TECHNICAL REPORT

ESPON

June 2013

A strategy for elaboration and update of coherent time series of hierarchical territorial units.

SUMMARY

This technical report proposes a coherent strategy for the collection of coherent time series of core data. This strategy involves the following topic :

1. Identification of problems encountered until now in the estimation of missing values and outlier check for time series of count data
2. Definition of conditions for a long terme strategy
3. Proposal of a coherent solution to be implemented before the end of M4D project and likely to be further developed and improved by next ESPON program.

ESPON M4D -
MULTI DIMENSIONAL DATABASE DESIGN & DEVELOPMENT

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

26 pages

LIST OF AUTHORS

Claude Grasland, UMS 2414 RIATE

Ronan Ysebaert, UMS 2414 RIATE

Timothée Giraud, UMS 2414 RIATE

Martin Charlton, NCG

Alberto Caimo, NCG

Hélène Mathian, UMR Géographie-cités

Christine Plumejeaud, UMR 7266 LIENSs

Contact

ronan.ysebaert@ums-riate.fr

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

TABLE OF CONTENT

Introduction

The Core Database Strategy is an important part of the ESPON M4D project, corresponding, in general, to activities developed in work package B (Thematic group) but also, to some extent, to activities developed in work package A (e.g. storage of time series, identification of core data in the interface) and work package C (identification of outliers value, quality check).

This activity which is normally supposed to cover half of the activity of the ESPON M4D project (according to contract definition) has been in practical terms delayed to the final period of activity 2013-2014. The reason of this delay was the priority decided by ESPON CU on storage and diffusion of data collected by ESPON project through the web interface. As a result, the major part of available workforce has been concentrated on this part of the work in 2011-2012.

The opportunity of this priority will not to be discussed here. But it is nevertheless clear that Core Database Strategy remains a major contractual obligation for M4D (deliverables related to this task has been delayed, not removed). And it is also very clear that, in a long term perspective, the Core Database Strategy is as important as the storage of data collected by ESPON 2013 projects.

If a new ESPON is launched for the period 2014-2020, the priority in terms of data collection will necessarily be the collection of the basic count data (population, activity, production, land use) and their elaboration in time series as long as possible. If such data are not available immediately, many difficulties will be encountered by new project, as we now by experience of what happened in the beginning of the programming period 1999-2006 and 2007-2013.

Moreover, if we consider a cross-programming period perspective, we can consider that a major added value of the ESPON program could be to produce cumulative efforts which mean, in practical terms, to enlarge past time series in order to be able in the future to propose more accurate previsions. With coherent time series covering the period 1990-2010, it is reasonably possible to expect accurate predictions for the period 2010-2030, which is a major wish of stakeholders and policy makers.

The problem is that building such long term time series is a very difficult and complicated task, that can only be engaged for a limited set of indicators. The aim of the Core Database Strategy is precisely to define what are the indicators to be completed in priority in order to derive many others by intelligent procedures of aggregation, disaggregation, spatial analysis, etc…

Because of pressure on other objectives in 2011-2012, the elaboration of the long term time series of core data has been until now limited to few indicators. This is not really a problem as long as we have demonstrated that a lot of information can be derived from a limited number of core indicators. But what is more tricky is the fact that:

- Estimation of time series of core data is actually based on a manual procedure that consumes a lot of time and is difficult to replicate
- Outlier check of time series of count data is difficult to realize because the indicators are not ratio but absolute count, which limit the use of numerous methods of outlier detection.

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 time1". 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 which remains, generally, constant during the time periods.

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

1 Diagnosis of the difficulties encountered with time series of count data

Time series of count data are very specific statistical objects that require various and different procedures of outlier check and estimation. To illustrate this point, let us start with a basic analysis of the time series of population at NUTS3 level from 1990 to 2010 delivered by M4D project.

1.1 Heterogeneous sources and heterogeneous methods of estimation

The elaboration of a complete table of population for all NUTS3 regions implies a very huge amount of empirical work by human specialist in order to remove every missing value from the table. Until now, the strategy developed by M4D has been (1) *to choose the best available data and the best method of estimation for the estimation of each missing value and (2) to store all metadata related to the various sources and the various methods of estimation*. This strategy is illustrated in Figure 1 where sample of data and metadata are displayed.

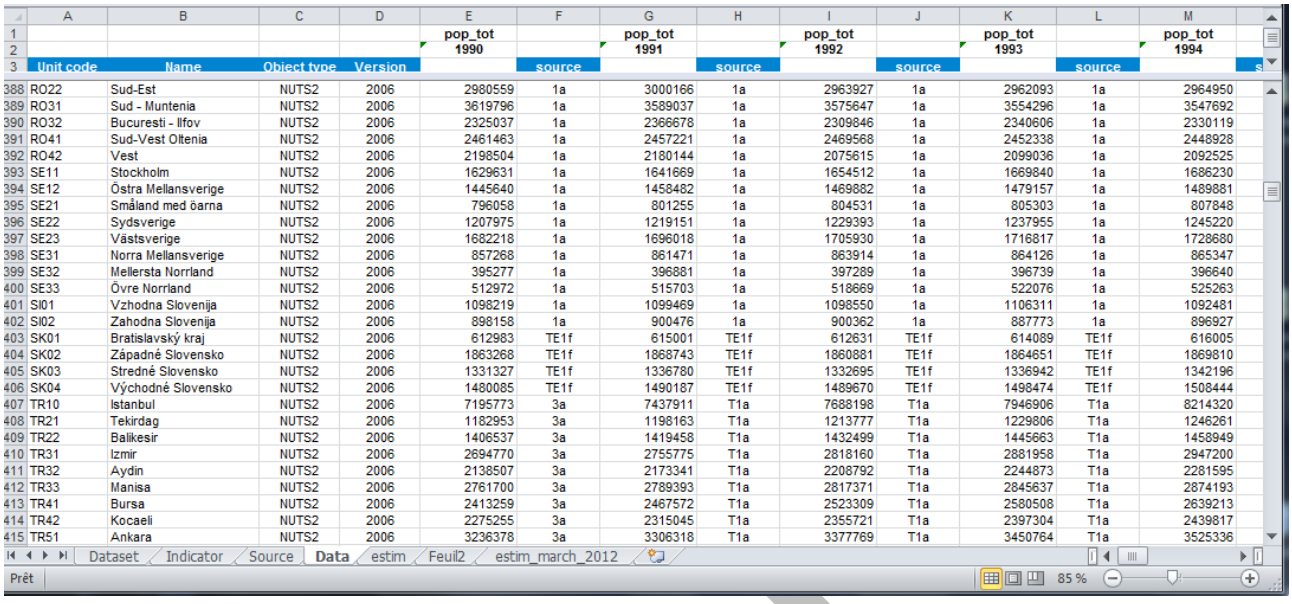
What are the strengths or weaknesses of this solution? The answer is not obvious because each strategic choice has a double face.

The multiplication of sources can't be avoided because no data provider is able to provide complete time series for the period at the targeted territorial level (NUTS3, Version 2006). Eurostat is generally chosen has prior provider but in many cases the missing data are necessarily collected through complementary sources provided by National Statistical Offices (NSI) of the countries. The problem is of course that NSI does not necessary use the same territorial division than Eurostat (risk of error) nor the same definition. Even if it is the case, it can happen that updates are made by NSI on data that are not transmitted to Eurostat, or only with delays. In this case different figures characterize the same territorial unit at the same time, according to NSI and Eurostat. And it is not obvious to decide what is the best one: Eurostat has a political legitimacy at EU level, but NSI are the highest legitimacy at national level and are at least the responsible of initial data collection. Our purpose is not to solve this theological question but to underline the fact that mixture of several sources can increase the risk of "breaks" in time series.

