Estimating International Migration on the Base of Small Area Techniques

Voineagu, Vergil and Caragea, Nicoleta and Pisica, Silvia

2013

Professor Vergil VOINEAGU, PhD
Academy of Economic Studies
E-mail: vergil.voineagu@yahoo.com
Associate Professor Nicoleta CARAGEA, PhD
Ecological University of Bucharest
E-mail: nicoletacaragea@gmail.com
Silvia PISICA, PhD
National Institute of Statistics
E-mail: silvia.pisica@insse.ro

ESTIMATING INTERNATIONAL MIGRATION ON THE BASE OF SMALL AREA TECHNIQUES

Abstract. *Population migration flow is a component of population facing difficulties in measuring in the inter-census period of time. The rationale of this study is that Romanian statistics on international migration flows are of very poor quality, the availability of data on past trends being strongly limited, provided only from administrative sources. For this reason, in the inter-census period, the variable of interest is provided by the labour force survey available at national and regional level every quarter of the year since 2004. The smaller disaggregation like localities level using direct estimators conducts to results of unreliable estimates and will surely lead to higher standard error and consequently, high coefficients of variation. The main reason for this is the insufficient number of respondents or no respondent at all in a small domain. Small area estimation techniques are able to carry out the estimation at the localities level (NUTS[1] 5).*
The main purpose is to provide methods able to estimate the population in Romania, based on the Labour Force Survey and also the results of 2002, respectively 2011 population census.

Key words: *international migration, population, demography, statistics, small area estimation*

JEL classification: C13, C15, C53

1. Introduction

[1] NUTS – Nomenclature of Territorial Units for Statistics

Based on small area estimation techniques, we intend to carry out a simulation of international migration using the results of the Population Census (2002 and 2011, respectively).

International migration consists of two components: emigration and immigration. Statistically speaking, according to Regulation (EC) No 862/2007 on Community statistics on migration and international protection, we define the two components of international migration as follows:

- 'immigration' means the action by which a person establishes his or her usual residence in the territory of a Member State for a period that is, or is expected to be, of at least 12 months, having previously been usually resident in another Member State or a third country;
- 'emigration' means the action by which a person, having previously been usually resident in the territory of a Member State, ceases to have his or her usual residence in that Member State for a period that is, or is expected to be, of at least 12 months.

Every EU citizen has the right to live and work in any country within the Community. This is one of the most tangible EU membership benefits its citizens enjoy. For some, this involved moving from poorer countries to richer countries, generally to northwestern Europe to benefit from higher wages and better living conditions. However, free movement of persons in Europe creates difficulties in the measurement of international migration in general and international emigration in particular.

The data on immigration accurately capture the phenomenon, as they are recorded through administrative sources and provided annually by IGI (Inspectorate General for Immigration) for foreigners establishing their residence in Romania and by DEPABD (Directorate for People and Database Administration) for Romanian citizens reestablishing their residence in Romania.

Emigration is, however, very difficult to quantify; it is more difficult to count people leaving than those arriving in a country. National legislation does not provide for citizens' obligation to notify authorities when establishing usual residence in another country. Registration in the Passports Directorate records is performed only if Romanian citizens request establishing their domicile (permanent residence) in another state, either an EU member or a non-member.

Thus in terms of emigration, the existing[2] data from administrative sources currently do not cover the whole phenomenon, as there is a severe underestimation in the number of emigrants; this deficit gives an overvaluation of the Romanian population.

Lack of availability of exact figures on emigration led to the need for new statistical thinking based on estimation methods. On the recommendation of the European Commission under Article 9 (1) of Regulation 862/2007, in the course of the statistical procedure, the National Statistical Institutes are allowed to use *"well-documented statistical estimation methods based on scientific data"*. Such estimations can be carried out when data observed directly is not available or, for

[2] At the end of 2012

example, when data from administrative sources must be adjusted to meet the definitions.

Statistical estimations have been used in the past in a number of countries as part of the production of official statistics on migration, especially when data sources from statistical sample surveys were used mainly. The intent of this Regulation provision was to ensure that where national authorities would still use estimations, the estimating procedures used are transparent and clearly documented.

2. Description of the estimation method (SAE) - conceptual framework

The basic idea of the method is to apply econometric models to estimate international migration for the lowest administrative units.

