ESPON 2013 DATABASE

QUALITY RATHER THAN QUANTITY…

FINAL REPORT –MARCH 2011

This report presents the final results of a Scientific Platform Project conducted within the framework of the ESPON 2013 Programme, partly financed by the European Regional Development Fund.

The partnership behind the ESPON Programme consists of the EU Commission and the Member States of the EU27, plus Iceland, Liechtenstein, Norway and Switzerland. Each partner is represented in the ESPON Monitoring Committee.

This report does not necessarily reflect the opinion of the members of the Monitoring Committee.

Information on the ESPON Programme and projects can be found on www.espon.eu

The web site provides the possibility to download and examine the most recent documents produced by finalised and ongoing ESPON projects.

This basic report exists only in an electronic version.

List of authors

UMS RIATE (FR)
Claude Grasland*
Maher Ben Rebah
Ronan Ysebaert
Christine Zanin
Nicolas Lambert
Bernard Corminboeuf
Isabelle Salmon

LIG (FR)
Jérôme Gensel*
Bogdan Moisuc
Marlène Villanova-Oliver
Anton Telechev
Christine Plumejeaud

UAB (ES)
Roger Milego
Maria-José Ramos

IGEAT (BE)
Moritz Lennert
Didier Peeters

TIGRIS (RO)
Octavian Groza
Alexandru Rusu

Université du Luxembourg (LU)
Geoffrey Caruso
Nuno Madeira

UMR Géographie-Cités (FR)
Anne Bretagnolle
Hélène Mathian
Marianne Guerois
Liliane Lizzi
Guilhain Averlant
François Delisle
Timothée Giraud

National University of Ireland (IE)**
Martin Charlton
Paul Harris
Stewart Fotheringham

National Technical University of Athens (GR)**
Minas Angelidis
Gabriella Karka
Epameinondas Tsigkas
Kostas Santimpantakis
Vivian Bazoula

Umeå University (SE)**
Einar Holm
Magnus Strömgren

UNEP/GRID (CH)**
Hy Dao
Andrea De Bono

* Scientific coordinators of the project
** Expert

TABLE OF CONTENT

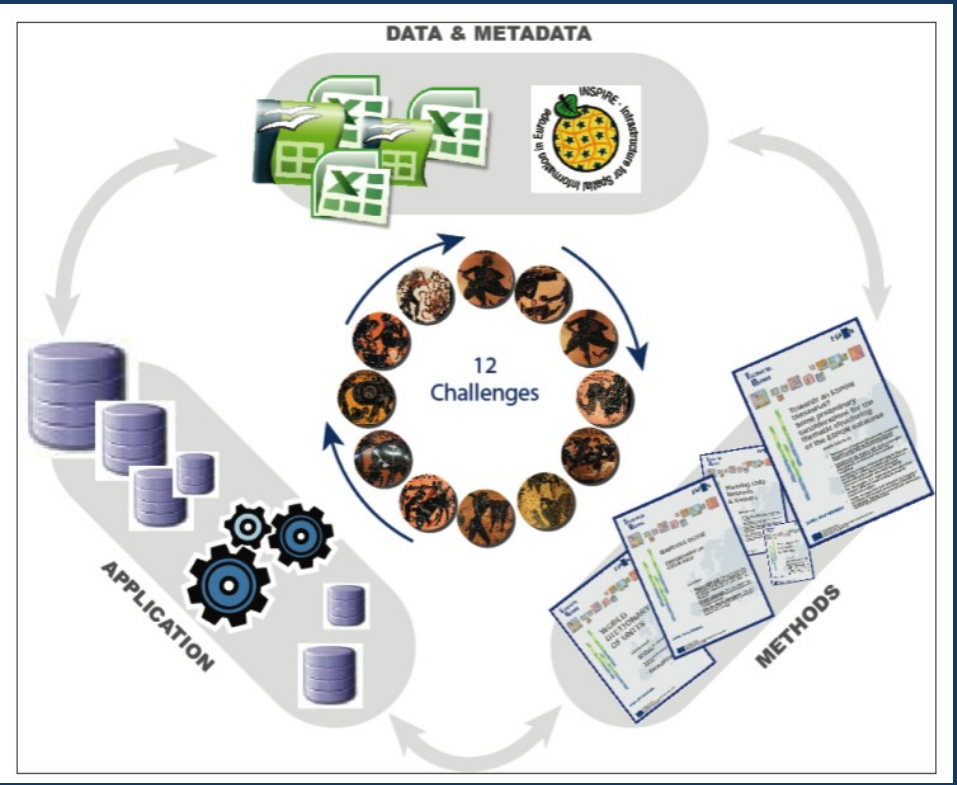
FOREWORDS

The document we deliver here is called the FINAL REPORT.
He that outlives this FINAL REPORT, and comes safe home,
Will stand a tip-toe when the PROJECT is named,
And rouse him at the name of ESPON 2013 DATABASE.
He that shall live this FINAL REPORT, and see old age,
Will yearly on the vigil feast his neighbours,
And say "I WAS IN ESPON 2013 DATABASE PROJECT"
Then will he strip his sleeve and show his scars.
And say "These wounds I had on ESPON DATABASE."
Old men forget: yet all shall be forgot,
But he'll remember with advantages
What feats he did in ESPON 2013 DATABASE: then shall our names.
Familiar in his mouth as household words
RIATE, LIG-STEAMER, UNIVERSITIES OF BARCELONA AND LUXEMBOURG
GEOGRAPHIE-CITES, TIGRIS, NTUA, NCG, UMEA, UNEP, IGEAT
Be in their flowing cups freshly remember'd.
This REPORT shall the ESPON CU teach his NEW PROJECTS;
And ESPON DATABASE 2013 PROJECT shall ne'er go by,
From this day to the ending of the world,
But we in it shall be remember'd;
We few, we happy few, we band of brothers;
For he to-day that sheds his blood with me
Shall be my brother; be he ne'er so vile,
This day shall gentle his condition:
And researchers in European Union now a-bed
Shall think themselves accursed they were not here,
And hold their manhoods cheap whiles any speaks
That fought with us upon ESPON DATABASE FINAL REPORT.

With Special thanks to William S. for inspiration.
Original version available at http://pagesperso-orange.fr/rhetorique.com/azincourt.htm

INTRODUCTION

A division of work in 12 challenges has been the core of the project since the beginning. These challenges provided a simple and efficient division of work between partners and experts, each of them being responsible for one challenge, possibly in association with others. But challenges had also to be integrated in a more synthetic way in the second part of the project, which can be illustrated on the figure below by the three work areas defined as Methods, Application, Data and Metadata.



1. Data and metadata. The amount of data present in the ESPON database is the most obvious output of a project called "Database". It is also the easiest way to evaluate progress made at ESPON level because it includes both basic data collected by ESPON DB project itself, and other data collected by all ESPON projects. But it is important, in our opinion, to insist on the fact that **metadata are probably more important than data themselves.** More precisely, it is not useful to enlarge the ESPON Database if data is not very accurately described (definition, quality, property copyrights). We acknowledge that the elaboration of such metadata was not an easy task, both for the ESPON DB project and for other ESPON projects and we apologized for that at the Malmö seminar. But we are convinced that, without this collective effort, the sustainability of the ESPON program will not be ensured.

2. Methods, presented in the form of standalone booklets called Technical Reports, are the necessary complement of data and metadata. They represent the second major contribution of the ESPON DB project. In the 12 challenges, we have explored a great number of options that could enlarge the scope of data collected and used in the ESPON project. This chunk of knowledge was produced by the ESPON DB project itself with many inputs from other ESPON projects dealing with specific geographical objects (e.g. FOCI for urban and local data; Climate Change and RERISK for Grid Data; DEMIFER or EDORA for time series at NUTS2 or NUTS3 levels; Priority 2 projects for local data). Technical Reports focus on questions that are regularly asked in ESPON projects and try to summarize collective knowledge. Some Technical Reports provide clear solutions. Some identify shortcomings or dead-ends. Others focus on questions of cartography, in particular the mapping guide elaborated by RIATE that has been made available on the ESPON website.

3. Applications are different computer programs elaborated by project partners for data management, data query or data control. It is important to understand that the ESPON database is not made of a single application doing everything, but of a set of interlinked applications with several purposes in the data integration process. Many misunderstandings appeared in the beginning of the project in relation with this issue and many efforts were made to clarify the vocabulary. A basic distinction has to be made between an interface for query that is now available on the ESPON website and an application for data management. The second one is the "back office" of the interface but it also fulfills more general objectives of data integration. These two major applications are designed and implemented by the Computer science research team LIG, but it is important to note that other partners and experts of the project contributed to this work. In particular, the UAB team has contributed to the elaboration of the metadata editor with LIG. It has also developed the OLAP program for NUTS to GRID conversion. The UL team has adapted a specific program of text mining for the elaboration of ESPON Thesaurus. The experts of NCG have developed application for outlier detection in R language.

The Final report of the ESPON 2013 Database project is therefore not limited to the present document but involves all the above mentioned material (technical reports, applications, data). What we try to present here is a short guide for accessing to this whole set of resources. We have divided this report in two parts:

- **Part 1 Application** presents the software oriented elements produced by the project and also some conceptual elements that drive the software implementation.
- **Part 2 Thematic** presents the different technical reports elaborated in order to improve the scope of the ESPON database in terms of space, time, scale, geographical objects and fields of policy action.

1.APPLICATION

INTRODUCTION – PART 1

The first part of this report presents the software oriented elements produced within the ESPON 2013 Database Project. This concerns not only software elements (e.g. the different components of the ESPON DB Application) but also conceptual elements (e.g.architecture, schemas) that drive the software implementation.

The first section of this part gives a brief overview of the ESPON DB Application and dataflow. The following sections describe, in their respective order, the different phases of the ESPON DB dataflow. Section 1.2 describes the upload phase (i.e. the ESPON DB metadata profile and editor). Section 1.3 follows the different stages of the data checking process. Section 1.4 offers some insights about the storage phase, what are the databases and ontologies that lay behind the ESPON Database Application. Section 1.5 shows the query and download phase, performed by the end users via the Web Download interface.

The next two sections shed more light on the coding scheme (1.6) and the thematic classification (1.7) which are of crucial importance for structuring the ESPON 2013 Database and making available the information for end users.

Then, the section 1.8 shows the methodology used for building the ESPON OLAP Cube which allows to combine information described on grid (Corine Land Cover) and socio-economic data in the NUTS nomenclature.

Finally the section 1.9 presents the different map-kits available for ESPON Projects, from local case studies to the World. On top of that, some basic rules of cartography are described in order to ensure harmonisation of maps in the ESPON Program.

1.1. THE ESPON DB APPLICATION AND DATAFLOW

The ESPON 2013 Database Application is a complex information system dedicated to the management of statistical data about the European territory, spanning over a long period of time. The overall architecture relies on two databases: one is used for storing ontological data, and the other, called the ESPON Database, is meant to be queried by end-users. The latter only is made accessible to users through Web interfaces (see figure on the right, above) that each corresponds to the four main functionalities offered by the ESPON 2013 Database Application: registration, administration, upload of data and metada, query and retrieval of such data and metadata.

The ESPON DB Application data flow describes the path followed by both data and metadata from the moment they are entered in the ESPON DB Application, until they are output as answers to queries expressed by end-users. Four phases are identified along this data flow:

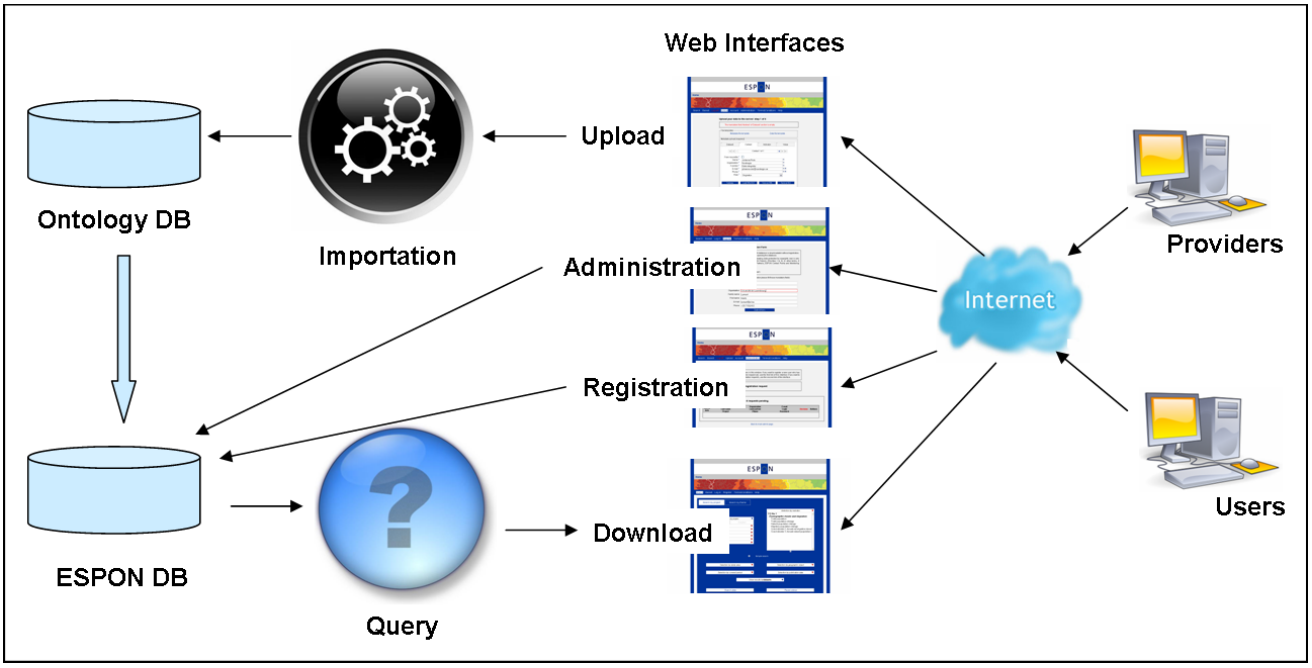
1. The upload phase is handled by the upload Web interface through which users (here, data providers) are guided in the preparation and the transfer of both their data and metadata files to the ESPON Database server. During this phase, users are helped in providing well formated and Inspire compliant metadata through the ESPON Metadata Editor. This phase is described in more detail in section 1.2.

2. Then, the checking phase follows; it aims at validating both data and metadata files provided by users before they are stored in the ESPON Database. The checking process alternates between automatic and manual steps performed either by the application itself or by the expert members of the ESPON DB 2013 Project. If some of the errors detected cannot be corrected or need some additional information and precisions, then both data and metadata files are sent back to providers in order to be fixed. When the checking phase succeeds, then the validated data and metadata files are ready to be stored in the ESPON Database. This phase is described in more detail in section 1.3.

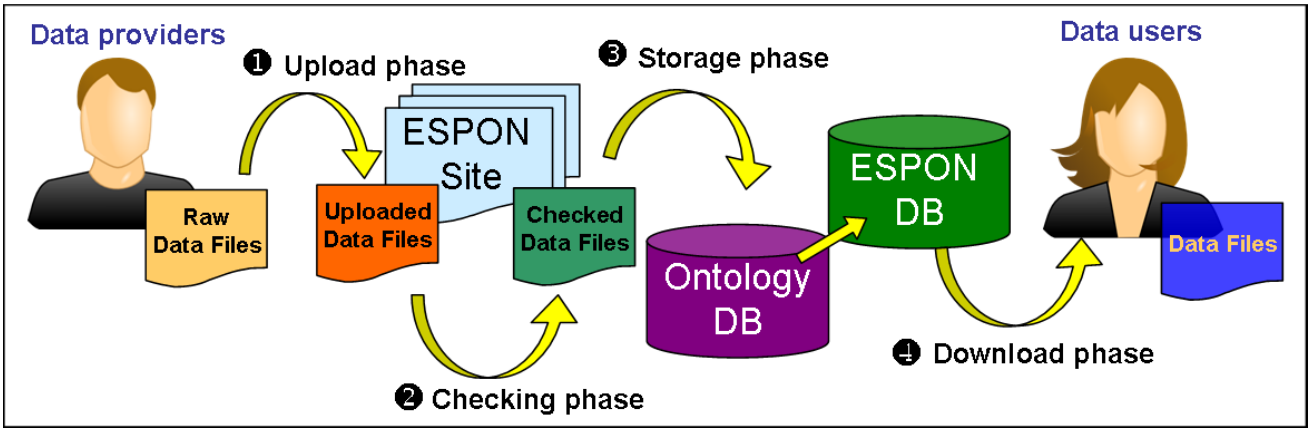
3. The storage phase deals with the management and the maintenance of both data and metadata in the ESPON Database. Flexible database schemas have been designed and built for handling long term storage of statistical and spatial data, considering that both data and metadata may evolve while stored in the ESPON Database, as a result of harmonization and gap filling processes. This phase is described in more detail in section 1.4.

4. During the download phase, end-users of the ESPON DB Application are invited to explore, search and retrieve both data and metadata through a Web interface. Free data and metadata can be accessed and downloaded by any end-user, while data and metadata subject to copyright restrictions are made available for authorized and registered users only. This phase is described in more detail in section 1.5

The ESPON DB Application Architecture And Data Flow



The ESPON DB Application relies on a Web-based architecture, including two databases (ontology DB and ESPON DB) for long term storage of statistical and spatial data. Data providers and end-users interact with the EPSON DB (register, upload files, query data and download files) via Web based interfaces.



The ESPON DB Application data flow allows receiving data from ESPON Projects (acting like data providers) and returning these data to other ESPON Projects (acting as data consumers). The intermediate phases allow checking and improving data quality and are performed without any interaction with the users.

1.2. THE UPLOAD PHASE

Data and metadata files entered by data providers (mainly ESPON Projects) have to be compliant with the ESPON DB data and metadata formats so that they can be uploaded on the ESPON DB Application server.

The ESPON DB metadata profile has been created because an indepth analysis of the state of the art has revealed that, so far, there is no standard metadata profile aimed at describing statistical territorial data. Indeed, existing spatial data standards (ISO 19115, the INSPIRE directive) offer very detailed description profiles for spatial data, but thematic and statistical descriptions of data are insufficient. The ESPON DB metadata profile covers 3 main purposes:

- Preserving the compatibility with the existing standards (by INSPIRE, ISO) by integrating the same main elements in the profile.
- Minimizing the quantity of work data providers have to do when filling metadata by, for instance, inferring automatically metadata from the associated data when possible (e.g. temporal or spatial coverage).
- Providing sufficient information about the content of data (indicators) and about their origin, by including indicator level and value level descriptors in the profile.

The Web metadata editor is an interactive application, which assists data providers in the creation of data descriptions compliant with the ESPON DB metadata profile. The editor can be used to create a new metadata file, or to edit and modify an existing one. It handles, opens, and saves files in both XML and XLS formats. It guides a data provider in filling the three categories of descriptors covered by the metadata profile:

1. Information about the dataset as a whole: contact information, dataset title and abstract, etc.
2. Information about each indicator in the dataset: name, description, indicator methodology, thematic classification, etc.
3. Information about each value in the dataset: the primary source of each individual value, the estimation or correction methods applied to it, the copyright constraints associated with it, etc.

The editor checks and underlines syntactical errors found in metadata and provides dropdown lists that ease the time consuming but valuable task of filling data description (e.g. for personal information, already described indicators, etc.).

The Metadata Profile And Editor

dataset

Dataset information	
name	Age Structure Data
upload date	27/04/2010
last update date	27/04/2010
Metadata point of contact	
name	Johanna Roto
email	johanna.roto@nordregio.se
organization	Nordregio
function	Data integrator
role	Originator

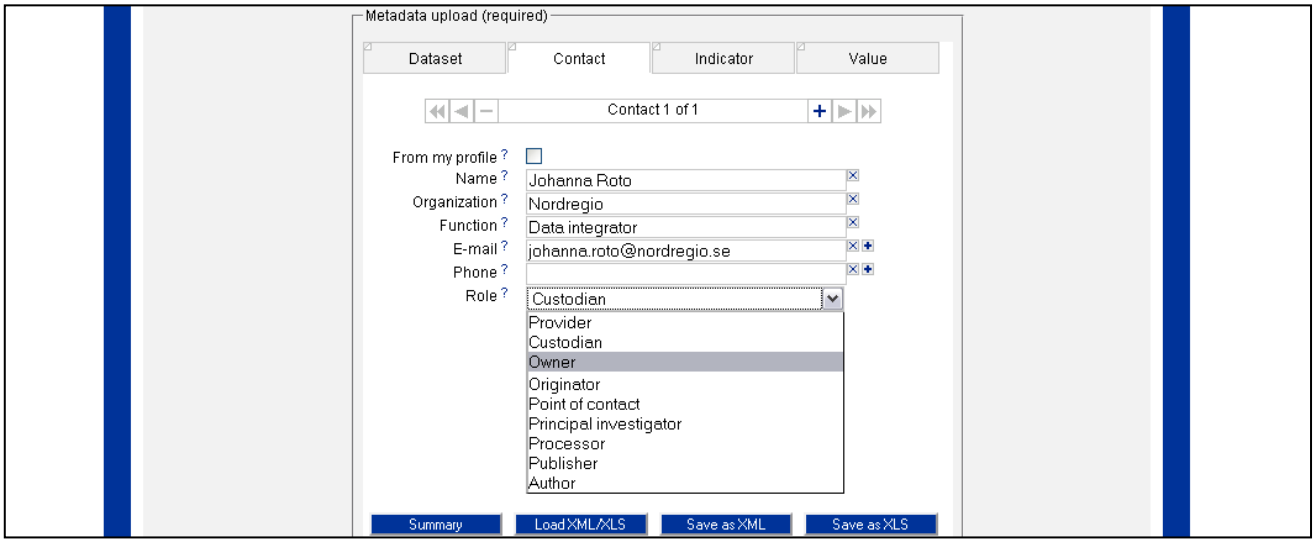
indicator

Identification		
code	CH_2039	
name	Change in Population Aged 20-39	
units	%	
abstract	Change in population aged 20-39 years, annual average change	
methodology		
classification		
theme		Demography
keywords		Population

value

Scope				
	label	1		
	lineage			
		provider	EUROSTAT	
		date	4/2010	
		URL	http://epp.eurostat.ec.europa.eu/por	
		methodology		
		methodology URI		
	reliability			
		estimation		FALSE
		quality	high	
	constraints			
		public data access		TRUE
		public metadata access		TRUE
		copyrights	EUROSTAT	

NUTS2	TYPE	ver	CH_2039	scope
validity_start			2001	
validity_end			2005	
AT11	NUTS2	2006	-1.47	1
AT12	NUTS2	2006	-1.20	1
AT13	NUTS2	2006	0.74	1
AT21	NUTS2	2006	-1.87	1
AT22	NUTS2	2006	-0.93	1
AT31	NUTS2	2006	-1.14	1
AT32	NUTS2	2006	-0.84	1
AT33	NUTS2	2006	-0.63	1
AT34	NUTS2	2006	-0.62	1
BE10	NUTS2	2006	1.39	1
BE21	NUTS2	2006	-0.81	1
BE22	NUTS2	2006	-1.30	1
BE23	NUTS2	2006	-1.06	1
BE24	NUTS2	2006	-1.15	1
BE25	NUTS2	2006	-1.52	1
BE31	NUTS2	2006	-0.44	1
BE32	NUTS2	2006	-0.99	1
BE33	NUTS2	2006	[illegible]	1
BE34	NUTS2	2006	-0.46	1
BE35	NUTS2	2006	-0.61	1
BG31	NUTS2	2006	-1.11	1
BG32	NUTS2	2006	0.15	1
BG33	NUTS2	2006	0.05	1
BG34	NUTS2	2006	-0.08	1
BG41	NUTS2	2006	0.27	1
BG42	NUTS2	2006	0.13	1



The ESPON DB metadata profile (upper figure) contains information about the dataset as a whole, about each individual indicator and about each individual value. Metadata and data files are strongly linked; all indicators and scopes described in the metadata file must be present in the data file, and *viceversa*. Metadata can either be provided in the shape of formatted Excel files, or created through theWeb Metadata Editor (lower figure), which adds the benefits of automaticly filling data and checking syntactical errors.

