Discussion Papers No. 440, December 2005
Statistics Norway, Statistical Methods and Standards

Li-Chun Zhang and Ib Thomsen

A prediction approach to sampling design

Abstract:
Standard approaches to sample surveys take as the point of departure the estimation of one or several population totals (or means), or a few predefined sub-totals (or sub-means). While the model-based prediction approach provides an attractive framework for estimation and inference, a model-based theory for the variety of randomization sampling designs has been lacking. In this paper we extend the model-based approach to the prediction of individuals in addition to totals and means. Since, given the sample, the conditional prediction error is zero for the selected units but positive for the units outside of the sample, it is possible to use the sampling design to control the unconditional individual prediction mean square errors. This immediately raises the need for probability sampling. It turns out that balancing between optimal prediction of the population total and control over individual predictions provides a fruitful model-based approach to sampling design. Apart from raising the need for probability sampling in general, it leads naturally to a number of important design features that are firmly established in the sampling practice, including the use of simple random sampling for homogeneous populations and unequal probability sampling otherwise, the division of a business population into the take-all, take-some and take-none units, the most common two-stage sampling designs, the use of stratification with proportional allocation, etc.. Most of them have not received adequate model-based treatment previously. Our approach enables us to give an appraisal of these methods from a prediction point of view.

Keywords: Individual prediction, business survey, unequal probability sampling, two-stage sampling, linear regression population, common parameter model

Address: Li-Chun Zhang, Statistics Norway, Statistical Methods and Standards.
E-mail: li.chun.zhang@ssb.no

Discussion Papers comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

Standard approaches to sample surveys take as the point of departure the estimation of one or several population totals (or means), or a few predefined sub-totals (or sub-means). Under the model-based prediction approach (Valliant, Dorfman, and Royall, 2000) the implications on the sampling design can be extreme as in the case of purposive selection for populations under the ratio model, where the sample consists of the n largest units and n is the sample size. There is clearly a legitimate concern over the sensitivity of such a purposive sample, because the n largest units no longer constitute an optimal sample if the linear predictor turns out to be misspecified. It is also clear that a purposive sample is not suitable for many other potential uses of the survey data, such as micro simulations of econometric models, or unplanned domain estimation, *etc.*. Other times the narrow focus on the population total may be too weak as in the case of homogenous population, with independent units of common mean and variance, where any non-informative sampling design is as good as another. The lack of a model-based theory for the variety of randomization designs is understandable because model-based inferences of population totals in principle do not require probability sampling, as long as the data are gathered noninformatively (Rubin, 1976; Sugden and Smith, 1984). Probability sampling is viewed as a robust and impartial way of achieving noninformative samples.

Neither is the situation satisfactory under the design-based approach. Särndal, Swensson, and Wretman (1992) gave a unified account of the model-assisted design-based estimation for finite populations. Under the ratio model with variances proportional to the auxiliary sizes, the general regression estimator (GREG) becomes the classic ratio estimator (Cochran, 1977, Chapter 6), provided the units are selected with equal probability, which is almost never applied in populations with size-dependent variances. The approximate optimal design for the GREG under this particular variance structure is to select the units with probabilities proportional to the square root of the auxiliary sizes, which of course yields a GREG that is different from the ratio estimator in return. The same conflict exists when the variance is assumed to be constant: the classic ratio estimator is a GREG only if the units are selected with probabilities proportional to the auxiliary sizes, whereas the optimal design for the GREG is equal-probability sampling in this case. The common practice of combining unequal probability sampling with the classic ratio estimator is therefore inconsequent from a design-based point of view.

There are many such examples in survey sampling where tried-and-trusted practices are apparent anachronisms of the theories of both schools. Take e.g. the creation of take-none units in business surveys, i.e. the ones that are excluded from the sample by design. For model-based optimal prediction of the population total, the population is divided into the take-alls (i.e the self-inclusion units) and the take-nones through purposive selection. Under the design-based approach, sampling with probability proportional to some size variable may divide the population into the take-alls and the take-somes (i.e. the units having inclusion probabilities strictly between 0 and 1), provided sufficient variation in the size variable. In neither case, however, there is theoretical ground for distinguishing between the take-somes and the take-nones.

Now, probability sampling can be introduced under the model-based approach provided concerns other than the optimal prediction of the population total are included as design criteria. For instance, Valliant, Dorfman, and Royall (2000) gave a detailed account of balanced samples, which provide bias protection when the true linear predictor is a polynomial of a single auxiliary

variable. Probability sampling is still not a theoretical necessity. But it is used as practical means to 'zoom in' towards the various balanced samples. However, there remains considerable gap between the theory of balanced samples and the sampling practice, e.g. with respect to some of the issues mentioned above.

A different approach is studied in the sequals. We place emphasis on general database-like uses of survey data, in addition to the prediction of population totals (or means). This leads us to consider individual prediction, which is descriptive inference at the most dis-aggregated level, in addition to the prediction of population total, which is descriptive inference at the most aggregated level. Since, given the sample, the conditional prediction error is zero for the selected units but positive for the units outside of the sample, it is possible to use the sampling design to control the *unconditional* individual prediction MSEs. This immediately raises the need for probability sampling. If we consider the individual unconditional prediction MSE as a measure of expected sample information about the corresponding unit, the sampling design becomes crucial for the distribution of this information over the population. To facilitate the exposition, we focus on equal individual prediction, and derive equal prediction designs for linear regression populations as well as clustered populations under the intracluster correlations model. We notice that, while equal prediction seems a natural choice for multiple uses of survey data, it is by no means the only criterion that may be considered. Thus, one should treat the particular designs in this paper as examples of a general approach, rather than explicit guidelines to be imposed in sampling practice.

It turns out that balancing between optimal prediction of the population total and control over individual prediction provides a fruitful model-based approach to sampling design. Apart from raising the need for probability sampling in general, it leads naturally to a number of important design features that have been firmly established in the real world of sampling, including the use of simple random sampling for homogeneous populations and unequal probability sampling otherwise, the division of a business population into the take-all, take-some and take-none units, the most common two-stage sampling designs, the use of stratification with proportional allocation, *etc.*. At the same time this enables us to give an appraisal of these methods from a prediction point of view.

The rest of the paper is organized as follows. In Section 2 we lay out the basic theory for linear regression models. In Section 3, we consider sampling designs under the ratio model, which is the most common situation in business surveys. In Section 4, we present some general results for the intracluster correlations model. In Section 5, we study sampling designs for clustered populations with common mean and variance, which is the simplest model for clustered populations. We consider both two-stage cluster sampling and direct sampling of elements, depending on the sampling frame and mode of data collection available. Finally, Section 6 contains a summary and some discussions.

2 Linear regression population

2.1 Prediction of population total

Denote by $U = \{1, ..., N\}$ the finite population of N units. Consider the following linear regression model for the population

$$y_i = x_i^T \beta + e_i \qquad \text{where} \quad E(e_i) = 0 \quad \text{and} \quad V(e_i) = \sigma_i^2 \quad \text{and} \quad Cov(e_i, e_j) = 0 \qquad (1)$$

for $i \neq j \in U$. The independence assumption makes a special case of the general linear model (Valliant, Dorfman, and Royall, 2000, Theorem 2.1.1). The corresponding best linear unbiased predictor (BLUP) of the population total, denoted by $Y = \sum_{i \in U} y_i$, is given by

$$\tilde{Y} = \sum_{i \in s} y_i + X_r^T \tilde{\beta}$$

where $\tilde{\beta} = (\sum_{i \in s} x_i x_i^T / \sigma_i^2)^{-1} (\sum_{i \in s} x_i y_i / \sigma_i^2)$, and $r = U \setminus s$ contains the units outside the sample, and $X_r = \sum_{k \in r} x_k$. The conditional prediction MSE given the sample is

$$\Delta_r = \sum_{k \in r} \sigma_k^2 + X_r^T \gamma_s^{-1} X_r$$

where $\gamma_s = \sum_{i \in s} \gamma_i$, and $\gamma_i = x_i x_i^T / \sigma_i^2$.

Thus, for optimal prediction of Y one would choose, with certainty, the particular sample s which minimizes Δ_r over all possible samples, i.e. purposive selection. For example, under the ratio model with a single covariate x_i and variance $\sigma_i^2 \propto x_i^a$ for some constant $a \geq 0$, purposive selection leads to the cutoff sample of the n units having the largest x-values (Royall, 1970), provided $a \leq 2$. On the other hand, in the special case of $x_i = 1$ and $\sigma_i^2 = \sigma^2$, we have $\Delta_r = N(N-n)\sigma^2/n$, which is a constant of the sample, such that optimal prediction of Y not at all depends on the sampling design, as long as it is noninformative.

For the unconditional MSE of the BLUP under probability sampling, we have

$$\begin{aligned} \Delta_r &= \sum_{k \in r} \sigma_k^2 + \sum_{k \in r} x_k^T \gamma_s^{-1} x_k + \sum_{k \neq j \in r} x_k^T \gamma_s^{-1} x_j \\ &= \sum_{k \in U} (1 - I_k)(\sigma_k^2 + x_k^T \gamma_s^{-1} x_k) + \sum_{k \neq j \in U} (1 - I_k)(1 - I_j) x_k^T \gamma_s^{-1} x_j \end{aligned}$$

where $I_k = 1$ if $k \in s$ and $I_k = 0$ if $k \in r$. An approximation to the MSE is then given by

$$\text{MSE} \approx \sum_{k \in U} (1 - \pi_k)(\sigma_k^2 + x_k^T \Gamma^{-1} x_k) + \sum_{k \neq j \in U} (1 - \pi_k - \pi_j + \pi_{kj}) x_k^T \Gamma^{-1} x_j$$

where π_k is the inclusion probability of the kth unit, and π_{ij} is the joint inclusion probability of the kth and jth units, and $\Gamma = \sum_{k \in U} \pi_k x_k x_k^T / \sigma_k^2$.

2.2 Individual prediction

Consider now individual prediction under the linear model (1). For any $k \notin s$, the BLUP is given by $\tilde{Y}_k = x_k^T \tilde{\beta}$, with $\tilde{\beta}$ given above. The conditional prediction MSE of $\tilde{Y}_k$ is

$$\Delta_k = \sigma_k^2 + h_k \qquad \text{where} \quad h_k = x_k^T \gamma_s^{-1} x_k$$

Let E_p denote expectation with respect to the sampling design. The unconditional MSE of the BLUP, i.e. the expectation of Δ_k with respective to the sampling, is given as

$$\begin{aligned} \text{MSE}_k &= \sum_{s;k\in s} p(s) \cdot 0 + \sum_{s;k\notin s} p(s)\Delta_k = (1-\pi_k) \sum_{s;k\notin s} P(S=s|k \notin s)\ \Delta_k \\ &= (1-\pi_k) E_p(\Delta_k | k \notin s) = (1-\pi_k)(\sigma_k^2 + H_k) \end{aligned}$$

where $p(s) = P(S=s)$ is the probability of selecting the sample s, and $H_k = E_p(h_k|k \notin s)$. Notice that $(1-\pi_k)\sigma_k^2$ is the MSE of the best predictor (BP) $x_k^T\beta$ provided β is known, and the additional term $(1-\pi_k)H_k$ is due to the estimation of β.

