ESPON 3.2 - Data Navigator 2 - Final Report

Joël Boulier, Claude Grasland, Timothée Giraud, Nicolas Lambert, Jérôme Gensel, Bogdan Moisuc, Marlène Villanova-Oliver, Sylvia-Sorana Mogosan, Moritz Lennert, Pablo Medina, et al.

ESPON

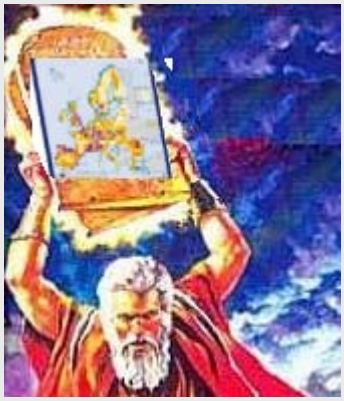
EUROPEAN SPATIAL PLANNING OBSERVATION NETWORK

ESPON 3.2

DATA NAVIGATOR 2

Final Report

Part 1 – Handbook for Data Collection



Co-financed by the European Community through the Interreg III ESPON Programme

This report represents the final results of a research project conducted within the framework of the ESPON 2000-2006 programme, partly financed through the INTERREG III ESPON 2006 programme.

The partnership behind the ESPON programme consists of the EU Commission and the Member States of the EU25, plus Norway and Switzerland. Each country and the Commission are 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.lu

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

ISBN number:
This basic report exists only in an electronic version.
Word version:

List of contributors

Teams members of the Hypercarte Research Network

UMR Géographie-cités (FR)

Joël Boulier*

Claude Grasland*

Timothée Giraud

Thimothée Giraud

UMS RIATE (FR)

Nicolas Lambert

IMAG (FR)

Jérôme Gensel*

Bogdan Moisuc

Marlène Villanova-Oliver

Sylvia-Sorana Mogosan

Other partners

IGEAT (BE)

Moritz Lennert
Pablo Medina Lockard
Gilles Van Hamme

TIGRIS (RO)

Octavian Groza
Ionel Muntele

* Scientific coordinators

Part 1. Handbook for data collection

Part 2. Experiments on database integration

Table of contents

TABLE OF FIGURES

TABLE OF MAPS

TABLE OF BOXES

Part 1 - Handbook for Data Collection

Introduction

The initial aim of the scientific support study *ESPON 4.1.2. Data Navigator II* was to produce a **handbook on data acquisition and harmonization** giving guidance to future ESPON projects, which should be presented as a standalone document to be used in a practical project implementation. The basic idea was therefore to **consolidate** the ESPON program and to reinforce the quality control of data in future the ESPON II Program (*see . Box .1*)

Therefore it should take into consideration the type of data needed for carrying out the ESPON applied research projects in particular the regional breakdown reflected in the ESPON Database. This project should include orientations and procedures for data acquisition, harmonization and quality assessment and also include recommendations for creating new datasets for the ESPON territory based on data from national sources. The underlying objective is to ensure that the ESPON programme follows common standards on data acquisition and harmonization. It is particularly important for the credibility of the regionalised datasets created for the ESPON space that different aspects of data quality are checked before displaying the information on a European map or having the data included in the ESPON Database. Thus the overall aim of this project is to produce a handbook giving guidance to future ESPON projects based on the lessons learned within ESPON so far, which should be presented as a standing alone document to be used in a practical project implementation."

Box 1. Extract from the initial terms of reference of Data-Navigator II

Our research group did not answer to this initial tender because (1) the terms of reference was not realistic and (2) we did not agree with the idea of simply consolidating the existing way of collecting data in ESPON 2002-2006. As a matter of fact, we have considered that the elaboration of the handbook for data acquisition and harmonization would not be useful if not linked to a discussion on two other topics:

- **The enlargement of data sources in ESPON II**
- **The choice of the data model to be used in ESPON II**

It is only once this modification of the objectives of the project has been officially accepted by ESPON CU and ESPON MC that our consortium decided to try to propose some preliminary ideas on a very difficult topic. Views we expose in this document should be considered rather as explanatory proposals than final solutions.

This report is divided in four parts:

1. The **Data circuit** part describes briefly the system of data collection set up in the first phase of the ESPON program by giving an global overview (1.1). We present the dual role of Transnational Project Groups which are, at the same time, users and producers of data and, consequently, both receive input data (1.2) from multiple sources and send output data (1.3) towards the ESPON DB.
2. The **Rules for data collection** part firstly gives some general principles for data collection which are summed up in the form of "10 commandments" (2.1). These commandments come from an analysis of the data circuit presented in section 1. They also enlighten the necessity of re-thinking the data circuit and the way data are stored and managed. We therefore propose a new data model, called Long Term Data Base, whose main aim is to guaranty the right application of the commandments (2.2). Then, we show that this model also serves as a support for both the estimation of missing values and the elaboration of practical rules for quality control (2.3).
3. The **recommendations for ESPON 2013** concludes this first part and summarises our proposals for further research.
4. The **Experiences** part presents, in a separate part, a set of concrete examples built by the experts of our group and which deal with the thematic harmonization of data (4.1), the extensive use of national sources (4.2) and the integration of environmental and socio-economic data (4.3).

1 The Data Circuit

We strongly believe that (1) the elaboration of a better statistical system is a necessary condition for the political construction of Europe and (2) the development of regional policies in Europe is hampered by the lack of harmonized and high quality databases at regional level. As noticed by the geographer C. Raffestin[1], the building of a state is always related to the building of a harmonized statistical system. The "monopoly on the legitimitate use of statistical information within a given territory" is an attribute of state which is as important as the "monopoly on the legitimate use of physical force within a given territory" proposed by M. Weber as an essential characteristic of political construction. Being an emerging political construction, the European Union suffers from many weaknesses in its technical bodies. However, it can be noticed that each progress in the building of Europe is somehow related to progresses in the quality of statistical information provided by European agencies like the European Environment Agency (EEA) and Eurostat. In the early 1980's, the quality of socio-economic data was quite poor as Eurostat was most of the time constrained to include data sent by national statistical offices "as they were", without any possibility of correction or, worse, no way to decline… Step by step, the power of the European statistical system has increased: it can now refuse or criticize data sent by national offices, it can elaborate its own definitions and launch specific production of harmonized data (e.g. Labour Force surveys, Corine Land Cover) and it has launched specific initiatives for quality control (Joint Research Center) or global accessibility to information through the INSPIRE directive (see. Box. 2). Still, these progresses in data quality and harmonization are rather slow but they are

[1] **Raffestin C., 1980**, *Pour une géographie du pouvoir*, Paris, Litec.

a positive signal of the increasing integration of the statistical system and, consequently, of the political construction of Europe.

The INfrastructure for SPatial InfoRmation in Europe initiative (INSPIRE) aims at making available relevant, harmonized and quality geographic information for the purpose of formulation, implementation, monitoring and evaluation of Community policy-making.

INSPIRE Principles

- Data should be collected once and maintained at the level where this can be done most effectively.
- It should be possible to combine seamlessly spatial data from different sources and share it between many users and applications.
- Spatial data should be collected at one level of government and shared between all levels.
- Spatial data needed for good governance should be available on conditions that are not restricting its extensive use.
- It should be easy to discover which spatial data is available, to evaluate its fitness for purpose and to know which conditions apply for its use.

Source : http://inspire.jrc.it/

Box 2. INSPIRE: INfrastructure for SPatial InfoRmation in Europe

One question to be asked so far concerns the position of ESPON in this general picture? To our opinion, the importance of ESPON should be neither over-estimated nor under-estimated in this general story of progress of the European statistical system:

• **Why the action of ESPON should not be over-estimated**. Looking back at the period 2002-2006, it is clear that ESPON is 90%

dependant of data provided by European statistical institutes – and especially Eurostat- in its global activity. Without a strong partnership with Eurostat and EEA, the ESPON program would not be able to produce operational political results. Not only because of data availability (we could theoretically imagine that ESPON collects data from national statistical offices without using official European system) but mainly because of data legitimacy (if ESPON was using different figures than the ones provided by Eurostat or EEA when available, the political conclusions derived from the work would be considered as non consistent and non useful by the Commission). In practical terms, ESPON is really "compelled" to use the existing official data available at European level. It is only when ESPON can demonstrate that requested data are not officially available at European level that it can engage new data collections. But even in this case, ESPON should be very cautious and, ideally, develop contact with official institutions in order to obtain an "official stamp or label" on the new database that it elaborates.

- **Why the action of ESPON should not be under-estimated**. ESPON has the major advantage of performing applied studies on existing official statistics. This way, it provides some added value at European level and offers a kind of scientific and political quality control. The mistakes or weakness of any database are difficult to identify by simple statistical procedures of control, even if some of these procedures of control can be applied automatically (e.g. control of extreme values, exploration of outliers by data mining techniques, …). In many cases, it is only when experimented practitioners are using the data that problems or discrepancies can be discovered. Normally, Eurostat and EEA should have received many feed-backs from ESPON on a great diversity of topics. If these feed-backs are considered as advices and not as criticisms, they should have contributed to the improvement of the Eurostat and EEA databases, which can benefit to all the end-users, not only ESPON members. The problem is that such a positive loop between ESPON and

official European agencies has probably not been sufficiently developed in ESPON I, due to lack of time and money for the support of networking.

The new period of existence of ESPON (2007-2013) should be an opportunity for enhancing cooperation between ESPON and European statistical agencies but with the awareness that ESPON is only able to present modest proposal in its specific field of territorial planning and is certainly able to solve alone a problem which is much wider. Don't overestimate ESPON capacity and keep humoristic distance… (see Figure 1).

Figure 1 Allegory of ESPON providing rules for data collection.

1.1 Schema

It is important that the future leaders of the ESPON Transnational Project Groups (TPG) be fully aware that they play a crucial role in the elaboration of the ESPON database seen as a collective production. Each research team involved in ESPON can **benefit** from the work of researchers which have previously contributed by preparing new indicators, typologies, etc. In turn, the research teams engage themselves to **develop** the common database by introducing their own production. Espon database is a **collective work** or, better said in german language, a "**kollektiv Werke[2]**".

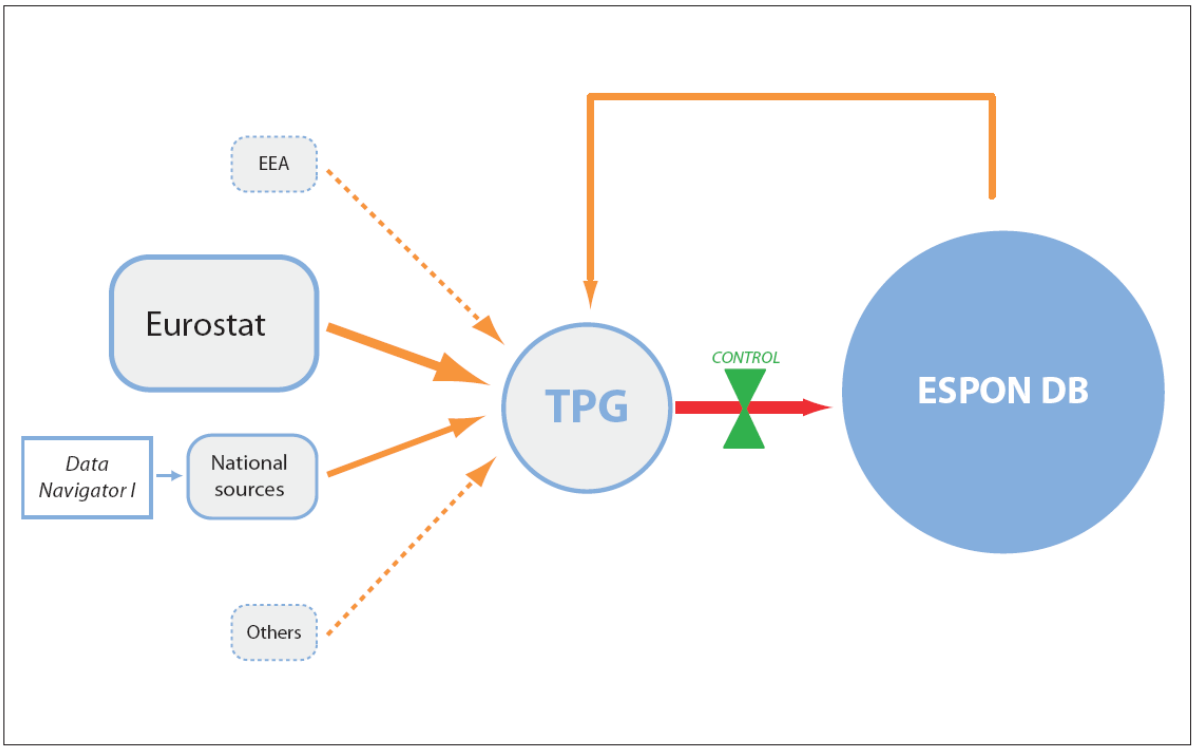


Figure 2 The data circuit in ESPON I (2002-2006)

As shown on Figure 2, during the first period of ESPON, the majority of data used by ESPON came from Eurostat (probably 80 to 90%). National

[2] In English, but more obviously in German, the words work/Werk can be related to (1) the simple production/fabrication of something and (2) the elaboration/creation of something exceptional...

censuses were used primarily for removing holes from the Eurostat database and for countries which are not members of the EU (like Norway, Switzerland) or for candidate countries for which the Eurostat database was not always complete (Romania, Bulgaria). But very few TPG tried to produce datasets based only on national sources, despite the existence of the *Data Navigator I,* a large survey that had been performed on national sources at the beginning of ESPON I. It is true that the obligation to cover the 29 countries of ESPON area made very difficult such building of data from national sources, but in many cases (e.g. case studies, zoom), these national sources revealed to be very precious. More surprising, very few TPG used the datasets from the EEA, except in the case of land use (Corine Land Cover). This limitation of the use of environmental data in ESPON was certainly related to the obligation of producing databases in the geometric framework of administrative units (NUTS2 or NUTS3). The ESPON database could have been based on a grid model and, in this case, the use of EEA data would have been certainly more intensive. Other sources used by some projects are in many cases at the origin of interesting and original productions of ESPON (e.g. the measures of accessibility derived from information on networks of transport infrastructures …). But these original sources were unfortunately often characterised by a high degree of complexity for their production which restricts the reproductibility of such a work.

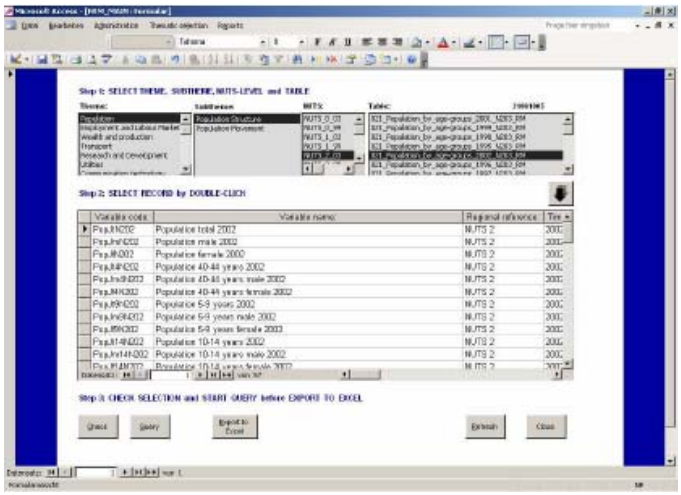
1.2 TPG Inputs

In this section, we describe what are the different possible sources of data collected/used by TPG, including data previously collected by ESPON and reexploited by TPG. We stress on the fact that Eurostat is not the single source of information and explore in more details EEA and National Sources.

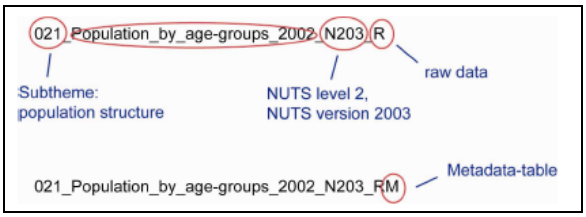
1.2.1 ESPON DB

The EPSON Database is a collection of data, indicators or typologies provided by the TPGs. This joint database includes two kinds of tables: data tables and metadata tables. The ESPON DB is a database stored as a .mdb file (Microsoft Access software is needed).

The structure of the database is oriented towards thematic tables classified in themes (19 thematic categories from Data Navigator I) and subthemes. For accessing data tables, is it possible to use a thematic selection tool, while the result of the query can be exported towards Microsoft Excel software.



The table name gives some information about theme, subtheme and NUTS level, NUTS version and data type (raw data or indicator).



The quality control of the data sets is made by the managers of the ESPON DB before provision to the database.

1.2.2 EUROSTAT

Despite existing missing values or gaps, the EUROSTAT database is an important source of data for socio-demographic themes. The indicators it contains are relatively simple. These indicators cover all the EU members. One can also find indicators for EU state candidates and – not systematically - some neighbour countries (Swiss, Norway).

On the EUROSTAT Web site (http://epp.eurostat.ec.europa.eu/), various socio demographic indicators are available together with their spatial description (see Figure 3)

Principles guiding the REGIO bases:

- Origin of data

The oldest data date from the middle of years 70. In principle, variables are available for each year. The used geographical units are NUTS levels 1 to 3. The version of the NUTS used is normally the last being in use (2003, at present time). Data are provided to EUROSTAT by the national statistical institutes. After a process of verification (and possibly a round-trip between offices and EUROSTAT), data are implemented in the EUROSTAT databases.

- Values update

Most indicators are continuously updated. There is no general update. Dates of data collection (censuses, investigations) by the national offices can be different: this explains partly hiatuses in series. In case of

modification of the NUTS contours, the reconstitution of the statistical sets in the past in the new grid is an explicit work performed by national institutes (for this task may require data for NUTS unit at higher levels). These updates are uncertain.

It is very important to register the date of a data extraction because this date only corresponds to a state of indicators series for a precise update of values and a precise delimitation of territorial units Then, a new extraction of information from EUROSTAT necessitate to store the old value you obtained previously because some work has certainly been done by ESPON TPG on the basis of this old value and they should be present if someone was interested in the verification of previous research.

- Contains

The EUROSTAT database is organized in hierarchical themes:

- economy and finance,
- agriculture and fisheries,
- population and social conditions,…

Most of data available in Eurostat are related to national level and it is only in the theme “General and regional statistics” that statistics can be obtained at regional level. Extraction of data from the EUROSTAT database is possible through the Web site of EUROSTAT. The query is formulated through the choice of units and dates. Without a user registration, extractions are limited to 100000 values. Thus, it is not always easy to get a complete statistical set covering every level of NUTS.

TPG should be aware that many other information on regional statistics are available on Eurostat Web site, especially diffusion of preliminary results (*Eurostat news release*), methodological papers (*Statistics in focus*), compilations (*yearbooks*), etc. Direct contacts with the services of regional statistics of Eurostat are highly recommended for ESPON members.

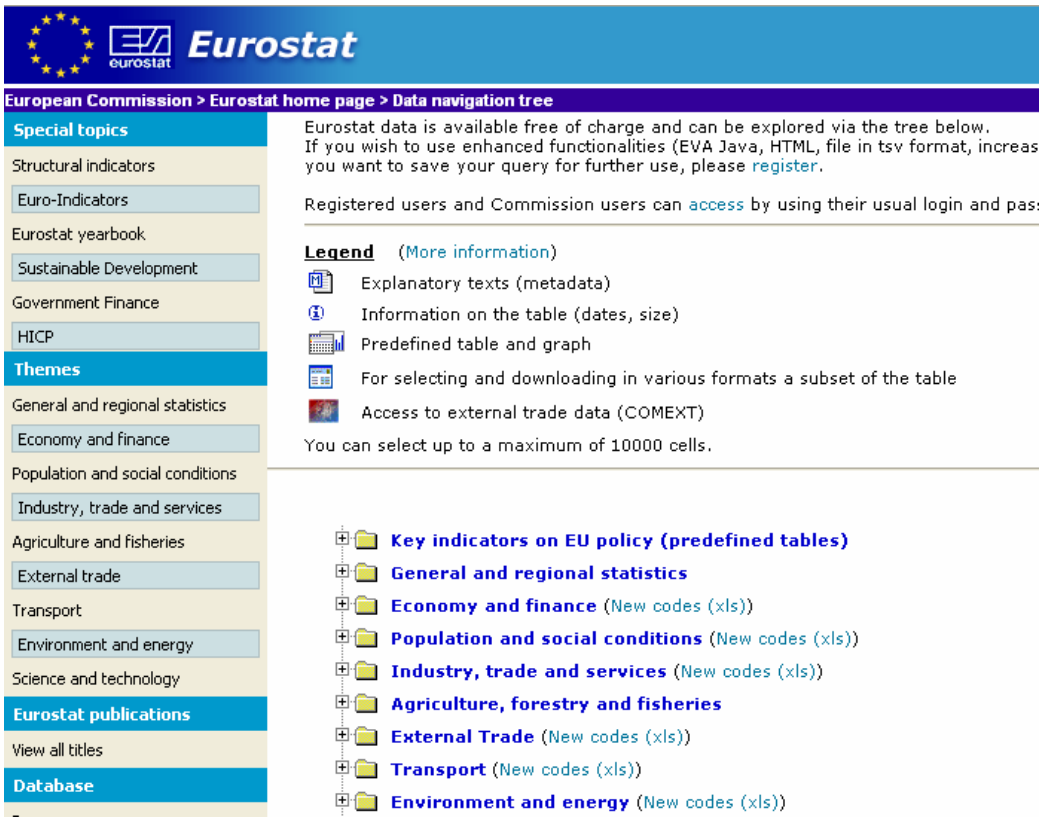
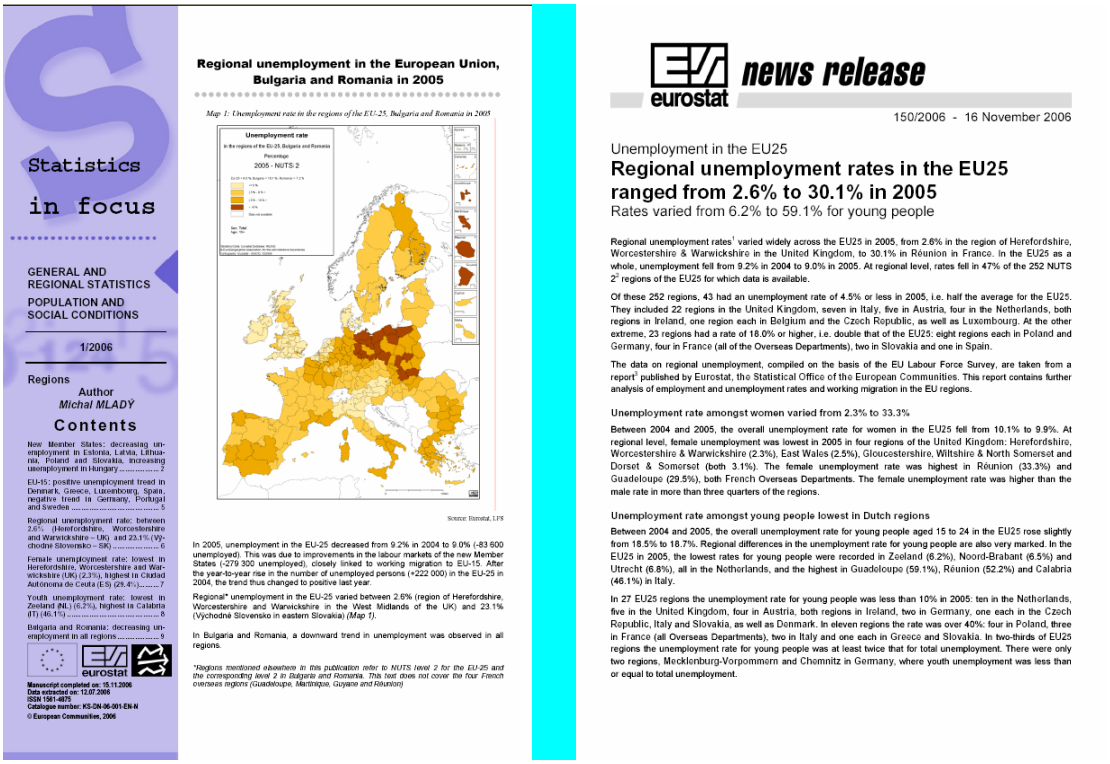


Figure 3 EUROSTAT Website



Statistics in focus

GENERAL AND REGIONAL STATISTICS
POPULATION AND SOCIAL CONDITIONS

1/2006

Regions

Author
Michal MLADÝ

Contents

New Member States: decreasing unemployment in Estonia, Latvia, Lithuania, Poland and Slovakia, increasing unemployment in Hungary 2

EU-15: positive unemployment tread in Denmark, Greece, Luxembourg, Spain; negative tread in Germany, Portugal and Sweden 5

Regional unemployment rate: between 2.6% (Herefordshire, Worcestershire and Warwickshire – UK) and 23.1% (Východné Slovensko – SK) 6

Female unemployment rate: lowest in Herefordshire, Worcestershire and Warwickshire (UK) (2.3%), highest in Ciudad Autónoma de Ceuta (ES) (29.4%) 7

Youth unemployment rate: lowest in Zeeland (NL) (6.2%), highest in Calabria (IT) (46.1%) 8

Bulgaria and Romania: decreasing unemployment in all regions 9

eurostat

Manuscript completed on: 15.11.2006
Data extracted on: 12.07.2006
ISSN 1561-4875
Catalogue number: KS-DN-06-001-EN-N

Regional unemployment in the European Union, Bulgaria and Romania in 2005

Map 1: Unemployment rate in the regions of the EU-25, Bulgaria and Romania in 2005

Source: Eurostat, LFS

In 2005, unemployment in the EU-25 decreased from 9.2% in 2004 to 9.0% (-83 600 unemployed). This was due to improvements in the labour markets of the new Member States (-279 300 unemployed), closely linked to working migration to EU-15. After the year-to-year rise in the number of unemployed persons (+222 000) in the EU-25 in 2004, the trend thus changed to positive last year.

Regional* unemployment in the EU-25 varied between 2.6% (region of Herefordshire, Worcestershire and Warwickshire in the West Midlands of the UK) and 23.1% (Východné Slovensko in eastern Slovakia) *(Map 1)*.

In Bulgaria and Romania, a downward trend in unemployment was observed in all regions.

**Regions mentioned elsewhere in this publication refer to NUTS level 2 for the EU-25 and the corresponding level 2 in Bulgaria and Romania. This text does not cover the four French overseas regions (Guadeloupe, Martinique, Guyane and Réunion)*

eurostat **news release**

150/2006 - 16 November 2006

Unemployment in the EU25

Regional unemployment rates in the EU25 ranged from 2.6% to 30.1% in 2005

Rates varied from 6.2% to 59.1% for young people

Regional unemployment rates[1] varied widely across the EU25 in 2005, from 2.6% in the region of Herefordshire, Worcestershire & Warwickshire in the United Kingdom, to 30.1% in Réunion in France. In the EU25 as a whole, unemployment fell from 9.2% in 2004 to 9.0% in 2005. At regional level, rates fell in 47% of the 252 NUTS 2[2] regions of the EU25 for which data is available.

Of these 252 regions, 43 had an unemployment rate of 4.5% or less in 2005, i.e. half the average for the EU25. They included 22 regions in the United Kingdom, seven in Italy, five in Austria, four in the Netherlands, both regions in Ireland, one region each in Belgium and the Czech Republic, as well as Luxembourg. At the other extreme, 23 regions had a rate of 18.0% or higher, i.e. double that of the EU25: eight regions each in Poland and Germany, four in France (all of the Overseas Departments), two in Slovakia and one in Spain.

The data on regional unemployment, compiled on the basis of the EU Labour Force Survey, are taken from a report[3] published by Eurostat, the Statistical Office of the European Communities. This report contains further analysis of employment and unemployment rates and working migration in the EU regions.

Unemployment rate amongst women varied from 2.3% to 33.3%

Between 2004 and 2005, the overall unemployment rate for women in the EU25 fell from 10.1% to 9.9%. At regional level, female unemployment was lowest in 2005 in four regions of the United Kingdom: Herefordshire, Worcestershire & Warwickshire (2.3%), East Wales (2.5%), Gloucestershire, Wiltshire & North Somerset and Dorset & Somerset (both 3.1%). The female unemployment rate was highest in Réunion (33.3%) and Guadeloupe (29.5%), both French Overseas Departments. The female unemployment rate was higher than the male rate in more than three quarters of the regions.

Unemployment rate amongst young people lowest in Dutch regions

Between 2004 and 2005, the overall unemployment rate for young people aged 15 to 24 in the EU25 rose slightly from 18.5% to 18.7%. Regional differences in the unemployment rate for young people are also very marked. In the EU25 in 2005, the lowest rates for young people were recorded in Zeeland (6.2%), Noord-Brabant (6.5%) and Utrecht (6.8%), all in the Netherlands, and the highest in Guadeloupe (59.1%), Réunion (52.2%) and Calabria (46.1%) in Italy.

In 27 EU25 regions the unemployment rate for young people was less than 10% in 2005: ten in the Netherlands, five in the United Kingdom, four in Austria, both regions in Ireland, two in Germany, one each in the Czech Republic, Italy and Slovakia, as well as Denmark. In eleven regions the rate was over 40%: four in Poland, three in France (all Overseas Departments), two in Italy and one each in Greece and Slovakia. In two-thirds of EU25 regions the unemployment rate for young people was at least twice that for total unemployment. There were only two regions, Mecklenburg-Vorpommern and Chemnitz in Germany, where youth unemployment was less than or equal to total unemployment.

Figure 4 EUROSTAT publications

1.2.3 National Sources

Case study for the use of the national statistical sources: Romania

The first phase of the ESPON program (2002-2006) was a good occasion for testing the European databases (like EUROSTAT), as well as the ones made by the research teams involved in different ESPON projects. The remarkable results obtained have validated this database but also, simultaneously, have shown its limits, among which the most frequent ones are:

- very short periods covered by the time series of statistical data. When this is the case, a long term study of the evolution tendencies of the territorial systems becomes almost impossible;
- frequent gaps in the time series of data. When this is the case, it is often difficult to make a synchronic comparison between administrative units from different states;
- temporal gaps between census in different countries (different census years). This causes problems in the harmonization and the control of national data, even at levels present over the entire European territory (NUTS 2 and NUTS 3);
- unequal administrative subdivisions of the European states, which explains the difficulties in their comparison;
- different levels of administrative organisation from state to state; for example Romania has only 2 official levels (*comuna* – the commune, NUTS 5 and *judeţul* – the county, NUTS 3), to which an informal level is added (*regiunea de dezvoltare* – development region, NUTS 2, see Map 4-A). This fact makes the international analysis of the territorial dynamics impossible at certain levels of administrative organisation.

Partial solutions to those problems consists in using the existing national database. This has two major advantages: a greater variety of scales and a greater diversity of data than in the EUROSTAT database.

1.2.4 EEA

The Land Cover classification often reveals the deep functioning of our societies. This is the reason why a wide aspect of environnement can be approached by considering the Land Cover of the European territory. A very important and precise source of information is provided by the European Environment Agency (EEA). This organisation provides a very precise information for the land cover (available around 1/50000° scales).

The Web site of the European Environment Agency (EEA) provides the most important part of information on Land Cover within Europe. Two major data formats are available:

Vector source

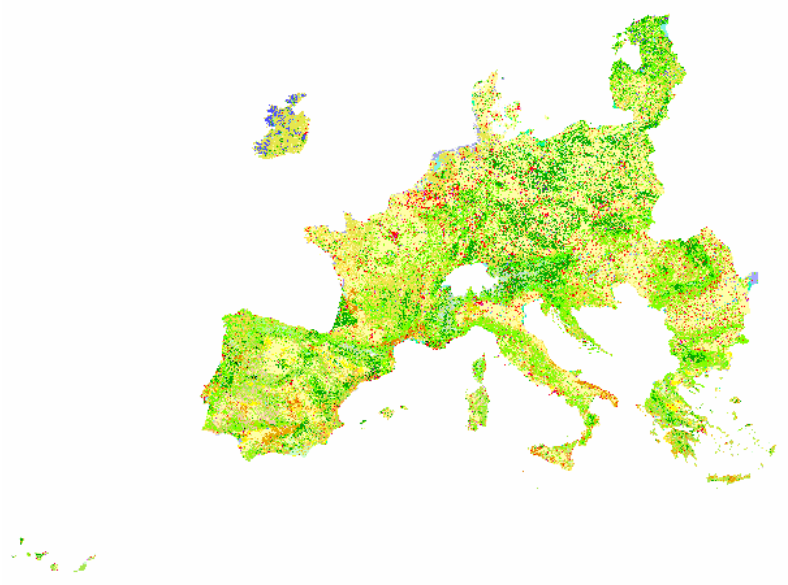
This is the native format used by the EEA. Data of Corine Land Cover 2000 are freely accessible at state levels (except for Germany and Austria). Each acquisition of data is condionned by state autorisation. The land cover is described for each state with a wider buffer zone around its boudaries (few kilometers wider than national boundaries). This buffer is needed in order to check the consistency of transnational data production. The dataset is composed of 1990's and 2000's land cover raw data. In parallel, seamless harmonized data are provided by EEA: those data represent the Europe wide dataset for which a harmonization of land cover polygons has been achieved: same typed polygons, on both sides of a border, are correctly linked.

