DRAFT TECHNICAL REPORTS
Annexed to the Second Interim Report

❖ Application

- Thematic structuring and variables labeling within the ESPON 2013 Database: An empirical method *(27 p)*
- Metadata guide - Acquisition and storage of Data and Metadata in ESPON 2013 Database *(30 p)*
- ESPON Database application - Towards a web interface for the ESPON 2013 Database *(22 p)*

❖ Methodological issues

- Towards an approach of time series data issues: from empirical methods to application (23 p)
- Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase *(32 p)*
- Naming UMZ: Methods and results *(12 p)*
- Spatial analysis for quality control *(81 p)*
- Using downscaled population in local data generation – A country level examination *(16 p)*
- Mapping guide for ESPON Projects and the external community *(32 & 24 p)*

❖ Scales & data availability

- World database – Towards a World Dictionary of units (46 p)
- Analysis of the availability and the quality of data on Western Balkans and Turkey (45 p)
- Local data – First investigations in Romania, Bulgaria, Czech Republic and Slovakia (26 p)

Note: **The Draft Technical Reports (DTR)** are full part of the present SIR but are also considered as non final versions or "work in progress". Each challenge is improving this document and it is only with the Final Report that they will be considered as definitively achieved. Only the mapping guide has the status of Final Technical Report

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 *Agriculture*). 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	I	_	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.

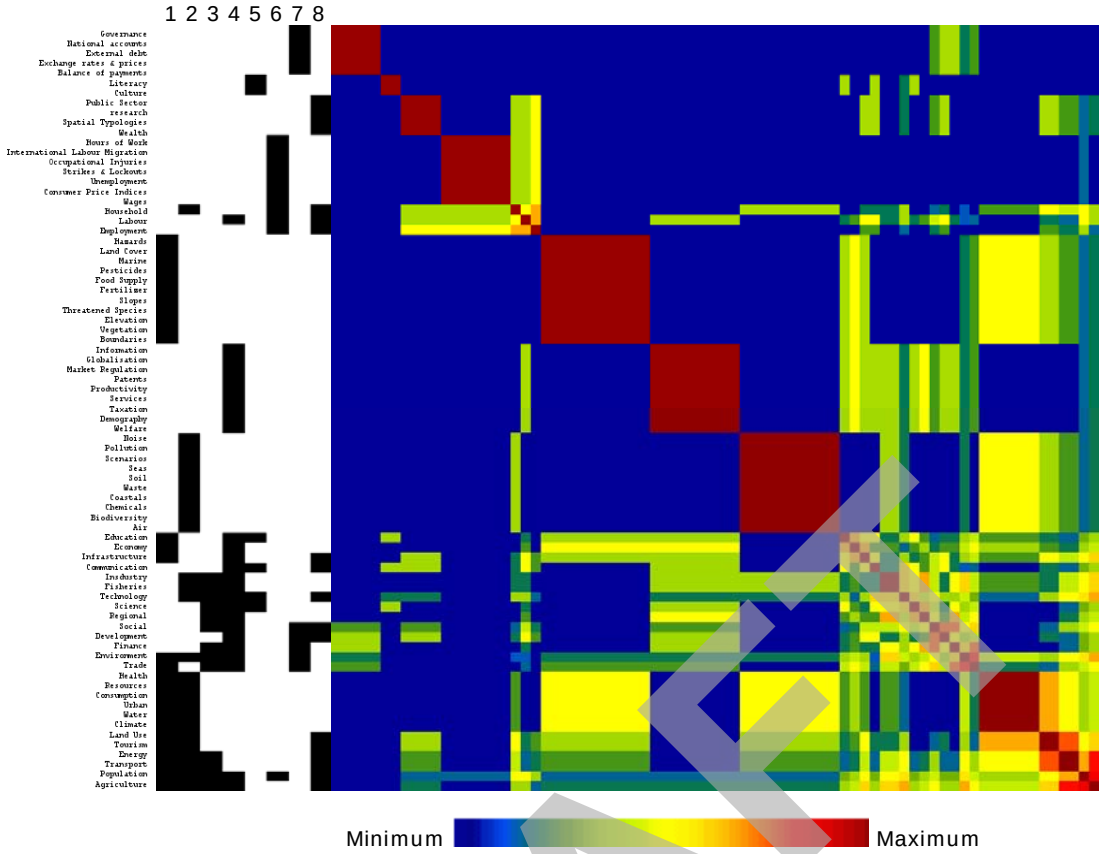
Forrest, D. (1999). Geographic Information: Its Nature, Classification, and Cartographic Representation. *Cartographica*, 36(2), 31-53.

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

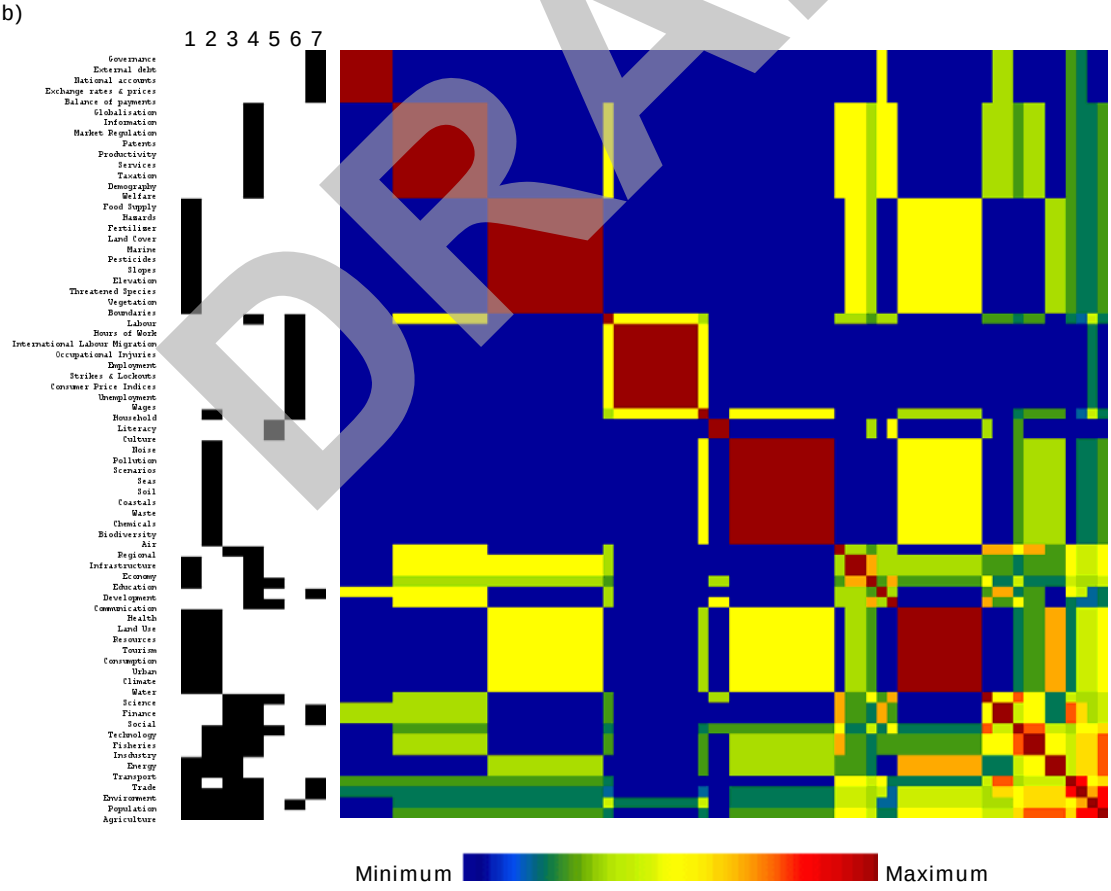
Slocum, T., et al. (2005). *Thematic Cartography and Geographic Visualization*. New Jersey: Pearson Prentice Hall.

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

(a)



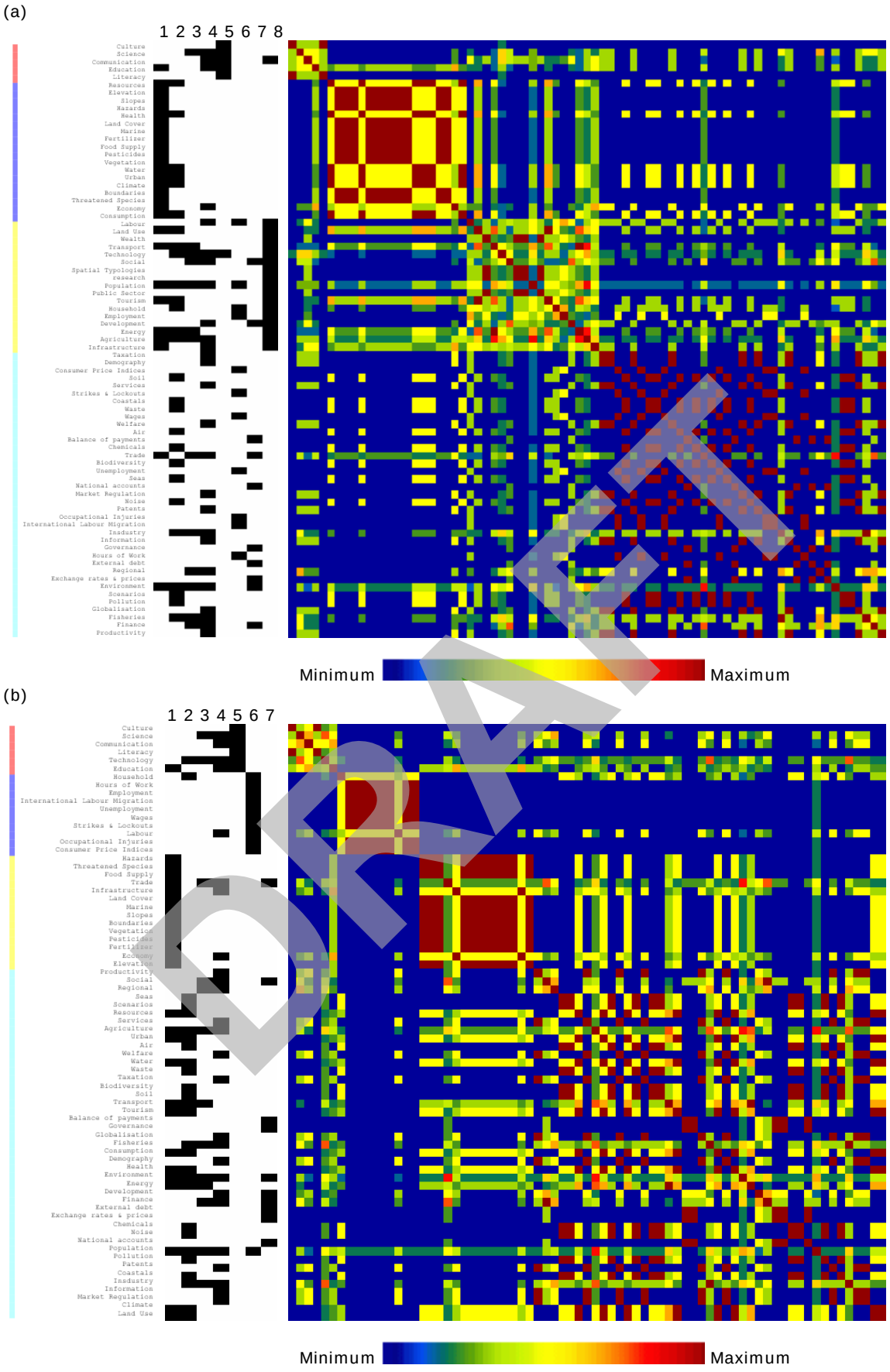
(b)



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

Figure 1

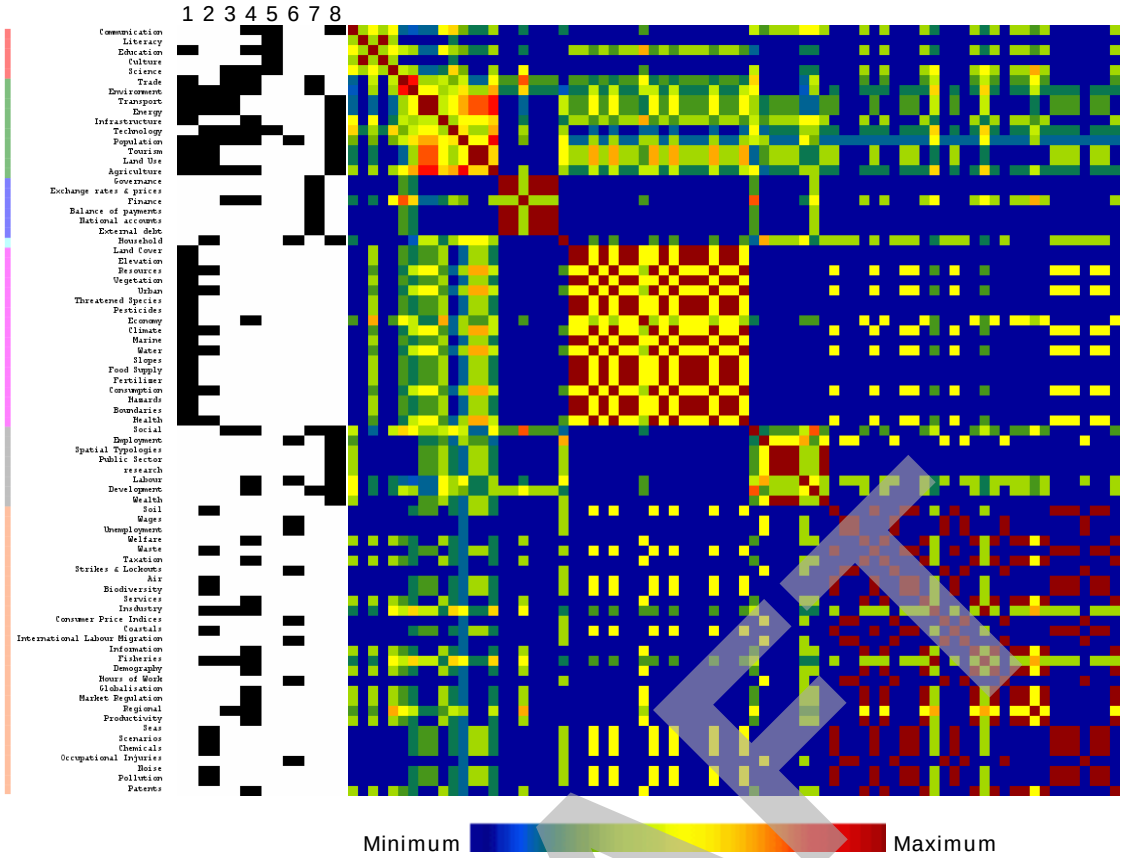
Matrix visualisation in GAP environment of nomenclatures (subjects) used by statistical databases (variables). The first sorted matrix (a) includes ESPON 2006 structure whereas the second one (b) ignores it.



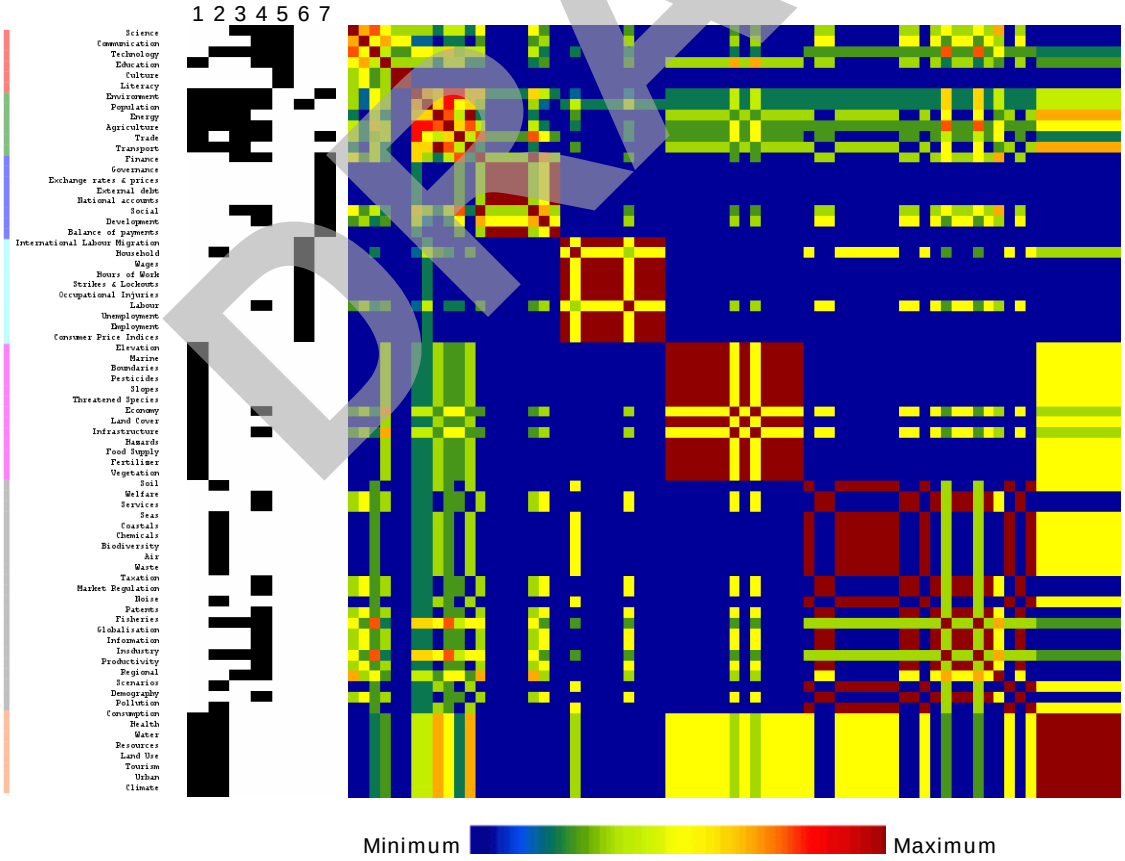
Note: (1) UNEP, (2) EEA, (3) EUROSTAT, (4) OECD, (5), UNESCO, (6) ILO, (7) WPI, (8) ESPON 2006.

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

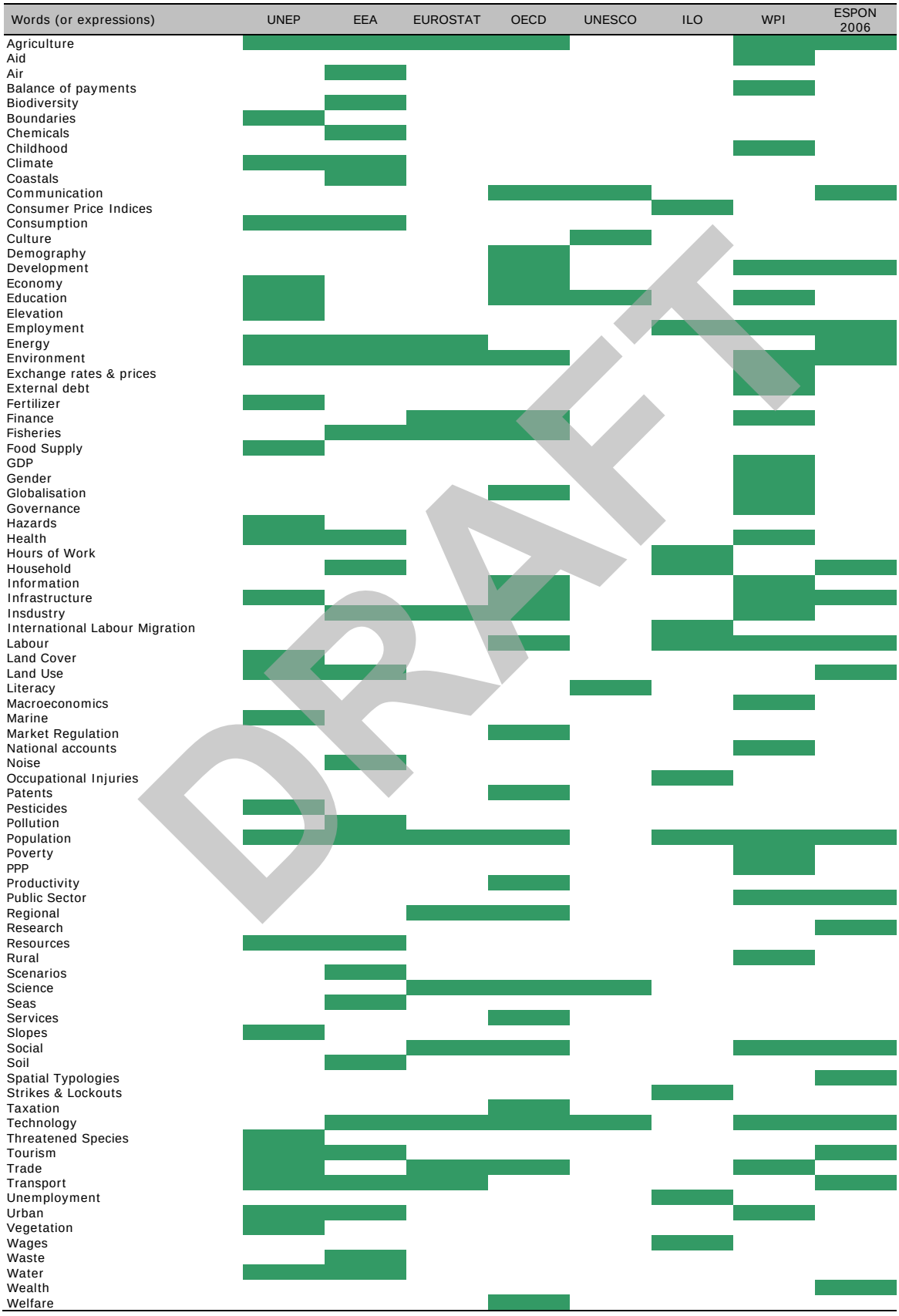
OECD
1 General Statistics
2 Agriculture and Fisheries
3 Demography and Population
4 Development
5 Economic Projections
6 Education and Training
7 Environment
8 Finance
9 Globalisation
10 Health
11 Industry and Services
12 Information and Communication Technology
13 International Trade and Balance of 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 _	_ 5	9 2	N 0201POPtrtc_ __95N2
Total population	02.Demography	01.Population Structure	1999	NUTS2/3**	pop_t_1999	0 2	0 1	POPtrtc _	_ 9	9 2	N 0201POPtrtc_ __99N2
Total population change	02.Demography	02.Population Movement	1995-1999	NUTS2/3**	Pop_ch_95_99	0 2	0 2	POPtrtc _	5 9	9 2	N 0202POPtrtc_ 9599N2
Natural population change	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	Nat_ch_96_99	0 2	0 2	NATtrtc _	6 9	9 2	N 0202NATtrtc_ 9699N2
Migratory population change	02.Demography	02.Population Movement	1996-1999	NUTS2/3**	Mig_ch_96_99	0 2	0 2	MIGtrtc _	6 9	9 2	N 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_ _	5 9	9 2	N 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_ _	6 9	9 2	N 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_ _	6 9	9 2	N 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_ _	6 9	9 2	N 0202POPdev__ 9600N2
Total population	02.Demography	01.Population Structure	2000	NUTS2/3**	Pop_t_2000	0 2	0 1	POPtrtc _ _	0 0	0 2	N 0201POPtrtc_ __00N2
Total population	02.Demography	01.Population Structure	2000	NUTS2/3**	pop_t_2000	0 2	0 1	POPtrtc _ _	0 0	0 2	N 0201POPtrtc_ __00N2
Sum of births	02.Demography	01.Population Structure	2001-2005	NUTS2/3**	Birth_01_05	0 2	0 1	BTHtrtc _	0 1	0 2	N 0201BTHtrtc_ 0105N2
Sum if death	02.Demography	01.Population Structure	2001-2005	NUTS2/3**	Death_01_05	0 2	0 1	DTHtrtc _	0 1	0 2	N 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	0 2	N 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	0 2	N 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	0 2	N 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	5 2	N 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	0 5	N 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	0 5	N 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	0 5	N 0202POPdev__ 0105N3
Multimodal potential accessibility, absolute level	03.Transport	03.Accessibility	2001,2006	NUTS3**	MM	0 3	0 3	ACCmmab s	_ 6	0 3	N 0303ACCmmabs __06N3
Multimodal potential accessibility, standardised	03.Transport	03.Accessibility	2001,2006	NUTS3**	MM_i	0 3	0 3	ACCmmst d	_ 6	0 3	N 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	0 3	N 0303ACCmmstd 0106N3
Multimodal potential accessibility, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	MM_r	0 3	0 3	ACCmmre l	0 1	0 3	N 0303ACCmmrel 0106N3
Multimodal potential accessibility, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	MM_a	0 3	0 3	ACCmmab s	0 1	0 3	N 0303ACCmmabs 0106N3
Potential accessibility by air, absolute level	03.Transport	03.Accessibility	2001,2006	NUTS3**	Air	0 3	0 3	ACCaiab s	_ 6	0 3	N 0303ACCaiabs __06N3
Potential accessibility by air, standardised	03.Transport	03.Accessibility	2001,2006	NUTS3**	Air_i	0 3	0 3	ACCaist d	_ 6	0 3	N 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	0 3	N 0303ACCaistd 0106N3
Potential accessibility by air, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	Air_r	0 3	0 3	ACCaire l	0 1	0 3	N 0303ACCairel 0106N3
Potential accessibility by air, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	Air_a	0 3	0 3	ACCaiab s	0 1	0 3	N 0303ACCaiabs 0106N3
Total Population	02.Demography	01.Population Structure	2001,2006	NUTS3**	Pop_t	0 2	0 1	POPtrtc _ _	0 6	0 3	N 0201POPtrtc_ __06N3
Potential accessibility road, standardised	03.Transport	03.Accessibility	2006	NUTS3**	ROAD_Index	0 3	0 3	ACCrdst d	_ 6	0 3	N 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	0 3	N 0303ACCrdstd 0106N3
Potential accessibility road, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	ROAD_Absolutechange	0 3	0 3	ACCrdab s	0 1	0 3	N 0303ACCrdabs 0106N3
Potential accessibility road, index change	03.Transport	03.Accessibility	2001-2006	NUTS3**	ROAD_Indexchange	0 3	0 3	ACCrdix _	0 1	0 3	N 0303ACCrdix_ 0106N3
Potential accessibility rail, standardised	03.Transport	03.Accessibility	2006	NUTS3**	RAIL_Index	0 3	0 3	ACCrlst d	_ 6	0 3	N 0303ACCrlstd __06N3
Potential accessibility rail, relative change	03.Transport	03.Accessibility	2001-2006	NUTS3**	RAIL_Relativechange	0 3	0 3	ACCrlre l	0 1	0 3	N 0303ACCrlrel 0106N3
Potential accessibility rail, absolute change	03.Transport	03.Accessibility	2001-2006	NUTS3**	RAIL_Absolutechange	0 3	0 3	ACCrlab s	0 1	0 3	N 0303ACCrlabs 0106N3
Potential accessibility rail, index change	03.Transport	03.Accessibility	2001-2006	NUTS3**	RAIL_Indexchange	0 3	0 3	ACCrlix _	0 1	0 2	N 0303ACCrlix_ 0106N2
Total population	02.Demography	01.Population Structure	2000-2007	NUTS0/1/2**	POP	0 2	0 1	POPtrtc _ _	0 0	0 0	N 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	0 0	N 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	0 0	N 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	0 0	N 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	0 0	N 0201POP75t__ __00N0
Annual average population change	02.Demography	02.Population Movement	2000-2007	NUTS0/1/2**	POP_ch	0 2	0 2	POPartc _	_ 0	0 0	N 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	0 0	N 0202POP2064a ch00N0
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	0 0	N 0202POP2064a ch00N0
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	0 0	N 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	0 0	N 0202POP75ach __00N0
Share of 20-39 years	02.Demography	01.Population	2005	NUTS0/1/	POP2039_sh	0 0	0 0	POP2039	_ 0	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	N	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	3 N	0301PDTinrtd __30N3
Productivity of airports	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_E2_Prod_air	0 3 1	0	PDTairt d	_ 0	3 N	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	3 N	0703ECOgwhix __30N3
Congestion costs	03.Transport	03.Accessibility	2005,2030	NUTS3*	PIM_E4_Cong_level	0 3 3	0	COScgti x	_ 0	3 N	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	3 N	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	3 N	0402CO2rdix_ __30N3
Safety of roads	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_Q3_Traffic_seg	0 3 1	0	SFTrdix _	_ 0	3 N	0301SFTrdix_ __30N3
Market opportunities	07.Economy	03.Income and Consumption	2005,2030	NUTS3*	PIM_Q4_GDP_3h	0 7 3	0	MKToppi x	_ 0	3 N	0703MKToppix __30N3
Landscape fragmentation	03.Transport	01.Transport Infrastructure	2005,2030	NUTS3*	PIM_I1_HCI_dens	0 3 1	0	LANfrgi x	_ 0	3 N	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	3 N	0302EXTvstix __30N3
Regional integration	03.Transport	03.Accessibility	2005,2030	NUTS3*	PIM_I3_Road_N2_n	0 3 3	0	REGinti x	_ 0	3 N	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	_ 0	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	EROrski x	_ 4	0 2	0401EROrskix __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	0 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 0 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 0	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 0	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	MIGntrtc	_ 7	0 2	0202MIGntrtc__07L2
Number of persons born abroad	02.Demography	01.Population Structure	2000-2007	LAU2**	D-NS_5	0 2	0 1	POPbrnrtc	_ 7	0 2	0201POPbrnrtc_07L2
Total number of unemployed persons	07.Economy	02.Unemployment	2007	LAU2	E-NS_4a	0 7	0 2	UMPtrtc	_ 7	0 2	0702UMPtrtc___07L2
Number of unemployed persons, female	07.Economy	02.Unemployment	2007	LAU2	E-NS_4b	0 7	0 2	UMPfrtc	_ 7	0 2	0702UMPfrtc___07L2
Number of unemployed persons, male	07.Economy	02.Unemployment	2007	LAU2	E-NS_4c	0 7	0 2	UMPmrtc	_ 7	0 2	0702UMPmrtc___07L2
Active population, total	07.Economy	01.Employment	2007	LAU2	E-NS_1a	0 7	0 1	ACTtrtc	_ 7	0 2	0701ACTtrtc___07L2
Active population, males	07.Economy	01.Employment	2007	LAU2	E-NS_1c	0 7	0 1	ACTmrtc	_ 7	0 2	0701ACTmrtc___07L2
Active population, females	07.Economy	01.Employment	2007	LAU2	E-NS_1b	0 7	0 1	ACTfrtc	_ 7	0 2	0701ACTfrtc___07L2
Total persons working in agriculture, hunting	07.Economy	01.Employment	2005	LAU2	E-NS_2a	0 7	0 1	AGRtrtc	_ 5	0 2	0701AGRtrtc___05L2
Total persons working in fishing	07.Economy	01.Employment	2005	LAU2	E-NS_2b	0 7	0 1	FIStrtc	_ 5	0 2	0701FIStrtc___05L2
Total persons working in mining and quarrying	07.Economy	01.Employment	2005	LAU2	E-NS_2c	0 7	0 1	MINtrtc	_ 5	0 2	0701MINtrtc___05L2
Total persons working in manufacturing	07.Economy	01.Employment	2005	LAU2	E-NS_2d	0 7	0 1	MANtrtc	_ 5	0 2	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	_ 5	0 2	0701GAStrtc___05L2
Total persons working in construction	07.Economy	01.Employment	2005	LAU2	E-NS_2f	0 7	0 1	CONtrtc	_ 5	0 2	0701CONtrtc___05L2
Total persons working in wholesale and retail	07.Economy	01.Employment	2005	LAU2	E-NS_2g	0 7	0 1	RETtrtc	_ 5	0 2	0701RETtrtc___05L2
Total persons working in hotels and restaurants	07.Economy	01.Employment	2005	LAU2	E-NS_2h	0 7	0 1	HOTtrtc	_ 5	0 2	0701HOTtrtc___05L2
Total persons working in transport, storage	07.Economy	01.Employment	2005	LAU2	E-NS_2i	0 7	0 1	TRAtrtc	_ 5	0 2	0701TRAtrtc___05L2
Total persons working in financial intermediation	07.Economy	01.Employment	2005	LAU2	E-NS_2j	0 7	0 1	FINtrtc	_ 5	0 2	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	_ 5	0 2	0701REStrtc___05L2
Total persons working in public administration and defence	07.Economy	01.Employment	2005	LAU2	E-NS_2l	0 7	0 1	PADtrtc	_ 5	0 2	0701PADtrtc___05L2
Total persons working in education	07.Economy	01.Employment	2005	LAU2	E-NS_2m	0 7	0 1	EDUtrtc	_ 5	0 2	0701EDUtrtc___05L2
Total persons working in health and social work	07.Economy	01.Employment	2005	LAU2	E-NS_2n	0 7	0 1	HEAtrtc	_ 5	0 2	0701HEAtrtc___05L2
Total persons working in other community activities	07.Economy	01.Employment	2005	LAU2	E-NS_2o	0 7	0 1	COMtrtc	_ 5	0 2	0701COMtrtc___05L2
Total persons working in activities of households	07.Economy	01.Employment	2005	LAU2	E-NS_2p	0 7	0 1	HOUtrtc	_ 5	0 2	0701HOUtrtc___05L2
Total persons working in extra-territorial organizations	07.Economy	01.Employment	2005	LAU2	E-NS_2q	0 7	0 1	ORGtrtc	_ 5	0 2	0701ORGtrtc___05L2
Number of unemployed persons by age	07.Economy	02.Unemployment	2007	LAU2	E-NS_5	0 7	0 2	UMPtrtc	_ 7	0 2	0702UMPtrtc___07L2
Long-term unemployment	07.Economy	02.Unemployment	2007	LAU2	I-NS_1	0 7	0 2	UMPlngrtc	_ 7	0 2	0702UMPlngrtc_07L2
Part-time unemployment	07.Economy	02.Unemployment	2007	LAU2	E-NS_7	0 7	0 2	UMPprtrtc	_ 7	0 2	0702UMPprtrtc_07L2
Number of employees by size of the company	07.Economy	01.Employment	2007	LAU2	E-NS_8	0 7	0 1	EMPcpnrtc	_ 7	0 2	0701EMPcpnrtc_07L2
Number of persons with secondary education degree	06.Social Affairs	01.Education	2007	LAU2	E-NS_10a	0 6	0 1	EDUscdrtc	_ 7	0 2	0601EDUscdrtc_07L2
Number of persons with tertiary education degree	06.Social Affairs	01.Education	2007	LAU2	E-NS_10b	0 6	0 1	EDUtrtrtc	_ 7	0 2	0601EDUtrtrtc_07L2
Number of students of higher education institutions	06.Social Affairs	01.Education	2007	LAU2	E-NS_11	0 6	0 1	EDUhghrtc	_ 7	0 2	0601EDUhghrtc_07L2
Number of companies created	07.Economy	01.Employment	2007	LAU2	E-NS_12	0 7	0 1	FIRopnrtc	_ 7	0 2	0701FIRopnrtc_07L2
Number of companies closed	07.Economy	01.Employment	2007	LAU2	E-NS_13	0 7	0 1	FIRclsrtc	_ 7	0 2	0701FIRclsrtc_07L2

TECHNICAL REPORT

ESPON

Metadata GUIDE

Acquisition and storage of Data and Metadata in ESPON 2013 DB

CONTENT

Definition of an ESPON 2013 DB profile of ISO 19115 standard. The differences between the edition of socio-economical or environmental are handled through the use of various templates inside the geonetwork editor.

Design of the ESPON data and metadata flow which ensures the importation, the integration, the query, and the exportation of data with their associated metadata.

Adaptation of the open-source Geonetwork V2.4 software, as the first prototype of ESPON metadata editor.

ESPON 2013 DATABASE

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

30 PAGES

LIST OF AUTHORS

Jérôme Gensel, University Pierre Mendès-France, UMR 5217 LIG

Claude Grasland, Université Paris I et VII, UMR 8504

Benoit LeRubrus, UMR 5217 LIG

Roger Milego, UAB

Bogdan Moisuc, UMR 5217 LIG

Christine Plumejeaud, University Joseph Fourier, UMR 5217 LIG

Maria José Ramos, UAB

Anton Telechev, UMR 5217 LIG

Ronan Ysebaert, UMS RIATE

Contact

christine.plumejeaud@imag.fr
mariajose.ramos@uab.es
ronan.ysebaert@ums-riate.fr

tel. + 33 4 76 82 72 80

TABLE OF CONTENT

Introduction

The initiative of creating and organising metadata is considered as a transversal challenge for the ESPON 2013 Database project in the First Interim Report (2009 February 2007). This transversal challenge is a crucial point, either for INPUT in the ESPON database or OUTPUT, since a rich database would be useless without providing the users with appropriate information or tools to allow them to have full advance knowledge of the existence and characteristics of the gathered information.

In order to ensure a long-term use of the ESPON database and its interoperability with other systems, the metadata schema is based on the INSPIRE initiative and ISO standards.

The INSPIRE (Infrastructure for Spatial Information in Europe) directive states that the Member States should provide descriptions in the form of metadata to guarantee users to find data and to establish whether they may be used and for what purpose.

Technically, the INSPIRE directive recommends the use of the standard ISO 19115 or Dublin Core for data with geographic references, which is the case of all datasets (environmental or socio-economic) within ESPON DB 2013. The ISO 19115 can be easily mapped with the Dublin-Core, which is actually an extension of the ISO, and it is most used within different institutions such as the EEA and JRC among others.

The aim of this document is to present the ESPON schema for socio-economic metadata and the data & metadata flow inside the ESPON system. They are the results of many internal discussions that have been held within ESPON 2013 database project, based on the theoretical background established in the First Interim Report.

1 Definition of a profile for socio-economic data

Description of the socio-economic data and the associated extent of the ISO 19115 standard.

1.1 The socio-economic data format

Data provider delivers his/her statistical information to ESPON 2013 DB using spreadsheet data files with a standard structure specially defined to be integrated in the ESPON database.

The objective of a standard structure data files is to obtain all the necessary elements defining an indicator value. The structure of the files has been built taking into account the compromise between computer requirements (automation mechanisms and integration in the system) and user requirements (for experts as well as for non-expert users). This structure is an adaptation of previous files used by statisticians to exchange their data, but we have added some explicit information: the temporal validity, the geometry version used for the statistical unit, its scale level, etc.

Figure 1 shows an example of a fragment from a standard input data file where it is possible to identify some of their elements.

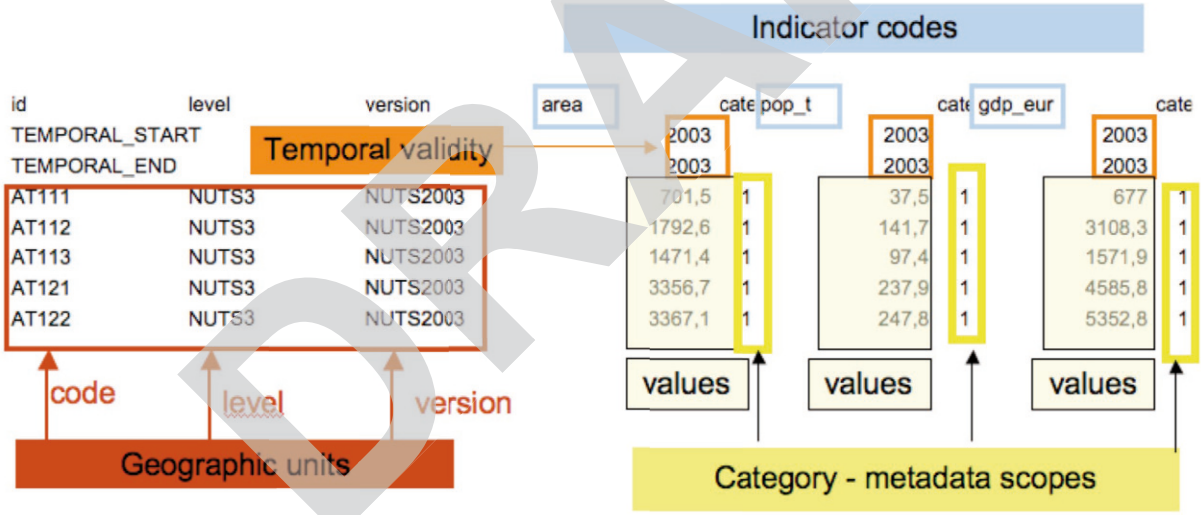


Figure 1: Data organisation in Excel files

- **Indicator codes**: these elements are valid for an entire column so they can be put into the header of the column that contains the indicator value. The codes follow the EUROSTAT naming conventions
- **Temporal validity**: can be an instant (the start equals the end) or an interval (used when measuring flows). It gives the validity time of the values stored in the column.
- **Geographic units**: each dataset is associated with a certain nomenclature describing the spatial composition of the geographical space: NUTS, WUTS, UMZ, etc. The first column gives the code of the unit in the nomenclature, the second column precises the

level of the unit inside the nomenclature, the third column presents the version of the geometry of the unit used to compose this dataset.

- **Category**: each value of the indicator is associated with a metadata label refering to quality information that is described in the metadata file. The reason for these labels is that each indicator's value can have a different quality, and values having the same quality are grouped by scope (a spatial and temporal context).

1.2 Three levels of metadata information

The metadata information has been organized at three different levels establishing a mechanism of specialization and refinement for the information, specially the quality information: dataset, indicator and value.

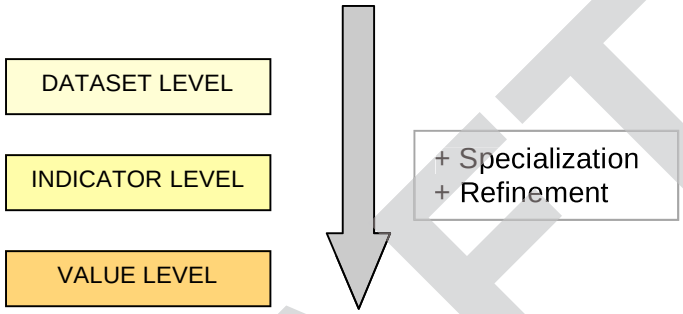


Figure 2: Inheritance mechanism: the information defined at one level can be refined at the lower level.

The ISO 19115 standard has been extended to reflect this information organisation. In the First Interim Report (2009 February 2007), we presented the various topics that the standard proposes to fill information, as shown in *Figure 3*.

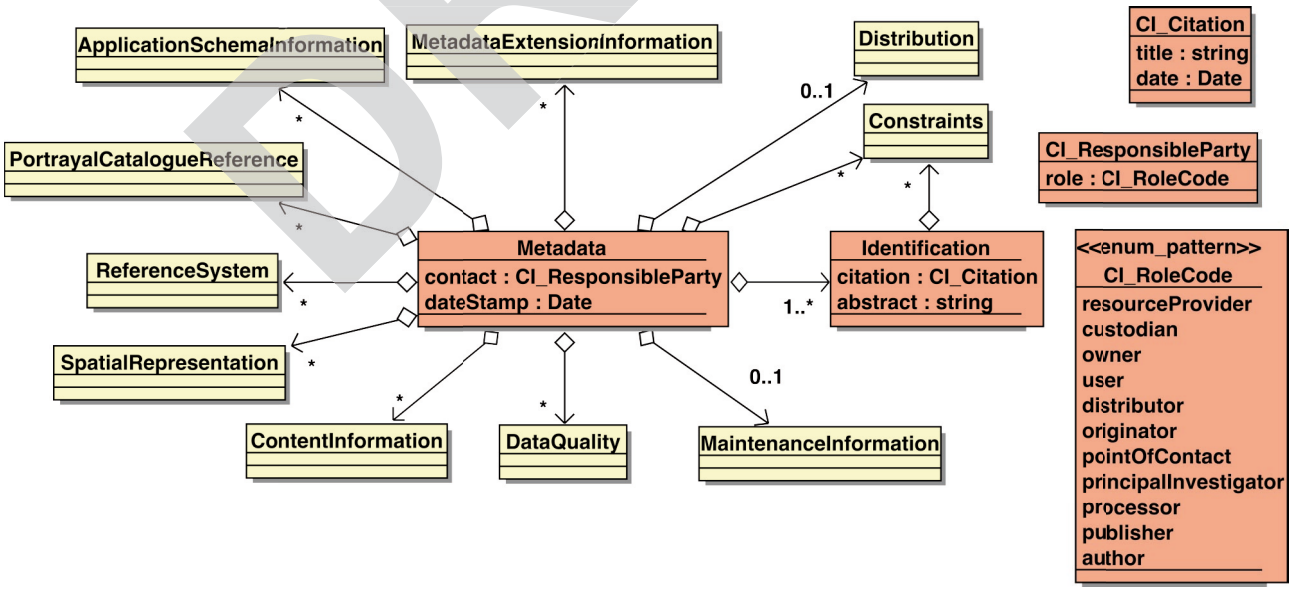


Figure 3: ISO 19115 schema – main topics.

Through a deeper study of the standard, it appears that the standard has a recursive structure and can be used to define metadata on series of dataset, as shown on the *Figure 4*. The ESPON extension re-uses this facility, and uses it to define dataset metadata at the *series* level of the standard, and indicator

metadata at the *dataset* level of the standard. The schema defining the extension is fully included inside the *Annexe 1* of this document.

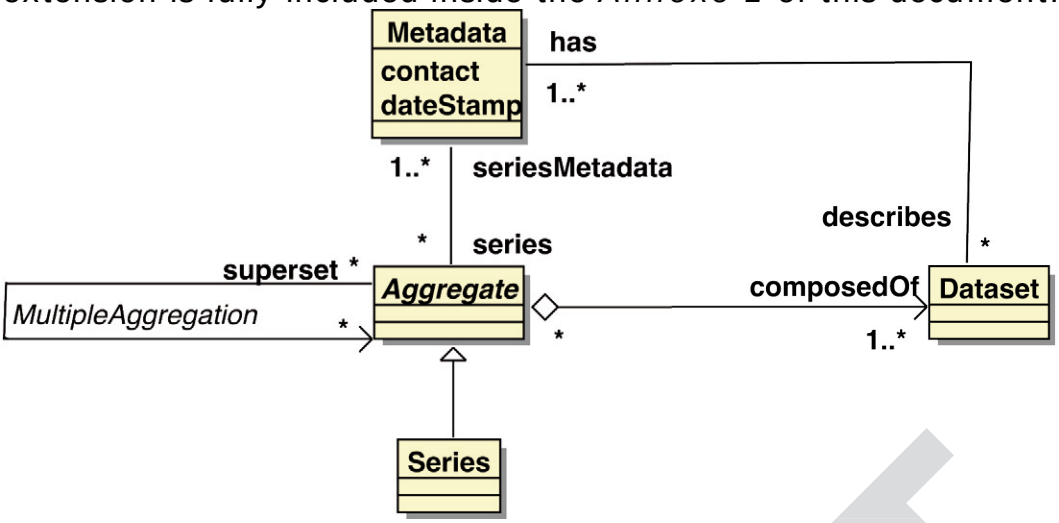


Figure 4: Recursive structure of the ISO 19115 schema.

1.2.1 Dataset level

The first level defines the main characteristics of the whole dataset. It includes all the common points for all the elements inside the dataset and also the metadata description like the name of the delivery, the authors, etc. The fields in blue will be exported from the database, but shouldn't be filled by the user, since they can be deduced from the data file.

Information	Description
Dataset name	Name of the data file grouping the collection
Acquisition date	Date of record of metadata/data into DB
Contact Point	Name, email, etc of the author of this metadata
Maintenance	Any useful information (frequency, next update)
Distribution information	This topic mentions the distributor of the data (generally, the ESPON 2013 DB project), as well as the conditions for obtaining those data: fees, URL for download, etc. Whenever a specific project does not distribute the data for free, its conditions must be mentioned through this topic. Otherwise, the topic is optional, and filled by the ESPON 2013 Database Project by default.
Metadata standard	Names of the metadata standard, and metadata profile
Spatial representation	The geometry layer name used for filling in the value: by example, NUTS. Many scales can be used (from NUTS-5 to NUTS-0) A dataset can mix various versions of a geometry : for example, mixing NUTS2006 with NUTS1999 is authorized.

Tableau 1. Information of the dataset level

1.2.2 Indicator level

The second level of metadata information is based on the characteristics of each socio-economics indicator. Inside a dataset, each indicator has its particular specifications and characteristics. The selection of appropriate themes and keywords is a key point, at this level, to obtain a good description of the resource.

In order to ease the edition of indicators and to harmonize definitions and codes inside the database, the editor will provide a list of well-known indicators (that may be already present in the database).

Information	Description
Code	Code of the indicator (harmonized with existing nomenclature)
Name	Name of the indicator
Abstract	Full text description of the indicator's meaning
Unit of measure	Can be meter / square, thousands, etc: free text
Classification (themes, keyword)	Various themes and keywords found into thesauri to describe the resource. Multi-indexes are allowed: you can choose a theme, attach 1 to n keywords, and then select another theme, and attach as well your keywords.
Language	Language used to describe indicator field. ENGLISH by default in all ESPON metadata
Temporal extent	The oldest and more recent dates for the values to be found

Tableau 2. Information of the indicator level

1.2.3 Value level

The value level is based on the characteristics and description of the indicator values and it is linked with the "category column" of the data files (see point *3. The socio-economic data*). This is the most specific level because it defines the main characteristics of the value of a specific indicator that belongs to one dataset. The lineage and reliability contribute to give information about the quality, the procedures for data collection, the sources and the methods used for delivering and transforming data.

The information for each value is grouped by "scopes". A scope defines a subset of data that have the same lineage. For example, if a dataset for the EU 27 contains official data coming from Eurostat, from the Romanian and from the Bulgarian national statistical institutes, the values metadata will be grouped in three scopes, one for EU 25 values (source EUROSTAT), one for Ro values (source: Romanian national statistical institute) and one for Bg values (source: Bulgarian national statistical institute).

Information		Description	
Scope			
	Label	Label in front of each value in dataset, that allows to map metadata information with data information	
	Spatial validity	The list of geographical units (the highest scale level is used) attached to this value	
	Temporal validity	A list of validity instant or interval for the value	
Lineage			
	Provider	The statistical institute or ESPON project that provides the values	
	Source	The document reference (file name or URL) and date of extraction of the data.	
	Methodology	URL for report	A pointer towards a more complete description of the methods used to obtain these data. This is required whenever modifications has been applied by the provider to the original data, for instance when data are complex indicators that were built from less complex indicators, when data are result of corrections or estimations of the original data, etc. One or more reports can be inserted, including intermediate data in excel sheets in order to illustrate more clearly the methodology
		Formula	A formula expressed in a semi-formal language based on the indicator codes (**Math ML** for instance)
		Text	A free text description of the methodology used for estimating, calculating or correcting the data in this scope
Reliability			
	Official	*true* if the value has been certified by an official institute	
	Estimation	*true* if the value comes from an estimation	
	Quality	A level of confidence to the reliability of the value; *low, medium, high, no opinion*	
Constraint			
	Data access	*true* (is labelled "public") : the data is public which means all users, without any restriction, can view and download the data *false* (is labelled "ESPON community") : means that only members of ESPON projects will be able to download those datasets	
	Metadata access	*true* : users can know whether these data are collected in the database, whatever the data access's value. *false* : nobody can learn about the existence of these data inside the database, because their metadata are hidden to every body. It is a special issue for purchased data that can only be used on the demand of an expert.	
	Copyrights	A text area will allow users to specify the rights (for use, for access) associated with their data.	

Tableau 3. Information of the value level

1.3 Environmental versus socio-economic information

In the process of creating an appropriate metadata schema for Espon, it is necessary to take into account the two different types of data that will be gathered and integrated: environmental and socio-economic data.

Although there are some shared points between the information required to describe environmental and socio-economic data, there are also some crucial differences that will influence the final structure.

Shared points:

- The information can be organised in three levels: dataset, indicator and values.

  Note: Although raster format and a single indicator with homogenous quality on the whole set are more common, vector format and heterogeneous quality levels inside the dataset also exist.
- The information required to describe the dataset must be: metadata information, maintenance, constraint and distribution.
- The information required to describe the indicator must be: identification, extent, content.

Main differences:

- For environmental data some specific information is required to have an appropriate description of the dataset. In this case, the spatial representation and spatial reference attributes are more detailed compared to socio-economic data.
- For socio-economic data the quality, the lineage and reliability information are essential to have a good understanding of the production and validity of the dataset.

2 Metadata and data flow

2.1 Description of the flow

The ESPON data flow and structure ensure the importation, the integration, the query, and the exportation of sets of statistical indicators linked to geographical units.

The metadata or data file provide criteria to implement the search data functionality. The criteria should be able to answer four basic questions:

- What? : What is the information gathered? Or does a dataset on a specific topic exist?
  - Provided by the indicator information and dataset name.
- Where?: Does a dataset/indicator on a specific topic exist for a specific place?
  - Provided by geographical extent information.
- When?: Does a dataset on a specific topic exist for a specific date or period?
  - Provided by temporal extent
- Who?: Which is the point of contact to learn more about or order the dataset?
  - Provided by data providers and dataset contact information.

The next figure (*Figure 5)* shows the simplified data flow of data and metadata files in the ESPON system.

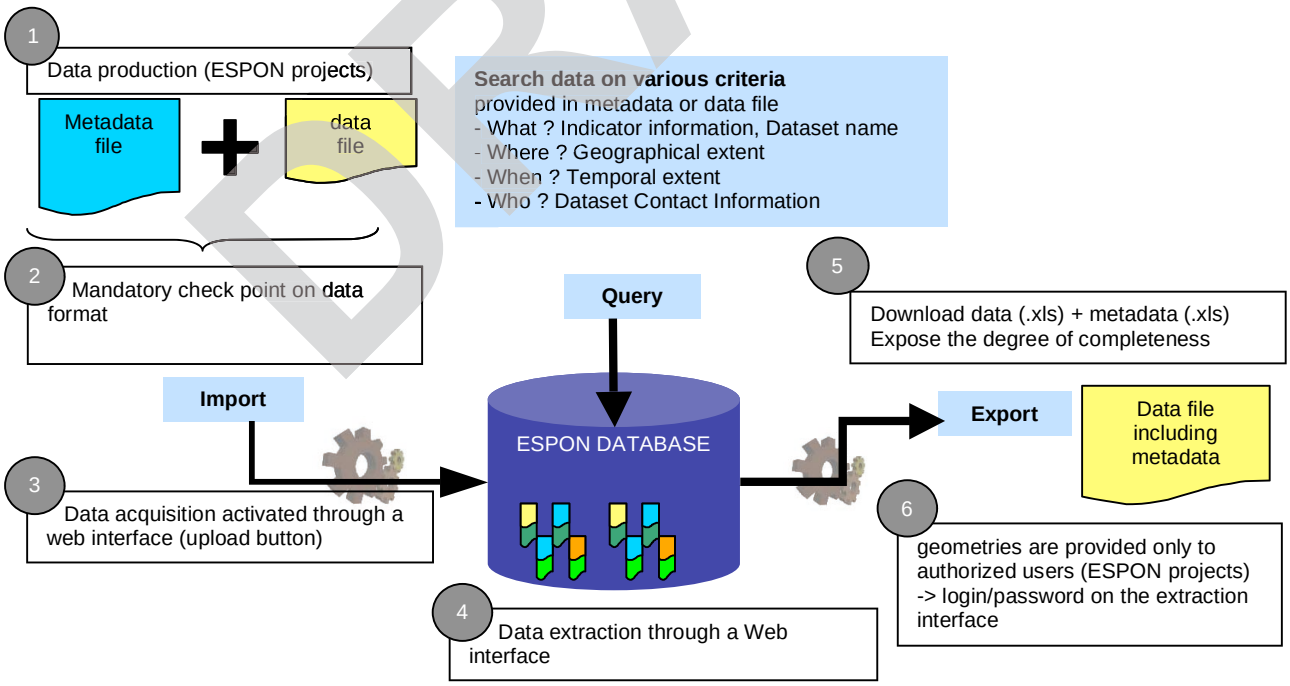


Figure 5: Data and Metadata flow inside ESPON system

First, data are prepared by ESPON internal projects or external providers and upload on the ESPON server: through a Web interface (see *Figure 6)*, authenticated users

upload the pair of data and metadata files, plus extra files giving detailed information on lineage and transformation process (PDF, DOC, XLS, PPT, or any format are accepted, provided they are zipped in a file, less than 20 Mo in size).

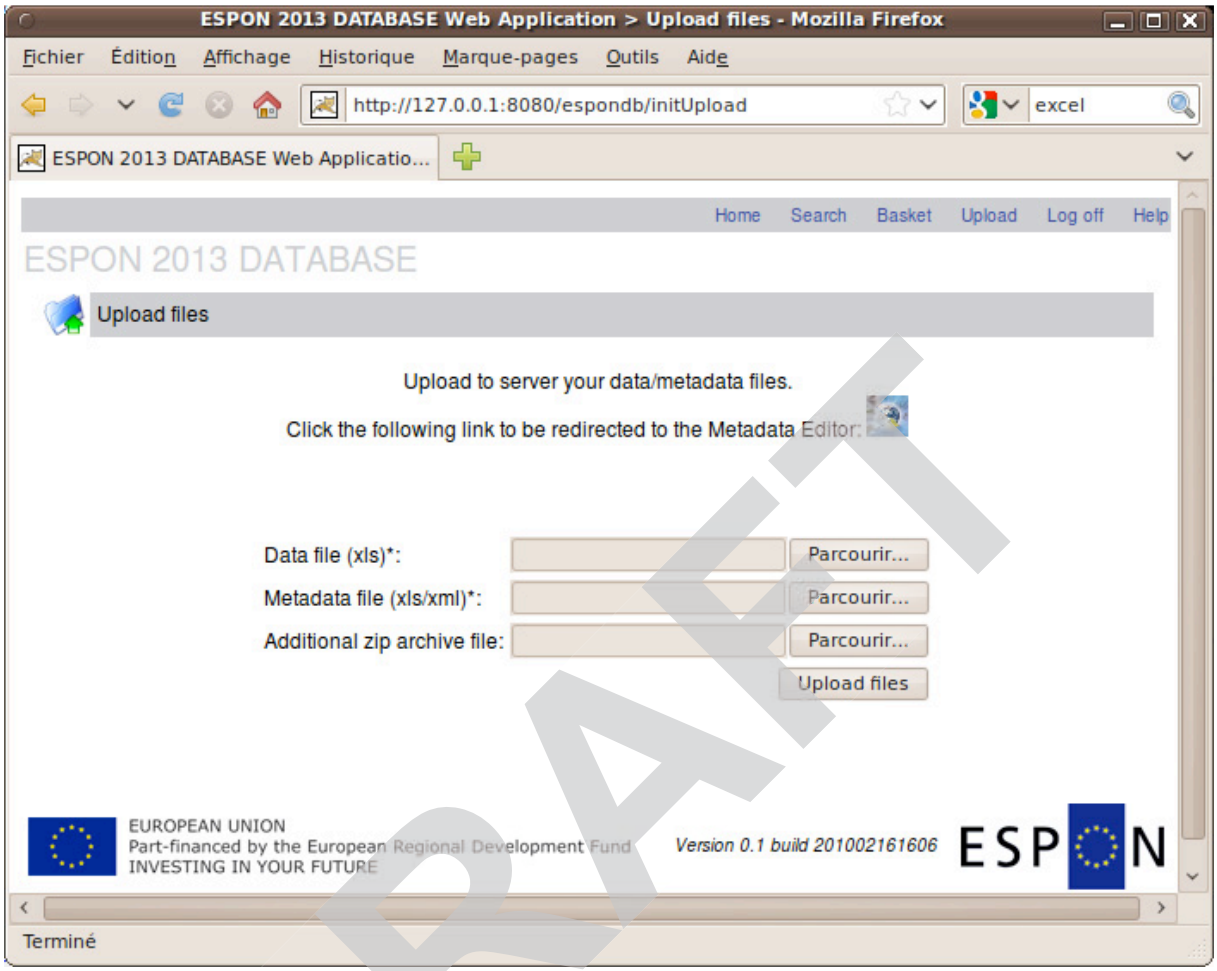


Figure 6: Web interface for data and metadata upload

The metadata file can be either in the Excel format, like the one defined in the ESPON technical report sent in June 2009 to ESPON partners and entitled "*Guidelines for metadata and data in ESPON 2013*", or in the XML format. In order to build a conformant XML file, the project gives access to an adapted version of geonetwork, supporting our ISO 19115 profile, through which the user can edit, check and export a valid XML file of metadata.

Then, a checking is made by ESPON 2013 dabase project to validate the data: ESPON 2013 dabase project inform the data provider about the acceptation/rejection of the data. It can be due to a non-conformant metadata or data file, regarding the format that has been defined. In case of acceptation, the ESPON 2013 DB project import data into the ESPON database. During this IMPORT phase, the system enriches metadata descriptions.

Finally those data are then redistributed together with their associated metadata, through the export tool. The export tool allows users to query data on metadata criteria, and retrieve data with their associated metadata, inside the same XLS file.

For the metadata acquisition process, the ESPON 2013 dabase project has evaluated the workload and technical limitations to find the most appropriate solution. Due to the great number of inconsistencies, format errors, and mismatches that can be found inside data files, it was decided that all specific metadata that could be deduced from data files, would be shown at EXPORT phase, and not during the metadata acquisition phase, which occurs before the IMPORT phase.

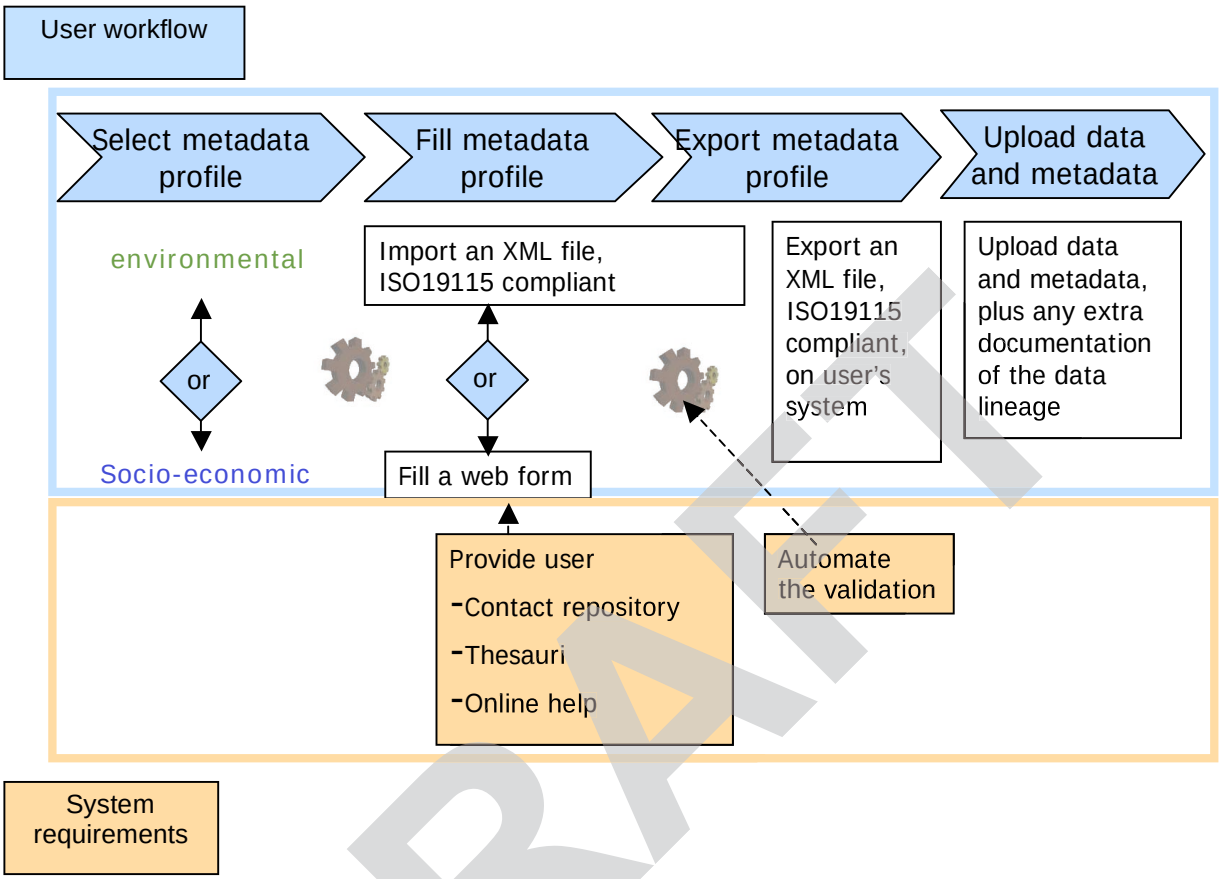


Figure 7: System requirements for the proposed user workflow

1. First of all, the user connects onto the metadata editor, and selects the specific metadata profile (environmental or socio-economic).

2. The user can import an existing XML metadata file, or fill a Web form of the metadata profile. To ease this work, the system should provide some tools to the user, such as contact repository, thesauri and online help. For now, *geonetwork* does not provide Contact repository neither Thesauri, which are elements to get from the ESPON database. Those elements should be connected before the end of 2010 year.

3. Once the metadata is filled in, the system validates the information automatically. If the information introduced is correct, the metadata xml file, which is 19115 ISO compliant, is created. User save the file on its own system using an Export button.

4. Finally, the user uploads the data and metadata files (and additional files) through the ESPON Web portal. When parsing the uploaded files (if the check has found no format error), the system is able to compute as much as possible metadata information automatically. For example, when reading a data file, the category column associated with the indicator column allows the system to deduce the spatio-temporal extent of each Quality scope (validity time and code of concerned units).

2.2 Filling a socio-economical profile

The process of filling in a profile has to be consistant with the different designed levels of information: dataset, indicator and value. Following the different levels, the first information required is about the dataset, then about the indicator and, finally, about the values (quality information).

This has the drawback of forcing the user to enter a quality topic for each indicator, even though the user would naturally consider that a certain quality topic could apply equally for various indicators. This is why the user can specify such common lineage at dataset level, using the labels to identify the scope of such information.

Moreover, the constraints information can also be specified on specific area, using the scope label. Otherwise, the full dataset might have been in restricted access for all data, due to a limitation access specified on a few part of the dataset. For example, on Poland, data are not in free access for certain employment indicators, whereas there is no restriction in other European countries for those data. Defining constraint at the value level is more flexible, since constraint, like lineage, can be specified for any spatio-temporal part of the dataset. For example, the previous dataset could be available on the full Europe area, minus the Poland's data. This mechanism matches better the main goal of ESPON's project : the collect, documentation, and dissemination of data for a broader audience.

The *Figure 8* shows a schema of the filling process of the socio-economic profile

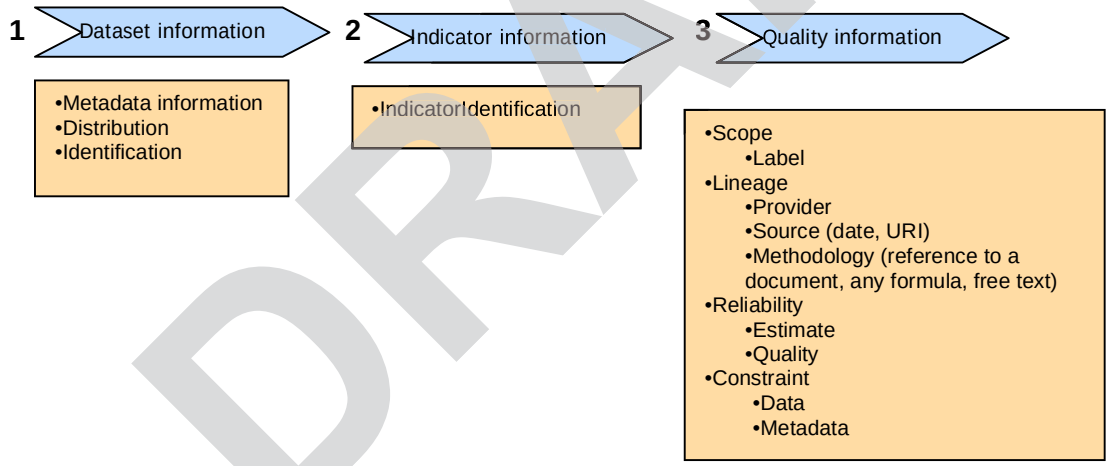


Figure 8: Edition steps for the ESPON profile

At the dataset level, as it can be seen in the *Figure 8*, information about the metadata, its identification, and distribution is required.

The indicator level inherits properties of the dataset level concerning the distribution, maintenance, metadata information, and so on. It is necessary to identify the indicator, using new fields (name, code, abstract, units) that are added to the extension of the ISO 19115.

Finally, the value level, which is the most specialized level, requires information about the lineage and reliability of the values. In this case, at least one quality topic is mandatory, that applies by default to the whole set of values. Still it is possible to add another lineage to refine the information for a sub-set of values.

2.3 Storage model of metadata associated to data

UML model has been developed for the socio-economic metadata profile. This model describes the elements that belong to the profile and their relationships.

It is important to highlight that this model has some **limitations concerning composite indicators**. Indeed, the lineage is the association of **one** indicator and **one** source. If a value is the result of the combination of many indicators of various sources, the user can only specify this fact in the methodology field. By means of this field, the user can get the documentation (report, formula, text, or URI) associated with this socio-economic composite indicator, knowing that it has been produced by ONE source (the project or institute that establishes this composite indicator). The formulas of normalization, weighting, and so on, that can exist between various indicators are very difficult to formalize.

The *Figure 9* presents the UML model. The green colour elements refers to the Dataset level, the pink elements to the Indicator level, and the yellow ones to the Value level.

Some points about the UML model:

- Each Dataset element has a relation one to one with: Contact, Distribution.
- Each Dataset is associated to one Maintenance object, that records the dates of the versions of Dataset (the name of a Dataset is used to identify a Dataset).
- Each Indicator element has a relation many to many with Theme. The Keywords are attached to the Theme.
- The Lineage is the association of one Indicator and one Dataset, and it handles all genealogical information for one Indicator of a given Dataset. It is thus qualified by one Reliability, one Source, and one Constraints applying to the values attached to this Lineage. This means that one Indicator can be linked with one Dataset, producing as many Lineage as necessary to qualify the values. It means also that one Lineage can be associated with many Indicators, coming from the same Dataset, or from other Dataset (this case is less frequent).
- Each Lineage element has a relation many to one with the Values: a Lineage can apply to one or more Values but one Value only refers to one Lineage.
- Each value is associated to one GeographicUnit. All the spatial information (that is to say the name of the nomenclature in use, the level of the unit in this nomenclature, the version of the geometry file used to define the spatial extent of the unit) is hidden into this restricted view of the ESPON database schema.

