Statistics Norway
Department of Coordination and Development

Li-Chun Zhang

Documents

Some Norwegian Experience with Small Area Estimation

Preface The increasing demand on small area statistics has generated considerable research interest in the past years at national statistical agencies around the world. As the main provider of social economic data in the Norwegian society, Statistics Norway regards small area statistics as an integral part of its statistical production; and throughout the years small area estimation, has been the subject of a number of studies. In this paper we examine and communicate some of the Norwegian experience in light of updated investigations. Strategically we strongly support a wholesome approach to small area statistics, since success often does not so much depend on a particular technique, but the context/environment under which it is applied. Methodologically we emphasize the need to view small area estimation under the predictive framework, where modeling provides means rather than ends in themselves. It is our intention to highlight some of the more problematic aspects of small area estimation, i.e. places where we feel further developments are urgent. In particular, several issues and obstacles in the coming Norwegian Census 2000 will be included in our discussion.

1 Introduction

The increasing demand on small area statistics has generated considerable research interest in the past years at national statistical agencies around the world (e.g. Brackstone, 1987; Ansen, Hallen, and Ylander, 1988; Schaible, 1993; Decaudin and Labat, 1997). As the main provider of social economic data in the Norwegian society, Statistics Norway regards small area statistics as an integral part of its productive activity; and throughout the years small area estimation has been the subject of a number of studies. For instance, Laake (1978, 1979) contained some early attempts at the synthetic estimator in combination with post-stratification. Heldal, Swensen, and Thomsen (1987) carried the discussion further in connection with the Norwegian Census 1990, while Spjøtvoll and Thomsen (1987) concentrated on the composite estimator and proposed an efficient empirical Bayes method. Also Neural Network (Nordbotten, 1996) has been studied in the same context. Here we re-examine some of these experiences together, or in comparison, with alternative methods including generalized raking (Deville, Särndal, and Sautory, 1993), general semi-parametric regression models (Green, 1987), nested-error regression model (Battese, Harter, and Fuller, 1988) and linearized area-effect model — related to the works of Gelman and Little (1997) and Lazzeroni and Little (1998). The discussion aims primarily at highlighting some of the more problematic aspects of small area estimation, where we feel further developments are urgent. It is organized in four case studies, following a general formulation which unifies the various approaches involved.

2 SMAREST: SMall ARea ESTimation

Methodological developments of SMAREST in the recent twenty years or so have witnessed a growing emphasis on modeling. The reason is understandable: SMAREST arises when direct estimation based on corresponding subsamples become unfeasible/unreliable; assumptions across area/domain are then necessary in order to make use of "indirect data". Modeling in SMAREST as such provides means by which we arrive at small area statistics. Indeed, small area prediction is in many ways more appropriate a term than SMAREST. The difference is not trivial as users often question the validity of the underlying model, which may not always be convincingly resolved, even when the model does seem plausible in many respects. It reminds us of "the need to look at the problem of small area data in its entirety and not only as an estimation problem" (Singh, Gambino, and Mantel, 1993), and the importance of clarification and integration of the basic model assumptions into the design already at the planning stage. Likewise is sensitivity analysis in case the model fails essential to the practice of SMAREST.

Without losing generality, modeling in SMAREST has two considerations: to make use of indirect data, it must contain structural features which are common across area/domain; whereas to account for between-area variation, it must also deal with area-specific deviation from these common, synthetic features. Denote by y the survey variable. Let a be the area index. Denote by μ_a the *mean-parameter* of the survey variable from area a, i.e. $\mu_a = E[y_a]$. The synthetic features common to the population can often be summarized in the following manner, i.e.

$$h(\mu_a) = \xi_a = g(\xi, x_a), \tag{1}$$

where $h()$ is the link function, and ξ some global parameter, and x_a the relevant auxiliary information. Notice that $h()$, ξ and $g()$ are all independent of a. Typically $g()$ is of the linear form which, through the link function $h()$, makes way for generalized linear models. Whereas the general form, among others, allows for non-parametric approach as well. In particular, (1) is specified at the area-/domain-level if y_a is a scalar, in which case the auxiliary x_a is a vector in general; whereas it is specified at the individual-level if y_a is a vector, i.e. $y_a = (y_1, ..., y_{N_a})^T$ where N_a is the size of the small area, in which case x_a is a matrix in general.

To account for overdispersion an additional deviation part is sometimes necessary, which can similarly be summarized in $\eta_a = \eta(z_a, e_a)$, where z_a contains relevant auxiliary information, and e_a random errors with prior distribution $\pi(e_a)$ — often though $\pi(e_a)$ is only specified up to the first two moments of e_a. Notice that e_a is postulated by the model, and has nothing to do with sampling error. Combining synthetic feature with local deviation, we have, for ξ_a and η_a defined as above,

$$h(\mu_a) = \xi_a + \eta_a = g(\xi, x_a) + \eta(z_a, e_a). \tag{2}$$

We call ξ_a the *synthetic-parameter* and η_a the *deviation-parameter*. The mean-parameter μ_a is obtained from $\xi_a + \eta_a$ through a transformation defined by the link function $h()$. In particular, (1) is the special case of (2) where $\eta_a \equiv 0$.

A number of models for SMAREST can be expressed by means of (1) and (2). Let $g()$ take the linear form, i.e. $g(\xi, x_a) = x_a^T \xi$, and $h(\mu_a) = \mu_a$. Setting $\eta_a = 0$, i.e. (1), gives us the linear model, whereas allowing for area-specific random η_a, i.e. (2), leads to the variance-component approach. For instance, the group-mean model (Holt, Smith, and Tomberlin, 1979) which motivates the synthetic estimator based on post-stratification can be defined in terms of (1), where ξ is the post-stratum mean across the areas, and x_a the known post-stratum proportions within the relevant small area. The random area-effect model (Ghosh and Rao, 1994) follows from (2), where $\eta_a = z_a e_a$ for some fixed z_a and random effect e_a. Whereas in the nested-error regression model (Battese, Harter, and Fuller, 1988), we have $\eta_a = e_a$ and $z_a = 1$. Moreover, setting $h()$ to be some non-identity link, we obtain the generalized linear model (McCullagh and Nelder, 1989) by (1), and the generalized linear mixed model (Breslow and Clayton, 1993) by (2) where $\eta_a = z_a^T e_a$ with normally distributed e_a. Finally, in non-parametric approach, $g()$ can be left unspecified, an example of which we shall discuss later in connection with the Norwegian Census 2000.