1.3 THE CHECKING PHASE

In order to insure data input in the ESPON DB are error-free, the data and metadata files are first subject to a thorough process of checking. The checking process is fourfold:

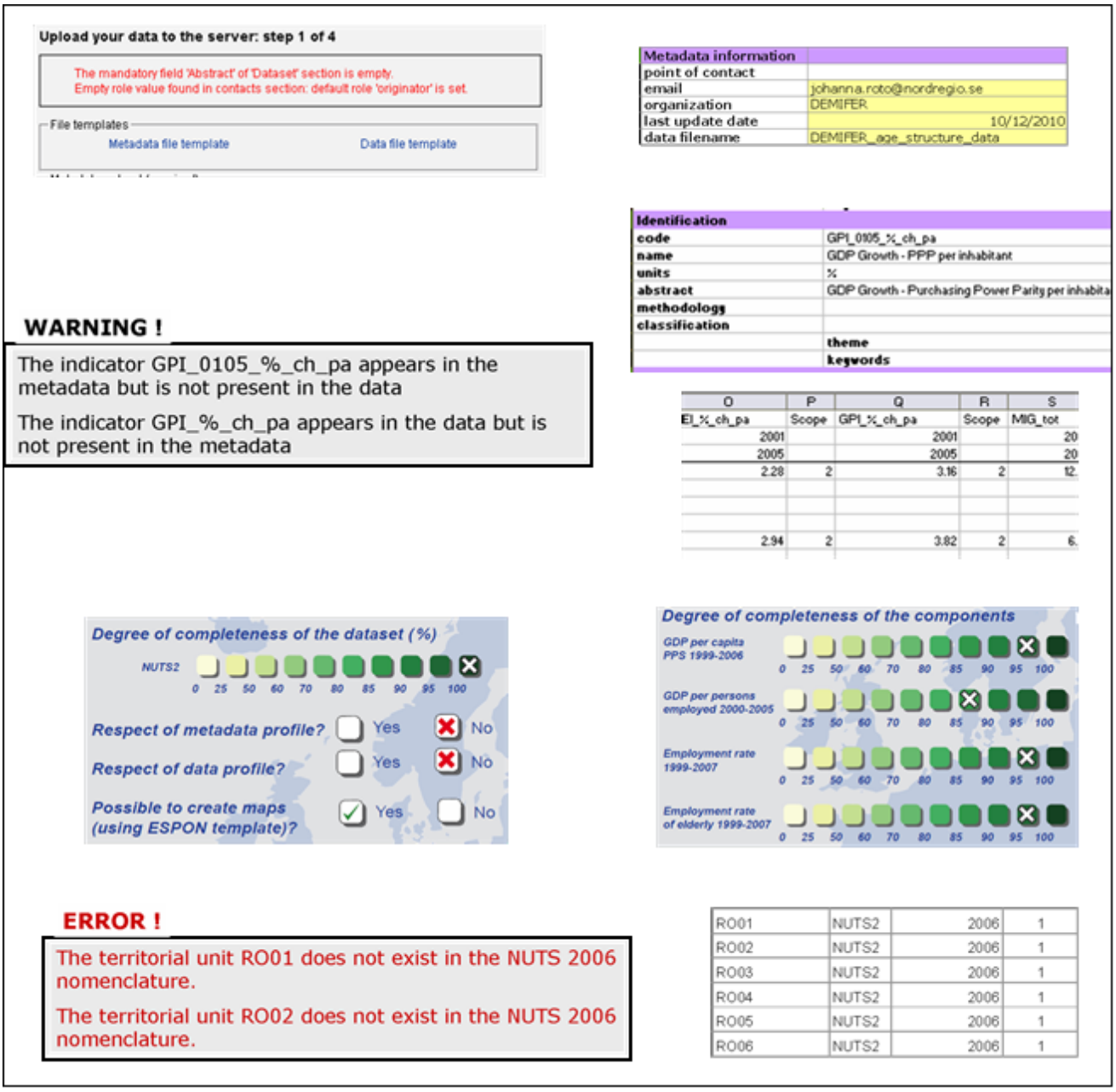
1. The syntactic checking is an automated process that aims at finding and correcting syntactical errors in both data and metadata. It is launched when providers upload their data and metadata files through the metadata editor. There are four categories of errors to be corrected: empty mandatory fields, format errors (e.g. when indicator values are text instead of numbers), typing errors (e.g. when typing the names of metadata descriptors) and data/metadata correspondence errors (e.g. when indicators described in the metadata are not present in the data or *viceversa*). During this phase, the application interacts with the user, and then it is possible to solve all syntactical errors before uploading files to the ESPON DB server.

2. The thematic checking is a manual process performed by thematic experts (i.e. lead partner RIATE), which consists in assessing the thematic relevance and completeness of the dataset related to the studied topic and assessing the compliance of the metadata and data submitted with regard to ESPON requirements. In this phase, the thematic expert assesses whether the indicators and values present in the dataset are well described, whether the geographical completeness of the dataset (i.e. the covered area) is satisfactory, whether the statistical completeness is sufficient (e.g. the NUTS resolution is sufficient for describing the phenomenon or if a finer NUTS level should be sought). Obviously, there can be no automatic correction for the thematic shortcomings, so if a dataset is considered as unsatisfactory, the data provider is required to make the necessary adjustments.

3. The outlier checking is an automated checking phase aimed at detecting possible errors in individual indicator values. A set of statistical, spatial and temporal analysis methods is triggered by ESPON DB thematic experts in order to find outliers, values that are potentially incorrect. Outliers may result either from data manipulation errors, or from exceptional, but correct, values. The difference between the two cases is established by a human expert. If some value errors are detected, the data provider may be required to make the necessary adjustments.

4. The final checking is performed when data and metadata are included in the database by the acquisition tools. If the acquisition is successful, that means that all the integrity constraints of the database are satisfied. This phase consists in checking the consistency of the dataset with itself, but also against the rest of the data already stored in the database. Additional data (especially, spatial and thematic ontologies) helps in detecting whether false entities exist in the dataset (e.g. inexistent territorial units), or if duplicated entities appear in the dataset (e.g. the same indicator with different names), or if ambiguous entities are present (e.g. different indicators having the same name, code or abstract).

Outputs Of The Data Checking Process



An illustration of different types of errors and outcomes of the checking process. On the first row, see two missing metadata fields reported by the metadata editor. On the second row, a mismatch of indicator code between the data and the metadata file, reported by the upload interface. On the third row, fragments of data quality and completeness assessments, reported by thematic experts. On the fourth row, detection of territorial units assigned to the wrong NUTS version, reported by the acquisition tools upon importation in the megabase.

1.4 THE STORING PHASE

The ESPON DB Application uses two databases for the long term storage of statistical data. The separation is done in order to obtain an application optimized for two different (and conflicting) purpose:

- The ontology database is based on a conceptual schema optimized for data harmonization. This conceptual schema imposes more separation between entities, and separation implies more effort at query time (thus, query processing performance is decreased).
- The ESPON DB is based on a snapshot schema optimized for query performance in the Web interface. The data are structured in such a way that fast query answer is privileged (see a short explanation in the figure to the right, below).

The ESPON DB Application also integrates a standalone Java application that allows inserting the content of paired data and metadata files into the ontological database.

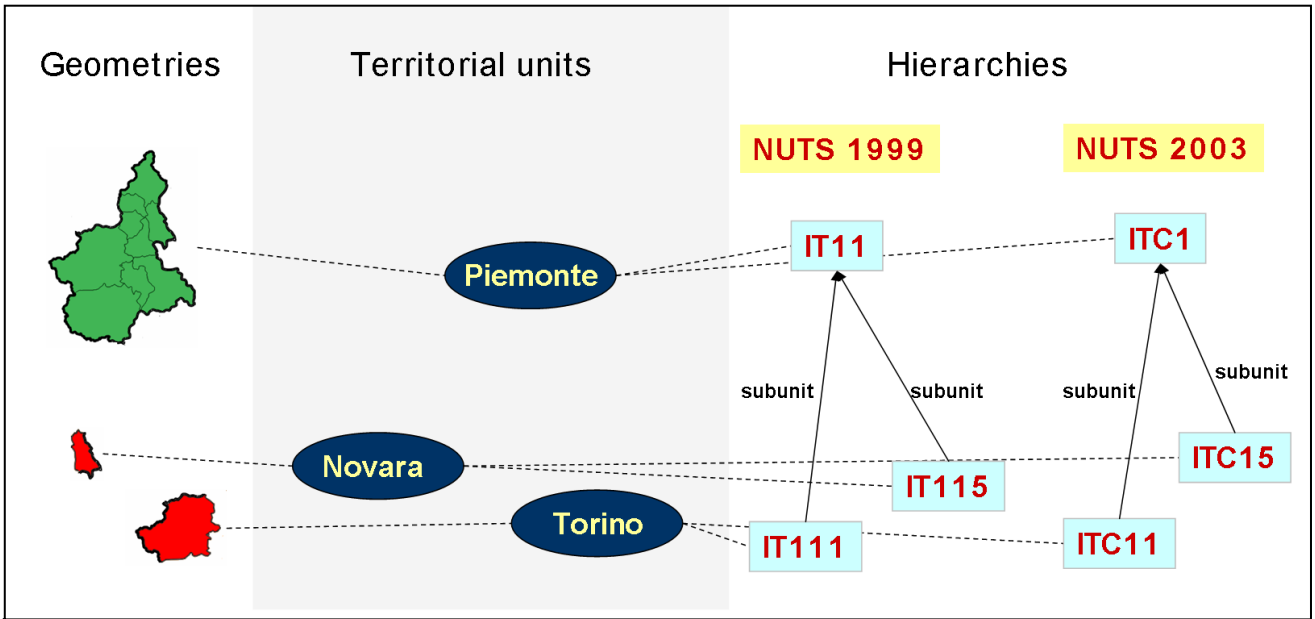
In order to enforce data consistency, this ontology database contains two ontologies, a spatial ontology (dictionary of territorial units and changes, see the figure on the right, above for a small example) and a thematic ontology (a dictionary of indicators).

Relying on such ontologies makes it possible to detect fake entities (e.g. a territorial unit code that doesn't exist in a given NUTS revision), duplicated entities (e.g. two codes for the same indicator) and ambiguous entities (e.g. the same code for two different indicators). The existing spatial ontology covers NUTS data and follows the evolution of the different NUTS versions from NUTS 1995 to NUTS 2006. In order to insure database consistency, this ontology is to be extended to higher levels (world/neighbourhood) but also, as much as possible, towards lower levels (local). The thematic ontology (see Indicator coding and classification section for more clarifications) aims at giving a comprehensive dictionary of indicators stored into the ESPON DB.

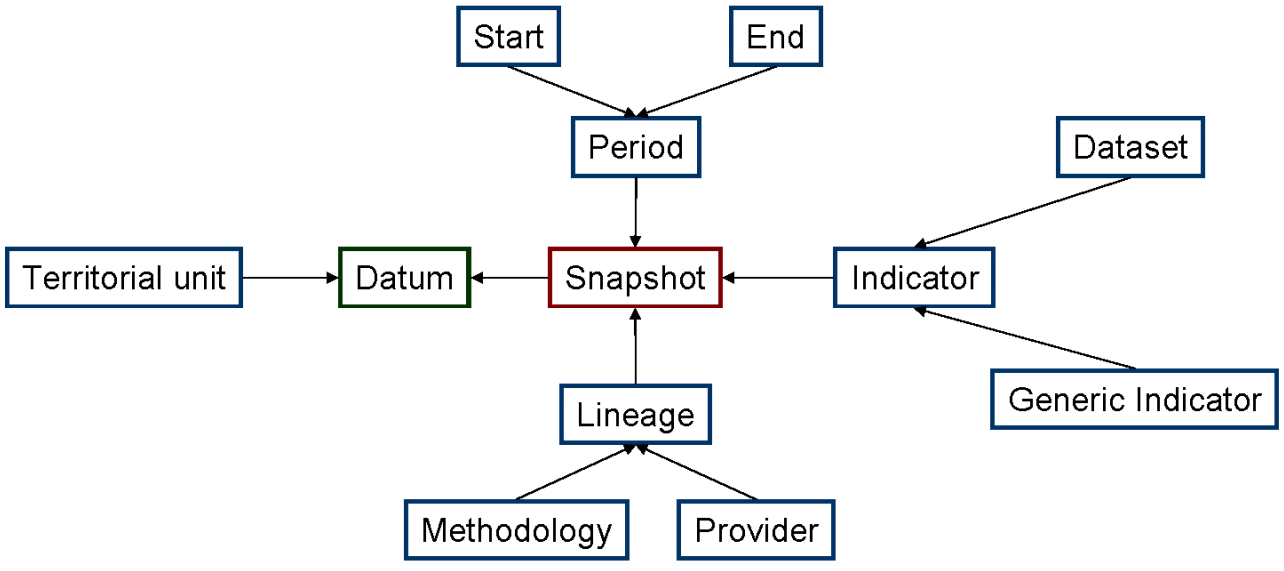
Data and metadata that have been made consistent and harmonized in the ontological database are transferred towards the ESPON DB. The ESPON DB is a PostgreSQL database implementing a schema targeted at offering high, scalable performance for online exploration and querying of big data quantities (see the figure to the right, below, for a brief presentation of the schema). It is designed for storing thematic or environmental data associated with discrete spatial divisions (e.g. NUTS and similar, LAU, etc.).

The schema of the ESPON DB allows storing and retrieving all the content described by the metadata profile. Additionnally, it integrates a user management facility, required for differentiating access to free and copyrighted data.

ESPON DB Application Databases And Ontologies



The spatial ontology makes a clear separation between territorial units and territorial division hierachies. One territorial unit can be part of many hierachies and it may have a different code in each hierarchy. Within each hierarchy, it can have different "subunit" relations with other units. Every attribute (name, geometry, indicators) can evolve in time. This allows a very clear view of territorial division changes.



The ESPON DB schema is optimised for fast querying and for reduced database size. On this simplified representation, we can see how three of the four dimensions of an indicator value (*datum* table) have been merged. Introducing the "snapshot" table allows more than halving the size of the datum table (which is the main table of the database, holding millions of records). It also allows fastening queries, by introducing an additional indexing level.

1.5 THE DOWNLOAD PHASE

MAIN FEATURES

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

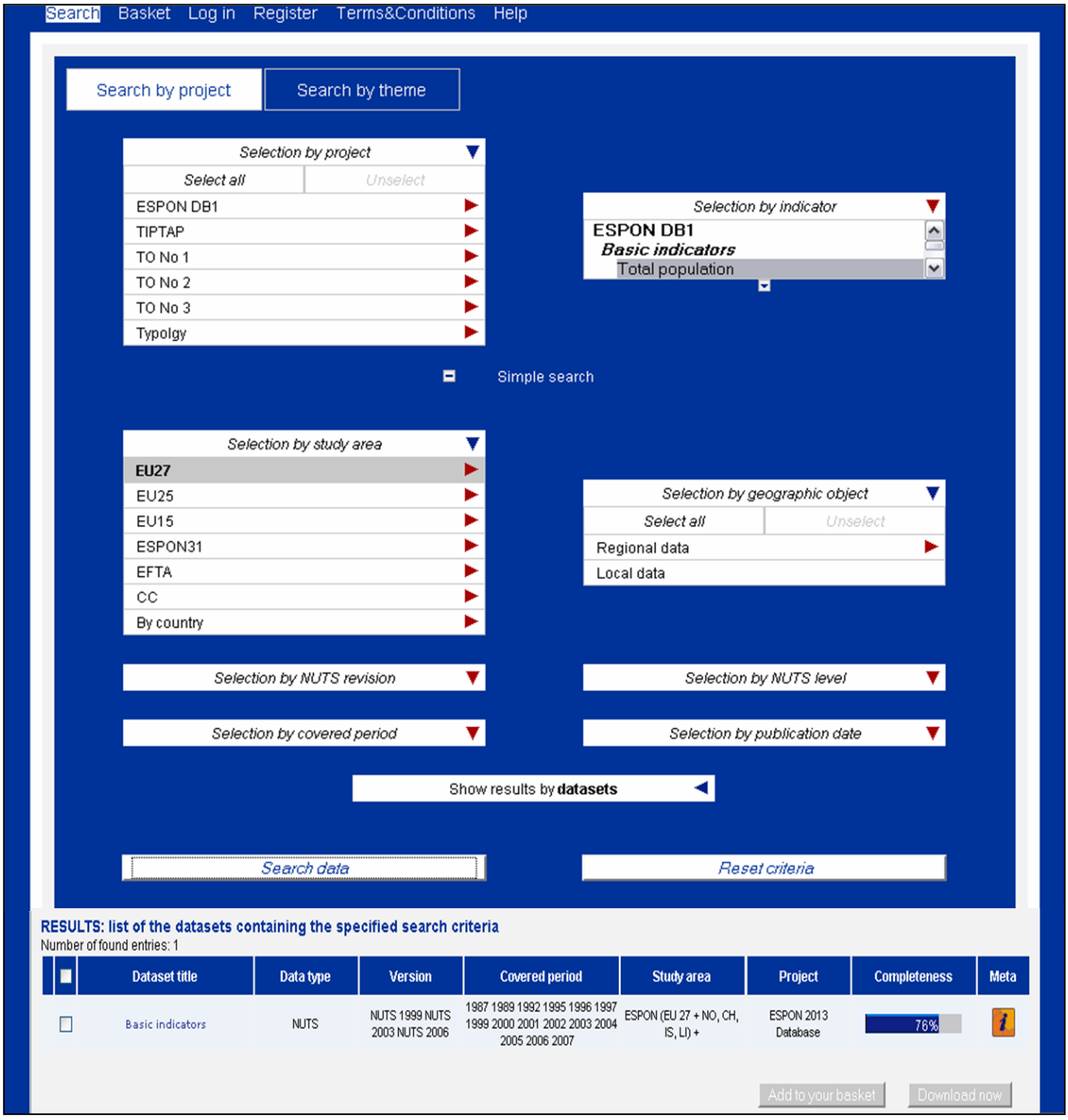
- The use of a server-side application cache system allows the application to avoid querying the database for all browsing tasks excepting the advanced search. This insures fast data searching, whatever the database size.
- The use of an XML exchange format for the answer to queries allows decreasing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

The Web download interface (see figure to the right) allows users to search and explore data in two ways: either by project (data provider and dataset) or by theme. In each type of search, an advanced mode is also available, allowing users to add more research and filter criteria: study area (country groups or countries), covered time period, object type (nomenclature versions and levels), and publication date. The search results can be listed as datasets or as individual indicators.

The table of results that is generated as a first answer can be further filtered in order to match yet better the user's needs, by removing unwanted indicators, territorial units, years or versions. Selected search results can be progressively added to a basket as in most of e-commercial Web applications. The basket can be downloaded at the end of the session, under the form of a zip file containing all the datasets selected by the user.

The table of results lets the users see the completeness of the dataset as a whole and also by nomenclature level, under the shape of a percentage bar (see figure). The interface also gives the possibility to the users to consult all the metadata related to the dataset. The three levels of metadata can be viewed: dataset, indicator and value levels. The completeness can be displayed by nomenclature level on a map.

The Web Query And Download Interface



The Web Query and Download interface allows users to formulate two types of basic search: 'by project' and 'by theme'). For each basic search, advanced search criteria can be added. This search interface is dynamic: search criteria lists are expanded only if they are used. On the example, two additional search criteria have been added, (study area: "EU 27" and geographic object type "NUTS and similar"). The Web Query and Download interface has been optimized so that building complex queries takes as little space as possible in the Web browser.

The web interface contains all regional data delivered by ESPON Projects (e.g in March 2011: EDORA, DEMIFER, Climate Change, ReRisk, TIPTAP, Accessibility Update, Demography Update, Lisbon Update, Telecomm Update, the Typology Compilation and some indicators of the ESPON 2006 Database.

1.6 CODING SCHEME

KEY FINDINGS

- The harmonisation of coding schemes is of crucial importance for the ESPON 2013 DB. With this regard, TPGs involved in applied research projects are increasing the level of ambiguity when put into practice their own scheme to code indices, indicators and other measures
- To a certain extent, coding schemes are not used to express the content of data but rather an attempt to homogenise codes. However, some information needs to be provided and, most importantly, it needs to be arranged in a consistent way to avoid conflicts with the web-based user interface
- Despite the diversity of approaches to code data, standards used by ESPON projects were taken into account in the analysis that allowed the creation of the coding scheme

DESCRIPTION

The coding scheme has been elaborated in the context of the ESPON 2013 DB project to provide TPGs with a unique code. Against this background, research teams are encouraged to apply a scheme that comprises three fields. The information to be added in each field corresponds to the subject, restrictions and/or derivations, and level of measurement. Other elements that might be used to classify data should not be considered as they already appear in the metadata file (e.g. time, space).

