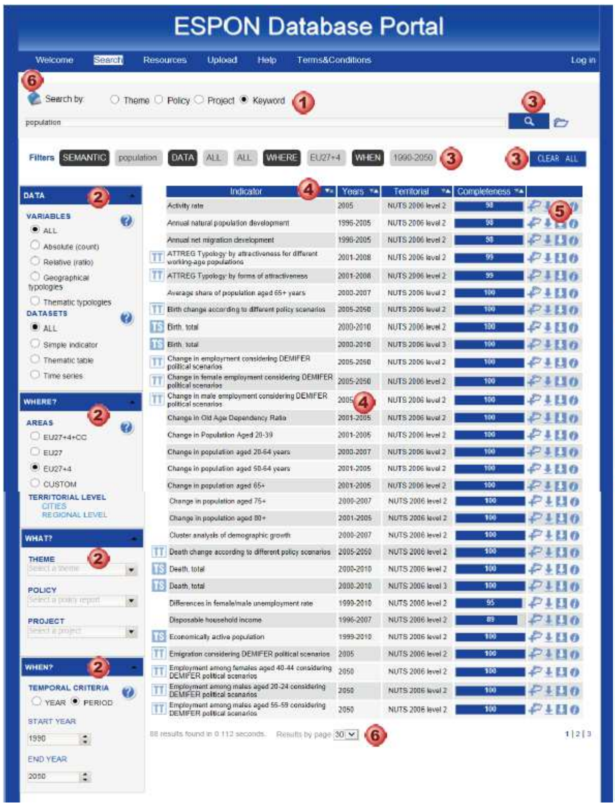
ESPON

ESPON M4D
Multi-dimensional Data Design and Development

SECOND INTERIM REPORT

28th June 2013



Search by: ○ Theme ○ Policy ○ Project ● Keyword

population

Filters: SEMANTIC population | DATA ALL ALL | WHERE EU27+4 | WHEN 1990-2050 | CLEAR ALL

DATA – VARIABLES: ● ALL ○ Absolute (count) ○ Relative (ratio) ○ Geographical typologies ○ Thematic typologies

DATASETS: ● ALL ○ Simple indicator ○ Thematic table ○ Time series

WHERE? – AREAS: ○ EU27+4+CC ○ EU27 ● EU27+4 ○ CUSTOM

TERRITORIAL LEVEL: CITIES, REGIONAL LEVEL

WHAT? – THEME: Select a theme; POLICY: Select a policy report; PROJECT: Select a project

WHEN? – TEMPORAL CRITERIA: ○ YEAR ● PERIOD; START YEAR: 1990; END YEAR: 2050

	Indicator	Years	Territorial	Completeness
	Activity rate	2005	NUTS 2006 level 2	98
	Annual natural population development	1995-2005	NUTS 2006 level 2	98
	Annual net migration development	1996-2005	NUTS 2006 level 2	98
TT	ATTREG Typology by attractiveness for different working-age populations	2001-2008	NUTS 2006 level 2	99
TT	ATTREG Typology by forms of attractiveness	2001-2008	NUTS 2006 level 2	99
	Average share of population aged 65+ years	2000-2007	NUTS 2006 level 2	100
TT	Birth change according to different policy scenarios	2005-2050	NUTS 2006 level 2	100
TS	Birth, total	2000-2010	NUTS 2006 level 2	100
TS	Birth, total	2000-2010	NUTS 2006 level 3	100
TT	Change in employment considering DEMIFER political scenarios	2005-2050	NUTS 2006 level 2	100
TT	Change in female employment considering DEMIFER political scenarios	2005-2050	NUTS 2006 level 2	100
TT	Change in male employment considering DEMIFER political scenarios	2005	NUTS 2006 level 2	100
	Change in Old Age Dependancy Ratio	2001-2005	NUTS 2006 level 2	100
	Change in Population Aged 20-39	2001-2005	NUTS 2006 level 2	100
	Change in population aged 20-64 years	2000-2007	NUTS 2006 level 2	100
	Change in population aged 50-64 years	2001-2005	NUTS 2006 level 2	100
	Change in population aged 65+	2001-2005	NUTS 2006 level 2	100
	Change in population aged 75+	2000-2007	NUTS 2006 level 2	100
	Change in population aged 80+	2001-2005	NUTS 2006 level 2	100
	Cluster analysis of demographic growth	2000-2007	NUTS 2006 level 2	100
TT	Death change according to different policy scenarios	2005-2050	NUTS 2006 level 2	100
TS	Death, total	2000-2010	NUTS 2006 level 2	100
TS	Death, total	2000-2010	NUTS 2006 level 3	100
	Differences in female/male unemployment rate	1999-2010	NUTS 2006 level 2	95
	Disposable household income	1996-2007	NUTS 2006 level 2	89
TS	Economically active population	1999-2010	NUTS 2006 level 2	100
TT	Emigration considering DEMIFER political scenarios	2005	NUTS 2006 level 2	100
TT	Employment among females aged 40-44 considering DEMIFER political scenarios	2050	NUTS 2006 level 2	100
TT	Employment among males aged 20-24 considering DEMIFER political scenarios	2050	NUTS 2006 level 2	100
TT	Employment among males aged 55-59 considering DEMIFER political scenarios	2050	NUTS 2006 level 2	100

88 results found in 0.112 seconds. Results by page 30 | 1 | 2 | 3

The ESPON Database portal offers the possibility to search among hundreds of indicators produced by ESPON Projects.

The search query interface allows policy makers or scientists to quickly find indicators for various areas (EU27 + Switzerland, Liechtenstein, Iceland and Norway) at various geographical levels (NUTS or city levels).

The aim of this section is to explain step by step the way to find indicators contained in the multidimensional ESPON database.

1 Search by theme, policy, project or keyword

The first step consists in selecting a semantic search criterion. It may be a theme (population and living conditions, education, governance, territorial structure etc.), a EU Cohesion Policy Document (EU2020, Territorial Agenda...), an ESPON project which has delivered the indicator or a keyword. During that step, user is helped with dropdown lists.

2 Filter the results

The indicators contained in the ESPON Database are both numerous and heterogeneous: available at various geographical levels, time period and with heterogeneous statistical types. In order to comply both with policy makers and scientists needs, various filter options are now operational.

- Filter *Data* to query by statistical properties the indicator (typology, count data etc.) or dataset type: simple indicator, thematic table or time-series.
- Filter *Where* to select specific study areas (EU27, ESPON Area, a given country). Please note that the completeness calculation depends on the study area chosen. It is also possible to choose a specific territorial level (NUTS 0, NUTS1, NUTS2 or NUTS3 for regional data and UMZ, FUA or MUA for cities).
- Filter *What* to add other semantic criteria. For instance, if a user has selected a thematic criterion, he may add a policy or a project criterion.
- Filter *When* to filter the results for a given year or time-period.

3 Launch the query

Taking into account the search criteria, the query is performed by clicking on the launch search button (🔍). All the query parameters chosen are displayed in the *filters toolbar*. At every moment, the user can choose to refine her/his query by adding or deleting some filters.

By clicking on the *clear all* button, all the filter parameters are refreshed.

What is a thematic table and a time-series?

TT A thematic table is considered as a set of indicators which are semantically linked. It means that it is impossible to download them separately. In practical terms, it relates to two kind of figures: contingency table (the sum of the columns makes sense, (e.g. age structure, employment by branch etc.) or indicators which are semantically linked (e.g. a typology and its related raw data, indicators contained in the 5th cohesion report).

TS A time-series is considered as an indicator available at different temporal extents (5 years at least) and sharing the principles of the core database strategy, e.g.:

- Most of the missing values have been estimated
- Each time-series has been statistically checked (outlier check)
- It is possible to easily update the indicator in the future.
- The statistical type of time-series indicators is necessarily absolute stocks or absolute flows (excepted data on temperatures).

4 Search result output

The search result table contains five columns.

- The first one relates to the *dataset type of the indicator* (TT for thematic table, and TS for time-series).
- Then, the *name, the year of reference and the territorial level* of the indicator are displayed.
- The last column reveals the *degree of completeness* of the indicator, calculated as regard to the study area chosen (ESPON Area by default).

Results are alphabetically sorted by default (indicator name). But they may also be sorted by year of reference (oldest or latest temporal extent available), by territorial level or by degree of completeness.

5 Download options

Three download options are available:

- *Basket* - Put the selected indicators into a coherent .zip file.
- *Download* - Download directly the indicator in a .xls file.
- *Metadata* - Open the metadata page of the indicator. It gives information related to the data provider, the methodology behind the indicator, the data sources etc. This metadata page can be exported in a .pdf file.

6 Additional options

The overview option - icon - displays key figures (number of indicators, number of themes etc.) and the alphabetical list of all the integrated indicators.This list can be exported in .pdf format.

The number of results displayed by page is 10 by default, but can also be 5, 20 or 30.

We hope that every users will enjoy these new functionalities and will retrieve the ESPON data he/she is looking for!
We are waiting for your suggestions and feedbacks..

To consult the search interface of the ESPON Database Portal:
http://database.espon.eu/db2/search

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

This second interim report represents the first results of a research project conducted within the framework of the ESPON 2013 programme, partly financed through the INTERREG III ESPON 2013 programme.

The partnership behind the ESPON Programme consists of the EU Commission and the Member States of the EU25, plus Norway, Switzerland, Iceland and Liechtenstein. 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.eu

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

ISBN number:
This basic report exists only in an electronic version.
Word version: 2013

List of contributors to the FIR

UMS RIATE (FR)

Claude Grasland

Ronan Ysebaert

Isabelle Salmon

Nicolas Lambert

Timothée Giraud

NCG (IE)

Martin Charlton

Alberto Caimo

LIG-IMAG (FR)

Jérôme Gensel

Benoit Le Rubrus

Camille Bernard

Laurent Poulenard

Marlène Villanova-Oliver

UAB (ES)

Roger Milego

César Martinez

Maria-José Ramos

UMR Géographie-cités (FR)

Anne Bretagnolle

Antonin Pavard

Hélène Mathian

Marianne Guérois

TIGRIS (RO)

Octavian Groza

Alexandru Rusu

Daniel Tudora

Aurelian Nicolae Roman

TABLE OF CONTENT

List of figures

Introduction

The Seasons of the ESPON Database

With the second interim report of the ESPON project Multidimensional Database Design and Development (M4D), we arrive at a turning point in the lifecycle of the project as well as the ESPON 2013 program. If we use the metaphor of the season to describe this lifecycle, we can say that the transition between the first and the second ESPON program (2007-2008) correspond to an Autumn-Winter period, the first ESPON database project (2008-2011) to a Spring-Summer Period and the current ESPON M4D project to (2011-2013) to a Summer-Autumn period, before the return of Winter that will correspond to the interruption of activities between ESPON II and ESPON III period (Figure 1)

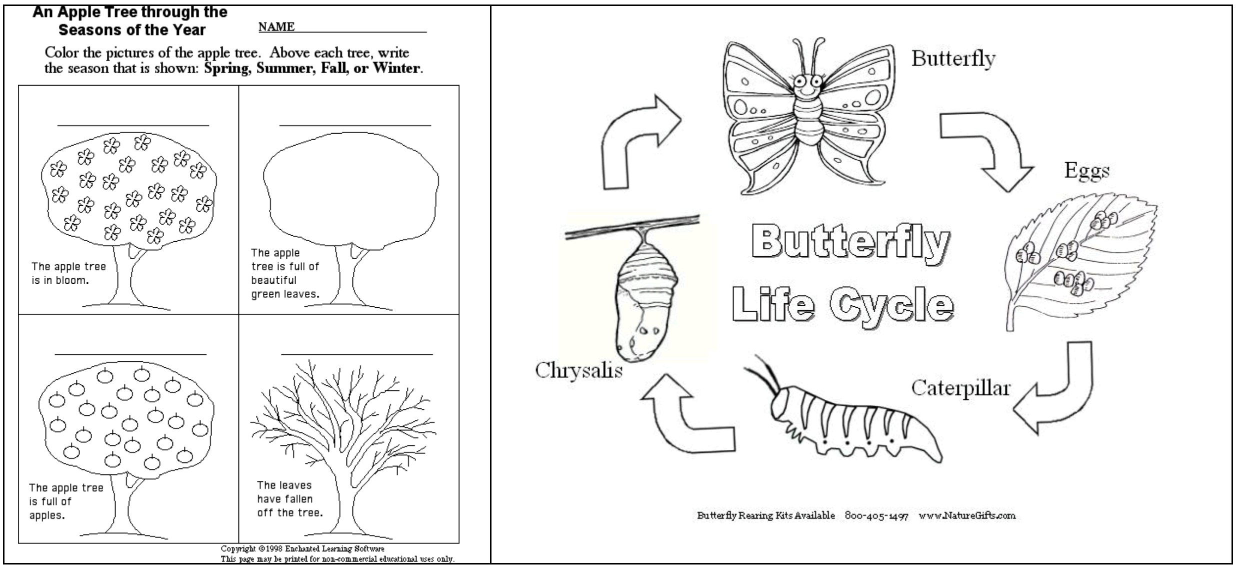

An Apple Tree through the Seasons of the Year
NAME
Color the pictures of the apple tree. Above each tree, write the season that is shown: Spring, Summer, Fall, or Winter.
The apple tree is in bloom.
The apple tree is full of beautiful green leaves.
The apple tree is full of apples.
The leaves have fallen off the tree.
Copyright ©1998 Enchanted Learning Software
This page may be printed for non-commercial educational uses only.
Butterfly
Eggs
Butterfly Life Cycle
Chrysalis
Caterpillar
Butterfly Rearing Kits Available 800-405-1497 www.NatureGifts.com

Figure 1 - A metaphor of the lifecycle of database in ESPON programming period

Let us clarify briefly the meaning of this metaphor of seasons, in terms of data collection in the ESPON program:

Wintertime is a moment of interruption of activities in terms of data collection. In the case of the ESPON program, this interruption of database activity is related to the discontinuities of funding that was observed in 2007-2009 and will come back again in 2013-2015. The fact that the responsibility of the database is allocated by tendering procedure to TPG (projects 3.1 and 3.2 in the first programming period; projects database and M4D in the second programming period) implies necessarily such a temporal break in the activity of data collection. It is not necessarily a negative thing because it makes possible to evaluate what was good or bad in previous period, what can be improved, etc.

Springtime is this very short moment, at the beginning of an ESPON programming period, where it is possible to propose innovations in terms of data collection (scales, time period, geographic object, thematic…). This short period of innovation in terms of data collection is absolutely crucial because it has a deep influence on all the production realized later by the program as a whole. In the first ESPON period (2001-2006), the innovation at stake was clearly the enlargement of the geographical coverage from Old Member States (EU15) to New Member States and associated countries. The strong interest for ESPON maps and ESPON results in this period was related to this crucial innovation. In the case of the second ESPON programming period (2008-2013) the time allocated to innovation was shorter because the database project started without any advance on the other projects. Regional data was immediately requested on the same area than in the previous programming period (EU31 at NUTS2 or NUTS3). Innovative ideas was proposed (combination of regions, cities, grid, local data[1]…) but certainly too late in the Spring Time period. And it is therefore not surprising to observe that the very large majority of data collected in the period 2008 has been of the same type (NUTS2 or NUTS 3 regions).

Summertime is this period of abundance where all the efforts made in wintertime and springtime are transformed in abundant harvest of data, maps and reports. The ESPON M4D has been clearly designed to focus on the harvest of results of ESPON 2013 and the dissemination of the products of this harvest. The effort made on the data check and data storage is precisely targeted to store everything that has been produced by ESPON projects. The elaboration of a powerful and efficient interface for data extraction or visualization is designed to insure the complementary task of dissemination of the results of the harvest toward the consumers. And last but not least, the connections established between the database project and the other priority 3 projects is a metaphor of the transformation of raw materials into more sophisticated products with higher added value in the chain (monitoring, automatic cartography, city benchmarking…). At the moment of the current SIR of M4D, we are in the middle of this harvest period, with a lot of workforce engaged in data storage, transformation and diffusion. It is a heavy task, no doubts… and we are not really discussing the quality of what is stored, but only the exact place of production and indication of technics employed (metadata!).

Autumn correspond to the final year of an ESPON programming period. It is a crucial moment because some original production that was launched late in the Spring Period produces their fruits too. Like grapefruits that will produce wine in September, or mushrooms that grow in the forests in October… we expect in the final months of the M4D project to deliver some original and precious products like database on cities, final version of Olap cubes, harmonized time series of core indicators, coverage of ESPON area with local data for some specific indicators, proposal of new geometries between LAU2 and NUTS3 … Of course, this production is minor as compared to the giant harvest of regional data realized in the summer period (NUTS2 and NUTS3 are the corn or the wheat in our metaphor). But their value is great because they complete the point of view deliver by regional data and offer new opportunities and perspective on territorial cohesion and spatial dynamics at various scales.

[1] Cf ESPON Database 1 final report.

Working Package A – Application

1.1 Computer (WPA1)

Since the First Interim Report (30th June 2012), innovative functionalities have been added to the ESPON Database Portal. This work has been done in conformance with the **ESPON Web Interface Specifications** (produced in 2012), integrating the inputs of the **Thematic Group of the M4D Project** (WP B), and taking into account the on-going adjustments asked by the **ESPON Coordination Unit**.

Although the ESPON Web Interface Specifications covered most of the integration functionalities, some adjustments were necessary to take into account the heterogeneity of the ESPON datasets and the complexity of the ESPON Data flow.

For instance, the integration of new nomenclatures (various NUTS versions, urban objects) or functionalities (new filters) involved a strong cooperation between both developers and thematic experts.

To increase the efficiency of their cooperation, RIATE has adopted the Agile Scrum method that LIG was already using.

The first part of this chapter describes how the method was useful to smooth and ease the work done by RIATE and LIG on the improvement of the Portal.

The second part details what has been done since the First Interim Report.

The last part presents the forthcoming functionalities that will be delivered before the Final Report of the ESPON M4D Project.

1.1.1 Methodology: Agile Management Method Scrum

Aware of the fact that the ESPON Web Interface Specifications were neither exhaustive nor definitive, it was necessary that RIATE thematic experts and LIG developers kept collaborating closely all along the development process.

This collaboration relied on the Scrum methodology. The aim of that report is not to present that methodology, but you can find it in more details here:

http://en.wikipedia.org/wiki/Scrum_%28software_development%29

In short, ‘Scrum’ is an iterative and incremental agile software development framework for managing software projects and products or application development. Its focus is on "a flexible, holistic product development strategy where a development team works as a unit to reach a common goal" as opposed to a "traditional, sequential approach".

The difficulties to predict by advance the nature of data delivered by the ESPON Project and their compliance with the ESPON Metadata model was a major problem to deal with.

The ESPON Coordination Unit feedbacks, led to many adjustments in the user interface, increasing its user-friendliness.

Those tasks needed frequent interactions between RIATE and LIG teams. This cooperation was eased through the use of Scrum artefacts, roles and meetings, adding transparency to the overall process for both partners (see Figure 2).

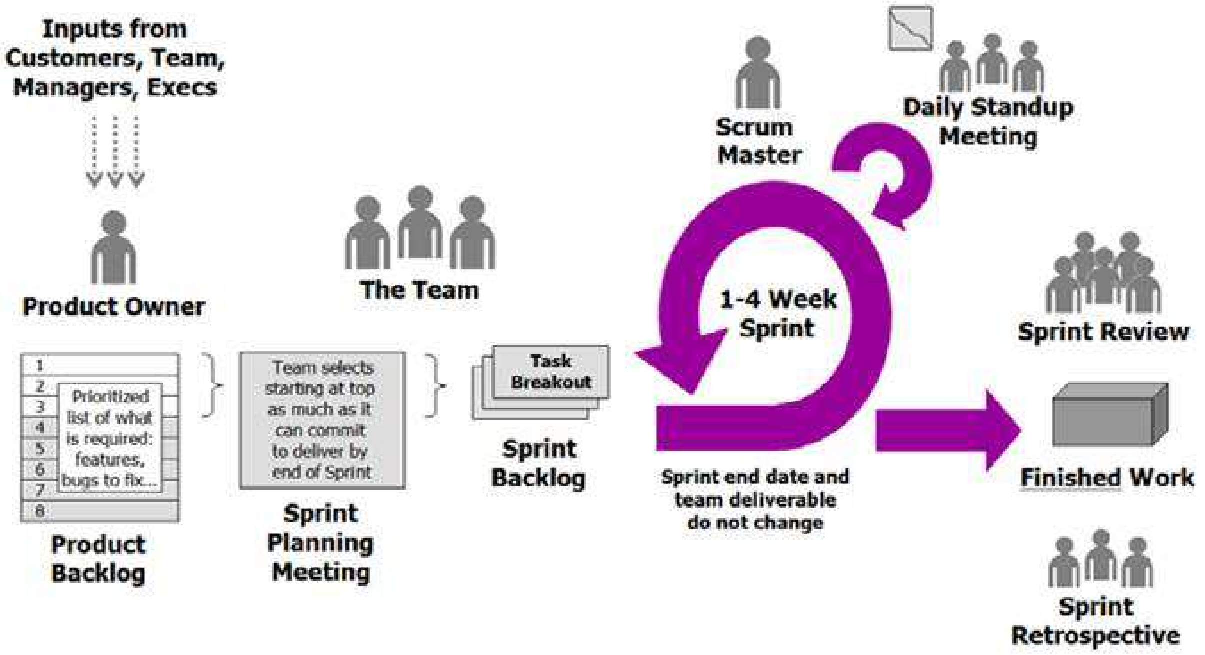

Inputs from Customers, Team, Managers, Execs
Product Owner
The Team
Scrum Master
Daily Standup Meeting
1-4 Week Sprint
Sprint Review
1
2
3
4
5
6
7
8
Prioritized list of what is required: features, bugs to fix...
Product Backlog
Team selects starting at top as much as it can commit to deliver by end of Sprint
Sprint Planning Meeting
Task Breakout
Sprint Backlog
Sprint end date and team deliverable do not change
Finished Work
Sprint Retrospective

Figure 2: SCRUM. Source: http://www.qikkwit.com (2013-06-20)

1.1.2 New features

This section lists the features of the ESPON Database Portal delivered since the First Interim Report (June 2012). These deliveries have been made through three releases[2]:

- Report "Delivery December 2012" on the 21st of December 2012;
- Report "Delivery February 2013" on the 8th February of 2013;
- Report "Delivery June 2013" on the 5th June 2013.

These deliveries include three different types of features:

- New functionalities of the ESPON Database Portal, such as improvement of the search query interface;
- New contents, work done by the thematic group WP B, involving all the ESPON M4D partners. That part is detailed in WPA4 section;
- New "Platform" functionalities such as datasets integration tracking tool or news edition.

1.1.2.1 Search interface

All the options specified in the Web Interface Specifications are available in the June 2013 release of the ESPON Database Portal, exception of the multiple keywords search.

From now on, the database can be queried through four different strategies:

- **Semantic:** strategy based on various classifications: thematic, project, policy or keyword;
- **Temporal:** strategy based on the temporal properties of the indicator;
- **Spatial**: strategy based on study areas (EU27…) and territorial levels (Regional, Cities). Experts can even specify the nomenclatures they wish (NUTS, MUA…).
- **Data:** strategy based on the nature of the indicator or its data.

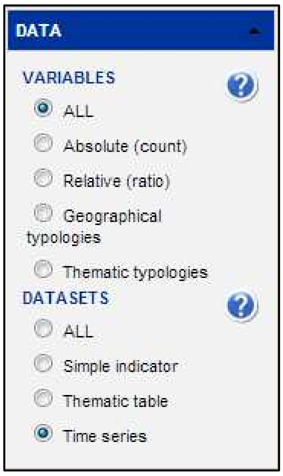
[2] For a complete listing of the new functionalities, please refer to the Annex 1 of the report.

All these strategies may be combined, making it easier to find ad hoc indicators, their number getting bigger. It must be noted that it is also possible to sort the results, accordingly to the user preferences.

The section below describes in more details the implemented filters.

1.1.2.1.1 Filter "Data"

In the Web Interface Specifications, this filter was entitled ***general.*** This was not precise enough to understand what was behind this strategy. It has been renamed « ***Data*** », since it relates to the statistical properties or the structure of the data (thematic table? ratio?)



DATA

VARIABLES

● ALL
○ Absolute (count)
○ Relative (ratio)
○ Geographical typologies
○ Thematic typologies

DATASETS

○ ALL
○ Simple indicator
○ Thematic table
● Time series

Figure 3 - The "DATA" filter (previously called "General")

This filter proposes two criteria:

- **Variables properties**

  To filter the results by the statistical properties of the indicator[3], five possibilities:

  - **Absolute** (count): relates to real quantity (a sum makes sense), e.g. total population
  - **Relative** (ratio): relates to the ratio between two quantities (a sum does not make sense), e.g. GDP per capita, share of elderly people etc.
  - **Geographical typologies**: relates to classifications structured in geographical types, e.g. mountainous areas, coastal areas etc.
  - **Thematic typologies**: relates to classifications structured in thematic types, e.g. demographic trends, typology on natural hazards etc.
  - **All**: statistical properties are not taken into account.

- **Datasets type**

  To filter the results by the dataset properties of the indicator(s):

  - A **thematic table** is considered as a set of indicators being semantically linked. It means that it is impossible to download them separately. In practical terms, it relates to two kinds of figures: contingency table (the sum of the columns makes sense, e.g. age structure, employment by branch etc.) or indicators semantically linked (e.g. a typology and its related raw data used, or indicators in a coherent dataset)[4]
  - A **time-series** is considered as an indicator available at different temporal extents (5 years at least) and having the following attributes (principles of the core database)[5]:

    (1) Most of the missing values have been estimated

[3] In ESPON Metadata, it corresponds to the NAT Type of the indicator.

[4] In ESPON Metadata, it corresponds to indicators declared as an aggregation.

[5] In ESPON Metadata, it corresponds to indicators which have the value « true » in the Core field.

(2) Each time-series has been statistically checked (outlier check)

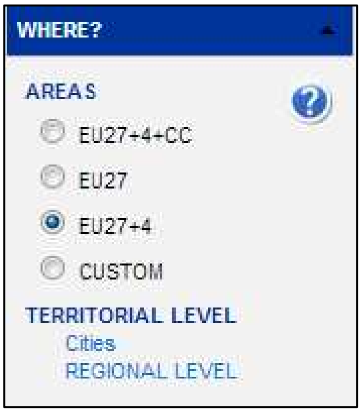
(3) It is possible to update the indicator quite easily in the future. It implies though a very good description of the indicator in the metadata.

(4) The statistical type of time-series indicators is certainly absolute stocks or absolute flows (excepted data on temperatures).

- **Simple indicator** is not a thematic table or a time-series.
- **All** dataset type is not taken into account (default choice).

It is important to note that some indicators can be both time-series and thematic table.

1.1.2.1.2 Filter "Where?"



WHERE?

AREAS

○ EU27+4+CC

○ EU27

● EU27+4

○ CUSTOM

TERRITORIAL LEVEL

Cities

REGIONAL LEVEL

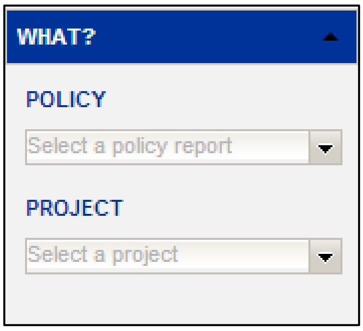
The filter "*Where?*" has been available since December 2012. It mainly consists in filtering indicators by an area of interest (EU27, EU27+4, EU27+4+CC) or a combination of countries.

It is also possible to filter indicators considering territorial level:

- Regional (NUTS);
- Cities (MUA, UMZ, FUA).

Figure 4 – The filter Where?

1.1.2.1.3 Filter What?



WHAT?

POLICY

Select a policy report

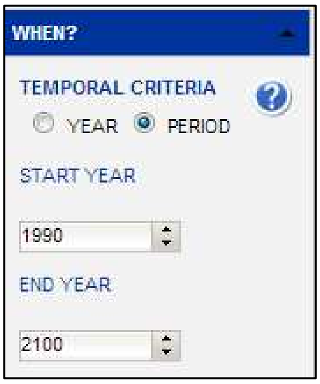
PROJECT

Select a project

This filter allows adding more semantic criteria. For instance, if the main semantic filter is "search by keywords", it is then possible to filter by theme, policy and/or name of an ESPON Project.

Figure 5 – The filter What?

1.1.2.1.4 Filter When?



WHEN?

TEMPORAL CRITERIA

○ YEAR ● PERIOD

START YEAR

1990

END YEAR

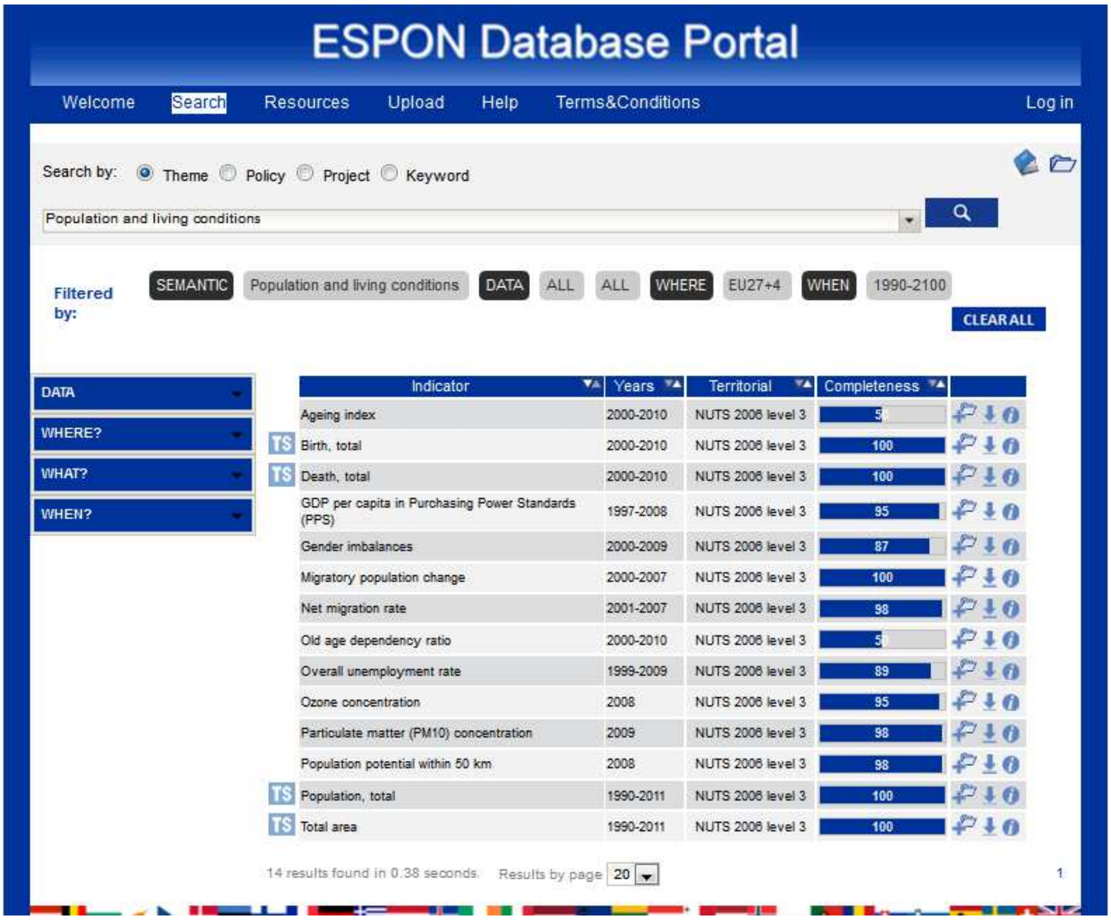
2100

Indicators contain data that is related to a year, or a period, or combinations of both. The when filter permits to select indicators on that criterion.

