Urban Audit since the Pilot Phase (Athina, Thessaloniki et Patras) **(1)**, then in UA II and UA III (Volos, Iraklion, Kavala, Kalamata, Ioannina et Larisa) **(2; 3)**. In total, nine cities are concerned.

□ LUZ definitions have changed between UA II and UA III. For UA II, NUTS 3 were used as proxy ("*NSSG* [National Statistical Service of Greece] *decided to use this NUTS3 level as a proxy for the large urban zone*" **(5)**).

○ In UA III, a new definition is based on labour market areas ("*There was a new delimitation of the existing LUZ areas of the 9 cities according to the Labour Market Areas and the suggestion of Eurostat*" **(4)**). LUZ are defined with a functional approach, but the methodology is not specified in the National Report. According to documentation sent by Urban Audit[29], commuting data have been used at LAU 1 level, with a threshold of 15% people commuting from a "suburb" to the central city (demos). Construction data come from 2001 Census.

Building blocks

◇ LAU 1, aggregation (i.e. *Demoi*/*Koinotites*)

□ Links: apparently based on commuting data (threshold 15%).

○ Athina: the LUZ corresponds to one NUTS 3, elementary administrative unit (i.e. *Nomoi*). Some outlying islands have been excluded.

Correspondence with GISCO

Same number of LUZ [30]

http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.

2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.

3. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.

4. General Secretariat of the National Statistical Service of Greece. (2008). *Urban Audit 2006/2007 Data Collection - Final Operation Report.*

5. (Undated). *Urban Statistics - Final Report*.

[29] According to information transmitted by Urban Audit, January 2009.

[30] Files downloaded and checked July 5, 2010

HUNGARY

Summary

◇ Hungarian cities have been included in UA II (Budapest, Miskolc, Pecs, Nyiregyhaza **(1)**) and in UA III (Debrecen, Szeged, Győr, Kecskemét, Székesfehérvár**(2)**). In total, nine cities are concerned.
□ LUZ definitions have not changed between UA II and UA III **(3)**.
○ LUZ correspond to LAU 1 aggregation, and the methodology is not fully specified:
"*LUZs consist of the NUTS 4 level units, i. e. statistical subregions of these cities. A statistical subregion is primarily a functional unit, established on the basis of actual working, residential, transport and secondary provisional (education, health care, and trade) connections between the central city and the urban zone around*" **(4)**.

Building blocks

◇ LAU 1, aggregation (i.e. *Statisztikaikistérségek*)
□ Links: the methodology is not fully specified (see above)
○ Budapest: LAU 2, aggregation (i.e. *Települések*). The methodology is partly specified:
"*In the case of Budapest, the Larger Urban Zone is made up of the 79 settlements of the legally defined Budapest agglomeration, where the 78 settlements surrounding the capital are tightly connected with the centre. A part of the settlements in the urban zone show the morphological aspects too of an agglomeration. In 25 settlements, more than 50% of the resident population live near – the incidental distance is not more than 200 meters in the corresponding categories – urban (built-in) areas of the capital, although a morphological connection to a lesser extent can be observed in other settlements as well. However, there is a functional relation in the case of all settlements, that is why the former agglomeration, which had consisted of 44 settlements, was extended. The Budapest agglomeration with today's boundaries and the list of the settlements included were most recently published in the Government Decree Nr. 89/1997. (V.28.). A specific feature of the past decade in Hungary was that suburbanisation started in the proximity of Budapest too. The migration loss of the capital was over 100 000 persons, and a significant proportion of those who moved out of Budapest went to live in settlements of the agglomeration*" **(4)**.

Correspondence with GISCO

Same number of LUZ [31]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.

[31] Files downloaded and checked July 19,2010

3. Hungarian Central Statistical Office. (Undated). *Urban Audit III - Final Report.*
4. Hungarian Central Statistical Office. (2004). Urban Audit II - Final Report.

IRELAND

Summary

◇ Irish cities have been included in Urban Audit since the Pilot Phase (Dublin,Cork**(1)**), then in UA II (Galway, Limerick **(2)**), and in UA III (Waterford **(3)**). In total, five cities are concerned.
□ LUZ delineations have changed between UA II and UA III. Ireland did not transmit National Report for UA III, but some information was provided by Urban Audit.
For UA II, "*the definition of the larger urban zones was based on local area development plans or in consultation with local authority planners. This applied to the LUZ zones used in the cities of Limerick and Cork*" **(4)**.
○ In UA III, LUZ are defined through 3 different ways:
Two LUZ are based on functional definitions: "*the LUZs for Limerick and Waterford cities are based on commuting data from the most recent census and a threshold of 20% commuters from the surrounding areas to the central city has been applied*" [32]. They correspond to an aggregation of LAU 2, probably built with the same commuters patterns than in UA II (threshold 20%).
LUZ of Cork and Galway correspond to planning regions (*Cork Area Strategic Plan (CASP)*[33] and *Galway Transport and Planning Study*[34]).
Dublin[35] corresponds to an aggregation of NUT 3.

Building blocks

◇ LAU 2, aggregation (i.e. *Electoral districts*)
□ Links: diversity of situations (see above). The methodology is not fully specified.
○ Dublin: Aggregation of 2 NUTS 3 (i.e. *Regional Authority Regions*).

Correspondence with GISCO

Same number of LUZ [36]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.
2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.

[32] Information transmitted in October 14, 2010
[33] http://www.corkcity.ie/casp/strategicplan/
[34] http://www.galwaycity.ie/AllServices/RoadsandTraffic/StudiesandSchemes/GalwayTransportandPlanningStudy
[35] Information transmitted in October 14, 2010
[36] Files provided by Urban Audit October 5, 2010.

3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
4. (2004). *Urban Audit - Irish National Report 2001*.

ITALY

Summary

◇ Italian cities have been included in Urban Audit since the Pilot Phase (Roma, Milan, Naples, Turin, Palermo, Genoa, Florence, Bari) **(1)** then in UA II (Bologna, Catania, Venice, Verona, Cremona, Trento, Trieste, Perugia, Ancona, L'Aquila, Pescara, Campobasso, Caserta, Taranto, Potenza, Catanzaro, Reggio di Calabria, Sassari, Cagliari) **(2)**, and in UA III (Padova, Brescia, Modena, Foggia, Salerno) **(3)**. In total, thirty-two cities are concerned.

□ LUZ definitions have changed between UA II and UA III.

During UA II, each LUZ corresponds to a NUTS 3: "*Provinces are administrative areas in some case without a credible geographical or statistical significance. There is no a credible alternative to consider Province as proxy of LUZ in Urban Audit II, at least at this stage*" **(4)**.

○ During UA III, LUZ definition is functional and corresponds to Local Labour System (LLS, i.e. *systemi locali del lavoro*). The methodology aggregates LAU 2 on the basis of employment (1000 jobs minimum in the LLS) and residence ("*occupied people working in A, occupied people resident in A, and occupied people resident and working in A*" **(5)**). Commuters are taken into account with a threshold of 10% (outflow) and 1% (inflow). The methodology uses a "*self-containment criteria*" with a threshold of 75% for building the LLS. This criteria is defined by *"occupied people resident and working in A/ occupied people resident in A"* **(5)**. Construction data come from *"the 1991 census intra-municipality daily commuting flows matrix"* **(4)**.

Building blocks

◇ LAU 2, aggregation (i.e. *Comuni*)

□ Links: based on commuting data (threshold (10%). The methodology is provided in the National Report **(5).**

Correspondence with GISCO

Same number of LUZ [37]

http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.

2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.

3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.

4. ISTAT. (Undated). *URBAN AUDIT II THE IMPLEMENTATION IN ITALY Final Report submitted to the European Commission*.

[37] Files downloaded and checked July 5, 2010

5. Istat. (Undated). *Urban Audit III - Final Country Report ITALY - Grant n. 72501-2006-001-2006-492*.
6. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.

LATVIA

Summary

◇ Latvian cities have been included in Urban Audit since UA II (Riga,Liepaja) **(1)**, and in UA III **(2)**. In total, two cities are concerned.
□ LUZ definitions have not changed between UA II and UA III: "*The delineation of spatial units of Latvia's territories participating in the project remained the same as in previous Urban Audit (UA) data collections*" **(3)**.
○ Liepaja LUZ corresponds to one LAU 1: "*Also* [LUZ] *level was defined easily because functional urban zones for both cities were known. As they could be approximated with NUTS level 3 or 4 units, it was decided to create LUZ using NUTS 4 units*" **(4)**. Riga LUZ is an aggregation of LAU 1

Building blocks

◇ Liepaja: LAU 1, elementary administrative unit (.i.e *Rajoni* and *republikas pilsētas*[38])
○ Riga: aggregation of LAU 1 (Riga and Ogre districts, and Jurmala city **(3)**).

