ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 182, 2004

Czesław Domański[*]

APPLICATION OF SMALL AREA STATISTICS FOR INTERNATIONAL COMPARISONS

Small is beautiful

1. Introduction

Statistics is the science that has great applicability and this fact is of great significance for research relating to both economic and social life. Any organised community cannot exist without a suitable system of gathering information as well as system of handling information on a scale of the entire country, isolated communities, distinguished communities or distinguished populations. This results in demand for data gathering methods and inference on the ground of them. Costs of conducting empirical statistical investigation are usually rather high. That is why we are looking for methods that would give an opportunity of making full use of the gathered information. The problem appears in the case of analyses relating to subpopulations having data obtained for the entire population with the representative method. The branch of statistics called small area statistics is the one that deals with the problems mentioned above. In the world economy permanent territorial changes take place. The process can cause additional difficulties in continuing statistical investigation in, for example, countries of the former Soviet Union. A similar situation takes place in particular countries. It seems that small area statistics will be able to solve the problem of lack of data for past periods.

These problems have become more and more popular recently both among statisticians and statistical data users. They result from the limitation of financial means and time for conducting investigation, as well as refusal of participating in research, incorrect answers, or the quality of statistical data.

[*] Prof., Institute of Econometrics and Statistics, University of Łódź.

Small area statistics provides methods of using data gathered for the entire population in complete research (censuses, current registration) and fragmentary research (usually representative) to infer about phenomena occurring in subpopulations. The methodology relating to estimation of small area parameters in case of different sample sampling schemes and procedures of data obtaining has been mainly developed. The conception of fragmentary investigation of finite populations with the representative method presented by Neyman (1934) is primary.

The concept of "small area statistics" was developed in the 1970-ies of the former century but problems that it includes had been engaged in much earlier when the fragmentary statistical surveys had begun to be undertaken. In Poland we can number information gathering and handling for tax purposes in the 16th century, the first census in 1789, representative surveys relating to school-age children and number of men of military age which were made on the ground of data originating from the census taking place in 1921 among fragmentary statistical surveys (see: Kordos 1993). The essential development of methodology of small area statistics can be observed for about thirty years. Published papers refer mainly to statistical information making and problems of distribution parameters of subpopulation estimation.

Generally, in small area statistics we can make an assumption that relations occurring between considered quantities or measures for the entire population are retained in distinguished subpopulations that is small areas.

To estimate numerical characteristics of investigated phenomena in small area we can use not only data originated from censuses, i. e. current registration but also information obtained by representative method as a result of suitable samples sampling.

Surveys in which methods of small area statistics are used can often be related to demographic phenomena. In many countries birth and death registrations are run. However, the problem arises from determining regional size of migration. The basis of estimating the level of the phenomenon method on the particular area is an assumption that the proportion of migration levels in the entire population and distinguished subpopulation (the population inhabiting distinguished area) is the same as proportion of the number of school-age children in the entire population and this subpopulation (see: Dol 1991).

Another method of migration size estimation for distinguished subpopulation uses the difference between the real number of school-age children (obtained from suitable registrations) and the number of school-age children estimated on the ground of life duration table for a suitable age group. Knowing the migration level in subpopulation we can estimate the amount of population in subpopulation.

Another method uses the estimation of number of households in subpopulation and average number of people per household for the entire population to estimate the amount of population in this subpopulation. The difference between the numbers of households obtained on the ground of the last census made before the investigated period multiplied by the average number of people in a household is the quantity which is used for estimating migration level and then the number of people in subpopulation in a particular period.