The multiplication of estimation methods for missing values lead to the same dilemma. M4D project has proposed a catalogue of solutions that are well documented and helpful for the human experts in order to choose the best one in each particular situation. But some of these methods are very sophisticated and can only be applied without error by very few human experts. Moreover, the work of estimation is actually done manually (by "click" in excel sheets) and cannot be automatically reproduced, even when the detailed method is precised in metadata file. This is a real problem when it comes to update time series because new information are added (e.g. publication by Eurostat of new figures of population for 2010) or when old information are modified (e.g. replacement of provisional figures by definitive ones). A classic example is the "break" in time series introduced by the result of a census: the estimation used between two census dates should normally be modified but in practical terms, it is generally not the case, creating automatically a time outlier at census date.

Figure 1 : The M4D estimated time series of population (1990-2010)

(a) sample of data



	A	B	C	D	E	F	G	H	I	J	K	L	M
1					pop_tot		pop_tot		pop_tot		pop_tot		pop_tot
2					1990		1991		1992		1993		1994
3	Unit code	Name	Object type	Version		source		source		source		source	
388	RO22	Sud-Est	NUTS2	2006	2980559	1a	3000166	1a	2963927	1a	2962093	1a	2964950
389	RO31	Sud - Muntenia	NUTS2	2006	3619796	1a	3589037	1a	3575647	1a	3554296	1a	3547692
390	RO32	Bucuresti - Ilfov	NUTS2	2006	2325037	1a	2366678	1a	2309846	1a	2340606	1a	2330119
391	RO41	Sud-Vest Oltenia	NUTS2	2006	2461463	1a	2457221	1a	2469568	1a	2452338	1a	2448928
392	RO42	Vest	NUTS2	2006	2198504	1a	2180144	1a	2075615	1a	2099036	1a	2092525
393	SE11	Stockholm	NUTS2	2006	1629631	1a	1641669	1a	1654512	1a	1669840	1a	1686230
394	SE12	Östra Mellansverige	NUTS2	2006	1445640	1a	1458482	1a	1469882	1a	1479157	1a	1489881
395	SE21	Småland med öarna	NUTS2	2006	796058	1a	801255	1a	804531	1a	805303	1a	807848
396	SE22	Sydsverige	NUTS2	2006	1207975	1a	1219151	1a	1229393	1a	1237955	1a	1245220
397	SE23	Västsverige	NUTS2	2006	1682218	1a	1696018	1a	1705930	1a	1716817	1a	1728680
398	SE31	Norra Mellansverige	NUTS2	2006	857268	1a	861471	1a	863914	1a	864126	1a	865347
399	SE32	Mellersta Norrland	NUTS2	2006	395277	1a	396881	1a	397289	1a	396739	1a	396640
400	SE33	Övre Norrland	NUTS2	2006	512972	1a	515703	1a	518669	1a	522076	1a	525263
401	SI01	Vzhodna Slovenija	NUTS2	2006	1098219	1a	1099469	1a	1098550	1a	1106311	1a	1092481
402	SI02	Zahodna Slovenija	NUTS2	2006	898158	1a	900476	1a	900362	1a	887773	1a	896927
403	SK01	Bratislavský kraj	NUTS2	2006	612983	TE1f	615001	TE1f	612631	TE1f	614089	TE1f	616005
404	SK02	Západné Slovensko	NUTS2	2006	1863268	TE1f	1868743	TE1f	1860881	TE1f	1864651	TE1f	1869810
405	SK03	Stredné Slovensko	NUTS2	2006	1331327	TE1f	1336780	TE1f	1332695	TE1f	1336942	TE1f	1342196
406	SK04	Východné Slovensko	NUTS2	2006	1480085	TE1f	1490187	TE1f	1489670	TE1f	1498474	TE1f	1508444
407	TR10	Istanbul	NUTS2	2006	7195773	3a	7437911	T1a	7688198	T1a	7946906	T1a	8214320
408	TR21	Tekirdag	NUTS2	2006	1182953	3a	1198163	T1a	1213777	T1a	1229806	T1a	1246261
409	TR22	Balikesir	NUTS2	2006	1406537	3a	1419458	T1a	1432499	T1a	1445663	T1a	1458949
410	TR31	Izmir	NUTS2	2006	2694770	3a	2755775	T1a	2818160	T1a	2881958	T1a	2947200
411	TR32	Aydin	NUTS2	2006	2138507	3a	2173341	T1a	2208792	T1a	2244873	T1a	2281595
412	TR33	Manisa	NUTS2	2006	2761700	3a	2789393	T1a	2817371	T1a	2845637	T1a	2874193
413	TR41	Bursa	NUTS2	2006	2413259	3a	2467572	T1a	2523309	T1a	2580508	T1a	2639213
414	TR42	Kocaeli	NUTS2	2006	2275255	3a	2315045	T1a	2355721	T1a	2397304	T1a	2439817
415	TR51	Ankara	NUTS2	2006	3236378	3a	3306318	T1a	3377769	T1a	3450764	T1a	3525336

Dataset | Indicator | Source | Data | estim | Feuil2 | estim_march_2012

Prêt 85 %

(b) sample of metadata

925	**Source Reference**		
926	**Label**	TE1f	
927	**Date**	2011-11-07	
928	**Copyright**	© ESPON	
929	**Provider**	**Name**	ESPON M4D
930		**URI**	
931	**Publication**	**Title**	
932		**URI**	
933		**Reference**	
934	**Methodology**	**Description**	Estimation based on the time and space dimensions (E1) 1/ Time dimension - power retropolation. This method uses the two closest neighbours placed in time after (1996 and 1997) the value estimated. 2/ Space dimension – The estimated values have been adjusted in a way that the sum of values of children units (e1,e2…en) are equal to the value of the parent unit (parent(e1,e2..en)). In this case, the NUTS3 values coming from the estimation are adjusted to the NUTS0 values
935		**URI**	
936	**Access Rule**	public	
937	**Estimation**	true	
938	**Quality Level**	medium	
939	**Source Reference**		
940	**Label**	TE1g	
941	**Date**	2011-11-07	
942	**Copyright**	© ESPON	
943	**Provider**	**Name**	ESPON M4D
944		**URI**	
945	**Publication**	**Title**	
946		**URI**	
947		**Reference**	

Dataset | Indicator | Source | Data | estim | Feuil2 | estim_march_2012

1.2 The dilemma: local precision versus global homogeneity

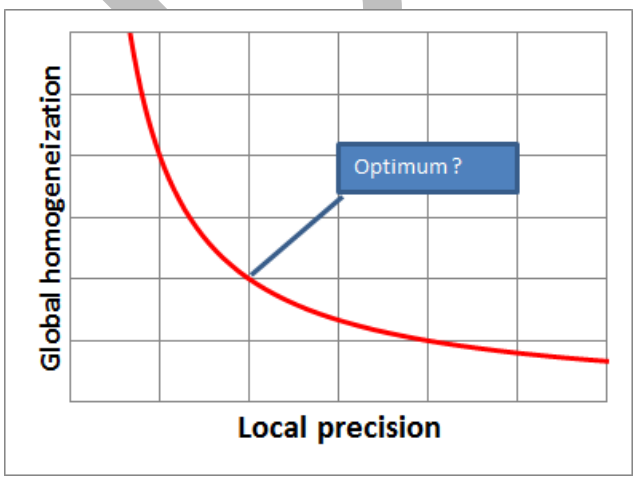
We propose to reconsider the strategic choice made until now to use the best available data or methods for the *local precision* of estimation of missing value and to examine an alternative where the focus is made on the global homogeneity of the solution. The main problem with the initial strategy (local precision) is the fact that all sources or used methods can be perfectly correct, but at the end the emerging global result is not so good. The problem clearly appeared when UMS RIATE started to realize a very basic outliers check of the time series of population, before to transmit to NCG for an in depth analysis of outliers combining all criteria. The very simple method we used has revealed so much anomalies in time series that we decided to postpone the transmission of data to NCG and also decided to reconsider the opportunity to disaggregate or aggregate data with OLAP cube as long as we would not have understood the reason of the apparition of such a big number of time outliers (ex. Figure 2)

