International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:3, No:1, 2009

Model-Based Small Area Estimation with application to unemployment estimates

Hichem Omrani, Philippe Gerber, and Patrick Bousch

Abstract—**The problem of Small Area Estimation (SAE) is complex because of various information sources and insufficient data. In this paper, an approach for SAE is presented for decision-making at national, regional and local level. We propose an Empirical Best Linear Unbiased Predictor (EBLUP) as an estimator in order to combine several information sources to evaluate various indicators. First, we present the urban audit project and its environmental, social and economic indicators. Secondly, we propose an approach for decision making in order to estimate indicators. An application is used to validate the theoretical proposal. Finally, a decision support system is presented based on open-source environment.**

Index Terms—**small area estimation, statistical method, sampling, empirical best linear unbiased predictor (EBLUP), decision-making.**

I. Introduction

THIS study is carried out under the framework of Urban Audit project that applies to small area estimation (SAE) to evaluate several indicators. There are different methods (e.g. direct, Greg, composite, synthetic, Empirical Bayesian (EB), Hierarchical Bayes (BH), Empirical Best Linear Unbiased Prediction (EBLUP) and Spatial EBLUP (S-EBLUP) estimators) available that provide estimation to small area levels [3].

This study use census files obtained from the Statistical Offices STATEC (Central service of statistics and economic studies of Luxembourg) and INSTEAD (International Network for Studies in Technology, Environment, Alternatives and Development). A stratified random design is used to draw samples from the population. Then, Small area estimators of unemployment estimates are calculated using two commonly used methods (GREG and EBLUP) for all small areas and compared by using different performance measures (relative to Bias (B), Average Relative Bias (ARB), Mean-Squared Error (MSE), Average Empirical Mean Squared Error (AEMSE) and Deviance information Criteria). The obtained results confirm the accuracy of EBLUP estimator which takes into account the spatial similarity according to a Simultaneously Autoregressive (SAR) specification and spatial structure of the data [7]. Thus, the EBLUP is applied under the framework of Urban Audit project with application to unemployment estimates. This study allows to combine information from different sources: geographical, census and administrative data. In addition to, this paper proposes a decision support system based on EBLUP for the estimation of several indicators of Urban Audit project. A web-based tool called DSS-SAE (Decision Support System for Small Area Estimation) has been developed that treats complex data collected from various information sources. The tool has several functionalities starting from data integration (import of data), small area estimation and finishes by graphical and geographic display of results.

The paper is structured as follow. First, we present the frame of the work. Then we describe briefly the recent advances in SAE Modeling and especially the EBLUP estimator. Next data and variables used for unemployment estimates are presented. The results are presented and schematically illustrated with figures and maps. Finally, the estimation assessment is done by comparing the results of EBLUP estimator with GREG estimator of SAE according to several performance measures.

H. Omrani is with the International Network for Studies in Technology, Environment, Alternatives and Development (INSTEAD) in GEODE (Geography and Development) department, Luxembourg (corresponding author, phone: +352.58.58.55.657; (e-mail:hichem.omrani@ceps.lu, web site: see http://www.hds.utc.fr/ omranihi)

P. Gerber is with INSTEAD office, Luxembourg. He is with the GEODE department (e-mail: philippe.gerber@ceps.lu).

P. Bousch is the director of GEODE department in INSTEAD office, Luxembourg (e-mail: patrick.bousch@ceps.lu).

The research was supported by the European Community, CEPS-INSTEAD and National Research Fund of Luxembourg

II. Framework: urban audit project

The Urban Audit is a response to growing demand for an assessment of the quality of life in European towns/cities, where a significant proportion of European Union citizens live. The Urban Audit is a joint effort by the Directorate-General for Regional Policy (DG REGIO) and Eurostat (statistical office of the European communities) to provide reliable and comparative information on selected urban areas in Member States (MS) of the European Union (EU) and the Candidate Countries. The joint effort enabled the collection of information concerning over 500 variables at three points in time (1981, 1991 and 1996) in 21 topical domains and at three spatial levels for 58 towns/cities (in their administrative boundaries) excluding Paris and London, 7 conurbations and 20 wider territorial units comprised by several regions in all 15 Member States. Information on a small number of indicators was collected for 2500 sub-city areas in 54 of the towns/cities to provide information on disparities within cites. Nearly 100 indicators were calculated and published, both in a printed format as well as in the Internet.

