Model based estimation at Statistics Netherlands

Bart Buelens[1], Peter-Paul de Wolf[2], Kees Zeelenberg[3]

[1] *Statistics Netherlands, Heerlen, Netherlands; b.buelens@cbs.nl*
[2] *Statistics Netherlands, The Hague, Netherlands; pp.dewolf@cbs.nl*
[3] *Statistics Netherlands, The Hague, Netherlands; k.zeelenberg@cbs.nl*

Abstract
In official statistics, models are sometimes used when there enough data for using design-based techniques are available. In those cases, models can provide alternative, in some sense, better estimates. National statistical institutes (NSIs) have always been reluctant to use models, apart from specific cases. Based on the experience at Statistics Netherlands we argue that NSIs should not be afraid to use models. In order to aid statisticians at Statistics Netherlands in the use of models, we have set up a number of guidelines based on the principles of objectivity and reliability of the European Code of Practice, and the idea that the primary purpose of an NSI is to describe, and not to prescribe or judge. In particular we require that the use of models is documented, made transparent to users, and refrains from making forecasts; also the models should rely only on actually observed data and they should be validated extensively.
After a description of these guidelines, we discuss how several examples, such as small-area estimates, seasonal adjustment, capture-recapture for hidden populations, corrections of mode effects with mixed-mode data collection, and corrections for non-response, fit into the guidelines. We also discuss some cases that do not seem to fit the guidelines.
We also look at the relatively new area of big data in official statistics. Here problems of coverage and selectivity may be severe, and we discuss how models may be used to produce more accurate statistical information. We also discuss some examples where modeling of big data was not successful.

Keywords: models, official statistics, big data.

1. Introduction

At Statistics Netherlands there is a tendency to increase the use of models in producing official statistics. This has stimulated the need for a policy concerning the use of models. The current document wants to provide some guidelines for producing statistics with the aid of models.

Firstly, we will explain what we mean with model based estimation. Secondly, some guidelines will be presented that one should follow when using models in producing official statistics.

2. Model based estimation

With "model based estimation methods," we mean a family of methods where statistical models are fundamental to the estimation method (Van den Brakel en Bethlehem, 2008; Van den Brakel, 2012). In classic sample surveys, design based estimation is more often used as opposed to model based estimation. Knowledge about the sampling design usually suffices to construct design unbiased estimators. This is also the main way estimators are constructed at Statistics Netherlands.

In model based estimation, a model is usually constructed where some dependent variables are expressed as a function of some independent variables. Under certain assumptions, the model could then be used to estimate missing values of the dependent variables. This can be done at the micro level as well as on an aggregated level. In case the data are obtained using a sampling design, the design could be included in the model to account for some consequences of the design, just like in design based estimation. For example, in case there is some selectivity to be expected, the design variables and the sampling probabilities could be included in the model.

Models are not completely excluded from design based estimation. Sometimes models are used to take care of specific problems, like imputation of missing values, seasonal adjustment or adjusting for selectivity of non-response.[1] The properties then usually still resemble those of pure design based estimators. This contrary to the situation of model based estimators, where e.g., misspecification of the model will lead to biased estimates.

[1] When models are used to correct for non-response in design based estimation, this is usually called *model-assisted* estimation.

Under certain circumstances, design based estimators may give rise to problems. For example, when estimates on small subpopulations are needed, the sample size is often too small to achieve estimates that are accurate enough using the traditional design based estimators. Small area estimation is a model based method that can increase the accuracy[2] of the estimates by making use of information on other, related, subpopulations. This technique is used at Statistics Netherlands in producing e.g., the monthly unemployment rate (based on LFS) and in producing regional estimates according to the Key Subject Regionalization (based on LFS, HEALTH and CVS).[3] Moreover, models are used when modifications or redesigns of production processes lead to discontinuities (e.g., changes in trends), to compensate for those discontinuities.

3. Guidelines related to model based estimation

Models are used in situations where not all the data that are needed for obtaining (accurate) estimates by means of design-based techniques are available. The models are then used to estimate the missing data. A model is usually characterized by a number of parameters. Those parameters are estimated using some available data; this is called identifying and *fitting* the model. To that end, often data is used that is related, but unequal to the variable that has to be estimated. The fitted model is then used in the estimation of the variable of interest.

In the following, we will provide some guidelines that can be used to evaluate the use of model based estimation in official statistics. The guidelines can be divided into two groups: the first group deals with the question when a model based approach is valid (general principle), the second group deals with the question how to assess and account for the use of model based estimation.

---

[2] The variance of model based estimators is usually calculated conditionally on the model. I.e., the variance is calculated under the assumption that the model is valid and applicable.