Equal prediction accuracy implies that $\text{MSE}_k = \lambda$ for some constant λ, or $(1-\pi_k) = \lambda w_k$ where $w_k^{-1} = \sigma_k^2 + H_k$. Provided $\sum_{i\in U} \pi_i = n$, we have $\lambda = (1-n/N)\bar{w}^{-1}$ where $\bar{w} = \sum_{i\in U} w_i/N$, and

$$\pi_k = 1 - (1-n/N) w_k/\bar{w} \tag{2}$$

The inclusion probabilities (2) takes into account the estimation of the regression coefficients. If we ignore this piece of uncertainty, then we arrive at the inclusion probabilities that would have yielded equal prediction by the BP, denoted by

$$\pi_k^0 = 1 - (1-n/N)\{\sigma_k^{-2}/(\sum_{i\in U} \sigma_i^{-2}/N)\}$$

Consider the special of $x_i = 1$ and $\sigma_i^2 = \sigma^2$. We have $h_k = H_k = 1/n$ and $w_k = \bar{w}$, such that $\pi_k = \pi_k^0 = n/N$. That is, equal prediction implies equal probability sampling, which theoretically justifies the intuition behind simple random sampling (srs) from any homogeneous population with common mean and variance. Otherwise, equal prediction requires *unequal* probability sampling. Generally, unequal probability sampling is needed if we wish to control the MSEs in any unequal way, say, $\text{MSE}_k = \lambda_h$ for $k \in U_h$ and $U = \cup_{h=1}^H U_h$, or $\text{MSE}_k a_k = \lambda$ for fixed constants $\{a_k; k \in U\}$. Next, consider single x_i and $\sigma_i^2 \propto x_i^2$, we have $h_k = H_k = \sigma_k^2/n$ and $w_k^{-1} = \sigma_k^2(1+1/n)$, such that $\pi_k = \pi_k^0$. That is, the design is the same whether knowing β or not, which makes sense because the variance of $\tilde{\beta}$ is a constant of the sample.

In general, since H_k depends on the π_i's, the inclusion probabilities are not explicitly given by (2). Consider the inverse of a square matrix as a smooth function of its elements, we obtain $E_p(\gamma_s|k \notin s)^{-1}$ as a first-order Taylor linear approximation to $E_p(\gamma_s^{-1}|k \notin s)$. We have

$$E_p(\gamma_s|k \notin s) = \sum_{i\in U_{(k)}} \gamma_i P(i \in s|k \notin s) = \sum_{i\in U_{(k)}} \gamma_i(\pi_i - \pi_{ik})/(1-\pi_k)$$

where $U_{(k)} = U \setminus \{k\}$. Thus, H_k depends on the π_{ik}'s as well. In Poisson sampling (PS), we have $\pi_{ik} = \pi_i\pi_k$ for $i \neq k$, which is convenient since we then have

$$E_p(\gamma_s|k \notin s) = \sum_{i\in U_{(k)}} \pi_i\gamma_i \qquad \text{and} \qquad H_k \doteq x_k^T (\sum_{i\in U_{(k)}} \pi_i\gamma_i)^{-1} x_k$$

A drawback with the PS is that the sample size is not fixed. Approximate PS with fixed sample size can be achieved by the method of sequential Poisson sampling (Ohlsson, 1998). Another potentially useful approximation of π_{ij} in terms of the first-order inclusion probabilities was given by Hartley and Rao (1962) for systematic πps sampling based on random listing.

Sufficient conditions for solutions to (2), viewed as a fixed points equation of the π_k's, follow as a special case of the Contraction Mapping Theorem (e.g. Ortega, 1972). In particular, for the existence of a set of proper solutions, the right-hand side should map any set of proper π_k's onto the interval $(0, 1)$. Now that the w_k's are strictly positive, π_k can be arbitrarily close but never attain the unity. However, negative values arise whenever $w_k/\bar{w} > N/(N-n)$, which is easily the case when σ_k^2 is small. In this way, equal individual prediction leads to the creation of the *take-none* units. The primary reason is that these units have so small 'intrinsic variation' (i.e. σ_k^2) compared to the rest of the units, that the prediction remains less uncertain about them even if they are excluded from the sample by design.

Numerically, we set $\pi_k = 0$ for the take-none units which are then removed from the sampling frame. Proper inclusion probabilities are sought for the units that remain in the frame, i.e. the *take-some* units. We may need to repeat the adjustment several times before the take-some units are settled and the corresponding π_k's found for them. Whenever take-none units are generated in this way we need to check their uniqueness. This can be done by varying the starting values. Some obvious choices include π_k^0 for equal BP design, or equal probability n/N, or probabilities proportional to a chosen size variable. Notice that, for a given population, the take-none units depend on the variance assumption as well as the sample size.

3 Ratio regression population

3.1 Constrained equal prediction design

The ratio model is a special case of the linear regression model (1). It is often used as a reasonable model for business survey planning. As mentioned before, for optimal prediction of the population total, the purposive selection amounts to take the n units having the largest x-values, provided $0 \leq a \leq 2$. Such an extreme design, however, is only used in rare cases. In practice, one finds typically in business surveys some *constrained* probability sampling design as follows: (i) the population is divided into the take-none, the take-some and the take-all units; (ii) the take-some units are selected either using a probability proportional to size (pps) scheme or stratified srs. Indeed, the stratified srs design can be formed to emulate the pps design (Wright, 1983). Both the take-alls and the take-nones are parts of the constraint.

Now, the cut-off limit between the take-all and take-some units can be explored with respect to the efficiency for the prediction of population total. But when it comes to the cut-off limit between the take-some and take-none units, the choice will apparently be based on experiences or conventions, of course, together with considerations of response burdens and other practical concerns. The choice of the pps scheme depends on the variance assumption. In theory one should select with probability proportional to σ_i (i.e. $x_i^{a/2}$). This is approximately optimal for the GREG from a design-based perspective (Särndal, Swensson, and Wretman, 1992, Result 12.2.1). Whereas it makes the first step towards the so-called root(v) weighted balance (Valliant, Dorfman, and Royall, 2000, Chapter 3), where v denotes the individual variance. Typically, one

assumes the variance to be proportional to either x or x^2. We refer to the pps scheme as the *root pps (rpps)* design if the probability is proportional to $\sqrt{x}$, and the (direct) pps design if the probability is proportional to x.

There are thus at least three choices one needs to make, i.e. the cutoff limits between the three sub-populations and the variance parameter a. It is possible to explore these issues in terms of a balance between optimal prediction for the population total and control of the individual prediction. Take first the equal prediction design. As explained before, given sufficient variation in the individual variances, equal prediction leads to the creation of the take-none units and, thereby, a theoretical cut-off limit between the take-none and the take-some units. Since the largest units have lower inclusion probabilities than in the purposive selection, the equal prediction design entails loss of efficiency for the population total. The efficiency can easily be improved by, firstly, imposing a user-specified number of take-all units and, then, applying the equal prediction approach to the adjusted frame. Such a *constrained* equal prediction (cep) design leads to the creation of the take-all, take-some and take-none units, where the take-some units will receive unequal inclusion probabilities, depending on the choice of two design parameters: the number of take-all units, denoted by N_1, and the variance parameter η (i.e. assuming $\sigma_i^2 \propto x_i^\eta$ at design stage). For any fixed choice of η, the cep designs can be arranged in a nested set according to N_1 and studied in a systematic fashion.

3.2 An example based on Norwegian business register data

To illustrate, we use a data set (of 4 industrial groups) extracted from the Norwegian business register, where $N = 5077$. Let x_i be the number of employees plus 1, which will be used as an allround measure of the size of the business units. Table 1 gives the main characteristics of the skewed distribution of the x-values in the population.

Table 1: Characteristics of 4 industrial groups from the Norwegian business register

		Quantile of x_i								
$\min(x_i)$	0.10	0.25	0.50	0.75	0.90	0.95	0.975	0.99	$\max(x_i)$	N
1	1	2	2	5	15	33	80	178	1737	5077

Table 2 gives the theoretical cutoff limits between the take-none and the take-some units under the unconstrained equal prediction design (i.e. $N_1 = 0$), for various combinations of (η, f), where η is the variance parameter and $f = n/N$ is the sampling fraction. The take-nones are the smallest units in the population. Let N_0 be the number of take-none units; and let X_0 be the total of x from the take-none units. Both N_0 and X_0 are increasing in η and decreasing in f. For instance, at $\eta = 1$ and $f = 0.2$, the take-none sub-population contains about 68% of all the units, and a coverage of about 11% in terms of x, i.e. X_0/X.

In the top-left plot of Figure 1, the inclusion probabilities of the direct pps design are compared to those of the cep design with variance assumption $\sigma_i^2 \propto x_i$ and $N_1 = 310$, which is the number of self-inclusion units implied by the direct pps scheme. The sampling fraction is 20%. The inclusion probability is seen to increase quicker in x for the take-some units under the

Table 2: Proportion of take-none units N_0/N and coverage X_0/X under equal prediction design with variance assumption $\sigma_i^2 \propto x_i^\eta$ and sampling fraction $f = n/N$.

	N_0/N				X_0/X			
f	0.05	0.1	0.2	0.3	0.05	0.1	0.2	0.3
$\eta = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\eta = 0.5$	0.866	0.732	0.131	0.131	0.207	0.130	0.011	0.011
$\eta = 1$	0.907	0.824	0.681	0.131	0.252	0.175	0.112	0.011
$\eta = 1.5$	0.920	0.842	0.681	0.596	0.273	0.187	0.112	0.091
$\eta = 2$	0.928	0.854	0.732	0.596	0.287	0.196	0.130	0.091

cep design than under the pps design. In the rest three plots, the individual prediction MSEs by the pps and purposive selections are compared to those under the cep designs with respective design variance parameter $\eta = 1, 1.5$ and 2. The true population variance is set at $\sigma_i^2 \propto x_i^a$ for $a = 1.5$ in all the three cases. While there exist clear differences between the alternative designs for the take-some units (identified by the cep designs), the MSEs for the take-none units vary little from one design to another, which justifies the creation of the take-none units.

More systematic comparisons between the alternative designs are given in Figure 2. Given the population and a fixed sample size ($n = 1015$ and $f = 0.2$ in this case), the rpps design implies 92 self-inclusion units (i.e. $N_1 = 92$). It is approximately optimal for the GREG under the variance assumption $\sigma_i^2 \propto x_i$. In the top-left plot of Figure 2, the rpps design is compared to the cep design with the same constraint (i.e. $N_1 = 92$) and design variance assumption, with respect to the following measures: (a) the TMSE ratio, i.e. the MSE of the BLUP for population total under the cep design against that under the rpps design, (b) the MMSE ratio, i.e. the mean of all the individual prediction MSEs under the cep design against that under the rpps design, (c) the coefficient of variation (CV) of the individual MSEs under the cep design, i.e. $\{\sum_{k\in U}(\text{MSE}_k - \text{MMSE})^2/(N-1)\}^{1/2}/\text{MMSE}$, and (d) the CV of the individual MSEs under the rpps design. These four measures are evaluated as the underlying population variance structure $\sigma_i^2 \propto x_i^a$ varies for $a \in [0, 2]$. It is seen that the cep design is considerably more efficient than the rpps design, both in terms of the TMSE and MMSE ratios, especially for $a \geq 1$. Also, the variation among the individual prediction MSEs is much less under the cep design than the rpps design for $a > 0.5$. Clearly, the rpps design assigns unnecessarily low inclusion probabilities to the larger ones among the take-some units for this population.