Small area estimation method involves producing estimators for geographical areas for which the direct results obtained from statistical sample surveys are not reliable (of trust). Conceptualization of "small area estimation" can create confusion, because this technique does not require that the domains must be small, but the number of statistical units selected from the respective geographical areas must be low. Therefore, distrust on direct estimates for small areas (by localities, for example) results from the fact that the samples comprise a too small number of statistical units, or - in some cases – this even do not exist.

Small area estimation "borrows" relevance and accuracy by combining data from sample surveys with covariates from other data sources (census or administrative sources). However, the estimates should be treated carefully due to estimation errors. It is known the idea that any estimation generates suspicions concerning the way to fit the model and the accuracy of the results and thus suggests the existence of potential errors resulting from the difference between the estimates and the true values.

In this sense, particular interest will be given to the diagnosis of the models applied; respectively the sources of error[3] with impact on the results of the estimation based on the small area techniques will be examined. Even so, one should consider a clear message: "the results of estimation methods cannot guarantee that these are the true values of each range deemed small". Prediction errors give an indication of model's reliability in terms of estimated values closeness to the reality, but neither these errors cannot be certain, because in practice, the true values of the variable of interest are unknown.

Typology of estimators obtained by applying estimation techniques on small areas and description of significance thereof.

We distinguish between three types of estimators:

[3] There are mainly three types of errors: sampling errors, auxiliary variables errors and model generated errors.

a) Direct Estimators - are derived from data obtained from the sample survey (specifically, direct estimators consist in grossing up the sample data at locality level based on the weighting coefficients - Horwitz-Thompson estimators)
b) Synthetic Estimators - are determined by applying a regression model combining the covariates corresponding to the sample survey;
c) Aggregated Estimators - are obtained based on a linear combination between the Direct Estimator and the Synthetic Estimator.

To ensure representativeness on small areas, the estimators must be unbiased (the estimated average of the variable of interest must represent all the statistical units in the sample). Unbiased Estimators are usually obtained by selecting very large samples, the selection comprising statistical units distributed in all the small domains (*design-unbiased estimators*). Direct Estimators can be used successfully in this case.

It is not always possible to reach the representativeness of the data only by using samples. Therefore, it is necessary to use econometric methods to determine some unbiased estimators (*model-unbiased estimators*).

GREG Estimator

The generalized regression estimation (GREG) is a design-based model-assisted approach with numerous applications to domain estimation and it is sometimes used also in small area estimation. GREG is obtained by adjusting the direct estimator with the differences between covariates provided by two sources of data.

The model that adjusts the direct estimator is based on the correlation between the y variable of interest and covariates x_i.

The GREG estimator is model-assisted in the sense that regressing y on x is done only for removing the unexplained variation from y to increase estimation accuracy.

Regression equation can be written as:

$$\hat{\bar{Y}}_d^{GREG} = \frac{1}{\hat{N}_d}\sum_{i \in s_d} w_{id} y_{id} + \left(\bar{X}_d - \frac{1}{\hat{N}_d}\sum_{i \in s_d} w_{id} x_{id}\right)^T \qquad (*)$$

GREG includes direct estimator, calculated exclusively on the basis of data obtained from the survey (LFS). The direct estimator is a sum of the Horvitz-Thompson estimator:

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d}\sum_{i \in s_d} w_{id} y_{id} \qquad (**)$$

Where:

$\hat{N}_d$ - represents the total population for each area

w_{id} - weight of selection / inverse value of the inclusion probability of unit i in area d

y_{id} - variable of interest for unit i in area d

d - number of small areas (such as localities, according to identifier code of the domain)

$\hat{\beta}$ - regression coefficients

$\bar{X}_d$ - covariates contained in census file

x_d - covariates contained in LFS file

Based on formulas (*) and (**) is obtained the mathematical expression of GREG estimator:

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d^{DIRECT} + \left(\bar{X}_d - \frac{1}{\hat{N}_d}\sum_{i \in s_d} w_{id} x_{id}\right)^T \hat{\beta}$$

As a result, we get GREG estimator (estimated values of the y variable of interest, based on regression between the directly estimated y values and covariates values obtained from two data sources). It will be observed that the GREG estimator adjusts the direct estimator in terms of a higher degree of homogeneity, i.e. a smaller variation of the data obtained at area level. Will compare variances for the two or more estimators. It is expected that the dispersion for GREG estimator to be less than the one of direct estimator.

