Annex II

Methodology for production of Small Area Estimates at Local Administrative Units level

Introduction

Small area estimation methods are used to overcome the problem of small samples sizes to produce small area estimates that improve upon the quality of direct survey estimates obtained from the sample in each small area. Sample surveys are the only source for the part of statistical data required in the frames of the Urban Audit. Generally, the production of small area data is limited by two serious conditions. The first is connected to the expenditure on data production. The statistical surveys at lowest levels are most expensive due to the necessity to provide accurate and reliable information. The second is connected to the necessity to ensure the statistical data confidentiality. Tendency is observed all over the world to use more and more sample surveys instead exhaustive ones. The sample surveys are designed to be representative up to NUTS 3 level. It was considered until recently that representative results for smaller units cannot be produced based on such surveys. Usually a two-stage sample is used in the NSI, stratified by administrative-territorial regions (NUTS 3) and residence (urban, rural) and on this basis is created 56 strata. According to some researchers one of the possibilities to ensure reliability applying the sample data collection methods in surveys is to increase the sample size or strengthen the samples in separate regions, which increases the statistical product value but does not solve the problem of production of data for statistical units that are not included into the sample. The stratification by administrative-territorial division and by the settlements status, which is to a great extent subjective, does not take into account the heterogeneity of the territorial units. There is no survey, mentioned in the available literature, taking into consideration the heterogeneity of the settlements.

There is no enough experience and empirical knowledge in Bulgarian practice on the methods for small area estimation. In 2003, within the framework of the World Bank project “Monitoring, estimation and elaboration of poverty politics”, the analysis of poverty at municipal level is done. It is based on data from the Multitopic Households Survey and it is conducted by team of experts from the National Statistical Institute, Ministry of Labour and

Social Policy and the Bulgarian Academy of Sciences. This is the first attempt in Bulgaria for SAE at such territorial level. The method for small area estimation used by the authors is known as “Mapping of poverty”. In essence a regression model is developed based on the sample survey, which is “placed” on the 2001 population census data. Thus are “assigned values” for the target variables to each census unit (unit level model). Auxiliary census variables are included in the regression model, named by the authors “potential factors”. Based on the poverty estimates produced for the existing during that time 262 municipalities, the cluster analysis is applied for their typologisation. In essence, this survey uses the clustering only for summarizing and presenting of small area estimation results and not to increase the accuracy and reliability of the estimates themselves. The authors make a conclusion that “regarding to the national peculiarities it turned out that by mapping could be received reliable poverty estimates at district and municipal level, but not at settlement level” ([1], page 93). Solution of this problem is offered below.

At the current project stage, for 33 variables on the employment, economic activity of the population, the education and household’s income and living conditions there is no another source of information different from the sample surveys (excluding the census year). This involves the selection of an appropriate mathematical and statistical tool for estimation of the target variables data at municipal and settlement level.

Cluster and structural analysis as an effective solution for production of data for small territorial units

For the purposes of the Urban Audit project the application of an approach, offering a solution for the above mentioned problems and creating possibility to estimate the missing data from sample surveys, is planned. The idea is based on the hypothesis that there is heterogeneity in each district and it exists between the districts also. The approach consists in clustering of the settlements in Bulgaria by several auxiliary variables in four main thematic directions: demography, economic activity, education and economy thus creating possibility to produce estimates at municipal and settlements level in several steps:

1. **Selection of auxiliary variables** whose correlation with the variables to be estimated is statistically significant. The auxiliary variables are chosen so as to have good predictive power. The variables used by the clustering are age - structure, fertility, mortality, immigration, emigration, number of persons employed (over fifteen, divided into five-years age groups), employment in 21 industries (according to NACE Rev.2), structure of the educational qualification and net sales revenues per capita.

Demographic information is a particular form of auxiliary information and the most important one. It is especially appropriate if the population size or demographic composition of small areas varies considerably. In Bulgaria, with its extreme variation in population densities (from 2.4 people per sq. km. in Treklyano municipality to 3315.8 people per sq. km in Plovdiv municipality), this is a very common issue.