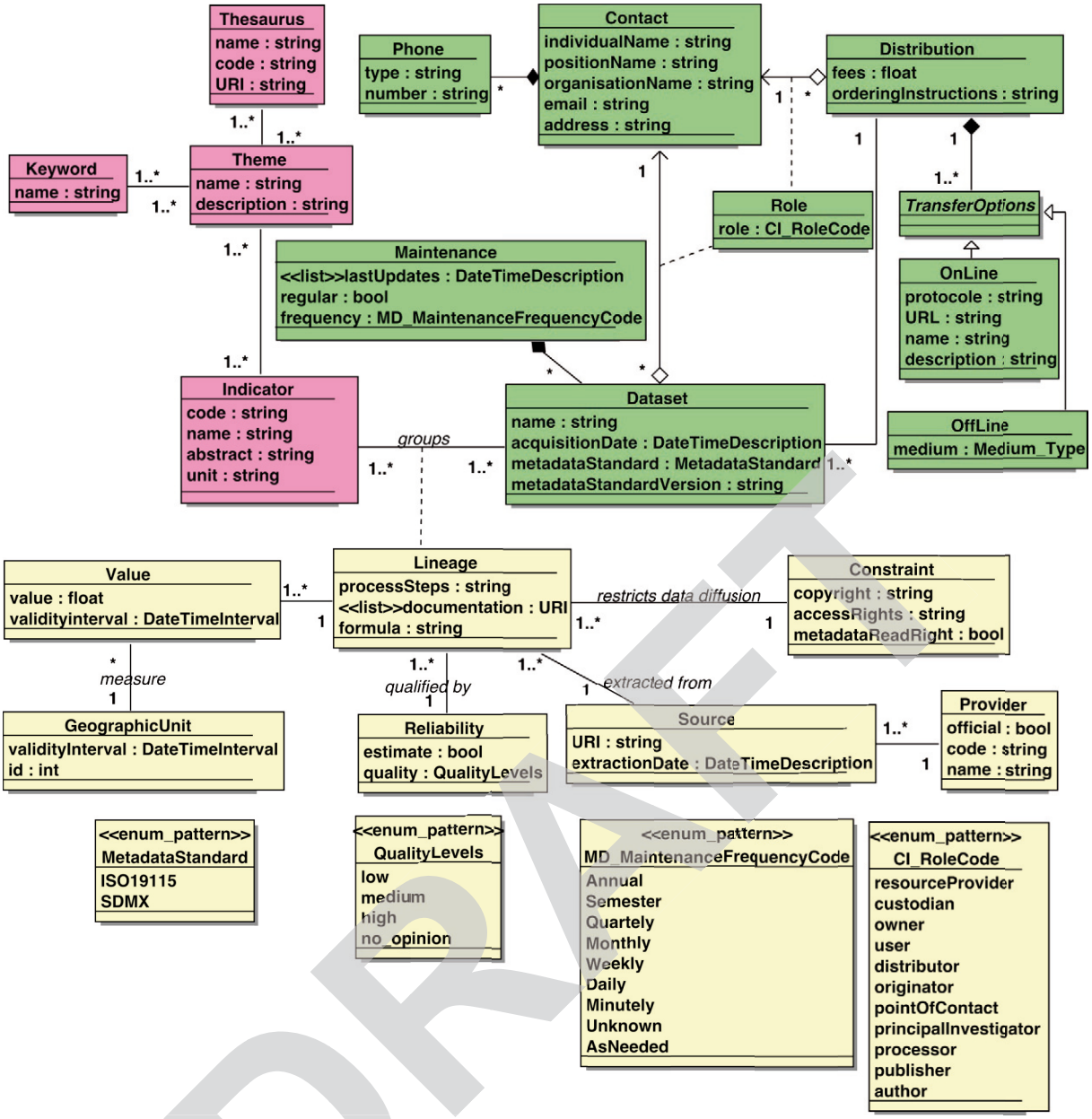


Figure 9: UML model of metadata/data storage into ESPON database, restricted to the thematic part

It appears then that a lot of information (such as extent of dataset, maintenance frequency, spatio-temporal extent of indicators, geometry version and levels for spatial representation) can be deduced automatically at EXPORT phase.

- Maintenance frequency: we use the Maintenance object, which is linked to various updated version of a Dataset (a Dataset is identified uniquely by its name, a Dataset version by the name and the associated acquisitionDate).
- Spatio-temporal extent of an Indicator in a Dataset: by selecting the corresponding Lineage, and associated values, we can compute the temporal Extent by aggregation of validityPeriods; as well, the spatial extent by geometric aggregation of GeographicUnit associated to each values.
- Spatial representation of a Dataset: by selecting all Lineage and corresponding values and their associated GeographicUnits, we can compute (using the hidden part of the model) the geometry version and geographic scale of the selected GeographicUnits.

This will help to simplify the metadata acquisition process by reducing the number of mandatory fields.

3 Presentation of the first prototype of metadata editor

A metadata editor is a key point for creating, importing/exporting metadata and searching data. Several application concepts can be used while developing such a tool, but a Web-based editor seems to be the best option to fulfil the requirements of ESPON 2013 DB project. Web-application architecture makes unnecessary to provide users with special hardware/software and facilitates future updates of the software. Geonetwork v2.4 tools have been chosen as the base for creating a first prototype version of a Web editor for metadata.

In February 2009, the First Interim Report for ESPON 2013 DB established a list of basic functionalities that should be provided by the metadata editor. Later, in June 2009, some additional requirements were expressed: supply tools for editing **indicator and provider information**, when that information is already present into the database. That possibility can help to avoid duplicating entries in the database. For instance, in order to provide an indicator that is already present in the database, the user selects it in a list, and all indicator fields are pre-filled: name, code, abstract, unit of measure, and classification. In the same way, to indicate a provider as the source of a value, user should select the provider inside a list, and has only to indicate the download URI, and the date of data acquisition. When the indicator or the provider are not known inside the DB, and are not available inside their respective lists, the user can specify that the information is new, and then edit the provider's data, or the fields of the indicator (name, code, abstract, unit of measure).

3.1 Adaptation of geonetwork

The base Web application for the metadata editor is Geonetwork application (http://geonetwork-opensource.org), that is an open source GPL software. It lets users manage spatial information and manipulate several standards of metadata specifications for geospatial datasets. Its handy Web GUI gives a possibility to optimize metadata creation and concentrate user's attention on metadata accuracy rather than on editing XML metadata schemas. Geonetwork used Jetty HTTP server (http://jetty.codehaus.org/jetty) and extends Jeeves API framework (http://sourceforge.net/projects/jeeves) with numerous servlets to ensure application functionality and interaction with Mckoi Distributed Database (http://www.mckoi.com). Client-side browser Web content is produced by Jeeves framework using XSL transformations.

In spite of the fact that Geonetwork already supported several metadata standards, it had to be extended to be adapted for managing socio-economic metadata profiles in the context of ESPON 2013 DB project.

The first step to adapt Geonetwork to ESPON 2013 DB needs was the creation of an XML data grammar and scheme that extend ISO rules. The grammar is available in the **Annexe 1**, and will be published with the following namespace `xmlns:esponMD="http://www.espon.eu/esponMD".` It also entailed all necessary configurations of Jeeves java servlets and modifications of XSL transformation rules to adapt them to new Geonetwork functionalities. This was done using the following

developper's guide (http://geonetwork-opensource.org/documentation/how-to/geonetwork-v2-2-shema-template-howto) designed for the previous Geonetwork version (v 2.2), which helped for the first steps of the work, but should be adapted for the new version of Geonetwork v2.4 published in October 2009.

After implementing in Geonetwork the possibility to edit and stock ESPON 2013 DB compliant metadata files, the rest of the development work was dedicated to adding several new functionalities as well as to eliminating minor errors and issues that arose from recent changes.

When the user is working with a metadata form in our extended Geonetwork platform, he now has a possibility to download directly the XML file that corresponds to his metadata. That option was partially present in search results list of the earlier version of the platform, where the user could show the metadata XML code and copy/paste it in his text editor. Now, an 

button appears onto metadata editing form and lets the user obtain the XML code at any moment. After a click on the button, the browser receives the XML file from the Geonetwork server and opens a file save dialog. That scenario is very useful regarding the use of metadata files in the main ESPON 2013 DB application: after saving the XML metadata file in Geonetwork, the user will only have to import it into ESPON platform.

Another useful functionality that was added to Geonetwork is the revision of the search results form. In the previous version, all metadata DB entries found according to any search criteria appeared in the results list titled with the name of their corresponding template. This was very confusing, because the user had to read additional information about the metadata found or even open the preview form to find the entry that he was looking for. In our extended version, ESPON metadata entries appear in the search results form with their custom titles; searching the necessary metadata entry is obviously easier now.

Some other minor optimizations of metadata edit form have been also made. They mainly concern correcting and extending the user interface: adding text areas (instead of text fields) for the form entries that are capable to contain larger amounts of text data entered by the user; adding file chooser buttons to facilitate the indication of paths to uploaded files; removing ambiguous "save as template/metadata" combo box; removing blanc entries into fixed-list enumerations in combo boxes, and many other minor changes.

3.2 Using the metadata editor

According to *Figure 7* of the present report, we can detail the following use cases of metadata editing in Geonetwork:

Case 1. Creating a new ESPON 2013 DB metadata

In administration panel (see *Figure10*), click on *New metadata* link

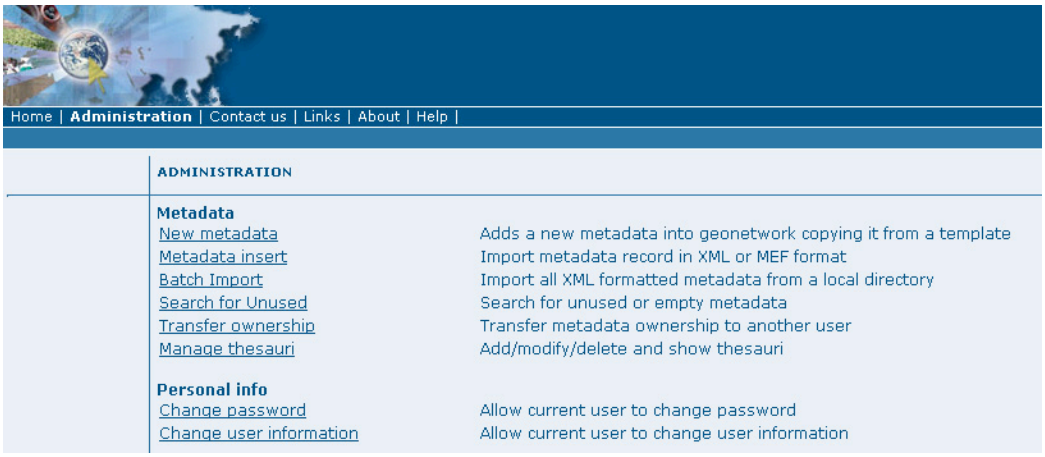


Figure 10: Administration menu for metadata files management

The metadata creation dialogue should appear (*Figure 11*). Choose *Metadata template for ESPON socio-economic basic indicators* in *Template* combo box and click on *Create* button.

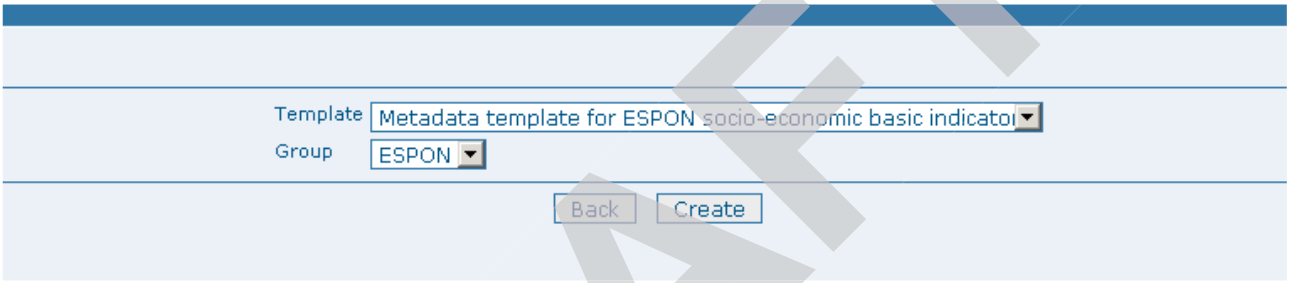


Figure 11: Selecting a template for metadata edition

Geonetwork application will now create a new metadata entry in its database and will show the form corresponding to case 3 of the present tutorial.

Case 2. Importing a metadata file or a metadata template into Geonetwork

In administration panel, click on *Metadata insert* link. This should bring out the *Import metadata* dialogue (*Figure 12*).

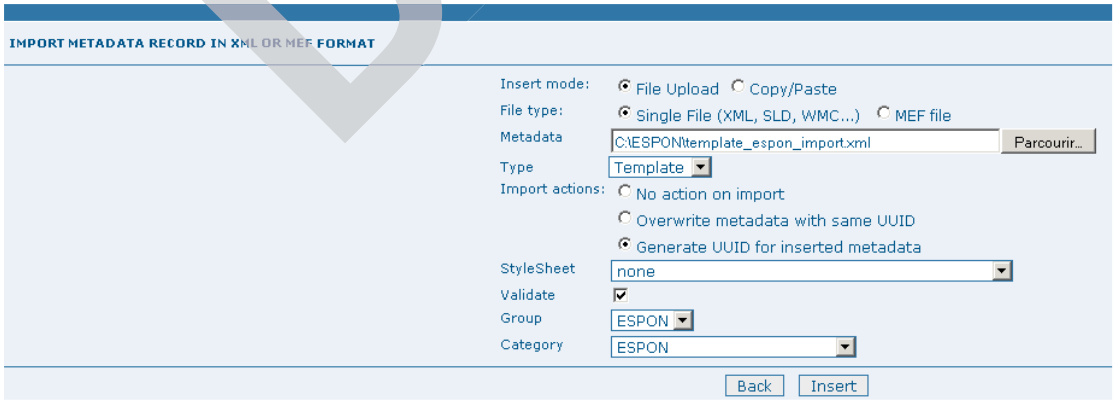


Figure 12: Importing an XML metadata file

Specify insert mode properties, file type and path. Don't forget to choose “template” in *Type* combo box if you want to import a template. In that case, you will be able to use imported metadata as a template in *New metadata* creation dialogue described above.

Case 3. Editing a metadata / template entry

After creating a new metadata or choosing *Edit* option for an entry in search results form, metadata edit form is shown (*Figure 12*):

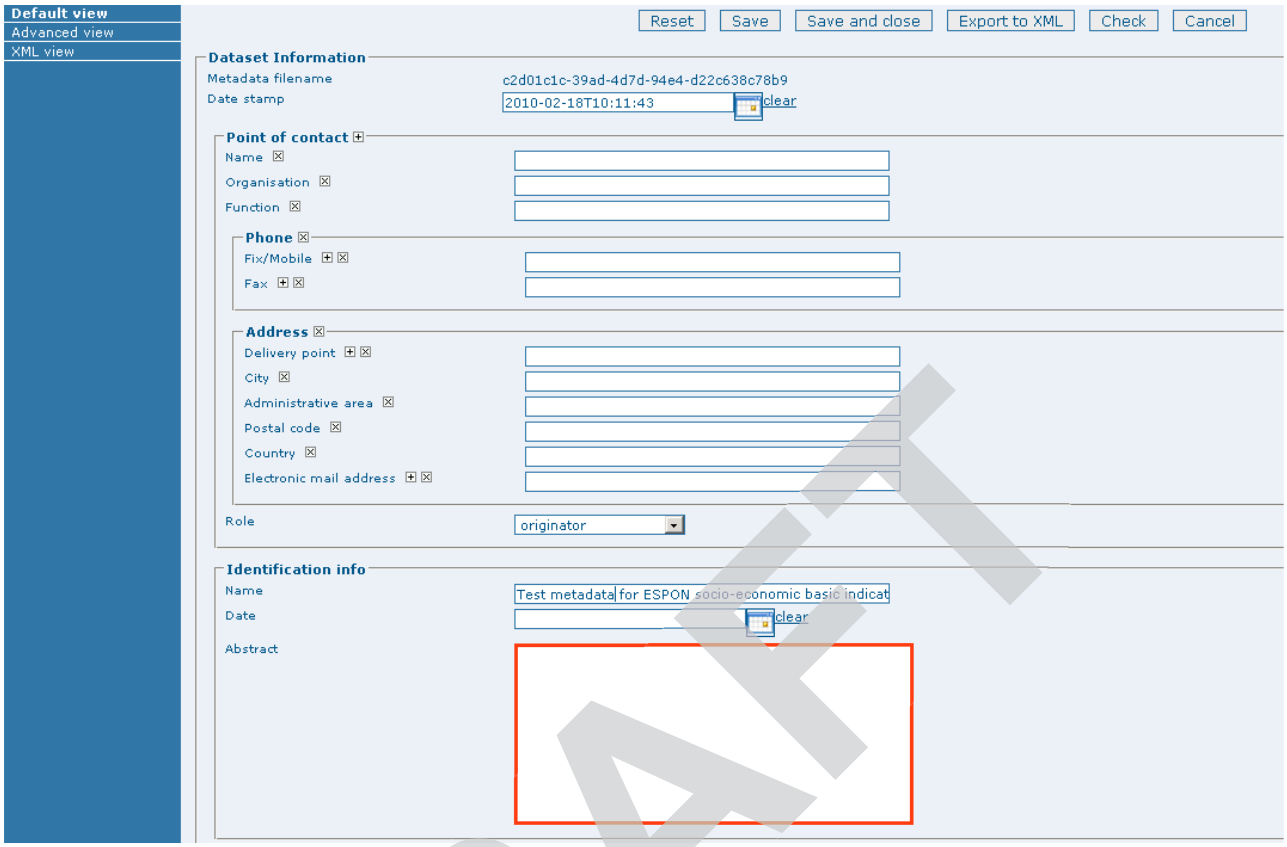


Figure 12: Editing a form to build an XML metadata file

Here, you should carefully specify all the characteristics of your metadata. The mandatory fields that must be specified are highlighted in red colour. Don't forget to fill them if you want your metadata to be valid.

In order to optimize your editing of metadata for further importing into ESPON 2013 DB application, you can directly obtain the XML file for your metadata from that dialogue. After a click on *Export to XML* button, the current metadata entry will be updated into Geonetwork database and the server will generate an send the corresponding XML file. A download dialogue will appear to let the user choose save options for the received file.

Here after we present some screenshots of editing XML metadata process, based on true use case for metadata and data.
Figure 13 shows the dataset level edition panel, *Figure 14* shows the Distributor information (filled by default), *Figure 15* shows the Indicator information and *Figure 16* shows the Quality information.
Note that the field *methodology* has been removed from the Indicator level, and belongs to the Lineage topic of the value level.

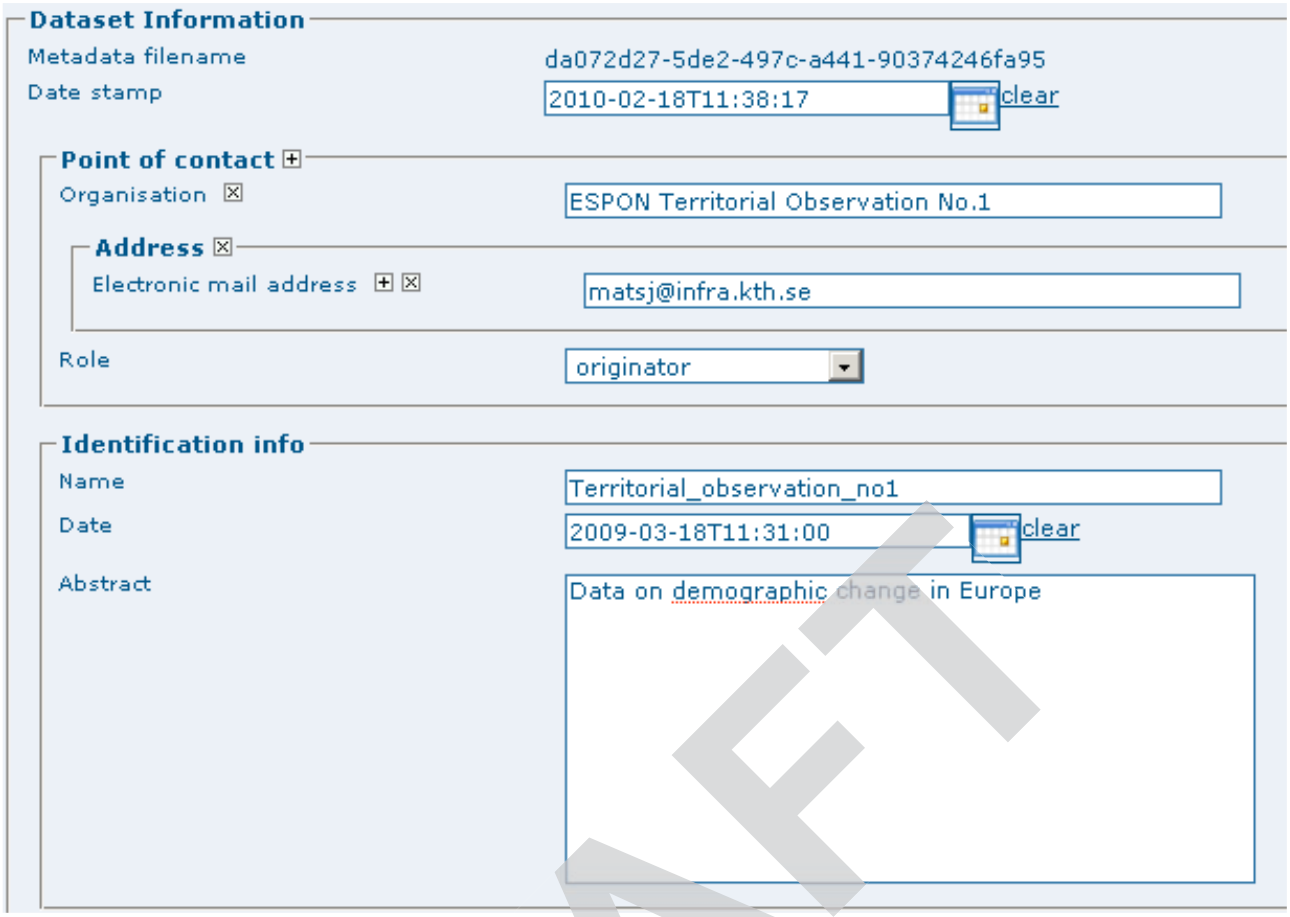


Figure 13: Editing the dataset

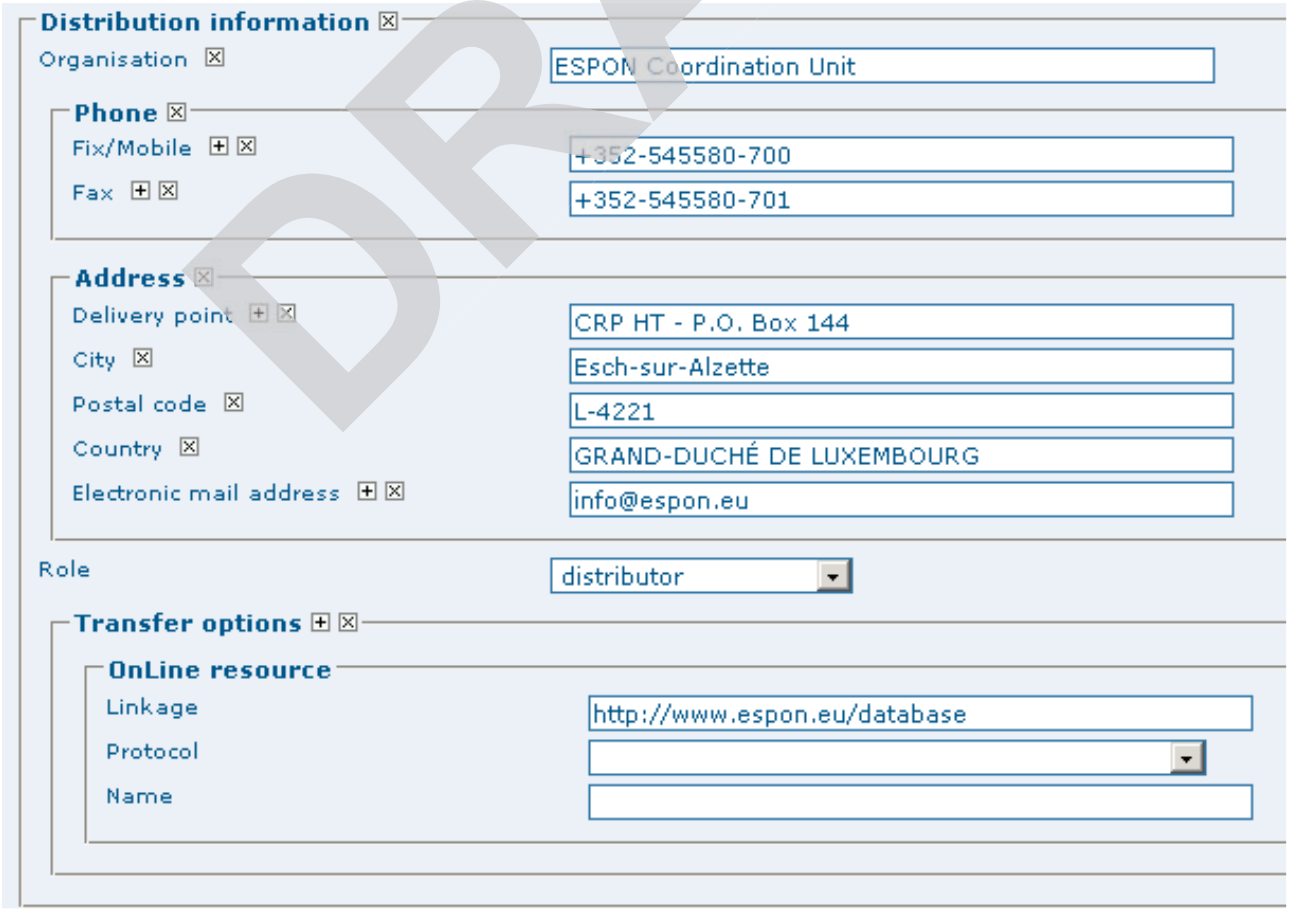


Figure 14: Editing the distributor information for the dataset: filled by default

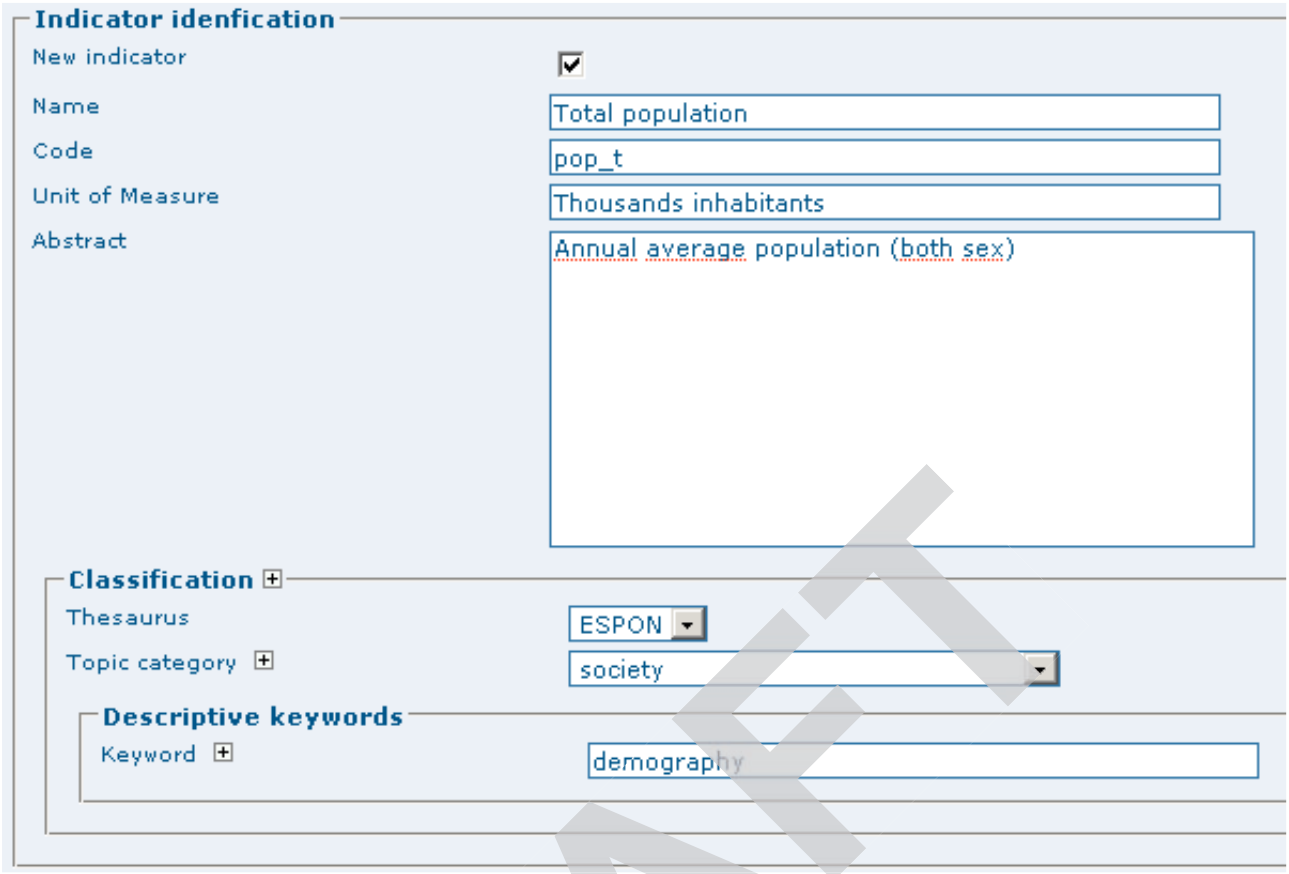


Figure 15: Editing the indicator information

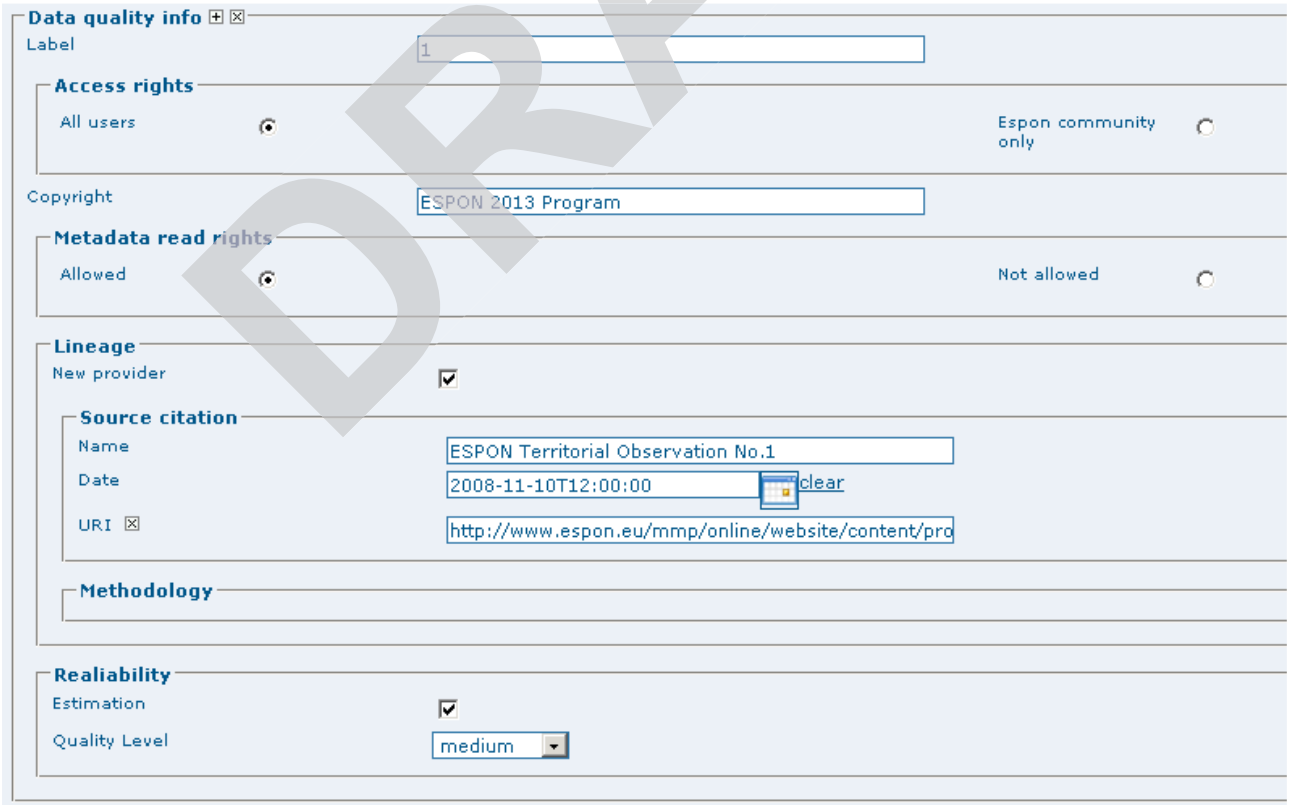


Figure 16: Editing the quality information of the value level

ANNEXES

ANNEXE 1 – Schema of the extension of ISO 19115 : esponMD

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:gco="http://www.isotc211.org/2005/gco" xmlns:gmd="http://www.isotc211.org/2005/gmd"
xmlns:esponMD="http://www.espon.eu/esponMD" targetNamespace="http://www.espon.eu/esponMD" elementFormDefault="qualified"
version="0.1">
	<!-- ================================== Annotation ================================== -->
	<xs:annotation>
		<xs:documentation>This file was generated by Christine Plumejeaud, 29 October 2009, using eclipse Europa 3.1.1
		This make an extent of the ISO 19115 standard for the ESPON 2013 DB project,
		allowing for the definition of new elements (all are prefixed with espon ),
		The template ESPON is provided to show how to use this extension. A word document explains this template.
		The principle is as follows: we use "series" element for our DATASET, and dataset element for our INDICATOR
			- at DATASET level, use the gmd:MD_Metadata element to describe. But replace inside the element
gmd:MD_DataIdentification by esponMD:EsponDatasetIdentification
			- at INDICATOR level, use the esponMD:EsponIndicatorIdentification instead of gmd:MD_DataIdentification
		The template and the XSD schema have been tested using the Xerces validator.
		It was then edited with XMLSpy v2009 sp1 (http://www.altova.com) by maria ramos (EMBRACE)
		</xs:documentation>
	</xs:annotation>
	<!-- ================================== Imports ================================== -->
	<xs:import namespace="http://www.isotc211.org/2005/gco" schemaLocation="gco/gco.xsd "/>
	<xs:import namespace="http://www.isotc211.org/2005/gmd" schemaLocation="./gmd/gmd.xsd "/>
	<!-- ============================================================================== -->
	<xs:simpleType name="EsponThesaurusCode_Type">
		<xs:annotation>
			<xs:documentation>List the available thesauri grouping High-level geospatial data thematic classification to assist in the
grouping and search of available geospatial datasets</xs:documentation>
		</xs:annotation>
		<xs:restriction base="xs:string">
			<xs:enumeration value="ESPON"/>
		</xs:restriction>
	</xs:simpleType>
	<!-- ..... -->
	<xs:element name="EsponThesaurusCode" type="esponMD:EsponThesaurusCode_Type" substitutionGroup="gco:CharacterString"/>
	<!-- ..... -->
	<xs:complexType name="EsponThesaurusCode_PropertyType">
		<xs:sequence minOccurs="0">
			<xs:element ref="esponMD:EsponThesaurusCode"/>
		</xs:sequence>
		<xs:attribute ref="gco:nilReason"/>
	</xs:complexType>
	<!-- ============================================================================== -->
	<xs:complexType name="EsponClassification_Type">
		<xs:annotation>
			<xs:documentation>This group the thesaurus, thems and keywords inside an elements,
			so that we know from which thesaurus themes and keyWords are extracted. </xs:documentation>
		</xs:annotation>
		<xs:complexContent>
			<xs:extension base="gco:AbstractObject_Type">
				<xs:sequence>
					<xs:element name="thesaurus" type="esponMD:EsponThesaurusCode_PropertyType"/>
					<xs:element name="topicCategory" type="gmd:MD_TopicCategoryCode_PropertyType"
maxOccurs="unbounded"/>
					<xs:element name="descriptiveKeywords" type="gmd:MD_Keywords_PropertyType" minOccurs="1"
maxOccurs="1"/>
				</xs:sequence>
			</xs:extension>
		</xs:complexContent>
	</xs:complexType>
	<!-- ..... -->
	<xs:element name="EsponClassification" type="esponMD:EsponClassification_Type"/>
	<!-- ..... -->
	<xs:complexType name="EsponClassification_PropertyType">
		<xs:sequence minOccurs="0">
			<xs:element ref="esponMD:EsponClassification"/>
		</xs:sequence>
		<xs:attributeGroup ref="gco:ObjectReference"/>
		<xs:attribute ref="gco:nilReason"/>
	</xs:complexType>
	<!-- ============================================================================== -->
	<xs:complexType name="EsponDataConstraints_Type">
		<xs:annotation>
			<xs:documentation>Restrictions and legal prerequisites for accessing and using the dataset.</xs:documentation>
		</xs:annotation>
		<xs:complexContent>
```

```
                        <xs:extension base="gmd:MD_Constraints_Type">
                                <xs:sequence>
                                        <xs:element name="copyright" type="gco:CharacterString_PropertyType"/>
                                        <xs:element name="indicatorfreeAccess" type="esponMD:freeAccess_PropertyType"/>
                                        <xs:element name="accessConstraints" type="gmd:MD_RestrictionCode_PropertyType"
minOccurs="0" maxOccurs="unbounded"/>
                                        <xs:element name="freeUse" type="gco:Boolean_PropertyType" minOccurs="0"/>
                                        <xs:element name="useConstraints" type="gmd:MD_RestrictionCode_PropertyType" minOccurs="0"
maxOccurs="unbounded"/>
                                        <xs:element name="otherConstraints" type="gco:CharacterString_PropertyType" minOccurs="0"
maxOccurs="unbounded"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:complexType name="freeAccess_PropertyType">
                <xs:sequence>
                        <xs:element name="Allusers" type="gco:Boolean_PropertyType" minOccurs="1" maxOccurs="1"/>
                        <xs:element name="Esponcomunity" type="gco:Boolean_PropertyType" minOccurs="1" maxOccurs="1"/>
                </xs:sequence>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponDataConstraints" type="esponMD:EsponDataConstraints_Type"/>
        <!-- ..... -->
        <xs:complexType name="EsponDataConstraints_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponDataConstraints"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ========================================================================== -->
        <xs:complexType name="MetaDataConstraints_Type">
                <xs:annotation>
                        <xs:documentation>Restrictions and legal prerequisites for accessing and using the metadata.</xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:MD_Constraints_Type">
                                <xs:sequence>
                                        <xs:element name="readrights" type="esponMD:readrights_PropertyType"/>
                                        <xs:element name="accessConstraints" type="gmd:MD_RestrictionCode_PropertyType"
minOccurs="0" maxOccurs="unbounded"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <xs:complexType name="readrights_PropertyType">
                <xs:sequence>
                        <xs:element name="Allowed" type="gco:Boolean_PropertyType"/>
                        <xs:element name="NotAllowed" type="gco:Boolean_PropertyType"/>
                </xs:sequence>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponMetaDataConstraints" type="esponMD:MetaDataConstraints_Type"/>
        <!-- ..... -->
        <xs:complexType name="EsponMetaDataConstraints_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponMetaDataConstraints"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ========================================================================== -->
        <xs:complexType name="EsponConstraints_Type">
                <xs:annotation>
                        <xs:documentation>Restrictions and legal prerequisites for accessing and using the data and the
metadata.</xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gco:AbstractObject_Type">
                                <xs:sequence>
                                        <xs:element name="dataConstraints" type="esponMD:EsponDataConstraints_PropertyType"/>
                                        <xs:element name="metadataConstraints"
type="esponMD:EsponMetaDataConstraints_PropertyType"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponConstraints" type="esponMD:EsponConstraints_Type"/>
        <!-- ..... -->
        <xs:complexType name="EsponConstraints_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponConstraints"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ========================================================================== -->
```

```
	<xs:simpleType name="EsponScopeCode_Type">
		<xs:annotation>
			<xs:documentation>Description of the class of information covered by the information</xs:documentation>
		</xs:annotation>
		<xs:restriction base="xs:string">
			<xs:enumeration value="any"/>
			<xs:enumeration value="all"/>
			<xs:enumeration value="specifiedExtent"/>
		</xs:restriction>
	</xs:simpleType>
	<!-- ..... -->
	<xs:element name="EsponScopeCode" type="esponMD:EsponScopeCode_Type" substitutionGroup="gco:CharacterString"/>
	<!-- ..... -->
	<xs:complexType name="EsponScopeCode_PropertyType">
		<xs:sequence minOccurs="0">
			<xs:element ref="esponMD:EsponScopeCode"/>
		</xs:sequence>
		<xs:attribute ref="gco:nilReason"/>
	</xs:complexType>
	<!-- ============================================================================ -->
	<xs:simpleType name="EsponNomenclatureCode_Type">
		<xs:annotation>
			<xs:documentation>Description of the class of information covered by the information </xs:documentation>
		</xs:annotation>
		<xs:restriction base="xs:string">
			<xs:enumeration value="any"/>
			<xs:enumeration value="WUTS0"/>
			<xs:enumeration value="WUTS1"/>
			<xs:enumeration value="WUTS2"/>
			<xs:enumeration value="WUTS3"/>
			<xs:enumeration value="NUTS0"/>
			<xs:enumeration value="NUTS1"/>
			<xs:enumeration value="NUTS2"/>
			<xs:enumeration value="NUTS2-3"/>
			<xs:enumeration value="NUTS3"/>
			<xs:enumeration value="NUTS4"/>
			<xs:enumeration value="NUTS5"/>
			<xs:enumeration value="LAU1"/>
			<xs:enumeration value="LAU2"/>
		</xs:restriction>
	</xs:simpleType>
	<!-- ..... -->
	<xs:element name="EsponNomenclatureCode" type="esponMD:EsponNomenclatureCode_Type"
substitutionGroup="gco:CharacterString"/>
	<!-- ..... -->
	<xs:complexType name="EsponNomenclatureCode_PropertyType">
		<xs:sequence minOccurs="0">
			<xs:element ref="esponMD:EsponNomenclatureCode"/>
		</xs:sequence>
		<xs:attribute ref="gco:nilReason"/>
	</xs:complexType>
	<!-- ================================== Classes ================================== -->
	<xs:complexType name="EsponMaintenanceInformation_Type">
		<xs:annotation>
			<xs:documentation>Information about the scope and frequency of updating</xs:documentation>
		</xs:annotation>
		<xs:complexContent>
			<xs:extension base="gmd:MD_MaintenanceInformation_Type">
				<xs:sequence>
					<xs:element name="regularUpdates" type="gco:Boolean_PropertyType"/>
					<xs:element name="scopeCode" type="esponMD:EsponScopeCode_PropertyType"/>
					<xs:element name="scopeExtent" type="gmd:EX_SpatialTemporalExtent_Type" minOccurs="0"
maxOccurs="unbounded"/>
				</xs:sequence>
			</xs:extension>
		</xs:complexContent>
	</xs:complexType>
	<!-- ..... -->
	<xs:element name="EsponMaintenanceInformation" type="esponMD:EsponMaintenanceInformation_Type"/>
	<!-- ..... -->
	<xs:complexType name="EsponMaintenanceInformation_PropertyType">
		<xs:sequence minOccurs="0">
			<xs:element ref="esponMD:EsponMaintenanceInformation"/>
		</xs:sequence>
		<xs:attributeGroup ref="gco:ObjectReference"/>
		<xs:attribute ref="gco:nilReason"/>
	</xs:complexType>
	<!-- ============================================================================ -->
	<xs:complexType name="EsponDatasetIdentification_Type">
		<xs:annotation>
			<xs:documentation>This is used to fill the Identification topic of the series (DATASET level)
		- all information are common to a set of the indicators </xs:documentation>
		</xs:annotation>
		<xs:complexContent>
			<xs:extension base="gmd:AbstractMD_Identification_Type">
				<xs:sequence>
					<!--  INHERITED AND MANDATORY ELEMENTS-->
					<!--
					<xs:element name="citation" type="gmd:CI_Citation_PropertyType"/>
					<xs:element name="abstract" type="gco:CharacterString_PropertyType"/>
```

```
                                        -->
                                        <xs:element name="maintenance" type="esponMD:EsponMaintenanceInformation_PropertyType"
minOccurs="0"/>
                                        <!– at IMPORT, maintenance should not be filled –>
                                        <xs:element name="dataQualityInfo" type="esponMD:EsponDataQuality_PropertyType"
minOccurs="0" maxOccurs="unbounded"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!– ..... –>
        <xs:element name="EsponDatasetIdentification" type="esponMD:EsponDatasetIdentification_Type"
substitutionGroup="gmd:AbstractMD_Identification"/>
        <!– ..... –>
        <xs:complexType name="EsponDatasetIdentification_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponDatasetIdentification"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <xs:complexType name="EsponSpatialRepresentation_Type">
                <xs:annotation>
                        <xs:documentation>Information about the spatial objects in the dataset : you can specify a nomenclature, and precise
the levels used. It can also be vector or grid (see the URL
http://www.isotc211.org/2005/resources/codeList/gmxCodelists.xml#MD_SpatialRepresentationTypeCode) </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:AbstractMD_SpatialRepresentation_Type">
                                <xs:sequence>
                                        <xs:element name="spatialRepresentationType"
type="gmd:MD_SpatialRepresentationTypeCode_PropertyType"/>
                                        <xs:element name="spatialResolution" type="gmd:MD_Resolution_PropertyType" minOccurs="0"/>
                                        <xs:element name="spatialNomenclatureName" type="gco:CharacterString_PropertyType"
minOccurs="0" maxOccurs="unbounded"/>
                                        <xs:element name="nomenclatureLevel" type="esponMD:EsponNomenclatureCode_PropertyType"
minOccurs="0" maxOccurs="unbounded"/>
                                        <xs:choice>
                                                <xs:element name="vectorSpatialRepresentation"
type="gmd:MD_VectorSpatialRepresentation_Type" minOccurs="0"/>
                                                <xs:element name="gridSpatialRepresentation"
type="gmd:MD_GridSpatialRepresentation_Type" minOccurs="0"/>
                                        </xs:choice>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!– ..... –>
        <xs:element name="EsponSpatialRepresentation" type="esponMD:EsponSpatialRepresentation_Type"
substitutionGroup="gmd:AbstractMD_SpatialRepresentation"/>
        <!– ..... –>
        <xs:complexType name="EsponSpatialRepresentation_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponSpatialRepresentation"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!– ========================================================================== –>
        <xs:complexType name="EsponIndicatorIdentification_Type">
                <xs:annotation>
                        <xs:documentation>Use this at INDICATOR level instead of gmd:MD_DataIdentification to give the minimum but
necessary informations
                        - contact information is  mandatory //XSL rule
                        - but name, classification, code, etc., are mandatory elements;
                         </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:AbstractMD_Identification_Type">
                                <xs:sequence>
                                        <!–  INHERITED AND MANDATORY ELEMENTS–>
                                        <!–
                                        <xs:element name="citation" type="gmd:CI_Citation_PropertyType"/>
                                        <xs:element name="abstract" type="gco:CharacterString_PropertyType"/>
                                        –>
                                        <xs:element name="newIndicator" type="gco:Boolean_PropertyType" minOccurs="1"
maxOccurs="1"/>
                                        <!– CP 12102010 - true means it is not present in EUROSTAT nomenclature –>
                                        <xs:element name="code" type="gco:CharacterString_PropertyType"/>
                                        <xs:element name="unitOfMeasure" type="gco:CharacterString_PropertyType"/>
                                        <xs:element name="language" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                                        <!–  NOT EDIT : ENGLISH –>
                                        <xs:element name="characterSet" type="gmd:MD_CharacterSetCode_PropertyType" minOccurs="0"
maxOccurs="1"/>
                                        <!–  NOT EDIT : UTF8 –>
                                        <xs:element name="classification" type="esponMD:EsponClassification_PropertyType"
maxOccurs="unbounded"/>
                                        <xs:element name="extent" type="gmd:EX_SpatialTemporalExtent_PropertyType" minOccurs="0"
maxOccurs="unbounded"/>
                                        <!– set extent.minOccurs to 0 for IMPORT step –>
```

```
                                </xs:sequence>
                        </xs:extension>
                        <!– gmd:MD_DataIdentification_Type –>
                </xs:complexContent>
        </xs:complexType>
        <!– ..... –>
        <xs:element name="EsponIndicatorIdentification" type="esponMD:EsponIndicatorIdentification_Type"
substitutionGroup="gmd:AbstractMD_Identification"/>
        <!– gmd:MD_DataIdentification –>
        <!– ..... –>
        <xs:complexType name="EsponIndicatorIdentification_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponIndicatorIdentification"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!– ========================================================================== –>
        <xs:complexType name="EsponScope_Type">
                <xs:annotation>
                        <xs:documentation>Give the scope of the information given for quality :
                        - can be the whole coverage of the indicator (many years, the full study area),
                        - either a small part (one year, just one geographic unit by example)
                        Label is mandatory and unique for one indicator.
                        Informations givens for the smaller extent prevail above the wider extent.
                        </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:DQ_Scope_Type">
                                <xs:sequence>
                                        <!– INHERITED AND MANDATORY ELEMENTS –>
                                        <!–
                                        <xs:element name="level" type="gmd:MD_ScopeCode_PropertyType"/>–>
                                        <!– XSL FILL gmd:level with a foo value such as series or dataset –>
                                        <xs:element name="spatioTemporalExtent" type="gmd:EX_SpatialTemporalExtent_PropertyType"
minOccurs="0"/>
                                        <!– replaced gmd:EX_Extent_PropertyType by gmd:EX_SpatialTemporalExtent_PropertyType, and set
it as optional (for IMPORT step) –>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!– ..... –>
        <xs:element name="EsponScope" type="esponMD:EsponScope_Type" substitutionGroup="gmd:DQ_Scope"/>
        <!– ..... –>
        <xs:complexType name="EsponScope_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponScope"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!– ========================================================================== –>
        <xs:complexType name="Espon_SourceCitation_Type">
                <xs:annotation>
                        <xs:documentation>Standardized sourceCitation reference (extends the gmd:CI_Citation_Type)
                        The new and mandatory element is the date of the data acquisition;
                        and you can precise the dataSetURI (optional) which indicates the location (web link) of the data to download
                        You MUST add the elements of gmd:CI_Citation_Type and date (of extraction)
                        </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:CI_Citation_Type">
                                <xs:sequence>
                                        <!– INHERITED AND MANDATORY ELEMENTS–>
                                        <!–
                                        <xs:element name="title" type="gco:CharacterString_PropertyType"/>
                                        <xs:element name="date" type="gmd:CI_Date_PropertyType" maxOccurs="unbounded"/>
                                        –>
                                        <xs:element name="dataSetURI" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                                        <!– if available –>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!– ..... –>
        <xs:element name="Espon_SourceCitation" type="esponMD:Espon_SourceCitation_Type" substitutionGroup="gmd:CI_Citation"/>
        <!– ..... –>
        <xs:complexType name="Espon_SourceCitation_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:Espon_SourceCitation"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!– ========================================================================== –>
        <xs:complexType name="EsponSource_Type">
                <xs:annotation>
                        <xs:documentation>Espon source gives a provider name (the title of citation), a date of acquisition and an
optional URI
```

```
                </xs:documentation>
            </xs:annotation>
            <xs:complexContent>
                <xs:extension base="gco:AbstractObject_Type">
                    <xs:sequence>
                        <!-- name of the provider is filled with the Espon_SourceCitation.title instead -->
                        <xs:element name="newSource" type="gco:Boolean_PropertyType" minOccurs="1"
maxOccurs="1"/>
                        <xs:element name="sourceCitation" type="esponMD:Espon_SourceCitation_PropertyType"/>
                    </xs:sequence>
                </xs:extension>
            </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponSource" type="esponMD:EsponSource_Type"/>
        <!-- ..... -->
        <xs:complexType name="EsponSource_PropertyType">
            <xs:sequence minOccurs="0">
                <xs:element ref="esponMD:EsponSource"/>
            </xs:sequence>
            <xs:attributeGroup ref="gco:ObjectReference"/>
            <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ============================================================================ -->
        <xs:complexType name="EsponMethodology_Type">
            <xs:annotation>
                <xs:documentation>
                This new element defines the process steps linked to an indicator in a simplified way.
                You can :
                - link one to many files with the metadata file, documenting the process steps
                - express the formula you could have applied on the data to compute them from source data.
                - describe in a text the process steps that lead to this indicator
                - list of components separated by a semi-colon separator
                </xs:documentation>
            </xs:annotation>
            <xs:complexContent>
                <xs:extension base="gco:AbstractObject_Type">
                    <xs:choice maxOccurs="unbounded">
                        <xs:element name="description" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                        <xs:element name="formula" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                        <xs:element name="file" type="gco:CharacterString_PropertyType" minOccurs="0"
maxOccurs="unbounded"/>
                    </xs:choice>
                </xs:extension>
            </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponMethodology" type="esponMD:EsponMethodology_Type"/>
        <!-- ..... -->
        <xs:complexType name="EsponMethodology_PropertyType">
            <xs:sequence minOccurs="0">
                <xs:element ref="esponMD:EsponMethodology"/>
            </xs:sequence>
            <xs:attributeGroup ref="gco:ObjectReference"/>
            <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ============================================================================ -->
        <xs:complexType name="EsponLineage_Type">
            <xs:annotation>
                <xs:documentation>Espon lineage is made of 2 parts :
                - the source,
                - the methodology,
                This element extends the ancient gmd:LI_Lineage_Type but if forces some new mandatory simplified elements
                </xs:documentation>
            </xs:annotation>
            <xs:complexContent>
                <xs:extension base="gmd:LI_Lineage_Type">
                    <xs:sequence>
                        <xs:element name="esponSource" type="esponMD:EsponSource_PropertyType"/>
                        <xs:element name="methodology" type="esponMD:EsponMethodology_PropertyType"/>
                    </xs:sequence>
                </xs:extension>
            </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponLineage" type="esponMD:EsponLineage_Type" substitutionGroup="gmd:LI_Lineage"/>
        <!-- ..... -->
        <xs:complexType name="EsponLineage_PropertyType">
            <xs:sequence minOccurs="0">
                <xs:element ref="esponMD:EsponLineage"/>
            </xs:sequence>
            <xs:attributeGroup ref="gco:ObjectReference"/>
            <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ============================================================================ -->
        <xs:simpleType name="EsponQualityLevelCode_Type">
            <xs:annotation>
                <xs:documentation>Give a human estimate level of the indicator quality : high is the best quality
level</xs:documentation>
            </xs:annotation>
```

```
                <xs:restriction base="xs:string">
                        <xs:enumeration value="high"/>
                        <xs:enumeration value="medium"/>
                        <xs:enumeration value="low"/>
                        <xs:enumeration value="no opinion"/>
                </xs:restriction>
        </xs:simpleType>
        <!-- ..... -->
        <xs:element name="EsponQualityLevelCode" type="esponMD:EsponQualityLevelCode_Type" substitutionGroup="gco:CharacterString"/>
        <!-- ..... -->
        <xs:complexType name="EsponQualityLevelCode_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponQualityLevelCode"/>
                </xs:sequence>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ========================================================================== -->
        <xs:complexType name="EsponReliability_Type">
                <xs:annotation>
                        <xs:documentation>This new element could be assimilated to a report : indicates id data are issues from an official
source
                        (in an document, ESPON will provide a list of official providers (or considered as)),
                         if data have been estimated, and the quality level that the provider estimates </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gco:AbstractObject_Type">
                                <xs:sequence>
                                        <xs:element name="official" type="gco:Boolean_PropertyType" minOccurs="0"/>
                                        <!-- official is not mandatory at IMPORT step : will be computed through a rule -->
                                        <xs:element name="estimation" type="gco:Boolean_PropertyType"/>
                                        <xs:element name="qualityLevel" type="esponMD:EsponQualityLevelCode_PropertyType"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponReliability" type="esponMD:EsponReliability_Type"/>
        <!-- ..... -->
        <xs:complexType name="EsponReliability_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponReliability"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ========================================================================== -->
        <xs:complexType name="EsponDataQuality_Type">
                <xs:annotation>
                        <xs:documentation>
                        Quality is extended with a reliability element and a constraint element,
                        and the scope element is extended to include an SpatioTemporalExtent element
                        Informations givens for the smaller extent prevail above the wider extent.
                        The constraint at value level allow user for a finest control of data dissemination rights.
                        Indeed, constraints are semantically linked with the lineage of the data : if data are extracted from a source
                        that doesn't allow public data dissemination, this fact can be expressed here.
                        </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:DQ_DataQuality_Type">
                                <xs:sequence>
                                        <!-- INHERITED MANDATORY
                                        <xs:element name="scope" type="gmd:DQ_Scope_PropertyType"/>
                                        but esponMD:EsponScope_PropertyType can be used instead of gmd:DQ_Scope_PropertyType
                                        -->
                                        <!--
                                        label points on a set of values inside the dataset
                                        scope is mandatory, but can be typed with EsponScope_Type instead of DQ_Scope_Type
                                        For the scope, only a label should be provided at IMPORT, valued with "series" or "dataset" value
                                        lineage is mandatory: gives the source, the methodology, the temporal lineage, the spatial lineage
(NUTS version, NUTS level)
                                        constraint is mandatory: gives the access rights, copyrights of data, and metadata access
                                        (a metadata access to false would mean that the existence of this value should be hidden to public)
                                        -->

                                        <xs:element name="label" type="gco:CharacterString_PropertyType"/>
                                        <xs:element name="lineage" type="esponMD:EsponLineage_PropertyType"/>
                                        <xs:element name="reliability" type="esponMD:EsponReliability_PropertyType"/>
                                        <xs:element name="constraints" type="esponMD:EsponConstraints_PropertyType"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!-- ..... -->
        <xs:element name="EsponDataQuality" type="esponMD:EsponDataQuality_Type" substitutionGroup="gmd:DQ_DataQuality_Type"/>
        <!-- ..... -->
<xs:complexType name="EsponDataQuality_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponDataQuality"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
```