The procedure is not constrained to a limit of characters, but it is important to respect the above-mentioned structure. As a consequence, the first field should integrate information about the subject. The second part refers to widely used abbreviations that impose restrictions and/or use derivations. Ultimately, the third field specifies the level of measurement so that users can understand the statistical operations that have been carried out on the data. In ascending order of precision, the different levels of measurement are nominal, ordinal, interval, and ratio.

For each field, a non-exhaustive list of acronyms and abbreviations is provided to encourage harmonisation. In some cases, adaptations will be necessary, especially to obtain more degree of freedom when facing rather complex, but similar, data. The coding scheme has been implemented and tested for datasets delivered by the first round of ESPON projects under Priority 1, 2, and 3.

Additional improvements will be needed to further increase the quality of this proposal. At this point, it is not possible to anticipate many of the indices and indicators that will be delivered. That will require the involvement of the ESPON research community through a continuous, dynamic process.

Illustrative examples of harmonised coding schemes

(a)

Subject(s)								Derivations / Restrictions							Level of measurement					
m	**i**	**g**	**.**	**p**	**o**	**p**	**_**	**c**	**h**	**.**	**t**				**_**	**r**	**t**	**c**			

(b)

Subject(s)								Derivations / Restrictions							Level of measurement					
a	**c**	**c**	**.**	**a**	**i**	**r**	**_**	**a**	**b**	**s**					**_**	**r**	**t**	**e**			

(c)

Subject(s)								Derivations / Restrictions							Level of measurement					
e	**d**	**u**	**.**	**s**	**c**	**d**	**_**	**t**							**_**	**r**	**t**	**c**			

(d)

Subject(s)								Derivations / Restrictions							Level of measurement					
p	**o**	**p**					**_**	**2**	**0**	**-**	**3**	**9**	**.**	**t**	**_**	**r**	**t**	**c**			

(e)

Subject(s)								Derivations / Restrictions							Level of measurement					
C	**O**	**2**	**.**	**r**	**o**	**d**	**_**	**v**	**o**	**l**					**_**	**r**	**t**	**e**			

(f)

Subject(s)								Derivations / Restrictions							Level of measurement					
t	**y**	**p**					**_**	**r**	**u**	**r**	**a**	**l**			**_**	**n**	**o**	**c**			

The following examples provide a better understanding of the rationale behind the coding scheme, where (a) reflects 'Migratory population change', (b) 'Potential accessibility by air [absolute level]', (c) 'Persons with secondary education degree', (d) 'Population aged 20-29 years', (e) 'CO2 emissions by road traffic', and (f) 'Typology of rural regions'. Each field of the coding scheme should be separated by the underscore symbol. In addition, it suggests a number of cells to be filled in by TPGs.

1.7 THEMATIC STRUCTURE

KEY FINDINGS

- Database structures adopted by international organisations with in-house data constitute an important source of information. Therefore, we apply a visual grouping technique to illustrate, by means of correlation matrices, homogeneous clusters of words that identify those themes.
- The rationale for sub-themes derives from text mining methods. We assume that the ESPON 2006 Programme introduced new vocabulary. This assumption is investigated by extracting keywords from a large corpus of textual data. In order to improve the interpretation of the results, we employ visualisation tools of data co-occurrence to understand similarities.
- The results obtained suggest that the ESPON 2013 DB should be structured in 7+1 themes and 29 sub-themes (available in **Annex 2**).

DESCRIPTION

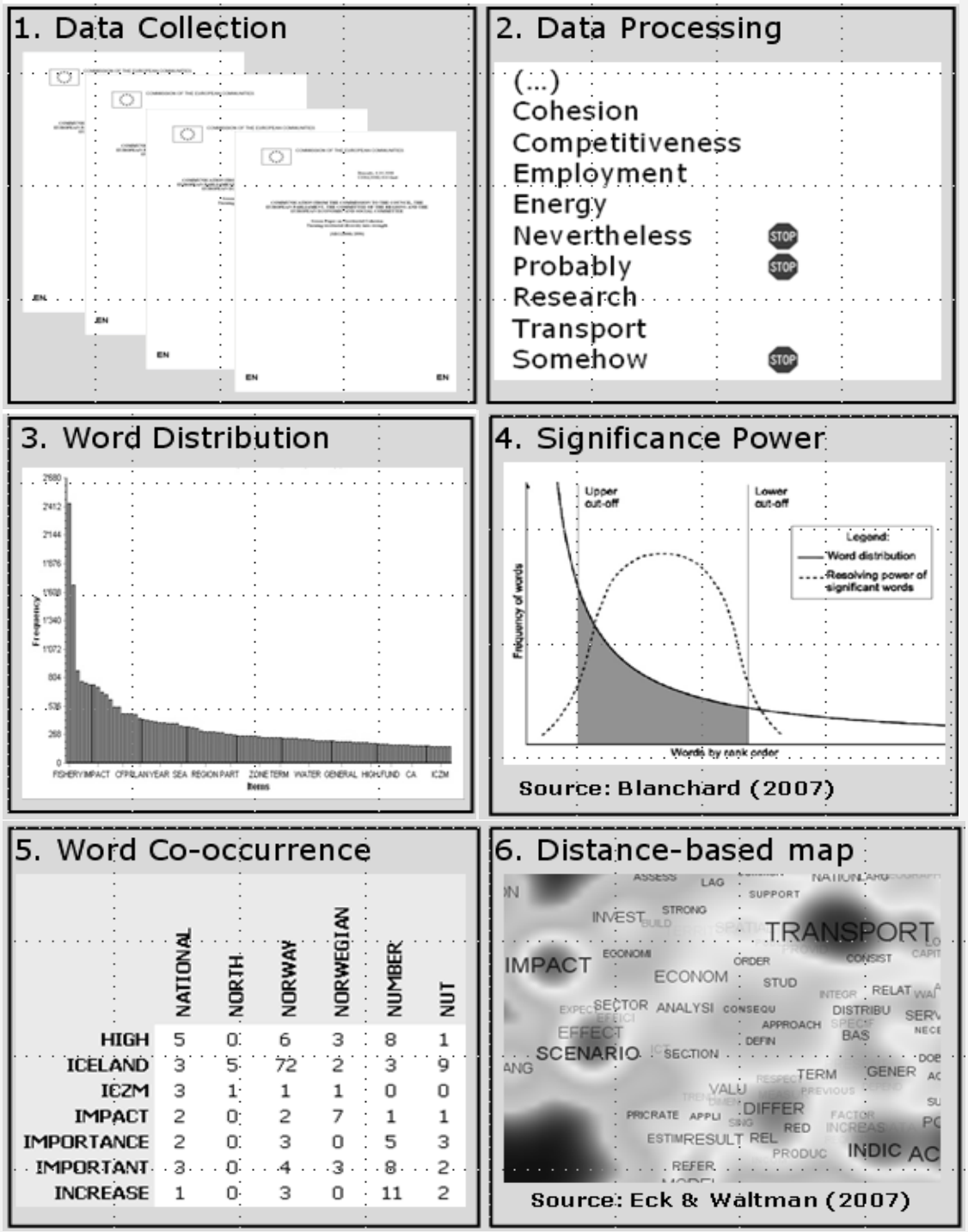
A two-step approach has been developed to structure the ESPON 2013 DB by themes and sub-themes. We argue that database structures adopted international organisations should support the definition of themes. This assumption lies on the fact that, very often, database structures define common topics to allocate data. For this purpose, we employ correlation matrices to analyse similarities and consequently interpret the results through visual grouping techniques. The proposal suggests seven themes. In addition, we add a theme to cover cross-thematic and non-thematic data.

The demand from the ESPON 2013 DB end users will be characterised by immediate, easy and practical access to data. A properly structure is therefore the key to meet this request. The next step comprised the definition of sub-themes. In order to achieve this goal, we explore the potentialities offered by text mining methods. This approach is used to find patterns across textual data that, inductively, create thematic overviews of text collections.

According to Dühr (2010), ESPON introduced new vocabulary of shared spatial concepts in Europe. We investigate this assumption by extracting keyword co-occurrence from texts with ESPON evidence and results.

In order to achieve concrete groups of keyword co-occurrence, textual data needs to be carefully prepared. Similarly, one of the crucial needs in text mining is the ability to visualise the relation of words. Hence, we apply a visualisation tool to construct and view maps of keywords based on co-occurrence and therefore better explore the results obtained from the information extraction phase. The results obtained constitute the basis for decision-making on sub-themes that eventually will facilitate the allocation of variables delivered by TPGs.

Short description of data preparation and visualization



	NATIONAL	NORTH	NORWAY	NORWEGIAN	NUMBER	NUT
HIGH	5	0	6	3	8	1
ICELAND	3	5	72	2	3	9
ICZM	3	1	1	1	0	0
IMPACT	2	0	2	7	1	1
IMPORTANCE	2	0	3	0	5	3
IMPORTANT	3	0	4	3	8	2
INCREASE	1	0	3	0	11	2

The methods used to identify sub-themes on text collections with ESPON evidence considered the above-mentioned steps. These steps have been performed for each of the seven themes that came out from our analysis on database structures.

1.8 THE OLAP CUBE

KEY FINDINGS

- OLAP stands for **On-Line Analytical Processing**. It consists on a **multidimensional** data model, allowing complex analytical and ad-hoc **queries** with a rapid execution time. OLAP technology has been proven to be a useful way to integrate NUTS-based data together with continuous data, such as land cover, over different time frames.
- The **ESPON OLAP Cube** consists on a stand-alone database which includes some socio-economic variables that are integrated and combined within a set of **dimensions** (**Spatial** dimensions (e.g. NUTS regions), **Thematic** dimensions (e.g. land cover), **Temporal** dimensions (e.g. 2003, 2006…)).
- The **ESPON OLAP Cube** can be used connecting MS Excel to the .cub file or by means of an online connection to a remote server (in this case, it can be queried from MS Excel and also from ArcGIS).

METHODOLOGICAL ISSUES

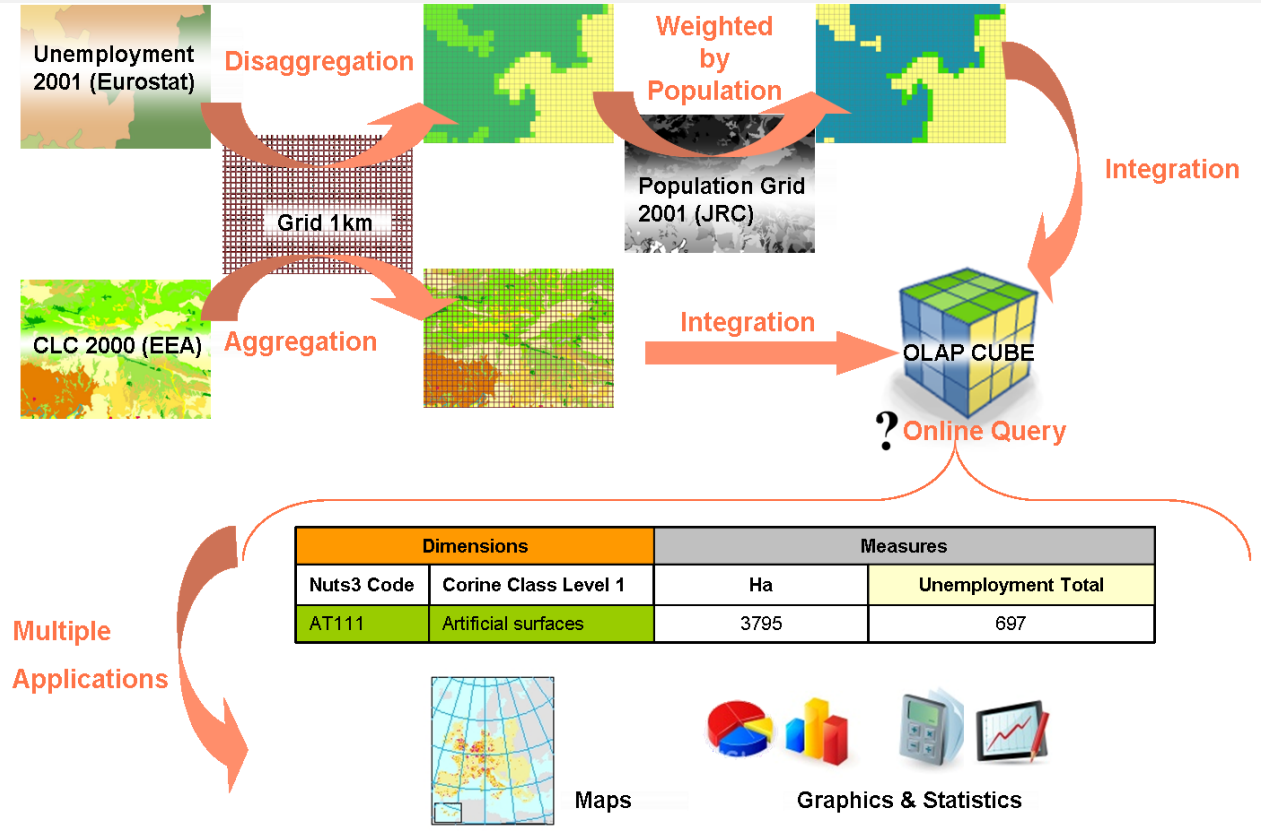
An OLAP Cube can be queried online and offline. So far, the online connection has not been implemented. In order to test the cube, we provide a single file .CUB which works offline. The .CUB file can be connected to and queried from Microsoft Excel with a few easy steps. A user manual has been provided, attached to the Technical Report.

THEMATIC ISSUES

The current version 3.0 of the ESPON OLAP Cube includes the following variables and dimensions:

- Socioeconomic variables : GDP 2003, GDP 2006, Active population 2003, Active population 2006, Unemployment 2003, Unemployment 2006.
- Land cover: Corine Land Cover 1990, 2000, 2006.
- Land cover changes: Land Cover Flows 1990-2000, 1990-2006, 2000-2006.
- Measures : Population density 2001 ; Area (ha)
- Geographical dimensions : Elevation Breakdown ; Biogeographic Regions ; Large Urban Zones and City Names ; Massifs ; Nuts 2006 ; Nuts 2003 ; River Basin Districts UE.

Methodological schema of disaggregation of socioeconomic data and combination with other data types



This schema summarises the whole methodology implemented by the ESPON DB project in order to combine socioeconomic data, usually reported by administrative units, together with other types of data, mainly continuous, such as land cover. The OLAP Cube is the core of the methodology as it is used as the integration repository for all the datasets, and allows a prompt querying of the data to get interesting analytical results.

The diagram shows, on the one hand the process of aggregation/disaggregation of data by means of the 1 km Reference Grid, and the weighting by population density whenever it is possible, to add some value to the disaggregation of the source data.

Finally, all the variables reported by 1 km grid cell are integrated into the OLAP Cube, which facilitates their combination and querying as it has been explained, making the creation of maps and graphs straight-forward.

1.9 CARTOGRAPHY IN ESPON

KEY FINDINGS

- The ESPON mapkit is a set of mapkits according to the geographical levels.
- It ensures harmonization of all the maps produced in ESPON projects.
- It is compliant with ESPON Database application.
- It is available at different format (ArcGis, QGIS, Philcarto).
- A mapping guide is provided to explain main rules for mapping in ESPON.

DESCRIPTION

As a general rule, maps are used to visualize geospatial data and enhancing statistical data to understand phenomena. In ESPON Program, there is a need to produce a lot of maps. This part presents the mapkit developed by the ESPON DB project and follows **3 main objectives:**

i) Ensuring harmonization of maps. Maps are produced by researchers, engineers or students involved in each ESPON projects. Consequently, we need to ensure graphical harmonization of all maps produced by different authors, with different software. The mapkit tool (consisting of specific mapkits collection) contains geometries, cartographical templates and graphic elements (logos, disclaimers). In every case, these different elements are available in Arcgis format (mxd + shapfiles), Quantum GIS[1] (a user friendly Open Source Geographic Information System licensed under the GNU General Public License), and Philcarto[2] which is a free software for thematic cartography.

ii) Ensuring compatibility with the ESPON 2013 database application. The ESPON DB application provides indicators at local, regional and global level. It also provides data on different geographical objects (*e.g. dots and grids*). The mapkit ensures the mapping of data on these different kinds of objects. It is compliant with the ESPON Database application and permits to visualize, on a map, the data extracted from the application whatever the kind of data.

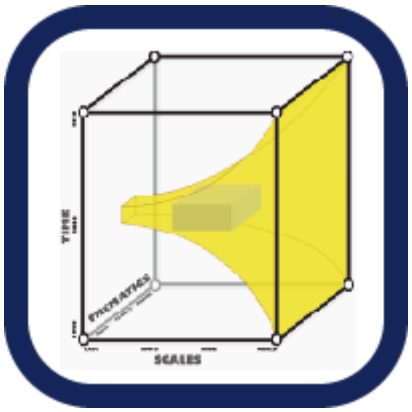
iii) Enhancing information (How to make good maps). Many possibilities exist to show data on map. Choosing relevant representation is not an obvious task and has to be considered seriously. Indeed, choosing the wrong way of mapping can completely misrepresent the data. For this reason, a mapping guide was realized to help people to follow "good rules" of cartography. Moreover, it is important to keep in mind that choice in cartography is always dependent on the type of data (and targeting the right audience) and that there is never an optimal solution. Map is always a compromise.

[1] http://www.qgis.org/
[2] http://philcarto.free.fr/

The set of map kits

This picture is an overview of the ESPON mapkit. Actually, it is composed by a set of 6 specific mapkits adapted to different geographical levels, from local to global.

To download the map templates and get the last versions of the geometries (LAU2, NUTS, Countries of the World), go to ESPON Intranet (http://intranet.espon.eu)



2.THEMATIC ISSUES

INTRODUCTION – PART 2

The second part of this report presents the different solutions elaborated by ESPON Database project in order to enlarge the possibilities of territorial data exploration, mapping and prospective. Each of the 10 sections is related to one or several technical reports that are referenced and can be downloaded on the ESPON website. For each topic, we firstly summarize the key findings, the methodological issues and thematic issues on the left page. Then we illustrate the interest of this technical report by a significant and demonstrative example of application on the right page.

The first topic is related to time series harmonization (2.1) and describes how to combine NUTS data with different geometries and, more generally, how to remove holes in time series and enlarge them back to past or toward future.

The second topic is related to urban data, which are becoming more and more a key issue for territorial cohesion and competitiveness. We describe first how to cope with two different definitions of cities produced by EEA or Eurostat. A first report demonstrates how to make better use of Urban Morphological Zones, in particular through a naming process (2.2). A second report focuses on the specification of Larger Urban Zones and tries to clarify the national definition of cities used by each country in the different periods of elaboration of urban audit (2.3). Finally, an attempt is made to delineate Functional Urban Areas with a harmonized criterion of polarization (2.4).

The third topic is related to the thematic enlargement of ESPON database in order to overcome the classical focus on economic dimensions. The combination of social and environmental data can be strongly developed through the elaboration of aggregation and disaggregation methods, making possible to transfer information from NUTS to Grid or from Grid to NUTS (2.5). The same idea is applied for individual data based on surveys that can be used for the elaboration of more innovative information on social dimension of territorial cohesion, but with a great attention paid to the problems of sampling errors when data are estimated at regional or city levels (2.6).

The fourth topics is related to the enlargement of the possibilities of ESPON database to upper or lower geographical scales, in order to make more easy the objective of the "five level approach" described in the first ESPON Synthesis Report (p. 17). We analyze firstly the possibility to improve the collection of local data at LAU1 and LAU2 levels (2.7). Then, we propose solutions for the coverage of regional data related to candidate countries and the rest of western Balkans (2.8). Finally, we describe the procedure of data collection for supporting the elaboration of a world database (2.9).

2.1 TIME-SERIE HARMONISATION

KEY FINDINGS

- Review of literature on the various possible solutions for the harmonization of times series. Benchmarking of these solutions and proposal of a general solution that is a development and improvement of the "ESTI" model previously proposed in ESPON 2006 Data Navigator Project.
- Compilation and inclusion in the ESPON 2013 database of data using old NUTS version (1995, 1999, 2003, 2006). In particular, data elaborated in ESPON 2006 program and historical data from Eurostat.
- Elaboration of a systematic dictionary of change of NUTS units based on the concept of "lineage". The concept of lineage is more general than a simple review of modification (as provided by Eurostat) and offers the possibility to follow a regional unit through time, even when names, geometry, codes, etc. are changing.