2. **Measurement of the distance between the structures** in the profiles of the population for its basic demographic, social and economic characteristics, listed above, through the formula:

$$\cos\alpha = \frac{\sum_{i=1}^{n} p_{i1} p_{i2}}{\sqrt{\sum_{i=1}^{n} p_{i1}^2 \sum_{i=1}^{n} p_{i2}^2}} ,$$

where:

$\{p_{i1}\}_{i=1}^{n}$ is the structure in respect to the observed indicator in a given settlement;

$\{p_{i2}\}_{i=1}^{n}$ is the structure in respect to the same indicator of the country average;

p_{i1} и p_{i2} are the corresponding relative shares of the two structures;

i is the consecutive share;

n is the number of the relative shares;

α is the angular distance between two vectors, which are points of normalized Euclidean space and represent the structures that are compared;

$\cos\alpha$ is a standardized measure, functionally dependent on the Euclidean distance between the two structures.

This general formulation of the issue is specified by replacing the country average in the previous formula with the hypothetical uniform structure acting as a starting point, the beginning of the co-ordinates. The usage of uniform structure increases the analytical possibilities of the model and allows production of comparisons between the EU countries. Applying this approach for the calculations a concrete measure was used. It reflects the Euclidean distance between the structure $\{p_{ij}\}_{i=1}^{n} = \{p_{ij}, i = 1,...,n\}$ of the j^th settlement and the hypothetical uniform structure $\{{}_e p_i\}_{i=1}^{n} = \left\{ {}_e p_i = \frac{1}{n}, i = 1,...,n \right\}$:

$$\cos\alpha_j = \frac{\sum_{i=1}^{n} p_{ij\,e}\, p_i}{\sqrt{\sum_{i=1}^{n} p_{ij}^2 \sum_{i=1}^{n} {}_e\, p_i^2}} = \frac{\sum_{i=1}^{n} p_{ij} \frac{1}{n}}{\sqrt{\sum_{i=1}^{n} p_{ij}^2 \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} p_{ij}}{\sqrt{n \frac{1}{n^2} \sum_{i=1}^{n} p_{ij}^2}} = \frac{1}{n\sqrt{n \frac{1}{n^2} \sum_{i=1}^{n} p_{ij}^2}} = \frac{1}{\sqrt{n \sum_{i=1}^{n} p_{ij}^2}}$$

where:

$$\sum_{i=1}^{n} p_{ij} = 1,$$

p_{ij} are the shares of the separate units *i*, (*i=1, ..., n)* in relation to the total number of all units in the respective aggregate (jth LAU2);

n is the number of all relative shares.

3. **Definitions of homogeneous groups of settlements** with a lowest intra-group and largest between-group variance. Intra-group and between-group dispersing is determined by the formula:

$$\sigma^2_{\text{intra-group}} = \mathrm{E}(\xi_i - \mu)^2 \approx \frac{1}{n-k}\sum_{i=1}^{n}(\xi_i - \mu)^2 \approx \frac{1}{n-k}\sum_{i=1}^{n}\left(\cos\alpha_i - \frac{1}{n-k}\sum_{j=1}^{n}\cos\alpha_j\right)^2 = \hat{\sigma}^2_{\text{intra-group}}$$

And

$$\sigma^2_{between-groups} = \mathrm{E}(\xi_i - \mu)^2 \approx \frac{1}{k-1}\sum_{i=1}^{n}(\xi_i - \mu)^2 \approx \frac{1}{k-1}\sum_{i=1}^{n}\left(\cos\alpha_i - \frac{1}{k-1}\sum_{j=1}^{n}\cos\alpha_j\right)^2 = \hat{\sigma}^2_{between-groups}$$

Where:

σ^2 is the theoretical – "true" variance value;

$\hat{\sigma}^2$ is the estimate of the theoretical – "true" variance value;

E is a symbol for mathematical expectation;

μ is the theoretical – "true" value of the mathematical expectation;

ξ_i are random values;

$\cos\alpha_i$ are realizations of the random values – measurements of the distances between structures;

α_i are angles between the structures;

i, j indicate the intervals;

k is the number of the groups;

n is the number of the LAU2s;

$n-k$ and $k-1$ are the corresponding degrees of freedom.

If $\hat{F}_{empirical} = \frac{\hat{\sigma}^2_{\text{intra-group}}}{\hat{\sigma}^2_{between-groups}} > F_{theoretical}$ then it is considered that the difference between group means is statistically significant, where the theoretical value $F_{theoretical}$ is taken from the F-distribution table by level of significance $\alpha = 0.05$ and degrees of freedom $(k-1)$ and $(n-k)$.