3 Some Norwegian experience of SMAREST

The following cases have been taken from the practice of Statistics Norway. The first one deals with the use of auxiliary information in the Norwegian Labour Force Survey (Laake, 1978; Heldal, Swensen, and Thomsen, 1987; Zhang, 1999). Next we discuss treatment of overshrinkage using an empirical Bayes approach (Spjøtvoll and Thomsen, 1987) and a penalized (quasi) likelihood approach under the general semi-parametric regression models (Green, 1987). The third case will be based on data from the Effectivity Survey 1996 conducted by the Norwegian Fishery Directory — Statistics Norway was consulted on the matter of a re-design. The techniques involved are nested-error regression model (Battese, Harter, and Fuller, 1988) and linearized area-effect model, which is related to the work of Gelman and Little (1997) and Lazzeroni and Little (1998). Finally, we shall discuss a number of issues and obstacles in the Norwegian Census 2000.

3.1 Use of auxiliary information in the Norwegian Labour Force Survey (LFS)

"An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small area have the same characteristics as the large area, we identify these estimates as synthetic estimates" (Gonzalez, 1973). In presence of auxiliary information it is seldom to apply the mean of a large area directly to *all* the small areas. Synthetic estimates are often formed in combination with post-stratification, the assumption being that the post-stratum mean does not vary over the small areas which cut across the post-strata. As noted earlier, such a synthetic estimator can be motivated by model (1), i.e. let h be the post-stratum index,

$$\mu_a = \sum_h (N_{ah}/N_a)\mu_h \qquad \text{where} \quad N_a = \sum_h N_{ah} \quad \text{and} \quad \mu_h = (\sum_a \sum_{i \in U_{ah}} y_i)/(\sum_a N_{ah}),$$

where U_{ah} denotes the corresponding sub-population and N_{ah} its size.

Laake (1978) evaluated this synthetic estimator in the context of the Norwegian LFS. The post-stratification was based on auxiliary information of Sex and Age. No improvement, measured in terms of the mean square error, was found in the resulting synthetic estimator at the county level. Heldal, Swensen, and Thomsen (1987) brought new light to the conclusion, by which time a population Register on employment-status had been established which could be linked to the LFS at the individual level. Due to the high correlation between the Register-Employment and the LFS-Employment, post-stratification according to the former substantially improves the efficiency of the corresponding synthetic estimator at the county level. However, post-stratification based on Register-Employment, Sex and Age (12 groups), though feasible at the national level, would lead to a large number of empty cells within each county. Instead Zhang (1998) adopted the generalized raking (Deville, Särndal, and Sautory, 1993), and found that the gain in efficiency at the national level (through post-stratification) was largely preserved at the county level (through calibration). The result applies to stratified random sampling in general (Zhang, 1999), in which case the conditional inference of Holt and Smith (1979) can be carried out by means of the generalized raking within each stratum.

Again, this example reminds us of the importance of a wholesome approach towards small area statistics. The success of a technique depends always on the context in which it is applied. The use of quality auxiliary information was first made possible through the establishment of the Register on employment-status. Without the Register, the counties would have to be treated as "small" areas from which direct estimators are too unreliable to be accepted. Whereas the generalized raking is applied within each county, i.e. only direct data are used for the county estimates and the counties are no longer "small" areas. Moreover, to enhence the performance of the county estimates, the Norwegian LFS underwent a re-design period between 1996 and 1998. Among other things stratified sampling has replaced an approximate self-weighting two-stage sampling, and smaller counties are now given a relatively higher representation in the LFS.

3.2 Overshrinkage among within-stratum municipality estimates in the Norwegian LFS

Under a Bayesian framework, overshrinkage in SMAREST basically implies that the small area estimates in general display less between-area variation than what is postulated *a priori*. Morris (1983), Spjøtvoll and Thomsen (1987) and Ghosh (1992) developed so-called constrained approaches to address the problem. Denote by s the sample and s_a the sub-sample from area a, with the respective size n and n_a. Let $\bar{y}_a = \sum_{i \in s_a} y_i/n_a$. Spjøtvoll and Thomsen (1987) adopted a simple empirical Bayes (EB) model which, in our present notation (2), can be given as

$$\mu_a = \mu + e_a \qquad \text{where} \quad E[e_a] = 0 \quad \text{and} \quad Var(\mu_a) = \sigma^2 \quad \text{and} \quad E[\bar{y}_a|\mu_a] = \mu_a,$$

i.e. the sub-sample mean is an unbiased estimator of the corresponding mean-parameter. Assume that the EB estimator belongs to the linear class, i.e. $\hat{\mu}_a = \alpha_a \bar{y}_a + \beta_a$. In the constrained EB (CEB) estimator, (α_a, β_a) are such that $E[\hat{\mu}_a] = \mu$ and $E[(\hat{\mu}_a - \mu)^2] = \sigma^2$. This leads to a composite estimator of the following form, i.e.

$$\hat{\mu}_a = w_a^\gamma \bar{y}_a + (1 - w_a^\gamma)\mu \qquad \text{where} \quad w_a = (n_a\sigma^2)/[(n_a - 1)\sigma^2 + \mu(1-\mu)], \tag{3}$$

where $\gamma = 1/2$. It was noticed that $\gamma = 1$ corresponded to the EB estimator such that $E[(\hat{\mu}_a - \mu)^2]$ was minimized, which would result into overshrinkage among $\hat{\mu}_a$. Whereas $\gamma = 0$ leaves us with the direct estimator $\bar{y}_a$ alone. Since $E[\bar{y}_a] \equiv \mu$ by stipulation, estimator (3) satisfies $E[\hat{\mu}_a] = \mu$. The parameter γ can be considered a *shrinkage factor* which balances between the direct estimator $\bar{y}_a$ and the indirect population mean μ. Often neither μ nor σ^2 is known, in which case (3) is applied after substitution of their respective estimates based on the sample.

