Measuring change of poverty estimates on small area level

Jan Pablo Burgard, Ralf Münnich and Stefan Zins
University of Trier, Economic and Social Statistics Department,
Universitätsring 15, 54296 Trier, Germany
Corresponding author: Jan Pablo Burgard, e-mail: JPBurgard@uni-trier.de

Abstract

Recent developments in small area estimation methods facilitate the production of poverty and inequality indicator estimates at local level. This information is highly demanded by scientist from a wide variety of research fields like social sciences and economics. And so it is increasingly requested by politicians for evidence based politics.

Even more important is measuring the evolution of an indicator over time. On national level, classical design-based estimators perform well. However, on local level the classical methods may perform poorly due to very low sample sizes. Furthermore, estimates of change often rely on rotational samples. This imposes difficulties for measuring the variance of the change due to a possibly extreme small number of overlapping observations in certain areas. Hence, almost no reliable information may be deduced there. In contrast to classical methods, small area methods may cope much better with small sample sizes and small overlapping sample sizes. Thus, small area methods may foster the production of reliable estimates on local level.

In this paper we present a small area approach for measuring the change over time of poverty and inequality indicators in the presence of rotational samples. This approach will be compared with classical design-based estimators. This comparison is done via a wide scale Monte-Carlo study relying on a realistic data set. The focus lies on assessing the performance of the classical and small area estimates of change under complex survey designs.

Keywords: Small Area Estimation, Estimation of Change, Rotational Samples, Poverty and Inequality Indicators

1 Indroduction

The assessment of poverty is more and more in the focus of the European Union and its member states. In recent years several large scale EU framework program (FRP) were conducted addressing the estimation of poverty and inequality indicators. For example in the 7. FRP the research project AMELI (`http://ameli.surveystatistics.net`) and SAMPLE (`http://www.sample-project.eu/`) were financially supported. The core indicator set used here fore are the Laeken-indicators. One of the most used Laeken indicators is the at-risk-of-poverty rate (ARPR), which is the proportion of persons with an income lower or equal to 60% of the median income of the population. Besides the mere indicator estimate for one point in time it is of fundamental importance to gain knowledge over the evolution of the indicators.

For reasons of clarity and space restrictions the focus will lie on the estimation of the ARPR and the estimation of the change of the ARPR over time. After introducing rotational design, the design-based and model-based estimators for the

estimation of the ARPR will be presented. Then the estimation of the evolution of the ARPR will be discussed.

2 Rotational Designs

To analyze the change in indicator values over time, surveys are needed that are repeated within the observation period $\mathcal{T} = \{1, \ldots, T\}$. At time t a sample is selected from $\mathcal{U}^t$, where $\mathcal{U}^t$ denotes the sampling frame at time t and $N^t = |\mathcal{U}^t|$ the population size at time t. Further, $\mathcal{U} = \cup_{t=1}^{T} \mathcal{U}^t$ with $N = |\mathcal{U}|$, whereby $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$.

In most cases repeated surveys involve sample co-ordination to implement a certain longitudinal design of the study. A rotational design can be considered as mediation between two scenarios, sampling independently at each occasion and a panel survey. In most rotational designs the repeated samples are co-ordinated in such a way that there is a systematic pattern for units to repetitively enter and leave the current sample. Sample co-ordination induces a stochastic dependency, usually correlation, between samples that affects the analysis of change, especially when estimating the variance of a measured change. Following the notation of Nedyalkova et al. [2009] for sampling at several occasions we have to distinguish between the two kinds of samples without replacement: $\mathbf{s}^t = (I_1^t, \ldots, I_k^t, \ldots, I_N^t)', \ \forall\, t \in \mathcal{T}$, is called a *cross-sectional sample* of size $n^t = \sum_{k=1}^{N} I_k^t$, $\mathbf{s}_k = (I_k^1, \ldots, I_k^t, \ldots, I_k^T)', \ \forall\, k \in \mathcal{U}$, is called a *longitudinal sample* of size $n_k = \sum_{t=1}^{T} I_k^t$ and $\mathbf{s} = (\mathbf{s}^1, \ldots, \mathbf{s}^t, \ldots, \mathbf{s}^T)'$ is called a *joint sample*. Where $I_k^t = 1$ if element k is selected into the sample at time t and $I_k^1 = 0$ else. Further we define the sampling designs [Tillé, 2006, Cap. 2] $p^t(\cdot)$ on support $\mathcal{S}^t$ as the *cross-sectional sampling design* at time t, $p_k(\cdot)$ on support $\mathcal{S}_k$ as the *longitudinal sampling design* of element k, and $p(\cdot)$ on support $\mathcal{S}$ as the *joint sampling design*. Where $\mathcal{S}^t$ is the set of all cross-sectional samples in t, $\mathcal{S}_k$ the set of all longitudinal samples for element k, and $\mathcal{S}$ the set of all joint samples. The designs produce the following inclusion probabilities: $\pi_k^t = \sum_{\mathbf{s}^t \in \mathcal{S}^t} I_k^t p^t(\mathbf{s}^t)$, $\pi_{kl}^t = \sum_{\mathbf{s}^t \in \mathcal{S}^t} I_k^t I_l^t p^t(\mathbf{s}^t)$, $\pi_k^{tu} = \sum_{\mathbf{s}_k \in \mathcal{S}_k} I_k^t I_k^u p_k(\mathbf{s}_k)$, and $\pi_{kl}^{tu} = \sum_{\mathbf{s} \in \mathcal{S}} I_k^t I_l^u p(\mathbf{s})$.