It is decided to cluster all the LAU2s in Bulgaria, not only the Urban Audit units, in order to enable the production of small area estimates in the next project phases when the city list probably will be expanded. More over, working with all the settlements, we are able to examine the entire population.

4. **Inclusion in the model of non-weighted, aggregated data from sample surveys** on the settlements included into a given sample. The homogeneity ensured based on clustering allows the application of the theoretical definition of Laplace for calculation of probabilities of occurrence of a given event by the realization of a statistical experiment. This definition requires the total number of equally possible events to be put in the denominator of this famous formula. The Laplace's definition is applicable to the cases of finite spaces of elementary events. In this case, the denominator is the sum of all units observed in the cluster, and the numerator is the sum of all observed cluster units that meet the criterion of the target variable. According to the definition of Kolmogorov (which generalizes the Laplace's one and which is fundamental, because it is a base of the axiomatic development of the calculus of probability) the probability measure has the properties as positiveness, normalization and additivity. These properties imply that the probability, calculated in this way, is one and the same in all LAU2s that compose a definite cluster. This probability has to be multiplied by unit's number in each settlement in a given cluster. The result will be an estimate of the respective target variable.

   Scheme:

   $\boxed{\text{Clusterisation}} \Rightarrow \boxed{\text{Homogeneity}} \Rightarrow \boxed{\text{Equally possible events}} \Rightarrow$

   $\boxed{\text{Calculation of the probability measure}} \Rightarrow \boxed{\text{Use of its properties}} \Rightarrow$

   $\boxed{\text{Application of the model to each LAU2 in the respective cluster}}$.

5. **Estimation, validation and presentation of the results** on small territorial units, part of the Urban Audit through thematic maps. Visualisation of the results of estimates will provide a visual and quantitative error assessment and will allow analysts and users to discover the unexpected spatial patterns or anomalies in the territorial distribution of the estimates values.

In the scientific literature the optimum number of the clusters is under discussion. Generally it is considered that their number should be at least 3 and as a maximum 15. Since we are working with a great number of territorial units (5302 settlements existing during the Census ‘2011) using significant number of auxiliary variables, clusters became more sensitive, the spectrum of observations expands and it is possible to have as a result more than 15 homogenous groups.

Based on number of experiments it was found out that 3 clusters with optimal homogeneity should be created for each separate variable. After that the clusters obtained according to all 9 selected auxiliary variables have to be intersected. Because of the variety of data used clusters are split to smaller groups of LAU2s which are of the same nature. The theoretical maximum number of the clusters seems to be 3^9 which is much more than the number of settlements and the real needs, but it indicates that the clustering done has enough disjunctive capacity. We were seeking for a scale that allows enough observations from the sample surveys to be available in each cluster.