```
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
        <!-- ================================================================================ -->
        <!-- not used for the moment -->
        <xs:complexType name="EsponIndicatorMetadata_Type">
                <xs:annotation>
                        <xs:documentation>
                        Use this element for INDICATOR level. This forces the quality and new indicatorIdentification elements to be mandatory.
                        But contact and dataStamp are still mandatory
                        User could add as many elements as wished (this gives more flexibility to the extension)
                        The extension keeps all the ancient elements, in order to allow for the copy-paste of metadata set information
                        This would be usefull for example when extracting a set of indicators coming from various datasets
                        </xs:documentation>
                </xs:annotation>
                <xs:complexContent>
                        <xs:extension base="gmd:MD_Metadata_Type">
                                <xs:sequence>
                                        <!-- INHERITED MANDATORY
                                        <xs:element name="contact" type="gmd:CI_ResponsibleParty_PropertyType"
maxOccurs="unbounded"/>
                                        <xs:element name="dateStamp" type="gco:Date_PropertyType"/>
                                        -->
                                        <xs:element name="indicatorIdentification"
type="esponMD:EsponIndicatorIdentification_PropertyType"/>
                                        <!-- quality can be specified either at dataset level, either on each indicator
                                                The quality element will be mandatory at EXPORT step, but can be skipped at IMPORT step, if
the user fill quality elements at Dataset level.
                                         -->
                                        <xs:element name="dataQualityInfo" type="esponMD:EsponDataQuality_PropertyType"
minOccurs="0" maxOccurs="unbounded"/>
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
        <!-- ..................................................................... -->
        <xs:element name="EsponIndicatorMetadata" type="esponMD:EsponIndicatorMetadata_Type"
substitutionGroup="gmd:MD_Metadata"/>
        <!-- ..................................................................... -->
        <xs:complexType name="EsponMetadata_PropertyType">
                <xs:sequence minOccurs="0">
                        <xs:element ref="esponMD:EsponIndicatorMetadata"/>
                </xs:sequence>
                <xs:attributeGroup ref="gco:ObjectReference"/>
                <xs:attribute ref="gco:nilReason"/>
        </xs:complexType>
</xs:schema>
```


References

- *Websites*

Geonetwork: http://geonetwork-opensource.org

Jetty HTTP server: http://jetty.codehaus.org/jetty

Jeeves framework: http://sourceforge.net/projects/jeeves

MCKoi distributed database: http://www.mckoi.com

TECHNICAL REPORT

ESPON

ESPON DATABASE APPLICATION

Towards a web interface for the ESPON 2013 Database

CONTENT

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 below description is based on a typical user session, from the authentication to the download of results

ESPON 2013 DATABASE

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

22 PAGES

LIST OF AUTHORS

Benoit Le Rubrus, LIG Steamer

Bogdan Moisuc, LIG Steamer

Contact

Jerome.Gensel@imag.fr
Marlene.Villanova-Oliver@imag.fr
Bogdan.Moisuc@imag.fr
Chirstine.Plumejeaud@imag.fr
Benoit.Le-Rubrus@imag.fr

tel. + 33 4 76 82 72 25

TABLE OF CONTENT

Forewords

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 below description is based on a typical user session, from the authentication to the download of results.

Note: as this document was written before the availability of the application on the Internet, the displayed URL on following screenshots, http://127.0.0.1:8080/espondb, is obviously not the real URL to access the application. Moreover, displayed results and values were extracted from a test database, hence they must be considered as irrelevant and for illustrating purpose only.

1 Login

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.

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 a registered account.

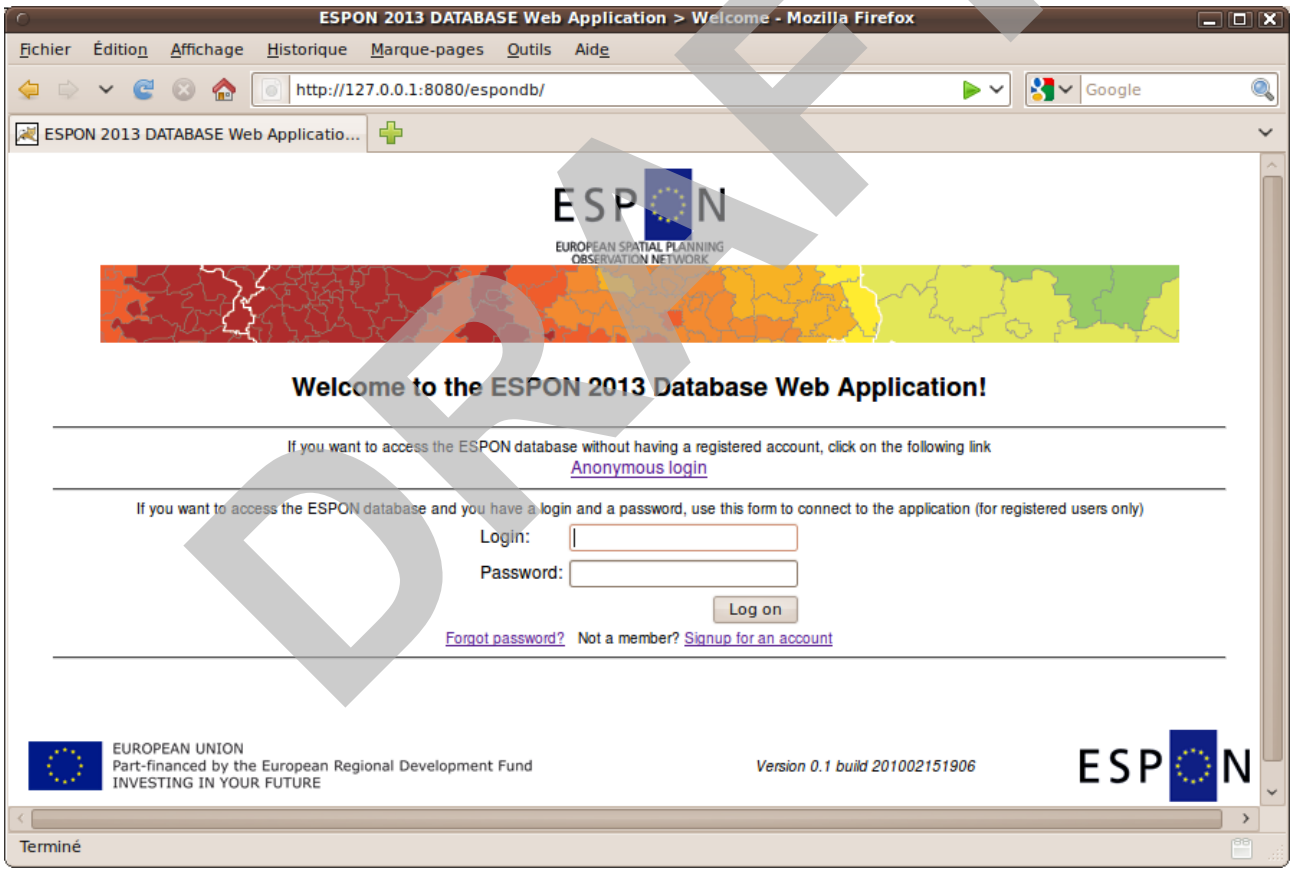


***Figure 1**: Authentification*

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. Both links asking for an account or a forgotten password redirects the user to a simple page that invites him/her to contact the manager by mail for registration. This simple page is shown on figure 2.

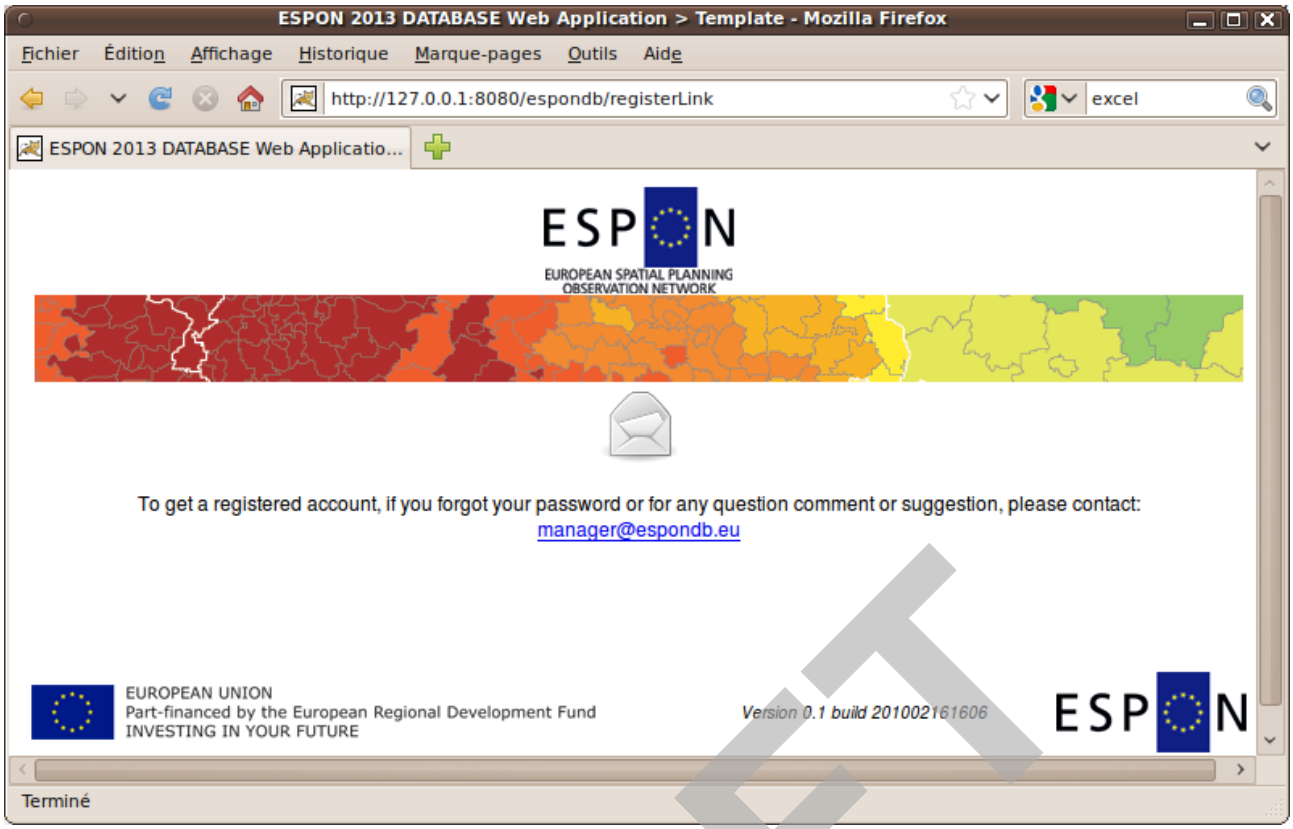


***Figure 2**: Contact page*

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 password;
- 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.

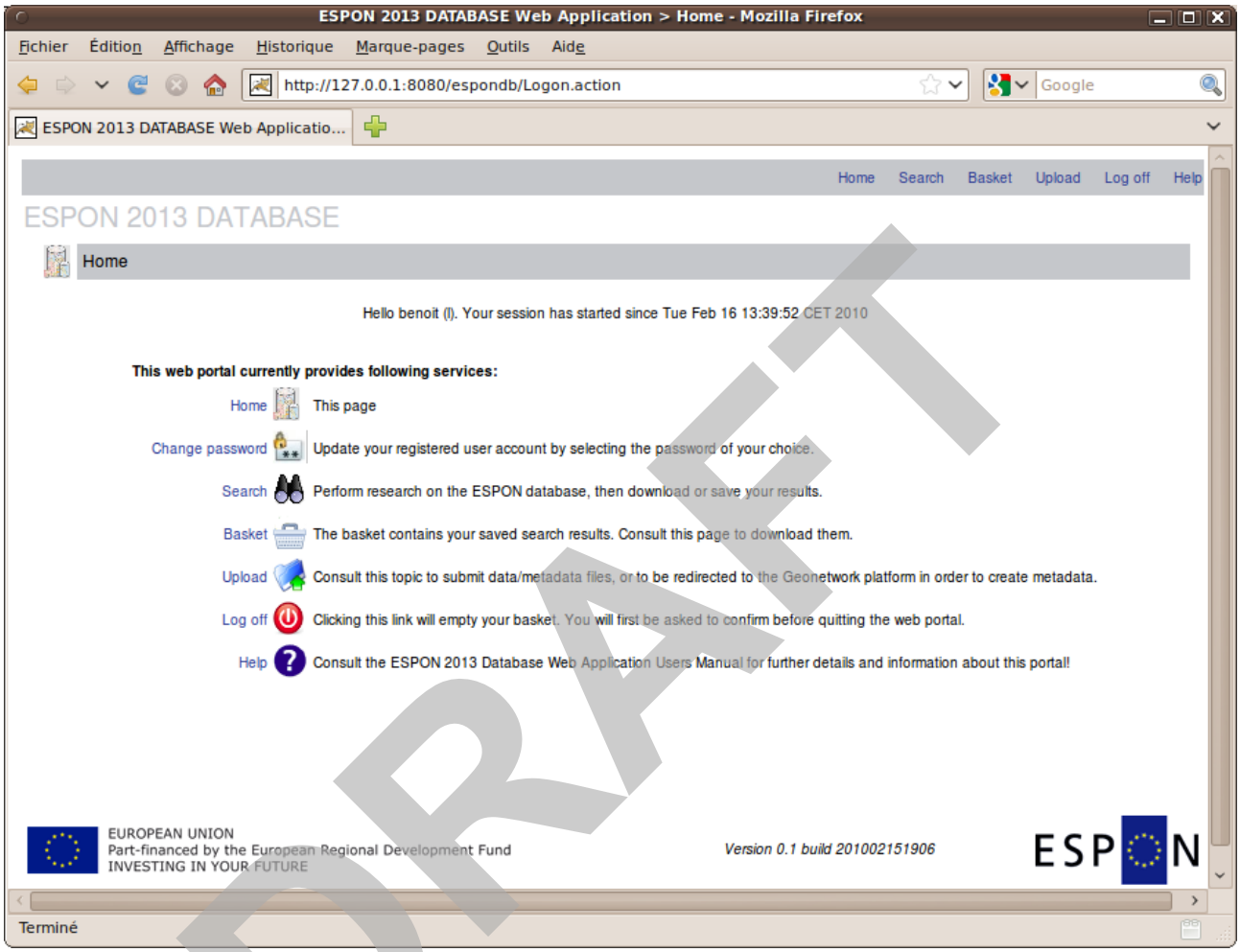


***Figure 5**: Home page, for registered users only*

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

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.

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.

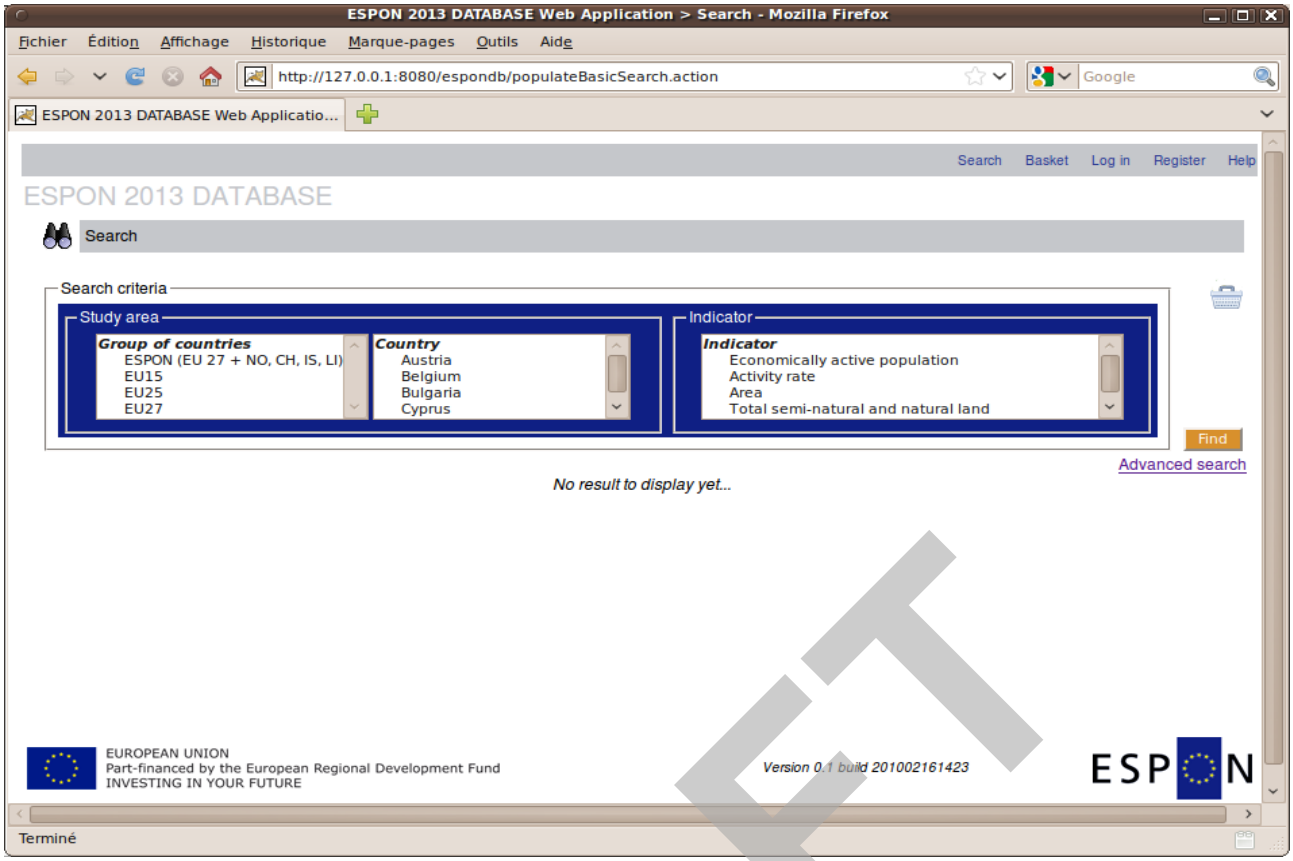


***Figure 6**: Search form page: basic mode*

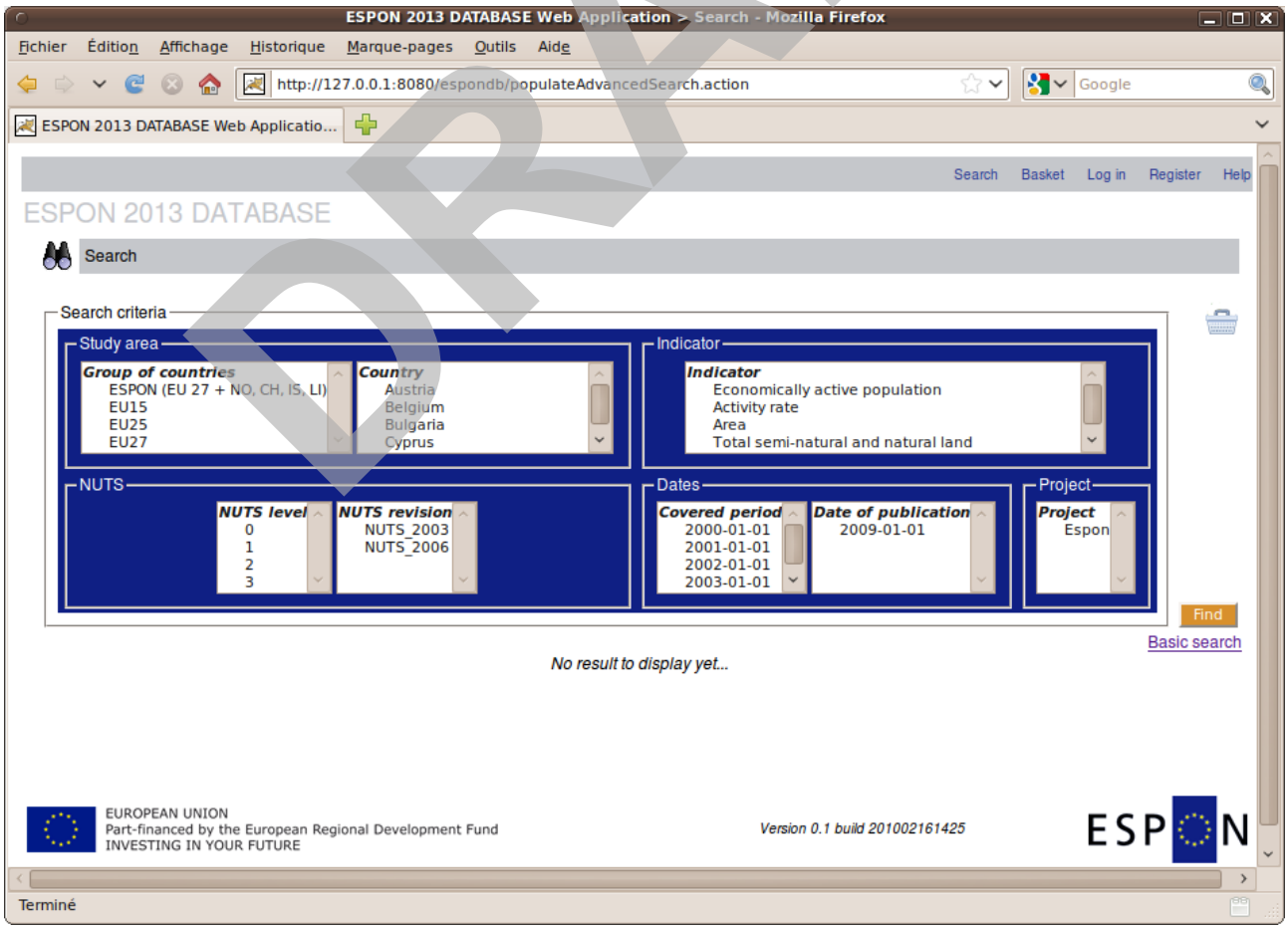


***Figure 7**: Search form: advanced mode*

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.*

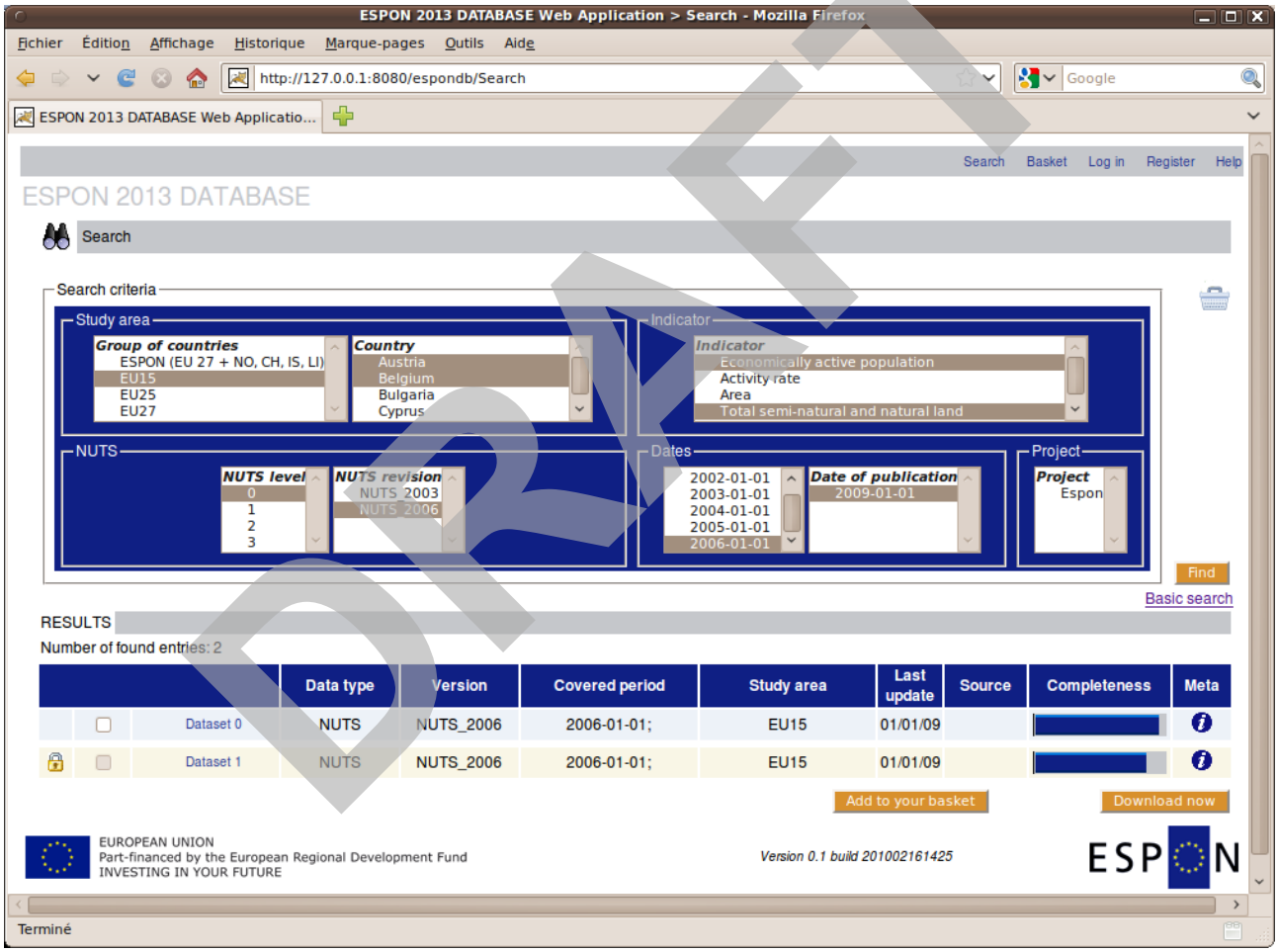


***Figure 8**: Search results*

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:

- 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.
- 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.

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.

- 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.
- Data type column: the current version of the application only manages NUTS, future versions may also include WUTS, GRID, UMZ data types for example.
- Version: as only NUTS data type is currently managed, this field displays the NUTS revision.
- Covered period: this field shows the temporal coverage of the current result.
- Study area: this field shows the spatial coverage of the current result.
- Last update: this field shows when this dataset was last updated.
- Source: this field shows the provider of this result.
- 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.

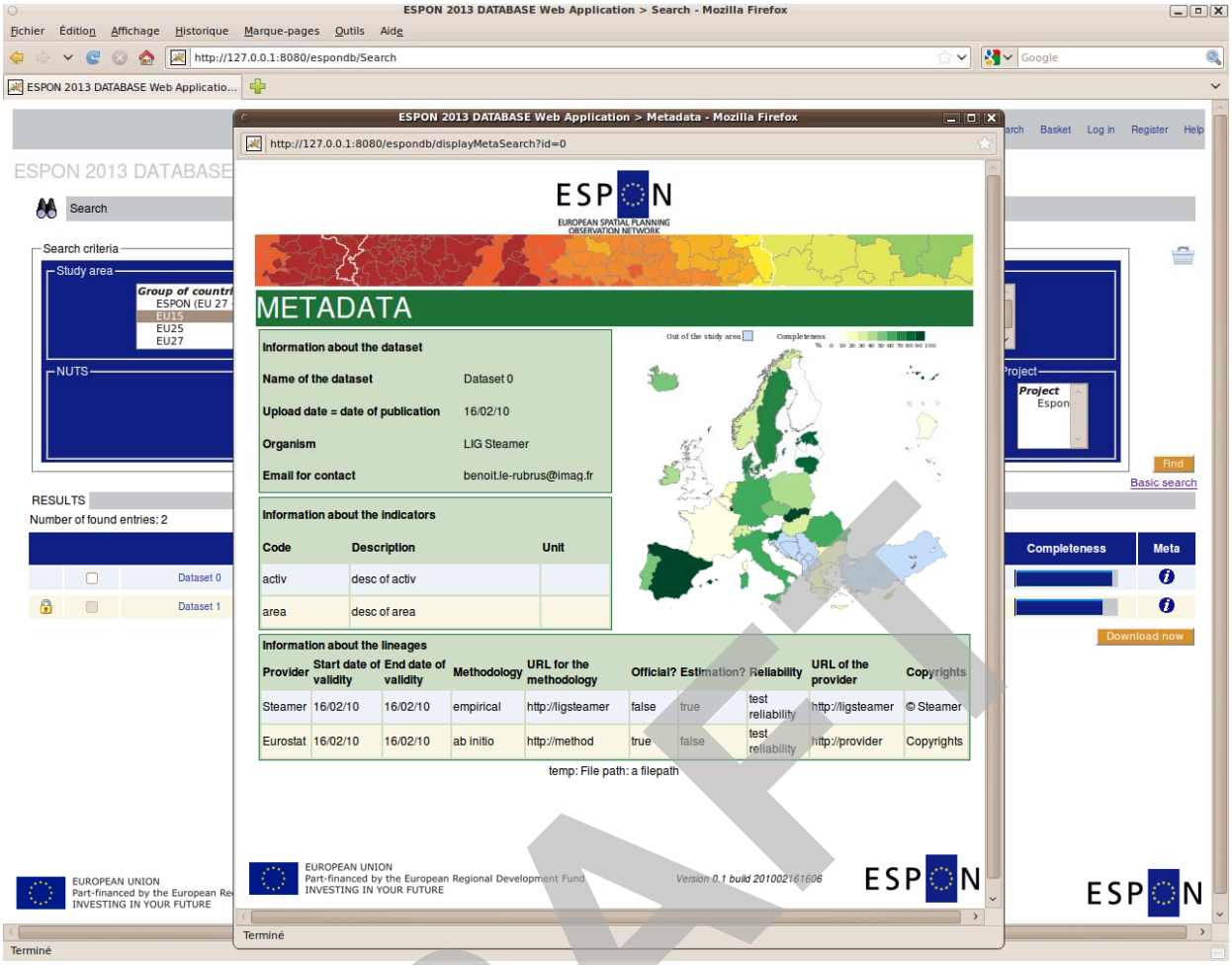


***Figure 9**: Metadata popup for a search result*

This screenshot shows the popup window which is opened over the main window when the user clicks the icon in the "Meta" column cell of a search result. The provided information concerns the dataset, its indicators, its lineage. Furthermore, a map of Europe shows the completeness rate of data for each concerned country. White, yellow then light green colours are used for low rate values (example: France), dark green colours, for high rate values (Spain). Countries which are located out of the study area are filled in blue (Turkey).

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.

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.

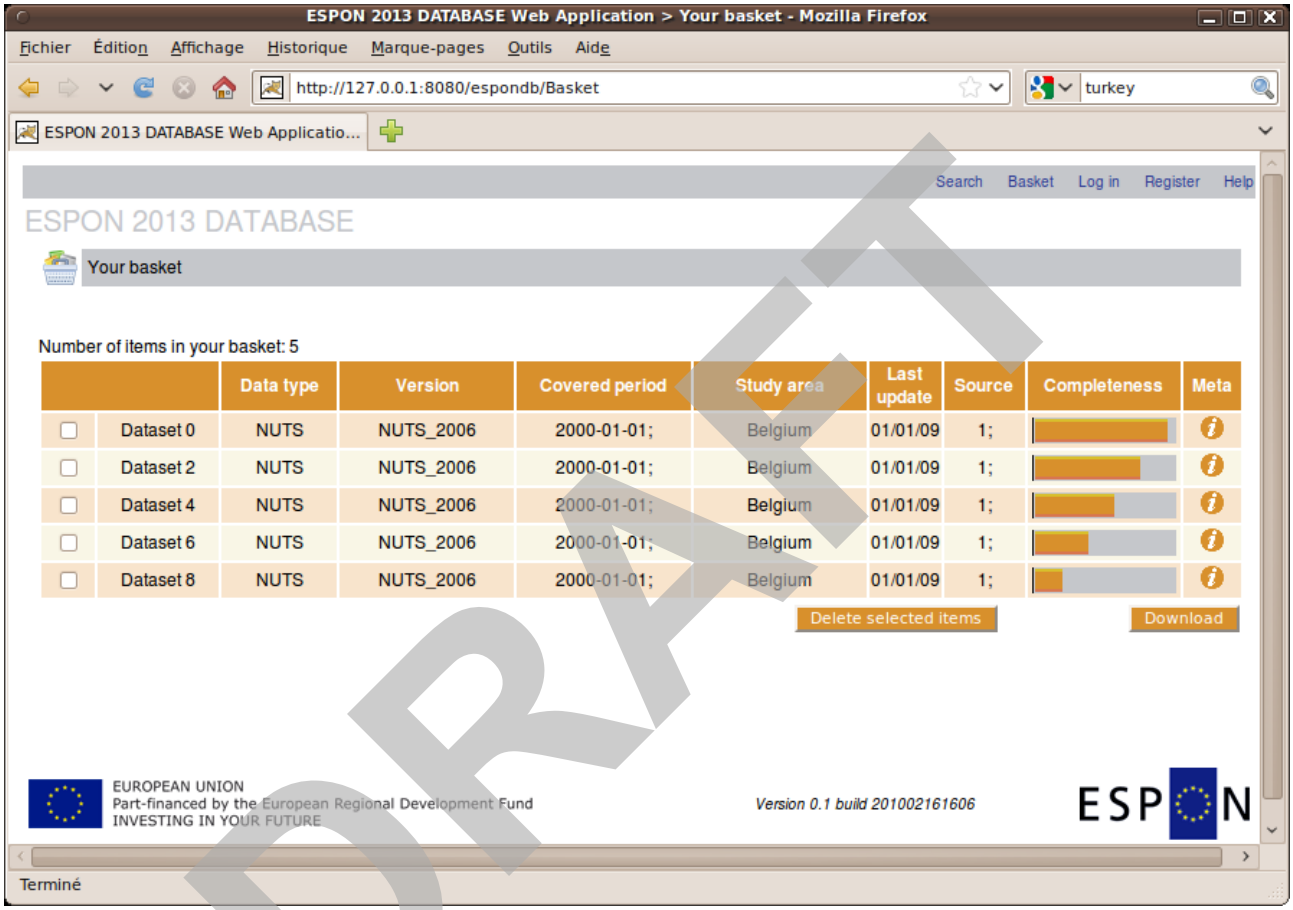


***Figure 10**: Basket page*

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:

- the first sheet is untitled “Dataset”, it provides general information about the current dataset (name, contact, etc.);
- the second sheet is untitled “Indicators”, it provides metadata information about included indicators in this dataset: a description, the unit, the methodology, etc.;
- the third sheet is untitled “Lineage”, it provides metadata information about the lineage of the dataset: validity start, methodology, etc.
- 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.

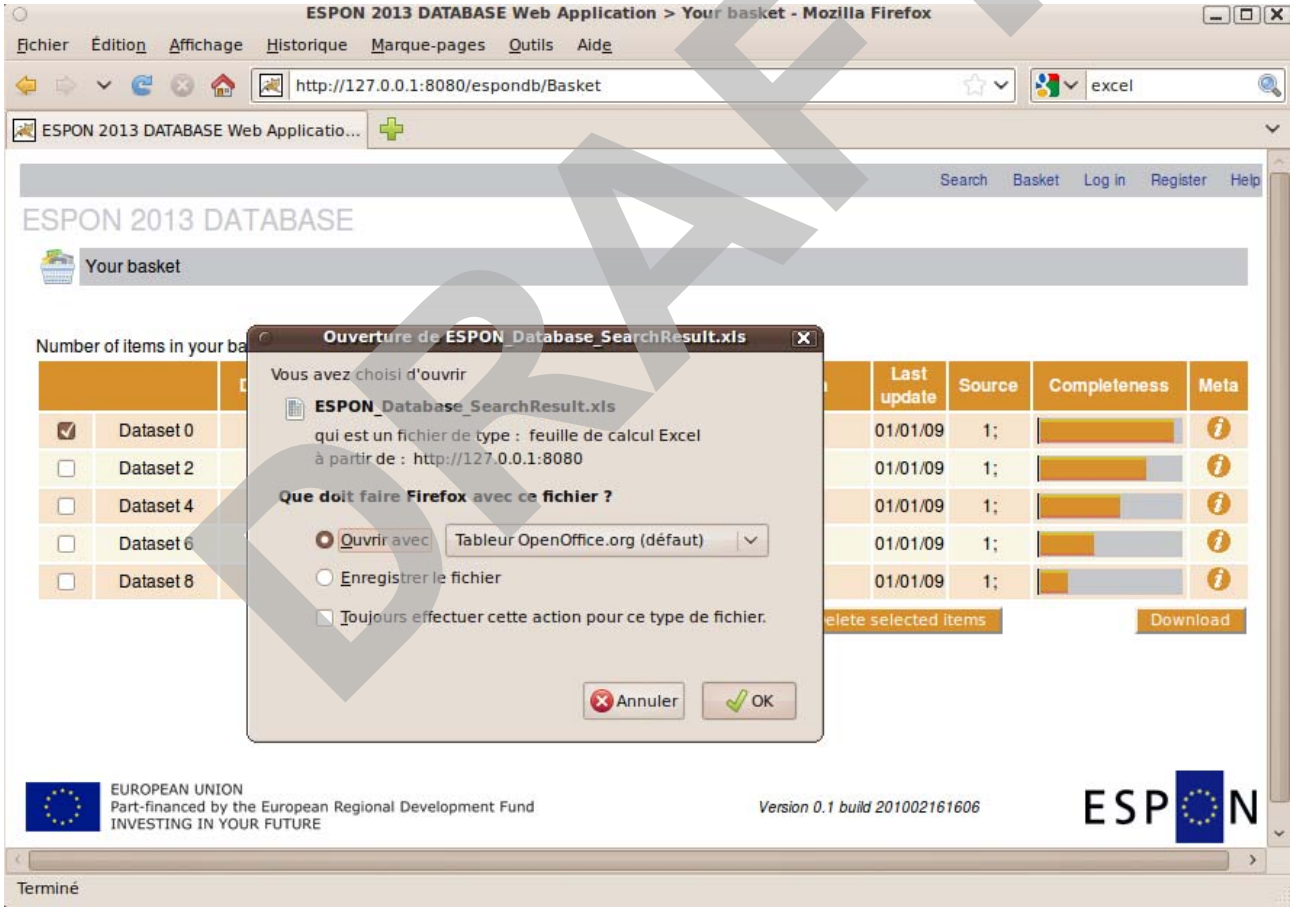


***Figure 11**: Download file browser invitation*

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.

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).

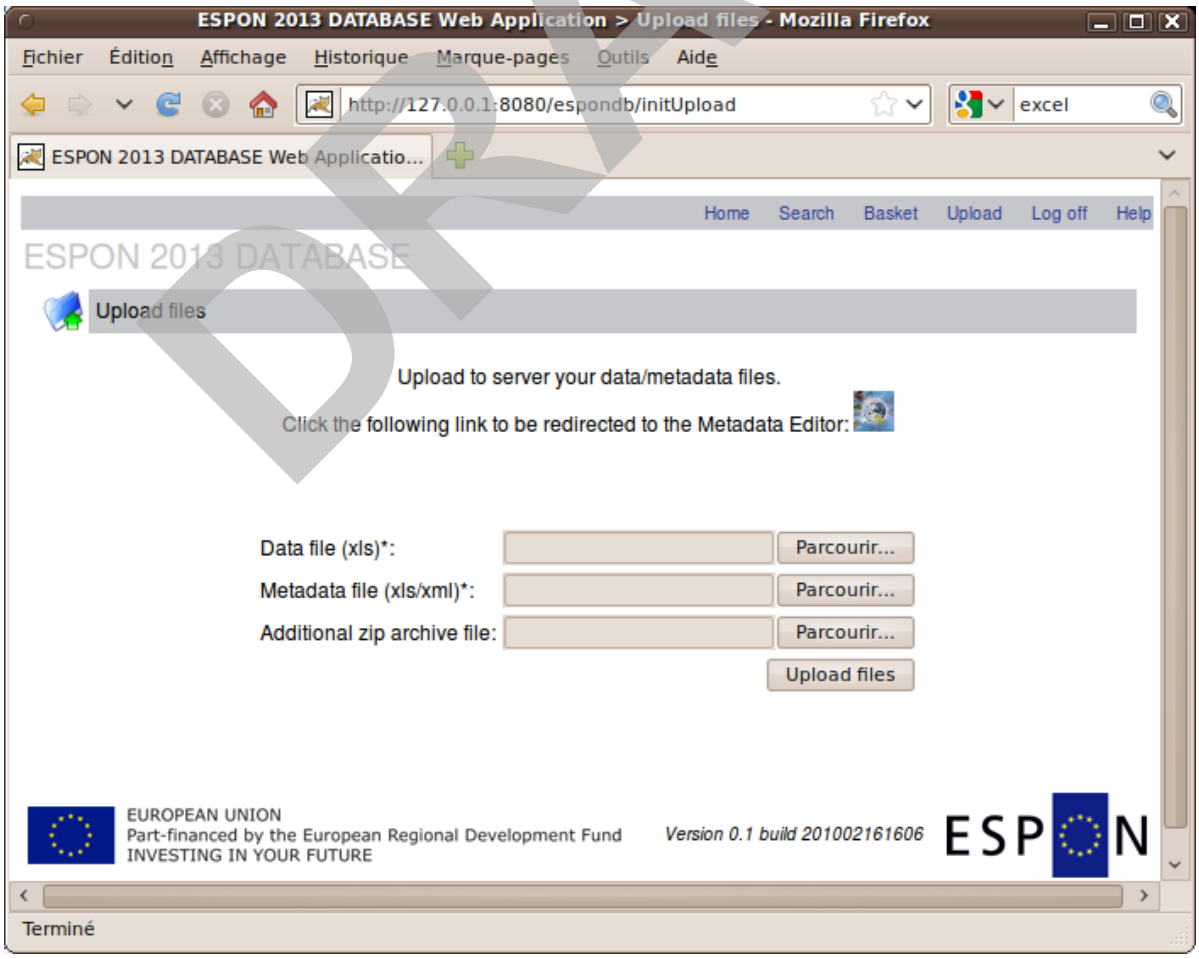


***Figure 12**: Upload page*

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

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.

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.

5 Forbidden action page

The “Forbidden” page was designed to be returned to users who are not allowed to access a page of the application: for example, an anonymous user is not expected to access the Upload page. For security reasons, reserved pages to registered users are not accessible via an hypertext link wherever on the available pages of the anonymous session. Moreover, typing the URL of the Upload page in the browser address bar is not enough to access it, as a double check is performed. In such a case, the user is redirected to the “Forbidden” page shown on figure 13.

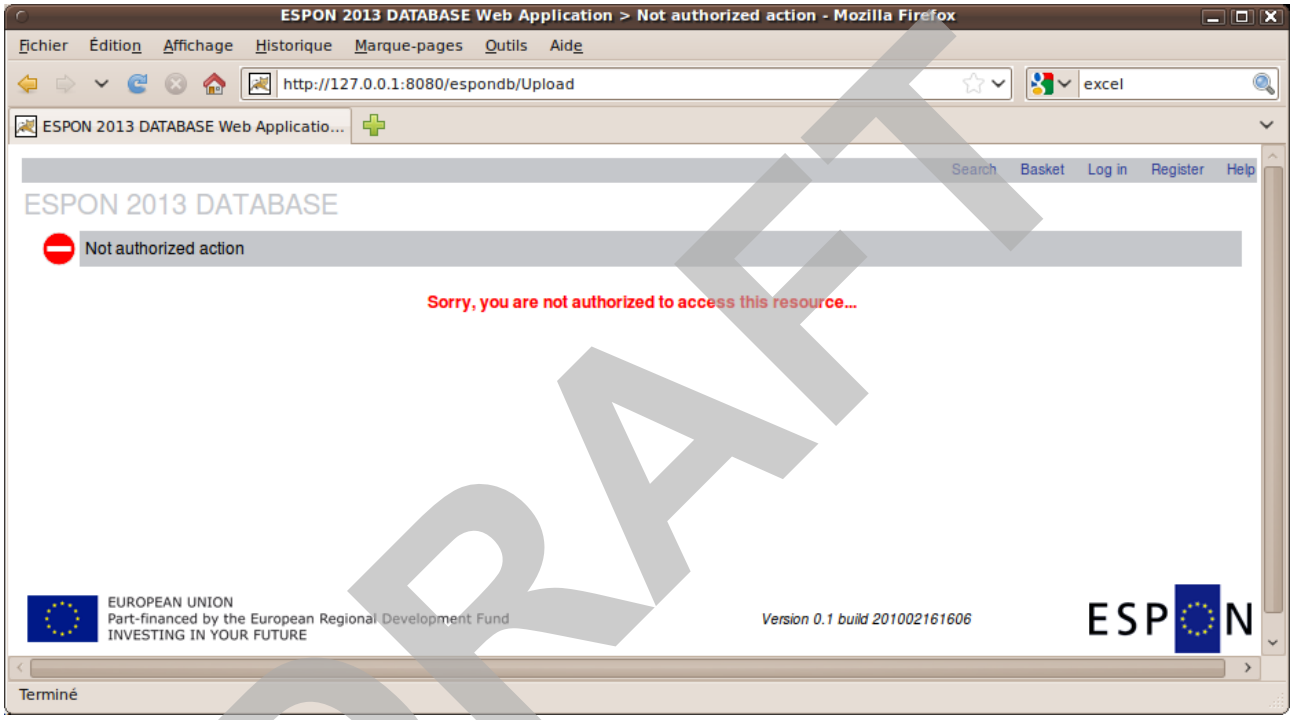


***Figure 13**: Forbidden action*

6 Database models

The database model has been extended to accommodate all the information that can be collected via the data and metadata files defined by the ESPON vector metadata profile. The ESPON 2013 Database Application contains 2 databases; the back office of the application is based on a PostgreSQL database which is much more complex than the ESPON database itself in order to achieve long term data integration. We present hereinafter in detail the schema of the megabase and then we only present the ESPON database based on its differences with the megabase.

Due to the size of the megabase model, we will split its presentation into several pieces. Each piece is equivalent to one dimension of the data. From a conceptual point of view, the megabase can be viewed as a hypercube with 5 dimensions to any statistical data:

1. The spatial dimension identifies precisely the spatial unit described by the indicator value;
2. The thematic dimension identifies precisely what type of statistical indicator is described by a given indicator value;
3. The lineage dimension identifies precisely the source of the indicator value (the database or the organization that published the value), as well as the transformations (corrections, estimations, etc.) that were applied to the value until its actual state;
4. The temporal dimension defines the moment in time or the period described by the indicator value;
5. The dataset dimension describes the publication (name, author, etc.) of which the indicator value is part of.

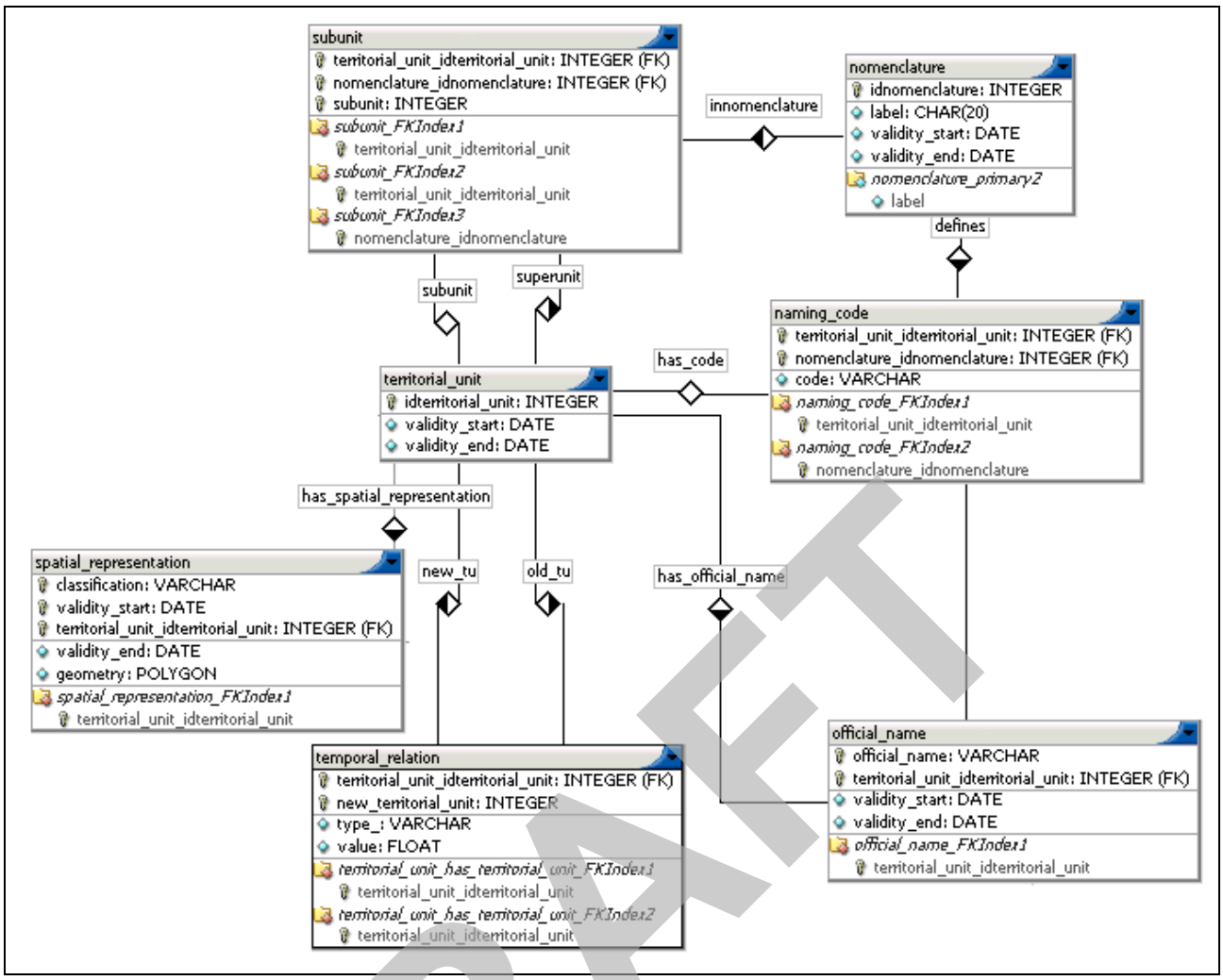


***Figure 14**: Megabase tables describing the spatial dimension*

The spatial description of the data is given via a spatiotemporal ontology. This ontology describes a complete list of the administrative units (for now only NUTS units are stored but the model can accommodate other type of discrete spatial units, e.g. cities, work catchments, etc.), stored in the *territorial_unit* table. Territorial units are described as parts of a nomenclature. A nomenclature (see table *nomenclature* in Figure 14) defines two main properties for territorial units: their code, which is an abstract, standard and unambiguous name based on certain conventions (see table *naming_code* in Figure 14), and their hierarchical inclusion relations (see table *subunit* in Figure 14). Nomenclatures can evolve in time (see attributes *validity_start* and *validity_end* in the table *nomenclature*), and, as such, units codes can also change. Territorial units also have official names that can change in time (see table *official_name* in Figure 14). The temporal relations between territorial units are depicted via the table *temporal_relation*, which contains an attribute (see type attribute in Figure 14) that allows describing the specific type of relation linking two territorial units. Each territorial unit can also have several spatial representations (see table *spatial_representation* in Figure 14), different geometries being fit for different purposes (e.g. more detailed geometries for applying geometrical operations and simplified geometries for on screen display). The spatial representation of a territorial unit can also evolve in time.

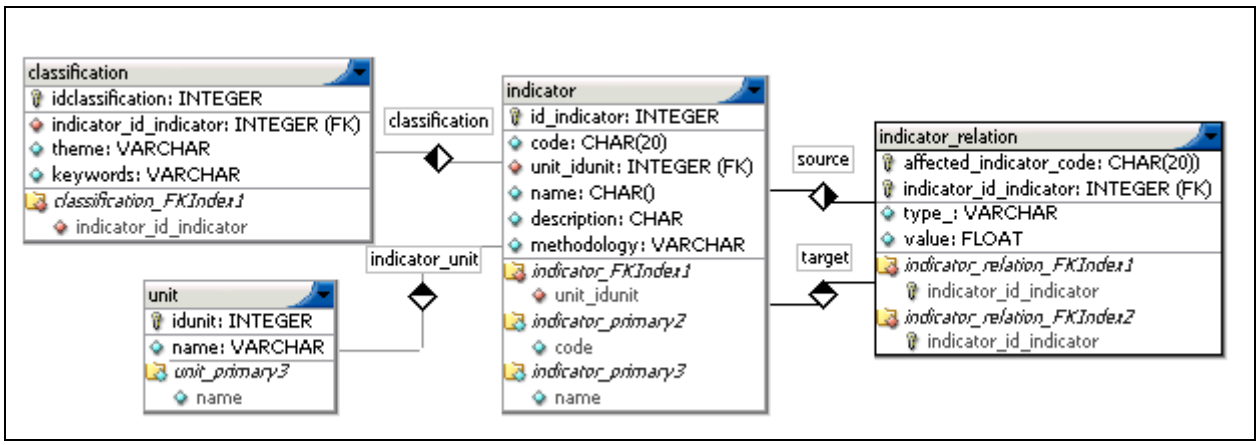


***Figure 15**: Megabase tables describing the thematic dimension*

The thematic description of the indicator values is given via a thematic ontology. However, as the thematic ontology is still a work in progress, for now the table structure reflects the structure of the fields found in the data files. The table *classification* (see Figure 15) simply stores a description of the theme and of the keywords attached to each indicator. When the ESPON DB thematic classification will be ready, the content of this table will be restructured and organized hierarchically in themes, sub themes, groups of indicators (like age pyramid) and individual indicators. The table *units* is required in order to keep track of the different measure units for the same indicator. This covers conversions between different measure systems and also multiples. For instance, the same indicator, total population, can be expressed in thousands of inhabitants at NUTS0 level and in inhabitants at NUTS3 level. The table *indicator_relation* is a generic table (can be used for a specific type of relation by changing the *type* attribute) designed to store vertical and horizontal relations between indicators.

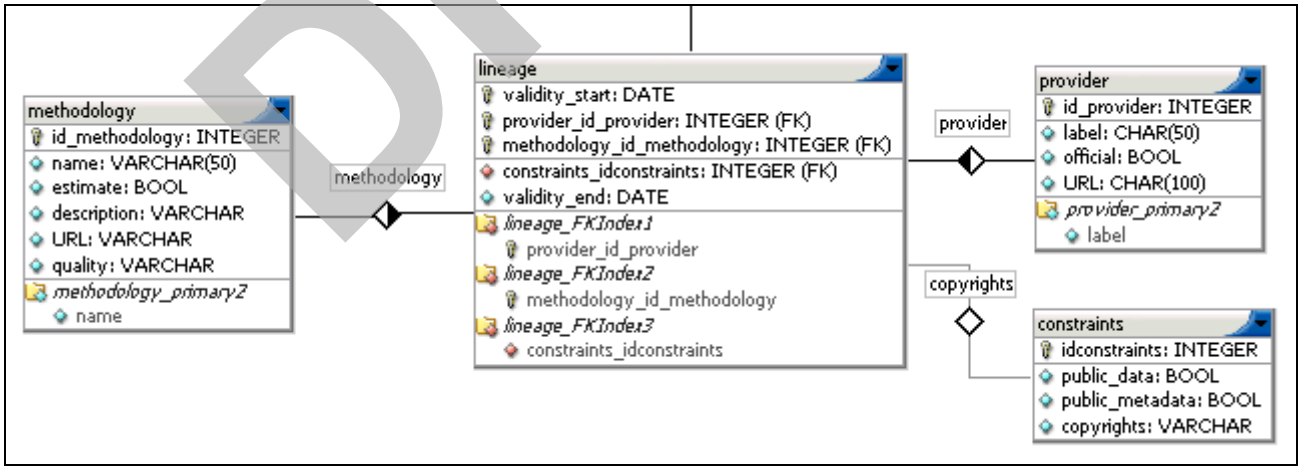


***Figure 16**: Megabase tables describing the lineage dimension*

The description of the lineage of an indicator value is given in the *lineage* table (see Figure 16). The lineage of an indicator value is described by three elements: the original data provider (table *provider*), the methodology for the transformations applied to the data (table *methodology*), and the copyright constraints attached to the indicator value (table *constraints*).

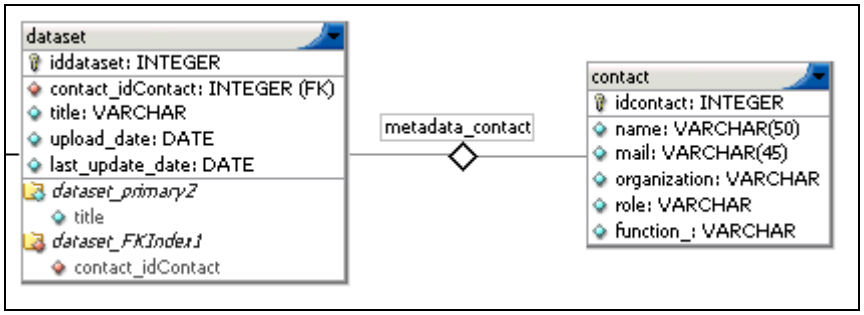


Figure 17: *Megabase tables describing the publication dimension*

A publication is described in the megabase through the table *dataset* (see Figure 17). Another type of information about the dataset maintained in the megabase is about the person who submitted the publication (see table *contact*). Besides keeping a trace of data deliveries from ESPON projects, the coordinates of the contact person from projects are kept with a double purpose, in order to help with the fast filling of the metadata in the Web metadata editor (where, via a connection to the megabase a list of contacts will be displayed) and in order to manage user access to the ESPON database (as most of the data providers will be members of ESPON projects that will also need privileged access to the ESPON database).

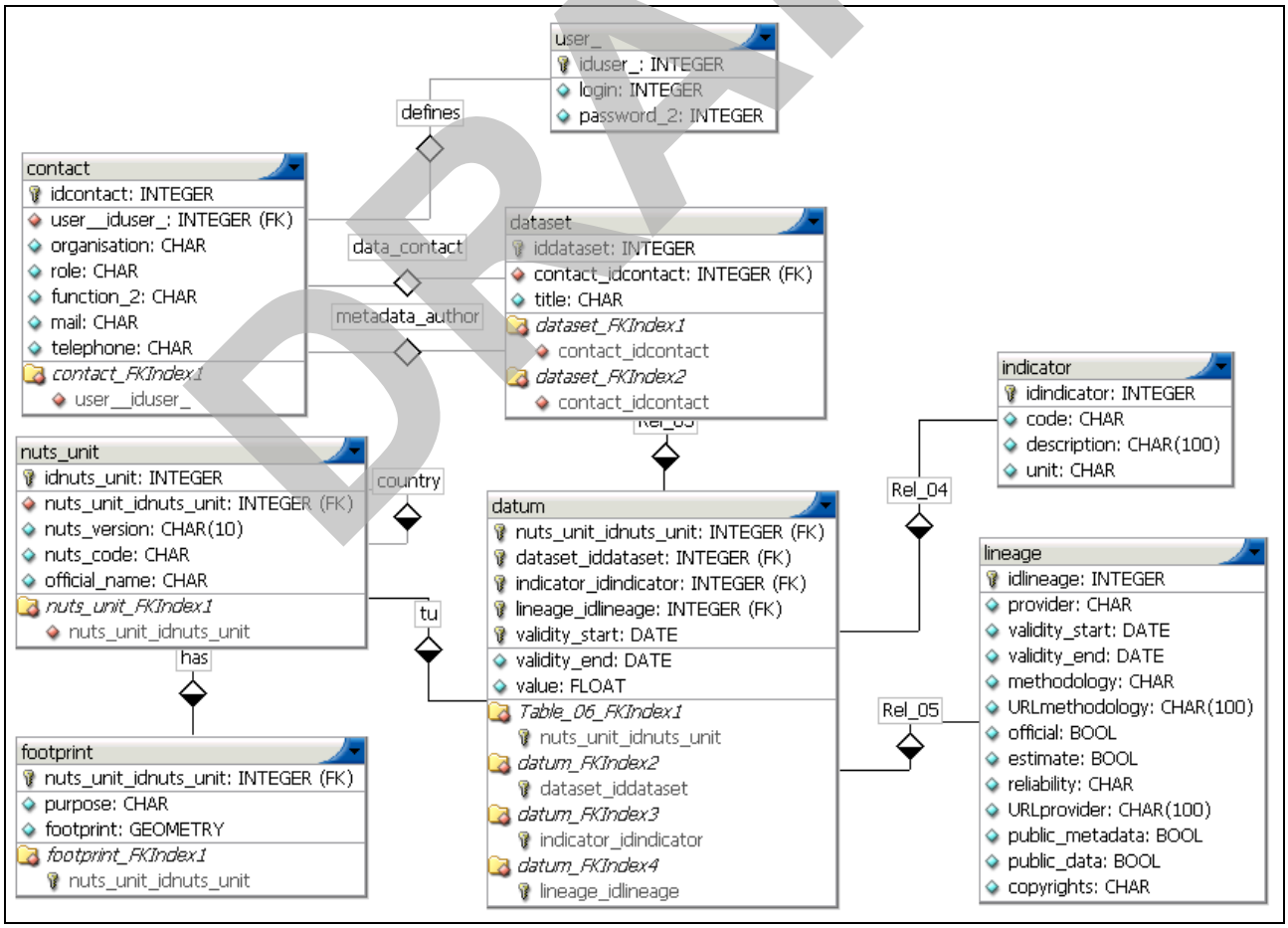


Figure 18*: Schema of the ESPON Database*

In contrast with the megabase, the ESPON Database has a simpler structure (see Figure 18) which makes answering queries faster. Each dimension described in the megabase is compressed in a single table in the ESPON Database. The spatial and thematic ontologies are left aside (no relations between spatial units or indicators are present). The only extra content in the ESPON Database compared to the megabase is the table *user_*, which allows representing information (typically the login and the password) for managing the access of the users ESPON Database to copyrighted data (i.e. what is considered the internal part of the ESPON Database).

