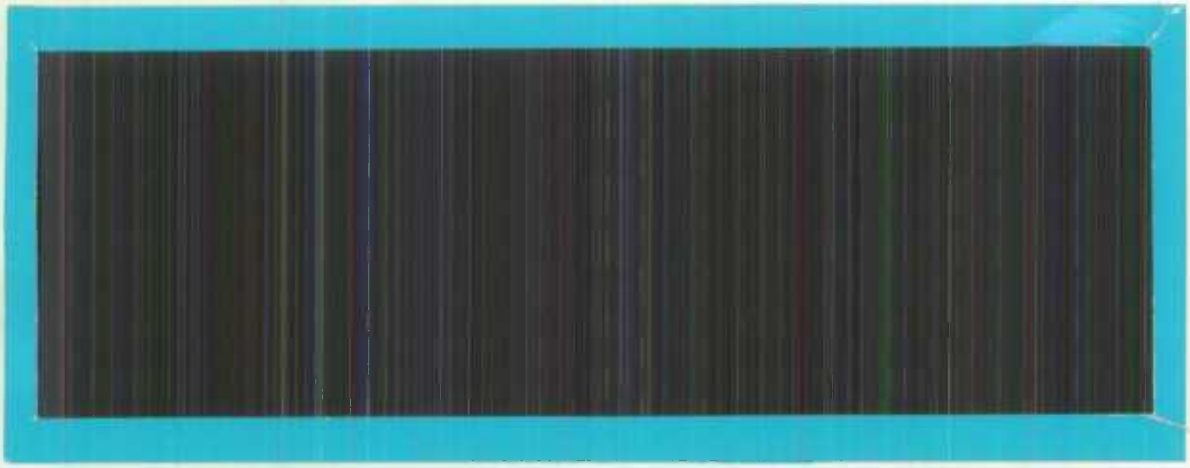
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER NO. SSMD 91-018E

METHODOLOGY BRANCH

TIME SERIES GENERALIZATION OF FAY-HERRIOT ESTIMATION FOR SMALL AREAS

A.C. SINGH[1], H.J. MANTEL[1], AND B.W. THOMAS[2]

[1] *Social Survey Methods Division, Statistics Canada*
[2] *Business Survey Methods Division, Statistics Canada*

ABSTRACT

In estimation for small areas it is common to borrow strength from other small areas since the direct survey estimates often have large sampling variability. A class of methods called composite estimation addresses the problem by using a linear combination of direct and synthetic estimators. The synthetic component is based on a model which connects small area means cross-sectionally (over areas) and/or over time. The Fay-Herriot estimator is a composite estimator which provides empirical best linear unbiased predictors for cross-sectional data under a linear regression model with uncorrelated small area effects. In this paper we consider three models to generalize Fay-Herriot estimation to more than one time point. In the first model, regression parameters are random and serially dependent but the small area effects are assumed to be independent over time. In the second model, regression parameters are nonrandom and may take common values over time but the small area effects are serially dependent. The third model is more general in that regression parameters and small area effects are assumed to be serially dependent.

The resulting estimators, as well as some cross-sectional estimators, are evaluated using bi-annual data from Statistics Canada's National Farm Survey and January Farm Survey.

KEY WORDS: Composite estimation; State space models; Kalman Filter; Best linear unbiased prediction.

RÉSUMÉ

Lors de l'estimation des données régionales, on utilise couramment les données d'autres petites régions, puisque les estimations directes d'enquête ont souvent une grande variabilité d'échantillonnage. Une catégorie de méthodes, que l'on appelle estimation composite, élimine ce problème en utilisant pour cela une combinaison linéaire d'estimateurs directs et synthétiques. La composante synthétique repose sur un modèle qui relie transversalement (selon les régions) et/ou dans le temps les moyennes des petites régions. L'estimateur de Fay-Herriot est un estimateur composite qui donne les meilleurs prédicteurs sans biais linéaires empiriques pour les données transversales dans un modèle de régression linéaire avec effets régionaux non corrélés. Nous examinons ici trois modèles afin de généraliser l'estimateur de Fay-Herriot à plus d'une période. Dans le premier modèle, les paramètres de régression sont aléatoires et ont une dépendance sériale, mais on suppose que les effets régionaux sont indépendants dans le temps. Dans le deuxième, les paramètres de régression sont non aléatoires et peuvent prendre des valeurs communes dans le temps, mais les effets régionaux ont une dépendance sériale. Le troisième modèle est plus général, en ce sens que les paramètres de régression et les effets régionaux ont par définition une dépendance sériale.

1. INTRODUCTION

There exists a considerable body of research on small area estimation using cross-sectional survey data in conjunction with supplementary data obtained from census and administrative sources. A good collection of papers on this topic can be found in Platek, Rao, Särndal and Singh (1987). The basic idea underlying all small area methods is to borrow strength from other areas by assuming that different areas are linked via a model containing auxiliary variables from the supplementary data. It would also be important to borrow strength across time because most surveys are repeated over time. Recently time series methods are being employed to develop improved estimators for small areas; see Choudhry and Rao (1989) and Pfeffermann and Burck (1990). It is interesting to note that after the initiative of Scott and Smith (1974) on the application of time series methods to survey data, there has been only lately a resurgence of interest in developing suitable estimates of aggregates from complex surveys repeated at regular time intervals; see e.g. Bell and Hilmer (1987), Binder and Dick (1989), Tiller (1989), and Pfeffermann (1991).

In this paper we consider some natural generalizations of the Fay Herriot (FH) estimator for small areas when a time series of direct small area estimates is available. The important work of Fay and Herriot (1979) shows how direct estimators can be smoothed by cross-sectional modelling of small area totals. The resulting estimators are composite estimators (i.e. convex combinations of direct and model-based synthetic estimators) and are also empirical best linear unbiased predictors (EBLUPs). With the use of structural models, we derive time series EBLUPs which combine both cross-sectional and time series data. The main purpose of this paper is to compare time series EBLUPs with cross-sectional estimators such as post-stratified domain, synthetic, FH and sample size dependent estimators.

An empirical study based on Monte Carlo simulations from real time series data obtained from Statistics Canada's biannual farm surveys was conducted to investigate potential gains in efficiency with time series EBLUPs. The main findings of the study are

(i) There can be substantial gains in efficiency with time series EBLUPs over cross-sectional estimators.

(ii) Within the class of time series methods considered in this paper, introduction of serial dependence in the random small area effects is found to be considerably more beneficial than dependence of the parameters of the synthetic component.

(iii) Although any smoothed version of the direct small area estimator is expected to be biased, the time series EBLUPs exhibit less bias than other methods including FH estimator.

