Using Register Information from Multiple Aggregation Levels for the Prediction of Small Area Counts and Means in the Swiss Structural Survey

Burgard, Jan Pablo
E-mail: JPBurgard@uni-trier.de

Münnich, Ralf
E-mail: Muennich@uni-trier.de

University of Trier, Social and Economical Statistics Department
Universitätsring 15
54296 Trier, Germany

Abstract

For the European Census Round 2010, the methodology for Swiss Census has been changed towards a register-assisted census. Within the register-assisted census, the population figures are drawn from the population register. Since the population registers in Switzerland do not contain all information of interest, an additional sample will be drawn, which is the Structural Survey.

When using model-based estimation methods, such as small area estimation, the predictive power of auxiliary information is of major importance. Since registers, in general, contain only demographic variables which have little predictive power for many variables of interest, for instance unemployment or mother tongue, external information sources may be interesting. Further, the aggregation level of modelling may influence the modelling considerably. In this paper, we discuss opportunities to improve the accuracy of small area modelling by using external information sources and different aggregation levels.

Keywords: Small Area Estimation; Aggregated Register Covariates; Binomial Model; Swiss Structural Survey;

The Swiss Structural Survey

From 2010 on, the traditional census in Switzerland is replaced by a register-assisted census. In addition to the population register, the Structural Survey (`http://www.bfs.admin.ch/bfs/portal/en/index/news/02/03/02.html`) will be conducted. Within this survey, a stratified sample of 200.000 persons is drawn in all municipalities proportionally to their size. Within the areas simple random sampling is applied. Because of the low sampling fraction (under 3%) direct design based estimators may yield imprecise results. Further, many of the variables of interest in the survey are binary. As such, the assumptions of classical linear methods may be violated and estimation may be poor. Thus, it is of interest to investigate whether the application of a categorical estimator may be of use. Further, some register information may be available only on aggregated levels, which leads to the question on which level to estimate the small area counts.

In the next Section some estimators will be presented which can be applied in this context. Then, simulation results will be presented for some of the estimators. This paper concludes with a summary and outlook.

Prediction of Small Area Counts

A well known and widely used estimator for small areas and domains is the GREG estimator (cf. Särndal et al, 2003). For the estimation of area totals in small area context, the GREG estimator is given by

$$(1) \quad \widehat{\tau}_{d,\mathrm{GREG}^*} = \sum_{i=1}^{N_d} \widehat{\theta}^*_{id} + \sum_{i=1}^{n_d} w_{id}\big(y_{id} - \widehat{\theta}^*_{id}\big) \quad ,$$

where d denotes the $d'th$ area, i indicates the $i'th$ person and n_d and N_d denote the sample size or the area size respectively. $\widehat{\theta}^*_{id}$ is an estimator of the variable of interest for the person i in area d, and w is the inverse inclusion probability for the sampled unit. In the classical GREG case $\widehat{\theta}^{\mathrm{lin}}_{id} = x_{id}\widehat{\beta}$ with $\widehat{\beta}$ being the solution to a linear regression. For a thorough discussion of the application of linear models in this context we refer to Särndal et al (2003). Lehtonen and Veijanen (2009) propose to use for $\widehat{\theta}^*_{id}$ a non-linear estimator if the dependent variable is categorical. As example they present the LGREG estimator which uses a logistic regression estimator $\widehat{\theta}^{\mathrm{logit}}_{id} = [1 + exp(-x_{id}\widehat{\beta})]^{-1}$ or the MLGREG which uses a logistic mixed model estimator $\widehat{\theta}^{\mathrm{mlogit}}_{id} = [1 + exp(-x_{id}\widehat{\beta} - \widehat{u}_d)]^{-1}$.