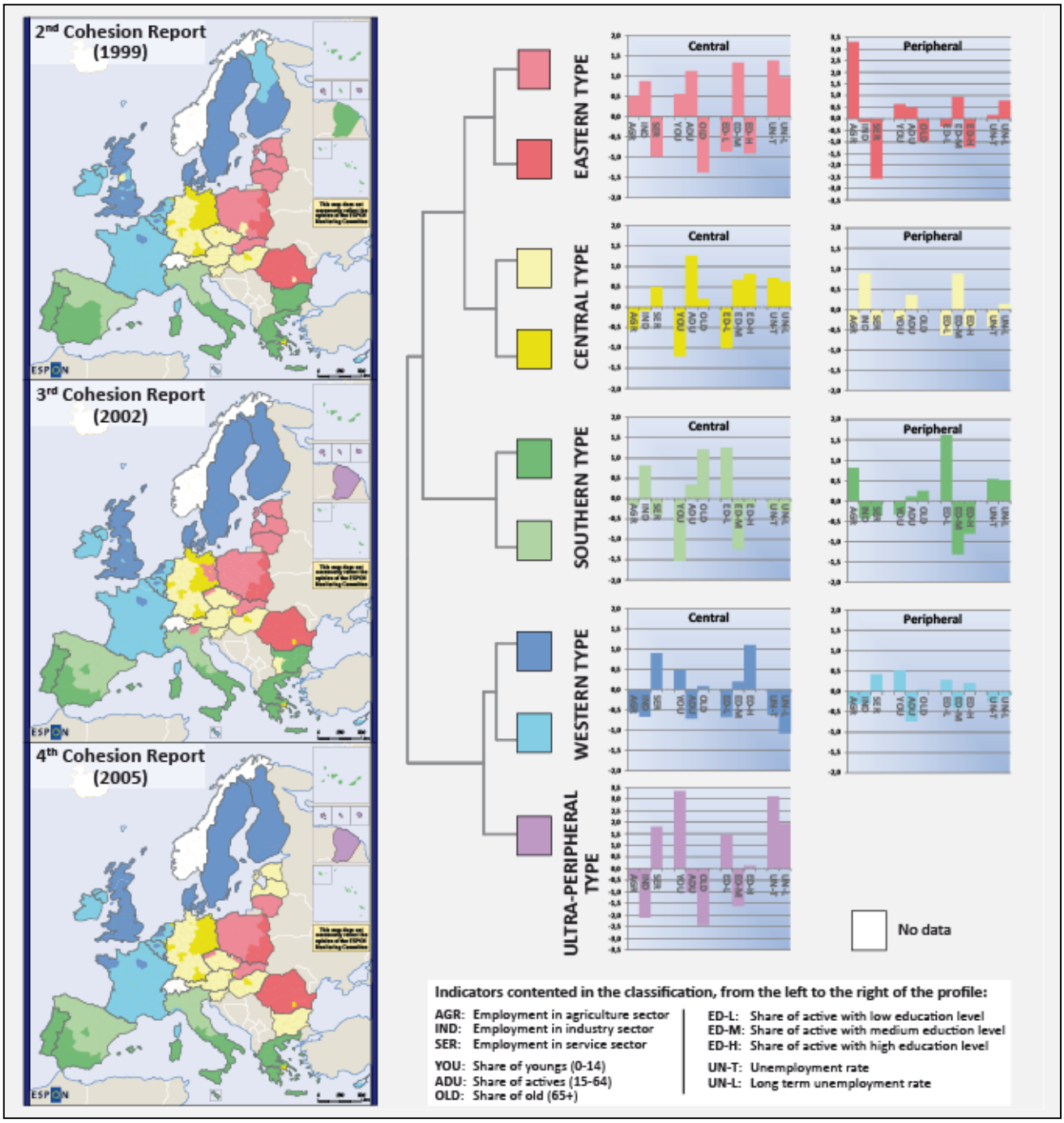
METHODOLOGICAL ISSUES

Time series approach can be organized in two main steps. Firstly, the search and the exploration of historical databases (New Chronos from EUROSTAT, cohesion reports from DG-REGIO…) are performed. This step aims at providing a survey of continuous time-data series that could be built from these databases. Additionally, we have explored NUTS changes between 1995 and 2006. This exploration resulted in the compilation of the dictionary of NUTS changes, which allows the survey of territorial changes (codes, names and geometries). But the most important contribution of the dictionary is the identification of the genealogy (lineage) of NUTS which proves 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 will provide the basis for a future computer implementation of automatic procedures for estimating the values of missing data.

THEMATIC ISSUES

Some of the methods proposed for the estimation of missing values in time series have been directly used in the ESPON Database in order to remove holes or to propose "provisional estimation" of indicators of interest like the Unemployment rate in March 2010 at NUTS2 level (ESPON First Synthesis Report, p. 21). Moreover, we have demonstrated that it is possible to analyze the evolution of spatial patterns of regional unequalities with changing territorial units, through the example of a cross analysis of statistical annexes of 2nd, 3rd and 4th Cohesion Reports (see on next page).

Typology of EU Regions according to Cohesion Reports (out of GDP…)



A typology of EU regions, based on structural phenomena used in the cohesion report (demography, economy, education, labour force) but excluding GDP per capita, reveals the existence of 4 basic patterns of strength and weakness in northwestern, southern, central and eastern parts of EU. Interestingly, the exclusion of GDP per capita provides a division of EU territory which does not follow the classical division between old and new member states. Despite some minor changes, this structural pattern is very stable through time and could be therefore used as a basis for "taylor-made" policies of regional development.

- ***Dictionnary of NUTS changes (1995-2006):*** *Stable units, genealogy, types of changes*
- ***NewCronos database:*** *Eurostat historical regional data, 270 tables extracted in csv files*

- ***2nd , 3rd and 4th Cohesion Reports data:*** *Basic indicators policy oriented for the EU27 regions (on the period 1999-2004)*
- ***NewCronos data:*** *In the NUTS 1999 delineation, population 1970-2001, age pyramids 1990-1995-2000*

2.2 NAMING URBAN MORPHOLOGICAL ZONES (1)

KEY FINDINGS

- A European data base operational for urban studies, containing 4437 cities over 10 000 inhabitants, defined from CLC2000 with harmonized criteria (EEA, last version of Urban Morphological Zone shapes), population (JRC, last version of Population Density Grid), names (ESPON Data Base, see Technical Report) and metadata.
- An automatised process for naming UMZ which allows quick updating with new versions of sources or methods (EEA and JRC) or new dates (2006, 2010…). The methodology is validated through expertise at national scale and matching to other sources (Eurostat, Geopolis, Google Earth). It is fully explained and documented in the Technical Report.
- An exploration of the common features of European cities in 2000 (population, surface, density) that confirms the very high regularity of the hierarchical structure at the European level, a major North-South density gradient (lower than 2000 inh./km^2 in Sweden, Denmark, Finland, above 4000 inh./km^2 in Italy, Spain or Greece), but also a strong and regular relationship with city size levels (densities exceeding 5700 inh./km^2 in cities larger than 2 millions inhabitants, then decreasing regularly until 3000 inh./km^2 for cities between 10 000 and 25 000 inhabitants).
- An exploration of international cities of Europe from a very local point of view, which is based on three different indicators computed for each concerned UMZ (number of crossed countries, total population, share of population living outside the main country). Crossing the two latter indices enlightens for instance some cases of small but much internationalized cities, with more than 40% of their population abroad, at the Poland/Germany, Slovakia/Hungary or Austria/Germany border.

METHODOLOGICAL ISSUES

Elaborating a methodology for naming physical entities that are automatically built from satellite images raises a series of redoubtable and very classical issues in data base modelling. The inputs have to take into account a huge set of data, the diversity of sources, technologies and national approaches, and evolving contents. The research that has been developed is constantly based on two main principles: international harmonizationof processes and automation of each step of the process, i.e.:

- Automation of the computation of populations intersecting the different sources (UMZ, Population dentity grid and national data base selected for giving the name)

- Automation of the attribution of names given to each UMZ, according to the way it overlaps the national data base elementary units: clearly concentrated inside one unit (then receiving one name), or expanding clearly on 2 or more units (then receiving 2 or more names, like in industrial or littoral conurbations).

From morphological delimitations (UMZ) to a European set of cities : naming process

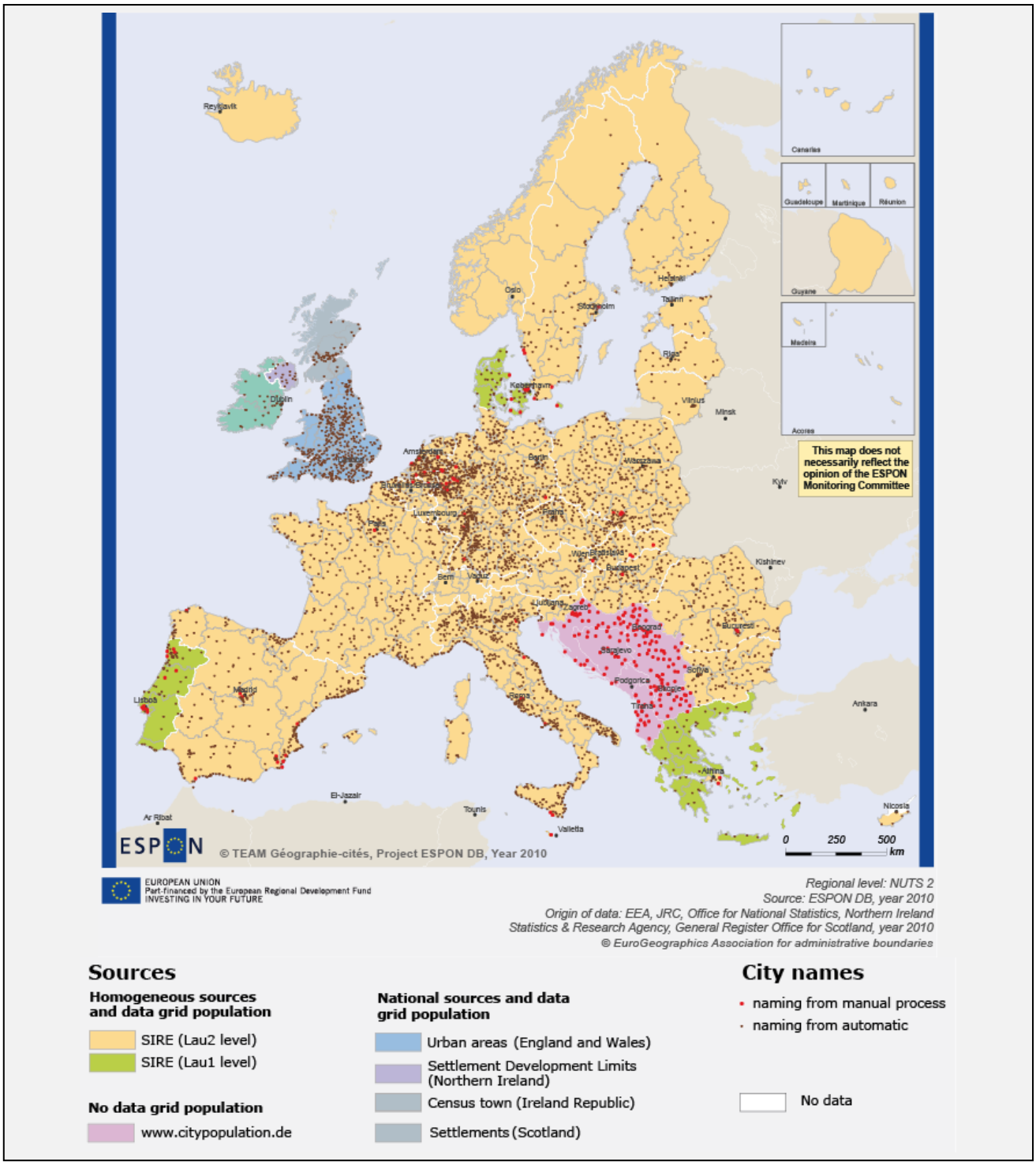


Figure 2.2.1 illustrates one step of the methodology used for naming UMZ, consisting to choose the relevant sources for naming cities (LAU2, LAU1 or national data sets), according to an urban expertise and to the avaibility of the Population density grid.

2.2. NAMING URBAN MORPHOLOGICAL ZONES (2)

- Automation of the final checking, by comparing names to Eurostat compilation of national city names, Geopolis data base (François Moriconi-Ebrard 1993) and Google Earth. A typology of the main inadequations (about 90 cases on 4437) has been realized and solutions proposed, that will make easier the future checks.

THEMATIC ISSUES

The UMZ data base is now fully operational for a deep exploration of the common features and diversity of European urban settlements. Three types of analyses have been presented in the Technical Report:

- *City size distribution*: the classical rank-size distribution, plotted for the 4437 cities over 10 000 inhabitants, confirms the very high regularity of the hierarchical structure at the European level. The absolute value of the slope, used as an indicator of city size inequality level, is very close to other values computed by European researchers with former databases.

- *Density patterns*: a multiscalar analysis of density levels in Europe gives striking results, with a major North-South gradient but also a strong and regular relationship with city size levels. National specificities appear also very strongly, as revealed by the higher densities of Dutch cities, the strong discontinuities observed for instance at the Franco-Spanish frontier and at the German-Polish border, or as suggested by the high densities of some Eastern countries like Poland. These studies are of high interest for the future, for example in urban planning issues (transportation or environmental topics). Interoperability with other geo-referenced data bases (urban transport infrastructures, urban mobility, socio-economic LAU data…) opens a wide range of environmental and social studies

- *International UMZ*: two indices have been computed for these cities, the number of crossed countries and the share of population living in one or more countries different than the main one. It allows to qualifye in a comparable way to what extent the city is embedded in a multi-national context. For exemple, the most populated international UMZ is Brussel/Anvers/Gand, but it extends in a very small part in Netherlands (population living there is only 1%). At the opposite, some UMZ located at the Poland/Germany, Slovakia/Hungary or Austria/Germany frontiers are less than 50 000 inhabitants but more than 40% of their population lives abroad.

European City sizes and Densities (UMZ/ CLC 2000)

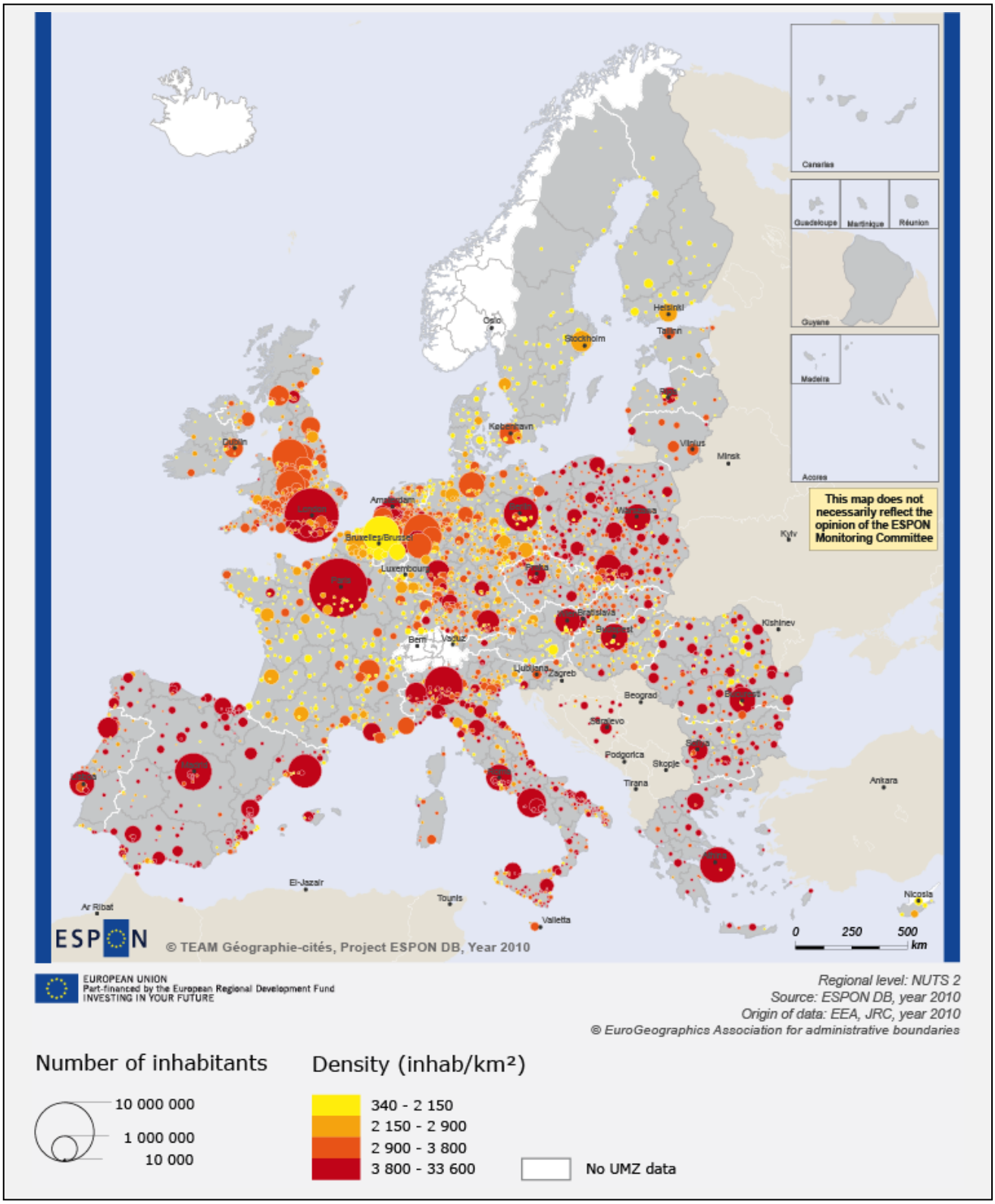


Figure 2.2.2 enlightens one of the thematic explorations that can be lead now with the UMZ database: the map of the density indicator reveals a major North-South gradient, but also national and size effets.

2.3 LUZ SPECIFICATIONS

KEY FINDINGS

- Larger Urban Zones, collected in each country by Urban Audit 2004, are based on various national definitions (functional areas, planning regions, local administrative units, etc.). Using indicators computed in these perimeters requires first a good knowledge of national specifications, in order to be able to identify possible bias resulting from the national heterogeneityof LUZ definitions.
- Based on an expertise from national reports sent to Urban Audit by the different countries and completed by other sources of information, the Technical Report presents a clear synthesis of the LUZ specifications, through 4 synthetic typologies and maps but also an Annex containing 30 country-sheets which describe in a common vocabulary and "syntax" the general rules used by each country to define its LUZ.
- The results enlighten a very large heterogeneity in the national approaches used to define LUZ (Figure 2.3.1) and engage the users of the database to be very cautious when interpreting some statistical results. But it also enlightens a very interesting evolution between UA 2001 and UA 2004, towards more functional definitions, mainly based on commuters, even if the criteria remain very different from one country to another (see thresholds, Figure 2.3.2).

METHODOLOGICAL ISSUES

After collecting and expertizing documentation on national specifications, a common "syntax" has been used for categorizing specifications, i.e the building blocks used in such perimeters, the links between these building blocks when LUZ is an aggregation of elementaty units, the diversity of thresholds in commuting-based approaches, the possible evolution of the LUZ definition since UA 2001, and some particular cases (for example when Capital cities use another type of definition).

THEMATIC ISSUES

The typology of LUZ delineations enlightens a large heterogeneity not only at the European scale, with six different types of definitions (Figure 2.3.1), but also at the national scale (the majority of Capital cities being processed differently than the other cities of the country). However, in the same time, a clear dynamic towards more functional approaches, largely encouraged by Urban Audit and Eurostat, is undergoing. Between UA 2001 and UA 2004, six countries have changed to adopt functional definitions, and two others will probably do the same for the next Urban Audit. All the new participant members to UA 2004 have chosen a functional definition.

Figure 2.3.1: Typology of LUZ delineations (Urban Audit 2004)

Figure 2.3.2: Variety of commuting thresholds in LUZ functional delineations (Urban Audit 2004)

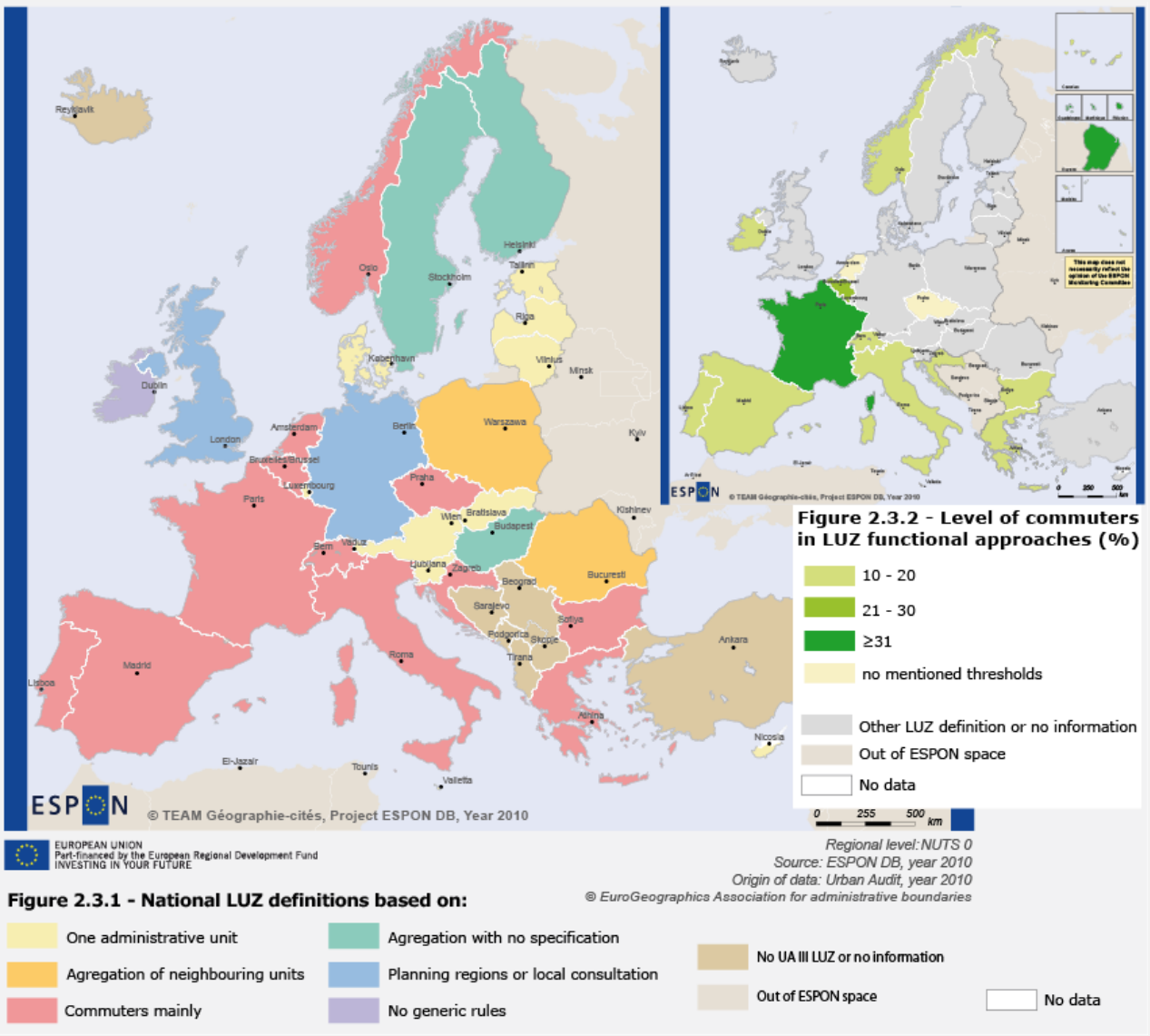


Figure 2.3.1 gives the result of the general typology of LUZ definitions in Europe, based on a compilation of national reports, which enlightens the diversity of national approaches in LUZ definitions. Figure 2.3.2 shows the variability in the choice of commuter thresholds for countries that explicitely mention this criteria in the rules of LUZ delineations. The map does not enlighten any gradient or regional structure, and statistical analysis did not reveal any correlation between commuting threshold and the average size of administrative units.