Green (1987) studied the general semi-parametric regression models, and suggested to base estimation on the penalized (quasi) likelihood (PQL). The following simple model in the notation of (2) derives directly from such general semi-parametric regression models. Suppose binary survey variable y, in which case μ_a becomes the parameter of the Binomial distribution in area a, and

$$h(\mu_a) = \log \mu_a - \log(1 - \mu_a) = \xi + e_a \qquad \text{where} \quad E[e_a] = 0. \tag{4}$$

Should the distribution of e_a be normal, the model would be a special case of the generalized linear mixed models (GLMM, Breslow and Clayton, 1993) and, more generally, the hierarchical generalized linear models (HGLM, Lee and Nelder, 1996). Unlike Green (1987), model (4) specifies the first moment to be zero, and can be called an *empirical* generalized linear mixed model (EGLMM). The estimation procedure described in Green (1987) remains applicable, however, the additional assumption can perhaps be used to motivate the following PQL, i.e.

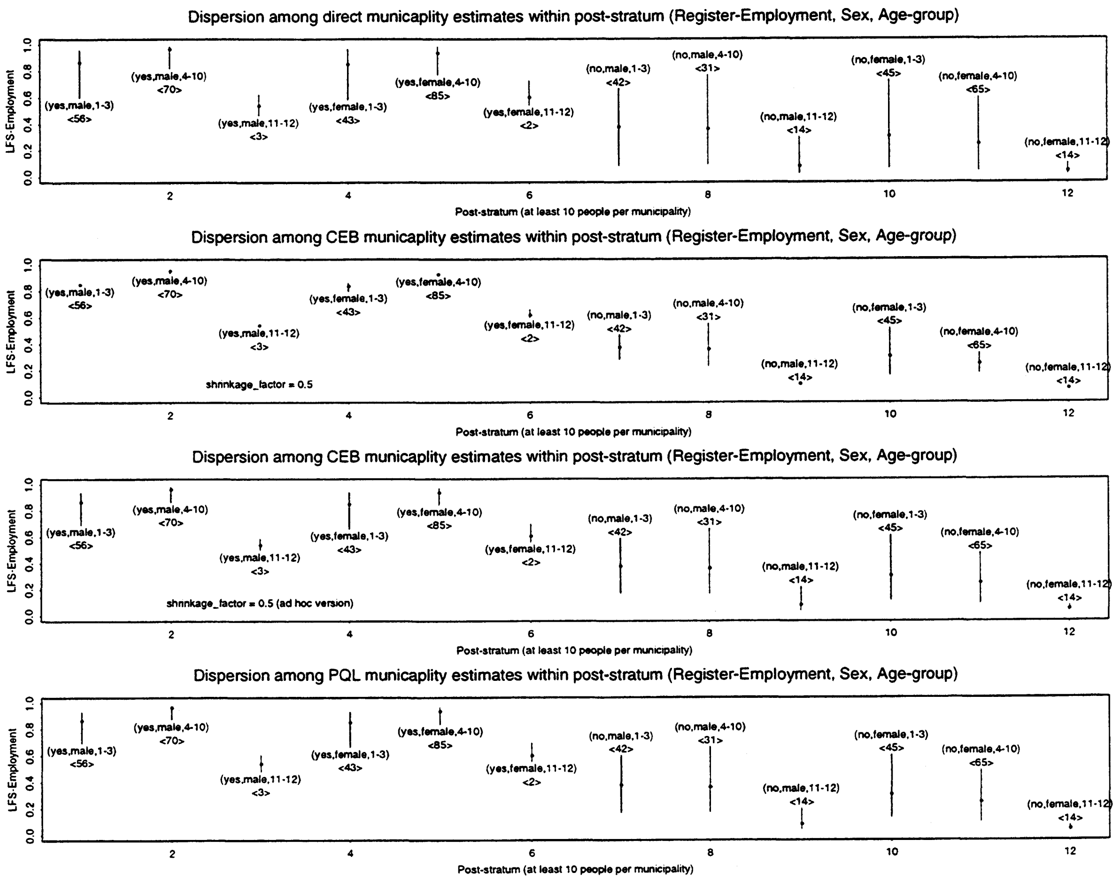
$$l_e(\xi; y) = l(\xi; y, e) - \frac{1}{2}\sum_a e_a^2 = [\sum_a y_a(\xi + e_a) + n_a \log(1 - \mu_a)] - \frac{\lambda}{2}\sum_a e_a^2,$$

under which $\sum_a \hat{e}_a = 0$. The term $l(\xi; y, e)$ measures the ability of $\hat{\mu}_a$ to account for the data. Whereas large values of e_a can be considered the penalty to be paid for the increment in $l(\xi; y, e)$, without which e_a can be chosen such that $\hat{\mu}_a$ yield perfect fit to the data. Tuning parameter λ (≥ 0) has similar effect on the PQL estimates as that of the shrinkage factor γ in (3). It controls the roughness of penalty: $\lambda = 0$ gives us the direct estimates, whereas $\hat{e}_a \to 0$ as $\lambda \to \infty$. Let $\psi_a = y_a - n_a\mu_a$ and $v_a = n_a\mu_a(1 - \mu_a)$. Let $e = (e_1, ..., e_A)^T$ and $\psi = (\psi_1, ..., \psi_A)^T$ where A is the number of small areas, and V the diagonal matrix with diagonal $(v_1, ..., v_A)^T$, and I the $A \times 1$ unity vector. We update $\theta = (\xi, e_1, ..., e_A)^T$ as $\theta + j^{-1}u$ and iterate, where

$$u = \begin{pmatrix} I^T\psi \\ \psi - e \end{pmatrix} \qquad \text{and} \qquad j = \begin{pmatrix} I^TVI & I^TV \\ VI & II^T + V \end{pmatrix}.$$