Next, in the top-right plot, the cep design (with $N_1 = 310$ and $\eta = 1$) is compared to the pps design, which implies 310 self-inclusion units in this case. The cep design is more efficient, because the largest take-some units have higher inclusion probabilities (top-left plot, Figure 1). For instance, when the true variance is $\sigma_i^2 \propto x_i^{1.5}$, the pps design entails about 25% loss of efficiency in terms of the MSE for total, and about 20% loss in terms of MMSE. Also, the variation among the individual MSEs is much lower under the cep design for $a > 0.5$ — see the top-right plot of Figure 1 for details at $a = 1.5$.

The pps design is approximately optimal for the GREG under the variance assumption $\sigma_i^2 \propto x_i^2$. In the bottom-left plot of Figure 2, we compare it to the cep design with $N_1 = 310$ and

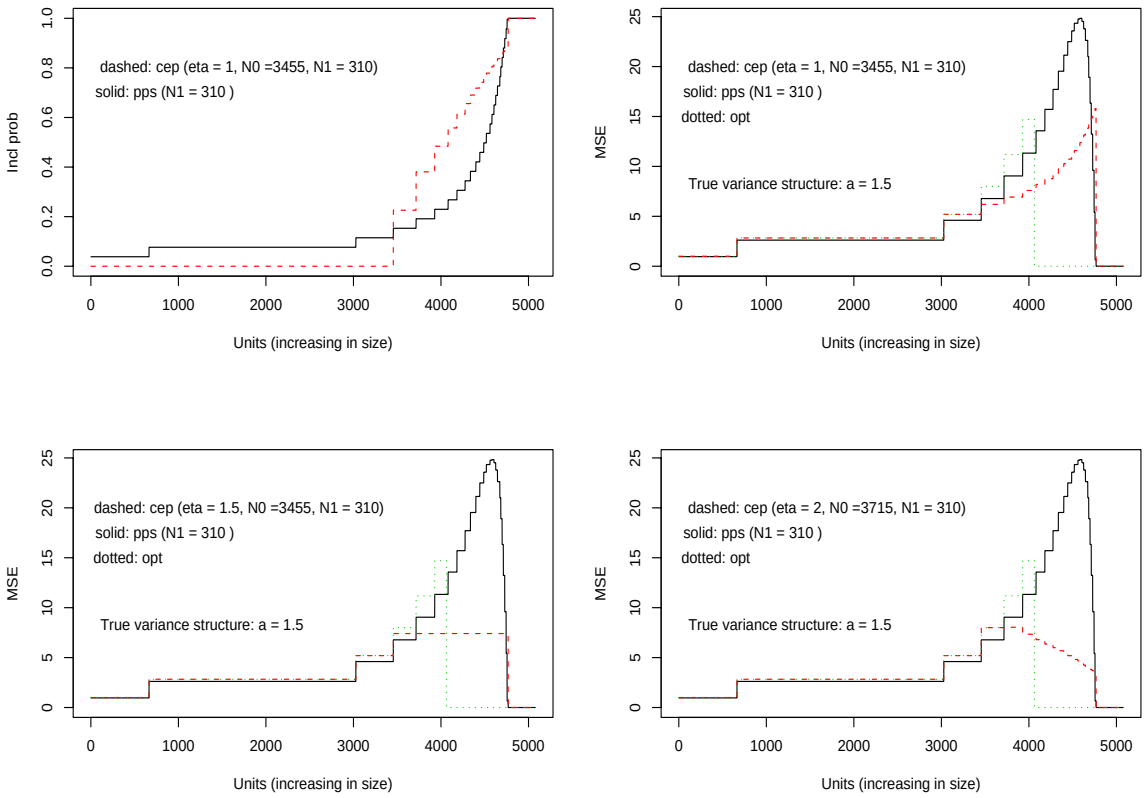

Figure 1: Top left: Inclusion probabilities by cep and pps design. Top right, bottom left and bottom right: Individual prediction MSE by cep, pps and purposive selection (opt). Sampling fraction 20%.

$\eta = 2$. Raising the design parameter η from 1 to 2 increases the inclusion probabilities of the larger take-some units. The cep design becomes therefore even more efficient. It is interesting to see that the variation among the individual prediction MSEs has become more stable (and reduced) when the true variance structure is $1 \leq a \leq 2$. The reason is apparent from Figure 1.

Finally, the cep design (with $N_1 = 310$ and $\eta = 2$) is compared to the optimal design for population total, i.e. the purposive selection, in the bottom-right plot of Figure 2. The maximum loss of efficiency is about 15% for total and about 30% in terms of the MMSE. It is possible to balance the loss of efficiency against the advantages of probability sampling through the choice of N_1. For instance, at $N_1 = 500$ (i.e. about half of the sample) and $\eta = 2$, the maximum loss of efficiency is reduced to 10% and 20%, respectively.

In summary, balancing between optimal prediction of the population total and control over individual prediction yields nested classes of constrained probability sampling designs, providing theoretical motivations for the common use of such designs in practice. The properties of the standard pps or rpps design can be examined with reference to the cep designs for the given population. This also shows us whether and how potential improvements over the pps designs can be achieved in light of the available prior knowledge of the variance structure. The plots in Figure 2 are especially helpful in situations where one needs to comprise between multiple Y of

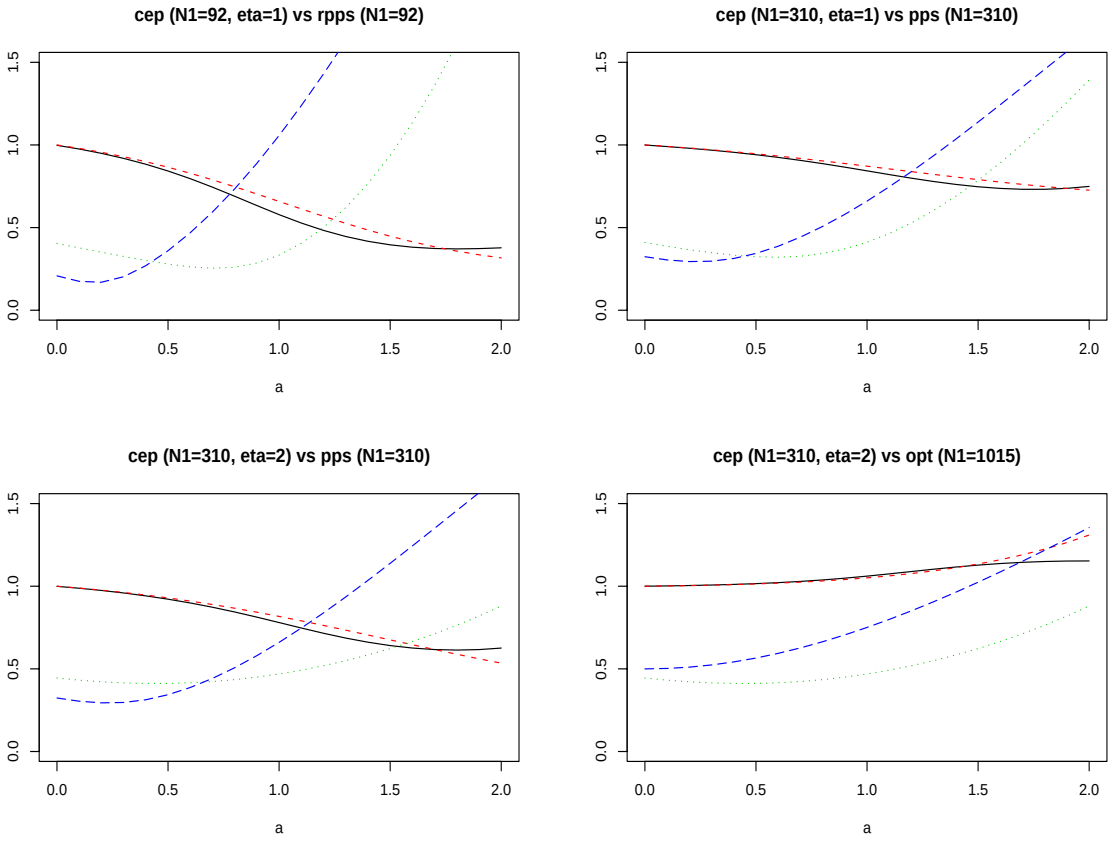

Figure 2: Comparison between cep and alternative designs when the underlying population model varies with respect to the variance structure $\sigma_i^2 \propto x_i^a$: TMSE ratio (solid), MMSE ratio (dashed), CV of individual prediction MSEs under cep design (dotted), and CV of individual prediction MSEs under alternative design (long dashed). Sampling fraction being 20% in all cases.

interest. The different Y's can have different variance inflation measured against x, all of which are now summarized in a single plot for comparison.

4 Clustered population

4.1 Intracluster correlations and variance components

Some degree of clustering among "nearby" units tends to exist in all natural populations. Scott and Smith (1969) and Royall (1976) considered model-based estimation under the following variance assumptions for clustered populations, i.e.

$$V(Y_{ij}) = \sigma_i^2 \qquad \text{and} \qquad Cov(Y_{ij}, Y_{ik}) = \rho_i \sigma_i^2 \qquad \text{and} \qquad Cov(Y_{ij}, Y_{gk}) = 0 \tag{3}$$

where (ij) denotes the jth unit of the ith cluster, for $i = 1, ..., M$ and $j = 1, ..., N_i$ and $\sum_{i=1}^{M} N_i = N$. In particular, the parameter ρ_i is known as the intracluster correlation. Notice

that in standard texts on survey sampling (e.g. Cochran, 1977), N denotes the number of clusters and M denotes the number of elements, contrary to our notation.

Scott and Smith (1969) motivated the assumptions (3) by means of variance components. Suppose that the variance of Y_{ij} is the sum of that of two independent random components, denoted by $\sigma_i^2 = \Omega + \phi_i$, where Ω is the between-cluster variance and ϕ_i is the within-cluster variance. The intracluster correlation is then given by

$$\rho_i = Cov(Y_{ij}, Y_{ik}) / \sqrt{V(Y_{ij})V(Y_{ik})} = \Omega/(\Omega + \phi_i)$$

Such variance components models are standard in small area estimation (Fay and Herriot, 1979). The model (3) appears more general because it allows for negative intracluster correlations, although this is not usual. Indeed, ρ_i is bounded from below by

$$\rho_i \geq -1/(N_i - 1)$$

from noting that the variance of any cluster total must not be negative. Thus, ρ_i is virtually nonnegative for any cluster of reasonable size. Meanwhile, a variance components model can be more general than the intracluster correlation model if the within-cluster variance is unit-specific. Suppose that the variance of Y_{ij} is given by $\Omega + \phi_{ij}$, then the correlation is no longer constant for all pairs of observations from the same cluster. We refer to Rao (2003) for more general variance components models with an emphasis on small area estimation. In this paper we consider only the intracluster correlations model.