TECHNICAL REPORT

ESPON

Towards an approach of time series data issue

From empirical methods to applications

CONTENTS

- The issue of time series is strongly linked to the question of missing values management, either because the territorial unit has changed in the course of time or because data are simply missing.
- Two parallel and complementary ways have been investigated: data collection in historical regional databases and NUTS changes modeling.
- The New Chronos database, the historical database from Eurostat, allows downloading data in the old NUTS versions (namely 1999 and 2003). In fact, this information is not available on the Eurostat website. However a work of reorganization is necessary before using the indicators contented in this database.
- Thanks to data collection in the previous NUTS versions and the formalization of NUTS changes, we aim to propose in a near future solutions in order to build continuous time series from 1995 to present.

ESPON 2013 DATABASE

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

23 PAGES

LIST OF AUTHORS

Ben Rebah Maher, UMS 2414 RIATE

Peeters Didier, IGEAT, Université Libre de Bruxelles

Plumejeaud Christine, LIG

Ysebaert Ronan, UMS 2414 RIATE

Contact

maher.benrebah@ums-riate.fr / tel. (+ 33) 1 57 27 65 35
dpeeter1@ulb.ac.be / tel. (+32 2 650 50 77)
christine.plumejeaud@imag.fr / tel. (+33) 4 76 82 72 11
ronan.ysebaert@ums-riate.fr / tel. (+ 33) 1 57 27 65 35

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 EUROSTAT. In spite of it importance, this process has just been initiated by the previous ESPON DB project (2006)

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. During the ESPON 2006 Program, some projects have experimented the limits raised by the change of territorial unit delineations (text boxes 1, 2 and 3). The problem is indeed well-known. To resolve that situations, some innovative methods have been elaborated: The ESTI framework for estimating missing values is one of them[1]. However, that kind of methodology support does not directly try to understand the nature of territorial units changes, it delivers a panel of methods for overcoming the problem of missing values. Modelling NUTS changes is the next step to ensure the development of long-term databases, which are key-issues for a lot of investigations.

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

Text-boxes 1,2 and 3: *Review of problems raised by the time-series issue in the ESPON 2006 Literature*

[1] ESPON 3.2, Data Navigator 2, Part 1 – Handbook for Data Collection, pp. 50-54

The aim of this Technical Report is to provide a background to time series challenge of ESPON DB project by demonstrating what is behind NUTS changes and what could be done starting from this point.

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.

The first section of the report shows the different sources useful for understanding NUTS changes (National Statistical Institutes, Eurostat); the second one focuses on the methodology used to model NUTS changes. The third part presents some methods to overcome these changes by using various conceptual methods. Finally, the fourth section shows some concrete application which can be deducted from this conceptual story.

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:

ŠБЪЛГАРИЯ

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 Eurostat[2] 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 described from the 1999 version, 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, is a full-time job because of the number of variables and parameters. We are thinking about a definition of a methodology to use the maximum potential of this useful source of information. But the work is still in progress.

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

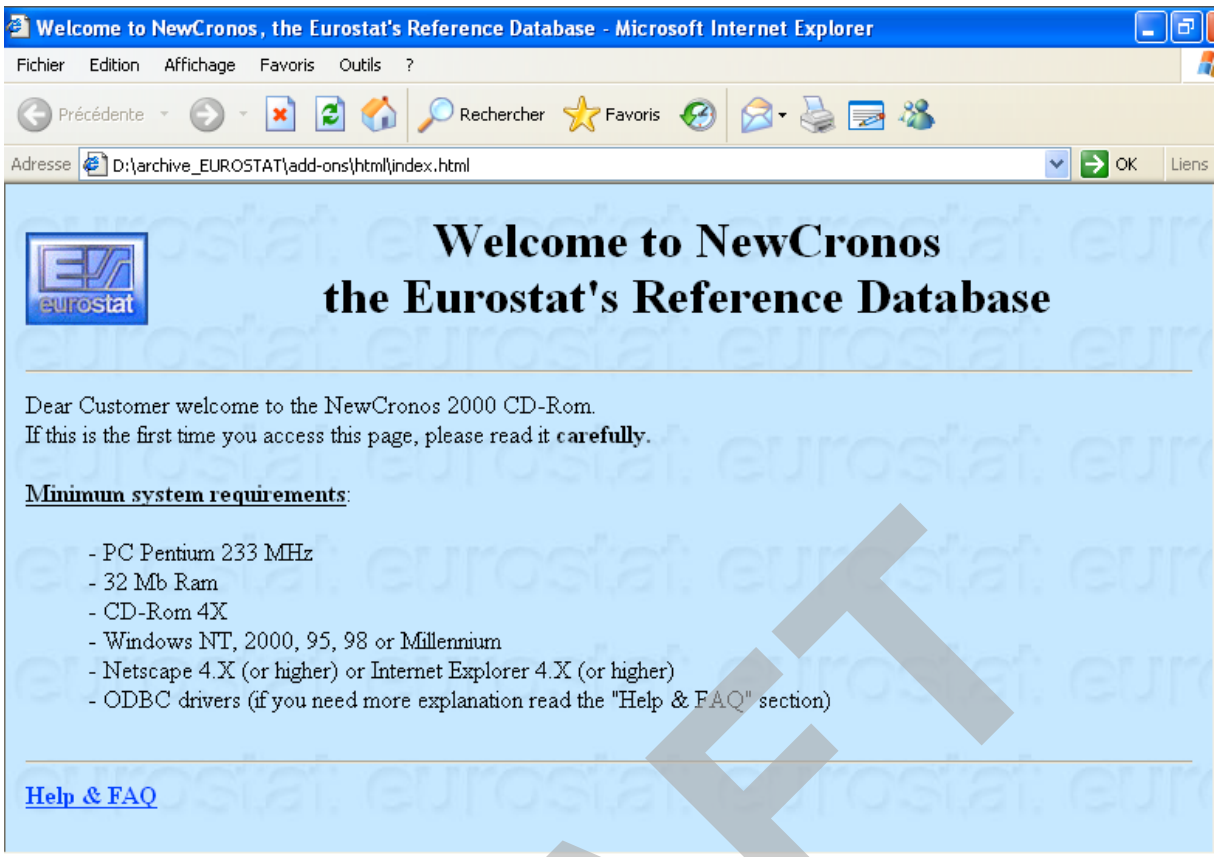


Figure 2: *NewCronos Web interface*

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 (figure 4). Earlier 2003 territorial nuts changes data sets are also available. That kind of data presentation is very useful to re-build a complete time-series.

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 the effect of national administrative boundaries change on the NUTS geometry and hierarchy.

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

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

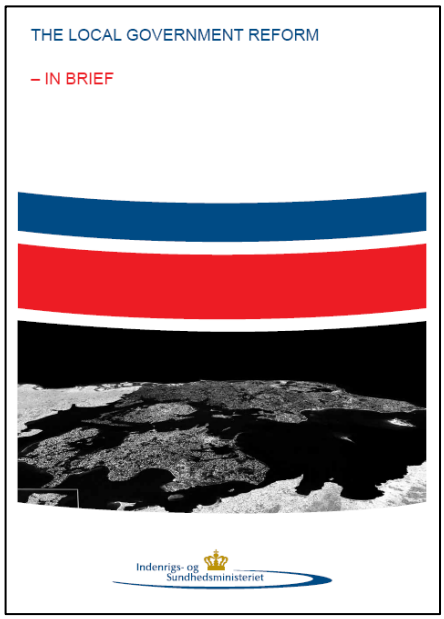
31 dicembre 2008
- Regioni
- Province
- Comuni

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



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 (translated in English) is not usually available for all the countries of the ESPON Area.

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

1.4 Conclusion

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

- NUTS changes vary greatly from country to country due to their different administrative structures;
- The available data and documents referencing the NUTS changes 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, (2001) 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 8).

- 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 (figure 7).

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

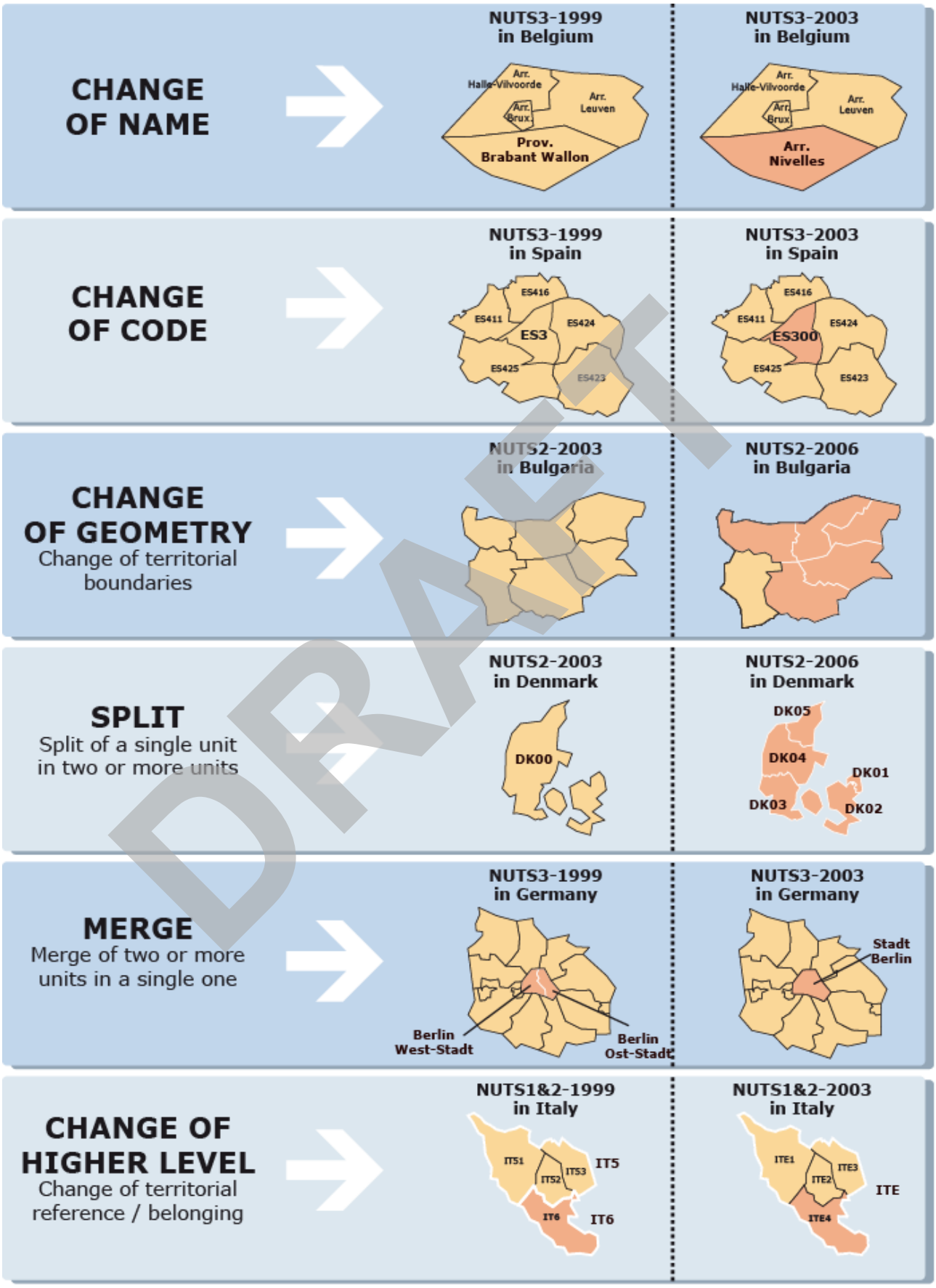


Figure 8: *Examples of elementary NUTS changes*

- 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. The modification of the geometry can occur in two ways: by merging (two territorial units become a single one, like in Berlin between 1999 and 2003) or splitting (a single territorial unit is divided in different units, like in Denmark between 2003 and 2006)

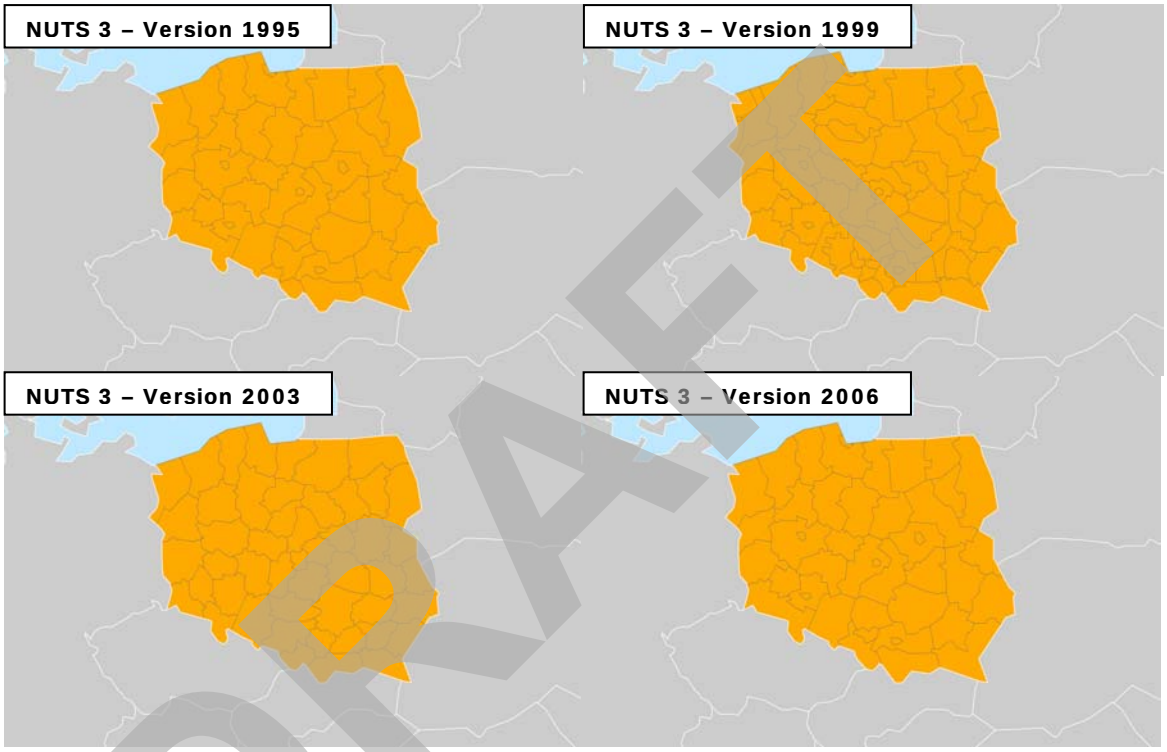


Figure9: *Complex geometry change in Polish nuts 3 units*

Considering the characteristics of NUTS units which are mentioned, limiting the investigation to basic (elementary) changes dos not allow to reconstruct genealogy of NUTS. For this we had considered NUTS structure as a system and we focused on the relationship between changes.

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 (figure 10) 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 emphasises 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.

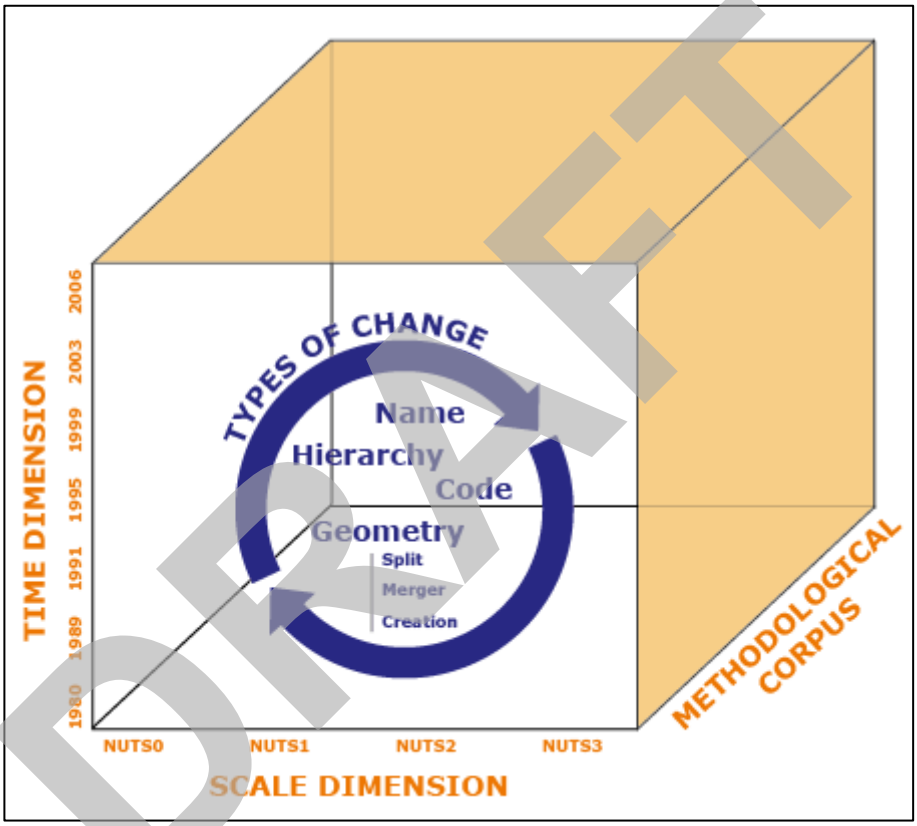


***Figure 10**: Cube structure of NUTS formalisation*

We demonstrate our approach through the analysis of the example of Italian NUTS between 1995 and 2006:
Concerning the temporal dimension, two 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 be also 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 (Langran 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).

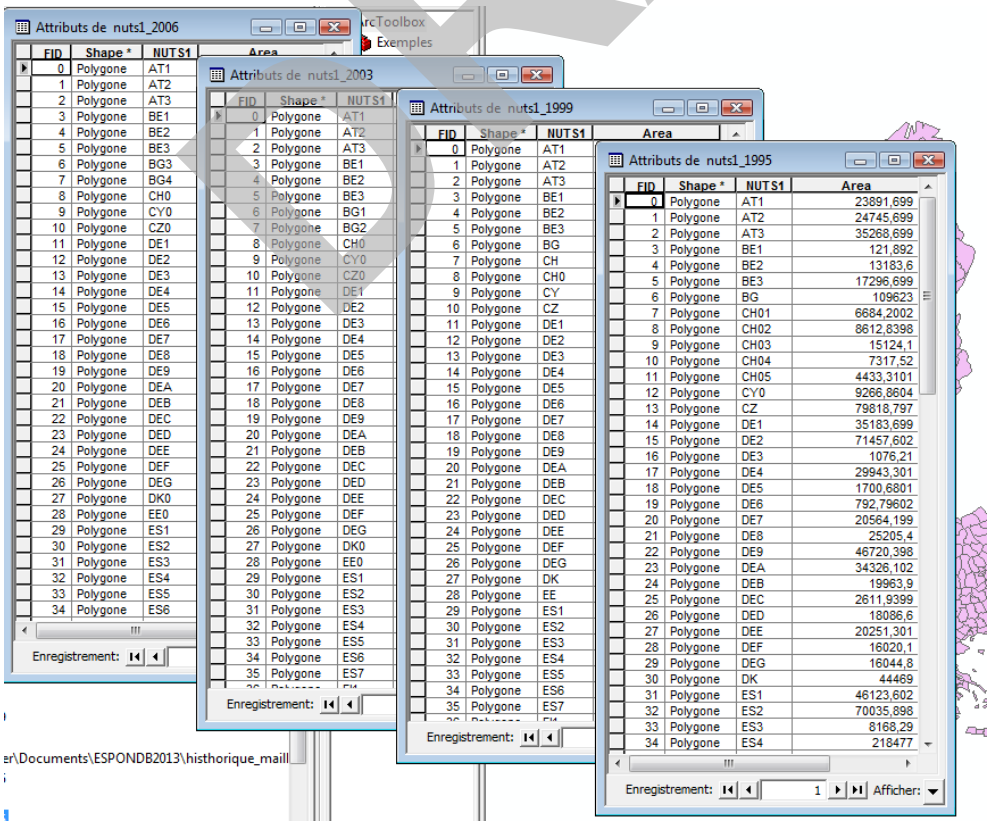


Figure 11: *Snapshot of NUTS versions between 1995 and 2006 (example of Nuts 1 level)*

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 12 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 13 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.

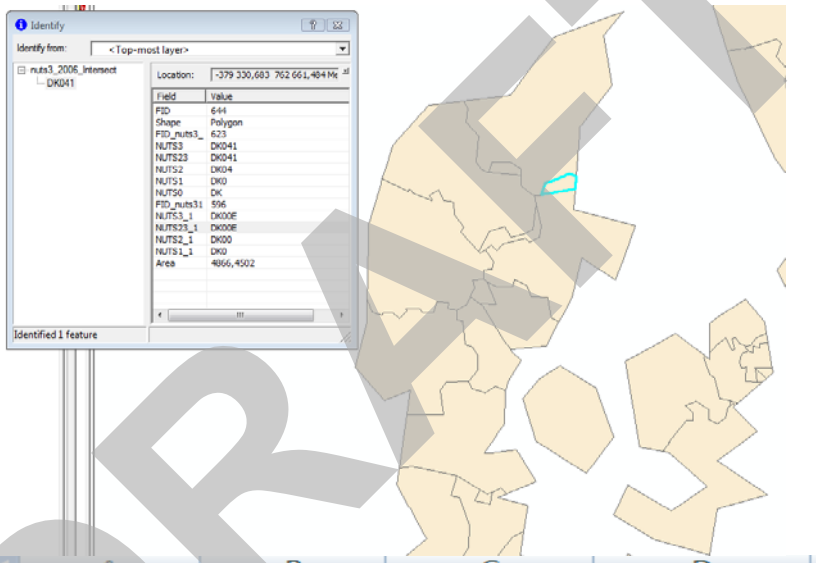
Code 2003	Code 2006	Country	NUTS level 1	NUTS level 2	NUTS level 3	CHECK	CHANGE	LIFE	HIERARCHY	GEOMETRY
	DK01			Hovedstaden		changed	GEOM	N	0	GEOM+
	DK011				Byen København	changed	GEOM	N	0	GEOM+
	DK012				Københavns omegn	changed	GEOM	N	0	GEOM+
	DK013				Nordsjælland	changed	GEOM	N	0	GEOM+
DK007	DK014				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 14 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 figure 15 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*

The proportion of the population transferred will also be tested. This method is very useful for estimating missing data because of territorial changes and could be calculated by passing through the demographic characteristics of the different NUTS versions. However, the best way to proceed stays to have information at very local scale: using LAU2 population to estimate what is the demographic importance of the change. Another source of information could be grid data (1km raster format). In the next step of the project, it is expected to go in that direction.

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 this applications were developed in the following time periods: 2006-2003; 2006-1999; 2006-1995, 2003-1999; 2003-1995 and 1999-1995.

3.3 Towards to cartographic display of NUTS changes: in progress

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. At this stage of work, we are limited to distinguish static units (no change observed) from changed units as shown by the figure 16. Units yellow coloured are static. However, white units have been changed since 1995. It is expected to create a typology of NUTS change in a near future.

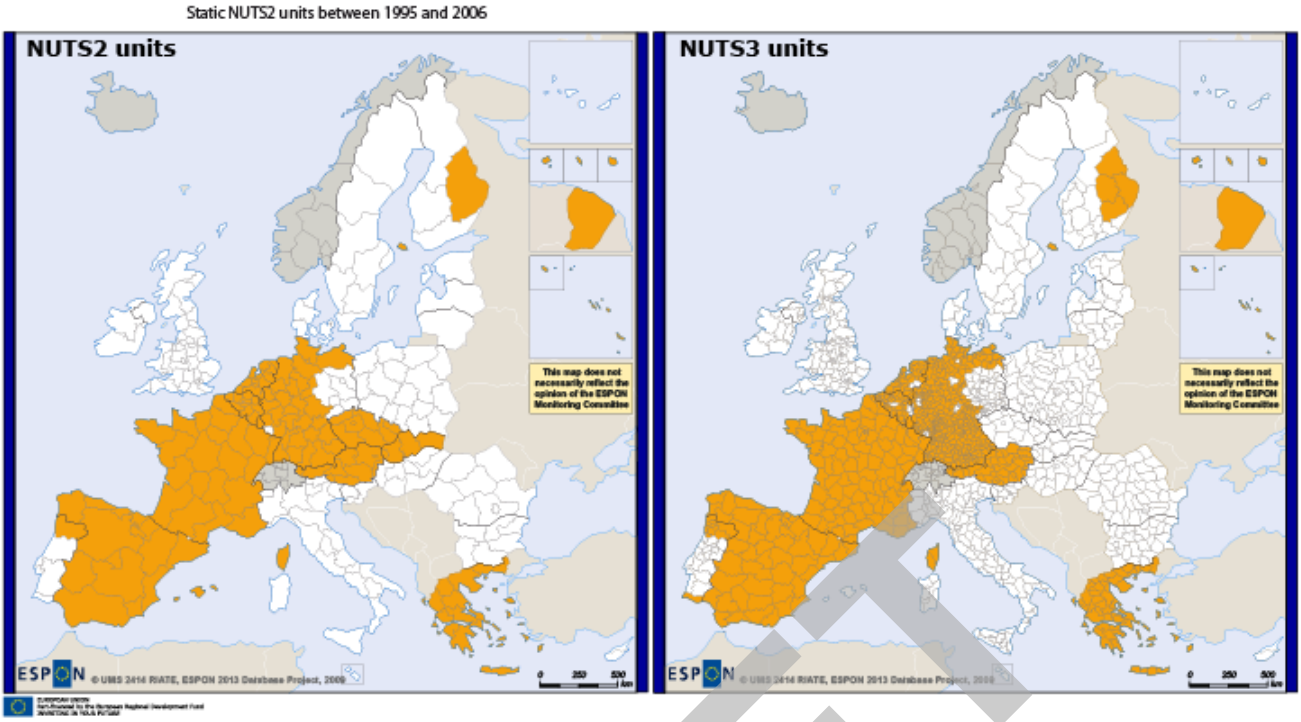


Figure 16: *Static NUTS 2 and Nuts 3 units between 1995 and 2006 (all criteria)*

4 First applications

Based on the results of exploration of NUTS changes, implemented time series data was made from data used in Cohesion Reports (2th, 3th and 4th). The work consisted to collect data and to identify, thanks to the code and the name of the territorial unit, what was the version of NUTS where the information was integrated. In fact, data from Cohesion Reports combines territorial units from various NUTS versions.

The aim of this work is to make possible the comparison of indicators in heterogeneous NUTS version in order to produce comparative thematic maps (figure 17 and figure 18). 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 17**: NUTS version identification of spatial units used by the 4th Cohesion report*

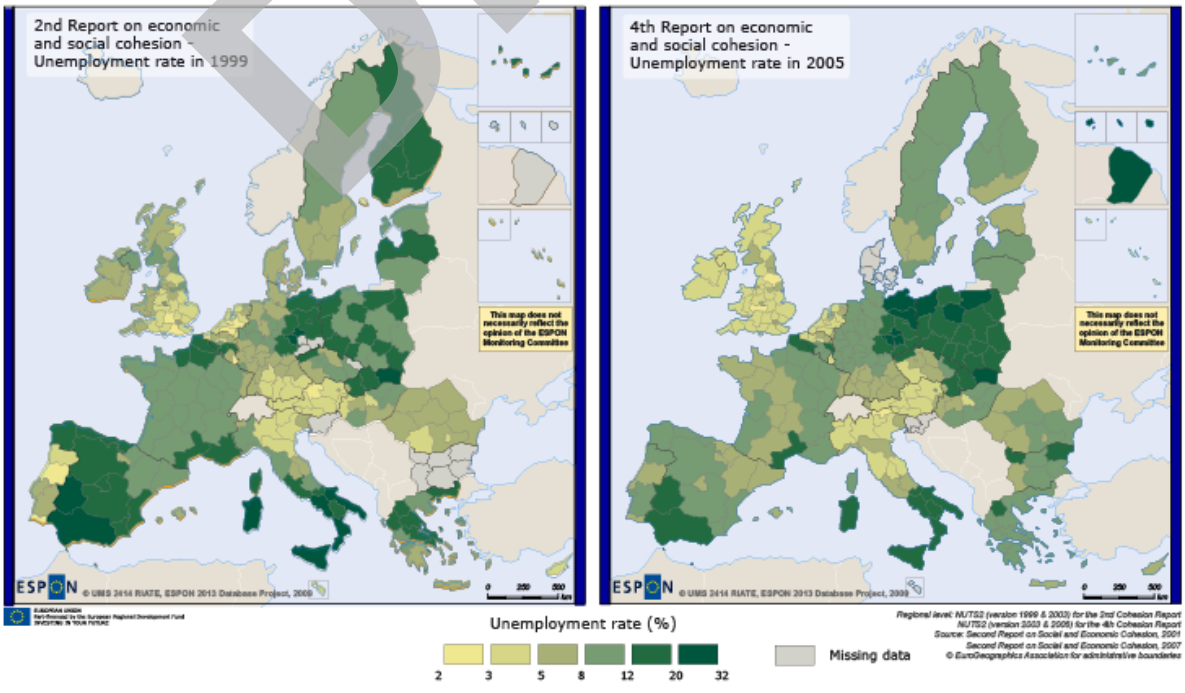


***Figure 18**: Mapping the Unemployment rate in 1999 and 2005 in heterogeneous NUTS versions*

Another work was undertaken to develop historical NewCronos database results exploration. At the stage of work, no deliverable data sets are available because of the complexity of the conversion tool to mix them with the current Eurostat data.

Conclusion

This formalization and these first results should not be seen as a normative approach, but rather a descriptive one which would be improved in the next steps of the project. Actually, these results will serve mainly as a background information (working paper) for the ESPON DB project. Meetings and exchanges with Eurostat and other institutions, interested by time series issue, will allow improving and validating the results.
The figure 19 summarizes the working progress of the times series challenge.

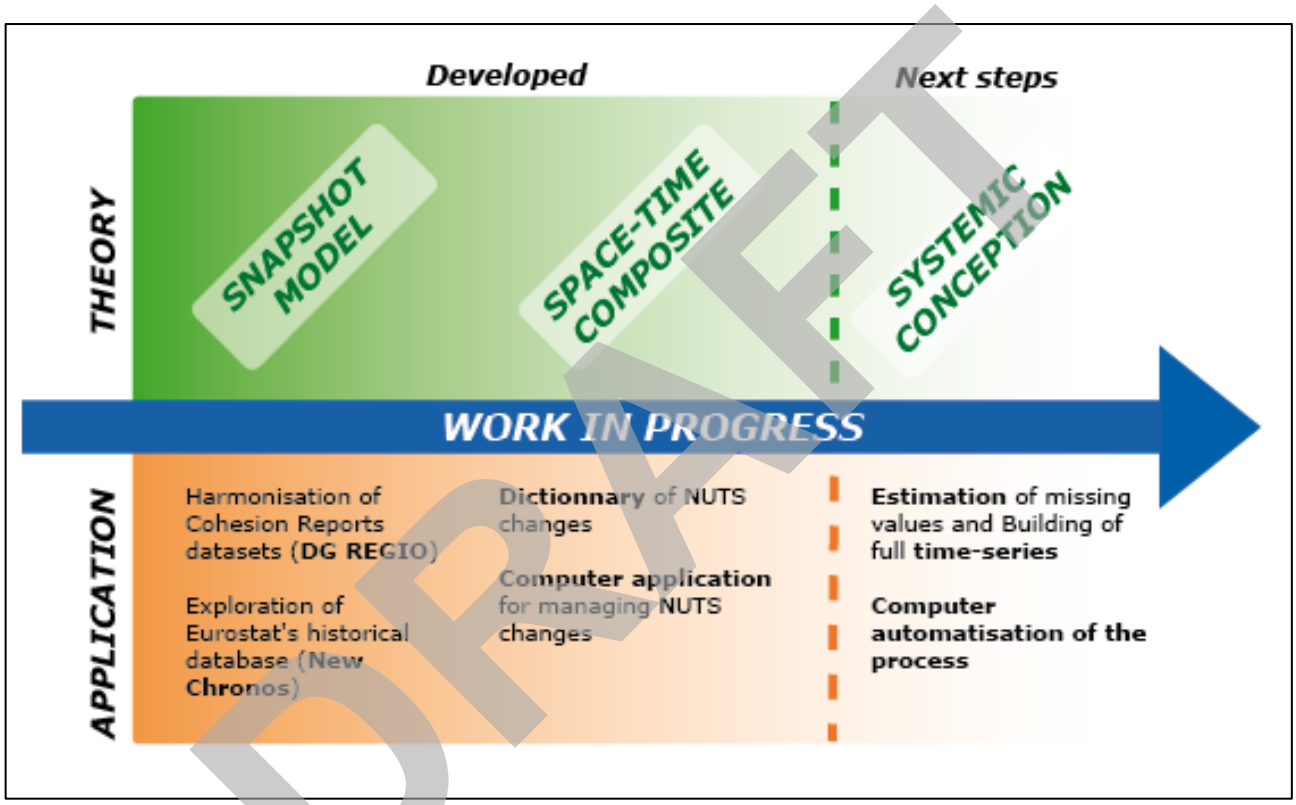


Figure 19: Synthesis of the work of the "time-series" group

References

LANGRAN G., Time in Geographic Information Systems, London: Taylor & Francis, 1992.

YUAN M., Wildfire Conceptual Modeling for Building GIS Space-time Models, Proceeding of GIS/LIS'1994, PP. 860-869, 1994.

OTT T. and SWIACZNY, F., 2001, "*Time-integrative Geographic Information Systems, management and spatio-temporal data*". Springer, 2001.

TECHNICAL REPORT

ESPON

Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase.

CONTENT

- Methodology description. This section describes the details of the three integration methods defined to face the challenge of combining data measured using different reporting units.
- Testing the methodology and automatic tools. This part presents the steps undertaken and results achieved for a series of tests made in order to test the methodology proposed. It also presents some automatic tools developed to ease the implementation of the integration methods.
- Future steps. Taking into account the results form the testing process and the needs of the challenge, this section describes the next steps to be developed.

ESPON 2013 DATABASE

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

32 PAGES

LIST OF AUTHORS

Roger Milego, University Autonomous Barcelona, UAB 08193 Bellaterra

Maria José Ramos, UAB 08193 Bellaterra

Contact

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

tel. + 34 93 581 35 46

TABLE OF CONTENT

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]**.**

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

[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

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, 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.

[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

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.

1 Methodology description

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

According to what it has been explained, 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.

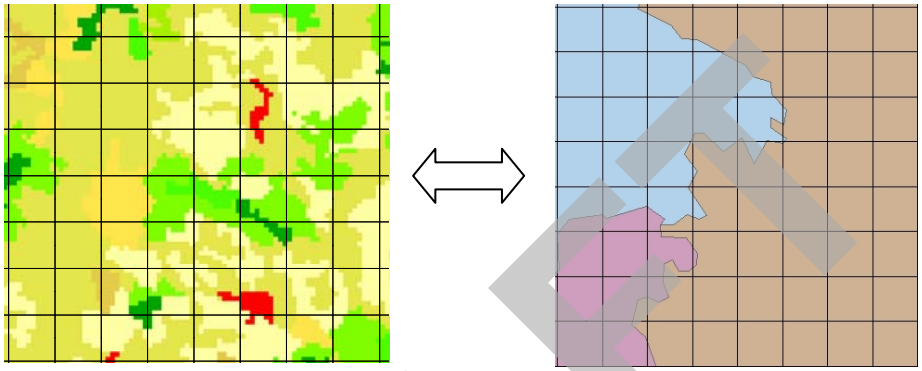


Figure 1. *The 1 km European Reference Grid will hold both environmental and socioeconomic information*.

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.

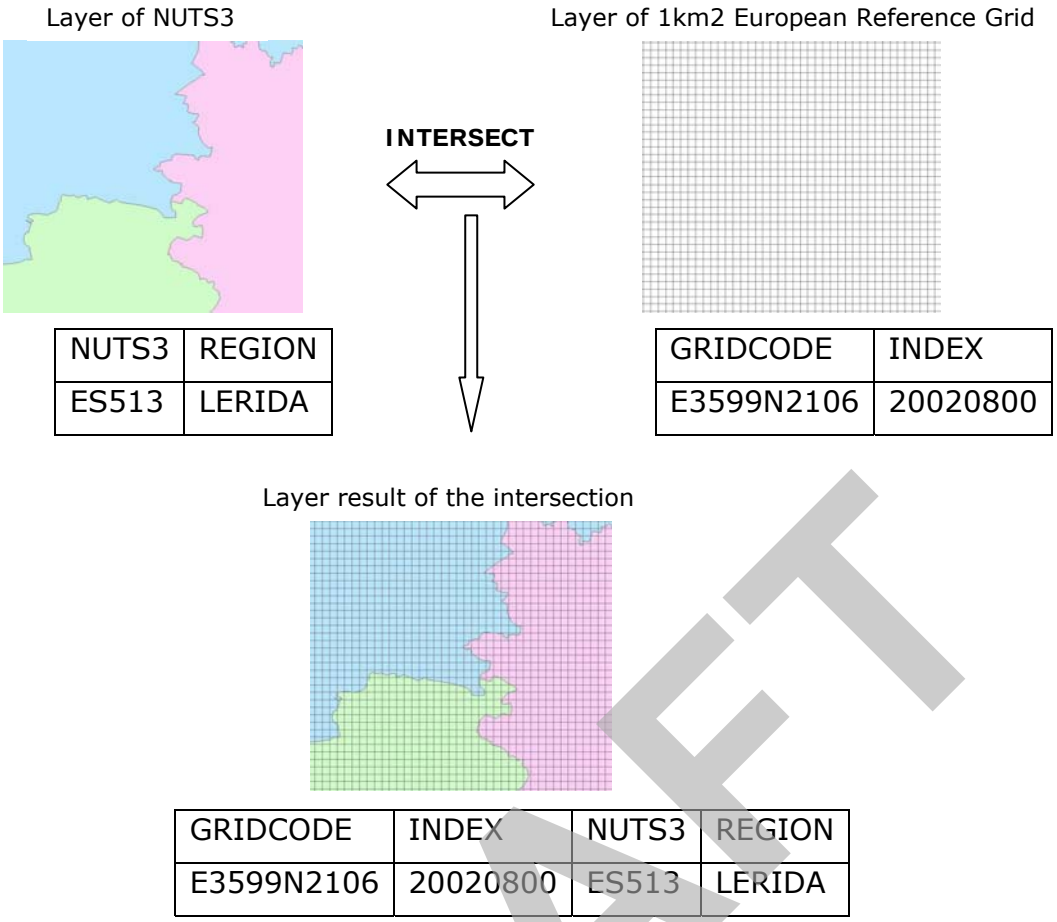
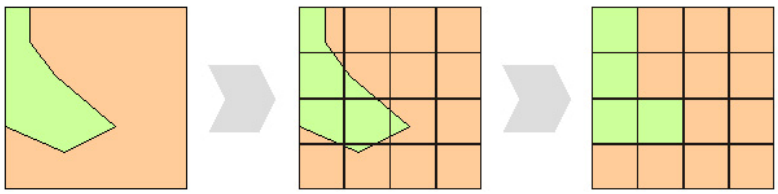


Figure 2. *Example of the intersection tool.*

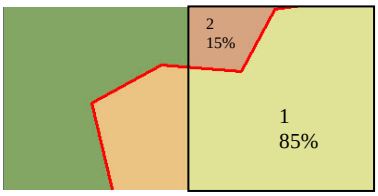
Once the intersection has been computed, a way of computing a single figure by each reference grid cell should be defined. It has been stated that depending on the nature of each indicator or variable, a different kind of integration procedure should be defined. In this regard, we have defined and tested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

1.1 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.



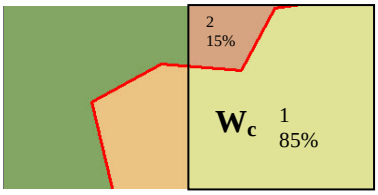
Cell value = Σ (V_i * $Share_i$)

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.

2 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.

2.1 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**[7] 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 area and to relate to other descriptors.

a)Source data:

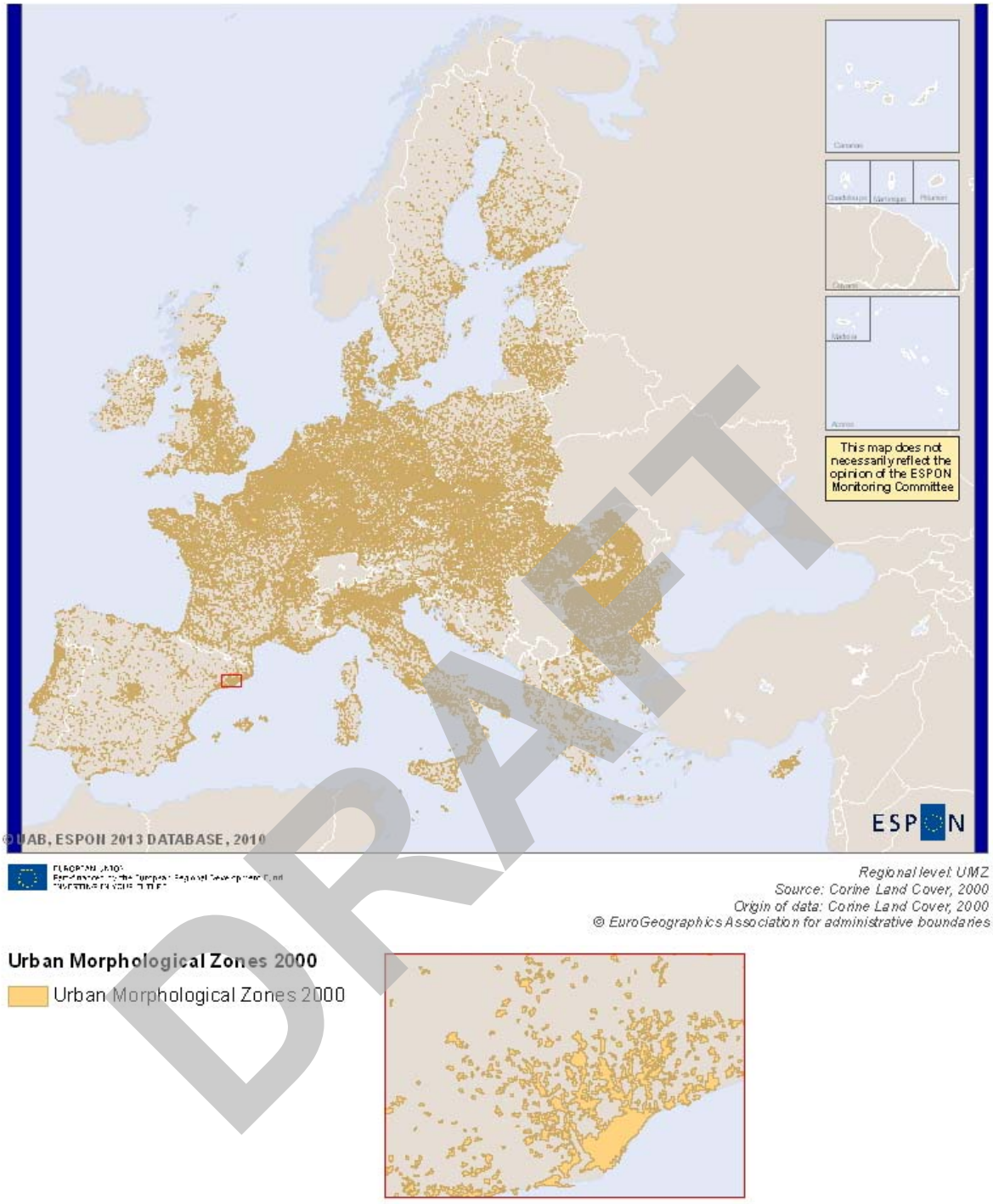
In order to make this test we have used the Urban Morphological Zones 2000 (Map1. Urban Morphological Zones 2000), an EEA's dataset which is a delimitation of urban areas according to a functional definition. The UMZ2000 come from a reclassification of different land cover classes of CLC2000 following 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

[7] 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.

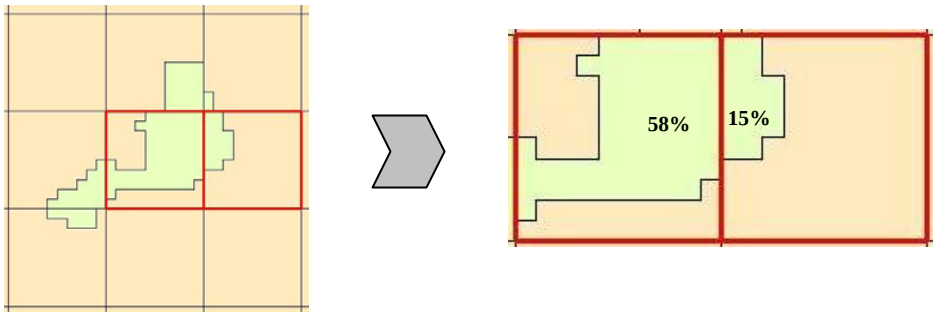


Figure 3. *Intersection of UMZ2000 and the Reference Grid*.

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.

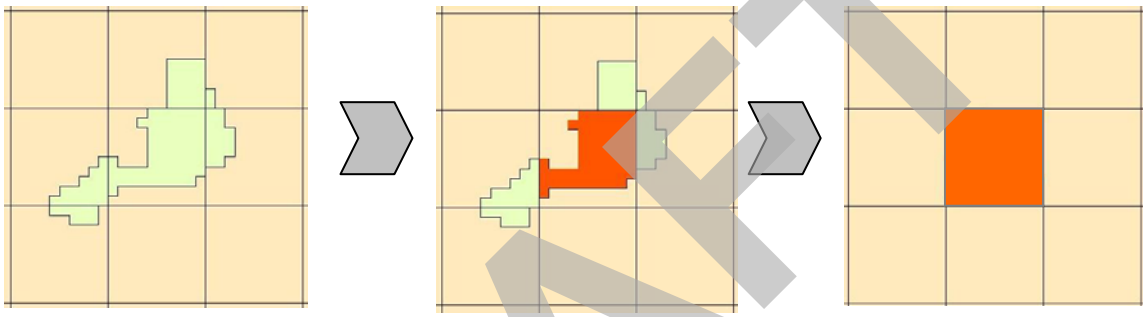
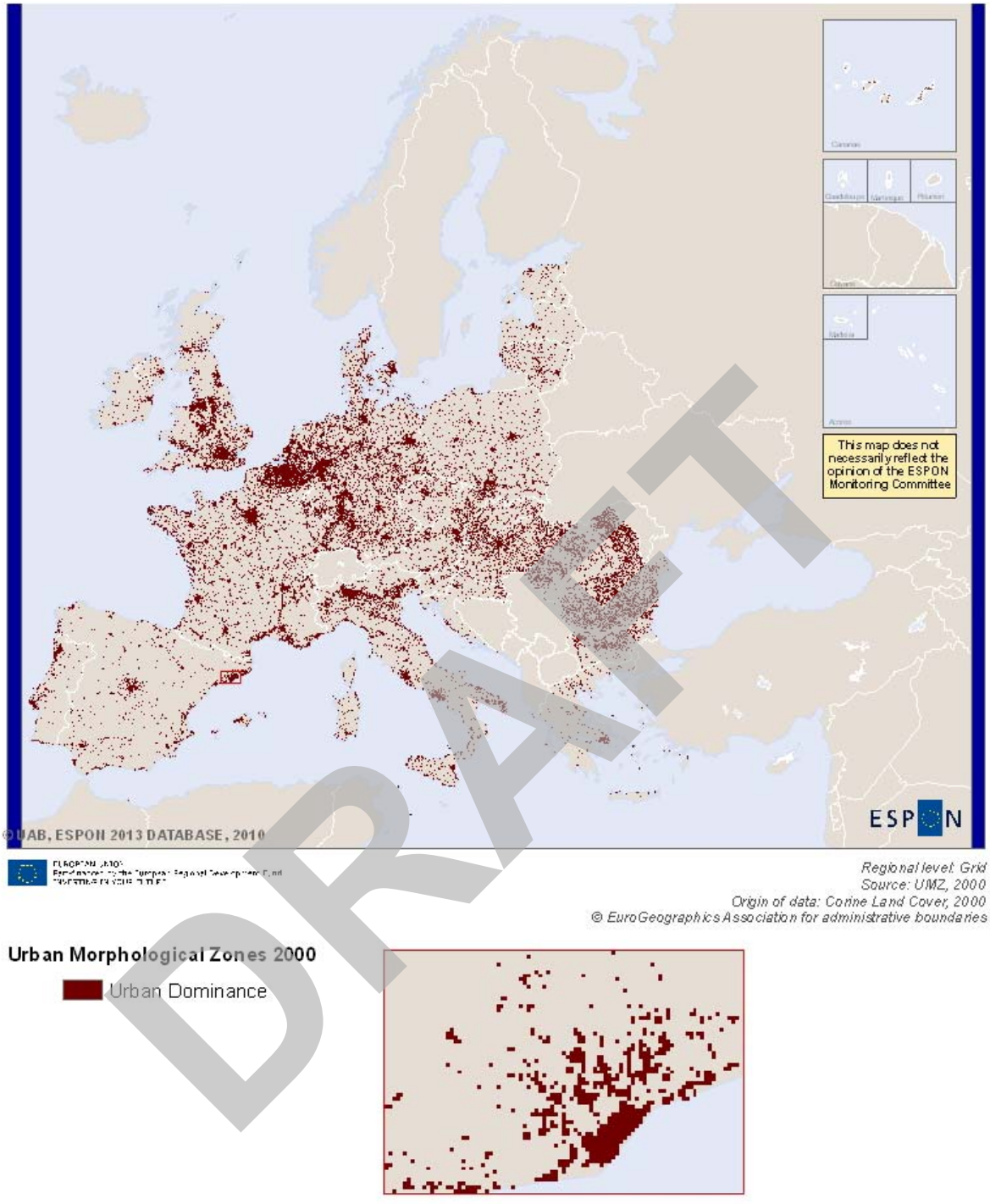


Figure 4. *Urban dominance definition process*.

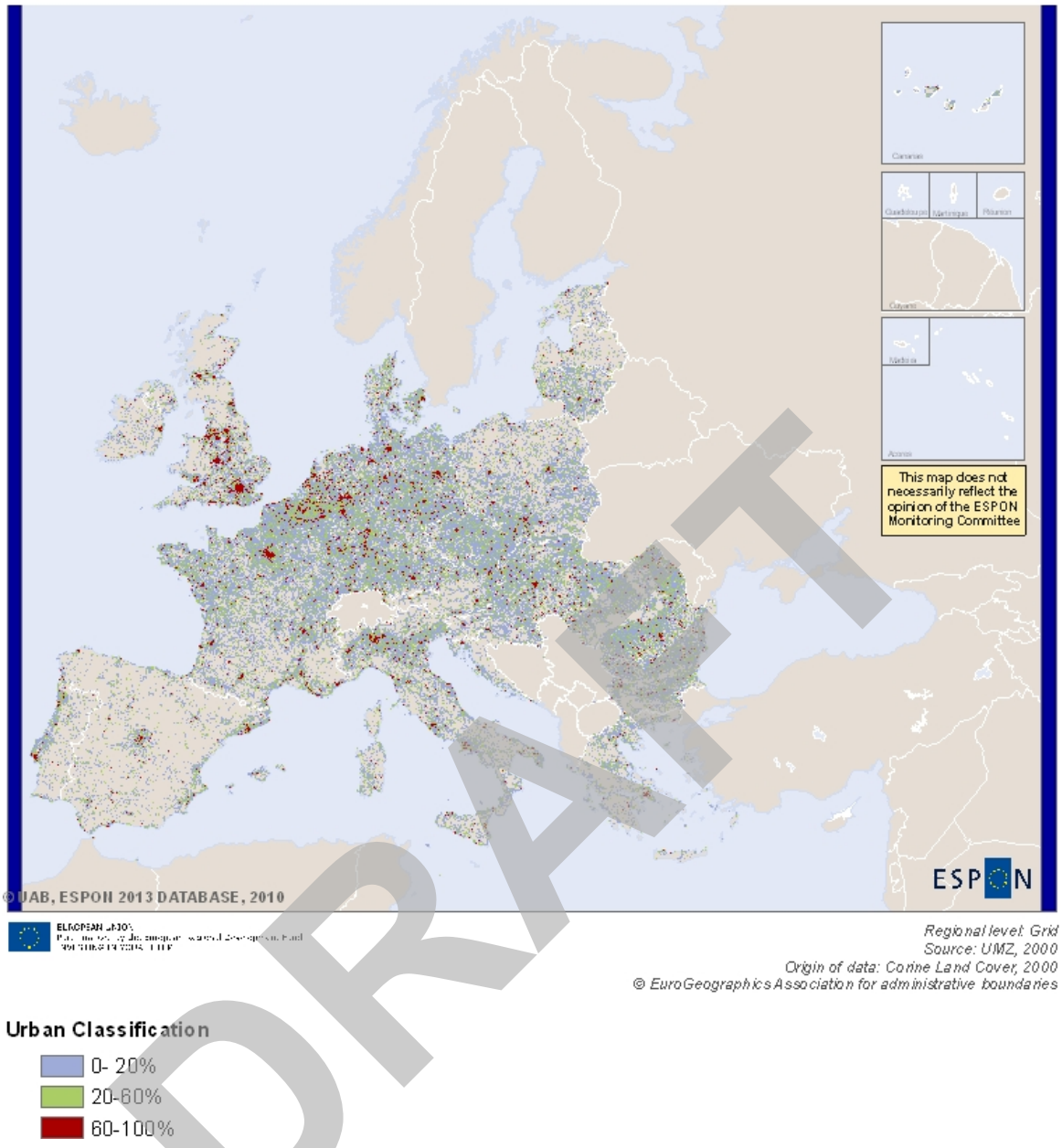
Finally, this kind of map can be elaborated, where we can see the urban dominance in Europe by 1 km grid cell being able to identify quicker the main points of urban surfaces in Europe:

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



Conclusions: 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.

2.2 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 percentage points	in percentage points	in percentage points	in percentage 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	7,8	-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.