Correspondence with GISCO

Same number of LUZ [39]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

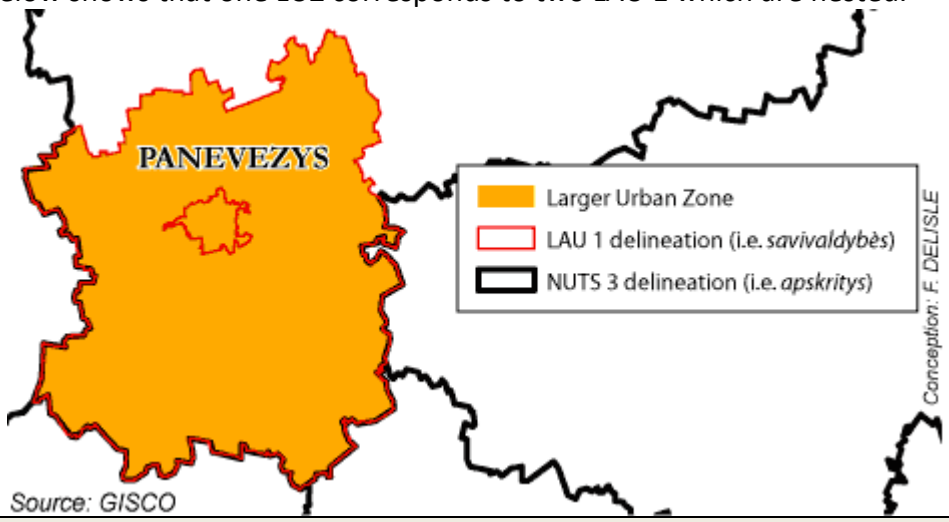
1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
3. Latvijas statistika. (2008). *Urban Audit 2006 - Grant agreement No. 72501.2006.001-2006.478 - FINAL TECHNICAL IMPLEMENTATION REPORT LATVIA*.
4. Supe, Jolanta. (2004). *Data Collection Project - Urban Audit II - Latvia*. CSB of Latvia.

[38] A territorial reform occurred in Latvia in 2009. Before that (when UA II and UA III National Reports were written), LAU 1 were made of "districts" (i.e. *rajoni*) and"cities" (i.e. *republikaspilsētas*).
[39] Files downloaded and checked July 21, 2010

LITHUANIA

Summary

◇ Lithuanian cities have been included in UA II (Kaunas, Panevezys, Vilnius) **(1)** then in UA III **(2)**. In total, three cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ Each LUZ corresponds to one LAU 1. The National Report mentions some analyses before defining LUZ, but without more details: "*In co-operation with the cities, possible ways to define LUZ and SCD were examined. On the basis of analyses, the following territorial units [i.e. LAU 1] in the Urban Audit II were agreed to be used*" **(3)**.
The map below shows that one LUZ corresponds to two LAU 1 which are nested.



Building blocks

◇ LAU 1, elementary administrative unit (i. e. *Savivaldybės*)
□ No aggregation
○ Vilnius: LAU 1 aggregation (City Core (Vilnius City) + Vilnius district + Elektrenai district + Trakai district)

Correspondence with GISCO

Same number of LUZ [40]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
3. Lisauskaite, Jolita. (Undated). *FINAL REPORT ON URBAN AUDIT II.* Statistics Lithuania.

[40] Files downloaded and checked July 21, 2010

LUXEMBOURG

Summary

◇ Luxembourg city has been included in Urban Audit since the Pilot Phase (Luxembourg) **(1)**, then in UA II **(2)**, and in UA III **(3)**.
□ LUZ definition has changed between UA II and UA III.
For UA II, the LUZ corresponds to an aggregation of LAU2 ("*we have defined the large urban zone (LUZ) regrouping 14 communes"***(2)**). However, according to Eurostat, the LUZ is underestimated **(4)**. For UA III, LUZ is enlarged to NUTS 0, as it can be deduced from GISCO.
○ In UA III, Luxembourg LUZ corresponds to Luxembourg country.

Building blocks

◇ NUTS 0, elementary administrative unit (i.e. whole country)
□ No aggregation

Correspondence with GISCO

Same number of LUZ [41]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.
2. CEPS/INSTEAD. (2003). *Urban Audit II Interim Report Luxembourg August 2003*.
3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
4. Carlquist, Torbiörn. (2006, August 30). The Larger Urban Zones in the Urban Audit data collection*. Globalisation Impact on Regional and Urban Statistics*. Wroclaw.

[41] Files downloaded and checked July 19, 2010

MALTA

Summary

◇ Maltese cities have been included in UA II (Valetta, Gozo) **(1)**, then UA III **(2)**. In total, two cities are concerned. The National Report for UA III was not available.
□ LUZ definitions have not changed between UA II and UA III (deduced from GISCO observations).
○ Each LUZ corresponds to one NUTS 3 **(3)**.

Building blocks

◇ NUTS 3, elementary administrative unit (i.e. *Gzejjer*)
□ No aggregation

Particular cases

Gozo: The City Core and the LUZ are both defined as the same NUTS 3

Correspondence with GISCO

Same number of LUZ [42]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
3. National Statistics Office - Malta. (Undated). *Urban Audit Final Report.*

[42] Files downloaded and checked July 19, 2010

NETHERLAND

Summary

◇ Dutch cities have been included in Urban Audit since the Pilot Phase (Amsterdam, Rotterdam) **(1)**, then in UA II (Arnhem, Eindhoven, Enschede, Groningen, Heerlen, s'Gravenhage, Tilburg, Utrecht) **(2)**, and in UA III (Breda, Nijmegen, Apeldoorn, Leeuwarden) **(3)**. In total, fourteen cities are concerned.

□ LUZ definitions have changed between UA II and UA III.

For UA II, each LUZ corresponds to one NUTS 3: "*Statistics Netherlands nevertheless proposed to use the Dutch NUTS 3 regions (called COROP Regions) as a proxy for the larger urban zones*" **(4)**.

○ For UA III, LUZ correspond to the functional region called "stadsgewest":

"*More data came available [since UA II] for another proxy being 'stadsgewest', which represents better the relation of the city with its surroundings*"**(5)**. The methodology used is described in a reference document **(6)** and on the Statistics Netherlands website **(7)**, but in Dutch langage. It seems that two steps are followed:

The agglomeration is first delineated, starting from built-up area criteria (land use map from 1996), and then using some thresholds (population > 100 000 inh., employment > 50 000 persons and market area > 150 000 persons).[43]

The functional area surrounding the city is then built using following criteria[44]: commuters (data from Labour Force Survey realized between 1995 and 1997); residential migrations (Survey realized between 1996 and 1997); urban infrastructure information.

Building blocks

◇ LAU 2, aggregation (i. e. *Gemeenten*)

□ Links: based on commuting data and residential migrations. The methodology is not specified in the National Report.

Particular cases

Almere: "*The exception is Almere, which territory is part of the stadsgewest Amsterdam, and by consequence has not a stadsgewest of its own*"**(5)**.

Correspondence with GISCO

Same number of LUZ[45]

http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés

[43] http://www.cbs.nl/nl-NL/menu/methoden/begrippen/default.htm?ConceptID=876

[44] http://www.cbs.nl/nl-NL/menu/methoden/begrippen/default.htm?ConceptID=877

[45] Files downloaded and checked July19, 2010

européennes.
2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
4. Statistics Netherlands. (2003). *Urban Audit II - The implementation in the Netherlands*.
5. Statistics Netherlands. (2008). Urban Audit 2006 - The implementation in the Netherlands. 2008.
6. Vliegen, Mathieu. (2005). *Grootstedelijke agglomeraties en stadsgewesten afgebakend.* Voorburg/Heerlen : Centraal Bureau voor de Statistiek.
7. Centraal Bureau voor de Statistiek. (2010). *CBS - Home*. Checked July 19, 2010, on CBS: http://www.cbs.nl/nl-NL/menu/home/default.htm.

NORWAY

Summary

◇ Norwegian cities have been included in UA III (Oslo, Bergen, Trondheim, Stavanger, Kristiansand and Tromsø). A total of six cities are concerned.
□ Norway did not participated to UA II
○ LUZ correspond to the functional region called "storbyregioner" **(1)**. The documentation relating to LUZ has been transmitted by Statistics Norway to Urban Audit in Norwegian language **(2)**. It seems that LUZ definitions are largely based on criteria related to travel times and commuting data. According to information transmitted by Urban Audit, "*Commuting was the leading criterion, travel times the adjustment instrument applied for the creation of the regions. A rule states that municipality's center with over 10 percent commuting levels to a bigger center is affiliated with this center if it does not represent a separate commuting area*"[46].

Building blocks

◇ LAU 2, aggregation (i.e. *Kommuner*)
□ Links: probably mainly based on commuting data (threshold 10%)

Correspondence with GISCO

Same number of LUZ [47]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ROG. (2005). *Data for storbyregioner/Urban Audit – avgrensing og inndeling av regionene*.
2. Det kongelige Kommunal og- regionaldepartement. (2003). *Storbymeldingen: Om utvikling av storbypolitikk.*

[46] Information provided by Urban Audit on December 3, 2010.
[47] Files downloaded and checked July 21, 2010.

POLAND

Summary

◇ Polish cities have been included in UA II (Warsaw, Lodz, Krakow, Wroclaw, Poznan, Gdansk, Szczecin, Bydgoszcz, Lublin, Katowice, Bialystok, Kielce, Torun, Olsztyn, Rzeszow, Opole, Gorzów Wielkopolski, Zielona Gora, Jelenia Gora, Nowy Sacz, Suwalki, Konin, Zory **(1)**), then in UA III (Czestochowa, Radom, Plock, Kalisz, Koszalin) **(2)**). In total, twenty-eight cities are concerned.

□ LUZ definitions have not changed between UA II and UA III.

○ LUZ definition is based on an aggregation of neighboring units (LAU 2 and LAU 1).

"*Due to the lack of flow statistics (i.e. journeys to work) (…)*" **(3)**, functional urban region could not be built.