Raster format

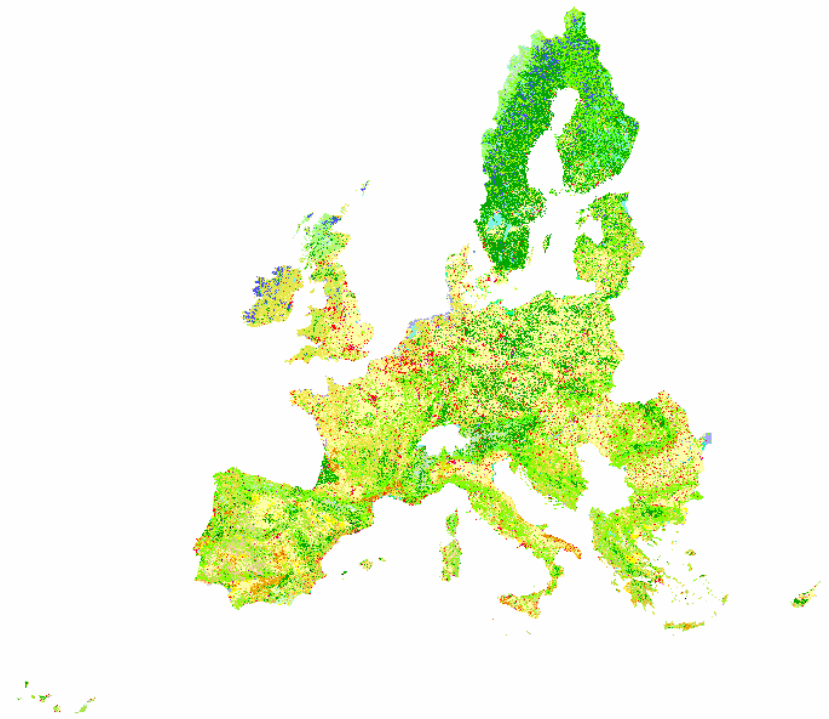
This dataset is derived from the vector data. This spatial continuous description is available for all the Europe boundaries. Two spatial resolutions are delivered:

- 100 * 100 meters cells
- 250*250 meters cells.

The land cover is harmonized both in space and time. The current time serie is composed by 1990's and 2000's raster data. The two grids set have the same spatial references[3], and the same nomenclature. All the grids are perfectly superposable. However, the geographical extension is more narrow for the 1990's cover (see Maps 1 and 2).



Map 1 **Corine Land Cover – 1990 – Spatial extension.**



Map 2 **Corine Land Cover – 2000 – Spatial extension**

[3] ETRS89 Ellipsoidal Coordinate Reference System.

1.3 Outputs

Whatever its source (EUROSTAT, EEA, National sources, etc.), each datum collected by a TPG can then serves to supply the ESPON DB. Such TPG outputs have been systematically **controled** by the BBR[4] before being registered. This control has several dimensions explained in the successive "*Guidance Papers*" of ESPON and in the report delivered by the BBR in project ESPON 3.1 and ESPON 3.2. Two crucial dimensions have been introduced for the control of data:

- **Geometric dimension**: Each TPG had to indicate precisely the geometry of spatial units (generally NUTS3 or NUTS2) together with the year of delimitation of the regional units (1999 or 2003). When a TPG used a different geometry (e.g. project 3.4.1. on Europe in the World), it was asked to deliver the precise definition of the spatial units constituting this geometry and the map files related to them.
- **Statistical dimension**: Each TPG had to provide two Excel files, one with the statistical data and another one with all metadata, giving precise description of the variables, their origin, their eventual modification, their potential use, etc.

The data circuit used in ESPON phase I for supplying the ESPON Database has been sometimes criticized by TPG which have found it too difficult (too much controls…) or too simple (not enough control …) and sometimes by the same people! As a matter of fact, the solution adopted for data storage in ESPON I was pragmatic, efficient and fully adapted to the situation of a program of operational research whose main expected outputs were synthetic indicators elaborated considering official regions.

[4] The BBR was the team in charge of the organisation and diffusion of ESPON data during the period 2002-2006.

As a matter of fact, the ESPON Data Base should, in a close future, be able to cope with several issues:

- Territorial units (identified by the NUTS nomenclature) for which statistical indicators are provided by TPG are heterogeneous and evolve in space and time. A track of this possible evolution in space and time should be kept;
- Datasets or indicators provided by the TPG are sometimes uncomplete and therefore should be completed as much as possible by means of reliable estimation methods;
- Datasets or indicators provided by the TPG may come from heterogeneous sources of information (statistical institutes) and may require some uniformisation or, at least, some quality control.

In its current form, the ESPON Data Base can *not* fulfill all these requirements. A restructuration is needed. Challenges addressed here deal with data consistency, data completeness, data evolutivity, data comparativity, as well as data query formulation.
This restructuration aims at better fit users needs whether they be policy makers or scientists.

2 Rules for data collection

In this section, we detail the good practices for data collection according to our experience and the recommendation of national and international statistical institutes like for example Eurostat (see box 3)

EUROSTAT defines the quality of statistics with reference to seven criteria:

1. **Relevance:** an inquiry is relevant if it meets users' needs. The identification of users and their expectations is therefore necessary.
2. **Accuracy**: accuracy is defined as the closeness between the estimated value and the (unknown) true value.
3. **Timeliness and punctuality in disseminating results**: most users want up-to-date data which are published frequently and on time at pre-established dates.
4. **Accessibility and clarity of the information**: statistical data are more valuable when they can be easily accessed by users, when they are available in standard formats, and when they are adequately documented.
5. **Comparability**: statistics for a given characteristic have the greatest usefulness when they enable reliable comparisons of values through space and time. The comparability component emphasizes the comparison of the same statistics between countries in order to evaluate the meaning of aggregated statistics at the European level.
6. **Coherence**: when originating from a single source, statistics are coherent in that elementary concepts can be combined reliably in more complex ways. When originating from different sources, and in particular from statistical surveys of different frequencies, statistics are coherent as long as they rely on common definitions, classifications and methodological standards.
7. **Completeness:** domains for which statistics are available should reflect the needs and priorities expressed by the users of the European Statistical System.

Source : http://forum.europa.eu.int/irc/dsis/coded/info/data/coded/en/gl011043.htm

Box 3. Definition of quality of statistics by Eurostat

We firstly present some general rules to be followed in every case (2.1). Then, we propose a new data model, called Long Term DataBase, compatible with a process that would follow our rules for data collection

(2.2). This model also serves as a support for estimation of missing values (2.3) and for quality control procedures as well (2.4). The model, together with the definition of estimation and controle procedures, forms the basis for a new framework increasing the capabilities of ESPON data base in terms of data acquisition, harmonization, completion and control.

2.1 The 10 Commandments for data collection

Previous experiments led by many European research projects have clearly shown that the **quality** of the database is much more important than the **quantity** of information. It has been proved many times at European level that a small database perfectly integrated is much more useful and efficient than a wide database which is a simple compilation of heterogeneous indicators built without applying any rules of quality control.

2.1.1 The 10 commandments

Derived from the practice of ESPON 2002-2006, we can propose 10 “commandments” which define the general rules to be followed in order to build an objective, efficient, harmonized and evolutive database.

AN OBJECTIVE DATABASE

Commandment 1: “You shall always use the most primary sources of information”

Commandment 2 : “If you can not use primary sources, you shall indicate precisely the path leading to the data collection”

The precise identification of the **initial sources** of information (census data, survey…) is an absolute necessity for the quality control of information. In any case, it should be possible to identify **the path** of elaboration of any figure of the ESPON database, from tertiary or secondary sources to primary (initial) sources of information. Each transformation or modification of primary sources should be clearly identified and registered when secondary sources are introduced in the ESPON database.

For instance, most data used by ESPON are secondary sources, because they were download on Eurostat website but are derived from initial sources which are national census elaborated in each member state. In this case, a precise identification of sources would imply the storage of the references to each national statistical institute in charge of the production of census data and the precise time of these censuses. But it is not necessary because this work of documentation has most of the time been perfectly done by EUROSTAT. Then, in this very precise case, it is sufficient to indicate the reference of EUROSTAT where all those precisions are available. But in other cases, databases are tertiary or quaternary sources.

For example, some data about European regions can be download on the website of the French Observatoire des Territoires[5] but these data are derived from Eurostat data which are derived from national sources … and are therefore tertiary sources. This data are correctly documented but should not be used by a TPG which should use original sources.

In some cases, data are not documented at all, as the famous data of states of the world proposed by CIA under the name of "*The World*

[5] http://www.territoires.gouv.fr/

Factbook[6]". This publication of the american government proposes an extraordinary database with 200 to 300 indicators describing all states of the world (at least, the one which are recognised as such by United States …). But the origin of data are never precisely described and only administration of the United State is mentioned as data provider. The real producers from first level (states) or second level (United Nations) are not indicated. It means that the user of this data has no possibility to check and verify the figures: either he believes what is said by the US Government, either he does not. This is a typical violation of our 1st and 2nd commandment (see Box 4).

The World Factbook is prepared by the Central Intelligence Agency for the use of US Government officials, and the style, format, coverage, and content are designed to meet their specific requirements. Information is provided by Antarctic Information Program (National Science Foundation), Armed Forces Medical Intelligence Center (Department of Defense), Bureau of the Census (Department of Commerce), Bureau of Labor Statistics (Department of Labor), Central Intelligence Agency, Council of Managers of National Antarctic Programs, Defense Intelligence Agency (Department of Defense), Department of Energy, Department of State, Fish and Wildlife Service (Department of the Interior), Maritime Administration (Department of Transportation), National Geospatial-Intelligence Agency (Department of Defense), Naval Facilities Engineering Command (Department of Defense), Office of Insular Affairs (Department of the Interior), Office of Naval Intelligence (Department of Defense), US Board on Geographic Names (Department of the Interior), US Transportation Command (Department of Defense), Oil & Gas Journal, and other public and private sources

Source : https://www.cia.gov/cia/publications/factbook/docs/contributor_copyright.html

Box 4. Example of criticable sources : the CIA World Factbook

[6] https://www.cia.gov/cia/publications/factbook/index.html

AN EFFICIENT DATABASE

Commandment 3: "You shall always collect the raw count variables rather than ratio or other indexes derived from their combination"

Commandment 4: "When you are not able to provide raw count variables, you shall indicate the weight to be used for aggregation".

In many cases, the indicators used for territorial planning are a mathematical combination (addition, substraction, division, product…) of raw count variables which are not directly useful but are, in practice, the kernel information from which all indicators are directly or indirectly derived. A good database structure should absolutely store those kernel variables (real information) and not necessarily store the derived indicators (virtual information) which can be automatically computed when request.

In a short term perspective, this principle may seem heavy to apply. For instance, if one want to use a variable like the median age of population in NUTS3 regions, one has (1) to store all the age structure of those NUTS 3 regions and (2) to store the formula of median age computation in a SGBD. Apparently, this solution is time and human resources consuming, but, eventually it produces a very important gain of time, ressources and quality in a long term perspective because:

- The spatial aggregation or disaggregation of data is then much easier. In the case of the median age, it is impossible to estimate the values at NUTS 2 level if we have stored only the median age at NUTS 3 level, even if the result is weighted by population (it would be possible with mean age of population, but not with median). But it is very easy to

aggregate all age structure (which are count variables) from NUTS 3 to NUTS 2 and then to apply the formula of median age computation which has been stored for NUTS 3 and remain available at NUTS 2.

- Many indicators are derived from the same kernel variables: which means that with a limited number of good kernel variables, it is possible to produce a very wide set (probably infinite) of indicators and derivated variables. Storing kernel variables can favour the production of new indicators which will not be possible if this kernel information had not been stored. Imagine for example that a TPG has produced the indicators Z1=(A/B) and another TPG the indicator Z2 = (C-D)/E, the strategy of kernel indicators (storage of A,B,C,D,E) make possible the construction of many other indexes like (C-D)/B or A/E which would not have been possible if we had only stored the indexes Z1 and Z2.

- Statistical tests are generally not available or biased if the kernel information is not available in the database structure: in the very simple example of GDP/inh., a good statistical evaluation of heterogeneity can not be made by a simple comparison of regional ratios but implies a direct examination of the unequal repartition of the raw count values of population and wealth. Generally speaking, the use of ratio is very dangerous in statistical analysis because results are non weighted and subject to random variations in small areas. Keeping the initial count variables from which ratio are derived is a necessary condition for a correction of those biases.

A HARMONIZED DATABASE

Commandment 5: "You shall always explain precisely your procedure for time harmonization".

Commandment 6: "You shall always explain precisely your procedure for territorial harmonization".

Commandment 7: "You shall always explain precisely your procedure for thematic harmonization".

Time harmonization

In the European context, especially if we take into account the enlargement of databases from 15 to 27 countries, it is not possible to use primary sources without modifications and harmonization. An obvious example is related to the census year and date which are different in most European states. Thus, if we want to evaluate the regional distribution of population at 1st January 1990 or 2000 for all European regions, we will be necessarily obliged to introduce estimations for many states which have a different census time. Those estimations are not a problem as far as the estimation procedure is clearly indicated in the database structure.

An ideal situation would be the storage of data and formula used in order to produce harmonized information derived from kernel information. For instance, if we want to evaluate the population of France regions in 1980 starting from the census variables of 1975 and 1982, we should store (1) the regional populations of France in 1975 and 1982 (primary kernel information) and (2) the precise formula used for the estimation (linear, exponential, …) of population in 1980 from population in 1975 and 1982.

It is necessary to keep in mind that those rules applies only to raw count variables and are not necessary for indicators which are derivated from

combination of raw count variables variables. If we take the example of a regional index of median age of population in France in 1980, the database should indicate that it is the result of a formula applied to age structure in 1980, which is derived from a formula of interpolation derived from the age structure in 1975 and 1982 …

Territorial harmonization

Another crucial problem is related to the harmonization of European territorial divisions, which is not only a technical problem but also a political problem, as far as the attribution of structural funds is related to particular values (or thresholds) established at a particular level of territorial division. Practices of “gerrymandering” or simple evolution of regional divisions are responsible for a very difficult problem in the establishment of long term time-series at regional level. Users of regional statisticis in Europe unfortunately have to face with gaps, discontinuities and breakdowns in regional time series, which leads to a very pernicious lack of efficiency in the production of study on trends at European level. It has been many times underlined (especially in ESPON project 3.2) that it is impossible to make good territorial prospective (prospective trends) if it is impossible to compare present situations to past ones (retrospective trends)

It is thus necessary to distinguish between “official maps at a given time t” which are required to use the official delimitation of the period t, and “trend maps on a long period [t1, tn]” which can not use official delimitations because of regular changes in the official delineation of political and administrative boundaries. If the structure of the ESPON database is not able to take into account the evolution of territorial division, then one has to face the dilemma presented in Figure 5: either a great diversity of indicators with high spatial resolution, but only for a

very short period of time, is available, or longer time series, but only for a limited number of indicators and with low spatial resolution are present.

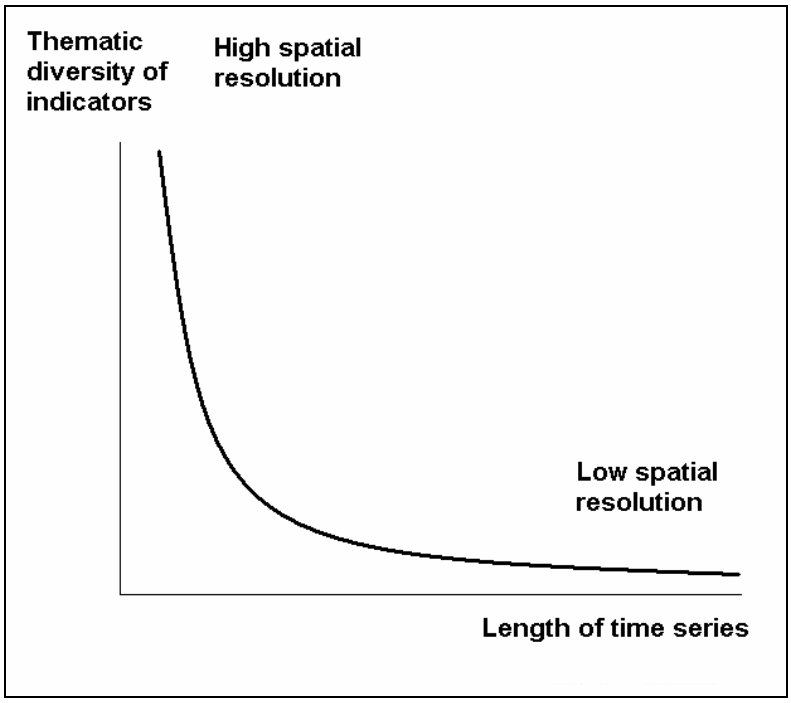


Figure 5 The dilemma of time vs thematic harmonization

The structure of the ESPON database used in the period 2002-2006 was typically subject to this dilemma. And the choice was made to have a high spatial resolution (NUTS 3) and a great diversity of indicators but with very short length of time series (generally less than 5 years). An alternative structure of database can be proposed for ESPON 2013 called "Long Term Data Base" which can allow to avoid the dilemma and to produce at the same time long term time series for a limited number of indicators.

Thematic harmonization

In addition to the problem related to temporal and spatial harmonization of European databases, it is necessary to focus also on the problems linked to the harmonization of definitions used in European states.

Possible biases may be introduced by differences in the definition of variables and/or collection of information procedures.

A classical example of those biases is related to the measure of the infant mortality rate. At first sight, this index is based on a very precise definition (ratio between death between 0-1 year and number of birth) but in fact many problems of harmonization have been revealed by demographers. In certain states, for instance, children who die in their 1st day of life were not recognised as “birth” (and thus, neither as “dead between 0 and 1”). Accordingly, they were classified as “birth dead” and not taken into account in the computation of the infant mortality rate, which introduces an important reduction of the ratio, when compared to other states.

In the case of ESPON, one has to be very careful about those possible biases and it is important to store in the database the possible biases related to a lack of harmonization in the definition, as well as statistical system practices when those biases are established by experts. The researchers of the different TPG will probably indicate those biases in their reports and an important work is to store those experts advises in the ESPON database. For many crucial subjects (unemployement, R & D, accessibility, …) this precise criticism of indexes is, in a sense, more important than the value of indexes stored in the database and represent the real added value of the ESPON Program. We have illustrated this point through several examples in Chapter 3.1.

AN EVOLUTIVE DATABASE

Commandment 8: "You shall propose indicators that can be further collected by others than yourself".

Commandment 9: "You shall propose indicators that can be further update in the future"

Commandment 10: "You shall propose indicators that can be further collected in new countries"

According to the previous principles, we intend to propose the design and development of an open and evolutive database which will be probably limited in a first step, but can be further developped geographically, historically and thematically in order to produce a cumulative knowledge base on European Spatial Developement.

An open database means that every researcher involved in the ESPON Program will be able to contribute to its development (as a data provider or as an expert) and that, conversely, all researchers involved in the ESPON Program will have the right to use this database for the purpose of the program. This interactive open access of all ESPON research members to the ESPON database may be technically complicated (problems of security). However, to our opinion, it is above all a guarantee of quality of the results since all indexes involved in the ESPON database will be subject to the collective evaluation by a community of more than 200 researchers of all the European Union and accessing or neighbouring countries.

Reseach teams working for the ESPON program should therefore avoid the use of "private" databases which are their own property and that they are not ready to deliver to the community of other ESPON researchers.

Moreover, they should always contact the ESPON coordination unit whenever they discover a database of particular interest that could be integrated into the ESPON database.

To illustrate, when the TPG ESPON 3.4.1. Europe in the World has decided to use an harmonized historical database on population and GDP of states of the World published by OECD and achieved by the independent researcher Angus Madison, the decision has been taken to negociate the use of this database for all ESPON members and not only for the lead partner or research teams of the ESPON project 3.4.1. As a result, future TPG of ESPON 2013 which will produce research on Europe in the World or historical evolution of European economy will benefit from the integration of this data in ESPON database.

A more difficult question is related to indicators which are derived from mathematical model or specific databases belonging to research units. A good example is provided by the indicators of accessibility produced by ESPON project 1.2.1 *Transport*. What is at stake here is how such indicators could be updated in the future if there are some evolution in the transportation network, or if the regional delimitations changes, or if the territorial scope of ESPON is enlarged (inclusion of Turkey or Balkanic countries), etc. It is obvious that what should be stored in the ESPON database is not only the regional scores of accessibility but also the transportation network and the program/algorithm used for the computation of accessibility.

In ESPON 2013, the focus should not be put on the **quantity** of indicator produced by TPG's but rather on their **quality** which includes the possibility to update values regularly in the past and the future, and to enlarge geographically their coverage when needed.

2.1.2 Towards a "10 commandments compatible" Data Model

The 10 commandments proposed above make no sense without a Data Model making their application possible. The long-term storage of thematic and geometric data for territorial units of the European area, at different resolution levels (ranging from the state level (NUTS 0) to the local authorities level (NUTS 5)) implies tackling several issues:

- *Evolutivity issues* require a flexible schema, so that new data (or data types) can be easily added up. This will allow experts to use the model, and consequently the database, for an extended period of time and to keep it up to date, without having to build a new database for each new application.
- *Data quality issues*: the data model has to keep track of the quality of the data it contains (data validation by statistics institutes, genealogy of data sets, automatic detection of data inconsistencies...)
- *Usability issues:* the data model has to be usable as a shared resource. It should help users to easily understand which data are available (by using thematic and geographic ontologies), while semi-automated data acquisition mechanisms will make it possible to easily update the data while preserving the overall coherence of the structure.

What is at stake for ESPON II is therefore not to propose a complete revolution of previous practices of data collection and storage but rather to imagine a solution which combines pragmatism and ambition in a realistic framework. What we basically suggest (Figure 6) is to develop two databases in an integrated way: the Territorial Indicator Data Base (TIDB) and the Long Term Data Base (LTDB).

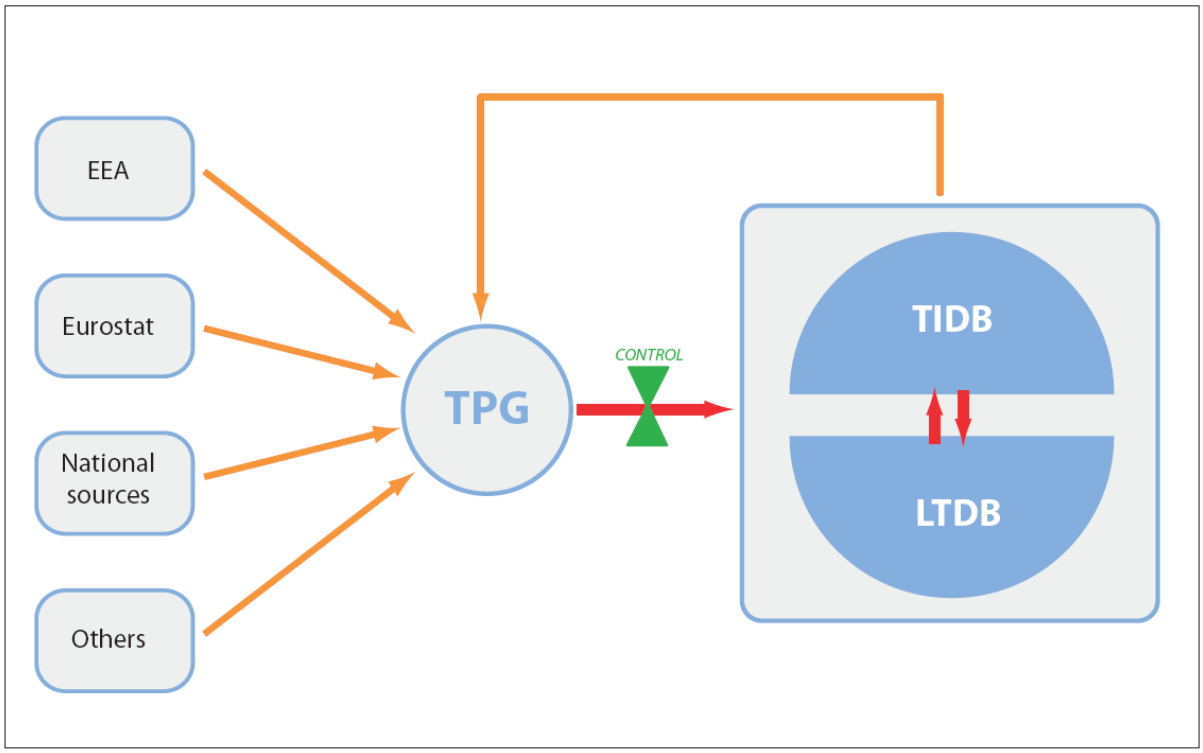


Figure 6 Proposals for data circuit in ESPON II (2007-2013)

- **The Territorial Indicator Data Base (TIDB)** will be a political oriented database which has the mission to deliver standard information to decision and policy makers in the field of spatial planning and territorial cohesion. These indicators are generally used for the **monitoring of regions** and they should be specifically organized according to general political objectives (competitiveness, cohesion, sustainability…) or sectorial political questions (agriculture, transport …). This database should be very easy to query and use and is in a sense the visible part of the ESPON data iceberg.

- **The Long Term Data Base (LTDB)** will be a scientific oriented database which has the mission to store and improve all data of interest for the ESPON program and to fulfil specific tasks like quality control, estimation of missing values, harmonization of time series… It is in a sense the "back office" of the TIDB, the part of the iceberg of data which is not necessarily visible from a political point of view but which is crucial for the sustainability of the ESPON program.

The ESPON database which was elaborated in ESPON I (2002-2006) was clearly something in between with properties of both TIDB and LTDB. What we propose is to have a clearer distinction between these two objectives in ESPON II. We here focus on the LTDB.

2.2 The LONG-TERM DATABASE

In this section, we present the data model we propose for the LTDB. We then focus on the 4 dimension-framework, called ESTI, which allows a complete charaterization of data considering Space (E), Source(S), Time (T) and Thematic (I). We also show how the model, together with the ESTI framework, is a good candidate to support both estimation method for missing values and quality control procedures.

2.2.1 The LTDB DataModel

In the project ESPON 3.2, we have proposed a data model (see Figure 7) which aims at supporting a long-term storage of spatial and thematic data concerning geographic units. This data model copes with any change in the structure of geographical units can: their name, their spatial representation, their position in the administrative hierarchy, their codename and their thematic part (indicators).
The main components of the data schema are:

- *Geographic Unit* (GU), this central element in the schema describes the identity of the geographic units (an internal identifier to be generated by the system), as well as its lifespan, because the existence of geographic units may be limited in time (e.g. Western and Eastern Germany).
- *GU Spatial Representation* describes the geometrical shape corresponding to the footprint of the geographic unit (attribute geom

Obj). Since it can change in time, it is associated with a validity interval.

- *GU Name* describes the official name of the geographic unit (attribute Name Value). Since this name can change in time, it also has a validity interval.
- *Code nomenclature* describes the code of the geographic unit. This code depends on a nomenclature (e.g. the NUTS 2000 territorial unit code system) and it is stored because for most databases the codes of the territorial units are used as an identifier rather than their names.
- *Nomenclature* represents the different code systems that are used to identify territorial units in statistical databases. Nomenclatures have a validity interval and are produced by well-known organisms of statistics like Eurostat.
- *Composition* represents the composition relations between geographic units, related to a certain hierarchy. The Level defines the level of the composing geographic units in the respective hierarchy.
- *Hierarchy* defines the set of composition relations existing between the geographic units on a certain territory for a given period of time. It can change in time and has a validity interval. Its number of levels is also stored.
- The *Similarity* association defines different kinds of similarity/proximity measure that might be needed for a more advanced phase of the application.
- The *Evolved from* association allows storing genealogy relations between geographic units, necessary for some estimation methods.
- *Indicator* defines a thematic part of the geographic units.
- The *Value* association allows storing the value of a certain indicator for a certain geographic unit and for a certain period (attribute validity interval). A track is kept of the production moment and of other metadata necessary for assessing the quality of the data.
- The *Source* class allows keeping track of the genealogy of the data stored in the database. The attribute Code source defines the code of

the database from which the data were retrieved or of the process by which the data were produced.

- *Provider* represents the statistical organisms from which the data were retrieved.
- *Validation* allows storing information about the confirmation of the quality of the data issued by trustworthy statistics organisms.

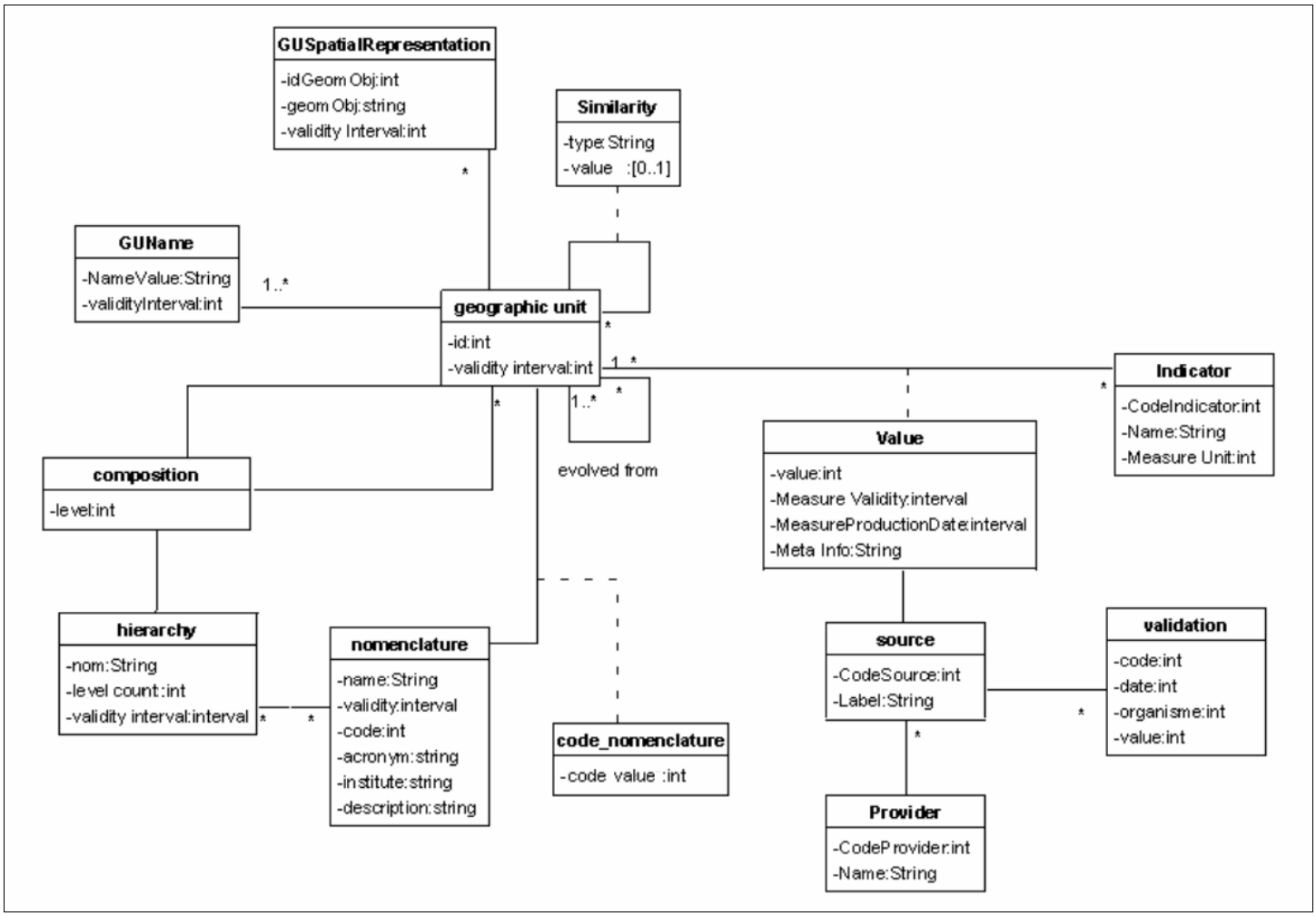


Figure 7 The data model of the future LTDB.