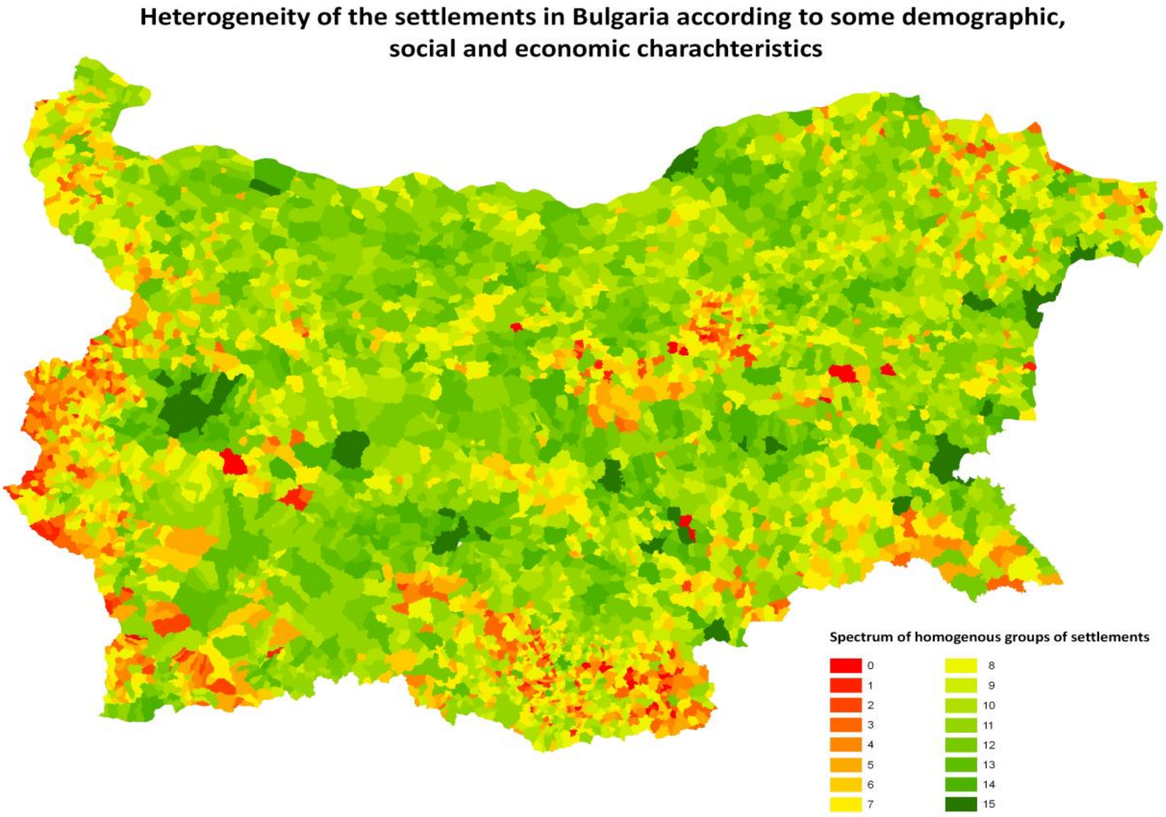
As a result of the clustering 16 homogenous groups of LAU2s were obtained. They are presented on the figure below (only the settlements that have their own land are mapped).

The application of this approach will increase the accuracy of estimates in the following directions:

1. The impact of the sample stratification on the spatial coverage of data will be reduced. Through in the clustering the influence of the geographic factor is taken into account. In each cluster we examine identical LAU2s regardless of their spatial location (whether they are in neighbourhood or in different districts). The LAU2s are considered to be identical although there is heterogeneity in the clusters also but of another order. By the application of cluster analysis we bridge the distances and overcome the restrictions imposed by the physical and temporal dimensions of the geographic space working in another topological space composed by abstract objects (the researched structures themselves represent points that belong to the n-dimensional Euclidean space). Assuming that these homogenous groups of settlements served as strata of the sample, it would be meaningless exactly where in the stratum the sampled households are. This is a possibility to reduce the sample size and to obtain representative results at LAU2 level. In conditions of heterogeneity in the units of observation a systematic error arises. As much the units of observation are, more the systematic error multiplies. Many researchers strive for larger sample size not taking into account the effect on data quality. In the methodology proposed by us the systematic error is lower. We would recommend the number of households observed

in a sample survey to be slightly over their minimum theoretical number required. Of course this is another important field of research;

2. The problem with the small sample sizes and non sampled territorial units is solved. The method proposed ensures data on the non-sampled LAU s and data quality improvement for the LAU s which are included into the sample;
3. Created groups of settlements are homogeneous in respect the auxiliary variables, that correlate to the target variables which will give the opportunity for production of sufficiently reliable inexpensive small area estimates;
4. The pointed approach enables the different sample surveys to work synchronised in a united system.


Heterogeneity of the settlements in Bulgaria according to some demographic, social and economic charachteristics
Spectrum of homogenous groups of settlements
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Bibliography:

1. Bulgaria – Challenges of the poverty, Analysis from the Multitopic Household Survey, National Statistical Institute, 2003.

2. A guide to small area estimations, Australian Bureau of Statistics, 2006.

3. Cameron, A.C., P.K Trivedi, Regression Analysis of Count Data, Cambridge: Cambridge University Press, 1998.

4. Chambers, R.., Calibrated Weighting for Small Area Estimation, Southampton Statistical Sciences Research Institute, Methodology Working Paper, M05/04, 2005.

5. Christov, E., Influence of the Changes in Mortality According to Age and in the Age Structure of the Population on the Changes in Gross Mortality (Methodological Solutions and Empirical Analysis), Population, Review of the Institute of Demography, Bulgarian Academy of Sciences, 22-47, 2000.

6. Trewin, D., Small Area Statistics Conference, Survey Statistician, 41, 8-9, 1999.