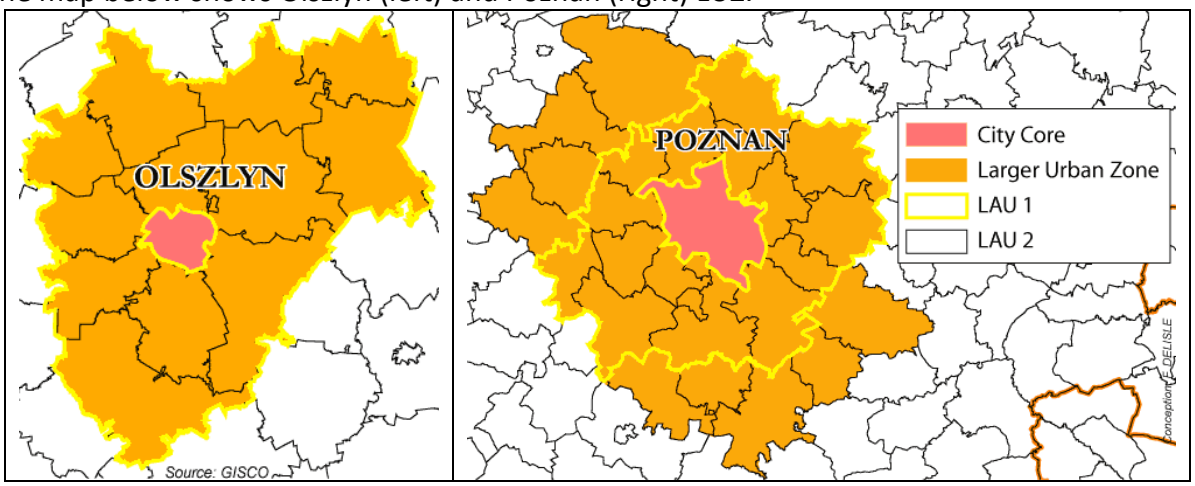
The methodology depends on City Core size **(3; 4; 5)**:

Population up to 100 000 inh.: LUZ corresponds to the surrounding LAU2 contiguous to the City Core (see for example, Suwalki at North-East);

Population between 100 and 250 000 inh.: LUZ corresponds to the surrounding LAU 1 that is contiguous to the City Core or to the LAU 1 containing the City Core (in most of the cases, the City Core is itself one LAU1, included inside another bigger LAU 1). (see for example Olszlyn, map below).

Population above 250 000 inh.: LUZ corresponds to the surrounding LAU 1 that is contiguous to the City Core, except if in some part of the ring, LAU 1 outer limit fits with local LAU 2 outer limit (see map below at the north-west and at the south-east of Poznan). Then, new LAU 2 are locally added (4 at the north-ouest and 5 at the south-east).

The map below shows Olszlyn (left) and Poznan (right) LUZ.



Probable changes in LUZ definitions will occur for next Urban Audit (UA IV).

Building blocks

◇ LAU 2, aggregation (i.e. *Gminy*) and/or aggregation of LAU 1 (i.e. *Powiat*)

□ Links: based on contiguity and city core size.

Correspondence with GISCO
Same number of LUZ [48] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1
References
1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities. 2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities. 3. CENTRAL STATISTICAL OFFICE. (2008). *Urban Audit 2006 - Final Country Report (Poland) - Grant agreement no. 72501.2006.001-2006.484*. 4. Felczak, Dominika. (2003). *Urban Audit II Poland - Intermediate report summarising tasks 1 and 2 of Urban Audit program*. Central Statistical Office of Poland. 5. Młodak, Andrzej. (Undated). *Polish experiences and possibilities in realisation of the URBAN AUDIT programme*. Central Statistical Office.

[48] Files downloaded and checked June 24, 2010

PORTUGAL

Summary

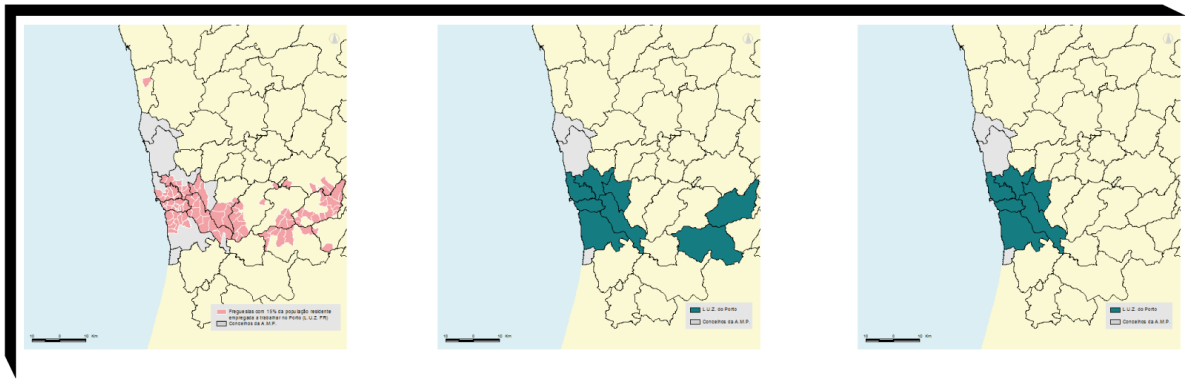
◇ Portuguese cities have been included in Urban Audit since the Pilot Phase (Lisbon, Oporto, Braga **(1)**), then in UA II (Aveiro, Coimbra, Setubal, Funchal, Ponta Delgada **(2)**), and in UA III (Faro **(3)**). In total, nine cities are concerned.

□ LUZ definitions have changed between UA II and UA III.

For UA II, each LUZ except Lisbon and Opporto, corresponds to one LAU 1: "*For all medium sized cities, it seems preferable to use NUTS 4 as a proxy for the functional urban region*" **(4)**. Lisbon and Opporto are defined as functional areas, with the methodology that has been used for all cities in UA III. For UA III, all LUZ are defined as functional areas.

○ LUZ are defined as functional areas, following three different steps, described on the map below **(5)**. The construction data come from Census 2001 **(5)**.

LARGER URBAN ZONE



Etape 1

Identification of the "freguesias" (LAU level 2) that have a working commuting intensity to the urban audit's cities (município) of more than 15% of the working resident.

Etape 2

Quantify the proportion of working people in all those "freguesias" belonging to a same "município" in the total of working people of the "município" they belong to. We have selected only the "municípios" where there was at least 50%

Etape 3

When needed, we have imposed the continuity spatial criterion to the core city

Source : Instituto Nacional de Estatśstica. Urban Audit III in Portugal, Final Report for the European Commission within the Framework of the Grant Agreement for an Action. 2008. Agreement Number 72501-2006-001-2006-485.

Building blocks

◇ LAU 1, aggregation (i.e. *Municipios*)

□ Link: based on commuting data (threshold 15% at LAU 2 level, see map above)

Particular cases

Setúbal: "*According to results Setúbal is the only city that still remains with "Luz identical to city*" " **(5)**.

Correspondence with GISCO
Same number of LUZ [49] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1
References
1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes. 2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities. 3. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities. 4. (Undated). *Urban Audit II in Portugal, Final Report for the European Commission within the Framework of the Grant Agreement for an Action*. 5. Instituto Nacional de Estatística. (2008). *Urban Audit III in Portugal, Final Report for the European Commission within the Framework of the Grant Agreement for an Action*.

[49] Files downloaded and checked June 24,2010

ROMANIA

Summary

◇ Romanian cities have been included in UA II (Alba Iulia, Arad, Bacau, Braila, Bucharest, Calarasi, Cluj-Napoca, Craiova, Giurgiu, Oradea, Piatra Neamt, Sibiu, TarguMures, Timisoara **(1)**) then in UA III **(2)**. In total, fourteen cities are concerned.
□ No information concerning evolution between UA II and UA III in National Reports
○ LUZ definition is based on an aggregation of neighboring units (LAU 2).
For UA II, LUZ are built with "*bordering communes situated around 15 km in the near vicinity of the cities*" **(3)**, except Bucharest (see below). These communes must be relevant with a law from 1968, "*abrogated in 1989, in which was considered suburban communes (term that is not used anymore) for each city of county*" **(3)**. "*In conclusion, LUZ were formed adding to the city the bordering suburban communes that are urbanized, industrialized and other well economic develop communes*"**(3)**.
For UA III, "*Romania proposed to be attached to the cities the bordering urbanized communes situated near vicinity of the municipalities selected as Urban Audit cities*"**(4)**.

Building blocks

◇ LAU 2, aggregation (i.e. *Comuni*/*Orase*/*Municipiu*)
□ Links: distance criteria (15 km) and juridic criteria (law from 1968, see above)
○ Bucharest: aggregation of LAU 2, using a distance criteria (20 km) and a selection of LAU 2: "*we selected the communes and cities around 20 km from Bucharest, that are nearly integrated into city and well economically developed*"**(3)**.

Correspondence with GISCO

Same number of LUZ [50]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
3. (Undated). *Urban Audit II for Candidat Countries Romania - Intermediate Report*.
4. Romania National Institute of Statistics. (2007). *URBAN AUDIT HISTORICAL DATA*.
5. Romania National Institute of Statistics. (2008). *Urban Audit - Final Operational Report Romania*.

[50] Files downloaded and checked July 21, 2010

SLOVAKIA

Summary

◇ Slovakian cities have been included in UA II (Bratislava, Košice, B. Bystrica, Nitra **(1)**), and in UA III (Prešov, Žilina, Trenčín, Trnava**(2)**). In total, eight cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ Each LUZ corresponds to one LAU 1 ("*District at the level of LAU1 was selected as LUZ in all cities*" **(3)**), except the LUZ of Bratislava (see below) **(4)**. According to the National Report, "*all proposals for LUZs are based on the employment zones, which were specified within the research work done for the SO SR* [Statistical Office of the Slovak Republic] *last year and we have agreed them also with the cities contact persons*"**(4)**.

