SSC Annual Meeting, July 1995
Proceedings of the Survey Methods Section

BAYESIAN SAMPLING ERROR MODELLING WITH APPLICATION

W.R. Bell[1]

ABSTRACT

Certain analyses of published survey estimates have recently begun to account for sampling error in the published estimates. This includes time series signal extraction estimation in repeated surveys, small area estimation, seasonal adjustment, and benchmarking. These analyses require variances and covariances of the sampling errors. In some cases direct estimates of these quantities are used; in other cases the direct estimates are used to estimate a model for the sampling errors. In either case the analysis typically proceeds conditional on these results, thus failing to recognize uncertainty about the sampling error model. This paper presents a Bayesian approach to addressing this problem. The approach uses estimates of sampling error variances and covariances to develop a "posterior" for parameters of a model for the sampling errors. In one version of the approach this posterior becomes a "prior" for the subsequent analysis. This combines information from the estimated sampling error variances and covariances with information from the published survey estimates in a sensible way that recognizes uncertainty about the sampling error model parameters. This version assumes that the survey estimates are stochastically independent of the estimates of sampling error variances and covariances. An alternative version of the approach is presented that avoids this independence assumption. Classical (frequentist) approaches are also discussed. The approach is illustrated with an application that involves time series signal extraction estimation in a repeated survey.

RÉSUMÉ

Certaines analyses d'estimations provenant d'enquêtes ont récemment commencé à rendre compte de l'erreur échantillonnage dans les estimations publiées. Ceci inclut l'estimation de l'extraction du signal d'une série chronologique dans des enquêtes répétées, l'estimation pour de petites régions, l'ajustement saisonnier et l'étalonnement. Pour ces analyses, il faut connaître les variances et les covariances des erreurs d'échantillonnage. Dans certains cas des estimations directes de ces quantités sont utilisées; dans d'autres cas les estimations directes sont utilisées pour estimer un modèle pour les erreurs d'échantillonnage. Dans chaque cas, typiquement l'analyse procède conditionnellement à ces résultats, négligeant donc de reconnaître de l'incertitude dans le modèle sur les erreurs d'échantillonnage. Cet article présente une approche bayesienne pour faire face à ce problème. L'approche utilise des estimations des variances et covariances d'erreur d'échantillonnage pour développer un "a posteriori" pour les paramètres d'un modèle pour les erreurs d'échantillonnage. Dans une variante de cette approche, cet a posteriori devient un "a priori" pour des analyses subséquentes. Ceci combine d'une façon raisonnable, l'information contenue dans les variances et covariances estimées des erreurs d'échantillonnage avec l'information provenant des estimations publiées, reconnaissant ainsi l'incertitude sur les paramètres du modèle d'erreur échantillonnage. On suppose, dans cette variante, que les estimations d'enquête sont stochastiquement indépendantes des estimations des variances et covariances de l'erreur d'échantillonnage. On présente une autre variante de l'approche qui permet d'éviter cette hypothèse d'indépendance. On examine aussi les approches classiques (fréquentistes). L'approche est illustrée par une application qui comporte l'estimation de l'extraction de signaux chronologiques dans une enquête répétée.

1. INTRODUCTION

Considerable attention has been given in recent years to applying modelling techniques to improve direct estimates from sample surveys. The application of time series models and signal extraction results to estimates from repeated surveys, originally proposed by Scott and Smith (1974) and Scott, Smith, and Jones (1977), has been more recently investigated by, for example, Binder and Dick (1989, 1990), Bell and Hillmer (1990, 1994), Pfeffermann (1991), and Tiller (1992). Small area estimation using linear models with best linear unbiased prediction or Bayesian or empirical Bayesian analysis has been investigated by a number of authors -- see Ghosh and Rao (1994) for a survey. Dempster and Hwang (1993) and Ghosh and Nangia (1993) have investigated the combination of time series modelling and small area estimation techniques.

The various investigations into these problems have several things in common. Most start with direct sample survey estimators (such as weighted estimates of totals), or composite estimators derived from these, that are thought to be not accurate enough because of high sampling variance due to inadequate sample sizes. In small area applications there may be no sample at all in some small areas. A model is then developed for the direct survey estimators that incorporates models for both the true underlying population quantities and the sampling errors. Using this model, improved estimators of the true quantities are produced by standard results that can

[1] William R. Bell, Statistical Research Division, Bureau of the Census, Washington, DC 20233-9100.

alternatively be thought of as best linear unbiased prediction, signal extraction, conditional expectation under multivariate normality, empirical Bayesian smoothing, or computation of Bayesian posterior means. Estimates of variances of the error of the improved estimators are also produced, either on the assumption that the fitted model is the truth, or using asymptotic approximations or Bayesian calculations to reflect uncertainty due to not knowing some model parameters.

Relatively little attention has been paid in this process to the development of sampling error models. In small area estimation it is more or less standard for the sampling error "model" to involve no more than direct estimates of the sampling error variances (and sometimes certain covariances). In both small area and time series applications, it has been standard to assume the sampling error model is known when the model for the direct survey estimators (including the model for the true population quantities) is estimated, and when linear prediction calculations are made. Thus, no account is taken of uncertainty about the sampling error model when variances of the errors in the improved estimates are calculated.