When it comes to the mean structure, two special cases are worth noting: (i) auxiliary information may be available cluster-wise as a proxy measure of the cluster mean (or total), (ii) $E(Y_{ij})$ is a constant. The latter is the primary case considered by Scott and Smith (1969) and Royall (1976), and will be studied in Section 5. For the results in this section, we allow a slightly more general mean structure by assuming that $E(Y_{ij})$ is related to a single auxiliary variable through a multiplying constant, i.e.

$$E(Y_{ij}) = x_{ij}\beta \tag{4}$$

We assume single auxiliary variable because this is the most common situation at the design stage. We allow for unit-specific mean because models combining (3) and (4) have been successfully used in survey context (Battese, Harter, and Fuller, 1988).

4.2 Prediction of population total

We start with the prediction of population total. Denote by $\tilde{Y} = \sum_{i=1}^{m} \sum_{j \in s_i} a_{ij} y_{ij}$ the BLUP conditional on s, where s_i is the ith sample cluster for $i = 1, ..., m$. Again, the BLUP and its conditional MSE follow from the theory of general linear model. In the first place, the conditional MSE is minimized only if $a_{ij} = a_i$, such that

$$V(\tilde{Y}_r - Y_r | s) = V_r + \sum_{i=1}^{m} n_i \tau_i^{-1} a_i^2 - 2 \sum_{i=1}^{m} n_i (N_i - n_i) \rho_i \sigma_i^2 a_i$$

where n_i is the size of s_i, and Y_r is the total of y outside the sample, and $\tilde{Y}_r$ is its BLUP, and $\tau_i^{-1} = (1-\rho_i)\sigma_i^2 + n_i\rho_i\sigma_i^2 = \sigma_i^2[1+(n_i-1)\rho_i]$, and $V_r = \sum_{i=1}^{M}(N_i-n_i)\sigma^2\{1+(N_i-n_i-1)\rho_i\}$ is the variance of Y_r. Next, the Lagrange method gives us

$$a_i = b_r\bar{x}_i\tau_i/(\sum_{g=1}^{m} n_g\bar{x}_g^2\tau_g) + (N_i-n_i)\rho_i\sigma_i^2\tau_i$$

where $b_r = x_r - \sum_i n_i(N_i-n_i)\bar{x}_i\rho_i\sigma_i^2\tau_i$, and x_r is the total of x outside the sample, and $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$. The conditional prediction MSE is then

$$\Delta_r = V_r + b_r^2/(\sum_{i=1}^{m} n_i\bar{x}_i^2\tau_i) - \sum_{i=1}^{m} n_i(N_i-n_i)^2\rho_i^2\sigma_i^4\tau_i$$

It is unclear how the purposive selection looks like in general, although in principle the solution can be determined numerically for the given population by going through all possible sample cluster sizes $\mathbf{n} = (n_1, ..., n_M)$ and sample cluster means $\bar{x}_i$. (The problem will be dealt with more closely under the common mean model in Section 5.) When it comes to the unconditional prediction MSE, i.e. $E_p(\Delta_r)$, we have

$$E_p(V_r) = \sum_{i=1}^{M}\sigma_i^2\sum_{j=1}^{N_i}(1-\pi_{ij}) + \sum_{i=1}^{M}\rho_i\sigma_i^2\sum_{j\neq k=1}^{N_i}(1-\pi_{ij}-\pi_{ik}+\pi_{ij,ik})$$

where π_{ij} is the inclusion probability of (ij), and $\pi_{ij,ik}$ is the joint inclusion probability of (ij) and (ik). The rest two terms of the MSE can easily be approximated when $\mathbf{n}$ is fixed by the design. Otherwise, Monte Carlo evaluation provides a straightforward option, although it can be computationally intensive.

4.3 Individual prediction

For any $(gk) \notin s$, consider the BLUP $\tilde{Y}_{gk} = \sum_{i=1}^{m}\sum_{j\in s_i} a_{ij}y_{ij}$. Unbiased prediction conditional on s implies that $\sum_i\sum_j a_{ij}x_{ij} = x_{gk}$, and the conditional MSE under the model (3) is

$$V(\tilde{Y}_{gk} - Y_{gk}|s) = \sigma_g^2 + \sum_i(1-\rho_i)\sigma_i^2(\sum_j a_{ij}^2) + \sum_i\rho_i\sigma_i^2(\sum_j a_{ij})^2 - 2\rho_g\sigma_g^2(\sum_j a_{gj})$$

because $V(\sum_j a_{ij}Y_{ij}) = (1-\rho_i)\sigma_i^2(\sum_j a_{ij}^2) + \rho_i\sigma_i^2(\sum_j a_{ij})^2$. Notice that the last term on the righ-hand side exists only if the cluster g is represented in the sample. It follows that, for any value of $\sum_j a_{ij}$, the conditional MSE is minimized only if $a_{ij} = a_i = \sum_j a_{ij}/n_i$. Next, by the Lagrange method, we find that

$$a_i = b_{gk}\bar{x}_i\tau_i/(\sum_l n_l\bar{x}_l^2\tau_l) + \delta_{ig}\rho_g\sigma_g^2\tau_g$$

where $b_{gk} = x_{gk} - n_g\bar{x}_g\rho_g\sigma_g^2\tau_g$, and $\delta_{ig} = 1$ if $i = g$ and 0 otherwise. The conditional MSE of the BLUP $\tilde{Y}_{gk}$ is then given by

$$\Delta_{gk} = \sigma_g^2 + h_{gk} \qquad \text{where} \quad h_{gk} = b_{gk}^2/(\sum_{i=1}^{m} n_i\bar{x}_i^2\tau_i) - n_g\rho_g^2\sigma_g^4\tau_g$$

It follows that the unconditional prediction MSE of the BLUP is

$$\text{MSE}_{gk} = (1 - \pi_{gk})(\sigma_g^2 + H_{gk}) \qquad \text{where} \quad H_{gk} = E_p\{h_{gk}|(gk) \notin s\}$$

Provided $\sum_{(ij)\in U} \pi_{ij} = n$, equal prediction implies the following fixed-points equation

$$\pi_{gk} = 1 - (1 - n/N)w_{gk}/\bar{w} \tag{5}$$

where $w_{gk}^{-1} = \sigma_g^2 + H_{gk}$ and $\bar{w} = \sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij}/N$. To actually derive the π_{gk}'s we need to set the ρ_i's, which is usually difficult. An immediate use of the results of this section is then to check how a particular design works as the population intracluster correlations vary in some plausible ways.

5 Common mean population

5.1 Common parameter model

The common mean assumption is an important special case, especially when studying the sampling design for general purposes. It follows that the expectation of the cluster total is proportional to the size of the cluster. There is clearly a connection to the ratio model when a unit with mean $x_i\beta$ can be a 'cluster' made up of x_i elements, all having a common mean β. Now, when the elements have a common mean, it is often reasonable to assume that they also have the same variance, i.e. $\sigma_{ij}^2 = \sigma^2$. Indeed, the *common parameter model* assumes that $\rho_i = \rho$, which can often be motivated by means of variance components, i.e. $\sigma^2 = \Omega + \phi$. But it is easy to conceive other covariance structures. For instance, under what may be referred to as the *clustered ratio model*, we assume that the cluster total has variance proportional to x_i^a. The case of $a = 1$ is then equivalent to the assumption of $\rho_i = 0$. Whereas $1 < a < 2$ implies that $\rho_i = (x_i^{a-1} - 1)/(x_i - 1)$ if $\sigma_{ij}^2 = \sigma^2$, i.e. positive intracluster correlation between the elements that decreases towards 0 as the cluster size increases. Notice that such covariance structures can not arise from a variance components assumption.

Table 3: Characteristics of municipality household numbers in Norwegian Census 2001.

	Quantile of N_i								
$\min(N_i)$	0.10	0.25	0.50	0.75	0.90	0.95	0.99	$\max(N_i)$	M
94	541	917	1805	3826	8241	14098	42518	266856	434

In the sequals, we will focus on the common parameter model. Our approach will be illustrated in the setting of the so-called Master Sampling Plan for household surveys, which serves as a point of departure for all surveys of households and persons conducted at Statistics Norway. The design clusters are the municipalities with a total $M = 434$, and the elements are the households with a total $N = 1.962$ million, based on the Norwegian census 2001. Table 3 gives the main characteristics of the distribution of N_i, i.e. the population cluster sizes.

We shall consider both two-stage cluster sampling and direct sampling of elements, because both types are being used at the statistical offices, depending on the available sampling frame and mode of data collection. Two- or multistage sampling designs are necessary when a sampling frame does not exist for the ultimate sampling units, but are more readily available for the primary sampling units. They may also be preferred due to cost considerations or other procedural concerns that are important in practice. A key factor here is the mode of data collection. Face-to-face interviews call for careful planning at the design stage, where two- or multistage sampling can greatly reduce the cost required. Indeed, the desire to equalize work loads for interviewers often leads to equal allocation of the ultimate sampling units to the sample clusters. The availability of a complete sampling frame for the ultimate sampling units, as well as alternative mode of data collection such as computer assisted telephone interview (CATI), allows us to sample elements directly without prohibitive increase in cost. For example, the Norwegian Labour Force Survey uses a single-stage sampling design. Direct sampling of elements is generally more efficient than cluster sampling.

The variance assumptions (3) is often used for the study of two-stage sampling designs. It is less often used when studying direct sampling of elements. Stratified sampling with the municipalities as the strata is probably more standard. However, detailed statistics are often of interest either at the municipality level or some regional level, where the regions consist of neighbouring municipalities. Variance component models with the municipalities being the "small areas" are frequently used for such purposes (Rao, 2003), which have the same variance structure. Moreover, it turns out that the choice is not critical for our main findings there.

We notice that in reality Master sampling plan will surely involve some form of stratification among the municipalities. To limit the scope of the investigation, however, we shall keep away from this issue and only come back to it in the discussions at the end of the paper. It is also worth mentioning that the results presented in this Section need to be distinguished from the situation where the clusters are much smaller and the number of clusters much larger, e.g. when the households themselves are treated as the clusters, and the members of the households as the elements. Again, more discussions on this will be given later.

5.2 Two-stage cluster sampling

We consider two-stage sampling designs with equal sample cluster sizes, i.e. $n_i = n/m$, where m is the number of sample clusters. For simplicity and without loss of generality, we assume that n/m is naturally an integer. Under the common parameter model, the conditional prediction MSE for the population total, given any such two-stage cluster sample, is

$$\Delta_r \propto (N-n)(1-\rho) + \sum_{i=1}^{M}(1-\gamma)(N_i - I_i n/m)^2\rho + \{\sum_{i=1}^{M}(1-\gamma)(N_i - I_i n/m)\}^2/(m\psi)$$

where $\gamma = (n/m)\rho/\{(n/m)\rho + 1 - \rho\}$, and $\psi = (n/m)/\{(n/m)\rho + 1 - \rho\}$, and $I_i = 1$ if the ith cluster is selected and $I_i = 0$ otherwise. Royall (1976) showed that the purposive first-stage sample consists of the m largest clusters in the population. Nothing specific is implied for sampling within the clusters: any noninformative scheme is as good as another.