Building blocks

◇ LAU 1, elementary administrative unit (i.e. *Okresy*)
□ No aggregation
○ Bratislava: NUTS 3 (i.e. *Kraje*) **(4)**.

Correspondence with GISCO

Same number of LUZ [51]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
3. The Statistical Office of the Slovak Republic. (2008). *Final Technical Report of the Grant 72501.2006.001-2006.487 Urban Audit data collection 2006/2007*. Bratislava.
4. Olexa, Michel. (2005). *Urban Audit II - PHARE 2001 Project - Intermediate Report for the Slovak Republic*. Bratislava.

[51] Files downloaded and checked July 19, 2010

SLOVENIA

Summary

◇ Slovenian cities have been included in UA II (Ljubljana, Maribor) **(1)**, and in UA III **(2)**. In total, two cities are concerned.
□ LUZ definitions have not changed between UA II and UA III.
○ Each LUZ corresponds to one NUTS 3 (*"NUTS3 level* [equals LUZ]*"* **(3)**). In Slovenia, NUTS 3 correspond to statistical regions, used as functional and planning areas[52] (*"Data were collected for city Ljubljana (large size city) and for city Maribor (medium size city) namely for the following spatial units: Podravska statistical region (Ljubljana LUZ) and Osrednjeslovenska statistical region (Maribor LUZ)"***(4)**).

Building blocks

◇ NUTS 3, elementary administrative unit (i.e. *Statistične regije*).
□ No aggregation

Correspondence with GISCO

Same number of LUZ [53]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
2. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities.
3. Statistical Office of the Republic of Slovenia. (Undated). *Urban Audit II - Quality Report on Data Collection - Country: Slovenia*.
4. Statistical Office of the Republic of Slovenia (SORS). (Undated). *Final Report - URBAN AUDIT 2006 - (Grant agreement NO 72501-2006-001-486)*.

[52] http://www.stat.si/vodic_oglej.asp?ID=360&PodrocjeID=2
[53] Files downloaded and checked July 19,2010

SPAIN

Summary

◇ Spanish cities have been included in Urban Audit since the Pilot Phase (Madrid, Barcelona, Valence, Seville, Saragossa, Malaga) **(1)** then in UA II (Badajoz, Las Palmas, Logrono, Murcia, Oviedo, Palma di Mallorca, Pamplona/Iruña, Santander, Santiago de Compostela, Toledo, Valladolid, Vitoria/Gasteiz) **(2)**, and in UA III (Bilbao, Cordoba, Alicante/Alacant, Vigo, Gijon, Santa Cruz de Tenerife) **(3)**. In total, twenty-four cities are concerned.
□ LUZ definitions have changed between UA II and UA III.
For UA II, each LUZ correspond to one NUTS 3 **(2)**:
"*The administrative Spanish units were adapted to the three spatial levels defined in the Project. The Spanish statistical information has been obtained for these different spatial units: national, region (autonomous communities), provinces, municipalities, districts and sections*"**(4)**. However, in the years after UA II, some analyses showed that Spanish LUZ were overbounded ("*Countries where the LUZ were over-bound include (…) Spain*" **(5)**).
○ In UA III, LUZ definition is functional and based on commuting data:
"*A new design for the LUZ has been done for the 25 core cities, according to the community census table program for 2000/2001 which contains data on commuting between municipios. In some cases we tried to approach the non official Spanish NUTS level 4 division*" **(3)**. This definition was adopted specifically for Urban Audit needs.
Construction data come from census 2001.

Building blocks

◇ LAU 2, aggregation (i.e. *municipios*)
□ Links: based on commuting data (threshold 15%)[54]
○ Madrid: NUTS 3, elementary administrative unit (i.e. *Provincias*)[55]

Particular cases

L'Hopitalet de Llobregat: this city does not correspond to a LUZ because it is part of Barcelona LUZ.

Correspondence with GISCO

Same number of LUZ [56]
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

[54]This threshold is deduced from our computations and observations, based on an Excel file sent by Urban Audit on December 7, 2010.
[55] According to information transmitted by Urban Audit January 2009. However, two LAU 2 (i.e. *Municipios*) are missing on GISCO.
[56] Files downloaded and checked July 5, 2010.

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.
2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities.
3. (2004). *Final Report, Grant Agreement of Urban Audit II in Spain Agreement number-2002 CE 16 0 AT 186*.
4. (Undated). *Final Country Report Spain*.
5. Carlquist, Torbiörn. (2006, August 30). The Larger Urban Zones in the Urban Audit data collection*. Globalisation Impact on Regional and Urban Statistics*. Wroclaw.

SWEDEN

Summary

◇ Swedish cities have been included in Urban Audit since the Pilot Phase (Stockholm, Göteborg) **(1)**, then in UA II (Malmö, Jönköping, Umeå) **(2)**, and in UA III (Uppsala, Linköping, Örebro) **(3)**. In total, eight cities are concerned.

□ LUZ definitions have changed between UA II and UA III.

For UA II, LUZ definition of Jönköping and Umea correspond to "Local Labour Market Areas" ("*for the last two cities (Jönköping and Umeå) 'Local Labour Market areas' were used to form the LUZ*" **(2)**).

○ In UA III, LUZ definition corresponds to the Local Labour Market Areas defined by Eurostat (*"For the other seven cities 'Local Labour Market areas' defined by Eurostat were used to form the LUZ"* **(3; 2)**). By checking GISCO data and maps, one can deduce that these LUZ delineations do not correspond to the two kinds of functional areas used by the Swedish statistics, the A-Region (*arbetsmarknadsregioner*) and FA-Region (*funktionella analysregioner*) **(4;5)**.

Building blocks

◇ LAU 2, aggregation (i.e. *Kommuner*)

□ Links: No information

○ Stockholm: NUTS 3, elementary administrative unit (i.e. *län*) **(3)**

Correspondence with GISCO

Same number of LUZ [57]

http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1

References

1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes.

2. Statistika centralbyràn. (Undated). *Urban Audit II - Final Report.* Stockholm : s.n., Undated.

3. Final Country Report [Sweden].

4. COMMIN. (2010). *COMMIN | The Baltic Sea Conceptshare*. Checked October 29, 2010, on COMMIN | The Baltic Sea Conceptshare*:* http://commin.org/en/bsr-glossaries/national-glossaries/sweden/arbetsmarknadsregion.html.

5. Statistics Sweden. (2010). *Lokala arbetsmarknader - – egenskaper, utveckling och funktion.* Orebro : Statistiska centralbyrån.

[57] Files downloaded and checked June 24, 2010.

SWITZERLAND

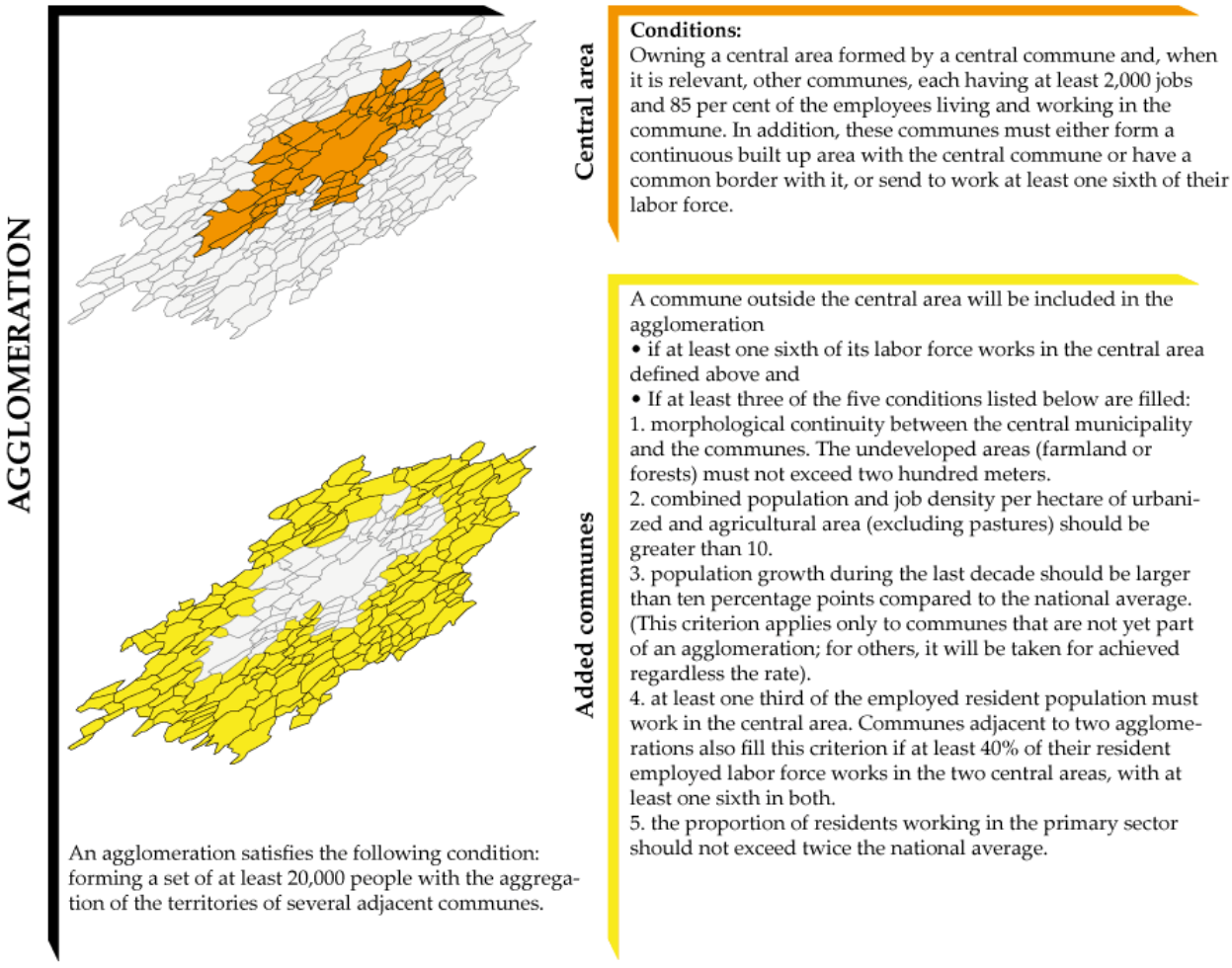
Summary

◇ Swiss cities have been included in UA III as a pilot phase (Zürich, Genève, Bern, Lausanne), and five more cities have been added (Basel, Winterthur, St Gallen, Luzern et Lugano) **(1)**. In total, nine cities are concerned.