The Urban Audit is a useful tool for decision-making at European, national, regional and local level. Table 1 presents the 9 statistical fields and 25 domains of the Urban Audit II.

The Urban Audit II has been based on 258 participating ”cities” (see Fig. 1), of which 189 were from the 15 EU Member States. It aims to provide information at three spatial levels:

- the Core City (administrative definition), as the basic level (Label "A");
- the Larger Urban Zone (Label "LUZ"), which is an approximation of the functional urban zone centered around the town/city;
- the Sub-City District (Label "SCD"), which is a subdivision of the city according to strict criteria (5000-40000 inhabitants in each sub-town/city district).

TABLE I
STRUCTURE OF THE URBAN AUDIT II STATISTICS

1. Demography	
1.1	Population
1.2	Nationality
1.3	Household structure
2. Social aspects	
2.1	Housing
2.2	Health
2.3	Crime
3. Economic aspects	
3.1	Labor market
3.2	Economic activity
3.3	Income disparities and poverty
4. Civic involvement	
4.2	Local administration
5. Training and education	
5.1	Education and training provision
5.2	Educational qualifications
6. Environment	
6.1	Climate / geography
6.2	Air quality and noise
6.3	Water
6.4	Waste management
6.5	Land use
6.6	Energy use
7. Travel & Transport	
7.1	Travel patterns
8. Information society	
8.1	Users & infrastructure
8.2	Local e-Government
8.3	Information and Communication Technology (ICT) sector
9. Culture and recreation	
9.1	Culture and recreation
9.2	Tourism

In the Urban Audit II project, neraly 100 statistical indicators are defined and their categories are described in Table 1. The indicators have been produced for the Core Cities, Cub-City Districts and Larger Urban Zones. It is important that indicators produced for these spatial units are based on reliable data sources and they are obtained by using adequate statistical techniques. In most cases, data obtained from Censuses, different administrative and statistical registers and national and local databases are used in a given country or a spatial unit.

A benefit of the proposed approach in this paper is that sufficient data are available to calculate indicators even for small spatial units and indicators can be obtained without any sampling error. In some cases, data are only available in a sample survey, or data are not necessarily readily available for some indicators according to the required definitions and covering the given spatial unit. For these situations, statistical estimation methods need to be applied to overcome the problems of gaps in the database. A variety of estimation methods are discussed in next section.

Fig. 1. Participating Cities in the Urban Audit II (256 cities)

III. RECENT ADVANCES IN SAE MODELING

Small area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors [9]. It is now widely recognized that direct survey estimates for small areas are likely to yield unacceptably large standard errors due to the smallness of sample sizes in the areas. This makes it necessary to "borrow strength" from related areas to find more accurate estimates for a given area or, simultaneously, for several areas [4]. This has led to the development of alternative methods such as synthetic, sample size dependent, generalized regression (GREG) [5], empirical best linear unbiased prediction (EBLUP) [1], [6] empirical Bayes [2] and hierarchical Bayes estimation [8]. The present article is an appraisal of two of these methods. The performance of these methods is also evaluated using real data. EBLUP estimator, for most purposes, seems to have a distinct advantage over other methods. In this research, we focus on the study of EBLUP estimator with application to unemployment estimates.

IV. PROPOSED APPROACH

The proposed process of simulation for SAE is schematized according to Fig 2. It is composed of three steps: sampling, estimation models and measures quality.

Auxiliary information and statistical models have a crucial role in small area estimation. This paper studies the accuracy of two conventional small area estimators. A set of samples are drawn up starting from the database of the population with different techniques of sampling and the most commonly used SAE estimators GREG and EBLUP, utilize auxiliary information, which were applied for each sample, to predict unemployed by commune in Luxembourg.