Systematic designs Tillé [2006, § 7.1] are in some ways particularly suitable to serve a longitudinal design, as their underling implicit stratification [see Brewer, 2003], results in more equally distributed selection over $\mathcal{T}$ and given its last selection a unit can be told that it will not be contacted again of a given time period.

For the purpose of small area estimation we take that the population is partitioned into D sub-populations, called areas (or domains). Let therefore $\mathcal{U}^t = \dot{\bigcup}_{d \in \{1..D\}} \mathcal{U}_d^t$ be the sampling frame at time t, where $\mathcal{U}_d^t$ is the sampling frame at time t in area d. Further let $N_d^t = |\mathcal{U}_d^t|$ and therefore $N^t = \sum_{d=1}^{D} N_d^t$. The variable of interest is the income (usually the equalized disposable household income) in area d, $\mathbf{y}_d^t = (y_{1d}^t, \ldots, y_{kd}^t, \ldots, y_{N_d^t d}^t)'$, where y_{kd}^t denotes the income of person k in area d at time t.

In the context of small area estimation a rotation design can now become problematic if one is interested in estimating change. Because if the rotation does not take place separately within areas, which will be the case if one estimates unplanned areas, we may have $Pr(\mathbf{s}_d^{t\,'} \mathbf{s}_d^u = 0) > 0$, with $\mathbf{s}_d^t$ as the cross-sectional sample in area

d at time t of size n_d^t. Although $\mathbf{s}_d^t$ and $\mathbf{s}_d^u$ are correlated taking into account a possible dependency between y_{id}^t and y_{id}^u for producing design based point and variance estimates of change in area d may not be possible as no tuple (y_{kd}^t, y_{kd}^u) has been observed.

In contrast to the design-based approach, a model-based approach may use the information form other areas in order to overcome this problem. The classical approach in model-based small area estimation is to build a model, containing all information available in the sample. Therefore, also predictions for unplanned areas or domains are possible, where effectively no common sample has been drawn.

3 Design-based and Model-based Estimation of the At-Risk-of-Poverty Rate at Local Level

The ARPR in area d at point in time t is then defined by

$$\text{(1)} \quad \text{ARPR}_d^t := \frac{1}{N_d^t} \sum_{i \in \mathcal{U}_d^t} \mathbb{I}(y_{id}^t < PT^t) \quad ,$$

where PT^t represents the poverty threshold (PT) of the population at time t, which is defined as $PT^t = 0.6F_{y^t}^{-1}(0.5)$, where F_{y^t} is the distribution functions of the income at time t. A design based estimator for the ARPR is given by

$$\text{(2)} \quad \widehat{\text{ARPR}}_d^t = \hat{F}_{y_d^t}(\widehat{PT}^t) \, ,$$

where $\hat{F}_{y_d^t}(x) = \sum_{k \in s_d^t} w_k^t \mathbb{1}(y_k^t \leq x)(\sum_{k \in s_d^t} w_k^t)^{-1}$, $\widehat{PT}^t = 0.6\hat{F}_{y_d^t}^{-1}(0.5)$ with $\hat{F}_{y_d^t}^{-1}(p) = \inf\left\{x \in \mathbb{R} : p \leq \hat{F}_{y_d^t}(x)\right\}$, and $w_k^t = {\pi_k^t}^{-1}$ is the survey weight of person k at time t. s_d^t denotes the set of sampled units in area d for the point in time t. A variance estimator for (2) can obtained by linearization, more precisely through the values of the influence function of (1). Deville [1999] gave some practical rules to derive influence functions for varied functionals, which can be used to derive it for (1). Then the asymptotic variance of $\widehat{\text{ARPR}}_d^t$ can be estimated by

$$\text{(3)} \quad \widehat{\text{V}}(\sum_{k=1}^{N} I_k^t w_k^t \hat{z}_k^t) = \sum_{k \in s_d^t} \sum_{l \in s_d^t} \frac{\pi_{kl}^t - \pi_k^t \pi_l^t}{\pi_{kl}^t} \frac{\hat{z}_k^t}{\pi_k^t} \frac{\hat{z}_l^t}{\pi_l^t} \, ,$$

where $\hat{z}_k^t$ are the estimated values of the influence function of ARPR_d^t at point y_k^t.