(iv) Within the class of cross-sectional methods, the performance of the FH estimator is best overall followed by that of the sample size dependent estimator. Nevertheless, in the context of the empirical study presented here, the gains over direct estimation in mean squared error due to cross-sectional smoothing are marginal at best and there is considerable cost in terms of bias. How can we determine when the benefits of smoothing outweigh the costs?

Section 2 contains a description of various cross-sectional methods for small area estimation. Time series EBLUPs are described in Section 3 and the details and results of the Monte Carlo comparative study are given in Section 4. Finally, some directions for future work are mentioned in the Section 5.

2. METHODS BASED ON CROSS-SECTIONAL DATA

In this section, we assume that information is available only for a particular point in time t. Let $\underset{\sim}{\theta}_t$ denote the vector of small area population totals $\theta_{kt}, k=1, \ldots K$, at time t. Here we define briefly some well known small area estimators under the assumption that the underlying sampling design is stratified simple random; for more details, see Rao (1986). Särndal and Hidiroglou (1987) and Pfeffermann and Burck (1991) also contain a good survey of various small area estimators.

2.1 Method 1 (Expansion or Horvitz-Thompson estimator for domains)

This method of estimation is defined by

$$g_{1kt} = \sum_h (N_{ht}/n_{ht}) \sum_{j \in s_{hkt}} y_{hjt} \,, \tag{2.1}$$

where at time t, y_{hjt} is the jth observation in the h stratum, s_{hkt} denotes the set of n_{hkt} sample units falling in the kth small area in the hth stratum and n_{ht}, N_{ht} denote respectively the sample and population sizes for the hth stratum. The above estimator is often unreliable because the random sample size n_{hkt} may be small in expectation and could have high variability. Conditional on the realized sample size n_{hkt}, g_{1kt} is biased. However, unconditionally, it is unbiased for θ_{kt}.

2.2 Method 2 (Post-stratified domain estimator)

We will refer to this estimator also as the direct small area estimator. Suppose the population size N_{hkt} is known for each (h,k,t). The efficiency of the estimator g_{1kt} could be improved by post-stratification. Suppose small areas themselves constitute post-strata within stratum h. We have

$$g_{2kt} = \sum_h (N_{hkt}/n_{hkt}) \sum_{j \in s_{hkt}} y_{hjt} = \sum_h N_{hkt} \bar{y}_{hkt}. \tag{2.2}$$

However, this estimator also may not be sufficiently reliable because of the possibility of n_{hkt}'s being small in expectation. If $n_{hkt} = 0$, the above estimator is not defined. In practice, some ad hoc value such as 0 is often chosen for $\bar{y}_{hkt}$ when $n_{hkt} = 0$. In the empirical study presented in this paper, we have set $\bar{y}_{hkt}$ as $(\bar{X}_{hkt}/\bar{X}_{ht})\,\bar{y}_{ht}$ whenever $n_{hkt} = 0$, where X is a suitable covariable.

The estimator g_{2kt} is both conditionally and unconditionally unbiased. Its conditional variance v_{kt} (whenever $n_{hkt} > 0$ for all h at time t) is given by

$$v_{kt} = \sum_h N_{hkt}^2 \,(n_{hkt}^{-1} - N_{hkt}^{-1})\, \sigma_{hkt}^2 \tag{2.3}$$

where σ_{hkt}^2 is the population variance for the intersection of the hth stratum with the kth small area at time t. The variance σ_{hkt}^2 can be estimated by the usual estimator s_{hkt}^2 for $n_{hkt} \geq 2$. Note that the estimate of the conditional variance v_{kt} also provides an estimate of the unconditional variance of g_{2kt}.

If $n_{hkt} = 1$, then we can use a synthetic value as an estimate of σ^2_{hkt} which can be defined as $\sum (n_{hkt}-1)\ S^2_{hkt} / \sum (n_{hkt}-1)$, the summation being over all k for which $n_{hkt} \geq 2$ within each (h,t). If $n_{hkt}=0$, v_{ht} of (2.3) is of course not defined. With the synthetic value of $\bar{y}_{hkt}$ used in this case, we need a synthetic value of its mean squared error. For each (h,t), it can be defined as

$$(\bar{X}_{hkt}/\bar{X}_{ht})^2\ (n_{ht}^{-1}-N_{ht}^{-1})\ S^2_{ht} + (\hat{\text{bias}})^2 \tag{2.4}$$

where $(\hat{\text{bias}})^2$ will be taken as

$$\sum_{n_{hkt}>0} ((\bar{X}_{hkt}/\bar{X}_{ht})\bar{y}_{ht} - \bar{y}_{hkt})^2/m_{ht}. \tag{2.5}$$

where m_{ht} is the number of small areas with sample in stratum h at time t.

2.3 Method 3 (Synthetic estimator)

It is possible to define a more efficient estimator by assuming a model which allows for "borrowing strength" from other small areas. This gives rise to synthetic estimators. For instance, suppose different small area totals are connected via the auxiliary variable X_{kt} by a linear model as

$$\theta_{kt} = \beta_{1t} + \beta_{2t} X_{kt}, k=1, \ldots K, \tag{2.6a}$$

or in matrix notation

$$\underset{\sim}{\theta}_t = F_t \underset{\sim}{\beta}_t , \tag{2.6b}$$

where $F_t = (F_{1t}, F_{2t}, \ldots, F_{Kt})'$, $F_{kt} = (1, X_{kt})'$. The above model may not be realistic because no random fluctuation or random small area effect (a_{kt}, say) is allowed. In other words, the error term a_{kt} is assumed to have both mean and variance zero. Now consider a model for the direct small area estimators g_{2kt}'s as

$$\underset{\sim}{g}_{2t} = F_t\ \underset{\sim}{\beta}_t + \underset{\sim}{\epsilon}_t \tag{2.7}$$

where $\underset{\sim}{g}_{2t} = (g_{21t}, \ldots, g_{2Kt})'$, $\underset{\sim}{\epsilon}_t = (\epsilon_{1t}, \ldots \epsilon_{Kt})'$, ϵ_{kt}'s are uncorrelated as k varies with mean 0 and variance v_{kt} defined earlier by (2.3).

Denoting by $\hat{\underline{\beta}}_t$ the weighted least squares (WLS) estimate of $\underline{\beta}_t$, we obtain the regression-synthetic estimator of θ_{kt} under the assumed model as

$$\underline{g}_{3t} = F_t \hat{\underline{\beta}}_t. \tag{2.8}$$

The above estimator could be heavily biased unless the model (2.6) is satisfied reasonably well.