Figure 6 – The filter When?

1.1.2.1.5 Search results

The figure below shows the search query interface.



Search by: ● Theme ○ Policy ○ Project ○ Keyword

Population and living conditions

Filtered by: SEMANTIC Population and living conditions DATA ALL ALL WHERE EU27+4 WHEN 1990-2100

	Indicator	Years	Territorial	Completeness
	Ageing index	2000-2010	NUTS 2006 level 3	50
TS	Birth, total	2000-2010	NUTS 2006 level 3	100
TS	Death, total	2000-2010	NUTS 2006 level 3	100
	GDP per capita in Purchasing Power Standards (PPS)	1997-2008	NUTS 2006 level 3	95
	Gender imbalances	2000-2009	NUTS 2006 level 3	87
	Migratory population change	2000-2007	NUTS 2006 level 3	100
	Net migration rate	2001-2007	NUTS 2006 level 3	98
	Old age dependency ratio	2000-2010	NUTS 2006 level 3	50
	Overall unemployment rate	1999-2009	NUTS 2006 level 3	89
	Ozone concentration	2008	NUTS 2006 level 3	95
	Particulate matter (PM10) concentration	2009	NUTS 2006 level 3	98
	Population potential within 50 km	2008	NUTS 2006 level 3	98
TS	Population, total	1990-2011	NUTS 2006 level 3	100
TS	Total area	1990-2011	NUTS 2006 level 3	100

14 results found in 0.38 seconds. Results by page 20 1

Figure 7 – Search query interface

In addition to search filters, results table has been improved:

- **A column with icons identifying the dataset type**

  A column has been added to results table. It shows the dataset type, with the help of icons. There are four possibilities:

  - No icons, it is a simple indicator;
  - TS indicates it is a time-series;
  - TT indicates it is a thematic-table;
  - both icons, it is a time-series AND a thematic-table.

- **Sort the results**

  It is now possible to sort the results by clicking on columns of the table:

  - Indicator name;
  - Years;
  - Territorial (e.g. geographical object);
  - Indicator completeness (depends on the study area chosen).

- **Choose the number of results by page**

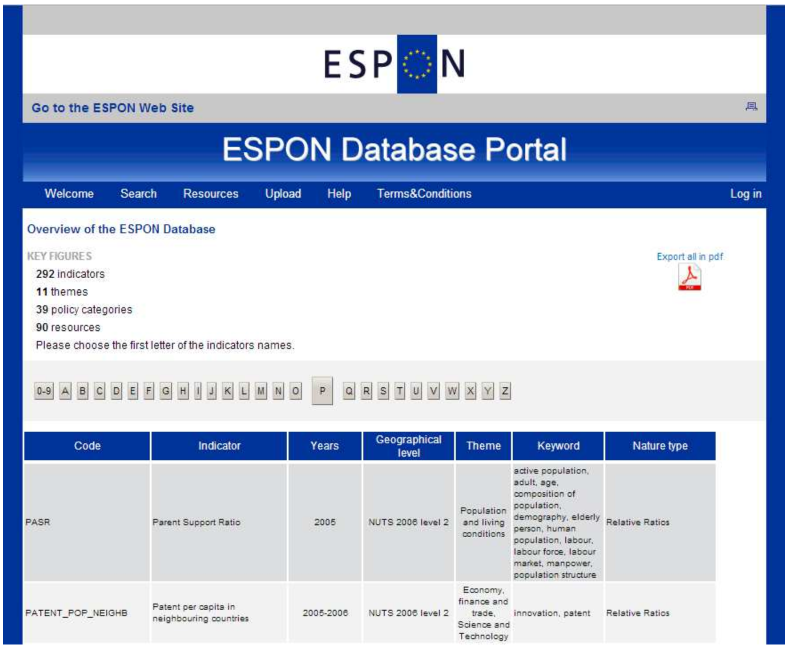
  It is now possible to define the number of results displayed by page.

1.1.2.2 The Overview page

The overview page has been available since February 2013. It provides the full list of key indicators.

Features of indicator displayed in this page are code, name, temporal extent, geographical level, themes, keywords and nature type. The indicators are alphabetically classified.

It is possible to download a summary of the content of the database (*pdf* format).



Overview of the ESPON Database

KEY FIGURES

292 indicators
11 themes
39 policy categories
90 resources

Please choose the first letter of the indicators names.

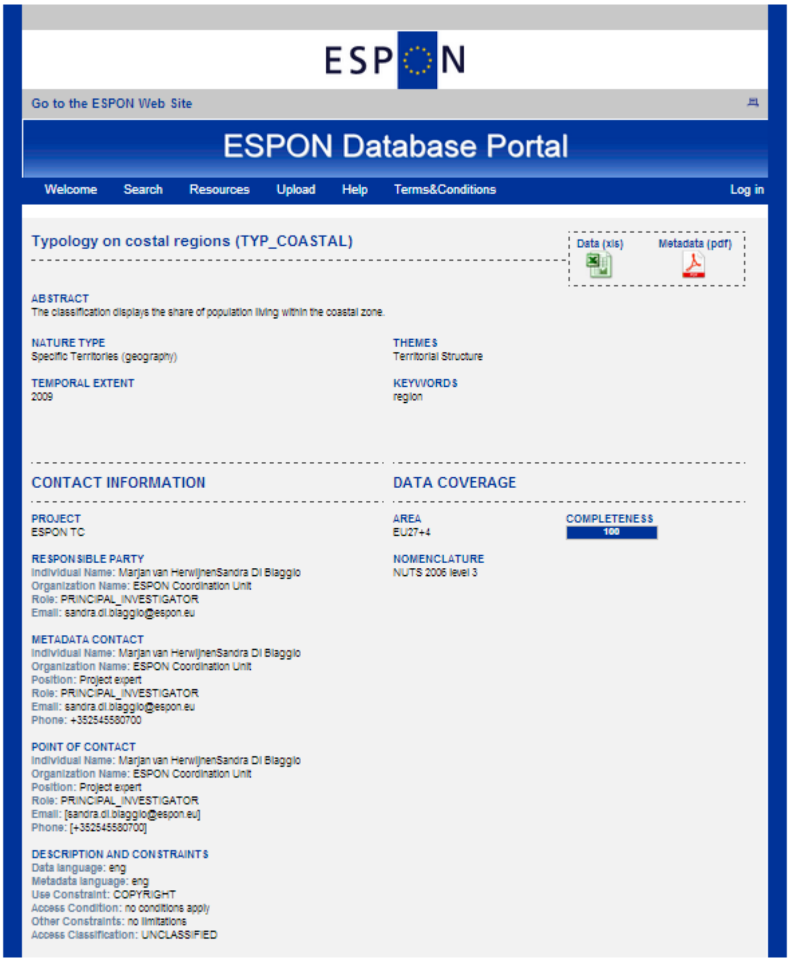
Code	Indicator	Years	Geographical level	Theme	Keyword	Nature type
PASR	Parent Support Ratio	2005	NUTS 2006 level 2	Population and living conditions	active population, adult, age, composition of population, demography, elderly person, human population, labour, labour force, labour market, manpower, population structure	Relative Ratios
PATENT_POP_NEIGHB	Patent per capita in neighbouring countries	2005-2006	NUTS 2006 level 2	Economy, finance and trade, Science and Technology	innovation, patent	Relative Ratios

Figure 8 - Overview page

1.1.2.3 Metadata page

The metadata page has been improved. Its content is now richer, and is based on the filters that were selected. For instance, if only one country was selected in Where filter, Metadata will concern only the data related to that country (sources, completeness…).

User can download a pdf version of that page, or download the data it is related to.



Typology on costal regions (TYP_COASTAL)

ABSTRACT
The classification displays the share of population living within the coastal zone.

NATURE TYPE
Specific Territories (geography)

THEMES
Territorial Structure

TEMPORAL EXTENT
2009

KEYWORDS
region

CONTACT INFORMATION

PROJECT
ESPON TC

RESPONSIBLE PARTY
Individual Name: Marjan van HerwijnenSandra Di Biaggio
Organization Name: ESPON Coordination Unit
Role: PRINCIPAL_INVESTIGATOR
Email: sandra.di.biaggio@espon.eu

METADATA CONTACT
Individual Name: Marjan van HerwijnenSandra Di Biaggio
Organization Name: ESPON Coordination Unit
Position: Project expert
Role: PRINCIPAL_INVESTIGATOR
Email: sandra.di.biaggio@espon.eu
Phone: +352545580700

POINT OF CONTACT
Individual Name: Marjan van HerwijnenSandra Di Biaggio
Organization Name: ESPON Coordination Unit
Position: Project expert
Role: PRINCIPAL_INVESTIGATOR
Email: [sandra.di.biaggio@espon.eu]
Phone: [+352545580700]

DESCRIPTION AND CONSTRAINTS
Data language: eng
Metadata language: eng
Use Constraint: COPYRIGHT
Access Condition: no conditions apply
Other Constraints: no limitations
Access Classification: UNCLASSIFIED

DATA COVERAGE

AREA
EU27+4

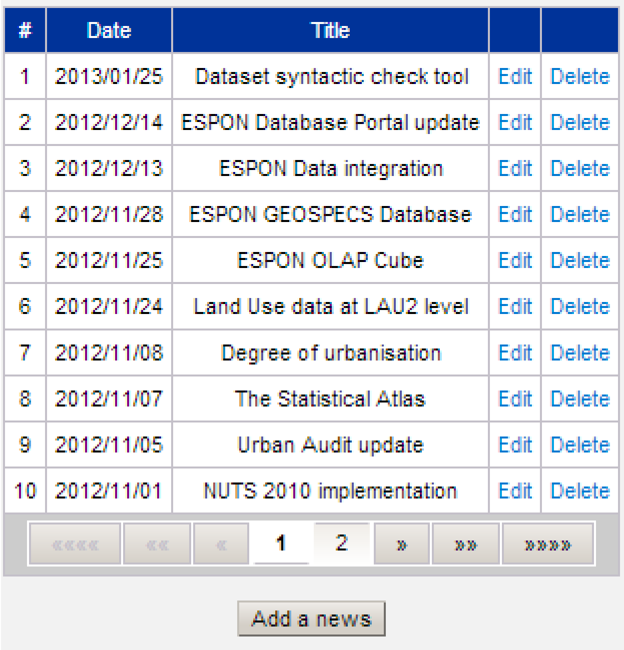
COMPLETENESS
100

NOMENCLATURE
NUTS 2006 level 3

Figure 9 – The Metadata page

1.1.2.4 A handy "News" management tool

For authorized users only, it is now possible to add or edit or delete the list of the "News" displayed on the home page of the Portal.

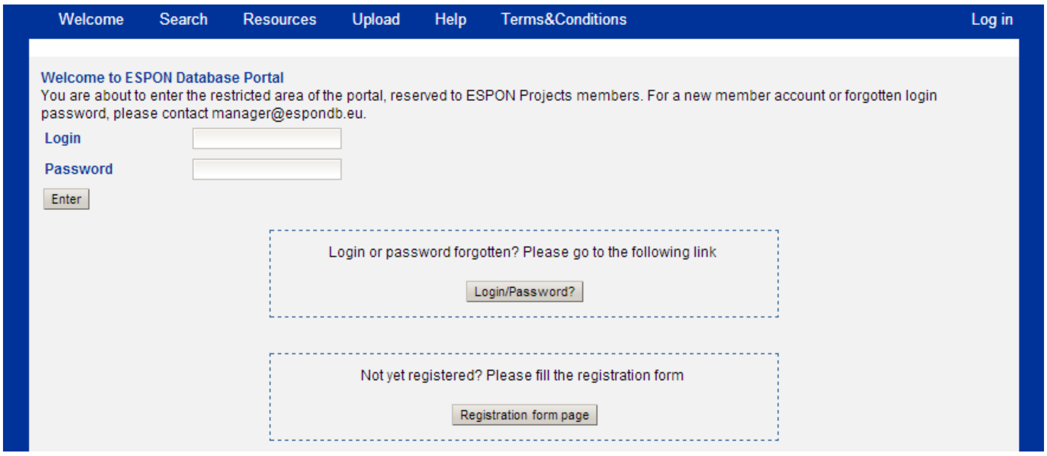


#	Date	Title		
1	2013/01/25	Dataset syntactic check tool	Edit	Delete
2	2012/12/14	ESPON Database Portal update	Edit	Delete
3	2012/12/13	ESPON Data integration	Edit	Delete
4	2012/11/28	ESPON GEOSPECS Database	Edit	Delete
5	2012/11/25	ESPON OLAP Cube	Edit	Delete
6	2012/11/24	Land Use data at LAU2 level	Edit	Delete
7	2012/11/08	Degree of urbanisation	Edit	Delete
8	2012/11/07	The Statistical Atlas	Edit	Delete
9	2012/11/05	Urban Audit update	Edit	Delete
10	2012/11/01	NUTS 2010 implementation	Edit	Delete

Figure 10 – The "News" management tool

1.1.2.5 Creation of a "user's registration page"

Another management tool helps registering new users. By the past, it was hand-made and took a lot of time and emails. This is the reason why this feature, initially not specified in the Web Interface Specifications, has been finally implemented (figure 11).



Welcome to ESPON Database Portal

You are about to enter the restricted area of the portal, reserved to ESPON Projects members. For a new member account or forgotten login password, please contact manager@espondb.eu.

Login

Password

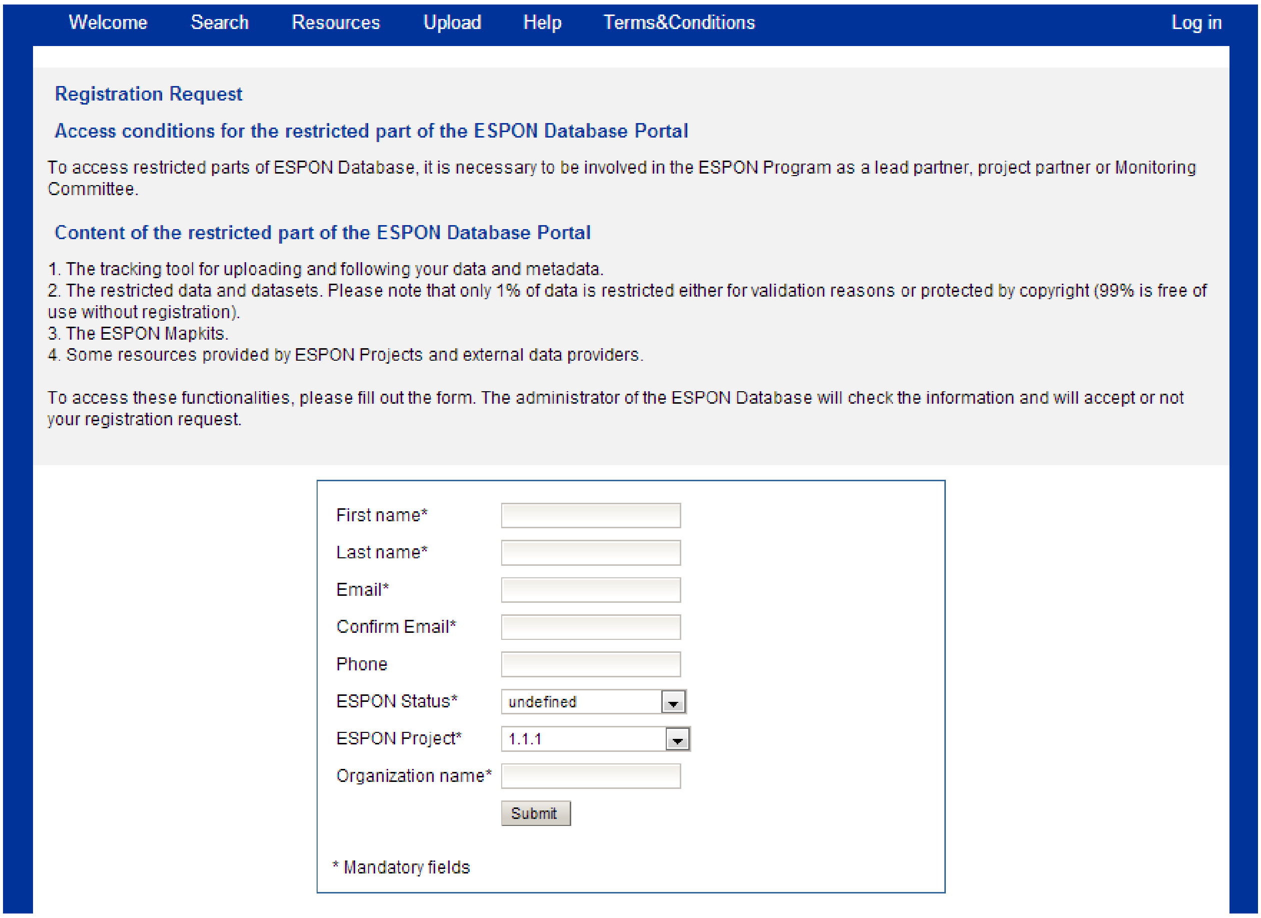
Login or password forgotten? Please go to the following link

Not yet registered? Please fill the registration form

Figure 11 – User's registration page

On the registration page, a user can:

- Login and access the restricted area of the portal;
- Recover his/her forgotten login/password pair;
- Ask for a member account (see figure 12).



Registration Request

Access conditions for the restricted part of the ESPON Database Portal

To access restricted parts of ESPON Database, it is necessary to be involved in the ESPON Program as a lead partner, project partner or Monitoring Committee.

Content of the restricted part of the ESPON Database Portal

1. The tracking tool for uploading and following your data and metadata.
2. The restricted data and datasets. Please note that only 1% of data is restricted either for validation reasons or protected by copyright (99% is free of use without registration).
3. The ESPON Mapkits.
4. Some resources provided by ESPON Projects and external data providers.

To access these functionalities, please fill out the form. The administrator of the ESPON Database will check the information and will accept or not your registration request.

First name*	
Last name*	
Email*	
Confirm Email*	
Phone	
ESPON Status*	undefined
ESPON Project*	1.1.1
Organization name*	
	Submit

* Mandatory fields

Figure 12 – Online registration form

When this form is submitted, the ESPON Database administrator and the ESPON Coordination Unit are notified, and they can status on the registration request.

1.1.3 Web Services SWOT Analysis

As mentioned in the contractual document entitled "Specification – ESPON Applied Research Project [2013/3/3] – ESPON Database and Data Development – Phase II (2011-2014)" (ESPON M4D Subsidy Contract Annex IV), "*the access to the ESPON Database should not only be considered through a Web query interface, but also via direct integration of the data into third party applications, including spatial data portals. First a SWOT analysis should clarify this aspect and provide the ESPON CU with the pros and cons of the different possibilities*".

To respond this need, the delivered document entitled "Access to ESPON Database by third-party applications" is the result of a technical expertise conducted in Spring 2013 by UNEP/Grid-Geneva in the context of the M4D project. This expertise includes a review of currently available solutions to query the ESPON Database and their evaluation with respect to solutions based on Web services, and recommendations for designing a solution based on Geographic Web services.

1.1.4 Next steps

This section lists the main functionalities to be implemented.

Regarding the ESPON Database Portal Web Application:

- Metadata page of indicators: map of completeness for indicators that are neither a time-series nor a thematic table.
- Case studies:
  - Integration of case studies.
  - Search interface dedicated to case studies (to be further discussed between M4D Partners and ESPON CU).
  - Output metadata page for case studies.
- Availability of the M4D Newsletter June 2013 in the Welcome page.
- Regarding the documents available in the "Help" section:
  - The User Manual needs to be completely reviewed and detailed according to the new functionalities of the portal.
  - The Dictionary of Spatial Units is a draft version and will be soon completed.

1.2 Statistics (WP A2)

1.2.1 Outlier handling

We are exploring the applicability of Bayesian methods to provide estimates of the forecast uncertainty. If a user is attempting to determine whether the impact of some development or policy is likely to have a 2% impact on employment, but the forecasts are ±3%, then the impact will not be detectable reliably.

Since the submission of the First Interim Report in June 2012, datasets have started to flow from the ESPON Projects, mainly Priority 1 projects. A methodology for identifying potentially anomalous values was developed during the 1st Phase of the Database project, using data on the evolution of the GDP as test data. In essence, a number of alternative approaches to the identification of anomalous values was used on the data. If a NUTS region was identified as having an anomalous value for a particular variable, then it was flagged. The more flags a region accrued than the more likely it would be a candidate for checking. The tests included (i) aspatial distributional tests for variables taken one at a time (ii) aspatial tests for variables taken two at a time (iii) aspatial tests for variables taken several at a time (iv) spatial tests.

Following the commencement of the 2nd Phase of the project, further test data was supplied to the NCG by RIATE, and the specification of the data organisation within an Excel spreadsheet was finalised. The checking software was extensively re-written to deal with the spreadsheets – an important decision was that the data should be read directly from the spreadsheet into the data structures used by the checking software. There are 4 worksheets with each spreadsheet:

Identifier:	supplier information
Indicator:	data organisation and data dictionaries
Source:	data lineage
Data:	the values for testing

The metadata in the Indicator worksheet is used to create a header file which then drives the data check procedure. The output is a printed report, with optional exploratory graphics (histograms, scatterplot matrices, boxplots, and maps). The printed reports are then re-organised into a Word document, in which interpretative comments are added by the NCG team to reflect the manner in which the anomalous values might be handled. In many cases this is a request to check that value that appears to be anomalous, as there can be false positives arising from the check process. We have picked up a small number of disagreements between the metadata and the data, which can only be picked up during the quality check process.

Once the datasets started to arrive from the ESPON Projects – with a warning about their availability delivered by email from the tracking tool, it became clear that there was much greater variety in terms of both the data organisation and the closeness with which project partners interpreted the M4D Team specifications. Early on it became clear that a strategy to deal with missing data was required. A second concern arose about the nature of the spatial tests. These will be expanded on below.

Several sections of the code have again been re-written, and the guidance supplied in the report to the data suppliers is as specific as possible concerning the values and locations which have been identified as anomalous.

Missing values

Several of the datasets which have been received from the suppliers have missing values. This is not itself anomalous, as the dataset specifications permit the suppliers to nominate a unique value (say -99 or NA) to represent a missing value. A missing value is supplied when it has not been possible for the supplier to obtain or calculate a value for the indicator in question. This raises the problem of how to deal with missing data. We can observe two types of missing information: item-missing and unit-missing. Item missing data are those values which are not present for part of the data for a particular NUTS region. Unit missing data are data missing for every indicator for a particular NUTS region. In several cases data lines have been supplied for the French départements et territoires d'outre-mer but the entry for every indicator is supplied with a missing value code.

For the process of anomaly detection we made the decision to omit NUTS regions with missing data. We do not know the generating process for the data, and we cannot assume that any imputation technique we might apply would be context-independent. The process that generates the missing values is unlikely to be random, but using the available cases is perhaps the expedient course of action. This has two consequences. First, for spatial tests we lose potential neighbours if A is a neighbour of B and A is missing, then there can be no result for A; however, in considering B the number of neighbours is reduced by 1, decreasing the precision of the test statistic we compute. For multivariate tests we are forced to ignore any case for which any indicator in the multivariate set is missing; if the missing data were randomly located, the more variables in the multivariate set, the more cases that will be missing, and we lose the ability to identifying potentially anomalous data.

Data distributions

Many of the tests assume a roughly symmetrical distribution – we avoid tests which rely on a particular model such as the Normal distribution. The boxplot statistics are used as a basis for many tests, but the notches (±1.58 * IQR/n0.5) appear to be based on the asymptotic normality of the median. Several indicators have exhibited long tails on the right hand side (this is typical of ata relating to income distribution), so the boxplot tests have returned many what appear to be false positives. The underlying philosophy of the detection approach is to make no changes to the data, so transformations such as square root or log to reduce the effects of the long tail have been avoided. When we produce the written report for the supplier, we make comment on the distribution shape and the consequences for anomaly detection.

Spatial tests

The spatial tests that were developed for the Phase 1 work included a test due to Hawkins for detecting spatial outliers. It turns out that Professor Hawkins was unaware that his name was attached to a test, but confirmed the characteristics of the test statistic named in the earliest paper to describe the test. He has suggested some modifications, and commented that the 5% cutoff we have used in a number of tests to determine an anomalous value is perhaps not sufficiently far into the tail of the distribution to identify something as an 'outlier'. Until we have done some investigating on the amended test, we are using a Local Moran's I to identify spatially unusual measures. This gives an indication of whether the observed value is high or low in comparison with values in the neighbouring zones.

Written report

The output from the detection exercise is a written report in PDF form which is returned to the M4D team and uploaded for transmission to the supplier. Much of the detection exercise requires detailed statistical knowledge, in particular multivariate and spatial statistics. While some projects have members with a high level of statistical expertise, this is not so across all ESPON projects. For this reason we have added explanatory commentary to the lists of apparently anomalous zone codes detailing the reasons for their appearance. The final section of the report provides advice as to where the supplier needs to check the supplied data, and whether it needs to have a second pass through the data check procedure.

The timing between the arrival of datasets has had some consequences for our ability to meet the return time target of two weeks. It would appear that a smooth and steady flow of datasets is not what happens in reality, and arrival has been irregular. When several arrive with a few days of each other, it is challenging to meet the target. One supplier sent 14 datasets simultaneously!

1.2.2 Outlier handling

An issue has arisen with regard to the Hawkins' Test for spatial outliers. Hawkins 1980[6] monograph "Identification of Outliers" defines an outlier as 'an observation that deviates so much from other observations that it was generated by a different mechanism'. Hawkins is credited by Rossi et al 1992[7] with the development of a test thus:

> All values, z(x), are considered to be suspect. Because spatial outliers are not necessarily the largest or smallest values encountered, each value is compared with its neighbouring values. Let n be defined as the number of neighbouring values excluding z(x), let M equal the arithmetic mean of the n values, and let σ^2 denote the average variance for equivalently sized neighbourhoods over the sample space. Assuming the neighbourhood values are normally distributed, Hawkins has shown:

[6] Hawkins D, 1980, Identification of Outliers, London: Chapman and Hall

[7] Rossi RE, Mulla DJ, Journel AG and EH Franz, 1992, Geostatistical tools for modelling and interpreting ecological spatial dependence, Ecological Monographs, 62(2), 277-314

$$h = \frac{n[z(x) - M]^2}{(n+1)\sigma^2}$$

to be distributed as χ^2.

Thus if the computed value of h in the equation above is greater than the tabulated value of χ^2 then that z(x) is likely to be a spatial outlier. At the 5% significance level, χ^2 is 3.842. Rossi et al cite his 1980 monograph as the source of the test; they also cite Krige and Magri 1982[8] with a demonstration of the test, who in turn cite Hawkins 1980. Krige and Magri describe the test in appendix 1 of their paper and give some more clues: in particular that n should be small "between 4 and 10; 4 for low data variability and 10 for high variability".

Hawkins 1980 has no reference to any spatial tests and, having exchanged emails with Professor Douglas Hawkins[9], it turns out he has never published such a test, although he was working with Krige and Magri around that time and would have engaged in discussions[10]. Professor Hawkins confirmed that the statistic is distributed as χ^{21}, but he also pointed out that if the statistic was rewritten as:

$$(\frac{z(x) - M}{\sigma})\sqrt{\frac{n}{n+1}}$$

it would preserve the sign of the difference and could be treated as a normal variate. This is rather more useful than a test which merely suggests 'anomaly' as it would indicate whether the value for the zone in question was higher than that in the neighbouring zones or lower. A related issue is that of 'winsorising'. Hawkins suggested also that any z(x) values which are greater than the corresponding C = M + 1.96σ√(n/(n+1)) should be replaced by the corresponding C {and also M - 1.96σ√(n/(n+1)) } before the spatial tests are carried out.

There is also a question relating to the 5% significance test – Hawkin's view of 'outlier' is that it implies something unusual and therefore an outlier would be expected much further into the tail of the distribution. This means that with 5% we would identify false positives ("outliers" that are not actually particularly unusual). In practice, few of the datasets have yielded what seems to false positives, but it would be prudent to consider a more stringent significance level.

This then raises the awkward problem of multiple testing; if there are 100 tests with a 5% significance level, we would expect 5 tests to be significant at random. That is, 5 regions would be identified as false positives. This implies that a Bonferroni[11] style adjustment to the test level should be employed. The Bonferroni adjustment adjusts the testwise error rate to α/m so that the familywise error rate is α. With 100 tests, to achieve a familywise 0.05 significance level the individual critical values are raised from 1.96 to 3.48. The 0.01 significance level requires a chnage in the critical value from .58 to 3.89. It has been observed that Bonferroni adjustment can be too conservative (too many false negatives), so true outliers may be missed. Another correction for multipl,e testing is due to Šidák: 1-$(1-\alpha)^{1/n}$. This adjustment assumes independent tests, and as we are using the same data over and over again for the tests, this adjustment is not appropriate. An adjustment for multiple dependent tests, such as that by Benjamini and Yekutieli 2001[12,] is perhaps more appropriate.

[8] Krige, DG, and EJ Magri, 1982, Studies of the effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body, Mathematical Geology, 14, 557-564

[9] Professor Douglas M Hawkins, School of Statistics, University of Minnesota

[10] Hawkins, D, 2012, personal communication

[11] Bonferroni, CE, 1935, Il calcolo delle assicurazioni su gruppi di teste, Studi in Onore del Professore Salvatore Ortu Carboni, Rome: Italy, 13-60

[12] Benjamini, Y and Yekutieli D, 2001, The control of the false discovery rate in multiple testing

In the light of this we have decided that, for the moment, spatial outliers will be tested using Anselin's 1995[13] Local Moran I statistic. This is a local version of Moran's I statistic used as a measure of spatial autocorrelation. The value of a variable in some location is compared with the values in the immediately neighbouring locations.

$$I_i = (z_i - \bar{z}) \sum_{j=1}^{n} w_{ij} (z_j - \bar{z})$$

Whilst the global statistic can be used to make some statement about whether values of the variable of interest are similar or not to those in neighbouring cells, it will not indicate where this is a characteristic. The local neighbourhood relations may be distance based or adjacency based. The range of the local statistic is from -1 to +1, with zero indicating the presence of no local autocorrelation. Positive local spatial autocorrelation would indicate that the value of the variable in a given zone is similar to those in its neighbourhood, and negative local spatial autocorrelation would indicate that the value is different from that in the neighbouring zones. Significant negative values would suggest potential outliers. The derivation of the p-values for the test require randomisation tests, and the interpretation requires care.

The choice of significance level should take into account the number of tests being carried out. Goovaerts and Jacquez 2005[14] suggest an adjustment in which the testwise level is at the □/b level, where b is the average number of neighbours in the set of zones. This allows for the non-independence of the tests. In the R implementation of the local Moran I the adjustment for multiple testing is over the number of neighbours+1.

The implementation of both tests requires some care, since the variance estimates with very small neighbourhoods may be unstable. This means that with adjacency based neighbourhoods, islands need to be removed. This is consistent with the treatment of islands in Biomedware's[15] Spacestat package.

under dependency, Annals of Statistics, 29, 1165-1188

[13] Anselin, L., 1995, Local indicators of spatial association – LISA, Geographical Analysis, 27, 93-115.

[14] Goovaerts P and Jacquez G, 2005, Detection of temporal changes in the spatial distribution of cancer rates using local Moran's I and geostatistically simulated spatial neutral models, Journal of Geographical Systems, 7(1), 137-159

[15] Biomedware, 2012, SpaceStat | Biomedware, http://www.biomedware.com/?module=Page&sID=spacestat

1.3 MapKit (WP A3)

Since June 2012 and on the basis of the work done within the WP B5 (Neighbourhood), UMS RIATE has updated the Mapkit on the Regional Neighbourhood. This work has been done thanks to the feedbacks provided by the ESPON ITAN[16]. Thus, the new version of the Mapkit on the Regional Neighbourhood contains the following changes:

- Simple revision of the SNUTS nomenclature for Russia, Tunisia and Algeria (code change or aggregation of SNUTS units).
- Complete revision of the SNUTS nomenclature for Morocco (split of territorial units at level 3).
- Addition of Armenia and Azerbaijan in the SNUTS nomenclature.
- Addition of a precise layer for GIS calculations.