Two approaches for the estimation of the ARPR in the field of small area estimation are discussed. The first approach models the probability of being at-risk-of-poverty via a logistic generalized regression estimator (LGREG) [Lehtonen and Veijanen, 2009]. That is, the indicator $z_{id}^t := \mathbb{I}(y_{id}^t \leq PT_d^t)$ is modelled via a logistic regression model:

$$\text{(4)} \quad \begin{aligned} z_{id}^t &\sim \text{Bern}(p_{id}^t) \\ \text{logit}(p_{id}^t) &= x_{id}^t \beta \end{aligned}$$

The estimation of the ARPR is then performed in the fashion of an GREG estimator

$$(5) \quad \text{ARPR}_d^t = \frac{1}{N_d^t}\left[\sum_{i\in\mathcal{U}_d^t}\widehat{p}_{id}^t + \sum_{i\in s_d^t} w_{id}^t(z_{id}^t - \widehat{p}_{id}^t)\right]$$

The second approach is to model the income variable, and then estimate the ARPR via a Monte-Carlo approximation. Molina and Rao [2010] propose to estimate indicators on basis of an empirical best prediction approach. Here fore, the income variable is modelled via a mixed model of the following form for a fixed point in time:

$$(6) \quad y_{id}^t \sim \text{N}(x_{id}^t\beta + u_d^t, \sigma_e^{t,2})$$

with

$$(7) \quad u_d^t \sim \text{N}(0, \sigma_u^{t,2}) \quad .$$

Hereby, x_{id} in (6) is a $1 \times p+1$ row-vector of p covariates with preceding intercept for the person i in area d and u_d denotes the area effect.

Molina and Rao [2010] show, that the empirical best predictor for (1) is given by

$$(8) \quad \widehat{\text{MR}}_d^t = \frac{1}{N}\left(\sum_{i\in s_d^t}\mathbb{I}(y_{id}^t < PT_d^t) + \frac{1}{L}\sum_{l=1}^{L}\sum_{i\in\mathcal{U}_d^t\setminus s_d^t}\mathbb{I}(y_{id}^{*(l)} < PT_d^t)\right) \quad ,$$

where $y_{id}^{*(l)}$ being the l-th from L vector generated by the model

$$(9) \quad y_{id}^{*(l)} = x_{id}^t\beta + u_d^{*(l)} + e_d^{*(l)} \quad .$$

The $u_d^{*(l)} \sim \text{N}(0, \sigma_u^{t,2}(1-\gamma_d^t))$ and the $e_{i,d}^{*(l)} \sim \text{N}(0, \sigma_e^{t,2})$ are the l-th draw from their distributions.

Instead of drawing L random vectors, one can perform a numerical integration e.g. via the Gauß-Hermite quadrature. This is much faster, and more stable for additive indicators as the ARPR.

In the case of very small sample sizes one can simplify the equation (8) by dropping the left part of the sum. Then the approach is simply to model the distribution of the income and calculate the ARPR given the parametric distribution with the estimated parameters. Assuming that the income is e.g. log-normally distributed, one may estimate the parameters of the log-normal distribution form the sample. In analogy to Molina and Rao [2010] one takes the log of the income and models it via a normal mixed model. The ARPR can now be obtained directly from the log-normal distribution by plugging in the estimated parameters, as the scale and location parameters are identical to those from the normal distribution. Alternatively, one may obtain the ARPR from the normal distribution with the same parameters by shifting the poverty threshold by $\log(0.6)$. The ARPR for a predefined poverty threshold PT_d given the estimated distribution of the income is then

$$(10) \quad \text{ARPR}_d^t = \int_{-\infty}^{log(0.6)+PT^t} \frac{1}{\sqrt{2\pi\widehat{\sigma}_d^{t,2}}} e^{-\frac{(x-\widehat{\mu}_d^t)^2}{2\widehat{\sigma}_d^{t,2}}}\, d\,x \quad .$$

This is basically the point $\log(0.6) + PT^t$ of the normal distribution with the estimated parameters $\widehat{\mu}_d^t$ and $\widehat{\sigma}^{t,2}$. This value is easily and fast obtainable from standard statistic software. In contrast to the design-based approach, where in each area enough samples have to be drawn to obtain a reliable estimate, in the case of this small area estimator even with no sample size in an area, a sensible predictor may be produced.