2.4 Method 4 (EBLUP - empirical best linear unbiased predictor)

Using the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor (BLUP) approach; see e.g. Battese, Harter, and Fuller (1988), and Pfeffermann and Barnard (1991), the bias of the synthetic estimator can be reduced considerably by using a composite estimator. This is obtained as a convex combination of $\underline{g}_{2t}$ and a somewhat modified $\underline{g}_{3t}$. For this purpose, it is assumed that

$$\underline{\theta}_t = F_t \underline{\beta}_t + \underline{a}_t , \tag{2.9}$$

where a_{kt}'s are uncorrelated random small area effects with mean 0 and variance w_{kt}. Thus we have a somewhat modified model for $\underline{g}_{2t}$ as

$$\underline{g}_{2t} = F_t \underline{\beta}_t + \underline{a}_t + \underline{\epsilon}_t. \tag{2.10}$$

Here $\underline{a}_t$ is also assumed to be uncorrelated with $\underline{\epsilon}_t$. Let $\underline{g}^*_{3t}$ denote the modified synthetic estimator of $\underline{\theta}_t$ under (2.10). The BLUP of $\underline{\theta}_t$ under the model defined by (2.9) and (2.10) is

$$\begin{aligned} \underline{g}_{4t} &= \underline{g}^*_{3t} + \Lambda_t(\underline{g}_{2t} - \underline{g}^*_{3t}) \\ &= \Lambda_t \underline{g}_{2t} + (I-\Lambda_t)\ \underline{g}^*_{3t} \end{aligned} \tag{2.11}$$

where

$$\begin{aligned} \Lambda_t &= (V_t^{-1} + W_t^{-1})^{-1}\, V_t^{-1} = W_t U_t^{-1}, \\ I - \Lambda_t &= (V_t^{-1} + W_t^{-1})^{-1}\, W_t^{-1} = V_t\, U_t^{-1}, \\ U_t &= V_t + W_t,\ V_t = \mathrm{diag}(v_{1t}, \ldots v_{kt}), \\ W_t &= \mathrm{diag}(w_{1t}, \ldots, w_{kt}). \end{aligned} \tag{2.12}$$

The expression (2.11) follows from the general results on linear models with random effects, see e.g., Rao (1973, p. 267) and Harville (1976). The BLUP or BLUE of $F_t\,\underline{\beta}_t$ is $\underline{g}_{3t}^{*}$ and BLUP of $\underline{a}_t$ *is* $\Lambda_t(\underline{g}_{2t} - \underline{g}_{3t}^{*})$. It may be of interest to note that the formula for BLUP does not change regardless of whether or not $\underline{\beta}_t$ is known. However, its MSE does change as expected due to estimation of $\underline{\beta}_t$. It can be shown that,

$$\begin{aligned} \mathrm{MSE}(\underline{g}_{4t} - \underline{\theta}_t \,|\, \underline{\beta}_t \text{ known}) &= W_t U_t^{-1} V_t \\ &= V_t - V_t U_t^{-1} V_t\,, \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} &\mathrm{MSE}(\underline{g}_{4t} - \underline{\theta}_t \mid \underline{\beta}_t \text{ unknown}) \\ &= \mathrm{MSE}\{(A_t + W_t\, U_t^{-1}(I - A_t))\ \underline{\epsilon}_t - V_t\, U_t^{-1}(I - A_t)\ \underline{a}_t\} \end{aligned} \tag{2.14}$$

where $A_t = F_t(F_t' U_t^{-1} F_t)^{-1}\, F_t'\, U_t^{-1}$. The MSE matrix of (2.14) can be easily obtained from MSEs of $\underline{\epsilon}_t$ and $\underline{a}_t$.

When V_t and W_t are replaced by their estimates, the estimator $\underline{g}_{4t}$ is termed EBLUP. Note that the model (2.9) is more realistic than (2.6), and therefore, the performance of $\underline{g}_{4t}$ is expected to be quite favourable. The estimator $\underline{g}_{4t}$ approaches $\underline{g}_{2t}$ when v_{kt}'s get small, i.e. when n_{hkt}'s become large. However, it remains biased in general, conditional on $\underline{\theta}_t$.

2.5 Method 5 (Sample Size dependent estimator)

An alternative composite estimator which can considerably attenuate bias of the synthetic estimator $\underline{g}_{3t}$ as compared to the EBLUP $\underline{g}_{4t}$ is given by the sample size dependent estimator of Drew, Singh, and Choudhry (1982). It is defined as

$$g_{5t} = \Delta_t g_{2t} + (I-\Delta_t)\, g_{3t}\,, \tag{2.15}$$

where $\Delta_t = \mathrm{diag}(\delta_{1t}, \ldots, \delta_{kt})$,

$$\delta_{kt} = \begin{cases} 1 & \text{if } \sum_h \hat{N}_{hkt} \geq \lambda \sum_h N_{hkt}\,, \\ \sum_h \hat{N}_{hkt} / \lambda \sum_h N_{hkt} & \text{otherwise} \end{cases} \tag{2.16}$$

$\hat{N}_{hkt}$ being $n_{hkt}(N_{ht}/n_{ht})$, and the parameter λ is chosen in an ad hoc manner as a way of controlling the contribution of the synthetic component. In practice, λ is generally chosen as 1, 1.5 or 2. The above estimator takes account of the realized sample size n_{hkt}'s and if these are deemed to be sufficiently large according to the condition in (2.16), then it does not rely on the synthetic estimator. This property is somewhat similar to that of g_{4t}. However, the condition in (2.16) could be satisfied even if some or all n_{hkt}'s are small, and then, unlike g_{4t}, the above estimator fails to borrow strength from other small areas even though g_{2t} is unreliable.