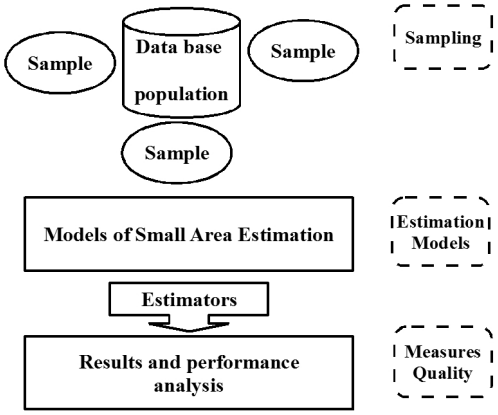

Fig. 2. Process of simulation for SAE

1) Notation:

- U=1,2,. . .,N: population (fixed, finite)
- U_1,. . .,U_d,. . .,U_D: domains of interest (non-overlapping)
- $Y(d) = \sum_{U(d)} y_k$, d=1,. . .,D; target parameters
- $X = (X_{1k},\ldots,X_{pk})$: auxiliary variable vector. We assume that the vector value X_k and domain membership are known for every population unit $k \in U$.
- s: sample, ω: weights
- Horvitz-Thompson estimator:

$$\hat{Y}_d = \sum_{s(d)} \omega_i \times y_i \tag{1}$$

It is simple but inefficient, for that efficient estimators are presented below.

2) Efficient estimators: The most commonly used SAE estimators GREG (G) and EBLUP (E) utilize auxiliary information x, which is used to predict study variable (unemployed). The two estimators are computed as mentioned in (2, 4):

- Model-assisted Generalized regression estimator GREG: this model combines direct information from the sample with aggregated data in order to improve the quality of estimates.

$$\hat{Y}_G(d) = \sum_{j\in s} \frac{1}{\pi_j} \times y_j + \sum_k \beta_k \times (\sum_{p=1}^{N} x_p - \sum_{j\in s} \frac{1}{\pi_j} \times x_j) \tag{2}$$

For SRS (Simple Random Sampling) without remplacement: $\pi_j = \frac{n}{N}$ (inclusion probability) and β_k are estimated using weighted linear regression. This estimator is based on linear predictor (computed using weighted regression of the sample data) plus some weighted correction terms based on the sampled units and the difference between the observed and the predicted value of each individual.

- Model-dependent EBLUP: It is a linear combination of the observed value and it can be expressed as follow:

$$y = X \times \beta + Z \times u \tag{3}$$

where X is a set of covariates, β their associated coefficients, u area rondom effects and Z a matrix that models the structure of u. The random effects u are distributed with zero mean and variance σ_u^2. The estimates of β are obtained using standard Generalised Least Square techniques, whilst the estimates of u are computed using their Empirical Best Linear Unbiased Predictor (EBLUP). The EBLUP estimator of y is then defined as:

$$\hat{y} = X \times \hat{\beta} + Z \times \hat{u} \tag{4}$$

3) Performance and quality measures: properties of GREG and EBLUP estimators of totals for domains: The evaluation of the estimators is based on the quality of the following measures: Bias (5, 6), precision (7), accuracy (8, 9) and design effect (10).

- Bias:
  - The Bias is measured by the following equation:

$$Bias(\hat{Y}_d) = E(\hat{Y}_d) - Y_d \tag{5}$$

  - The absolute relative bias: ARB (%) is measured as follow:

$$ARB(\hat{Y}_d) = \frac{|E(\hat{Y}_d) - Y_d|}{Y_d} \tag{6}$$

- Precision:
  - It is measured by the variance:

$$V(\hat{Y}_d) = E(\hat{Y}_d) - E(Y_d))^2 \tag{7}$$

- Accuracy:
  - It is measured by mean squared error (MSE):

$$MSE(\hat{Y}_d) = E(\hat{Y}_d) - Y_d)^2; with\ d = 1,\ldots,D \tag{8}$$

This measure describes the difference between the real average and the value of the estimator for each studied area.
  - Another similar measure for the accuracy is defined. It is the relative root mean squared error (RRMSE (%):

$$RRMSE(\hat{Y}_d) = \frac{\sqrt{E(\hat{Y}_d) - Y_d)^2}}{Y_d} \tag{9}$$

- Design effect:
  - The design effect is used to compare the variability of the same estimator for a particular sampling scheme $p(s)$. Usually; SRS is taken as the refrence. This measure is computed as follow:

$$DEFF_{p(s)} = \frac{V_p(s)[\hat{\bar{Y}}]}{V[\hat{\bar{Y}}]} \tag{10}$$

This measure is used in the case study in section V.

V. Application

A. Case Study

The GREG and EBLUP estimators were applied to estimate the unemployment by commune at Luxembourg where various data are collected from the Statistical Office STATEC. In Luxembourg, we find 116 communes (as shown in Fig. 3, in the right). These communes are grouped in 6 areas (North, East,

South-eastern, Southern, Western, Center, southern Center, see Fig. 3 in the left). The communes are treated here like units. To estimate the average of unemployement rates by sex and area, a sample is drawn from the set of communes.