□ Switzerland did not participate to UA II

○ LUZ definitions correspond to the Swiss functional region called « agglomération », which is described in 2000 by the Office Fédéral de la Statistique **(2)**. Criteria were formulated in 1990, actualized in 2000. Construction data come from Census 2000[58].

The figure below describes the different steps for building the "agglomérations", translated in English by us:



Source : Schuler, Martin, Dessemontet, Pierre et Joye, Dominique. **Recensement fédéral de la population 2000 - Les niveaux géographiques de la Suisse.** *Neufchâtel : Office fédéral de la statistique, 2005.*

[58] http://www.bfs.admin.ch/bfs/portal/fr/index/regionen/11/pro/01.html

Building blocks
◇ LAU 2, aggregation (i.e. *Gemeinden/Communes/Comuni*) □ Links: based on commuting data (threshold 16,66%) ○ Geneva: Aggregation of one NUTS 3 (the *Canton* of Genève) and one LAU 1 (the *District* of Nyon, in the canton of Vaud).
Correspondence with GISCO
Different number of LUZ: 9 LUZs but only 4 in Gisco[59]. The five cities included in the second phase are missing. http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1
References
1. Eurostat. (2008). *European Regional and Urban Statistics Reference Guide.* Luxembourg : Office for Official Publications of the European Communities. 2. Jeanneret, Barbara. (2009). *Audit urbain 2006/07. Rapport final pour l'Union européenne (Suisse).* Office fédéral de la statistique OFS. 3. Schuler, Martin, Dessemontet, Pierre et Joye, Dominique. (2005). *Recensement fédéral de la population 2000 - Les niveaux géographiques de la Suisse.* Neufchâtel : Office fédéral de la statistique.

[59] Files provided by Urban Audit September 5, 2010

UNITED KINGDOM

Summary

◇ English cities have been included in Urban Audit since the Pilot Phase (Birmingham, Leeds, Glasgow, Bradford, Liverpool, Edinburgh, Manchester, Cardiff) **(1)**, then in UA II (Aberdeen, Belfast, Bristol, Cambridge, Derry, Exeter, Leicester, Lincoln, London, Newcastle upon Tyne, Portsmouth, Sheffield, Worcester, Wrexham) **(2; 3)**, and in UA III (Coventry, Kingston-Upon-Hull, Stoke-on-Trent, Wolverhampton, Nottingham) **(4; 5)**. In total, twenty-seven cities are concerned.

□ LUZ definitions have not changed between UA II and UA III.

○ LUZ are defined according to a local consultation process:

"*The Larger Urban Zone or LUZ (…) are built up from LAU1 (Local Authority District) spatial units*" **(5)**. "*ONS* [Office for National Statistics] *sought the recommendation of relevant Local Authorities and Government Office Regions when constructing the LUZ area for each of the 24 cities under analysis. This produced a range of forms for the LUZs, that reflect the diversity of urban experience in the UK (…). The majority of cities assembled LUZs consisting of other Local Administrational Units. These were not necessarily arranged around the city in a 'doughnut' formation*" **(3)**.

Building blocks

◇ LAU 1, aggregation (i.e. Lower tier authorities (districts) or individual unitary authorities, Individual unitary authorities or LECs (or parts thereof), Districts).

□ Links: Local consultation

Particular cases

Derry: "*For Northern Ireland, Derry City was not given a LUZ. This was due to its rural surroundings as to have included neighbouring LAUs would not have been representative of the area*"**(3)** ;

Stevenage and Gravesham: "*Stevenage and Gravesham are included within the London LUZ as well as being Urban Audit cities on their own account*"**(5)**;

Lincoln: "*Lincoln City Council (…) opted not to have a LUZ comprised of neighbouring LAUs. Instead, a number of electoral wards were specified as representing the true reach of the city of Lincoln in its rural locale*"**(3)**;

Leeds and Bradford: "*Leeds and Bradford share a single LUZ*" **(5)**; "*It was virtually impossible to distinguish the urban reach of Leeds, discounting the effect of Bradford, and vice versa. There was also the pull which Bradford had on Leeds, etc.*"**(3)**.

Aberdeen: "*The LUZ for Aberdeen contained only one other LAU, Aberdeenshire. The total area of this LUZ, however, came to nearly 6500 km². This is second only to the LUZ for London. It was decided that the reach of Aberdeen's LUZ into its surrounding area was in part due to the relatively sparse population density in the vicinity*"**(3)**.

Wirral: This city does not correspond a LUZ[60].

[60]According to information transmitted by Urban Audit January 2009 and GISCO September 5, 2010.

Correspondence with GISCO
Same number of LUZ [61] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco/popups/references/administrative_units_statistical_units_1
References
1. ERECO. (2000). *L'Audit Urbain, Vers un référentiel pour mesurer la qualité de la vie dans 58 villes européennes.* Luxembourg: Office des publications officielles des Communautés européennes. 2. Eurostat. (2004). *Urban Audit, Methodological Handbook, 2004 edition.* Luxembourg: Office for Official Publications of the European Communities. 3. Office for National Statistics - Regional & Local Division. (Undated). *URBAN AUDIT II Implementation for the United Kingdom Final Report for the European Commission (Eurostat)*. 4. Eurostat. (2008). European Regional and Urban Statistics Reference Guide. 5. (2008). *Urban Audit 3 Final Country Report: United Kingdom*.

[61] Files downloaded and checked July 5, 2010

Figures list:

Tables:

Insert:

Annex:

TECHNICAL REPORT

ESPON

THE FUNCTIONAL URBAN AREAS DATABASE

CONTENT

- As a joint venture between 3 challenges of the Espon DB project (Urban data, Local data and Time series) and starting from the results of the previous Espon program we provide an update of the database of the Functional Urban Areas (FUAs) and Morphological Urban Areas (MUAs), as well as their inter-relations. Not only is it enhanced, it is also fundamentally enriched by the quality of the data provided, as the Functional Urban Areas (FUAs) are now delineated for most of the European countries of the Espon space at the LAU2 level.
- The FUAs are defined as the labor basins of the MUAs which are defined as densely populated areas.
- The main quality and advantage of the FUAs are their simple and universal definition throughout Europe, making them comparable in all the countries where they were delineated.
- Moreover we have also produced a list of indicators for these FUAs.

ESPON 2013 DATABASE

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

16 PAGES

LIST OF AUTHORS

Didier Peeters, Free University of Brussels, IGEAT

Contact

dpeeter1@ulb.ac.be
tel. +32 2 650 50 77

TABLE OF CONTENTS

1 General Methodology

Theoretically each MUA has a FUA, each FUA has a MUA near its center. But reality can be less straightforward and we find a good number of slightly more or much more complex cases.

First a MUA can be a secondary MUA in another's FUA. This happens when the extent of a FUA takes in a small MUA or when the population of the MUA has a commuting rate higher than the chosen threshold to another MUA. Most cases are obvious but a few are not, since the commuting flows are measured at the level of the LAU2 and a MUA can have several of these: one can therefore see that part of a MUA is commuting toward another one while the rest of it is not. Those cases were solved by first checking if the secondary MUA was landlocked into the rest of the «main» FUA which would lead to consider the secondary MUA as actually part of the main FUA, second considering the literature and the commonly accepted facts. We don't see any major rule that could lead us to elaborate a methodology to classify these rare cases.

Second a FUA can have two «twin» MUAs (or even more) when the commuting flows from one to another are crossing each other (they are usually touching each other) and the LAU2s around them send commuters on a relatively equivalent level to the both of them. An example of this is Douai and Lens, or Béthune and Bruay-la-Buissière near Lille in France, or Locarno-Bellinzona in Switzerland.

We must here mention an unfortunate limitation of the database: except for the French residents there are no transnational commuters data provided, which is a non-sense for a European perspective. We have trans-border FUAs (see project Espon Metroborder) around the French frontier but only for the French residents working in another country, not for the opposite although we know that there are commuters from Germany or Italy working in French FUAs (Strasbourg, Nice, ...). In the Espon Metroborder project this weakness was bypassed by assuming that the FUAs on both sides of the border are forming one entity, considering the results of Espon 1.4.3 based themselves on the literature.

Besides all the data limitations mentioned above we have encountered here and there difficulties in some areas like Scotland or former East-Germany, due to too big incoherencies between the different database, and this because of radical reshaping of the LAU2s.