Data from the Norwegian LFS in the 4th quarter of 1995 have been used to study both methods. Post-stratification was applied to the LFS-sample according to Register-Employment — (yes, no), Sex — (male, female), and Age — 12 age-groups each of five years. For a clearer overview, we collapsed the 12 age-groups into 3, i.e. (1-3, 4-10, 11-12), thereby reducing the number of post-strata from 48 ($= 2 \times 2 \times 12$) to 12. Within each post-stratum, sub-sample from a municipality was admitted provided (i) it contained at least 10 people, and (ii) it is non-degenerated in the sense that the observed LFS-Employment Rate there was neither 0 nor 1. Based on all such admitted sub-samples we calculated the LFS-Employment Rate using, (a) the direct within-stratum municipality estimates $\bar{y}_{ah}$ for binary LFS-Employment y, (b) the CEB estimates (3) with shrinkage factor $\gamma = 0.5$ where the method of moment (Copas, 1972) was used to estimate (μ, σ^2), (c) the CEB estimates ($\gamma = 0.5$) with the sample variance of the observed LFS-Employment Rate as the *ad hoc* estimator of σ^2, and (d) the PQL estimates under model (4) where $\lambda = 1$. Dispersion among these within-stratum municipality estimates has been plotted in Figure 1, where the numbers in "<>" indicate how many municipalities were involved within each post-stratum, and the marks the estimated overall post-stratum LFS-Employment Rates.

Figure 1: Dispersion among within-stratum municipality estimates in the Norwegian LFS



The method of moment estimator (Copas, 1972) of (μ, σ^2) was such that, in several post-strata, we had to truncate negative $\hat{\sigma}^2$ to zero. The shrinkage factor had no effect on the CEB estimates there, which consequently had no dispersion at all. This extra contingency was mainly caused by the relatively high homogeneity among the within-stratum municipality sub-samples. In other words, there is an inherited threshold for overdispersion in the CEB approach: unless the observed dispersion gets over this limit, the CEB estimates simply degenerate into the overall mean. The *ad hoc* estimator of σ^2 would have been appropriate if the admitted sub-samples were the population. The results seem to suggest that one should check, if possible, whether some fixed prior σ^2 could be used to stablize the CEB estimates. For the present case, such prior σ^2 may be based on previous population census. Deviation from the true σ^2 is likely to be out-weighted by the advantage of avoiding degenerating cases. It is interesting to see that the dispersion among the *ad hoc* CEB estimates are rather similar to that among the PQL estimates. This raises the question whether the latters are not too easily affected by the random variation in the observed LFS-Employment Rates. The PQL approach can of course be made more resilent by larger values of λ, which controls the dispersion among the estimates in a smooth fashion. We may, if possible, simulate the PQL approach based on previous census data to empirically choose some suitable λ. Such sensitivity analysis would certainly add to our credibility if the 'stablized' CEB approach agrees with the 'empirical' PQL approach.

3.3 Linearized area-effect in the Effectivity Survey (ES) 1996

In the Effectivity Survey 1996 conducted by the Norwegian Fishery Directory, the sample contained 394 fish boats (from a total of 1283 in the population). Classified according to (i) the length of the boat — 4 classes, (ii) the type of license granted — 22 types, and (iii) the county in which the boat was registered — 9 counties, there were altogether 166 non-empty domains in the population, of which 109 were represented in the sample. Let the amount of fish Catch be the survey variable, denoted by y. We notice that the Catch is in fact known for all the fish boats, so that various methods can be checked empirically against the true values. Based on monthly report to the Directory, a yearly fishing income, denoted by x, is available for each boat, which will be used as the auxiliary variable. This gives us a simple regression model, i.e.

$$\mu_a = x_a \beta_a \qquad \text{where} \quad \mu_a = (\mu_1, ..., \mu_{N_a})^T \quad \text{and} \quad x_a = (x_1, ..., x_{N_a})^T,$$

where β_a is the domain-specific 'slope', and N_a the size of the corresponding domain, denoted by U_a. We refer to this as the *domain model*, which motivates the direct within-domain ratio estimator.

Battese, Harter, and Fuller (1988) applied nested-error regression model for prediction of County Crop Areas. In the present context and notation of (2), let μ_a, x_a and y_a all be vector-valued, and

$$\mu_a = x_a \beta + e_a \qquad \text{and} \qquad y_a = \beta x_a + e_a + \epsilon_a \quad \text{where} \quad \epsilon_a = (\epsilon_1, ..., \epsilon_{N_a})^T$$

with independent e_a and ϵ_a, and $E[e_a] = 0$, and $E[\epsilon_a] = 0$, and $Var(e_a) = \sigma^2$, and $Var(\epsilon_j) = \tau^2$ for $j \in U_a$. The individual error, i.e. $e_a + \epsilon_j$ for $i \in s_a$, has a nested structure calling for variance-component approach (Fuller and Battese, 1973). The nested-error regression model makes use of indirect data through global β and accounts for between-domain variation through random e_a.