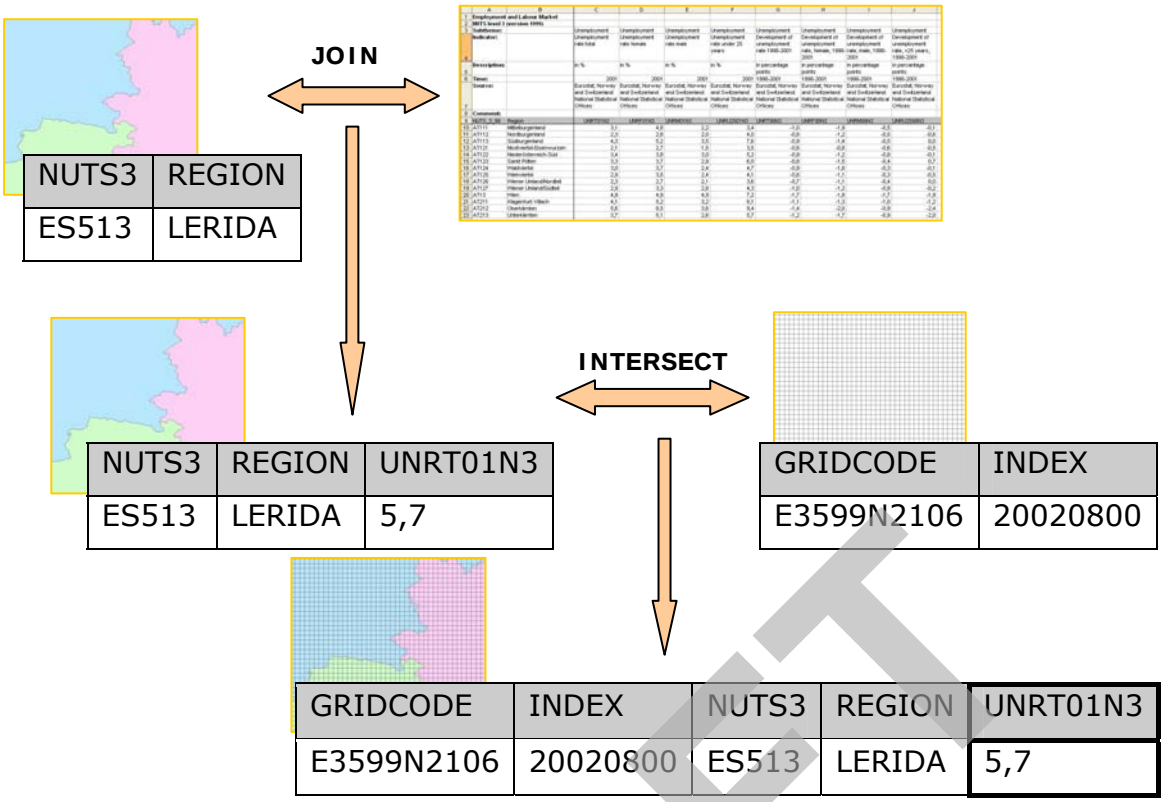
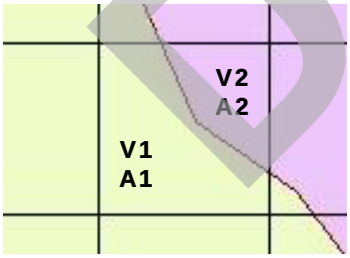


Figure 6. *Creation process schema for unemployment downscaling.*

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:

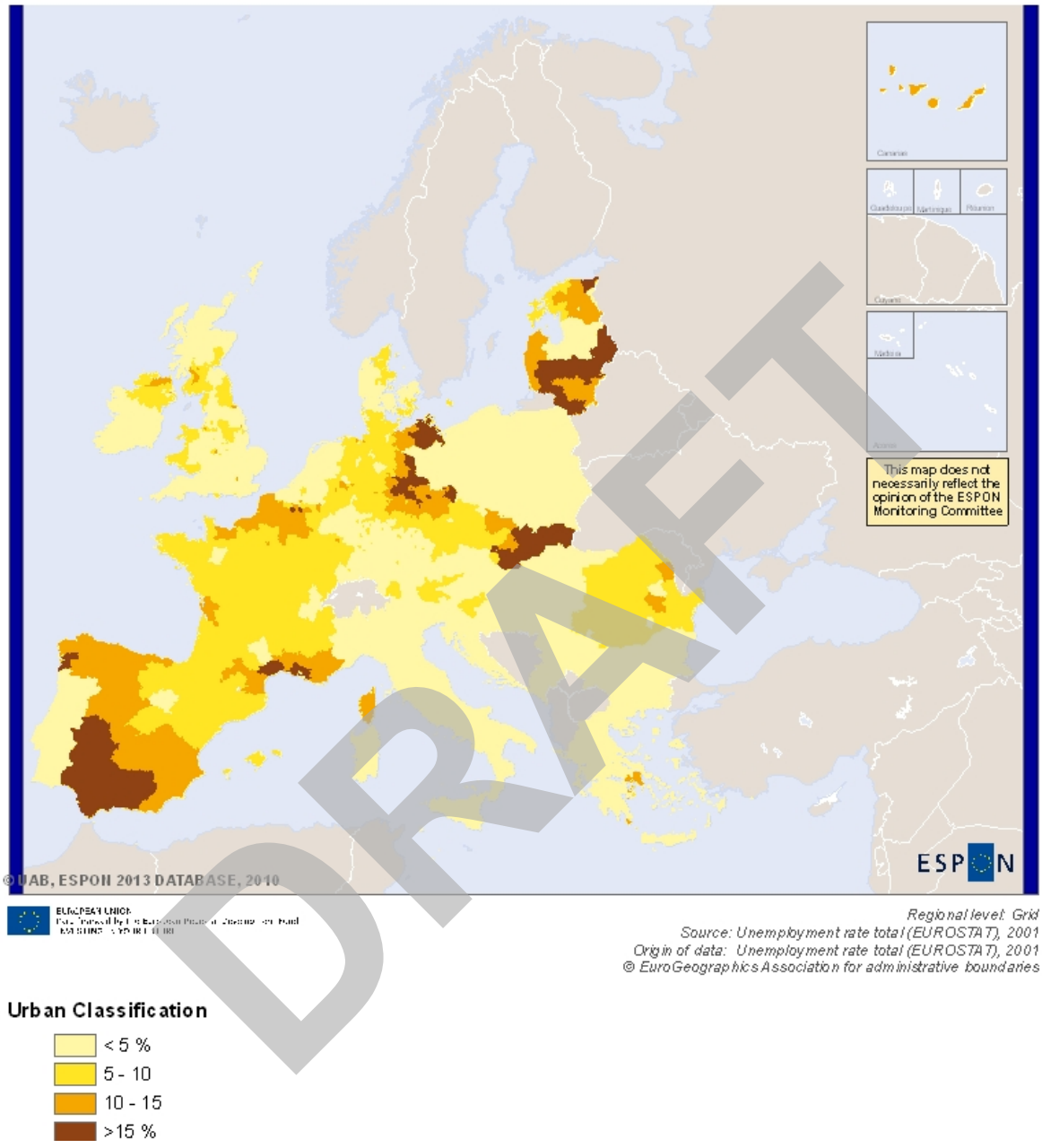
V1= 8,8 * 0,6 = 5,28
V2 = 8,6 * 0,4 = 3,44

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

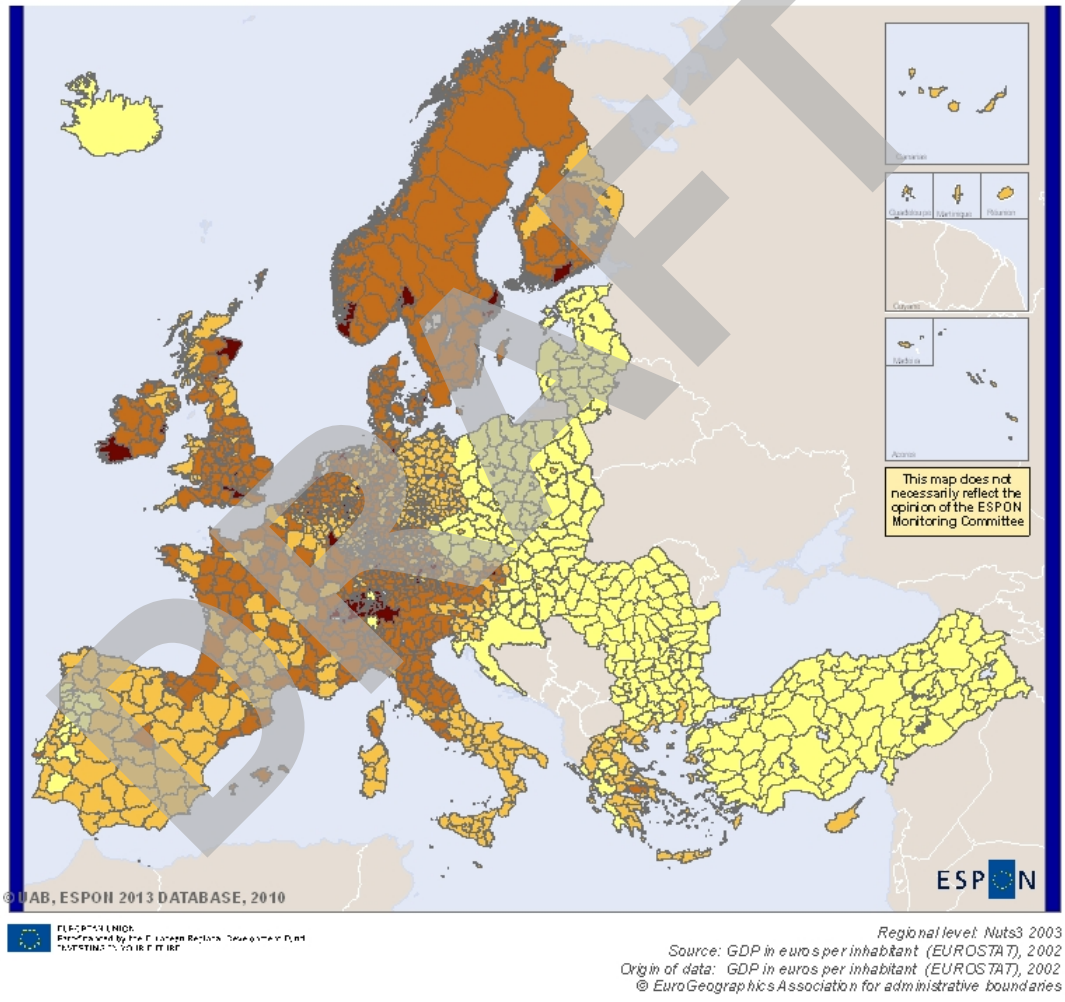


Although it seems a map where NUTS3 have been coloured as such, the data are stored by 1 km grid cell and, therefore, they can be compared with other data stored under the same grid coding, such as land cover data.

2.3 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[8] for the year 2001. In this grid, population data for communes is remapped based on Corine Land Cover classes and a quite complex algorithm[9].

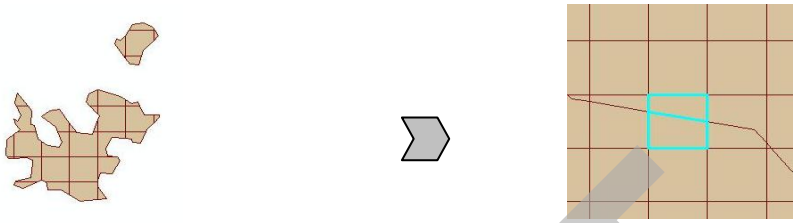
Map 5 . GDP €/inhab. 2002 distributed by Nuts3 2003



8 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

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

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.

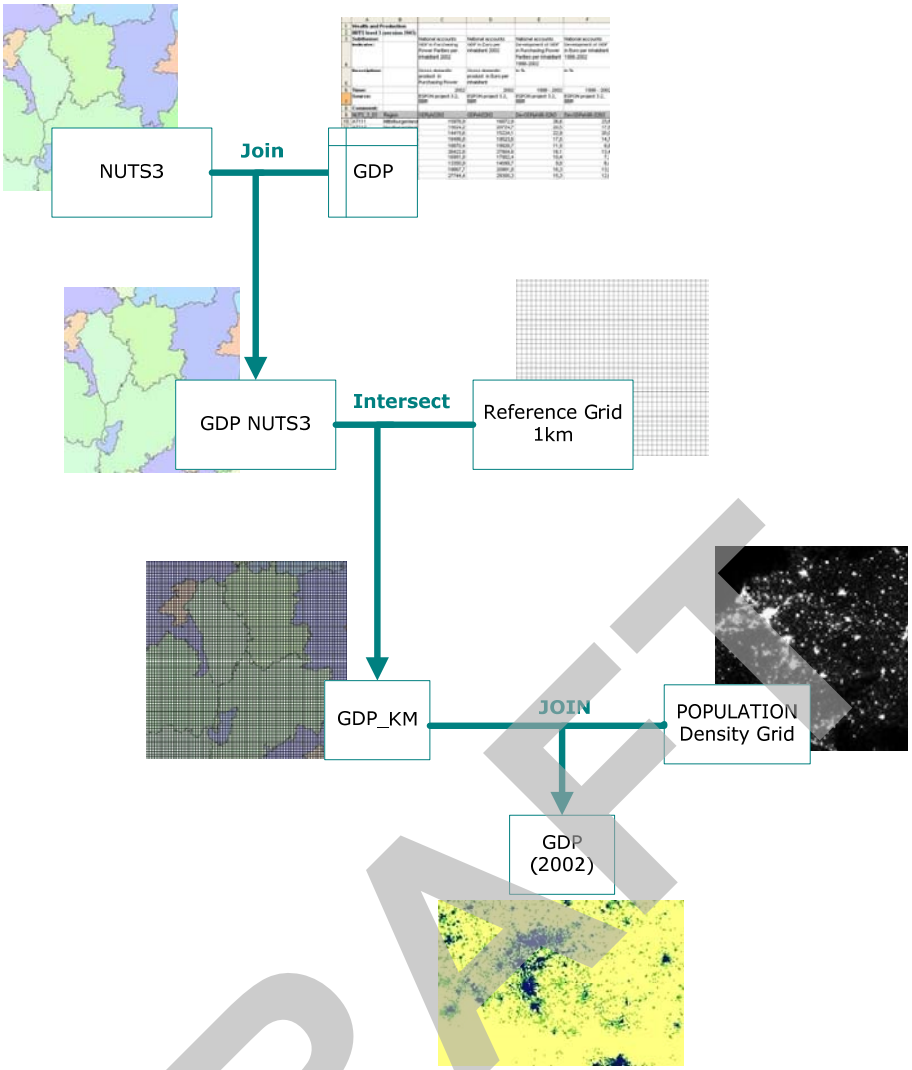


Figure 7. *Creation process schema for GDP downscaling.*

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.

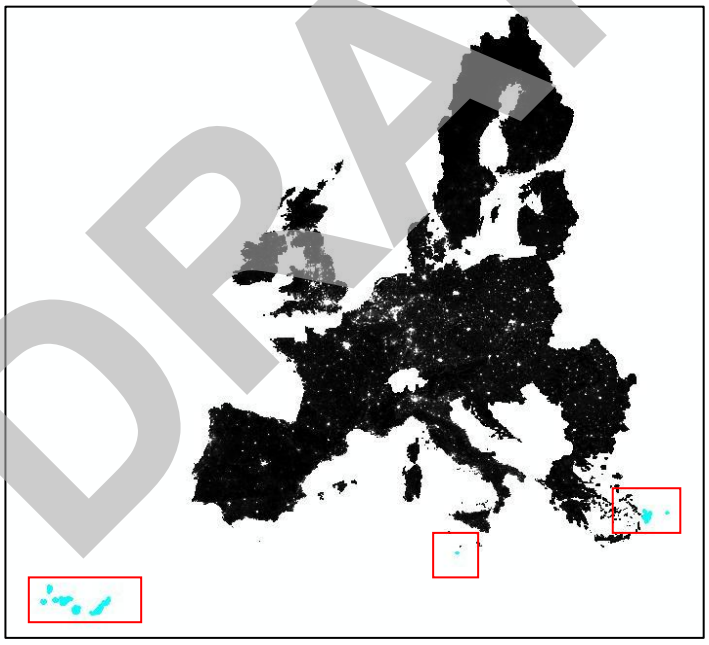
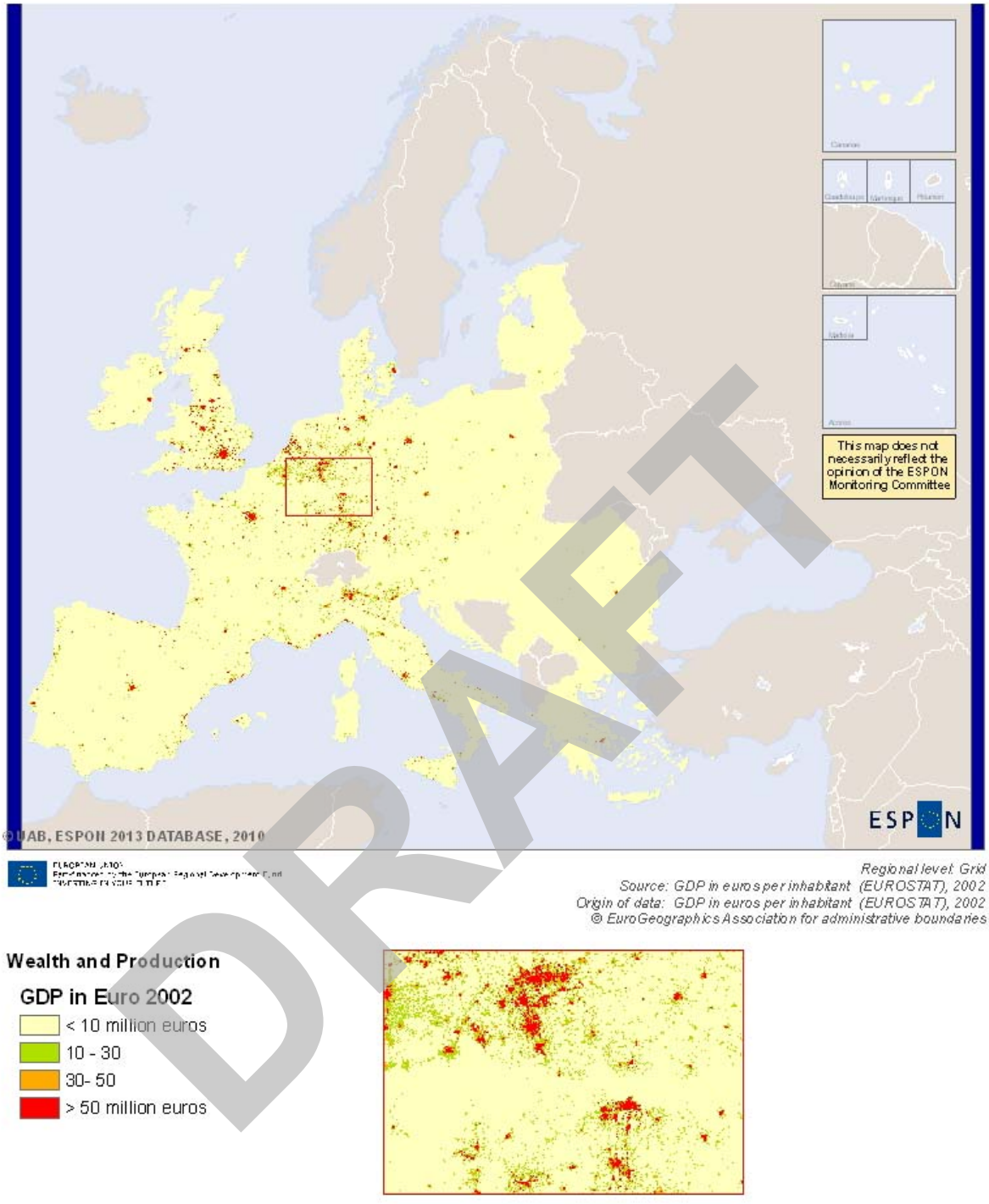


Figure 9. *Location of areas out of population grid's scope.*

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

Map 6 . Distribution of GDP in Euro 2002 by grid



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:

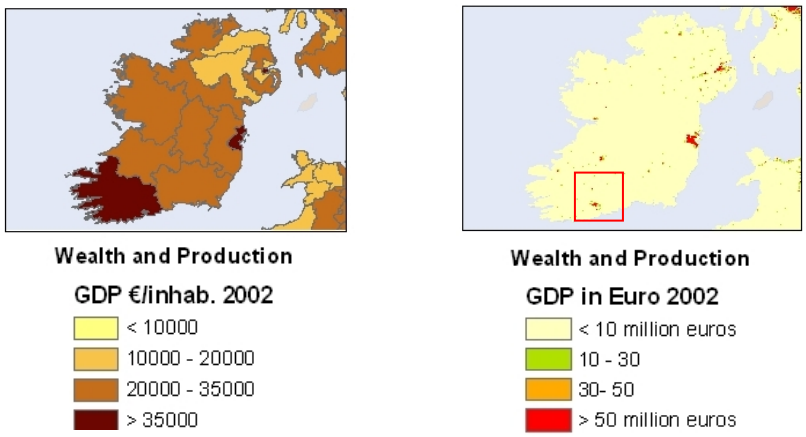


Figure 10. *Original GDP €/inhab distributed by Nuts3 2003 vs. GDP in Euro distributed by grid at the South-west of Ireland.*

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).

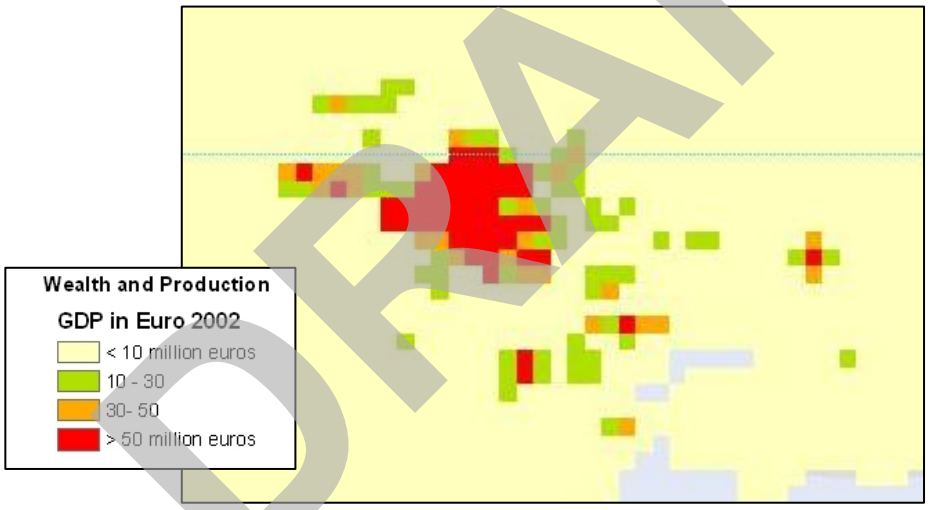


Figure 11. *Zoom in on the Cork City at the south-west of Ireland*

b) North of Italy:

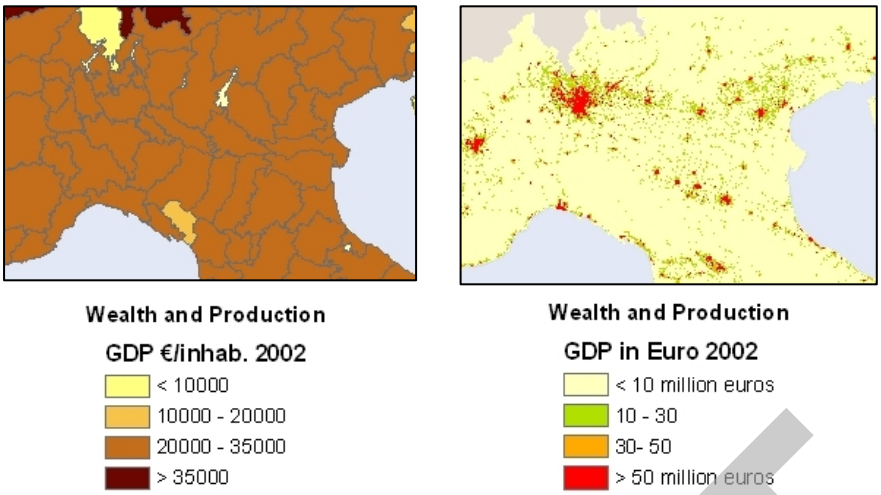
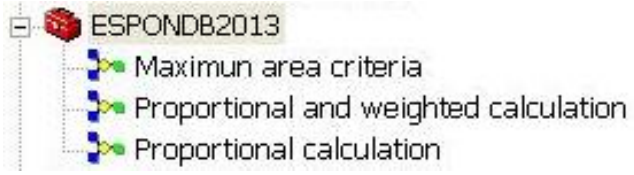


***Figure 12**. Original GDP €/inhab distributed by Nuts3 2003 vs. GDP in Euro distributed by grid at the North of Italy.*

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.

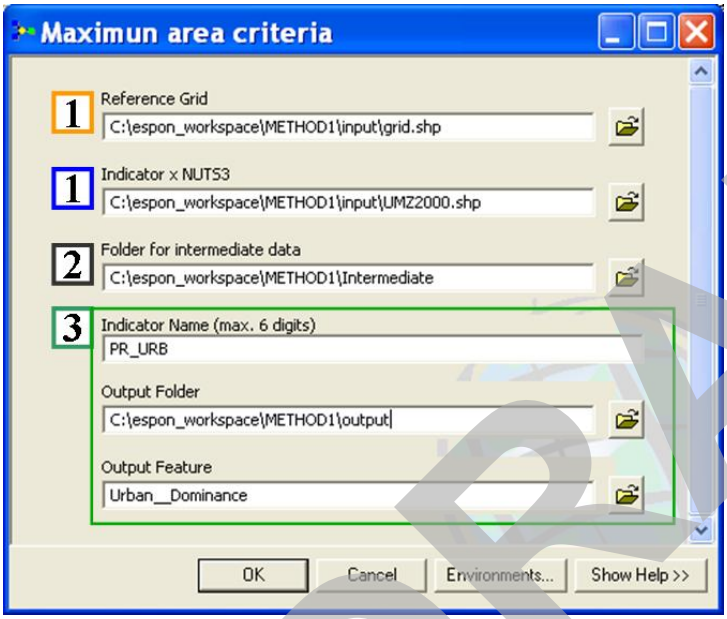
3 Automatic tools

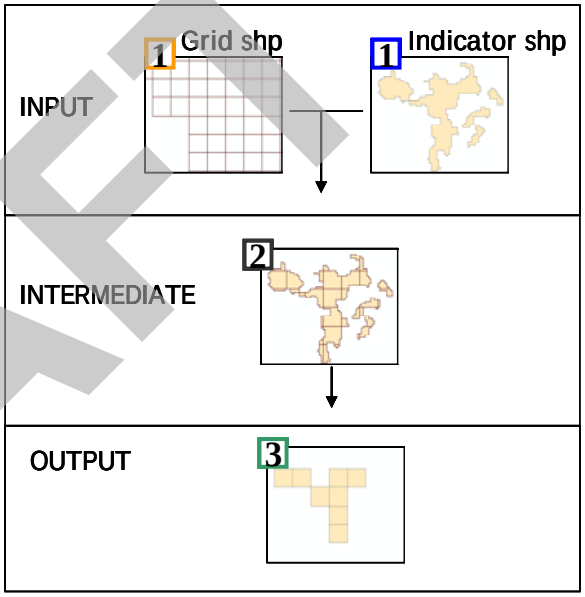
In order to facilitate the testing processes an ESPONDB toolbox within ArcCatalog has been develop for each methodology described before.



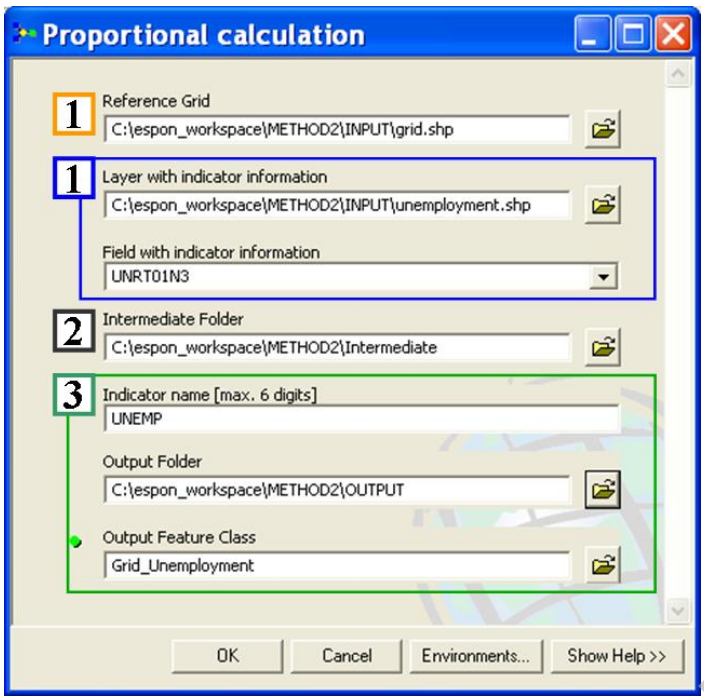
The next figures present the tools created:

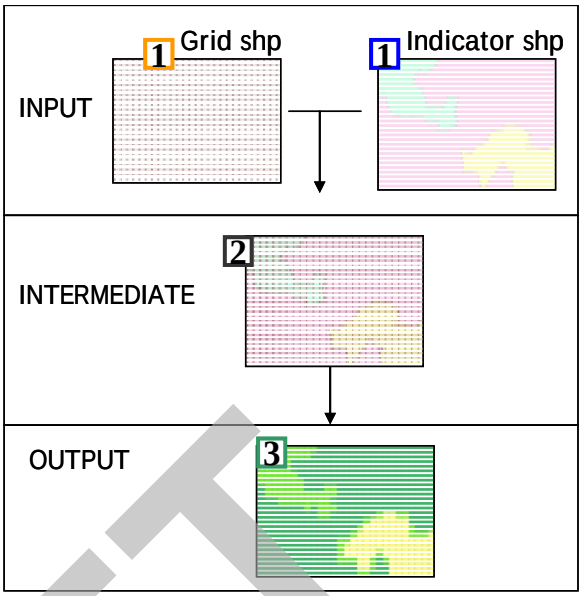
a) Maximum area criteria



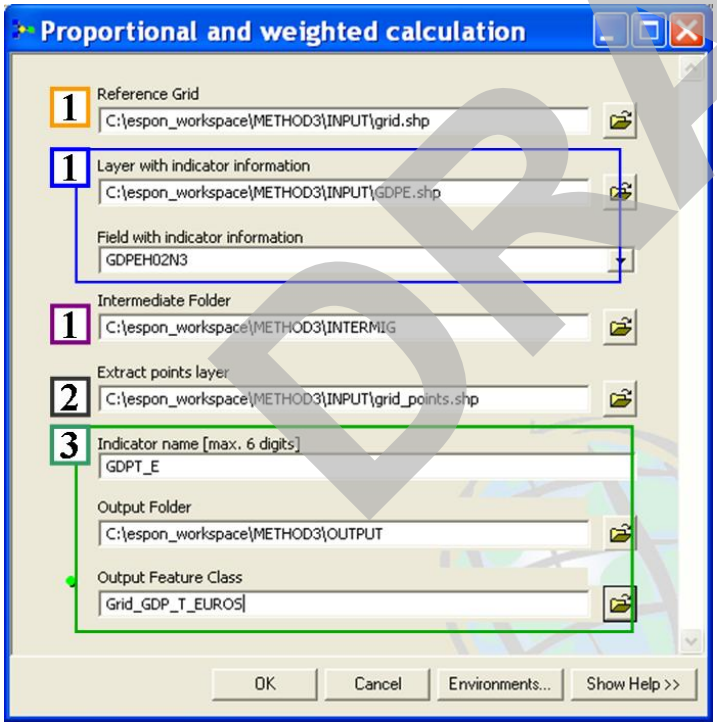


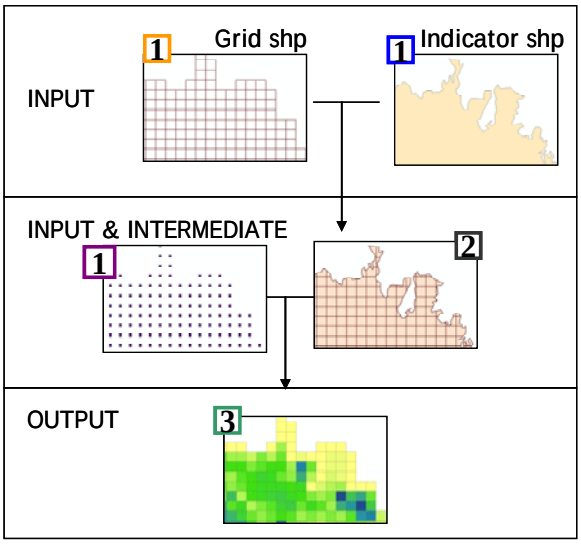
b) Proportional calculation





c) Proportional and weighted calculation





4 Future steps

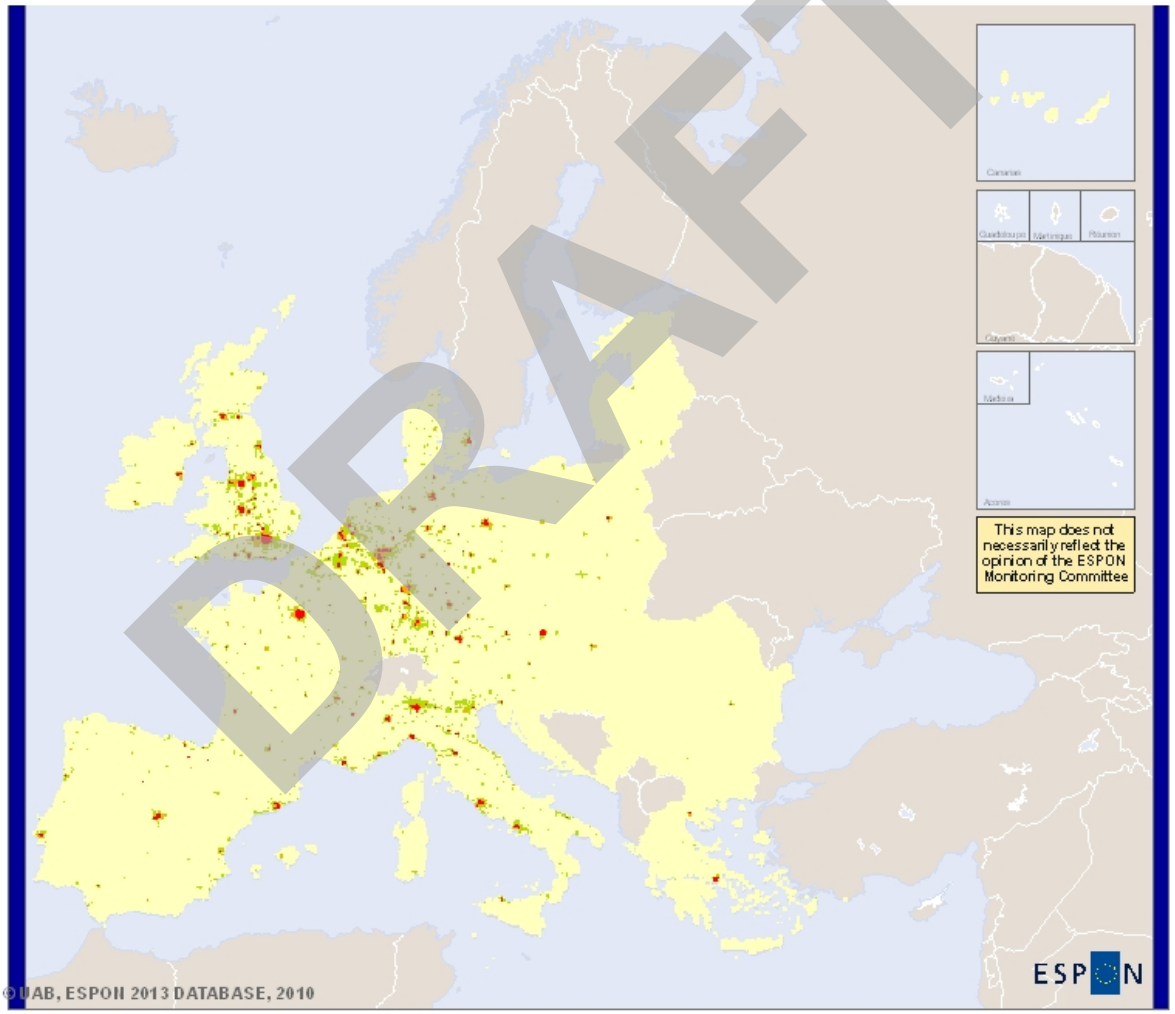
a) New tests based on the most updated information and ESPON projects results.

New tests have been started using the most updated territorial information (Nuts3 and Nuts2 2006) and new information coming from ESPON 2013 projects.

The next maps show some of these new tests developed. In both cases the methodology used is the proportional and weighted calculation where the values have been weighted by the JRC's population density grid dataset for the year 2001.

1 GDP 2001 in Million Euros distributed by 1km Grid. Source: GDP 2001 in Million Euros distributed by Nuts 3 2006 (ESPON 2013 DB).

Map 7 . Wealth and Production. GDP in Euro 2001 distributed by 1km grid



Regional level: Grid
Source: GDP in euros per inhabitant (EUROSTAT), 2001
Origin of data: GDP in euros per inhabitant (EUROSTAT), 2001
© EuroGeographics Association for administrative boundaries

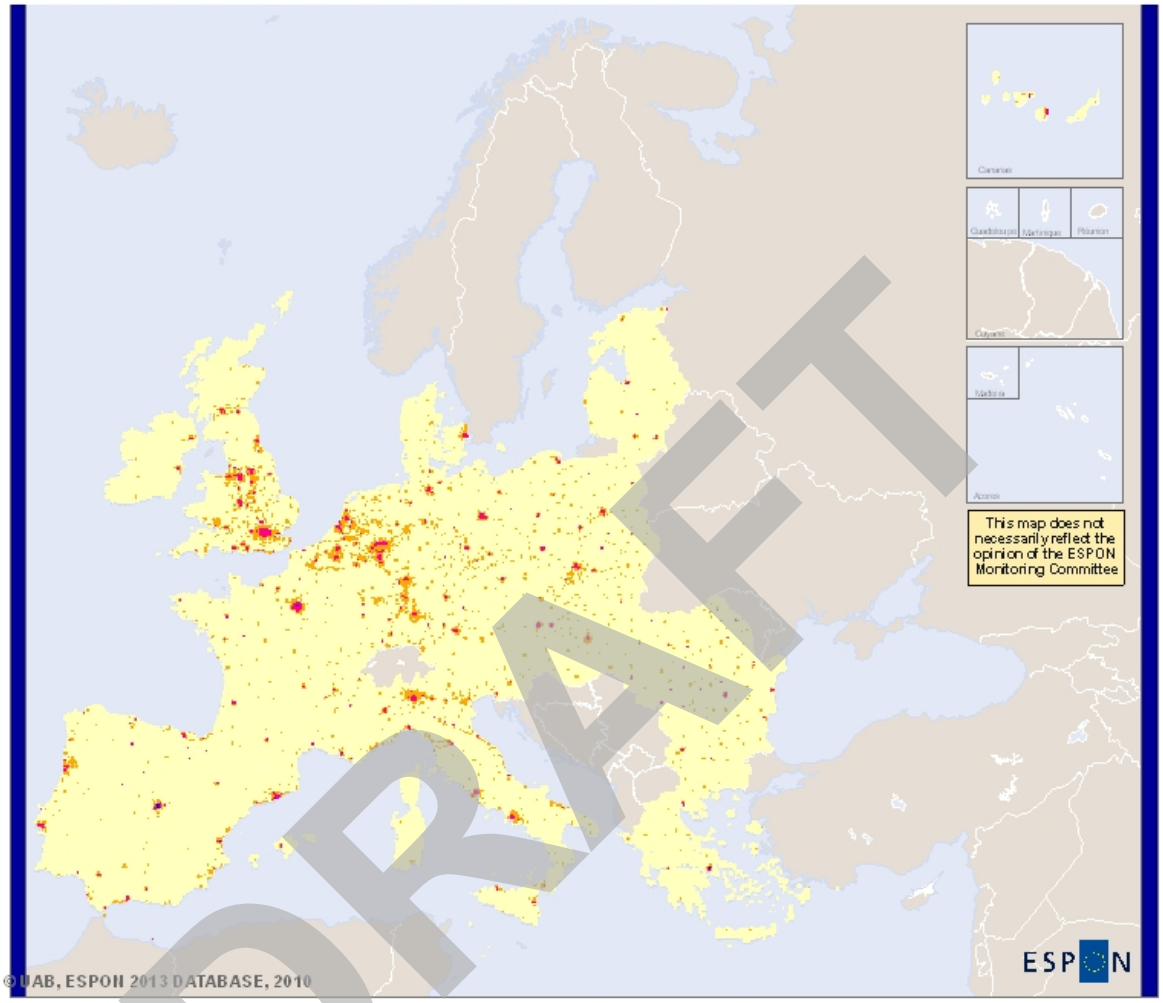
Wealth and Production

GDP in Euro 2001

- < 10 million euros
- 10 - 30
- 31 - 50
- > 50 million euros

2. Active people 2001 distributed by 1km Grid. Source: Active people 2001 in thousand inh. distributed by Nuts 2 2006 (ESPON 2013 DB)

Map 8. Active Population 2001 distributed by 1km grid



Regional level: Grid
Source: Active people 2001 in thousand inh. by Nuts 2 2006 (ESPON 2013 DB), 2006
Origin of data: Active people 2001 in thousand inh. by Nuts 2 2006 (ESPON 2013 DB), 2006
© EuroGeographics Association for administrative boundaries

Active Population 2001

- < 0.2 Thousand inhab.
- 0.2 -0.8
- 0.8 - 5
- > 5 Thousand inhab.

B) Integration of socio-economic and environmental information.

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

The OLAP technology[10] use a multidimensional data model, allowing complex analytical and ad-hoc queries with a rapid execution time.

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

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

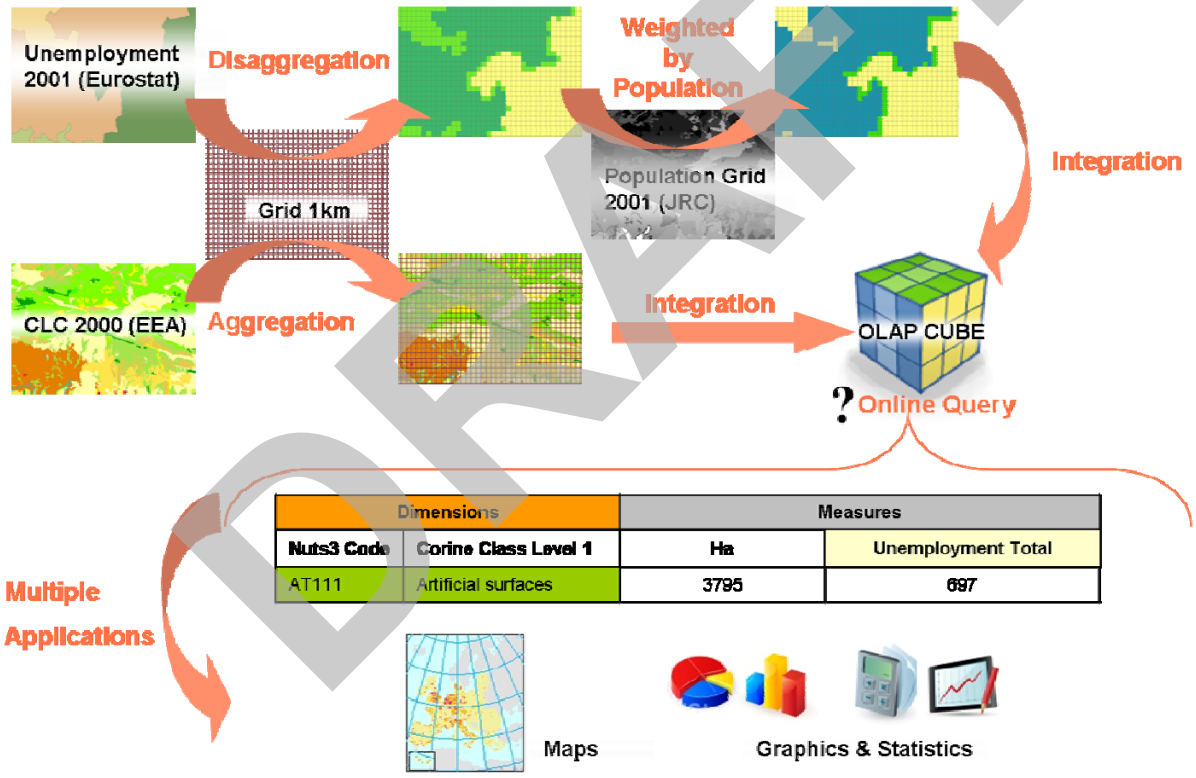
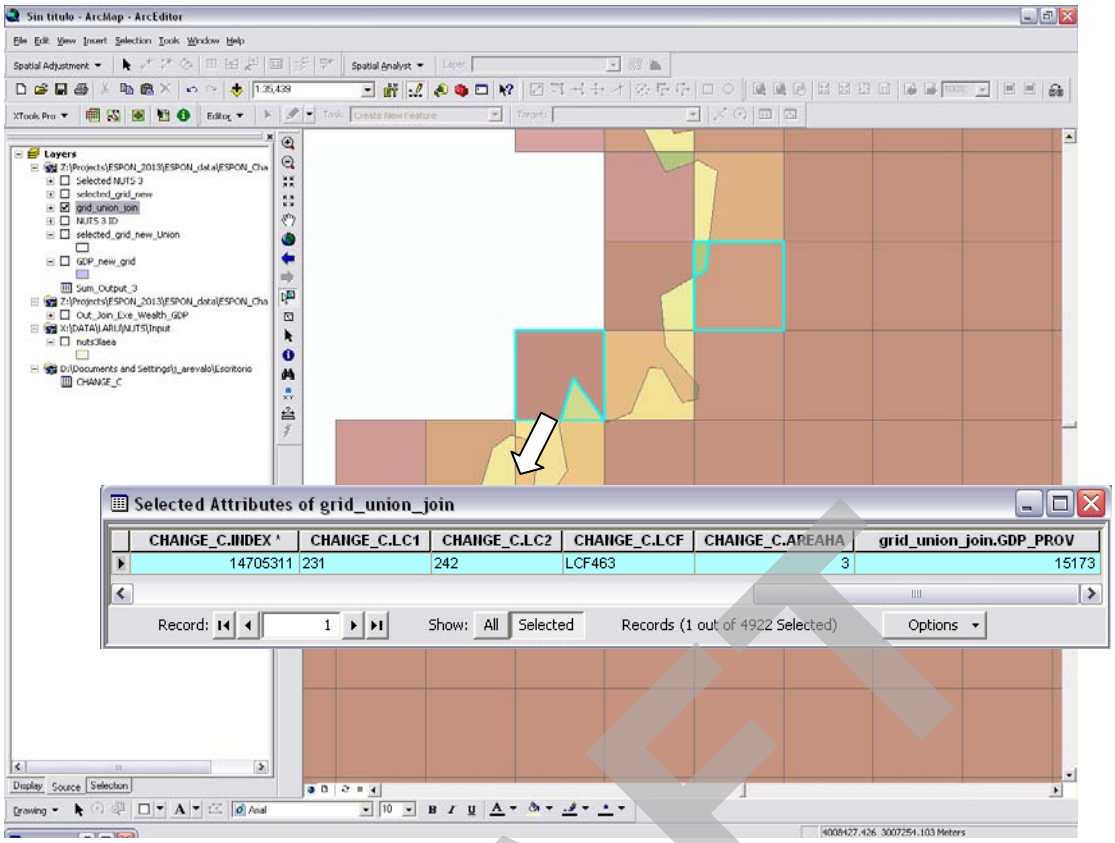


Figure 13. *Simplified schema of data integration in an OLAP Cube*

This last option will allow the user, for example, to put together a "GDP in purchasing power" value, originally measured by NUTS3 region, with the land cover flows between 1990 and 2000 or 2000 and 2006, coming by the Corine Land Cover changes:

[10]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



In this way, we will be able to ask to the ESPON-OLAP cube more questions taking also into account socioeconomic variables or indicators. For example, if we integrate the GDP and CLC in an 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 basis.

Therefore, in the following months we should **produce more disaggregated variables with more recent data, and integrate the most interesting results into an OLAP cube**, in order to show the possibilities of analysis behind this technology.

c) Improvement of the methodologies proposed.

The results of our testing processes and new projects developed in the European context will be the inputs to introduce new variables in the proposed methodologies. The addition of new methodological approaches is not excluded.

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.

The European Grid Club project is aimed at developing a grid-based system of small area statistics for Europe. It is regarded as a one of four taskforces working for the European forum for Geostatistics [11]

It is foreseen to follow all the advances, studies and methodologies proposed by the European Grid Club project. We will also try to be in direct contact with the project to allow a valuable exchanging of information and to introduce new variables in our methodologies.

d) Integration with the ESPON 2013 Database.

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

One possibility that will be analyzed is the creation of a direct link to the OLAP cube from the ESPON 2013 Database web application, although this case will be studied amongst other possibilities.

[11] 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.

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

TECHNICAL REPORT

ESPON

NAMING UMZ: METHODS AND RESULTS

CONTENT

- UMZ 2000 database. Currently, the objects are not described by a name attribute. This attribute is essential for creating a semantic link to the territory
- A new method for naming UMZ. This method gives one or several names to UMZ, according to their spatial configuration
- Urban expertise. For some cities or countries, the automatic algorithms are not relevant and must be combined with an expertise
- Discussion. Further improvements and developments are suggested

ESPON 2013 DATABASE

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

12 PAGES

LIST OF AUTHORS

Anne Bretagnolle, University Paris 1, UMR Géographie-cités

Timothée Giraud, UMR Géographie-cités, UMS 2414 Riate

Marianne Guerois, University Paris 7, UMR Géographie-cités

Hélène Mathian, C.N.R.S., UMR Géographie-cités

Contact

anne.bretagnolle@parisgeo.cnrs.fr
timothee.giraud@gmail.com
mguerois@yahoo.fr
mathian@parisgeo.cnrs.fr

tel. + 33 1 01 40 46 40 00

TABLE OF CONTENT

Introduction

The *Urban Morphological Zones* have been created in 2004 by the Environment European Agency in order to analyse "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 areas, 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]. The geographical coverage of the UMZ 2000 database is the same than CORINE Land Cover data: 27 countries of the European Union plus 4 countries in the Balkan region (Albania, Bosnia-Herzegovina, Croatia, and Macedonia)[2].

The UMZ dataset can be downloaded freely on EEA website. Different attributes are available:

- Commune codes (for each UMZ, list and attributes of all LAU 2 overlaying the unit entirely or partly)

- Population (estimated from JRC's 2001 Population density grid, version V3 made by Javier Gallego, *Joint Research Center*)[3].

- Area and perimeter

Different adjustments have been made to UMZ database in order to facilitate its use by ESPON partners. In accordance with propositions made in FIR 2009, ESPON database team in charge of Challenge 6 ("Urban data") has prepared a new version of the UMZ database which improves in two different ways the current one available on the EEA website.

- Updated Population: using automatic methods, we have updated the population of all the UMZ with the last version (version V4.1) of the Population density grid built in 2007 by *Joint Research Center*[4]). The scale used for this grid is 100x100 meters.

- Assessment of a centroïd[5] for each UMZ.

However, the principal difficulty concerns the identification of the UMZ. On the EEA website, each UMZ is characterized by a numeric identifier but not by a name. This is not a trivial point in the sense that giving a name raises different theoretical and methodological problems, which are described (with solutions and discussions) in this report.

[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] The new version of CLC2000 (september 2009) includes Norway, Lichtenstein and Island, and CLC2006 will also cover Switzerland.

[3] For further details, see Downscaling population density in the European Union with a land cover map and a point survey, http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=1018.

[4] 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.

[5] The centroïd is the centre of gravity computed as the average of the coordinates of all the UMZ's vertices.

1 Giving a name to UMZ: stakes and matter

UMZ are “physical” zones without any identity. A basic operation is to give them a name so that zones become “cities”. Only the name allows the link between physical objects delineated in the database and the concept of city. For example, an UMZ located around Berlin centre can be used as a representation of Berlin city (i.e. a state of “Berlin” at time *t* as a result of past evolution of urban sprawl) only if the user can identify the zone as “Berlin” zone.

Giving a name to UMZ implies to define a correspondence with a reference list. One could think that this reference list could consist in Local Administrative Units (LAU) and that it is enough, for example, to give the name of the main LAU 2 on which the UMZ is laying. Two objections can be made: first, one has to take into account the polycentric cities, which seem to develop more and more often in Europe and that are under ESPON study scopes. Secondly, which administrative unit must we use: LAU 1, LAU 2, NUTS 3 or other? The answer depends on historical backgrounds of countries. For example, the name “Leipzig” fit with LAU 2 level whereas the name “Dublin” fits with LAU 1 level.

The methodology we have chosen combines automatic processes and urban expertise.

Automatic processes are useful in the sense that they are the same on the whole Europe, like the automatic process of the UMZ building. Furthermore, the processes automation allow their extension to smaller UMZ (useful for some ESPON Projects), even if we have tested our methods only on the largest (more than 10 000 inhabitants).

Urban expertise is necessary to take into account national and historical urbanization contexts, before making choices that will modify some results of automatic processes.

2 Methodology

2.1 General presentation

The methodology that has been chosen is inspired by the one used by French Census Board (INSEE) to give names to French urban areas (*unités urbaines*)[6]. Rules and criteria have been elaborated to differentiate three types of spatial configurations:

- *Situation 1 : UMZ with a strong core* (the city receives one name)
- *Situation 2: UMZ with several cores* (the city receives several names)
- *Situation 3: UMZ with a less strong core* (the city receives one name)

In the first situation, the major part of the UMZ population is located inside one administrative unit, and the city extents around a clear morphological centre. We have retained, like INSEE, the minimal threshold of 50 % inhabitants, which gives rather goods results (these results are discussed later).

In the second and third situation, no administrative unit concentrates more than 50 % of the UMZ population: we retain therefore the main administrative unit AND we examine the other administrative units that largely contribute to the UMZ population. If they represent more than 50% of the main administrative unit contribution, we retain them and the UMZ is considered as "UMZ with several cores" (Situation 2). If not, we keep only the main administrative unit for naming UMZ. It is then considered as "UMZ with a less strong core" (situation 3).

2.2 Algorithms and examples

The methodology can be presented as a succession of steps or algorithms (see Figure 1). Each step involves automatic calculations.

2.2.1 Geometrical and statistical sources

Three different types of objects are overlaid.

- UMZ

- Local administrative units (LAU 1 and LAU 2, EuroBoundaryMap 2006 v2.0 from EuroGeographics, validity: 2006).

- Population density grid from EEA (see above)

[6] Composition communale des unités urbaines, Population et délimitation 1999, Nomenclatures et codes ; INSEE mars 1999

2.2.2 Computation steps

We compute the population intersecting LAU and UMZ, for each LAU and we observe the maximal value. Three different situations can occur (Figure 1).

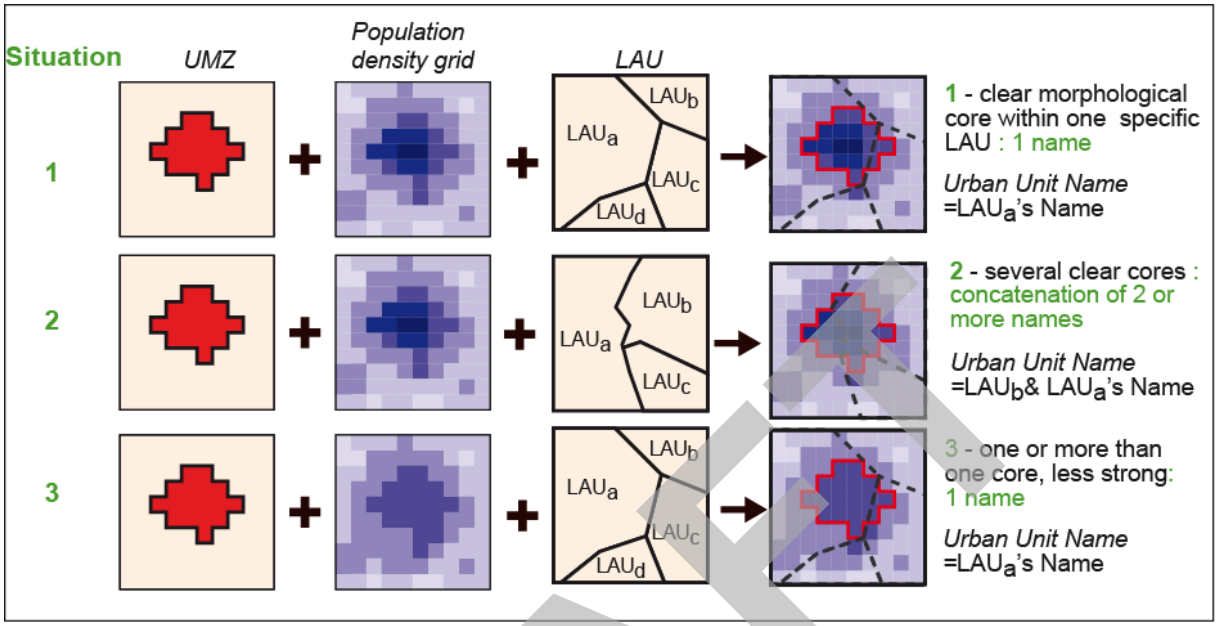


Figure 1: *Naming methodology (Situation 1, 2 and 3)*

SITUATION 1: The maximal value is larger than 50% (in Figure 1, it is the case for LAU_a). We have an UMZ with one strong core, clearly organized around a center located in LAU_1. We give the name of the LAU_a" to the UMZ. This is the case of Leipzig example (Figure 2).

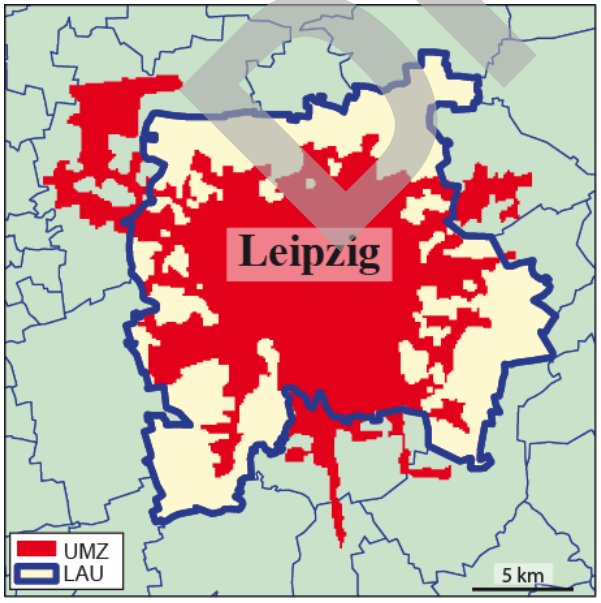


Figure 2: *Leipzig (Germany), an UMZ with one strong core (Situation 1)*

UMZ population: 534 896 inh.
UMZ population in Leipzig LAU 2: 483709 inh.
(Population figures: Population density grid V.4-1)

SITUATION 2 and 3: There is no one main core as defined above, thus the unit containing the main part of population of the UMZ is retained (LAU_b in Figure 1). Then the administrative units that contribute the most to the UMZ population are examined (here LAU_a and LAU_c). The population located in the intersection of these secondary LAU and the UMZ is then considered and shown against the population of the main LAU within the UMZ. Two different situations can occur.