For individual prediction of $(gk) \notin s$, we have

$$\Delta_{gk} = \Delta_g \propto 1 + (1 - I_g\gamma)^2/(m\psi) - I_g\gamma\rho = \{1 + 1/(m\psi)\} - I_g D$$

where $D = \gamma(2-\gamma)/(m\psi) + \gamma\rho$. The unconditional prediction MSE is then given by

$$\mathsf{MSE}_{gk} \propto (1-\pi_{gk})\{1+1/(m\psi)\} - (\pi_g - \pi_{gk})D$$

because

$$E_p(I_g|I_{gk}=0) = P(I_g=1|I_{gk}=0) = (\pi_g - \pi_{gk})/(1-\pi_{gk})$$

where π_g is the first-stage selection probability of the gth cluster. It is seen that equal individual prediction implies equal-probability selection within the clusters regardless of π_g, which provides a theoretical justification for the standard practice. Let the second-stage inclusion probability be $p_g = \pi_{gk}/\pi_g = (n/m)/N_g$ for $(gk) \in U_g$. Provided $\sum_{i=1}^{M} \pi_i = m$, we have

$$\pi_g = (m/M)(\xi_g/\bar{\xi}) \tag{6}$$

where

$$\xi_g^{-1} = \{1+1/(m\psi)\}p_g + D(1-p_g) \qquad \text{and} \qquad \bar{\xi} = \sum_{i=1}^{M} \xi_i/M$$

The following observations are worth noting. (1) Equal prediction implies equal probability selection in the case of $N_i = N/M$. (2) In the case of $\rho = 0$, $\xi_g^{-1} \propto p_g$ such that $\pi_g = mN_i/N$. That is, equal prediction implies first-stage pps sampling in the case of independent elements. The pps-srs two-stage design is a so-called "equal probability selection method" (epsem, Kish, 1965) where $\pi_{ij} = n/N$. (3) In the case of $\rho = 1$, $\xi_g^{-1} \propto (1+1/m)$ such that $\pi_g = m/M$. That is, first-stage equal-probability selection, giving rise to the srs-srs (i.e. twice equal-probability) two-stage design. Together, the pps-srs and srs-srs designs provide much of the basis for two-stage sampling in practice (Cochran, 1977). (4) Provided $\rho \in (0,1)$, equal prediction implies that the larger clusters have larger ξ_i^{-1} and, thus, larger inclusion probabilities than average (i.e. m/M). However, due to the D-term in ξ_g^{-1}, π_g is not as high as mN_i/N.

Figure 3 compares the pps first-stage selection probabilities to the that of the equal prediction design (epd), with alternative specifications of the intracluster correlation for the Norwegian household population, where $n = 3000$ and $m = 100$. Clearly, the epd under-samples the larger cluster and over-samples the smaller ones, compared to the pps scheme. Indeed, there is very little difference among the selection probabilities for the larger clusters under the epd even in very weakly clustered population, say, $\rho = 0.05$, which converge quickly to the srs scheme. Thus, unconstrained epd entails great loss of efficiency for the population total. We can improve this by adopting a constrained approach as follows: (a) choose a number of self-inclusion clusters, denoted by M_1, and (b) apply the equal prediction design to the remaining take-some clusters. More difficult is the choice of ρ. The results in Figure 3 show that equal prediction implies almost equal first-stage selection probability for the larger clusters. In situations where plausible value of ρ is lacking, one may start by looking at the constrained srs (csrs) schemes.

Systematic comparisons between the alternative first-stage schemes are given in Figure 4 with respect to the TMSE and MMSE ratios, and relative standard deviation (RSD), i.e. standard deviation of the individual MSEs in relation to the overall variance σ^2. Random sampling is used at the second-stage in all the cases. In the top-left plot, the purposive selection is compared to the pps scheme. Clearly, there is a loss of efficiency for the population total due to the departure from the optimal design. Next, in the top-right plot, the pps design is compared to

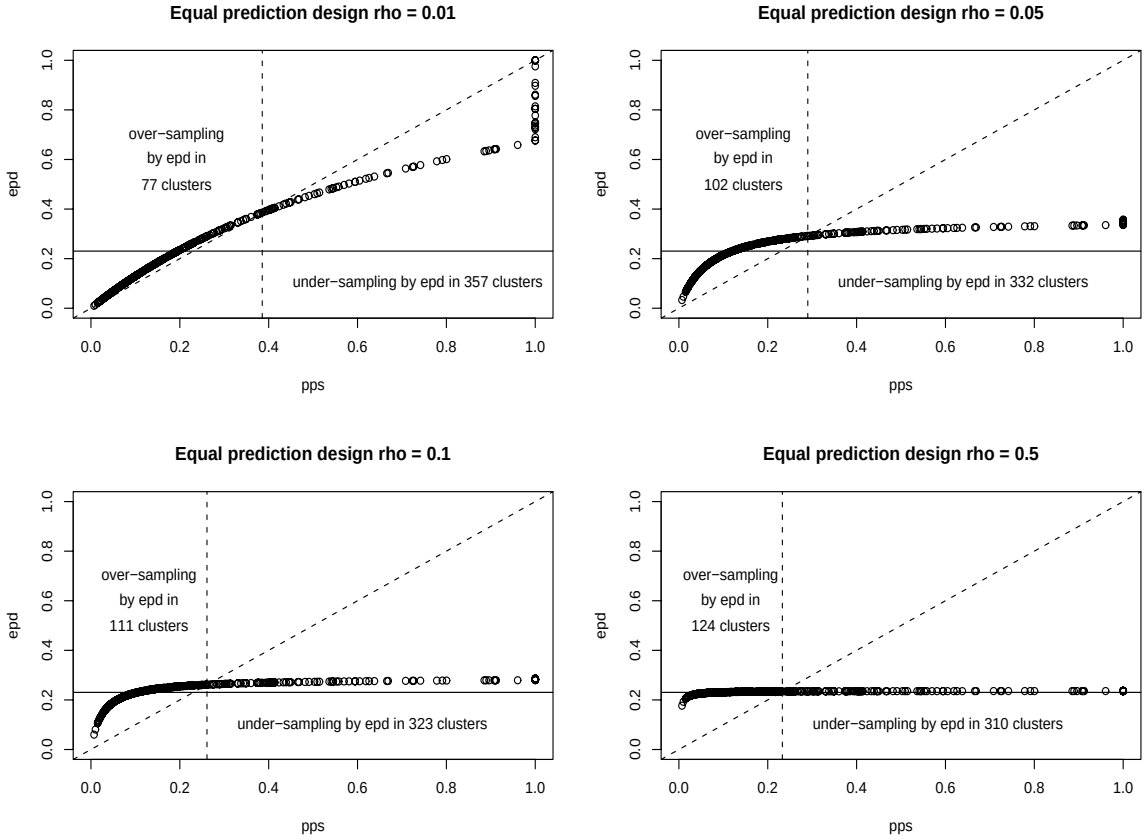

Figure 3: First-stage selection probability by pps and equal prediction design (epd) for Norwegian household population ($m = 100$ and $n = 3000$). Solid horizontal line marks equal probability selection $\pi_i = m/M$. Dashed vertical line marks where pps selection probability exceeds that of epd.

the csrs design with $M_1 = 23$, which is the same number of self-inclusion clusters implied by the pps scheme. Equal probability selection among the take-some clusters leads to further loss of efficiency. We can improve the efficiency of the csrs by increasing M_1. It can be seen from the bottom-left plot that, at $M_1 = 50$ which is half of the number of clusters in the sample, the csrs scheme becomes almost as efficient as the pps scheme throughout the range of ρ.

Finally, in the bottom-right plot, details are given for the cluster-wise individual prediction MSEs under the various first-stage designs in the case of $\rho = 0.2$. The csrs scheme yields almost equal individual prediction for all but the few smallest take-some clusters. The variation is due to the extreme small sizes of the smallest clusters. If we want, we can reduce the variation by, first, grouping the smallest clusters together into a primary sampling unit (PSU) and, then, applying an extra stage of cluster sampling if this regrouped PSU is selected at the first stage. In other words, multistage cluster sampling, with equal probability selection at evey stage, can be applied to appropriately chosen sub-populations to emulate the two-stage equal prediction design. Meanwhile, under the pps first-stage design, the individual MSE decreases with the population size of the take-some cluster. There is thus a trade-off between the efficiency for elements from small and large take-some clusters.

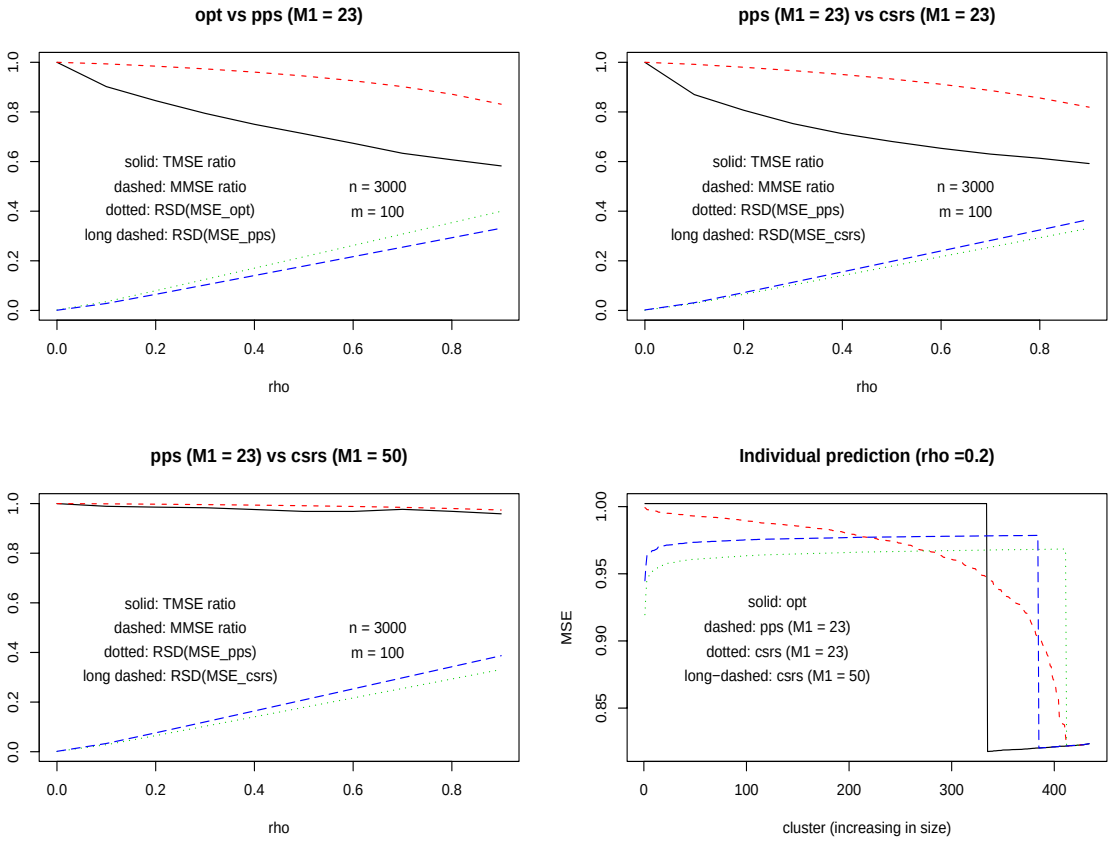

Figure 4: Comparison of alternative first-stage designs for Norwegian household population: purposive selection (opt) vs pps (top-left), pps vs csrs with $M_1 = 23$ (top-right), and pps vs csrs with $M_1 = 50$ (bottom-left). Bottom right: Cluster-wise individual prediction MSE in case of $\rho = 0.2$.