2.2.2 The ESTI framework

Data in the LTDB can therefore be characterized according to four dimensions E, S, T and I where:

(E) represents the **spatial dimension** and refers to one or several territorial units;

(S) represents the **source dimension** and refers to one or several organism of statistics;

(T) represents the **time** dimension and refers to one or several instants or/and periods

(I) represents the **thematic** dimension and refers to one or several indicators.

For instance, the value 83859 could be the result of a query concerning the dimension I and where:

E='Austria'

S='EUROSTAT'

T='1999'

I='Area in km^2'.

Although the query is answered a lot of implicit assumptions are made here:

- Regarding the E dimension, the query is supposed to refer to the territorial unit whose shape corresponds to the "official" delimitation of Austria.
- Regarding the T dimension, the value returned corresponds to the value provided by the 'EUROSTAT' for the period ranging from 01/01/99 to 31/12/99.
- Regarding the I dimension, areas of lakes and rivers in Austria are supposed to be included…

In the previous example, a value is returned for the corresponding value of E, S, T and I, namely the tuple ('Austria', 'EUROSTAT', '1999', 'Area in km^2'). It should be noted that values of E, S, T or I, or a combination of

them might be missing when formulating the query. By default, such a missing value in one or more dimensions handled by the query, should be considered equivalent to a wildcard operator "*" meaning "all values".
However, unavoidable incompleteness in the LTDB will lead to unanswered queries. This occurs when, no value is present in the LTDB, for a given tuple of values (included wildcard "*"). The objective is to overcome this case of missing value by proposing one or several estimation methods in order to compute the more probable (although not measured or sure) value and return it as an answer to the query. LTDB users should be warned when the returned value is a computed estimated value replacing a missing one.

Then, information in the LTDB (that is the future database implanting the data model presented in Figure 7) can be represented as a four-dimensional hypercube (the ESTI framework) with holes corresponding to missing values. Estimation methods help in filling-up these holes by considering information provided by the neighborhood of these holes. We present the main principles of the estimation methods which could be implemented as procedures together with the future LTDB. Then we describe why and how the model could moreover allow the elaboration of quality control procedures.

2.3 The ESTI framework for estimating missing values

Below is described a formalization of the principles of some of the estimation methods which could be implemented in the LTDB.

2.3.1 Notations

Let X(e,s,t,i) be a value, possibly missing (not known or not defined) in the hypercube where e describes one territorial unit, s describes one source, t describes one instant, and i describes one indicator.
For the sake of simplicity, only this elementary case is considered here. However, it can be shown that more complex cases, where e,s,t, and i are each described by set of values, can be decomposed in elementary ones.
For a missing value X(e,s,t,i), estimation methods exploit known values X(e',s',t',i') in the neighbourhood of X(e,s,t,i). This neighbourhood can be defined according to one or more dimension among E, S, T and I. Some one and two-dimensional estimation methods are described.
Let child be the operator which, for a spatial unit e, returns the set of all the spatial units e1,e2,…,en which are spatially included in e, at the first level of spatial inclusion.
Let parent be the operator which, for a spatial unit e, returns the spatial units p which includes e, at the first level of spatial inclusion.

2.3.2 One-dimensional estimation methods

a) Estimation methods based on the spatial dimension (E)

In this case, for a missing value X(e,s,t,i), s,t,and i being fixed, the idea is to use spatial units e' being at an upper, lower or same hierarchical spatial level as e, and for which X(e',s,t,i) is known, in order to obtain information on the missing value.

Different cases are to be considered:

Case 1) if X(parent(e),s,t,i) is known and if X(e′,s,t,i) is known for all e′ so that parent(e′) = parent(e) and e′ ≠ e (e′ is at the same hierarchical level as e and has the same parent as e), then

$$X(e,s,t,i) = X(parent(e),s,t,i) - \sum_{e'} X(e',s,t,i)$$

Case 2) if X(parent(e),s,t,i) is known and if there exists at least one e' so that parent(e′) = parent(e) and e′ ≠ e (e′ is at the same hierarchical level as e and has the same parent as e), and X(e′,s,t,i) is not defined then three methods can be used.

Sub-case 2.1) Min-max

$$X(e,s,t,i) = X(parent(e),s,t,i) - \left(\frac{\sum_{e'' \in child(e)} X(e'',s,t,i) + X(parent(e),s,t,i) - \sum_{e'} X(e',s,t,i)}{2} \right)$$

Sub-case 2.2) Average of children (where e′′ is so that e = parent(e′′) and X(e′′,s,t,i) is defined)

$$X(e,s,t,i) = \frac{|child(e)|}{\left| \bigcup_{e''} \{e''\} \right|} \sum_{e''} X(e'',s,t,i)$$

Sub-case 2.3) Average of the values of spatial units of the same hierarchical level (where e′ so that parent(e′) = parent(e) and e′ ≠ e and X(e′,s,t,i) is defined),

$$X(e,s,t,i) = \frac{\left(X(parent(e),s,t,i) - \sum_{e'} X(e',s,t,i) \right)}{|child(parent(e))| - \left| \bigcup_{e'} \{e'\} \right|}$$

b) Estimation methods based on the source dimension (S)

Alternative sources of information can be used when the main source does not provide the target information. The idea here is to replace the missing value X(e,s,t,i) by a known value X(e,s',t,i) where s' is another organism of statistics. X(e,s,t,i) = α X(e,s',t,i) where α is a correlation factor empirically fixed.

c) Estimation methods based on the temporal dimension (T)

Various time interpolation methods using linear or non-linear assumption, prospective or retrospective computations of tendency can be used. Three of them are described here. The idea here is to estimate the missing value X(e,s,t,i) by using two known values X(e,s,t1,i) and X(e,s,t2,i).

Case 1) Interpolation 1-1 method (where t1 < t < t2)
This method uses the two closest neighbours placed in time before and after t.

$$X(e,s,t,i) = X(e,s,t_1,i) + \frac{X(e,s,t_2,i) - X(e,s,t_1,i)}{t_2 - t_1}(t - t_1)$$

Case 2) Retrospective 2 method (where t < t1 < t2)
This method uses the two closest neighbours placed in time after t.

$$X(e,s,t,i) = X(e,s,t_1,i) - \frac{X(e,s,t_2,i) - X(e,s,t_1,i)}{t_2 - t_1}(t_1 - t)$$

Case 3) Prospective 2 method (where t1 <t2 < t)
This method uses the two closest neighbours placed in time before t.

$$X(e,s,t,i) = X(e,s,t_2,i) + \frac{X(e,s,t_2,i) - X(e,s,t_1,i)}{t_2 - t_1}(t - t_2)$$

d) Estimation methods based on the thematic dimension (I)

Alternative indicator(s) can be a value is missing for the given indicator. The idea here is to replace the missing value X(e,s,t,i) by a known value X(e,s,t,i') where i' is another indicator. X(e,s,t,i) = α X(e,s,t,i') where α is a correlation factor empirically fixed.

2.3.3 Multi-dimensional estimation methods

Multi-dimensional estimation methods are result from the combination of two or more one-dimensional methods. Generally they are more accurate and capitalize on more information

Example: estimation method (ET)

This method is based on a combination of a spatial estimation method (E) with a temporal estimation method (T)
Let us suppose that the value X(e,s,t,i) is not known while the value X(parent(e),s,t,i) is known, as well as are known the values X(e,s,t1,i) and X(e,s,t2,I) of the two closest neighbours placed in time before and after t (t1 < t < t2).

We compose a spatial estimation method

$$X(e,s,t,i) = X(parent(e),s,t,i) \times Freq(e,s,t,i)$$

where

$$Freq(e,s,t,i) = X(e,s,t,i) \Big/ X(parent(e),s,t,i)$$

Yet, X(e,s,t,i) is not known for computing Freq(e,s,t,i) but X(e,s,t,i) can be at its turn estimated using a temporal estimation method (interpolation 1-1)

$$Freq(e,t,i) = Freq(e,t1,i) + \frac{Freq(e,t2,i) - Freq(e,t1,i)}{t2 - t1}(t - t1)$$

2.4 Empirical methods for quality control

The main interest of the E.S.T.I. framework is to make very clear the different solutions for (a) estimation of missing values and (b) quality control of data. Both problems are indeed very similar because quality control is based on the research of "exceptional" values which are non consistent with the neighbouring values in the 4 dimensions of E.S.T.I.. We provide here a set of simple methods of quality control illustrated by typical examples:

2.4.1 Method of quality control based on spac(E) dimension

In most cases, data can be collected at different levels of spatial aggregation and a simple way to validate the data is to verify if the **whole is equal to the sum of the parts**.

Example : verification of GDP of Spanish region

If we want to obtain the GDP of the regions of Spain (NUTS1), we will also download the GDP of the whole Spain (NUTS0) in order to check that the sum of regional GDP is equal to the national value. The small differences observed in 1996 and 1999 are not a problem because they are only the result of the normal margin of error due to rounded values.

Gross domestic product (GDP) at current market prices at NUTS level 2
Source EUROSTAT / Date of extraction: Thu, 11 Jan 07 11:18:18
mio_pps Millions of Purchasing Power Parities

		1995	1996	1997	1998	1999	2000
es	**Spain**	**534425.2**	**563539.1**	**593999.8**	**633978.5**	**691476.6**	**746116.9**
es1	Noroeste	49630.4	51872.3	54019.1	57612.2	62074.7	64429.1
es2	Noreste	64456.6	67710.1	71564.2	76250.5	82954	88917.4
es3	Comunidad de Madrid	89843.5	94538.7	100200.7	108892.7	118949.8	131758.4
es4	Centro (ES)	60709.5	63790.2	66146.2	70082.4	75890.6	79034.2
es5	Este	163978.3	173803.1	183573.9	194889.2	213116.3	232486
es6	Sur	85489.7	90336.6	95641.4	101400.1	110447.2	119500.5
es7	Canarias (ES)	20317.2	21488.2	22854.3	24851.4	28044.1	29991.3
Sum of (es1..es7)		**534425.2**	**563539.2**	**593999.8**	**633978.5**	**691476.7**	**746116.9**
Difference		0	0.1	0	0	0.1	0

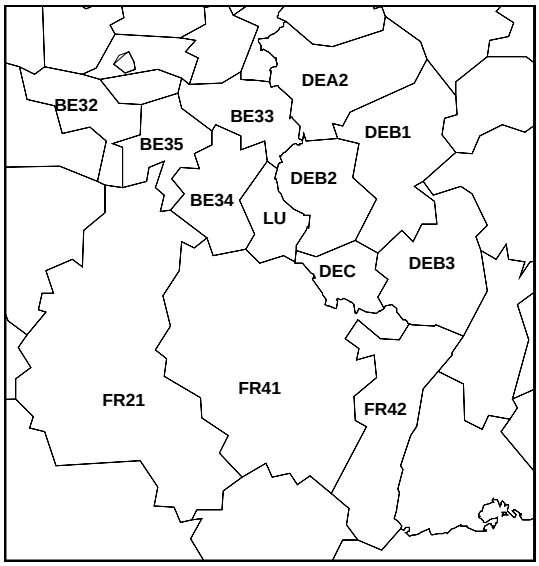
This method is very simple but also very important because there are many empirical situations where the whole is not equal to the sum of parts from geographical point of view. For example, the GDP of Norway is sometimes not equal to the sum of regional GDP because one part of the income of oil exploitation is not allocated to immediate use but reserved for future generations and is therefore not allocated to any inhabited region. We have therefore to add a specific statistical unit which is not geographical in the usual sense ("Oil of North Sea") and which is ambiguous from a time point of view (related to present or future?).

In most cases, the spatial distribution of phenomena is characterised by positive **spatial autocorrelation** i.e. by the fact that neighbouring regions are generally more similar than distant regions. It is therefore interesting to **produce maps of discontinuities or to compute local index of spatial autocorrelation (LISA) in order to identify exceptional differences at local level.** The interest of this method of quality control is the fact that we perform at the same time a research on the most interesting empirical situations (if there are no mistakes…).

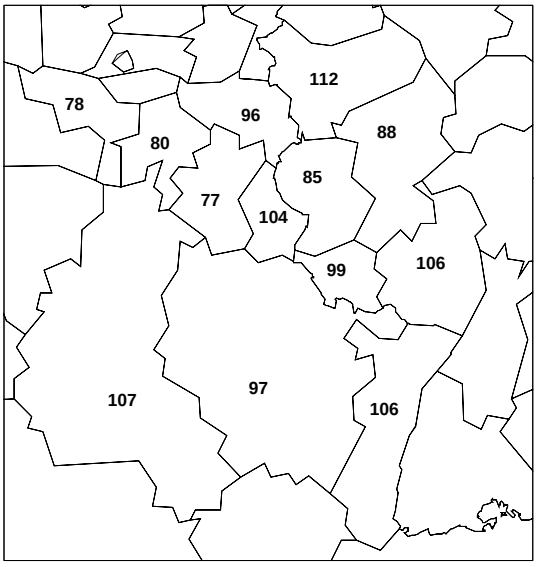
(a) Statistical information

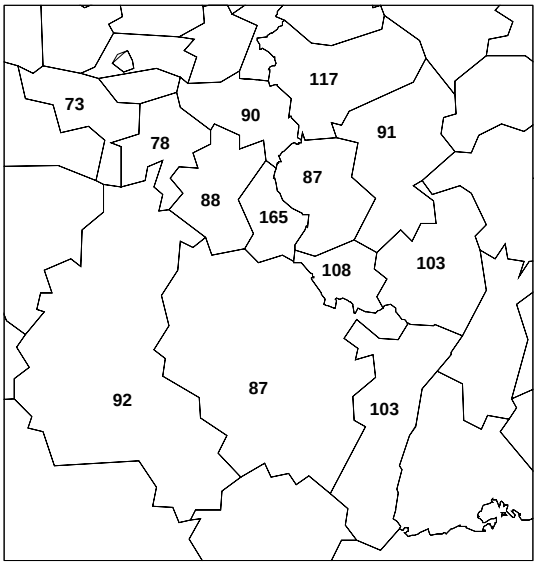
Nuts_2	Region	PIBH80	PIBH96
BE32	HAINAUT	6722	14852
BE33	LIEGE	8298	18203
BE34	LUXEMBOURG (B)	6645	17738
BE35	NAMUR	6918	15882
DEA2	KOELN	9602	23615
DEB1	KOBLENZ	7579	18491
DEB2	TRIER	7283	17660
DEB3	RHEINHESSEN-PFALZ	9138	20959
DEC	SAARLAND	8546	21867
FR21	CHAMPAGNE-ARDENNE	9242	18695
FR41	LORRAINE	8347	17616
FR42	ALSACE	9087	20885
LU	LUXEMBOURG (GRAND-DUCHE)	8972	33456

(b) Geographical information



(c)Local deviation index in 1980 and 1996





In the situation of 1980, the differences between neighbouring regions are very small. The local deviation index is minimum -23% (Belgium region of Luxemboug) and maximum +12% (German region of Koblenz). In the situation of 1996, we observe an exceptional value with the G.D. of Luxembourg which is +65% above neighbouring regions. In this case, the method of control quality has not discovered a statistical mistake but an exceptional situation from empirical point of view.

Figure 8 Evolution of discontinuities of GDP around Luxembourg (1980-1996)

2.4.2 Method of quality control based on (S)ource dimension

In most of the cases, ESPON studies are based on regional statistics published by EUROSTAT or by national institutes of statistics and it is not possible to propose alternative sources. The only possible control of quality is precisely the benchmarking between statistics published by national and European statistical offices which can help to discover problems or to solve difficulties.

But there are also situations where alternative sources of information can be combined in order to check the accuracy of the results.

If we consider, for instance, the elaboration of a database at state level covering all the world, as it was done in ESPON project 3.4.1. Europe in the World, we have different possible sources (different agencies of United Nation, OECD, CIA…) and the results are not necessary the same according to the different sources. As an example, we have compared the population in 1999 for 168 states according to the *World Development Indicator 2003* (which is an annual publication of United Nation) and *The World Economy* (a historical database built by an independent researcher, Angus Madison, and published by OECD). We can observe on Figure 8, that many differences can be observed between the two sources, especially in development countries where the most important differences lie. Sometimes, the figure of population 1999 is higher according to WDI than to TWE (e.g. Myanmar or India) and sometimes it is the contrary (e.g. Egypt or Brazil). They are no general rules and the differences can

generally not be explained by differences in territorial delimitation[7] but by differences in initial sources or methods of estimation.

Another example derived from ESPON project 1.3.3. Cultural heritage is the choice of sources for the elaboration of an index of cultural resources derived from various sources (tourism guides, official census, websites, etc…). In such situation of lack of official data, it is very important to compare alternative sources with sample tests. For example, if we decide to use tourism guide A, we compare the list of cultural resources in tourism guide B, C or D for a representative sample of region in different states. If the list converges, we can use guide A. If not, we are obliged to collect several sources ….

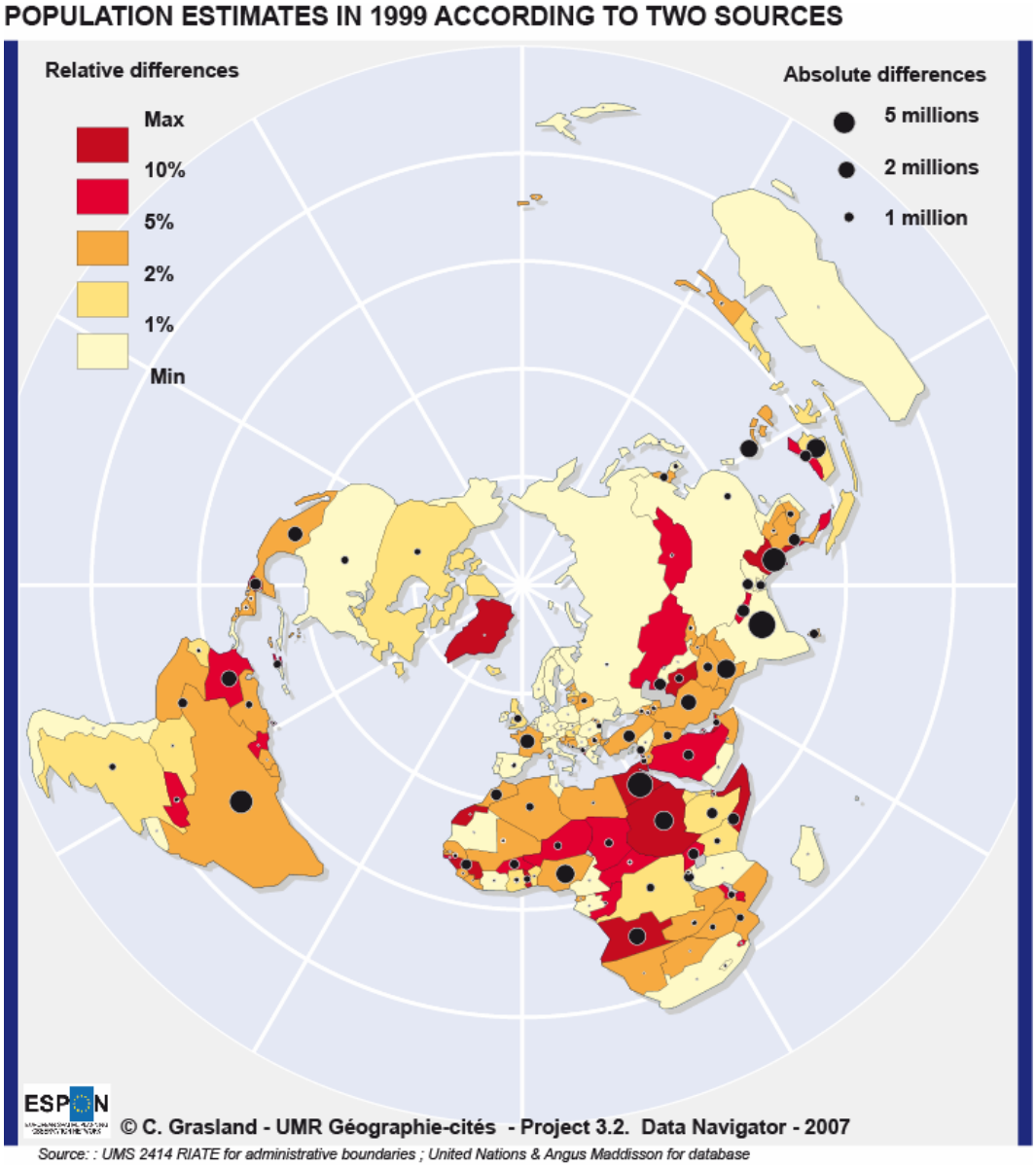
State	WDI	TWE	Abs. Dif	Rel. Dif
Egypt	62.8	69.1	-6.3	-10%
Brazil	168.0	173.3	-5.3	-3%
Indonesia	203.6	207.4	-3.9	-2%
Pakistan	134.8	138.5	-3.7	-3%
Sudan	30.6	34.1	-3.5	-12%
…	…	…	…	…
Colombia	41.5	39.0	2.5	6%
Angola	12.8	9.9	2.8	22%
Nigeria	123.9	120.4	3.5	3%
Myanmar	47.2	41.5	5.7	12%
India	999.0	991.7	7.3	1%

WDI : World Development Indicator, 2003, United Nations

TWE : The World Economy, 2001, OECD - Angus Maddison

Figure 9 Comparison of population estimates at world level according to different sources in 1999

[7] In the case of France, the difference between the population proposed by WDI (58.6 millions) and TWE (60.8 millions) were related to the fact that remote territories was not include in France in the WDI database but where included in TWE. But such explanation can not be applied in the other examples of states like Egypt, Brazil, Myanmar of India …



Map 3 Population estimated in 1999 according to 2 sources

2.4.3 Method of quality control based on (T)ime dimension

Method T1: Control of results by analysis of exceptional variations

In most of the cases, the variation of variables through time is slow and gradual. If we focus on raw count of variable (Xt), we can compute easily the derivate X't between period of time which is, in most simple formulation, the annual growth rate. We can then establish a set of variation rates (the same territorial unit at different period of time; different territorial units for the same period of time ; different territorial units at different period of time) in order to estimate the parameters of the distribution of time-variation (mean, standard deviation) and finally propose a **standardised index of variation X"t** which helps to find immediately the exceptional variations in a dataset. As in the case of the methods E2 based on spatial autocorrelation, the methods T1 based on time autocorrelation are very interesting because what they help to identify is either errors or exceptional situations of high empirical interest.

Xt	1996	1997	1998	1999	2000	2001	2002	2003
SK010 Bratislavský k.	618.6	619.0	618.3	617.0	617.2	599.1	599.0	599.8
SK021 Trnavský k.	548.6	549.1	550.0	551.1	551.3	551.0	550.9	551.5
SK022 Trencianský k.	610.1	610.2	610.0	609.4	609.1	605.5	604.2	602.8
SK023 Nitrianský k.	717.4	717.3	716.9	716.0	715.3	713.2	711.6	710.4
SK031 Zilinský k.	686.6	688.8	690.3	691.9	693.4	692.3	692.7	693.3
SK032 Banskobystrický k.	664.0	663.9	663.6	663.0	662.3	662.1	660.8	659.5
SK041 Presovský k.	771.2	775.3	779.2	782.7	786.0	790.3	792.4	794.0
SK042 Kosický k.	757.3	759.6	762.3	764.0	766.2	766.2	767.0	768.4

X't= [Xt+1-Xt) / Xt]	V96-97	V97-98	V98-99	V99-00	V00-01	V01-02	V02-03
SK010 Bratislavský k.	0.1%	-0.1%	-0.2%	0.0%	-2.9%	0.0%	0.1%
SK021 Trnavský k.	0.1%	0.2%	0.2%	0.0%	-0.1%	0.0%	0.1%
SK022 Trencianský k.	0.0%	0.0%	-0.1%	0.0%	-0.6%	-0.2%	-0.2%
SK023 Nitrianský k.	0.0%	-0.1%	-0.1%	-0.1%	-0.3%	-0.2%	-0.2%
SK031 Zilinský k.	0.3%	0.2%	0.2%	0.2%	-0.2%	0.1%	0.1%
SK032 Banskobystrický k.	0.0%	0.0%	-0.1%	-0.1%	0.0%	-0.2%	-0.2%
SK041 Presovský k.	0.5%	0.5%	0.4%	0.4%	0.5%	0.3%	0.2%
SK042 Kosický k.	0.3%	0.4%	0.2%	0.3%	0.0%	0.1%	0.2%

mean	0.00%	std-dev	0.45%

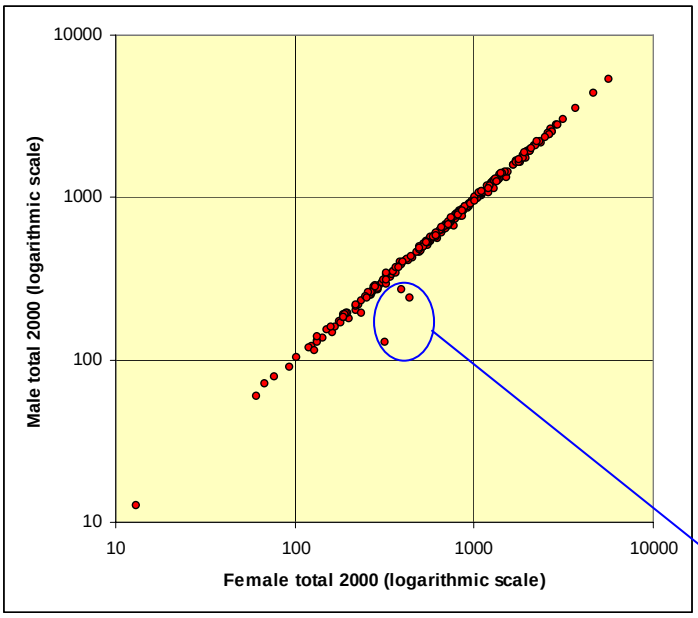
X"t = [X't - mean] / std.dev.	V96-97	V97-98	V98-99	V99-00	V00-01	V01-02	V02-03
SK010 Bratislavský k.	0.1	-0.2	-0.5	0.1	**-6.5**	0.0	0.3
SK021 Trnavský k.	0.2	0.4	0.4	0.1	-0.1	0.0	0.2
SK022 Trencianský k.	0.0	-0.1	-0.2	-0.1	-1.3	-0.5	-0.5
SK023 Nitrianský k.	0.0	-0.1	-0.3	-0.2	-0.6	-0.5	-0.4
SK031 Zilinský k.	0.7	0.5	0.5	0.5	-0.3	0.1	0.2
SK032 Banskobystrický k.	0.0	-0.1	-0.2	-0.2	-0.1	-0.4	-0.4
SK041 Presovský k.	1.2	1.1	1.0	0.9	1.2	0.6	0.4
SK042 Kosický k.	0.7	0.8	0.5	0.6	0.0	0.2	0.4

Figure 10 Example : Evolution of population of regions of Slovakia (ESPON Database, May 2006)

In this example, the method reveals an exceptional variation of population in the territorial unit of Bratislava between 2000 and 2001. 17300 inhabitants has disappeared in 1 year and this variation of -2.9% is fully exceptional in the context of Slovakian region between 1996 and 2003. As we have no concrete empirical explanation, we can suspect a mistake or an administrative change in the delineation of the region of Bratislava.

2.4.4 Method of quality control based on (I)ndicator dimension

Keeping in mind that the data collection is based on raw variables, we can use the creation of combination of raw variables as a method for testing the quality of the raw variable. If we want to check a raw variable Y, we can use another raw variable X which is correlated to Y in order to find exceptional value which could be mistakes. Many solutions are available from statistical point of view. The most simple is to build a ratio Z=Y/X and then to analyze the distribution of the Z values according to the mean and standard deviation of Z. But we can also use more accurate models on the form Y=f(X) for which we will analyse the residuals which are out of the boundaries of statistical interval



Code	Region	PFT00N2	PMT00N2	Sex-Ratio
NO01	**Oslo Og Akershus**	497.767	475.978	1.05
NO02	**Hedmark Og Oppland**	186.380	183.051	1.02
NO03	**Sør-Østlandet**	438.037	238.681	1.84
NO04	**Agder Og Rogaland**	316.430	127.834	2.48
NO05	**Vestlandet**	393.001	270.383	1.45
NO06	**Trøndelag**	195.987	193.637	1.01
NO07	**Nord-Norge**	194.449	195.506	0.99
NO	**Norway**	**2222.051**	**1685.070**	**1.32**

Figure 11 Example: Distribution of population by sex in European regions in 2000 (ESPON Database, May 2006)

In this example, the graphic crossing absolute number of male and female in European regions in 2000 reveals the existence of three regions which are clearly "outliers" from statistical point of view. If we compute the ratio of number of females per male, we obtain exceptional values for three regions of Norway and, after verification of figures on the website of statistical office of Norway, it appears that these figures were mistakes.

2.4.5 Methods of quality control based on several E.S.T.I. dimension

We have limited our presentation to very simple methods of quality conrol of data which are based on only one isolated dimension of E.S.T.I. but it is clear that better results can be obtained through the combination of several dimensions.

3 Recommendations for ESPON 2013

3.1 Database as mirror of ESPON

The recommendations that are proposed in this section rely on the following assumptions:

- **ESPON is not INSPIRE** and its mission is not to solve all statistical and cartographical problems in Europe.
- **ESPON is an applied research program** which means that it has to produce data of high quality.
- **ESPON is a political oriented program** which means that it produce evidences of practical interest for political action
- **ESPON is a spatial oriented program** which means that cartography and more generally spatialization of phenomena under investigation is crucial.
- **ESPON tries to build long term scenarios** for European territory which implies to avoid analysis limited to present situation and to develop long term times series in past and future.
- **ESPON tries to integrate sectorial policies** which means that it should cover all dimensions of territorial cohesion and propose a combination of indicators of economic competitiveness, social cohesion and sustainable development.

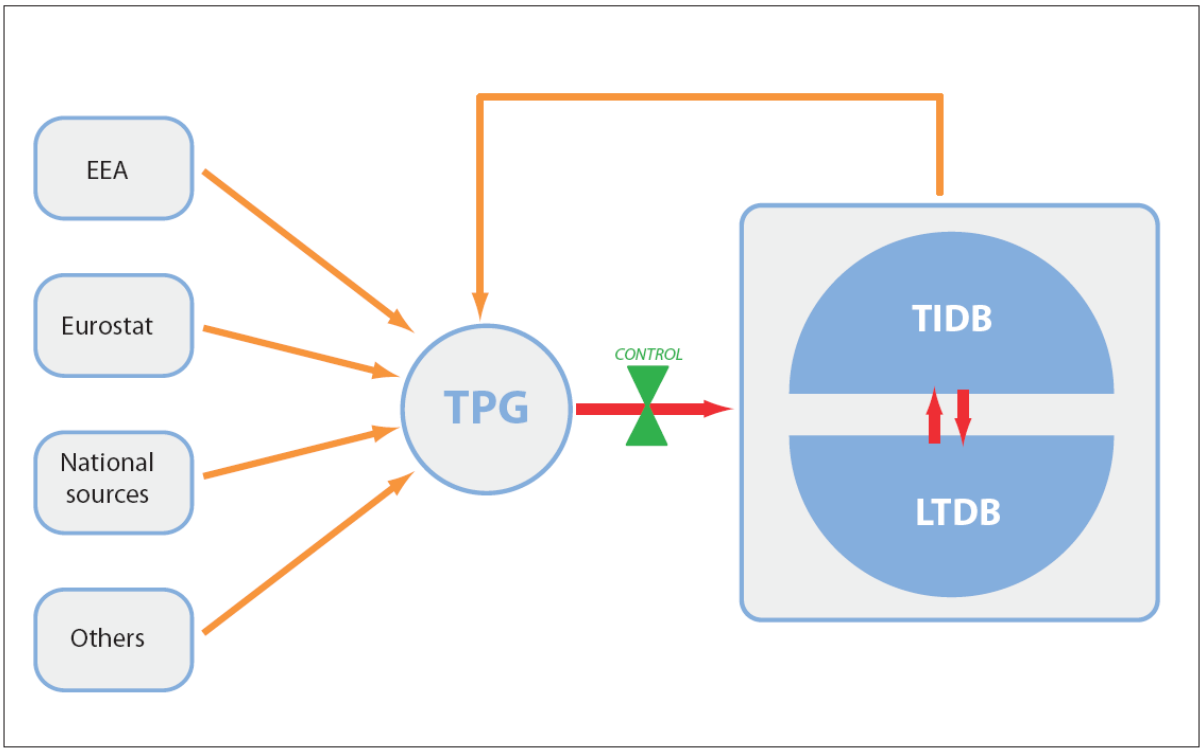
The database is a **symbolic mirror** of the ambition of a program like ESPON and looking at its structure and content it is possible to check its success and failures, to deduce its choices and constraints, and more generally to evaluate its performances.