LUZ specifications: *Documentation about Larger Urban Zones (LUZ) covering Urban Audit III area.*

2.4 THE FUNCTIONAL URBAN AREAS DATABASE

KEY FINDINGS

- As a joint venture between 3 challenges of the Espon DB project (Urban data, Local data and Time series) and starting from the results of the previous Espon program we provide an update of the database of the Functional Urban Areas (FUAs) and Morphological Urban Areas (MUAs), as well as their inter-relations. Not only is it enhanced, it is also fundamentally enriched by the quality of the data provided, as the Functional Urban Areas (FUAs) are now delineated for most of the European countries of the Espon space at the LAU2 level.

- The FUAs are defined as labor basins of the MUAs which are themselves defined as densely populated areas, all this independently from any national, administrative or political definitions, but based instead on pure statistics.

- The main quality and advantage of these FUAs are their simple and universal definition throughout Europe, making them comparable in all the countries where they were delineated.

- Finally we have also produced a list of indicators for these FUAs.

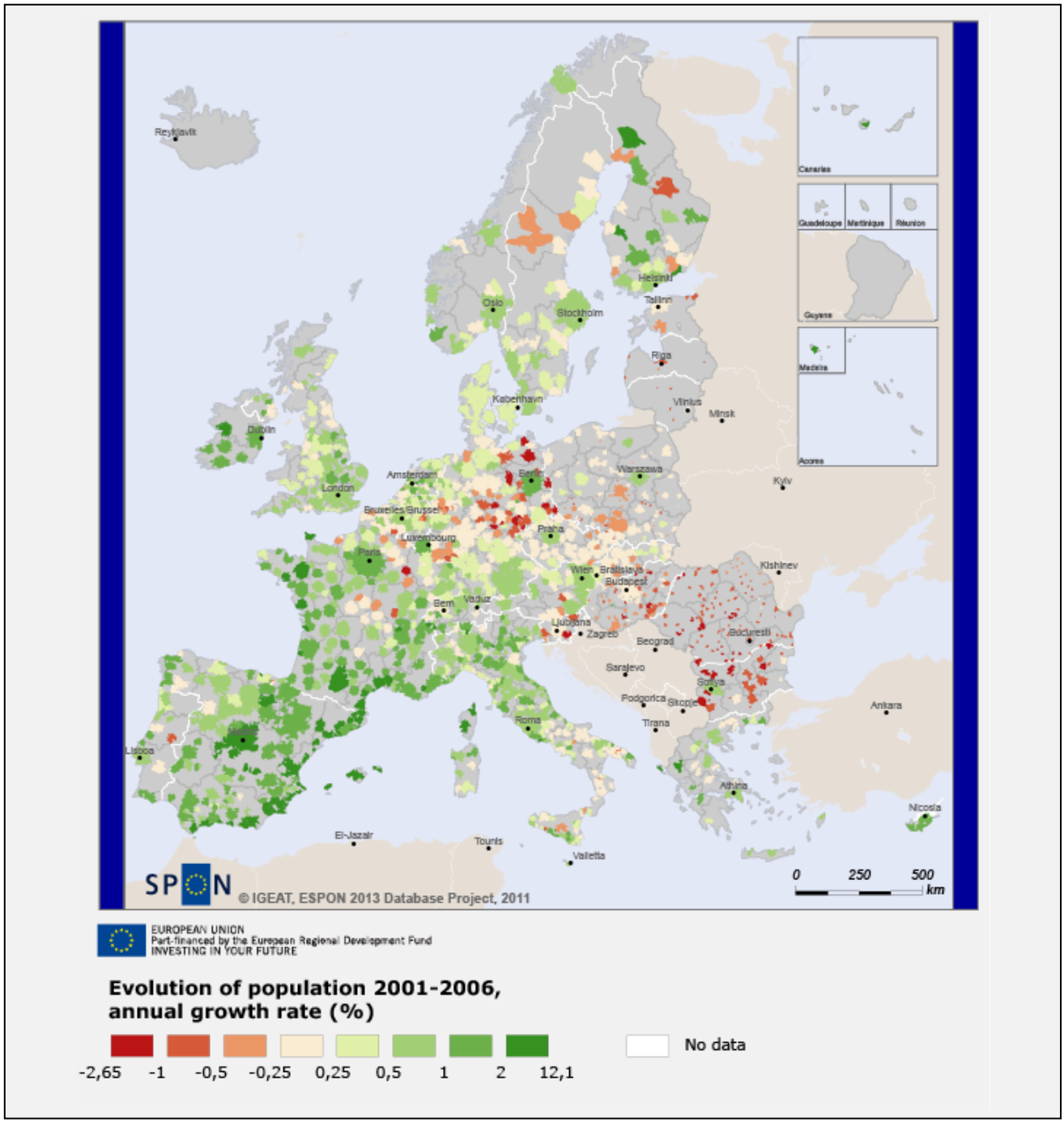
METHODOLOGICAL ISSUES

The MUAs are built by agglomerating the LAU2s having a population density higher than 650 inhabitants per km2. There can be one single LAU2 or hundreds of them. The MUAs kept in the list have a total population of at least 20 000, or actually was it made so by the Espon 1.4.3 project on Urban Functions in 2006.
We have transposed the LAU2s references from the old nuts-5 system from 1997 into the LAU2s of 2008, but without actualizing the population numbers that are still from 2001. The FUAs are the labour basins of the MUAs, obtained by agglomerating the contiguous LAU2s having 10 % of their «economically active population» working in the nearby MUA, so to say in the employment place of the neighborhood.
The LAU2s with a population higher than 20 000 but with a population density lower than 650 were also selected, like some LAU2s in the northern countries or in the south of Spain.

THEMATIC ISSUES

The MUAs are thus independently defined from any cultural, political or administrative definition of the cities throughout the European countries, each having their own urban system. The MUAs have a simple definition which is coherent at the European scale and as densely populated areas they are considered as probable employment places, which will be verified (or not) in the next step, the building of the FUAs.

Evolution of population 2001-2005 in the FUAs



To illustrate the key findings and the thematic application of this work let's display here a map of the European FUAs with the population evolution between 2001 and 2006. This was made with the population figures at the LAU2 level for 2001 and for 2006 when available, otherwise by applying the evolution of the NUTS3 figures to the 2001 FUAs population. We provide in the database other indicators as listed in the technical report.

FUA database: *Geometries and 15 basic indicators in different thematic fields, covering the entire ESPON Area*

2.5 SOCIAL/ENVIRONMENTAL DATA

KEY FINDINGS

- 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 due to have an European coverage and follow Inspire specifications.
- The "proportional and weighted" aggregation method is the one that gives better results, plus some added value to the downscaling.
- Different methods are independent from the source data format and can be applied to vector and raster format.
- 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.

METHODOLOGICAL ISSUES

Depending on the nature of each indicator or variable, a different kind of integration procedure must be applied. In this regard, we have defined and tested with different data three integration methods. The "proportional and weighted" aggregation method is the one that gives better results, plus some added value to the downscaling. Thus, it is the recommended one:

Proportional and weighted calculation: the cell takes an area proportionally calculated value, and 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.

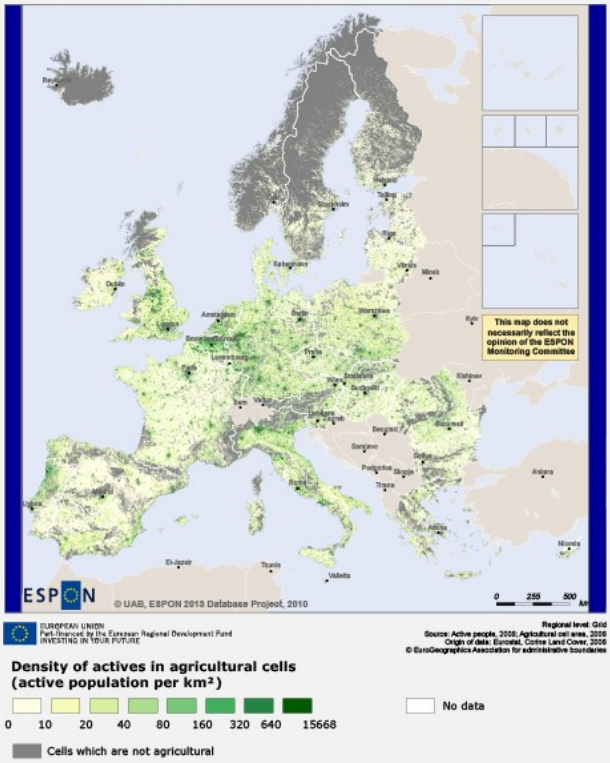
THEMATIC ISSUES

The methodology that has been defined under the Challenge 5 of the ESPON 2013 DB project provides useful tools to combine data provided by administrative units, such as NUTS 3 divisions, together with continous data, namely gridded variables. In the end, it allows the user to go back and forth from one type of data to the other one and viceversa, depending on the purpose of the analysis to be made.

The following maps are just an example to illustrate how gridded data can be used to report by NUTS 3, and how data originally reported by NUTS 3 can be reported by grid, giving an added value to the source data.

Disaggregation and aggregation of data

Figure 2.5.1 - Active people 2006 in agricultural grid cells (CLC 2006)



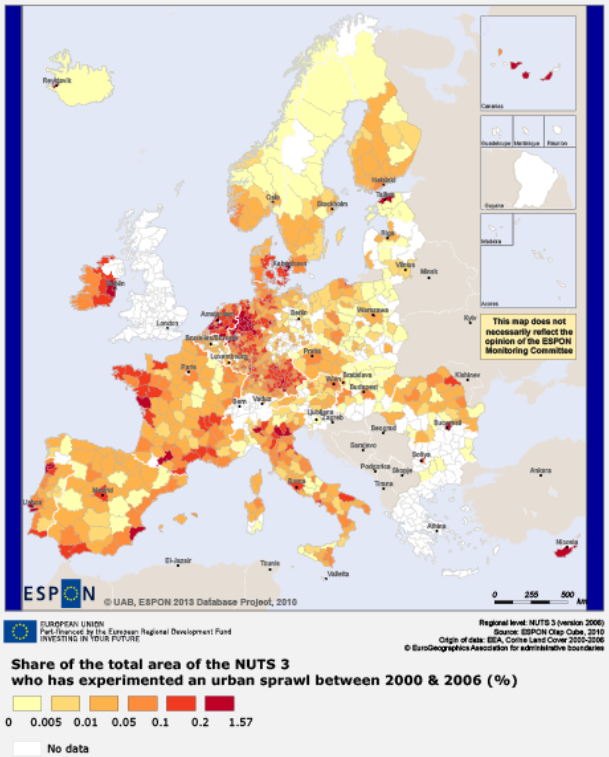


Figure 2.5.1 shows the distribution of active people by 1km grid cells with more than 50 ha. of agricultural landcover. This is an example of **disaggregation** of data by administrative units and the combination of such data with Corine Land Cover 2006 (level 1 agricultural class).

Figure 2.5.2 shows the urban residential sprawl process between 2000 and 2006. This map has been produced by means of the **aggregation** of Corine Land Cover Changes (as Land Cover Flows) by nuts3 regions using the ESPON Olap Cube v3.0.

Corine Land Cover regional data: *Corine Land Cover 1990, 2000 and 2006 (3 hierarchies) in the NUTS 2006 delineation (levels 0, 1, 2 and 3)*

Grid indicators: *GDP, active population and unemployed persons (2003 and 2006) desaggregated in the European Reference grid (1 km)*

2.6 INDIVIDUAL DATA ANDS SURVEYS

KEY FINDINGS

- European-wide sources for fine-grained spatial socioeconomic data are almost non-existing. Data is seldom presented with a high degree of spatial resolution, partly due to issues related to survey sample sizes and sampling errors. In fact, published tables in general present data at NUTS levels 1 and 2 (and occasionally NUTS 3). The project aims at creating new, more fine-grained data, based only on what is readily available from public European-wide data sources.
- The project demonstrates one possible route towards improvements based on systematically merging available information. By departing from the JRC 1 ha population grid (partly based on CORINE land cover data), and simultaneously utilizing information from many different tables, data sets with a higher geographical resolution can be produced. This can be achieved since the tables together contain more information about the joint distribution over space and attributes than do the single tables side by side.
- The project shows that different tables from Eurostat, such as ones based on the Labour Force Survey (LFS), can be combined with each other and with the JRC population grid, enabling the construction of – otherwise not existing – spatial socioeconomic datasets at the km^2 and NUTS 3 levels.
- Despite obvious limitations, the JRC population grid is a quite reasonable tool in its own right for certain purposes. For instance, it provides fairly good estimations of the population of Urban Morphological Zones (UMZ).

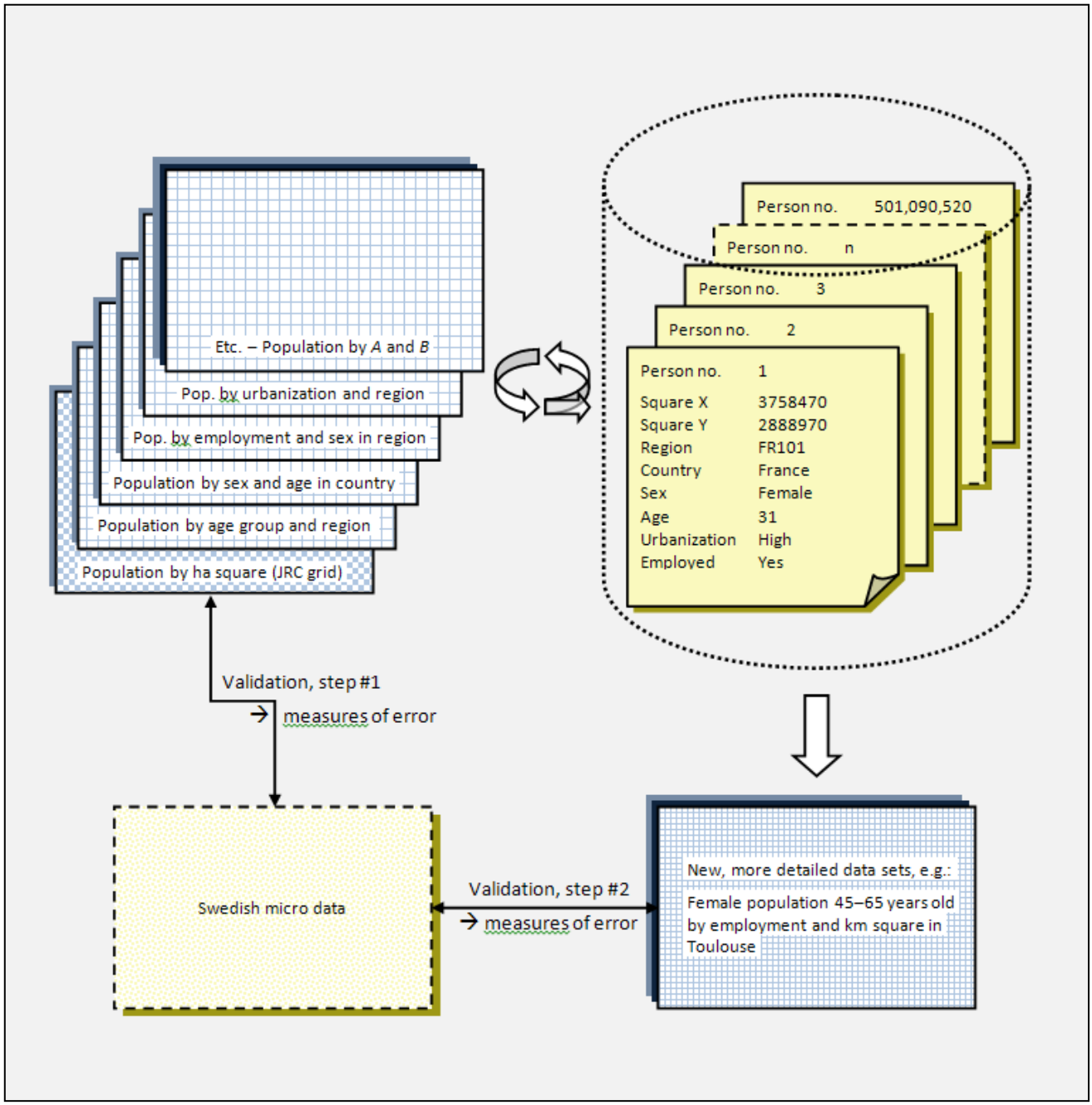
METHODOLOGICAL ISSUES

The first step is the creation of a synthetic individual database of the European population. The attributes of the artificial individuals are then assigned randomly and after that systematically iterated so as to become consistent with all employed tables. The end result is synthetic individuals that are jointly consistent with all supplied information.

THEMATIC ISSUES

The majority of the population is located in urban areas covering a tiny fraction of space, whereas a small share of the (aging) population is distributed over substantial parts of the countryside; exhibiting a shortage of qualified labour for emerging jobs. Information for analysis and counteraction related to such problems are effectively hidden by current aggregation levels in available data.

New data sets from multiple tables via a synthetic population



The figure shows, schematically, how disaggregated population data is used together with different survey tables to produce a synthetic individual database – as well as new data sets. Despite shortcomings related to disaggregation, survey sampling errors and inconsistencies arising from the method of iterative attribute exchange, the end result represents an efficient way to increase the amount of local information using presently available data resources.

2.7 LOCAL DATA

KEY FINDINGS

- The use of the classical spatial patterns (points, surfaces, networks), mobilized at the local scale and managed with specific methodologies, allows us to provide basic indicators at local level (LAU2), for selected case studies.
- The construction of indicators at local scale using information that is based on some specific geographical objects (grid information or networks) could function as an option to the traditional data sources.
- The exploration of the main sources of data at this level (NSI) is an opportunity to design methods that are able to properly match indicators and geometry. The collection of several LAU2 codes is strongly needed, in this case.

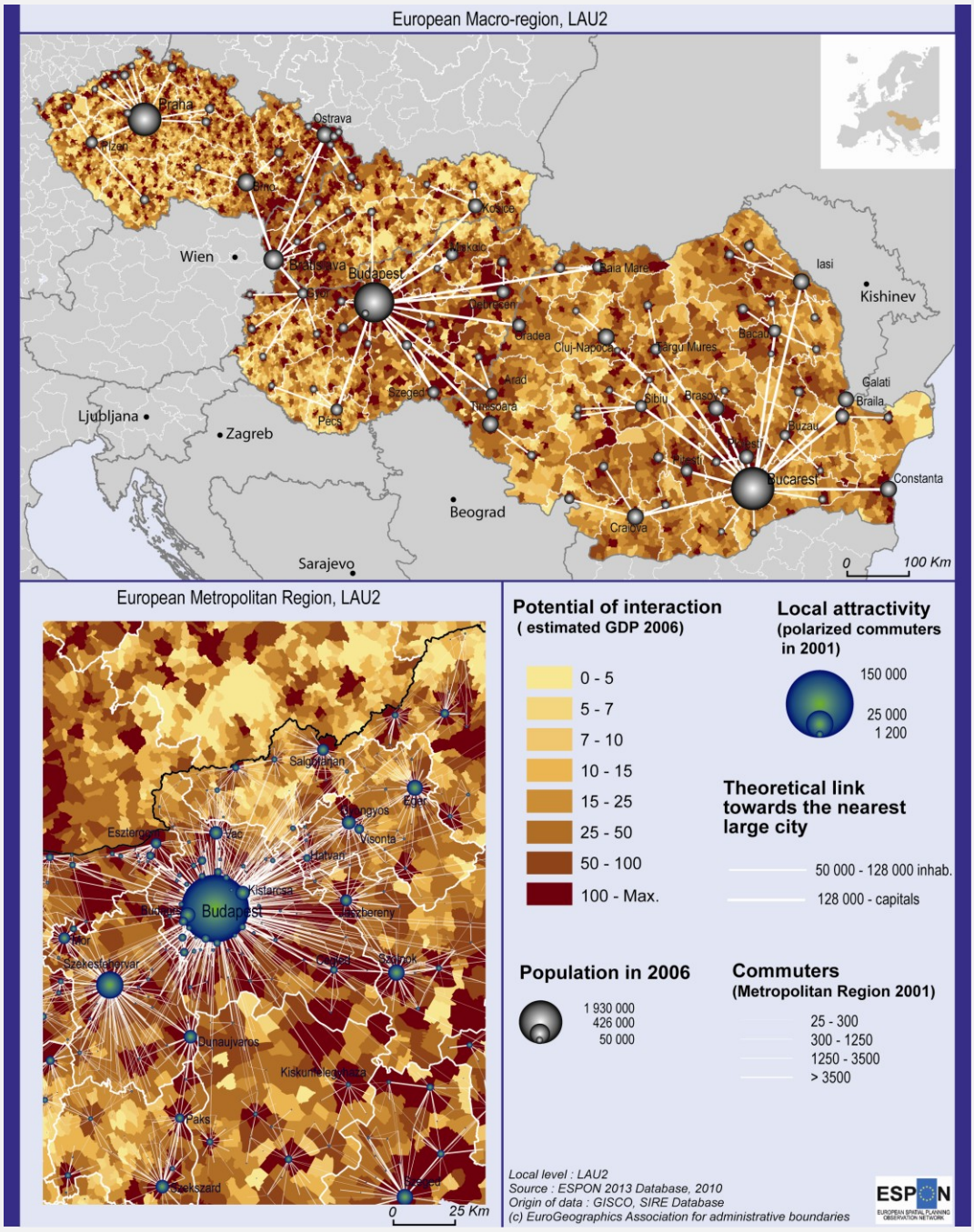
METHODOLOGICAL ISSUES

Using the LAU2 as scale of reference for data integration is a double challenge. As a first task, the variety and the quantity of geometries demand not only data mining and GIS techniques of manipulation, but also a good knowledge of the "terrain", sometimes in the classical sens of the expression. Secondly, the data production is time consuming, even when the implementation of simple models is needed. However, solutions and methods to overcome these difficulties can be figured and good practices in the local data manipulation are now available. Experimenting the methods is partially depending on the number of LAU2, but in some cases (potential model) is just a matter of time.