Methods of small area statistics can be used in different fields of socio-economic life (see for example Falorsi and others (1993), Platek (1993), Falorsi and others (1999), Fay, Herriot (1979), Domański, Pruska (1998, 1999), Gambio, Dick (1999), Gambio and others (1998), Gołata (1996, 1997, 2002), Kordos (1997, 1999), Kordos, Paradysz (1999), Witkowska (1999), Witkowski (1992), Rao (2002). The subject literature is very wide and it is hard to present the most important publications even in general profile. However, the above list shows that Polish scientists have made a great contribution in this branch of science. Kordos (1994, 1997, 1999, 2002) presented achievements on this scale most widely. In recent years the most significant moments for the development of this branch of science were three international conferences, which took place in turn: in Ottawa in 1985, in Warsaw in 1992 and Riga in 1999. In 2000 The International Consortium of "Small Area Statistics" was created within The Fifth Framework Project in which statisticians from seven countries took part. They came from Finland (Risto Lehtonem), Spain (Carlos Ballano Fernández), Norway (Li-Chun Zhang), Poland (Jan Kordos) Sweden (Sixten Lundstrőm), Great Britain (Patrick Heady and Kerry Ellis from The United Nations, and Ray Chambers from Southhampton University, N. T. Longford from Medical Statistics Department, Harvey Goldstein from Education Institute) and Italy (Stefano Falorsi). The organizers assumed that partners taking part in consortium should:

a) understand what problems of small area estimation ate to be solved,

b) know basic methods of small area estimation,

c) have basic data sets on the ground of which these methods could be simulated.

The general purpose of the research scheme is to provide European countries (and EU as a whole) with estimation methods directed at small areas that would enable to obtain solid estimations of these areas.

More specific purposes were formulated as follows:

1. Development and improvement of suitable statistical methods which are directed at small area estimation and the estimation of their practical aptitude.

2. Development or improvement of administering data systems used with small area estimation.

3. Investigation of conditions for these methods application in each interested European country.

2. Basic issues of small area statistics

Small area statistics as a branch of statistics deals with methods of using statistical information obtained for the entire population (based on representative research of population, censuses, current registration and other auxiliary data) for inferring about investigated features in distinguished subpopulations which are called small areas, fields or domains.

One of small area classifications was proposed in the paper of Purcell and Kish (1979). They distinguish four kinds of small areas:

– main small area (if its population totals at least 10% of entire population size)

– secondary small area (if its population totals from 1% to 10% of the population size)

– mini small area (the area which has the size of the order from 0,01% to 1% of the population size),

– rare small area (if the number of its elements is smaller than 0,01% of the population size).

The requisition for information and statistical methods for small area is notified both by the government (because of distribution problems and detecting regions of characteristics which differ much from the countrywide indexes) and self-government as well as individual businessmen who are interested in local economic situation for planning and administrative purposes.

A huge interest in small area statistics results from, among others, the need of:

– obtaining statistical data for suitable geographical levels in sections,

– working out programmes of regional, economic and social development, as well as their realization estimation,

– monitoring various phenomena and processes for distinguished subpopulations (for example districts, communities and pensioners' households).

Statistical information gathering on regional scale, when the requisition for the data appears, is very expensive. That is why the problem of possibility of using data gathered for nation-wide research purposes in regional research arises. It is of particular importance in case of data gathered using the representative method. That is because the representative method for the entire population does not have to provide representative information for distinguished subpopulations, or there is too little information for small area to use in the process of various economic decision making.

Small area statistics research concentrates mainly on such construction of estimators or functional characteristics of investigated variables distributions which are to provide precise estimations of these distributions parameters. New possibilities of increasing the size of the sample referring to small area or increasing the representativeness of the sample are created by simulation methods.

Methodological papers about small areas refer to developing procedures of gathering reliable data (special sample sampling schemes, obtaining extra information) and statistical inference methods for subpopulations. The average and global value are parameters which are estimated most often. Much attention is paid to error analysis of estimators. The sample for small area is made by elements of sample sampled from the entire population, which pertain to small area. Features of estimators for samples specified in this way generally differ from those, which are sampled from the entire population.

3. Short characteristics of construction of small area estimator

Distribution of investigated characteristic in a given population can be characterized by various methods. Most often we estimate chosen parameters and precision estimation of the estimation. In order to do that we can use direct or indirect methods (see: Schaible and Casady 1994, Domański and Pruska 2001).

Direct estimation is the estimation of small area parameters on the basis of random sample whose elements are units pertaining to sample sampled from the entire population and small area, or on the basis of sample sampled specially from small area. If we use information referring to other subpopulations or the entire population in the inference process, we talk about indirect estimation. It is aimed at increasing estimation efficiency.

On the ground of Koros's paper (1999) we can give the following classification of indirect estimation methods.