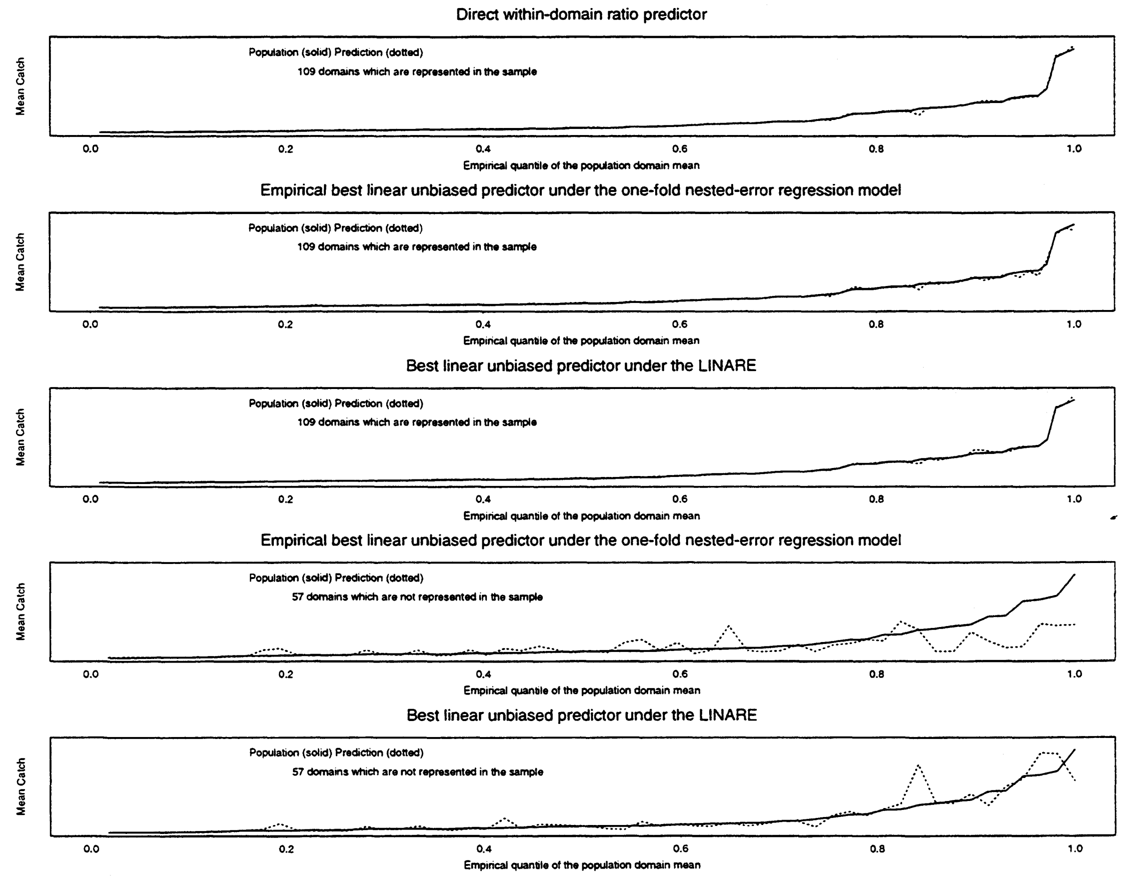
Linearized area-effect model (LINARE) seeks to account for the area-/domain-effects through a suitable linear structure. Denote by $B = B(x)$ the $A \times q$ *design matrix* whose ath row corresponds to the ath domain, and ξ the $q \times 1$ parameter vector. Domain-specific 'slopes' are now given by the $A \times 1$ vector $B\xi$, and we obtain a simple synthetic model in the sense of (1) as

$$\mu_a = x_a (B\xi)_a \qquad \text{and} \qquad y_a = x_a (B\xi)_a + \epsilon_a \quad \text{where} \quad \epsilon_a = (\epsilon_1, ..., \epsilon_{N_a})^T. \tag{5}$$

In particular, the design matrix B may arise from the defining variables for the domains in the same way calibration arises from post-stratification through suitable dummy-indexing. The domain model is recovered given $A \times A$ identity design matrix B; otherwise reduction in the rank of B typically corresponds to deletion of higher-order "interactions" among the defining variables. Whereas Lazzeroni and Little (1998) and Gelman and Little (1997) went on to fit a hierarchical model, the LINARE (5) remains synthetic and no deviation-parameter is introduced. Inference employs standard techniques for the linear models, after deletion of redundant columns in the $n \times q$ *sample design matrix*, whose ith row is given by the ath row of B provided $i \in s_a$. Above all, this calls for specification of variance-constants, denoted by $v_a = v_a(x) = (v_1, ..., v_{N_a})^T$, such that $Var(\epsilon_j) = v_j \tau^2$.

The LINARE is motivated for the ES 1996 by the fact there were 57 domains not represented in the sample, where any model has to be synthetic in the sense of (1) since random effect can not be 'predicted' there. We applied the LINARE to the data of the ES 1996, where B was set up in the same way as calibration towards the marginal totals of x, i.e. 35 $(= 4 + 22 + 9)$ of them. The variance-constants were defined as $v_j = x_j^r$, where we chose $r = 1$ — varying r had very little effect on the results. We calculated (i) the direct within-domain ratio predictor under the domain model for the 109 domains which were represented in the sample, and (ii) the empirical best linear unbiased predictor (EBLUP) under the nested-error regression model for all the 166 domains, and (iii) the best linear unbiased predictor (BLUP) under the LINARE for all the 166 domains. Empirical comparison with the known population values have been summarized in Figure 2, separately, for the 109 domains represented in the sample and the rest 57 domains.

Figure 2: Domain mean predictors for the Effectivity Survey 1996



As we have remarked earlier, modeling in SMAREST serves above all as means of small area prediction. The existence of the 57 unrepresented domains therefore necessarily affects the angle of modeling — "the approach involving components of variance has arisen from the need to take account of the between-small-area components of variance. However a much more rewarding approach is to seek to explain why small areas differ" (Holt and Moura, 1993). We might add that sometimes "to explain" means to be able to predict. Whereas the domain model is intuitive, the nested-error regression model is flexible in treating the domain-effects as random. The LINARE, on the other hand, seeks explanation in a linear structure. In this case all the three models gave results close to each other for the 109 domains represented in the sample. The differences became first clear for the rest 57 domains. The domain model was not feasible at all; neither can the random effects be 'predicted'. Although the LINARE seems to adequately account for the observed between-domain variation, it does not necessarily follow that we have found the explanation to "why small areas differ" — its performance was also noticeably worsened for the 57 domains. Better results could be achieved through the LINARE, or perhaps also the nested-error regression model, in combination with a design, under which all the domains are ensured representation in the sample.

