STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 603–610

SAE TEACHING USING SIMULATIONS

Jan Pablo Burgard[1], **Ralf Münnich**[2]

ABSTRACT

The increasing interest in applying small area estimation methods urges the needs for training in small area estimation. To better understand the behaviour of small area estimators in practice, simulations are a feasible way for evaluating and teaching properties of the estimators of interest. By designing such simulation studies, students gain a deeper understanding of small area estimation methods. Thus, we encourage to use appropriate simulations as an additional interactive tool in teaching small area estimation methods.

Key words: small area estimation, teaching, simulations, design-based simulations, model-based simulations.

1. Challenges in Teaching SAE

Small area estimation (SAE) methods are becoming increasingly valuable for both methodologists and practitioners, and are used quite regularly in the production of official statistics. The last two decades have witnessed an explosion of small area estimation methods. However, the advances are mostly in the theoretical field, and practitioners still lack adequate knowledge of all the advancements in SAE methodology.

The classical way to present the benefits and drawbacks of SAE methods is using slides. Graphs and tables are used for illustration, and often simulation results are presented on the slides as well. From experience, however, for many students the understanding which estimator is preferably applicable is still lacking. These students, in order to obtain a good result in the exam, will memorize mainly the advantages and disadvantages of the respective methods. This is certainly not the didactic goal, and holds the further restraint that many of them are not able to transfer their knowledge to new methods developed later on.

We emphasize using simulations as an interactive tool to teach SAE methods. A large list of literature exists concerning the use of computers in statistical classes (McKenzie, 1992) and some papers focus directly on the use of simulations

[1] Trier University. E-mail: burgardj@uni-trier.de.
[2] Trier University. E-mail: muennich@uni-trier.de.

(Kalsbeek, 1996, Hesterberg, 1998, DelMas et al., 1999). Hesterberg (1998) describes simulations as follows:

> The basic idea in simulation is to emulate real life, where one collects a sample of random data (using a survey or an experiment), and summarizes the data graphically or numerically. In simulation one generates a sample of random data on the computer in a way that mimics a real problem and summarizes that sample in the same way. However, instead of doing this only once, one may do it many times, to investigate how much summaries vary.

In the context of statistical education Mills (2003) states that

> Regardless of how clearly a teacher explains a concept, students will understand the material only after they have constructed their own meaning for the new concepts, which may require restructuring and reorganizing new knowledge and linking it to prior or previous knowledge.

Further he points out that

> [...] meaning is acquired through a significant interaction with new knowledge.

An educational concept for teaching mathematics and statistics that addresses these aspects is *discovery learning* proposed, e.g., by Bruner (1961). The key idea is to provide students with materials needed to solve the imposed questions - rather than providing simply their solutions. However, as Mayer (2004) points out, an unguided form of discovery learning is not recommendable. Kirschner et al. (2006) state that the learners need guidance to reach a certain level of knowledge, from which point on they can increasingly learn from discovery. In an empirical evaluation of different teaching methods Alfieri et al. (2011) find that *Enhanced Discovery Learning* shows to have a positive effect on learning. In enhanced discovery learning, the teacher accompanies the discovery process by instructional guidance, or feedback or other merits. In our view, simulations provide a platform for such enhanced discovery learning with a built-in feedback system.

In the following section it is discussed how simulations can be used in the special context of SAE to support the students in the process of understanding the merits of the different methods at hand. In Section three, an example simulation used in graduate classes is provided. We conclude with a summary and outlook.

2. The Use of Simulations for Teaching SAE

In SAE, two major types of simulations can be considered, design-based and model-based simulations (for a more detailed discussion, see e.g. Burgard, 2013, and Münnich, 2014).

In model-based simulations random samples from a superpopulation model are drawn. The methods of interest are then applied to these random samples. This is an effective procedure to check particularly whether (a) under optimal

conditions, that is when all model assumptions hold, a method yields the theoretically expected results and (b) a method is programmed correctly. Usually, it is far more sophisticated to derive a real world behaviour in the context of survey statistics. As Graham Kalton stated in Malay Ghosh's honorary symposium in 2014

> In case we want to apply small area methods in official statistics, we have to consider the sampling design.