Equal prediction means that the sample information is evenly distributed everywhere in the population on repeated sampling. When the elements are independent with common mean and variance, this implies equal probability sampling. The pps-srs design is an epsem. It is also an equal prediction design when $\rho = 0$. But it will no longer be so provided small departures from independence. Approximate equal prediction can be achieved through the srs-srs design, possibly with extra stages of cluster sampling among the smallest clusters, without knowing the 'true' intracluster correlation. This is an important advantage over the exact epd. It also means that the srs-srs design is robust for multiple Y of interest. The loss of efficiency can be controlled by the number of self-inclusion clusters, leading to the csrs-srs design. Whether or not equal prediction is important in a given situation, a systematic comparison between the nested set of csrs-srs designs and the alternative design can be helpful for making a choice between them.

5.3 Sampling of elements

5.3.1 Prediction of population total

Under the common parameter model, the conditional prediction MSE for the population total is

$$\Delta_r \propto (N-n)(1-\rho) + \sum_{i=1}^{M}(1-\gamma_i)(N_i-n_i)^2\rho + \{\sum_{i=1}^{M}(1-\gamma_i)(N_i-n_i)\}^2/(\sum_{i=1}^{M}\psi_i)$$

where $\gamma_i = n_i\rho/(n_i\rho+1-\rho)$, and $\psi_i = n_i/(n_i\rho+1-\rho)$, based on any sample configuration of $(n_1,...,n_M)$. Since the contribution is the same for any element within a cluster, all purposive samples of elements that minimizes Δ_r have the same allocation of sample cluster sizes, to be referred to as the *purposive allocation*. Thus, optimal prediction of the population total implies that we should treat the clusters as strata at the design stage, despite we are working here under an intracluster correlations model, and apply stratified sampling provided we make the practical extension of the terminology to allow for take-none strata, just as in the business surveys. Nothing specific can be said about the within-stratum sampling: any noninformative scheme is as good as another.

How the purposive allocation looks like is unknown except in two special cases: (i) any allocation is as good as another if $\rho = 0$, and (ii) one element for each of the n largest clusters if $\rho = 1$. Otherwise, provided an exhaustive search through all possible allocations is too much an undertaking, we turn to the following *greedy* algorithm:

1. Sort the clusters such that $N_1 \leq N_2 \leq \cdots \leq N_M$.
2. Assign the first element to the largest cluster, and denote the initial allocation by $\mathbf{n}_{(1)}$.
3. Iterate (a) - (b) for $k = 2,...,n$.
   (a) For each $j = m_k,...,M$, calculate Δ_r that results from allocating the kth element to the jth cluster (i.e. if possible), where m_k is the largest cluster with $n_i = 0$ according to $\mathbf{n}_{(k-1)}$, which is the allocation of the first $k-1$ elements.
   (b) Choose the allocation of the kth element that has the smallest Δ_r from (a). In cases of ties, choose the largest cluster. Set this allocation to be $\mathbf{n}_{(k)}$.
4. Set $\mathbf{n}_{(n)}$ to be the purposive allocation.

The greedy algorithm is not guaranteed to find a global optimum. The outcome can be controlled using additional repeated random search as follows:

1. Randomly select a cluster j with $n_j > 0$ and another cluster k where $k \neq j$.
2. Calculate Δ_r that would result from re-allocating one element from j to k.
3. Accept the re-allocation if the new Δ_r is smaller than before.

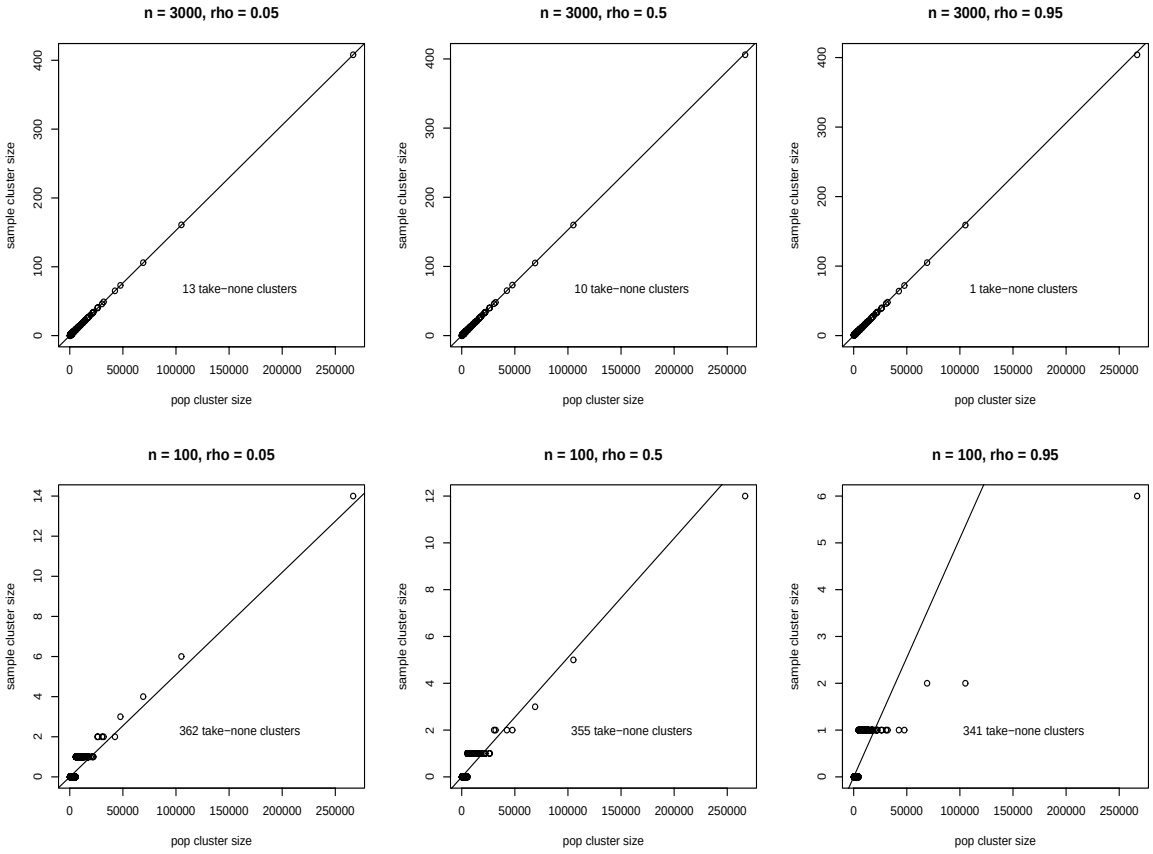

Figure 5: Purposive allocation for Norwegian household population by greedy algorithm, with varying sample size n and design intracluster correlation rho. Solid line marks the proportional allocation.

The outcome of the greedy algorithm seems a plausible solution, if no re-allocation can be found after a fair number (say, 10000) of repeated random search.

Some results of the greedy algorithm for the Norwegian household population are given in Figure 5. The sample size is 3000 in the plots of the top row, and it is 100 in the bottom row. The design intracluster correlation is 0.05 in the first column, and 0.5 in the second column, and 0.95 in the last column. No re-allocation was found after 10^4 random searches in any of the cases shown. Clearly, the most striking feature is the proportionality between the greedy solution and the population cluster size. In the case of $n = 3000$, the purposive allocation is almost strictly proportional for $0.05 \leq \rho \leq 0.95$ (marked by the solid lines in the plots). In other words, the proportional allocation is highly robust as an approximate optimal allocation here. In the case of $n = 100$, where the sample size is much smaller than the number of clusters $(M = 434)$, the proportional allocation holds almost for ρ up to 0.5. For instance, the largest cluster gets assigned 14 elements at $\rho = 0.05$, which is the proportional allocation after rounding. It gets only two fewer at $\rho = 0.5$. At $\rho = 0.95$, the purposive allocation looks more like the theoretical solution at $\rho = 1$: only the 3 largest clusters gets, respectively, 6, 2 and 2 elements each, with the next 90 largest clusters getting one each. In practice, the intracluster correlation is positive but small in many populations, such that the proportional allocation seems robust also here.

The result is interesting because it coincides with the optimal allocation for stratified populations with equal stratum variance (Valliant, Dorfman, and Royall, 2000, Section 6.1). The BLUP under the stratified model is the usual stratified expansion estimator, for which the proportional allocation is well known as the Neyman allocation (Neyman, 1934) in the design-based theory, provided the stratum population variances are equal. There is certainly a connection between the stratified model with equal stratum variance and the common parameter model. When the clusters are considered as strata, the stratified model treats the difference between a cluster mean and the mean of all the cluster means as a fixed effect, whereas the common parameter model treats the same quantity as a random effect, i.e. the variance component assumption. Meanwhile, under the random-effect model for the cluster means, the intracluster correlations are equal only if the stratum variances are equal. In other words, the only difference here is that in one model the cluster mean effect is treated as fixed, whereas in the other it is treated as random. Our results above suggest that this difference in models has little practical consequences on the optimal design for population total, i.e. we should treat the clusters as design strata and apply stratified sampling with proportional allocation no matter which model, with the corresponding BLUP, is used for estimation.

5.3.2 Constrained equal allocation design

When it comes to individual prediction, the conditional prediction MSE for $(gk) \notin s$ is

$$\Delta_{gk} = \Delta_g \propto 1 + h_g \qquad \text{where} \quad h_g = (1-\gamma_g)^2/(\sum_{i=1}^{m} \psi_i) - \gamma_g \rho$$

Clearly, equal prediction requires $\pi_{gk} = \pi_g$, i.e. equal probability sampling within the clusters. Moreover, $\rho = 0$ implies that $\pi_{gk} = n/N$, i.e. equal probability sampling everywhere. Whereas $\rho = 1$ implies that we need at most one element from any cluster. Thus, for $0 < \rho < 1$, equal prediction implies that the expected sample cluster size is lower than nN_i/N for large clusters but higher than that for small clusters.

Since Δ_g depends on the sample only through the sample cluster sizes, equal individual prediction can be achieved straightforwardly using stratified simple random sampling, where the clusters are set as the design strata. Thus, while optimal prediction of population total points towards stratified sampling, equal individual prediction implies random sampling within stratum in addition. Given $\mathbf{n}$, we have $H_g = E_p[h_g|(gk) \notin s, \mathbf{n}] = h_g$, and

$$\mathsf{MSE}_{gk} = \mathsf{MSE}_g \propto (1 - n_g/N_g)(1 + h_g)$$

It follows that, given proportional allocation, the individual prediction MSE is decreasing as the population cluster size N_i increases. Whereas, given *equal* allocation, i.e. $n_i = n/M$ provided $n > M$, the MSE is increasing with N_i because h_g is now a constant across the clusters. In other words, the equal prediction allocation is in general somewhere between proportional and equal allocation. Actually, rather close to the latter in situations where the overall sampling fraction is negligible, because $1 - n_i/N_i = 1 - n/(MN_i)$ will then be very close to 1 except perhaps for the few smallest clusters.