– estimation based on data from another small area but coming from the same period in which a given small area is investigated,

– estimation using investigated variable value from a given small area but from another period than the analysed one,

– estimation based on variable values from another small area than investigated one, and another period than considered one,

– estimation based on data from other sources.

Estimators whose values result from transformation of only values of investigated variable observed in a sample are called common ones. If we use additional information (for example about auxiliary variables) while estimators constructing the special names, which are typical of given statistics (for example quotient or regressive estimator), will be assumed.

Within indirect estimators we can distinguish synthetic estimators class. They are constructed by assumption that population structure does not differ much from small area structure. One of the possible approaches is the division of population, which contains H small areas to G separable layers, however distinguished populations and layers are not separable. Constructing synthetic estimators we often assume that G layers in a small area have the same numerical characteristics as G analogous layers in the entire population, for example in particular layers the relation of global values of two characteristics in the entire population is the same as the one in distinguished small areas.

Methods of indirect estimators for small area construction are to lead to increasing their efficiency in comparison with direct estimators. We also consider estimators which are linear combination or more complex functions of other estimators. We should also notice that statistics' characteristics are influenced by sample sampling scheme.

Small area estimation issues concentrate mainly on estimation of global value, average and obtained estimations precision. We can distinguish a few main types of estimators of global value for small area (see: Falorsi and others (1999)). In case of dependent sample sampling the following estimators belong to them:

1) Common estimator:

$$\hat{T}_{h.} = \frac{N_{h.}}{n_{h.}} \sum_{i \in \lambda_{h.}} y_i , \tag{1}$$

where: $N_{h.}$ and $n_{h.}$ – number of elements, adequately from h-th small area and the entire population belonging to h-th small area.

2) Quotient estimator:

$${}_q\hat{T}_{h.} = \frac{X_{h.}}{\hat{X}_{h.}} \hat{T}_{h.} , \tag{2}$$

where: $X_{h.}$ – global value in h-th small area, $\hat{X}_{h.}$ – global value estimator in h-th small area, ${}_q\hat{T}_{h.}$ defined by the formula (1).

3) Synthetic estimator:

$${}_s\hat{T}_{h.} = \frac{\hat{T}_{A.}}{\hat{X}_{A.}} X_{h.} , \tag{3}$$

where: ${}_{s}\hat{T}_{h.}$ – global value estimator adequately in A group of small areas and in h-th small area.

4) Complex estimator:

$$ {}_{com}T_{h.} = w_{d}\ {}_{q}\hat{T}_{h.} + (1 - w_{d})_{s}\hat{T}_{h.}, \tag{4} $$

where: w_{d} is a constant from $\langle O;1\rangle$,

5) Sample size-dependent estimator:

$$ {}_{sample}\hat{T}_{h.} = \alpha_{h}\ {}_{q}\hat{T}_{h.} + (1 - \alpha_{h})_{s}\hat{T}_{h.}, \tag{5} $$

where: $\alpha_{h} = 1$ if $n^{*}_{h.} \geq N^{*}_{h.} = N_{h.} / N$, and at the same time $n^{*}_{h.} = n_{h.} / n$, and otherwise $\alpha_{h} = n^{*}_{h.} / N^{*}_{h.}$ (n- population sample size).

Empirical Best Linear Predictors (EBLUP) make another group of small area estimators.

One of them is the estimator of the following form:

$$ {}_{emp}\hat{T}_{h.} = \gamma\ {}_{h}\hat{T}_{h.} + (1 - \gamma_{h}) X_{h.}\hat{\beta}, \tag{6} $$

where: $\hat{\beta}$ is BLUE estimator that is Best Linear Unbiased Estimator of parameter β which occurs in model:

$$ \hat{T}_{h.} = X_{h.}\beta + v_{h.} + e_{h.}, \tag{7} $$

in which $v_{h.}$ is such a random component that $E(v_{h.}) = 0$ and $D^{2}(v_{h.}) = \sigma^{2}_{v}$ and $e_{h.}$ stands for error resulting from using estimator $\hat{T}_{h.}$ and at the same time $\boldsymbol{E}(e_{h.} | T_{h.}) = 0, D^{2}(e_{h.} | T_{h.}) = \sigma^{2}_{e}$. The average γ_{h} in formula (7) stands for the relation of variance σ^{2}_{v} estimation and the sum of variances σ^{2}_{v} and σ^{2}_{e} estimations.