3.4 Some issues and obstacles in the Norwegian Census 2000

Due partly to the improvements on the quality of register-based information, including those mentioned earlier in the context of the Norwegian LFS, the Norwegian Census 1990 was based on a large-scale sampling survey, in combination with traditional population census. The design was such that census was carried out only among the smallest municipalities, whereas random samples were taken from the larger ones. Statistics there were produced using methods such as post-stratification and raking, in combination with the various register sources (Lillegård, 1993). The coming Norwegian Census 2000 will be divided into two parts, i.e. the household and personal census. The latter has basically been conceived as register-based, with one or two special small supplementary surveys. Traditionally, population census, administrative registers, as well as sampling surveys have been the sources of small area statistics; and SMAREST concerns only the last of these three. The strategy of the Norwegian Census 2000 has given rise to both conceptual and technical challenges.

Above all conceptual clarification is needed between register-based and population census. While both provide information at the individual level, there exist irreconcilable gaps due to the nature of the two production methods. This introduces the following dilemma: without adjustment, the register-based census is incomparable with the previous population censuses, nor current survey-based statistics; whereas in adjusting the register-based census to be more population-census-alike, one runs the risk against concealed incompatibility despite all the effort, which no longer can be attributed to conceptual differences alone. It is essential that the public users are made sufficiently aware of such differences *in advance*, in order to avoid possibly irreparable damages to the future.

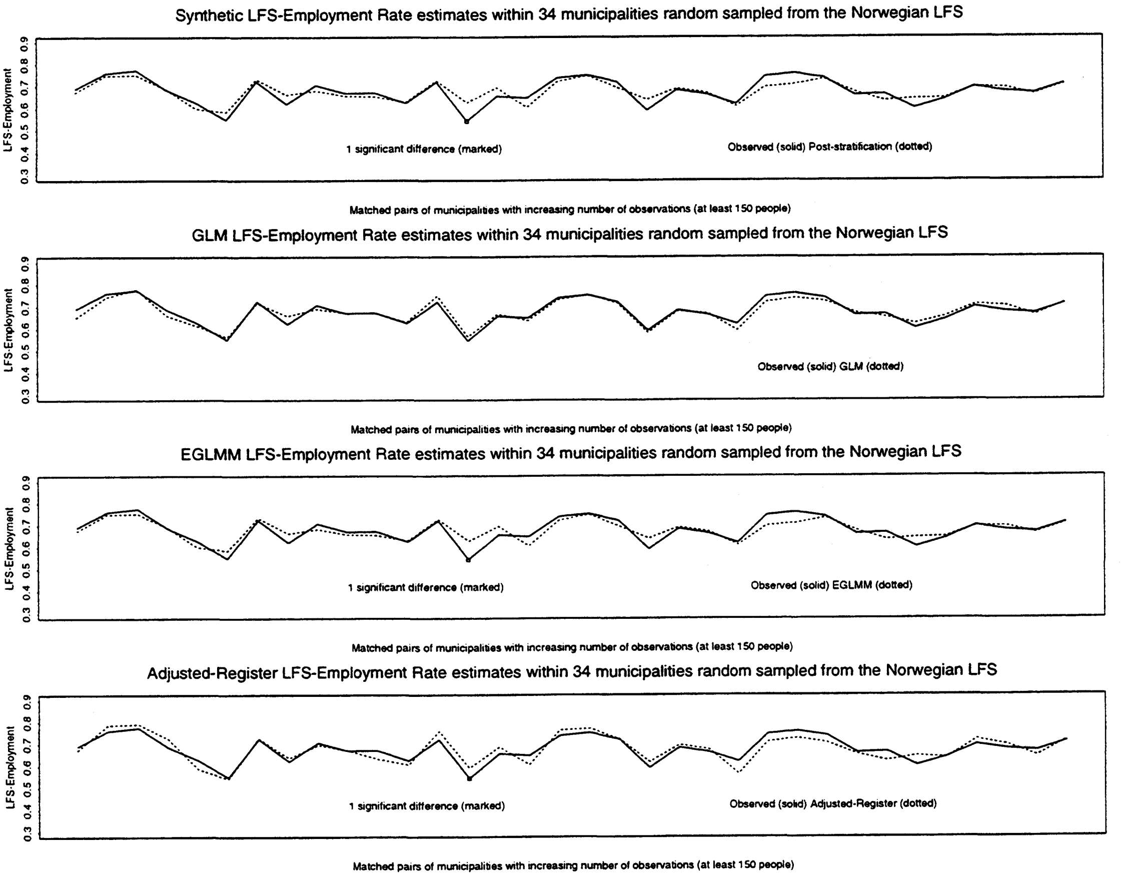
For instance, method of classification has been studied to produce register-based Census Employment statistics which secures agreements with the LFS estimates at the national level. While it would be inappropriate to get into the details here, it suffices to notice that the procedure is of the synthetic nature in the sense of (1), where $g()$ is left unspecified and x a number of register sources including the Register used in the LFS estimation. The method contains no area-specific adjustment, and the deviation-parameter (2) is absent. This differs from the traditional population census, as well as the Norwegian Census 1990, where direct estimation was employed for the sampling municipalities. Though there are no intrinsic reasons why a synthetic approach should not be able to produce good small area statistics, it is not a simple issue to evaluate its performance in this respect, unless we take its validity for granted. Due to the developments since 1990, neither the availability nor the quality of the register sources remains the same. This has also made it difficult to empirically test any register-method on the data from the Census 1990.