Figure 2 : Example of time outlier check for population (1990-2010)

code	Name	vpr1990	vpr1991	vpr1992	vpr1993	vpr1994	vpr1995	vpr1996	vpr1997	vpr1998	vpr1999	vpr2000	vpr2001	vpr2002	vpr2003	vpr2004	vpr2005	vpr2006	vpr2007	vpr2008	vpr2009	vpr2010	vpr2011
SK010	Bratislavský kraj	-1	1	0	0	0	0	0	0	-4	4	0	0	0	0	0	0	0	0				
SK021	Trnavský kraj	-4	3	0	0	0	-1	1	0	-1	0	0	0	1	0	0	0	1	1	-1	0		
SK022	Trenciansky kraj	-3	3	0	0	0	-1	0	0	1	-2	1	0	0	0	0	0	0	0	-1	0		
SK023	Nitriansky kraj	-4	3	0	0	0	-1	0	0	0	-1	1	0	0	1	0	0	0	0	0	0		
SK031	Zilinský kraj	-4	3	0	0	-1	1	-1	1	-1	-2	1	0	0	0	0	0	0	0	0	0		
SK032	Banskobystrický kraj	-4	3	0	0	0	-1	0	-1	0	1	-1	0	0	0	-1	0	0	0	0	0		
SK041	Presovský kraj	-4	3	0	0	-3	2	0	0	0	1	-1	0	0	0	0	0	0	0	1	0		
SK042	Kosický kraj	-4	4	0	0	-2	2	0	0	0	-2	1	0	0	0	0	0	0	0	1	-1		

The provisional diagnosis that we have made on population data lead to an interesting but striking conclusion: **the more we try to obtain exact value of isolated figure, the more we increase the number of outliers in time series**. To be sure, the objective of local optimization is to some extent contradictory with the objective of global homogenization of time series and we have to explore the possibility to define an optimum which is necessarily a compromise (Figure 3)

Figure 3: The compromise between local precision and global homogeneity of time series



1.3 A strategy based on the joint operation of data estimation and outlier check.

Our solution to the dilemma is not to choose one strategy against the other because both approaches are admittedly possible, depending on the user's needs.

- **The strategy of best local estimation** is typically convenient for users looking for official isolated figures and trying to answer to precise question like "*what was the population of Flanders in 1991?*". What is important for such users is to have very precise metadata defining the original sources (*e.g. Belgium NSI*) or the method of estimation used (*e.g. interpolation between year 1990 and year 1995 under the assumption of exponential growth*). For such a user, the discontinuities in time series are not a problem, precisely because they are looking for a single time period or a single unit.
- **The strategy of global homogenisation** is typically convenient for users not interested in the analysis of specific situation but the examination of global trends in space or time. For example, a user trying to answer to a question like "*What has been the profile of population growth or decline of EU regions between 1990 and 2010*". In this case the degree of precision of a specific figure is not important at all. But discontinuities or outliers in time series are on the contrary a real threat for the analysis because they can introduce the apparition of specificities in time series that are purely artificial and related only to a change of sources or methods of estimation.

It is very clear that for the majority of data currently involved in the ESPON database, the strategy of local estimation appears as the best solution. But it is not the case for the long term time series called core data where the major interest is precisely to produce global evolutions and prospective results. We suggest therefore adopting a new approach for this specific group of data that will be explained and developed in the next section.

The major originality of the approach proposed in this strategy is the fact that estimation of missing values and outliers' checking are not realized in separate steps but together. We try to obtain free time series that are eventually simplified or that are fully consistent in temporal and territorial terms. Equally, we would like the method reproducible so that datasets can be easily updated and recomputed when new information are made available or when boundaries changes occur in territorial divisions (like the reform of NUTS).

1.4 Relation with previous technical report or work done by M4D

The strategy proposed in the next section is the result of previous research done by NCG and RIATE. The reader can find more details in the following technical reports or annex of the first interim report of M4D.

Work previously done by NCG

Work previously done by RIATE

2 Solution proposed for homogenization and update of times series of core data

2.1 General rules and objectives

The solution that will be developed in the following section is based on a limited number of rules that should normally be followed without exceptions, in order to fulfill precise objectives

1. *Only one primary source is normally used for the production of time series*. The fact to use different sources for the same territorial unit is indeed a major factor of creation of "breaks" or heterogeneity. It means that we will normally prefer to estimate values rather than use alternative data source.

2. *All times series should be perfectly consistent in terms of hierarchical aggregation of territories*. The different subdivision of data provided by a primary producer should be perfectly exact. If data provided by the initial producer does not follow this rule, they will be modified in order to fulfill perfectly the aggregation rules of the nomenclature.

3. *All time series should be free of time outlier, except when the outlier can be explained by concrete and real facts*. It means that we prefer to obtain values that are different from the official one when an obvious statistical bias is present in time series of the data producer. Typically, when a new census creates a discontinuity in the time series, we will recalculate the values between theses census and the previous one. More generally we will try in the majority of case to obtain stationary time series as long as we have no reason to suspect that specific event has created discontinuities.

4. *All estimation of missing values should be made by mean of an automatic procedure that can be repeated quickly and – ideally - without manual intervention*. This rule is the most difficult but also the most important because time series should be regularly modified for different reasons : (1) introduction of recent data provided by data producer ; (2) discovery of errors in existing data or modification of provisional values in definitive ones ; (3) discovery of new estimation methods that could improve previous ones.

5. *All procedures and methods used in the estimation should be transparent and added in the metadata field*. This general rule of the ESPON database is just reminded here but remains very important. The user of time series should be perfectly aware of the fact that data that are sometime different from "official statistics" because of the target of global homogeneity.

6. *An estimation of uncertainty should be ideally added to all figures of time series*. In principle, we do not need to introduce here an outlier check of this data because we have precisely decided to remove outliers. But we should ideally indicate the 95% confidence interval of values present in time series, not only for estimated values but also for the other ones.

2.2 A data model combining time and territorial hierarchy

The application of previous rules (1) and (2) lead us to propose a specific data model for the storage of time series describing hierarchical territorial units like NUTS. To illustrate the strategy, we will take the example of estimation of missing data for active population of Bulgaria between 1999 and 2010 on the basis of EUROSTAT data at NUTS0, NUTS1 and NUTS2 levels (version 2006). This specific data model can be firstly presented in tabular format (Figure 4) but is more clear if presented in form of hierarchical trees of data linked through time (Figure 5).