Disclaimer: using JoSAE package from R, it is obtained the GREG estimator without additional calculations.

For finer adjustments are used other estimators that are based on more complex econometric models *(model-unbiased estimators),* as follows in the next section.

SYNTH Estimator

The synthetic estimator is based on assuming a (linear) model for the data so that the values of the areas that have not been sampled are estimated from the model using only information for available covariates. In other words, the Synthetic estimator is set up so that involves linearity between the variable of interest / dependent variable and independent variables / influence factors for all areas, including those that were not included in the sample, the values on area level are estimated using the additional information fields known to the entire population statistics (census, for example, or other administrative sources).

The general equation of synthetic estimator can be written as:

$y_{id} = x_d^T \gamma + u_d + e_d$

Where:

x_d^T - is the transposed matrix composed of covariates values obtained in the areas / localities for the entire population (known values of census)

γ - regression coefficients

u_d, e_d - the residuals of the regression whose average is zero and variances are σ_u, σ_e (u_d, e_d have normal distribution, centered, with variances σ_u, σ_e)

u_d - the random-effect residuals

e_d - residuals due to fixed effect

An equivalent equation can be written as:

$$y = x\beta + zu + e$$

Where:

y and e - are vectors of size (m x 1)

x - is a matrix of size (m x p)

β - is a vector of size (px 1)

u - is a vector of size (D x 1)

z - is a matrix of size (m x D)

m - number of localities / small areas contained in the sample

p - number of covariates

D - number of small areas ranging from the entire population (in this case equals the number of localities, according to area identifier code = m).

Using JoSAE package from R, the Synth estimator is obtained.

EBLUP Estimator (Empirical Best Linear Unbiased Predictor)

In statistics, BLUP is the resulting value of a predictor based on linear mixed models to estimate parameters due to regression's random effects. As we mentioned in the previous section, linear mixed regression models may be applied if the individual data *(unit level)* can be organized into groups / areas / clusters *(area level)*. In these cases, the theoretical values of the dependent variable / variables of interest are correlated with factor variable / independent variables through a regression function whose parameters can be estimated by different methods [2]. These regression parameters can be generated by fixed effect regression (number of coefficients is equal to the number of factors in the model), or can be generated by random effect (number of coefficients are multiplied by the number of areas / groups).

EBLUP is the sum of sample observations and predicted values of non-sampled observations of variable y.

EBLUP estimator, at statistical unit level *(unit level),* is calculated as:

$$EBLUP\_A = \bar{X}_d^T \hat{\beta}_{unit} + \gamma_d(\bar{y}_d - \bar{x}_d^T \hat{\beta}_{unit})$$

Or an equivalent formula:

$$EBLUP\_A = (\bar{X}_d^T + \gamma_d \bar{x}_d^T)\hat{\beta}_{unit} + \gamma_d \bar{y}_d$$

The estimates of β are obtained using standard Generalized Least Square techniques, whilst the estimates of *u* are computed using their Empirical Best Linear Unbiased Predictor (EBLUP).

EBLUP estimator in the domain level/ area level *(area level)* is calculated as:

$$EBLUP\_B = \bar{X}_d^T \hat{\beta}_{area} + \gamma_d(\bar{y}_d - \bar{x}_d^T \hat{\beta}_{area})$$

Or an equivalent formula:

$$EBLUP\_B = (\bar{X}_d^T \cdot \gamma_d \bar{x}_d^T)^{\downarrow}_{area} \cdot \gamma_d \bar{y}_d$$

Where:

$\bar{X}^T$ - Is the transposed matrix composed of covariates values obtained in the localities for the entire population (from census).

x_d^t - Transposed of covariates at localities level for units in the sample (from AMIGO/LFS).

$\bar{y}_d$ - variable of interest at area level contained in the sample

The fundamental difference between GREG, Synth and EBLUP estimators is that EBLUP estimators reduce the noise / enhance large dispersions of the mean values estimated by using γ_d regressors. Use of this regressor has as a result the modeling in u_d residual values (unknown) of population (from census) based on e_d residual values of individuals in the sample (from LFS).

Using JoSAE package from R, we obtain the output for EBLUP, GREG and Synth estimators.