3. METHODS BASED ON POOLED CROSS-SECTIONAL AND TIME SERIES DATA

Suppose information is available for several time points, $t=1 \ldots T$, in the form of direct small area estimators g_{2t} and also the small area population totals for the auxiliary variable. We will now introduce some estimators which generalize the Fay-Herriot estimator g_{4t} in different ways by taking account of the serial dependence of the direct estimates $\{g_{2t} : t=1 \ldots T\}$. Recall that for the Fay-Herriot estimator, the model for $\underline{\theta}_t$ has two components, namely, the trend component $F_t \underline{\beta}_t$ and the area component $\underline{a}_t$. The estimator g_{4t} borrows strengh over areas for each t and is given by the sum of two components, each being BLUP (BLUE) for the corresponding random (fixed) effect, i.e.,

$$g_{4t} = F_t \hat{\underline{\beta}}_t + \hat{\underline{a}}_t. \tag{3.1}$$

Methods based on time series data could, however, borrow strength over time as well. There are several ways one could build serial dependence in the series $\{g_{2t}\}$. Our main purpose, as mentioned in the introduction, is to illustrate that substantial gains in efficiency could be realized with time series EPLUPs. Moreover, it might be that an estimator with a simpler time-dependence structure could perform almost as well as one with a more complex structure. To this end,

we introduce three estimators $\underset{\sim}{\theta}_{6t}$, $\underset{\sim}{\theta}_{7t}$ and $\underset{\sim}{\theta}_{8t}$ corresponding to three interesting scenarios which are motivated from specific structural models for serial dependence. More specifically, first we let $\underset{\sim}{\beta}_t$ evolve over time (e.g. according to a random walk), but assume that $\underset{\sim}{a}_t$ is serially independent. This will give rise to a composite estimator

$$\underset{\sim}{\theta}_{6t} = F_t \tilde{\underset{\sim}{\beta}}_t + \tilde{\underset{\sim}{a}}_t \tag{3.2}$$

Note that $\tilde{\underset{\sim}{\beta}}_t$ in (3.2) would now be based on all the small area estimates up to time T and therefore would be different from $\hat{\underset{\sim}{\beta}}_t$ of (3.1) which is based on only direct estimates at time t. The estimator $\tilde{\underset{\sim}{a}}_t$, as a result, would also be different from the corresponding component $\hat{\underset{\sim}{a}}_t$ of (3.1).

For the second estimator, we let $\underset{\sim}{\beta}_t$ be fixed (it may or may not be common for different time points) and let the area effects $\underset{\sim}{a}_t$ be serially dependent according to, for example, a random walk. This time series generalization could be viewed as an analogue of the model proposed by Choudhry and Rao (1989). The resulting composite estimator will have the same form as (3.1) i.e.

$$\underset{\sim}{\theta}_{7t} = F_t \bar{\underset{\sim}{\beta}}_t + \bar{\underset{\sim}{a}}_t . \tag{3.3}$$

but the component estimates $\bar{\underset{\sim}{\beta}}_t$ and $\bar{\underset{\sim}{a}}_t$ would be different.

Finally, for the third estimator, we let both $\underset{\sim}{\beta}_t$ and $\underset{\sim}{a}_t$ evolve over time. This will have more complex serial dependence than either (3.2) or (3.3). Its form will be similar to (3.1) and can be represented as

$$\underset{\sim}{\theta}_{8t} = F_t \bar{\bar{\underset{\sim}{\beta}}}_t + \bar{\bar{\underset{\sim}{a}}}_t \tag{3.4}$$

All the three types of generalizations of Fay-Herriot estimator yield estimators that are members of the general class proposed by Pfeffermann and Burck (1991) using structural time series models. They used maximum likelihood method under normality assumption to estimate model parameters. However, as will be seen later, for the serial dependence considered in this paper for

illustrative purposes, the method of moments can be used for estimating parameters.

Each of the above three estimators is described below in more detail.

3.1 Method 6 (Time Series EBLUP-I)

In this case, the structural time series model for the direct small area estimates $\{\underline{y}_{2t} : t=1,..T\}$ is specified by the following state space model. Let α_t denote $(\underline{\beta}_t', \underline{a}_t')'$ and H_t denote (F_t, I).

Observation Equation

$$\begin{aligned} \underline{y}_{2t} &= \underline{\theta}_t + \underline{\epsilon}_t \\ \underline{\theta}_t &= F_t \underline{\beta}_t + \underline{a}_t \equiv H_t \underline{\alpha}_t \end{aligned} \tag{3.5a}$$

Transition Equation

$$\underline{\alpha}_t = G_t \underline{\alpha}_{t-1} + \underline{\zeta}_t \tag{3.5b}$$

where

$$G_t = \begin{pmatrix} G_t^{(1)} & 0 \\ 0 & 0 \end{pmatrix}, \ \underline{\zeta}_t = \begin{pmatrix} \underline{\xi}_t \\ \underline{\eta}_t \end{pmatrix}, \tag{3.5c}$$

along with the usual assumptions about random errors, i.e., $\underline{\epsilon}_t$, $\underline{\zeta}_t$ are uncorrelated, $\underline{\zeta}_t$ is uncorrelated with $\underline{\alpha}_s$ for $s<t$, and that $\underline{\epsilon}_t \sim (\underline{0}, V_t)$, $\underline{\zeta}_t \sim (\underline{0}, \Gamma_t)$ where Γ_t = block diag (B_t, W_t). The covariance matrices V_t, B_t, and W_t are generally diagonal. If $\underline{\beta}_t$ evolves according to a random walk, then $G_t^{(1)} = I$. The second diagonal submatrix of G_t is zero because $\underline{a}_t$'s are assumed to be serially independent.

The estimator $\underline{y}_{6T}$ is BLUP of $\underline{\theta}_t$ given all the direct estimates up to time T. To find $\underline{y}_{6T}$, first we will find BLUP $\tilde{\underline{\alpha}}_T$ of $\underline{\alpha}_T$, from which BLUP of $\underline{\theta}_T$ can be simply obtained as $H_t \tilde{\underline{\alpha}}_T$. Since $\underline{\alpha}_t$'s are connected over time according to the transition equation, it is possible, albeit cumbersome, to get $\tilde{\underline{\alpha}}_T$ directly from the theory of linear models with random effects for the complete data . However, it could be convenient to compute it recursively using Kalmar Filter (KF). Traditionally KF is viewed as a Bayesian technique in which at each time t, the prior distribution of $\underline{\alpha}_t$ given data up to t-1 is updated to get the posterior distribution of $\underline{\alpha}_t$ given data up to time t. Although it is instructive to view

KF in this manner, it is not necessary under mixed linear models. Suppose $\tilde{\underline{\alpha}}_{t|s}$ denotes the BLUP of $\underline{\alpha}_t$ based on s observations, $s < t$. It is known that the BLUP $\tilde{\underline{\alpha}}_t$ of $\underline{\alpha}_t$ based on t observations is the same as the BLUP of $\underline{\alpha}_t$ based on $\tilde{\underline{\alpha}}_{t|s}$ and the last $t-s$ observations. In other words, information in the previous data can be condensed into an appropriate BLUP before augmenting more current data points. The recursion algorithim for obtaining $\tilde{\underline{\alpha}}_{t|s}$ is given as follows.