Figure 4 : Illustration of the strategy of hierarchical data reconstitution

Step1 : Initial data

code	name	level	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
BG	Bulgaria	NUTS0		2794.7	2702.8	2741.0	2834.7	2922.6	2981.9	3110.0	3252.6	3360.7	3253.6	3052.8
BG3	Severna i iztochna Bulgari	NUTS1					1412.0	1445.6	1476.3	1529.5	1581.7	1632.2	1571.9	1465.9
BG4	Yugozapadna i yuzhna tse	NUTS1					1422.8	1477.0	1505.6	1580.5	1670.9	1728.5	1681.7	1586.9
BG31	Severozapaden	NUTS2					315.7	318.3	314.6	327.7	345.4	359.3	341.3	313.7
BG32	Severen tsentralen	NUTS2					335.1	344.9	344.4	352.0	368.3	374.4	365.6	336.0
BG33	Severoiztochen	NUTS2					350.5	361.3	389.3	405.0	413.4	429.1	409.5	387.5
BG34	Yugoiztochen	NUTS2					410.6	421.1	428.0	444.8	454.6	469.4	455.6	428.7
BG41	Yugozapaden	NUTS2					855.4	894.5	920.7	974.1	1025.3	1060.2	1042.4	991.3
BG42	Yuzhen tsentralen	NUTS2					567.3	582.5	584.9	606.4	645.6	668.3	639.2	595.7

Step2 : Estimation and check

code	name	level	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
BG	Bulgaria	NUTS0	**2794.7**	**2794.7**	**2702.8**	**2741.0**	**2834.7**	**2922.6**	**2981.9**	**3110.0**	**3252.6**	**3360.7**	**3253.6**	**3052.8**
BG3	Severna i iztochna Bulgari	NUTS1	0.4981	0.4981	0.4981	0.4981	0.4981	0.4946	0.4951	0.4918	0.4863	0.4857	0.4831	0.4802
BG4	Yugozapadna i yuzhna tse	NUTS1	0.5019	0.5019	0.5019	0.5019	0.5019	0.5054	0.5049	0.5082	0.5137	0.5143	0.5169	0.5198
BG31	Severozapaden	NUTS2	0.2236	0.2236	0.2236	0.2236	0.2236	0.2202	0.2131	0.2143	0.2184	0.2201	0.2171	0.2140
BG32	Severen tsentralen	NUTS2	0.2373	0.2373	0.2373	0.2373	0.2373	0.2386	0.2333	0.2301	0.2329	0.2294	0.2326	0.2292
BG33	Severoiztochen	NUTS2	0.2482	0.2482	0.2482	0.2482	0.2482	0.2499	0.2637	0.2648	0.2614	0.2629	0.2605	0.2643
BG34	Yugoiztochen	NUTS2	0.2908	0.2908	0.2908	0.2908	0.2908	0.2913	0.2899	0.2908	0.2874	0.2876	0.2898	0.2924
BG41	Yugozapaden	NUTS2	0.6013	0.6013	0.6013	0.6013	0.6013	0.6056	0.6115	0.6163	0.6136	0.6134	0.6199	0.6246
BG42	Yuzhen tsentralen	NUTS2	0.3987	0.3987	0.3987	0.3987	0.3987	0.3944	0.3885	0.3837	0.3864	0.3866	0.3801	0.3754

Step3 : Core data

code	name	level	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
BG	Bulgaria	NUTS0	**2794.7**	**2794.7**	**2702.8**	**2741.0**	**2834.7**	**2922.6**	**2981.9**	**3110.0**	**3252.6**	**3360.7**	**3253.6**	**3052.8**
BG3	Severna i iztochna Bulgari	NUTS1	1392.0	1392.0	1346.3	1365.3	1412.0	1445.6	1476.3	1529.5	1581.7	1632.2	1571.9	1465.9
BG4	Yugozapadna i yuzhna tse	NUTS1	1402.7	1402.7	1356.5	1375.7	1422.7	1477.0	1505.6	1580.5	1670.9	1728.5	1681.7	1586.9
BG31	Severozapaden	NUTS2	311.3	311.3	301.0	305.3	315.7	318.3	314.6	327.7	345.4	359.3	341.3	313.7
BG32	Severen tsentralen	NUTS2	330.4	330.4	319.5	324.0	335.1	344.9	344.4	352.0	368.3	374.4	365.6	336.0
BG33	Severoiztochen	NUTS2	345.6	345.6	334.2	338.9	350.5	361.3	389.3	405.0	413.4	429.1	409.5	387.5
BG34	Yugoiztochen	NUTS2	404.8	404.8	391.5	397.0	410.6	421.1	428.0	444.8	454.6	469.4	455.6	428.7
BG41	Yugozapaden	NUTS2	843.4	843.4	815.6	827.2	855.4	894.5	920.7	974.1	1025.3	1060.2	1042.5	991.2
BG42	Yuzhen tsentralen	NUTS2	559.3	559.3	540.9	548.6	567.3	582.5	584.9	606.4	645.6	668.3	639.2	595.7

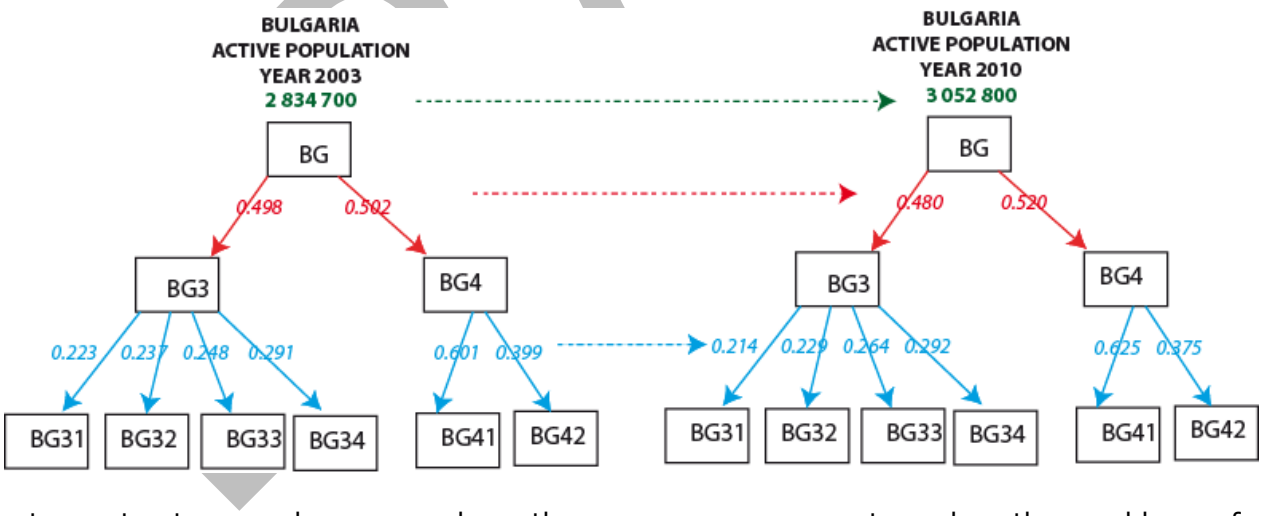
- Count variable is estimated only for the top level of hierarchy: for example, the population of Bulgaria in 1999 is estimated in number of active.
- All other territorial units are described by a frequency of the unit of upper level: for example the unit BG3 is described by the evolution of its share of BG and the unit BG31 as a share of BG3.
- Consistency of hierarchy is imposed. When the sum of frequency of child units depending from the same parent is different from 1.00, the value is adjusted. For example, the sum of BG31+BG32+BG33+BG34 in 2003 is equal to 1411.9 when BG3 is declared as 1412.0. This is not an error but simply a question of rounded value. Nevertheless it is corrected.

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.

Figure 5 : Tree representation of the data model



The tree structure make more clear the way we propose to solve the problem of homogenization and estimation of missing values by dividing a big problem in smaller parts more easy to solve, according to René Descartes' method: "*The second [principle] is to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution*". Basically, the problem that we have to solve is reduced to a combination of vertical and horizontal analysis of the trees.