In design-based simulations the random samples are drawn according to a sampling design from a fixed finite population. It is basically an attempt to reproduce the true survey process of interest. A major emphasis has to be laid on a realistic population that mimics all important characteristics of the real population. This realistic population could be for example an older version of the actual population. The design-based simulation then is useful for comparing different methods on their applicability in a certain survey context with regards to the sampling design.

Thus, when teaching SAE methods, model-based simulations are a good starting point to study the properties of SAE. However, for studying real world behaviour, the design-based simulation approach seems considerably more appropriate for applications, at least for official statistics.

As the field of SAE encompasses several statistical disciplines and applications, there are multiple decision criteria to acknowledge for when choosing appropriate methods. Some central but non-exhaustive aspects to consider are the classical statistical properties, user acceptance, as well as computational complexity and stability. Performing simulations in either way helps to understand advantages and disadvantages of the statistical methods given the relevant decision criteria, e.g. triple-goal (Shen and Louis, 1998), and further enables the students to evaluate new methods later on their own.

For most estimators in SAE, classical statistical properties are proven. These are generally based on asymptotic theory, regarding sample size, or the number of areas or domains. Both asymptotic arguments, however, have to be used carefully in SAE, as the typical setting is a small sample size and a finite number of areas (Pfeffermann, 2006). By varying the sample sizes within a simulation, the effect of small sample sizes or small number of small areas can be visualized. An example will be given in the next section.

An important hurdle is the acceptance of the published small area estimates by data users. This argument is specifically important in official statistics, where the users of the published data are not necessarily proficient in SAE. In practice, one major reservation against many small area estimators is that they are not design unbiased. However, as design unbiasedness and small variance of small area estimators are usually antagonists, the demand for design unbiasedness may better be dropped in favor of reducing the mse of the estimators. This can be visualized by using simulations.

SAE methods are often computationally very complex. Computation times may be prohibitive for too large data sets, and computational stability may depend severely on the data structure. Hence, the computability of many programs depends on the present sample. Using simulations, in general, a large set of different samples is provided and applied. Since many computer codes may fail in single samples, the simulation yields a realistic view on possible computational issues. Those *special* samples can be analysed into more detail which might lead to a reformulation of the estimator or an improvement of the computer program.

Additionally, in order to tackle in depth the before mentioned specific issues, simulations are a useful tool in the lecture to recapitulate the learned materials.

3. An Example for Using a Simulation in SAE Teaching

In general, when teaching SAE we start with the presentation of a new estimator and describe its statistical properties. Within the next step, students shall generate a superpopulation that fulfills all the assumptions of this estimator. The teacher accompanies the process of finding an appropriate superpopulation by asking supporting questions. By gradually deviating from the *optimal* superpoplation that fulfills all model assumptions of the estimator, the impact from deviations on the performance of an estimator can be observed.

Design-based estimation methods such as the direct estimator (Cochran, 2007, p. 21 et seqq.) rely on asymptotic arguments, and have good performance in large sample settings. Their performance, measured in terms of accuracy, is indirect proportional to the sample size. However, the sample size tends to be very small in SAE applications (Rao, 2003, p. 1). The following example simulation will tackle the following questions in this context. How do small sample sizes affect the outcome of direct estimators? Are there sample sizes under which we should prefer SAE methods to design-based methods? How much can we gain from using model-assisted and model-based estimation?

The students are asked to generate a superpopulation which shows the advantages of model-assisted and model-based estimation over the direct estimator without auxiliary variables. The discussion generally leads to the idea that the correlation between the dependent variable and the covariates, the ratio of between area variation and residual error, as well as the sample size will have an impact on the outcome of the different estimators.

The estimators of interest are the direct estimator without auxiliary information, the model-assisted direct estimator *GREG* (Särndal et al., 1992, §6.4), and the model-based Battese-Harter-Fuller estimator (BHF, Battese et al., 1988). From the viewpoint of official statistics, this may be seen as *from design towards model-based methods* (cf. Münnich et al., 2013). Holding the residual error constant, the superpopulation for a model-based simulation can be constructed with

- one dependent variable y as linear function of the realizations x of an arbitrary random variable X with

- normally distributed unit level error terms e with $\mathrm{E}(e) = 0$ and $\mathrm{Var}(e) = \sigma_e^2$,
- and normally distributed area level error terms u with $\mathrm{E}(u) = 0$ and $\mathrm{Var}(u) = \sigma_u^2$.