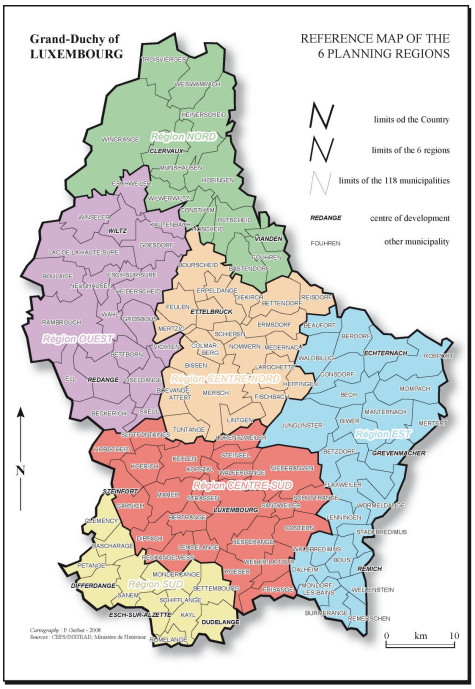

Fig. 3. Luxembourg and its 116 communes and 6 areas

The Table 2 summarizes the data and variables used for the case study. This application treats the unemployed by sex in the 116 communes of Luxembourg in 2007 (see Fig. 3). The data used, from the census file, about the number of

TABLE II
VARIABLES OF APPLICATION

Code	Identifier
$Unempl_m$	number of unemployed males
$Unempl_f$	number of unemployed females
Comm	code and name of commune
Reg	geographic indicator of area (6 areas) or region

unemployed males and females are shown in Figs. 4 and 5.

We used also auxiliary data of Luxembourg active population as shown in the map in Fig. 6.

B. Sampling design and GREG estimator

To estimate the regional mean (number of unemployed by area) we sample from communes. We use here 3 sampling techniques as described below and the results are shown in Fig. 7.

1) Simple Random Sampling Without Replacement (SRSWOR):
   - Sample is made of 18 communes ($\cong 15\%$)
   - Equal probabilities for all the communes

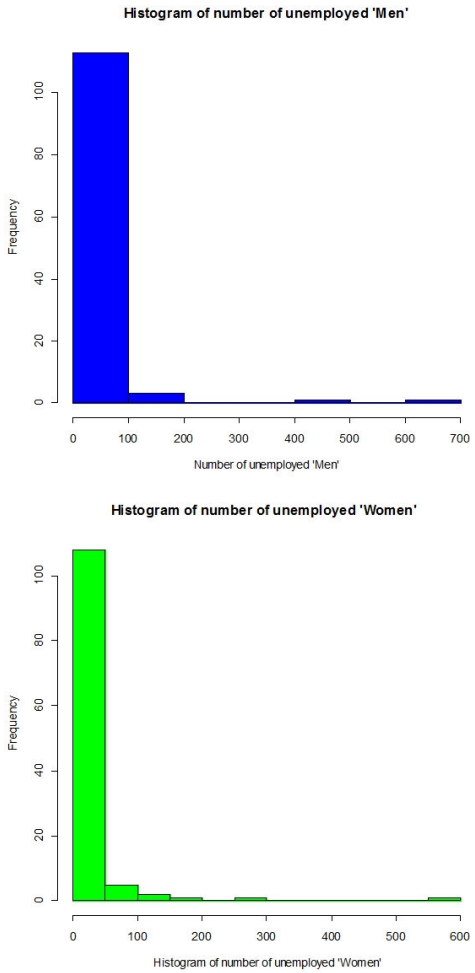

Fig. 4. Communes according to number of unemployed males (in the top); Communes according to number of unemployed females (in the bottom)

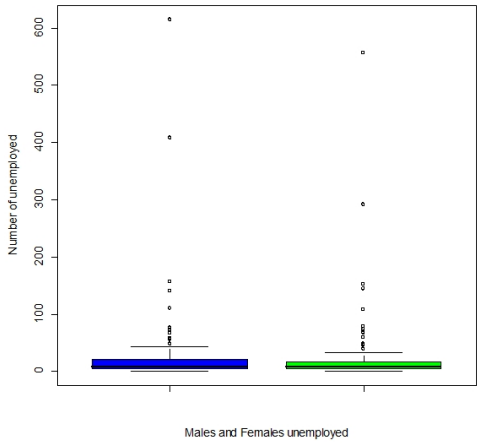

Fig. 5. Number of unemployed males (in the left); Number of unemployed females (in the right)