[3] LFS = Labour Force Survey = EBB = Enquête Beroepsbevolking, HEALTH = Health Survey = GEZO = Gezondheidsenquête, CVS = Crime Victimization Survey = IVM = Integrale Veiligheidsmonitor.

3.1. General principle

The general principle when using model based estimation in official statistics, is the principle that official statistics give a description of society as it is (or as it has been). This is reflected in the explanatory memorandum of the Dutch Statistics Act which says: "… statistics provide an objective view of the social and economic reality". Moreover, the European regulation on statistics (Regulation (EC) No 223/2009) has as one of its principles: "*reliability*: … statistics must measure as faithfully, accurately and consistently as possible the reality that they are designed to represent" and the European Code of Practice has the principle: "*accuracy and reliability*: European Statistics accurately and reliably portray reality."

When using model based estimation, we will use the notion of objectivity such that the data used to estimate the model, should be related to the subject of the statistic of interest. That is, both the entities and the populations related to the model should reflect those of the statistical phenomenon in question. Moreover, the model should only be used for the time period for which the data are available (i.e., no forecasting). Data from the past can be used to identify (or fit) the model, but estimation should not exceed the present (i.e., at most nowcasting). Note that extrapolation and forecasting in the Netherlands is a task of governmental bureaus like SCP (Netherlands institute for social research) and CPB (Netherlands bureau for economic policy analysis).

We will use the principle of reliability such that a failure of the model (e.g., misspecification) should not lead to changes in the (conclusions based on the) estimate of the statistical phenomenon. Hence, the estimation procedure should be robust with respect to model-failure. Especially in case of time series this is an important point: model failure could lead to situations where turning points in the time series are found at the wrong places.

So far we have spoken about model based estimation in relation to "end-products". Obviously, model based estimation could just as well be used to estimate "intermediate products" that in themselves are not published. Whether model based estimation is a valid approach in those cases, depends on the impact those intermediate products may have on the end-products that will be published.

3.2. The use of models

In case model based estimation is used to produce official statistics, the following issues should be taken into account:

1. **Goal.** The goal of using model based estimation should be to estimate data that is not available, and as such to improve the overall estimation process. The general principle as discussed in 3.1 should be fulfilled.
2. **Data**. Models are used to estimate missing data. Both for fitting the model as well as for the final estimation procedure, only data that are available and relevant, should be used.
3. **Standard**. Model based methods that are used at Statistics Netherlands should follow any general consensus in the literature on similar situations. In case one wants to deviate from such standards, profound arguments should be given ("comply or explain").
4. **Model selection**. Alternative models should be considered, in order to find the most appropriate model. With model selection, the aim is to choose between families of models and related estimation methods.
5. **Model fit**. Diagnostics could be used to evaluate the appropriateness of the model and model specification. Analyzing residuals could be helpful. Try to use objective measures to choose within a family of models (think about R2, AIC, BIC and the like). Be aware of over fitting.
6. **Robustness**. The model should be robust against changes in the underlying data, outliers and sudden events (e.g., change in reporting period of the underlying data or legal changes that affect registered data). Sensitivity analysis leaving out part of the data could be considered (cross validation) as well as simulation studies.
7. **Stability**. Methods used for statistics that are published on a periodic basis (e.g. quarterly statistics) should be stable over time. In model based estimation, the model should stay valid and applicable for each issue of a statistic. Assumptions should be tested periodically.
8. **Mean Square Error**. Variance and bias are both components of the Mean Square Error (MSE). The MSE of an estimator based on a model should be evaluated and, if possible, estimated. Ideally the distinction between variance and bias should be made: the variance

should be small enough and the bias should be acceptable. If possible, compare with the variance of design based estimators for the same situation.9.

9. **Assumptions**. Assumptions should be stated explicitly. Think about assumptions related to the specification, validity and applicability of the chosen model and about assumptions on the distribution of certain variables related to the model based estimation. It should explicitly be evaluated whether the assumptions are plausible in the situation at hand. Whenever possible, assumptions should be tested with other data or evaluated using studies in the relevant literature. Assumptions should be evaluated on a regular basis.
10. **Publication**. Whenever official statistics are published that have been produced using a model based approach, this should explicitly be mentioned in the publication. The used models should be documented and be easily available to the interested user of the published statistics. This may help to avoid the situation where the model that was used to produce the statistics is being "discovered" by an outside researcher, which in turn could lead to incorrect conclusions. For the interested user, it should also be easy to obtain information on the details of the model selection, the assumptions and the analyses regarding model fit and robustness.

3.3. Assessment

When developing model based estimation of official statistics, the abovementioned points should be taken into account. Such an assessment could help in deciding whether or not to use model based estimation in a specific situation. In general a complete assessment will not be possible in advance: evaluation of the points mentioned in section 3.2 will usually be done during the process of developing the method and will be completed afterwards. This implies that the use of models is the responsibility of the statistician and the statistical division (line management), where the methodology and quality departments can offer advice and/or consultancy.