The resulting settings are as follows

- Setting 1: lower σ_u^2 lower cor(y,x)
- Setting 2: higher σ_u^2 lower cor(y,x)
- Setting 3: lower σ_u^2 higher cor(y,x)
- Setting 4: higher σ_u^2 higher cor(y,x)

By assuming higher and lower values for both, the correlation between y and x and for σ_u^2, the magnitude of the gain in efficiency of one estimator over the others can be visualized. As can be seen from Figure 1, the improvement of using the model-assisted as well as the model-based estimator over the direct estimator is the larger the higher the correlation between y and x. Additionally, the smaller the variance σ_u^2, and therefore the smaller the ratio $\frac{\sigma_u^2}{\sigma_e^2}$, the higher is the improvement over the direct estimator. Further, it becomes apparent that in the case of rather small sample sizes (n=4) the improvement of using the model-assisted and model-based estimators over the direct estimator without auxiliary covariates is larger than in the case of n=40. Especially the gain from using the BHF over using the GREG is more pronounced in the case of low sample sizes (n=4).

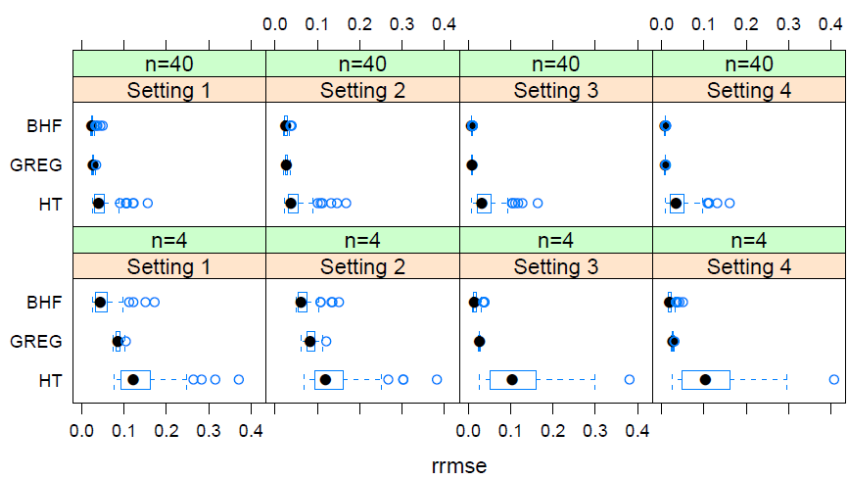

Figure 1. Rrmse of the estimators in the settings 1–4

Another perspective on the performance of estimators rather than looking at the rrmse, which is more convincing to many practitioners, is to look at the Monte-Carlo probability of lying within an acceptable interval. Such an acceptable interval can be defined as an interval in which the estimates should at least lie in. For instance, in Figure 2 an absolute distance of 1 from the true value is defined as acceptable. The Monte-Carlo probability of lying within the interval is then simply the rate of samples with successes within the Monte-Carlo simulation. The gain of using a model to not using auxiliary variables is immense. However, if sample size is larger, the gain from using the BHF over the GREG is not that pronounced as in the case of low sample sizes (n=4).

Certainly, in this context a considerable number of measures and their impact on the selection of adequate estimators can be investigated via simulations, which furnishes a better understanding of the entire methodology.