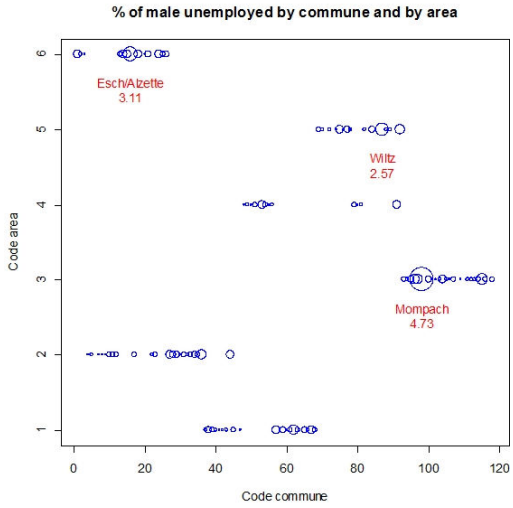

Fig. 6. Males unemployment rate by commune and by area

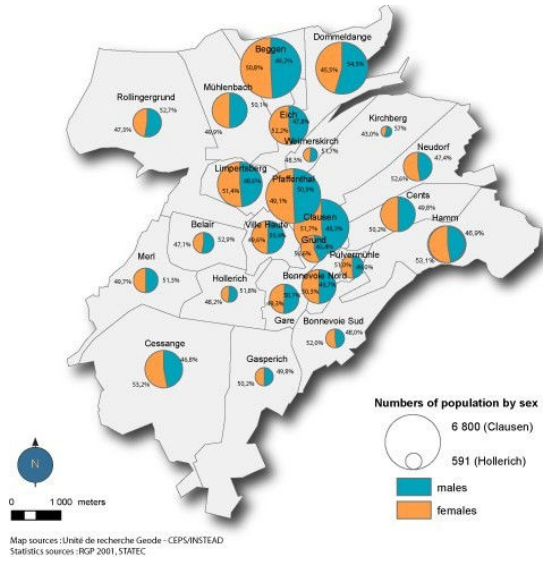

Fig. 7. Repartition of the population of Luxembourg city by sex, in 2007

2) Stratified SRS Without Replacement (CLSWOR):
   - Sample is made of 18 communes ($\cong 15\%$)
   - 3 communes per area (region)
   - Equal probabilities for all communes within strata (i.e. area)
3) Stratified SRS Without replacement (Two-stage Sampling) (CLSWOR2):
   - Sample is made of 18 communes ($\cong 15\%$)
   - 6 communes sampled by area
   - Some areas do not contribute to the survey sample

From the variances by area of the GREG estimator and according to three techniques of sampling (SRSWOR, CLSWOR, CLSWOR2), we deduce that the GREG estimator with equal probabilities for all communes within areas

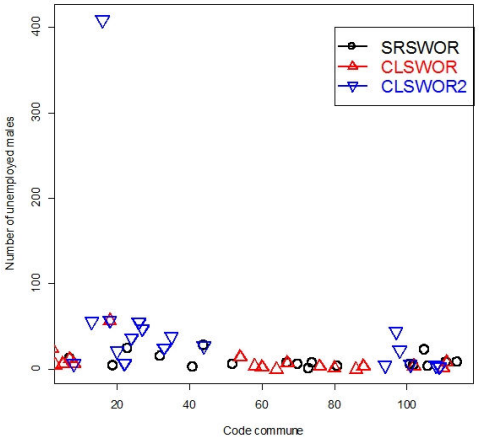

Fig. 8. Number of unemployed males by commune according to three techniques of sampling (SRSWOR, CLSWOR, CLSWOR2)

(CLSWOR) is of good quality (see Fig. 8).