3.2 ESPON II (2007-2013) should adopt a new database model

The final version of the database produced by ESPON I (2002-2006) is the mirror of a **pioneer period** which has explored many ways and produce impressive results. But which has also revealed many problems, expected or not, in the compilation of spatial statistics. It means that the solution which was chosen in ESPON I should not necessarily be followed in the future and that, starting from what we have learned, we should propose a new database model for ESPON II (2007-2013)

The future program ESPON 2013 will need a secure and efficient tool for its data. This point is a major one because a large part of the successful lies in the capacity of the TGP to mutualise useful, updated and limpid data and methods. In particular (*see. The 10 commandments*), the data have to be collected with a perfect knowledge of the sources and eventual transformation and the data should be easily analysed in time, space and thematic dimensions (*see. The ESTI framework*).

The database model used in 2002-2006 was certainly the most efficient in the pioneer period because at this moment we ignore many of the problems that has been discovered by the concrete practice of ESPON. But now we are aware that a simple data model structure can't fulfil all expectations of ESPON. But it is not necessary obvious to propose immediately a perfect solution to all problems with an "ideal data base model". What we propose therefore is more pragmatic and would be to introduce a **dual structure** in the future ESPON database divided into a **Territorial Indicator Data Base (TIDB)** and a **Long Term Data Base (LTDB)** .



Proposal of dual structure for future ESPON II data base model

This idea of a **dual core** for the future ESPON database is not based on an opposition between the so-called "political" and "scientific" dimensions but rather on the idea that **ESPON has to face two very different objectives** as research program of applied research for political decision:

- **Monitoring territorial cohesion in Europe in real time** is a crucial mission of ESPON which implies necessarily a limited set of indicators, updated regularly and displayed in a standard format from territorial point of view (NUTS units). This is the main objective of the **Territorial Indicator Data Base (TIDB)**

- **Building scenarios and exploring new political dimensions of territorial cohesion** is another mission of ESPON which has to make possible the development of new policies through the elaboration of new indicators (economic, social, environmental), at different scales (from local to global) and for different time periods

(from past to future). As explained by J.C. Leyghe at the final meeting of ESPON in Espoo "*the regional policy will not be maintained if we are not able to propose better statistical indexes*". Enlarging the spatial, thematic and geographic dimensions of statistical information is the main objective of the **Long Term Data Base (LTDB).**

3.3 The advantage of a network structure of ESPON database

In ESPON 2002-2007 the database was elaborated in a relatively centralised way and we can really ask if it is reasonable to propose a model which implies a more decentralised approach where the TPG's are really at the centre of the network of data base production. Is it not a risk, a danger to propose such a solution?

- **In terms of quality control** the advantage of a dual structure are obvious. There will be necessarily exchange between LTDB and TIDB and the data will be subject to a double check by the research teams responsible of this two parts of the ESPON database. The "*lead partner principle*" is certainly not relevant for this question of quality control and ESPON should avoid it for future tenders on database which should be under the direct authority of ESPON Knowledge Platform.
- **In terms of internal networking**, we strongly suggest that a network solution is feasible and offers even guarantee of efficiency than a centralised solution. TPG's will be more likely to contribute to the common elaboration of ESPON database if they consider it as their common work. Even if two research teams are specifically responsible of LTDB and TIDB, the whole ESPON community should

be considered as responsible of the elaboration of this common work or *Kollektiv Werke*.

- **In terms of external networking,** we strongly believe that ESPON should develop a higher diversity of partnerships and avoid an over concentration on Eurostat data. More contacts should be established with OECD, EEA, JRC, UN…. The specific added value of ESPON is not to produce map with Eurostat data (even with some external countries) but to analyse this data in a political way and to integrate them with other sources at national or international level.

3.4 An agenda for research on database for ESPON 2013

The objective of the future ESPON 2013 implies research on databases covering many topics. We propose here a preliminary set of topics which should obviously be completed

- How to store efficiently primary sources of information and solve the copyright problems for use by all ESPON members?
- how to keep the precise model and path of data collection, specially for the non primary indicators (storage of the estimation formulas, harmonization methods, weight used for aggregation, …)
- How to offer procedures for time harmonization with a databank of expert methods?
- How to rebuild territorial unities in order to face changes in administrative levels or territorial divisions?
- How to store the procedure for thematic harmonization which are not technical but also conceptual?
- How to secure the possibility of updating and completing data in the future ?

- How to combine data covering all dimensions for territorial cohesion when they are derived from different sources with different geometric supports (nuts, grid, point patterns …) ?

ESPON

EUROPEAN SPATIAL PLANNING OBSERVATION NETWORK

ESPON 3.2

DATA NAVIGATOR 2

Final Report

Part 2 – Experiments on database integration

Co-financed by the European Community through the Interreg III ESPON Programme

4 Experiments on database integration

4.1 Thematic harmonisation (IGEAT)

4.1.1 Introduction

It is impossible to provide general methods of how to harmonize data thematically as the issues differ quite significantly from case to case. We, therefore, chose to give some examples with very common concepts which should allow to give an idea of the types of problems encountered and the types of solutions that can be found.

Through our examples, we will show that even the use of concepts which are generally considered as standard and universal raises big difficulties, notably in international comparisons. In the framework of this study, we could not tackle all the difficulties related to the use of these concepts. The objective is only to make understandable the caveats to consider when collecting and harmonizing data.

4.1.1.1 Thematic harmonization in brief: an overview of the theoretic steps

Obtaining statistics in a given field should be tackled ideally in several steps.

The first step consists in defining in a very rigorous way the thematic concept(s) to be quantified (for example migration, development, polycentricity, metropolitanization,….). This step is strictly theoretical and should not deal *a priori* with any statistical difficulties that could appear from the concepts.

The second step aims at conceiving an ideal measurement of the concept, without considering the availability of the data since we set on an ideal perspective.

The third step consists in looking for the available statistics and estimating the limits of the available data regarding to the ideal measurement. These limits can be of various nature: unavailability of the data for some dates or some areas; problems of comparability of the data coming from different sources, or even from one synthetic source (international statistical institute); statistical bias provoked by the use of imperfect indicators. In all cases, it requires a detailed examination of the procedure of data collection, in particular by EUROSTAT when one works on European data at regional level.

The fourth and last step consists in improving the data in order to come nearer to the ideal measurement of the concept. We can imagine some more or less complex indirect indicators (proxies) to estimate the processes we are studying (see examples here under), whether you use these estimations for all data, or for some missing data (missing date or regions). In all cases, one should be aware of the heterogeneity such estimations could create.

Before giving some examples of this general approach, we would like to insist on the fact that all data are not usable: in some cases, estimations are too far from the ideal measurement of the concept, or data are not comparable enough, and it is better not to use them at all.

4.1.1.2 Some examples of well-known indicators

a) Development and economic growth

Development concept

Development and economic growth have long been assimilated in the economic literature. Authors such as Rostow or Braudel clearly assimilate the development process to a long term economic growth. However, progressively, the development concept has become larger. For François Perroux[8], growth is "a sustained increase of an indicator for a long period: for a nation, the gross domestic product in real terms", while "development is the combination of mental and social changes which make the nation capable in increasing its gross domestic products in real terms". Beyond the socio-cultural dimension of development, economic development could not be reduced to economic growth, since economic structures, technological skill or economic command are also essential dimensions.

An ideal measurement of development concept

An ideal measurement of development should deal with classical estimations of economic wealth but also with the qualities of economic

[8] Cited by Bernard Conte, « le concept de développement », http://conte.u-bordeaux4.fr/Enseig/Lic-ecod/docs_pdf/LeConceptDeDeveloppement.pdf

structures and the socio-cultural dimension of the development. We will not enter into detail of such complex matter, and will focus on the simple measurement of regional economic wealth and economic inequalities between regions. An ideal measurement of economic wealth should enable us to compare the production and consumption potentials of the territories.

Available statistics to estimate the territorial wealth

The production potential of a given territory is estimated by the ***gross domestic product,*** while the wealth of its inhabitants is evaluated by the ***net income***.

The gross domestic product (GDP) adds up all added values produced by the economic actors of a given territory. However, some choices have to be made to better define the production potential of the territory. For example, we will opt for factor cost domestic product rather than market prices product, which include taxes and subsidies. The power purchasing estimation is also more appropriate for international comparisons, since it takes into consideration the price differences between countries. In the same way, economic growth, to be rigorously comparable, should be evaluated in constant prices, in order to compensate for inflation differences between countries. All these estimations are available in the international comparisons, among others from EUROSTAT. **However, and this is a central difficulty when one considers interregional rather than international comparisons, estimations of domestic growth in purchasing power standards or economic growth at constant prices are conceived on a national basis and do not consider the important differences of prices and inflations from one region to another in a given country.**

The net income is a different measurement: we subtract the consumption of fixed capital from the GDP and take into consideration the transfers between the area and the rest of the world (for example by workers in another area, or transfers of capital incomes). In general, the net income raises the same problems of measurement than those we underline for the gross domestic product. **But the net income also neglects to take into consideration incomes transfers provoked by the redistribution policies of the national states.** An interesting attempt to take these transfers into consideration has been made by Axel Behrens[9].

b) Unemployment and underemployment

Unemployment and underemployment concepts

Unemployment and underemployment concepts are very complex and often give rise to doubtful international comparisons. Unemployment and underemployment can be defined easily as a state of involuntary partial or total non-activity for persons present in the employment market.

An ideal measurement of unemployment or underemployment

An ideal measurement of unemployment should take into consideration all the persons of the employment market who have no job, or only have a partial job while they would like to work full time. This estimation would remain imperfect since it would not take into consideration persons who are out of the employment market because they lost hope of finding a job. And saying that, we still neglect aspects related to the quality of the jobs.

[9] see Behrens Axel, « How rich are Europe's regions ? », in Statistics in Focus, EUROSTAT, 06-2003.

For example, persons could be forced to accept poor jobs in countries which have the most restrictive policies of unemployment benefits: in these countries, the statistics of unemployment improve but not always the well being of these persons!

The real measurements of the unemployment

According to EUROSTAT using the workforce enquiries, the following definitions are used:

1°) Unemployed persons are all persons 15 to 74 years of age who were not employed during the reference week, had actively sought work during the past four weeks and were ready to begin working immediately or within two weeks;

2°) Unemployment rate is the share of unemployed persons in the total number of active persons in the labour market. Active persons are those who are either employed or unemployed. Unemployment is expressed as a rate of the total active population.

We see how much such definitions are not sufficient to have a comparable estimation of unemployment, since employment markets regulations can be different from one country to another. For example, taking into account partial unemployment (on the basis of which full time job definition?), with the distinction between desired and forced part time job, may induce important differences in the global unemployment figures. Furthermore, the presence or absence in the employment markets is very dependent of national regulations and social behaviours: Irish unemployment remains weak among others because many women do not enter in the employment market; by contrast, east German unemployment is very high not only because of economic crisis but also because women activity rate is amongst the highest in Europe, a tradition inherited from the communist period. And we could give many more examples of the

problems related to the estimation of unemployment rate. Thus, when one compares unemployment rates, what does one compare exactly?

Such difficulties lead us to refuse international comparisons of unemployment, or at least to refuse the dominant interpretations of these figures in terms of economic dynamism or social well being. Interregional comparisons inside each country raise less problems since employment market regulations are mostly national. A measurement of gaps between regional and national unemployment rates appears to us to be more honest to give an image of underemployment at regional level in Europe than a simple map of unemployment rate.

c) Migrations

Migration concept

The concept of migration refers to the change of the main residence place of a person for a sufficiently long period. Such definition raises two main problems: the notion of “main residence place” - the one where the person has lived most of the time during the last year -, and the notion of sufficiently long period (in general the arbitrary limit of one year is considered). For example, main and secondary residence places are not always easy to distinguish: some could spend more or less the same time in both residence, for example some wealthy retired people; others could register their main residence in a certain place for fiscal reason.

An ideal measurement of migration

An ideal measurement of migratory movements for a given area should include all immigrants and emigrants, as well as their origins or destinations. On this basis, we could evaluate migratory balance

(immigrants less emigrants), migratory rates and quotients (respectively the migratory balance regarding to the average population of the considered year and to the population at the beginning of the year) and the directions of migratory flows.

Indirect estimations of migratory movements

Available statistics about migrations are very insufficient. The number of immigrants and emigrants are in general very incomplete and unsatisfactory figures, since many movements are not registered, among others the clandestine immigration. A massive regularization process can lead to major disruptions in the statistics. Moreover, origin-destination statistics are even weaker: interregional movements only exist for some countries, but are totally non existent between regions of different countries in Europe and between European regions and the rest of the World.

The absence of population registers in many countries can explain a part of the insufficiencies. But even in countries which do hold one, migrants often do not announce their movements, particularly in the case of departures.

In this context, we could opt for indirect estimations of migratory balances and rates. For example, we can use the natural movement method: it is based on the simple fact that the growth of population in a limited period of time is the sum of natural and migratory balances. Since figures of population, deaths and births are relatively well known, we can estimate the migratory balance. Population data are not totally reliable, notably because of the clandestine immigration, but this estimation is nevertheless a relevant indicator of territorial attractiveness. Unfortunately, these estimations do not distinguish origins of the migrants and a positive migratory balances can hide contradictory movements. For

example, the positive migratory balance of some metropolitan areas is generally the result of a very positive migratory balance for foreigners, especially from outside the EU15, and a negative balance for nationals.

4.1.1.3 ANNEX: An overview of some of the most relevant experiences of some ESPON projects

a) Migratory movements (ESPON 1.1.4.)

The following table, extracted from the 1.1.4. final report, synthesises the problems related to indicators and data concerning migrations at regional level in Europe. It deals with the most important political aspects, the availability of the data and the evaluations realized to obtain complete sets of data about some aspects related to migrations.

Political important aspects related to migrations	**ideal indicators**	**existing indicators at regional level (EUROSTAT)***	**Data used and own evaluation***
depopulation of rural-peripheral regions	1Total migratory balance 2 population evolution 3 ageing	- Interior immigration and emigration is available at nuts 2 level for most of the countries - exterior immigration is available for some countries but is generally underestimated - external emigration is very incomplete and for most of the countries underestimated	our evaluation of migratory balance has been done with the natural movement method : it enables us to have a complete matrix at nuts 2 and nuts 3 level
depopulation of young and intellectual	Migratory balance of young and	Interior arrival and departures by ages are rather incomplete but	We evaluate the migratory balance by age classes level by the “age structure

for old industrial regions	active people	less than exterior arrival and departure	method" : it enables us to get a complete matrix at nuts 2 level
socio-professional insertion of foreigners immigrants	1 proportion of population originate from poor countries 2 exterior migratory balance	- exterior immigration is available for some countries but is generally underestimated - exterior emigration is very incomplete and for most of the countries underestimated	- exterior migratory balance has been evaluated at nuts 2 level - no data is available at regional level considering the origin of the migrants
depopulation and change of social structure of centre towns ; Suburbanization and space "spending"	metropolitan and intrametropolitan migratory balance segmented by ages and social classes	No data are available about this topic because nuts 2 and even nuts 3 level are inadequate to apprehend this problem	the geographic level (nuts 2, nuts 3) used for ESPON is in most cases not relevant to measure systematically this issue even if in many cases suburbanisation processes can be observed in the maps. The scale should be the metropolitan areas and these areas would have to be divided into core cities and suburbs with homogeneous criterion.
East west migration	Rate of immigration from eastern Europe	exterior immigration at nuts 2 level is incomplete and not available by country of origin	no data is available at regional level considering the origin of the migrants
mobility (temporary) of qualified person	1. proportion of population originate from rich countries 2- migratory balance of qualified	There is no data of migration segmented in function of the social status	no data is available about migration segmented in function of social status

	people		
Touristic mobility retreat migration	1 migratory balance of aged people 2 part of second residence owned by foreigners	Interior arrival and departures by ages are rather incomplete but less than exterior arrival and departure	We evaluate the migratory balance by age classes level by the “age structure method” : it enables us to get a complete matrix at nuts 2 level

Table 1 : Indicators about migrations at regional level

* We consider here only indicators directly related to migrations

Migratory balances

Global migratory balances at nuts-3 or nuts 2 level

The migratory analysis is confronted by various difficulties, both on a conceptual point and on a statistical basis.

The conceptual difficulties are increasing:

- Western Europe has increasingly more clandestine immigration;
- since the start of the 90s, a large number of nationals from Central and Eastern Europe work and live during most of the year in the European Union, covered by tourist visas, and even now as “tourists” without any need of a visa;
- the increasing mobility of the European population and the development of secondary residences, both in their country and abroad, can weaken the relevance of population count based on the so called main residence, which moreover can be chosen not in function of main residence but for fiscal reasons.

The methodology adopted here to make up an assessment of the migratory balances at the regional level is the natural movement method. The principle is simple: one calculates the difference between, on the one hand, population at the end and at the beginning of a period, and, on the other hand, the natural balance (births less deaths) during that very period. This method is relatively safe since the statistics on these three indicators are globally reliable. Nevertheless "some relatively small errors related to the population statistics at the beginning and the end of the period, above all in the countries without population register, can bring about a much bigger error on the assessment of the final balance, especially if they are of opposite mathematical signs[10]"

For this indicator as for the others, the territorial division is very important and may change if not the result at least its interpretation. For example, in some countries or some towns of a country, the central towns are separated from their suburbs while, for most of the cities, this is not the case. Most of these centres have negative migratory balances and therefore can give the impression that the metropolitan area is not attractive. There is no simple solution to the heterogeneity of the geographic divisions but we have to be very careful in the interpretation of the data and the maps.

Interior and exterior migratory balances

The internal migratory balance of each region was calculated from the data of migratory flows between regions within each country. It, therefore, evaluates the migratory balance of a region with all the other regions of the country.

---

10 J.M. Decroly & J. Vanlaer, *Atlas of European Population*, 1991

These data are furnished in an incomplete way by EUROSTAT. We had to complete it by national sources for Germany, United Kingdom, France, Norway and Switzerland. Moreover, these data concern different periods and different spatial levels. This is particularly problematic for some countries : Romania and Slovakia where data are only available for the year 2000; in France, the evaluation can only be done between two censuses, that is to say on the 1990-99 period (instead of 1996 to 1999). The data for Greece, Bulgaria, Ireland and Switzerland are still not available.

The external migratory balance

Data on the external migratory balance are very poor and not reliable. We, therefore, made an indirect evaluation based on a very simple equation :
Total migratory balance = external migratory balance + internal migratory balance.
The external migratory balance can be evaluated by the difference between the total and the internal balance, which are much more reliable data.

Migratory balances by age groups at nuts-2 level

We have assessed the migratory balances from the age structures (by groups of 5 years) and the mortality data by age. The principle consists in following an age group on a 5-year interval and deducting the deaths from the final population: the comparison between real and assessed final population represents the migratory balance by age. Nevertheless, the balance does not relate to the initial or final age group but to the average of both.

This estimation can be formalised as follows:

Migratory balance of the n age group = population (n+1,a+1) – population (n,a) + (deaths (n+1) + deaths (n))/2

n = age group
a = year

This method is quite indirect but the results are very coherent and the image provided is comparable with other, more direct, sources available in some countries. However, some problems persist when going into detail, especially in Slovakia and in Slovenia where the data on elderly mortality lack coherence.

International flows

The data concerning international migrations are relatively poor at regional level. In the EUROSTAT database, immigration data exist only for some countries but are not available for the most populated ones (UK, Germany, France). These data do not distinguish the immigrants in function of their origins, even only from inside or outside the EU.

The data of outmigration are incomplete but also much less sure and reliable. They are in most of the cases based on the declaration of the emigrants before they leave the countries ! For example, for all Spain, the outmigration doesn't exceed 400 persons in 1999!

The evaluation of the external migratory balance is a first approach of this question of international migrations at regional level. It gives an idea of the attractiveness of the regions at the international level and enables us to show the huge difference with the internal attractiveness.

Nevertheless, it is still very insufficient and one will have to use more sources and indicators in order to better measure and understand international migratory movements :

- data at national level which give also the country of origin of the migrants;
- data from national sources : for example, Spain has very good regional data on immigration from foreign countries ;
- case studies which focus on important aspects : pensioners, skilled workforce, regional impacts of international migrations,…

b) Economic structures at regional level in Europe (ESPON 3.4.2.)

In this project, a complete matrix of added value by sectors at nuts3/nuts2 level for all ESPON countries has been produced. The choice of such geographical division has been made to have a more homogeneous division than the pure NUTS3 or NUTS2 level. The sectoral division corresponds to sections and subsections of the NACE classification, that is to say 28 sectors.

These data are not existent for most of the countries at this level of spatial and sectoral disaggregation. The statistical basis comes from EUROSTAT, which produces at NUTS 2 level, added values at basic prices for all sections (17) of the NACE classification. Regarding to the initial ambitions, these data are insufficient: for some countries, NUTS 2 appears to be too large (for example in France or Spain) compared to some others countries ; the NACE classification into sections does not break down manufacturing industries as it was intended to do; in some countries, data are even weaker (for example in Germany). However, this matrix enables us to have a coherent and comparable matrix, to which we adjust all the evaluations.

The basic hypothesis of these evaluations is that the sectoral productivity is similar in the different regions. This hypothesis is relatively reasonable when one compares this type of evaluation with complete data of added values at regional level where it is available.

To complete the database, the following principles were used:

For some countries, data were complete (Norway, Finland, Cyprus, Malta, Luxembourg, Lithuania);
For some countries, only the disaggregation of manufacturing industries was lacking (Austria, Czech republic, Greece, Hungary, The Netherlands, Sweden, Slovakia, United Kingdom, Ireland, Estonia, Latvia). We evaluate the added values of the different manufacturing sectors by breaking down the national added values with employment data at regional level.
For Switzerland and Denmark, the same methodology has been used for nearly all the economic sectors;
For Spain, France, Italy and Portugal, globally the same methodology has been used but employment figures come from census data, generally in 2001 instead of 2002. Firstly, we broke down sectoral data of added value at NUTS 2 level (17 sectors), by employment data at NUTS 3 level. In a second step, added values of manufacturing sectors at national level have been broken down by employment data at NUTS 3 level. And finally, we adjust these last figures with the total manufacturing added values at NUTS 3 level obtained with the first step of our calculation.
For Germany, the basis was a matrix of added value by sections (17 sectors) at NUTS1 level, from the German national Statistics Institute (data at NUTS2 level from EUROSTAT were too incomplete). With the same methodology than point 4, these figures have been broken down by data of salaried employment into 28 sectors at NUTS2 level.

4.2 National Sources: the Romania Example (TIGRIS)

THE ADVANTAGES OF THE USE OF THE NATIONAL DATABASES

The national statistical offices and the scientific research teams from different European states hold complete time series of data, which can prove themselves very useful in the study of the long term structural tendencies.

This data can be aggregated by other administrative strucutures than the official ones. In the context of protecting the European cultural patrimony politics for example, the analysis of different indicators at the historical regions scale is extremely useful since, in many cases, these historical regions do not correspond to the current official administrative units. The ***map 4-B*** illustrates such a case: the county limits and development regions in Romania do not overlap over the historic regions outline although these last ones can be easily reconstructed by the re-aggregation of the data collected at the *smallest* scale

Another advantage in the use of the national statistics databases which derives from the diversity of their spatial scales is *the revelation of the "natural architecture"* of the national territories, most of the time ignored by the administrative divisions.

Exposing these fine territorial structures may remove the temptation to apply unique models (of organization, of development, of administration) or it can offer relevant explanations for the defections manifested by the official administrative. ***Map 5-A and 5-B*** show the weakness of the Romanian's NUTS 2 and NUTS 3 levels, which are indifferent to the territorial structures achieved in centuries by regional urban subsystems

(which as well represent the surest basis for the implementation of the politicies on called to enhance the polycentricism .
The use of the national statistic database makes it possible to take into consideration the territorial tectonics. The ***Map 6-A*** displays the new communes which shows the local community affirmations well as the fine study of the territorial organisation's realities (for example ***Map 6-B*** –the diversity and complexity of the urban phenomenon in Romania urban), which form solid basis for a scientific examination required by the decisions taken by the public authorities.

The local scale analysis of the demographic dynamics can show a phenomenon hidden by the statistic data aggregation at higher levels, as well as the case of the recent periurbanisation of the big Romanian cities (***Map 7***), phenomenon that is at the top in the administrative definitions of the metropolitan areas. Combined with the indicator's diversity *which can be found in the national statistic database*, this sort of analysis shows it's entire importance in the research of the complex territorial systems, vital in the implementation of the local development strategies, regional or inter-regional. Thus, the information about equipping the households and territory (***Table 2-A and Map 8***), collected at a NUTS 5 level, permit to reveal some territorial structures and of some spatial diffusion phenomenon which might slip out of the decision facts attention., used to thinking the territory from the point of view of a higher level administrative organisation.
The spatial diffusion, an extremely important phenomenon in the estimation of the natural tendencies or in the implemented politics, can only be made obvious with the help of the statistic series of data displayed in long periods of time, made possible by the national statistics service. ***Map 9 and map 10*** (the demographic diffusion of the Romanian demographic transition in the last 50 years) illustrate a convincing example of the interest shown by the national statistics in the long term modernisation of the spatial dynamics.

Difficulties and good rules of practical use of national sources

Braced in the dynamic realities of the territories, the national statistic series of data describe most of the times specific phenomena, with no correspondence established between the other states (as is the *mixt economic activity sector* in the Romania of transition, cf. ***Table 3*** - **Professional status by activity sector**), describe in different ways the collected indicators (unemployment, migration – cf. ***Map 11***), etc. From here one of the major problems in the usage of the national statistic database: *the extreme difficulties in harmonising data coming from 29 national statistical offices*.

The efficient and correct use of the national database as part of studies at the entire European Union scale presumes two types of actions:

1) *Radical actions*, that can have extremely favourable results at a medium and long period of time but are time and money consuming:

a) homogenisation according to an European standard of the national statistic systems ***or***

b) flexibility at a medium and long term of the community policies depending on the results of the research accomplished depending on the existing national statistic systems .

2) *Flexible actions* that can be more quickly adaptative for the ESPON program 2007-2013, although may also be getting ready for the radical actions mentioned above:

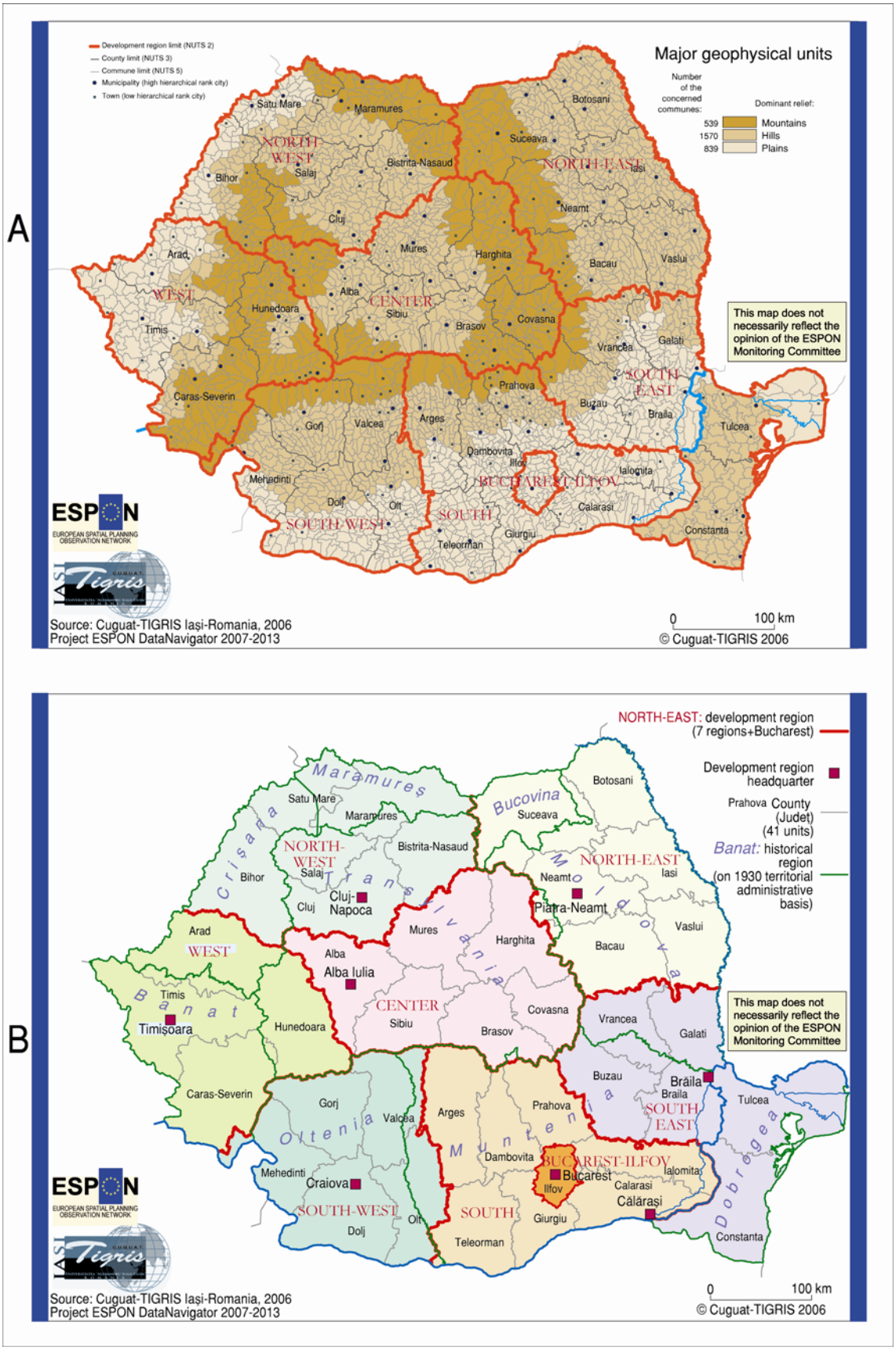
a) checking the EUROSTAT values by comparison with national sources for the existent data;

b) development of a European database of definitions used for indicators of each one of the 29 national statistical offices and the selection of the indicators with the same definitions in order to built up a European statistic base at levels NUTS 4 and NUTS 5;