The first results are actually "first step" in econometric modeling approach. Visualization of the results, their interpretation and statistical analysis creates the potential ways to improve the models used and the resumption of techniques of estimation. The main diagnosis indicators are generated by default by JoSAE packages used in the R software. For example, in case of linear regression model is used **nlme** package incorporating the function **lm** (linear model). A summary of the results obtained presents the regression coefficients (calculated using the least-squares method), standard deviations, t-student and F-statistic significance tests, and values of probability values of regressors estimation for various degrees of freedom.

3. Conclusions

The main conclusion is there are a set of difficulties encountered in the process of estimation. Because of them, the result of small area estimation of international migration is still in progress. Estimations will be used in near future in the calculation of population, migration stock being one of the demographic component.

Considering the methodological complexity of the small area estimation models, it is expected that the process of applying the methodology will take long and will set out a number difficulties.

1. The main difficulty arises from the fact that the sample of LFS (the only source with annual periodicity that provides information on the variable of interest: the number of emigrants left by 12 months and over) was not designed to rigorously estimate the international migration. Moreover, the sample does not fully cover all areas of Romania (about a quarter).

2. A second set of problems is generated by applying the econometric models from which migration is estimated. Moreover, small area estimation models use data from different sources, involving a large number of estimation stages, which could lead to displacement of estimators beside real values *(bias).* Rao [10] considers that "*we should take extra-precaution when using the sae method for approximating the mean squared errors, especially when the sample size is small and the variance components are large*".

3. Other limitations in applying estimation models are related to data availability. For example, an important factor with direct impact on international migration is the difference in economic welfare between countries of destination and countries of origin. Acquaintance with variables that quantify the economic welfare of the individual, such as income, could lead to increased significance estimates. Note that the variable should be available in both data sources (sample, respectively census).

4. Detailing the variables on disaggregation levels is not possible to estimate, but by deepening small area estimation procedures (requires sample segregation accordingly its diminution, depending on the variable of interest).

REFERENCES

[1] **Baltagi, H.B., (2011),** ***Econometrics***, Springer Publisher, ISBN 978-3-642-20058-8.

[2] **Battese, G. E., Harter, R. M. & Fuller, W. A. (1988),** ***An error-components model for prediction of county crop areas using survey and satellite data,*** Journal of the American Statistical Association, 83, 28-36

[3] **Breidenbach, J. and Astrup, R. (submitted 2011),** ***Small area estimation of forest attributes in the Norwegian National Forest Inventory***. European Journal of Forest Research.

[4] **C.-E. S¨arndal, B. Swensson, and J. Wretman.** ***Model Assisted Survey Sampling***.Springer-Verlag Inc., New York, 1992.

[5] **Caragea, N., Alexandru, A.C., Dobre, A.M. (2012),** ***Bringing New Opportunities to Develop Statistical Software and Data Analysis Tools in Romania***, The Proceedings of the VIth International Conference on Globalization and Higher Education in Economics and Business Administration, ISBN: 978-973-703-766-4, pp.450-456.

[6] **Gomez-Rubio (2008),** ***Tutorial on small area estimation***, UseR conference 2008, August 12-14, Technische Universitat Dortmund, Germany.

[7] **Jula, N. and Jula, D., (2009),** ***Modelare economica. Modele econometrice si de optimizare.***, Mustang Publisher, ISBN 978-606-8058-14-6

[8] **Michele D'Al ´o, Loredana Di Consiglio, Stefano Falorsi, Fabrizio Solari,** ***Course on Small Area Estimation***, ESSnet Project on SAE, Small area estimation

[9] **R Development Core Team (2005). R:** ***A language and environment for statistical computing***. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

[10] **Rao, J.N.K. (2003),** ***Small area estimation*****.**

[11] **Sarndal, C. (1984),** ***Design-consistent versus model-dependent estimation for small domains***, Journal of the American Statistical Association, JSTOR, 624-631

[12] **Schoch, T. (2011), rsae:** ***Robust Small Area Estimation. R package version 0.1-3***.

[13] **Voineagu, V., Pisica, S., Caragea, N., (2012)** ***Forecasting Monthly Unemployment by Econometric Smoothing Techniques****, http://www.ecocyb.ase.ro/32012/Virgil%20Voineagu.pdf*