Situation 2 : one or several secondary unit represents more than 50% of the main administrative unit population. 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).

Situation 3: no secondary unit represents more than 50% of the main administrative unit population. 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).

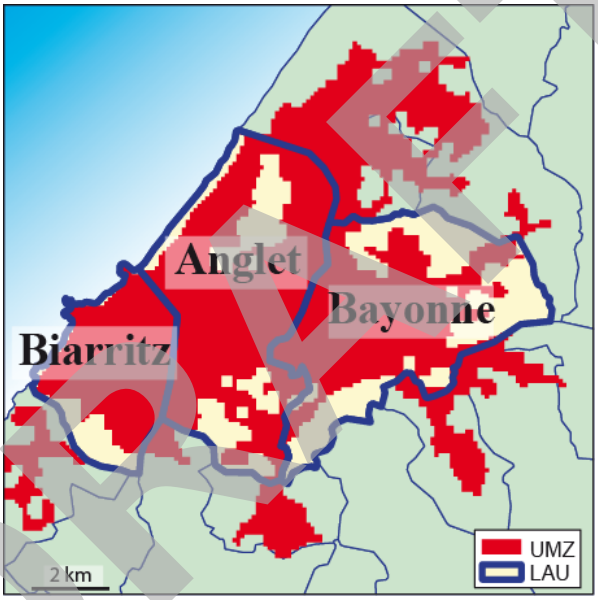


Figure 3: *Bayonne-Anglet-Biarritz (France), an UMZ with several cores (Situation 2)*

UMZ population: 128 554 inh.
Bayonne LAU 2 population inside UMZ: 39 733inh.
Anglet LAU 2 population inside UMZ: 35 032 inh.
Biarritz LAU 2 population inside UMZ: 30 156 inh.
Other LAU 2 population inside UMZ < 12 000 inh.
(Population figures: Population density grid V. 4-1)

2.2.3 Final data check

If two UMZ receive exactly the same name (Figure 4), we add a number after the name, according to the size of UMZ populations (Lódz - 1 and Lódz - 2).

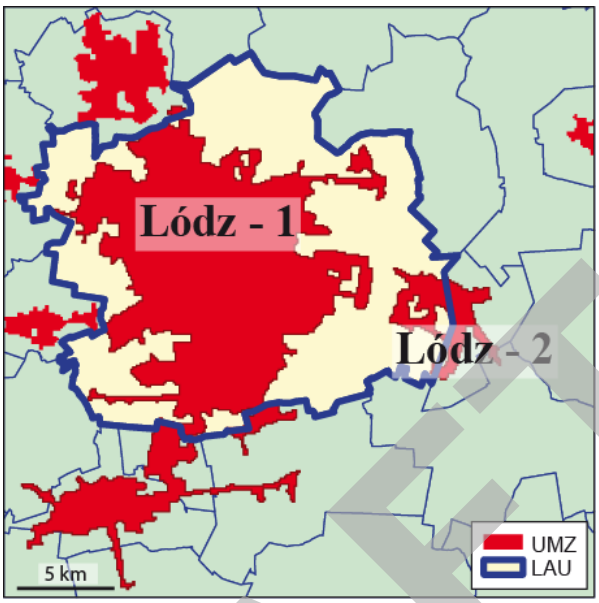


Figure 4: *Lódz (Poland), two UMZ receive the same name*

UMZ Lódz - 1: 834 112 inh.
UMZ Lódz - 2: 31 938 inh.
(Population figures: Population density grid V. 4-1)

3 Expertise and results

3.1 Which relevant administrative levels?

As recalled above, one basic question concerns the choice of administrative level used to give the names. We cannot use LAU 2 level for each country (for example, the name “Leipzig” fits with LAU 2 level whereas the name “Dublin” fits with LAU 1 level and the name “Paris” with NUTS 3). Thus, an expertise is necessary to select the best administrative unit level for each country.

3.1.1 LAU 1 instead of LAU 2 (United Kingdom, Ireland, Portugal)

In United Kingdom, Ireland, Portugal, the LAU 2 level does not fit with names given historically to cities. In Portugal, the status of city was given formerly by decree and most of the cities corresponded to LAU 1 capital cities (*capitais de distrito*). This heritage is still present, in the sense that the current names of the LAU 2 have no relationships with the names of the cities (Lisbon, Porto etc.). In United Kingdom and Ireland, LAU 2 correspond to very small units in urban areas and very large ones in rural areas, without any relationships with the names of the cities, better fitted at LAU 1 level. We will discuss again this point in Part 4.

3.1.2 NUTS 3 instead of LAU 2 (some capital cities)

In some countries, we have chosen the NUTS 3 level for capital cities, for different reasons.

- Paris, Bucharest, and Budapest: there are sub-city districts (called “arrondissement” or “sector”) at LAU 2 level. We have replaced the LAU 2 level by the NUTS 3 level in the automatic algorithms.

- London: we don’t find the toponym “London” at LAU 2 level (and the algorithm gives an “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 present 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 commune 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 of the UMZ is Brussel-Antwerpen-Gent.

- Valetta (Malta): there is just one administrative level below the national one (a LAU 2 level), and the eponym commune is too little to emerge from the automatic algorithm in the final name of the UMZ. We have then attributed by ourselves the name Valetta.

3.1.3 Other cases to be discussed

Some other particular cases have been identified.

- In Slovakia: Bratislava and Košice are divided in several districts at LAU 2 and LAU 1 levels, and the better level for the toponym (NUTS 3: “Bratislava region” and “Košice region”) is very large compared to UMZ extent. Here again, we have attributed by ourselves the names Bratislava and Košice to the UMZ.

Other particular cases will probably be detected later, through further developments and analysis, and will be noted in metadata information.

3.2 Balkan countries and Cyprus

In these countries, we have to face another type of particular cases. For the 53 UMZ larger than 10 000 inhabitants located in some of the Balkan countries (Albania, Bosnia-Herzegovina and Macedonia) and in Cyprus, the population has not been attributed by EEA using the Population density grid but using other sources (www.citypopulation.de)[8]. Consequently, we have not applied the automatic algorithms to these 53 UMZ but have kept the same name than the one given in UMZ database.

3.3 Results and map

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 (4357 objects). The results have been summarized in Table 1:

	SITUATION 1 “UMZ with one strong core”	SITUATION 2 “UMZ with several cores”	SITUATION 3 “UMZ with one core”
Total number	4108	80	169
Percentage	94	2	4

Table 1: *Naming UMZ through automatic methods and LAU2 level*

The typology presented in Table 1 has been mapped (Figure 5). The result allows a first validation of the naming processes that have been used. Indeed, if we focus first on “situation 2” (several cores), we recognise the industrial conurbations of the Midlands, of the French and Belgium basin, of the Ruhr basin, of Silesia and Galicia. We also identify the sea-side conurbations, for example in Portugal, Spain, Italy or France. A third type of “UMZ with several cores” consists in large cities sprawling and

[7] Urban Morphological Zones 2000 Version F1v0. (op. cit.)

[8] Urban Morphological Zones 2000 Version F1v0. (op. cit.)

connecting other large and closed cities, like in Belgium (around Brussels) or in Romania.

For the “situation 3” (one core but 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…), as if they could be interpreted as a step in a dynamic process from “one-core cities” toward “several-cores cities”. These results have obviously to be explored further.

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.

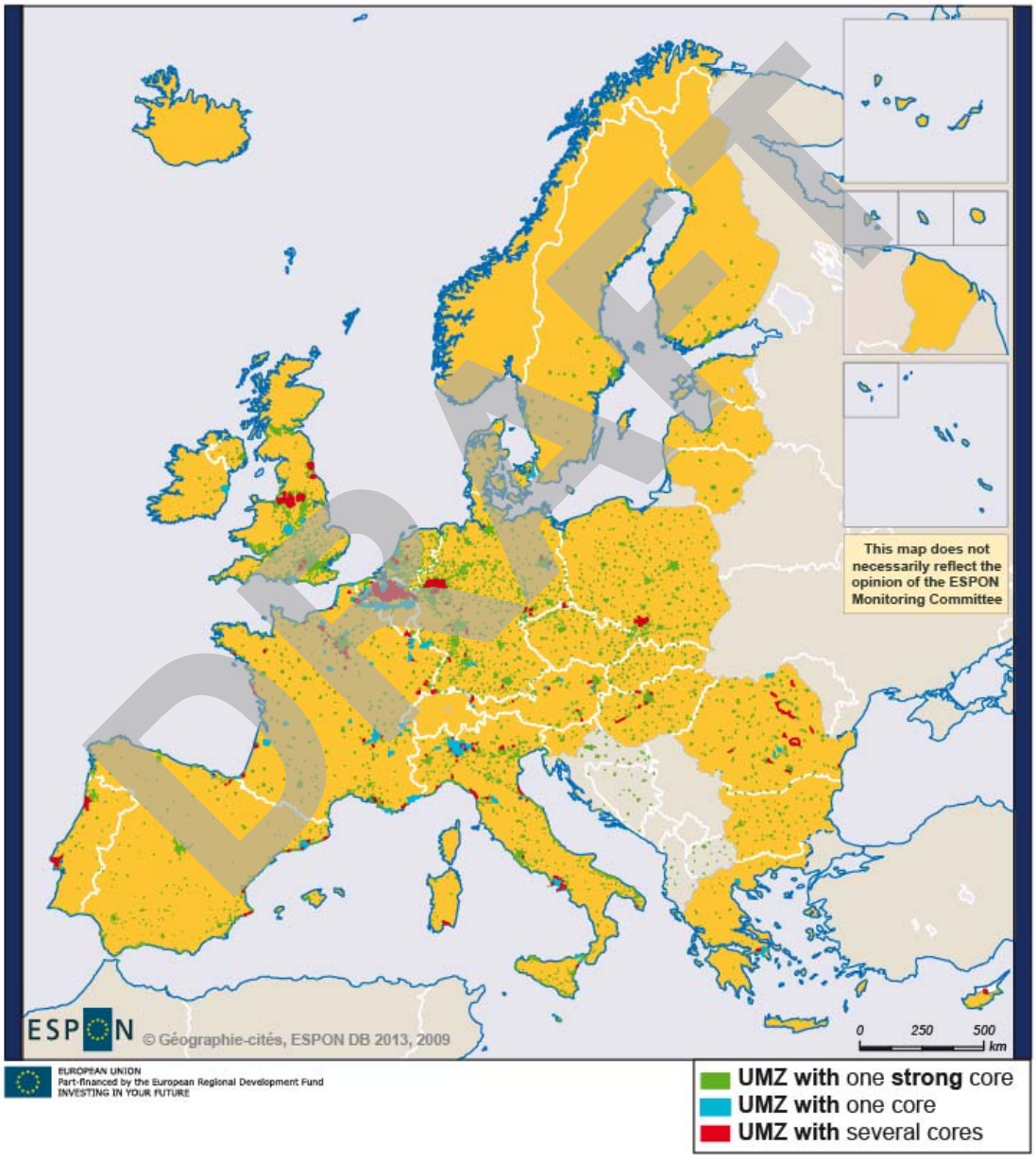


Figure 5: *UMZ typology according to naming results (“one strong core”, “several cores” and “one core”)*

4 Discussion and further developments

As mentioned above for some particular cases (see 3.1.1 and 3.1.2), the results presented here are not definitive and a further work is necessary to discuss and maybe improve the choices that have been made. For United Kingdom, Ireland and Portugal, LAU1 have been used but we could imagine another solution, selecting LAU2 in some cases and LAU1 in other[9]. We have begun a deeper work, using national sources and checking the names given to cities in national census. For other countries, we will check our results by comparing them to the list of national urban areas names, when they are available. For Cyprus and Balkan countries, we will look carefully at next versions of population density grid, in order to see if they cover these countries. If it is the case, we will try to apply the same methodology presented in this report.

The validation process of the naming method is still in progress, but the first results presented here are very promising. In particular, they illustrate the interest of a strong interaction between database formalization and thematic approaches. This intersection allows the elaboration of efficient methods (based on automatic processes coupled with expertise) but opens also new fields for urban research (see the map of UMZ's morphology according to administrative units), which interest ESPON partners not only for scientific considerations but also for applications in urban planning.

[9] For this deeper expertise, we will use the results of national census : in Portugal, the map of *cidades* available at http://sig.ine.pt, and for Great-Britain, the *census output area boundaries* that will be required at http://www.statistics.gov.uk/census2001. The first investigations show that this work will be relatively quick for Portugal and Ireland, but longer for Great-Britain.

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

$$\left(N\left(z(\mathbf{x}_i)-m_l\right)^2\right)\Big/\left((N+1)\bar{s}_l^{\,2}\right)\geq \chi^2_{crit-1} \qquad (1)$$

Here, $i=1,\ldots,n$; $\mathbf{x}$ 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 χ^2_{crit-1} 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 $z(\mathbf{x}_i) - m_l$. 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. $=[(1)/(3)]\times 1000$	2000	Euros
RATIO (6)	GDP_POP_2005_2006	GDP per inhabit. $=[(2)/(4)]\times 1000$	2005	Euros
RATIO	EVOGDP_2000_2005_2006	Evolution of GDP $=[(6)/(5)]\times 100$	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.

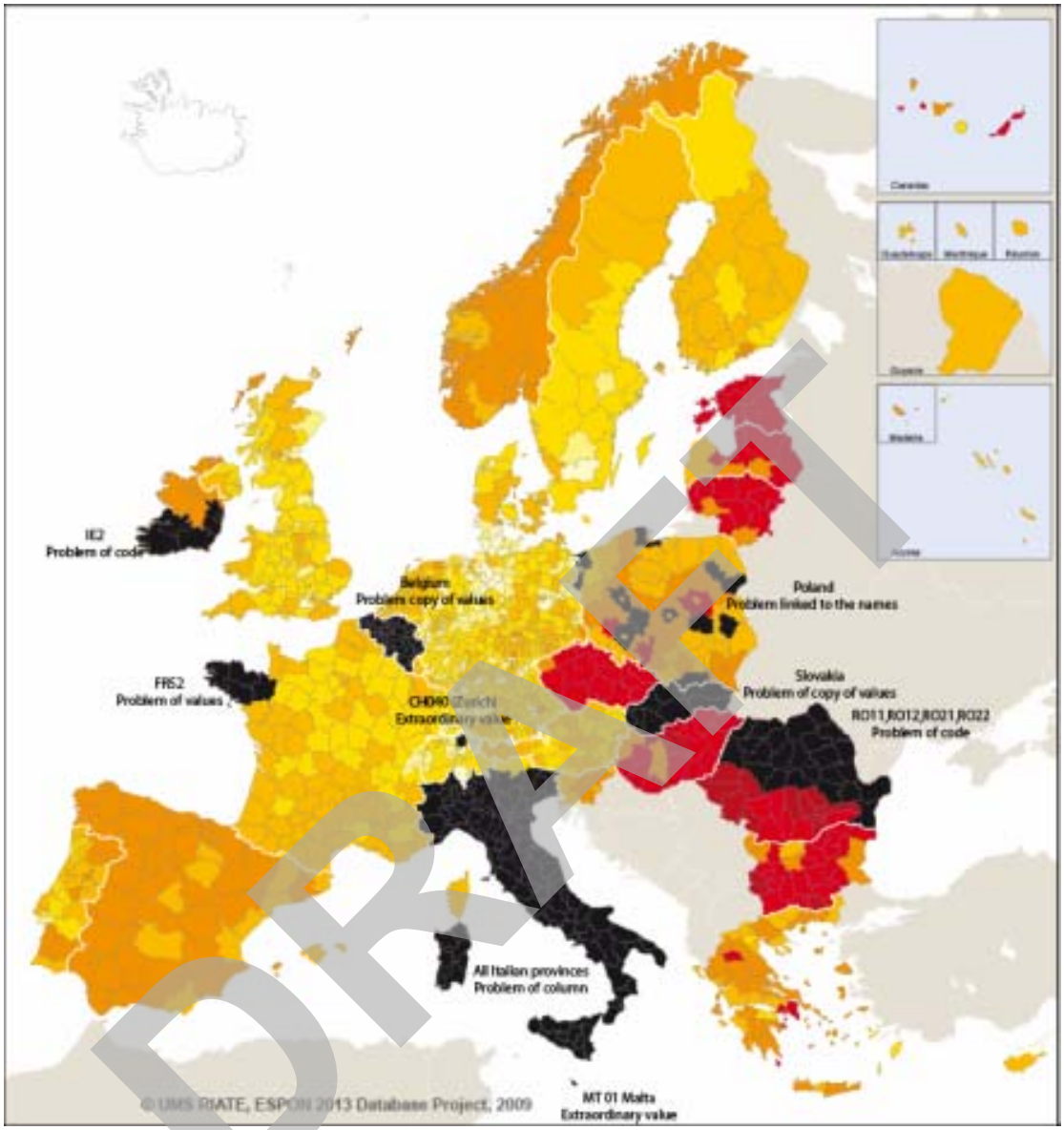


***Figure 1**: Location of input errors (in black) overlaid on the true EVOGDP_2000_2005_2006 data (see Fig. 2)*

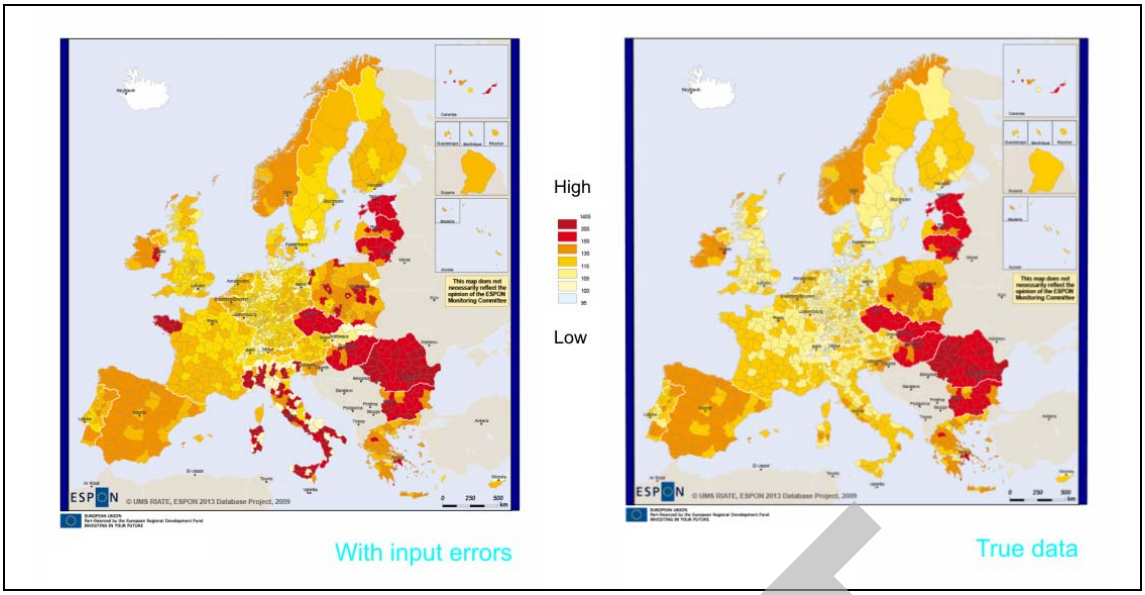


***Figure 2**: Maps of EVOGDP_2000_2005_2006 with and without input errors*

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.

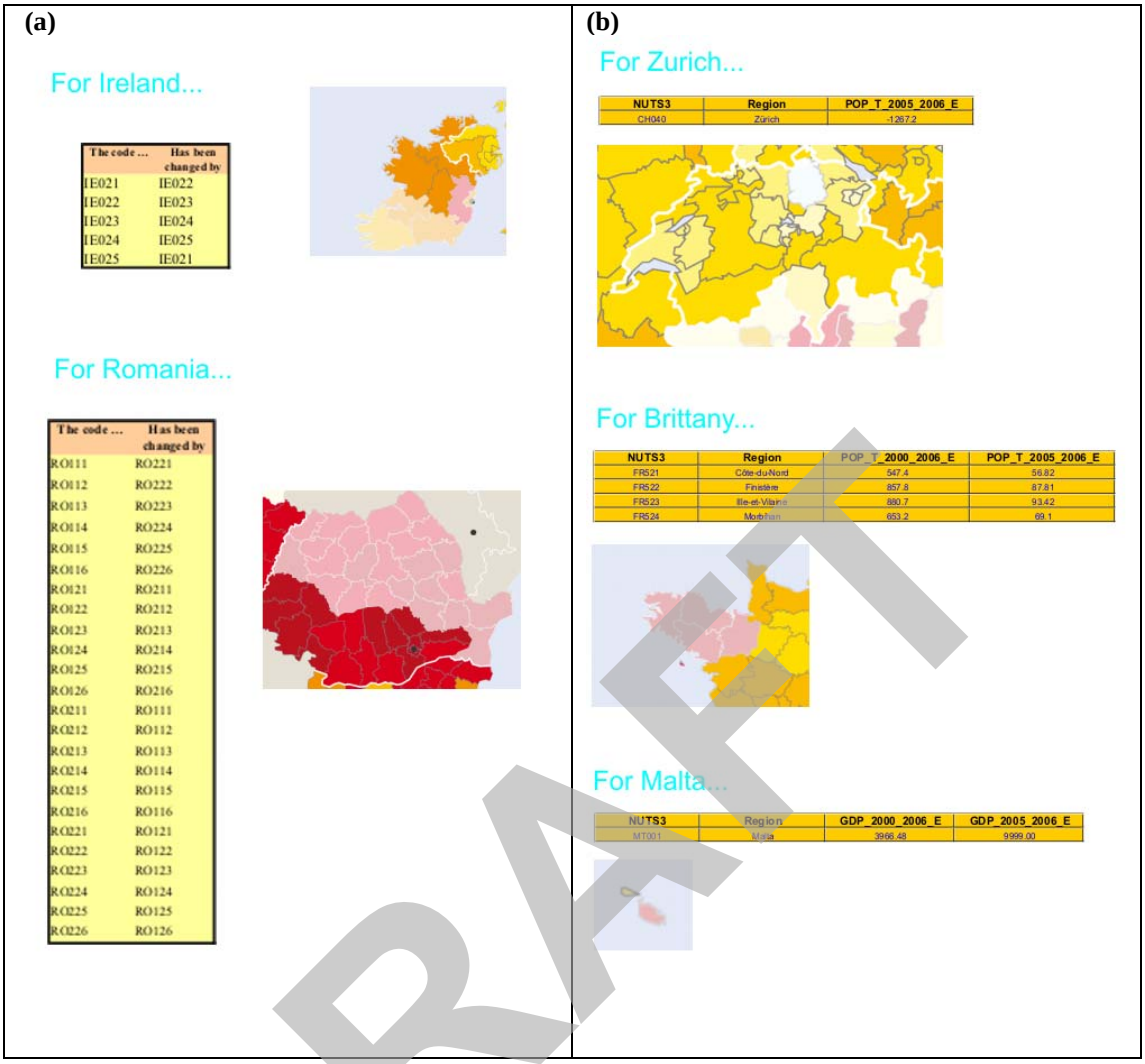


Figure 3 (a): *incorrect NUTS code entries (b) incorrect value entries*

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

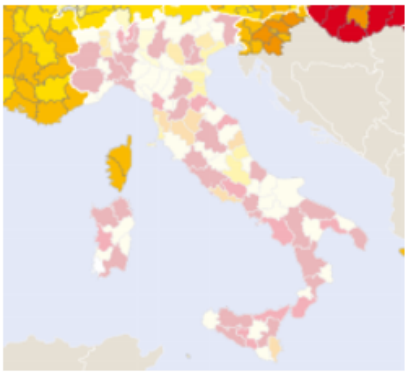
Figure 4: *Problems of copied or repeated data*

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...

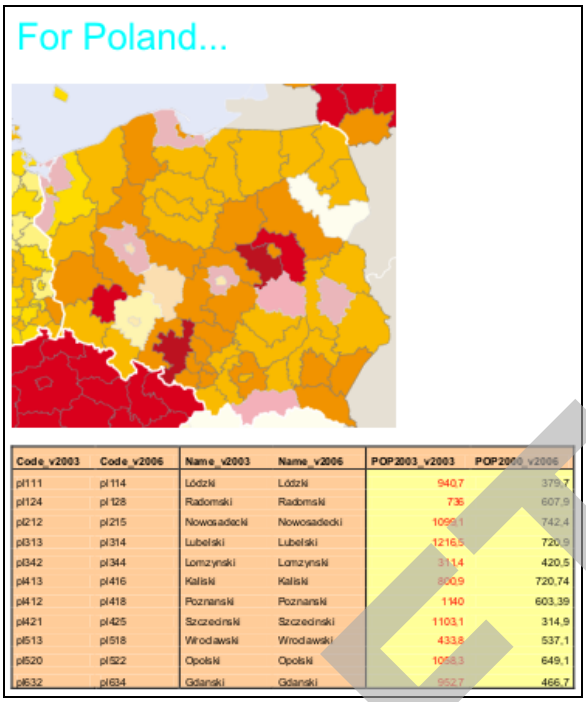


NUTS3	Region	GDP_2000_2006	GDP_2005_2006	POP_T_2000_2006	POP_T_2005_2006
ITC11	Torino	52029.49531	61725.00907	2172.6	177.2
ITC12	Vercelli	3730.69389	4584.692785	177.7	187.9
ITC13	Biella	4385.872603	4731.489232	187.7	161.6
ITC14	Verbano-Cusio-Ossc	3046.651604	3481.234948	159.3	354.5
ITC15	Novara	8224.855319	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	POP2006_v2006
pl111	pl114	Łódzki	Łódzki	940,7	379,7
pl124	pl128	Radomski	Radomski	735	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/

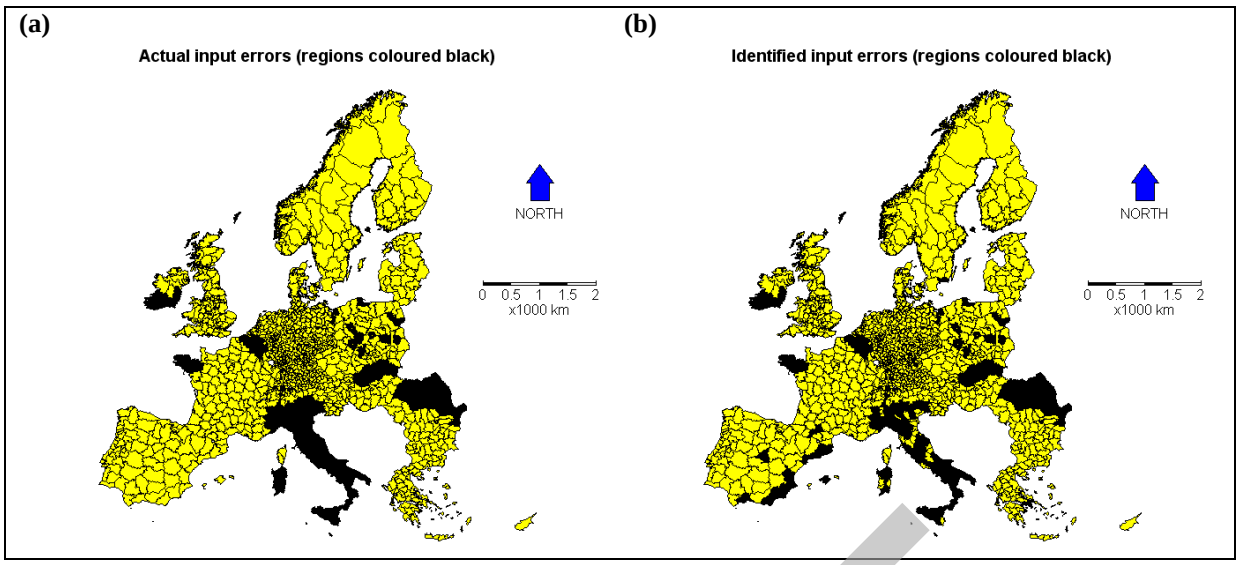


***Figure 7**: Location of (a) true (deliberate) input errors versus (b) identified input errors. Rates of false negatives, false positives and overall misclassification are 13.2%, 2.0% and 3.7% respectively*

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.

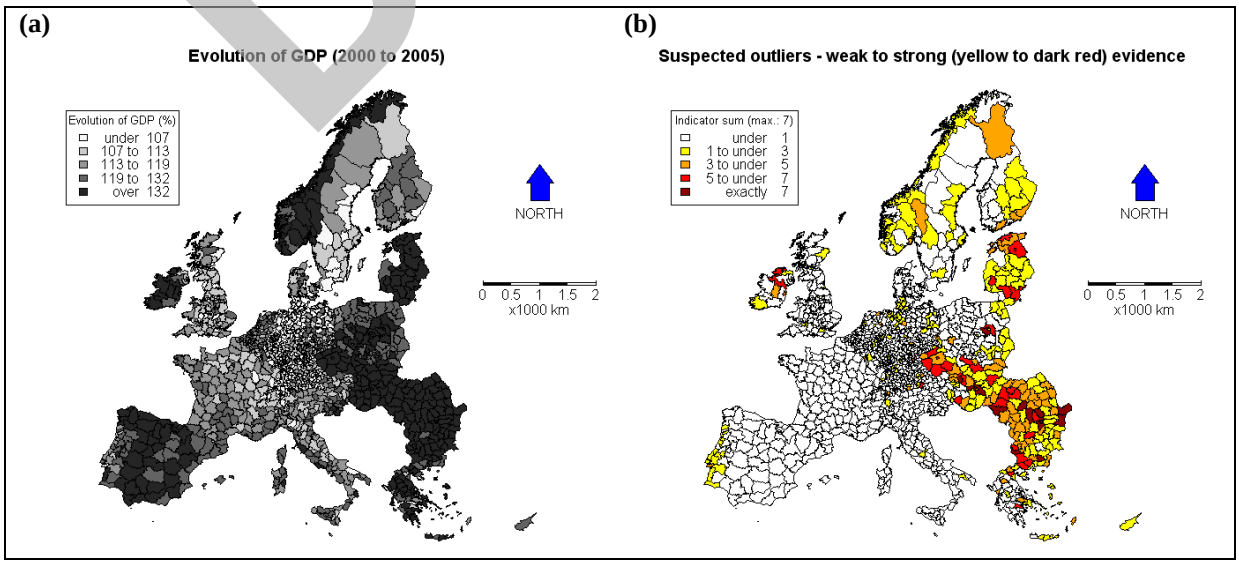


***Figure 8**: Spatial distribution of (a) EVOGDP_2000_2005_2006 and (b) suspected outliers for EVOGDP_2000_2005_2006 (seven univariate indicators)*

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.

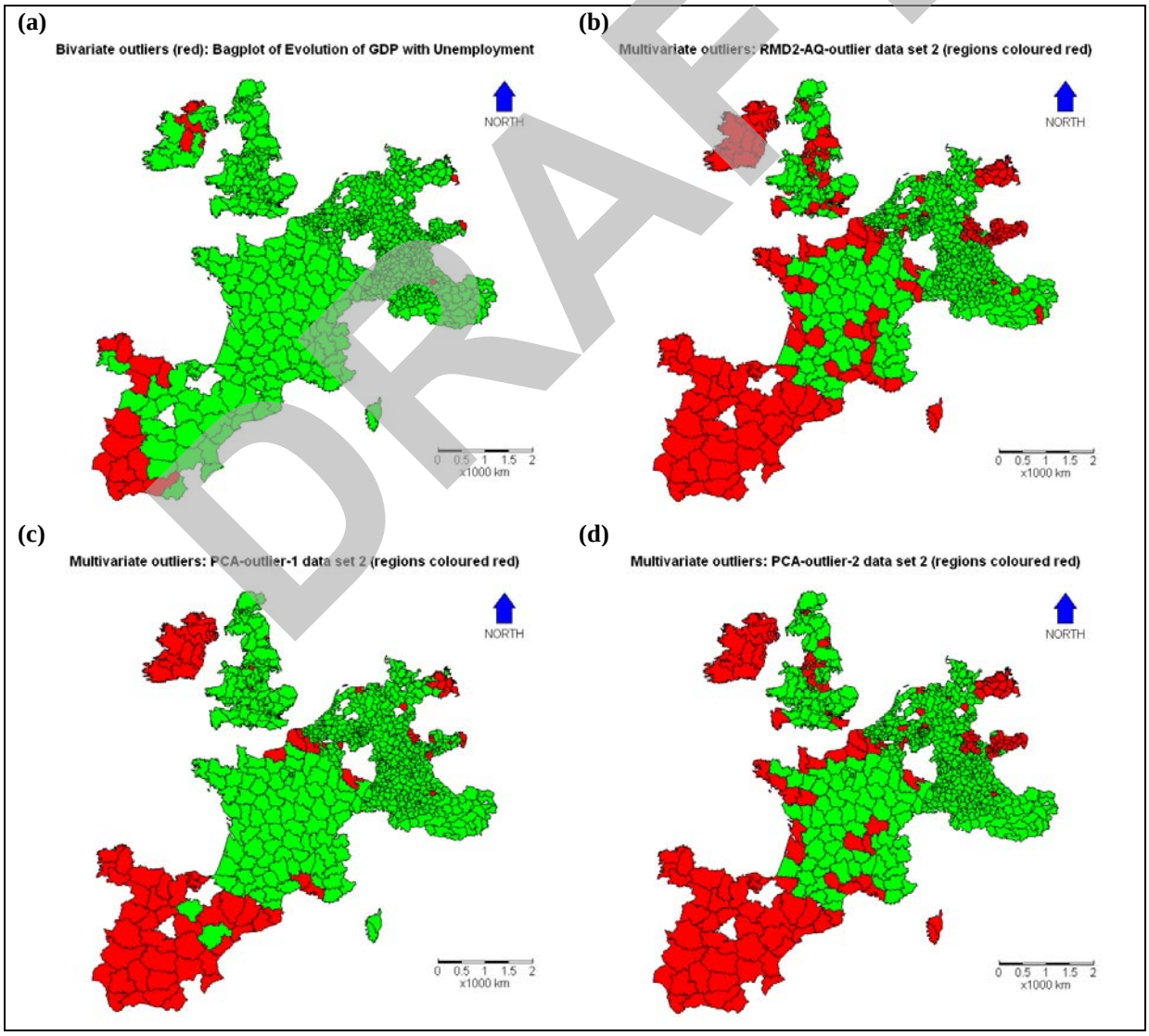


***Figure 9**: Spatial distribution of suspected (bivariate and multivariate) outliers according to (a) a bagplot of EVOGDP_2000_2005_2006 with UNEMP_R_2001_1999; (b) the RMD2-AQ-outlier technique; (c) the PCA-outlier-1 technique; and (d) the PCA-outlier-2 technique*

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.

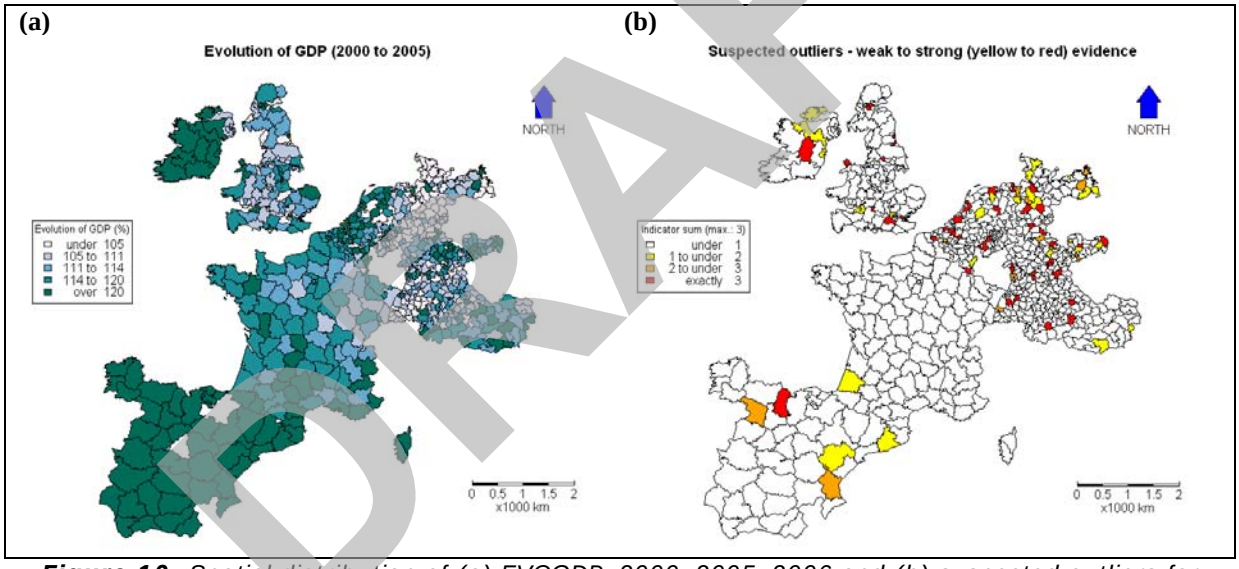


***Figure 10**: Spatial distribution of (a) EVOGDP_2000_2005_2006 and (b) suspected outliers for EVOGDP_2000_2005_2006 (three multivariate indicators).*

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.

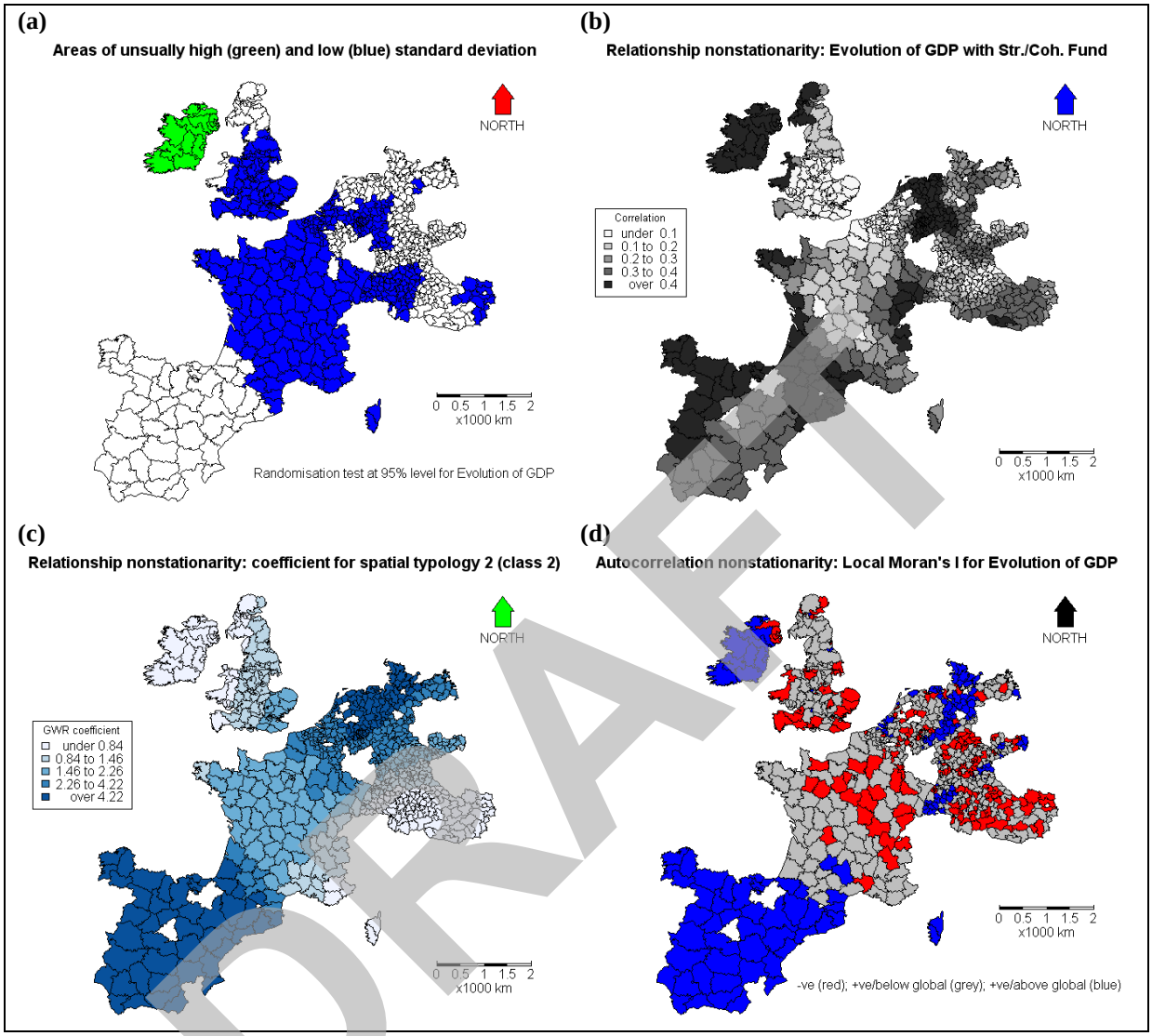


***Figure 11**: Identification of 'unusual' clusters in EVOGDP_2000_2005_2006 with respect to (a) its local variability (using GW standard deviations); (b) its local relationship to SF_CF_1999_1999 (via a GW correlation analysis); (c) its local relationship to class 2 of SPAT_TYPE_2_1999_1999 (via a 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.

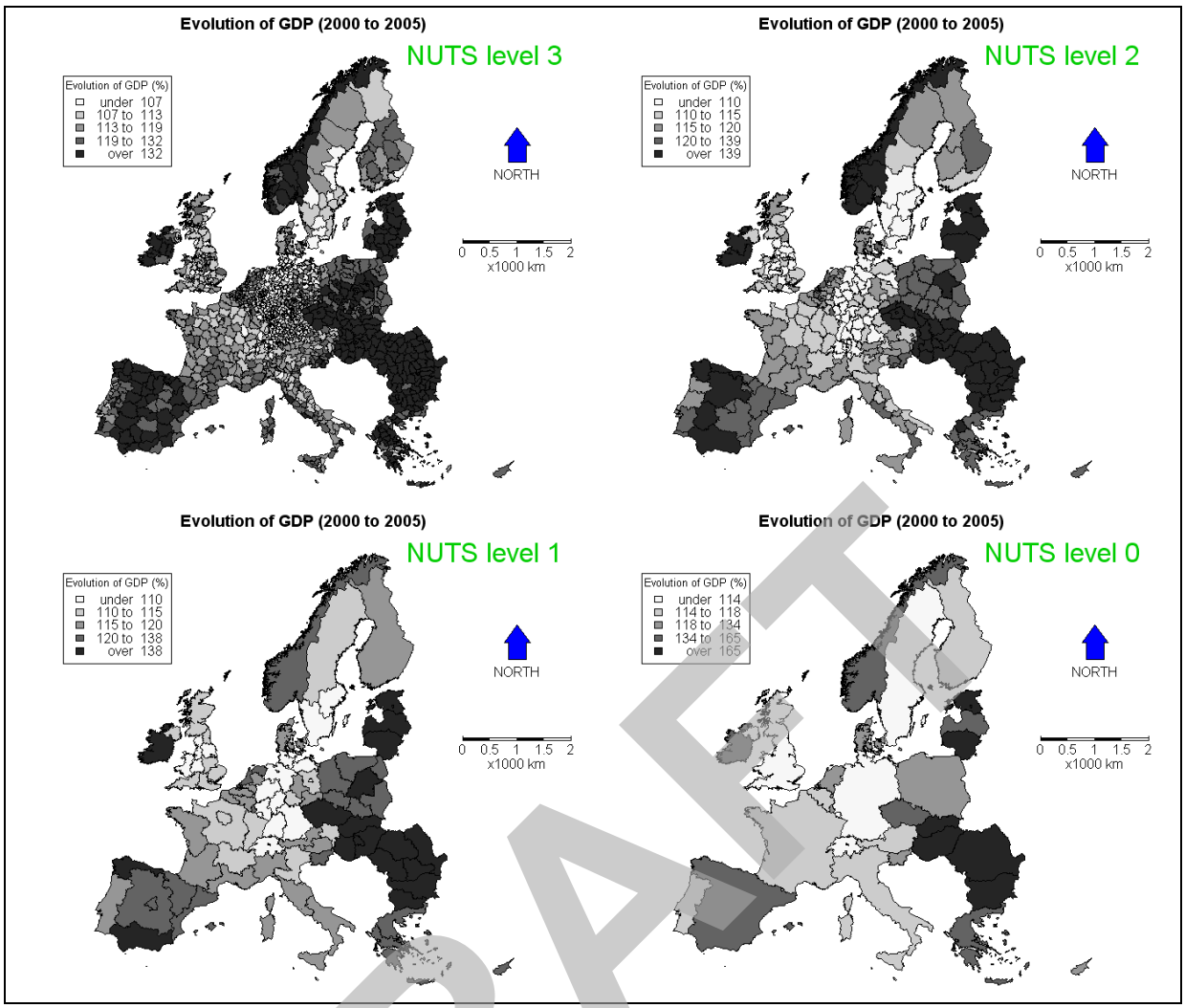


***Figure 12**: Spatial distribution of EVOGDP_2000_2005_2006 at four different NUTS levels (3, 2, 1 and 0)*

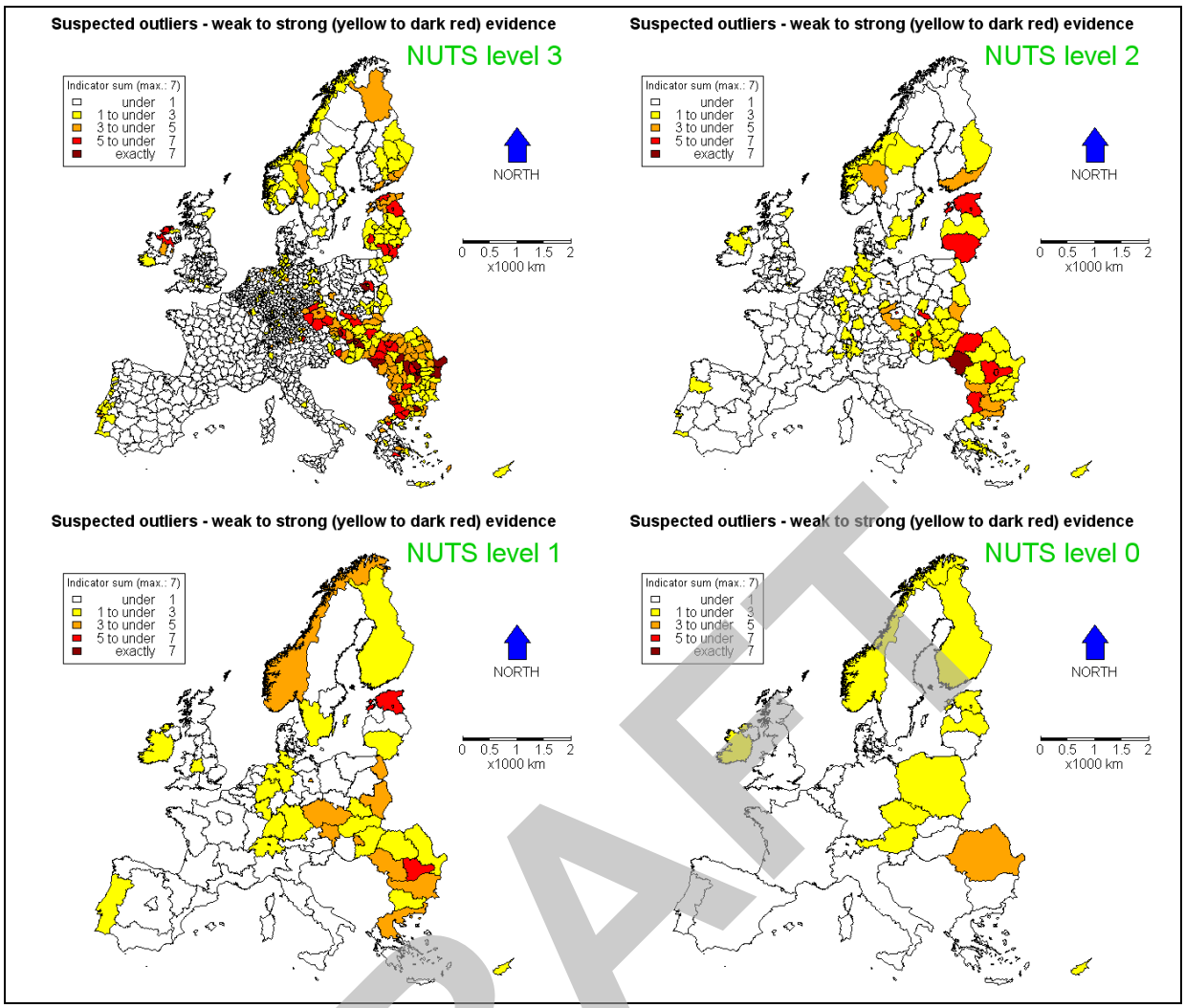


***Figure 13**: Spatial distribution of suspected outliers for EVOGDP_2000_2005_2006 (via seven univariate indicators), where outliers are identified at four NUTS levels (3, 2, 1 and 0)*

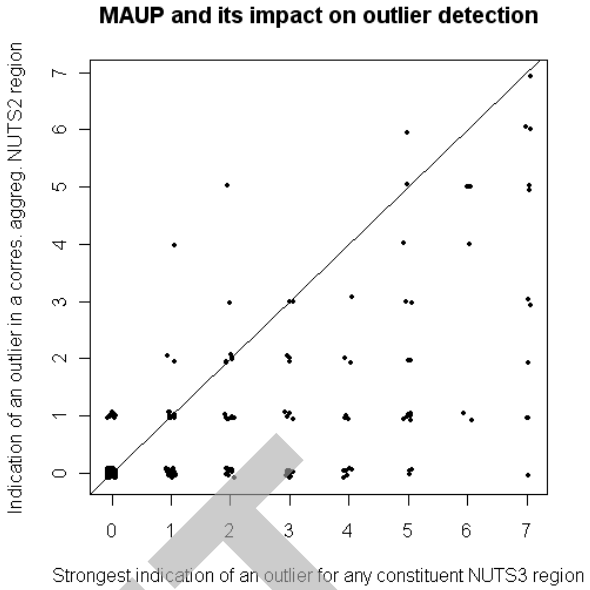
NUTS3 to NUTS2

Correlations:

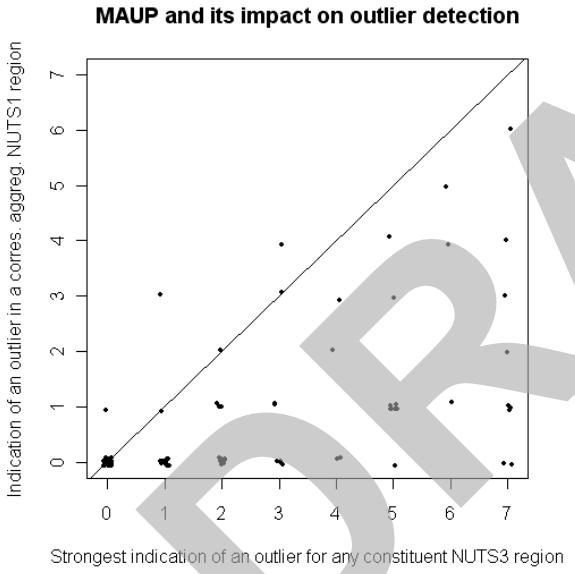
(b) NUTS3 to NUTS2: 0.67

(c) NUTS3 to NUTS2: 0.59

(d) NUTS3 to NUTS0: 023



NUTS3 to NUTS1



NUTS3 to NUTS0

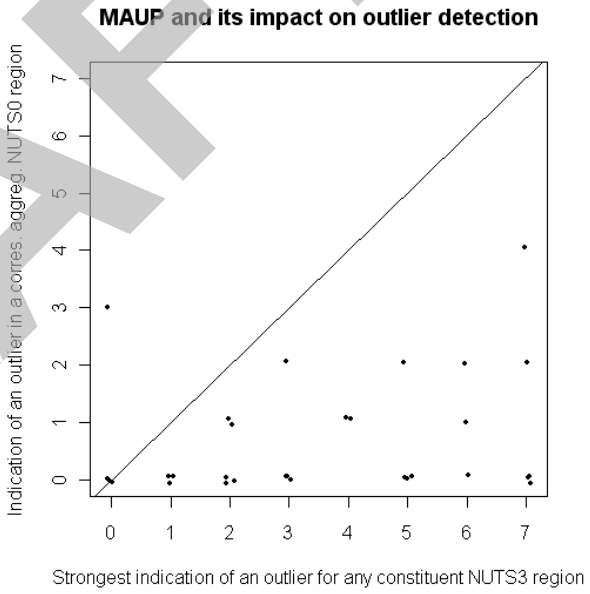


***Figure 14**: Scatterplots and correlations, where the "Strongest indication of an outlier for any constituent NUTS level 3 region" is related to the "Indication of an outlier in a corresponding aggregated NUTS level 2/1/0 region". Scatterplots are jittered to aide interpretation*

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

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.

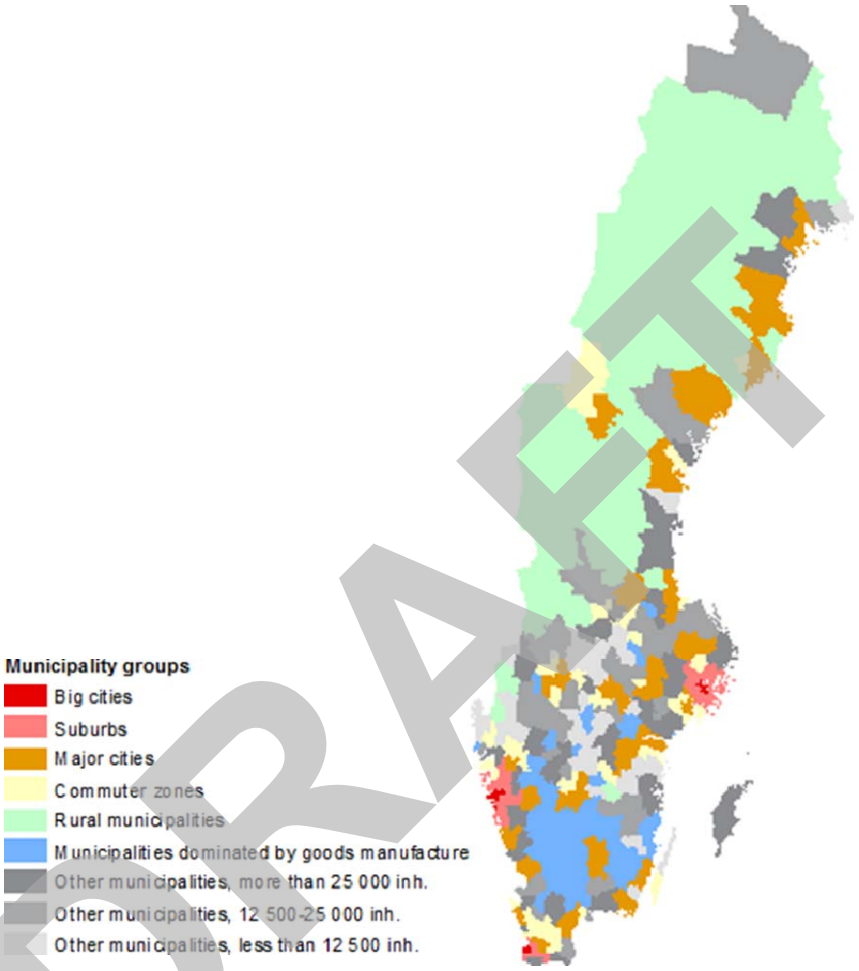


Figure 1: *Municipality group map*

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.

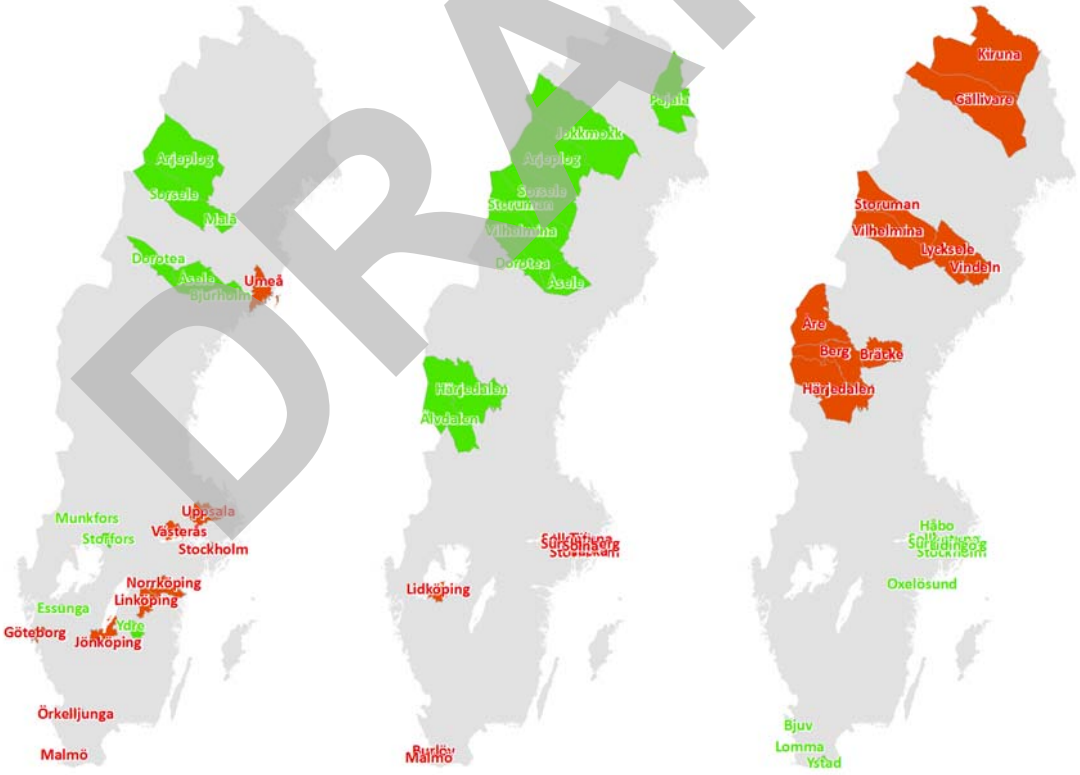


***Figure 2**: The ten municipalities with highest (red) and lowest (green) absolute residuals summarized (left) and in relation to area (middle) and population size (right)*

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.

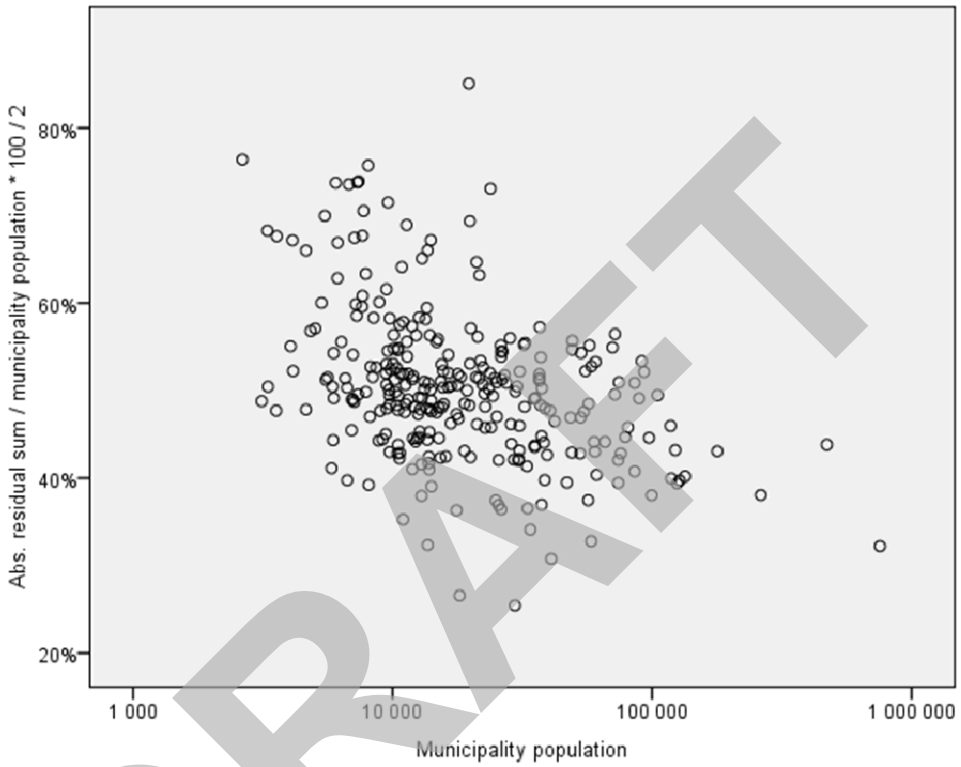


***Figure 3**: Municipality population size compared to absolute residual sum in relation to municipality population size*

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.

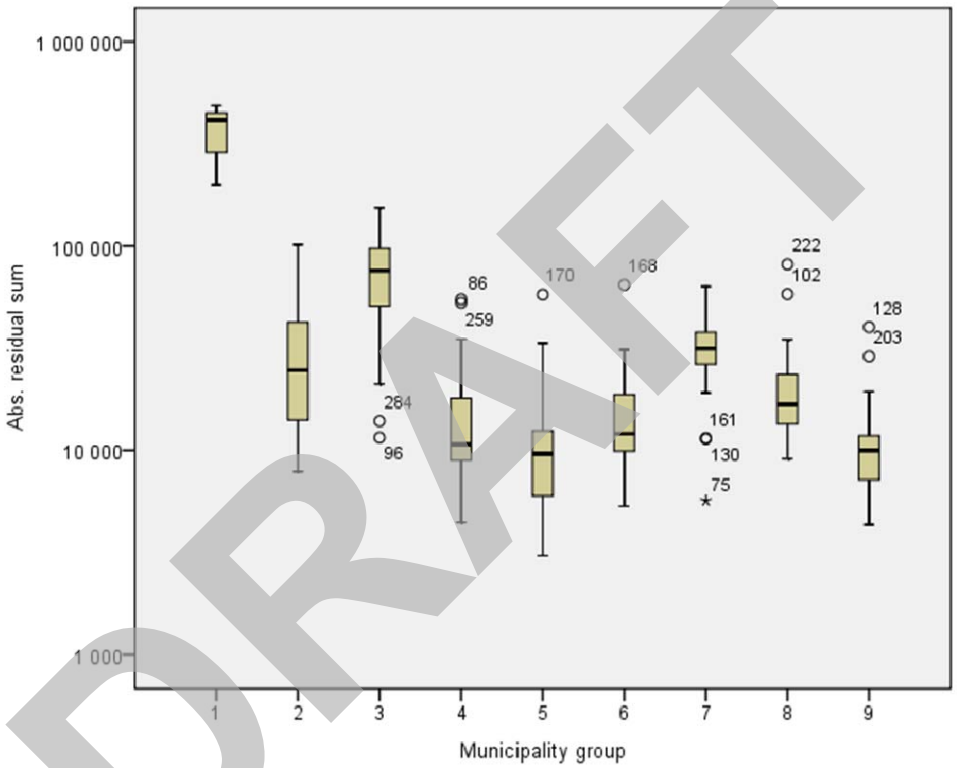


***Figure 4**: Absolute municipality residual sum subdivided by municipality group*

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.

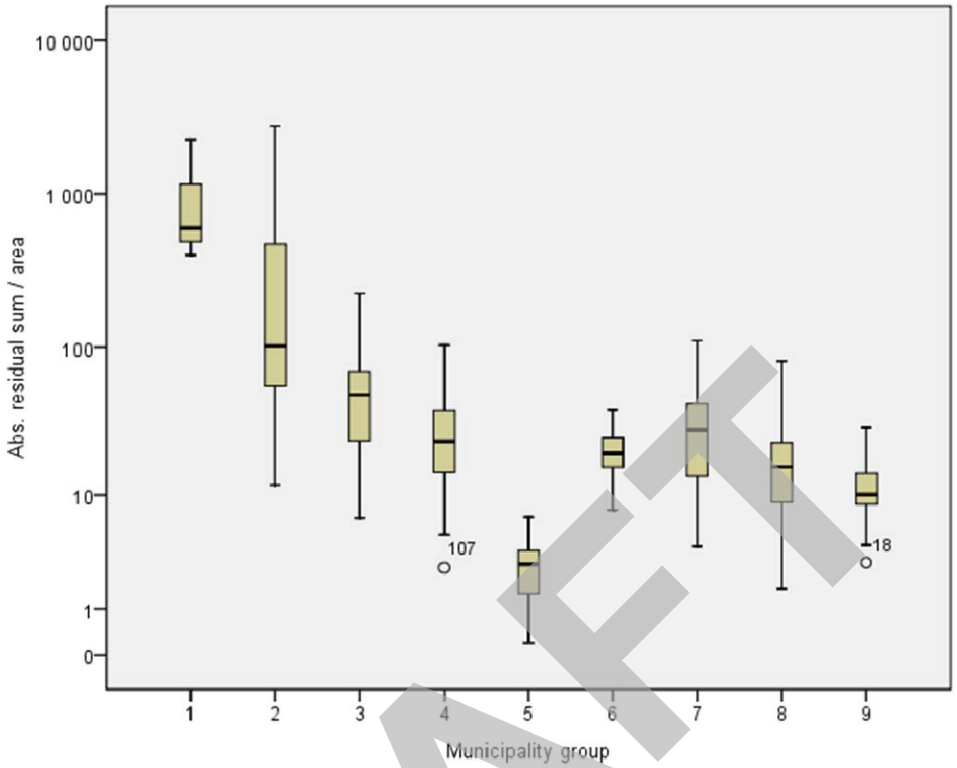


Figure 5 *: Absolute municipality residual sum in relation to area, subdivided by municipality group*

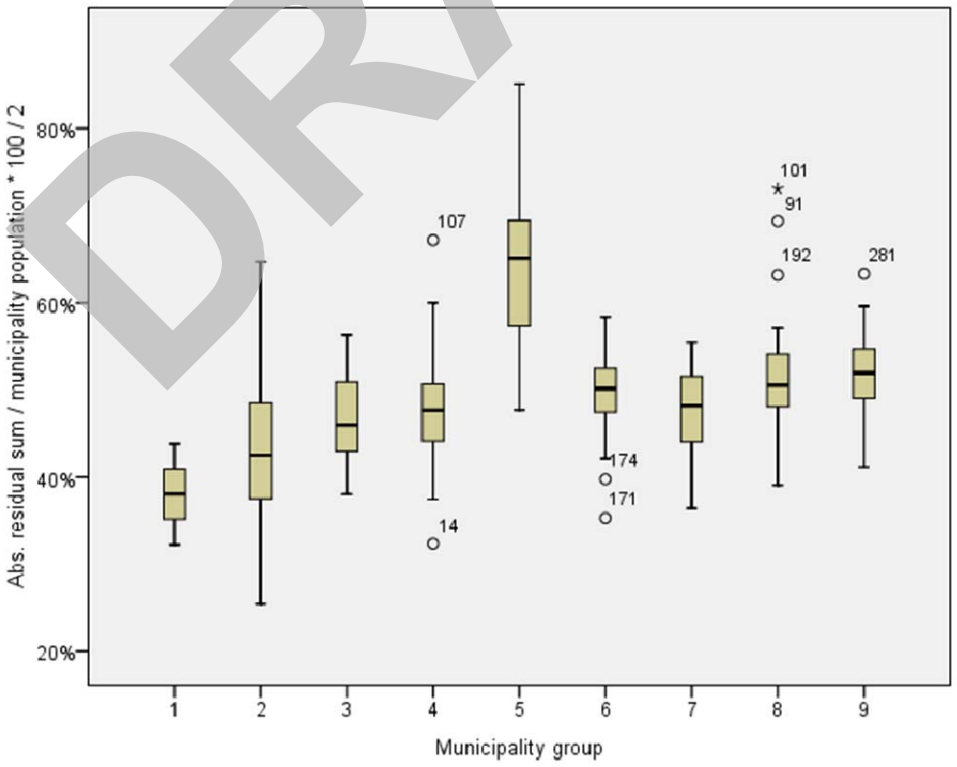


Figure 6*: Absolute municipality residual sum in relation to population size, subdivided by municipality group*

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).

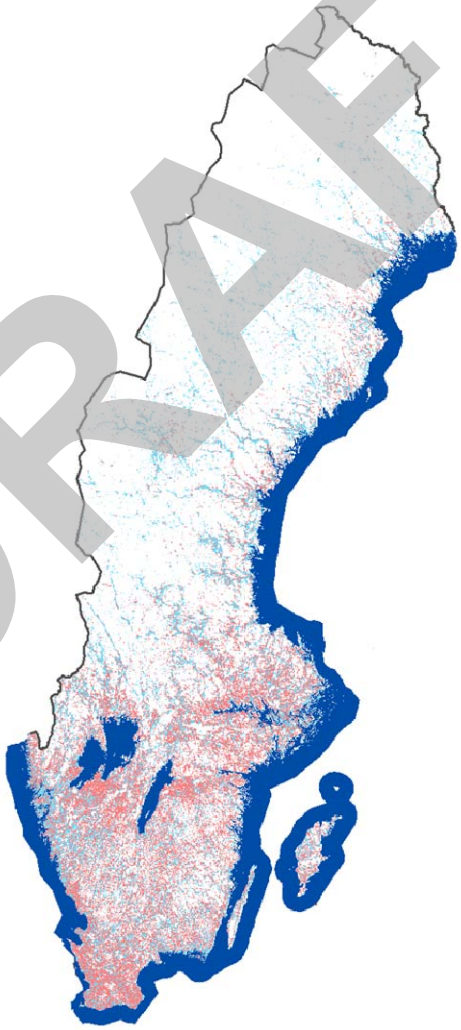


***Figure 7**: Absolute km^2 square residuals*

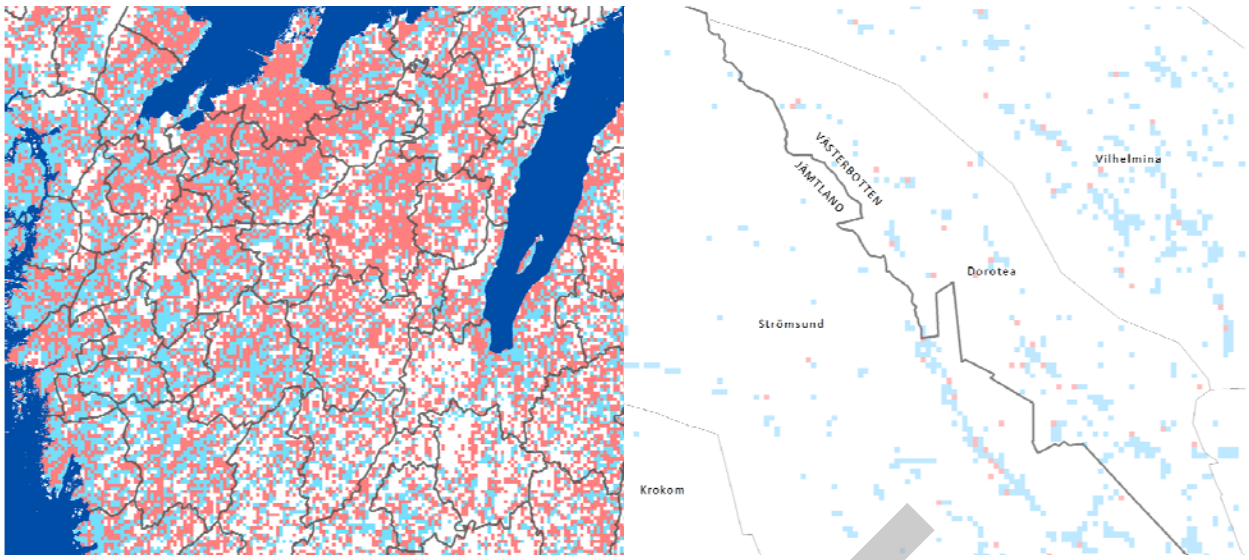


Figure 8: *Close-up of km^2 residuals in Southern (left) and Northern (right) Sweden*

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.

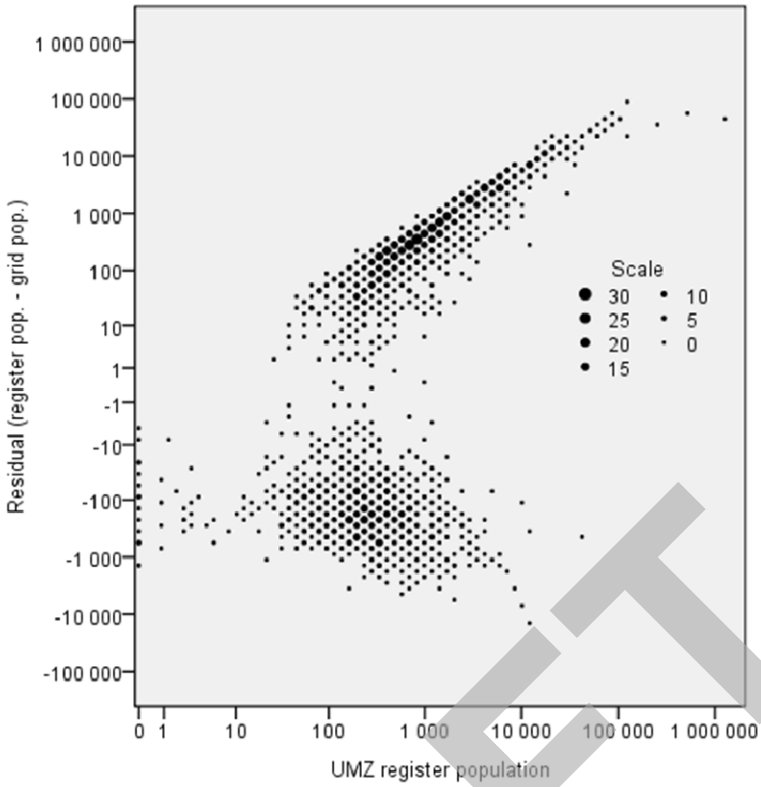


***Figure 9**: UMZ residuals by UMZ register population*

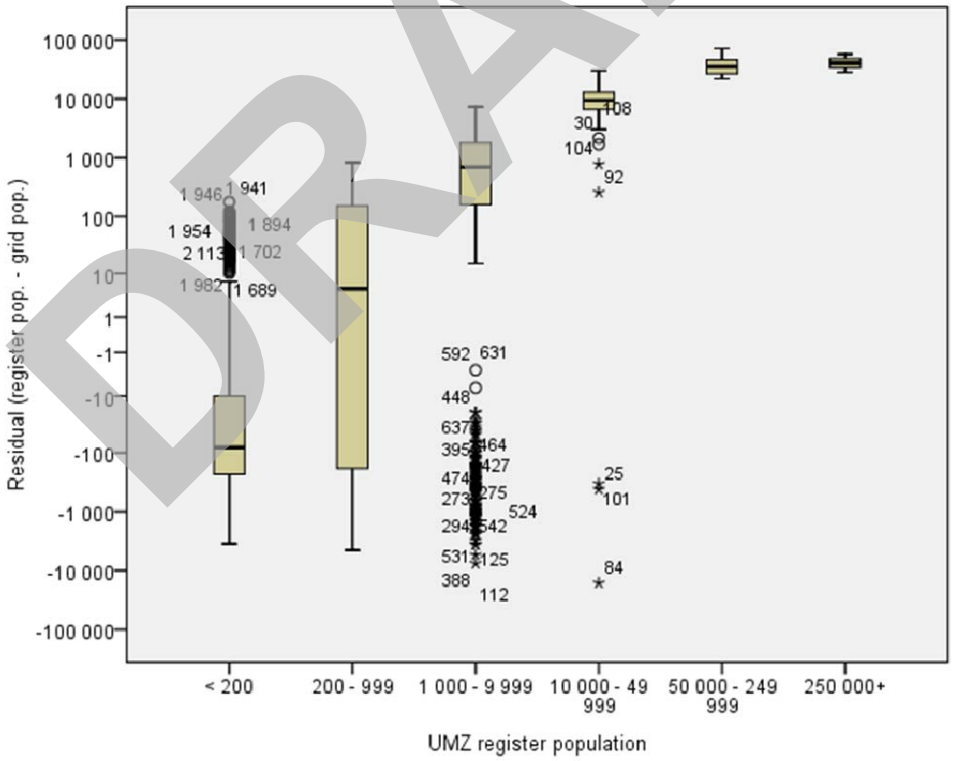


***Figure 10**: UMZ residuals by UMZ register population classes*

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.

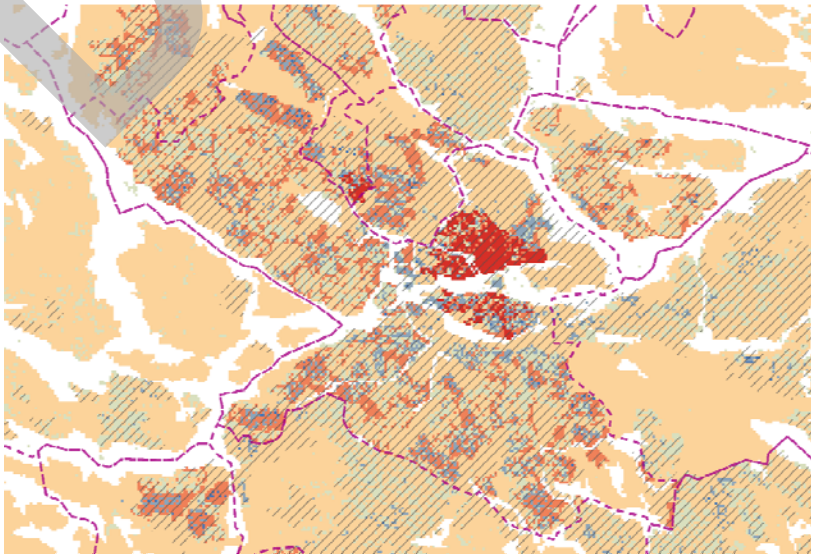


***Figure 11**: Close-up of 100 meter^2 residuals in UMZs in the Stockholm area*

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 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).

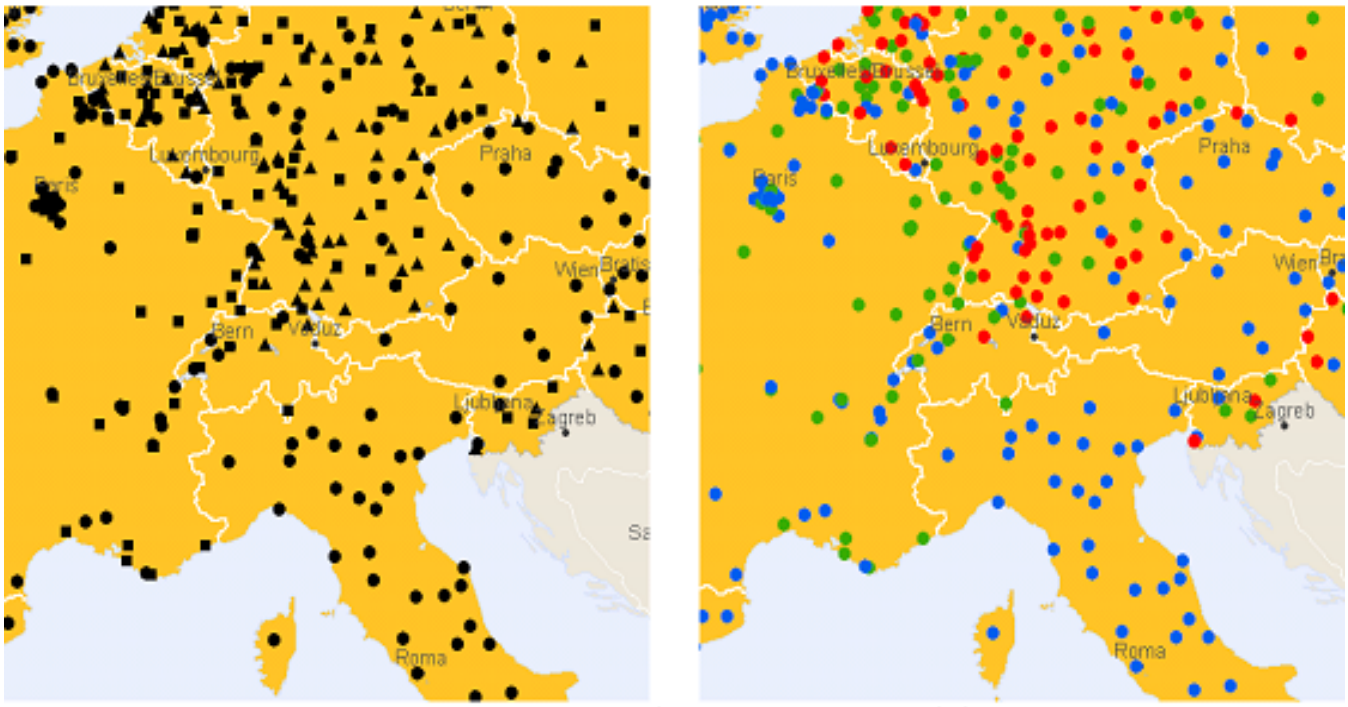


Figure 1 *– Universities by types – Two possibilities*
Good map = points + symbols or points + colours

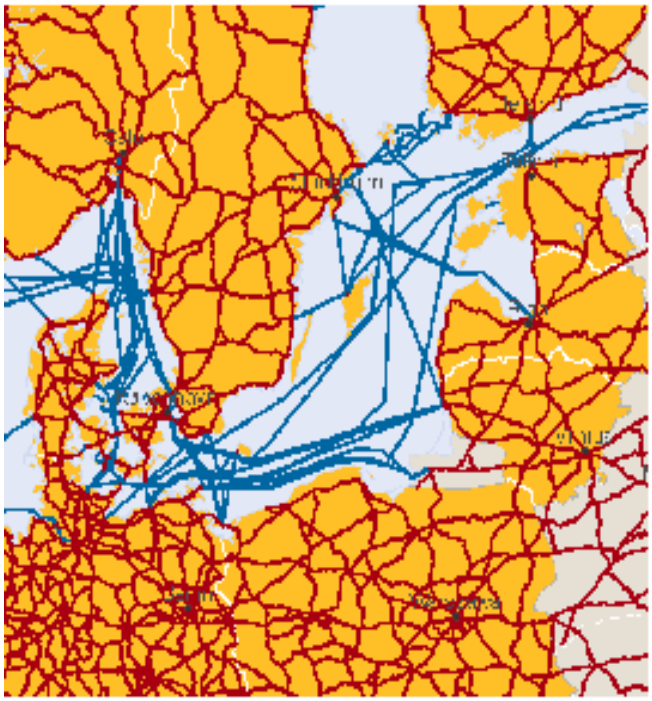


Figure 2 *– Mains roads and ferry connections*
Good map = line + colours

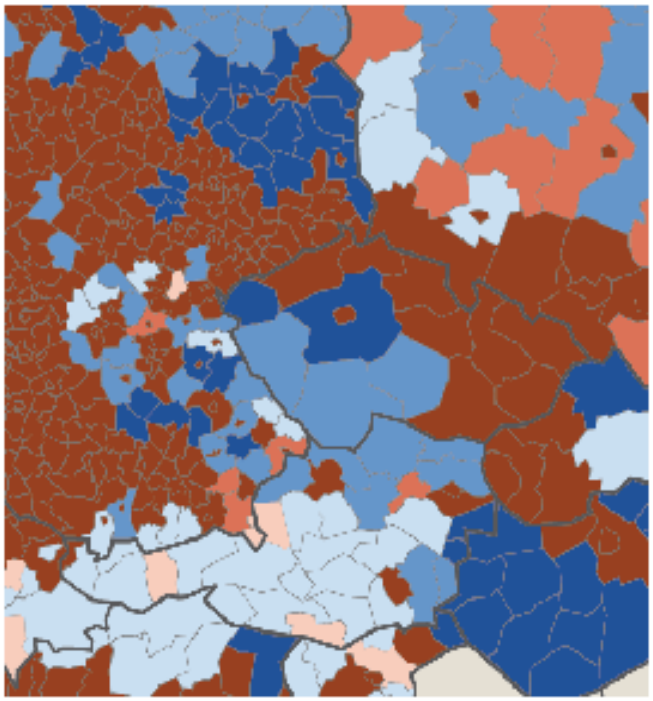


Figure 3 *– Results from Urban-Rural typology*
Good map = areas + colours

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.

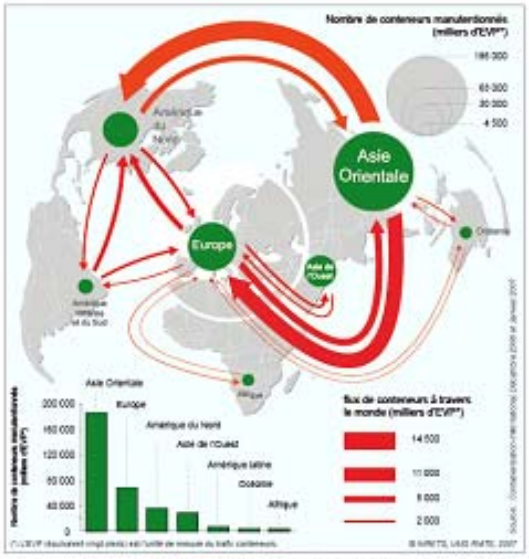


Figure 4 *- Containers flows across the World*
Good map = line + variation of line size

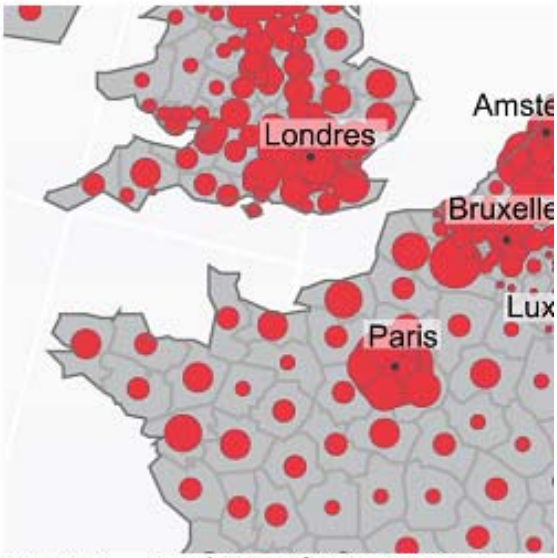


Figure 5 *– Population of NUTS 3 in Europe*
Good map = dot + proportional variation of size

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.