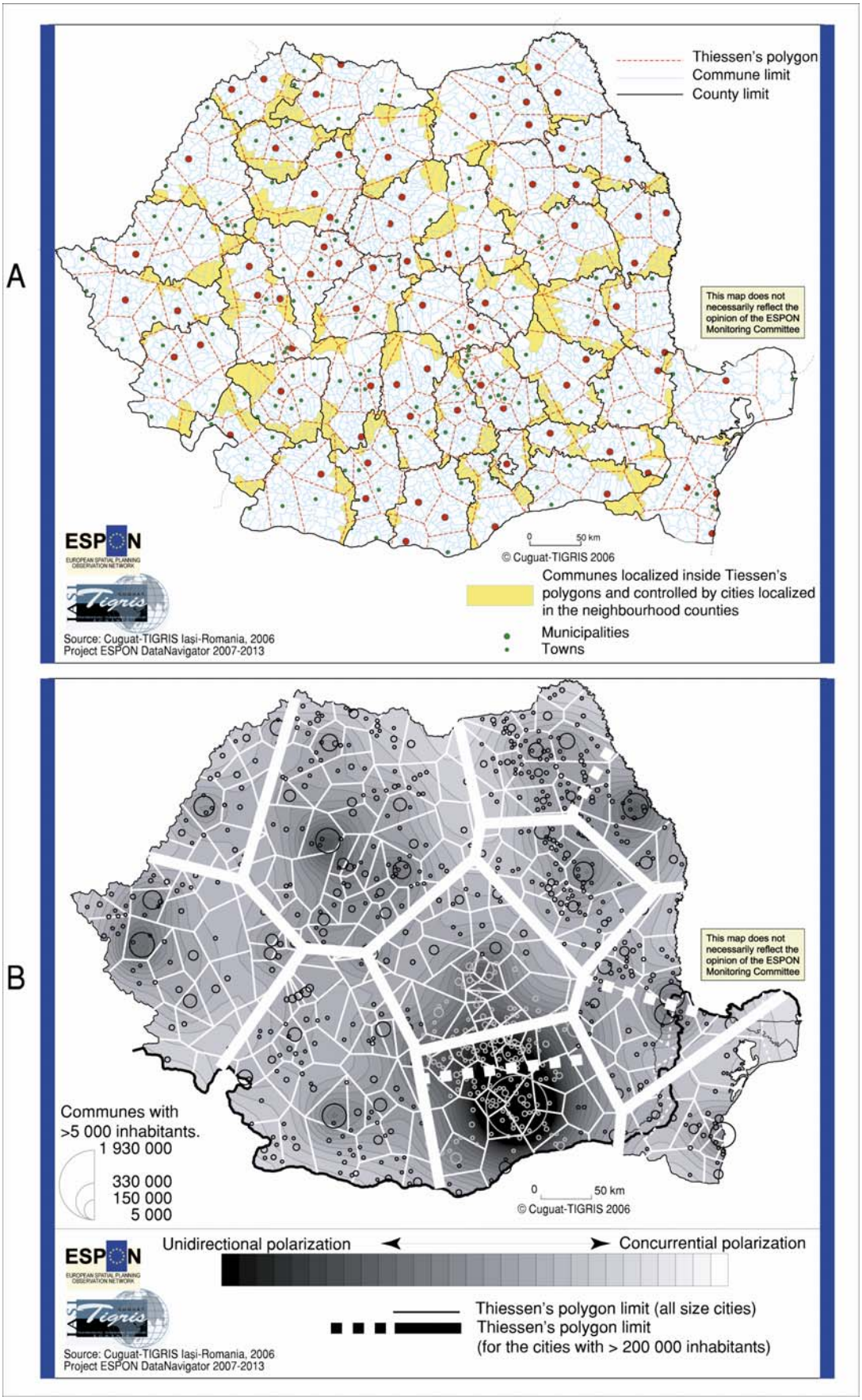
c) with the purpose of verifying the viability of this database, in depth case studies on selected countries or on selected cross-border regions;

d) verification and validation or invalidation of the projects achieved in ESPON1 by systematic case studies at a national, regional or cross border level;

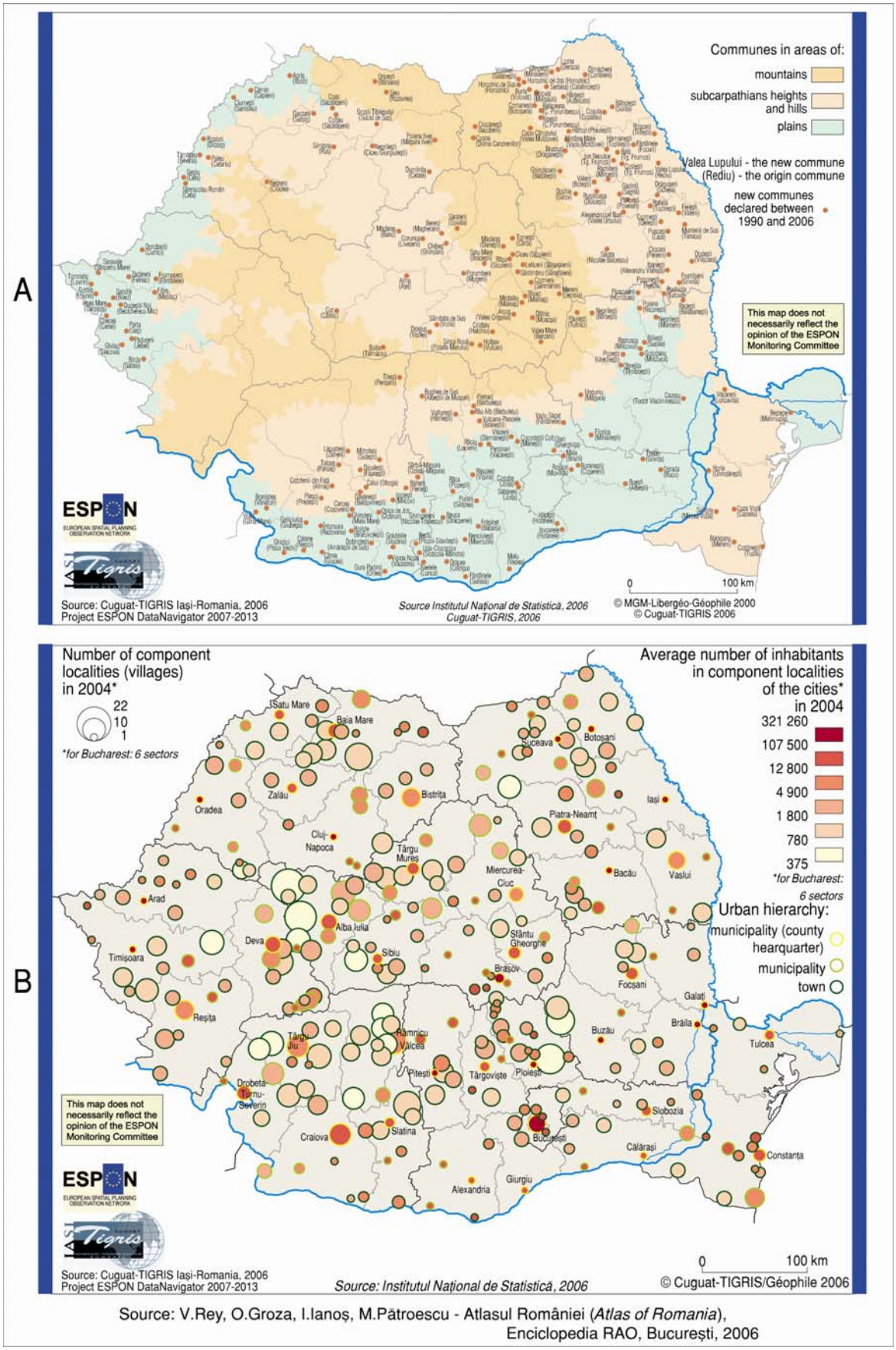
e) using of statistic techniques for adjusting the series of national statistic data and cartographic methods capable of lowering the danger in using collected data from territorial units of different size (especially different kinds of gridding – krigging, nearest neighbour, triangulation, etc.).



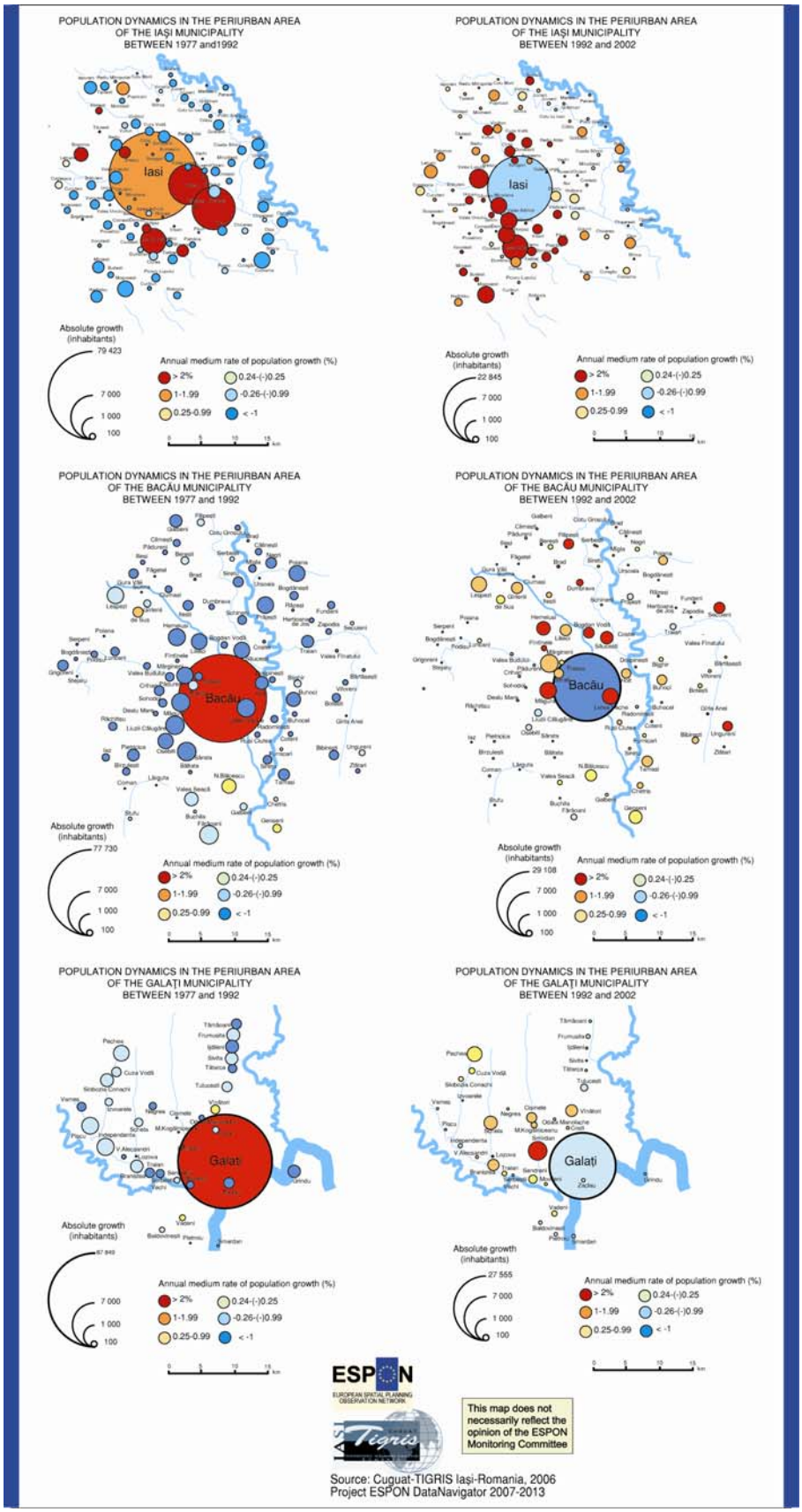
Map 4 The administrative structure of the Romanian territory



Map 5 The spatial structure of the Romanian territory



Map 6 Some administrative problems of the Romanian territory

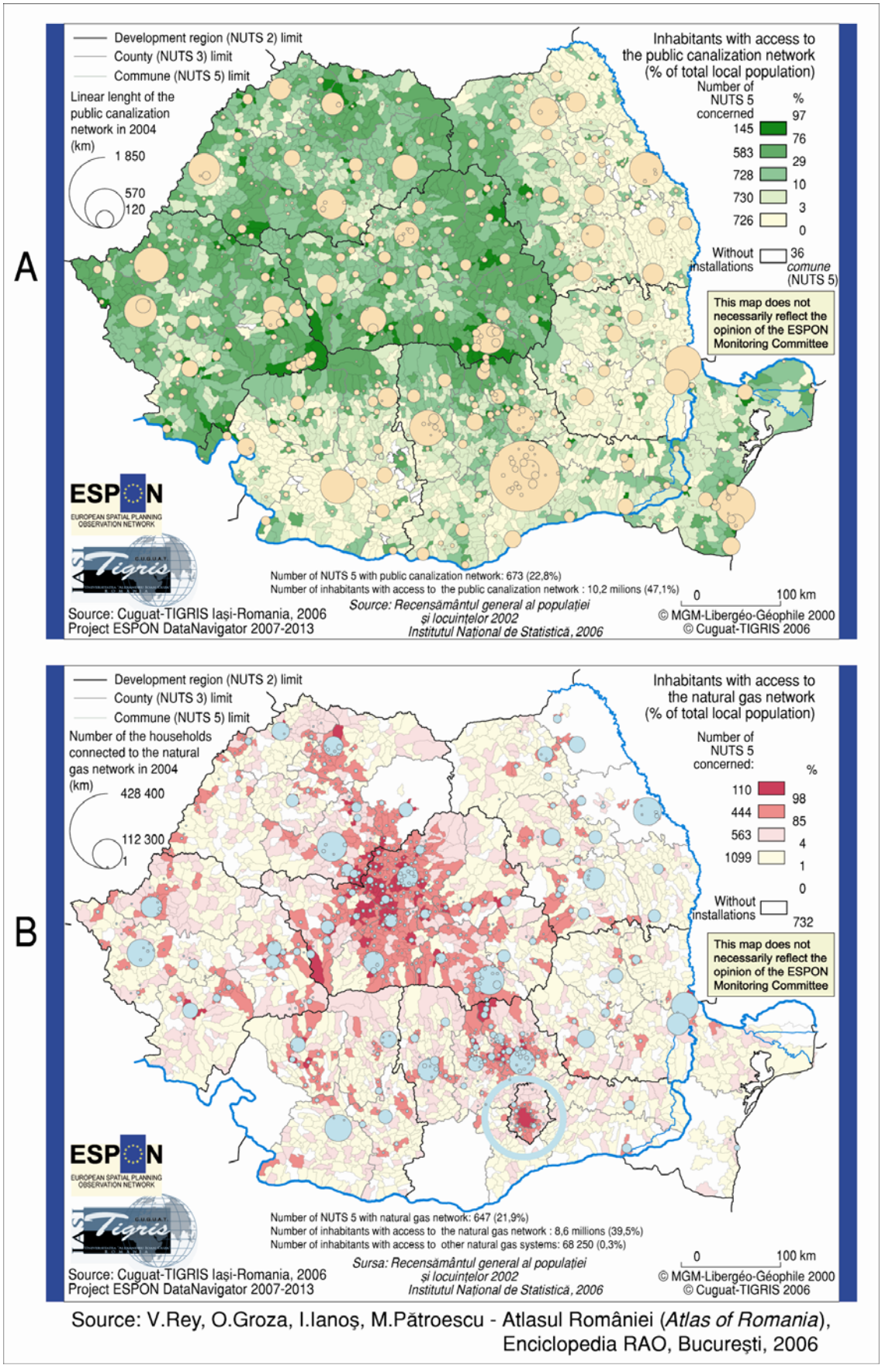


Map 7 **Local level cartography** (data availability : every 10 years on village level - census and every year - estimations)

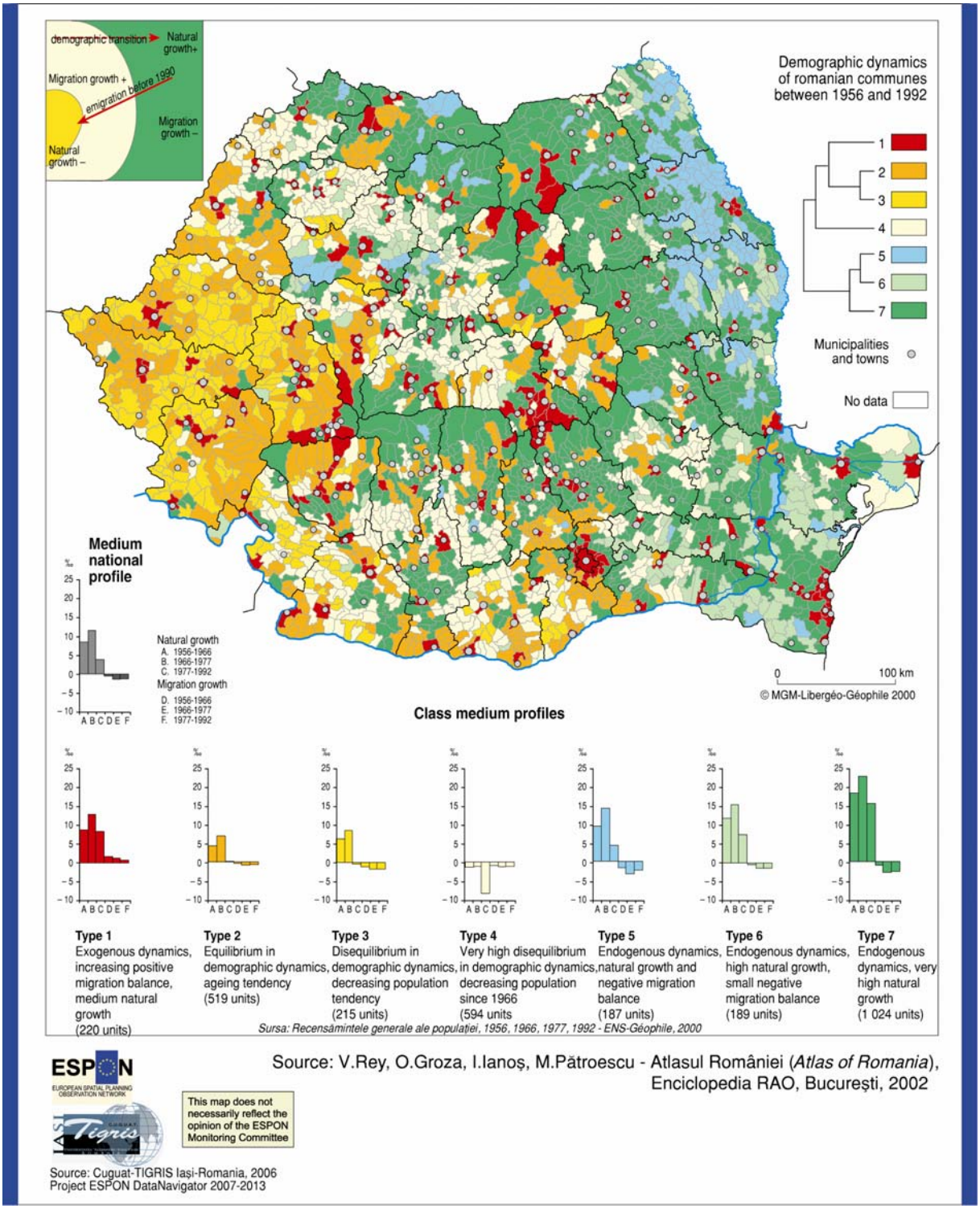
Table 1 Selected households equipment statistics (data availability : every 10 years on village level - census and every year on communal level - estimations)

(Availability: Census – every 10 years)

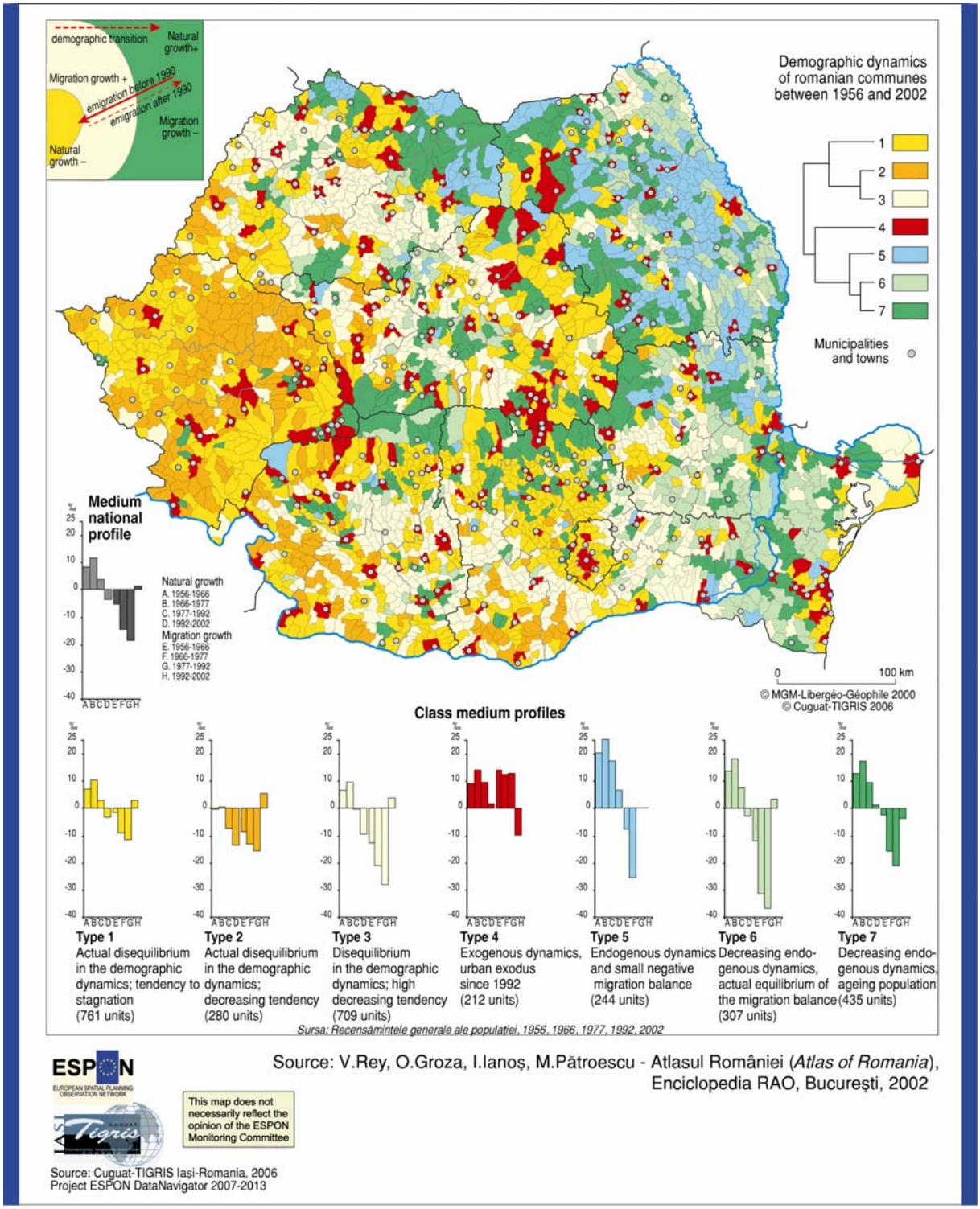
Households with:	electric installation
	water installation
	hot water installation
	canalization installation
	gas installation
	other situations
	without installations
Occupied households with:	electric installation
	water installation
	hot water installation
	canalization installation
	gas installation
	other situations
	without installations
Number of inhabitants	electric installation
in households with:	water installation
	hot water installation
	canalization installation
	gas installation
	other situations
	without installations



Map 8 Households equipment statistics: map examples



Map 9 **Mapping the spatial diffusion: Demographic transition between 1956 and 1992** (data availability: every 10 years on village level - census and every year on communal level - estimations)



Map 10 **Mapping the spatial diffusion: Demographic transition between 1956 and** 2002 (data availability : every 10 years on village level - census and every year on communal level - estimations)

Table 2 Selected demographic indicators (data availability on local level: census - every 10 years)

Active population

Occupied population

Unemployed searching another job

Unemployed searching the first job

Inactive population

Students (college/university)

Retired persons

Housewives

Supported by other persons

Supported by public or private organizations

Persons with a different status

Professional status

Employees

Owners

Free lancers

Members of an agricultural or a cooperative enterprise

Workers in their own households

Other situation

Undeclared

Professional status by activity sector

Persons occupied in the public sector

Persons occupied in the private sector

Persons occupied in the mixt sector

Persons occupied in their own households

Persons occupied in other households

Persons occupied in undeclared sector

Persons who have changed their professional status in rapport with the previous year:

Employees

Owners

Free lancers

Members of an agricultural or a cooperative enterprise

Workers in their own households

Other situation

Occupied population by professional status and by national economic activities:

Employees in:

Agriculture, forests and hunting

Fishing

Mining

Processing industry

Electric and thermal energy, natural gas and water

Constructions

Trade and technical services (cars, domestic equipment)

Hotels and restaurants

Transports, house wares, communications

Banking and financial transactions

(Real)Estate transactions, services to enterprises

Public administration

Teaching

Health and social care system

Other public, social or personal services

Workers employed in households

Extraterritorial organizations

Undeclared activity

Owners in:

Agriculture, forests and hunting

Fishing

Mining

Processing industry

Electric and thermal energy, natural gas and water

Constructions

Trade and technical services (cars, domestic equipment)

Hotels and restaurants

Transports, house wares, communications

Banking and financial transactions

(Real)Estate transactions, services to enterprises

Public administration

Teaching

Health and social care system

Other public, social or personal services

Workers employed in households

Extraterritorial organizations

Undeclared activity

Free lancers in:

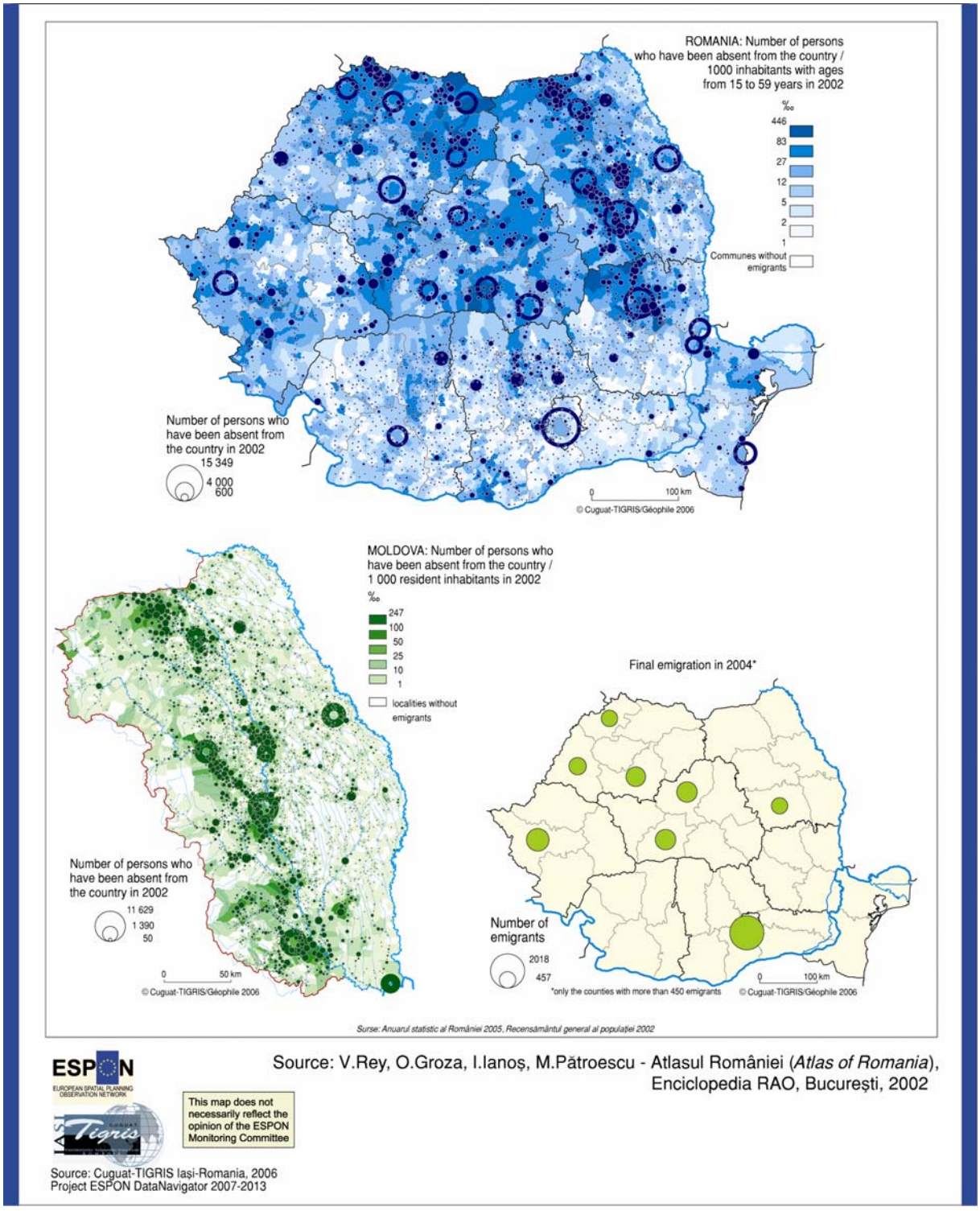
Agriculture, forests and hunting
Fishery
Mining
Processing industry
Electric and thermal energy, natural gas and water
Constructions
Trade and technical services (cars, domestic equipment)
Hotels and restaurants
Transports, house wares, communications
Banking and financial transactions
(Real)Estate transactions, services to enterprises
Public administration
Teaching
Health and social care system
Other public, social or personal services
Workers employed in households
Extraterritorial organizations
Undeclared activity

Members of an agricultural or a cooperative enterprise in:

Agriculture, forests and hunting
Fishing
Mining
Processing industry
Electric and thermal energy, natural gas and water
Constructions
Trade and technical services (cars, domestic equipment)
Hotels and restaurants
Transports, house wares, communications
Banking and financial transactions
(Real)Estate transactions, services to enterprises
Public administration
Teaching
Health and social care system
Other public, social or personal services
Workers employed in households
Extraterritorial organizations
Undeclared activity

Workers in their own households

Agriculture, forests and hunting
Fishing
Mining
Processing industry
Electric and thermal energy, natural gas and water
Constructions
Trade and technical services
Hotels and restaurants
Transports, house wares, communications
Banking and financial transactions
(Real)Estate transactions, services to enterprises
Public administration
Teaching
Health and social care system
Other public, social or personal services
Workers employed in households
Extraterritorial organizations
Undeclared activity



Map 11 Mapping emigration (data availability on local level: census - every 10 years)

4.3 Practical example of data integration between environmental and socio economic data (Géographie-cités & LSR-IMAG)

4.3.1 INTRODUCTION

To integrate data as different as socioeconomic and environmental data, enables us to meet two major needs:
can one currently build coherent complex indicators in time and space starting from these two families of indicators?
is the schema currently retained for the LTDB adapted to answer such requests?

The EU's Sustainable Development Strategy (SDS), adopted at the 2001 Gothenburg council
Consider that sustainable development and, more generally, environmental dimension are core components of territorial cohesion (as important as economic competitiveness and social cohesion). It is therefore important for ESPON to develop the elaboration of territorial index in this area and the project ESPON 2.4.1 has precisely tried to elaborate guidelines for future ESPON II. Many data are available on environment in the EEA which should be a partner for ESPON of equal importance than EUROSTAT in future ESPON II.
But an important technical problem, pointed in ESPON project 3.4.3 (MAUP) is the choice of territorial units and, more generally, spatial tools able to integrate socio-economical data with environmental ones. The solution which was adopted in ESPON I (transfer of environmental data

toward NUTS2 or NUTS3 units) is probably not the best, even if it was the only possible according to the structure of the ESPON database.
It was therefore of utmost importance, before to propose an handbook for data collection, to analyse in more detail the interest and the feasibility of integration of socio-economical data with environmental information. And to precise which modifications should be introduced in future ESPON database if ESPON II want to improve the combination of indicators derived from EUROSTAT and EEA in a coherent way

We chose to test LTDB and Europeans databases with a relevant set of thematic questions to test this integration of data according to various levels of difficulties:

Can one easily apprehend the evolution of population between two dates from the available sources ? The target period is 1990-2000 because the geographical units changed a lot during this period (identification, outline,…).

How important is the green framework for the European citizens : which is their green context? We look at this preoccupation through the possibilities offered to Europeans to have a close forest space: availability of forest per capita through various scales and administrative boundaries (NUTS) but also, in a spatial environment of 10 km (about 15 to 30 minutes accessibility); to obtain in fine a potential of forest per capita in the 10 km around the basic geographical units.

How to measure the concept of “true density” for a set of dates ? Following the example EEA, we try here to rebuild a regular grid with stock of population. The spatial resolution is 100 meters. The rule for assigning population rely on land uses and allowing a comparison for the two dates (1990 and 1999). The ventilation of the populations is obtained

by a desegregation calculated with a kind of "Look Up Table" based on the distribution of land uses given by Corine Land Cover
These complex questions mobilize many heteroclite informations in time and space. The socio-economic data come from the EUROSTAT databases and ESPON database. The environmental data result from the base Corine Land Cover (©EEA, 1990 and 1999).