- **Vertical analysis** will allow at the same time to check for logical errors (are sum of frequency of child of the same parent always equal to 1) and to estimate some missing values (definition of the value of a parent by sum of child or estimation of one missing child value by difference between the parent and the other childs …).

- **Horizontal analysis** will allow estimating missing values by mean of method of time series analysis and also to check for time outlier and provide margin of errors. But the important point is the fact that this estimation are made for small groups of time series that are typically the frequency of all the child of the same parent. This frequency is related to internal redistribution and not to external or general trends that are only taken into account for the estimation of the raw count at the top level of the tree.

With these reduction of problem in smaller parts, it appears more easy to propose automatic procedure of data check and data estimation that verify objectives (1) and (2) but can be implemented in a computer program, fulfilling the objectives (4),(5),(6). The most important difficulty remains the objective (3) related to the decision on what are real "breaks" in time series explainable by concrete fact and what are simple noise or biases to be eliminated by the procedure. To fulfill this final objectives, it would certainly be necessary to couple the estimation procedure with an expert system where human are invited to give advices on ambiguous cases where the algorithm cannot decide alone of the solution. For more details on this point, see the work realized by C. Plumejeaud (2010) in its Ph'D.

According to the time remaining in ESPON M4D project, we will focus on the production of a fully automatic solution without human expertise. The optimization of the procedure with expert intervention will be let opened for future work in ESPON 2014-2020.

2.3 Theoretical considerations

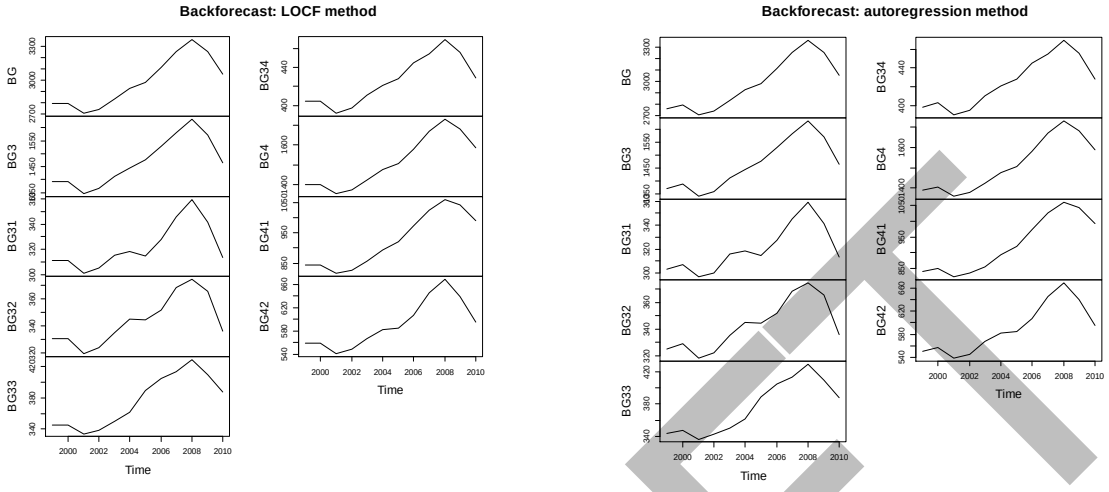
Of the several expedient strategies for dealing with data of this sort, one is do nothing and with the available data. A second strategy is to examine the relationships between the employment counts and the lower NUTS levels and their parent zones. There are also hierarchical constraints in that at any 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 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 ($2 \sim m \sim 4$); more of the supplied data is used with this approach.

Other strategies include hotdeck[2], k-nearest neighbour, and iterative model-based imputation[3]. Such strategies have been suggested for handling both unit and item non-response in longitudinal surveys[4]. 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.

[2] 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.
[3] 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
[4] 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

Figure 6: LOCF and autoregression methods for backforecasting



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.

After passing these two challenging tasks, it will be possible to define a strategy to automate the process of time-series creation/estimation. A first insight of what could be an automatic procedure is described in part 3.

3 Further insights in automation process

Up to now, inside ESPON 2013 database, checking and harmonization of datasets were made manually, (or semi-manually, by using Excel macros running on Excel files). Now, it is aimed to automate these processes, by using the spatiotemporal database from which a lot of benefits could be retrieved. At first, it can be useful to draw a diagram that summarizes the kind of activities to be done for achieving a good quality inside the database.

3.1 Data workflow

As seen on Figure 7, ESPON people first collect data, from Eurostat but also from National Statistical Institutes, or from the ESPON 2013 database itself. A set of *n* values (disregarding their associated unit or their associated indicator) is constituted. Some of them (*p, p<n*) can be missing values (or gaps), some of them (*q, q<n*) can be outliers, which means only unusual values, or abnormal values regarding a given statistical model. For instance, breaks in statistical time series are tagged as "outliers", in the sense they deviate from a smoothed model.

Those gaps and outliers have to be detected. Whenever all is fine (*p=0* and *q=0*), data can be disseminated as it. Most often, there are gaps and outliers (at least 20% of cases), and then some strategies (up to *m* different ones) must be chosen among a vast quantity of statistical methods, from simple ones to very complicated ones. Running a strategy consists in executing a statistical method based on assumption that will produce a set of values, up to *n*, but at least *p+q* values. Choosing a strategy requires analyzing the situation, with a human watch to feel which strategies could best suit to problems. After that, various strategies can be run simultaneously, and the best would be to compare their results in order to choose the best set of values (according criteria to be defined…which depend highly upon the final objective of this work!). At this stage, the process can be reiterated up to get a set of values that look convenient for the purpose.

Then a set of questions have to be answered:

- Do we stick to official data?
- Do we remove all values?
- Do we remove outliers?

Following the answers, we can replace all values, only gaps or gaps and outliers values. Then a set of metadata describing the transformation process and what have been diagnosed is produced. When statistical models produce an estimation of uncertainty (***confidence intervals)***, this should also be integrated inside metadata. After that, the new dataset can be disseminated.

Note that dissemination can be also the integration of the dataset inside the ESPON 2013 database, as a new version of the dataset or a replacement of a previous dataset.