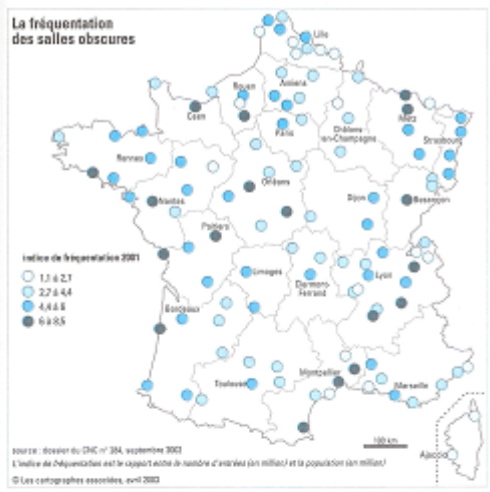


Figure 6: Cinema attendance Index in French main cities

Good map = dots + variation of colours

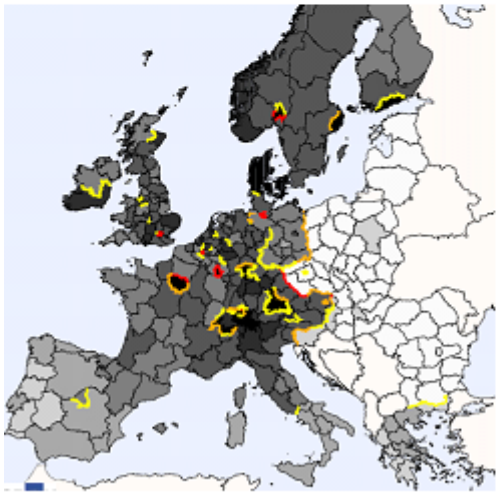


Figure 7: GDP per inhabitants discontinuities

Good map = lines + variation of colours

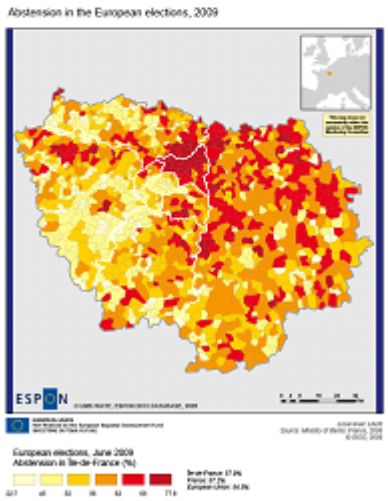


Figure 8: Abstention European votes in Île-de-France

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.

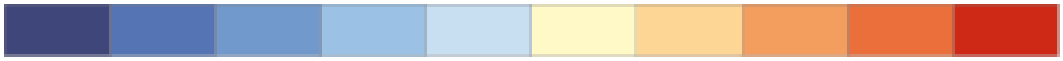
Figure 9: *Typology of Functional Urban Areas*
Good map: points + variation of colour (or size)

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:

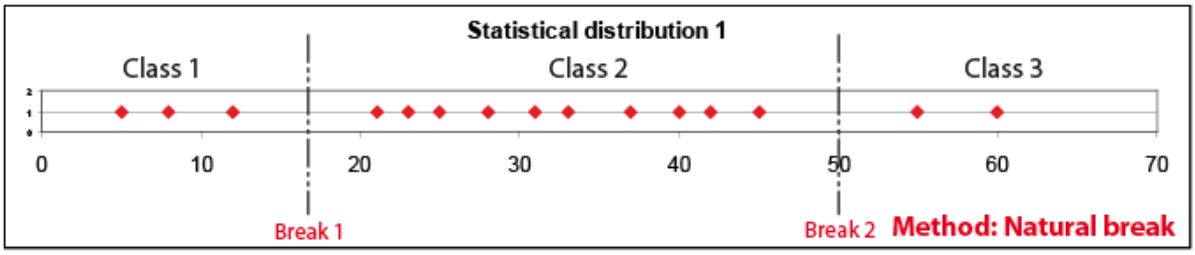
When combining red (warm colours) and blue (cold colours), **red is 'not good'/'negative'** and **blue is 'good'/'positive'"**

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.

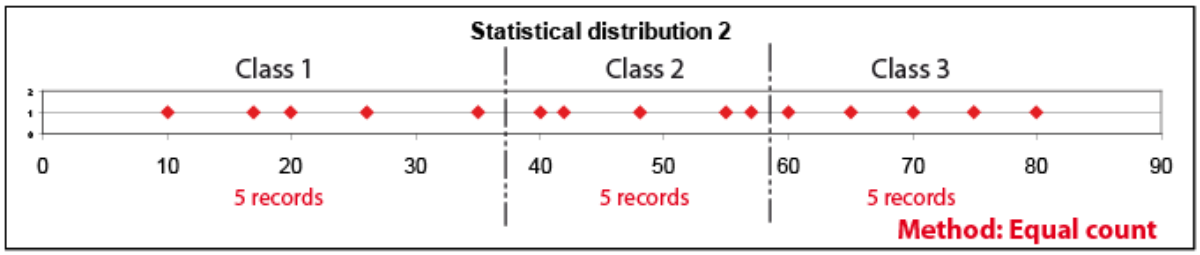
1.3.1 Natural Break

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.



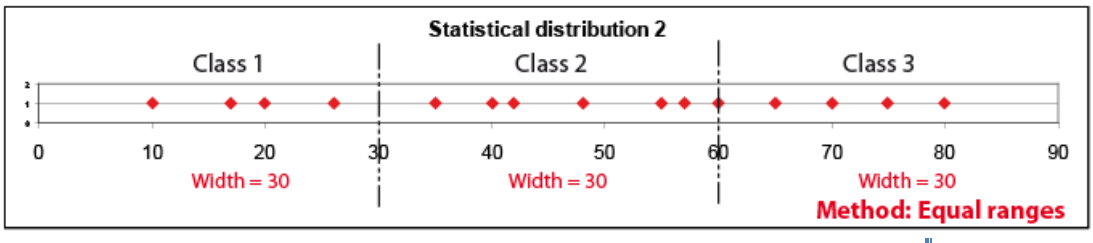
1.3.2 Equal Count or quantile

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.



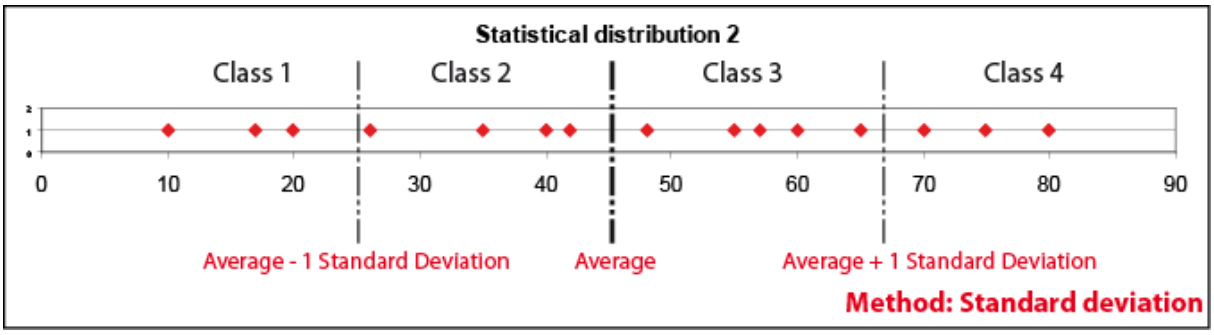
1.3.3 Equal Ranges

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

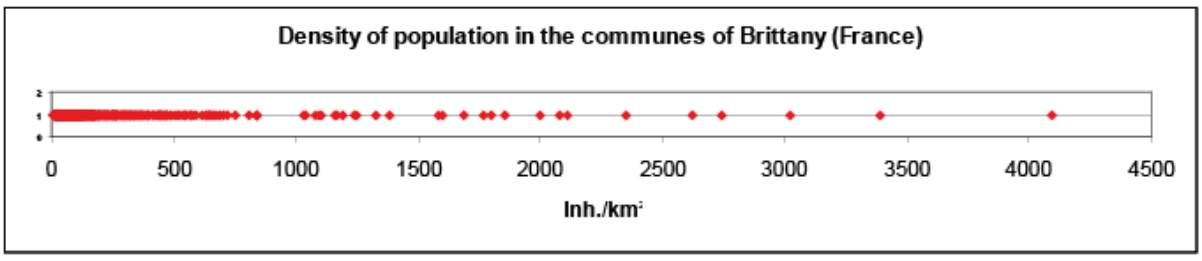
The widths of the class follow a geometric progression. To calculate the width of the different class, it is necessarily to estimate the geometric ratio, such as:

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.

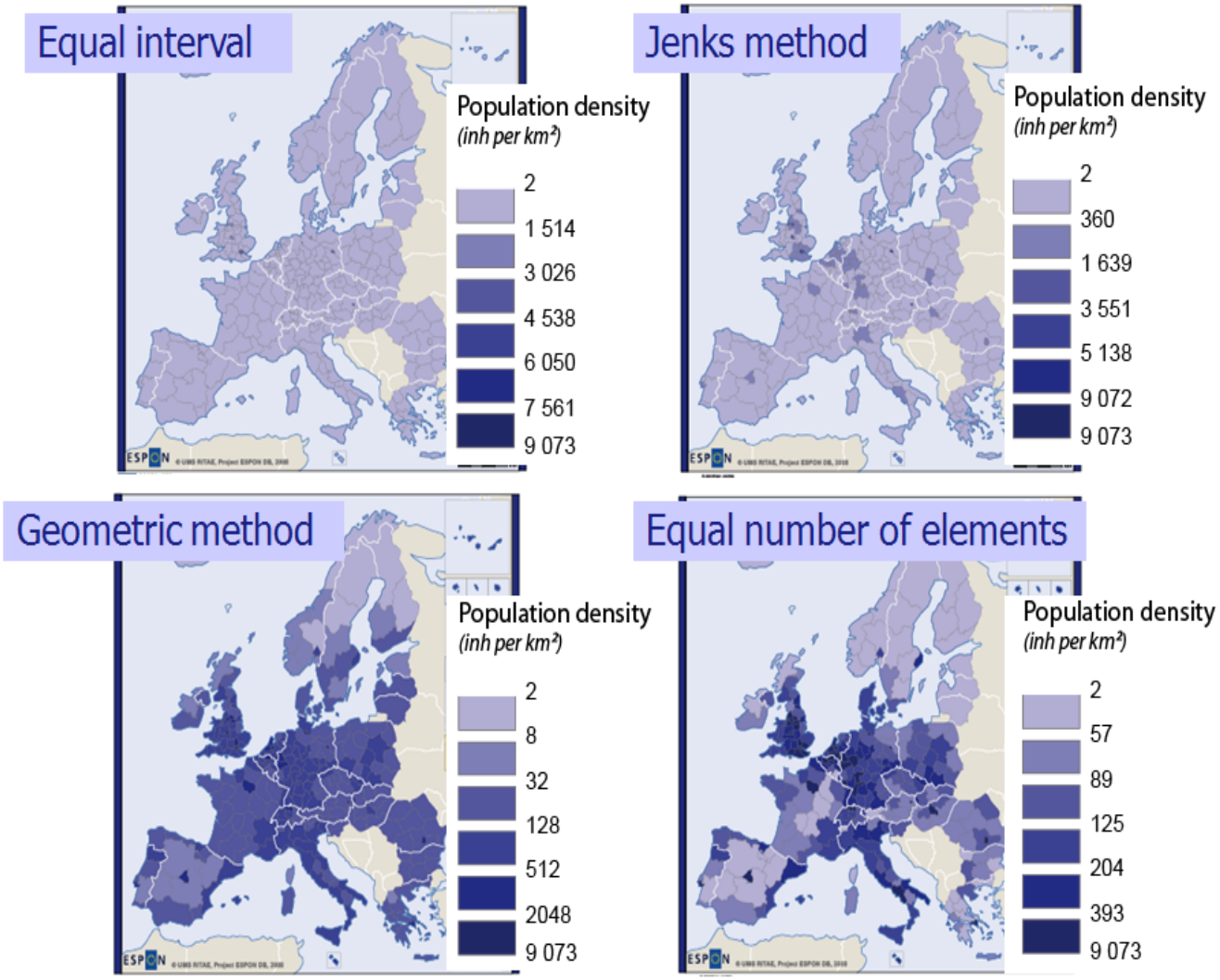


***Figure 12**: Result and efficiency are dependent upon the data classification method*

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.

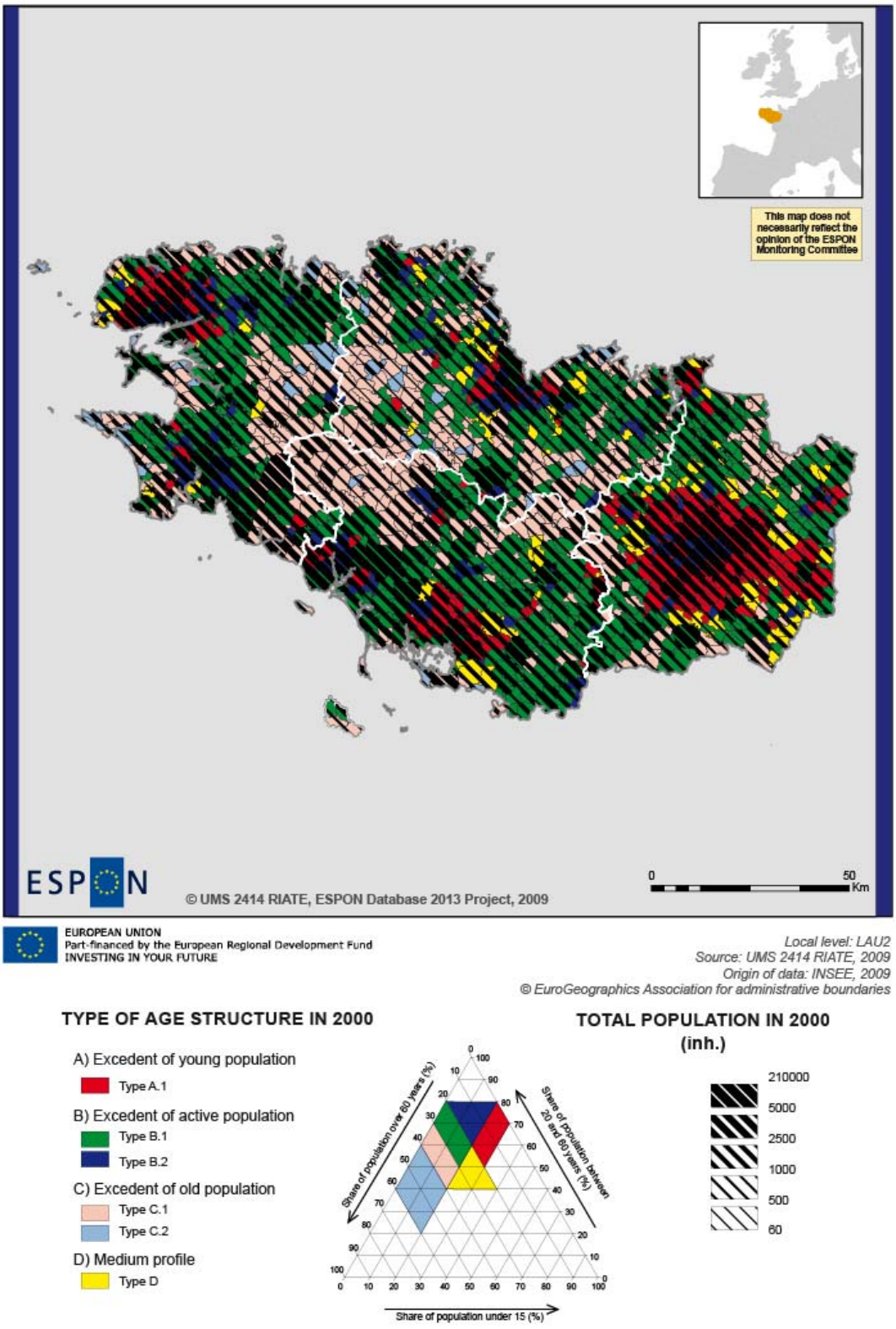


Figure 13**: Population and age structure in Brittany (France) – **with semiologic problems

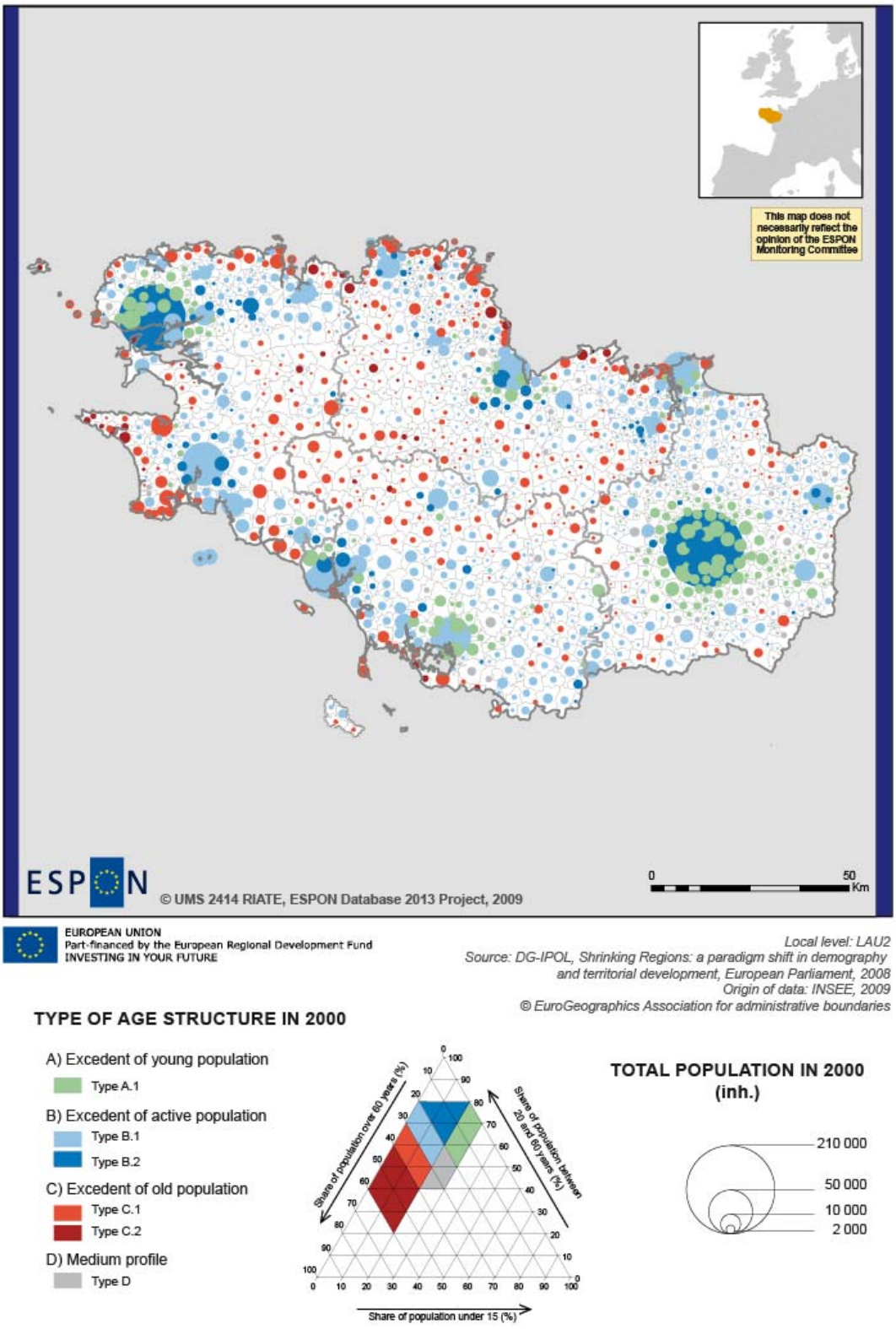


Figure 14**: Population and age structure in Brittany (France) – **without semiologic problems

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"!

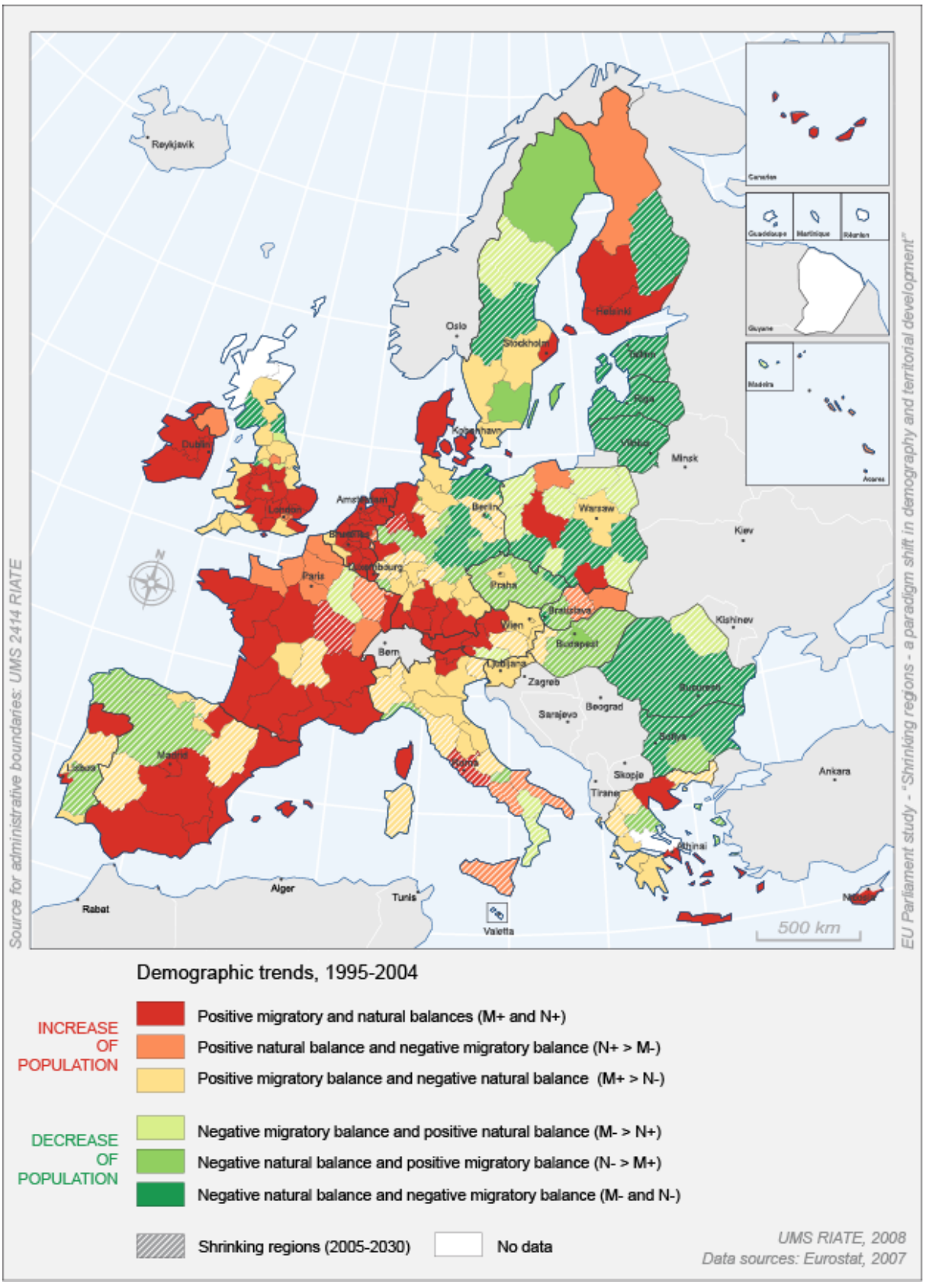


Figure 15**: Typology of regional growth patterns – **Possibility 1

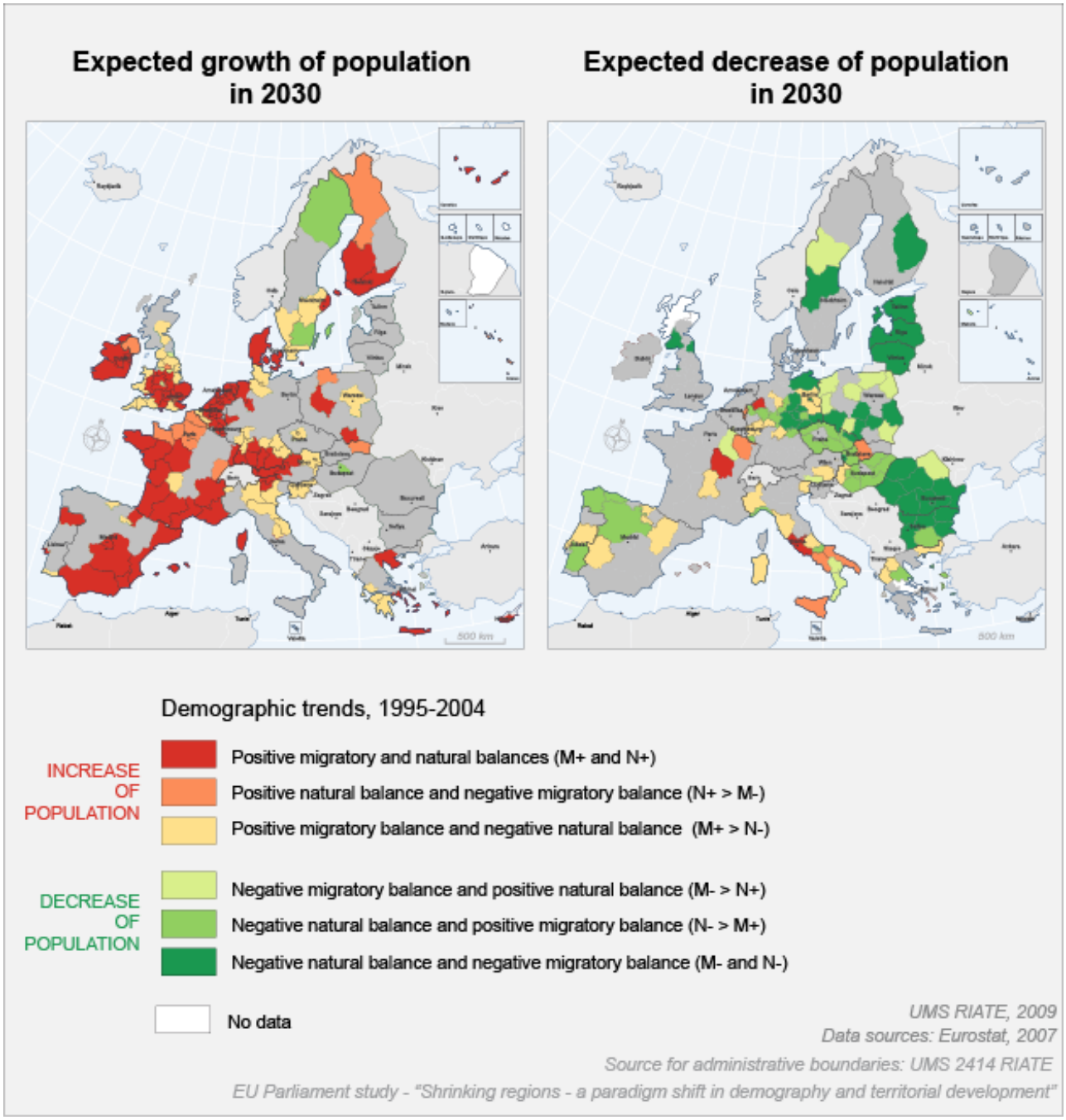


Figure 15**: Typology of regional growth patterns – **Possibility 2

ANNEXES

These annexes allow you to choose some efficient graphic variables to communicate differences in size, order or quality.

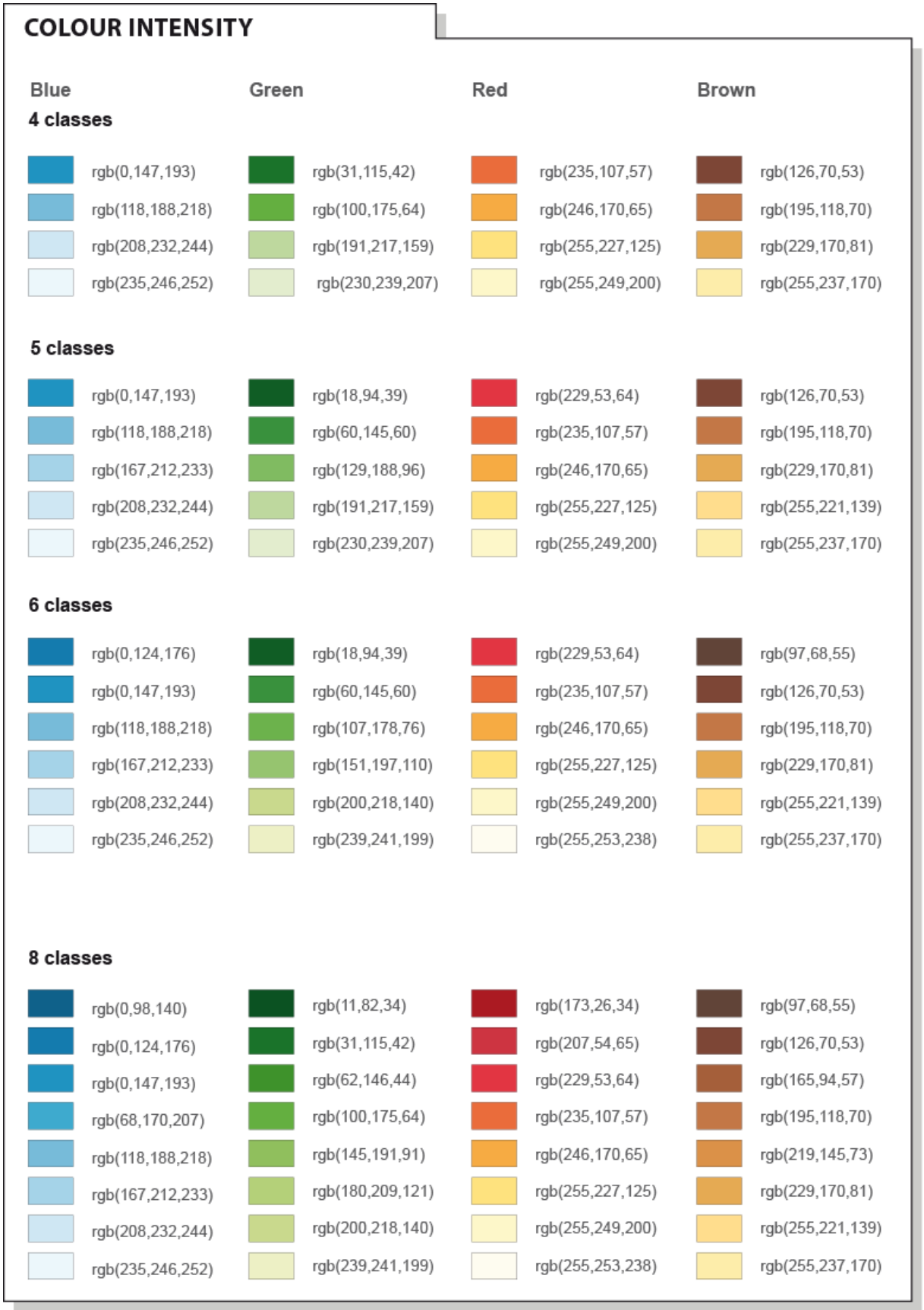
ANNEXE 1 - Relation of graphical variables to perceptual characteristics

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			

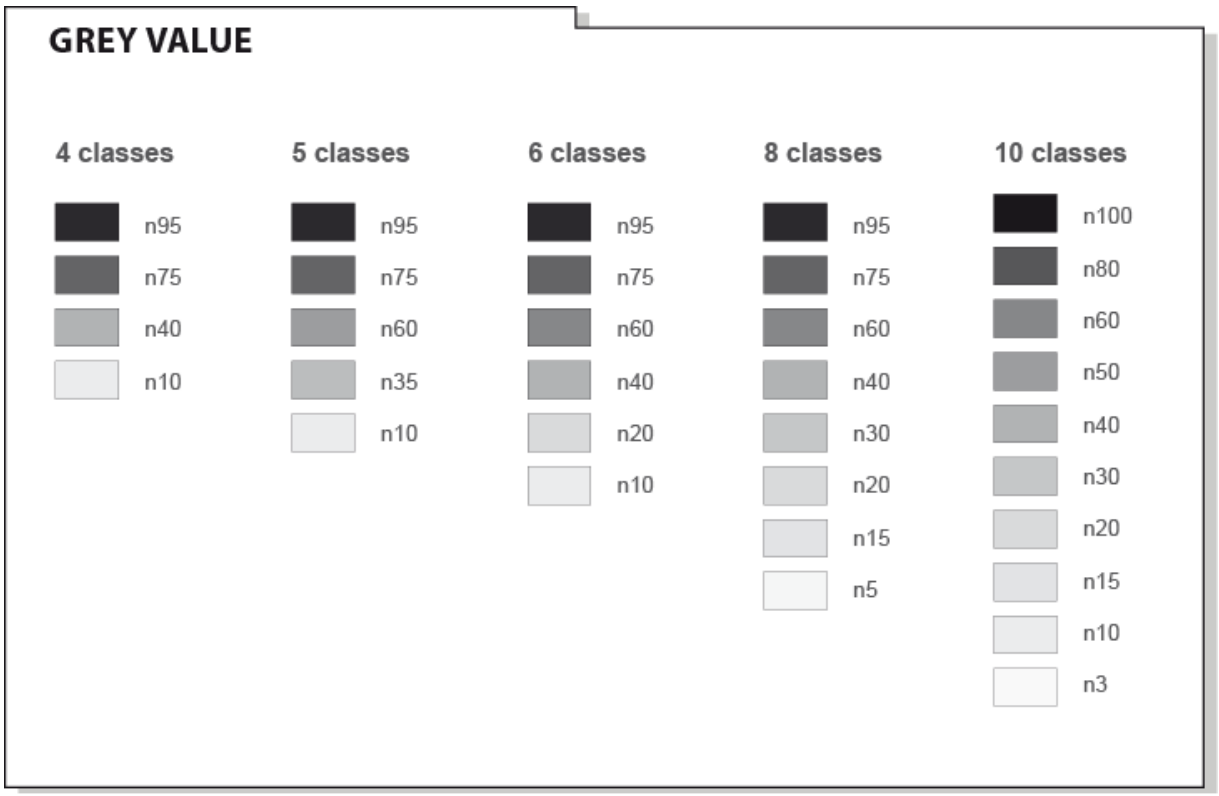
ANNEXE 2 - Numbers of categories that can be perceived at a glance

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

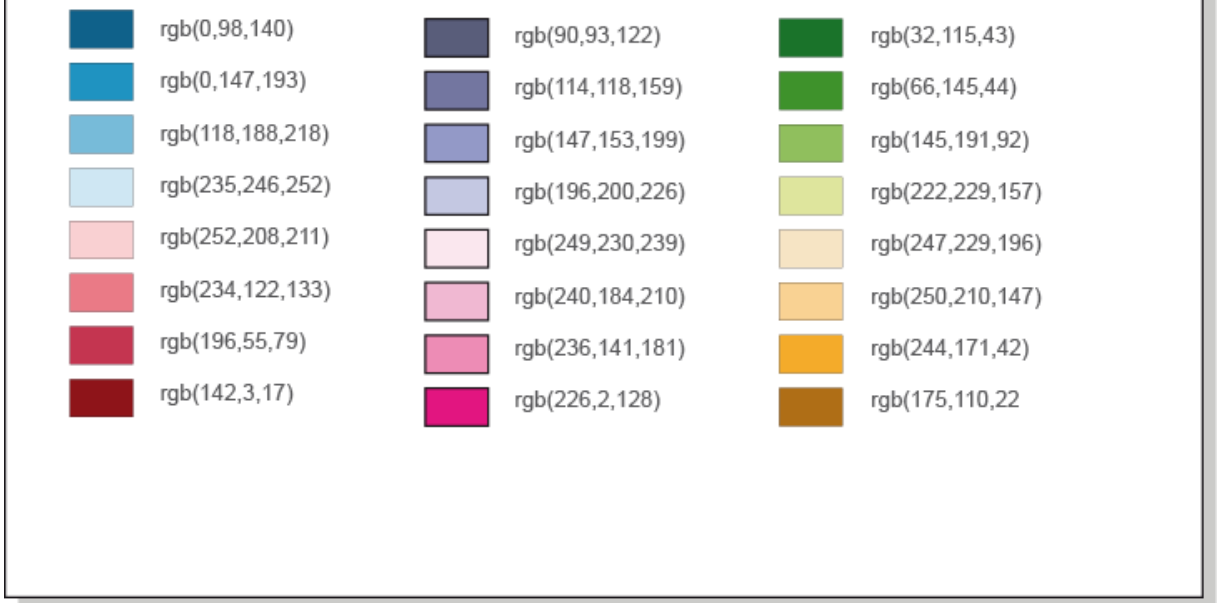
ANNEXE 3: Differences in value or lightness



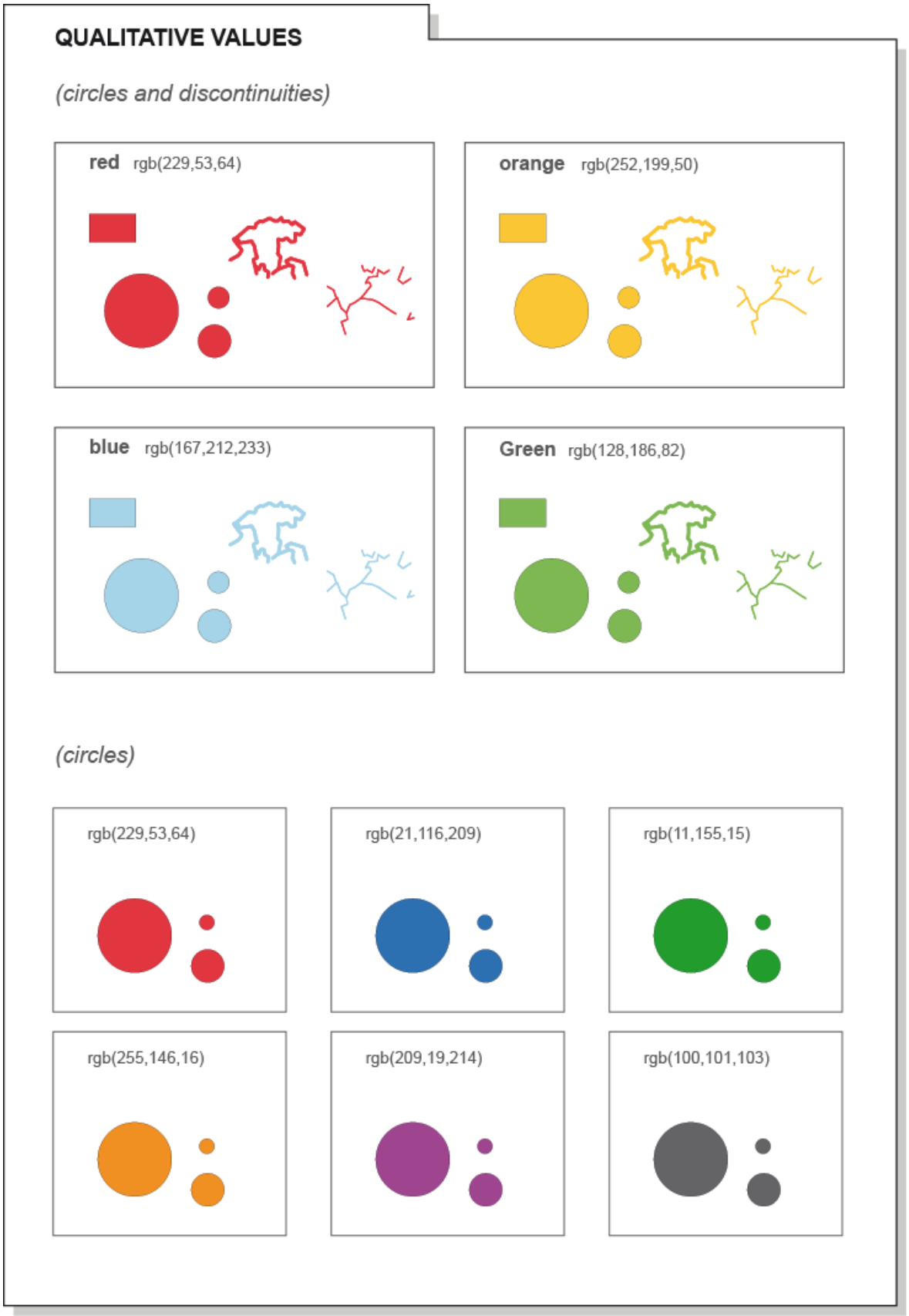
GREY VALUE



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/

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

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

32 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.

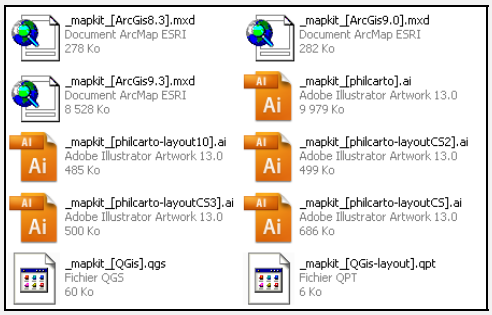
1 Maps and ESPON 2013

Description of the Map-Kit tool.

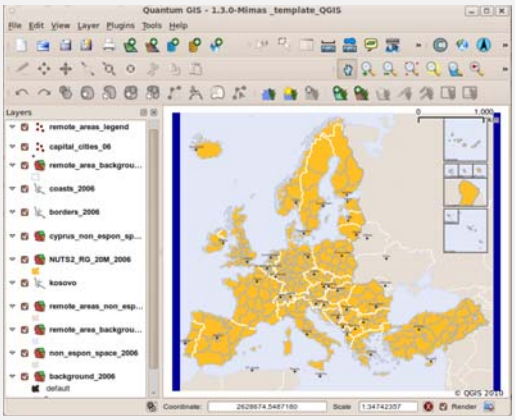
1.1 The “European” Map-Kit

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*

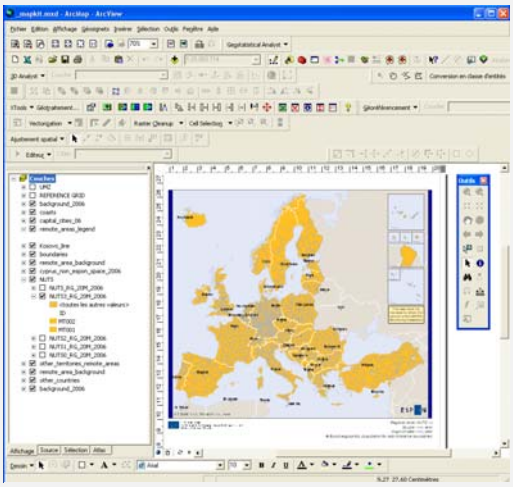
3 TOOLS FOR THE SAME MAPKIT



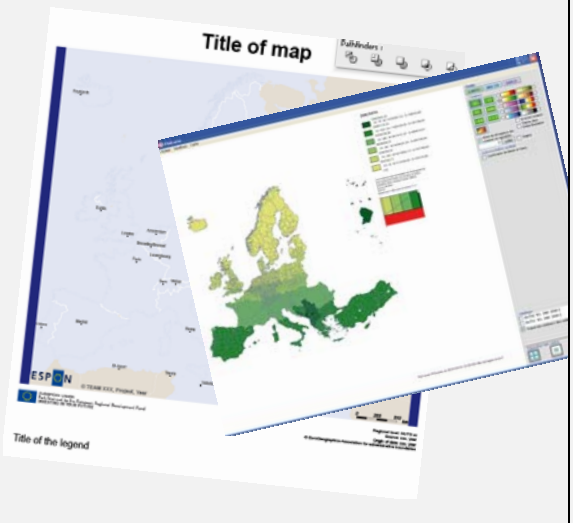
QGIS screenshot

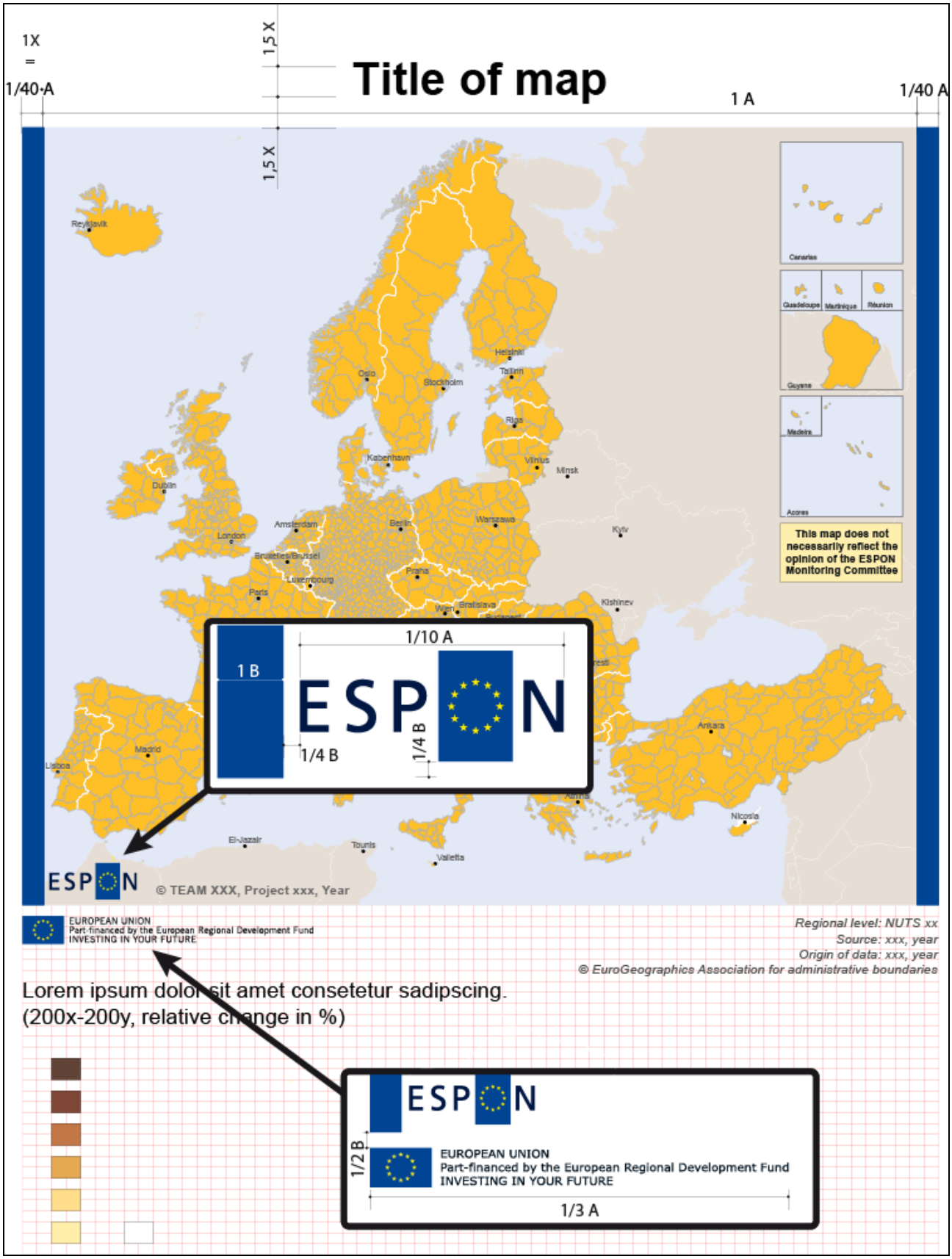


ArcGIS screenshot



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.

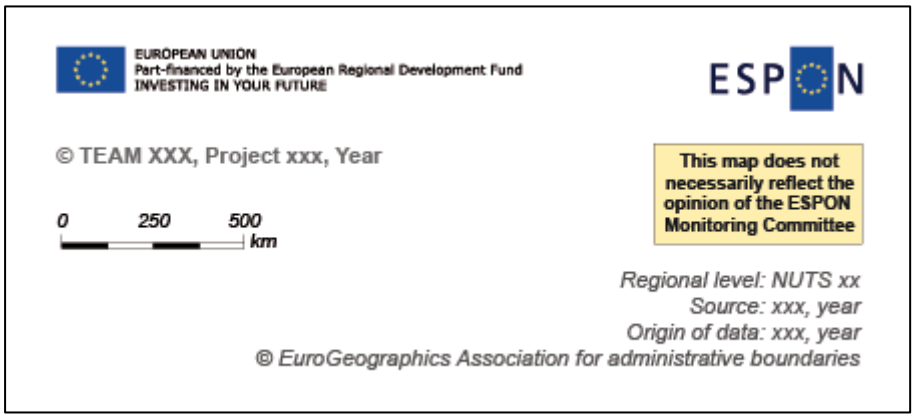
1.1.3 Projection and Ellipsoid

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.



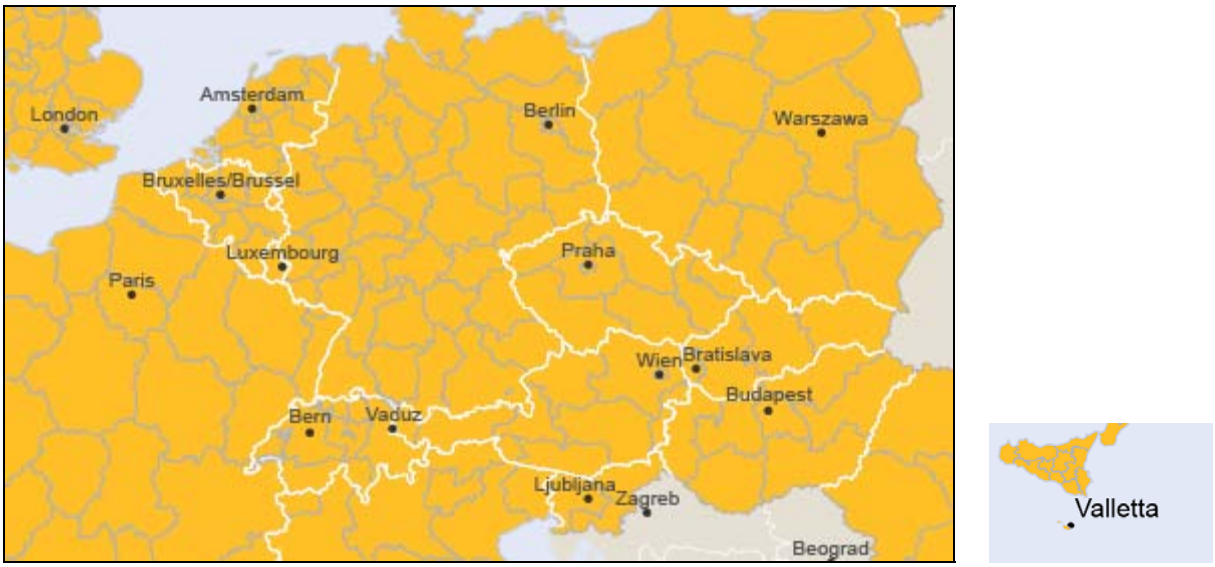
Some elements absolutely have to appear on the map layout:

- 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.

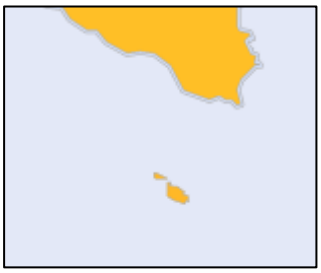
1.1.6 Remote territories

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

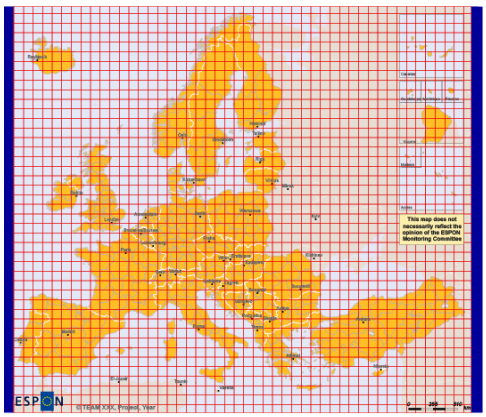
Cyprus is represented in two different colours. The North area appears in white as “no data”.

1.1.8 Coast of Malta



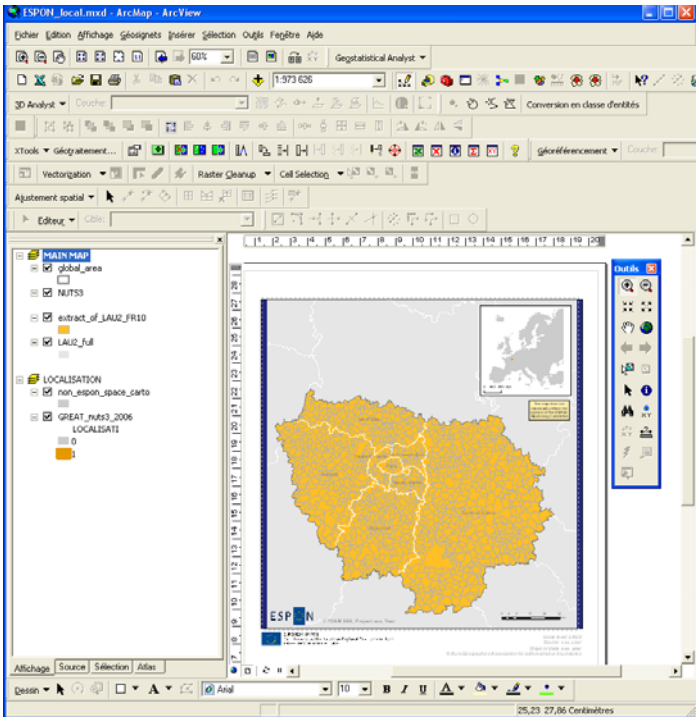
To ensure the visibility of Malta islands on the maps, we do not use light blue coast-line that could be reduce, on the map, the size of this country. Moreover, the line of the Malta polygon is drawn as thinner as possible.

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

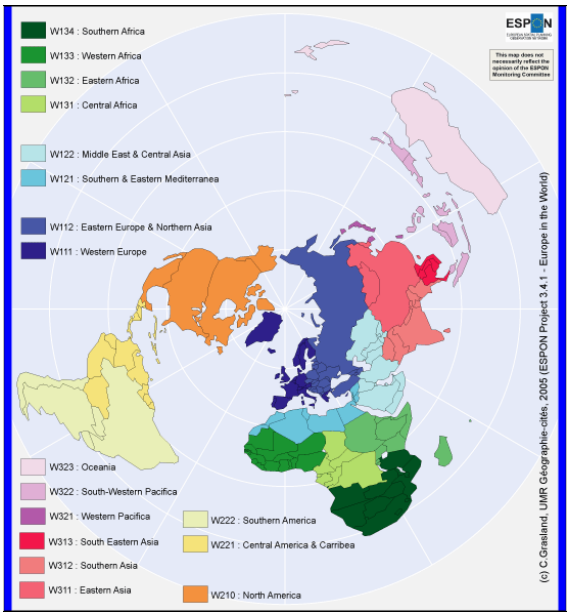


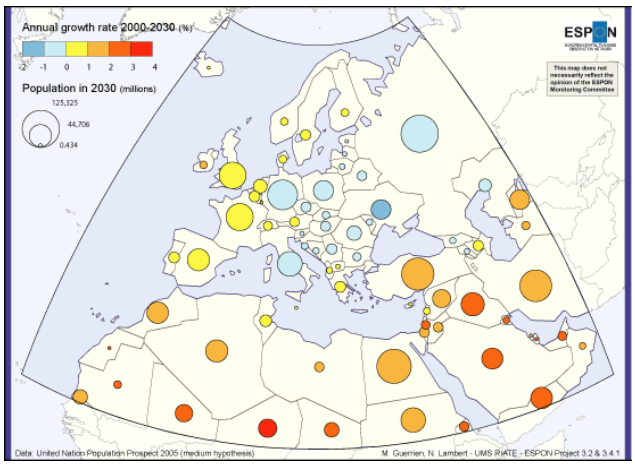
For Priority 2 projects, a “Local” map-kit has been created. As the same of the “European” Map-kit, some elements must figure on map: disclaimer, sources, logos, scale…

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)”.





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).

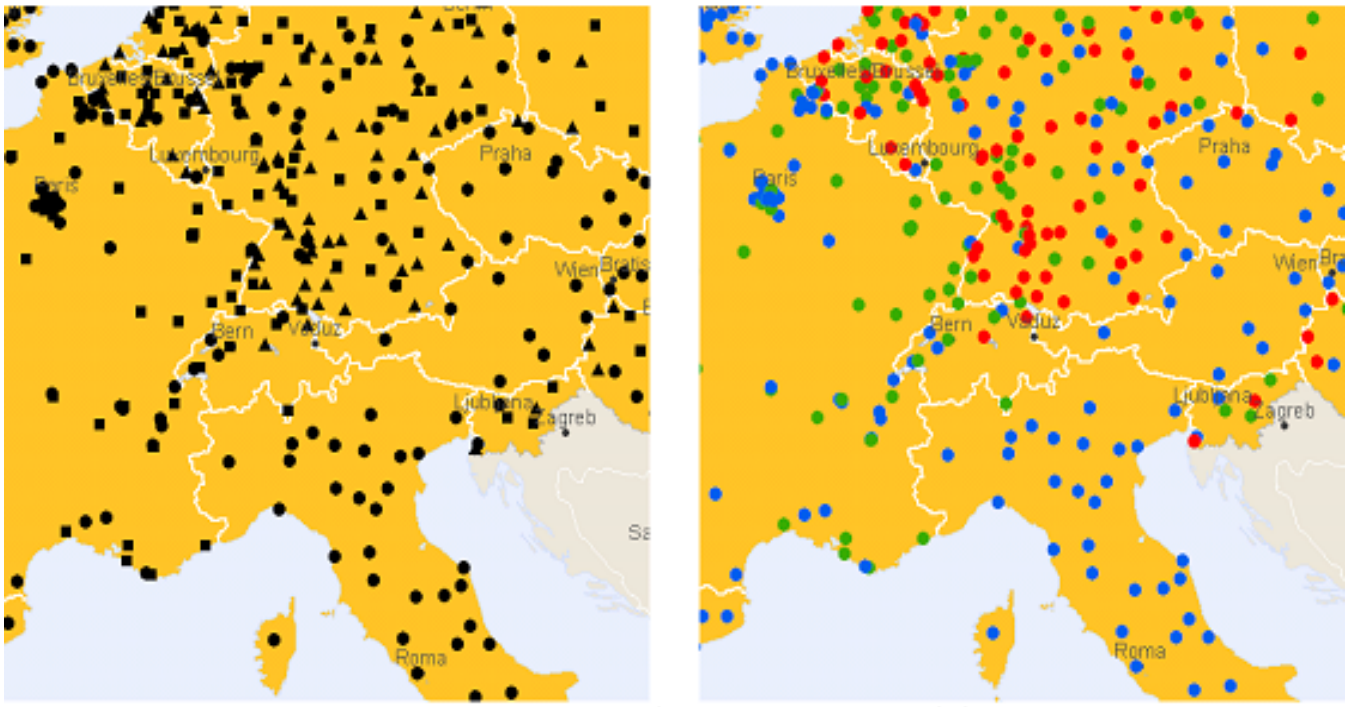


Figure 1 *– Universities by types – Two possibilities*
Good map = points + symbols or points + colours

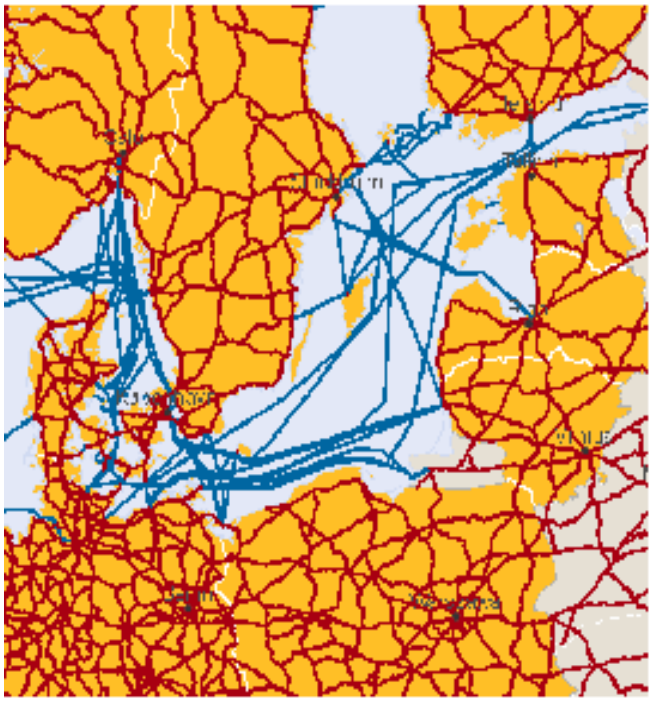


Figure 2 *– Mains roads and ferry connections*
Good map = line + colours

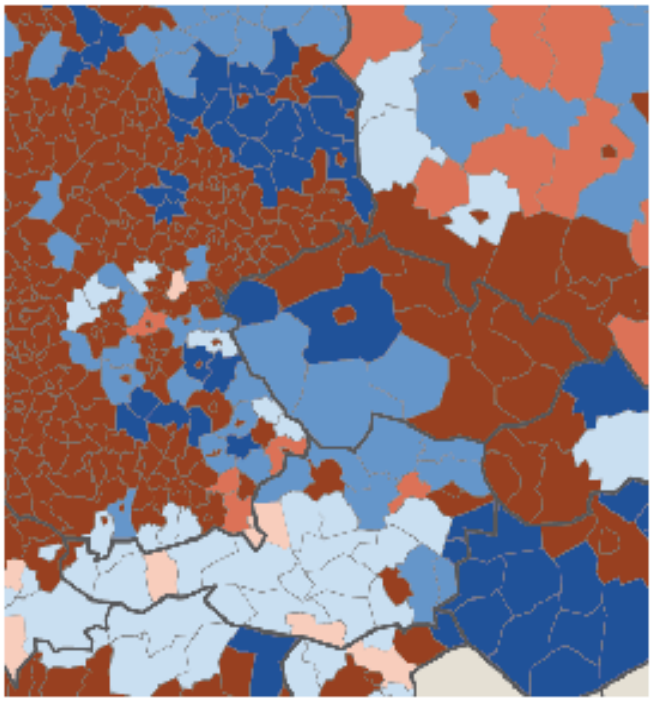


Figure 3 *– Results from Urban-Rural typology*
Good map = areas + colours

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.

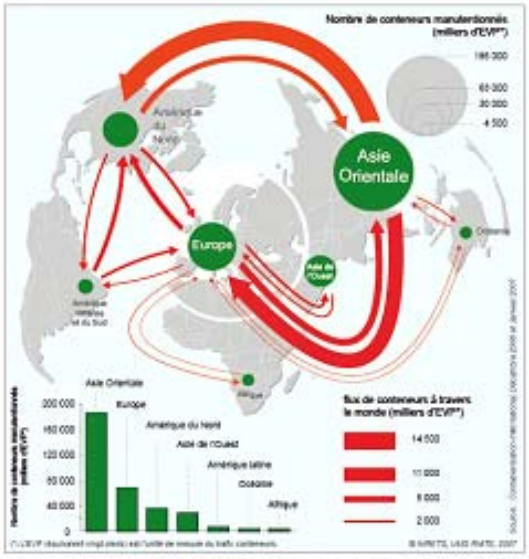


Figure 4 *- Containers flows across the World*
Good map = line + variation of line size

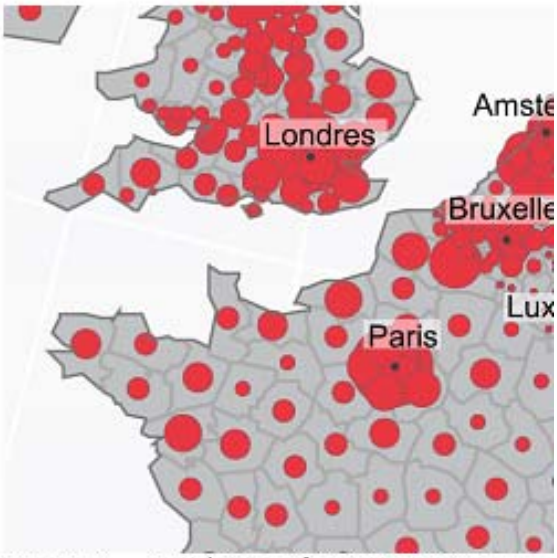


Figure 5 *– Population of NUTS 3 in Europe*
Good map = dot + proportional variation of size

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.

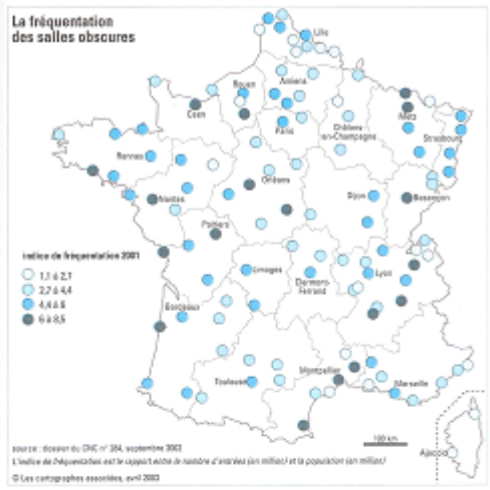


Figure 6: Cinema attendance Index in French main cities

Good map = dots + variation of colours

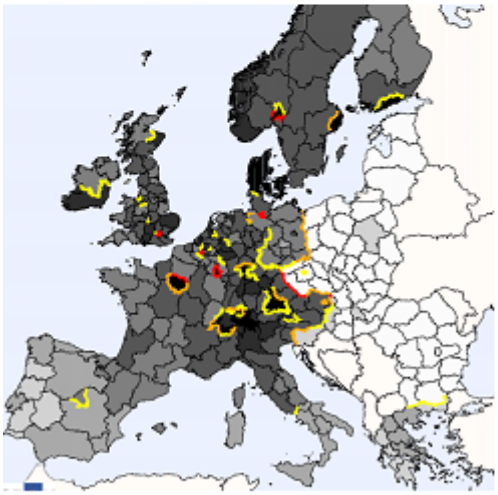


Figure 7: GDP per inhabitants discontinuities

Good map = lines + variation of colours

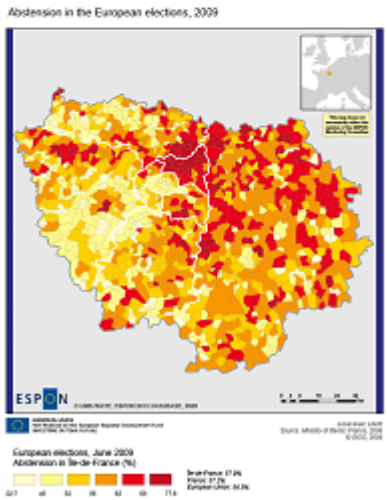


Figure 8: Abstention European votes in Île-de-France

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.

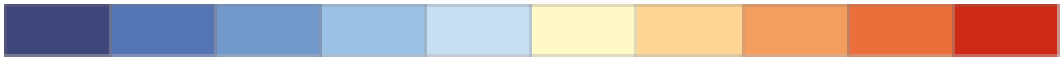
Figure 9: *Typology of Functional Urban Areas*
Good map: points + variation of colour (or size)

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:

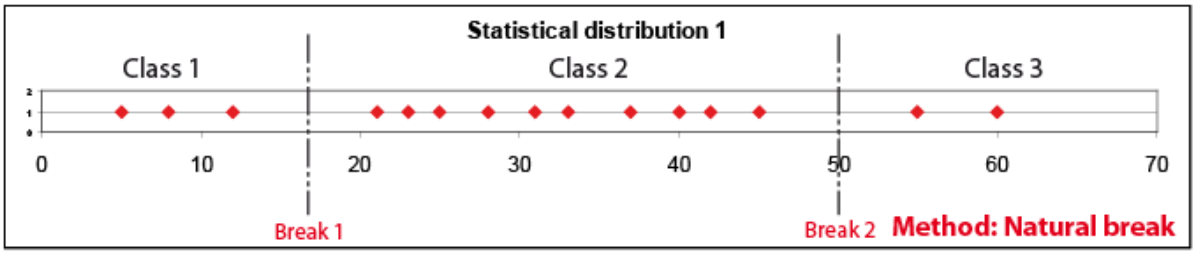
When combining red (warm colours) and blue (cold colours), **red is 'not good'/'negative'** and **blue is 'good'/'positive'"**

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.

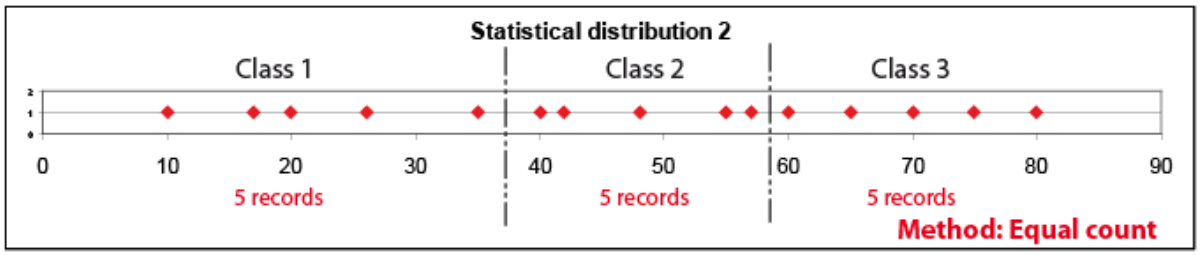
2.3.1 Natural Break

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.



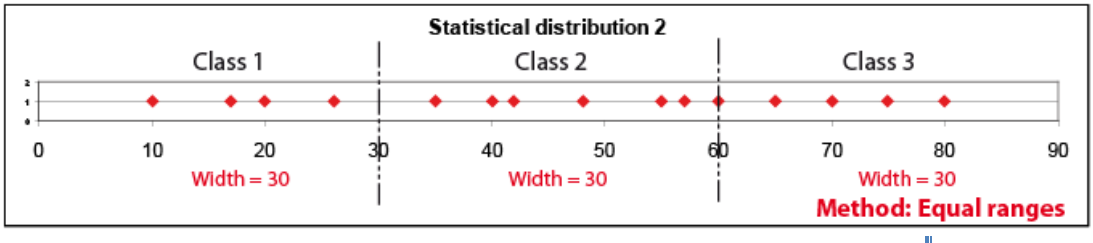
2.3.2 Equal Count or quantile

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.



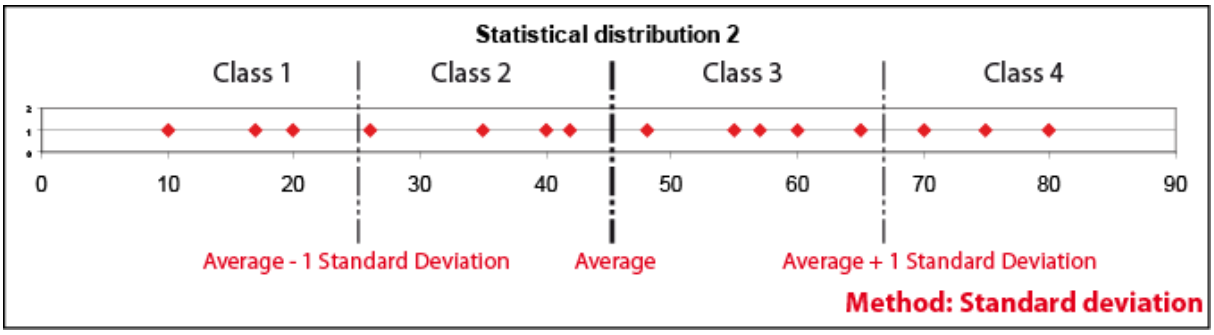
2.3.3 Equal Ranges

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

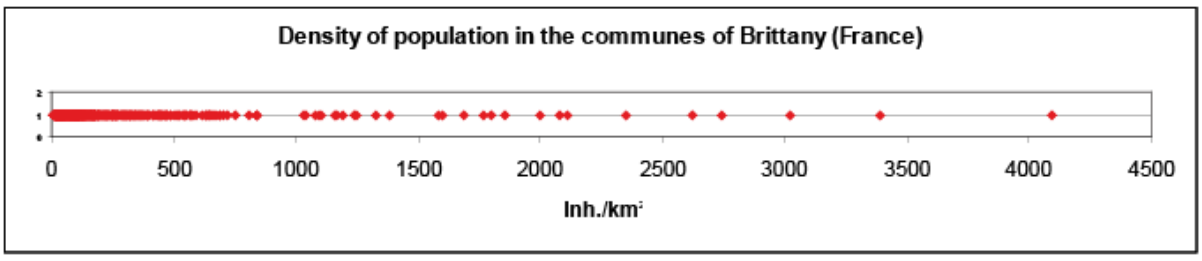
The widths of the class follow a geometric progression. To calculate the width of the different class, it is necessarily to estimate the geometric ratio, such as:

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.

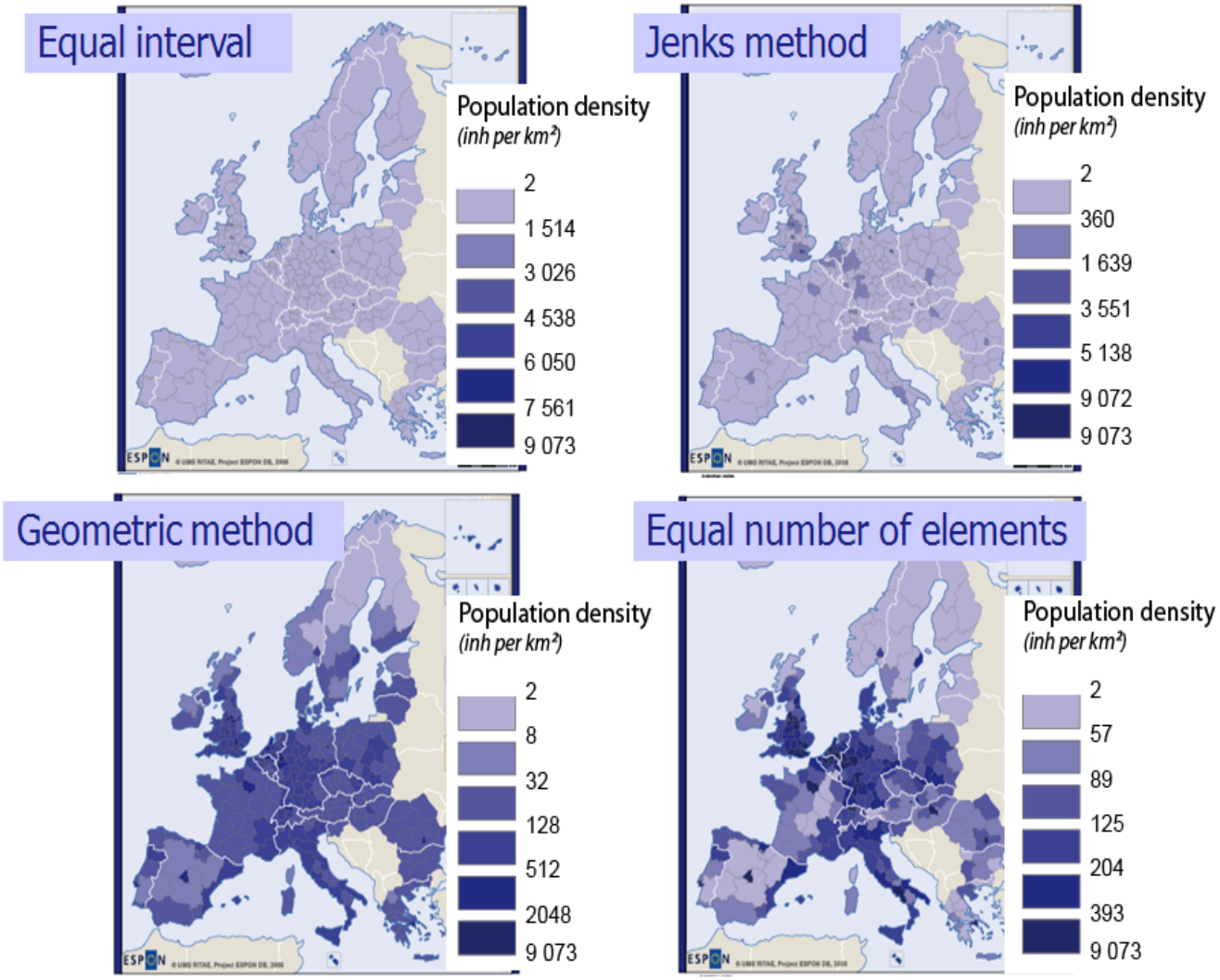


***Figure 12**: Result and efficiency are dependent upon the data classification method*

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.

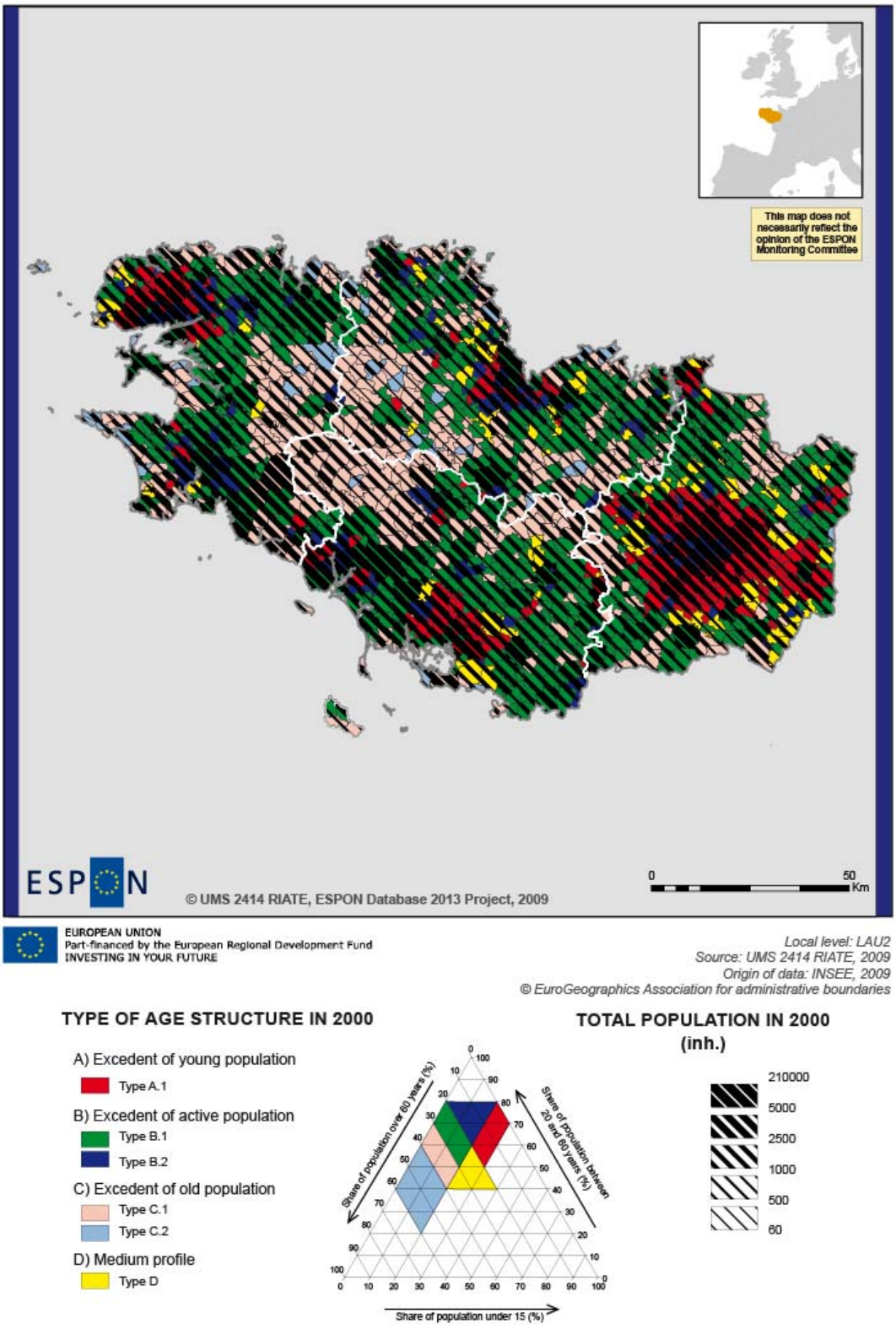


Figure 13**: Population and age structure in Brittany (France) – **with semiologic problems

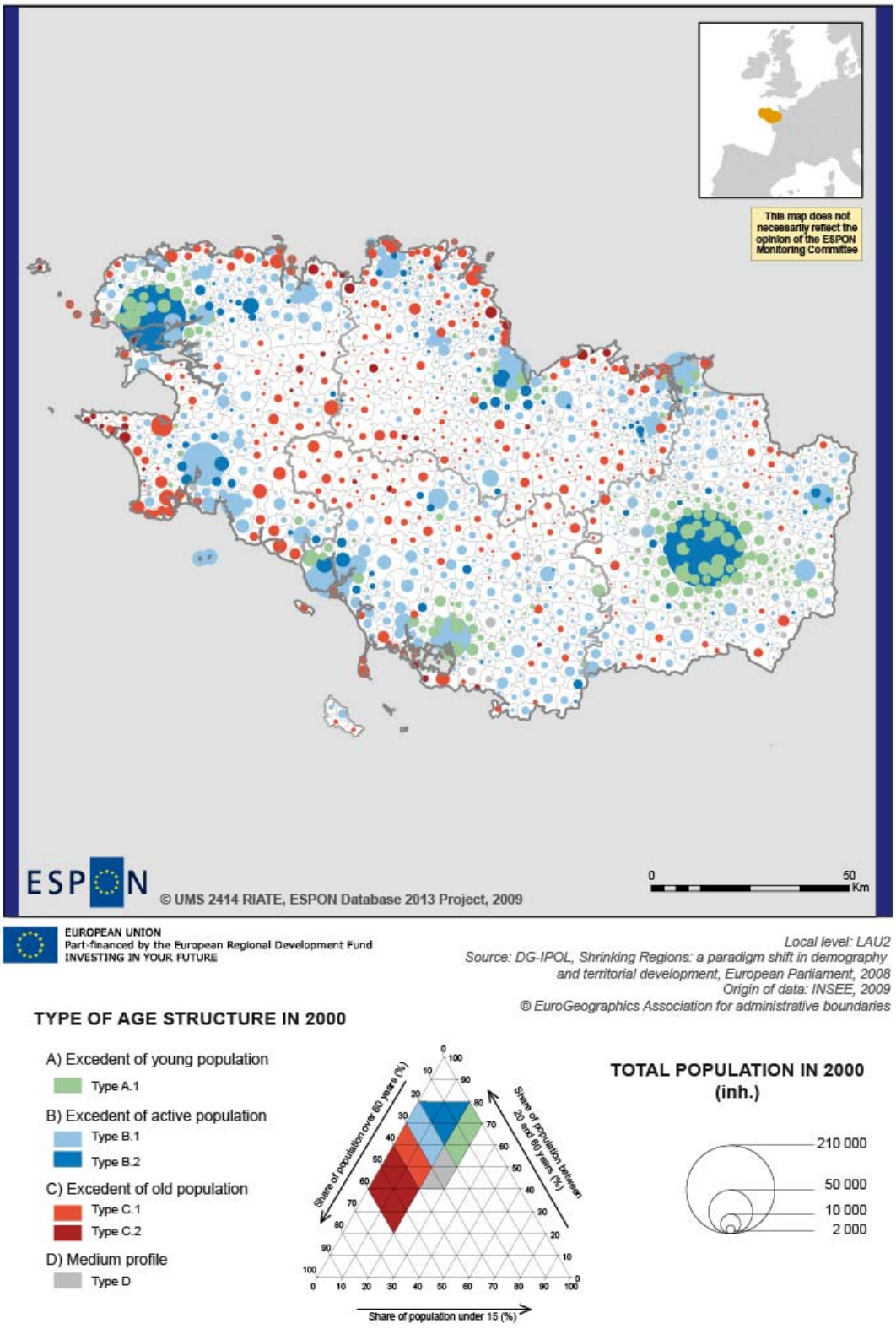


Figure 14**: Population and age structure in Brittany (France) – **without semiologic problems

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"!

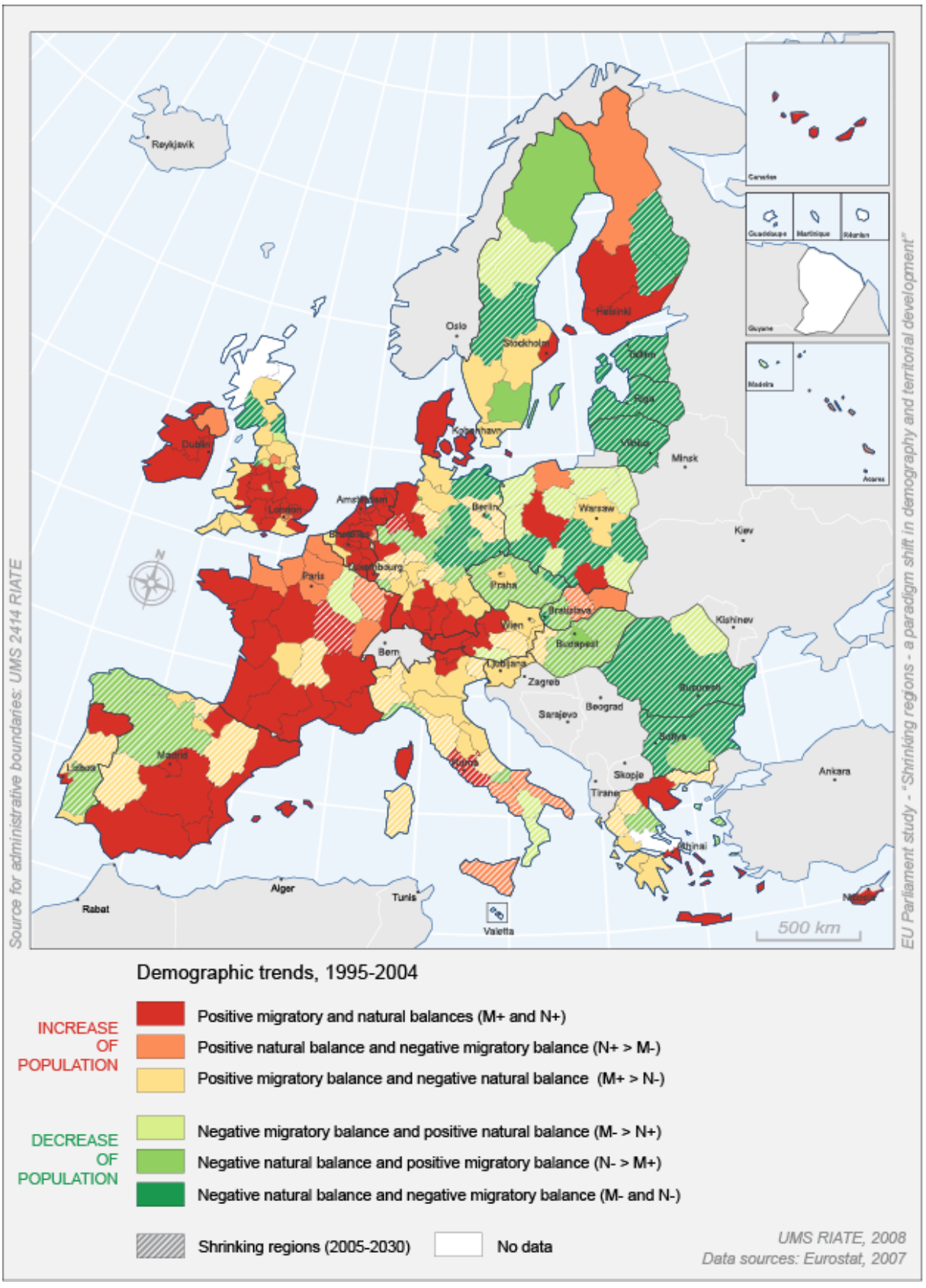


Figure 15**: Typology of regional growth patterns – **Possibility 1

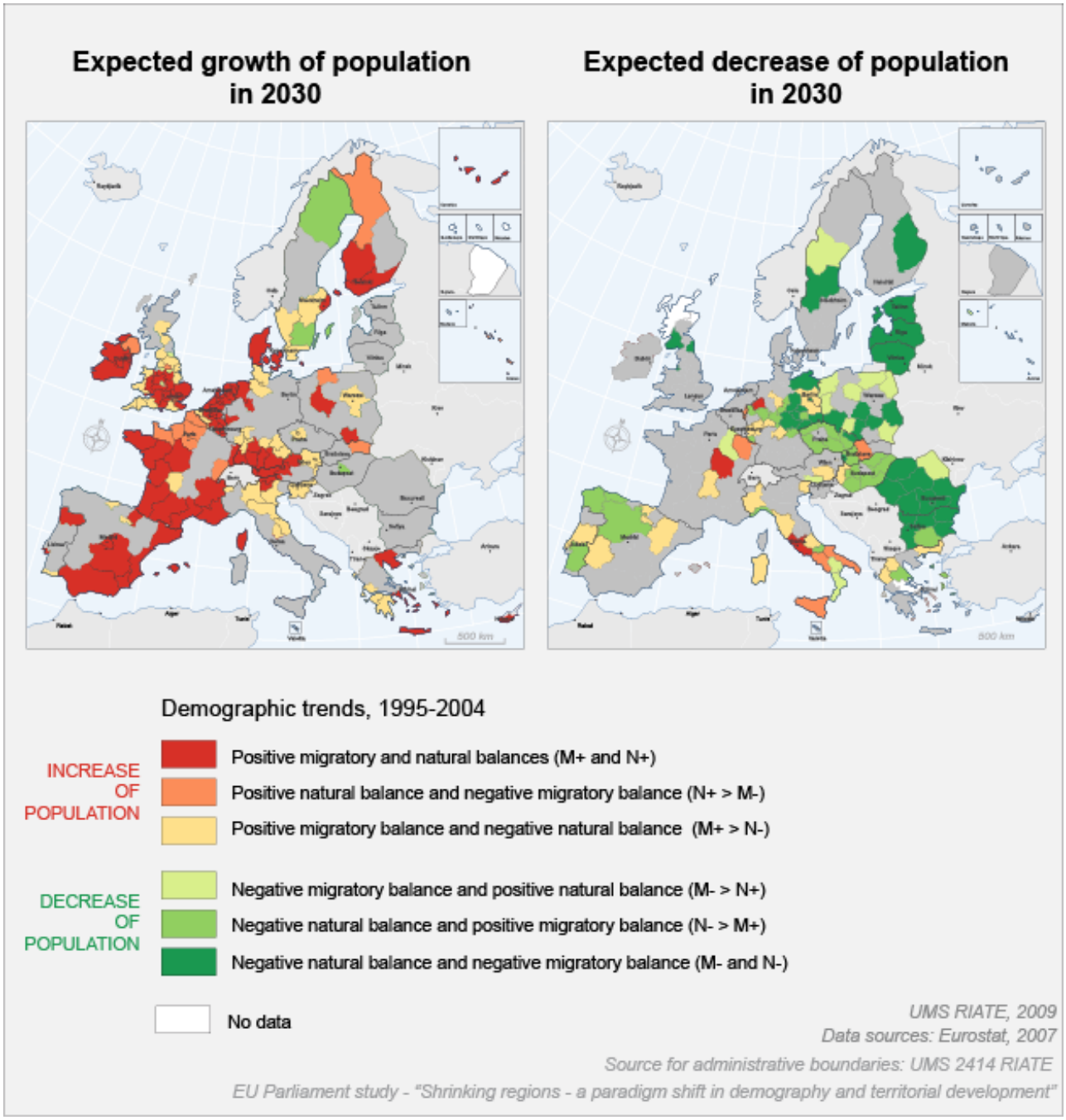


Figure 15**: Typology of regional growth patterns – **Possibility 2

ANNEXES

These annexes allow you to choose some efficient graphic variables to communicate differences in size, order or quality.

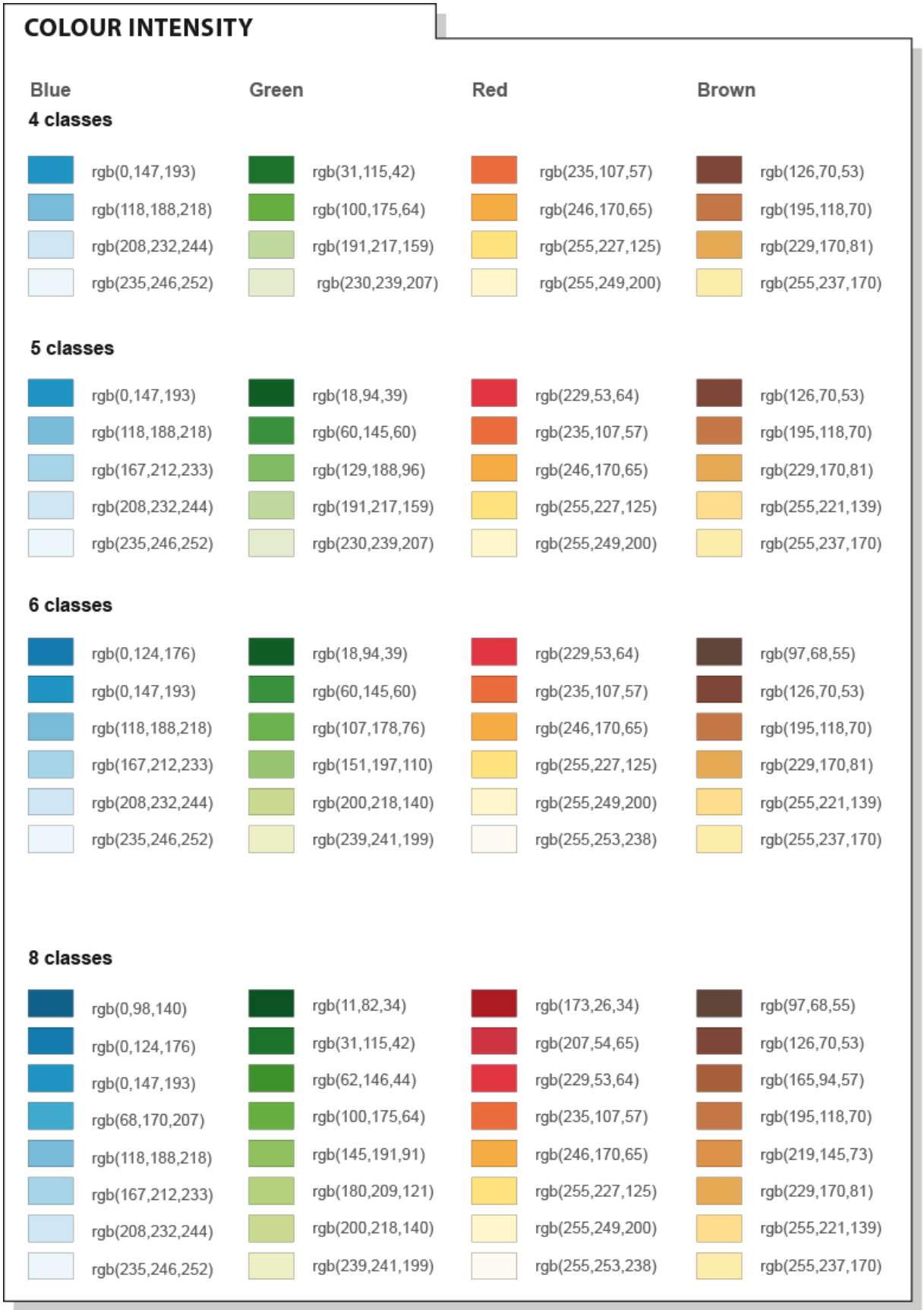
ANNEXE 1 - Relation of graphical variables to perceptual characteristics

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			

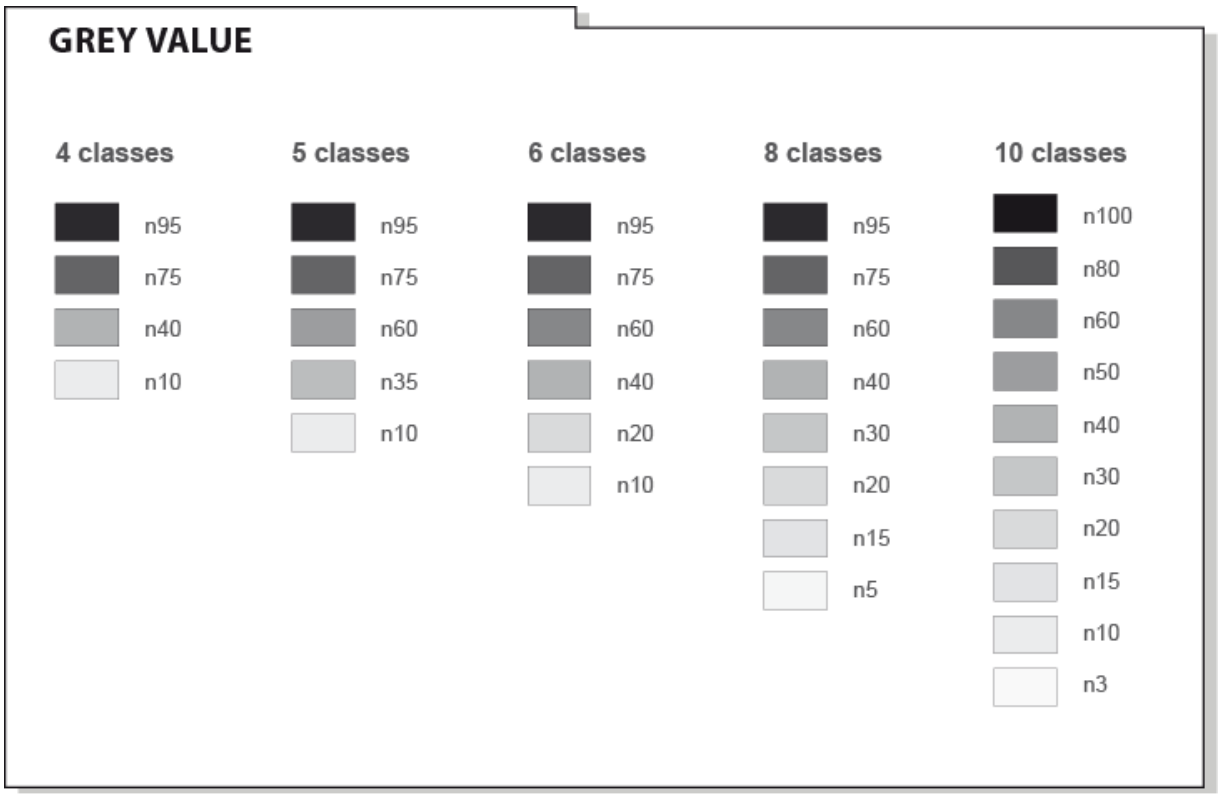
ANNEXE 2 - Numbers of categories that can be perceived at a glance

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

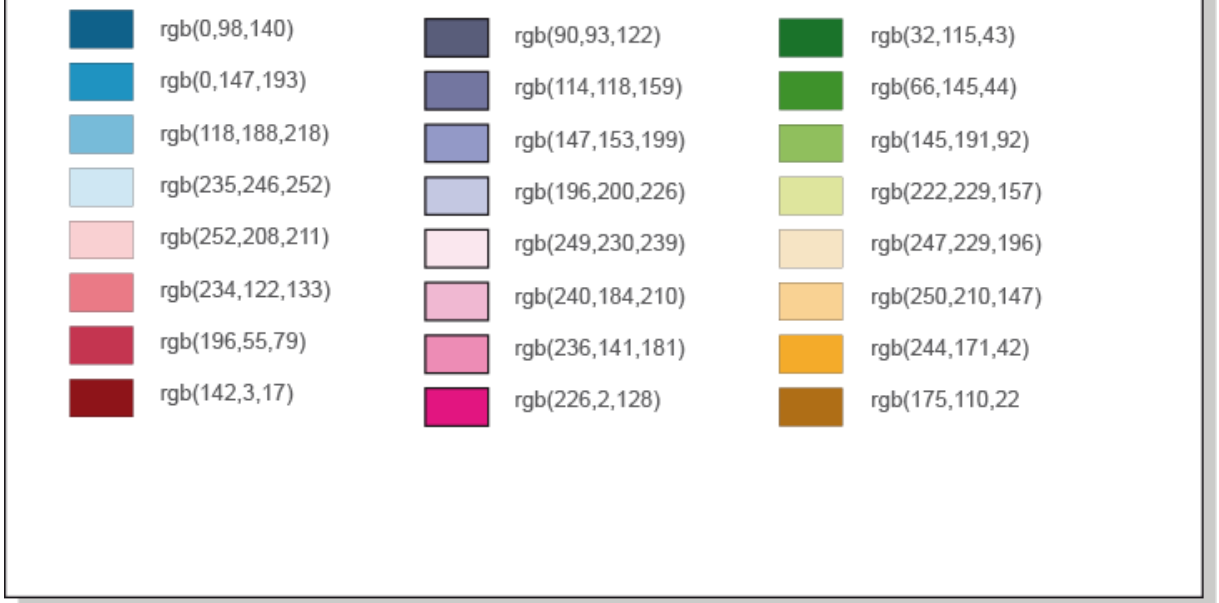
ANNEXE 3: Differences in value or lightness



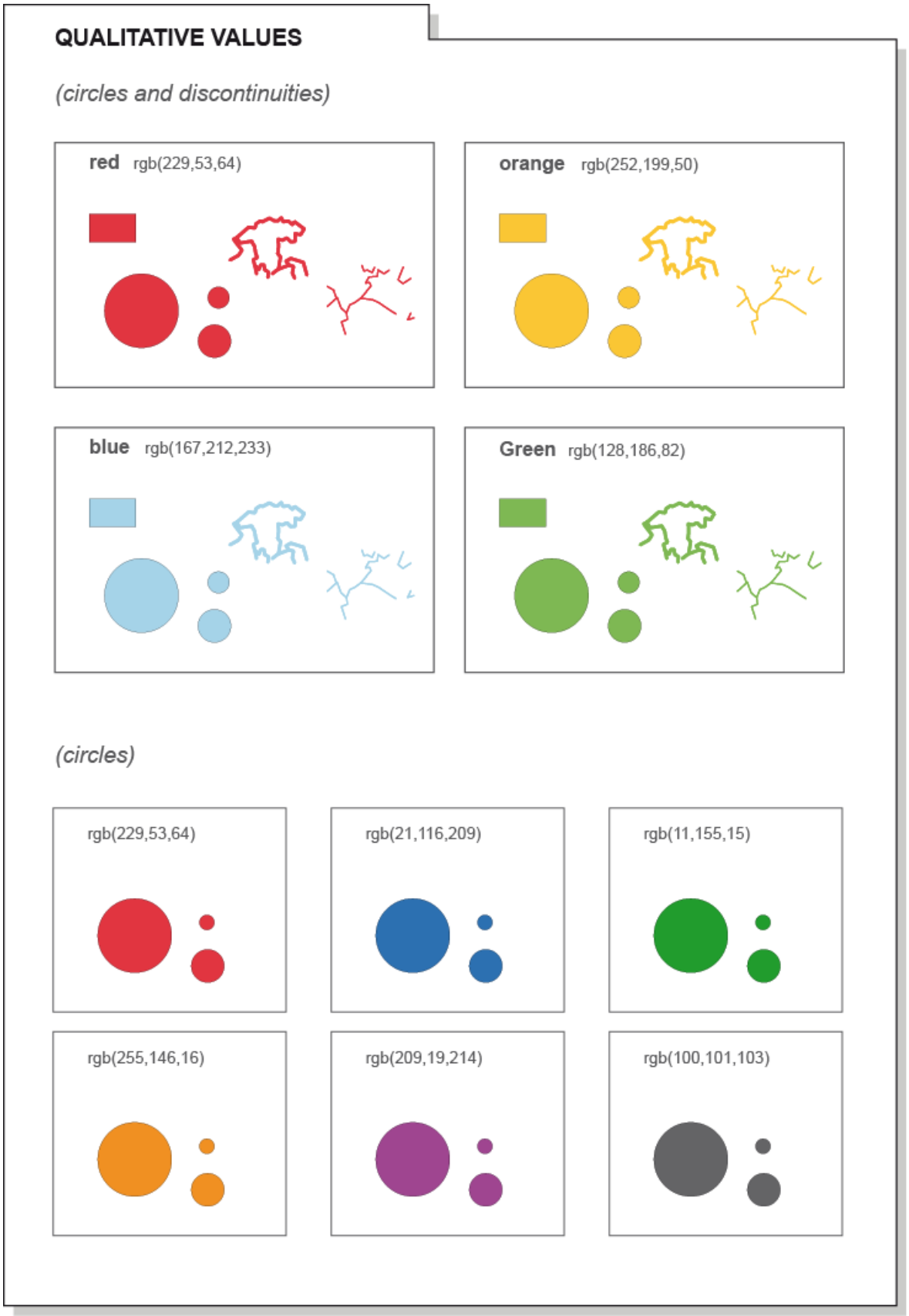
GREY VALUE



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/

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

_Toc254622453

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).

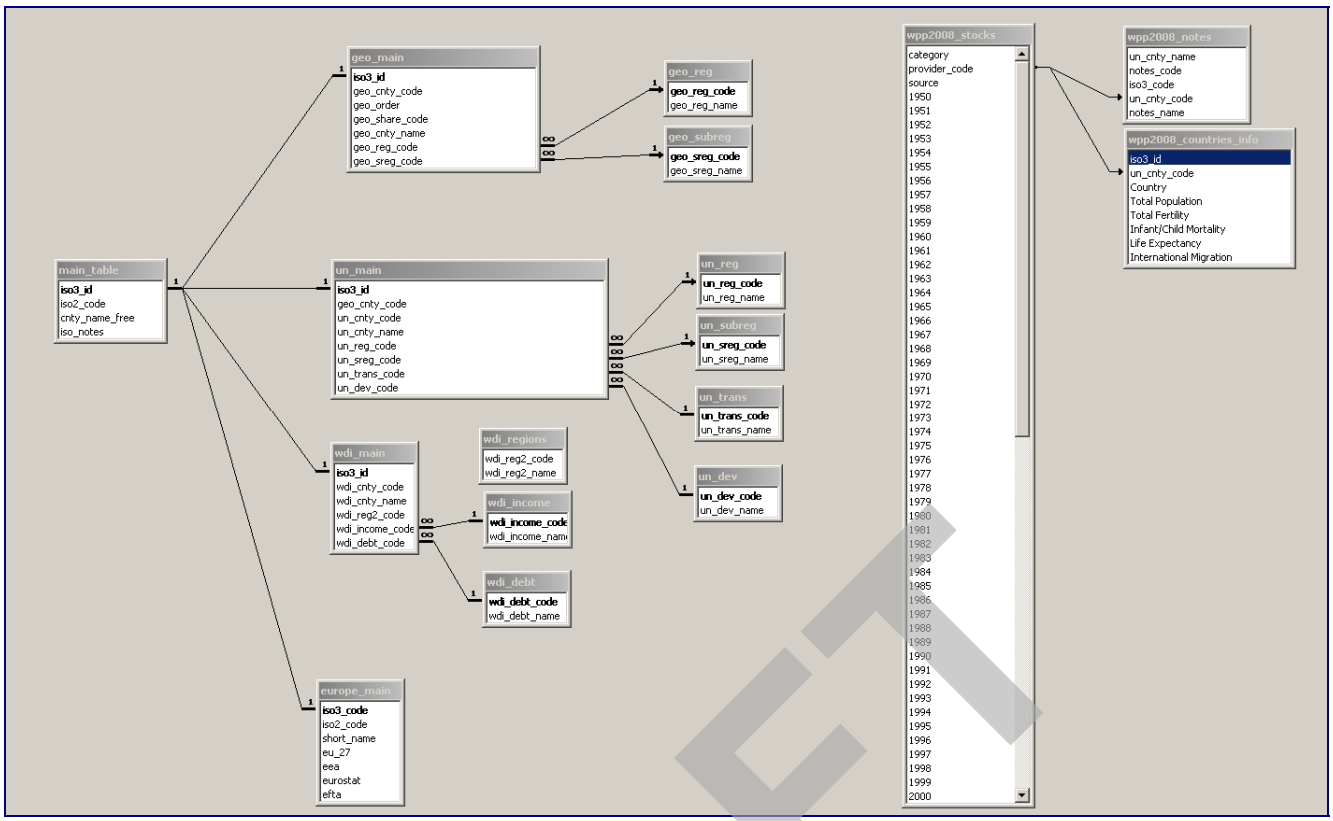


Figure 1 : countries/territories subdivisions and regional/thematic aggregations supplied by our main data providers

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.

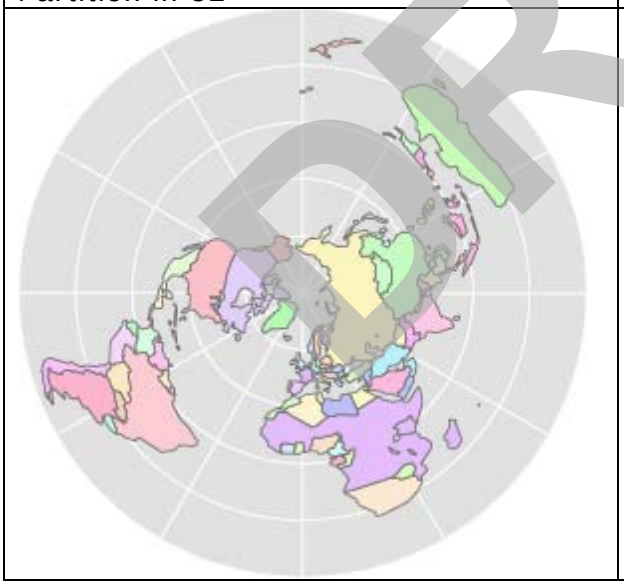
Partition in 82	Partition in 96
	

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).

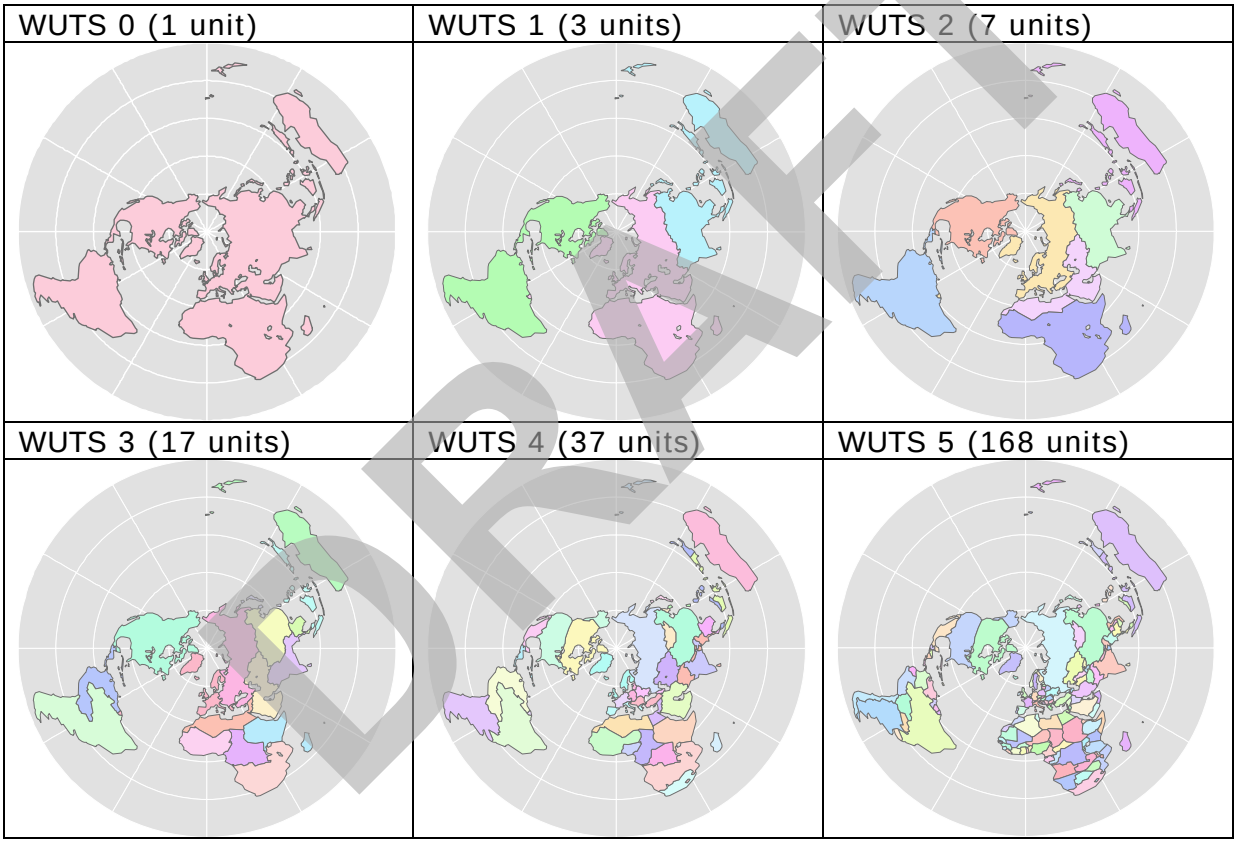


Figure 3 : Europe in the World subdivisions

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/

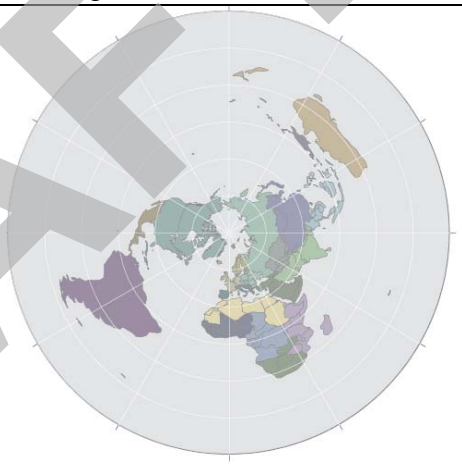
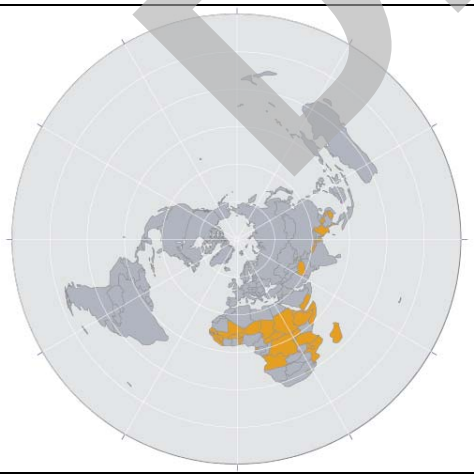
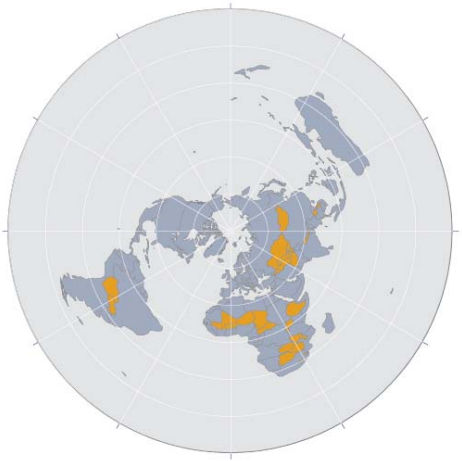
UN regional Breakdown	UN sub-regional Breakdown
	
UN Developing countries	UN Landlocked developing countries
	

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.

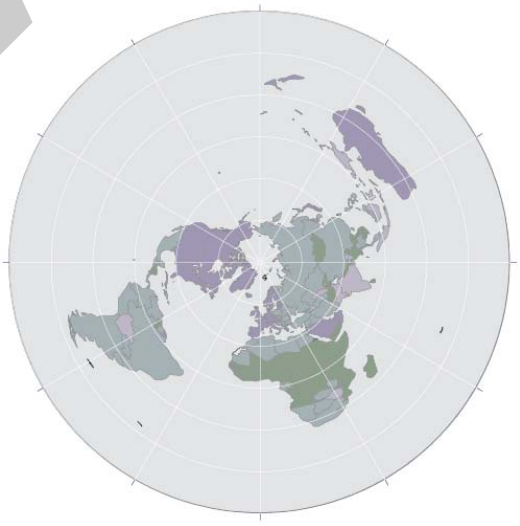
WDI aggregations by Income	WDI aggregations by Debt
	

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.

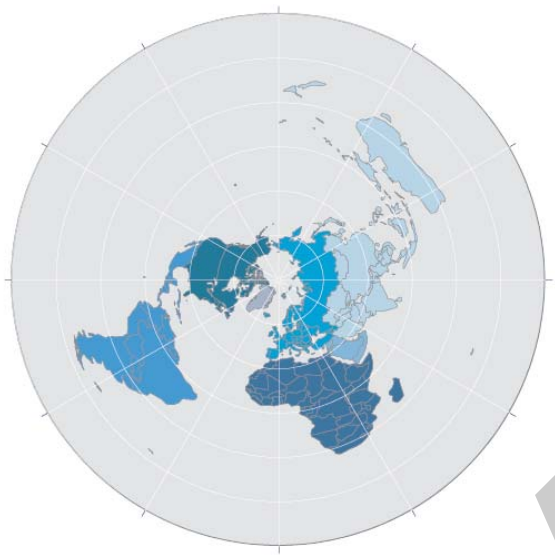
GEO regional Breakdown (7)	GEO sub-regional Breakdown (23)
	

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..)

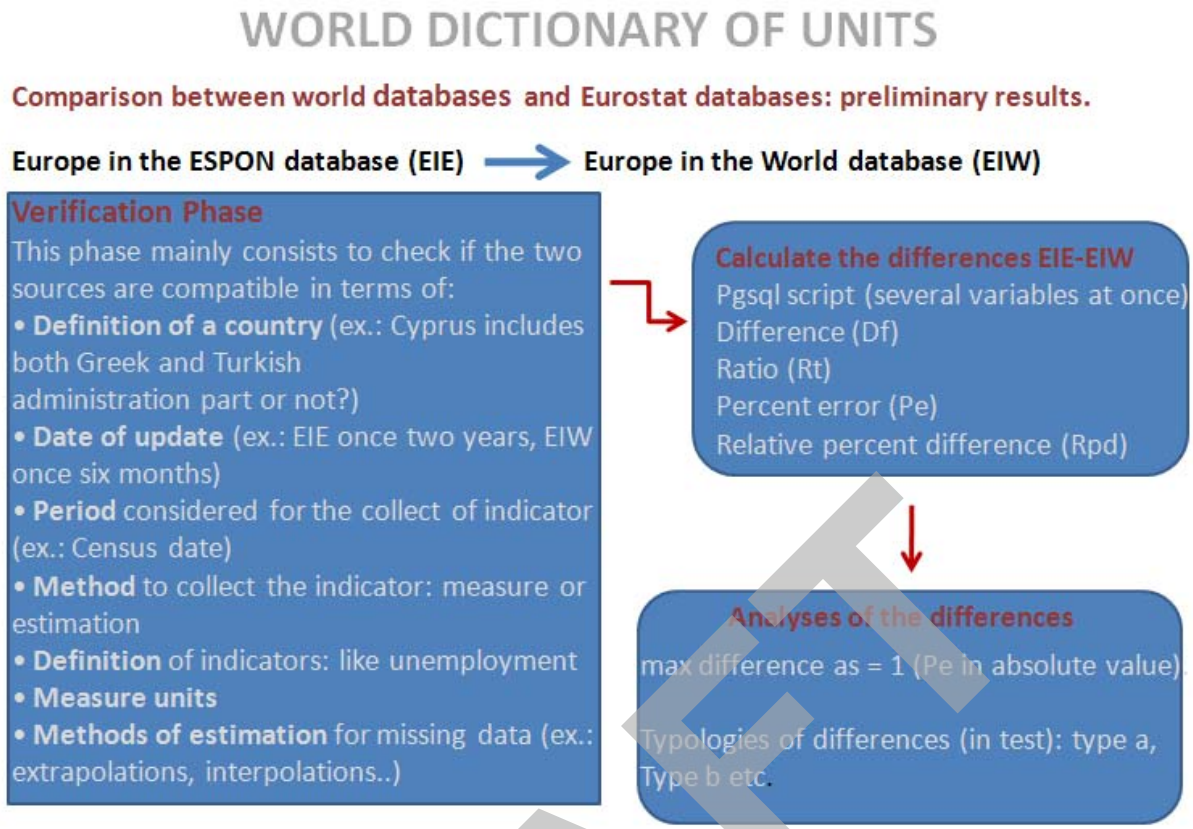


Figure 6 : Summary of the methodology used in the testing phase

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…

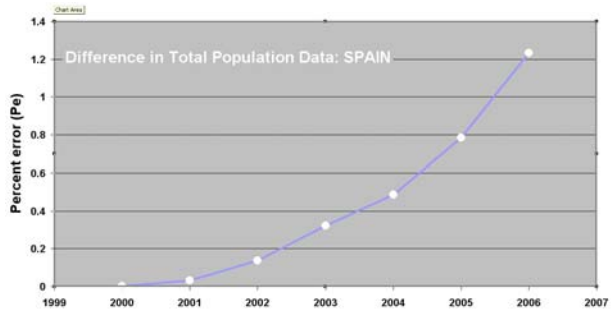
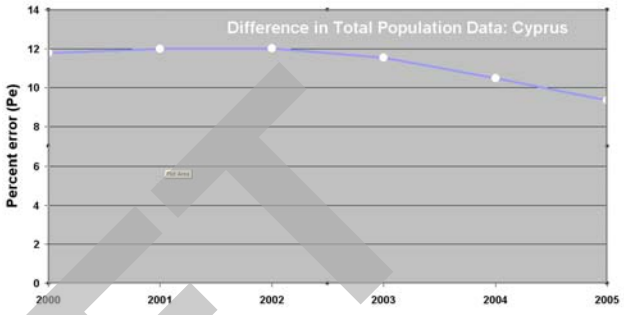
Example of exponential variability	Example of constant variability
	

Figure 7 : Different types of variability

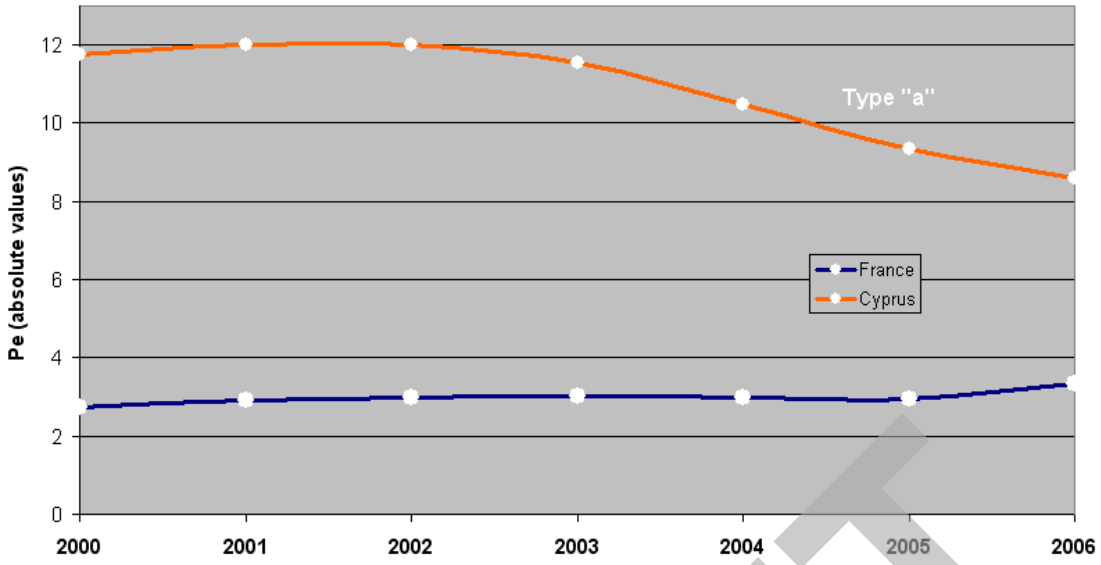
The typologies of difference include:

- 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…)

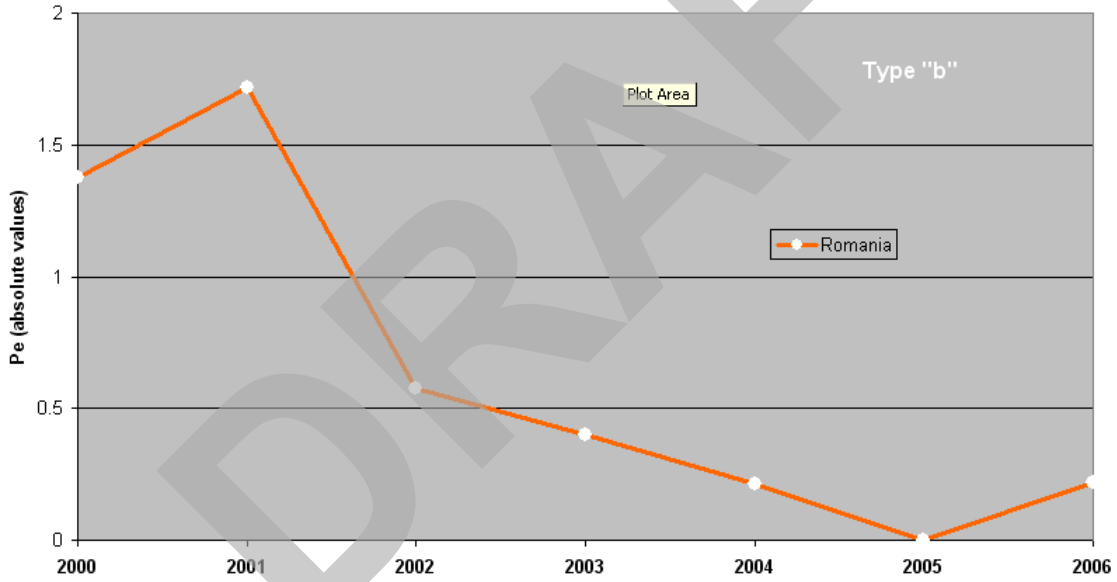


Figure 8 : Towards a typology of differences

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

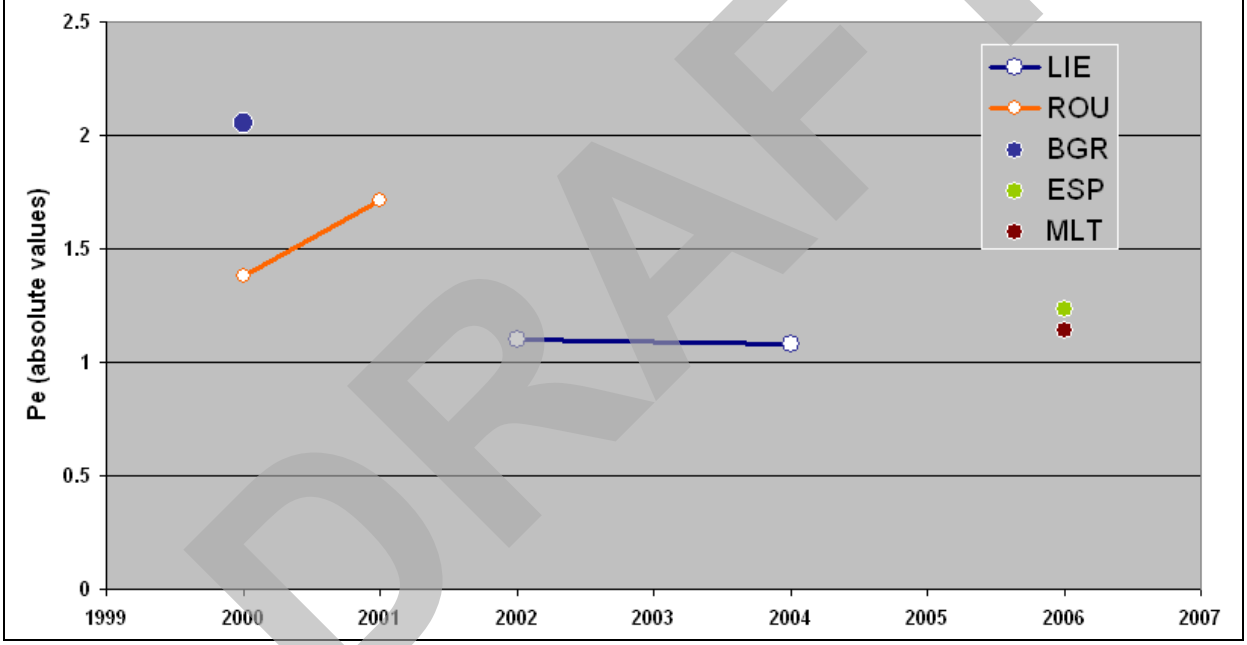


Figure 9 : Identification of countries where significant differences has been identified

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)

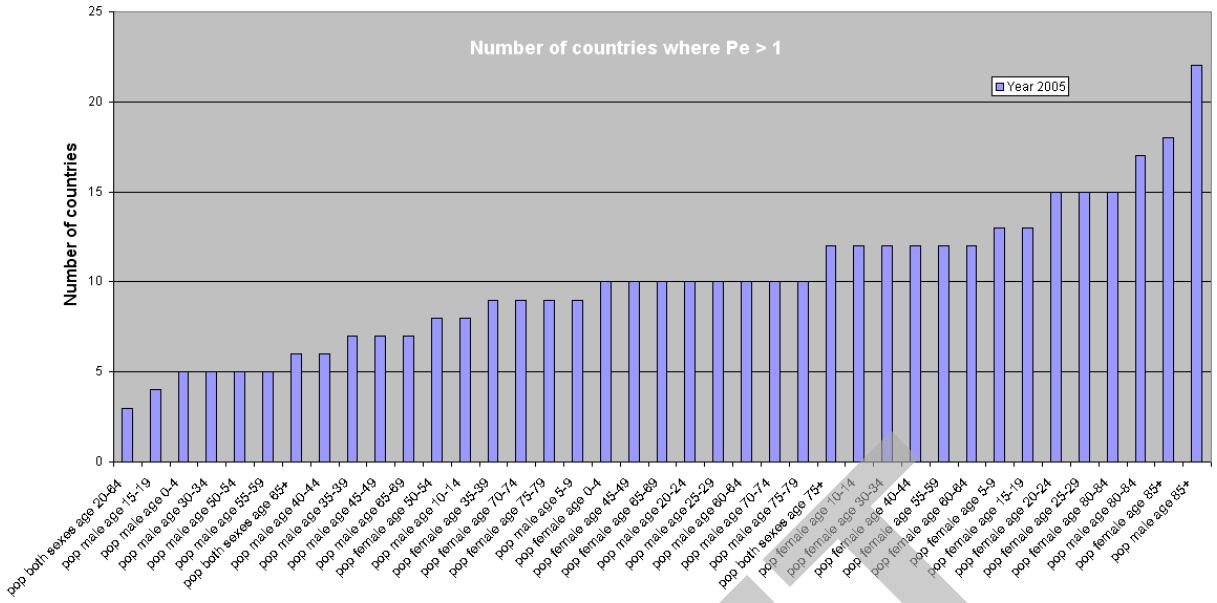


Figure 10 : Identification of age-classes where significant differences has been identified

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)

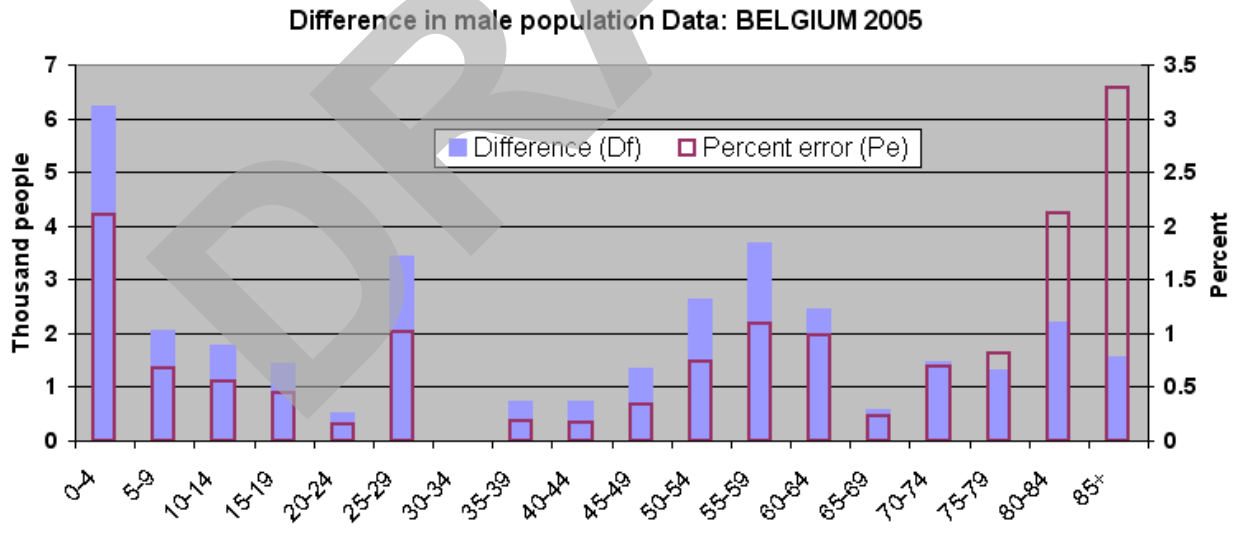
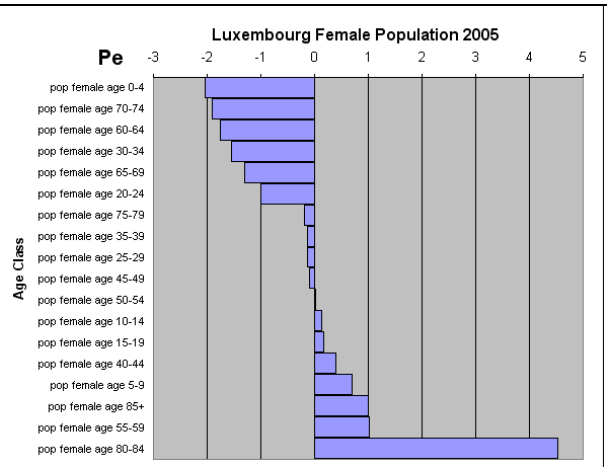


Figure 11 : Identification of age-classes where significant differences has been identified in Belgium

Per age groups for Luxembourg (year 2005)



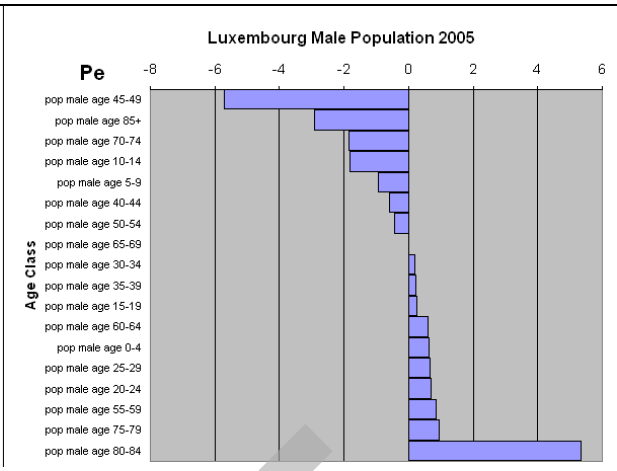


Figure 12 : Identification of age-classes where significant differences has been identified in Luxembourg

…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
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

Data related to GDP, exclude Transnistria

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 *is* 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 discuss the Spatial Administrative Divisions of WB and Turkey and presents the entire set of data required as well as a first set of "basic" 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

45 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

Contact

angelidi@central.ntua.gr

tel. + 30 210 7721731

TABLE OF CONTENT

Introduction and some general remarks

The **Western Balkans (WB)** countries -Albania, Bosnia and Herzegovina, Croatia, FYROM, Serbia, Montenegro and Kosovo (under UN Security Council Resolution 1244)- **and Turkey** are ***Candidate Countries (CC)*** **or** ***Potential Candidate Countries (PCC)***[1]*.*

The part of the ESPON Database 2013 project referred to **WB and Turkey** data is a first part of the Challenge 11 of the project. It aims 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.

According to the time-schedule of the project, we accomplished during 2009, 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. We initiated the discussion on this issue with Eurostat.

Apart from the discussion of the issues concerning the WB and Turkey data, we refer to the data delivered in 2009 to the Lead Partner of the project, which are gradually integrated in the ESPON 2013 database.

We comment here on the WB and Turkey data at NUTS0 to NUTS3 levels. We will comment concisely on the availability of data at LAU1 level during the stage of the project following the SIR (February 2010).

[1] According to the overall enlargement strategy document adopted by the Commission on November 8th 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; accession negotiations have not started. Albania, Bosnia and Herzegovina, Montenegro 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.

1. General assessment

1.1 The WB and Turkey Spatial Administrative Divisions

We first had to assess ***the conformity of the WB and Turkey spatial administrative divisions to the EU NUTS classification criteria.*** The NUTS Regulation lays down the following minimum and maximum thresholds for the average size of the NUTS regions.

NUTS 1: 3 - 7 million (of inhabitants)
NUTS 2: 800 000 - 3 million,
NUTS 3: 150 000 - 800 000.

Turkey, Croatia and FYROM have already adopted this 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* -see in Tables 1.1 and 1.2 and in detail in section 2. We have also examined, as possible, the administrative capacity of the spatial administrative divisions which fulfil better the respective NUTS population criteria. The results have not changed significantly the previous conclusion.

Table 1.1: NUTS1,2,3 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 NUTS1,2,3" 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 assesment
** 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 Candidate country, 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 used non official shapefiles for these countries downloaded from the Berkeley University website which have been appropriately adjusted by the RIATE workgroup. Obviously, after the establishment of official contacts among the ESPON 2013 Database project and the National Statistical Offices of these PCC, we could use their official shapefiles.

A first set of "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: Available data for the "2009 delivery" per CC / PCC (years 2000-2006*) and NUTS**

	Total area	Land area	Total pop	GDP in Euros and PPS**	Active population	Unemployment	Total pop. By sex-age 2005***	Pop. density
Albania	NUTS0,1,2,3		NUTS0 NUTS3(2001& 2004)	NUTS0***				NUTS0,1,2,3
Bosnia & Herzegovina	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)
Montenegro			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 Ta-

ble"). 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.

Diachronic comparisons