In given formulas sizes $N_{h.}$ occur. If they are not known they should be estimated.

Size $(N_{h}(h = 1,..., H)$ estimator has the form of the following statistics:

$$\hat{N}_h = \sum_{i\in\lambda_{h.}} \frac{1}{\Pi_i}. \tag{8}$$

Depending on the sample sampling scheme and information about auxiliary variables it is possible to construct other average and global value estimators. In each of those issues their variance estimation and distribution form defining becomes a big problem. Estimation precision issues are analysed in various papers. Very often these are simulation investigations. Dehnel (1997) in his paper presented results which indicated bigger precision of regressive estimators in comparison with direct and quotient estimators.

The comparison of estimators (1) and (6) precision on the ground of empirical data using Monte Carlo methods and individual dependent sampling was presented in the paper of Falorsi and others (1991). According to the presented results complex estimator (4) had the best precision measured, among others, by means of the following measure:

- the average of relative radical of mean square error:

$$RMSE = \frac{1}{H_a}\sum_{h=1}^{H_A}\frac{1}{T_{h.}}\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left({}_l\hat{T}_{h.} - T_{h.}\right)^2} \times 100 \tag{9}$$

where H_A stands for the number of small areas belonging to distinguished small areas set, L – the number of samples for which parameter $T_{h.}$ value was estimated, ${}_l\hat{T}_{h.}$ – the estimator of parameter $T_{h.}$ distinguished on the ground of l-th sample $l = 1, ..., l$.

In the discussed investigation complex estimator (4) had a bigger bias than estimators (1), (2) and (5).

Estimation of estimators' variances for small area can be developed by simplified methods for example by Mahalanobin method or jackknife method (see: for example Bracha 1998, Domański, Pruska 2001).

We should emphasise here that estimation of estimators' variances, used in small area statistics, is the problem which is being continually discussed and updated.

4. Examples of international solutions

In the USA billions of dollars are annually divided between states, smaller communities and, in particular, administrative units of the lower level – counties whose number totals 3143, depending on their economic and social characteristics. The criteria of division of some of funds depend on the population number of a particular county, income or the poverty size. The Census Office of the USA usually provides indispensable data. Data are based on information obtained from censuses, which take place every 10 years. Up till now the size of these funds has been defined mainly on the ground of information obtained from census data. In the USA it is already a traditional source of defining of income and poverty estimates on the local scale. Estimates refer to the following parameters:

- size of income median in households,
- number of people living below poverty limit,
- number of school-age children (age 5–17) living in poor families
- number of people over the age of 18 living below poverty limit.

As the economic conditions of particular regions have been changing fundamentally between census periods, it was necessary to develop methods allowing to obtain more current data. We can illustrate it by the following numbers: in the years from 1989 (the reference year for income data obtained from the census in 1990) to 1993 the median of households' income decreased by 7%, the number of people below poverty limit increased by 25%, number of school-age children living in poor families increased by 24%. Moreover, what is also very important the changes are not uniform in case of a particular region. Research results of the Current Population Survey (CPS) disclosed 52% increase in the number of the poor in Florida and 44% increase in California, but only 4% in Texas and 7% in Illinois. Considerable diversification of these parameters can also occur in districts.

To obtain more current data about poverty for states and districts it was necessary to use the representative survey results, which is CPS. However, the way of obtaining estimates on the local scale is different for the census and for CPS surveys. What is also different is the precision of the obtained estimations. This is caused by various reasons.

Developed statistical model links data coming from various sources including:

- march CPS survey in which income data are obtained,
- food stamps programme data,
- tax sets data, and
- population size estimation for investigated districts.

Four models, which allow defining quantities connected with describing poverty on the scale of a particular district, were outlined. The quantities are as follows:

a) global number of people living below poverty level,

b) number of school-age children living below poverty level,

c) number of people under the age of 18 living below poverty level, and

d) size of income median.

We showed only the general treatment of size estimation of school-age children poverty according to the American counties. Research team is still working so the final report has not been published yet. However, we can come to some conclusions on the ground of preliminary reports. We can see that various data sources were used here, obtained from both statistical surveys and administrative records. When using the term "small area" we understand the area for which, in some circumstances, we are not able to estimate demanded parameters estimations which could be regarded as reliable ones. For example, we can regard provinces of small number of investigated households on the ground of which we can obtain unreliable estimations as small areas.