4 Measuring the Change of Poverty Indicators at local level

The change in ARPR_d from time $t'-1$ to t' is given by

$$\Delta^1 \mathrm{ARPR}_d^{t'} = \mathrm{ARPR}_d^{t'} - \mathrm{ARPR}_d^{t'-1} \,. \tag{11}$$

A design based estimator for (11) is given by

$$\Delta^1 \widehat{\mathrm{ARPR}}_d^{t'} = \widehat{\mathrm{ARPR}}_d^{t'} - \widehat{\mathrm{ARPR}}_d^{t'-1} \,. \tag{12}$$

Goga et al. [2009] presented a solution to estimate the variance of (12). They expand the concept of using the influence function of a statistic to approximate its variance to the case of an estimator involving data from different time points collected by rotational samples. Because (12) estimates the change through separate cross-sectional estimates its approximate variance can be estimated by

$$\widehat{V}\Big(\sum_{t=t'-1}^{t'}\sum_{k=1}^{N} I_k^t w_k^t \hat{z}_k^t\Big) = \sum_{t=t'-1}^{t'}\sum_{u=t'-1}^{t'}\sum_{k\in s_d^t}\sum_{l\in s_d^u} \frac{(\pi_{kl}^{tu} - \pi_k^t\pi_l^u)}{\pi_{kl}^{tu}} \frac{\hat{z}_k^t}{\pi_k^t}\frac{\hat{z}_l^u}{\pi_l^u} \,. \tag{13}$$

For extending the idea of Molina and Rao [2010] to the estimation of the difference in time of the ARPR, we assume that the income in the points of time t and u follow a multivariate log-normal distribution. That is,

$$(y_{kd}^t, y_{kd}^u) \sim \log \mathrm{MVN}\left((\mu^t, \mu^u), \Sigma_{tu}\right) \quad . \tag{14}$$

Again we model the log-income via a normal mixed model, accounting this time for the multivariate dependencies. As sample fractions are taken to be very small the proposed small area estimator for the difference of the ARPR is given by

$$\begin{aligned}
\widehat{\Delta^1 \mathrm{ARPR}}_d^{t'} &= \mathrm{ARPR}_d^{t'} - \mathrm{ARPR}_d^{t'-1} \\
&= \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{N_d}\sum_{i\in\mathcal{U}_d}\mathbb{I}(y_i^{t',(l)} < 0.6\cdot \mathrm{T}^{t'}) - \frac{1}{N_d}\sum_{i\in\mathcal{U}_d}\mathbb{I}(y_i^{t'-1,(l)} < 0.6\cdot \mathrm{T}^{t'-1})\right) \\
&= \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{N_d}\sum_{i\in\mathcal{U}_d}\mathbb{I}(y_i^{t',(l)} < 0.6\cdot \mathrm{T}^{t'}) - \mathbb{I}(y_i^{t'-1,(l)} < 0.6\cdot \mathrm{T}^{t'-1})\right)
\end{aligned} \tag{15}$$

where the tuples $(y_i^{t',(l)}, (y^{t'-1,(l)})$ are drawn from the fitted multivariate distribution.

The MSE of the estimator (15) is estimated via a parametric bootstrap [cf. Efron, 1980, Hall and Maiti, 2006, Chatterjee and Lahiri, 2007] which is not further discussed in this short paper due to space constraints.

The concurrent approaches are compared within a large scale simulation study on a realistic data set.

5 Acknowledgements

This research is supported by the *Inclusive Growth Research Infrastructure Diffusion* (InGRID: FP7-INFRA-2012-1.1.1.-316291) and the *Forschungszentrum für Regional- und Umweltstatistik* (forumstat: `www.forumstat.uni-trier.de`).

References

Ken Brewer. *Combined Survey Sampling Inference*. Arnold, New York, 2003.

S. Chatterjee and P. Lahiri. A simple computational method for estimating mean squared prediction error in general small-area model. In *Proceedings of the Section on Survey Research Methods*, pages 3486–3493, 2007.

Jean-Claude Deville. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25(2):193– 203, 1999.

B. Efron. The jackknife the bootstrap and other resampling plans. Technical Report 63, Stanford University, California, December 1980.

C. Goga, J.-C. Deville, and A. Ruiz-Gazen. Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96 (3):691–709, 2009.

Peter Hall and Tapabrata Maiti. On parametric bootstrap methods for small area prediction. *Journal Of The Royal Statistical Society Series B*, 68(2):221–238, 2006.

Risto Lehtonen and Ari Veijanen. Design-based methods of estimation for domains and small areas. In C.R. Rao, editor, *Handbook of Statistics - Sample Surveys: Inference and Analysis*, volume 29, Part 2 of *Handbook of Statistics*, pages 219 – 249. Elsevier, 2009.

Isabel Molina and J. N. K. Rao. Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385, 2010.

Desislava Nedyalkova, Lionel Qualité, and Yves Tillé. General framework for the rotation of units in repeated survey sampling. *Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research*, 63(3):269–293, 2009.

Yves Tillé. *Sampling algorithms*. Springer Series in Statistics. Springer, New York, 2006.