Until December 2013, the M4D Project is waiting for the feedbacks from the ITAN Project concerning the following countries: Libya, Egypt, Occupied Palestinian Territories, Israel, Lebanon, Syria, Jordan, Ukraine and Belarus. If changes in the SNUTS nomenclature will be required, this mapkit will be adapted to that purpose.

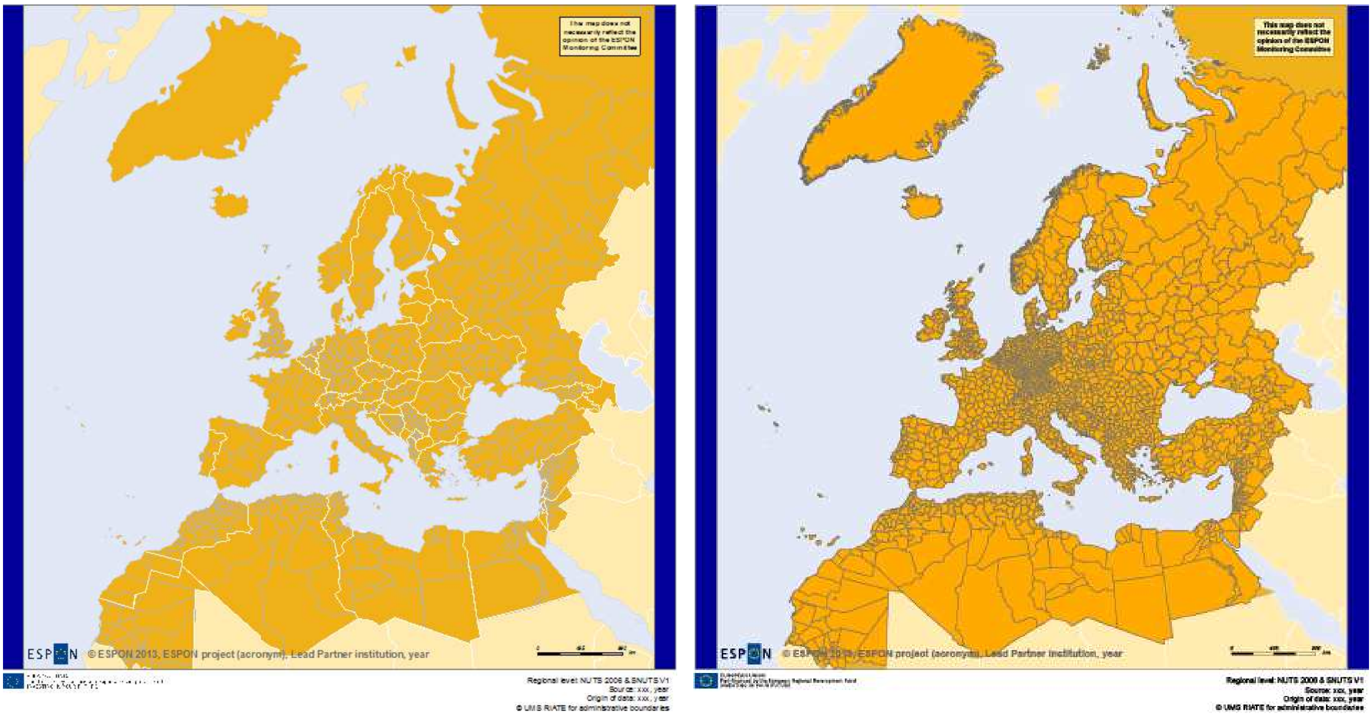

This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
ESPON
© ESPON 2013, ESPON project (acronym), Lead Partner institution, year
Regional level: NUTS 2006 & SNUTS V1
Source: xxx, year
Origin of data: xxx, year
© UMS RIATE for administrative boundaries
This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
ESPON
© ESPON 2013, ESPON project (acronym), Lead Partner institution, year
Regional level: NUTS 2006 & SNUTS V1
Source: xxx, year
Origin of data: xxx, year
© UMS RIATE for administrative boundaries

Figure 13 – Regional neighbourhood mapkit adapted to cartographic purpose (on the left) and layer adapted for GIS calculation (e.g. land cover calculation etc., on the right).

In the ESPON Coordination Unit response on revised version of the First Interim Report, it is mentioned that *"the generalised version of the geometries used to create the European Neighbourhood Map Kit tool should be replaced by the non generalised version from Eurogeographics".*

We strongly disagree with this approach for several reasons:

- With this geographical coverage and with the degree of precision of the boundaries, it is quite difficult (if not impossible) to distinguish the colours displayed in the polygons. **But this will be the most important information** for the maps that will be created within the ESPON ITAN Project.

[16] More precisions in the WPB5 part

- In graphical terms, the generalised map is more adapted for publications. Whatever the publication format (research papers, press etc.) a first step for creating maps consists by adapting the boundaries precision to the aim of the map in order to display the cartographic message in the most efficient way. This has been done with the generalised map, adapted to ESPON purposes. **Using the non-generalised version would be a non-sense for displaying socio-economic phenomenon at regional level for the European Neighbourhood.** It is important to remind that the boundaries of the ESPON Euromed Mapkit was more simplified, for a geographical coverage which was more or less the same. This Mapkit has been largely diffused to academics or practitioners; and much appreciated.
- When a NUTS revision will appear, Eurogeographics delivers a new shapefile with the adjusted boundaries. But by the past (NUTS 2003 to NUTS 2006 version), the degree of precision of the coastline has been modified also. It means that by following the solution of the Eurogeographics template, **if a NUTS change will occur, the update of the regional neighbourhood will imply a significant amount of human resources.**

Nevertheless, a layer combining Eurogeographics and GADM has been created. The geometries are seamless (no polygon intersection problems). **But this layer is adapted for GIS calculations only**.

1.4 Database updates (WP A4)

Besides the implementation of new functionalities (as described in part 1.1), the ESPON Database has been continuously fed with new ESPON TPGs key indicators datasets, and for some of them, with the new nomenclatures they are related to.

Regarding the data integration workflow, three status need to be distinguished:

- Datasets coming from former ESPON Projects (ESPON 2006 Program, beginning of the ESPON 2013 Program). They do not fit with the ESPON Metadata format.
- Datasets already integrated.
- Datasets to be integrated (in the tracking phase).

Recoding metadata

Aiming at being compliant with the INSPIRE directive, the ESPON Metadata has been largely modified since the ESPON Database 1 Project. New fields have been added (policy theme, core data, data type, GEMET keywords etc.). As a consequence, it was necessary to update a significant number of datasets:

- Those coming from the ESPON 2006 Program (ESPON Project Indicators[17]);
- Those provided by the projects of the ESPON 2013 Program before the metadata specifications were complete.
- Those with too little metadata information.

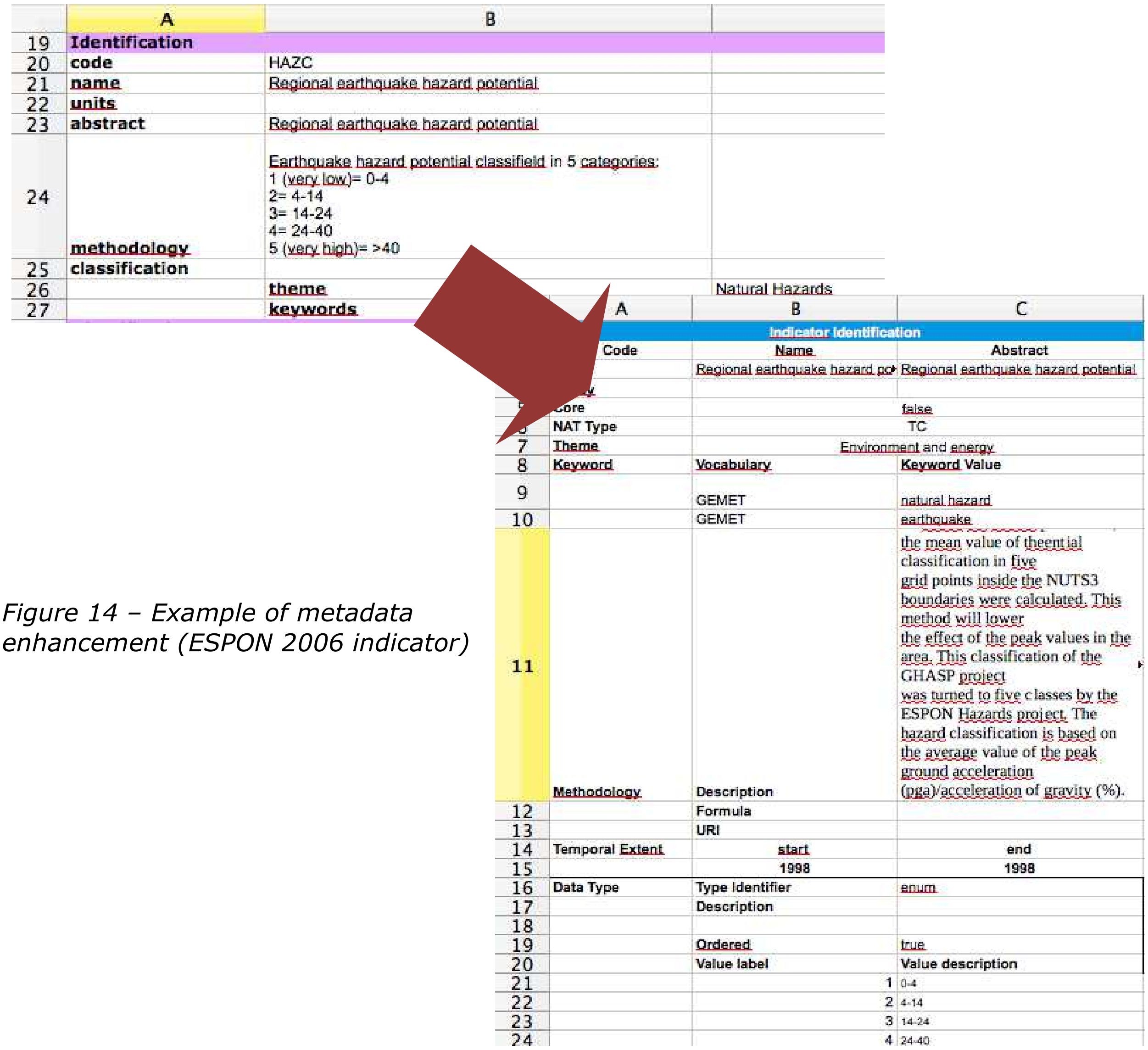
All in all, the adjustment of metadata has concerned 46 datasets (ESPON 2006 Program, INTERCO, TERCO, ATTREG and ESPON Database Project phase 1).

The first step consisted in complying with the new ESPON metadata specifications. Adjustments generally consisted in adding keywords, theme, NAT Type and a description of the data type (figure 14).

The second step was the enhancement of ESPON Metadata. Most of the time, the level of metadata was too poor to ensure a full understanding of the indicator for non-

[17] http://www.espon.eu/main/Menu_ToolsandMaps/ESPON2006Tools/DatabasePublicFiles/projectindicators.html

stakeholders (e.g. What is the regional earthquake potential?). Thus, it required intensive researches in the ESPON reports to improve the quality of metadata, as described in the figure 14 (the methodology field allows now to understand the methodology used to obtain the final typology).



	A	B	
19	Identification		
20	code	HAZC	
21	name	Regional earthquake hazard potential	
22	units		
23	abstract	Regional earthquake hazard potential	
24	methodology	Earthquake hazard potential classifield in 5 categories: 1 (very low)= 0-4 2= 4-14 3= 14-24 4= 24-40 5 (very high)= >40	
25	classification		
26		theme	Natural Hazards
27		keywords	

	A	B	C
		Indicator identification	
	Code	Name	Abstract
		Regional earthquake hazard po	Regional earthquake hazard potential
	y		
	ore		false
	NAT Type		TC
7	Theme	Environment and energy	
8	Keyword	Vocabulary	Keyword Value
9		GEMET	natural hazard
10		GEMET	earthquake
11	Methodology	Description	the mean value of theential classification in five grid points inside the NUTS3 boundaries were calculated. This method will lower the effect of the peak values in the area. This classification of the GHASP project was turned to five classes by the ESPON Hazards project. The hazard classification is based on the average value of the peak ground acceleration (pga)/acceleration of gravity (%).
12		Formula	
13		URI	
14	Temporal Extent	start	end
15		1998	1998
16	Data Type	Type Identifier	enum
17		Description	
18			
19		Ordered	true
20		Value label	Value description
21		1	0-4
22		2	4-14
23		3	14-24
24		4	24-40

Figure 14 – Example of metadata enhancement (ESPON 2006 indicator)

ESPON datasets delivered within the 2013 program required less work. Regarding the syntax, usually, the datasets were compliant with the expected format. The work consisted in checking the semantic of the metadata.

From now, the integration of datasets will follow the steps of the tracking process, syntactic check, semantic check, outliers check, ESPON CU validation.

The below section *"Pending Datasets"* permits to see the priority affected to these datasets.

Datasets integrated in December 2012

In December 2012, 35 datasets have been integrated into the ESPON Database. The following Priority 1 and 3 ESPON TPGs have produced these datasets:

- ESPON KIT
- ESPON Typology Compilation

- EDORA
- ReRisk
- DEMIFER
- TIPTAP
- ESPON Climate
- ESPON Database
- ESPON M4D (Core indicators, urban data)
- Map Updates on demography, accessibility, telecom, creative workforce

Besides the Key indicators datasets, covering the whole ESPON area with a good completeness of data, the data produced within Targeted Analysis are available in the "Resources" part of the ESPON Database Portal[18]. These datasets will be available in the Case-study Search interface (work in progress).

The ESPON Background data (e.g. all the data provided by each ESPON Project) are available in the "Resources" part of the ESPON Database Portal[19].

Datasets integrated in June 2013

In June 2013, 35 more datasets have been integrated into the ESPON Database. The following ESPON Projects produced these new datasets:

- ATTREG
- INTERCO
- M4D (cities)
- ReRisk (time-series on temperature)
- TERCO
- KIT

For a complete overview of the June 2013 delivery, please consult the document entitled "ESPON 2013 Database, June 2013 delivery".

Datasets uploaded in the tracking tool

Since December 2012, ESPON Projects have delivered their datasets with the ESPON Tracking Tool[20]. These datasets must pass the syntactic, the semantic and the outlier checks before being integrated. The time taken by the tracking process is time-consuming, but it considerably improves the quality of metadata.

A good example of the usefulness of this process implementation can be given by the delivery of the ESPON ESatDOR project. This P1 Project was the first one to successfully pass through the all steps of the tracking process. At first, ESatDOR has very early asked a lot of questions concerning the data and metadata specifications to the M4D team. Thanks to the exchanges of information (28 emails exchanged from the 16/10/2012 to the 15/04/2013), the syntactic check was not a problem and the semantics of the metadata were clearly defined. However, the outlier check has revealed some statistical outliers. From the main findings of this report, the ESatDOR project mentioned that *"As indicated by ESPON DB Manager by email, we reject the outliers check because we found some issues in the dataset that needed to be solved. Therefore, we start the process of uploading again with the mentioned problems corrected"*[21].

[18] http://database.espon.eu/db2/resource?idCat=72

[19] http://database.espon.eu/db2/resource?idCat=71 for ESPON priority 1 projects and http://database.espon.eu/db2/resource?idCat=73 for ESPON priority 3 projects.

[20] http://database.espon.eu/db2/tracking

[21] The outlier check of the ESPON ESatDOR project is available in Annex 1.

The ESatDOR dataset can be considered as an example of good practice in term of data integration: "*I think these experiences are good to improve both the tool and our knowledge on metadata problems*[22]".

In June 2013, the tracking tool contains datasets coming from 4 ESPON Projects:

- **SEGI (1 dataset)**: The outlier check has been uploaded on the tracking tool the 1st April 2013. We are waiting for the approval (or not) of the SeGI project to go to the next step of the integration.
- **SIESTA (2 datasets):** The outlier check has been uploaded on the tracking tool the 14th February 2013. We are waiting for the approval (or not) of the SIESTA project to go to the next step of the integration.
- **SEMIGRA (2 datasets)**: The semantics have been successfully passed. The Outlier check has been upload the 30th April 2013. Consequently to the outlier check, the SEMIGRA project has decided to resubmit its dataset;
- **Map update on Natural Hazards (14 datasets)**: A heavy data delivery. The semantics have been accepted and will be integrated soon in the ESPON Database after validation of the ESPON Coordination Unit (the outlier check is not adapted to typologies derived from raster data).

These **datasets will not be integrated into the ESPON Database before having passed all the steps of the tracking tool**[23]. Please note that the ESPON Coordination Unit has the right to validate instead of ESPON Projects the semantic or the outlier checks to speed up the process. But it is always better to let the ESPON Projects decide themselves. Therefore, better the reactivity of ESPON Projects is, faster the integration of their datasets will be.

Pending datasets

For December 2013, the M4D Project intends to make available in the Search Interface the following datasets:

- Key datasets produced within the ESPON 2006 Program[24].
- Data from the second, third, fourth and fifth EU Cohesion Reports (provided by the ESPON Database Project). These datasets are a bit complicated to integrate in the database since it combines data in various NUTS versions.

At the moment, all these datasets have been recoded accordingly to the latest version of the ESPON Metadata specifications.

To sum up, the remaining tasks regarding the database content are:

- Review and refactoring of the policy codes in the input Key Indicators datasets.
- Integration of new nomenclatures (sNUTS, LUZ).
- Integration of new datasets (ESPON 2006 Program and datasets based on above nomenclatures).

[22] Alberto Lorenzo Alonso, ESatDOR Project, ETC-SIA, University of Malaga, 15 April 2013.

[23] At least the syntactic and semantic checks, the outlier check when adapted.

[24] http://www.espon.eu/main/Menu_ToolsandMaps/ESPON2006Tools/DatabasePublicFiles/projectindicatorsterms.html

Working Package B - Thematic

2.1 Regions (WP B1)

2.1.1 Time-series

Strategy for harmonising time-series

One of the outputs from the ESPON M4D programme is an in depth analysis of the possible solutions for the update and harmonisation of times series data and for the eventual upload of the harmonised time series into the ESPON Database. Such times series are often incomplete or heterogeneous[25] at lower NUTS levels for a variety of reasons. The M4D team has considered different approaches to the imputation of the missing data in these time series. We make the distinction between unit non-response (no data available for a given NUTSx region) and item non-response (some but not all data are available for a NUTSx region).

The work is challenging for a series of reasons. First we should consider the nature of time series data in the context of the statistical theory which deals with the analysis of time series. A time series is "a collection of observations made sequentially in time26". In most cases there is a fixed time interval between the observations (1 hour, 1 day, 1 week, 1 month, 1 year…) but in rarer cases the observations may occur with differing time intervals, for example aviation accidents or railway accidents.

Activities surrounding time series concern the description of the main properties of series, the identification of unusual values in the series, attempts to explain the linkage between the series in question and other series (sea level and temperature), and prediction. We sometimes refer to predictions in the future as forecasting, and those backwards from the beginning of the series as backforecasting. Equally the terms extrapolation and retropolation can be used as synonyms for forecast and backforecasting. Interpolation is the filling in of values between known events for which there is data.

The description of a time series can include the decomposition of the series into its components sources of variation: trend, seasonal fluctuation, other cyclical variation, and residuals. The residuals themselves may not be random, but may require further modelling to detect any patterns – conventional models for this include moving average and autoregressive models. The challenge is that a time series in the sense that Chatfield and other authors have in mind is unlikely to be shorter than 50 elements; many 'classic' series have hundreds of observations.

This raises a problem for the description of the data series used in ESPON, since annual series (such as mid year population estimates) may have a few as 20 elements, some have fewer. Because of the shortness of the series, the data have more in common with longitudinal studies, and there are methods used in such studies for imputing data. There is another consideration to the ESPON time series and that arises because the 'time series' data have a strong and well-defined *cross-sectional component*[27] which remains, generally, constant during the time periods.

[25] In the case of time series, it is important to distinguish between *completeness* (all the table is filled with values) and *homogeneity* (all the following values are defining consistent time series, without discontinuities and unexplained time outliers).

[26] Chatfield c, 1989, *The Analysis of Time Series*, 4th edn, London:Chapman & Hall

[27] The *cross-sectional component* means that the time series stored by ESPON are short but not independent from each other. For example, the evolution of young population is not independent from the evolution of adult or old population. Or the evolution of a regional unit is related to the evolution of the state where this unit is

A starting point which helps to formalise an approach is to make some assumptions about the organisation of the data. We make the initial assumption that the data take the form of counts – for example, population, employees available for work, employees in employment. We can envisage a number of scenarios concerning the structure of the data, and this allows us to suggest appropriate strategies.

[1] In the simplest case data is only available as annual observations at NUTS0 level, for all or part of a time period between 1991 and 2011. We may have only one series, or we may have several series which are only available at NUTS0. No data is available for NUTS1, NUTS2 or NUTS3.

[2] The second scenario would involve a series at NUTS0 and counts available for NUTS1 and NUTS2 regions below the NUTS0 country for a single time period – this would usually correspond to the taking of a national population census.

[3] A third scenario would involve the presence of a series at NUTS0 for some or all of the time period together with cross-sectional counts available for two or more census periods, at NUTS1 and NUTS2.

[4] A fourth scenario would build on the third by including intercensal population counts for the NUTS1 and NUTS2 zones, but not for all time periods.

As an example, a type [4] test dataset for Bulgaria is used which has NUTS0 employment counts for 2000 through to 2010, with NUTS1 and NUTS2 counts for 2003 through to 2010. The task then is to complete the national series back 1 year to 1999, and then estimate the NUTS1/NUTS2 counts for the period 1999 to 2002 inclusive. A diagrammatic view of the data is show in the table below. The elements NA are those for which it is desired to estimate values. One of the outputs from the ESPON M4D programme is an in depth analysis of the possible solutions for the update and harmonisation of times series data and for the eventual upload of the harmonised time series into the ESPON Database. Such times series are often incomplete or heterogeneous[28] at lower NUTS levels for a variety of reasons. The M4D team has considered different approaches to the imputation of the missing data in these time series. We make the distinction between unit non-response (no data available for a given NUTSx region) and item non-response (some but not all data are available for a NUTSx region).

The work is challenging for a series of reasons. First we should consider the nature of time series data in the context of the statistical theory which deals with the analysis of time series. A time series is "a collection of observations made sequentially in time29". In most cases there is a fixed time interval between the observations (1 hour, 1 day, 1 week, 1 month, 1 year…) but in rarer cases the observations may occur with differing time intervals, for example aviation accidents or railway accidents.

Activities surrounding time series concern the description of the main properties of series, the identification of unusual values in the series, attempts to explain the linkage between the series in question and other series (sea level and temperature), and prediction. We sometimes refer to predictions in the future as forecasting, and those backwards from the beginning of the series as backforecasting. Equally the terms extrapolation and retropolation can be used as synonyms for forecast and backforecasting. Interpolation is the filling in of values between known events for which there is data.

located or to the neighbouring units… It is therefore possible to balance the problem of short time series by introduction of other methods of estimation.

[28] In the case of time series, it is important to distinguish between *completeness* (all the table is filled with values) and *homogeneity* (all the following values are defining consistent time series, without discontinuities and unexplained time outliers).

[29] Chatfield c, 1989, *The Analysis of Time Series*, 4th edn, London:Chapman & Hall

The description of a time series can include the decomposition of the series into its components sources of variation: trend, seasonal fluctuation, other cyclical variation, and residuals. The residuals themselves may not be random, but may require further modelling to detect any patterns – conventional models for this include moving average and autoregressive models. The challenge is that a time series in the sense that Chatfield and other authors have in mind is unlikely to be shorter than 50 elements; many 'classic' series have hundreds of observations.

This raises a problem for the description of the data series used in ESPON, since annual series (such as mid year population estimates) may have a few as 20 elements, some have fewer. Because of the shortness of the series, the data have more in common with longitudinal studies, and there are methods used in such studies for imputing data. There is another consideration to the ESPON time series and that arises because the 'time series' data have a strong and well-defined *cross-sectional component*[30] which remains, generally, constant during the time periods.

A starting point which helps to formalise an approach is to make some assumptions about the organisation of the data. We make the initial assumption that the data take the form of counts – for example, population, employees available for work, employees in employment. We can envisage a number of scenarios concerning the structure of the data, and this allows us to suggest appropriate strategies.

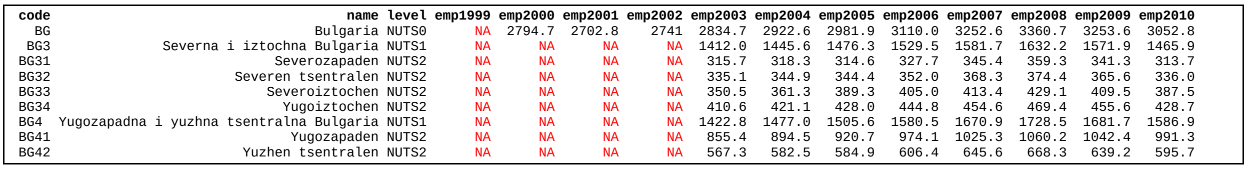
[1] In the simplest case data is only available as annual observations at NUTS0 level, for all or part of a time period between 1991 and 2011. We may have only one series, or we may have several series which are only available at NUTS0. No data is available for NUTS1, NUTS2 or NUTS3.

[2] The second scenario would involve a series at NUTS0 and counts available for NUTS1 and NUTS2 regions below the NUTS0 country for a single time period – this would usually correspond to the taking of a national population census.

[3] A third scenario would involve the presence of a series at NUTS0 for some or all of the time period together with cross-sectional counts available for two or more census periods, at NUTS1 and NUTS2.

[4] A fourth scenario would build on the third by including intercensal population counts for the NUTS1 and NUTS2 zones, but not for all time periods.

As an example, a type [4] test dataset for Bulgaria is used which has NUTS0 employment counts for 2000 through to 2010, with NUTS1 and NUTS2 counts for 2003 through to 2010. The task then is to complete the national series back 1 year to 1999, and then estimate the NUTS1/NUTS2 counts for the period 1999 to 2002 inclusive. A diagrammatic view of the data is show in the table below. The elements NA are those for which it is desired to estimate values.



code	name	level	emp1999	emp2000	emp2001	emp2002	emp2003	emp2004	emp2005	emp2006	emp2007	emp2008	emp2009	emp2010
BG	Bulgaria	NUTS0	NA	2794.7	2702.8	2741	2834.7	2922.6	2981.9	3110.0	3252.6	3360.7	3253.6	3052.8
BG3	Severna i iztochna Bulgaria	NUTS1	NA	NA	NA	NA	1412.0	1445.6	1476.3	1529.5	1581.7	1632.2	1571.9	1465.9
BG31	Severozapaden	NUTS2	NA	NA	NA	NA	315.7	318.3	314.6	327.7	345.4	359.3	341.3	313.7
BG32	Severen tsentralen	NUTS2	NA	NA	NA	NA	335.1	344.9	344.4	352.0	368.3	374.4	365.6	336.0
BG33	Severoiztochen	NUTS2	NA	NA	NA	NA	350.5	361.3	389.3	405.0	413.4	429.1	409.5	387.5
BG34	Yugoiztochen	NUTS2	NA	NA	NA	NA	410.6	421.1	428.0	444.8	454.6	469.4	455.6	428.7
BG4	Yugozapadna i yuzhna tsentralna Bulgaria	NUTS1	NA	NA	NA	NA	1422.8	1477.0	1505.6	1580.5	1670.9	1728.5	1681.7	1586.9
BG41	Yugozapaden	NUTS2	NA	NA	NA	NA	855.4	894.5	920.7	974.1	1025.3	1060.2	1042.4	991.3
BG42	Yuzhen tsentralen	NUTS2	NA	NA	NA	NA	567.3	582.5	584.9	606.4	645.6	668.3	639.2	595.7

Figure 15 – Concrete situation of time-series estimation

Of the several expedient strategies for dealing with data of this sort, one is to do nothing and just propose to limit the work to available data. **Doing nothing and excuding all estimations is a relevant strategy for a database which would be mainly used**

[30] The *cross-sectional component* means that the time series stored by ESPON are short but not independent from each other. For example, the evolution of young population is not independent from the evolution of adult or old population. Or the evolution of a regional unit is related to the evolution of the state where this unit is located or to the neighbouring units… It is therefore possible to balance the problem of short time series by introduction of other methods of estimation.

for normative and legal purpose (e.g. allocation of funds to regions with more than x% of unemployment; definition of cities by a threshold of y inhabitants; emission of CO2 lower to an objective decided by an international organisation). A second strategy is to try to estimate missing value and to fill the gap in the data. **Filling gaps is a relevant strategy for a database oriented to the definition of long terme political strategies, understanding of territorial impact of political decisions, etc.** The opposition between the two strategy is not an opposition between scientific/political approach but rather between normative/strategic use of data in a political context.

In the example of Bulgaria, it is possible to apply the second strategy (if relevant) and examine the relationships between the employment counts and the lower NUTS levels and their parent zones. There are also hierarchical constraints in that at any one time period, the count for BG4 should equal the sum of BG41 and BG42, and the proportional split between BG41 and BG42 should sum to 1. The proportions can be carried backwards according to a number of possibilities – last observation carried forward (LOCF) is one and fractional weighted imputation would be another. LOCF makes an assumption about the stationarity of the series which may or may not be reasonable. A fractional weighted approach would apply a prespecified set of weights to the previous m observation (2 ~ m ~ 4); more of the supplied data is used with this approach.

Other strategies include hotdeck[31], k-nearest neighbour, and iterative model-based imputation[32]. Such strategies have been suggested for handling both unit and item non-response in longitudinal surveys[33]. In our example, there is insufficient data for apply these approaches. Note that LOCF is one version of hotdeck.

In implementing an expedient strategy, the same code can be used for both LOCF and FWI: the weights are (1, 0, 0, 0) for the former, and (0.5, 0.3, 0.15, 0.05) for the latter. The differences between the backforecasts are shown in the diagrams below. The FWI approach is less conservative in its assumptions, although if the backforecast is long enough, the proportions will settle down to constant values.

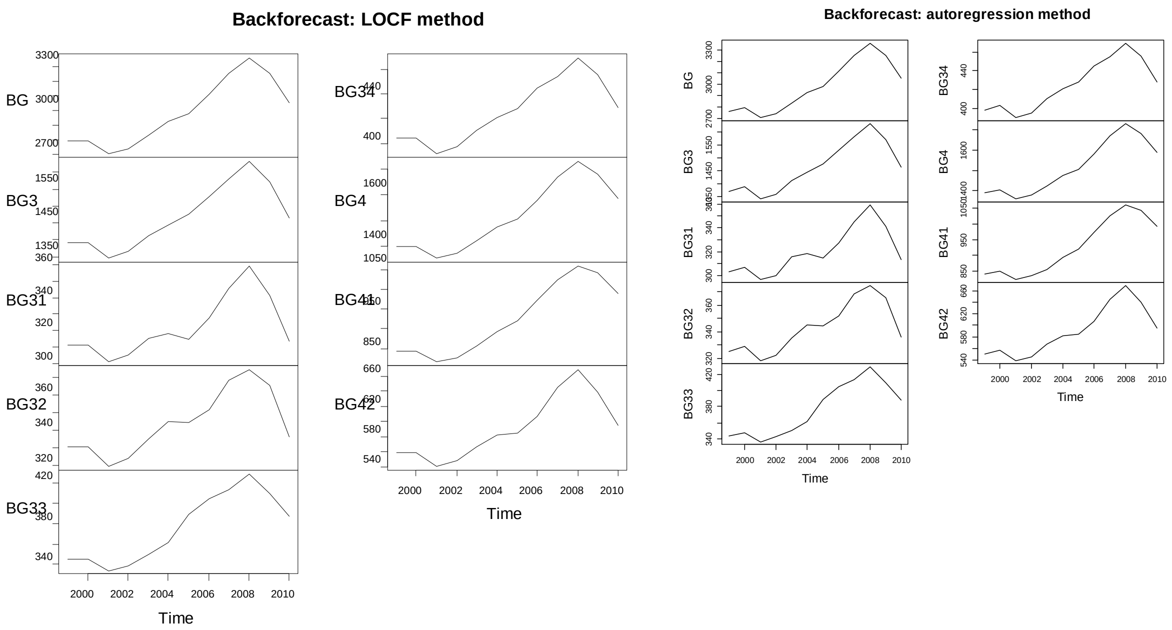


Figure 16 – LOCF and autoregression methods for backforecasting

[31] Ford BL, 1983, An overview of hot-deck procedures, in: Incomplete data in sample surveys, Madow WG, Olkin I, Rubin DB, (Eds.) , Academic Press, New York, pp.185-207.