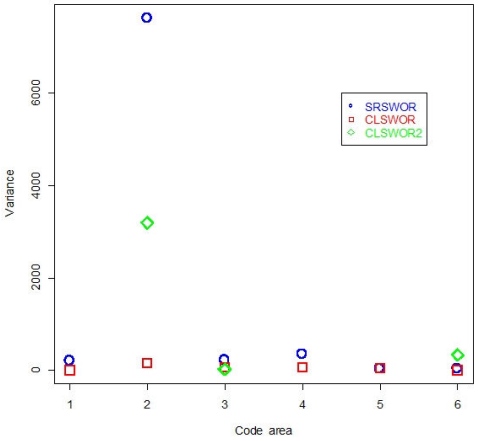

Fig. 9. Variance of GREG estimator according to three sampling techniques

We present in Fig. 9, the estimator GREG (CLSWOR) by area, while comparing the various estimates by area with the real average resulting from the population.

In order to improve the estimates taking into account axiliary data (covariates), we present in the next sub-section the results from EBLUP estimator with the same case study.

C. Results from EBLUP estimator

The mixed-effects model is used to improve the estimate. In this model, the random effects of measurements are due to the factors which were not measured. The space aspects can be measured using the random effects. We draw up the estimates of the unemployed on the communes by using EBLUP estimator (see Table 3 and Fig. 10)

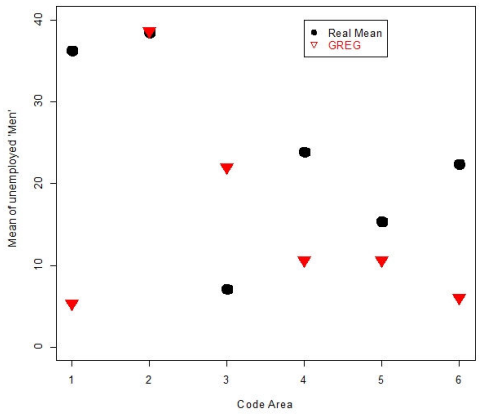

Fig. 10. Number of unemployed males according to (real mean, GREG (CLSWOR)) by area

TABLE III
REAL MEAN OF UNEMPLOYED BY AREA COMPARED TO GREG (CLSWOR) AND EBLUP ESTIMATORS

Area	1	2	3	4	5	6
Real mean	36.23	38.5	17.11	23.85	15.35	22.38
GREG	5.33	38.66	22	10.66	10.66	6
EBLUP	6.4	17.36	23.77	9.68	11.34	4.38
MSE-GREG	1.80	150.01	56.61	54.03	42.31	0.92
MSE-EBLUP	6.86	16.25	38.76	2.84	4.58	4.79

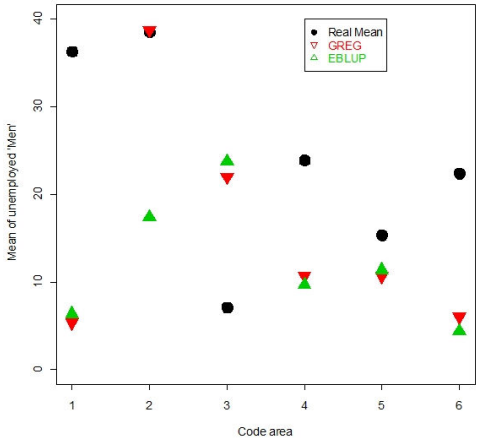

Fig. 11. Estimators (GREG (CLSWOR), EBLUP) for the 6 areas

D. Performance and quality measures

First, we mention that GREG estimator with CLSWOR sampling technique has the minimal variance (Table 4). Indeed, in this sub-section, we present the performance of GREG (CLSWOR) and EBLUP estimators.

TABLE IV
MSE FOR THE GREG ESTIMATOR

Estimator	MSE
GREG (SRCWOR)	1333.85
GREG (CLSWOR)	**272.12**
GREG (CLSWOR2)	377.04

In Fig. 11, we compare the performance of GREG (CLSWOR) and EBLUP estimators. From Table 11, we underline that EBLUP estimator is more effective than GREG (CLSWOR) estimator. The former has a distinct advantage over GREG estimator and its performance is justified by MSE, ERA and DEFF measures as shown in Table 5.