The full size test has been achieved for three states: Nederlands, Belgium and France:
For those states socio demographic datas are very present in data bases
They represent a sufficiently important zone so that tests have the sense
The Franco-Belgian border represents a good example of multisources juxtaposition (NUTS carvings, Corine Land Cover,...)
We dispose for this zone both natural borders (maritime coasts) and administrative borders
The morphology of the NUTS is there heterogeneous
One can raise many changes of grid cells between 1990 and 2000
Maillageses 1988 and 1990 have been redrawn (in the framework of this survey) in order to make them compatible with those of 2003 because of generalization differences

4.3.2 Choice of themas

We chose to work on indicators that permit to reach relevant thematic applications. To improve the European data base knowledge, we chose themas presenting increasing levels of complexity.
Three axes have been kept for the integration of data socio demographic and environnemental

4.3.2.1 Time Integration

It is here about representing of an information of population in a territorial maillage when reference dates change:
antecedence of the map grid on socio demaographic data : we led two tests (1) representing the population of 1990 in the NUTS 2-3 of 1988 and (2) 2000's in the NUTS 2-3 of 1999
antecedence of socio demaographic data on the grid of reference: to recover the population of 2000 in the NUts 2-3 of 1988 or the one of 1990 in the NUTS 2-3 of 1999
evolution of population in a indifferent NUTS : evolution of the population of 1990 to 2000, evolution of the population 1990-2000 transcribed in the NUTS 2-3 of one any date,

4.3.2.2 Integration to estimate a density of population

We try to construct a grid of densités to approach the real density notion, based on types of Land Cover. This real density should permit a easier representation of populations through changes of map entities. A "look up table" permits us to pass from an occupation of soil (EEA) to a grid of densities of regular population. We reconstituted two population densigrids: 1990 and 2000

4.3.2.3 Integration to fear the green life's framework of the European

What is the green context of the europeans life: in this part of the survey, wy tryed to raise a portrait of the green framework in which lives the European citizens. We make it through two themas:
- the part of forest existing in the European territorial unities (inventory, accounting)
- the forest space accessibility for the European citizens.

4.3.3 Data Sources

4.3.3.1 Maps data

Several carvings are necessary for the whole survey. We kept following maps:

- NUTS23 de 1980
- NUTS23 de 1988
- NUTS2 et NUTS3 de 1999
- NUTS0, 1, 2 et 3 de 2003

NUTS maps for the years 1980, 1988 and 1999 are drifted from a same origin. The 2003's map is currently and easily the alone available. But, it is not free of charge.

Two important problems present themselves since the exam of those maps:

The stake in relation of the vectorial sources and environnemental sources (Corine Land Cover). The geometric consistency between these maps and Corine Land Cover is not assured and the necessary geometric corrections are difficult because the whole distortion is not uniform.

Si Corine Land Cover et les fonds 2003 sont correctement calés, cependant, la généralisation des fonds est beaucoup plus forte en 2003 que pour les découpages des années précedentes. (see figure 31 and 32)

4.3.3.2 Environmental datas

Data that we chose to integrate are those of the EEA : Corine Land Cover (cf. infra)

4.3.4 Integration of data

4.3.4.1 Temporal integration

a) Simple Integration : Data of population in the NUTS grid

We look for the representation of population at different dates, despite of the possible geographical unit changes (identifications or geometries).

In case of coherent data, as for example NUTS23S of 1988 and the population of 1990, a simple joint here permitted to associate the set of data to the spatial units, without particular problem.

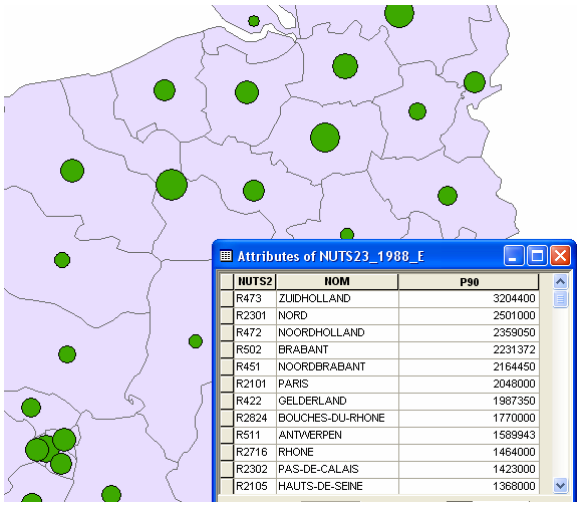


Figure 12 Population 1990 in NUTS 2-3 1988

On the other hand, that to make when the cartographic and statistical data are not compatible in the time evolution ? (example, population of 2000 in NUTS 23_88)

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

Method 1 :integration by evaluation based on a hypothesis of uniform distribution of population within NUTS: using a grid of population densities

Create a precise density grid simply based on the NUTS global population (sum of the values of all pixels of the nuts = nuts population and each pixel of the same nuts has the same value)

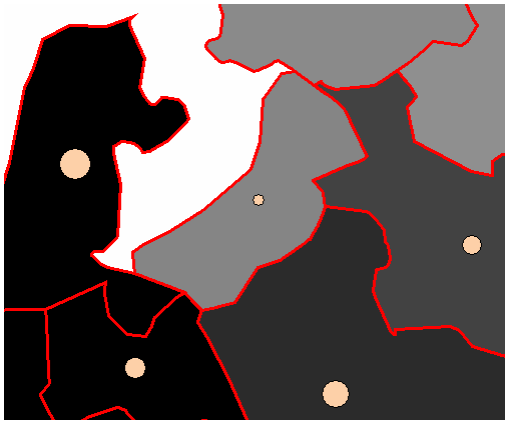


Figure 13 **Population 2000 for NUTS 2-3 1999 and population 2000 for a 100*100 meters grid**

And then sum the pixels values within the 1988 nuts.

Figure 14 Population 2000 for the NUTS 2-3 1988 and N23_88 and population 2000 for a 100*100 meters grid

Limits and distorsions:

This method of evaluation supposes an uniform distribution of the population in the space. However, as the shows the example of the Netherlands, the distribution of the population is very irregular

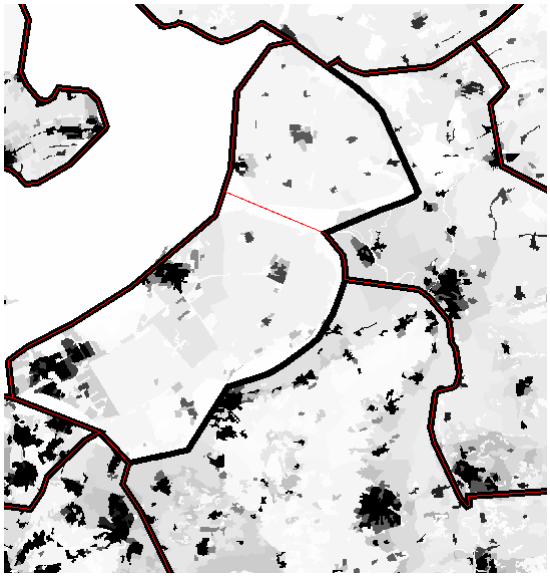


Figure 15 CLC P 2000, NUTS 2-3 1988 and NUTS 2-3 1999

Method 2: integration by evaluation based on a density grid: using the EEA grid of population densities

The EEA provides a grid of density of population according to a regular construction and a resolution of 100 meters. This grid has been build on a grid of population coefficients. These coeffciients are described for each Corine Land Cover type. The population indicators comme from eurostat (1990 and 2000).

From this grid of density of population and of the geographical grid NUTS 23 of 1999, we estimate the population by summurisation of pixels in each geographical unit.. A simple interpolation between values of density and pixel number in each unit permits to know the plausible total population by administrative unit.

Otherwise, a problem of mass conservation appears in this evaluation. The integration of land use and population come here up against a very different generalization level.

Because of the different generalization level between CLC and NUTS grids, a large part of the population is re-ventilated on the non habitable zones (sea,..). (cf. Marseille)

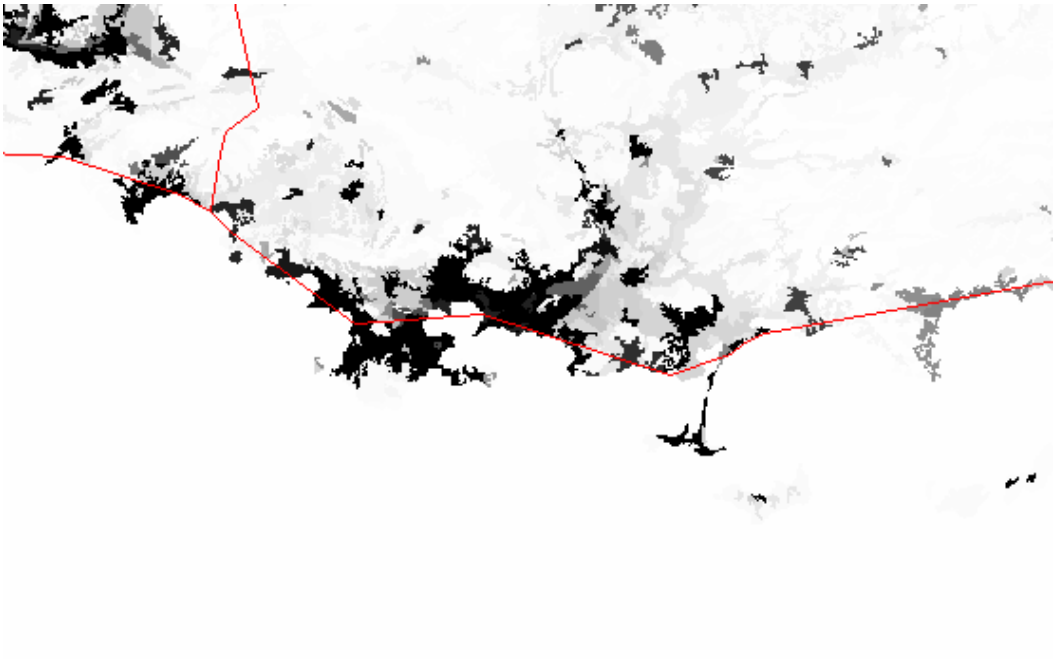


Figure 16 Medirterranean coast, near Marseille (France)

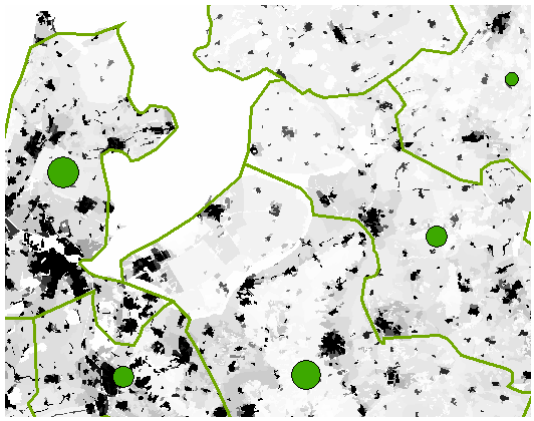


Figure 17 Population 2000 (issue de CLCP00) et CLCP00 dans N23_88

A such method is not usable in the time if we don't have grids of density population based on a same construction methods for the two dates (1990 and 2000).B

Method 3 : integration by an evaluation based on the reconstruction of density grids according to land use measured for two dates

Limits and slanted the method using directly EEA grids of density succeeds to a question very simple: can we construct our own grid of densities of population to assure a harmonization in the time ?

We suppose (1) that the population distribution is affected by the land use typology repartition and (2) that two similar land use modalities are homogeneous and offer similar potentialities of population for one date. EEA] provides, outre its own grids, des tableaux to translate types of land use to probable distribution of population. It is tables of coefficients of population by type of land use at the different dates.

Several setps are necessary for this construction:

1°) Every CLC type is coded according to coefficients of the EEA in 2000. We used some similar coefficients between 1990 and 2000 for reasons of facilités de recodage.

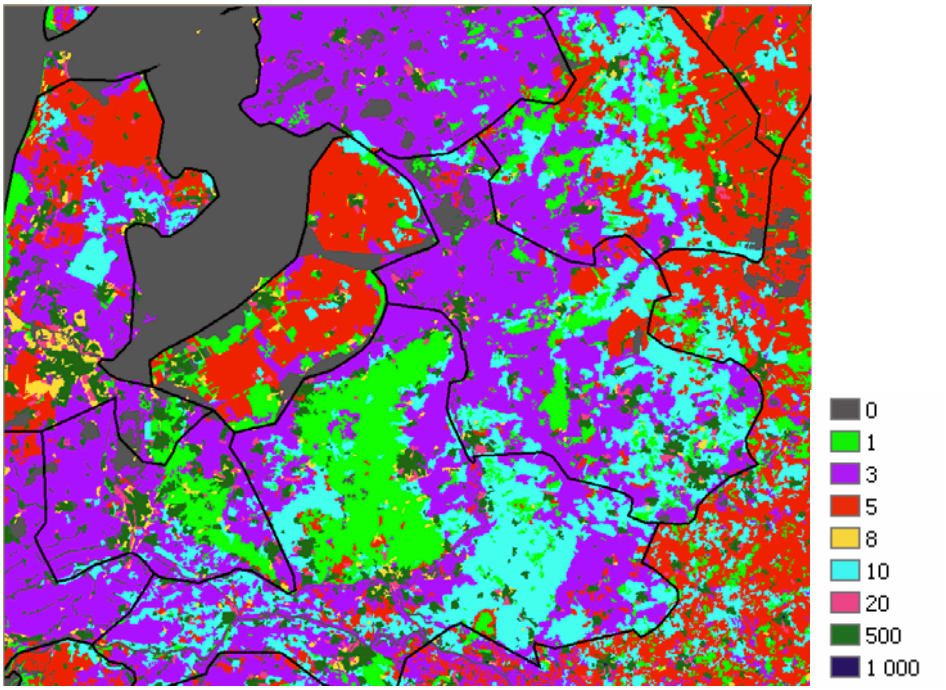


Figure 18 Grid of coefficients according to CLC 2000

2°) An adjustment grid is computed and allows to refine coefficients of polation distribution. Is obtained simply: we compare the is coefficients obtanied for the NUTS 23_99 with real populations in the same NUTS (ratio between the theoretical population and effective population in 1999).

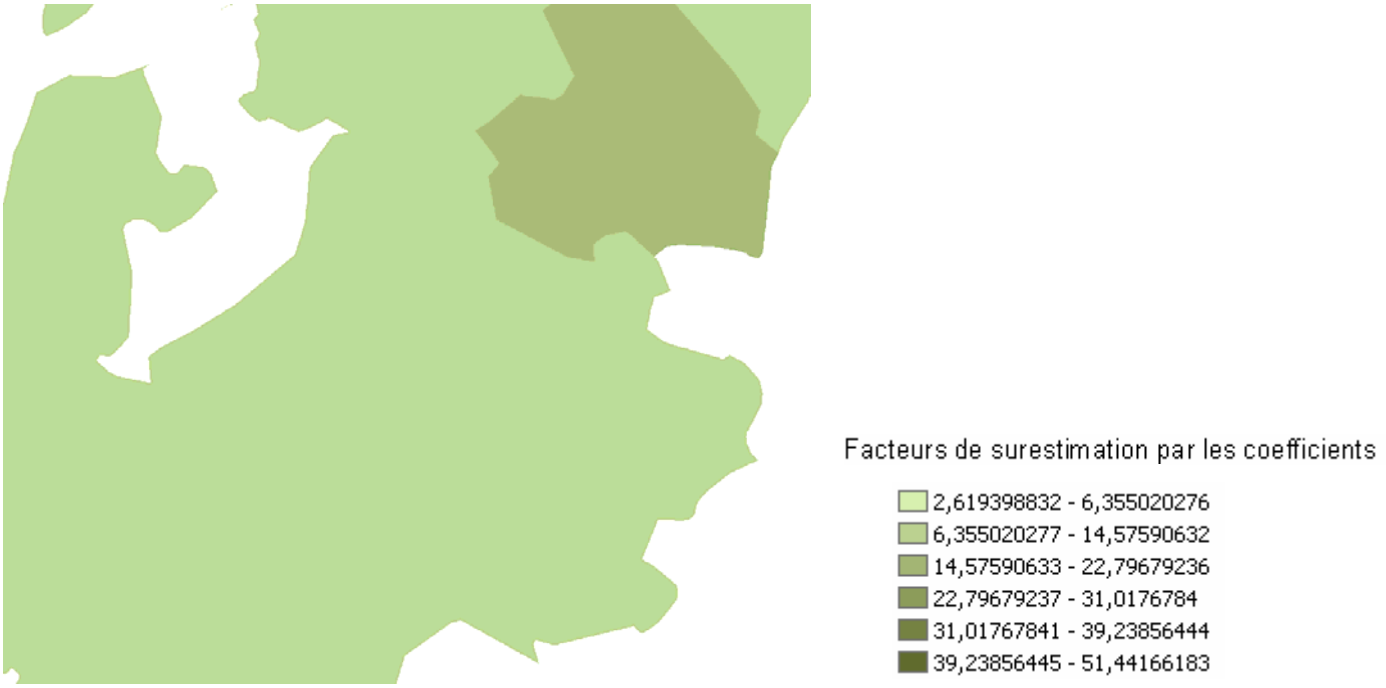


Figure 19 Grid of adjustment 2000

3°) Coefficients of distribution are straightened according to the calculated adjustment

4°) The population is computed for the NUTS according to this grid of densities

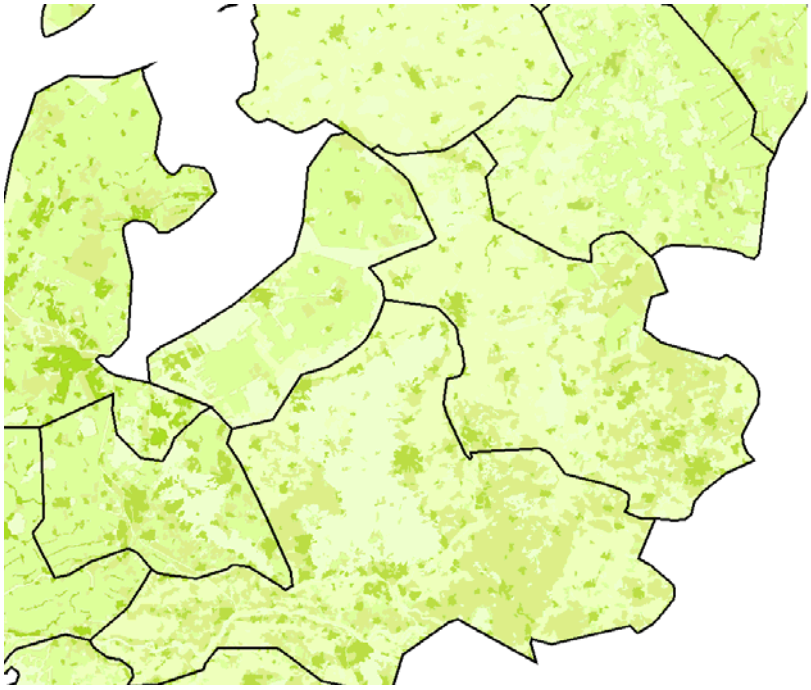


Figure 20 Population density grid (2000)

Limits and distorsions:

The twin sources (CLC and NUTS) have a very different spatial resolution. This implique important problems of exactness of pixel localization for a numbering.

A part of typology, wich represent high density of population, is not taken therefore in account in grids of adjustment and grids of densities.

Some land use types are then overestimated (cf. Marseille)

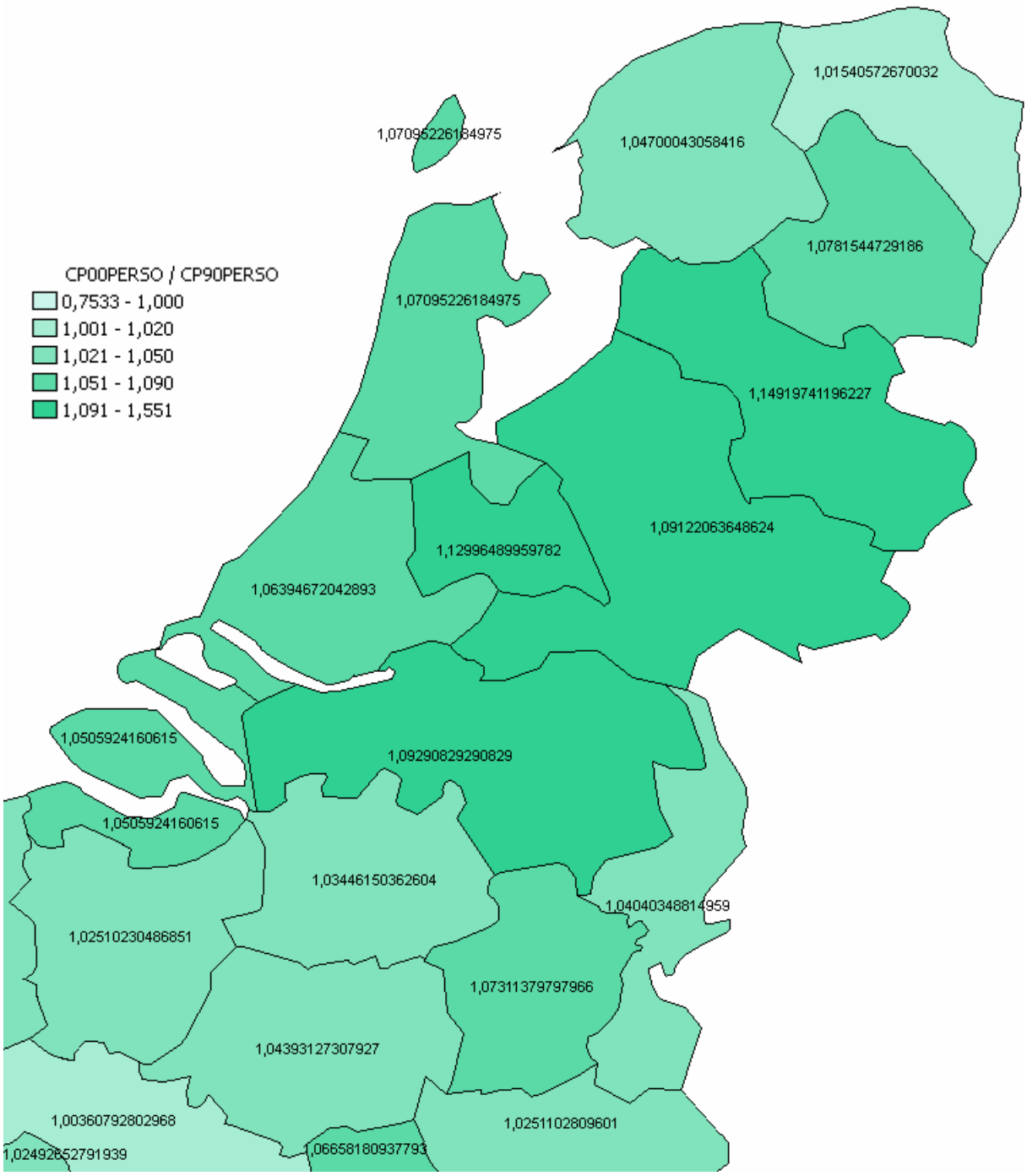
However, the global evaluation for the NUTS population is very acceptable and the population of every NUTS is rather coherent..

b) Encountered problems for a temporal integration : evolution of the population 1990 -2000

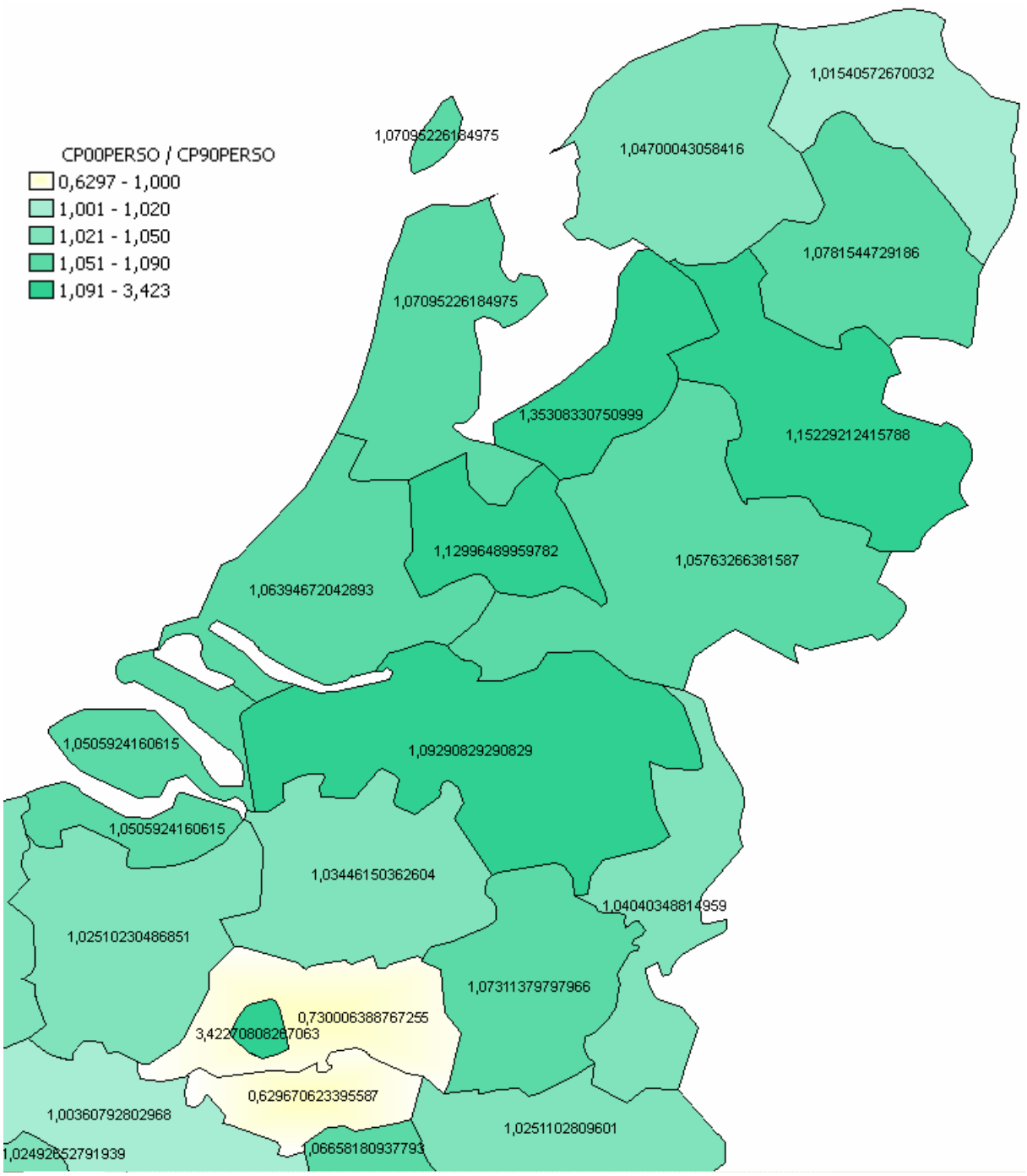
Changes of geometry and changes of units identification, don't permit to get directly evolutions of population basing on initial data, as the shows following example:
We estimate an evolution for a middle time (1990-2000) and represent it for different geographical grids (NUTS 23 1988 and NUTS 23 1999)
Whereas data of population initial are the similar, calculations of evolution (1990-2000) defer very strongly from a geographical grid to the other.

The very different cartographic representations imply some very different analyses
results seem rather little believable in zones of unit modification.
The hypothesis of a homogeneous distribution of the population entails these mistakes.
It is therefore necessary to use the harmonized data.

We try therefore with data produced by a same method (density grids based on land use repartition) .



Map 12 Evolution of the population 1999-2000 ; NUTS 2-3 in 1988



Map 13 Evolution of the population of 1990 to 2000 in the NUTS 2-3 1999 territorial division

The two assessments, and so the two representations, are enough similar and computed data look like reality. However, problems subsist for the NUTS near of Brussels. They are owed to slants of the method of calculation used for the grid of population density. The grid of adjustment doesn't take in account, in 1990, the very dense and very urban statute of Brussels. The adjustment is calculated for the whole NUTS (that includes Brussels and of other NUTS).

c) Experimentation to sortir des initial changing territorial divisions

The original territorial divisions are not satisfactory to transcribe an evolution of population. We are going to describe continuously the evolution of population in the space through a grid:

Using our own grids of population density: This solution is not satisfactory because zones that have changed of land use can't automaticly be populated or depopulate (instantanly change of the coefficient). Besides, zones without inhabitant, become populated (change of CLC type) undergo an infinite evolution and zones populated can subitly depopulate entirely in the same way.

Using data of origin disintegrated in a grid: the population is considered as homogeneously REPARTIE in the space and ithe computing is based on the previously calculated density grid. A simple ratio between the two grids of population give a simple information about evolution within the NUTS. The spatial precision is finest but we discern here only trends within NUTS (that are the basis of evaluations).

Figure 21 Grid of density of population in 1990

Figure 22 Grid of density of population in 2000

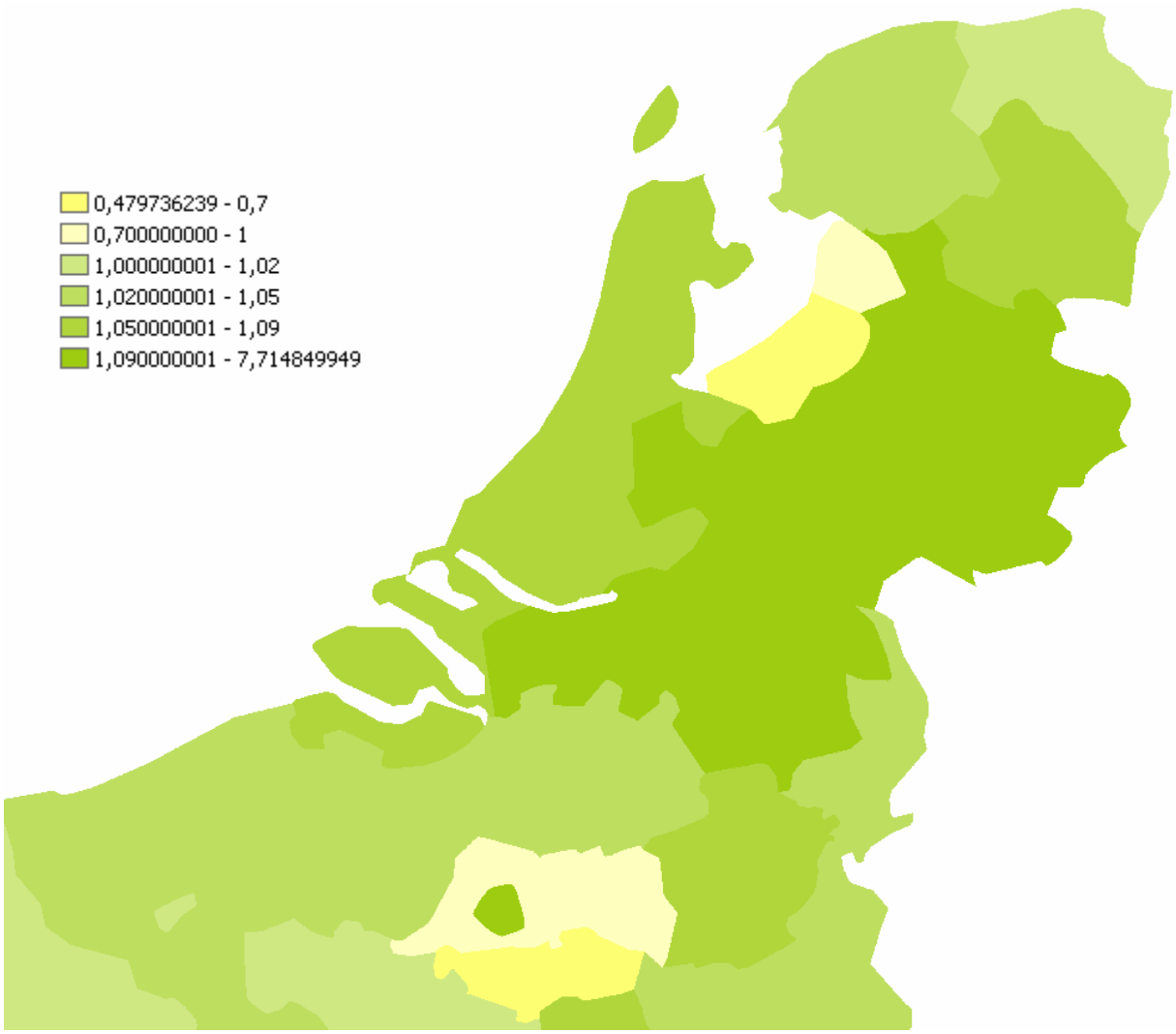


Figure 23 Evolution of the population of 1990 to 2000 in a 100*100m sized grid

First step: intersection of areas at the two dates = «smallest areas methodology"

Between two dates, it can be possible to land differences of spatial carvings to estimate a population. We create an new geographical unit : what we called the smallest common part. (obtained by a simple intersection). It is why it is indispensable:

- to have a good superposition of the two layers
- to have a similar generalization.