Clearly, equal prediction leads to under-sampling among the elements that belong to the larger clusters and, thus, loss of efficiency for the population total. This can easily be improved

by means of *constrained equal allocation (cea)* as follows: (i) choose a number of the largest clusters, denoted by M_1, to which we apply the proportional allocation, and (ii) apply equal allocation to the remaining clusters. In this way, we obtain a set of nested cea designs, whose efficiency for the population total increases monotonically with M_1.

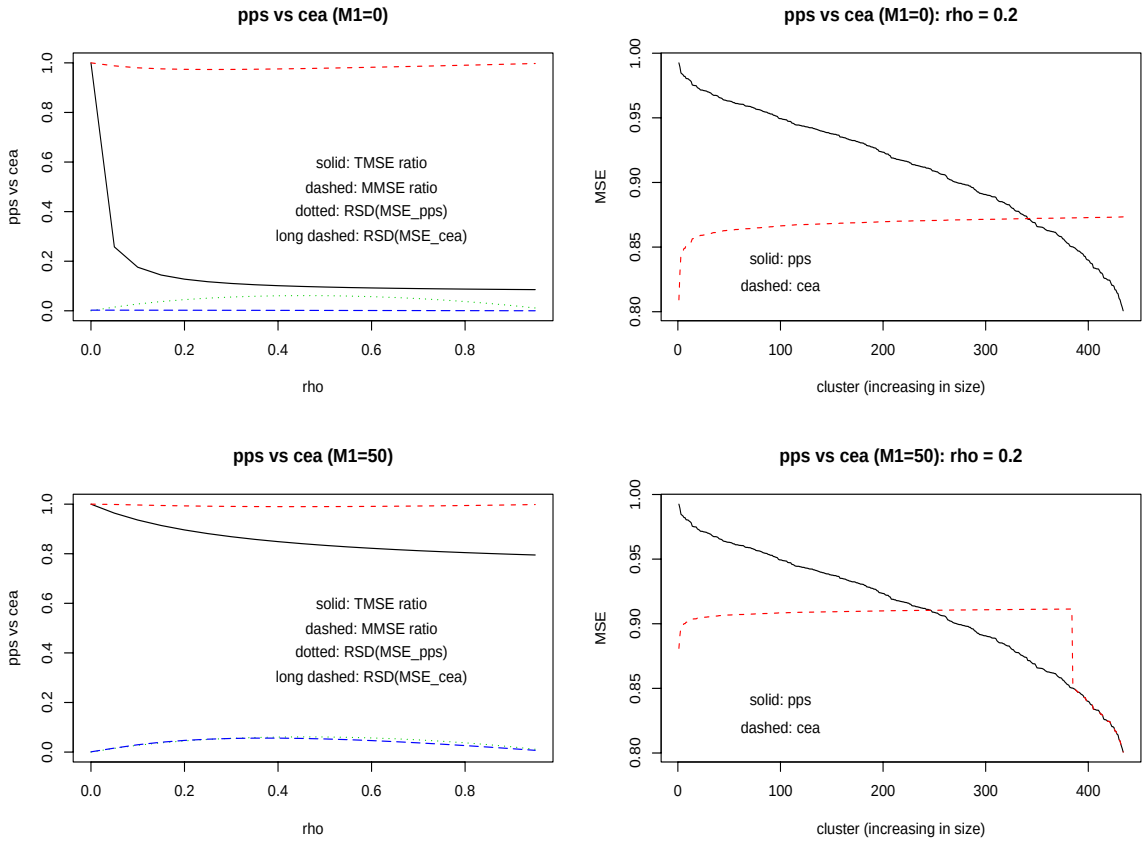

Figure 6: Proportional against constrained equal allocation. Left column: comparative performances as the underlying intracluster correlation ρ varies between 0 and 1. Right column: illustration of cluster-wise individual prediction MSEs for $\rho = 0.2$. Sample size being $n = 3000$.

In Figure 6 we show some results for the Norwegian household population. In the top row, the cea is unconstrained (i.e. $M_1 = 0$ and $n_i = n/M$ everywhere), where it entails great loss of efficiency for the population total even for very small intracluster correlation. For instance, the TMSE ratio is already below 20% at $\rho = 0.1$. On the other hand, the RSD under the equal allocation is so close to zero that it yields approximate equal prediction across the whole range of ρ. The RSD of the proportional allocation starts at zero when $\rho = 0$, it ends almost at zero when $\rho = 1$, because the number of take-none clusters is small in this case (Figure 5). Between the two ends, it increases to somewhere below 10% for $\rho \in (0.3, 0.6)$. In the top-right plot the individual prediction MSEs are given for the case of $\rho = 0.2$. Equal allocation yields almost constant MSE except for elements from the few smallest clusters. To further reduce the variation, we can reduce the sample sizes in the smallest clusters, and allocate the extra number of elements equally to the rest of the clusters. This would lead to an equal allocation design

constrained at the lower end.

In the bottom row of Figure 6, the equal allocation is constrained at $M_1 = 50$, which is just above 10% of all the clusters. The loss of efficiency is bounded to less than 20%. The variation among the individual MSEs is almost the same by the two allocations throughout the range of ρ. The details at $\rho = 0.2$ (bottom-right plot) show that the equal allocation leads to almost equal prediction for all the elements from the sub-population where equal allocation is applied, except for the very few smallest clusters.

Thus, approximate equal prediction can be achieved by means of stratified sampling with equal allocation of stratum sample size, where the clusters are treated as strata at the design stage. The strategy is robust towards the unknown intracluster correlation. The constrained equal allocation design can be used to control the efficiency for population total. There is a trade-off in terms of the individual MSE in the small and large clusters between the pps and equal allocation, which provides a basis for the choice of design in practice.

Again, we arrive at practically the same results under the common parameter model as we would have done under the stratified model with equal stratum variance. The BLUP for any element outside the sample is simply the corresponding sample stratum mean under the stratified model. Equal prediction implies then equal stratum sample size, provided equal stratum variance. To recover some of the lost efficiency, we would retain proportional allocation in a number of the largest strata, and use equal allocation in the rest of the population, i.e. constrained equal allocation. The only difference is that equal prediction is exact under the stratified model, regardless of the size of the strata.

6 Summary and discussion

We have introduced the control of individual prediction as a design criterion in addition to the efficiency of sampling. This gives rise to (unequal) probability sampling under the model-based framework of inference. Special attention has been given to the equal prediction design, under which the expected sample information (measured by the individual prediction MSE) is the same everywhere in the population, which seems a natural choice in anticipation of multiple database-like uses of the survey data. General results for the equal prediction design are given for linear regression populations and clustered populations under the intracluster correlations model. Various constrained equal prediction designs, which balance between prediction at the most aggregated and the most dis-aggregated levels, have been studied for the ratio model and the common parameter model, and illustrated using real-life data.

The constrained equal prediction approach provides theoretical motivations for a number of well-established sampling practices under a unified framework, and means by which these methods can be assessed for the given population from a prediction point of view. These include the use of simple random sampling for homogeneous population and unequal probability sampling otherwise, the division of take-all, take-some and take-none units in business surveys, the two basic schemes of two-stage sampling, *etc.*. None of them has previously received adequate model-based treatment. The constrained equal prediction approach does not lead to a single 'optimal' design. Rather, it generates sets of nested designs that form a systematic basis on which reasonable practical choices can be made.

We would like to close the paper with a few discussions. Firstly, any use of models must deal with the question of model misspecification. While model misspecification has been central in the debate between the model-based and design-based approaches to sample surveys, the usefulness of models for sampling design has been recognized on both sides (e.g. Hansen, Madow, and Tepping, 1983; Smith, 1994). Once control over individual prediction is put up as a design criterion, model becomes necessary. The design-based theory is simply unacceptable here because nothing can be inferred about the individual values outside of the sample, no matter how many observations or how much auxiliary information is available. Thus, the question is not whether models can be used, but how models can be used for sampling design.

Without dominant specific targets, sampling design must balance between a number of concerns. Combining optimal prediction of the population total and control over the individual prediction is a way of balancing between multiple uses of survey data. Another concern is multiple variables of interest. It is important to explore the performance of a particular design over a wide range of possible underlying populations, such as what happens if the variance of interest varies from proportional to $\sqrt{x}$ to x^2, or what if the intracluster correlation varies from 0 to 0.5, and so on. The measure-of-size variable should be relatively stable if the survey is to be repeated over time, despite more closely correlated auxiliary information may be available, if the latter fluctuates much more. We can explore alternative models, such as adding intercept to the ratio model. The risk of misspecifying the linear predictor can be alleviated by referring to the BP, i.e. to disregard the uncertainty due to the estimation of β. Apart from independence the only model assumption required for the BP-design is the variance structure, which is monotone increasing with x, say, $\sigma_i^2 \propto x_i^\eta$.

Stratification in one or another form is probably used in all sample surveys for statistical as well as practical reasons. It is necessary when the population is divided into a number of sub-populations, where the variables of interest follow different distributions. But it may also be needed in cases where the population follows a single model with only global parameters. Direct sampling of elements from clustered population with common mean and variance provides an example, where stratified sampling is necessary for optimal prediction of the population total and useful for control of individual prediction. However, not all the many good uses of stratification can be motivated from the two design criteria that we have focused on in this paper. Our intention here was to address some of the issues for which a theoretical treatment has been lacking. We have therefore deliberately avoided stratification as far as we could.

Another traditional topic in sampling which we did not consider here is single-stage cluster sampling. There are two situations. In the first one, the clusters are units of sampling as well as units of inference, in which case it can be dealt with using the theory for linear regression models. In the second one, the clusters are the sampling units but the elements are the units of inference, in which case we refer to the general results of Section 4. A typical example is inference of persons based on a sample of households. A reasonable design theory must involve stratification. It is also interesting to compare the performance of single-stage cluster sampling with that of two-stage cluster sampling. Neither of them, however, is the focus of this paper. There are many other important issues raised by the conflicting units of sampling and inference (Kish, 1965, Section 11.6), which again would take us beyond the scope of this paper.

References

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. *Journal of the American Statistical Association*, **83**, 28–36.

Cochran, W.G. (1977). *Sampling Techniques (third edition)*. New York: John Wiley and Sons.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, **78**, 776–793.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, **33**, 350–374.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–625.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, **14**, 149–162.

Ortega, J.M. (1972). *Numerical Analysis — A Second Course*. Academic Press, New York.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377–387.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, **71**, 657–664.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scott, A:.J. and Smith, T.M.F. (1969). Estimation in multistage surveys. *Journal of the American Statistical Association*, **64**, 830–840.

Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (With discussion). *International Statistical Review*, **64**, 3–34.

Sugden, R.A. and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495–506.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: Wiley.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, **78**, 879–884.