A simple illustration can be given through the following exercise. A random sample of municipalities were selected from the Norwegian LFS in the 4th quarter of 1995, among which we retained those sub-samples containing at least 150 people. The idea is to treat the admitted sub-samples, denoted by s, as our population, and see whether the observed municipality LFS-Employment Rates can be "explained" in some synthetic manner. More explicitly, we estimated the LFS-Employment Rate within each municipality, (i) using a synthetic estimator based on post-stratification according to Register-Employment (yes, no), Sex (male, female) and Age (12 age-groups) — the post-stratum means have been estimated based on s; (ii) under a generalized linear model (GLM), i.e.

$$h(\mu_a) = \log \mu_a - \log(1-\mu_a) = \bar{x}_a^T \xi \qquad \text{where} \quad \bar{x}_a = \sum_{i \in s_a} x_i / n_a,$$

and x_i the 16×1 dummy-indexed auxiliary vector w.r.t. Register-Employment, Sex and Age; (iii) under an EGLMM where we added a random area-effect e_a to the GLM above, and estimation was based on PQL with a penalty term $-\sum_a e_a^2/2$; and (iv) using a ratio adjustment of the corresponding observed Register-Employment Rate, denoted by $\bar{z}_a$, i.e. $\hat{\mu}_a = \bar{z}_a r$ where $r = \sum_{i \in s} y_i / \sum_{i \in s} z_i$. In Figure 3 the results were compared to the observed municipality LFS-Employment Rates, where we have marked the differences which were significant.

Figure 3: Some municipality LFS-Employment Rate estimates in the Norwegian LFS



Compared to the GLM, the EGLMM did not improve the agreement between the observed and fitted municipality LFS-Employment Rates in this case; and the idea of "explaining" the LFS-Employment through some synthetic structure alone seems not entirely implausible. In fact, no significant difference was found between the GLM estimates and the observed values. Despite such plausibility, it remains difficult to establish the validity of these methods for *all* the municipalities in Norway based on the data material available. Notice that we have used some of the largest sub-samples in the LFS, which also came from the larger municipalities in the population. Given the fact that Employment is certainly not among the worst census variables in this respect, evaluation of the register-based census statistics seems indeed a formidable task for the Norwegian Census 2000.

4 Summary

The demand on small area statistics seems to grow in pace with the ever quickening social economical changes around us. Research in SMAREST of the recent years have produced a variety of methods with clear emphasis on modeling. Through the use of indirect data, this may greatly improve the efficiency of the resulting estimator. Thus area-specific bias frequently becomes a central concern. We feel that evaluation of SMAREST should be placed under the predictive framework, where models are means rather than ends in themselves. More often than not, the predictive ability of a model may not be equated to the fact that it satisfactorily accounts for the observed area-dispersion, unless all the areas are adequately represented in the sample. Indeed, our experience seems to suggest that success often does not so much depend on a particular technique, than the context/environment in which it is applied. We therefore strongly call for a wholesome approach to small area statistics. Sometimes the combined effort could even turn SMAREST into direct estimation. As noted by several authors before us, lack of data characterizes SMAREST, which makes it difficult, if not impossible, to establish the validity of a model empirically based on the material available. Consequently, neither can sensible measures of error be provided without careful planning followed by reliable data collection. It is therefore particularly important that plausible assumptions, which may or may not be articulated by an explicit model, should be clarified and integrated into the design from the beginning.

References

Ansen, H., Hallen, S.-A., and Ylander, H. (1988). Statistics for regional and local planning in Sweden. *J. Off. Statist.*, **4**, 35–46.

Battese, G.A., Harter, R.M., and Fuller, W.A. (1988). A error-component model for prediction of county crop areas using survey and satelite data. *J. Amer. Statist. Assoc.*, **83**, 28–36.

Brackstone, G.J. (1987). Small area data: policy issues and technical challenges. In *Small Area Statistics*, ed. by R. Platek, J. Rao, C.-E. Särndal, and M. Singh, pp. 3–20. Wiley, New York.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9–25.

Copas, J. (1972). Empirical Bayes methods and the repeated used of a standard. *Biometrika*, **59**, 349–60.

Decaudin, G. and Labat, J.-C. (1997). A synthetic, robust and efficient method of making small area population estimates in France. *Survey Methodology*, **23**, 91–8.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.*, **88**, 1013–20.

Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *J. Amer. Statist. Assoc.*, **68**, 626–32.

Gelman, A. and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, **23**, 127–35.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *J. Amer. Statist. Assoc.*, **87**, 533–40.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, **9**, 55–93.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimators. In *Proceedings of the Social Statistics Section*, pp. 33–6. Amer. Statist. Assoc., Washington, DC.

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.*, **55**, 245–59.

Heldal, J., Swensen, A.R., and Thomsen, I. (1987). Census statistics through combined use of surveys and registers. *Statistical Journal of the United Nations ECE*, **5**, 43–51.

Holt, D. and Moura, F. (1993). Mixed models for making small area estimates. In *Small Area Statistics and Survey Designs*, ed. by G. Kalton, J. Kordos, and R. Platek, vol. **1**, pp. 221–31. Central Statistical Office, Warsaw.

Holt, D. and Smith, T.M.F. (1979). Post stratification. *J. Roy. Statist. Soc. A*, **142**, 33–46.

Holt, D., Smith, T.M.F., and Tomberlin, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *J. Amer. Statist. Assoc.*, **74**, 405–10.

Laake, P. (1978). An evaluation of synthetic estimates of employment. *Scand. J. Statist*, **5**, 57–60.

Laake, P. (1979). A predictive approach to subdomain estimation in finite populations. *J. Amer. Statist. Assoc.*, **74**, 355–8.

Lazzeroni, L.C. and Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *J. Off. Statist.*, **14**, 61–78.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. B*, **58**, 619–78.