Another approach to the estimation of small area totals is proposed by Battese et al (1988) (BHF). The statistical model behind the BHF can be seen as a unit-level mixed model (cf. Jiang and Lahiri, 2006), where, in general, a random intercept model is chosen. The unit-level variation is often referred to as sampling variance, and the area-level variation as area effect.

$$(2) \quad \widehat{\tau}_{d,\mathrm{BHF}} = \sum_{i=1}^{N_d} \big(x_{id}\widehat{\beta} + \widehat{u}_d\big) \quad .$$

An extension to binary data can be found in González-Manteiga et al (2007). The idea behind this is to use a generalized mixed model to predict the probability of a unit to fulfil a certain characteristic of interest, which finally yields

$$(3) \quad \widehat{\tau}_{d,\mathrm{BIN}} = \sum_{i=1}^{N_d} [1 + exp(-x_{id}\widehat{\beta} - \widehat{u}_d)]^{-1} \quad .$$

In contrast to the unit-level estimators stated above, area-level estimators need the covariates only on area-level. Due to its availability, in general, more auxiliary information can be incorporated in the models which likely allows building more powerful models. A well known area-level model is the estimator proposed by Fay and Herriot (1979) (FH). For its relation to the BHF estimator (see Datta and Lahiri, 2000). In the case of the Fay-Harriot model all information is only available on aggregate level:

$$(4) \quad \widehat{\tau}_{d,\mathrm{FH}} = x_d\widehat{\beta} + \widehat{u}_d \quad .$$

For the estimation of means or proportions one can divide the here mentioned estimates by the population total provided by the registers.

Estimation and Prediction on Different Aggregation Levels

In figure 1 the RRMSE of the point estimates over 1000 samples are shown for the classical GREG, the BHF and the BIN estimators. The red colour indicates that the RRMSE is above 25%. As can be seen, the GREG estimator behaves overall considerably well, except in Tessin and its surroundings. The BHF model performs slightly better in Tessin, and in many areas it reaches a considerably smaller RRMSE than the GREG estimator. However, also the BHF does not reach

good results in some areas. When estimating on municipality-level even the BIN estimator yields relatively high RRMSE, in general a higher RRMSE than the BHF. In some areas it yields poor results. Nevertheless, it performs better than the classical GREG estimate.

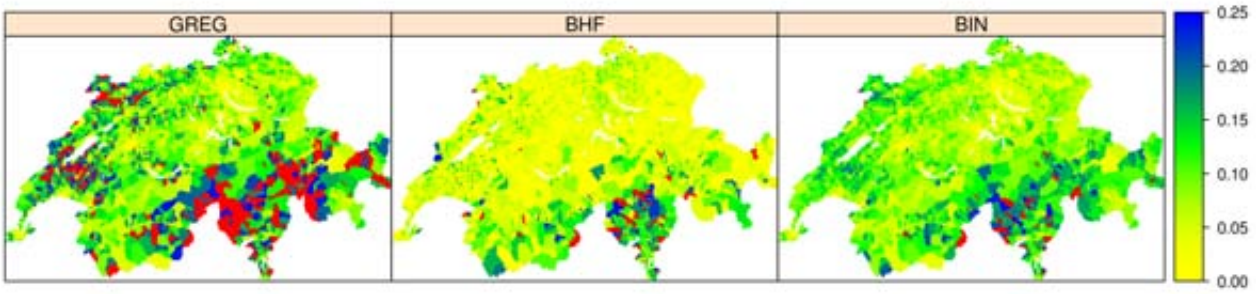

Figure 1: RRMSE of area estimates for the total of working population on municipality-level

In figure 2 municipalities smaller than 2000 inhabitants are merged with each other or a neighbouring municipality in order to construct areas which have at least 2000 inhabitants. All three estimators gain a lot from this aggregation of small municipalities. Again, the BHF model reaches for a large amount of areas the lowest RRMSE. But still the results for the areas in the south of Switzerland are partly quite high. It seems that the model does not hold for these areas. Even though the GREG does not get best results for many areas, it also only has one area with an unacceptable result. Also the BIN has for most areas a higher RRMSE than the BHF. But, in contrast to the BHF, it does not have any area with a very high RRMSE. Further, it outperforms also the GREG clearly. Thus, the aggregation of municipalities to areas of at least 2000 inhabitants improves estimates very much. Further, especially the GREG and BIN gain from this aggregation.