[32] Templ M, Kowarik A and Filzmoser P, 2011, Iterative stepwise regression imputation using standard and robust methods, Journal of Computational Statistics and Data Analysis, 55, 2793-2806

[33] Tang L, song J, Belin TR, and Unützer J, 2005, A comparison of imputation methods in a longitudinal randomized clinical trial, Statistics in Medicine, 24, 2111-2128

The LOCF approach is effectively a constant model, and this is shown in the flat sections of the line where the data has been imputed. The FWI approach might be more plausible – national population growth is assumed for the transition between the first and second time periods, and the proportional splits are similarly dynamic.

There are two challenges for the handling of time series data, given its characteristics in the ESPON context. The first is to be able to recognise the various patterns of unit non-response and item non-response and choose the appropriate strategies for handling these situations. A second challenge, perhaps greater, is to provide confidence intervals for these forecasts.

2.1.2 Functional indicators

As described in the First Interim Report (e.g. principles of the core database strategy), it is not necessary to gather too much indicators to create an innovative database : but it is important to select very carefully the more crucial one and to put a maximum of workforce on their elaboration and check.

With a limited number of consistent core indicators, e.g. count data, available in time-series, which have been carefully statistically checked (cf part 2.1.1) and using methods derived from spatial analysis (smoothing, discontinuities measures, multiscalar analysis), it is potentially possible to create an infinity of indicators derived from these basic data. As proposed by the ESPON HyperAtlas, based on multiscalar analysis, these methods are particularly adapted for proposing innovative views of the ESPON Space which may help the process of policy making[34].

These methods are also especially adapted to the EU policy context. A concrete example of the interest of the core database strategy can be given by the production of functional indicators (figures 17-18). With a single dataset (time-series of population) and having a travel-time matrix, it is possible to propose a high number of useful indicators for stakeholder and practitionners:

- What is the potential of population accessible in a given time-distance around my city/ my region ?
- What time-distance is necessary to reach a given amount of population if one decide to implement a production or service in this particular place ?

It is obvious that it is possible to extend functional indicators to other M4D Core indicators (GDP, employment etc.). We can therefore arrive to answers more sophisticated problems, at the top of the political agenda:

- What are the area of ESPON territory where opening of borders could create complementarities on labour market?
- What are the part of the European Territory where positive and negative spillover are observed in terms of economic growth at local scale ?

[34] Ysebaert R., Lambert N., Grasland C., Le Rubrus B., Villanova-Oliver M., Gensel J., Plumejeaud C., 2012, HyperAtlas, un outil au service du débat politique, application à la politique de Cohésion de l'Union Europénne in : Fonder les sciences du territoire, Bekouche et al., Karthala, 293 p. (an English version will be soon available).

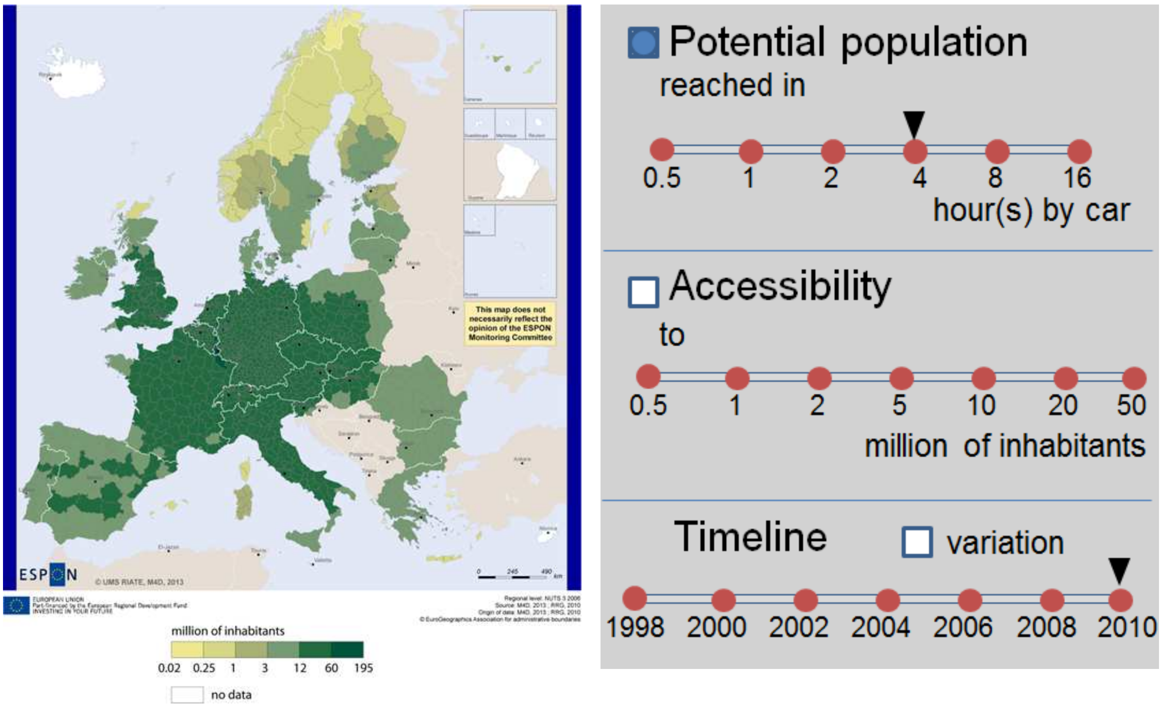


Figure 17 - Population reached in 4 hours by car

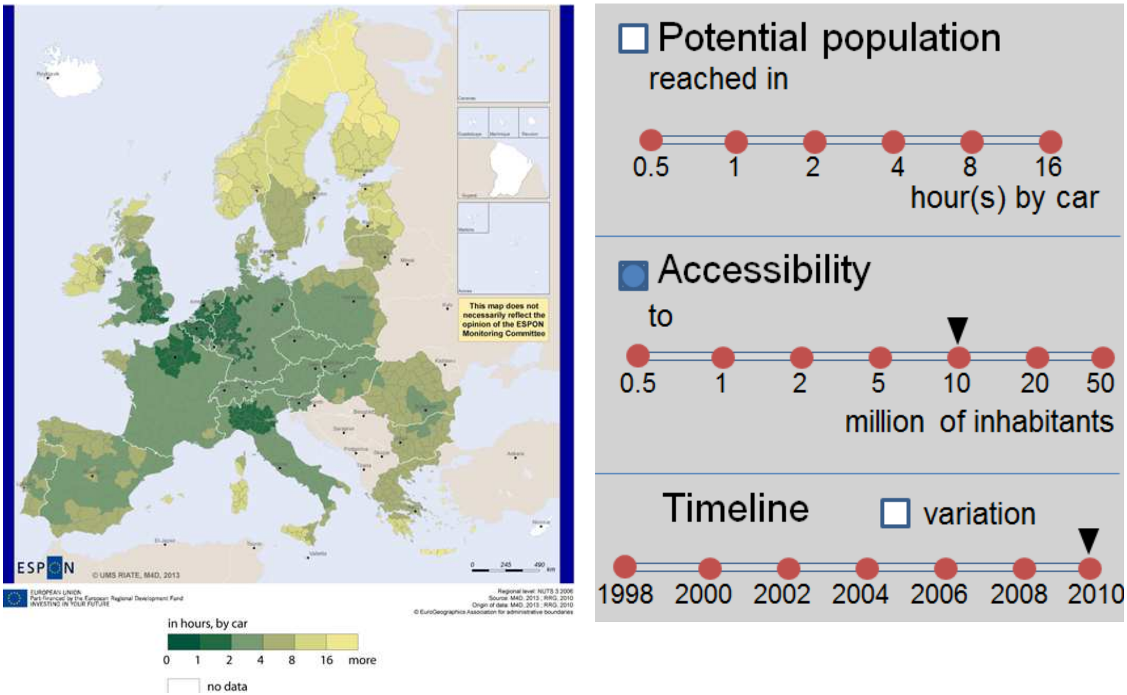


Figure 18 – Time needed to reach 10 million inhabitants

There are nevertheless three important conditions to generate these functional indicators (figure 19):

- **Development of time-series for count data**. Ideally, this time-series must be at a low geographical level (NUTS3) and must be regularly updated. Then, the work of time-series harmonisation is a huge work with a lot of methodological constraints. As a consequence, the selection, the collection and the harmonisation of time-series data is a huge work which must be considered as a project itself.
- **Acquisition and update of functional distance matrix between NUTS3.** The calculation of functional indicators cannot be delivered without having access to

the time-distance by road (but also rail, air, …) between each NUTS3. In the attempt done within the M4D Project, this has been made possible thanks to a contract signed out of ESPON between UMS RIATE and RRG[35]. In other terms, it means that the ESPON Community cannot explore the high potential provided by functional indicators without contracting with distance matrix providers. On top of that, it is important to remind that the distance matrix used in M4D is based on the 2006 transport network. To create time-series of functional indicators, it is better to update the distance matrix for present time and more generally to obtain time series of the functional distance between NUTS3 at regular time interval (1995, 2000, 2005, 2010,...).

- **Implementation of an interface for computation and request of these indicators.** These outputs can be computed through ad hoc computer programs written in R language by researcher of RIATE. But a computer interface could be certainly a more appropriate solution to manage heterogeneous information and to share this knowledge to non-specialists in computer programming.

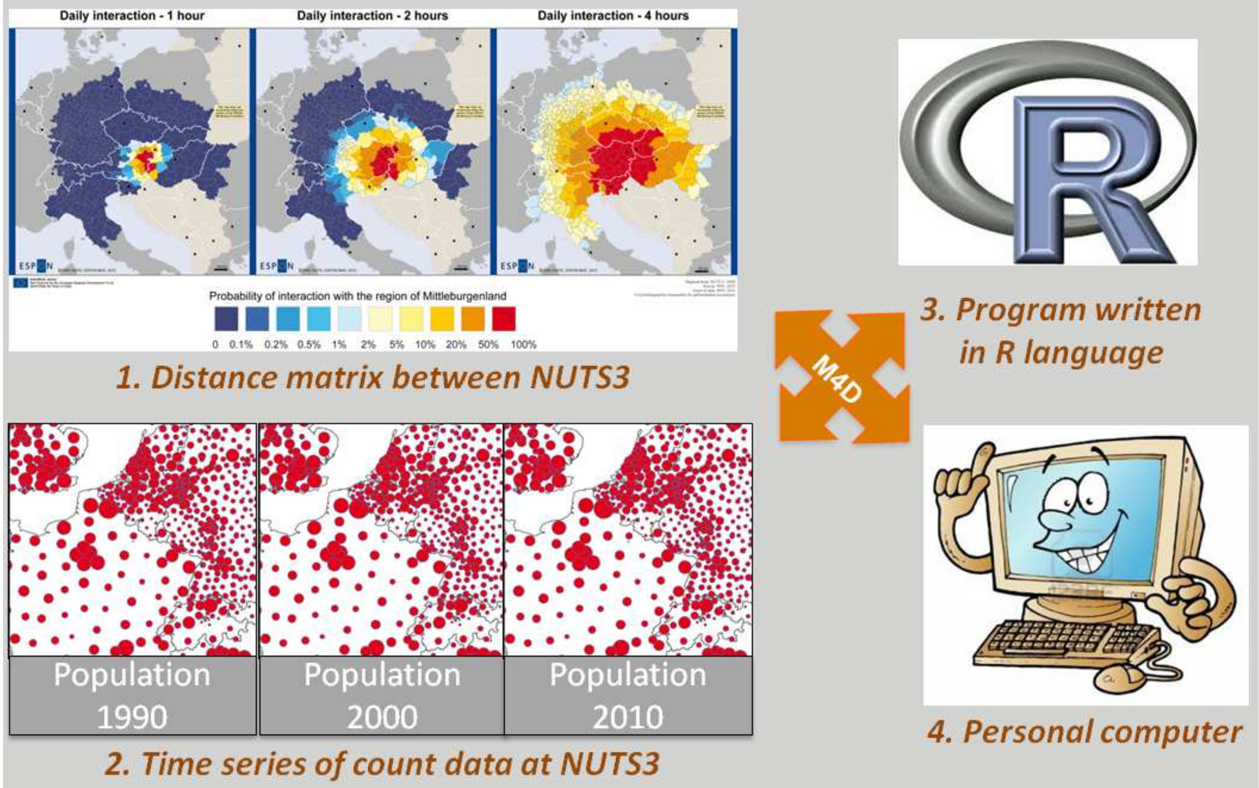

Daily interaction - 1 hour
Daily interaction - 2 hours
Daily interaction - 4 hours
Probability of interaction with the region of Mittleburgenland
0
0.1%
0.2%
0.5%
1%
2%
5%
10%
20%
50%
100%
1. Distance matrix between NUTS3
Population 1990
Population 2000
Population 2010
2. Time series of count data at NUTS3
M4D
3. Program written in R language
4. Personal computer

Figure 19 – Inputs needed to create functional indicators

Philosophical questions… Inside or outside the database?

The question of the integration of derivates from core indicators is a more generic question, not solved within the M4D project, but important to stress for the sustainability of the content of the ESPON Database in the future**.**

At the moment, each ESPON Project has to deliver its most representative indicators, the so-called "key indicators". But no general strategy regarding the content of the ESPON Database is assumed: to give a concrete example it is possible to find the total population 1990-2011 at each NUTS level (M4D core indicators), but it is only possible to query the population growth 1995-2005 (Map Update on demography). Based on M4D

[35] RRG GIS database, December 2010, Trans-European Transport Networks, NUTS3 Travel Time Matrices, ESPON Space, Cars & Lorries.

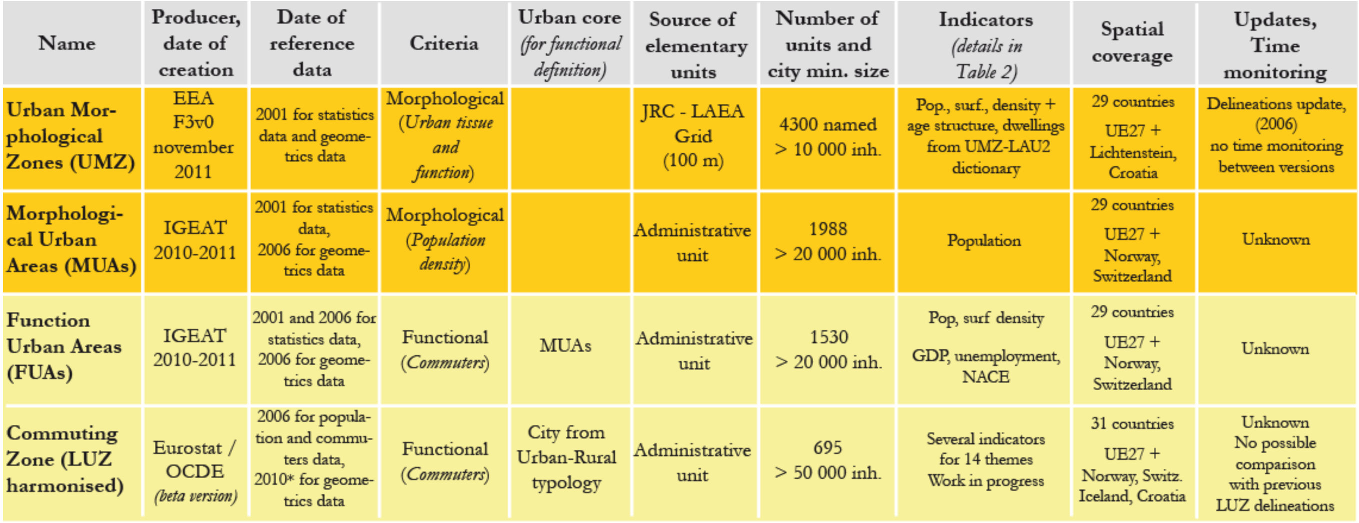
indicators, population growth 1991-1996, 2002-2008 can be computed also. This example is trivial, but it underlines the fact that it is possible to produce infinity of indicators derived from these core data. The situation is the same for functional indicators: do we have to store the potential population reached in 1 hour, 2 hours, 4 hours etc. in 1998, 1999 etc.?

The M4D project considers that ideally the ESPON Database should store a minimum of strategic information strongly verified and checked. Then and on the basis of this information available at different NUTS levels and for different time-periods, it should propose application using this core information to display indicators relevant for the policy debate. To be sure, we can imagine that we have decided to store in the ESPON database the amount of population in 1, 2, 3, 4 hours around each place. But a policymaker can consider that a threshold of 45 minutes or 1h30 is more adapted to its needs. Therefore, storing all thresholds does not make sense and it is more relevant to offer a specific interface making possible to compute in real time the indicator "*population located at a distance lower than X… minutes by road*" where X is interactively chosen by the final user.

2.2 Cities (WP B2)

2.2.1 Database integration: contribution to the Dictionary of spatial nomenclatures (M4D)

The M4D Dictionary of spatial nomenclatures is a new manual that is downloadable from ESPON DB portal and that explains the contents of the data bases integrated in that portal. The chapter 2 of this manual, which deals with cities, gives the main keys to help the ESPON users choosing the most appropriate urban DB, regarding their scientific targets. The first part provides a synthetic overview of the four urban databases selected, highlighting their complementarity in terms of urban definition, city size level, indicators or time monitoring (figure 20). In the second part, short presentations of each of these databases are exposed through a common analysis grid that helps comparisons between them and allows a comprehensible guidance for users.



Name	Producer, date of creation	Date of reference data	Criteria	Urban core *(for functional definition)*	Source of elementary units	Number of units and city min. size	Indicators *(details in Table 2)*	Spatial coverage	Updates, Time monitoring
Urban Morphological Zones (UMZ)	EEA F3v0 november 2011	2001 for statistics data and geometrics data	Morphological (*Urban tissue and function*)		JRC - LAEA Grid (100 m)	4300 named > 10 000 inh.	Pop., surf., density + age structure, dwellings from UMZ-LAU2 dictionary	29 countries UE27 + Lichtenstein, Croatia	Delineations update, (2006) no time monitoring between versions
Morphological Urban Areas (MUAs)	IGEAT 2010-2011	2001 for statistics data, 2006 for geometrics data	Morphological (*Population density*)		Administrative unit	1988 > 20 000 inh.	Population	29 countries UE27 + Norway, Switzerland	Unknown
Function Urban Areas (FUAs)	IGEAT 2010-2011	2001 and 2006 for statistics data, 2006 for geometrics data	Functional (*Commuters*)	MUAs	Administrative unit	1530 > 20 000 inh.	Pop, surf density GDP, unemployment, NACE	29 countries UE27 + Norway, Switzerland	Unknown
Commuting Zone (LUZ harmonised)	Eurostat / OCDE *(beta version)*	2006 for population and commuters data, 2010* for geometrics data	Functional (*Commuters*)	City from Urban-Rural typology	Administrative unit	695 > 50 000 inh.	Several indicators for 14 themes Work in progress	31 countries UE27 + Norway, Switz. Iceland, Croatia	Unknown No possible comparison with previous LUZ delineations

Géographie-Cités team, May 2013

Figure 20 - Which urban DB for which scientific target?

2.2.2 Populating ESPON urban database with local data

The delivery "Populating urban databases with local data", joined with the SIR, presents the detailed generic methodology for enriching an urban database (here, the Urban Morphological Zones) with local data (here, SIRE database 2008 from Eurostat). Three different steps are presented:

A) Join between the correspondence table and SIRE data

The join is based on LAU2 ID (from the correspondence table and from SIRE). Most of the unmatched cases are reduced by working on the ID codes or by taking into account the names of LAU2. Finally, only one hundred out of 23 000 LAU2 don't find a correspondence. An indicator of unmatched LAU2 has been created (number of unmatched LAU2 by UMZ).

B) Allocation and aggregation of SIRE data in each UMZ

The intensity of the link between one LAU2 and one UMZ is based on the population contribution (the area contribution has been tested but gives too much incoherent results, possibly because of the heterogeneity of the LAU2 area sizes in Europe). Three different cases are considered for aggregating the LAU2 contributions at the scale of one UMZ. a) There is only one UMZ inside the LAU2. Then, there is no need for aggregation. b) There are different UMZ inside one LAU2. Then, the populations of all these UMZ are added and this new total population is used for calculating the allocation/aggregation. c) The other UMZ. Then, the allocation is done on the basis of each LAU2 population contribution to the UMZ and all these contributions are aggregated.

C) Verification of the results

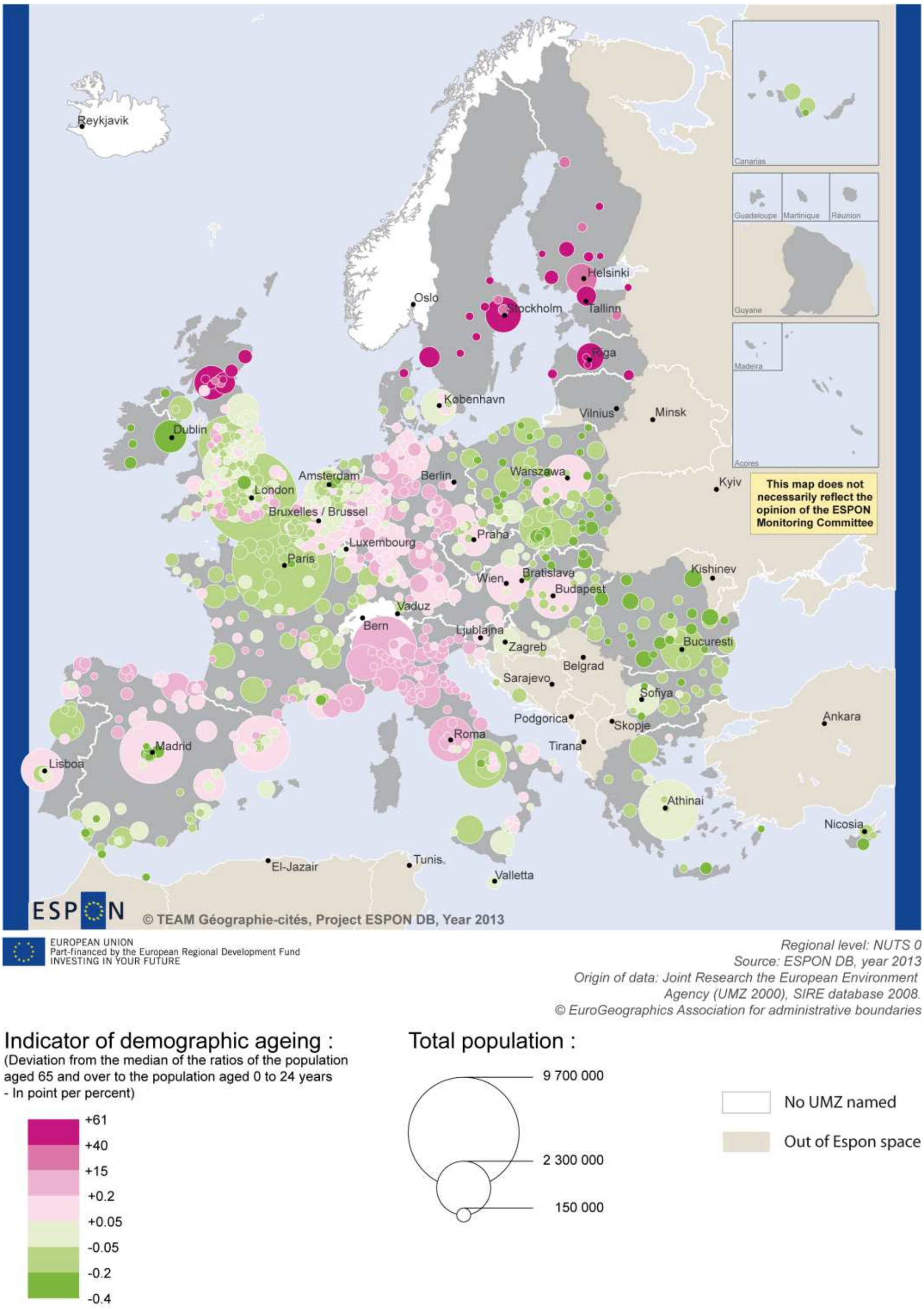
The verification is based on the population indicator, given first by the SIRE database and secondly by the population density grid (JRC). The population of each UMZ has been calculated using the populating method previously described and compared to the population given by JRC. A tolerance threshold of 10% has been chosen for qualifying the results. The results are good: only 95 UMZ (out of a total of 4304 UMZ) present deviations exceeding 10%. These results are expected for 45 UMZ that are concerned by missing LAU2. For the 50 remaining UMZ, most of the deviations are very concentrated near 10%, with exceptions mainly due to some inconsistency in the density grid (in Latvia, for unknown reason, and at the frontier with Switzerland where there is no data in the density grid). An indicator of validity has been created for these 95 UMZ.

This generic methodology has been tested on several indicators from SIRE database (education level, demographic data, housing…). The results depend on the quality of SIRE data. Some indicators are characterized by incompleteness or inconsistency (for instance, commuter data, education level, share of collective dwellings, see the detailed problems in the delivery). However, when the indicator sounds consistent, like the population by age classes, the results enlighten *the interest of populating urban data bases by using local data*. The results are markedly different from those generally mapped with nuts 3 levels (Figure 21, demographic ageing).

First, the information is much more precise, as the demographic indicator is not aggregated at Nuts3 level but in UMZ, most of them being very small and similar in size to one LAU2. Of course the map enlightens the national oppositions between countries with low demographic ageing indicator (and high fertility rates, like France, Ireland…) and countries with high one (Nordic countries, north western and central and eastern countries, Mediterranean countries). But this fine scale also allows enlightening interesting intra-regional contrasts, as the one between coastal regions ("rivieras") and interior ones (south Spain, South England, South France).

Furthermore, it makes much more obvious intra-national contrasts, as the one between Scotland and the rest of England, or between the north and the south of Denmark, of Italy and of Spain. Secondly, information concerning the urban hierarchy is not appearing on NUTS 3 maps whereas it is very striking here. The UMZ database covers the whole urban hierarchy (4300 cities larger than 10 000 inhabitants), which allows observing an opposition between small and large cities (especially capitals). The largest cities tend to be characterized by low demographic ageing indicator, probably because they are attractive for young adults, whereas the small cities of the rural counties or of the rivieras are characterized by more aged population, in particular retirees.

This generic method could be used for enriching urban databases (UMZ or other) with other SIRE indicators as soon as consistent local data will be available at European scale.



Regional level: NUTS 0
Source: ESPON DB, year 2013
Origin of data: Joint Research the European Environment Agency (UMZ 2000), SIRE database 2008.
© EuroGeographics Association for administrative boundaries

Figure 21 - Demographic ageing in European cities (UMZ) in 2008

2.2.3 Expertise on FUA construction methods

The definition of theoretical isochrones around city centres has been completed and refined through two important steps:

- a simultaneous work on 2 prototypes, Paris (Géographie-cités) and Barcelona (expert on transportation data, MCRIT), has provided the basis for a generic methodology. The MCRIT report on Barcelona is joined here as a delivery.
- an implementation to 9 other cities (Amsterdam, Firenze, Helsinki, Ljubljana, Madrid, Napoli, Praha, Stockholm, Wien) has been conducted in the context of Master theses and workshops and has contributed to improve this methodology.

A) Validation of ERM database consistency for this study

Following up on the work about transport data base comparisons (see December 2011 delivery), systematic comparisons between ERM road network and other national sources have proved the ERM database to be consistent for a study at this metropolitan scale. It also helped selecting the road levels that are useful for modelling isochrons. More specifically, this expertise helped to identify and record some missing data (Croatia) or some heterogeneity in data structure (the road hierarchy is not the same in Spain, for instance).

B) Methodology for modelling theoretical isochrones

Different choices have been made in the latest months in order to model theoretical isochrones around city centres:

- **City centre identification**: different methods have been compared and tested on a set of about 10 cities. Some of them could be automated (UMZ centroid, population density peak), but the most reliable one still remains the expert-based identification of a historical centre.
- **Shortest time travel paths computation**: the combined use of a road graph network and of a raster diffusion method has proved to be the most suitable method among others.
- **Speed parameters** estimation: **free-flow speeds** have been defined according to national legislations and to the road hierarchy.

Two main methods for estimating **peak-hour speeds** have been assessed and compared. The simplest one considers congestion as a discrete function of the distance to the centre and consists in implementing a peak-hour index inside the agglomeration (UMZ): this index has been first estimated by considering 50% free-flow speeds (see MCRIT delivery about Barcelona) and then refined by using congestion index defined by experts for each of the main European cities[36].

A second method currently tested for Paris, Helsinki and Stockholm (master theses) considers congestion as a continuous function of the distance to the centre. It relies on the construction of a congestion gradient that is based on the measure of different travel times for routes converging towards the city centre (figure 22), thanks to route calculation websites. It could help to identify either thresholds in the variation of congestion or constant parameter to implement more realistic congestion indices.