1.1 Available databases

Different databases had to be joined together to achieve this work:

- Eurogeographics 2008 geometries of the LAU2s with identification code "SHN"
- MUAs' compositions in NUTS-5s with identification code from "CMRGCD97"
- SIRE database from Eurostat, tables emp_place_tot_mat_01 and c_emp_place_tot_mat_01 for the commuting numbers between LAU2s, emp_place_tot_01 and c_emp_place_tot_01 for the non-commuters numbers (working in area), LAU2s with identification code "CODCOM"

The database providing the commuting numbers between LAU2s is named SIRE and is provided by Eurostat. Most of its content comes initially from the 2001 censuses.

Data are provided only at the LAU1 level for Bulgaria and Slovenia, the data seem to be insufficient for Hungary. 2001 is quite old now but we know by experience that the FUAs are relatively stable, they are built on a quite low commuting rate and made simpler and coherent by filling the holes in them (LAU2s that don't reach the commuting threshold but landlocked in the FUA) and by eliminating the LAU2s remotely detached from the rest of the FUA. So chances are that any changes in the commuting rates would merely affect the general shape of the FUA, moreover if any they would probably affect small LAU2s at the fringes that don't have too much influence. The objective is not to compute the populations of the FUAs with accuracy close to one inhabitant but to have a good magnitude and to establish functional relations. But still changes can happen and it would be a good thing to have the opportunity to make this work on more recent data.

Entity	Year	Source
LAU2s spatial definition	2008	Eurogeographics
MUAs population	2001	Espon 1.4.3
FUAs population	2001 and 2006*	LAU2s population, Eurostat
Commuting numbers	2001*	SIRE database, Eurostat

Table 1 - Sources of the different data used in the elaboration of the FUAs and the MUAs

1.2 Linking the different databases

The first step of the work was to link all these tables with their different identification codes. We got the help from Eurostat which provided us with a table of correspondence between CODCOM and 3 other codes (COMM_ID, LAU2_code and nsi_code), as well as the NUTS-3 code for 2004 and 2006, which was easier to match with the other available codes although not straightforward. A resulting table was produced, certainly not perfect, especially because of the difficulty to deal with complex cases were the matching was a n-to-n type. This is of course a time series issue, since we are dealing with data from 1997, 1999 and 2006 and it is not always possible to determine what became what. At first we made a spatial join between the old CMRGCD97 geometries and the 2008 geometries but the matching was not very good, due to some distortion in the projection. By checking the names and the code similarities we kept what appeared to be good and worked the imperfectly matched codes differently. For this we carefully used the names of the entities in the different sources and the code similarities, starting with the steadier methods and using progressively less steady solutions. We always used the code similarities and checked with the names. Progressively we eliminated most of the difficulties.

The objective of this first step was to transpose the LAU2s components of the old MUAs from the 1997 system into the new 2008 one, and also to match the CODCOM from the SIRE database with the EUROGEOGRAPHICS 2008 boundaries. So we stopped when we got a result satisfying for this objective but it should be improved by splitting this single big table into 3 or 4 more simple tables, but we didn't have enough time to produce something more rigorous.

1.3 The transposition of the MUAs into the new system

Espon 1.4.3 had produced the delineation of about 2000 MUAs (the 300 smaller not published) with a total of about 11000 NUTS-5 in them. All these NUTS-5 were transposed into the EUROGEOGRAPHICS 2008 LAU2s delimitations as explained above. Only a fistful of new MUAs was added during the FUA production but no further research was made on the identification of the MUAs. This might be done by inserting the UMZ database (produced in this same Espon Database project) into this urban areas delimitation tool.

1.4 The correspondence between the SIRE codes and the eurogeographics codes

CODCOM (SIRE) and SHN (EUROGEOGRAPHICS) share no similarities but the correspondence file provided by EUROSTAT was used and we have completed some relations due to differences between 2006 (SHN in the EUROSTAT file) and 2008 (EUROGEOGRAPHICS), by the same method based on the code similarities and the names correspondences.

2 The data processing

All the following steps are realized in SQL (MySQL or PotgreSQL) with postgis 1.5 for the spatial steps.

2.1 The selection of the LAU2s in relation to the MUAs

The initial work was done with the SIRE matrix. The table emp_place_tot_mat_01 (and c_ emp_place_tot_mat_01) is a simple matrix of about 2 millions lines containing the residence LAU2 (CODCOM), the working place LAU2 (DESCODCOM) and the number of commuters. The table emp_place_tot_mat_01 (and c_ emp_place_tot_mat_01) was completed with the active population working in their residence LAU2, in order to take into account all the working places. The MUAs were added in another column (this is not an orthodox way to proceed but it made things easier) in relation to the work place LAU2.

The process then consisted in computing the total number of commuters going in a MUA, for each CODCOM and each MUA. In the same time a commuting rate is computed by comparing this number to the «economically active population» from the tables emp_place_tot_01 (EU15) and c_ emp_place_tot_01 (former candidate countries). Only the rates of 10 % or higher were kept.

Then specific cases are processed because of the incoherencies between the different databases that were not solved by the correspondence table mentioned above, especially the countries were the commuting data are provided at the LAU1 level or cases of big cities considered as a single entity in one database and a set of multiple LAU2s in the other (Budapest, Bratislava, Paris), or transborder data provided by other sources (Luxembourg, Belgian border), or Hungary were there is no active population data provided.

We then obtain a list of LAU2s with their respective MUA work places and their commuting rates, all this forming kind of «proto-FUAs» since each MUA is considered as the seed of a potential FUA and since the LAU2s can be part of several «proto-FUAs». This is a raw material to build the real FUAs, with their complex elements.

2.2 The spatial operations

The geographical objects for a map representation are produced.

The MUAs are simply aggregated from the LAU2s in the Eurogeographic 2008 layer. This step is easy because the MUAs were previously delineated as explained above.

The FUAs are much more complicated. The objective is to obtain coherent areas, without holes in them and no isolated parts. Here are the different steps.

- Selection of the main destination of the commuters for each LAU2, among all the MUAs toward which the commuting rate is higher than 10 % . Actually the different destinations are ranked and the first one is kept. In case of equality (same number of commuters toward 2 MUAs) the bigger MUA is selected as main destination.
- The MUAs are grouped into FUAs, according to the main destination of the commuters of each MUA and/or according to the literature as explained in the main report and above.
- The exterior rings of the FUAs are created and the bigger part is kept. Several parts can be kept where the separation is due to geographical particularities (islands, both sides of a fjord,...).
- The LAU2s spatially enclosed in the exterior rings are selected and attributed to their FUA, so that the holes are filled.

3 Known issues

Besides the unavoidable incoherencies between the different databases already mentioned above or in the main report there are specific problems we could not bypass :

- The original definition of the FUA is based on a proportion of the 'occupied' active population but all we have is the 'economically' active population, which includes the unemployed population. So this distorts the commuting rate by overestimating it by a maximum factor of 1.1, considering a 10 % unemployment.
- The active population in Germany is wrong in at least half of the municipalities! This problem is somehow minimized because, first, when a LAU2 has a commuting flow toward more than one MUA (for example in Rhein-Ruhr region) what matters is the highest number of commuters whatever the active population is, considering that this happens in areas with a strong peri-urbanization and without free space between the FUAs. Second, the process of building the FUAs eliminates the LAU2s that are not among the others for a same FUA and homogenizes the area by making a ring around the LAU2s of the FUA. Third, the LAU2s where the commuters' number is higher than the active population are ignored. Fourth, the overestimation of the FUAs (there can't be underestimation) happens by including probably small municipalities at the fringes of the areas. But still, errors remain.
- Same remark for Slovakia, but apparently with a smaller proportion of erroneous data.
- In former East-Germany, especially in the Saxe-Anhalt, there were many municipalities merging after the reunification. This leads to many incoherencies between SIRE and EUROGEOGRAPHICS, i.e. we have commuting numbers for LAU2s that don't exist anymore and we didn't spend time on trying to redistribute the commuters in the new LAU2s. This might perhaps be done but not in this project.
- In the area between Glasgow and Edinburg we have the same problem than in Saxe-Anhalt (see above).
- In Portugal the commuting numbers in the core city areas are provided at the LAU1 level but this is not a problem because in every case we have a ring of municipalities (LAU2s) around this LAU1 that are included in the FUA. So any central municipality that would be out of the MUA but into that LAU1 area would be included in the FUA anyway.
- Technically speaking, instead of making a ring surrounding the FUAs with the Postgis ST_ExteriorRing function we could explore the possibility to use the ST_convexHull function that minimizes the quirks in the shape of some FUAs, but it could lead also to exaggerate their size.

Whenever new data would become available, like especially the active population in Germany, we would rebuild the FUAs immediately.

4 The indicators

In collaboration with the project Espon METROBORDER we have produced a list of indicators for all the FUAs.

Indicator	Year *	Source	Unit
Population	2001	LAU2, EUROSTAT	inhab.
	2006	LAU2, EUROSTAT	inhab.
Population variation	2001 - 2006	LAU2, EUROSTAT	%
FUAs Areas	2008	EUROGEOGRAPHICS	km^2
MUAs Areas	2008	EUROGEOGRAPHICS	km^2
Compactness	synthetic indicator		%
GDP	2006	NUTS-3, EUROSTAT	Euro
GDP/inhab.***	synthetic indicator		Euro/inhab.
Economical structure	2006	NUTS-3, EUROSTAT (6 big NACE sectors)	%
Unemployment**	2006	NUTS-3, EUROSTAT (6 big NACE sectors)	%

Table 2 - list of the indicators available in the Espon database

* : For some indicators some data are coming from different years. See the missing data issues below.
** : Unemployment values should be used «with caution» !
*** : see the map below

4.1 Methodology

For the Population and Area indicators we have simply computed the FUA values from the LAU2 values. The Compactness is the % of the population of the FUA actually living in the main MUA.