Lillegård, M (1993). Folke- og Boligtelling 1990 — Dokumentasjon av de statistiske metodene. Technical report, Statistics Norway (Rapporter 93/4), In Norwegian.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd edn). London: Chapman and Hall.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.*, **78**, 47–65.

Nordbotten, S. (1996). Neural Network imputation applied to the Norwegian 1990 population census data. *J. Off. Statist.*, **12**, 385–401.

Schaible, W.L. (1993). Use of small area estimators in U.S. federal programs. In *Small Area Statistics and Survey Designs*, ed. by G. Kalton, J. Kordos, and R. Platek, vol. **1**, pp. 95–114. Central Statistical Office, Warsaw.

Singh, M.P., Gambino, J., and Mantel, H. (1993). Issues and options in the provision of small area data. In *Small Area Statistics and Survey Designs*, ed. by G. Kalton, J. Kordos, and R. Platek, vol. **1**, pp. 37–75. Central Statistical Office, Warsaw.

Spjøtvoll, E. and Thomsen, I. (1987). Application of some empirical Bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, **2**, 435–49.

Zhang, L.-C. (1998). Dokumentasjonsrapport: Den nye estimeringsmetoden for Arbeidskraftundersøkelsen (AKU). Technical report, Statistics Norway (Notater 98/1), In Norwegian.

Zhang, L.-C. (1999). Post-stratification and calibration — A synthesis. *(Tentative accept). The American Statistician.*

Recent publications in the series Documents

97/11 S. Blom: Residential Consentration among Immigrants in Oslo

97/12 Ø. Hagen and H.K. Østereng: Inter-Baltic Working Group Meeting in Bodø 3-6 August 1997 Foreign Trade Statistics

97/13 B. Bye and E. Holmøy: Household Behaviour in the MSG-6 Model

97/14 E. Berg, E. Canon and Y. Smeers: Modelling Strategic Investment in the European Natural Gas Market

97/15 A. Bråten: Data Editing with Artificial Neural Networks

98/1 A. Laihonen, I. Thomsen, E. Vassenden and B. Laberg: Final Report from the Development Project in the EEA: Reducing Costs of Censuses through use of Administrative Registers

98/2 F. Brunvoll: A Review of the Report "Environment Statistics in China"

98/3: S. Holtskog: Residential Consumption of Bioenergy in China. A Literature Study

98/4 B.K. Wold: Supply Response in a Gender-Perspective, The Case of Structural Adjustments in Zambia. Tecnical Appendices

98/5 J. Epland: Towards a register-based income statistics. The construction of the Norwegian Income Register

98/6 R. Chodhury: The Selection Model of Saudi Arabia. Revised Version 1998

98/7 A.B. Dahle, J. Thomasen and H.K. Østereng (eds.): The Mirror Statistics Exercise between the Nordic Countries 1995

98/8 H. Berby: A Demonstration Data Base for Business Register Management. A data base covering Statistical Units according to the Regulation of the European Union and Units of Administrative Registers

98/9 R. Kjeldstad: Single Parents in the Norwegian Labour Market. A changing Scene?

98/10 H. Brüngger and S. Longva: International Principles Governing Official Statistics at the National Level: are they Relevant for the Statistical Work of International Organisations as well?

98/11 H.V. Sæbø and S. Longva: Guidelines for Statistical Metadata on the Internet

98/12 M. Rønsen: Fertility and Public Policies - Evidence from Norway and Finland

98/13 A. Bråten and T. L. Andersen: The Consumer Price Index of Mozambique. An analysis of current methodology – proposals for a new one. A short-term mission 16 April - 7 May 1998

98/14 S. Holtskog: Energy Use and Emmissions to Air in China: A Comparative Literature Study

98/15 J.K. Dagsvik: Probabilistic Models for Qualitative Choice Behavior: An introduction

98/16 H.M. Edvardsen: Norwegian Regional Accounts 1993: Results and methods

98/17 S. Glomsrød: Integrated Environmental-Economic Model of China: A paper for initial discussion

98/18 H.V. Sæbø and L. Rogstad: Dissemination of Statistics on Maps

98/19 N. Keilman and P.D. Quang: Predictive Intervals for Age-Specific Fertility

98/20 K.A. Brekke (Coauthor on appendix: Jon Gjerde): Hicksian Income from Stochastic Resource Rents

98/21 K.A.Brekke and Jon Gjerde: Optimal Environmental Preservation with Stochastic Environmental Benefits and Irreversible Extraction

99/1 E. Holmøy, B. Strøm and T. Åvitsland: Empirical characteristics of a static version of the MSG-6 model

99/2 K. Rypdal and B. Tornsjø: Testing the NOSE Manual for Industrial Discharges to Water in Norway

99/3 K. Rypdal: Nomenclature for Solvent Production and Use

99/4 K. Rypdal and B. Tornsjø: Construction of Environmental Pressure Information System (EPIS) for the Norwegian Offshore Oil and Gas Production

99/5 M. Søberg: Experimental Economics and the US Tradable SO_2 Permit Scheme: A Discussion of Parallelism

99/6 J. Epland: Longitudinal non-response: Evidence from the Norwegian Income Panel

99/7 W. Yixuan and W. Taoyuan: The Energy Account in China: A Technical Documentation

99/8 T.L. Andersen and R. Johannessen: The Consumer Price Index of Mozambique: A short term mission 29 November – 19 December 1998

99/9 L-C. Zhang: SMAREST: A Survey of SMall ARea ESTimation

Documents

PORTO BETALT VED INNLEVERINGA P.P. NORGE/NOREG

Tillatelse nr. 159 000/502

B

Returadresse:
Statistisk sentralbyrå
Postboks 8131 Dep.
N-0033 Oslo

Statistics Norway
P.O.B. 8131 Dep.
N-0033 Oslo

Tel: +47-22 86 45 00
Fax: +47-22 86 49 73

ISSN 0805-9411

Statistisk sentralbyrå
Statistics Norway