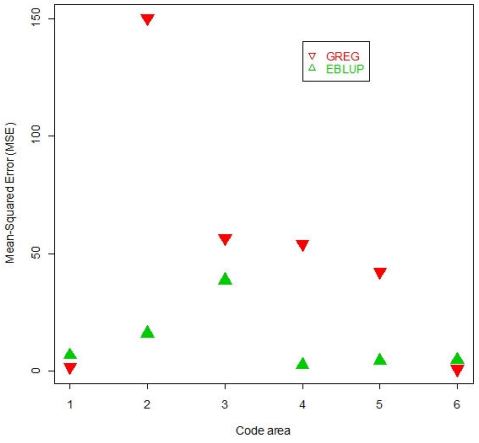

Fig. 12. MSE related to GREG (EASS SR) and EBLUP estimators for the 6 areas

TABLE V
QUALITY MEASURE RELATED TO GREG AND EBLUP ESTIMATORS

Measure quality	GREG	EBLUP
MSE	50.94	12.34
DEFF	100%	412.64%.

VI. DECISION SUPPORT SYSTEM FOR SAE: DSS-SAE

The theoretical propositions used for indicators estimation were implemented using a software platform in order to propose a complete application for decision making to manage and evaluate complex data related with urban audit project. The software developed is called DSS-SAE (Decision Support System for Small Area Estimation). It performs evaluations of

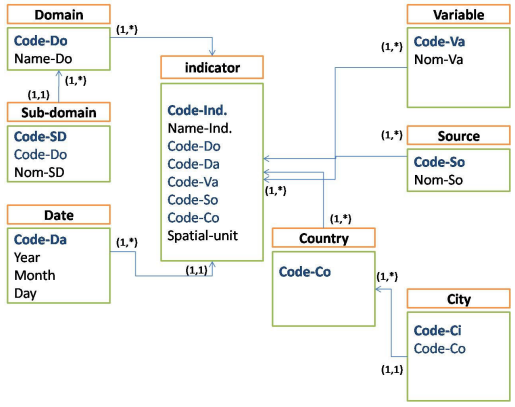

Fig. 13. Conception: multidimensional representation: OLAP

indicators at various levels: namely social, economic, environmental etc., see Table 1 and Fig. 12 with OLAP (Online Analytical Processing) representation.

The developed tool is illustrated in Fig. 13. It is based on a client server architecture which allows the estimation of diverse indicators (as shown in Table 1) of urban audit projects. This choice is justified by the fact that we involve multiple organisations, users and information sources which must interact for the purposes of data collection, management and estimation of small area for spatial unit.

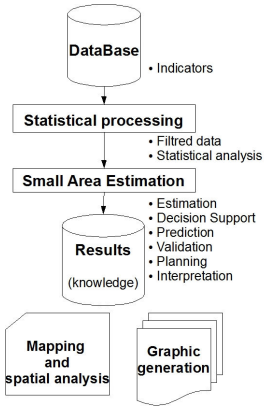

Fig. 14. Architecture of DSS-SAE

The developed platform uses the quadruple: Linux, Apache, MySQL and PHP (LAMP). It is web-oriented platform that collects ”on-line” data from various information sources (census, surveys, administration register, etc.). It includes varying interfaces described as follows:

- Web (HTML, Java Script, etc.);
- Linkage with statistical software R and the GIS Mapserver which are both ”open sources”;
- Client-server.

The tool DSS-SAE is composed of several modules described as follows:

1) User module: The management of users in such an application is fundamental. It allows knowing who uses the tool and controls the indicators of certain users on the application by blocking certain functionalities. We have therefore defined certain class of users (Administrator, Project Manager and Invited User). The users belong to a particular class and have access only to pre-defined functions.
2) Indicators module and statistical modelling: First, this module manages the indicators of urban Audit. Then, in DSS-SAE; the user can select and test various kind of approaches of SAE such as: GREG and EBLUP estimators, with optional pruning techniques. The user can test several parameters and validate the obtained results. Some validation techniques or quality measures are also available to test the quality of induced estimators.
3) Data sources module: This module includes all what is Addition/Deletion/Importing/Modification of information sources. We are convinced that it will be practical to export the results of different sources, whether it is for verifying the values contained in the database or for transferring data between different services (for e.g. between INSTEAD and STATEC). The storage of these data is therefore storage in a database with tables attributed for different type of data sources.
4) Graphics generation module: To do the graphics, we have used the graphical library ” JpGraph ” for PHP, entirely object programmed. It is one of the leader solutions in this domain (for PHP5). It displays all kind of graphics for e.g. histograms, point curves, Gantt charts, Pie charts, Radars, etc. This library recreates with PHP what can be done with Excel. The advantage lies in the fact that we use as source our database which allows the usage of a common source for various functionalities. With our tool we can for example import easily a file on unemployment and represent graphically the number of unemployed by sex over a selected period. A concrete example on this is presented in Figs 4 and 5. generated from the tool DSS-SAE.
5) Mapping/cartographic module: The Geographical Information System software MapServer is used for the display of maps. Moreover, with MapScript, which is a programming interface for PHP (Application Programming Interface or API), we have created dynamic maps (Figs. 6).