The assessment should obviously be part of the statistical process and should be included in the process documentation. Whenever the use of model based estimation is an essential part of

a published statistic, this should also be mentioned in the documentation available to external parties.

4. Examples

Currently, model based estimation is used at several places at Statistics Netherlands. Not only in research projects (assessing the validity of the approach) but also in published statistics. Examples are:

- Small area estimation (LFS)
- Monthly unemployment rate (modelling time series)
- Non-response correction (weighting/calibration under Missing At Random assumption)
- Correction of mode effects in mixed mode surveys (CVS)
- Capture-recapture models to estimate hidden populations
- Seasonal adjustment

A hot topic at the time of writing this document, is the use of non-sampled-data or non-probability-sampled-data. As an example, think of Big Data.

Whenever data are just available about a subset of the population, and these data are not obtained via a sample-survey nor stem from a data generating process with known parameters, there is no design, hence design based estimation is not a valid approach (Baker et al., 2013).

The use of such data within Statistics Netherlands is still limited, but may increase in the near future, see Struijs (2013) and the references therein. In that situation, models could be used to estimate population parameters. Moreover, the use of algorithmic models may then also be considered. With algorithmic models we mean models for which an explicit parameterization is not always available (Buelens et al., 2012).

5. Summary

At Statistics Netherlands as well as at other NSIs, there is a growing interest in model based estimation. For example, ONS uses model based estimation in their Neighbourhood Statistics.

In ONS (2013a) a quite extensive methodological justification of the use of models is given. ONS also has included some aspects of model based estimation in their quality guidelines (ONS, 2013b).

Statistics Canada is mentioning model based estimation in their guidelines (Statistics Canada, 2009) as well. In both the guidelines from ONS and those from Statistics Canada some of our ten points of section 3.2 are mentioned.

At Statistics Netherlands the use of models is increasing. Examples are time series models with the LFS, multi-level models with small area estimation in the Key Subject Regionalization and models describing continuous and skewed distributions of variables in the field of economic statistics. Moving towards mixed-mode surveys also stimulates research into models that take measurement error into account. Research in methodology that makes use of models, is part of the research program of Statistics Netherlands (CBS, 2014).

Model based estimation in the field of official statistics (re)fuels the discussion between design-based and mode-based approaches, including the philosophical aspects. Little (2012) advocates to combine both paradigms, using a Bayesian approach. However, this more fundamental discussion is not (yet) settled.

As usual with relatively new methods, there are a lot of possibilities, but one should also be aware of possible pitfalls. The current paper aims to guide the researcher when attempting to use model based estimation in official statistics. The principles and attention points mentioned in section 3, should aid those researchers. Each specific situation will give rise to its own specific questions and problems. To that end, we advise the statistical divisions to seek assistance from the methodology and quality departments.

6. References

Baker, R., Brick, J.M., Bates, N.A., Bbattaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Report of the AAPOR Task Force on non-probability sampling. American Association for Public Opinion Research.

http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Brakel, J. van den (2012). Models in official statistics. Inaugural lecture 27 April 2012. School of Business and Economics, Maastricht University. http://publications.maastrichtuniversity.nl/nl/page/body/record?identifier=oai%3Apublications.maastrichtuniversity.nl%3A27-M267149

Brakel, J. van den and Bethlehem, J. (2008). Model based estimation for official statistics. CBS Discussion paper 08002. Statistics Netherlands, The Hague/Heerlen. https://www.cbs.nl/nl-nl/achtergrond/2008/10/model-based-estimation-for-official-statistics

Buelens, B., Boonstra, H.J., van den Brakel, J. and Daas, P. (2012). Shifting paradigms in official statistics: From design-based to model-based to algorithmic inference. CBS Discussion paper 201218. Statistics Netherlands, The Hague/Heerlen. https://www.cbs.nl/nl-nl/achtergrond/2012/38/shifting-paradigms-in-official-statistics-from-design-based-to-model-based-to-algorithmic-inference

CBS (2014). Research programme 2015 (version December 18, 2014). PPM, Statistics Netherlands The Hague/Heerlen.

ONS (Office for National Statistics) (2013a). Neighbourhood statistics. Small Area Model-Based Households in Poverty Estimates. http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?page=analysisandguidance/analysisarticles/households-in-poverty-model-based-estimates-at-msoa-level.htm

ONS (Office for National Statistics) (2013b). Guidelines for measuring statistical quality. http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html

Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. http://publications.gc.ca/site/eng/361886/publication.html

Struijs, P. (2013). Big Data, Big Impact? Internal CBS report, Statistics Netherlands, Heerlen