At time $t-1$, let $\tilde{\underline{\alpha}}_{t-1}$ and P_{t-1} denote respectively the BLUP of $\underline{\alpha}_t$ and its MSE, i.e.,

$$\tilde{\underline{\alpha}}_{t-1} - \underline{\alpha}_{t-1} \sim (\underline{0},\ P_{t-1}) \tag{3.6}$$

Therefore, the BLUP $\tilde{\underline{\alpha}}_{t|t-1}$ of $\underline{\alpha}_t$ and its MSE $P_{t|t-1}$ based on $t-1$ observations is given by (in view of the relation 3.5b),

$$\tilde{\underline{\alpha}}_{t|t-1} = G_t\, \tilde{\underline{\alpha}}_{t-1},\quad P_{t|t-1} = G_t\, P_{t-1}\, G_t' + Q_t \tag{3.7}$$

Now, combining data at t, i.e., $\underline{g}_{2t}$ with $\tilde{\underline{\alpha}}_{t|t-1}$, one can get BLUP $\tilde{\underline{\alpha}}_t$ and its MSE as

$$\tilde{\underline{\alpha}}_t = \tilde{\underline{\alpha}}_{t|t-1} + P_{t|t-1}\, H_t' \,(H_t\, P_{t|t-1}\, H_t' + V_t)^{-1}\, (\underline{g}_{2t} - H_t\, \tilde{\underline{\alpha}}_{t|t-1}) \tag{3.8a}$$

and

$$\mathrm{MSE}(\tilde{\underline{\alpha}}_t - \underline{\alpha}_t) = P_t = P_{t|t-1} - P_{t|t-1}\, H_t' (H_t\, P_{t|t-1}\, H_t' + V_t)^{-1}\, H_t\, P_{t|t-1} \tag{3.8b}$$

In the usual *KF* terminology, (3.7) and (3.8) specify respectively the prior and posterior distributions of $\underline{\alpha}_t$ once the data at time t becomes available. Note that here distributions are specified only up to first two moments which is of course sufficient for linear Bayes estimation. The results (3.8a) and (3.8b) respectively give the posterior mean and variance of $\underline{\alpha}_t$ given data up to time t.

The above recursive algorithm or *KF* can be started at the initial time $t=1$ by noting that $\tilde{\underline{\alpha}}_1$ and P_1 are given by (2.11) and (2.14) respectively, i.e. the corresponding expressions for *FH* estimator at time $t=1$. The recursion is

continued until $t=T$ to obtain $\tilde{\underline{\alpha}}_T$ and the MSE matrix P_T. This, in turn, yields the BLUP of $\underline{\theta}_T$ at $t=1$ and its MSE as $H_T P_T H_T'$.

We will now illustrate method of moments for estimating model parameters in the special case when there is only one auxiliary variable X_{ht}, i.e. $F_t = (F_{1t}, \ldots F_{kt})'$, $F_{kt} = (1, X_{kt})'$, $\underline{\beta}_t = (\beta_{1t}, \beta_{2t})'$, $w_{kt} = \tau^2$ and $\underline{\beta}_t$ follows a random walk, i.e. $G_t^{(1)} = I$. Let $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$. Now, the unknown parameters τ^2, γ_1^2, γ_2^2 can be estimated by the method of moments as follows. The parameter τ^2 is obtained as the solution of

$$\sum_{t=1}^{T} \sum_{k=1}^{k} (g_{2kt} - F_{kt}' \hat{\underline{\beta}}_t)^2 / (v_{kt} + \tau^2) = T(k-2) \tag{3.9}$$

If there is no positive solution, we set τ^2 as zero. Here $\hat{\underline{\beta}}_t$ denotes the WLS estimate of $\underline{\beta}_t$ based on only the cross-sectional data at t. This is analogous to the method used in Fay and Herriot (1979) for cross-sectional data. An estimate of γ_i^2 can be obtained by solving (for $i=1,2$)

$$\sum_{t=2}^{T} (\hat{\beta}_{it} - \hat{\beta}_{i,t-1})^2 / (\gamma_i^2 + d_{ii}^{(t)}) = T-1 \tag{3.10}$$

where $d_{ii}^{(t)}$ is the (i,i) th element of $(F_{t-1}' U_{t-1}^{-1} F_{t-1})^{-1} + (F_t' U_t^{-1} F_t)^{-1}$.

When the above estimators of model parameters are substituted in the expression for $H_T \tilde{\underline{\alpha}}_T$ we get the time series EBLUP-I estimator $\underline{g}_{6T}$ at time T.

3.2 Method 7 (Time Series EBLUP-II)

The equations for the state space model for this case are similar to 3.5(a) and (b) except that the transition matrix G_t and the covariance matrix Γ_t are different. We have two cases.

3.2.1 Case 1 First suppose $\underline{\beta}_t$'s fixed and time-invariant but $\underline{a}_t$'s are serially dependent. Then the matrices G_t and Γ_t are given by

$$G_t = \begin{pmatrix} I & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \Gamma_t = \text{block diag } (0, Q_t) \tag{3.11}$$

For a given choice of Q_t, the *KF* can be run as in method 6 with the initial values $\tilde{\underline{\alpha}}_1$ and P_1 at $t=1$ obtained from the *FH* estimator at $t=1$. If $\underline{a}_t$ is

assumed to evolve according to a random walk, then $G_t^{(2)} = I$. Moreover, if Q_t is taken as $\nu^2 I$, then the only unknown parameter ν^2 can be estimated from an equation similar to (3.9). We will denote by $\underset{\sim}{g}_{7t}$ the EBLUP obtained in this case when the parameter estimate is substituted. Also we will denote by $\underset{\sim}{g}_{7t}^{a}$ the estimator in the special case when the common value of $\underset{\sim}{\beta}_t$ is assumed known.

3.2.2 Case 2 Here we assume that $\underset{\sim}{\beta}_t$'s are fixed but different for different time points. The area effects $\underset{\sim}{\alpha}_t$ evolve over time as before. The matrices G_t and Γ_t are

$$G_t = \begin{pmatrix} I & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \Gamma_t = \text{block diag}\{mI, Q_t\} \tag{3.12}$$

where m is a large integer. The expression of $\tilde{\underset{\sim}{\alpha}}_T$ and P_T obtained from the KF in this case approximately give the correct formulas as $m \to \infty$. If P_T is not required (as is the case in our Monte Carlo study), then a simpler alternative is the following. In the measurement equation, replace $\underset{\sim}{\beta}_t$ by $\hat{\underset{\sim}{\beta}}_t$ -- the cross-sectional WLS estimate and then treat $\underset{\sim}{\beta}_t$ as known. In other words, $\underset{\sim}{\beta}_t$ is no longer required to be a part of the state vector $\underset{\sim}{a}_t$. This strategy will give rise to the correct $\tilde{\underset{\sim}{\alpha}}_T$. However, the corresponding P_T would not be correct. The time series EBLUP in this case will be denoted by $\underset{\sim}{g}_{7t}^{b}$.