The economical indicators are computed by using the NUTS-3 values on which we apply a population ratio between the NUTS-3 and the intersection of the FUA and the NUTS-3. This is possible because we now have the LAU2 composition of the FUAs.

4.2 Examples

4.2.1 Basel

By way of illustration let's take an example: here, the Basel case.

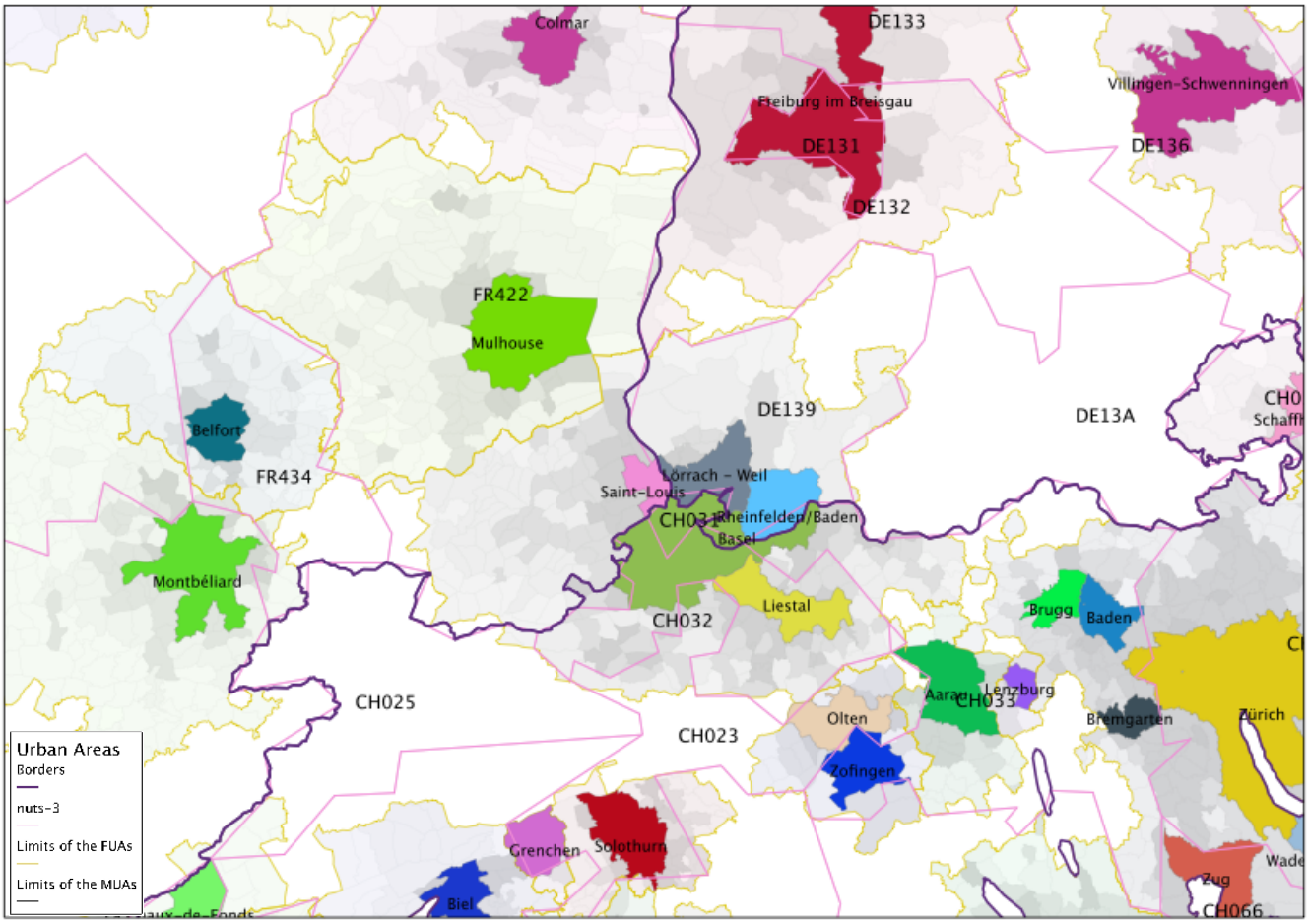


Figure 1 - The FUA of Basel in its region

The MUAs are shown in plain colors, the FUAs are in light colors, the more grayish the higher the number of commuters is.

nuts3_06	pop_nuts3	pop_fua	pop_fua in nuts3	coef
CH023	248	827,3	31,7	13
CH025	69	827,3	0,1	0
CH031	185	827,3	188,1	100
CH032	266	827,3	258,7	100
CH033	572	827,3	52,6	10
DE139	221	827,3	215,6	100
FR422	738	827,3	91,4	13

Table 3 - correspondence between the nuts-3 and the FUA of Basel.

We see here that the FUA of Basel extends over 7 nuts-3 in 3 countries, and we see in the table that this corresponds to different population values according to the nuts-3. We have computed for each of them a coefficient ('coef') in % giving what part of the nuts-3 indicator (for instance the GDP) we take from every nuts-3, the total giving the GDP of the FUA.

4.2.2 *The 4 different FUA types schematized in the main report*

In the main report we have presented 4 different FUA types (taken from Espon 1.4.3), here are 4 regions to illustrate them.

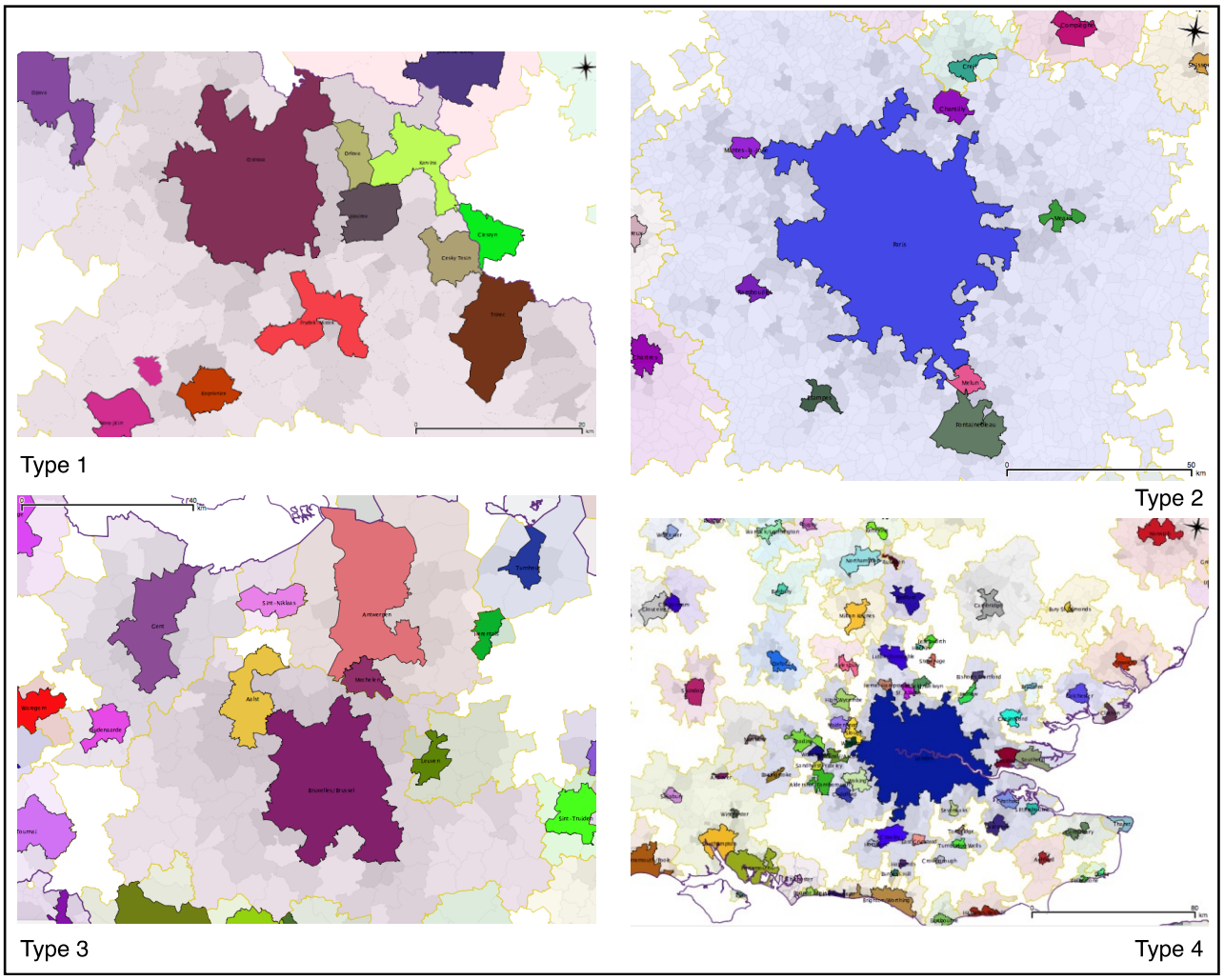


Figure 2 - Four different FUA types illustrated by the Ostrava region (1), the Ile-de-France Region (2), the Belgian central metropolitan region (3) and the London region (4).

4.2.3 Displaying FUAs database on maps, two examples

The FUAs database allows displaying innovative results for the all ESPON Area. For instance, the maps of GDP per capita 2006 (figure 3) and evolution of population 2001-2006 (figure 4) show strong contrasts between FUAs, and give an additional picture of the situation as compared to classical maps produced at NUTS3 level. However, it is important to have a look to keep in mind that some

of the values are estimated (cf section 4.3 below) and such results must be interpreted carefully.