Recent publications in the series Discussion Papers

348 S. Johansen and A. R. Swensen (2003): More on Testing Exact Rational Expectations in Cointegrated Vector Autoregressive Models: Restricted Drift Terms

349 B. Holtsmark (2003): The Kyoto Protocol without USA and Australia - with the Russian Federation as a strategic permit seller

350 J. Larsson (2003): Testing the Multiproduct Hypothesis on Norwegian Aluminium Industry Plants

351 T. Bye (2003): On the Price and Volume Effects from Green Certificates in the Energy Market

352 E. Holmøy (2003): Aggregate Industry Behaviour in a Monopolistic Competition Model with Heterogeneous Firms

353 A. O. Ervik, E.Holmøy and T. Hægeland (2003): A Theory-Based Measure of the Output of the Education Sector

354 E. Halvorsen (2003): A Cohort Analysis of Household Saving in Norway

355 I. Aslaksen and T. Synnestvedt (2003): Corporate environmental protection under uncertainty

356 S. Glomsrød and W. Taoyuan (2003): Coal cleaning: A viable strategy for reduced carbon emissions and improved environment in China?

357 A. Bruvoll T. Bye, J. Larsson og K. Telle (2003): Technological changes in the pulp and paper industry and the role of uniform versus selective environmental policy.

358 J.K. Dagsvik, S. Strøm and Z. Jia (2003): A Stochastic Model for the Utility of Income.

359 M. Rege and K. Telle (2003): Indirect Social Sanctions from Monetarily Unaffected Strangers in a Public Good Game.

360 R. Aaberge (2003): Mean-Spread-Preserving Transformation.

361 E. Halvorsen (2003): Financial Deregulation and Household Saving. The Norwegian Experience Revisited

362 E. Røed Larsen (2003): Are Rich Countries Immune to the Resource Curse? Evidence from Norway's Management of Its Oil Riches

363 E. Røed Larsen and Dag Einar Sommervoll (2003): Rising Inequality of Housing? Evidence from Segmented Housing Price Indices

364 R. Bjørnstad and T. Skjerpen (2003): Technology, Trade and Inequality

365 A. Raknerud, D. Rønningen and T. Skjerpen (2003): A method for improved capital measurement by combining accounts and firm investment data

366 B.J. Holtsmark and K.H. Alfsen (2004): PPP-correction of the IPCC emission scenarios - does it matter?

367 R. Aaberge, U. Colombino, E. Holmøy, B. Strøm and T. Wennemo (2004): Population ageing and fiscal sustainability: An integrated micro-macro analysis of required tax changes

368 E. Røed Larsen (2004): Does the CPI Mirror Costs.of.Living? Engel's Law Suggests Not in Norway

369 T. Skjerpen (2004): The dynamic factor model revisited: the identification problem remains

370 J.K. Dagsvik and A.L. Mathiassen (2004): Agricultural Production with Uncertain Water Supply

371 M. Greaker (2004): Industrial Competitiveness and Diffusion of New Pollution Abatement Technology – a new look at the Porter-hypothesis

372 G. Børnes Ringlund, K.E. Rosendahl and T. Skjerpen (2004): Does oilrig activity react to oil price changes? An empirical investigation

373 G. Liu (2004) Estimating Energy Demand Elasticities for OECD Countries. A Dynamic Panel Data Approach

374 K. Telle and J. Larsson (2004): Do environmental regulations hamper productivity growth? How accounting for improvements of firms' environmental performance can change the conclusion

375 K.R. Wangen (2004): Some Fundamental Problems in Becker, Grossman and Murphy's Implementation of Rational Addiction Theory

376 B.J. Holtsmark and K.H. Alfsen (2004): Implementation of the Kyoto Protocol without Russian participation

377 E. Røed Larsen (2004): Escaping the Resource Curse and the Dutch Disease? When and Why Norway Caught up with and Forged ahead of Its Neughbors

378 L. Andreassen (2004): Mortality, fertility and old age care in a two-sex growth model

379 E. Lund Sagen and F. R. Aune (2004): The Future European Natural Gas Market - are lower gas prices attainable?

380 A. Langørgen and D. Rønningen (2004): Local government preferences, individual needs, and the allocation of social assistance

381 K. Telle (2004): Effects of inspections on plants' regulatory and environmental performance - evidence from Norwegian manufacturing industries

382 T. A. Galloway (2004): To What Extent Is a Transition into Employment Associated with an Exit from Poverty

383 J. F. Bjørnstad and E.Ytterstad (2004): Two-Stage Sampling from a Prediction Point of View

384 A. Bruvoll and T. Fæhn (2004): Transboundary environmental policy effects: Markets and emission leakages

385 P.V. Hansen and L. Lindholt (2004): The market power of OPEC 1973-2001

386 N. Keilman and D. Q. Pham (2004): Empirical errors and predicted errors in fertility, mortality and migration forecasts in the European Economic Area

387 G. H. Bjertnæs and T. Fæhn (2004): Energy Taxation in a Small, Open Economy: Efficiency Gains under Political Restraints

388 J.K. Dagsvik and S. Strøm (2004): Sectoral Labor Supply, Choice Restrictions and Functional Form

389 B. Halvorsen (2004): Effects of norms, warm-glow and time use on household recycling

390 I. Aslaksen and T. Synnestvedt (2004): Are the Dixit-Pindyck and the Arrow-Fisher-Henry-Hanemann Option Values Equivalent?

391 G. H. Bjønnes, D. Rime and H. O.Aa. Solheim (2004): Liquidity provision in the overnight foreign exchange market

392 T. Åvitsland and J. Aasness (2004): Combining CGE and microsimulation models: Effects on equality of VAT reforms

393 M. Greaker and Eirik. Sagen (2004): Explaining experience curves for LNG liquefaction costs: Competition matter more than learning

394 K. Telle, I. Aslaksen and T. Synnestvedt (2004): "It pays to be green" - a premature conclusion?

395 T. Harding, H. O. Aa. Solheim and A. Benedictow (2004). House ownership and taxes

396 E. Holmøy and B. Strøm (2004): The Social Cost of Government Spending in an Economy with Large Tax Distortions: A CGE Decomposition for Norway

397 T. Hægeland, O. Raaum and K.G. Salvanes (2004): Pupil achievement, school resources and family background

398 I. Aslaksen, B. Natvig and I. Nordal (2004): Environmental risk and the precautionary principle: "Late lessons from early warnings" applied to genetically modified plants

399 J. Møen (2004): When subsidized R&D-firms fail, do they still stimulate growth? Tracing knowledge by following employees across firms

400 B. Halvorsen and Runa Nesbakken (2004): Accounting for differences in choice opportunities in analyses of energy expenditure data

401 T.J. Klette and A. Raknerud (2004): Heterogeneity, productivity and selection: An empirical study of Norwegian manufacturing firms

402 R. Aaberge (2005): Asymptotic Distribution Theory of Empirical Rank-dependent Measures of Inequality

403 F.R. Aune, S. Kverndokk, L. Lindholt and K.E. Rosendahl (2005): Profitability of different instruments in international climate policies

404 Z. Jia (2005): Labor Supply of Retiring Couples and Heterogeneity in Household Decision-Making Structure

405 Z. Jia (2005): Retirement Behavior of Working Couples in Norway. A Dynamic Programming Approch

406 Z. Jia (2005): Spousal Influence on Early Retirement Behavior

407 P. Frenger (2005): The elasticity of substitution of superlative price indices

408 M. Mogstad, A. Langørgen and R. Aaberge (2005): Region-specific versus Country-specific Poverty Lines in Analysis of Poverty

409 J.K. Dagsvik (2005) Choice under Uncertainty and Bounded Rationality

410 T. Fæhn, A.G. Gómez-Plana and S. Kverndokk (2005): Can a carbon permit system reduce Spanish unemployment?

411 J. Larsson and K. Telle (2005): Consequences of the IPPC-directive's BAT requirements for abatement costs and emissions

412 R. Aaberge, S. Bjerve and K. Doksum (2005): Modeling Concentration and Dispersion in Multiple Regression

413 E. Holmøy and K.M. Heide (2005): Is Norway immune to Dutch Disease? CGE Estimates of Sustainable Wage Growth and De-industrialisation

414 K.R. Wangen (2005): An Expenditure Based Estimate of Britain's Black Economy Revisited

415 A. Mathiassen (2005): A Statistical Model for Simple, Fast and Reliable Measurement of Poverty

416 F.R. Aune, S. Glomsrød, L. Lindholt and K.E. Rosendahl: Are high oil prices profitable for OPEC in the long run?

417 D. Fredriksen, K.M. Heide, E. Holmøy and I.F. Solli (2005): Macroeconomic effects of proposed pension reforms in Norway

418 D. Fredriksen and N.M. Stølen (2005): Effects of demographic development, labour supply and pension reforms on the future pension burden

419 A. Alstadsæter, A-S. Kolm and B. Larsen (2005): Tax Effects on Unemployment and the Choice of Educational Type

420 E. Biørn (2005): Constructing Panel Data Estimators by Aggregation: A General Moment Estimator and a Suggested Synthesis

421 J. Bjørnstad (2005): Non-Bayesian Multiple Imputation

422 H. Hungnes (2005): Identifying Structural Breaks in Cointegrated VAR Models

423 H. C. Bjørnland and H. Hungnes (2005): The commodity currency puzzle

424 F. Carlsen, B. Langset and J. Rattsø (2005): The relationship between firm mobility and tax level: Empirical evidence of fiscal competition between local governments

425 T. Harding and J. Rattsø (2005): The barrier model of productivity growth: South Africa

426 E. Holmøy (2005): The Anatomy of Electricity Demand: A CGE Decomposition for Norway

427 T.K.M. Beatty, E. Røed Larsen and D.E. Sommervoll (2005): Measuring the Price of Housing Consumption for Owners in the CPI

428 E. Røed Larsen (2005): Distributional Effects of Environmental Taxes on Transportation: Evidence from Engel Curves in the United States

429 P. Boug, Å. Cappelen and T. Eika (2005): Exchange Rate Rass-through in a Small Open Economy: The Importance of the Distribution Sector

430 K. Gabrielsen, T. Bye and F.R. Aune (2005): Climate change- lower electricity prices and increasing demand. An application to the Nordic Countries

431 J.K. Dagsvik, S. Strøm and Z. Jia: Utility of Income as a Random Function: Behavioral Characterization and Empirical Evidence

432 G.H. Bjertnæs (2005): Avioding Adverse Employment Effects from Energy Taxation: What does it cost?

433. T. Bye and E. Hope (2005): Deregulation of electricity markets—The Norwegian experience

434 P.J. Lambert and T.O. Thoresen (2005): Base independence in the analysis of tax policy effects: with an application to Norway 1992-2004

435 M. Rege, K. Telle and M. Votruba (2005): The Effect of Plant Downsizing on Disability Pension Utilization

436 J. Hovi and B. Holtsmark (2005): Cap-and-Trade or Carbon Taxes? The Effects of Non-Compliance and the Feasibility of Enforcement

437 R. Aaberge, S. Bjerve and K. Doksum (2005): Decomposition of Rank-Dependent Measures of Inequality by Subgroups

438 B. Holtsmark (2005): Global per capita CO_2 emissions - stable in the long run?

439 E. Halvorsen and T.O. Thoresen (2005): The relationship between altruism and equal sharing. Evidence from inter vivos transfer behavior

440 L-C. Zhang and I. Thomsen (2005): A prediction approach to sampling design