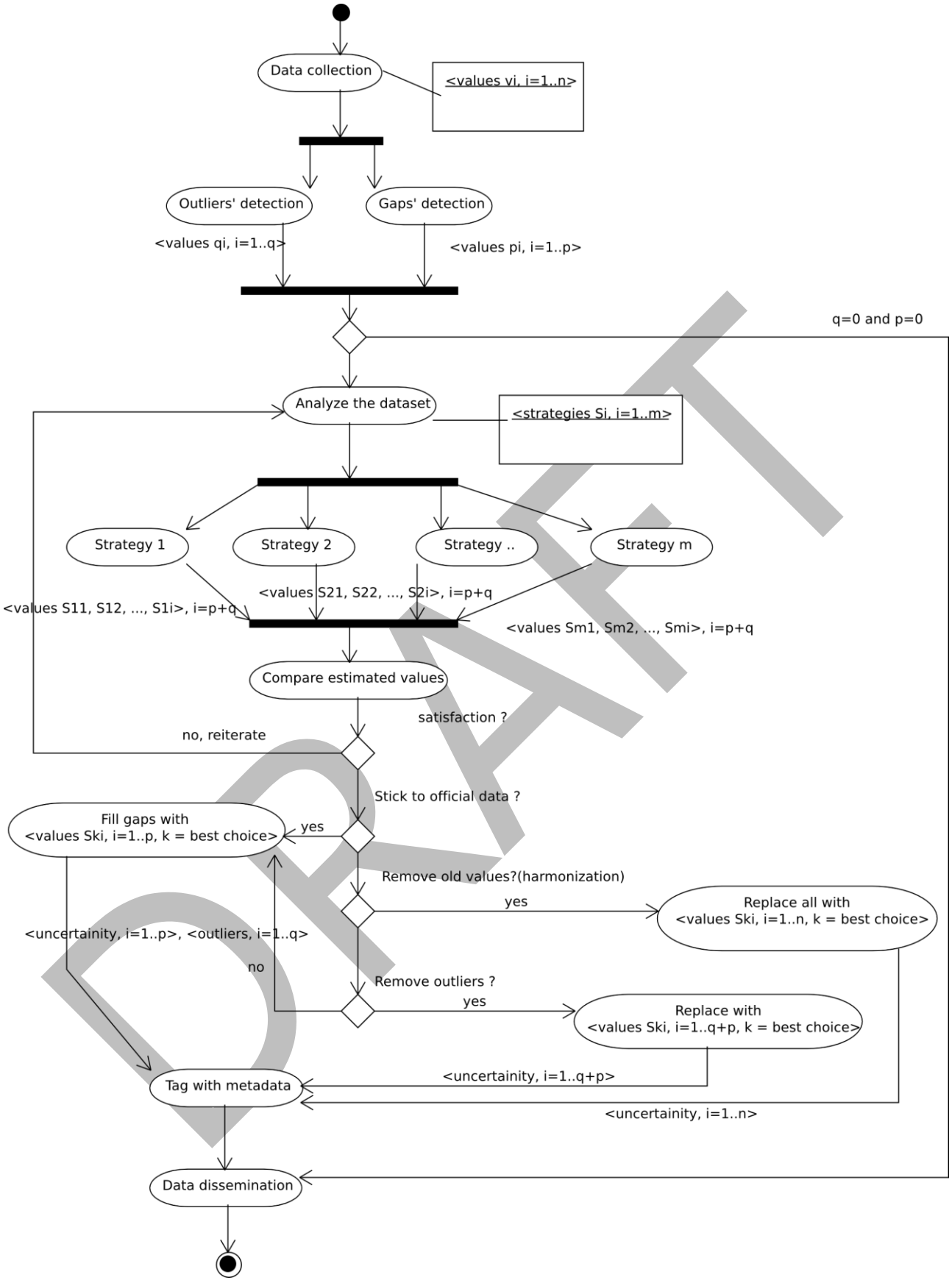


Figure 7: Activities diagram for a complete data workflow from collection to dissemination.

3.2 Tools

A good portion of these activities could be automated or at least semi-automatized, helping humans in detecting outliers, visualizing missing data, running various strategies, comparing their results, and making metadata. This leads us to search tools for exploring data and using statistical tools in the same time, in spatial, temporal and thematic dimension as well. Statistics can be univariate, bivariate or multivariate as well. Here, the advantage of having an ESPON 2013 database is manifest since the database can help to find ancillary data for regression or the control of statistical models. That's why one option could be that those tools connect (or are able) to read the database.

It happens that a first experiment for automating outliers checking was made during Christine Plumejeaud PhD Thesis in 2010, that had strong relationships with the problems the ESPON 2013 database was dealing with. A first state of the art in terms of Exploratory Spatial Data Analysis tools showed that clearly some tools could be plugged on the Espon 2013 database (Table1).

	Data visualization	**Statistical analysis**	**Database connection**	**Metadata visualization**
Sada [1997]	Yes	Yes	no	No
GeoDa [1998]	Yes	yes	No	No
CrimStat [2004]	yes	Yes	no	No
QuantumGis [2002]	Yes	Yes	yes	No
Grass [2010]	Yes	Yes	yes [R programing]	No

Table1. ESDA tools - a review.

It was also demonstrated that the R framework for statistical analysis would be suitable to run sophisticated statistical strategies. Moreover, all work already done in the project by Martin Charlton and its team was done with R (extensible and robust framework). However, concerning the usage of metadata, and the production of metadata, there was almost nothing (and there is still nothing to our best knowledge). Thus, a kind of ESDA tool, coupling spatio-temporal mapping capabilities with statistical capabilities was developed on top of the ESPON 2013 database, using Java and R. Despite these shortcuts, the tool, named QualESTIM, has shown that automating outlier' detection is feasible and useful (Figure 8). It has been acknowledged by various publications [Plumejeaud, 2012].

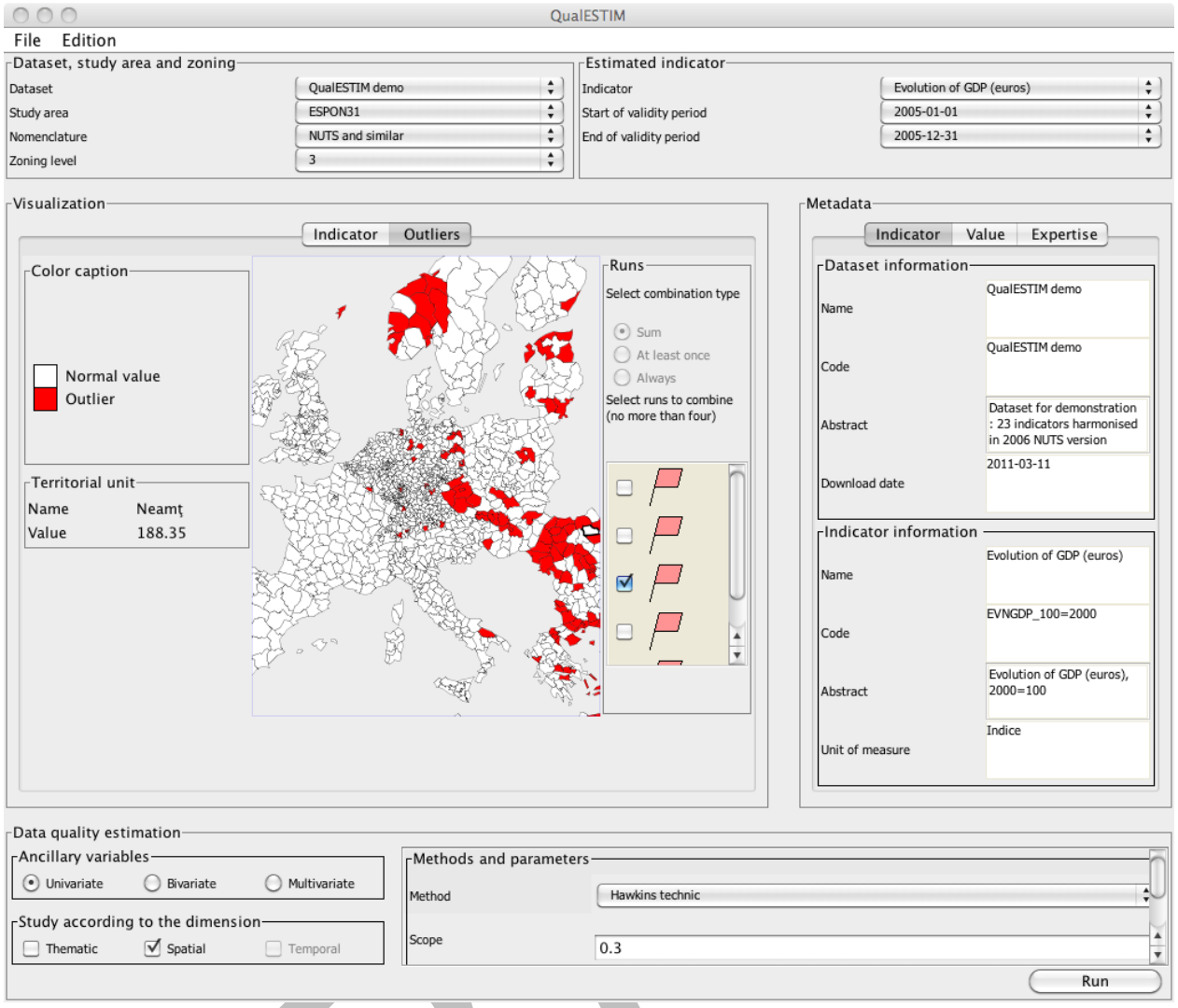


Figure 8. Overview of QualESTIM interface.