Kordos, Kubacki (1999) presented the application of the American approach to the estimation of poverty level for small areas in Poland.

However, we cannot directly apply the American approach to estimate poverty level according to new districts. In Poland we have access to completely different data sources. In our census we do not obtain information about the income from considerable fraction of households taking part in it. Instead we have household budget surveys and life conditions surveys in which population income data are gathered. Kordos and Kubacki propose solving the issue in a few stages. In the first one it would be necessary to use accessible data sources from statistical surveys and administrative records and try to estimate the poverty size for 49 former provinces. For most of these provinces the estimated precision of the poverty level on the ground of households' budgets survey is very low. We can also estimate poverty size for 10 macro-regions for which households budget surveys were estimated. It would be experimental research work. Experience from the work could be used in further stages which are aimed at estimating the poverty level for districts, and next, for 16 new provinces.

From the previous investigations of household budget and population's living conditions which took place in 1996 we are not able to estimate poverty size for former provinces. The estimation precision for most of provinces is very low. That is why the estimations do not have very big cognitive value. However, we can try to make them using small area statistics methods.

In the quoted paper authors presented approach based on the model which can be used to obtain estimations of number and frictions of poor households, for example in 1996. The estimation for province procedure uses two regression models which estimate poverty level – the previous provinces model (DW) and the 10 macro-regions model (MR) according to which household budget surveys (BGD) are developed.

Estimation procedure in 1996 contains:

1) creating and using DW model for 49 provinces aimed at obtaining preliminary estimations of the number of households living below poverty level (GUS, 1997). The estimations for previous provinces contain:

– using data obtained from administrative records and other sources which are accessible for all previous provinces and using them as predicted variable,

– defining and estimating regression equation referring to predicted variables in relation to dependent variable which is estimated as logarithm of the number of poor households in 1996 from household budget surveys for particular provinces,

– obtaining estimation of the number of poor households for 49 provinces using estimated regression coefficient from equation and predicted variables. For previous provinces which contained households in sample of household budget survey, estimations from the model are linked in some way with estimations obtained from BGD surveys for these provinces (that is estimations obtained from the model for a given province and surveys (BGD));

2) building and using macro-regional model MR to obtain estimations of the number of poor households according to macro-regions. Estimation procedure for macro-regions is similar to DW model for previous provinces although MR model differs from DW model in a few places:

3) correcting initial estimations of the number of poor households obtained from DW model (step 1) for the sake of compatibility of a given macro-region (step 2). In this way we obtain final estimations of the number of households living in poverty in 1996 for provinces of a given macro-region;

4) obtaining estimations for macro-regions in 1996 of general number of households using demographic data. We can use estimations from steps 3 and 4 to calculate fraction of poor households for previous provinces which can be useful in scientific research and socio-economic policy.

The equation of previous provinces uses, as predicted variables, estimations from records:

- individual taxes,
- people getting social welfare,
- registered jobless,
- the number of pensioners of Social Insurance Establishment,
- The level of PKB (GDP) from the nearest year.

The number of the poor is estimated from the household budget surveys in 1996. The equation is as follows:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + u_i + e_i$$

where:

y_i – logarithm (the number of poor households in DW i),

x_{1i} – logarithm (the number of people in tax registers of total income of specific value in province DW i),

x_{2i} – logarithm (the number of people getting social welfare in DW i),

x_{3i} – logarithm (the estimation of number of households in DW i),

x_{4i} – logarithm (the number of pensioners in DW i),

x_{5i} – logarithm (the number of the jobless registered in DW i),

x_{6i} – logarithm (PKB level estimated for DW i),

u_i – the model error for DW i ,

e_i – the random error of dependent variable for DW i .

Another example of the application of small area statistics for international comparisons is labour market. It can be analysed on the scale of the entire population (for example for particular country or particular groups of countries) or distinguished subpopulations. In the second case methods of small area statistics can be used to estimate sizes characterizing labour market, such as:

– number of occupationally active people,

– number of the jobless,

– number of people looking for a job because of losing one, resigning from a job, intending to take a job after a break or intending to take a job for the first time.