This paper focuses on the development of sampling error models, and on their use in making inferences about the true population quantities (in combination with models for the latter) for time series and small area estimation. A Bayesian framework is emphasized, though analogous calculations from a classical (frequentist) perspective are also discussed. Attention is focused primarily on posterior means and variances of the true population quantities. The Bayesian posterior variances allow for uncertainty about both the parameters of the model for the true quantities and the parameters of the model for the sampling errors. Under certain assumptions, our approach also allows combining information about the sampling error model parameters from two data sources: the direct estimates of sampling error variances and covariances, and the direct survey estimators.

Section 2 outlines a general model framework covering both time series and small area applications, and discusses classical and Bayesian analysis for this model in general terms. The details of the approach are illustrated in Section 3 with a time series example for which the sampling error model is quite simple. A primary objective in the analysis is to examine how uncertainty about all the model parameters, both those for the sampling error model and the model for the true population quantities, translates into uncertainty in the inferences about the true population quantities. For the example it is found that uncertainty about the parameters contributes a small amount of uncertainty to inferences about the population quantities.

2. GENERAL MODEL -- CLASSICAL AND BAYESIAN ANALYSIS

This section outlines a general model that covers both time series and small area applications. First, the following decomposition is assumed:

$$y_i = Y_i + e_i \qquad i = 1, \ldots, n \tag{2.1}$$

y_i = direct survey estimators

Y_i = population characteristics(" truth ")

e_i = sampling errors.

The index i could index small areas, or time points, or both. Alternatively, the y_i in (2.1) could be some transformation of original direct estimators z_i, such as $y_i = \log(z_i)$. This is particularly useful for removing dependence of the sampling error variance on the level of the corresponding population quantities being estimated. (See Bell and Hillmer 1990.) For simplicity, we shall work with (2.1), understanding that, if the y_i are transformed data, then reverse transformation will need to be applied to any improved estimates obtained.

In matrix-vector notation, assuming normality, the general model for $y = (y_1, \ldots, y_n)'$ is

$$y = Y + e \qquad Y = (Y_1, \ldots, Y_n)', \text{ etc.}$$

$$Y = X\beta + u \qquad u \sim N(0, \Sigma(\psi)) \tag{2.2}$$

$$e \sim N(0, V(\eta)) \quad \text{independent of } u.$$

The data assumed available are both y and C, the latter being a direct estimate of the variance-covariance matrix of the sampling errors e. The mean of e is taken as 0, an assumption that the original estimators y are design unbiased. The parameters of the model (2.2) are the $p\times1$ vector of regression parameters β, and the $r\times1$ and $m\times1$ vectors of parameters ψ and η that determine the $n\times n$ covariance matrices Σ and V. (I write $\Sigma(\psi)$ and $V(\eta)$ only when this is needed to emphasize the explicit dependence of Σ and V on the parameters ψ and η.) Having postulated a model of form (2.2), the task is to use the data y and C to make inferences about the parameters $\alpha = (\beta, \psi, \eta)$ and, ultimately, about the true population quantities Y.

2.1 Sampling Error Modelling

The first step is to use C to make inferences about η, the parameters of the sampling error model. As noted earlier, it has been standard in small area estimation to assume $V = C$, in which case η corresponds to all distinct elements of V. (This need not be $n(n+1)/2$ elements if certain covariances are known to be or assumed to be zero.) Two exceptions involve small domain estimation of census adjustment factors. Isaki, Huang, and Tsay (1991), for the U.S., and Dick (1995), for Canada, both replaced direct estimates of sampling error variances of adjustment factors by fitted values obtained from a regression equation. Isaki, Huang, and Tsay retained direct estimates of sampling error correlations; Dick assumed these all to be zero. Parametric sampling error models have been used more in time series applications, with C used to estimate η in some ad-hoc fashion to make $V(\eta)$ approximate C (*e.g.*, Bell and Hillmer (1990, 1994), Binder and Dick (1990), and Tiller (1992)).

The present paper pursues the following approach, suggested by Bell and Otto (1993), that is based on assuming a Wishart distribution (DeGroot 1970, Section 5.5) for C as a working model:

$$\nu C \sim \text{Wishart}\ (\nu, V(\eta)). \tag{2.3}$$

Thus, the mean matrix of the Wishart distribution, $V(\eta)$, is determined by the sampling error model parameters η. Generally, the degrees of freedom parameter ν will be unknown. In some cases ν can be set to an estimated value, and we then behave as if ν is known -- this is done in the example of Section 3. Alternatively, we could let ν, along with V, depend on model parameters included in η (expanding the definition of η). We would then write $\nu(\eta)$. In any case, for the remainder of this section we may regard η as defining the unknown parameters of the model (2.3).