THEMATIC ISSUES

The intersection of the grid information with the LAU2 geometry can function as a method of data collection and indicators development. As the GDP is generally available only at NUTS level, reagregating data at LAU scale represent an alternative way to analyze the territorial economic performance. The grid data was provided by ESPON DB project (Challenge 5) and for the moment only values for 2006 were agregated in the LAU geometry. To eventually smooth the high contrast between the spatial units, an option was made for the potential of interaction, with constraints regarding the moving-window and the distance decay. One of the model's parameter was the road network density at LAU2 and the values for Bulgaria are missing, explaining its absence on the map.

Economic performance and attractivity at local scale



Analyzed at the local scale, the economic performance is an archipel shaped by capitals, active frontiers, transportation corridors (Gyor-Budapest) and old industrial regions (Czech Silesia). In this territorial frame, the rural spaces are not so passive as one might expect, if we except the remoted areas or some specific cases (the Romanian eastern border). When zooming at a metropolitan scale, the same archipel model is this time articulated by a core-periphery gradient, complicated by the presence of privileged axis and secondary poles (Vac, Esztergom, Hatvan), the last ones linking Budapest to regional cities like Szekesfehervar, Szolnok or Szeged.

- ***Eastern Europe database:*** *Demographic and accessibility indicators for Czech Republic, Slovakia, Hungary, Romania and Bulgaria at LAU2 level with related geometries.*
- ***Eastern Europe Land Cover database:*** *Corine Land Cover 2006 at LAU2 level for Czech Republic, Slovakia, Hungary, Romania and Bulgaria with related geometries.*
- ***GDP in LAU2 units***: *Estimation of GDP in 2006 at LAU2 level in the all ESPON Area*

2.8 ENLARGEMENT TO NEIGHBOURHOOD

In this thematic issue we aimed to explore the possibilities to extend the pool of data on the ESPON countries on the Western Balkans (WB) countries and Turkey as well as to study how ensure that the relevant data be harmonized with the rest of the ESPON Database. From these countries, Croatia, FYROM (Former Yugoslav Republic of Macedonia), Turkey and Montenegro are Candidate Countries (CC) while Albania, Bosnia and Herzegovina and Serbia including Kosovo (Under UN Security Council Resolution 1244) are Potential Candidate Countries (PCC).

KEY FINDINGS

- Croatia, FYROM and Turkey, which have adopted the NUTS classification, available data covers the large majority of themes and years at NUTS3 level, while for Albania, Bosnia and Herzegovina, Montenegro and Serbia including Kosovo there are considerable gaps at NUTS2 and NUTS3 levels.
- We have also examined the availability and quality of data for the CC / PCC provided by Eurostat, NSO sites and other sources for a limited number of additional themes: Croatia, FYROM, Turkey and Serbia covers a large number of themes, years and levels and working on harmonisation is relatively easy, while for the other CC / PCC there are numerous gaps and working on harmonisation of the existing data from NSO is much more difficult.
- We should stress that Eurostat provides, for these additional themes, data for all CC / PCC at NUTS0 level. These data are not included in the Database as this last does not contain respective data for the ESPON countries.

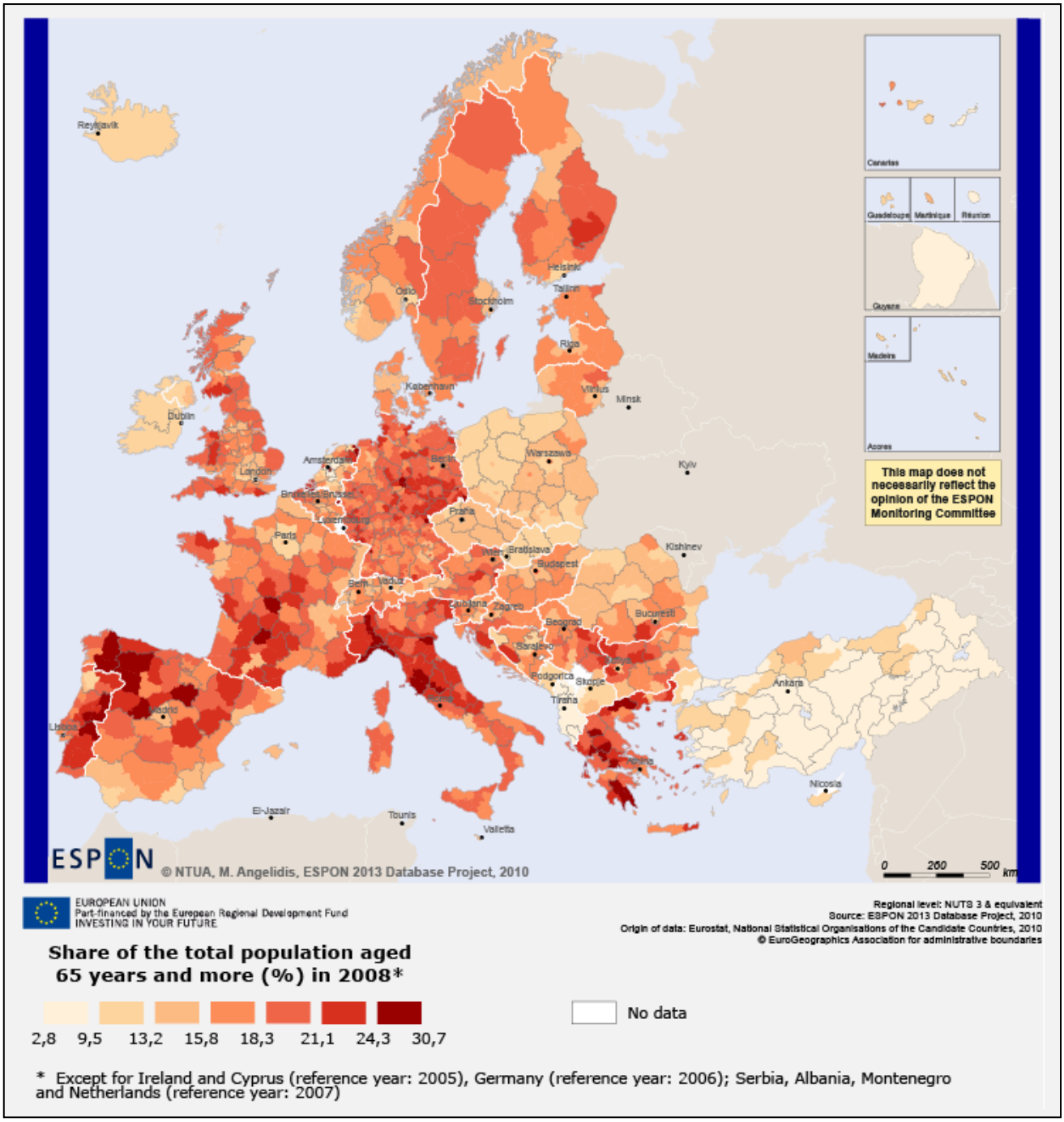
METHODOLOGICAL ISSUES

In order to ensure a sound comparability of data of the CC / PCC which have not adopted the NUTS classification, we have classified the existing administrative units of these countries at different territorial levels in **"similar to NUTS"** territorial units. We have used for this purpose the criterion of population potential of the EU NUTS classification as well as the overall structure of government in these countries with focus on the power of the respective regional and local authorities and the main features of territorial development in each administrative level per country. This method could be further used in the definition of "similar NUTS" divisions in the Eastern Neighbouring countries (ENC) and the Southern Mediterranean Neighbouring countries (MNC).

THEMATIC ISSUES

For the CC / PCC which have not adopted the NUTS classification, specific datasets and metadata at NUTS 0, 1, 2, 3 levels have been elaborated and included in the Database. Main sources of data are the NSO of the respective countries. The further development of both formal and informal collaboration of ESPON with Eurostat, DG Regio and the NSO could ensure a regular bilateral flow of territorial data for these countries.

Population aged above 65 years in the ESPON Area, Western Balkans and Turkey



The Map shows the population ageing in the Western Balkans and Turkey as well as in the ESPON space, at NUTS3 and "similar NUTS3" levels in 2008, using the population aged 65 years and over rate % as an example of the **close of the CC / PCC "gap".** This step is very important because it allows study the territorial particularities of these countries which should be taken into account in the future Cohesion and Neighbourhood Policies of the EU.

Basic indicators for Western Balkans: *139 indicators have been integrated in the database on the CC / PCC for the following "basic" themes: GDP, Area, Population, Population density, Age pyramid, Crude Births rate, Crude Deaths rate, Natural growth rate, Active population, Migration and Population sex, for the years 2000 – 2006 and NUTS 0, 1, 2 and 3 or "similar NUTS" 1, 2 and 3 levels.*

2.9 WORLD/REGIONAL DATA

KEY FINDINGS

- Elaboration of a provisional ESPON 2013 World Database with "indicators of reference" (Population, GDP, CO2 emissions, …) describing "units of reference"(states or territories) on a long period of time (1960-2010)
- Comparison of the official list of countries and related codification (ISO3) from main international "thematic" providers (UNEP, CHELEM, World Bank…). Elaboration of tables of correspondences between this database and the data previously collected by ESPON 2006 Europe in the World Project
- Linking of World data with Eurostat Regional data through a methodological tool (named "Gap Tracker") for explaining the differences between global databases and Eurostat data.
- Preparation of maps and graphics at global/regional scales in order to feed the first ESPON 2013 Synthesis Report

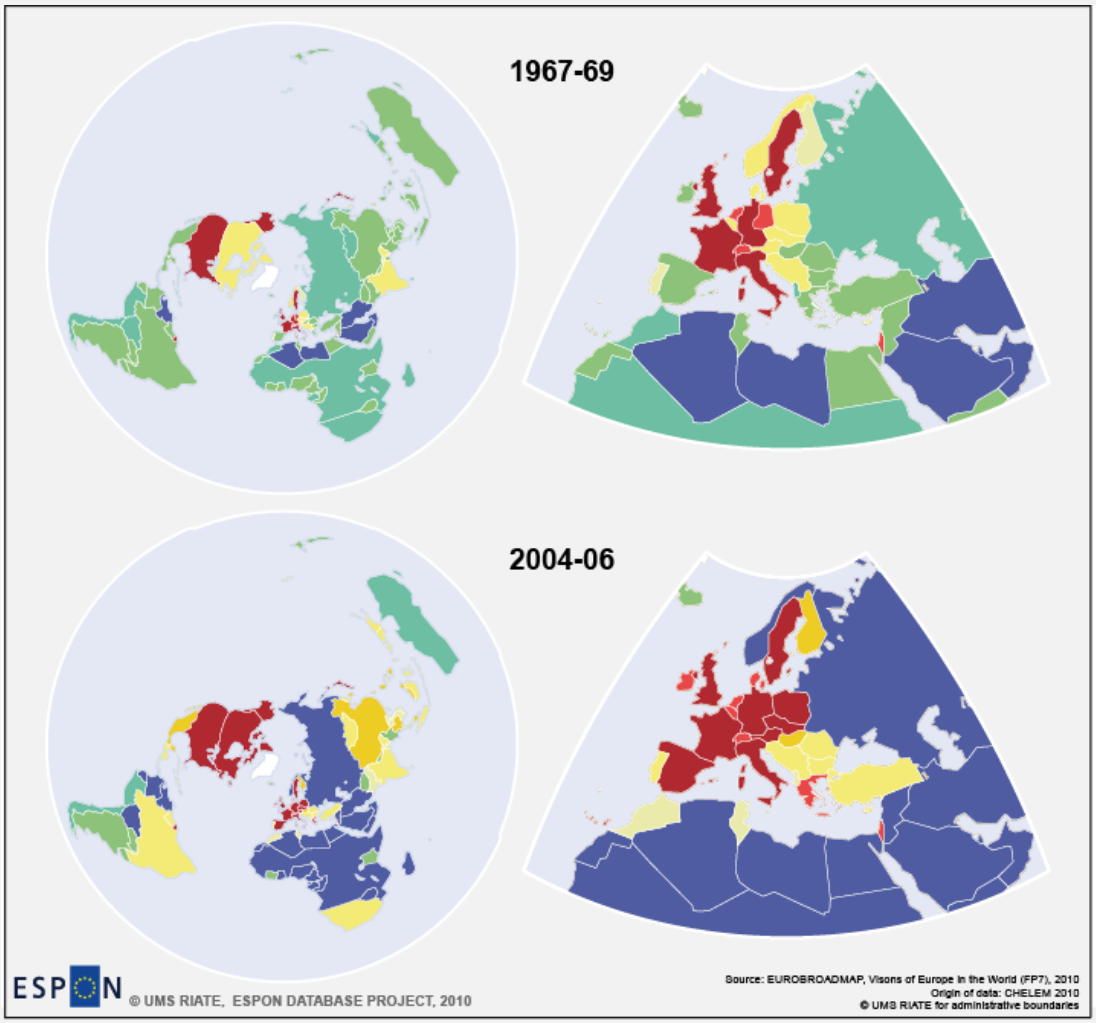
METHODOLOGICAL ISSUES

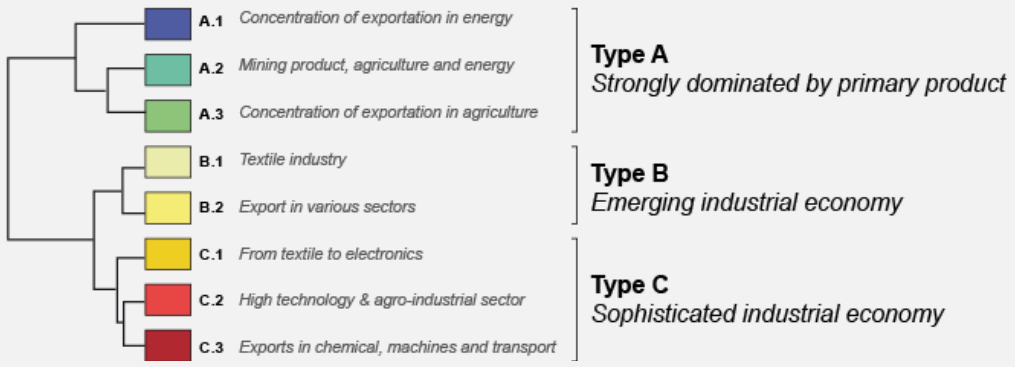
The ambition of the ESPON 2013 Program to support a "Five level approach" implies the elaboration of databases covering the "regional" scale (EU31+ southern and eastern neighborhoods) and the "global" scale (World) for different thematics (environment, demography, …) and different types of geographical objects (cities, countries, flows, …) at different periods of time (1960-2010). The main problem is not the collection of world data (more easy than NUTS data) but rather their internal harmonization (e.g. "France" can refer to various geographical objects, including or not the remote territories) and also the differences between World/European/National data providers (e.g. Population of UK in 2000 is very different according to UN, Eurostat or National Statistical Office).

THEMATIC ISSUES

The ESPON 2013 database project was not directly in charge of thematic exploration at world/regional scale but some experiments have been led in order (1) to evaluate how to map and store results of new projects like TIGER and (2) to support ESPON CU (Synthesis Report or SR). As an example, we shall mention an update of the map of discontinuities of GDP/inh in Europe in 2008 (SR, p. 57), a map of a global cities network in 2008 (SR, p. 32), a comparison of HDI and global footprint for world countries in 2006-2007 (SR, pp. 84-86), a typology of countries for trade export in a 40 year period (see Figure , right), and last but not least a complete illustration of the ESPON "Five level approach" for population growth 2001-2006 (Synthesis Report, pp. 15-16).

Typology of country profiles for trade export 1968 – 2005





At the beginning of the 21th century, we observe a very intense polarisation between the extended group of countries that export mainly energy or mineral ressources, and the group of industrialized countries which is more and more enlarged to new developping countries. The industrial cores of the 1960's are now clearly in competition with new industrial economy. In Asia as in America, the process of industrialization has clearly spread toward neighboring countries. The situation is fully different in EU, which is surrounded by countries with a lower level of industrialisation that export mainly energy or primary products, in both eastern (Russia, Central Asia) and southern (Africa) directions.

ESPON DB World indicators: *Urban and total land area, agricultural and total population, GDP, GNI and CO2 emissions in 2005 for 237 countries of the World + time serie (1980-2015) for total population and GDP + an Access database containing CO2 emissions and age pyramid data for a large time period.*

2.10 SPATIAL ANALYSIS FOR QUALITY CONTROL

KEY FINDINGS: DETECTING EXCEPTIONAL VALUES

- Exceptional values can arise from **(a)** data input or manipulation errors and **(b)** data values which are truly outlying (outliers).
- The accurate identification of an exceptional value is important because input errors should be treated differently to outliers.
- Input errors can usually be identified mathematically or sometimes, statistically. Outliers can only be identified statistically.
- A 'weight of evidence' approach to the statistical identification of outliers is proposed. The approach applies nine representative and complementary statistical outlier detection tests, where observations are flagged as outlying according the outcome of each test. This then builds up a 'weight of evidence' for the likelihood of a given observation being outlying (i.e. the evidence is strongest for an observation that is flagged as outlying for all nine tests). Key aspects of this approach are presented - summarising a more detailed presentation given in an accompanying technical report.

CONTEXT

It is paramount that the data to be stored in the ESPON 2013 Database should be as reliable as possible. It would be unwise to assume that data supplied to the Database is completely free from error. The activities under spatial analysis for quality control have led to the design and implementation of a battery of filters and tests against which potential input data can be tested. Exceptional values can arise for a number of reasons, so rather than employ a single test, several tests are applied. For outliers, this allows a judgement to be made as to the reliability of a sample observation based on the **weight of evidence** from an application of nine tests.

Exceptional values may arise during the coding, transmission, manipulation or editing of data. They may not be noticed until late in the day or not at all, particularly if the data that is a candidate for input to the Database has been the outcome from a series of sequential manipulations. Exceptional values may also arise in the measurement of data; perhaps a sensor on some measuring equipment has been incorrectly calibrated or is faulty; a respondent can tick the wrong box in completing a survey form. Exceptional values can also be true and valid observations.

DESCRIPTION

Logical input errors made during data input or manipulation in the ESPON 2013 Database can arise for a number of reasons. For example: incorrect NUTS codes might be entered or assigned in a table lookup; 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 process leading to the identification of an input error will follow some logical, mathematical approach that can be conveniently coded within the database architecture. For example, if a land use class could only take a positive integer value from 1 to 9 say, then an incorrect value 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 the distribution of values for the variable 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.

Statistical outliers can similarly arise for a number of reasons in the ESPON 2013 Database. In the accompanying technical report, we propose a novel 'weight of evidence' approach to the detection of outlying values, which has been developed since the Second Interim Report. As demonstration of this approach, it is applied to a real ESPON data set. The approach applies nine representative and complementary statistical outlier detection tests, where observations are flagged as outlying according the outcome of each test. This then builds up a 'weight of evidence' for the likelihood of a given observation being outlying (i.e. the evidence is strongest for an observation that is flagged as outlying for all nine tests).

The nine different tests deal with the identification of outliers with different characteristics which renders the value unusual. These characteristics are:

- **Aspatially outlying:** an unusually low or high value (i.e. an outlier in or near the tail of a statistical distribution).
- **Spatially outlying:** an unusually low or high value in a region when compared with values in neighbouring regions.
- **Temporally outlying:** an unusually low or high value with respect to a time series of observations.
- **Relationally outlying:** a pair (or group) of observations that relate to each other in an unusual manner. That is, the relationships are very different from what has been observed elsewhere in the database. For example, a particular region may have a high unemployment rate coinciding with a high GDP value (i.e. indicating a positive correlation) whereas we would normally expect high unemployment rates to coincide with low GDP values (i.e. a negative correlation). The statistical methods used to detect these kinds of unusual relationships are multivariate, whereas in the previous three cases, univariate statistical detection methods are used.

Combinated methods for detecting outliers

These four major forms of outliers are schematically depicted in the figure 10.1. Note that an observation may be outlying in more than one way; for example, it may be both aspatially and temporally outlying.

Individual and combination of variables in the database are evaluated against the nine detection tests. An observation that is found to be unusual on only one or two tests is considered less likely to b e outlying than an observation which is found to be unusual on all nine tests.

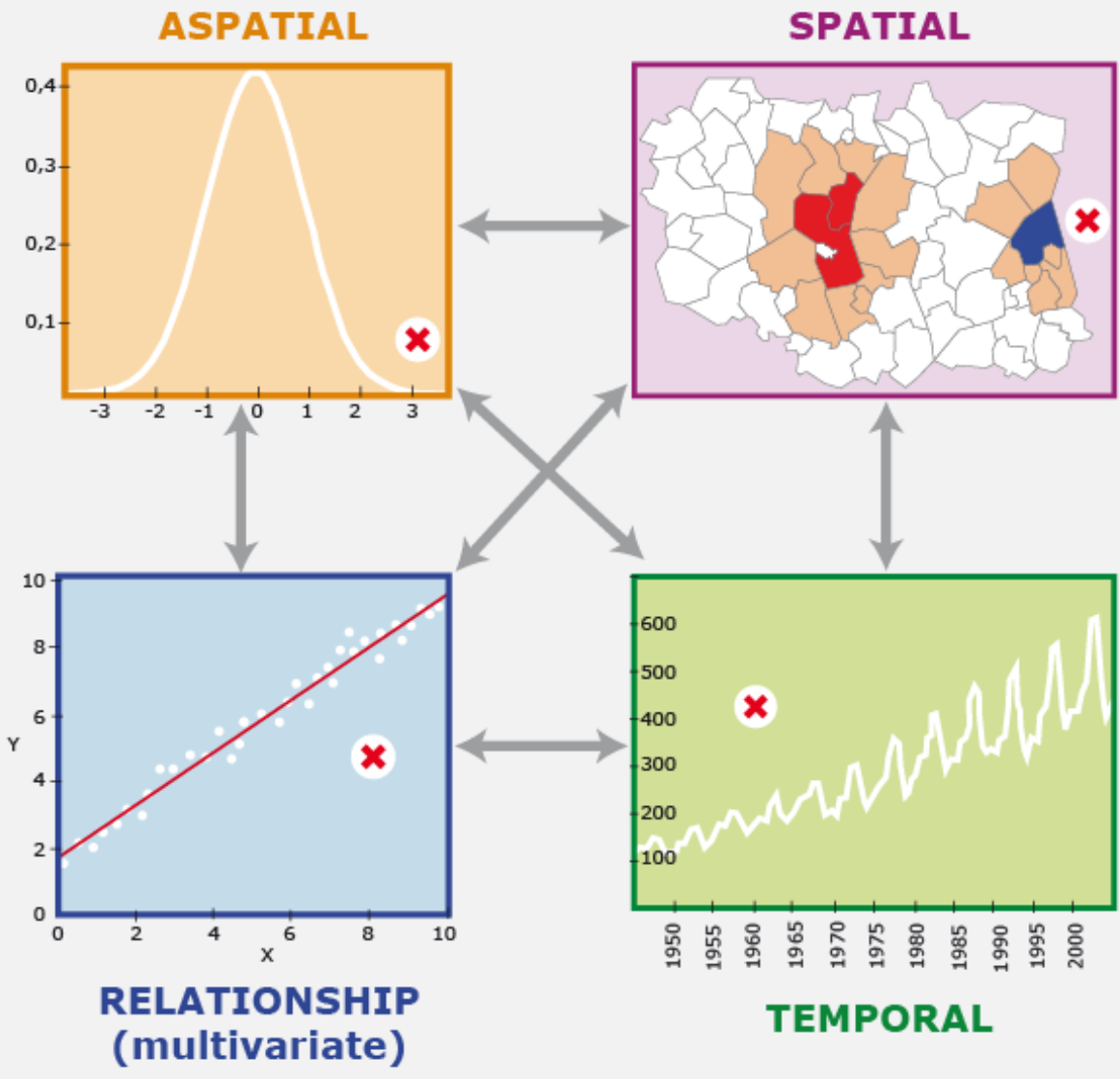


Figure 10.1 - Types of outliers detected

Borrowing from the accompanying technical report, **case study results** from the weight of evidence approach are given in Figs. 2 and 3. Here it is applied to ESPON unemployment rate data covering 709 NUTS23 regions for eight consecutive years 2000 to 2007. Fig. 10.2 presents maps for each of the nine outlier detection tests, whilst Fig. 10.3 presents the final 'weight of evidence' map. Thus as examples, there is strong evidence of at least one outlying unemployment rate (from the eight years) in a NUTS23 region in SW Spain and a NUTS23 region in N Corsica. Conversely, there is little evidence of at least one outlying unemployment rate (from the eight years) in a NUTS23 region in SW Ireland and all NUTS23 regions in Norway.