3.3 Method 8 (Time Series EBLUP-III)

As was the case with method 7, the equations for the state space model are similar to 3.5(a) and (b) except that the two matrices G_t and Γ_t are different. We have

$$G_t = \begin{pmatrix} G_t^{(1)} & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \Gamma_t = \text{block diag}\{B_t, Q_t\}, \tag{3.13}$$

If $\underset{\sim}{\beta}_t$ and $\underset{\sim}{\alpha}_t$'s follow the random walk-process, then both $G_t^{(1)}$ and $G_t^{(2)}$ are identity matrices. Moreover, as before, if $B_t = \text{diag}\{\gamma_1^2, \gamma_2^2\}$ and $Q_t = \nu^2 I$, then the model parameters ν^2, γ_1^2, γ_2^2 can be estimated in an analogous manner by the method of moments. The resulting EBLUP of $\underset{\sim}{\theta}_T$ will be denoted by $\underset{\sim}{g}_{8t}$.

4. MONTE CARLO STUDY

The cross-sectional and time series methods were compared empirically by means of a Monte Carlo simulation from a real time series obtained from Statistics Canada's biannual farm surveys, namely, the National Farm Survey (in June) and the January Farm Survey. Due to the redesign after the census of Agriculture in 1986, the survey data for the six time points starting with the summer of 1988 were employed to create a population for simulation purposes. To this, data from the census year 1986 was also added. Thus information at one more time point was available although this resulted in a 3-point gap in the series. The missing data points, however, can be easily handled by time series methods. It may be noted that although the data series is short, it is nevertheless believed to be adequate for illustrative purposes. The parameter of interest was taken as the total number of cattle and calves for each crop district (defined as the small area) at each time point. For simplicity, independent stratified random samples were drawn for each occasion from the pseudo-population although the farm surveys use rotating panels over time. The dependence of direct small area estimates over time was modelled by assuming that the underlying small area population totals are connected according to some random process. The auxiliary variable used in the model was the ratio-adjusted census '86 value of the total cattle and calves for each small area. This showed high correlations with the corresponding variable over time at the farm level. Specific details of the empirical study are described below.

4.1 Design of the simulation experiment

First we need to construct a pseudo-population from the survey data over six time points (June'88, Jan'89, ..., Jan'91). The actual design involves two frames (list and area) with a one stage stratified sampling from the list frame and a two stage stratified sampling from the area frame, for details see Julien and Maranda (1990). We decided to use survey data from the list frame only because the list frame corresponds to farms existing at the time of Census'86 and the chosen auxiliary variable for model building was based on Census'86 information. Moreover, we chose to use the data from the province of Quebec because its area sample is only a minor component of the total sample and the estimated coefficient variation for the twelve crop-districts (i.e. small areas of interest) of this province showed a wide range for the livestock variables.

It was decided to avoid variability due to changes in the underlying population over time by retaining only those farms which responded to all the six occasions. Also, farm units who belonged to a multiholding arrangement in any one of the seven time points (including the census) were excluded because of the problems in finding individual farm's data from the multiholding summary record and changes in their reporting arrangement over time.

The various exclusions described above were motivated from considerations of yielding a sharper comparison between small area estimators. The total count of farm units after exclusions was found to be 1160 out of a total of over 40,000 farms on the list frame. For the pseudo-population, we replicated the 1160 farm units proportional to their sampling weight so that the total size N of the pseudo-population was 10362 for micro-computer simulation.

The pseudo-population was stratified into four take-some and one take-all strata using Census'86 count data on cattle and calves as the stratification variable. The sigma-gap rule (Julien and Maranda, 1990) was used for defining the take-all stratum and the algorithm of Sethi (1963) was used for determining stratification boundaries for take-some strata. Neyman's optimum allocation was used for sample sizes for strata in order to achieve a high precision of the provincial estimate of total count. This resulted in a total sample size of 1036 (about 10% sampling rate) into allocations of 268, 325, 249 and 183 from takesome strata with 5001, 3188, 1850 and 312 farms, respectively, and the size of the take all stratum was 11. A total of 5000 simulations were performed. For each simulation, samples were drawn independently for each time point using stratified simple random sampling without replacement. The 5000 simulations were conducted in 2500 sets of 2 simulations where each set corresponds to a different vector of realized sample sizes in the twelve small areas within each stratum. This was required to compute certain conditional evaluation measures as described in the next subsection, see also Särndal and Hidiroglou (1989).

4.2 Evaluation Measures

Suppose m simulations are performed in which m_1 sets of different vectors of realized sample sizes in domains (h,k) are replicated m_2 times. The following measures can be used for comparing performance of different estimators at time T. Let i vary from 1 to m_1 and j from 1 to m_2.

(i) Absolute Relative Bias for area k

$$ARB_k = \left| (m^{-1} \sum_i \sum_j (est)_{ijk} - (true)_k) \,/\, (true)_k \right| \qquad (4.1)$$

The average of ARB_k over areas k will be denoted by $\overline{ARB}$.

(ii) Root Mean Square Conditional Relative Bias for area k

$$RMSCRB_k = \left\{ m_1^{-1} \sum_i (m_2^{-1} \sum_j (est)_{ijk} - (true_k))^2 \,/\, (true_k)^2 - B \right\}^{1/2} \qquad (4.2a)$$

$$B = m^{-1}(m_2 - 1)^{-1} \sum_i \left[\sum_j (est)_{ijk}^2 - \left(\sum_j (est)_{ijk} \right)^2 / m_2 \right] / (true)_k^2 \qquad (4.2b)$$

The correction term B adjusts for bias in the first term due to m_2 being finite. $\overline{RMSCRB}$ will denote the average of $RMSCRB_k$ over areas k.

(iii) Mean Absolute Relative Error for area k

$$MARE_k = m^{-1} \sum_i \sum_j \left| (est)_{ijk} - (true)_k \right| / (true)_k . \qquad (4.3)$$

and $\overline{MARE}$ denotes the average of $MARE_k$ over areas.

(iv) Root Mean Square Error for area k

$$RMSE_k = \left\{ m^{-1} \sum_i \sum_j ((est)_{ijk} - (true)_k)^2 \right\}^{1/2} \qquad (4.4)$$

and $\overline{RMSE}$ as before denotes the average over areas.

(v) Relative Root Mean Square Error for area k

$$RRMSE_k = RMSE_k \,/\, (true)_k . \qquad (4.5)$$

Again, we can define $\overline{RRMSE}$ as before.