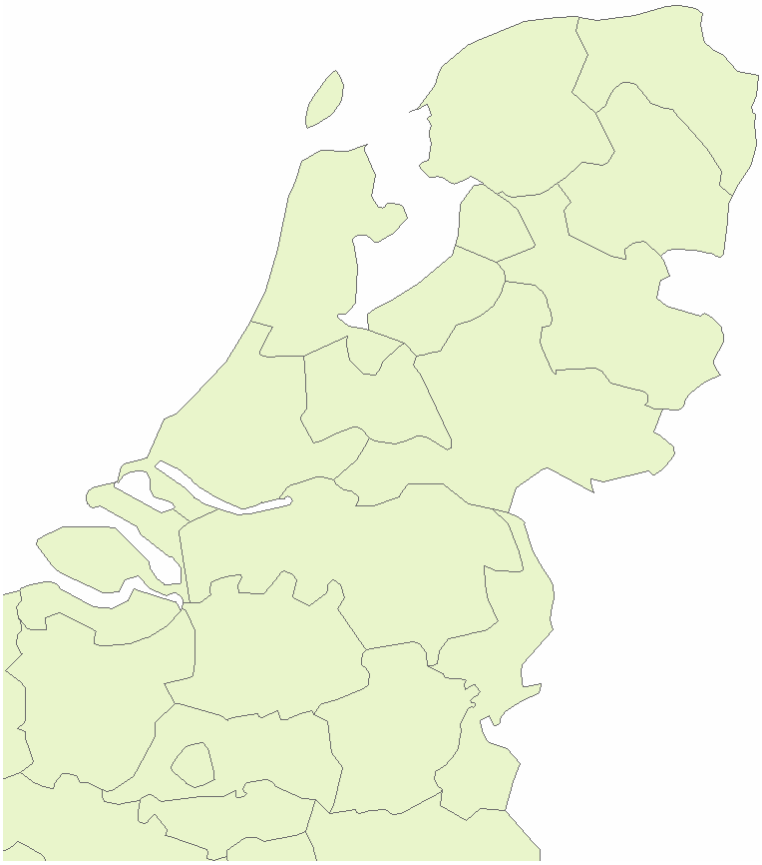


Figure 24 Obtained "smallest zones"

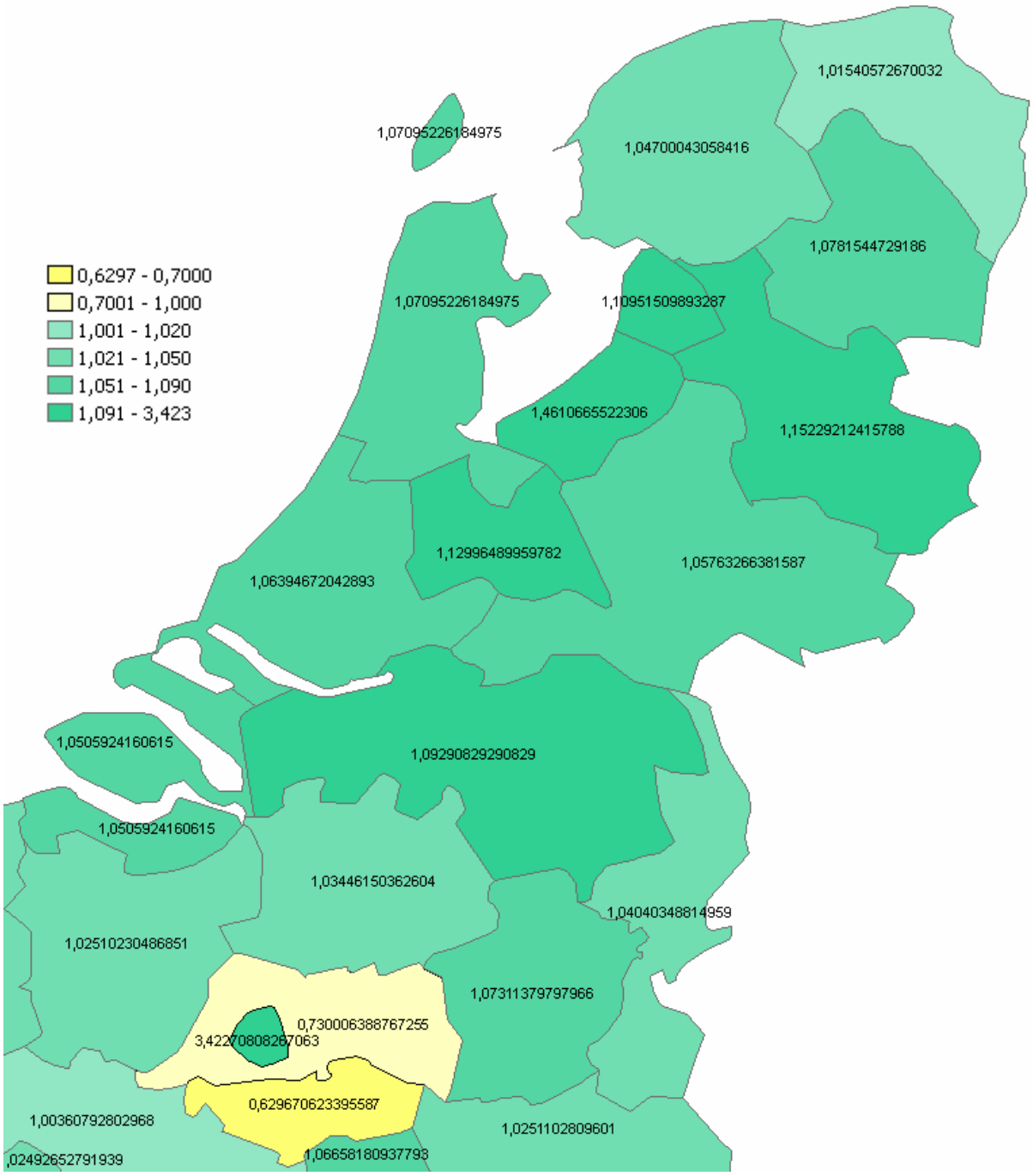
Second step: Numbering of density pixels (hypothesis of regular population distribution)

We get smallest areas/ We know their administrative unit belonging for both date 1 and date 2. . We comptons pixels of densities of population is sufficient (previously computed).

Every «smallest area», is described therefore by its mean density tfor the both dates.

1990 *2000*

Map 14 Smallest areas and grids of density of population



Map 15 Evolution of the density of population in the smallest zones

The evolution of populations, in this example 1990-2000, is vraissemblable: however, areas with very high density are certainly badly described (cf.construction of density grids).

reagregating «smallest areas" as original wider NUTS for the date 1 or date 2, it is then always possible to recover the evolution of population in one of the two initial carvings.

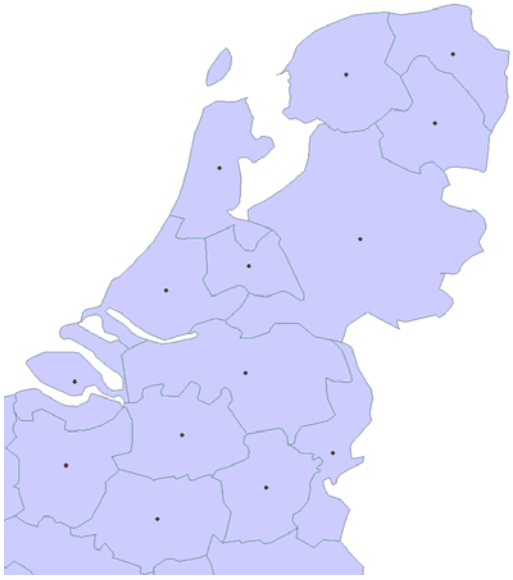


Figure 25 Intersections and their centroids

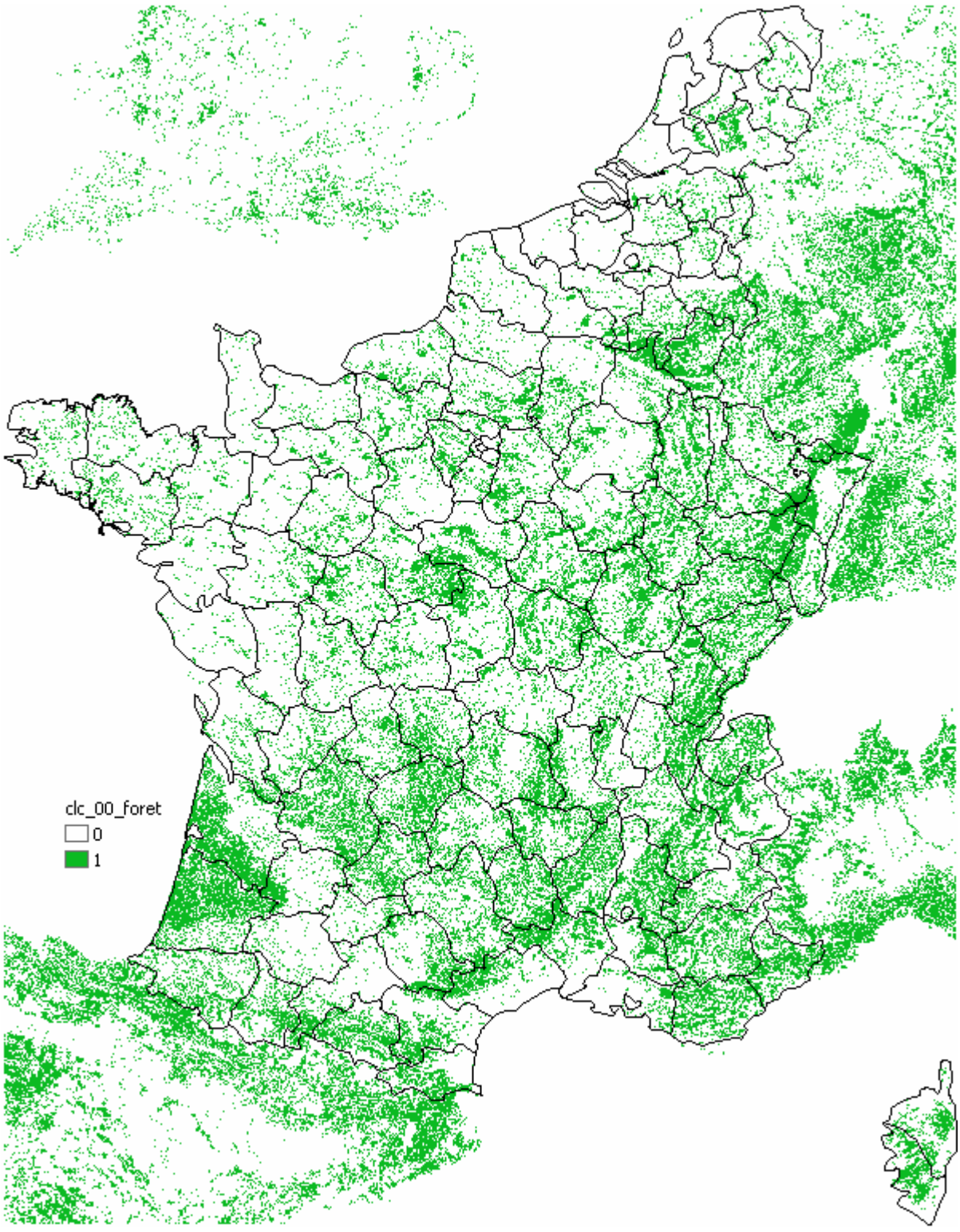
4.3.4.2 Multisource data integration: soial, economic and envinonmental data

a) Surface of forest per capita

The surface of forest per capita is an iexcellent ndicator to know the "good life framework" of the Europeans. A first approach consists in assigning to every NUTS , the part of its surface occupied by the forest.

This work is trivial and can be described easily:

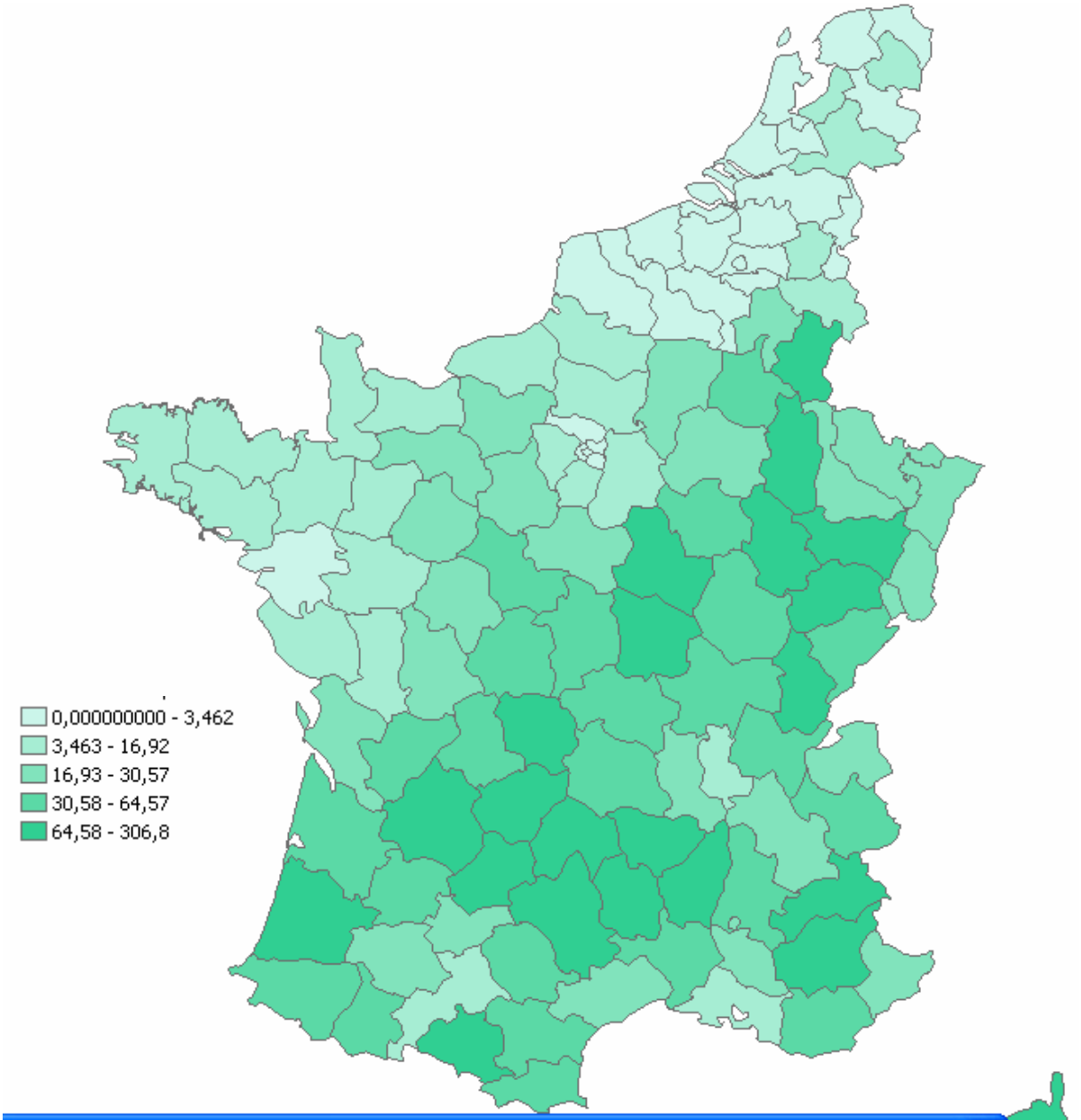
- to extract themes of forest of the grid Corine Land Cover (EEA)
- to generate a binary grid (0 or 1). (for example, spatial resolution of one hectare)



Map 16 NUTS 23 1999 and CLC 2000 (type = forest)

- To calculate the number of forest pixel for each NUTS23_99

- To establish the ratio forest surface per capita



Attributes of NUTS23_1999_E

NUTS_2	REGION	AREA	P00	CLCF00new
BE1	R,gion de Bruxelles	183719434,2	961900	210000
BE21	Prov. Antwerpen	2834223660,	1644700	3570000
BE22	Prov. Limburg (B)	2440749787,	793000	3740000
BE23	Prov. Oost-Vlaander	3087009031,	1362700	750000
BE24	Prov. Vlaams Braban	2026130832,	1016600	1540000

Map 17 Surface of forest per capita NUTS 2-3 1999 in 2000 (France)

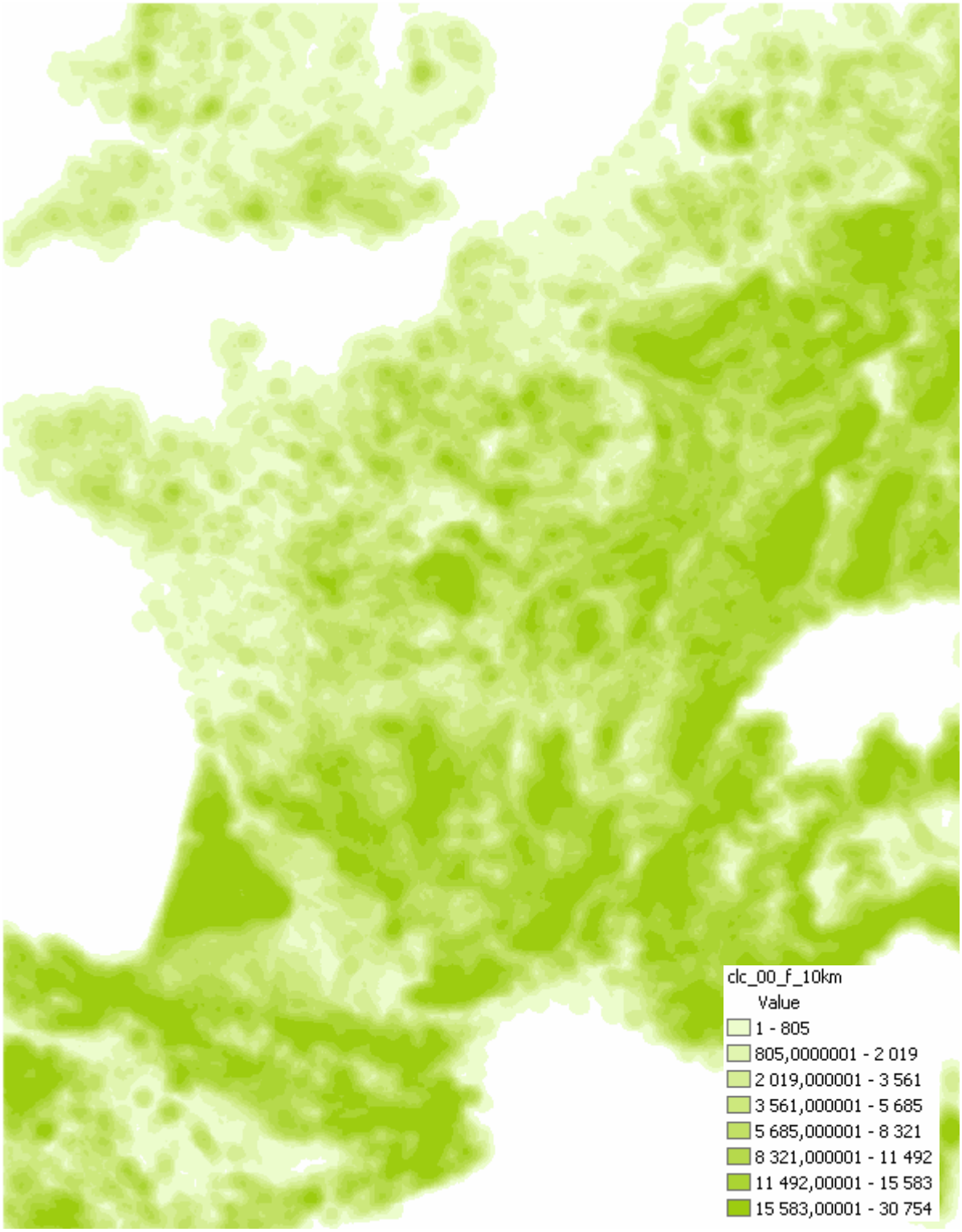
b) Potential of Surface of forest per capita within a 10 km radius

Another way to look at the state of the setting of life of the European is to estimate the availability of parklands (forests, here) close by the places of residence. In order to include concept of accessibility, we chose a ray of 10 km. It corresponds to a short access time by car (only some minutes) or a medium by-feet trajet about twenty minutes long (interesting with children).

This potential of available forest is calculated simply:

- the destination grid corresponds to a seedling of points (seedling of the Corine Land Cover pixel centers, for example)
- we estimate, for all point of the space, the number of forest pixels in the 10 km with a rectangular interaction spatial function. The shape of the function means that we fixe a maximal distance au delà de laquelle elle n'est plus acceptable. The question is: what is the maximum distance that the urban resident is able to browse in order to surrender in a place a few more close to the nature ?

This indicator, moreover simple, already permits to note important regional disparities with neighborhood forests.



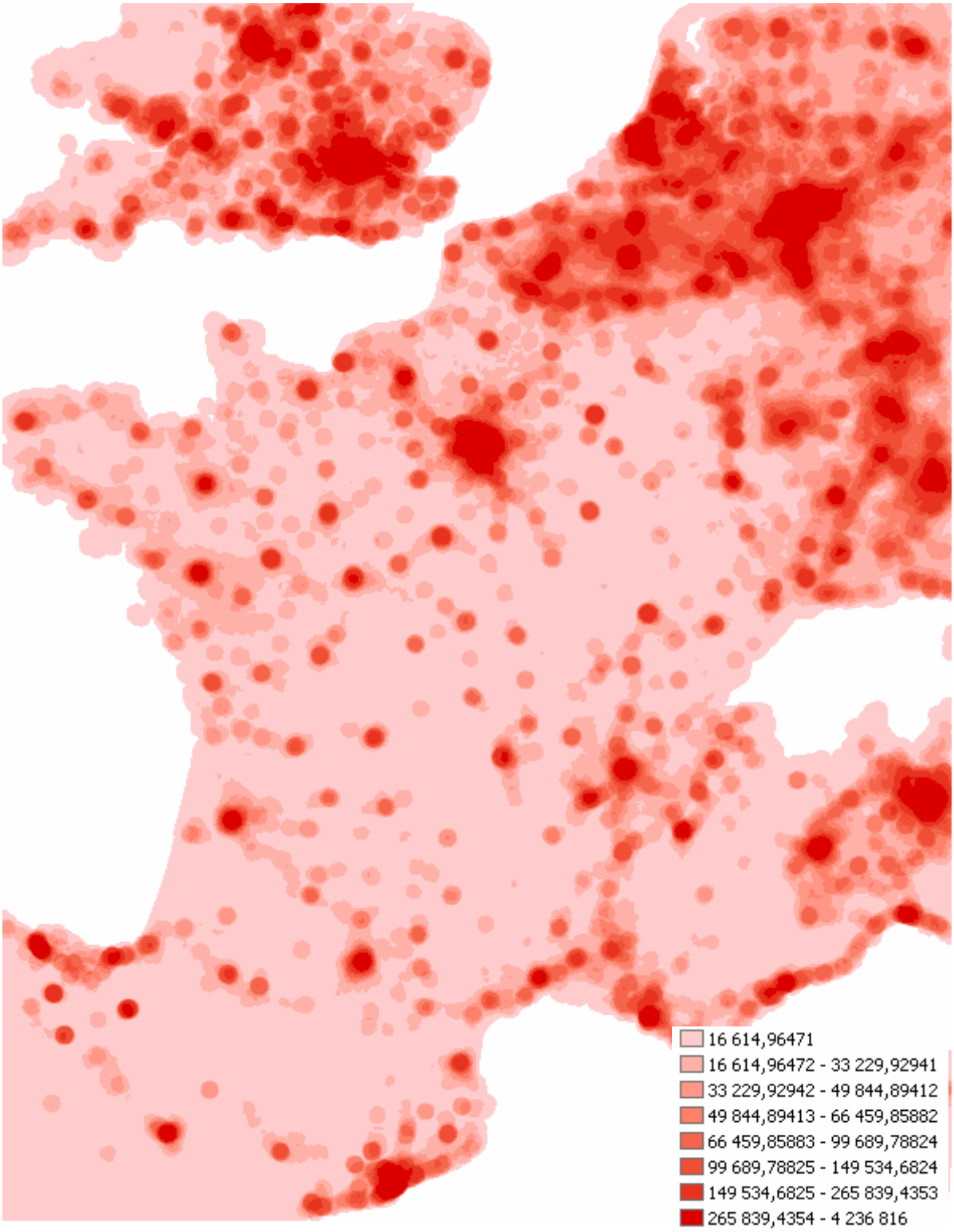
Map 18 Surface of forest in a 10 km radius (in hectares)

The "available forest" concept is a concept of social utility. The role of the forest is multiple (production, protection, recreation,…). But it is known

that the proximity of great units of human settlments encourages the function of public welcoming . Also, we estimated important to bring back this potential stock of wooded surfaces to a stock of population that could fréquenter the wooded spaces.

So, the potential of population must be estimated in the same conditions: radius, shape of the function, spatial resolution,...

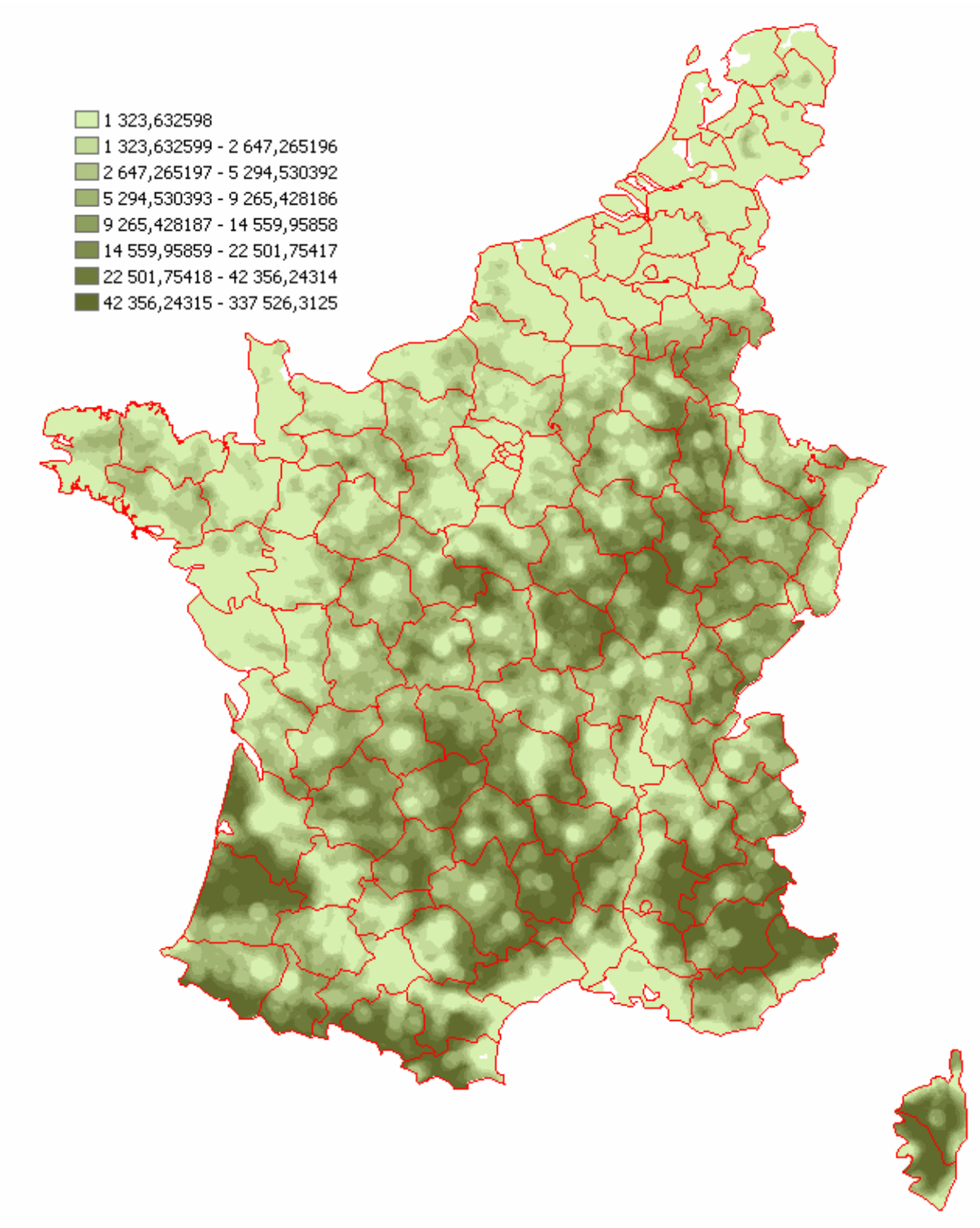
Therefore, we calculated the potential of population with a grid of one hectare cells and with a ten kilometers radius. The originla population grid was calculated for 2000.



Map 19 Population in a 10 km radius (in hectares)

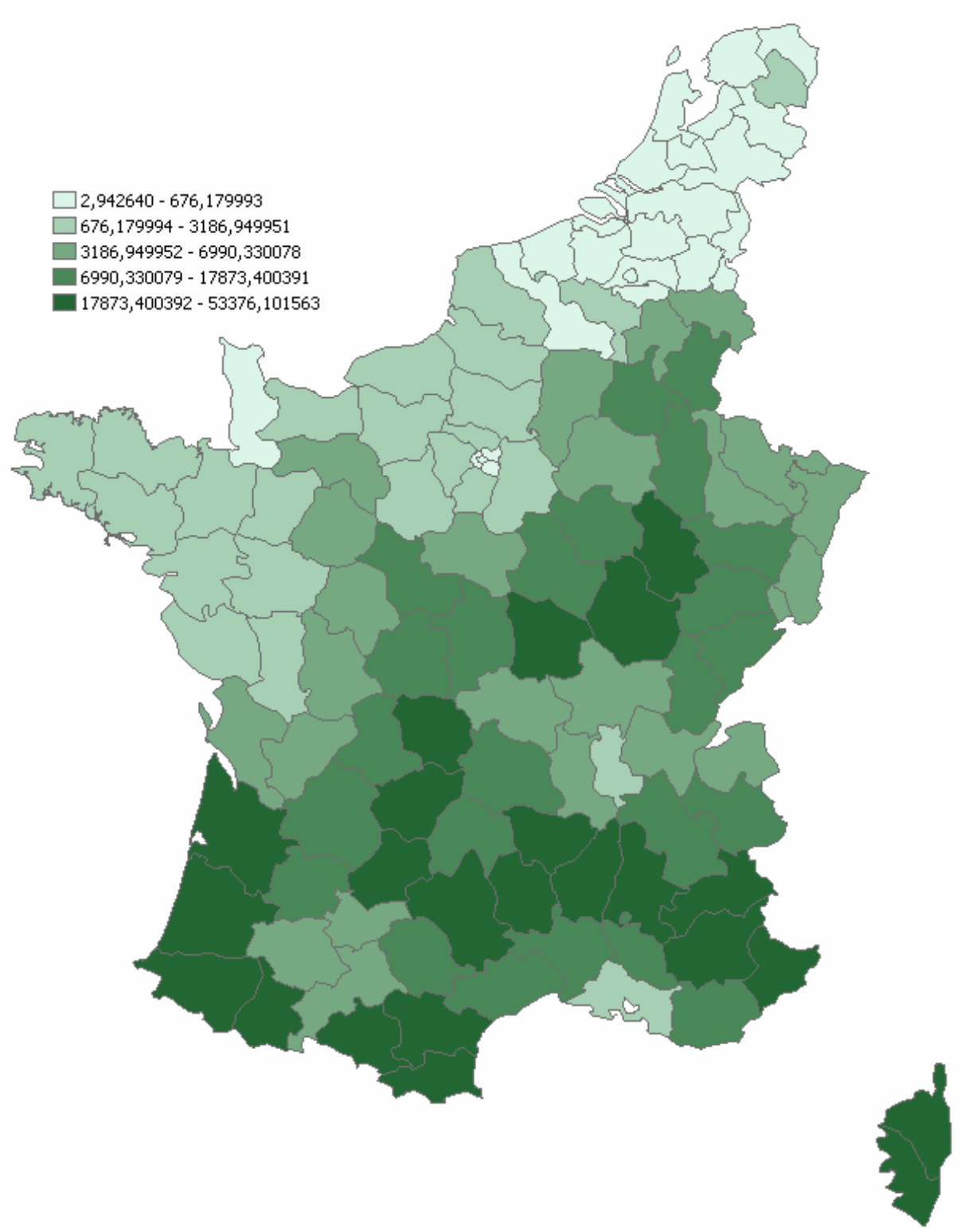
Lastly, the ratio to be obtained is an algebra grid processes: grid of potential of forest / grid of potential of population in 2000. In fine, we have an evaluation of the available forest surface per capita in the 10 km

in all point of the space. (independently of the survey of transportation networks that would refine this notion of access to the recreational forest)



Map 20 Surface of forest per capita in the 10 km (in m²)

While drifting this original information, we can enrich the description of the NUTS 1999 with the average of surfaces per capita in the 10 km.