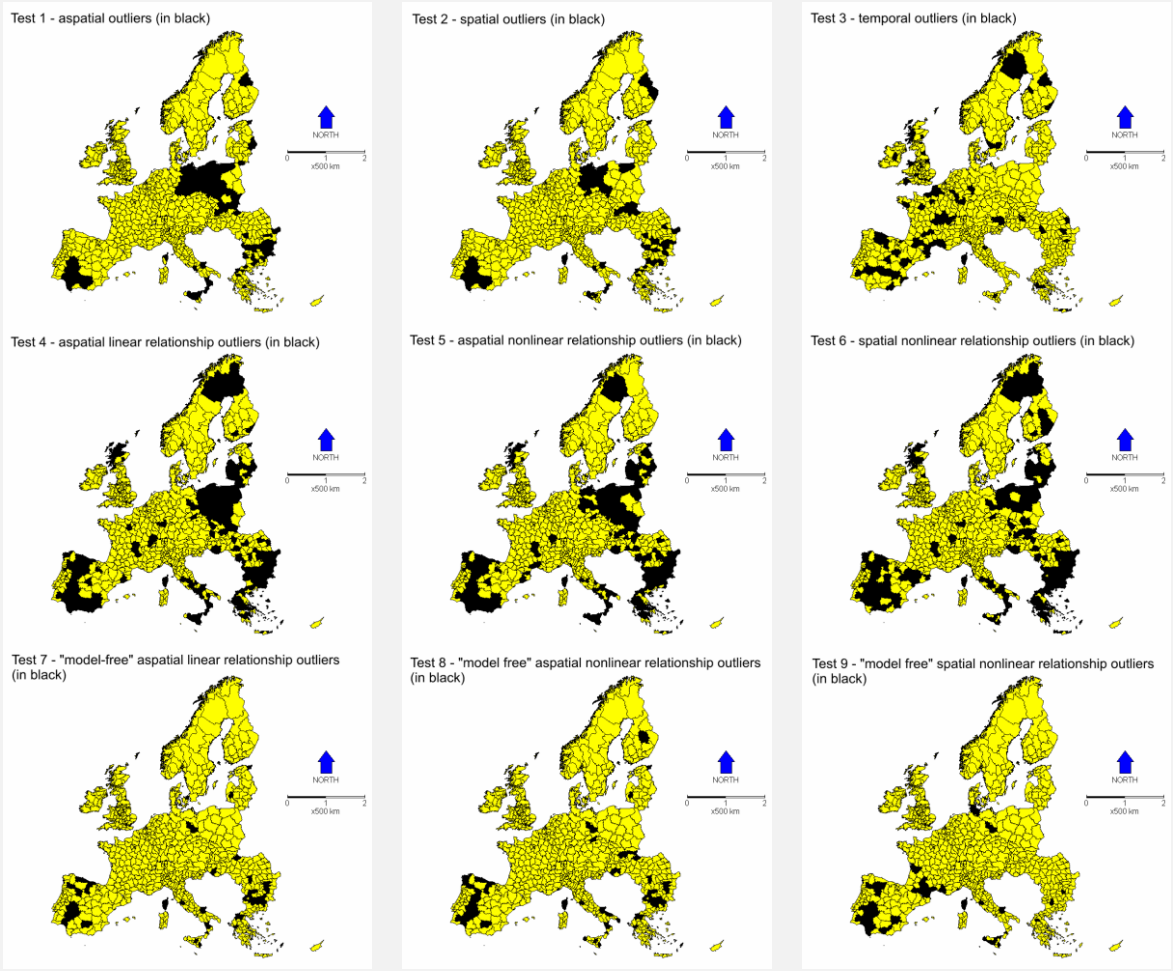


Figure 10.2 - Case study results for each of nine tests
(note that these maps are 'quick-look' screen dumps and are not intended as finished production quality outputs).

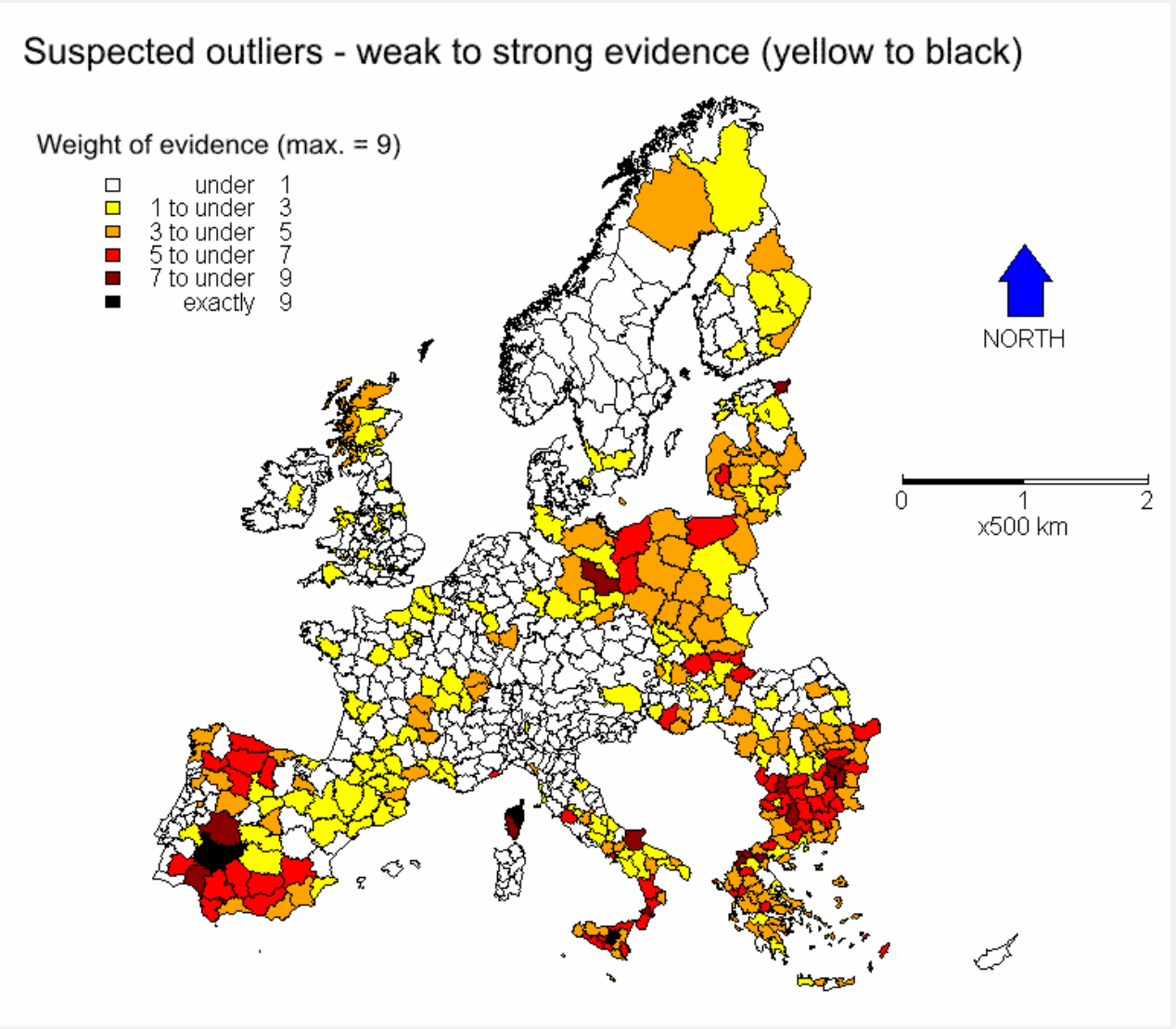


Figure 10.3 Case study 'weight of evidence' map
(note that this map is a 'quick-look' screen dump and is not intended as finished production quality output).

The decision of what action to take with regard to an identified logical input error or a statistical outlier (e.g. remove, replace, leave alone) should ultimately reside with an expert on the given data set (or its provider). Mechanisms can be put in place within the database architecture to do this in an efficient and effective manner.

For input errors, a simple removal and (if possible) a replacement will generally suffice. For outliers, considerably more attention is required. Outliers and their detection should not be naively viewed as a data cleaning or screening exercise, but can also be used to uncover interesting or unusual relationships in the database that has not been considered before.

CONCLUSION

In this conclusion, we summarize briefly the most important progress made during the project (section 1) and the most important difficulties encountered (section 2) before to address some proposals to our successors (section 3).

1. Progress made on ESPON Database during the project

The ESPON DB 2013 Project, in partnership with other projects from Priority 1 (TIPTAP, EDORA, DEMIFER, FOCI, RERISK) and Priority 3 (Demography, Accessibility, Lisbon Indicators, Typology, …), has elaborated a substantial database on European regions and cities, with very important added value for policymakers working on territorial cohesion. This database, that is now available on the ESPON website through an innovative computer application, will play a major role in the promotion of ESPON network and ensure a wider diffusion of results presented in the form of papers. At the same time, ESPON has developed stronger partnerships with data providers (Eurostat, EEA, National Statistical Agencies,) and data users (DG REGIO or DG AGRI). The ESPON 2013 Program as a whole is starting to be recognized as an important player in the field of databases at the European scale. The contribution of ESPON DB 2013 Project to this recognition has been crucial on several points:

A very strict definition of rules concerning metadata and quality check: this goal has been extremely time consuming (as INSPIRE directive and ISO norms were not directly applicable to many data used in ESPON). Even if it was a difficult constraint for our project, as for the other ESPON projects, the strict codification of metadata is absolutely crucial for ESPON external recognition.

The integration of various types of geographical objects : even if regional data (NUTS2 and NUTS3) remains actually dominant in the ESPON Database project, this one has been designed in order to open the door for data elaborated at upper and lower scales (World by states, local units) and for data using different geometries (cities, networks, …).

The attempt to enlarge time series towards past and future: as spatial planning is necessarily dynamic and prospective, we cannot limit our investigation to a short term period. But it has been demonstrated many times that it is impossible to enlarge future previsions (t+20 years) without an equivalent gain of information on past trends (t-20 years).

2. Difficulties that have been encountered

The first set of difficulties that we have faced within this project **was related to the ESPON agenda** and the fact that our Priority 3 projects started at the same time than other Priority 1 projects (DEMIFER, FOCI, TIPTAP, EDORA, RERISK) and data release (Demography, Accessibility, …). Starting 6 months before the other projects would have allowed delivering immediately basic data to the other ESPON projects and elaborating our metadata model or map kit tool, avoiding the use of an intermediate version that was imperfect and had to be modified several times. Therefore, starting earlier would have been better for all the parts involved.

The second set of difficulties was related to the difficulties (and the cost) of communication and networking within the ESPON Program in general and with external organization like EUROSTAT or DG REGIO. We had not anticipated the importance of this topic which revealed to be crucial but implies a lot of time of communication, explanation or discussion with a lot of actors. As an example, the ESPON CU asked to the ESPON Database Project to perform a manual data check of data delivered by Priority 1 or Priority 2 projects, which was a very time consuming task, to be done with strong time pressure (as it was a condition of payment for this projects). As a second example, it was also difficult to have regular contact with EUROSTAT or EEA because such a meeting should be jointly organized with ESPON Coordination Unit and not directly by ESPON Database Project.

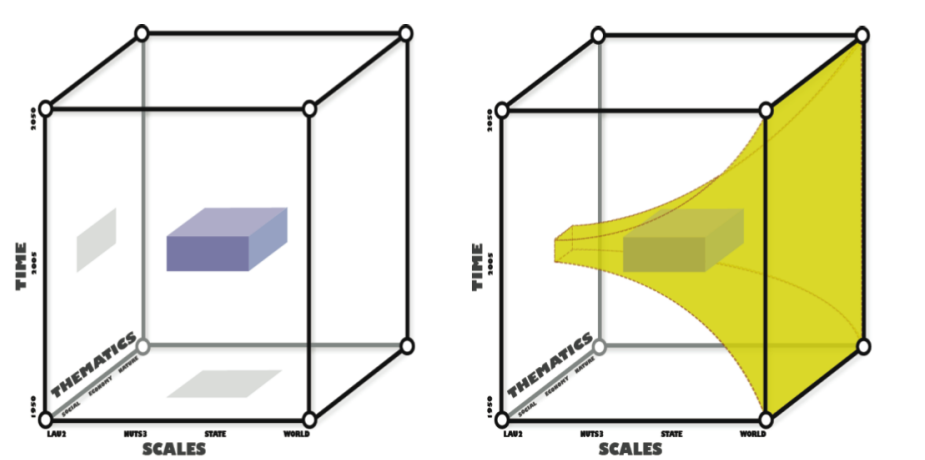
The third set of difficulties has been related to the excess of financial procedures. We know the rules of the ESPON program and they are valid until 2013. But we also know that the European Commission has insisted in 2008, after the crisis, on the necessity to make the financial rules easier and to avoid unnecessary administrative burdens. Our feeling as coordinators is that many times there was a danger of blocking the achievement of the ESPON DB 2013 Project. More and more work time, normally devoted to the productive part of the project, was in practice transferred to the management of administrative burdens related to "every-six-month-reports".

The fourth and final set of difficulties has been related to the lack of Knowledge and Support System. We were very frustrated to observe that our project did not benefit from a Knowledge Support System as the other major project of Priority 1. It was not necessarily because of the size of our project (more than all other Priority 1 project) but rather because of its strategic importance for ESPON that a KSS should have been established, with the best specialists of the domain (*e.g. G. Andrienko, M. Goodchild, …)* as a form of scientific recognition of the quality of the ESPON Database in the fields of computer science, cartography, geomatics.

3. Recommendation for the future

Based on the experience gained during the period 2008-2010, we suggest some recommendations to the successors of our project.

A) Maintain the ambition of enlarging database dimensions



Initial situation of ESPON 2006 Objectives for ESPON 2013

This objective remains fully accurate and a lot of work has still to be done in order to support innovative applied research on territorial cohesion.

B) Reinforce the networking dimension inside and outside ESPON

As explained in previous section, the ESPON Database project is necessarily connected to all other ESPON projects of Priority 1, 2 or 3 during their lifetime. In the initial stage, ESPON database provides new projects with data, map kits, technical reports … In the final stage, ESPON database receives data elaborated by this projects and checks it before integration in the database. All of this implies a strong cooperation and more contacts than the opportunities offered by the ESPON seminars every 6 months.

Outside ESPON, it is of crucial importance to have more regular communications with data providers at European, national or global levels. It is also important to be in contact with the scientific community working on advanced innovations in GIS, Cartography, Data Modeling…

C) Develop joint methods for automatic outlier detection and missing value estimation

One of the most important discoveries made during the ESPON 2013 Database project is the fact that estimation of missing values and detection of outliers are not two separated problems but a single one. To be sure, when you declare that a statistical value X is an "outlier", it is necessarily because you have an implicit model of estimation of X that indicates you that the model estimation X* is dramatically different from the observed value X. Therefore, all methods of outlier detection are also, by definition, methods of estimation for missing values. The four methods elaborated by the NCG team for outlier detection (see. 2.10) are in practice equivalent to the four dimension of the ESTI model proposed by LIG, RIATE and Géographie-cités for estimation of missing value and used in the challenge of time series (see. 2.1). At present time, we have mostly used simple method based on only one of the four possible dimensions (statistics, space, time, thematic) but better estimation of missing values or better detection of outlier can be expected by multidimensional methods.

D) Quality rather than Quantity

Repeating one more time the motto of our project, we would like to underline that many "big" databases has disappeared like dinosaurs because they did not make sufficient effort on the quality side and more precisely on the question of metadata. It is certainly true that the pressure made by our ESPON 2013 database project on the other ESPON projects for a very strict check of data and codification of metadata delivered (see. 1.2 and 1.3) has eventually limited the number of data they decided to deliver with their final report… But we fully assume this Malthusian perspective on data collection because it is the only sustainable strategy in the long run. Especially in the case of an applied research project where data can be used for political decision and should be fully subject to control of sources, estimation made, etc.

E) An open ESPON database

Better than storing a lot of Gigabytes, ESPON Database should offer a limited number of original and very high quality data which would offer him a specific place in the European and world network of data producers. This specialisation should be balanced by the opening of countinuous data flows and exchanges with Eurostat, EEA, Eurogeographics… but also OECD, UNEP, UNPP and more generally all National Statistical Institutes of the ESPON 31 area and the neighboring countries.

ANNEX 1 – Description of the footnotes

The ESPON Database is organised in several categories, taking into account the different types of data that it is possible to download from the application:

Regional data – Web interface: This category contains all regional data (in the NUTS delineation) available in the ESPON Database. These datasets are produced by ESPON Projects (mainly priority 1 projects) and it is possible to query the database through many kinds of requests: by theme, by project… (cf part 1.5 of the report)

Local data: This category contains all material – data and geometries - related to local territorial units (LAU2 level). In detail, this category contains both European databases at local level (restricted access) and also data derived from ESPON Priority 2 projects, which "zooms" on specified territories.

Neighborhood data: This category contains all data related to neighboring countries. At the moment, it contains only data of Western Balkans and Turkey at regional level (SNUTS units) but in a near future, data collection will be extended to Southern and Eastern Neighborhood of the ESPON Area.

Urban data: This category contains all data related to urban data: Functional Urban Areas, Larger Urban Zones, Morphological Urban Areas…

World data: This category contains all data related to countries of the World. It is namely in this field of the database where it is possible to download data from Europe In the World project (ESPON 3.4.1).

Grid data: This category contains all data available on grids (population grid from the Joint Research Center, Corine Land Cover of the European Environmental Agency…).

Historical data: This category contains all information related to the last versions of NUTS units (1999, 2003). It contents namely the dictionary of NUTS changes, Eurostat historical databases and the previous ESPON Database (Access tables).

Other data: This category contains databases delivered in various format (shapefiles, geodatabases), which not necessarily follows the ESPON Metadata templates (data coming from various institutions).

Technical reports: In this category it is possible to download all the technical reports produced by the ESPON Database Project. These technical reports describe how to solve specific problems of data integration that cannot be fully explained in the very brief description that is usually given in the metadata files.

Conseqently, the logos shown in the footnotes of the final report show in which category of the web application it is possible to download datasets and material produced by the ESPON Database Project.

The ESPON Database home page

ANNEX 2
Overview of the ESPON 2013 Database thematic structure

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

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

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

04 ENERGY AND ENVIRONMENT
0401 Energy and resources (e.g. renewable, nuclear, and fossil energies)
0402 Climate change (e.g. GHG emissions, air pollution)

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

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

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

99 CROSS-THEMATIC AND NON-THEMATIC DATA
9901 Integrative indices, indicators and scenarios (e.g. typologies, scenarios)
9999 Geographical objects (e.g. administrative units, grids, networks)

ANNEX 3
Survey on ESPON Database

The original survey on ESPON Database is available at the following:
http://survey.ums-riate.fr/index.php?sid=93326&lang=en

Answers have been collected between the 10th March and the 28th March, based on the ESPON Contact Point network.

Only the web interface (regional data) has been evaluated (it excludes indeed the extension of the interface (urban data, local data etc.).

TABLE OF CONTENT

A) SAMPLE DESCRIPTION

Due to the limited time available before the delivery of the final report, only 29 answers has been collected. These part analyze in detail the characteristics of this sample.

Age and Sex

	Women	Men	??	Total
<30			0	**0**
30-45	4	9	0	**13**
46-60	2	5	0	**7**
>60		1	0	**1**
??	5	2	1	**8**
Total	**11**	**17**	**1**	**29**

The answers are mainly from people aged 30-45 and eventually 46-60, but 8 people refused to indicate their age. Gender equilibrium is not perfect with a clear majority of men (17 on 28).

Countries of birth or activity

	COUNTRY OF …	
	… birth	**Professional activity**
Belgium	4	4
Cyprus	1	1
Czech Republic	2	1
Estonia	1	1
Finland	1	1
France	2	1
Germany	1	0
Greece	1	1
Hungary	1	1
Ireland	1	1
Italy	1	1
Latvia	2	2
Luxembourg	1	2
Malta	2	2
Netherlands	0	1
Poland	1	1
Romania	1	1
Slovenia	2	2
Spain	1	1
Suriname	1	0
Switzerland	2	1
United-Kingdom	0	1
No answer	0	2
Total général	**29**	**29**

The distribution of countries of birth or professional activity is rather equilibrated, except an over-representation of Belgium. As a whole 21 countries of the ESPON area are represented by at less one people.