These sizes can be regarded as global values of suitable variables and estimated by using estimators presented in the paper of Domański, Pruska (2001) if the sample sampling runs according to scheme of laminar, individual, dependent sampling. If the survey is run on the ground of information gathered by means of multistage sample sampling methods, global value estimators formulas have different form and become more complex (see: for example, Russo, Falorsi 1999; Bracha 1998; Gołata 1996).

It is a good idea to use methods of small area statistics for survey of unemployment phenomenon on the regional scale considering possibility of using data gathered in the entire population surveys (both in representative surveys and complete ones, for example current registration of people looking for a job). Small area statistics methods allow to avoid making special survey in order to gather needed information.

The analysis of unemployment was presented using the example of three macro-regions: Warsaw, Łódź, Lublin. Methods, which we used, can be applied analogously in other regional unemployment surveys, for example in particular UE countries or, in the future, in specified regions of Monetary Union.

Population of people at the age of 15 and more, according to international standards, is divided to three categories; the employed, the jobless, the occupationally passive. The employed and the jobless make a group of occupationally active people and all the rest of population is subsumed into the group of occupationally passive ones. Unemployment meter stands for the so called unemployment rate, which qualifies the unemployed share in the number of occupationally active people.

Economic activity analyses of population are made systematically and in various aspects in many countries. These types of investigations are also conducted in Poland (see: Kałaska, Witkowski 1993), for example on the ground of current register, in Employment Offices and special questionnaire. One of them is Survey of Economic Activity of Population (BAEL) made four times a year since 1992 (in February, May, August and November). Information is gathered on the ground of two-stage sampling sample in which the first-stage elements are sampled with stratification according to provinces. At the same time provinces are divided to rural layer and from 2 to 5 urban layers. The second-stage sampling layer contains flats. All inhabitants of sampled flat pertain to the sample.

According to definition accepted in BAEL employed people are those who worked or did not work in an investigated week but were employed by some employer.

According to BAEL, the jobless are people aged 15 or more who do not work but expect to start working in next 30 days or fulfil three conditions:

- in investigated week period were not employed,
- were actively looking for a job, that is took an action aimed at finding a job during four weeks previous to the survey,
- were able to take a job in investigated week as well as in the following one.

In our example we accepted a little different definition of the jobless. Among ones we number people who regard themselves as those who are looking for a job but expect to take one. The remaining ideas (the occupationally active, the employed and the occupationally passive) are conformable to version given above. However, we must remember that changing the definition of the jobless causes the change of the occupationally active group size. Introduced modifications were caused by intention of making a survey that differs from BAEL analyses, result of which is published in Poland by Central Statistical Office. Although in our investigation we use statistical data gathered in BAEL, our results give a little different unemployment estimation than analyses based on the definition accepted by BAEL.

We used data gathered in BAEL survey in November 1996 and November 1998.

In order to determine estimations of global value of some variables characterizing labour market in considered macro-regions we used direct estimator (5). It means that we applied estimators corresponding to single stratiform sampling (layers stand for provinces which pertain to particular macro-region, and auxiliary variable stands for demographic variable).

Czesław Domański

ZASTOSOWANIE STATYSTYKI MAŁYCH OBSZARÓW DO PORÓWNAŃ MIĘDZYNARODOWYCH

Statystyka jest nauką mającą ogromne zastosowanie, zarówno w życiu społecznym, jak i ekonomicznym, jednak nie często dysponujemy pełnymi informacjami zbiorowości generalnej, zachodzi więc potrzeba stosowania metod wnioskowania statystycznego. Koszty prowadzenia badań statystycznych są na ogół wysokie, stąd statystyka małych obszarów wychodzi naprzeciw potrzebom, co więcej, dynamiczne zmiany, w tym terytorialne, zachodzące w świecie uniemożliwiają prowadzenie ciągłych badań statystycznych. Artykuł zawiera propozycje stosowania metod statystyki małych obszarów w porównaniach międzynarodowych na przykładzie badania sfery ubóstwa i bezrobocia w ujęciu regionalnym ze szczególnym uwzględnieniem badań mikrospisów prowadzonych w USA i w Polsce.