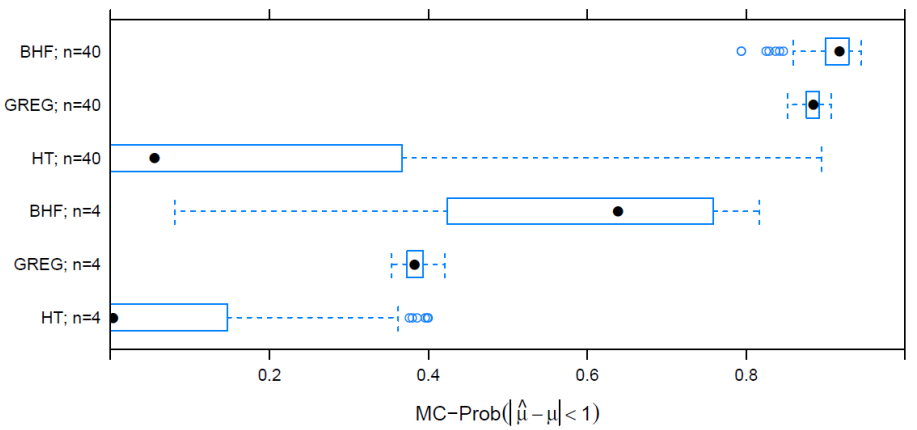

Figure 2. Monte-Carlo probability of lying within an acceptable interval in setting 3

4. Summary and Outlook

Teaching SAE methods covering both theory and applications is a challenging task. Students attending SAE classes rarely have a strong statistical education background with experience in applications. In this context we are convinced that the above presented approach of using simulation for teaching SAE methods is a very useful additional tool in teaching SAE. It provides a better and more sustainable understanding of applying and choosing appropriate SAE methods.

Acknowledgements

We thank Risto Lehtonen for inviting us to the Teaching SAE session at the SAE 2014 in Poznan. We also thank the audience for the feedback and the vivid

discussion after the presentation and throughout the conference, as well as the editor and an anonymous reviewer for very valuable comments that helped to improve the clarity of this paper considerably.

REFERENCES

ALFIERI, L., BROOKS, P. J., ALDRICH, N. J., TENENBAUM, H. R., (2011). Does discovery- based instruction enhance learning? Journal of Educational Psychology, 103(1): 1.

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(401): 28–36.

BRUNER, J. S., (1961). The act of discovery. Harvard educational review.

BURGARD, J. P. (2013). Evaluation of Small Area Techniques for Applications in Official Statistics. PhD Dissertation, Universität Trier.

COCHRAN, W. G., (2007). Sampling Techniques. Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley & Sons, New York, ISBN 9780471162407.

DELMAS, R. C., GARFIELD, J., CHANCE, B., (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. Journal of Statistics Education, 7(3).

HESTERBERG, T. C., (1998). Simulation and bootstrapping for teaching statistics. In American Statistical Association Proceedings of the Section on Statistical Education, pages 44–52.

KALSBEEK, W., (1996). The computer program called sample: A teaching tool to demonstrate some basic concepts of sampling (version 1.01). In American Statistical Association Proceedings of the Section on Statistical Education, Volume 103.

KIRSCHNER, P. A., SWELLER, J., CLARK, R. E., (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist, 41(2): 75–86.

MAYER, R. E., (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The Case for Guided Methods of Instruction. American Psychologist, 59(1): 14–19, Jan. 2004. ISSN 0003066X. doi: 10.1037/0003-066X.59.1.14.

MCKENZIE J. D., (1992). Why aren't computers used more in our courses. In Proceedings of the Section on Statistical Education, pages 12–17. The Association.

MILLS, J. D., (2003). A theoretical framework for teaching statistics. Teaching Statistics, 25 (2): 56–58.

MÜNNICH, R., (2014). Small area applications: some remarks from a design-based view. SAE 2014-Conference on Small Area Estimation in Poznan, http://sae2014.ue.poznan.pl/presentations/SAE2014_Ralf_Munnich_c330a31c0a.pdf.

MÜNNICH, R. T., BURGARD, J. P., VOGT, M., (2013). Small Area-Statistik: Methoden und Anwendungen. AStA Wirtschafts- und Sozialstatistisches Archiv, 6(3): 149–191.

PFEFFERMANN, D., (2006). Invited discussion of paper by J. Jiang and P. Lahiri: Mixed model prediction and small area estimation. TEST, 15: 65–72, URL http:// eprints.soton.ac.uk/38527/.

RAO, J. N. K., (2003). Small Area Estimation. Wiley series in survey methodology. John Wiley and Sons, New York.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN J., (1992). Model Assisted Survey Sampling. Springer, New York.

SHEN, W., LOUIS, T., (1998). Triple-goal estimates in two-stage hierarchical models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2): 455–471.