The precision (i.e. the Monte Carlo Standard Error) of each measure depends on m_1, m_2. It can be seen that for all measures except (ii), the optimal choice

of m_1, m_2 under the restriction that $m_2 > 1$ is $m_1 = \frac{m}{2}$, $m_2 = 2$. For the second measure, the appropriate choice of m_1, m_2 is less straightforward. In the simulation study, m was chosen as 5000 and the corresponding values of m_1, m_2 were set at 2500 and 2.

4.3 Estimators used in the Comparative Study

There were thirteen estimators included in the study, namely, *M1-M8* corresponding to g_{1T} to g_{8T}, *M3a - M5a, M7a* corresponding g^a_{3T} to g^a_{5T} and g^a_{7T} when β is assumed known, and finally *M7b* corresponding g^b_{7T}, - see section 3 for the definition of the estimators. We used a simple linear regression model for the synthetic component with the auxiliary variable defined as

$$X_{kt} = (\hat{\theta}_t / \theta_1)\theta_{k1} \tag{4.6}$$

where θ_{k1}, θ_1 respectively denote the population totals for small area k and the province at $t=1$, i.e. at Census'86. The estimator $\hat{\theta}_t$ denotes the post-stratified estimator of θ_t from the farm survey at time t at the province level. Thus, X_{kt} is simply a ratio-adjusted synthetic variable. The variances of error components in the regression model were assumed to be constant over areas. For time series models, it was assumed that the serial dependence was generated by a random walk. The above type of model assumptions have been successfully used in many applications and the main reason for our choice was simplicity. It was hoped, however, that the chosen models might be adequate for our purpose and might illustrate the differential gains with different types of models.

Since the Census'86 data was included in the time series, the direct estimate g_{21} corresponds to Census'86 and therefore the survey error $\underset{\sim}{\epsilon}_1$ would be identically $\underset{\sim}{0}$. Moreover, from the definition of X_{kt}, it follows that a reasonable choice of (β_{11}, β_{21}) would be $(0,1)$ which implies that $\underset{\sim}{a}_1$ must be $\underset{\sim}{0}$. Thus the covariance matrices B_t and W_t at $t=1$ are null and therefore, the distribution of $\underset{\sim}{\alpha}_t$ at $t=1$ would not require estimation as was suggested in section 3. The above modification in the initial distribution of $\underset{\sim}{\alpha}_t$ is natural in view of the extra information available from the census. Moreover, since the direct estimates g_{2t} were not available for $t = 2, 3, 4$; equations for estimating various model parameters were modified accordingly. For instance, for method 6, in equation (3.9) the index for the first summation would start from 5 (not 1)

and T on the right hand side would be replaced by $(T-4)$. The equation (3.10) was modified as , for $i = 1, 2$,

$$(\hat{\beta}_{i5} - \hat{\beta}_{i1})^2 / (4\gamma_i^2 + c_{ii}^{(4)}) + \sum_{t=6}^{T} (\hat{\beta}_{it} - \hat{\beta}_{i,t-1})^2 / (\gamma_i^2 + d_{ii}^{(t)}) = T-4, \tag{4.7}$$

where $C_{ii}^{(4)}$ is the $(i, i)th$ element of $(F_4' U_4^{-1} F_4)^{-1}$.

For method 7 (case 1), $\underset{\sim}{\beta}_t$ would in general have a common fixed valued only for $t \geq 2$ because at $t = 1$, $\underset{\sim}{\beta}_t = (0,1)'$. The k^{th} diagonal element of Q_{kt} would be $(t-1)\nu^2$ and the equation for estimating ν^2 can be expressed as

$$\sum_{t=5}^{T} \sum_{h=1}^{k} (g_{2kt} - F_{kt}' \hat{\underset{\sim}{\beta}}_t)^2 / ((t-1)\nu^2 + v_{kt}) = (T-4)(k-2) \tag{4.8}$$

For methods 7a and 7b, similar modifications apply. Modifications for estimating model parameters for method 8 are quite analogous and details are omitted here. For methods *M3a* - *M5a* and *M7a*, the value of $\underset{\sim}{\beta}_t$ for $t \geq 2$ was fixed at (0,1) i.e. the same value corresponding to $t = 1$. For the sample size dependent estimators, *M5* and *M5a*, the parameter λ was taken to be 2 since preliminary empirical results suggested that for this study that was better than either 1 or 1.5.

4.4 Empirical Results

Figures 1 through 4 display some of the empirical results. Figure 1 shows plots of the five evaluation measures averaged over small areas relative to the Fay-Herriot (*M4*) value. There is a clear pattern in the behaviour of various measures across different estimators. The direct estimator *M2* does very well with respect to $\overline{ARB}$ and $\overline{RMSCRB}$ and does not do badly on the other measures. In fact, it appears that the cross-sectional smoothing methods *M3*, *M3a* (synthetic), and *M5* and *M5a* (sample size dependent) do quite poorly in this study. The Fay-Herriot methods *M4* and *M4a* perform no better than post-stratified and are much worse in terms of bias. The time series methods *M7* (also *M7a*; *M7b*) and *M8* perform somewhat worse than *M2* with regard to bias, but overall they perform very well. Note that the expansion estimator *M1* has a very large conditional bias. We have not shown the Monte Carlo standard errors in the figures but they are all found to be quite negligible. Figures 2 to 4 show plots of $RMSE_k$ for small areas

divided into three size groups, namely low, medium and high, based on the ranking of their true population totals at time T. They are divided up into these three groups because the errors of estimation would be expected to be larger for the larger totals. Area 6 stands out as being most difficult to estimate by the smoothing methods. The reason for this is that, while there was an overall decline of about 16% in the total number of cattle and calves in the pseudo-population from June, 1986 to January, 1991, the decrease for area 6 was the most extreme at over 33% so that the ratio adjusted covariate would be least appropriate for area 6. Nevertheless, the time series methods *M7*, *M7a*, *M7b*, and *M8* do as well as the post-stratified estimator for area 6 and for most other areas they do much better. This is because the random walk model for the small area effects is able to track small areas which, like area 6, progressively deviate from the model. Note that the time series method *M6*, which assumes the small area effects to be independent over time, does not do as well.

The main conclusions are listed in Section 1 and will not be repeated here.

5. CONCLUDING REMARKS

It was seen by means of a simulation study that small area estimation methods obtained by combining both cross-sectional and time series data could substantially improve performance of estimators based only on cross-sectional data. However, it was also seen that the Fay-Herriot smoothing does not necessarily lead to appreciable improvements in mean squared error, even when there is substantial cost in terms of bias. A question of obvious importance is whether it is possible in practical situations to judge if the gains from any type of smoothing would outweigh the costs, and how to make this judgement.