Map 21 Average of forest surfaces per capita in the 10 km for NUTS 23 1999 (in [m²/hab])

This set of calculation is relatively simple but very gluttonous in computer dependances (time). Despite usingof an optimized software for this task, the sum of pixels in the ten kilometers for a "only" little portion of Europe (rectangle including the three studied countries) takes a considerable time (more than 24 hours).

4.3.5 CONCLUSION

This study shows well the assets, limits and constraints of the use of existing databases in space problems requiring a multi-source integration. The encountered problems are several orders:

- availability and coherence of the geometrical funds in the time (historical of the funds often absent, dates available reduced, unknown sources,…
- design and geometry of the boundaries maps: level of generalization very heterogeneous (cf. Norwegian coast and Normandy coasts or French nuts and British counties), cartographic projections multiple,
- update of the bases sometimes difficult to follow : the collections of data by the Member States aren't synchronized , the level used for the spatial restitution of indicators is the NUTS level in force at the time of the consultation and data are updated progressively in the time without history of the corrections, the un-updated data aren't stored for consultation.
- level of harmonization : the data are often collected within states boundaries and significant differences are shown in themes definition or in collection in across border areas (like land use for example).

The general problems met during these tests can be thus resumed:

1. Problems of maps compatibility

The incompatibility of the European maps are strongly related on:

- changes of indentificators: different table joints at the different dates is then impossible without expensive manipulation

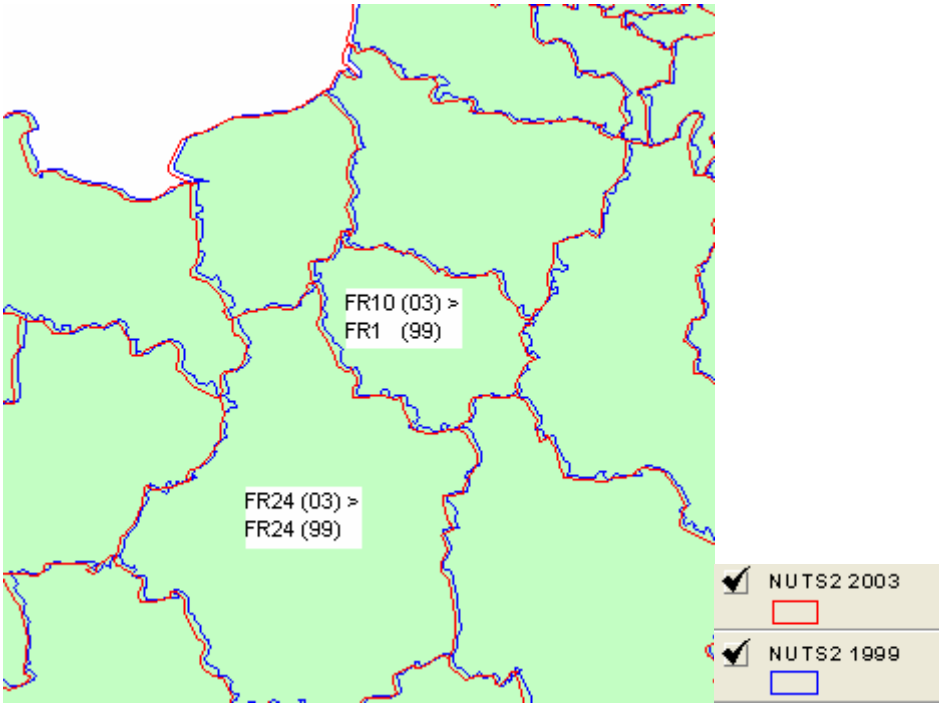


Figure 26 Code changes examples : NUTS 2 1999 and NUTS 2 2003

- simple changes of geometries: we can say these changes of geometries are often due to problems of generalization

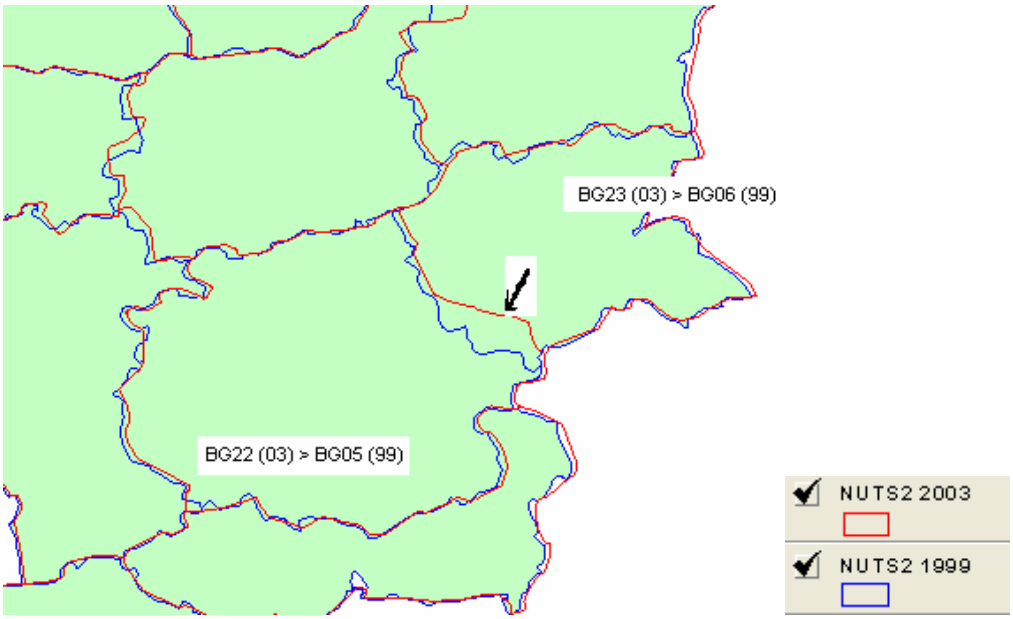


Figure 27 Geometry changes examples exemaples : NUTS 2 1999 and NUTS 2 2003

The only solution, in such a case, is to transform codes of these Bulgarian provinces and to displace the border between the two units ()

- Changes of geometry more hard: changes of limits between 99 and 03 are enough numerous and **complex.** (example of Portugal). To remedy this problem, there doesn't have any satisfactory solution because it would require:
    - to modify borders arbitrarily,
    - to create some new units,
    - to give a new name to units (compatibility of the identifying)
    - to proceed by hand (weight important of the cartographer's interpretation)

      However, it is relatively comfortable to mark all similar cases.

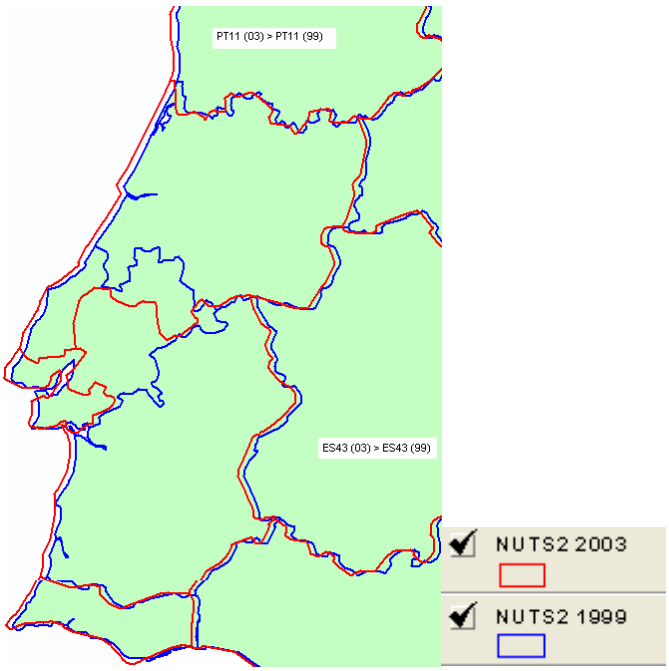


Figure 28 Geometry changes examples exemaples : NUTS 2 1999 and NUTS 2 2003 (Portugal)

If we could have finest levels, we could probably reconstitute easly (by aggregation) of the wider levels (NUTS 2 or NUTS 3) between two dates. The strong, and true hypothesis in the zone test, is that levels of carvings the finest are steadyer in the time. We tested it:

- for the NUTS 2S of 1999 for all the Europe,
- for the NUTS 23 of 1999, 1988 and 1980 for France, Belgium and the Netherlands

Results are encouraging. It is necessary in this case:

- to have well referenced vectorial units.
- To know precisely the origin of maps circulating (what is not apparently case for the 2003's NUTS whose generalization is so different with the one of the other years).

2. Environmental data: Corine Land Cover

For data of the EEA a certain number of conclusions can be expressed :

- Without the specific collaboration we benefitted, it could be very difficult to get a complete set of data on Europe. An authorization of countries concerned is often necessary and can take a lot of times.

- Harmonization of the CLC layers for all the european states: a seamless database has been build by harmonization of state level layers. The buffer zone of every state covered by the project Corine Land Cover has been suppressed and the adjustings of leaves harmonized in continuous on the European space. The contact zone between France and Belgium is a good example to indurstand the way followed to harmonize layers. A first harmonization is made to the national levels. In fact, the correspondence between vector and raster

is perfected within each country, but it is not assured beyond datacollection's units.

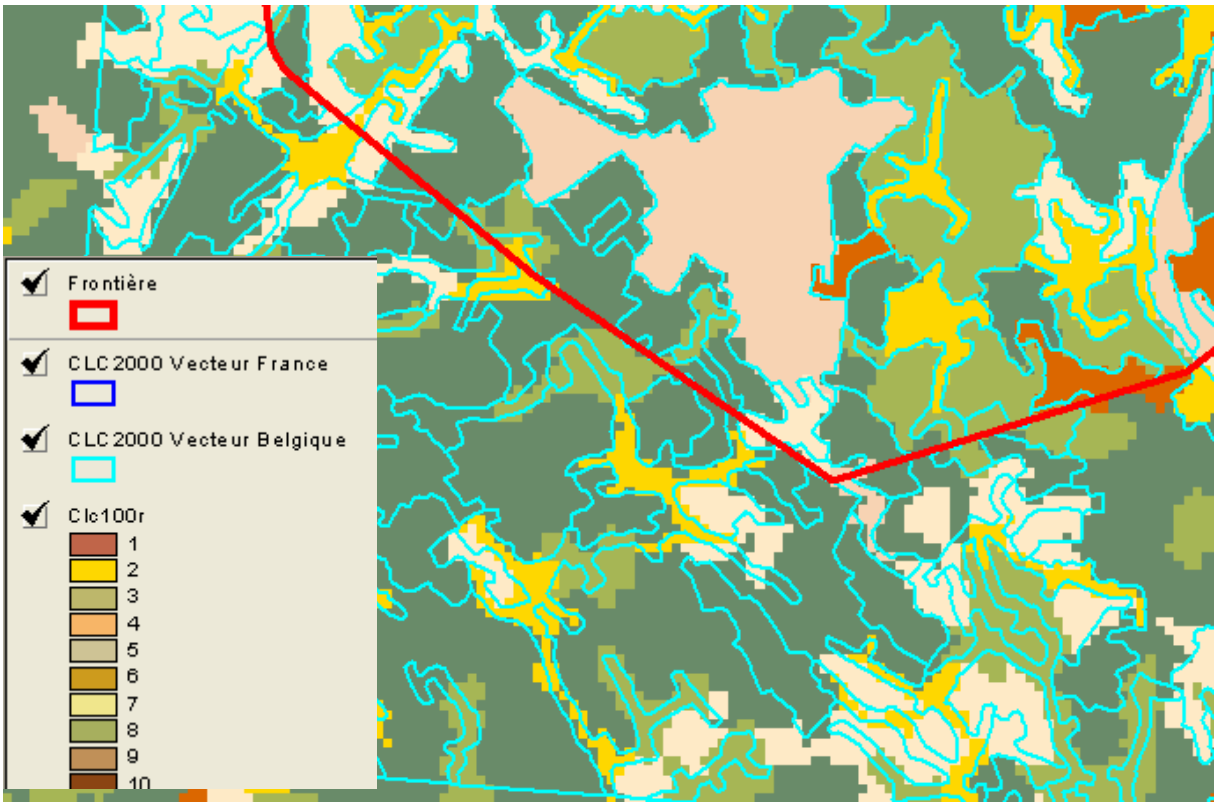


Figure 29 Vector and Raster overlays (Belgium)

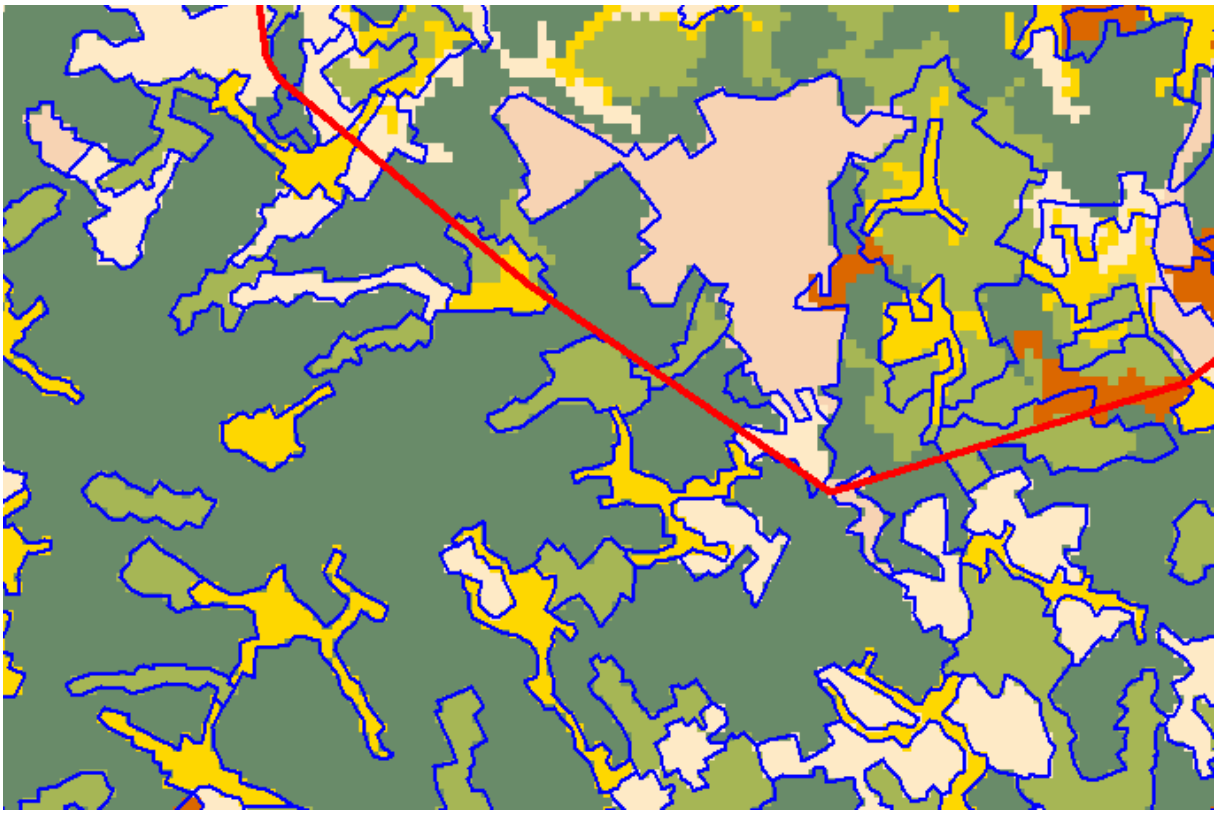


Figure 30 Vector and Raster overlays (France)

3. Données EUROSTAT :

The nature of problems met is related on the mission confided to EUROSTAT : the European statistical data harmonization. EUROSTAT harmonize data but don't recolte directly the finest raw national data. So, few important problems appears:

In principle all levels of NUTS should be accessible. But in facts we can acceed to data especially for the NUTS 2 (level that is frequently the finest existing in the EUROSTAT database).

To study some temporal sets in an evolutionary geographical grid, it would be necessary to have data of origin in their own original spatial unit. However, the territorial division given by EUROSTAT to restore sets of indicators is always dated with the last updating. (territorial division NUTS 2003 at the time of the writing of this report).

The ideal to treat the evolution in the time would be to have the raw data (grids as well as statistical sets) and in the finest scales, before harmonization,

The harmonized data are very important and rich but there is not conservation of the historic of data.

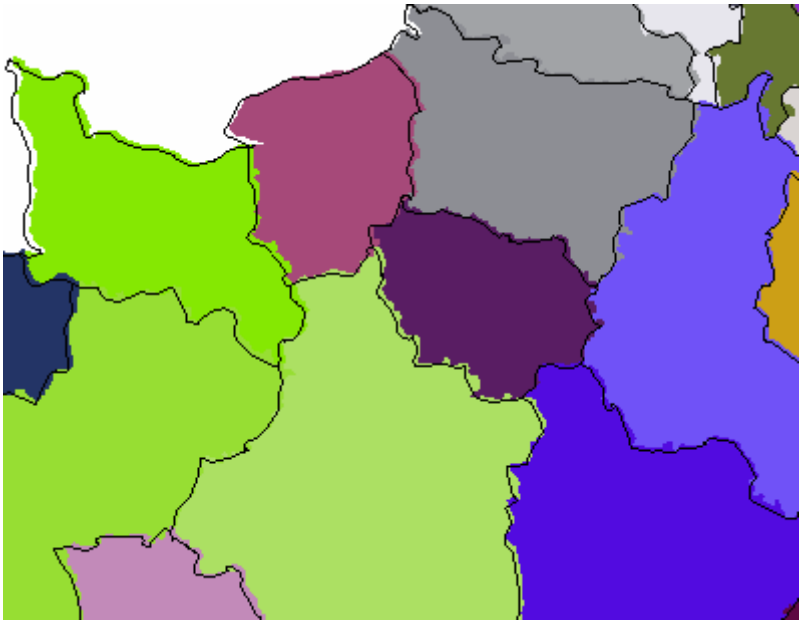
Spatio temporal hiatuses: it is very difficult to obtain some complete information for a wide area (ESPON space for example) for one date or, worse, for a temporal set.

4.3.6 Methodologic appendiceses : solution to make maps compatible ?

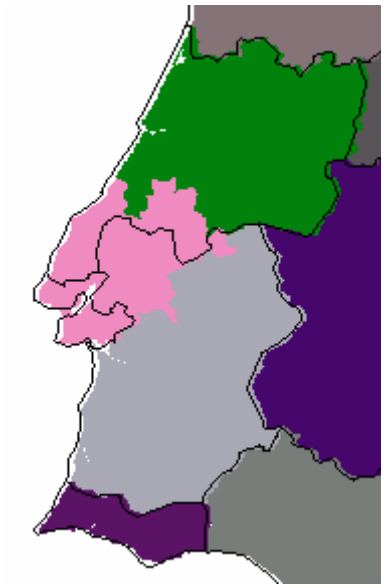
4.3.6.1 Obtaining a same identifier of the NUTS units for two dates (solution 1) : raster solution

As we saw the problem of maps is a major problem. Insofar as the 2003 NUTS map is correctly superposable with the CLC data , we can try to drift maps of it to the other (codes and geometries): that means to adapt 2003 map to transform them in 1999, 1988, 1980's one

1. Identification : to get map in 1999 or 1988 from 2003's one, it is necessary to preserve codes of origin and to recreate the geometries of the NUTS modified, created suppressed either.
2. To adjust codes between the several maps, we use raster spatial representation to recover codes of spatial units:
    - To transform first map (NUTS 1999 for example) in a grid for which pixels carry the code of the NUTS of which it is descended.
    - To recover for every NUTS for the second date (NUTS 2003 for example), the modal value of pixels found for the date 1 (code NUTS 2 1999).



This method permits to adjust problems of code change enough easily. It is however preferable to verify every NUTS, because cases of limit changes are not detected:



4.3.6.2 Recode the NUTS units for two dates (solution 2) : approximation by calculation of centroides

The method is simple (for the geographical relatively compact units) :

- calculation of centroïdeses for the NUTS2 1999
- affectation of the code NUTS 99 to the NUTS 2003 (that contain them).

This method adjusts correctly the case of code changes , but doesn't permit to mark problems of limit changes.

4.3.6.3 Limits changes detection by construction of a surface differential indicator

This indicator is simple : we compare simply areas of the NUTS2 1999 with NUTS2 2003 areas. If differences are weak (relatively), we can estimate that there were not modifications of limits. If differences are strong it is necessary to study the zone precisely and to do possible modifications.

For example we selects here (in blue on the map) the part of NUTS2 2003 whose surface varies the more (gain or loss).

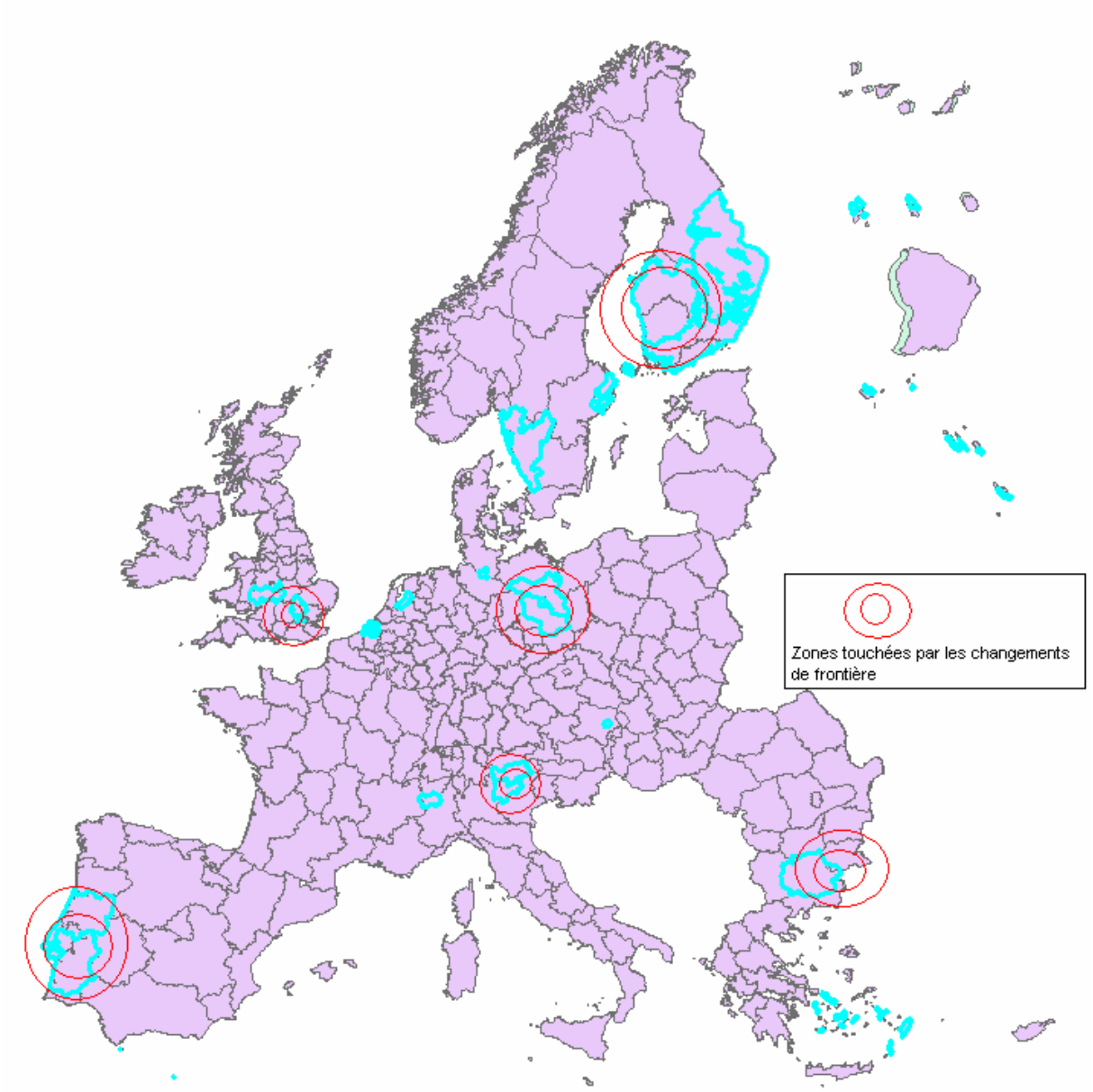


Figure 31 Advantages and limits of a surface indicator

- All zones of modifications are marked as well as zones where differences of generalization are very important.
- Once zones of border modification marked automatically, changes will benecessary manual
- The doorstep (%) to evaluate negligible changes is very strongly retlated ion user : to build a robust method, it's necessary to build a good method of assessment of the mistake

4.3.7 Appendiceses Illustrations

4.3.7.1 Problems of map generalization in Brittany 1999 and CLC 2000 superposition

These maps are globally distinctly more generalized (see figure 31 and 32) with the exception of certain inshore zones.

Figure 32 NUTS2 2003 et CLC1990 - Bretagne

Figure 33 NUTS3 2003 and CLC1990–Paris and small crown

4.3.7.2 Problems of data superposition (IDF, NUTS 3 99 and CLC 90)

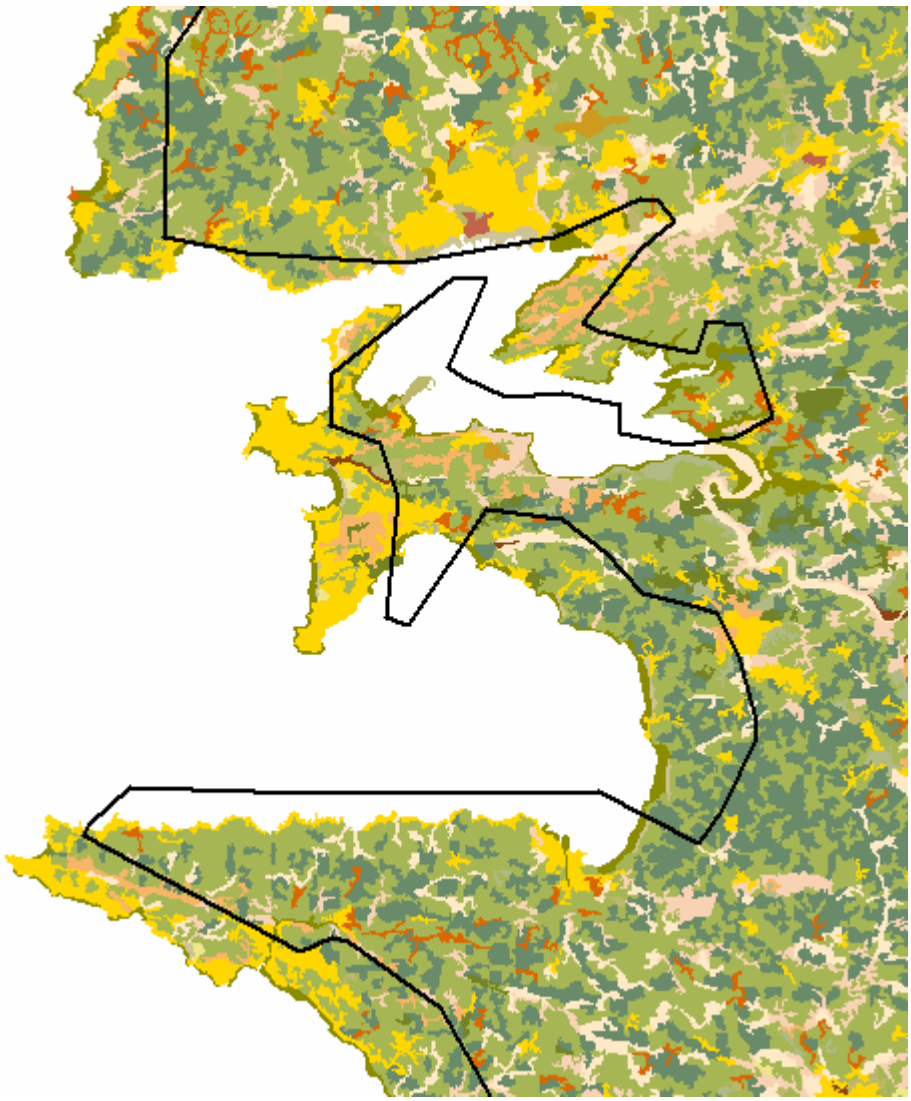


Figure 34 NUTS2 1999 and CLC1990 - Bretagne

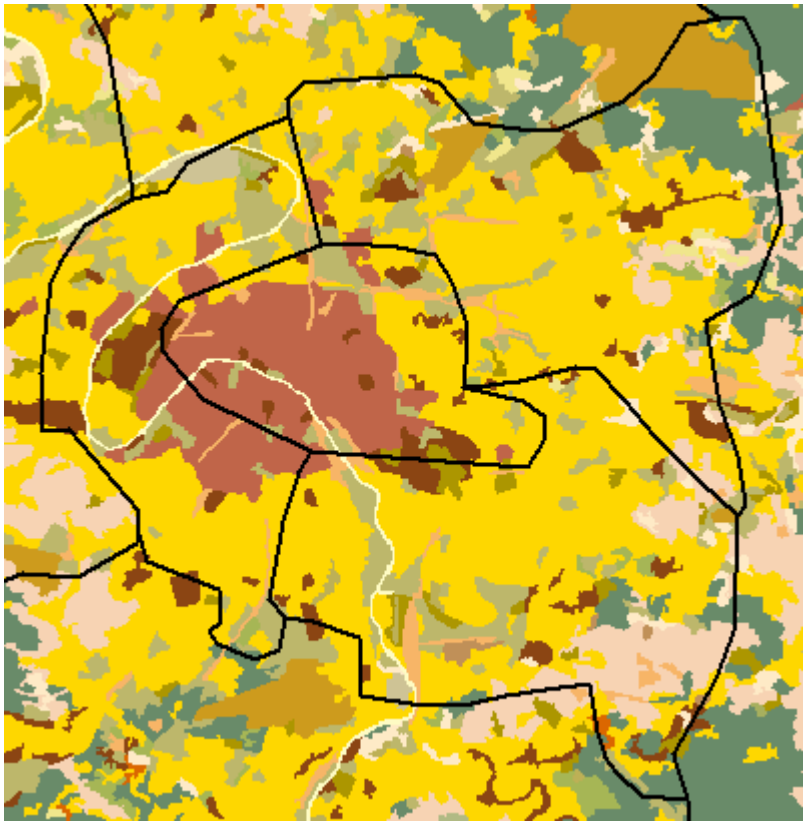


Figure 35 NUTS3 1999 and CLC1990–Paris and small crown