Finally, we conclude that the developed web based decision support tool for SAE ”DSS-SAE” of urban audit project, with its various functionalities, is helpful for decision making. The tool encompasses a module for visual representation of data allowing the user to estimate and to follow the evolution of spatiotemporal indicators (like unemployment, life quality, etc.) covering a given spatial unit.

VII. Conclusion

The prospect for this work consists in estimating indicators in various spatial units (city, larger urban zone and sub-city) by using small area estimation techniques. The major advantage of the approach proposed in this paper is its capability to combine, in efficient way, several information from census,

register, even in the case of incomplete data. It represents a reliable approach for small area estimation in order to evaluate environmental, economic and social indicators in several spatial units.

Finally, it enriches the development of a data-processing tool allowing the visualization of results. The tool is designed to be ergonomic, convivial, easy to use and to produce help with the estimation of the indicators by producing graphs and space-time dynamic map. Future work, will involve the development of novel techniques of SAE for handling missing data, fuzzy aspects and reliability of information sources.

ACKNOWLEDGMENT

This work is funded by the CEPS-INSTEAD (International Network for Studies in Technology, Environment, Alternatives and Development) of Luxembourg. It is under the framework of Urban Audit II project which is an European project funded by the European Commission. The research was financed by the National Research Fund of Luxembourg.

REFERENCES

[1] J. N. K. Rao. *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.
[2] M. Ghosh. *On some Bayesian solutions of the Neyman-Scott problem*. In: S.S. Gupta and J. Berger, Eds., Statist Decis. Theory and Related Topics, V. Springer, New York, 267-276, 1994.
[3] M. Ghosh, J.N.K. Rao. *Small area estimation: an appraisal*. Statist. Sci. 9, 5576, 1994.
[4] J. Jiang, P. Lahiri, *Mixed model prediction and small area estimation*, Test 15 (1) 196, 2006.
[5] C.E. Srndal, B. Swensson, J. Wretman. *Model Assisted Survey Sampling*, Springer, NewYork, 1992.
[6] G.K. Robinson. *That BLUP is a good thing: the estimation of random effects (with discussion)*. Statist, Sci. 6, 1551, 1991.
[7] R.M. Lark, B.R. Cullis, S.J. Welham. *On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML*. European Journal of Soil Science 57, 787799, 2006.
[8] L. Emily, *Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models*. Computational Statistics and Data Analysis, doi:10.1016/j.csda.2008.07.033, 2008.
[9] A. Sedeo-Noda. *Preemptive benchmarking problem: An approach for official statistics in small areas*, European Journal of Operational Research, doi:10.1016/j.ejor.2008.02, 2008.

Hichem Omrani, Doctor and Engineer in Information Technology from the Technology University of Compiegne-France. From July 2008, he is responsible of research at GEODE (Geography, development and spatial analysis) department of the CEPS-INSTEAD office. He is particularly interested in artificial intelligence techniques (treatment of information, data fusion). The issues in his studies consist of applying fuzzy logic theory, multicriteria analysis and belief theory for traffic, environmental, social and economic evaluation of impacts related to the urban mobility. He has published research papers at national and international journals and conference proceedings. (Phone: +352.58.58.55.657, e-mail: hichem.omrani@ceps.lu, web site: see http://www.hds.utc.fr/ omranihi).

Philippe Gerber, Doctor in geography and he is a responsible of research with GEODE department at INSTEAD office, Luxembourg (telephone: +352.58.58.55.601, e-mail: philippe.gerber@ceps.lu).

Patrick Bousch, Head of GEODE department since 2000. He has his master degree in 1989 in geography from the Louis Pasteur University at Strasbourg. He is president of superior Council of Town and country planning, since 2002. He teaches at Luxembourg University since 2004. (patrick.bousch@ceps.lu).