[36] TomTom Europen Congestion Index, 2013 : http://www.tomtom.com/lib/doc/congestionindex/2013-0129-TomTom%20Congestion-Index-2012Q3europe-mi.pdf

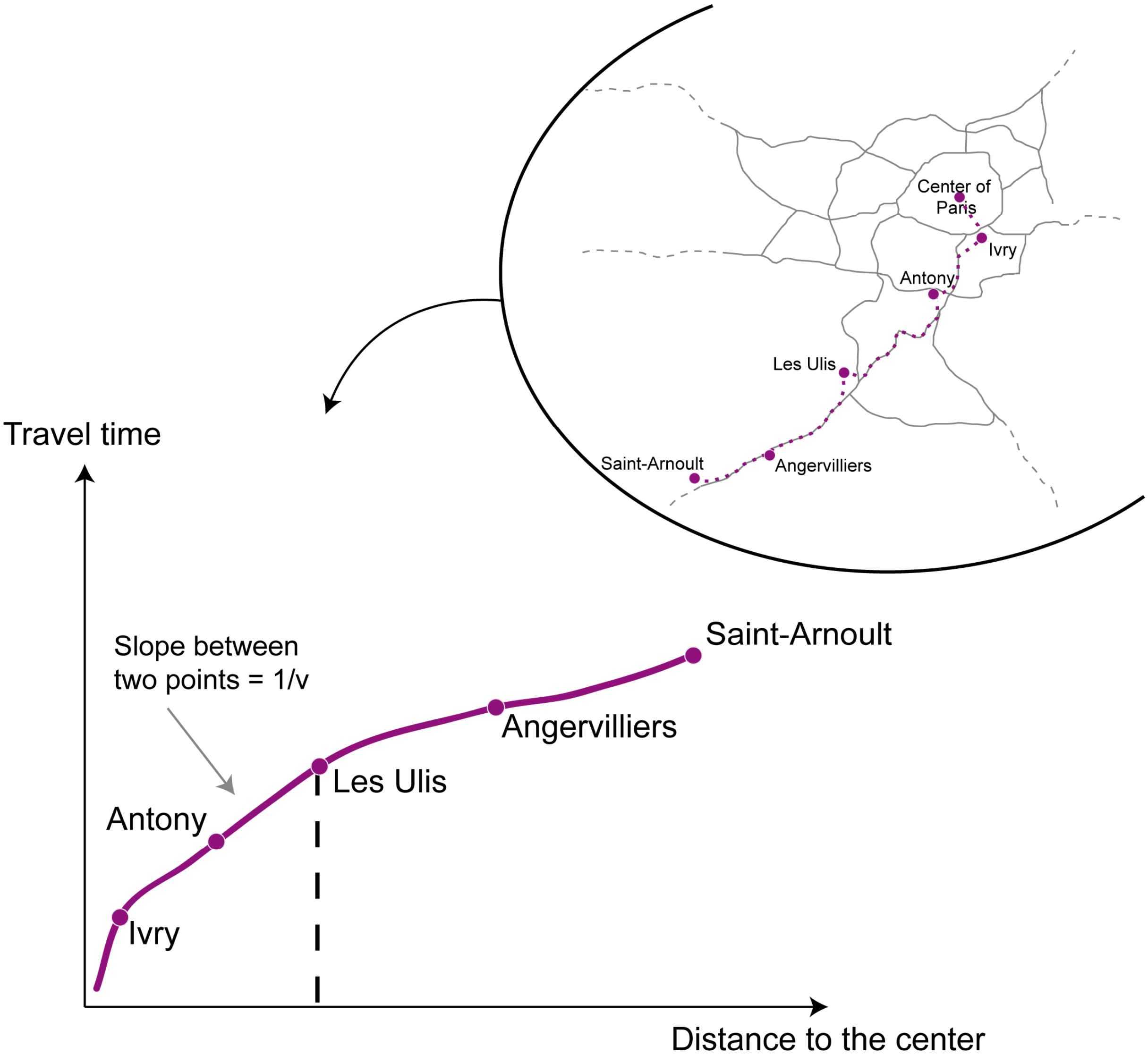


Figure 22 - Congestion gradient, from description to model
Source: Mathian, Pavard 2013

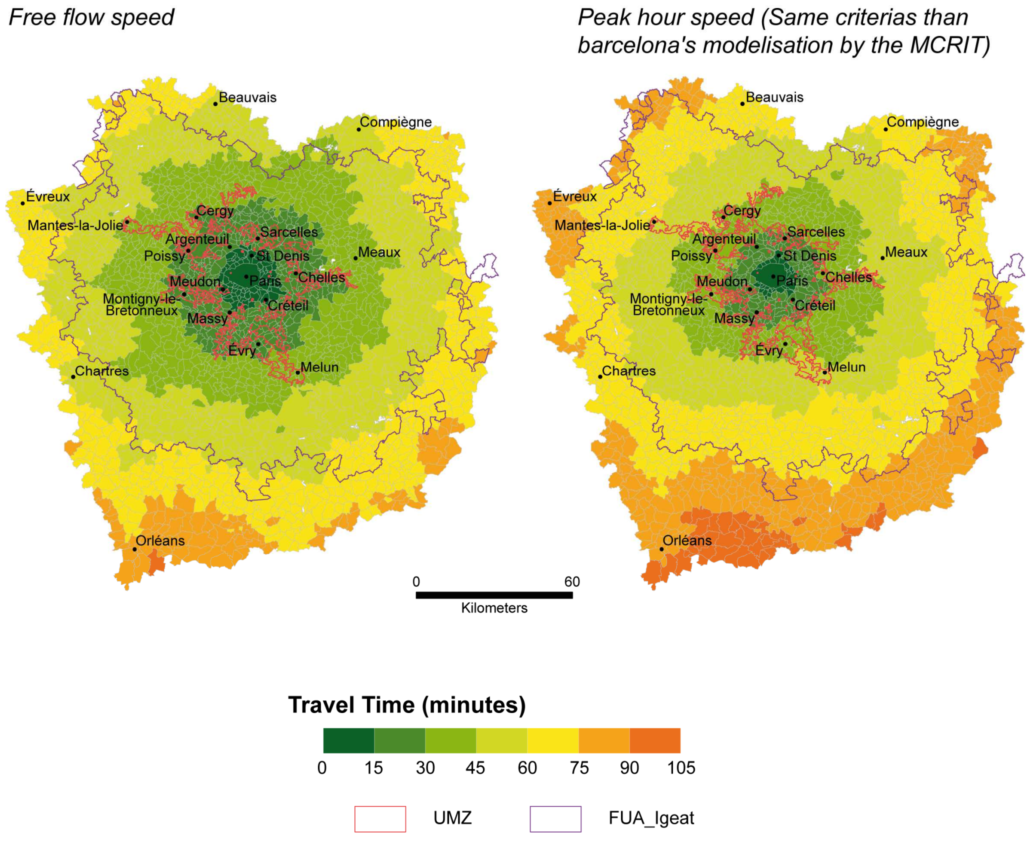
C) Isochron-based delineation as compared to other urban delineations

The comparison between the resulting isochron-based delineation and other urban delineations helped to better assess the results associated to that model.

A first assessment lies on the comparison of 1 hour isochrones resulting from the peak-hour model (with the simplest estimation of congestion) and a set of observed time travels in real peak-hour conditions, provided by route calculation websites. In the Paris case, the observed time travel is under-estimated of about 60% by our model for a set of 20 destinations distant from 40-50 km to the centre. This deviation led us to refine the estimation of congestion (see previous point B).

A comparison between isochrones results for Paris and Barcelona on one hand, and urban statistical delineations (UMZ, FUA_IGEAT) on the other hand, helped us to reconsider the estimation of maximal time budget for a centre-periphery travel, from 1 hour to 1h15 or 1h30 (Figure 23).

An important work has been initiated for 9 cities in order to compare isochron-based delineations with commuter-based delineations. Many difficulties were due to SIRE inconsistencies in commuters data (see June 2012 Deliveries) and to missing data in LAU2-SIRE dictionary. This work is still in progress.



Source : UMZ of 2000 (EEA, 2011), LAU2 of 2001 (GISCO, Eurostat 2001), FUA of 2006 (D. Peeters, IGEAT), EuroRegionalMap d'Eurogeographics, V3.1, 2011

Antonin Pavard

Figure 23 - *Modeling theoretical isochron in Paris: from free flow speed to elementary estimation of congestion*

2.2.4 Time series on urban data

The delivery "Time-series on urban data", joined with the SIR, presents the different choices that are available when building an urban database for integrating time but also the sources of variability that are internal to the building process. We first recall how it is complex to conceptualize an urban object that evolves through time:

- a city is the result of a dynamic process of concentration, which leads to a sprawl from the centre to the periphery, so that its geometrical delineation is changing over time.
- the city definition evolves over time, based first on morphological considerations (continuity of the built-up area), then based on employment polarization (commuters).
- a city is an aggregation of elementary entities for which statistics exist, and which evolve through time

We present in a second part an illustration of the different choices that precede the building of a time integrated urban database. These choices may be categorized (Figure 24) according to the reference selection for time harmonization. They have some repercussions on the way the database may evolve in the future. For instance, in case b) and c), the whole database will be re-evaluated at each new update. Instead of d) or e), where trajectories integrate the different phases of evolution and may integrate directly new updates.

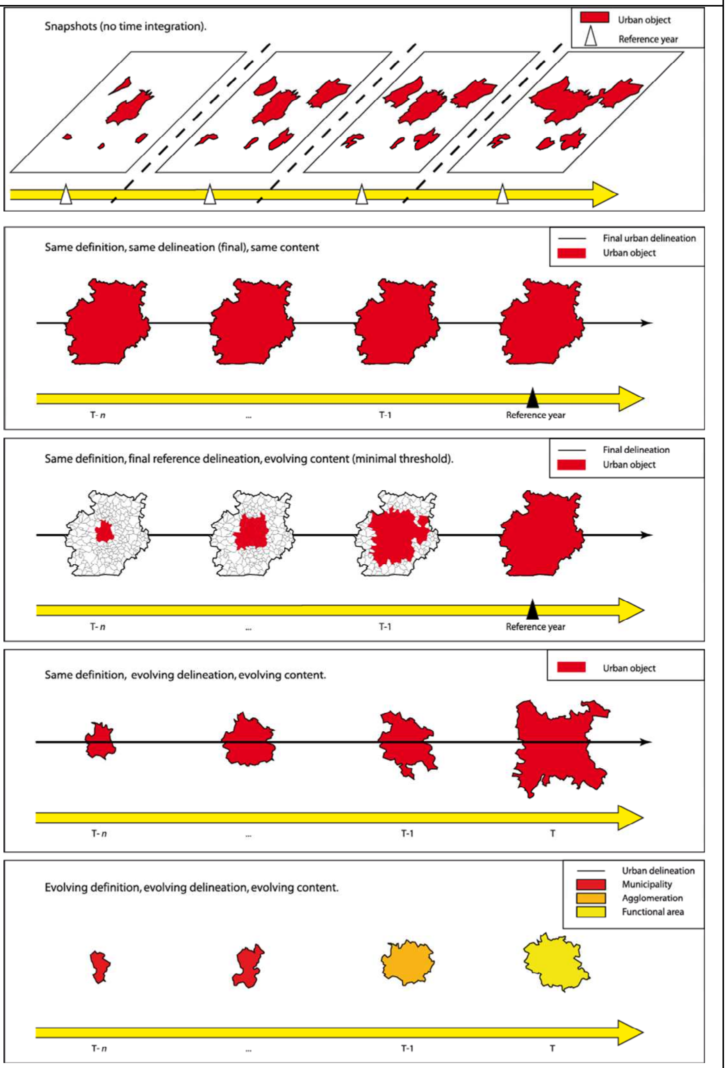
Type of time integration	Illustration
(a) **Sources driven:** the sources are the reference of the database (example of UMZ data base). Each survey gives a global image of urban pattern, without constructing urban objects that may be followed over time	 Snapshots (no time integration). Urban object Reference year Same definition, same delineation (final), same content Final urban delineation Urban object T-n ... T-1 Reference year Same definition, final reference delineation, evolving content (minimal threshold). Final delineation Urban object T-n ... T-1 Reference year Same definition, evolving delineation, evolving content. Urban object T-n ... T-1 T Evolving definition, evolving delineation, evolving content. Urban delineation Municipality Agglomeration Functional area T-n ... T-1 T
(b) **Final delineation driven:** the composition does not evolve over time	
(c) **Final delineation driven**: Semantic of the building blocks differentiates the composition;	
(d) **City entity driven**: delineation is built according to the semantic of the building blocks and centred on a defined and named urban entity. The semantic relation doesn't evolve.	
(e) **City entity driven**: delineation is built on aggregation of building blocks and centred on a defined and named urban entity, but semantic relation evolves over time.	

Figure 24 - Types of time integrations according to the definitions of urban objects

In the third part, we present the different sources of variability internal to the building process of the urban database, after phase of choosing the definition of the city. When building an urban database *in a diachronic perspective*, given the same model, time will affect the sources and the parameters. An illustration is given for the UMZ of Wien, in 2000 and 2006: a very local change in the sources (change of interpretation of one or two pixels in the south of the city) makes a global change in the city delineation (large reduction in the size of the city).

In conclusion, we recall the gap between the thematic questions raised by users, related to urban dynamics, and the answer possibilities given by the structures and contents of the urban databases. If we take, for instance, some frequently asked questions, such as "Is the urban sprawl intensifying in the two last decades or no?" or "can we estimate the rate of evolution of urban shrinking in Europe", or more simply "Is the urbanization level still in an exponential growth or in stagnation in the three or four last decades", we have to note that the answers are not simple. They will not only depend on the availability of the data (population of building blocks at different dates, urban delineations covering the whole countries in a harmonized way) but also on the choice in urban definition (regarding time integration), and on the variability internal to the building process of the database.

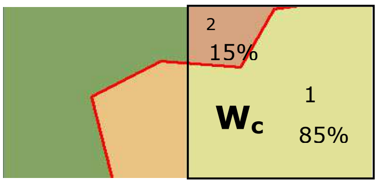
2.3 Grid and OLAP Cube (WPA3&B3)

2.3.1 Thematic part

OLAP Cube integration improvements

The process of building OLAP Cubes requires the use of heterogeneous tools and trained technicians in GIS technology and Databases. In most cases, it is a time-consuming task with an important implication of technical work. In order to improve this process, we have performed some tests by implementing part of the process in Geokettle[37], an Open Source Geospatial Business Intelligence ETL tool dedicated to the integration of different spatial data sources for building and updating geospatial data warehouses.

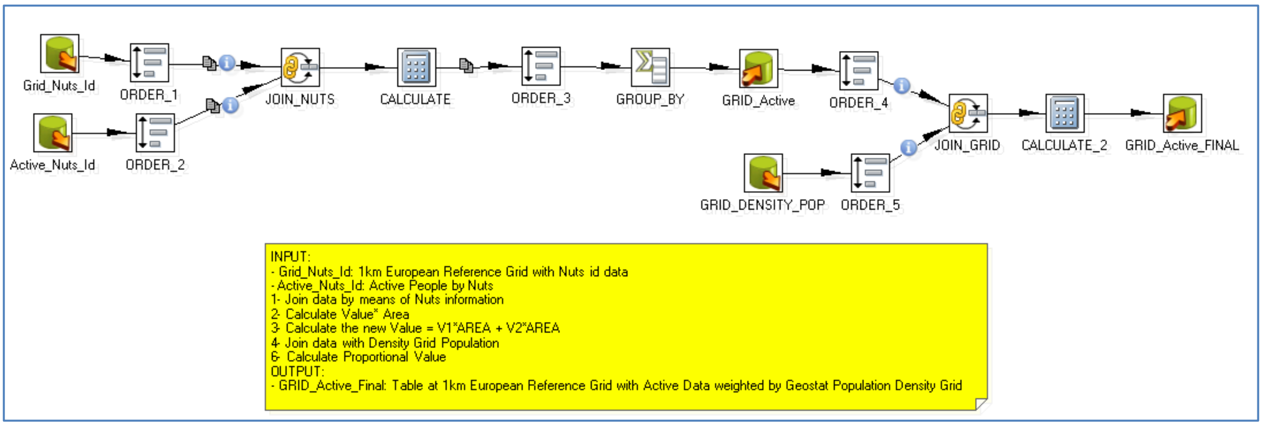
By means of user-defined transformations and jobs, Geokettle allows you to extract data from different platforms, transform the data and load it into different target formats. For sample, the proportional and weighted calculation, as it is shown in the next schema, can be implemented in a unique platform by creating a Geokettle transformation.



Cell value = Wc Σ (Vi * Sharei)

Vi = Value of unit i
Sharei = Share of unit i within the cell
Wc = weight assigned to cell c

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



INPUT:
- Grid_Nuts_Id: 1km European Reference Grid with Nuts id data
- Active_Nuts_Id: Active People by Nuts
1- Join data by means of Nuts information
2- Calculate Value* Area
3- Calculate the new Value = V1*AREA + V2*AREA
4- Join data with Density Grid Population
6- Calculate Proportional Value
OUTPUT:
- GRID_Active_Final: Table at 1km European Reference Grid with Active Data weighted by Geostat Population Density Grid

Figure 25 - Sample of transformation in Geokettle

As conclusion from the tests carried out, we can say that Geokettle has proven to be a useful tool to implement heterogeneous processes using a unique platform. It makes

[37] http://www.spatialytics.org/projects/geokettle/

easier to control all the processes and allows you to have the final results directly in Microsoft SQL without much technical work. Once the transformation or job is created it can be easily launched by any technician.

On the contrary, the heterogeneity of ESPON data and the addition of new indicators make quite difficult the creation of fixed transformations or jobs, and they have to be personalized before launching them. In addition, spatial tasks as spatial intersection have shown some limitations due to the great extension of data. In these cases, it is better to work directly with specific GIS software.

Advanced options as clustering or portioning has not been fully tested from the moment. New tests are required to finally set the most suitable platform to ease the creation of ESPON OLAP Cubes in terms of time-consuming and technician work.

New Updated Cube: ESPON OLAP Cube 6.0

A new updated Espon OLAP Cube has been delivered in June 2013 focused on urban geographical dimensions and the newest population statistics. The content of the ESPON OLAP Cube v. 6.0 is:

Measures:
Area in hectares
GDP 2000 Million Euros
GDP 2003 Million Euros
GDP 2006 Million Euros
GDP 2009 Million Euros
Population 2000 thousand inhabitants
Population 2003 thousand inhabitants
Population 2006 thousand inhabitants
Population 2009 thousand inhabitants

Spatial dimensions or LARUs (Land Analytical and Reporting Units):
CLC00 Hierarchical
CLC06 Hierarchical
CLC90 Hierarchical
Land Cover Flows 1990-2000
Land Cover Flows 2000-2006
Land Cover Flows 2000-2006
Nuts 1999 code
Nuts 1999 name
Nuts 2003 code
Nuts 2003 name
Nuts 2006 code
Nuts 2006 name
Nuts 2010 code
Nuts 2010 name
Functional Urban Areas (FUA)
Morphological Urban Areas (MUA)
Urban Morphological Zones (UMZ)
Large Urban Zones (LUZ)

Beside the ESPON OLAP Cube, the statistics of CLC2006 classes by NUTS 2010 breakdowns have been computed and included in the ESPON Database.

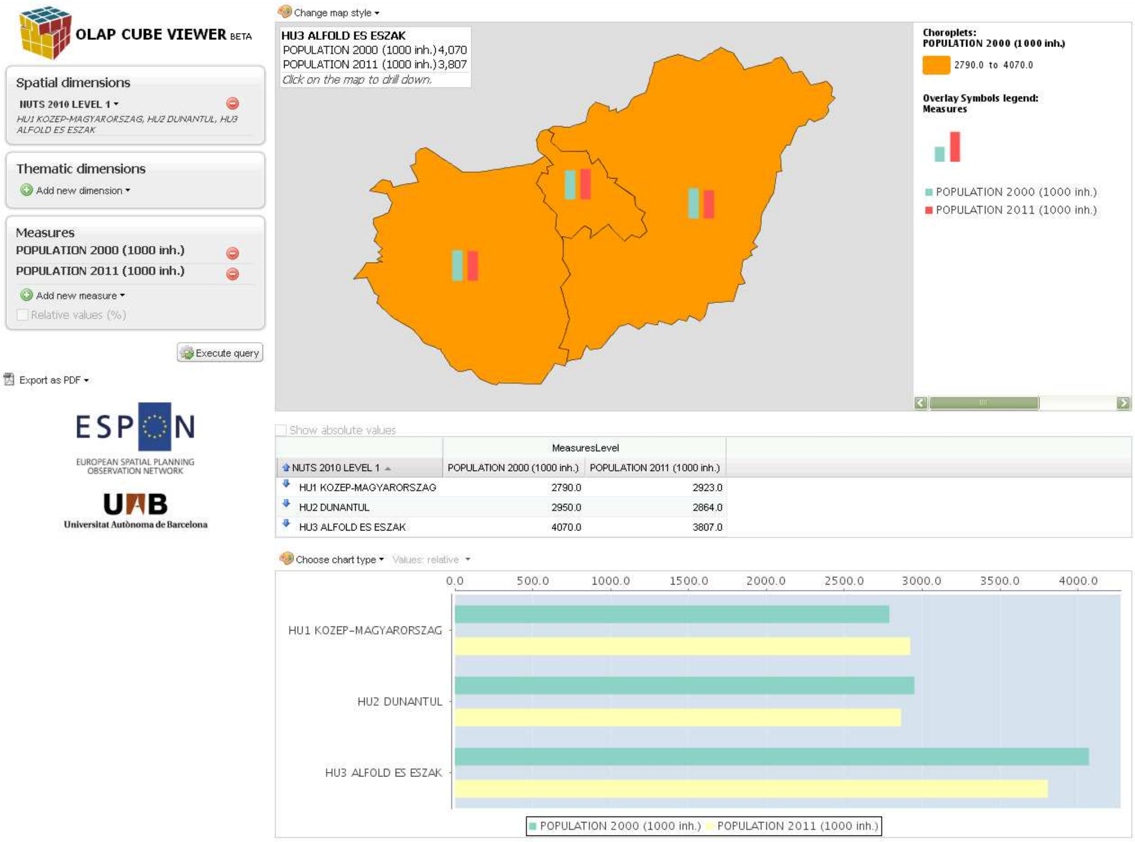
2.3.2 Technical part / Webtool

User friendly visualisation olap tool improvements

The web-based OLAP cube viewer has been steadily improved during this period, focused on improving the user experience, adding new possibilities and offering updated statistics.

The tool has been updated and connected to the last available offline ESPON cube (v. 5.0), which included population series from 1990 till 2011, all existing NUTS breakdowns (1999, 2003, 2006 and 2010) and Large Urban Zones. At the same time, an effort has been done to simplify the presentation of this cube on the web tool, which envisaged as a intuitive and simple access for the cubes. For this reasons, some OLAP dimensions have been hidden (Massifs, Biogeographical Regions, Sea Basins, etc) and could be offered as separate, thematic online cubes when appropriate.

As new measures (population) have been added to the cube, new requirements for the tool have emerged, such as the possibility to show different measures (e.g. the population for different years) on a single query, which was not necessary when the only measure was the area (hectares). The tool has been modified to satisfy this requirement, as illustrated on the snapshot. This feature enables the comparison of the population for different years using a specific reference NUTS version (which was potentially not available on the displayed population reference dates).



Spatial dimensions
NUTS 2010 LEVEL 1
HU1 KOZEP-MAGYARORSZAG, HU2 DUNANTUL, HU3 ALFOLD ES ESZAK

Thematic dimensions

Measures
POPULATION 2000 (1000 inh.)
POPULATION 2011 (1000 inh.)

□ Relative values (%)

HU3 ALFOLD ES ESZAK
POPULATION 2000 (1000 inh.) 4,070
POPULATION 2011 (1000 inh.) 3,807
Click on the map to drill down.

Choroplets:
POPULATION 2000 (1000 inh.)
2790.0 to 4070.0

Overlay Symbols legend:
Measures
POPULATION 2000 (1000 inh.)
POPULATION 2011 (1000 inh.)

□ Show absolute values

	MeasuresLevel	
NUTS 2010 LEVEL 1	POPULATION 2000 (1000 inh.)	POPULATION 2011 (1000 inh.)
HU1 KOZEP-MAGYARORSZAG	2790.0	2923.0
HU2 DUNANTUL	2950.0	2864.0
HU3 ALFOLD ES ESZAK	4070.0	3807.0

Figure 26 - Population in Hungary at NUTS level 1 for years 2000-2011 computed using NUTS version 2010

Some restrictions and checks have been applied to the Query Builder, in order to guide the user to useful results. For example, the No Data values are silently removed from the query as they can't be represented on the map. Similarly, a warning message is shown when too much members (rows and columns) are included in the query, as this kind of queries produces unreadable maps and slows down the application. In any case, the user will be able to decide whether the query should be performed anyway or not. A generic framework has been programmed in order to easily create new restrictions or warnings when required, enabling each produced cube to have customized limits.

The rendering of the map has also been improved, especially for small polygons such as islands, producing clearer maps which are easier to understand. The context countries are also included in this new version, which also increases the readability of the map.

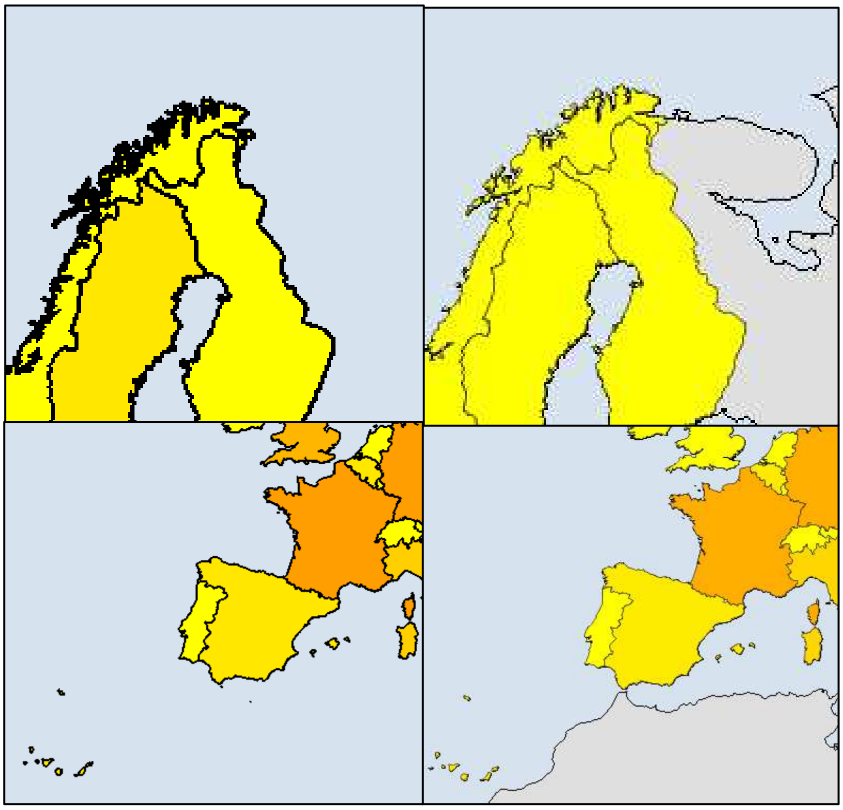
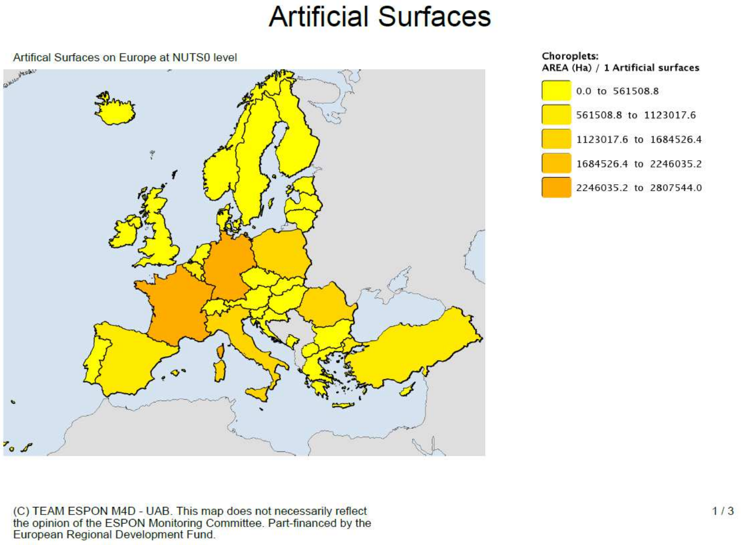


Figure 27 - Map rendering on 2012 version (left), map rendering on 2013 version (right)

The printing template has been customized to include ESPON references and disclaimers. At the same time, an error was corrected that produced the printing tool to fail on some situations.



Artificial Surfaces

Artifical Surfaces on Europe at NUTS0 level

Choroplets:
AREA (Ha) / 1 Artificial surfaces

- 🟨 0.0 to 561508.8
- 🟨 561508.8 to 1123017.6
- 🟨 1123017.6 to 1684526.4
- 🟧 1684526.4 to 2246035.2
- 🟧 2246035.2 to 2807544.0

(C) TEAM ESPON M4D - UAB. This map does not necessarily reflect the opinion of the ESPON Monitoring Committee. Part-financed by the European Regional Development Fund.

1 / 3

Figure 28 – Global overview of the map rendering on 2012 version (left), map rendering on 2013 version (right)

On the technical side, some open source software libraries used in the tool have been updated to latest versions, which corrects some internal errors and provides small performance improvements.

2.3.3 Future steps (until the end of the year)

Regarding the OLAP Cube and thematic contents

- Update of the OLAP Cube with new data and/or LARUs.
- Creation of an Urban OLAP Cube at a 100 m resolution to test its feasibility and usefulness.
- Further tests with advanced options of Geokettle will be undertaken to optimise the Cube production method.
- Enlargement of the Reference Grid to the neighbourhood, and harvesting with basic socioeconomic indicators (with the support of UNEP/GRID).

Regarding the webtool

Some improvements are considered relevant for ESPON M4D but could not be implemented till now. Therefore, they will be implemented for the end of the year:

- Ability to select the viewer components (chart, map, table) to be shown when executing a query.
- Improvements on the chart and table visualization (using pagination), as current visualization is suboptimal for big queries.
- Further improvements on symbolization will be performed in order to better adapt to Espon Mapping Guide (within the limits provided by the architecture and purpose of the application).
- Connection to the ESPON OLAP Cube v. 6.0. (foreseen for July 2013)
- The possibility to have several cubes available on the online tool.

2.4 Local data (WP B4)

Objectives

The workflow during this period focused on two major topics: the creation of an alternative geometry for the ESPON space and the production of relevant indicators at local scale. The need for an alternative derives from the problems produced by data representation at NUTS3 and/or LAU2 scale. These problems can be solved using a geometry that conciliates NUTS3 of relevant size with aggregated LAU2. The new geometry can be intersected with different layers and indicators and it will help better data visualization. The second objective is to provide indicators that describe the relation between the VIGO and the LAU2, in the ESPON area. As we don't intend to create accessibility indicators for the geographical objects of major importance, we describe in the text what kind of indicators can be derived by the analysis of this relation.

For the Draft Final Report, we expect to deliver 2 versions of the alternative geometry, one based on a location-allocation model and one based on the shortest distance towards LAU2 with central functions in their NUTS3. The list of indicators will also be enriched: relation between LAU2 and major nodes of transportation, universities, protected areas etc.

Work done

The differences between different administrative frames in the ESPON space reflect historical, political and natural backgrounds of the human action, shaping the geometries we see today on the maps. These differences are interfering with the mapping process of all kind of data, generally by a frustrating effect of mass (surface). Dealing with this mass effect is not an easy task. Comparing on the same map the NUTS3 information from France with the NUTS3 information from Germany could lead to conclusions and interpretations that are sensitive to this area effect. The same problem occurs at basic levels of territorial research - the LAU2 scale. For example, working with the LAU2 level of France means managing more than 36 000 spatial units. It is not the same situation in Norway, Denmark or Poland, countries where the number of LAU2 is reasonable (less than 3000). The solution to this set of problems is based on regionalization techniques that will partially eliminate the mass (area) disturbances, proposing instead an alternative geometry. Our intention was to aggregate the LAU2 level in new spatial units, for all the European countries, using a reduce number of methods, excluding the expert opinion in the selection, providing an alternative geometry with a low mass effect.

Since the beginning of the modern quantitative geography, the regionalization problems and algorithms occupied a central place in the academic debates. More than a scientific curiosity, the practical dimension of the regionalization acted like a magnet for different groups of interest in the debate (policy makers, planners, geographers, economists etc.). The searches for administrative efficiency, the need to create electoral districts based on equity, resource allocation at various spatial levels are the major motivations of regionalization exercises. The solutions we resume in this text are not numerous:

a. expert opinion. It is a solution that combines data and spatial analysis with the "geographical knowledge" of different places and regions. This approach cannot be applied to the large number of LAU2 functioning in the ESPON space.
b. the hierarchical clustering with a double constraint : contiguity and territorial belonging (NUTS3). This solution is technically difficult (time consuming) to apply it for all the ESPON area. In order to work, we need to analyze a dissimilarity matrix of 120 000 by 120 000 rows and columns weighted with a contiguity matrix that takes into account the NUTS3 of each LAU2. The final amount of data will be reduced to approximately 700 000 rows and 4 columns. It is not an impressive mass of information; the problem is that the aggregation of data is not feasible in a GIS (linkage and clustering). For some countries in the database, we have reserved this solution as an alternative to our routine.
c. the seek for functional regions constrained by NUTS3 belonging. The last model is a country by country approach of the problem. As we need to aggregate the LAU2 in larger units (pseudo-LAU1) and to respect the NUTS3 limits, the task is reduced to the search of candidate centers of aggregation. The distance between the LAU2 and the candidate centers will design the new perimeters of the merged LAU2 contours, as any LAU2 will be allocated to a single candidate. The number of candidates selected by NUTS3 will provide the number of the LAU2 allocated to each center. Basically speaking, many candidates means small pseudo-LAU1 regions, few centers produce large regions.