The architecture of QualESTIM was too much "hardcoded" within the ESPON 2013 database (a postgreSQL with postgis extension). This prevented any user to check datasets before they were imported inside the database. Thus, for the new framework, it seems highly desirable to be able to connect to ESPON 2013 database, but also to be able to read simple CSV or Excel files, (and corresponding SHAPE files ou MID/MID files in order to get the geometries of spatial units).

Furthermore, we can imagine this tool as a service offered on Web for various users, not only ESPON managers. A Web architecture offering a WPS processing framework for data processing could be very useful.

It is thus necessary to make a new review of tools, looking for statistical facilities, Web mapping facilities, capacities to read/write data sources of various formats: databases, CSV and spatial ones.

However, one of the major difficulties encountered by the use of statistical models and sophisticated statistical methods has to be underlined. It can be difficult to tune correctly the parameters of a model, such as the ones of the Geographically Weighted Regression for instance (cut off, bandwidth). Also the choice of the model requires a previous analysis because not all data are following a gaussian centered law (Figure 9).

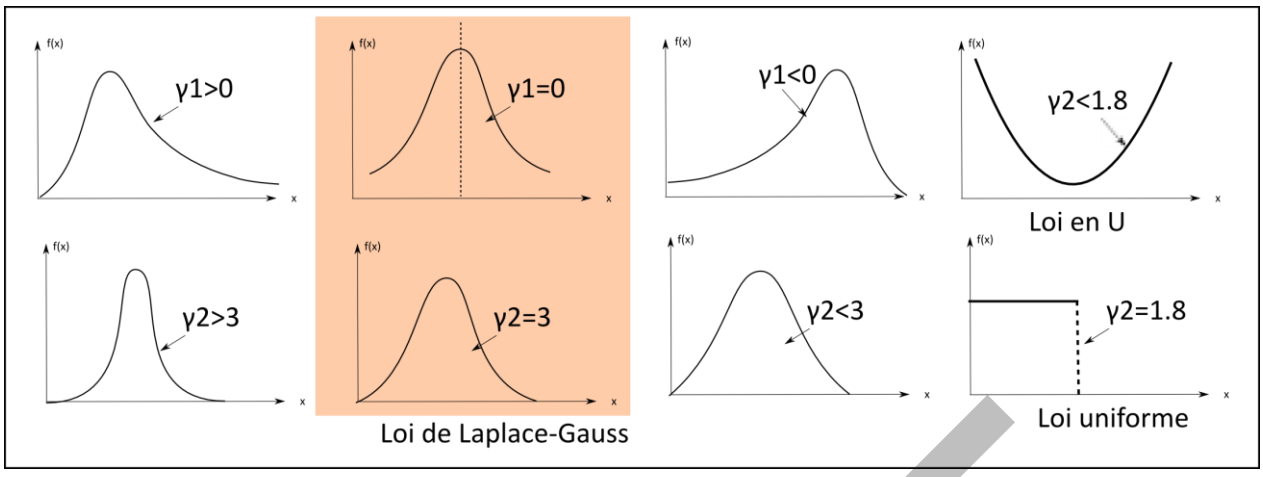


Figure 9. Various laws and distribution shapes leads to very different statistical models.

This is a well-known problem [Lantner, 1991] [Bastien, 2001] of usability of softwares. We may build a kind of knowledge database in order to help the user with the choice of methods and the choice of their parameters. Then we can imagine and dream about a fully automated tool for those tasks.

References

Ben Rebah M, Plumejeaud C, Ysebaert R, Peeters D, 2011, Modeling territorial changes and time-series database building process: empirical approaches and application, Technical report, ESPON Database project.

Christian B, Scapin DL, 2001, Évaluation des systèmes d'information et critères ergonomiques. In C. Kolski, editeur, Systèmes d'information et inter- actions homme-machine. Environnement évolués et évaluation de l'IHM. In- teraction homme-machine pour les SI, volume 2, pages 53–79. Hermès.

Charlton M, Harris P, Caimo A, 2012, Detecting and handling anomalous data in M4D, technical report, ESPON M4D project.

Charlton M, Caimo A, 2012, Time series analysis, technical report, ESPON M4D project.

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.

Grasland C, Ysebaert R, 2012, The core database strategy, a new paradigm for data collection at regional level, technical report, ESPON M4D project.

Lanter DP, Essinger R, 1991, User-Centered Graphical User Interface Design for GIS. Rapport technique 91-6, National Center for Geographic In- formation and Analysis.

Plumejeaud C., Exploration Spatio-Temporelle de l'Information statistique territoriale avec ses Métadonnées. Dans: Fonder les sciences du territoire, Ed. CIST; Karthala, 2012, à paraître.

Plumejeaud C., Villanova-Oliver M., 2012, QualESTIM: Interactive quality assessment of socio-economic data using outlier détection, 15th international conférence on Geographic Information Science (Agile'2012), 24-27 April, Avignon, (France)

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

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

Annexe

4.1 Draft example of automatic algorithm

(this part is to bedone in common by NCG and RIATE, if possible with expert advice of H.Mathian and C. Plumejeaud)

```
# Estimator_001.R
#
# C. Grasland, ESPON M4D, 2013
```

(A) IMPORT DATA

```
setwd("F:/claudegrasland/cg/projet/espondb2/estimator/test_bulgaria")
list.files()
act<-read.table("test_bulgaria_act.txt", sep="\t",dec=".",header=TRUE)
act
emp<-read.table("test_bulgaria_emp.txt", sep="\t",dec=".",header=TRUE)
emp
```

N.B. Datasets should be divided by homogeneous subsets in terms of original data producer. It is normally done by states (NUTS0) but they are exception in countries where different statistical offices exist (e.g. Scotland, Northern Ireland and Wales-England in UK)

(B) PREPARE DATA FORMAT

```
# (B.1) select a table
tab<-emp[,-3:-2]
head(tab)


# (B.2) define levels
lev<-nchar(as.character(tab$code))-2
lev
nblev<-max(lev)
nblev

# (B.3) Visualisation
mat<-t(as.matrix(tab[,-1]))
colnames(mat)<-tab$cod
tim<-ts(mat,start=1999, frequency=1)
plot(log(tim))
```

N.B. Insure that codes are harmonized to be use by the automatic procedures in next step.