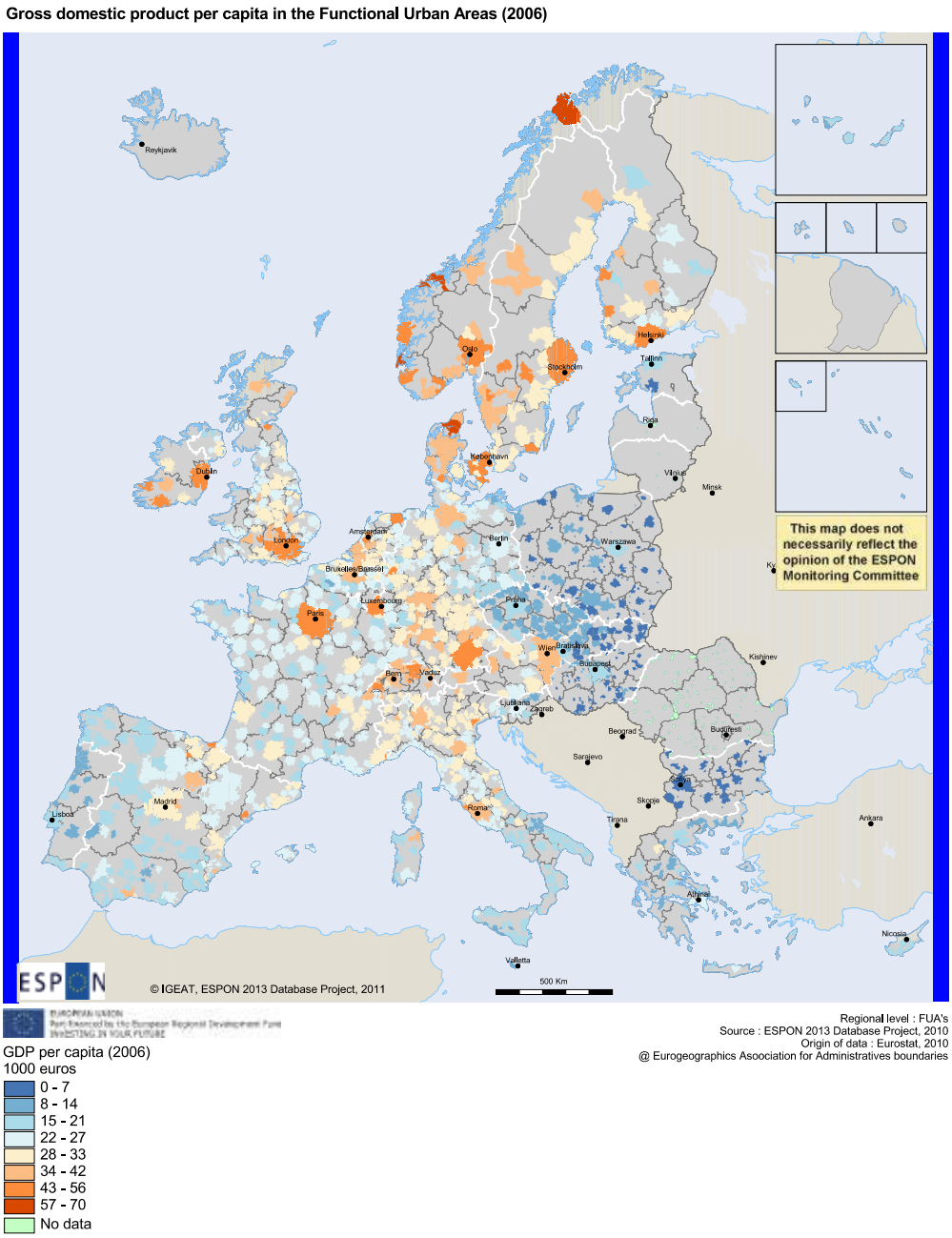


Figure 3 – Gross Domestic Product per capita 2006 in the FUA delineation

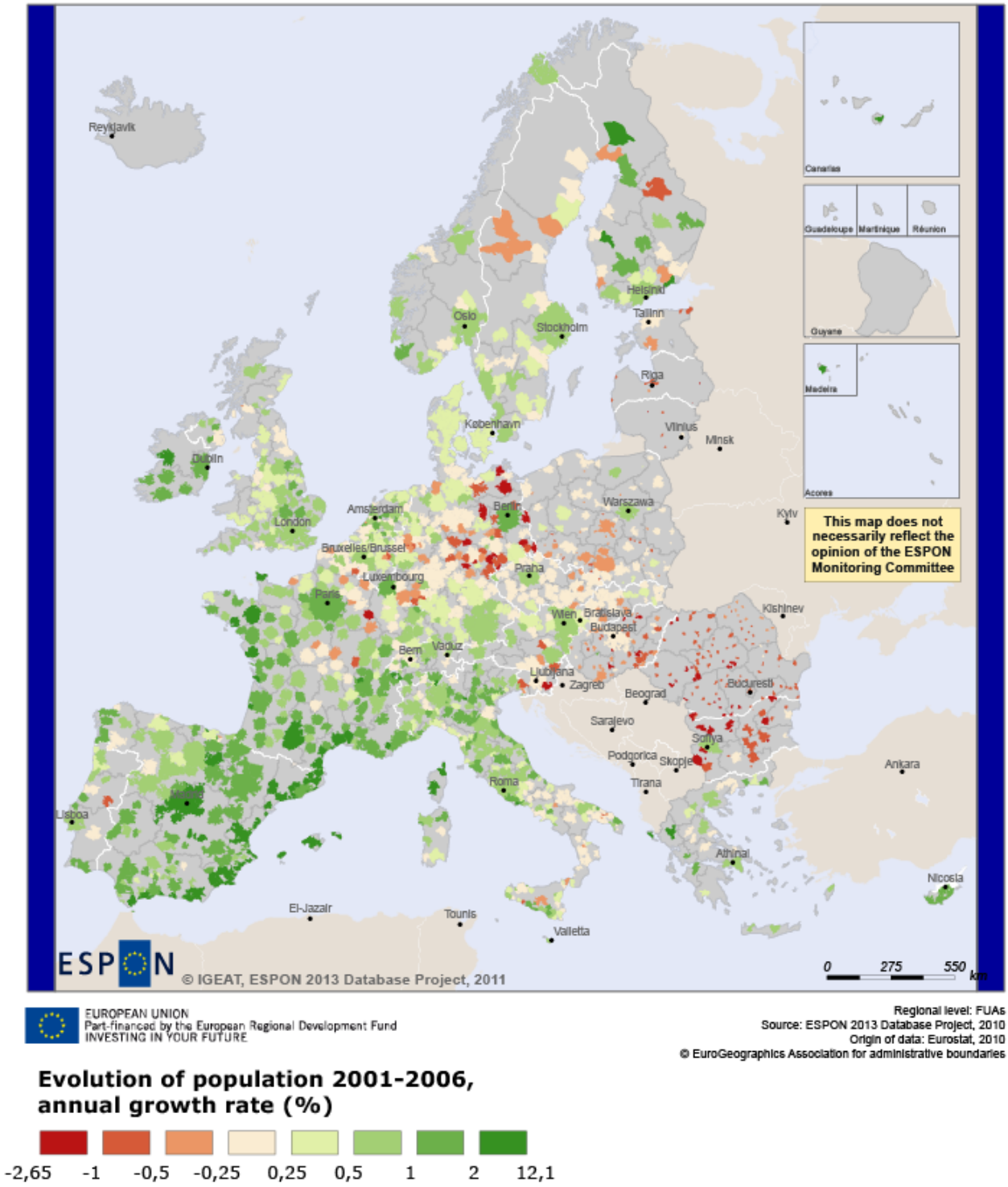


Figure 4 – Evolution of population 2001-2006 in the FUA delineation

4.3 Missing data issues

Poland

So far there are no commuting data in Poland. We received from Przemyslaw Sleszynski (Polish Academy of Sciences) a set of functional areas based on the last census data (2002) and other data from 2004 like the socio-cultural profiles of the population. We have included these data 'as is' in our database.

Romania, Latvia and Lithuania

No commuting data are provided for Romania, Latvia and Lithuania and we couldn't find any substitute, so we used the population data from Espon 1.4.3..

LAU2 Population in 2006

We don't have population numbers at the LAU2 level for Bulgaria, Cyprus, Denmark, United Kingdom, Lithuania, Latvia, Portugal and Romania. So we used the NUTS-3 replacement methodology described above to obtain the population of the FUA in 2006.

GDP in Switzerland and Norway

The GDP for Swiss and Norwegian FUAs are from 2005, due to missing data in Eurostat for the NUTS-3.

Unemployment

At the NUTS-3 level the unemployment values are not available in 2006 for Denmark, Sweden, Ireland, Switzerland and two Italian provinces (Sassari and Cagliari). See the synthesis table below for the replacement years.

Indicators	Countries	Substituting Years
Commuting data	Poland Romania Latvia Lithuania	-
2006 LAU2 population	Bulgaria Cyprus Denmark Lithuania Latvia Portugal Romania United Kingdom	NUTS-3 replacement
GDP	Switzerland Norway	2005
Unemployment	Denmark Sweden	2007
	Ireland	2004
	Switzerland	2005
	ITG25 (Sassari province) ITG27 (Cagliari province)	2008

Table 4 - synthesis of the missing data issues