In order to provide an alternative geometry, situated in an intermediate scale between the NUTS3 and the LAU2, we have made an option for the solution no. 3. There are two reasons for this choice: the combinatory potential of the resulting regions and the replicability of the solution from one country to another. The application of the solution and the algorithm are subject to some steps. These steps will be presented using the following approach: question, answer, justification.

Step 1
Q1: What is the optimum size of a pseudo-LAU1 unit?
A1: As surface, it should be 1000 sq. km.
J1: Comparing the size of the European NUTS3, we observe and homogeneous region formed by Germany, the Netherlands, Belgium and Luxembourg. These NUTS3 cannot be divided in smaller areas. What we can do is to merge LAU2 from the countries in an alternative geometry aiming to this surface. A combination between the NUTS3 from these 3 countries with the pseudo LAU1 will provide the first version of the alternative

geometry. There will be exceptions to this objective, basically in the Scandinavian countries (the Northern parts).

Step 2

Q2: Knowing the size of one alternative pseudo-LAU1, how should it be found the proper number of candidate centres?

A2: By dividing the country surface to 1000 km. The result is rounded. For example, a country about 246 000 sq. km. will have 250 candidate centres.

J3: It is impossible to obtain perfect surfaces of 1000 km, taking into account the variety of the LAU2 geometry. Putting 246 centres is a risk. What if, for some centres, the number of allocated LAU2 will be very reduced?

Step 3

Q3: How these centres will be chosen? How can we avoid that all the 250 centres are optimally located and homogeneously dispersed on the territory?

A3: The solution we propose is to choose these centres using a location-allocation model that maximize coverage for the facilities (candidate centres) in a 60 minutes limit.

J3: The method we proposed provides good results for different kind of facilities that need to have in their catchment area as many clients as possible and not to compete between them. The other solutions to select the centres - expert opinion, a rank-size approach at NUTS3 level, weighted accessibility models. The expert-opinion model is unreliable for this amount of LAU2 units, the rank-size delineation of candidates presents the risk of spatial concentration of candidate centres (all in the centre of the NUTS3, especially in mono-centric metropolized NUTS3) and the accessibility models are time consuming.

Step 4

Q4: The centres were located. What next?

A4: As all the LAU2 where allocated to a centre, the next step is to dissolve their internal limits (within the pseudo-LAU1) and merge them in a new brick of the alternative geometry. We verify for errors and eventually correct them.

J4: The errors are consubstantial to the research process. Bad location in the network, strange LAU2 shape, extra-territorialities, all can happen when we deal with more than 120 000 spatial units. As a matter of fact, no error is a sign that something is not right.

Step 5

Q5: The centres were located and the error sources evacuated. What next?

A5: Possessing a reliable alternative geometry, we intersect the new layer with basic LAU2 indicators, such as population, density, land cover etc. We obtain new data that we can map.

J5: Our objective is to obtain an alternative geometry, in order to verify if the mass effect is evacuated and if this new geometry is more versatile as a basemap. Basically, this new layer should function as a superior vector of information dissemination to policy makers and decision takers. Both elements should be tested.

Step 6

Q6 : How can one be sure that this alternative geometry serves better than the LAU2 frame and the NUTS3 basemap? How can we measure the mass (area) effect?

A6 : One simple way is to imagine a coefficient that can be related to a set of hypotheses that describes the mass effect. Like in the spatial autocorrelation coefficient, there are three hypotheses to expose:

a. Spatial units having the "same" size will present similar values of the mapped indicators. The value of the areal autocorrelation coefficient will be 1. Mass effect (area) is present and explains similitude between spatial units (pseudo-LAU1).

b. Spatial units having the "same" size will present dissimilar values of the mapped indicators. The value of the areal autocorrelation coefficient will be -1. Mass effect (area) is present and explains differences between spatial units (pseudo-LAU1)

c. Spatial units having the "same" size will present dispersed values of the mapped indicators(neither similar or dissimilar). The value of the areal autocorrelation coefficient will be 0. Mass effect (area) is theoretically absent and don't explains similitude or differences between spatial units (in our case, pseudo-LAU1). It looks like the Pearson's r in its aspect but it is just a formal illusion. In the case of the evolution of population between 2001 and 2006, for the alternative geometry of France, we have obtained a value of 0.19. About 20 % of the demographic trends can be explained by the size of the spatial units, all other things being equal or ignored.

J6: Measuring the area autocorrelation coefficient at different scale (LAU2, pseudo-LAU1, NUTS3) could work as a testing method for robustness of the alternative geometry. For example, it would be interesting to measure, for a single indicator (density of population or demographic evolution), how this coefficient varies with scale and area. In that case, we would be able to precise which administrative and geographical scale induces problems of map interpretation. However, this exercise involves matrix calculus based on 36 000 * 36 000 rows and columns, only for France LAU2 frame, demanding particular technical solutions.

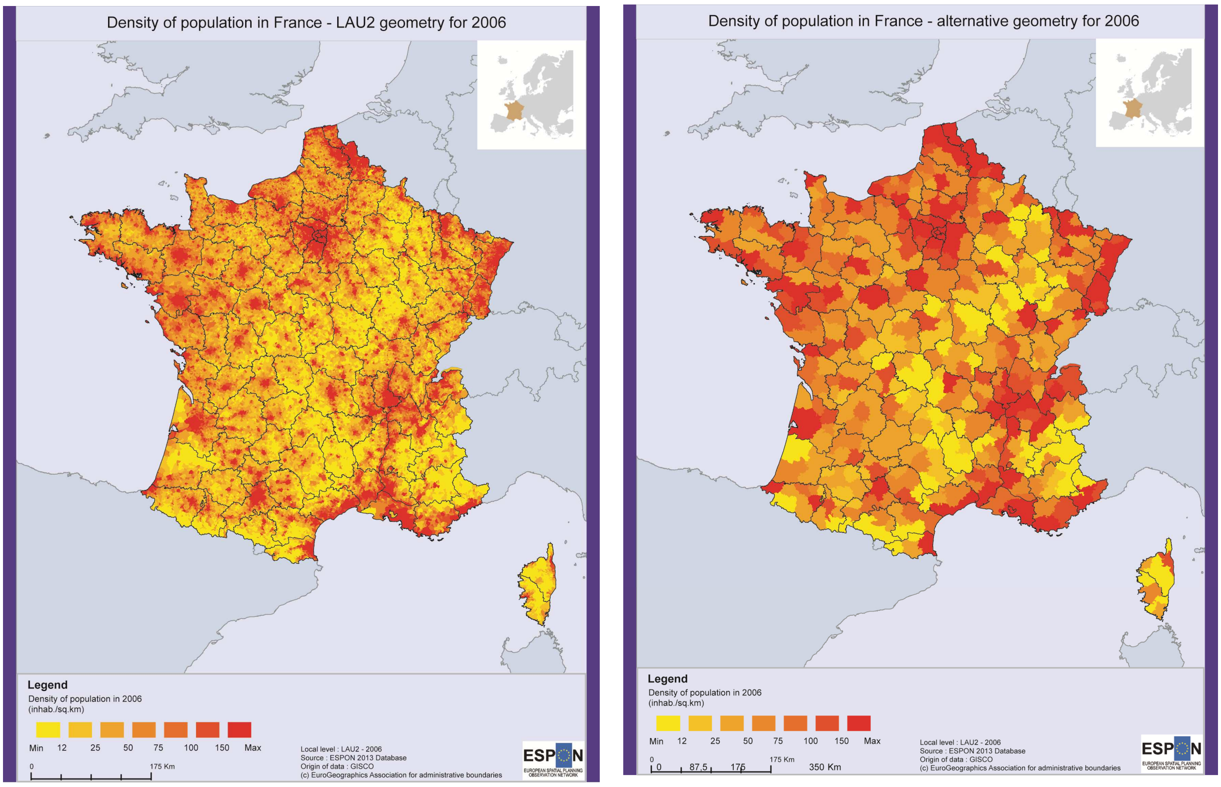

Density of population in France - LAU2 geometry for 2006
Legend
Density of population in 2006
(inhab./sq.km)
Min
12
25
50
75
100
150
Max
0
175 Km
Local level : LAU2 - 2006
Source : ESPON 2013 Database
Origin of data : GISCO
(c) EuroGeographics Association for administrative boundaries
ESPON
EUROPEAN SPATIAL PLANNING OBSERVATION NETWORK
Density of population in France - alternative geometry for 2006
Legend
Density of population in 2006
(inhab./sq.km)
Min
12
25
50
75
100
150
Max
0
87.5
175
175 Km
350 Km
Local level : LAU2 - 2006
Source : ESPON 2013 Database
Origin of data : GISCO
(c) EuroGeographics Association for administrative boundaries
ESPON
EUROPEAN SPATIAL PLANNING OBSERVATION NETWORK

Figure 29 - Density of population at LAU2 scale (France, 2006)

At this moment, we have managed to build the alternative geometry for almost all the countries in the ESPON area. The exceptions are Bulgaria, Greece, Cyprus, Malta, Switzerland, Island, Norway, Sweden and Finland. In the last four cases, a hierarchical clustering constrained by contiguity is a better option of regionalization than a location-allocation model. Bulgaria and Greece demand a different approach. The lack of a proper network (Bulgaria) and the insularity specificities of Greece demand prudence.

For the Eastern countries of the European Union, the method we proposed seemed to work properly. Adjustments might be needed in Hungary (Hajdu-Bihar) because the average distance between the LAU2 is considerable and interfere with the model of regionalization. In Poland we might need to rethink the role of the metropolitan areas in the delineation of the alternative geometry, probably by merging several pseudo-LAU1.

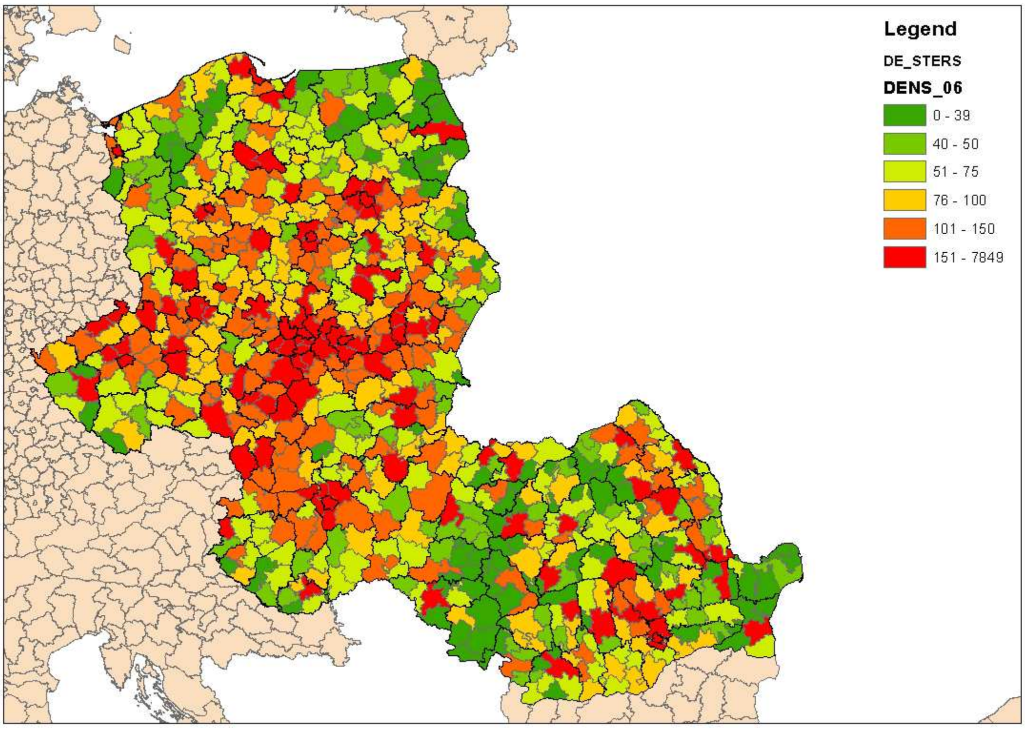


Figure 30 - Density of population expressed in an alternative geometry for countries in the Eastern Europe (data for 2006)

In order to conclude the presentation of the work done for this first objective, we recapitulate the major topics:

- elaboration of the regionalization method
- the model was applied to a set of countries and the results appears to be solid
- testing the mass-effect using the area autocorrelation indicator
- the selection of new aggregation centers is possible and this will provide a secondary version of the alternative geometry.

The geometry we used in this regionalization is based on the 2006 nomenclature of the LAU2. However, having the aggregation centers defined, we can easily switch between different versions (2001, 2008 etc.)

The second objective of this working period was to produce indicators at LAU2 scale, indicators derived by the relation between the VIGO (very important geographical objects) and the local elements of the administrative geometry. Defining this relation is very ambiguous, from a spatial point of view. In geography, the relation between the spatial units is generally understood as a form of spatial interaction and this interaction is described using gravity models. The output of these models treats the spatial and territorial organization as a sum of probabilities. Confronting these probabilities with empirical data (flows) helps researchers to calibrate the model's parameter. This approach is not possible to implement in the case of the VIGO because validation data is missing.

Our strategy was to simplify the sense of this relation, avoiding the problems generated by the spatial interaction models and their interpretation. One good example for this strategy is to describe the relation between the European airports and the LAU2. This relation can be described by various types of distance: euclidean, road distance (km) or time distance. Two different geometries were used in order to produce these results - the road network geometry and a layer with the location of the European airports (source: ETIS plus deliverables for 2010). The road network was weighted in order to approximate the time distance between the LAU2 centroids and the airports. Taking into account the road quality and a reasonable cruise speed for each edge, we can obtain a time distance matrix to exploit. All the centroids were located in the network and a time distance towards the closest three airports was calculated. The choice of three airports was suggested by the hypothesis of the interposed occasions acting as filters for the potential flows towards an air transport facility. The image illustrates the technical dimension of the method and the geometries involved in the analysis.

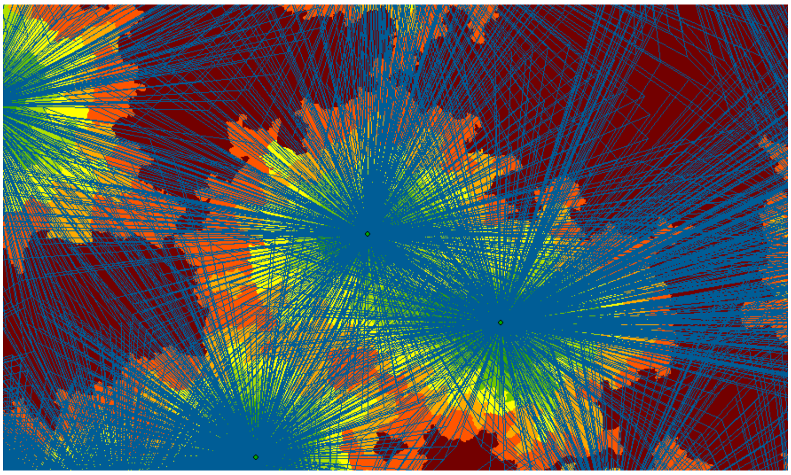


Figure 31 - Time distance towards the closest airport in the South of Poland (Katowice and Krakow) and links between the LAU2 centroids and the nearest airports.

Taking into account the large number of LAU2, the geometry was split in 5 subsets, introduced in the model and calculated. No administrative barriers were selected for the choice of the airports as facilities, having in mind the idea that the frontiers are highly permeable to the flows, in normal conditions. The results were remerged and the errors eliminated. Due to imperfections in the network dataset, results are not available for Greece and Bulgaria. The next step in our approach was to cumulate the population by distance to the closest airports and to map this indicator. The information was split in 5 equal classes (20%, 40%, 60 %, 80%, 100% of the cumulated ESPON area population). Despite the fact that a large number of LAU2 is placed in the last class, their share in the total population of the study is rather limited (only 20%). The extent of the red class should be interpreted in relation with the demographic concentrations, at ESPON space scale. This indicator could be considered as a tool of investigation aspects related to the territorial cohesion issues. One easy critic for this map (and the subsequent indicator) is the fact that airports and metropolitan agglomerations are spatially related. It is only partially true, for some metropolitan areas the distance to the airport is considerable. Moreover, some metropolitan areas collect more than one airport in their proximity, reshaping the space-time in their area of polarization.

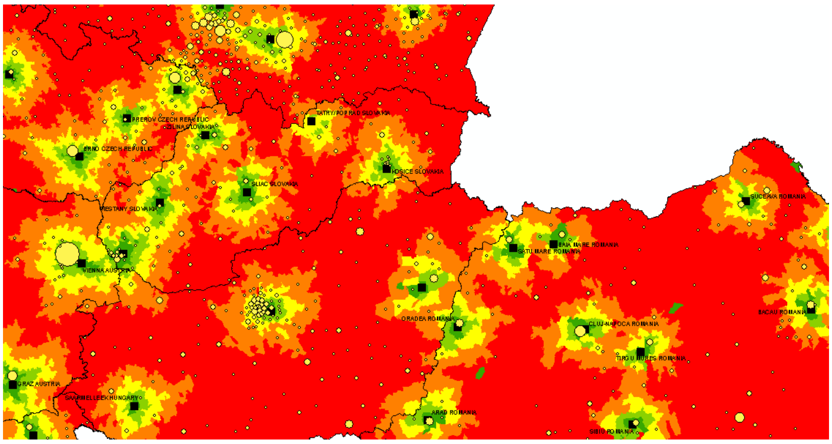


Figure 32 - Population cumulated by distance towards the closest airport (the red area cumulates only 20 % of the ESPON area population).

The proportional symbols on the map indicate the LAU2 with more than 10 000 inhabitants (2006). Their spatial distribution is interesting at national scale of analysis. In Hungary, some important elements of the urban system are in the red area - Miskolc, Szeged or Gyor. In Romania there are three cities with more than 200 000 inhabitants in the last class - Brasov, Galati and Braila. Situations like these could be found in other countries too and the interest of the map is to signal them. However, working with demographic data at local level is not an easy task. For instance, Portugal and UK present limited data at LAU2 level. The solution we adopted in order to calculate the cumulated population for all the LAU2 in the study area was an intersection between the LAU2 geometry of 2006 and two demographic grids, one provided by EEA (2001) and the other by GISCO (2006). This method allowed us to approximate the population at local level in areas where it was lacking.

The methodological approach of the relation between the VIGO and the LAU2 is interesting and provides scientific added value in the project. Seeing this relation as a cumulated sum (by distance) of the eventual users, for different territorial facilities, provides an exploration tool that was not available at this level before. It is certainly a perfectible indicator, the eventual errors being created by the quality of the networks, by the access to demographic data and by the definition of the VIGO. We resume the work done for this objective, reminding the key facts that describe this kind of indicator:

- calculation of time distances between LAU2 centroids (2006 nomenclature) and the nearest three airports. Another indicator can be derived from these values: average time to the nearest three airports, depicting areas where interposed occasions (airports) might compete. It also shows the role of the road network in the territory, at local level.
- calculation of cumulated population, using the distance as a filter. This indicator will eventually show how many persons live at a certain distance of one VIGO (airports).

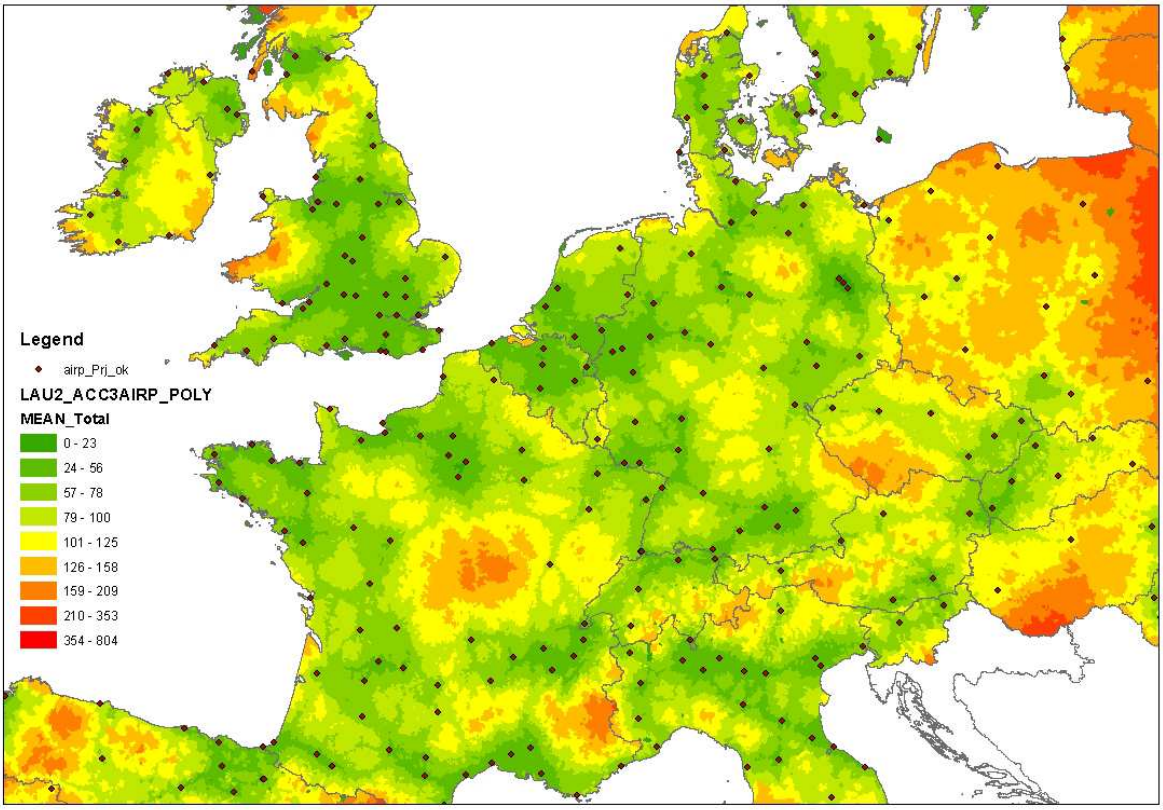


Figure 33 - Average time distance to the nearest three airports (LAU2 for 2006 nomenclature; road network for 2009)

Next steps

Having a tested methodological frame, the next steps in the workflow will be dedicated to three priorities:
1) elaborate two versions for the alternative geometry, in order to better compare the performance of this intermediate frame of spatial analysis. This process will also involve a geo-coding approach of the alternative objects.
2) creation of new indicators based on the relation between the VIGO and the LAU2 geometry. These indicators will take into account facilities such as major transportation nodes or protected natural areas of continental interest. We will integrate this environmental dimension because the VIGO label is not only related to the economic and social component of the European territory.
3) correction of datasets and geometries where needed and possible. Testing more the capacities of the network and LAU2 geometries provided some topological insufficiencies. It is not a matter of update, is a matter of intervention in specific areas. This aspect is needed in order to facilitate the work for the point 1 and 2.

2.5 European Neighbourhood (WP B5)

2.5.1 Regional neighbourhood

Following the division of tasks described in the Annex 1 of the First Interim Report (revised version of December 2012) written in order to fast the process of data collection and map creation, it has been decided together with the ESPON ITAN Project that:

- ESPON M4D takes in charge the definition of the nomenclatures and the creation of the Mapkit on Regional Neighbourhood. And this, for December 2012.
- ESPON ITAN takes in charge the collection of core data. M4D project gives support to ITAN for this task (choice of the indicators, metadata, quality checks etc.)

As a consequence, the ESPON M4D Project has created a first version of the Mapkit and a nomenclature for the territorial units located in the European Neighborhood (SNUTS) which must be used by the ITAN project for data gathering and for indicator building in December 2012.

For reminding, some main principles guide the building of the SNUTS nomenclature:

- Creating a nomenclature comparable to the NUTS nomenclature describing territorial division of the European Union. Consequently, the SNUTS (Similar to NUTS) nomenclature must follow the NUTS principles38.
- It is a multi-level and a hierarchical nomenclature: level 3 corresponds to the lower one, level 0 equals to the country level.
- The SNUTS nomenclature is based on the territorial divisions currently existing in neighbouring countries.

Three main sources were used for building the Mapkit and the nomenclature:

- The Global Administrative Area (GADM) database for the GIS layer. This resource provides seamless geometries for all the countries of the World at different administrative levels. However, some of the shapefiles available in the GADM are outdated, due to territorial reforms not taken into account.
- National Country Profiles (most of the time downloaded from National Statistical Institutes websites) in order to identify the levels 0, 1 & 2 of the SNUTS nomenclature, and, if possible, the level 3.
- External sources (Wikipedia, City Population, GeoHive…) when incoherencies where found between the GADM shapefiles and the National Country Profiles.

[38] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/principles_characteristics

At the end, the M4D Project has delivered to the ITAN Coordination team the following files:

- GADM geometries with SNUTS codes, with precise boundaries (adapted for GIS calculations).
- M4D geometries with SNUTS codes, with generalized boundaries (adapted for mapping purposes).
- A hierarchical nomenclature, with associated names and SNUTS codes.

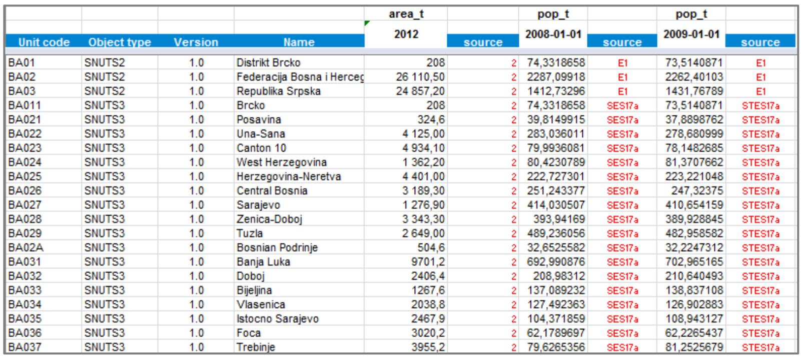
2 Time-series population 1979-2010 datasets: 1 coming directly from National Statistical Institutes (no estimations), 1 complete dataset (estimated missing values and adjusted to the United Nations data).

However, engineers who are not specialists of all the countries of the ITAN Area have created this Mapkit and this nomenclature. Consequently a validation process was necessary before gathering data for the European Neighbourhood. This has been made possible thanks to the Nomenclature Update Report, which is a template aiming to propose a procedure for referring possible mistakes existing in M4D files. The M4D Project has proposed a two steps approach for checking the regional neighbourhood nomenclature.

Step 1 – Depth analysis of the M4D Documents

For a given expertise area, the ITAN expert has to analyse the M4D nomenclature (figure 34) and have a look to the GADM geometries (figure 35), which were used to generate the simplified geometries available in the ITAN Mapkit.

On the basis of his/her knowledge and the official definition of the territorial division in his/her expertise area, the expert will generate the Nomenclature Update Report (cf step 2). On the basis of this report, the ITAN and the M4D lead partners will modify the GADM and ITAN geometries or attributes.



				area_t		pop_t		pop_t	
Unit code	Object type	Version	Name	2012	source	2008-01-01	source	2009-01-01	source
BA01	SNUTS2	1.0	Distrikt Brcko	208	2	74,3318658	E1	73,5140871	E1
BA02	SNUTS2	1.0	Federacija Bosna i Herceg	26 110,50	2	2287,09918	E1	2262,40103	E1
BA03	SNUTS2	1.0	Republika Srpska	24 857,20	2	1412,73296	E1	1431,76789	E1
BA011	SNUTS3	1.0	Brcko	208	2	74,3318658	SES17a	73,5140871	STES17a
BA021	SNUTS3	1.0	Posavina	324,6	2	39,8149915	SES17a	37,8898762	STES17a
BA022	SNUTS3	1.0	Una-Sana	4 125,00	2	283,036011	SES17a	278,680999	STES17a
BA023	SNUTS3	1.0	Canton 10	4 934,10	2	79,9936081	SES17a	78,1482685	STES17a
BA024	SNUTS3	1.0	West Herzegovina	1 362,20	2	80,4230789	SES17a	81,3707662	STES17a
BA025	SNUTS3	1.0	Herzegovina-Neretva	4 401,00	2	222,727301	SES17a	223,221048	STES17a
BA026	SNUTS3	1.0	Central Bosnia	3 189,30	2	251,243377	SES17a	247,32375	STES17a
BA027	SNUTS3	1.0	Sarajevo	1 276,90	2	414,030507	SES17a	410,654159	STES17a
BA028	SNUTS3	1.0	Zenica-Doboj	3 343,30	2	393,94169	SES17a	389,928845	STES17a
BA029	SNUTS3	1.0	Tuzla	2 649,00	2	489,236056	SES17a	482,958582	STES17a
BA02A	SNUTS3	1.0	Bosnian Podrinje	504,6	2	32,6525582	SES17a	32,2247312	STES17a
BA031	SNUTS3	1.0	Banja Luka	9701,2	2	692,990876	SES17a	702,965165	STES17a
BA032	SNUTS3	1.0	Doboj	2406,4	2	208,98312	SES17a	210,640493	STES17a
BA033	SNUTS3	1.0	Bijeljina	1267,6	2	137,089232	SES17a	138,837108	STES17a
BA034	SNUTS3	1.0	Vlasenica	2038,8	2	127,492363	SES17a	126,902883	STES17a
BA035	SNUTS3	1.0	Istocno Sarajevo	2467,9	2	104,371859	SES17a	108,943127	STES17a
BA036	SNUTS3	1.0	Foca	3020,2	2	62,1789697	SES17a	62,2265437	STES17a
BA037	SNUTS3	1.0	Trebinje	3955,2	2	79,6265356	SES17a	81,2525679	STES17a

Figure 34 – The M4D nomenclature (folder basic data, file M4D_basicsneighb_UN_20121002.xls)

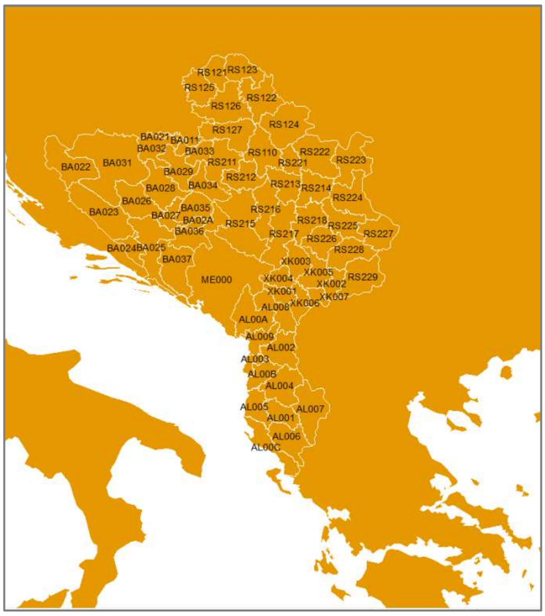


Figure 35 – GADM geometries including ESPON Codes, available at the lower SNUTS level (folder GADM final)

Step 2 – Generation of the nomenclature update report

The report must follow a precise generic procedure:

1. **A general comment** describing what kind of incoherencies were detected in the M4D files.
2. **Geometries part**: If SNUTS codes and/or geometries must be modified, a screenshot showing the reference map to be considered for updating the ESPON geometries. This map must be of high quality and must be sent to the ITAN and M4D managers in a .png format also. On top of that, it is necessary to write on the map the SNUTS codes you want to affect to each territory of the expertise area.
3. **Exhaustive listing of the modifications:** Based on the outputs of the technical report on time-series produced within the ESPON Database 1 project (figure 36), the M4D project has proposed a synthetic table organized in 7 columns to track the territorial changes. The column 1 and 2 identify the code and the name of the M4D nomenclature. The column 3 describes the nature of the change. The ITAN expert has to choose among the possibilities described in the figure 36 to describe the nature of the change. The column 4 describes what is the status of the territorial unit. Choose among these possibilities: no pb, creation, deletion or modification. The columns 5 and 6 aim to provide the new codes and names of the territorial units. Finally, you can optionally add a comment on the column 7.