Fields of professional activity

Fields of activity	Total
Geo	21
Geo/Eco	1
Geo/Law	2
Geo/Social Sciences and GIS	1
Geo/Sta/Eco	1
Journalist	1
Law	1
Management science, & regional development	1
Total général	**29**

With very few exceptions, the people who answered to the survey were working in the field of geography and spatial planning. Some of them combined it with Economics, Statistics and Politics.

Scale of territorial action

	Monoscalar	Multiscalar	Total
Local	2	12	**14**
National	4	11	**15**
Transnational	0	4	**4**
European	5	13	**18**
Global	0	0	**0**
Total	**11**	**40**	**51**

Concerning the scales of territorial action, the majority of answer is related to the "European" level followed by the "National" and the "Local". Very few declared the "Transnational" scale and none of them the "Global" one. It is interesting to observe that only a minority of people (11/27) declared only one scale. In the majority of case (18/27), people chose to declare multiple scale of interest like "European/National"(4 cases) , "European/National/Local"(3 cases) , or even "European/Transnational/ National/Local"(2 cases).

Professional status and relation with ESPON Program

	Policy maker	Researcher	no answer	Total
Monitoring Committee	4			**3**
Espon Contact Point	4	10		**13**
Lead Partner / Project Partner		6		**6**
Not ESPON	2	1	2	**5**
Total général	**10**	**17**	**2**	**29**

The majority of answers were provided by ESPON contact points. They have insured an efficient promotion of the survey and contact researchers when they were not able to answer directly themselves. The success was not equivalent on the Monitoring Committee side where only 4 answers was gathered (including one answer from European Commission). As a whole, the survey appears not perfectly equilibrated in terms of proportion of people that declared to be firstly "Policy makers" or "Researchers and experts".

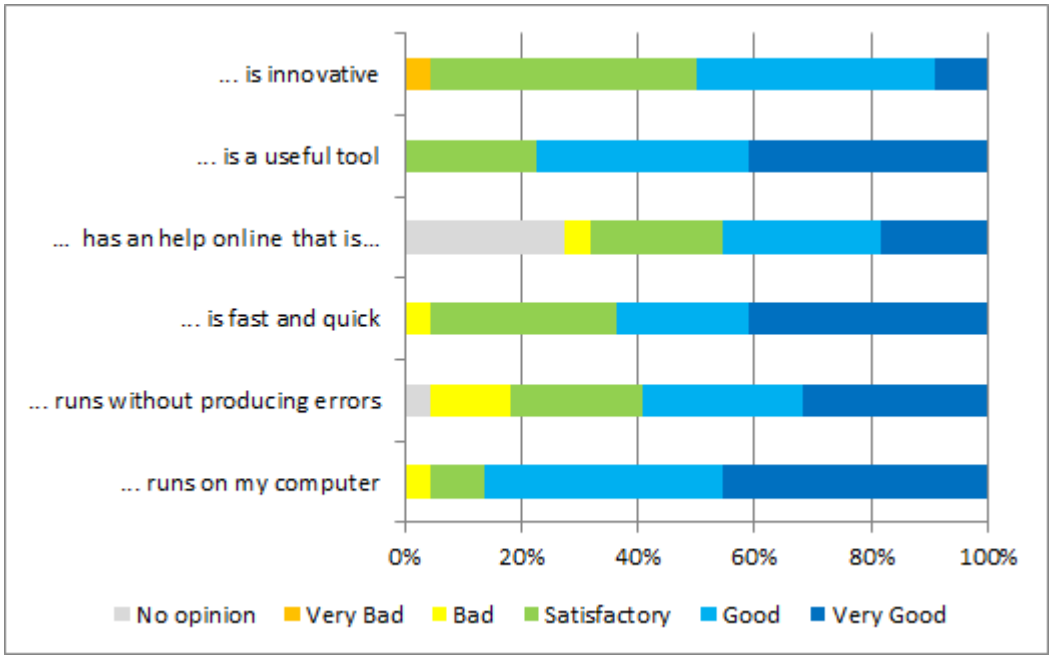
Conclusion: size and quality of the sample

The two surveys about ESPON Database and HyperAtlas V2 was realized jointly in order to save time and effort (e.g. same background questions). 19 people agreed to answer to the two parts of the survey (ESPON DB+HyperAtlas) but 4 participated only to ESPON DB part of the survey and 4 other participated only to HyperAtlas Survey. People generally tried 1 or 2 times time each application, but some of them used 3 times or eventually more each application. We can therefore conclude that in both scale the sample is reduced (20-25 people) but of good quality and sufficient to draft preliminary observations.

B) EVALUATION OF ESPON DATABASE

Overall appreciation

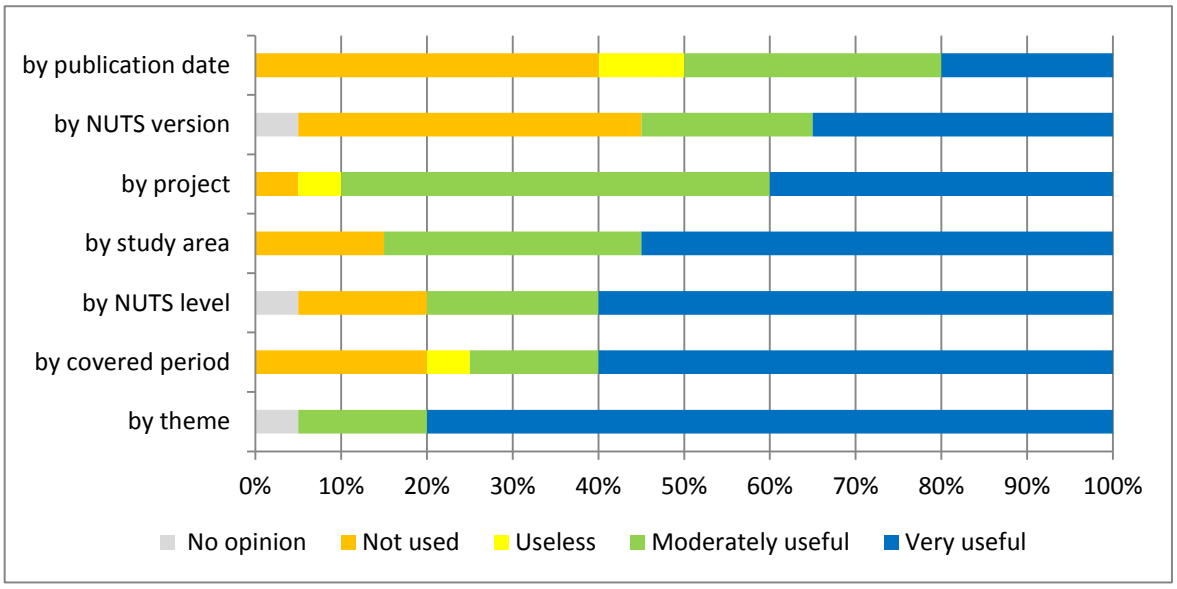
	You've tried ESPON DATABASE. Please would you say that the application...					
	No opinion	**Very Bad**	**Bad**	**Satisfactory**	**Good**	**Very Good**
... runs on my computer	0	0	1	2	9	10
... runs without producing errors	1	0	3	5	6	7
... is fast and quick	0	0	1	7	5	9
... has an help online that is...	6	0	1	5	6	4
... is a useful tool	0	0	0	5	8	9
... is innovative	0	1	0	10	9	2
Total	**7**	**1**	**6**	**34**	**43**	**41**



The overall appreciation of the ESPON Database Application is generally "good" to "very good" for all the criteria proposed. From technical point of view it appears that it generally ran well on the computers but sometime with errors as 10% of the users declared that it run "bad" and 20% declared only "satisfactory". This appreciation is correlated with the fact that only 60% declared "good" or "very good" for the item "fast and quick" but 30% declared only "satisfactory" and one answer found the performance "bad". The help on line was apparently not found by 6 people who delivered no appreciation. Concerning more general appreciation, the survey indicate that the application was perceived as very useful ("very good" for 40%, "good" for 40% and "satisfactory" in other cases) but not so innovative (only 50% of "good" and "very good", 45% of "satisfactory" and one single "very bad" quotation). As a whole, results are a bit lower than for ESPON HYPERATLAS.

STEP 1 : Usefulness of research criteria

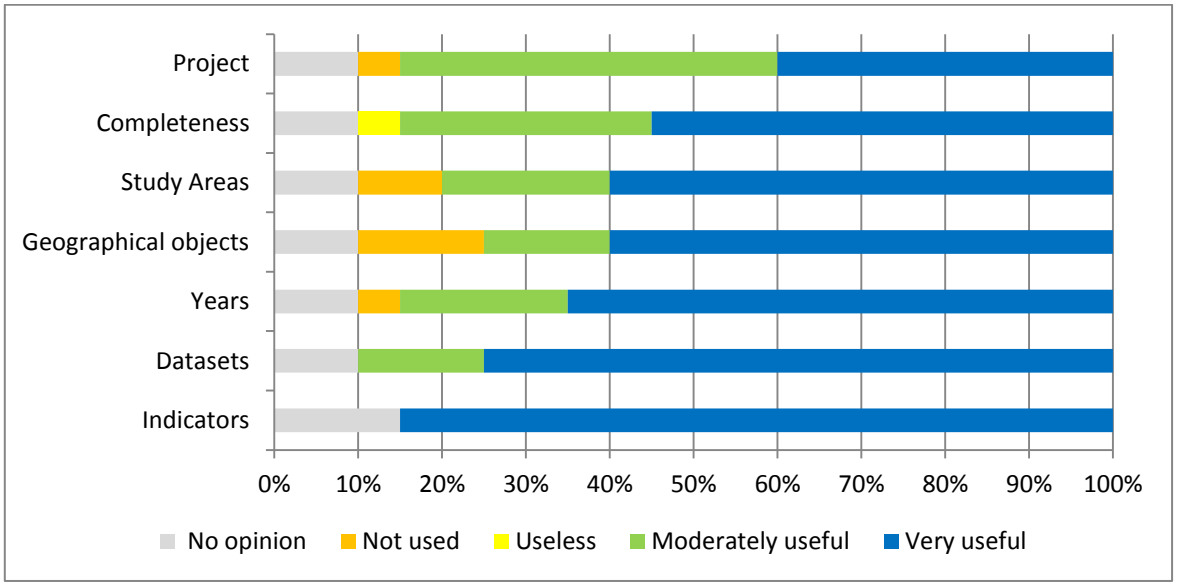
	ESPON DATATABASE 2013 APPLICATION How usefull dif you find the following research criteria?				
	No opinion	**Not used**	**Useless**	**Moderately useful**	**Very useful**
by theme	1	0	0	3	16
by covered period	0	4	1	3	12
by NUTS level	1	3	0	4	12
by study area	0	3	0	6	11
by project	0	1	1	10	8
by NUTS version	1	8	0	4	7
by publication date	0	8	2	6	4



The majority of people have tried both simple and advance criteria of research. Looking at their appreciation of the usefulness of the different criteria, we can notice that ***the favorite criterion is the research by theme****, which is considered as "very useful" in 80% of answers. People appears to be also* ***very attracted by space-time criteria*** *: covered period, nuts level and study area are declared as very useful by more than 50% of answers. The* ***research by project appears generally moderately useful*** *(50% of answers), even if it is declared "very useful" by 40% of answers. Finally, the less attractive criteria appeared to be the research by NUTS version or by publication date. But it is important to keep in mind that the ESPON DATABASE 2013 provided few possibilities to use these criteria at the moment of the survey and in many cases people simply did not used this criteria and did not really gave negative advice on it.*

STEP2: Previsualisation of data

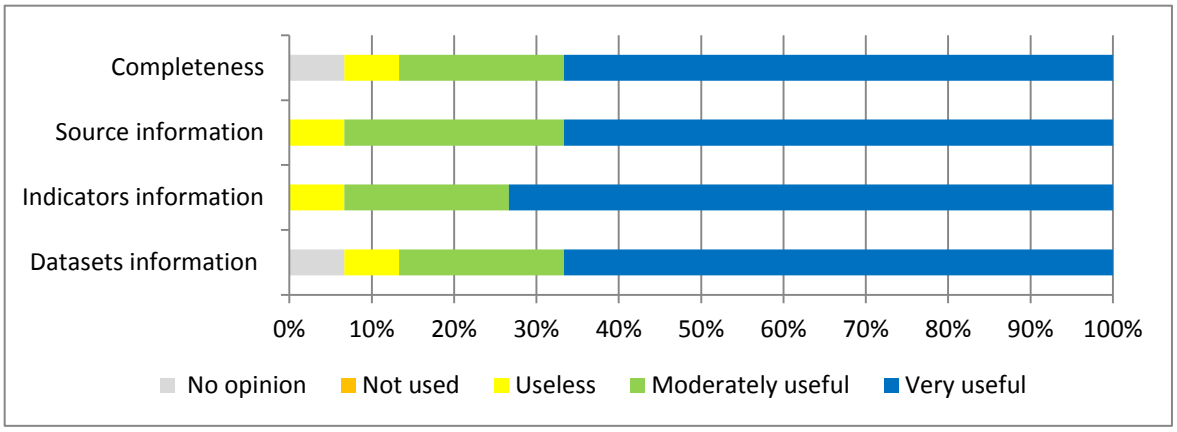
	ESPON DATATABASE 2013 APPLICATION How usefull dif you find the previsualisation informations ?				
	No opinion	**Not used**	**Useless**	**Moderately useful**	**Very useful**
Indicators	3	0	0	0	17
Datasets	2	0	0	3	15
Years	2	1	0	4	13
Geographical objects	2	3	0	3	12
Study Areas	2	2	0	4	12
Completeness	2	0	1	6	11
Project	2	1	0	9	8



The question related to the step 2 of previsualisation of data reveals that all criteria are appreciated by users, but with different degrees? The most useful information is without any doubt the definition of the indicators, followed by the name of the datasets. Then, people are further attracted by space-time information on the years, geographical object and study area. They are also interested generally interested by the Completeness of dataset, which is a good news for the author of the database because it was a huge work to develop this possibility Finally, we can observe that, as in the previous question on criteria of research, people who answered to the survey are clearly less interested by the project that has elaborated the data. This results could be positively interpreted as the fact that the users of the ESPON DATABASE consider the ESPON program as a whole and not as a collection of separated project.

STEP2: Previsualisation of metadata

	ESPON DATATABASE 2013 APPLICATION How usefull dif you find the metadata ?				
	No opinion	**Not used**	**Useless**	**Moderately useful**	**Very useful**
Datasets information	1	0	1	3	10
Indicators information	0	0	1	3	11
Source information	0	0	1	4	10
Completeness	1	0	1	3	10



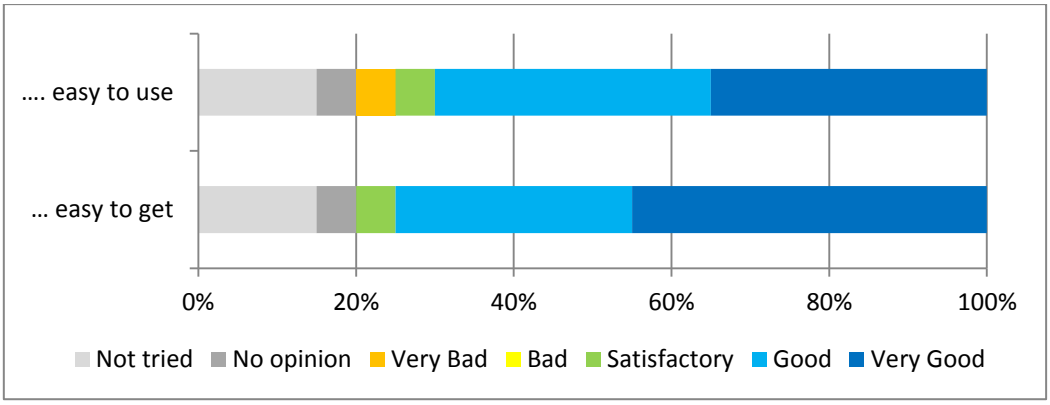
The previsualisation of metadata was clearly less used (15 answers) than the previsualisation of data (20 answers) and we can assume that it was mainly tested by specialist users. Anyway the level of satisfaction was pretty high with 65 to 70% of users that declared all the information on metadata as "very useful" and 20 to 30% "moderately useful". Only one single user considered that metadata information was "useless".

Looking at the more detailed answers, we can notice that many users are not very comfortable with the notion of completeness ("It is not clear to me what Completeness should represent"). Another answer notice that completeness is measured at the dataset level and would prefer more details by indicator ("On the completeness menu I realized that all the indicators have the same range - from 0 to 100 - this way many of them appears all in one color if the value is population, for example. It should be adapted to each indicator").

It is also mentioned that "The website address of the project could be useful" and that one should provide directly : one answer observe that " a more complete picture of the source of the data. What is the origin of the data? Eurostat? National datasets? ..." and another suggests that "it would be useful to provide more information about the original source of the datasets, clear definition or for indicators a link to a documentation how the indicator was constructed."

STEP3: Download of results

	ESPON DATABASE 2013 Would you say that the Excel files that you downloaded was …						
	Not tried	**No opinion**	**Very Bad**	**Bad**	**Satisfactory**	**Good**	**Very Good**
… easy to get	3	1	0	0	1	6	9
…. easy to use	3	1	1	0	1	7	7
Total	**6**	**2**	**1**	**0**	**2**	**13**	**16**

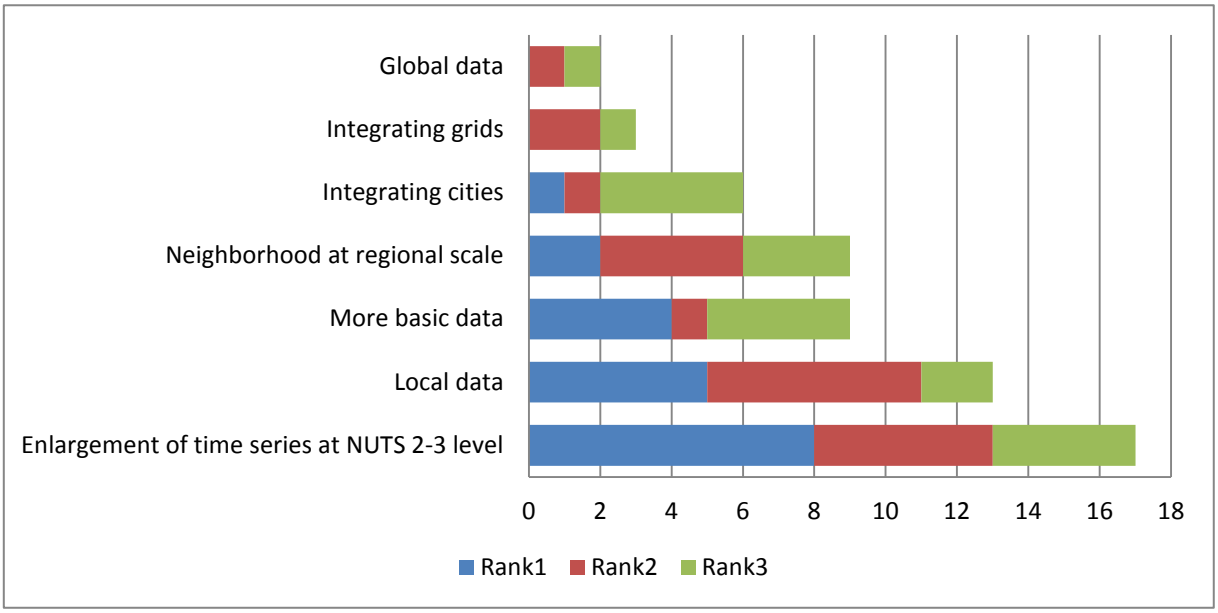


The final step of downloading data was realized by 16 users only. 3 users did not try and 1 declared to have no opinion (and probably did not try). Among these 16 users, the majority found data easy to use and easy to get, with a clear domination of "good" and "very good" answers. But at less one user declared that the application was just "satisfactory" on both criteria and another one found it "very bad" for the criteria of "easiness to use"?

The user that had provide the most negative advice explains that "if we are not aware of how it was created and what each topic means it results kind of difficult to deduce some information. (ex. on lineage or on the label of lisbon strategy data)". Another user that has ranked the application as "good" on both criteria observes that " It is easier to handle the data if no "Label column" is integrated in between the data columns. Since the data is the same for the whole column it is better if it is integrated into the metadata". This opinion is shared by another user that explains that "sheets containing the data should have only one heading row instead of three, in order to make them useable as database tables". Two other answers suggest that the format of the tables should be explained briefly in each Excel file that people download.

Priorities for further development of ESPON Database

	ESPON DATABASE 2013 Proposal of Improvement (3 maximum)			
	Rank1	**Rank2**	**Rank3**	**Total**
Enlargement of time series at NUTS 2-3 level	8	5	4	**17**
Local data	5	6	2	**13**
More basic data	4	1	4	**9**
Neighborhood at regional scale	2	4	3	**9**
Integrating cities	1	1	4	**6**
Integrating grids	0	2	1	**3**
Global data	0	1	1	**2**



The final question of the survey was related to the priorities for further data collection in ESPON Database. People were invited to choose 3 items among 7 proposals, and eventually to propose other ideas. Looking at the distribution of answers, it appears very clearly ***that enlargement of time series and development of local data*** *are the two main priorities for the majority of people that has answered. A second group of request is related to the* ***development of basic data collection, with enlargement to neighborhood at regional scale****. Surprisingly, we can observe* ***a limited number of answers for the development of data related to cities****. This result is tricking because the EU regional policy is more and more oriented toward the integration of data on metropolitan area. Less surprising is the fact that* ***very few answers was related to grid data and global data****, probably because these data has been very few used in ESPON until now.*

Looking at the complementary comments, one answer suggest that ESPON should "have the spatial data regulated according to the INSPIRE directive (metadata, services sharing, reporting, etc.), so it can be in line with other European projects. This way it should be even more useful. Another answer insist on the fact that "statistics get outdated quite fast. Thus the data should be added to the database as soon as possible". Finally one answer indicate that "The ESPON DATABASE should also provide acess to the shapefiles so that it is possible for users to map the information for themselves."