The models for the simulation study were chosen on general considerations. However, in practice, suitable diagnostics similar to those employed in Pfeffermann and Barnard (1991) should be performed before any model-based method can be recommended. Although we did not consider alternative stratifications or sample sizes in our simulation study, there is no reason to think that our conclusions would alter significantly if we were to do so. It should also be noted that the small area estimators can be modified to make them robust to misspecification of the underlying model; see e.g. the constraints used in Fay and Herriot (1979) and an alternative approach suggested by Pfeffermann and Burck

(1990). Further extension of the methods presented in this paper to the more realistic case of correlated sampling errors is currently being investigated.

ACKNOWLEDGEMENT

We would like to thank Jon Rao and Danny Pfeffermann for useful discussions and comments on an earlier version of this paper. The first author's research was supported in part by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa. Thanks are also due to Judy Clarke, Christine Larabie and Carmen Lacroix for their efficient processing the manuscript.

REFERENCES

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-components model for prediction of country crop areas using survey and satelite data. *Journal of the American Statistical Association*, 83, 28-36.

Bell, W.R., and Hillmer, S.C. (1987). Time series methods for survey estimation. *Proceeding of the American Statistical Association, Section Survey Research Methods*, 83-92.

Binder, D.A., and Dick, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

Choudhry, G.H., and Rao, J.N.K. (1988). Evaluation of small area estimators: an empirical study. Paper presented at the Symposium or Small Area Statistics, New Orleans, Aug. 26-27.

Choudhry, G.H., and Rao, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time*, eds. A.C. Singh and P Whitridge, 67-74.

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

Julien, C., and Maranda, F. (1990). Sample design of the 1988 national farm survey. *Survey Methodology*, 16, 117-129.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal Business Economics Statistics*, 9, 163-175.

Pfeffermann, D., and Barnard, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal Business Economics Statistics*, 9, 73-84.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

Platek, R., Rao, J.N.K., Särndal, C.E., and Singh, M.P. eds (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley & Sons.

Rao, J.N.K. (1986). Synthetic estimators, SPREE and best model-based predictors of small area means. Technical Report, Laboratory in Statistics and Probability, Carleton University, Ottawa.

Särndal, C.E., and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal American Statistical Association*, 69, 676-678.

Scott, A.J., and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal American Statistical Association*, 69, 674-678.

Sethi, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal Statistics*, 5, 20-33.

Tiller, R. (1989). A Kalman filter approach to labor force estimation using survey data. Paper presented to the Annual American Statistical Association Meeting, Washington, D.C.

Figure 1: evaluation measures relative to Fay–Herriot

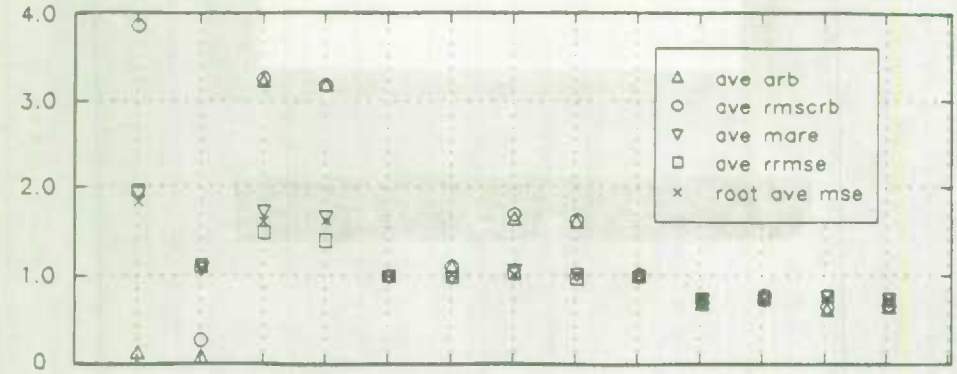


Figure 2: root mean squared errors, small small areas

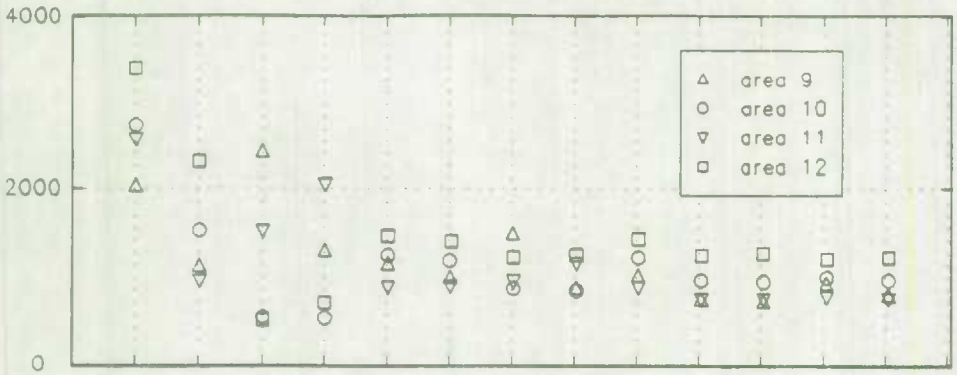


Figure 3: root mean squared errors, medium small areas

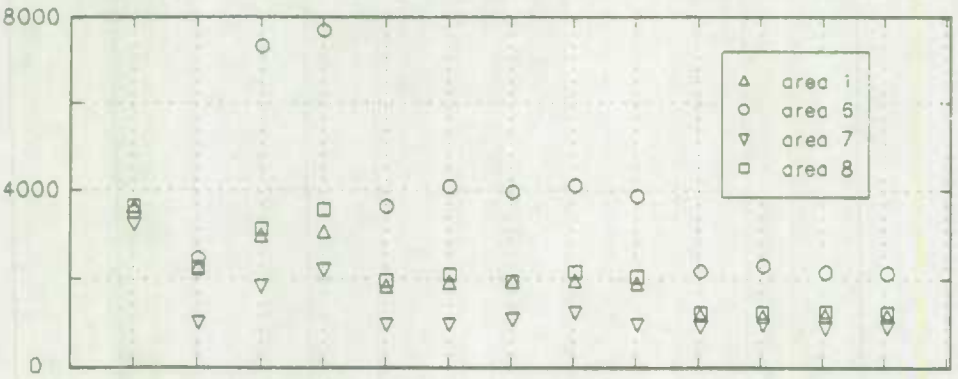


Figure 4: root mean squared errors, large small areas