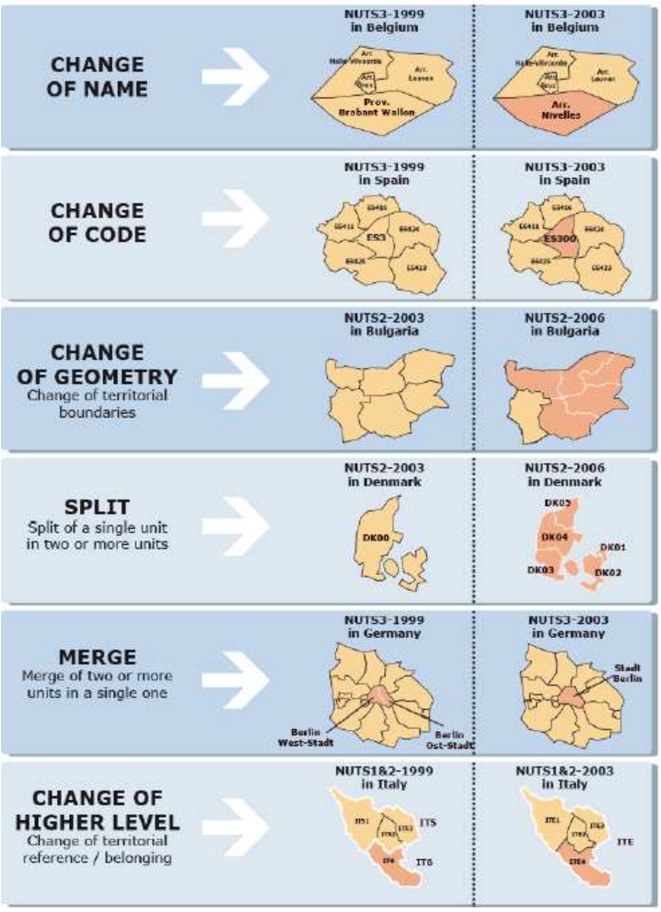
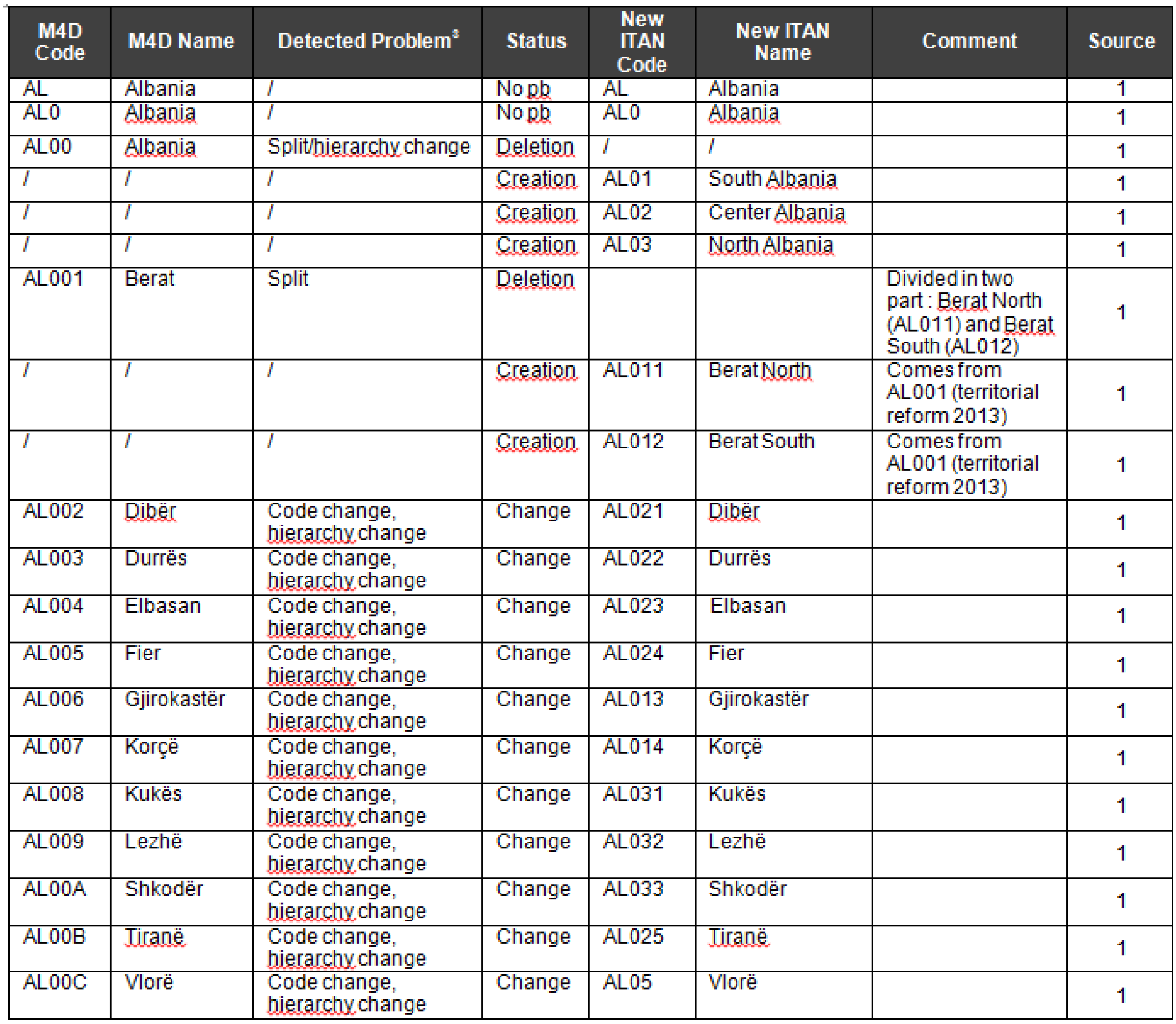


Figure 36 – Possibilities existing for the "detected problem" description

source: Ben Rebah et al., 2011, Empirical approach and applications for modeling NUTS changes and managing time series data, ESPON 2013 Database Project

In June 2013, the ITAN Project has provided to the M4D Project 4 Nomenclature Update Report for Morocco, Algeria, Tunisia and Russia[39]. The Mapkit and the nomenclature on regional neighbourhood have been updated consequently and sent to the ITAN Project. We are still waiting for the remaining reports. When it will be received, the mapkit and the nomenclature will be updated once again.

[39] The nomenclature update report on Morocco (filed by ITAN expert) is available on Annex 3



M4D Code	M4D Name	Detected Problem[3]	Status	New ITAN Code	New ITAN Name	Comment	Source
AL	Albania	/	No pb	AL	Albania		1
AL0	Albania	/	No pb	AL0	Albania		1
AL00	Albania	Split/hierarchy change	Deletion	/	/		1
/	/	/	Creation	AL01	South Albania		1
/	/	/	Creation	AL02	Center Albania		1
/	/	/	Creation	AL03	North Albania		1
AL001	Berat	Split	Deletion			Divided in two part : Berat North (AL011) and Berat South (AL012)	1
/	/	/	Creation	AL011	Berat North	Comes from AL001 (territorial reform 2013)	1
/	/	/	Creation	AL012	Berat South	Comes from AL001 (territorial reform 2013)	1
AL002	Dibër	Code change, hierarchy change	Change	AL021	Dibër		1
AL003	Durrës	Code change, hierarchy change	Change	AL022	Durrës		1
AL004	Elbasan	Code change, hierarchy change	Change	AL023	Elbasan		1
AL005	Fier	Code change, hierarchy change	Change	AL024	Fier		1
AL006	Gjirokastër	Code change, hierarchy change	Change	AL013	Gjirokastër		1
AL007	Korçë	Code change, hierarchy change	Change	AL014	Korçë		1
AL008	Kukës	Code change, hierarchy change	Change	AL031	Kukës		1
AL009	Lezhë	Code change, hierarchy change	Change	AL032	Lezhë		1
AL00A	Shkodër	Code change, hierarchy change	Change	AL033	Shkodër		1
AL00B	Tiranë	Code change, hierarchy change	Change	AL025	Tiranë		1
AL00C	Vlorë	Code change, hierarchy change	Change	AL05	Vlorë		1

Figure 37 – Theoretical example of a listing of modifications in Albania

2.5.2 Neighbouring cities

The Technical Report European Neighbouring Cities aims at collecting and expertizing the national or international urban definitions and indicators in 17 countries, in Maghreb40, Middle East41 and Eastern countries42. We have to note that there is no harmonized definition of cities in the European Neighbouring countries.
We have led a systematic collect of the different types of definitions, documentation, and urban indicators officially available in the neighbourhood countries (website of the Census Boards, official documents on Internet, questions sent to the Census Board contact points). For 12 countries amongst 17 we succeeded in finding a national official definition of urban objects (Figure 38).

[40] Algeria, Libya, Morocco, Tunisia.
[41] Armenia, Azerbaijan, Egypt, Georgia, Iraq, Israel, Jordan, Lebanon, Palestine, Syria.
[42] Belarus, Russia, Ukraine.

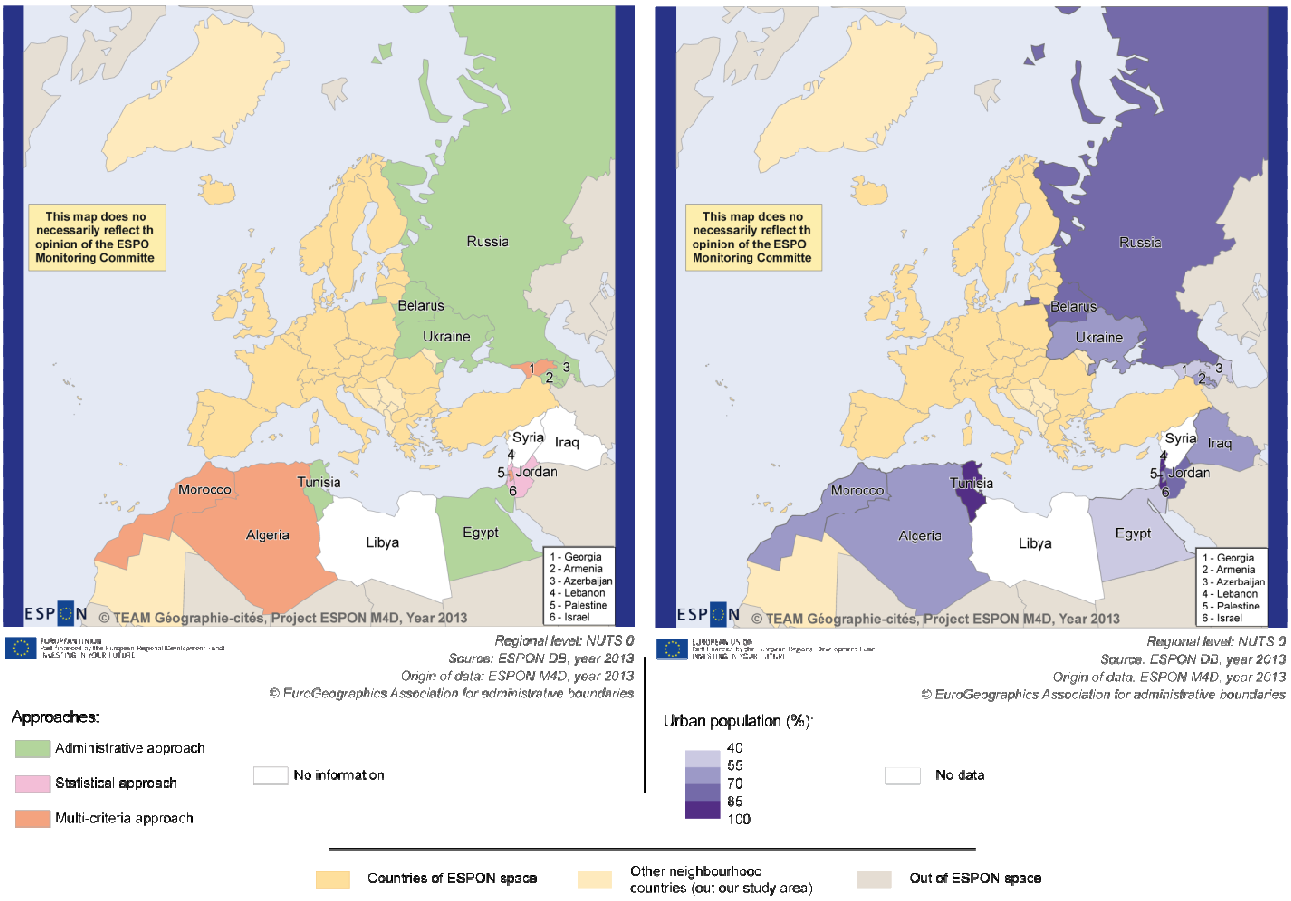


Figure 38 – National city definitions in the European Neighbourhood (on the left) and urban population levels in European Neigbourhood (on the right, based on national city definitions)

This documentation has been expertized and presented, country-by-country, in the Annex of the Technical Report, using a common "syntax" for categorizing the available information (administrative divisions, urban definition, data availability, references).

We also compared the results of this bottom-up approach to the ones obtained in Menapolis database for 5 of the 17 countries, using a top-down approach[43]. The Menapolis database is not freely available on the website, but some tables showing main indicators (total number of cities, level of urbanization, population of largest cities) may be downloaded.

A second delivery on European Neighbouring Cities, including the geometries of urban delineations (shape files) and the urban indicators that were collected (excel files with ID links with shape files) should be available in December 2013. For some countries we received an Excel list of cities without georeference codes and we have to fulfil the centroids in order to join the information with the shape files or to map the indicators.

[43] François Moriconi-Ebrard and its collaborators, http://e-geopolis.eu/menapolis.

Working Package C-D – Networking activities & support to the ESPON Coordination Unit

3.1 ESPON Database Portal and follow-up of ESPON Projects (WP C1-2-3)

For the First Interim Report, the ESPON M4D Project has provided some tools/documentation in order to help ESPON Projects in their data collection/creation/delivery within the ESPON Program:

- The tracking tool, which allows following the data integration process and the result of the syntactic, the semantic and the outlier checks.
- How to deliver my data documentation, available in the Upload section of the ESPON Database Portal.
- Delivery of the first M4D Newsletter.

Since the First Interim Report, continuous exchanges with ESPON TPGs has occurred regarding the ESPON Data and Metadata template, the way to deliver data and the ESPON Mapkits. In particular it concerned all the ESPON TPGs which have already delivered their data under the tracking tool (ESAtDOR, SeGi, Map update on natural hazards, SIESTA, SEMIGRA). Out of these ESPON Projects, it concerned also the follow-up of the SGPTD project (definition of the nomenclature, metadata creation), the City Benchmark project (overview of existing urban databases) and the ITAN Project (cf part on the European Neighbourhood).

Otherwise, the M4D LIG partner has been in particular in charge of the Priority 3 projects type. Under this action, since June 2012, LIG has mainly proposed its technical assistance to the ESPON Online Mapping Tool RIMAP project (AIDICO team conducted by Sergio Munoz). As agreed with Sergio Munoz during the ESPON Seminar in Aalborg (June 2012), LIG has regularly delivered new up-to-date versions of the ESPON Database to the AIDICO team. These deliveries (July 2012, March and May 2013) include all the materials to create the database (SQL scripts, software to integrate the ESPON TPGs Key Indicators datasets), and a relevant documentation proposing the description of all the tables and fields that compose the ESPON Database. Besides these deliveries, LIG has supported the RIMAP project by numerous technical emails exchanges. Though time-consuming, this support and feedbacks has the benefit, on the one hand, to encourage sharing functionalities and modules, on the other hand, to help at improving the design of the database for its better access and re-use, not only by the mean of the ESPON Database Portal Web Application, but also by the tools that are and will be developed by ESPON Projects (in particular, the Web Services for an access by third-party applications). This issue has also been clearly highlighted during the ESPON Tools Technical meeting on the 16th and 17th of May 2013 in Paris.

Thereby, all the partners of the M4D project have followed the ESPON Tools meeting. Exploiting cooperation synergy and linkages among ESPON tools was the topic of the meeting. As pointed out during these two days, the database is the "heart" of the ESPON toolbox. The importance of the LIG partner expertise and support regarding the ESPON Database access rose up and also the linkages between the Database and the others tools. Since this meeting, M4D has now shared the deliveries of the ESPON Database with the European Territorial Monitoring System (ETMS) project, conducted by Erik Gloersen. Some support by email has also been done towards ETMS since this delivery in May 2013. As a conclusion, the main beneficial effects of the networking activities for the June 2012-June 2013 period are the linkages between the ESPON tools around the database, and, on LIG side, the improvements of the technical documentations and delivered software for the needs of the current and next ESPON tools projects.

Out of this information flow, some functionalities/documentation has been added or updated in the ESPON Database Portal for making easier the understanding of the data process within the ESPON Program.

First of all, new functionalities developed within the WP A (cf above for more details) helps the access to the information related to the ESPON Database and namely the overview (cf part 1.1.2.2), the news management (cf part 1.1.2.4), the user's registration page (cf part 1.1.2.5), which ease the access to the ESPON Database Portal and its content.

Secondly, new documentation have been uploaded/updated in the ESPON Database Portal, and namely:

- **The Frequent Asked Section**, available under the Help section of the ESPON Database Portal was edited during the ESPON Database 1 project (2008-2011). Consequently, a significant number of questions/answers was outdated. The FAQ section has been totally restructured and is now available for external users and is divided in 8 sections: what is ESPON M4D? The ESPON Database Portal, Restricted part of the ESPON Database Portal, Data delivery, Metadata processing, Support to data creation and Local/urban data.
- **The Dictionary of Spatial Units documentation,** available under the help page of the ESPON Database Portal. It describes the choices made for integrating these nomenclatures within the Search Interface of the ESPON Database Portal. Among others, it describes the reference documents used for combining "official NUTS nomenclatures", for EU Members and non-official nomenclatures for EFTA and Candidate Countries in a single nomenclature in the Web Interface (one of the specificities of the ESPON Program). It makes also available all the units codes/names/hierarchies etc. for all the nomenclatures contained into the ESPON Database. This documentation may be especially useful for two kind of purpose: (i) Be transparent regarding the territorial units contained in the ESPON Database and the parameters used to estimate the completeness of the indicator. (ii) For ESPON TPG, give the inputs for delivering datasets covering all the ESPON Area.
- **How to deliver my data presentation,** made by the LIG and UMS RIATE for the ESPON Seminar in Paphos (December 2012). This presentation summarizes in a comprehensive way the How to deliver my data documentation. It is available in the tracking tool section of the ESPON Database Portal.

Last but not least, the M4D Project has proposed to the ESPON Community two **newsletters** (ESPON seminars in Paphos and Dublin), figure 39:

The **newsletter n°2** explained the structure of the ESPON Database Portal, the tracking tool and the aim of the quality check, described the data integrated in the last version of the ESPON OLAP Cube, synthetized the work done regarding the data collection at regional level for the European Neighbourhood, promoted the GEOSPECS database (new LAU2 figures for the ESPON Area), described the state of advancement of data integration and provided a methodological note on how to access to local data within the ESPON Program (SIRE database, Eurogeographics layers).

The **newsletter n°3** was more focused on the Search Interface and urban data. First of all, it described the way to retrieve data in the Search Interface of the ESPON Database Portal. Then, it described urban nomenclatures integrated in the Search Interface of the ESPON Database Portal and it has provided the latest information the M4D project had related to the definition of the new Large Urban Zones. Finally, it provided some milestones regarding the last developments of the core database strategy (functional indicators) and the state of advancement of the ESPON Data integration process.

ESPON

M4D NEWSLETTER

FOR ESPON PROJECTS

MULTI-DIMENSIONAL DATA DESIGN AND DEVELOPMENT

N°2 JUNE-DECEMBER 2012

EDITO ...

This second issue of the M4D newsletter (Multi Dimensional Database Design and Development) is the occasion to present to the ESPON Community the new version of the ESPON Database Web Portal, which is now on line. Since the M4D Project pays a great attention to data quality, we take you to the back office of the ESPON Database Web Portal and put the focus on the online tracking tool that monitors each step of the Data delivery process. Then, we explain what is behind the Data Quality Check, one of the milestones of this process, whose goal is to reveal unusual values. A new population time series 1990-2011 has been loaded in the ESPON OLAP Cube. This is another opportunity to navigate smoothly across multidimensional data. We believe that on going and next studies on the European Neighbourhood will probably take advantage of some new material (nomenclature, core indicators, and map kit) that the M4D Project has just released. Because keeping you informed about the last datasets integrated in the ESPON Database Portal is the other raison d'etre of this newsletter, this issue is dedicated to the GEOSPECS Project, which provides innovative indicators at local scale (LAU2). Lastly, we propose you a summary of very fruitful exchanges we had with EUROSTAT concerning the Sire Database. It might be useful for ESPON Projects that carry on or just start their activities around local data.
Enjoy your reading!

Jérôme Gensel, professor in Computer Science (LIG) ● ● ●

CONTENT

- ESPON M4D NEWS ..1
  Latest developments for the community
- ESPON DATA ..4
  The database and ESPON Projects data
- EXTERNAL DATA ..6
  Data from other European /international institutions

ESPON M4D NEWS - The ESPON Database Portal online!

The ESPON M4D project is very pleased to share the new version of the ESPON Database Portal with the ESPON Community. This portal - available at the following URL: http://db2.espon.eu - is composed with different key elements:

- Welcome page with a news section to inform the community about the last data integrated, last developments, etc.
- Query interface, the principle access to retrieve data produced by ESPON Projects (search Menu).
- Resources page, fed with technical reports, tools, data, mapkits, documentation.
- Upload page, to automatically upload data by ESPON Projects
- Help page, from where it is possible to retrieve all information concerning the ESPON Database Portal, ESPON Data and metadata etc.

The ESPON Database Portal homepage

Please note that this Portal is still under construction. New functionalities and indicators will be continuously integrated. You will be informed through the news section of the ESPON Database Portal and in the ESPON M4D Newsletter.

A user manual (http://db2.espon.eu/db2/usermanual) describes in details the content of each section of the ESPON Database Portal. We hope that this new tool will be useful for you!

Authors
Claude Grasland (UMS RIATE)
Ronan Ysebaert (UMS RIATE)
Isabelle Salmon (UMS RIATE)
Nicolas Lambert (UMS RIATE)
Timothée Giraud (UMS RIATE)
Benoît Le Rubrus (LIG)
Roger Milego (UAB)
Martin Charlton (NCG)
Alberto Caimo (NCG)
Jacques Michelet (Univ. Genève)

Want to share any information related to data for the ESPON Community? Please contact the ESPON M4D Manager: manager@espondb.eu

The electronic version of the M4D Newsletter is available on the ESPON Database website: http://db2.espon.eu

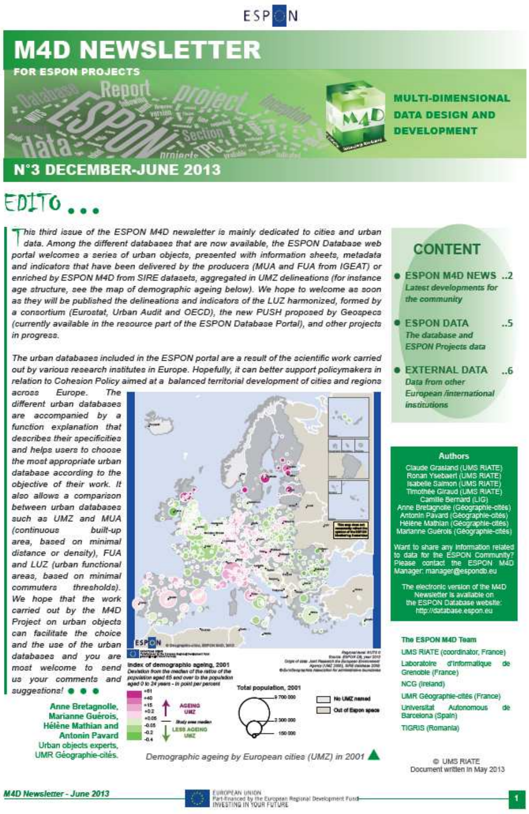
The ESPON M4D Team
UMS RIATE (coordinator, France)
Laboratoire d'Informatique de Grenoble (France)
NCG (Ireland)
UMR Géographie-cités (France)
Universitat Autonomous de Barcelona (Spain)
TIGRIS (Romania)

© UMS RIATE
Document written in November 2012

M4D Newsletter - December 2012

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

1



ESPON

M4D NEWSLETTER

FOR ESPON PROJECTS

MULTI-DIMENSIONAL DATA DESIGN AND DEVELOPMENT

N°3 DECEMBER-JUNE 2013

EDITO ...

This third issue of the ESPON M4D newsletter is mainly dedicated to cities and urban data. Among the different databases that are now available, the ESPON Database web portal welcomes a series of urban objects, presented with information sheets, metadata and indicators that have been delivered by the producers (MUA and FUA from IGEAT) or enriched by ESPON M4D from SIRE datasets, aggregated in UMZ delineations (for instance age structure, see the map of demographic ageing below). We hope to welcome as soon as they will be published the delineations and indicators of the LUZ harmonized, formed by a consortium (Eurostat, Urban Audit and OECD), the new PUSH proposed by Geospecs (currently available in the resource part of the ESPON Database Portal), and other projects in progress.

The urban databases included in the ESPON portal are a result of the scientific work carried out by various research institutes in Europe. Hopefully, it can better support policymakers in relation to Cohesion Policy aimed at a balanced territorial development of cities and regions across Europe. The different urban databases are accompanied by a function explanation that describes their specificities and helps users to choose the most appropriate urban database according to the objective of their work. It also allows a comparison between urban databases such as UMZ and MUA (continuous built-up area, based on minimal distance or density), FUA and LUZ (urban functional areas, based on minimal commuters thresholds). We hope that the work carried out by the M4D Project on urban objects can facilitate the choice and the use of the urban databases and you are most welcome to send us your comments and suggestions! ● ● ●

Anne Bretagnolle, Marianne Guérois, Hélène Mathian and Antonin Pavard
Urban objects experts, UMR Géographie-cités.

ESPON
Index of demographic ageing, 2001
Deviation from the median of the ratios of the population aged 65 and over to the population aged 0 to 24 years - in point per percent
+61
+40
+15
+0.2
+0.05
-0.05
-0.2
-0.4
AGEING UMZ
Study area median
LESS AGEING UMZ
Total population, 2001
9 700 000
2 300 000
150 000
No UMZ named
Out of Espon space

Demographic ageing by European cities (UMZ) in 2001

CONTENT

- ESPON M4D NEWS ..2
  Latest developments for the community
- ESPON DATA ..5
  The database and ESPON Projects data
- EXTERNAL DATA ..6
  Data from other European /international institutions

Authors
Claude Grasland (UMS RIATE)
Ronan Ysebaert (UMS RIATE)
Isabelle Salmon (UMS RIATE)
Timothée Giraud (UMS RIATE)
Camille Bernard (LIG)
Anne Bretagnolle (Géographie-cités)
Antonin Pavard (Géographie-cités)
Hélène Mathian (Géographie-cités)
Marianne Guérois (Géographie-cités)

Want to share any information related to data for the ESPON Community? Please contact the ESPON M4D Manager: manager@espondb.eu

The electronic version of the M4D Newsletter is available on the ESPON Database website: http://database.espon.eu

The ESPON M4D Team
UMS RIATE (coordinator, France)
Laboratoire d'Informatique de Grenoble (France)
NCG (Ireland)
UMR Géographie-cités (France)
Universitat Autonomous de Barcelona (Spain)
TIGRIS (Romania)

© UMS RIATE
Document written in May 2013

M4D Newsletter - June 2013

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

1

Figure 39 – First page of the M4D Newsletter n°2 and n°3

Next steps…
One of the main output of the ESPON Seminar held in Dublin (June 2013) regarding ESPON tools was the interest of external users to videos explaining how to use the tools produced within the ESPON Program. Taking into account that the user manual must be updated in the next steps of the project, we aim at providing a "user-friendly" manual, mainly organised around videos.

3.2 Networking with other institutions (WP C4)

Regarding the external networking activities, the M4D LIG partner has contracted a technical expertise with the **UNEP/Grid-Geneva team** (which has developed the Geodata Portal[44]). The objective of the expertise is three-fold: - a study of the access to the ESPON Database by third-party applications by the mean of OGC web services; - a conformity assessment of the ESPON Database with OGC standards and the INSPIRE directive; - recommendations for designing a solution based on OGC Web Services, allowing the exploitation of the ESPON metadata and data by third party applications, including geospatial data portals. This expertise has been carried out since a first meeting in September 2012 (in Grenoble), then in February 2013 (in Geneva), then via several Internet on-line meetings. Beyond this technical cooperation between LIG and UNEP/GRID, this external networking activity may help in a way at enlarging the accessibility to the ESPON Database, and in the meantime at promoting the visibility of the ESPON Program out of the "ESPON world".

[44] http://geodata.grid.unep.ch

In the working package related to urban objects, the M4D Géographie-cités partner has established strong contacts with the **OECD and Eurostat** regarding the definition of new Large Urban Zones (files, metadata, indicators). Thanks to these fruitful exchanges, it has been possible to propose a section on Large Urban Zones in the ESPON M4D Newsletter N°3.

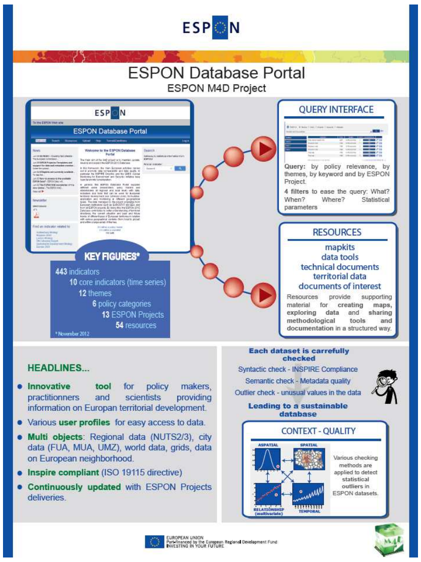
For populating urban databases through European reference grids, contacts have been established with the **Joint Research Center**, in particular to understand the main changes (methological framework etc.) between the population grid 2001 and the population grid 2006.

Finally, exchanges with the **COGIT** have been organized in order to compare the road network provided by Eurogeographics and the network provided by the French IGN.

3.3 Support to the ESPON Coordination Unit (WP D4)

The M4D Project provides continuous support to the ESPON Coordination Unit related to data collection and harmonisation, data expertise, mapping and publications/press releases. Out of continuous exchanges with the ESPON Coordination Unit on these topics, some events/reports have particularly mobilised the M4D Project. Since June 2012, it corresponds to the following elements:

Support for the organisation of the ESPON Seminar in Cyprus



ESPON

ESPON Database Portal
ESPON M4D Project

QUERY INTERFACE

Query: by policy relevance, by themes, by keyword and by ESPON Project.

4 filters to ease the query: What? When? Where? Statistical parameters

RESOURCES

mapkits
data tools
technical documents
territorial data
documents of interest

Resources provide supporting material for **creating** maps, **exploring** data and **sharing** **methodological** tools and **documentation** in a structured way.