Concerning the time period covered, while the 2009 Database tables contained data for 2000-2006, we added (in the CC / PCC tables) in some cases, data for the year 2007, which will be integrated in the Database in a future phase of the project.

We have advanced in 2009 in the compilation and processing of the data for the years 1989-1990-1991-1992 in order to make some first comparisons between the years 1991 and 2001 (2002 for Serbia), however this work is not yet finalized; we will include its results in a future version of the Technical Report in 2010.

2. Assessment per country

Albania

***Spatial units' levels*:**

The ***total population of the country*** amounted up to ***3.069.275 inhab.*** in ***2001*** (census).

Albania is divided into 12 prefectures (counties, Albanian: official qark/qarku, but often prefekturë/prefektura), 37 districts and 351 municipalities.

Concerning the EU regulation for NUTS 3: 150 000 - 800 000 inh.; all Albania' prefectures, except two, have 150 000 - 800 000 inh. in 2001. Prefectures have also a Council and considerable competences.

Therefore, ***Albania' prefectures could be eventually 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

Existing data at "similar NUTS3" level

(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.

2009 Delivery

The data is only on 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 ***3.836.920 inhab.*** in ***2007*** (Official estimate of the population).

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 inh.; 6 FBiH' 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 cantons of FBiH could be eventually assimilated to NUTS3.

[2] It officially belongs to both, but is governed by neither, and functions under a decentralized system of local government.

Map BA.1: NUTS3 units, Population per NUTS3 2001

Population per NUTS3 2001

- 0 - 49999
- 50000 - 100000
- 100000 - 200000
- 200000- 500000
- 500000 - 1000000
- NUTS2
- NUTS3

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

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.

2009 Delivery

- Total area No Data
- Land area No Data
- Total population NUTS 0 (2000-2006), similar NUTS1,2,3 (2007)

- 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 "similar NUTS0" (2007), "similar NUTS3" (2007)

We used as:

NUTS1: the entire country Bosnia and Herzegovina

NUTS2: the Federation of Bosnia and Herzegovina (FBiH), the Republic of Srpska (RS) and the Brčko District

NUTS3: the 10 cantons of FBiH, the Republic of Srpska, and the Brčko District

These should be reassessed later on in cooperation with the NSO of the country (Bosnia and Herzegovina) etc.

Croatia

Spatial units' levels

The ***total population of the country*** amounted up to ***4.437.460 inhab.*** in ***2001.***

- 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 inh. (in 2001).

Existing data at NUTS3 level

(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 HR.1: NUTS2 units / NUTS3 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

2009 Delivery

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 ***2.022.547 inhab.*** in ***2002***.

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).

Existing data at NUTS3 level

(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

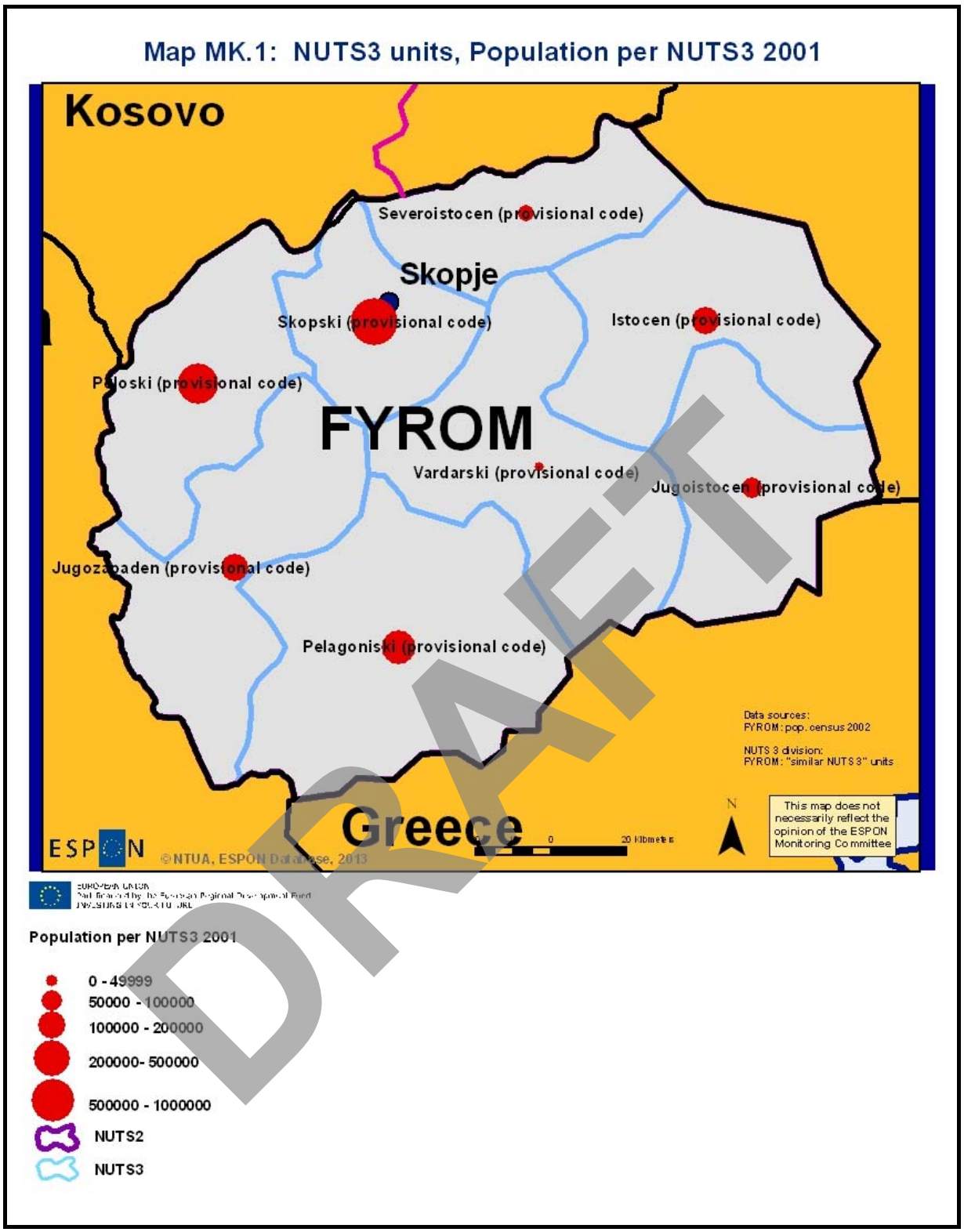
(b) Active population (total, distributions: per sex), number of employed and unemployed persons, employment per primary secondary and tertiary sector.

- From the population, *households and dwellings* census 2002

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.



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-inh. in 2002 (Table FY.1).

2009 Delivery

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 ***7.411.569 inhab.*** in ***2006*** (official estimate).

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 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 Low on Territorial Organisation, but are organised under the Government's Enactment of 29 January 1992.
In 2009, Serbian authorities adopted a law that formed seven new statistical regions in the territory of Serbia. The government is currently reviewing the decision and will probably amend the law and reduce the number of regions

Therefore:
- The two provinces (plus, eventually, the City of Belgrade) could be assimilated to NUTS 2 (and / or NUTS1).
- Districts could be reliably assimilated to NUTS3; 21 from the 25 districts had a population ranging between 150.000 and 800.000 inh, in 2002 (EU regulation limits for NUTS 3). The four remaining had 102.000-147.000 inhab. in 2002.

Most of the data concerning censuses of the population and building, specific surveys etc are aggregated and published on the level of *municipalities (LAU1).*

Existing data at "similar NUTS3" level

(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 is data from CLC2000 but there is not data on UMZ.

2009 Delivery

- 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 (eur, 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)

We assimilated the two parts of Serbia (Central Serbia and Vojvodina) to NUTS1 level as well as to NUTS2 level.

Map RS.1: NUTS2 units / NUTS3 units, Population per NUTS3 2001

Population per NUTS3 2001

- 0 - 49999
- 50000 - 100000
- 100000 - 200000
- 200000- 500000
- 500000 - 1000000
- NUTS2
- NUTS3

Montenegro

Spatial units' levels

The ***total population of Montenegro*** amounted in ***620.100 inhab.*** in ***2003,*** 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 inh. and only the municipality of Podgorica had more than 150.000 inh. (169.132) in 2003.

The country is divided in 21 *municipalities* which could be assimilated to *LAU1* level.

Alternatively: only the Municipality of Podgorica (with population over 150,00) could be assimilated to NUTS3.

Existing data at "similar NUTS3" level

(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.

2009 Delivery

- Total area No data
- Land area No data
- Total population NUTS 0 for 2000-2007
- 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 NUTS 0 for 2000-2007

Kosovo (Under UN Security Council Resolution 1244)

Spatial units' levels

The ***total population of the country*** amounted up to ***1.8-2.0 millions inhab. during the last years*** (different estimates are provided by Serbia and Kosovo).

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 inh., therefore ***the districts could be (difficultly) assimilated to NUTS 3 units.*** Municipalities could be assimilated to LAU1 level.

Existing data at "similar NUTS3" level

(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 is not data from CLC.

2009 Delivery

- 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 ***67.803.930 inhabitants*** in ***2000*** (census data), while according to an official estimate (Address Based Population Registration System) the country population amounted in ***70.586.260 inh.*** in December ***2007.***

Spatial units' levels:

Turkey, which 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 Turcish).

Seventy eight (78) of these last have (in 2000 and beyond) a population greater than 50.000 inhabitants.

Existing data at NUTS3 level

(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 is not data from CLC.

2009 Delivery

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

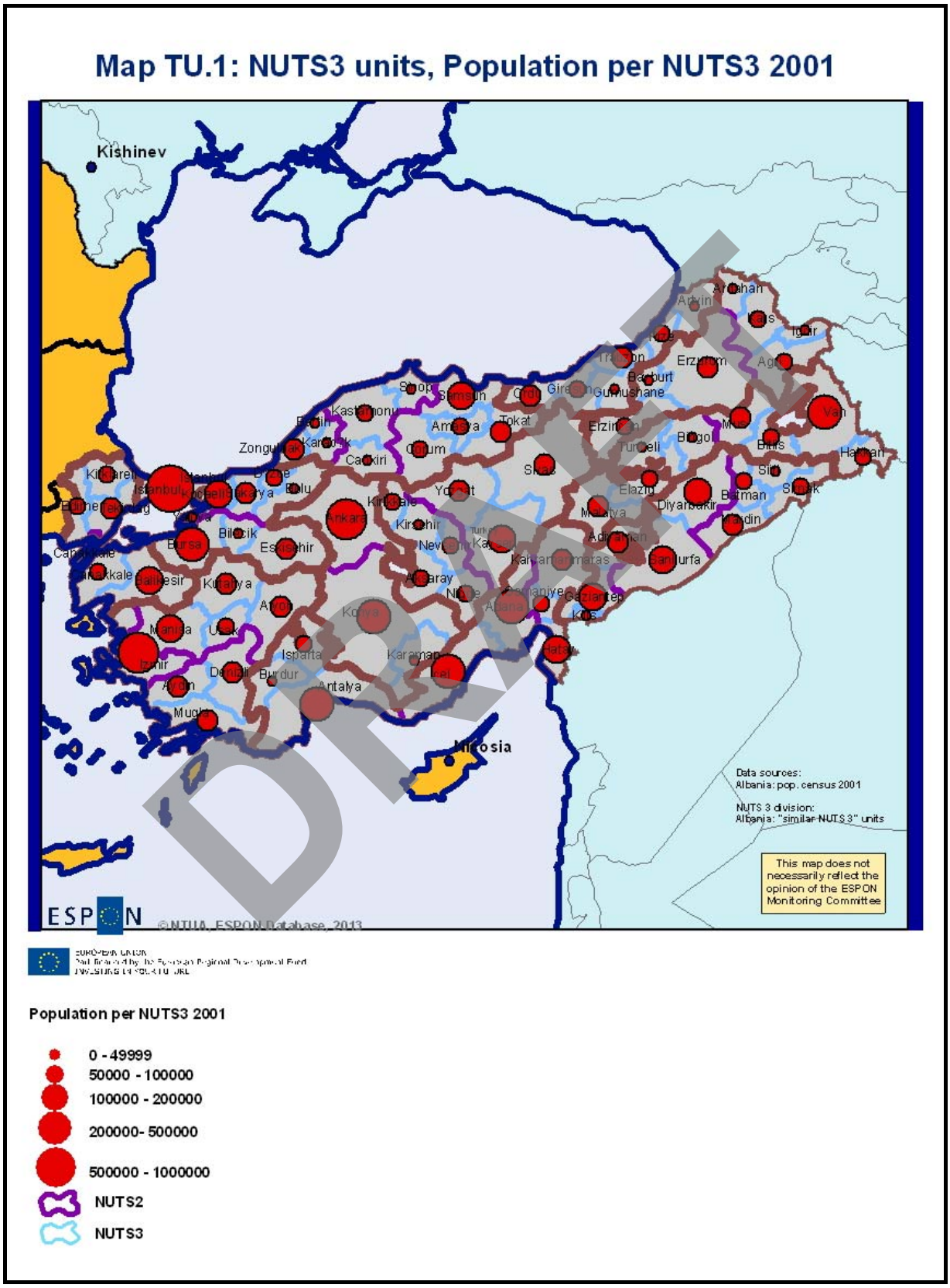
See in more detail in the "metadata" Excel table delivered to the LP

We used as:

NUTS1: the 12 Regions

NUTS2: the 26 Sub-regions

NUTS3: the 81 Provinces



3. Conclusions and work to be done in 2010

3.1 Data availability and quality at NUTS0 to NUTS 3 levels

(a) Compatibility of spatial administrative divisions with the EU NUTS classification

Turkey, Croatia and FYROM have already adopted this classification. For the rest WB countries, the criteria of the population weight (formal criterion) and the administrative capacity (informal criterion) are fulfilled in the majority of the existing administrative divisions (regions, districts etc) of these countries, therefore respective "similar NUTS" divisions could be used for the work on data without considerable problems.

(b) 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).

(c) 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, labor market. Some additional data for specific topics are provided by the NSO of these countries.

- It is less satisfactory for the *other Western Balkans;* relevant data are provided by the NSO.

In more detail, at NUTS3 level:

(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 CC.

3.2 Inclusion of the Western Balkans and Turkey in the scope of the ESPON Database

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 should be included at the moment in the Database.

3.3 Work to be done in 2010

Establishment of contacts and regular dataflow with the CC / PCC NSO, Eurostat and DG Regio

As it became clear from the above, the assessment of the availability and the quality of the data for the CC / PCC except Croatia, FYROM and Turkey, has almost advanced considerably on the basis of the data available in the official websites of the PCC NSO, other sources etc (a first basic set of data was included in the "2009 Delivery").

Further improvement of this work could be possible with the establishment of contacts and regular dataflow with the CC / PCC National Statistical Organisations (NSO), Eurostat and the DG Regio.

Especially this cooperation is useful for the following issues:

- Confirmation of the conformity of the administrative divisions of the CC / PCC (except Croatia, FYROM and Turkey)
- Clarification of the concepts used in some datasets provided by the NSO.
- Provision of some additional datasets.

We also assessed how it is possible to establish contacts with the National Statistical Organisations of these countries and ensure a regular dataflow among them and the ESPON 2013 Database Project.

A first possible strategy on this issue is to create a cooperation scheme among the ESPON Database project, the ESPON programme (MC, MA and CU), Eurostat, DG Regio and the CC / PCC NSOs. The LP has initiated the discussion on this issue with ESPON CU and Eurostat which has an official cooperation with the CC / PCC NSOs.

In case it appears that the implementation of this strategy is difficult, we could alternatively implement a second strategy of establishment of systematic non – official contacts between the ESPON Database project –and the ESPON programme- with the CC / PCC NSOs.

Evaluation of intermediate values

We will evaluate the values of several indicators (population etc) for the interval between the years for which we have official data.

Additional quality checks

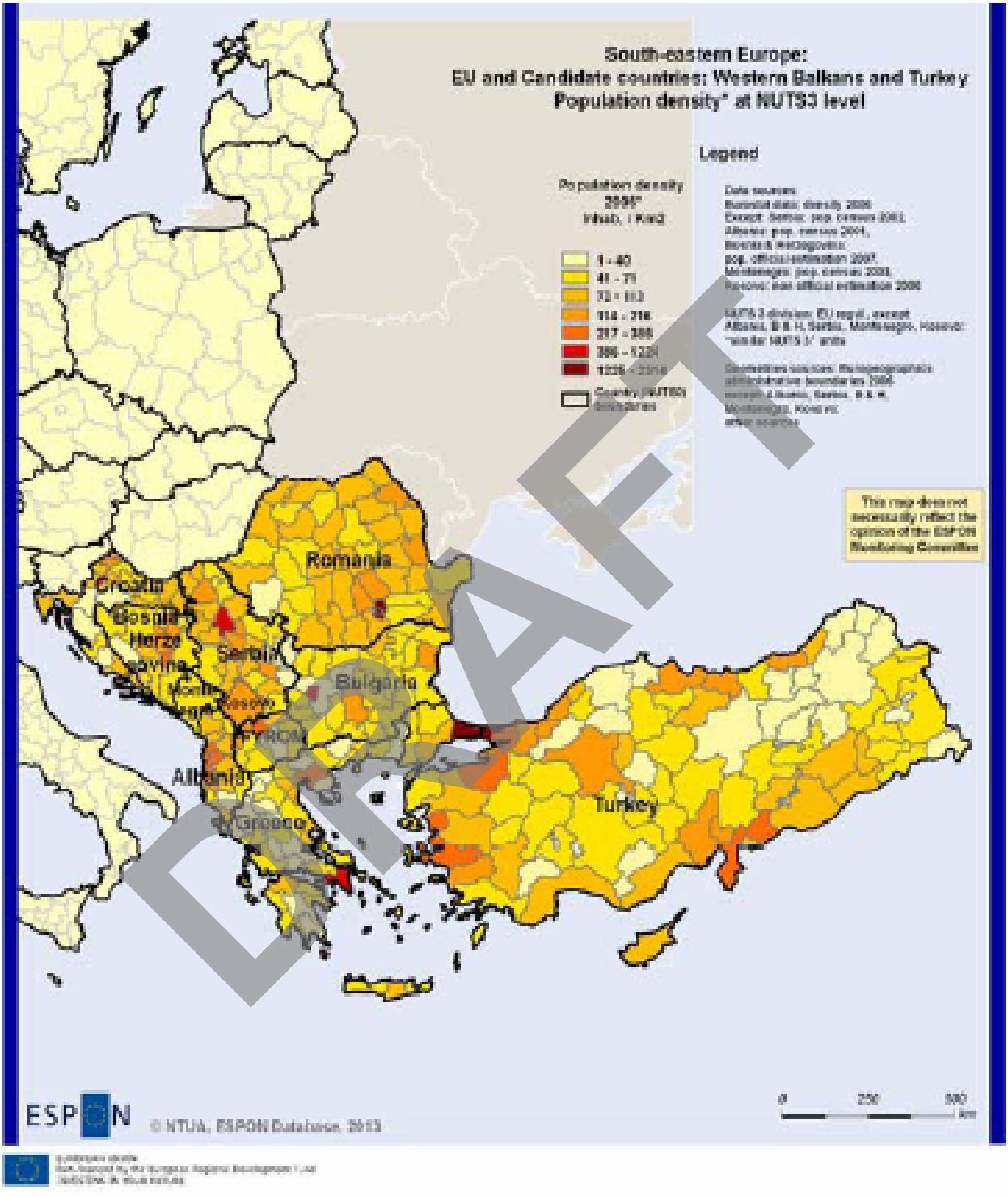
We have already compared the values of several indicators compiled from different sources. We will make additional checks/ comparisons, for example between the NSOs data and the respective ONU data.

Enlargement of the scope of the issues to be studied / Urban data

According to the enlargement of the content of the Database for the ESPON countries and the requests from other ESPON 2013 projects, we will collect statistics on similar issues for the CC/ PCC.

We will focus in particular on the urban data, making it possible to enlarge the urban database elaborated by ESPON 2006 project and further developed by ESPON 2013.

Annex -1 Maps



Map Annex.1: Population density in South-eastern Europe 2006

Annex 2 – Table 1: Western Balkans and Turkey available territorial data – from all sources

Data in "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 **households**								
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 from Eurostat / Short presentation

Albania, Bosnia and Herzegovina, FYROM, Serbia, Montenegro, Kosovo[3], 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)

[3] 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 from Eurostat / Detailed description

Albania, B and H, FYROM, Serbia, Montenegro, Kosovo[4], 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)

[4] 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 (2008), *Introduction to the NUTS and the Statistical regions of Europe*, http://ec.europa.eu/eurostat/ramon/nuts/introduction_regions_en.html as of 16.12.08

Eurostat (2008), *Statistical regions in the EFTA countries and the Candidate countries (CC) 2008* http://ec.europa.eu/eurostat/ramon/nuts/statistical_regions_en.html as of 16.12.08

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), *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), *Eurostat regional yearbook 2008*.

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 (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.*

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

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 Reports 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 (2008) 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.

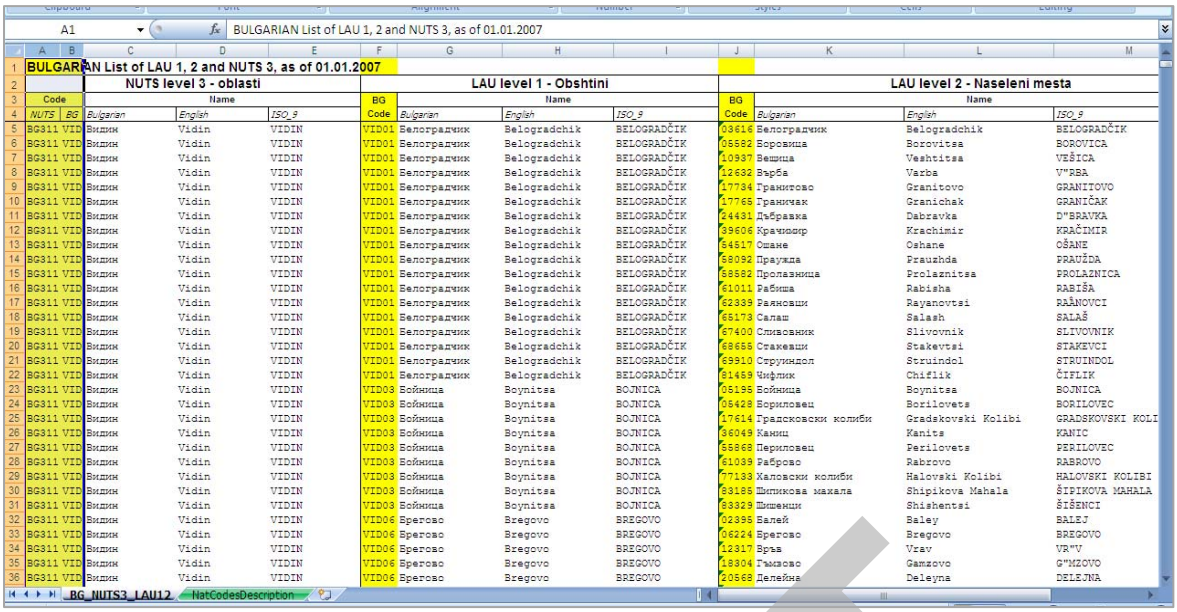
A second source of information used in our work consists in the official lists with spatial units (LAU1/2) in each country of the European Union. The list being available on the EUROSTAT website1, it's only a matter of proper downloading in order to get an image of the administrative organization of 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			Code	Name			Code	Name	
	NUTS	Bulgarian	English	ISO_9	BG	Bulgarian	English	ISO_9	BG	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

BG_NUTS3_LAU12 | NatCodesDescription

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.

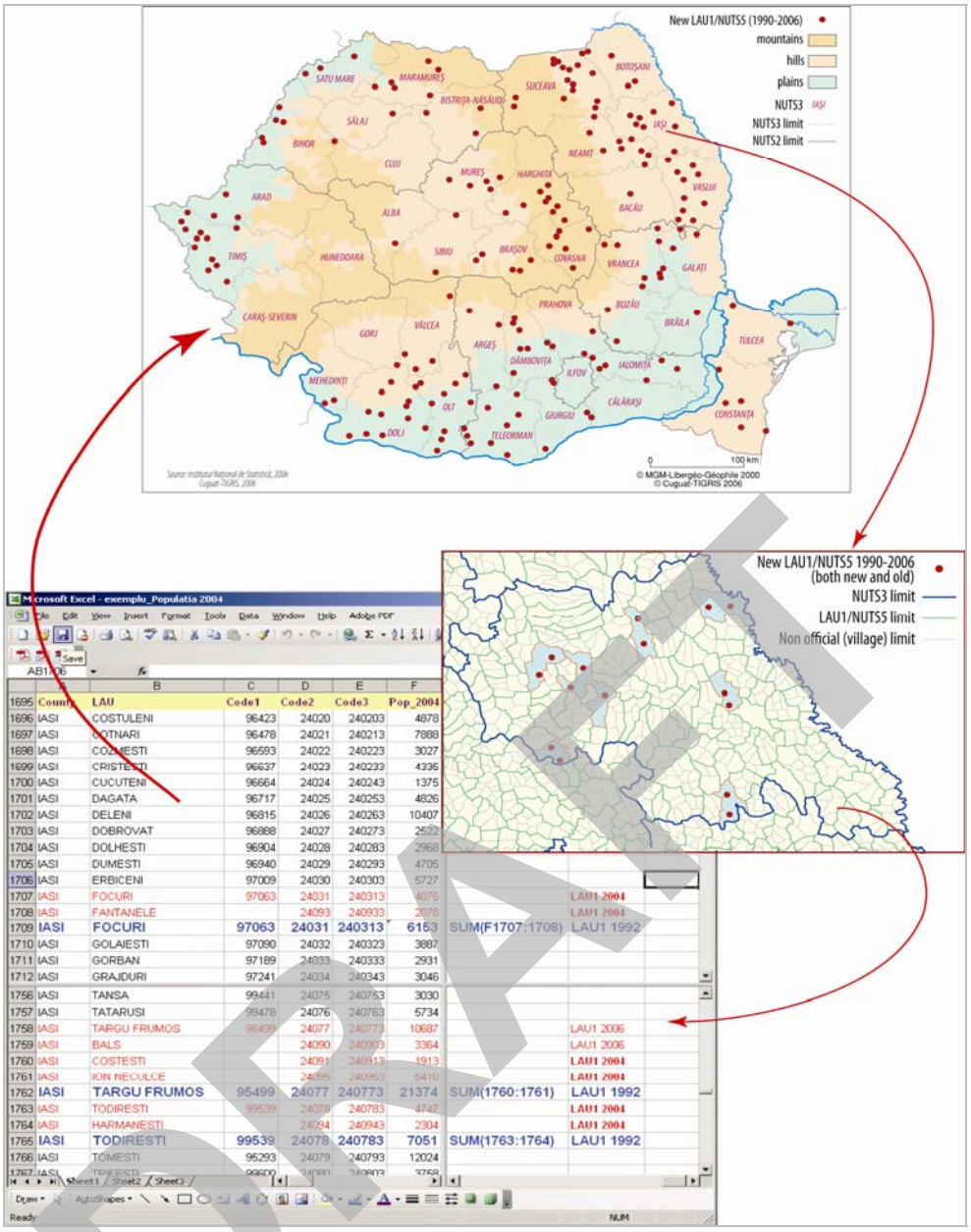


Figure 2: *New LAU2 [5] units in Romania (1990-2006)*

[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.

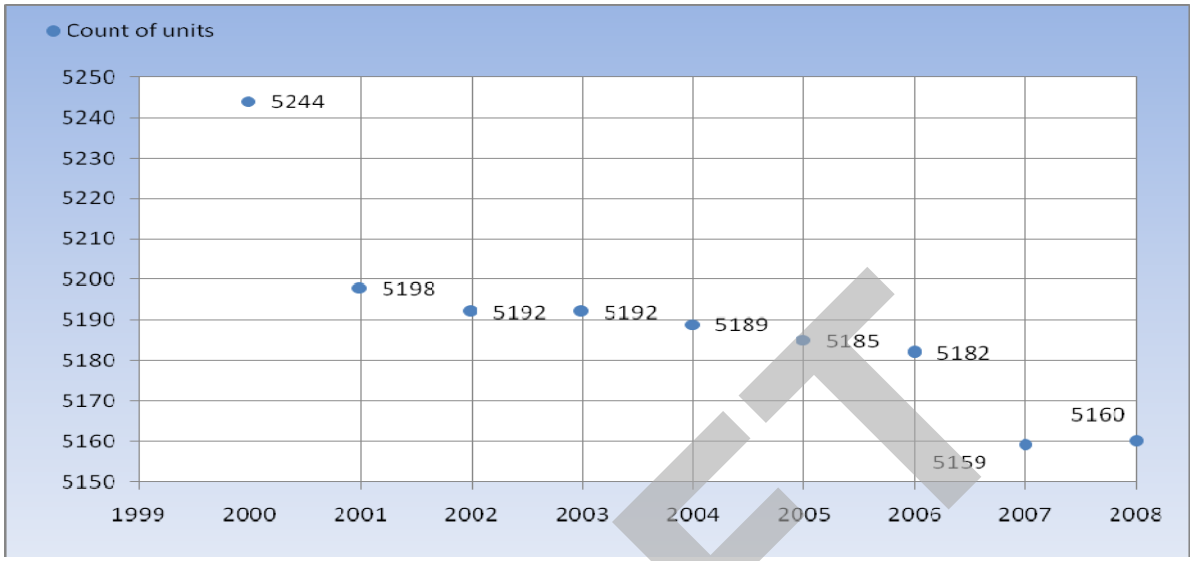


Figure 3: *The evolution of LAU2 in Bulgaria (x axis = time)*

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.

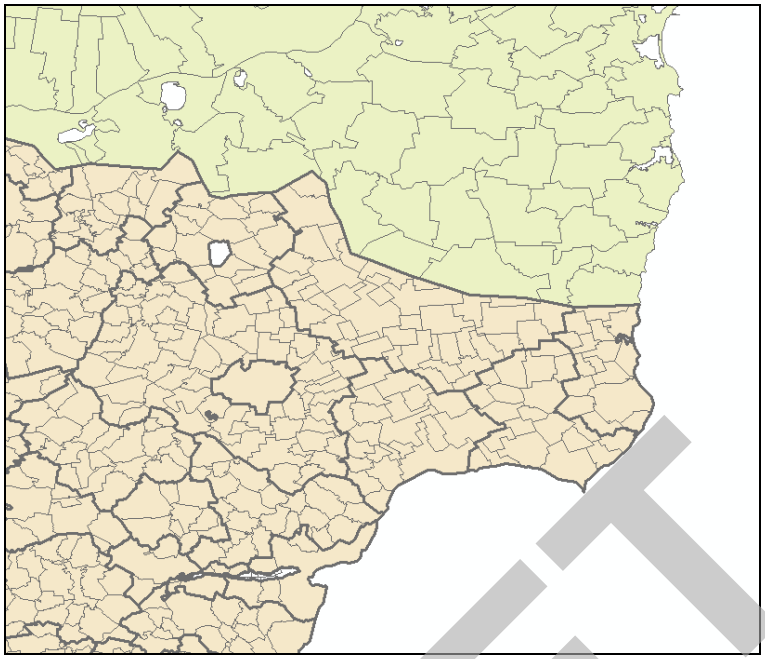


Figure no 4 – LAU2 in Romania and Bulgaria (LAU1)

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 useful[8] indicators.

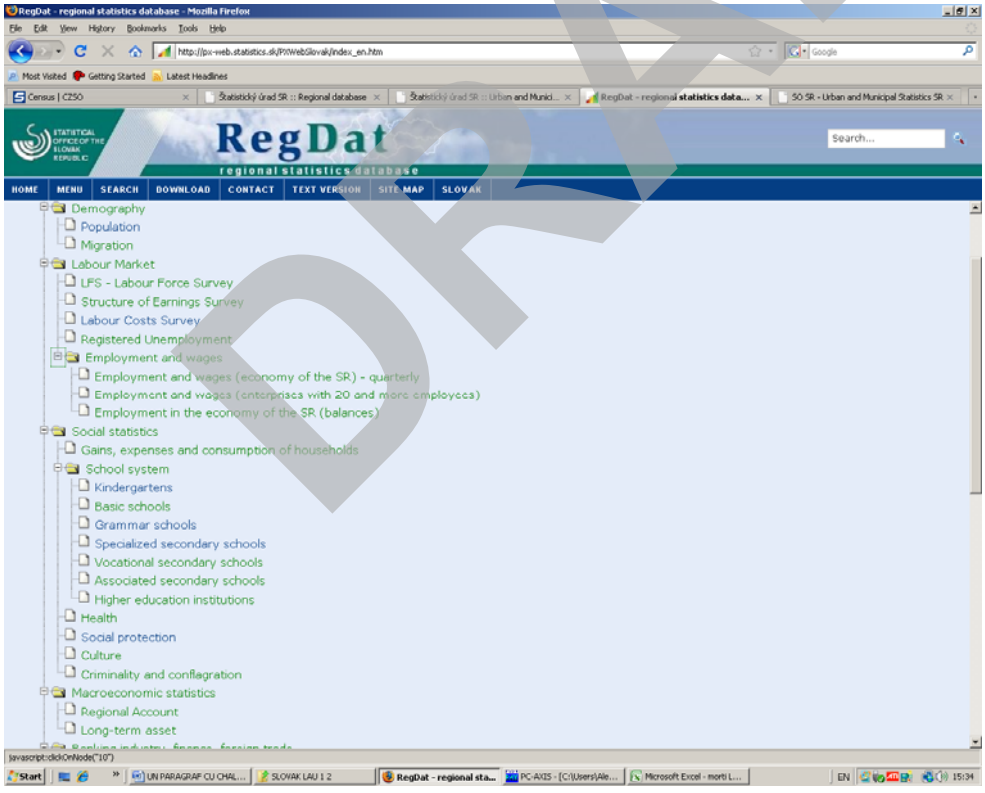


Figure 5: *The availability of statistical indicators in the case of the Slovak Republic - Sample view*

(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.

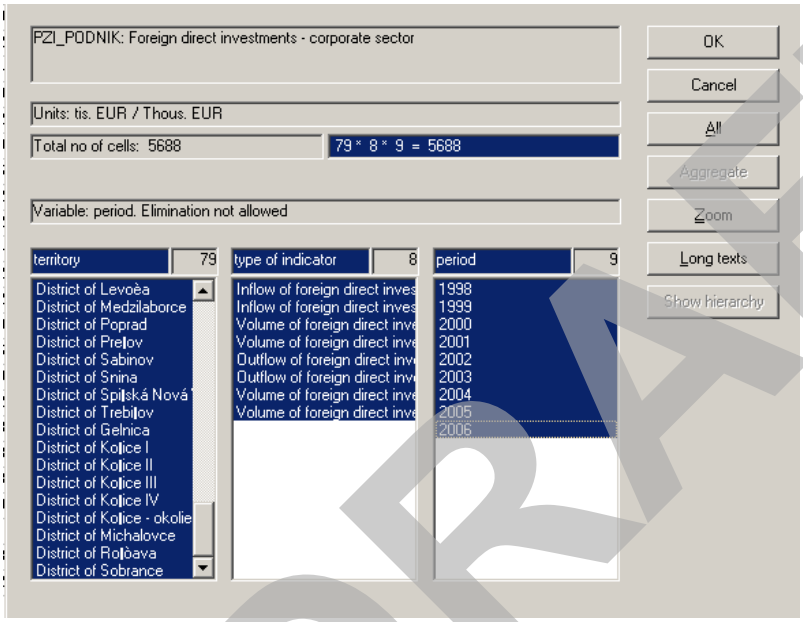


***Figure 6**: Foreign direct investments in Slovakia (LAU1 – 1998-2006, Pc-Axis software view)*

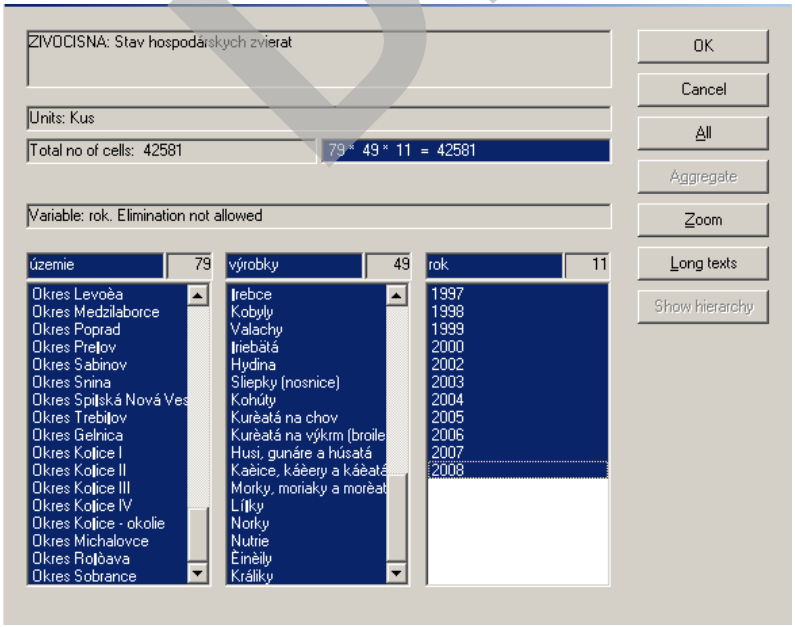
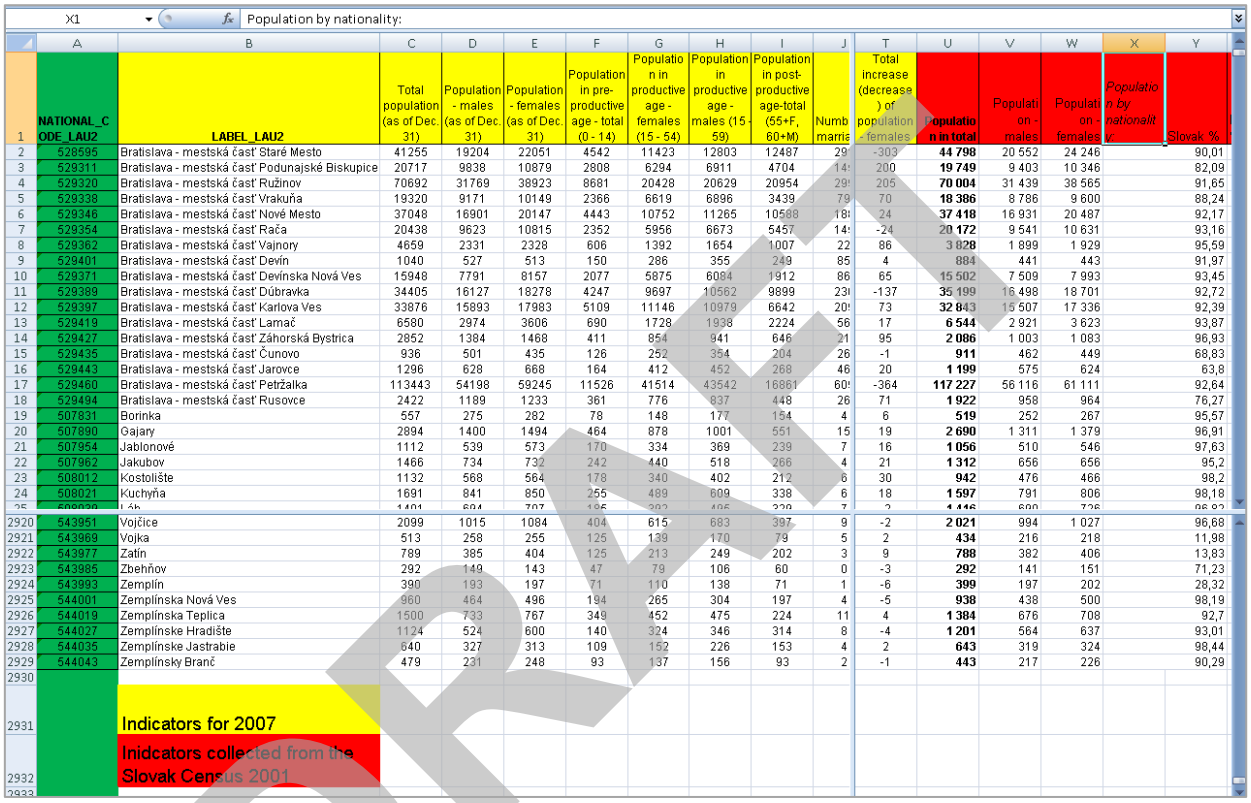


***Figure 7:** Stav hospodárskych zvierat by územie, výrobky and rok*

(**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.



X1 | Population by nationality:

	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 marria	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
25	508030	Láb	1401	694	707	186	393	495	329	7	-2	1 446	690	756		96,82
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.

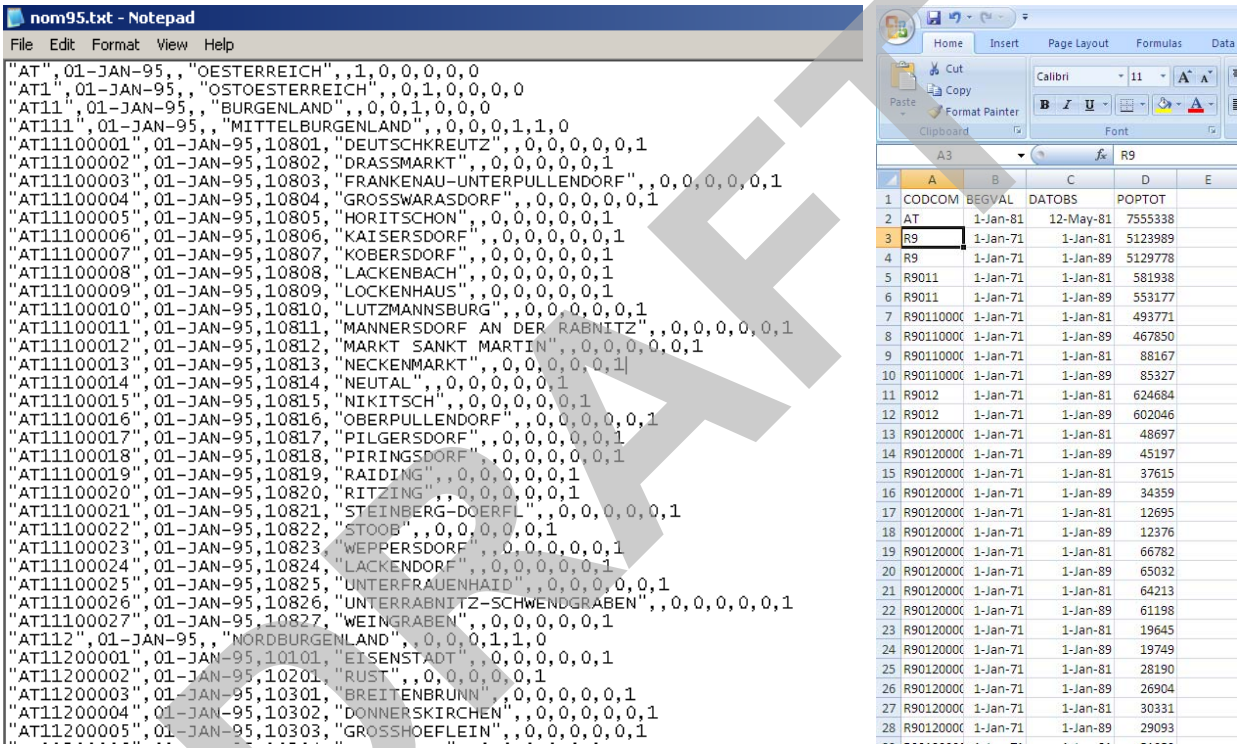


Figure 9: *SIRE database before (on the left – labels and codes) and after (on the right – population in 1991) data basic integration.*

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