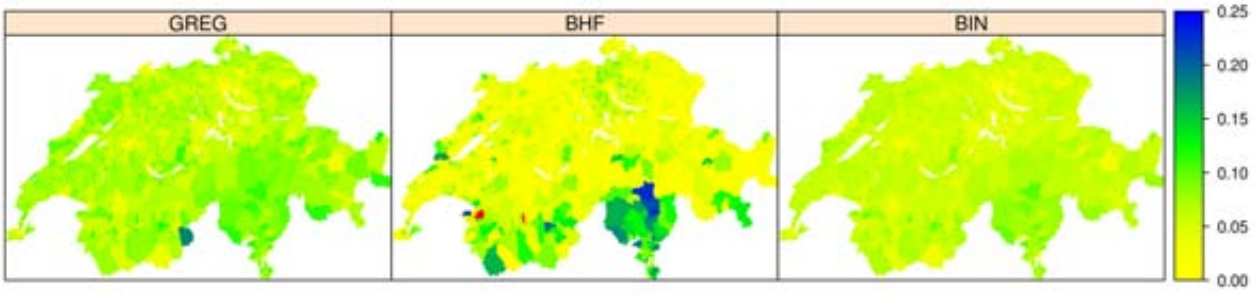

Figure 2: RRMSE of area estimates for the total of working population on aggregations of municipalities to areas with minimum size of 2000

For many variables of interest the registers do not provide good auxiliary variables. In this case aggregate information may be available, e.g. from other registers, which can not be used in disaggregated form by reason of disclosure. One way to use this information is to use the FH estimator. Further, a mixture of area-level and unit-level auxiliary variables may lead to improved results. In this case one approach is to replicate the aggregate auxiliary variable on unit-level length.

Here, the problem of the ecological fallacy stated by Robinson (1950) has to be considered, which indicates that the correlation of aggregate data does not imply the same correlation on unit-level. Anyhow, the interest here does not lie in making an inference about the correlation, but rather to use every correlation available for prediction purposes.

A major drawback of using aggregate information is the fact that area-level estimators suffer

much more from register errors than unit-level estimators. For a thorough simulation study in this context see Burgard and Münnich (2010).

Summary and Outlook

The aggregation level on which estimates are produced is of major importance also when using small area models. As could be seen, it depends on the aggregation of the areas whixh estimators is recommendet. Small area models outperform the GREG estimator in terms of RRMSE in this setting. With the BHF one could gain good results also in areas where the estimates from the GREG were rather poor. If areas are not too small the BIN estimator yields good results. However, unit-level binomial models are very computer-intensive, and sometimes run into convergence problems. In the Swiss Structural Survey the design weights do not play a major role as the design used is a proportional allocation design. For the influence of design weights on small area estimation see Münnich and Burgard (2011).

Within a large design based Monte Carlo simulation study on the full Swiss Census 2000 data set, the above mentioned small area estimators are compared under several scenarios. These include the estimation on different aggregation levels and the incorporation of aggregated covariates into the models.

Acknowledgements

This research is part of the research project *Simulation of the Structural Survey* which is financially supported by the Swiss Federal Statistical Office (SFSO). The authors thank Monika Ferster (SFSO) and Professor Partha Lahiri, University of Maryland and Joint Porgram of Survey Methodology, for very inspiring discussions.

References

Battese GE, Harter RM, Fuller WA (1988) An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83 (401):28 – 36

Burgard JP, Münnich RT (2010) Modelling over and undercounts for design-based monte carlo studies in small area estimation: An application to the german register-assisted census. Computational Statistics & Data Analysis In Press, Corrected Proof

Datta GS, Lahiri P (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10:613–627

Fay RE, Herriot RA (1979) Estimates of income for small places: An application of james-stein procedures to census data. Journal of the American Statistical Association Vol. 74, No. 366:269–277

González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2007) Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. Comput Stat Data Anal 51:2720–2733

Jiang J, Lahiri P (2006) Mixed model prediction and small area estimation. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research 15(1):1–96

Lehtonen R, Veijanen A (2009) Design-based methods of estimation for domains and small areas. In: Rao C (ed) Handbook of Statistics - Sample Surveys: Inference and Analysis, Handbook of Statistics, vol 29, Part 2, Elsevier, pp 219 – 249

Münnich R, Burgard JP (2011) On the influence of sampling design on small area estimates submitted

Robinson WS (1950) Ecological Correlations and the Behavior of Individuals. American Sociological Review 15(3):351–357

Särndal CE, Swensson B, Wretman J (2003) Model Assisted Survey Sampling. Springer