KEY FIGURES*

443 indicators
10 core indicators (time series)
12 themes
6 policy categories
13 ESPON Projects
54 resources
* November 2012

HEADLINES...

- **Innovative tool** for policy makers, practitionners and scientists providing information on Europan territorial development.
- Various **user profiles** for easy access to data.
- **Multi objects**: Regional data (NUTS2/3), city data (FUA, MUA, UMZ), world data, grids, data on European neighborhood.
- **Inspire compliant** (ISO 19115 directive)
- **Continuously updated** with ESPON Projects deliveries.

Each dataset is carrefully checked

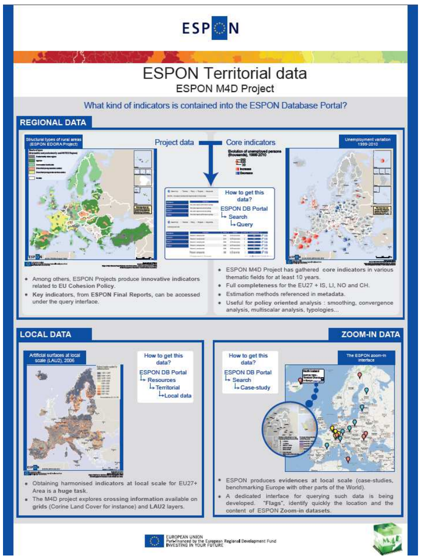
Syntactic check - INSPIRE Compliance
Semantic check - Metadata quality
Outlier check - unusual values in the data

Leading to a sustainable database

CONTEXT - QUALITY

Various checking methods are applied to detect statistical outliers in ESPON datasets.

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



ESPON

ESPON Territorial data
ESPON M4D Project

What kind of indicators is contained into the ESPON Database Portal?

REGIONAL DATA

Structural types of rural areas (ESPON EDORA Project)
Project data
Core indicators
Unemployment variation 1999-2010
Evolution of unemployed persons (thousands), 1999-2010
Increase
Decrease

How to get this data?
ESPON DB Portal
↳ Search
↳ Query

- Among others, ESPON Projects produce **innovative indicators** related to **EU Cohesion Policy**.
- **Key indicators**, from **ESPON Final Reports**, can be accessed under the query interface.

- ESPON M4D Project has gathered **core indicators** in various thematic fields for at least 10 years.
- **Full completeness** for the EU27 + IS, LI, NO and CH.
- Estimation methods referenced in **metadata**.
- Useful for **policy oriented analysis** : smoothing, convergence analysis, multiscalar analysis, typologies...

LOCAL DATA

How to get this data?
ESPON DB Portal
↳ Resources
↳ Territorial
↳ Local data

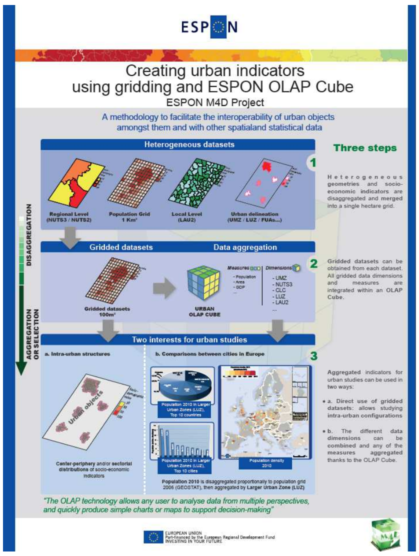
- Obtaining **harmonised indicators at local scale** for EU27+ Area is a **huge task**.
- The M4D project explores **crossing information** available on **grids** (Corine Land Cover for instance) and **LAU2** layers.

ZOOM-IN DATA

How to get this data?
ESPON DB Portal
↳ Search
↳ Case-study

- ESPON produces **evidences at local scale** (case-studies, benchmarking Europe with other parts of the World).
- A dedicated **interface** for querying such data is being developed. "**Flags**", identify quickly the location and the content of **ESPON Zoom-in datasets**.

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



ESPON

Creating urban indicators
using gridding and ESPON OLAP Cube
ESPON M4D Project

A methodology to facilitate the interoperability of urban objects amongst them and with other spatialand statistical data

Three steps

1 Heterogeneous geometries and socio-economic indicators are disaggregated and merged into a single hectare grid.

2 Gridded datasets can be obtained from each dataset. All gridded data dimensions and measures are integrated within an OLAP Cube.

3 Aggregated indicators for urban studies can be used in two ways:

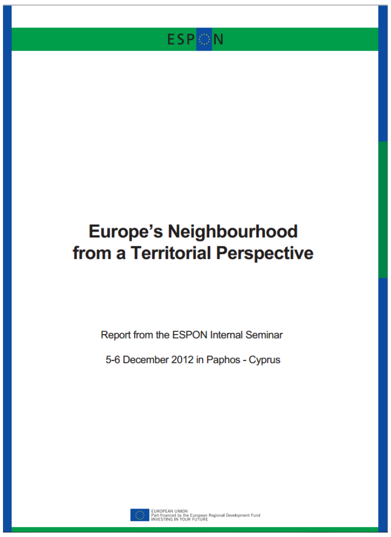
- a. Direct use of gridded datasets: allows studying intra-urban configurations
- b. The different data dimensions can be combined and any of the measures aggregated thanks to the OLAP Cube.

Heterogeneous datasets
Regional Level (NUTS3 / NUTS2)
Population Grid 1 Km²
Local Level (LAU2)
Urban delineation (UMZ / LUZ / FUAs...)
DISAGGREGATION
Gridded datasets
Data aggregation
Gridded datasets 100m²
URBAN OLAP CUBE
Measures
- Population
- Area
- GDP
Dimensions
- UMZ
- NUTS3
- CLC
- LUZ
- LAU2
AGGREGATION OR SELECTION
Two interests for urban studies
a. Intra-urban structures
b. Comparisons between cities in Europe
Urban objects
Center-periphery and/or sectorial distributions of socio-economic indicators
Population 2010 in Larger Urban Zones (LUZ), Top 10 countries
Population 2010 in Larger Urban Zones (LUZ), Top 10 cities
Population density 2010

Population 2010 is disaggregated proportionally to population grid 2006 (GEOSTAT), then aggregated by Larger Urban Zone (LUZ)

"The OLAP technology allows any user to analyse data from multiple perspectives, and quickly produce simple charts or maps to support decision-making"

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



ESPON

Europe's Neighbourhood
from a Territorial Perspective

Report from the ESPON Internal Seminar

5-6 December 2012 in Paphos - Cyprus

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

Figure 40 - Three posters related to the ESPON M4D Database activities and ESPON Seminar Report

Three posters have been realised in order to promote the ESPON Database Portal organisation, the data included in the portal (ESPON Territorial data) and the ESPON OLAP Cube (creating urban indicators using gridding and ESPON OLAP Cube).

On top of that, the M4D Project has participated to the ESPON Seminar report - *Europe Neighbourhood from a Territorial Perspective* – by delivering 8 maps especially for the report and harmonising all the rest of the maps of the document.

Support for the second ESPON Synthesis Report.

The ESPON M4D Project has taken in charge the map harmonisation of the 24 maps included in the report. Among others, the work has consisted by harmonising all the map templates and making easier the reading of some complex maps.

Support for the SIESTA Atlas.

The M4D Project has updated some indicators and estimating some missing values of 12 maps included in the SIESTA Atlas. The Regional EU2020S aggregate index has been also completed with missing values (EFTA) and the data/metadata have been corrected consequently. Finally, maps of the report have been harmonised.

Conclusion – From the Second Interim Report to the Final Report

Considering, on the one hand, the embedded functionalities in the June 2013 delivery for the Web Application and the state of advancement of the rest of the Working Packages (thematic and networking); and on the other hand, the terms of the contract and expected deliveries mentioned in the First Interim Report (June 2012), please find below an indicative agenda regarding the M4D deliveries until the Final Report (December 2014).

In between June 2013 and December 2013

- [WPA1] Regular updates of the Web Application with new-implemented features.
- [WPA3] Update of ESPON Mapkit on the European Neighbourhood based on ITAN feedbacks.
- [WPB5] Update of the nomenclature and basic indicators on European Neighbourhood based on ITAN feedbacks.
- [WPC] Continuous integration of ESPON data into the Web Application
- [WPC] Continuous updates of the News part of the ESPON Data Portal
- [WPC] Continuous support to ESPON Projects and ESPON CU
- [WPC] Integration of policy themes, keywords and thematic themes in close cooperation with the ESPON Coordination Unit
- [TRANSVERSAL] Continuous integration of new datasets via the new integration tools

June 2014

- [WPB5]: Final technical report on the ESPON Neighborhood database
- [WPB5]: Final version of the GlobCover 2009 by SNUTS units and by grid
- [WPB6]: Technical report on the evolutions of attractiveness between 2009-2012 and a sample of students located inside and outside EU
- [WPC3]: Implementation of the web-interface dedicated to ESPON Priority 2 projects
- [TRANSVERSAL] Draft Final Report
- [TRANSVERSAL] Continuous integration of new datasets via the new integration tools
- [TRANSVERSAL] Support to ESPON Coordination unit

December 2014

- [TRANSVERSAL] Final Report
- [TRANSVERSAL] Continuous integration of new datasets via the new integration tools
- [TRANSVERSAL] Support to ESPON Coordination unit

Annex 1 – Quality check of the ESatDOR indicators

Detail

Dataset: 2013-01-16-15-22-57-EsaTDOR_ESATDOR_Economic_Use_Composite v2_syntaxChecked.xls
Check date: 2013/03/20
Checked by: AC/MC
NUTS level 2:
NUTS date 2006:

Missing values

Values are coded as not present for some areas. The regions selected for analysis are those that have one or more borders which are coastal. Those NUTS 2006 regions, which are omitted from the analysis, have **n/r** as a missing value code for all data items. There are 317 rows in the dataset; 125 of these have valid data and 192 have missing data for all variables.

Non-Integer Count Data

The variables whose names start with **E09** are counts of employees. There are instances of non-integer values among the values for these variables. There are too many to list here but we can supply a list if necessary. For the variable **E09TOT** those regions with non-integer counts are:

```
[1] BE25 BG33 BG34 EE00 ES11 ES12 ES13 ES21 ES51 ES52 ES53 ES61 ES62 ES70 FR22 FR23 FR25 FR30
FR52
[20] FR53 FR61 FR81 FR82 FR83 GR11 GR12 GR14 GR21 GR22 GR23 GR24 GR25 GR30 GR41 GR42 GR43 IS00
ITC3
[39] ITD3 ITD4 ITD5 ITE1 ITE3 ITE4 ITF1 ITF2 ITF3 ITF4 ITF5 ITF6 ITG1 ITG2 LT00 MT00 NL11 NL12
NL32
[58] NL33 NL34 NO01 NO03 NO04 NO05 NO06 NO07 PT11 PT15 PT16 PT17 PT18 PT20 PT30 RO22 SE11 SE12
SE21
[77] SE22 SE23 SE31 SE32 SE33 UKC1 UKC2 UKD1 UKD2 UKD4 UKD5 UKE1 UKE2 UKF3 UKH1 UKH3 UKJ2 UKJ3
UKJ4
[96] UKK1 UKK2 UKK3 UKK4 UKL1 UKL2 UKM2 UKM3 UKM5 UKM6 UKN0
```

For example, the value of **E09TOT** for BE25 is 93430.50 and that for BE33 is 52049.79. Formatting the Excel cell to have 0 decimal places does not change the value stored in the cell.

Duplicated Rows

The following regions appear to have duplicated data for **ALL** variables:

```
NUTSCode NUTS Region Name
DK01 Hovedstaden
DK02 Sjælland
DK03 Syddanmark
DK04 Midtjylland
DK05 Nordjylland
```

Potentially Anomalous values

A number of regions appear to have potentially anomalous values – either unusually high or unusually low. These values may well be correct, even though the data check has identified them as unusual.

```
Data Check for Indicator: P_09TOT
NUTSCode Value NUTS Region Name
PT17 36.3533709981 Lisboa
------------------------------------------------------------------- 2
```

```
Data Check for Indicator: P_09SHIPBUI
NUTSCode Value NUTS Region Name
BG33 1.10737220652 Severoiztochen
DE80 0.809885442142 Mecklenburg-Vorpommern
ES11 0.797183568036 Galicia
GR42 0.921075455334 Notio Aigaio
ITC3 1.19966106917 Liguria
MT00 2.46951219512 Malta
NL34 0.854776366648 Zeeland
NO04 3.03467694132 NA
NO05 3.04944255938 NA
NO06 0.836057454185 NA
RO22 2.19980728802 Sud-Est
UKK4 1.39941804074 Devon
-------------------------------------------------------------------
Data Check for Indicator: P_09TRADSEC
NUTSCode Value NUTS Region Name
ITD5 5.61905493268 Emilia-Romagna
-------------------------------------------------------------------
Data Check for Indicator: P_09TRANSP
NUTSCode Value NUTS Region Name
DE50 4.12420830401 Bremen
EE00 3.5808073115 Eesti
ITC3 3.84178092744 Liguria
LV00 3.86749666518 Latvija
MT00 0 Malta
RO22 3.60949544499 Sud-Est
UKE2 3.67212262925 North Yorkshire
-------------------------------------------------------------------
Data Check for Indicator: P_09TOURISM
NUTSCode Value NUTS Region Name
PT15 28.6302314131 Algarve
PT17 29.415819209 Lisboa
-------------------------------------------------------------------
Data Check for Indicator: P_09FISHERI
NUTSCode Value NUTS Region Name
DE50 1.851513019 Bremen
ES11 3.77485209566 Galicia
ES13 2.20299309286 Cantabria
GR41 1.93758865248 Voreio Aigaio
GR42 2.0988725065 Notio Aigaio
IS00 8.27437056165 NA
NO05 2.51745031508 NA
NO06 2.22833085686 NA
NO07 3.15159574468 NA
PT15 2.28705071393 Algarve
PT20 2.93435251799 Região Autónoma dos Açores
UKM5 3.95500218436 North Eastern Scotland
UKM6 2.00776836158 Highlands and Islands
-------------------------------------------------------------------
Data Check for Indicator: P_09OTHER
NUTSCode Value NUTS Region Name
EE00 3.08012185834 Eesti
ES21 3.89550889179 País Vasco
ITD3 5.03596375307 Veneto
ITD4 6.10608962083 Friuli-Venezia Giulia 3
```

```
ITD5 4.26933906722 Emilia-Romagna
ITE3 4.83338433293 Marche
ITF5 3.07692307692 Basilicata
SE21 3.71865889213 Småland med öarna
SI02 3.27386808553 Zahodna Slovenija
-------------------------------------------------------------------
Data Check for Indicator: P_09OILGAS
NUTSCode Value NUTS Region Name
DK01 0.115075 Hovedstaden
DK02 0.115075 Sjælland
DK03 0.115075 Syddanmark
DK04 0.115075 Midtjylland
DK05 0.115075 Nordjylland
EE00 0.158415841584 Eesti
GR11 0.274374460742 Anatoliki Makedonia, Thraki
ITD5 0.0459734807761 Emilia-Romagna
ITE4 0.123651302424 Lazio
ITF1 0.134833698466 Abruzzo
ITF4 0.0454226723802 Puglia
ITF5 0.178461538462 Basilicata
NL11 0.522760115607 Groningen
NL12 0.0884067357513 Friesland (NL)
NL32 0.0480102502261 Noord-Holland
NL33 0.097709481009 Zuid-Holland
NL34 0.0844837106571 Zeeland
NO01 0.212450945416 NA
NO04 3.97302904564 NA
NO05 0.900387784779 NA
NO06 0.67112431897 NA
NO07 0.132092198582 NA
RO22 0.507533286615 Sud-Est
UKM6 7.18361581921 Highlands and Islands
-------------------------------------------------------------------
```

Potential Multivariate Outliers

An outlier check on the variables P_09SHIPBUI, P_09TRADSEC, P_09TRANSP, P_09TOURISM, P_09FISHERI, P_09OTHER and P_09OILGAS considered as a group indicates the unusually high or low values are to be found in the following NUTS regions. This may arise because these regions are unusual in comparison with the other regions in the study or it may arise because of incorrect data in one or more of the constituent variables.

```
NUTS Name
165 IS00 Ísland
172 ITD3 Veneto
173 ITD4 Friuli-Venezia Giulia
174 ITD5 Emilia-Romagna
208 NO04 Agder og Rogaland
209 NO05 Vestlandet
238 RO22 Sud-Est
245 SE21 Småland med öarna
318 UKM6 Highlands and Islands 4
```

Apparent internal summation inconsistencies

P_09TOT appears to be the sum of the other P_09 variables. In the following cases, the difference between the value of P_09TOT and the sum of the other p_09 variables is greater than 0.1:

```
UnitCode P_09TOT Sum Diff
84 DK01 16.81604 16.93111 -0.1150750
85 DK02 16.81604 16.93111 -0.1150750
86 DK03 16.81604 16.93111 -0.1150750
87 DK04 16.81604 16.93111 -0.1150750
88 DK05 16.81604 16.93111 -0.1150750
89 EE00 22.52719 22.68561 -0.1584158
140 GR11 15.03063 15.30500 -0.2743745
178 ITE4 22.01097 22.13462 -0.1236513
179 ITF1 22.00518 22.14001 -0.1348337
183 ITF5 17.31231 17.49077 -0.1784615
193 NL11 17.11344 17.63620 -0.5227601
205 NO01 25.08473 25.29718 -0.2124509
208 NO04 23.42442 27.39745 -3.9730290
209 NO05 24.06568 24.96607 -0.9003878
210 NO06 19.78108 20.45220 -0.6711243
211 NO07 19.18794 19.32004 -0.1320922
238 RO22 13.76603 14.27356 -0.5075333
318 UKM6 17.01836 24.20198 -7.1836158
```

Whilst these differences may be due to rounding errors, an outlier check on the differences suggests that the values for the entries listed above should be checked, particularly for **NO04**, **NO09** and **UKM6**.

Apparent internal computation inconsistencies

The P_09 variables are computed from the E09 variables. The unstated 'base' denominator values can be recovered by dividing each count by its corresponding rate. There are 290 instances where the row mean base value (i.e. the average for each NUTS unit) differs from the corresponding individual base values by more than -1 or +1. For example, for **BE25** the 'base' population values are 5000 (when rounded) with the exception of E09SHIPBUI which would appear to be 4992.

Annex 2 – Nomenclature Update Report, Morocco

Name of the expert**: Jean Yves Moisseron** & **H.Pecout**
Role: **CIST vice director & CIST engineer**
Organisation : **GIS CIST**
Email: **huges.pecout@gis-cist.fr**
Date of the report: **2013-04-04**

1) General comment

The territorial division used by M4D is an old one. Before 2005, Morocco has been 13 prefectures & 61 provinces.

But, **in 2005, The M'diq-Fnideq prefecture has been created.** It's a part of the old prefecture of Tétouan.
And**, in 2009, 13 provinces have been created:**

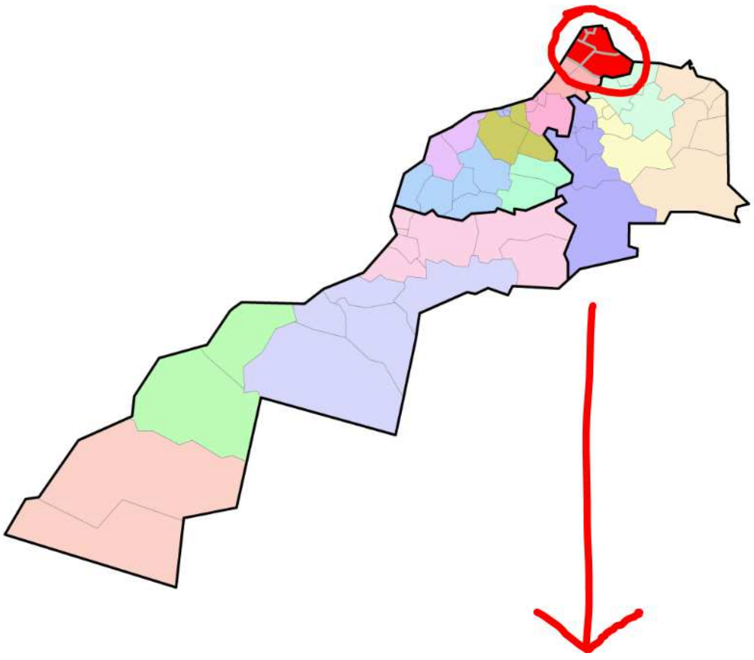
- **Tarfaya** by dividing up laayoune province (information to check)
- **Tinghir** by dividing up Ouarzazate province (part of Souss-Masse-Drâa region) & Errachidia province (part of Meknèse-Tafilalet region)
- **Sidi Ifni** by dividing up Tiznit province
- **Sidi Sliman** by dividing up Kénitra province
- **Rehanna** by dividing up El Kelâa Sraghna province
- **Driouch** by dividing up Nador province
- **Sidi Bennour** by dividing up El Jadida province (information to check)
- **Youssoufia** by dividing up Safi province (information to check)
- **Fquih Ben Salah** by dividing up Beni Mellal province
- **Midelt** by dividing up Khénifra & d'Errachidia province
- **Guercif** by dividing up Taza province
- **Ouerzane** by dividing up Sidi Kacem (part of Gharb-Chrada-Beni Hssen region) & Chefchaouen (part of tanger-tétouan region) provinces.
- **Berrechid** by dividing up Settat province

Ton conclude, we wants to add 14 SNUTS3 (province or prefecture). The News SNUTS3 have been created by dividing up old provinces. So, the delimitation of SNUTS2 (and maybe SNUTS1) have to be change too.

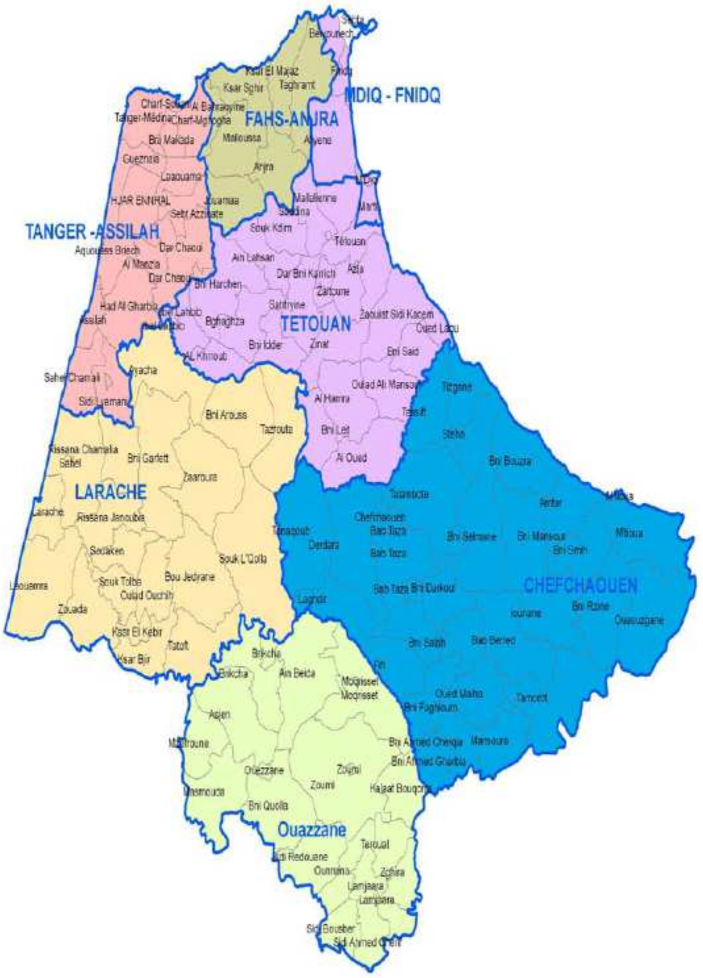
More info. :
"Le Maroc des region, 2010" from HCP. (eg. Document pdf)
http://fr.wikipedia.org/wiki/Organisation_territoriale_du_Maroc

2) Geometries

Old division => eg. The last MapKit from M4D. Ex :



New territorial division => All the map & delimitations in : « Maroc des régions, 2010.pdf ». EX :



3) Exhaustive listing of the modifications

LEVEL	M4D code	M4D NAME	Detected Problem	Status	New ITAN code	New ITAN Name
SNUTS0	MA	Morocco			MA	Morocco
SNUTS1	MA1	Nord-ouest			MA1	Nord-ouest
SNUTS1	MA2	Nord-est			MA2	Nord-est
SNUTS1	MA3	Sud-Occidental			MA3	Sud-Occidental
SNUTS2	MA11	Rabat Sale Zemmour Zear			MA11	Rabat Sale Zemmour Zear
SNUTS2	MA12	Doukkala Abda			MA12	Doukkala Abda
SNUTS2	MA13	Tadla Azilal			MA13	Tadla Azilal
SNUTS2	MA14	Tanger Tetouan			MA14	Tanger Tetouan
SNUTS2	MA15	Gharb Chrarda Beni Hssen			MA15	Gharb Chrarda Beni Hssen
SNUTS2	MA16	Chaouia Ouardigha			MA16	Chaouia Ouardigha
SNUTS2	MA17	Marrakech Tensift Al Haouz			MA17	Marrakech Tensift Al Haouz
SNUTS2	MA18	Grand Casablanca			MA18	Grand Casablanca
SNUTS2	MA21	Meknes Tafilalet			MA21	Meknes Tafilalet
SNUTS2	MA22	Fes Boulemane			MA22	Fes Boulemane
SNUTS2	MA23	Taza Al Hoceima Taounat			MA23	Taza Al Hoceima Taounat
SNUTS2	MA24	L'Oriental			MA24	L'Oriental
SNUTS2	MA31	Oued Ed dahab Lagouira			MA31	Oued Ed dahab Lagouira
SNUTS2	MA32	Laayoune Boujdour Sakia El Hamra			MA32	Laayoune Boujdour Sakia El Hamra
SNUTS2	MA33	Guelmim Es semara			MA33	Guelmim Es semara
SNUTS2	MA34	Souss Massa Draa			MA34	Souss Massa Draa
SNUTS3	MA111	Rabat			MA111	Rabat
SNUTS3	MA112	Khemisset			MA112	Khemisset
SNUTS3	MA113	Sale			MA113	Sale
SNUTS3	MA114	Skhirate Temara			MA114	Skhirate Temara
SNUTS3	**MA121**	**Al Jadida**			**MA121**	**Al Jadida**
SNUTS3	**MA122**	**Safi**			**MA122**	**Safi**
			Province created in 2009	**Split of El Jadida?**	**MA123**	**Sidi Bennour**
			Province created in 2009	**Split of Safi ?**	**MA124**	**Youssoufia**
SNUTS3	MA131	Azilal			MA131	Azilal
SNUTS3	**MA132**	**Beni Mellal**			**MA132**	**Beni Mellal**
			Province created in 2009	**Split of Beni Mellal**	**MA133**	**Fquih Ben Salah**
SNUTS3	MA141	Chefchaouen			**MA141**	**Chefchaouen**
SNUTS3	MA142	Fahs Anjra			MA142	Fahs Anjra
SNUTS3	MA143	Larache			MA143	Larache
SNUTS3	MA144	Tanger Assilah			MA144	Tanger Assilah
SNUTS3	**MA145**	**Tetouan**			**MA145**	**Tetouan**
			Prefecture created in 2005	**Spit of Tétouan prefecture**	**MA146**	**M'diq-Fnideq**
			Province created in 2009	**Split of sidi Kacem + Chefchaouen (from 2 differents regions)**	**MA147**	**Ouezzane**
SNUTS3	**MA151**	**Kenitra**			**MA151**	**Kenitra**

SNUTS3	**MA152**	**Sidi Kacem**			**MA152**	**Sidi Kacem**
			Province created in 2009	**Split of Kenitra**	**MA153**	**Sidi Sliman**
SNUTS3	MA161	Benslimane			MA161	Benslimane
SNUTS3	MA162	Khouribga			MA162	Khouribga
SNUTS3	**MA163**	**Settat**			**MA163**	**Settat**
			Province created in 2009	**Split of settat province**	**MA164**	**Berrechid**
SNUTS3	MA171	Al Haouz			MA171	Al Haouz
SNUTS3	MA172	Chichaoua			MA172	Chichaoua
SNUTS3	MA173	El Kelaa Sraghna			MA173	El Kelaa Sraghna
SNUTS3	MA174	Essaouira			MA174	Essaouira
SNUTS3	MA175	Marrakech			MA175	Marrakech
			Province created in 2009	**split of El Kelâa Sraghna province**	**MA176**	**Rehanna**
SNUTS3	MA181	Casablanca			MA181	Casablanca
SNUTS3	MA182	Madiouna			MA182	Madiouna
SNUTS3	MA183	Mohammedia			MA183	Mohammedia
SNUTS3	MA184	Nouaceur			MA184	Nouaceur
SNUTS3	MA211	Meknes			MA211	Meknes
SNUTS3	MA212	El Hajeb			MA212	El Hajeb
SNUTS3	**MA213**	**Errachidia**			**MA213**	**Errachidia**
SNUTS3	MA214	Ifrane			MA214	Ifrane
SNUTS3	**MA215**	**Khenifra**			**MA215**	**Khenifra**
			Province created in 2009	**Split of Khénifra & d'Errachidia province (from 2 differents regions)**	**MA216**	**Midelt**
SNUTS3	MA221	Boulemane			MA221	Boulemane
SNUTS3	MA222	Fes			MA222	Fes
SNUTS3	MA223	Sefrou			MA223	Sefrou
SNUTS3	MA224	Moulay Yacoub			MA224	Moulay Yacoub
SNUTS3	MA231	Al Hoceima			MA231	Al Hoceima
SNUTS3	MA232	Taounate			MA232	Taounate
SNUTS3	**MA233**	**Taza**			**MA233**	**Taza**
			Province created in 2009	**Split of Taza**	**MA234**	**Guercif**
SNUTS3	MA241	Berkane			MA241	Berkane
SNUTS3	MA242	Figuig			MA242	Figuig
SNUTS3	MA243	Jerada			MA243	Jerada
SNUTS3	**MA244**	**Nador**			**MA244**	**Nador**
SNUTS3	MA245	Ouajda Angad			MA245	Ouajda Angad
SNUTS3	MA246	Taourirt			MA246	Taourirt
			Province created in 2009	**Split of Nador**	**MA247**	**Driouch**
SNUTS3	MA311	Aousserd			MA311	Aousserd
SNUTS3	MA312	Oued Ed Dahab			MA312	Oued Ed Dahab
SNUTS3	MA321	Boujdour			MA321	Boujdour
SNUTS3	MA322	Laayoune			MA322	Laayoune
			Province created in 2009	**Split of Laayoune ?**	**MA323**	**Tarfaya**
SNUTS3	MA331	Assa Zag			MA331	Assa Zag

SNUTS3	MA332	Es Semara			MA332	Es Semara
SNUTS3	MA333	Guelmim			MA333	Guelmim
SNUTS3	MA334	Tan Tan			MA334	Tan Tan
SNUTS3	MA335	Tata			MA335	Tata
SNUTS3	MA341	Agadir Ida Ou Tanane			MA341	Agadir Ida Ou Tanane
SNUTS3	MA342	Chouka Ait Baha			MA342	Chouka Ait Baha
SNUTS3	MA343	Inezgane Ait Melloul			MA343	Inezgane Ait Melloul
SNUTS3	**MA344**	**Ouarzazate**			**MA344**	**Ouarzazate**
SNUTS3	MA345	Taroudannt			MA345	Taroudannt
SNUTS3	**MA346**	**Tiznit**			**MA346**	**Tiznit**
SNUTS3	MA347	Zaroga			MA347	Zagora
			Province created in 2009	**Split of Ouarzazate & Errachidia (from 2 differents regions)**	**MA348**	**Tinghir**
			Province created in 2009	**Split of Tiznit**	**MA349**	**Sidi Ifni**