(C) BUILD TREE OF HIERARCHICAL LEVELS

```
# (C.1) define table by level
tab0<-tab[lev==0,]
tab1<-tab[lev==1,]
tab2<-tab[lev==2,]

# (C.2) Compute freq tables

# PAS TRES BIEN ECRIT !!!

# Ratio between levels 1 and 0
```

```
sup1<-data.frame(tab1[,1])
names(sup1)<-"code"
sup1$code<-substr(sup1$code,1,2)
sup1<-merge(sup1,tab0,by="code")
freq10<-as.matrix(tab1[,-1])/as.matrix(sup1[,-1])
row.names(freq10)<-tab1$code
apply(freq10,FUN=sum,MARGIN=2)
agr10<-data.frame(rownames(freq10),freq10)
names(agr10)[1]<-"code"
agr10

# Ratio between levels 2 and 1
sup2<-data.frame(tab2[,1])
names(sup2)<-"code"
sup2$code<-substr(sup2$code,1,3)
sup2<-merge(sup2,tab1,by="code")
freq21<-as.matrix(tab2[,-1])/as.matrix(sup2[,-1])
row.names(freq21)<-tab2$code
apply(freq21,FUN=sum,MARGIN=2)
agr21<-data.frame(rownames(freq21),freq21)
names(agr21)[1]<-"code"
agr21

# (C.3) Synthetic table
syn<-rbind(tab0,agr10,agr21)
syn$code<-as.character(syn$code)
rownames(syn)<-syn$code
syn

# (C.4) Visualisation
mat<-t(as.matrix(syn[,-1]))
colnames(mat)<-syn$code
tim<-ts(mat,start=1999, frequency=1)
plot(tim)
```

N.B. Transform data into a tree of frequencies of upper node, with the exception of top level which is measured in the unit of count of initial data

(D) CHECK HIERARCHICAL CONSISTENCY

```
#(D.1) General check

test<-data.frame(syn[-1,-1])
sup<-substr(syn$code,1,nchar(syn$code)-1)[-1]

resul<-aggregate(test,FUN="sum",by=list(sup))
error<-round((as.matrix(resul[-1])-1),6)
row.names(error)<-resul[,1]
error

# At this step, we can decide to declare NA the nodes that are
# characterized by error superior to a given threshold

# (D.2) Correct minor mistakes ?

# Here, we can introduce a correction in order to eliminate
# minor errors. For each node, we force the sum of frrequency to be 1
```

N.B. We have to choose a level of error which is acceptable and related to "round" figures. In this case, we can automatically correct the frequency by a procedure. In other case, it is preferable to stop and send a report on problems in aggregation. The user is in this case oblige to decide what is false (parent or child ?). A good example is provided below.

code	name	level	une199	une200	une200	une200	une200	une200	une200	une200	une200	une200	une200	une201
BG	Bulgaria	NUTS0	126.5	122.6	140.7	609.0	449.2	400.0	334.4	305.7	240.2	199.7	238.0	348.0
BG3	Severna	NUTS1	329.9	362.0	398.8	353.2	270.3	235.6	186.5	184.4	159.9	131.9	143.0	199.3
BG31	Severoza	NUTS2	75.5	82.8	97.5	87.4	51.5	46.3	45.2	40.6	34.0	27.4	29.8	38.8
BG32	Severen	NUTS2	82.8	86.3	93.2	79.5	60.5	56.5	49.2	54.8	44.1	34.9	33.3	43.8
BG33	Severoizt	NUTS2	91.7	104.4	102.6	95.9	88.2	76.4	53.4	49.9	50.2	40.6	47.7	65.7
BG34	Yugoizto	NUTS2	79.9	88.5	105.5	90.4	70.2	56.4	38.6	39.1	31.6	28.9	32.2	51.0
BG4	Yugozap	NUTS1	204.1	204.8	265.2	239.2	178.9	164.4	147.9	121.4	80.3	67.9	95.0	148.7
BG41	Yugozap	NUTS2	101.2	104.6	136.1	121.4	108.5	92.3	75.4	67.2	42.2	32.0	44.9	72.1
BG42	Yuzhen t	NUTS2	102.9	100.2	129.1	117.8	70.4	72.1	72.5	54.1	38.1	35.8	50.1	76.6

(E) FILL MISSING VALUES OF PARENT NODES WITH COMPLETE CHILDS

```
# At this step, we fill the missing values for all superior nodes
# where inferior nodes values are available.
```

N.B. In the previous example of unemployment in Bulgaria, we can imagine that we have declared "missing" the values of BG in 1999-2002. If we consider that information on lower level is correct, we compute BG by sum of NUTS1 or NUTS2 units.

(F) FILL MISSING VALUES OF NODES WITH COMPLETE BROTHERS

```
# At this step we fill the missing values of units where all brothers
# are available
```

NB. Case where we have for example the values of BG31,BG32,BG33 and BG3 but not the value of BG34. It is normally possible to obtain the missing value by equation BG34= BG3-BG31-BG32-BG33

(G) ESTIMATION OF TIME SERIE OF COUNT DATA AT TOP LEVEL

```
top<-t(syn[1,-1])
ori<-ts(top,start=1999,frequency=1)

# Estimation function (basic example)
# part to be optimized ...

library(zoo)
ori<-zoo(ori)
est<-ori
est<-na.locf(est,na.rm=FALSE,fromLast=FALSE)
est<-na.locf(est,na.rm=FALSE,fromLast=TRUE)
est<-na.approx(est,na.rm=FALSE)


# Visualization
gra<-merge(est,ori)
gra
plot(gra, plot.type="single",col=c("red","blue"), lwd=2)


# Exportation of result
top<-t(est)
```

N.B. Estimations methods are not necessarily the same than in following step. The outliers can be related here to external events outside the country. This estimation should be very carefully done as long as all count value will depend from this single vector of count.

(H) ESTIMATION OF TIME SERIES OF FREQUENCY FOR EACH NODE

```
agr<-t(agr10[,-1])
ori<-ts(agr,start=1999,frequency=1)

# Estimation function (basic example)
# part to be optimized ...

library(zoo)
ori<-zoo(ori)
est<-ori
est<-na.locf(est,na.rm=FALSE,fromLast=FALSE)
est<-na.locf(est,na.rm=FALSE,fromLast=TRUE)
est<-na.approx(est,na.rm=FALSE)

# Correction for Sum=1
correct<-as.matrix(est)
tot<-apply(correct,FUN=sum,MARGIN=1)
correct<-correct/tot

# exportation of resilt
new_agr10<-t(correct)


# (G.3) Estimation of frequency vectors of nodes 2/1

# Placer une boucle pour parcourir tous les noeuds


# (G.3.1) Node BG3
agr<-t(agr21[substr(agr21$code,1,3)=="BG3",-1])
agr

ori<-ts(agr,start=1999,frequency=1)

# Estimation function (basic example)
# part to be optimized ...

library(zoo)
ori<-zoo(ori)
est<-ori
est<-na.locf(est,na.rm=FALSE,fromLast=FALSE)
est<-na.locf(est,na.rm=FALSE,fromLast=TRUE)
est<-na.approx(est,na.rm=FALSE)

# Correction for Sum=1
correct<-as.matrix(est)
tot<-apply(correct,FUN=sum,MARGIN=1)
correct<-correct/tot

# Resulting table
new_agr21_BG3<-t(correct)
new_agr21_BG3



# (G.3.2) Node BG4
agr<-t(agr21[substr(agr21$code,1,3)=="BG4",-1])
agr

ori<-ts(agr,start=1999,frequency=1)

# Estimation function (basic example)
# part to be optimized ...

library(zoo)
ori<-zoo(ori)
est<-ori
est<-na.locf(est,na.rm=FALSE,fromLast=FALSE)
est<-na.locf(est,na.rm=FALSE,fromLast=TRUE)
est<-na.approx(est,na.rm=FALSE)

# Correction for Sum=1
correct<-as.matrix(est)
```

```
tot<-apply(correct,FUN=sum,MARGIN=1)
correct<-correct/tot

# Resulting table
new_agr21_BG4<-t(correct)
new_agr21_BG4


# (G.3.3) Fusion of nodes of level 2/1

new_agr21<-rbind(new_agr21_BG3,new_agr21_BG4)
new_agr21
```

N.B. The methods used here for the estimation are the most stupid ones. It is certainly possible to do much better. But all methods has to keep in mind the constraint of conservation of the sum of frequency equal to 1.