If the sample covariance estimates (elements of C) were of the form $c_{ij} = K^{-1}\sum_{k=1}^{K}(y_{ik}-y_i)(y_{jk}-y_j)$, where the y_{ik} are observations on individual units k=1, ..., K for time point or small area i, the survey estimators y_i are the sample means ($y_i = K^{-1}\sum_{k=1}^{K} y_{ik}$), and $(y_{1k}, ..., y_{nk})'$ for k=1, ..., K are assumed *iid* multivariate normal, then C will indeed have a Wishart distribution with ν=K−1. Furthermore, C will be independent of the data y, a condition relevant to the discussion in Section 2.2. The same results obviously hold if the y_i are N times the sample means (estimating population totals Y_i), where N is the population size. Though estimates of sampling error variances and covariances are rarely of this simple form, the Wishart model may still prove useful since it provides an objective means of using the data C in making inferences about η. Following a classical approach, the model (2.3) can be estimated by maximizing the Wishart likelihood

$$L(\eta\mid C) = g(\nu)\,|V(\eta)|^{-\nu/2}\exp\{-(1/2)tr[V(\eta)^{-1}C]\} \tag{2.4}$$

where $g(\nu)$ includes those terms in the Wishart density not explicitly present in (2.4) -- these involve ν but not V. Following a Bayesian approach, (2.4) can be multiplied by a (possibly noninformative) prior distribution, $p(\eta)$, which yields something proportional to the posterior $p(\eta\mid C)$.

Knowledge of the sample design should be taken into account in specifying the sampling error model, i.e., in specifying $V(\eta)$. In particular, if samples for different areas, or for different time points, are drawn independently, then the corresponding sampling errors will be approximately uncorrelated, which constrains certain elements of $V(\eta)$ to zero. In some cases, with a suitable ordering of the observations y, $V(\eta)$, and correspondingly C, will be block diagonal. When this occurs the diagonal blocks of C can be assumed to have independent Wishart distributions analogous to (2.3).

2.2 Modelling the Direct Estimates y

The next step is specification of the model for Y, which completes specification of the model (2.2) for y. In small area applications, attention has focused on selection of the regression predictor variables in the X matrix. The usual assumption about $\Sigma = Var(u)$ in small area estimation has been $\Sigma = \sigma^2 I$, though Ghosh and Rao (1994) point out some exceptions. In time series applications regression variables may also be present in the model, but most of the attention focuses on modelling correlation between the $u_i s$ (autocorrelation). The resulting time series model for u_i essentially is the parametric representation for $Var(u)$ as $\Sigma(\psi)$. Generalizations are needed to handle nonstationary time series models, *i.e.*, models that involve differencing. For simplicity, we shall omit discussion of this here.

Once the full model (2.2) has been specified, the next step is to use the data y to make inferences about β and ψ, and possibly to update the inference about η. Four possible approaches are discussed below. These are defined by whether a classical or Bayesian approach is taken, and by whether y and C are assumed independent. As noted in Section 2.1, independence of y and C holds under the usual conditions leading to the Wishart distribution for C. The typical survey estimators and variance estimators do not satisfy these conditions, however, thus an assumption of independence of y and C should not be made simply as a matter of course, but should be considered in regard to its appropriateness for any specific application. The assumption of independence is what allows the combining of information from both y and C in making inferences about η. If independence is assumed when y and C are actually positively related, then the resulting inference would tend to understate uncertainty about η and, thus, about Y.

If y and C are not assumed independent, then a classical approach would take an estimate $\hat{\eta}$ of η

obtained from C in the sampling error modelling, compute $V(\hat{\eta})$, and hold V fixed at $V(\hat{\eta})$ when estimating (2.2) by maximum likelihood. In this approach, the idea behind the sampling error modelling is that the parametric assumptions about the form of V result in $V(\hat{\eta})$ being a better estimate of V than the direct estimator C. This approach assumes nothing about the relation between y and C, and does not use y to update the estimate of η obtained from C.

A Bayesian approach that does not assume y and C independent starts with simulations from the distribution corresponding to $p(\eta \mid C)$, or from a suitable approximation (*e.g.*, an asymptotic normal approximation) to this distribution. Assume such simulations can be produced. Then for each simulated η, we require simulations from the distribution corresponding to $p(\beta, \psi \mid y, \eta) = p(\beta \mid y, \psi, \eta) p(\psi \mid y, \eta)$. Simulation of β is straightforward because the conditional distribution of β given (y, ψ, η) is normal. (See Bell and Otto (1993)). Simulating ψ from its conditional distribution given (y, η) poses a problem, however. Even in the simplest case where the Y_i are assumed i.i.d $N(0, \sigma^2)$, the conditional distribution of $\psi = \sigma^2$ or of $\psi = \tau = 1/\sigma^2$ given (y, η) is not a standard distribution. Since $p(\psi \mid y, \eta)$ can be readily calculated, simulations of ψ for given η could be obtained by specialized simulation techniques as discussed in Gelfand and Smith (1990), or using an asymptotic normal approximation to $p(\psi \mid y, \eta)$. The difficulty is that these computations will need to be redone for each simulated η to simulate ψ and then β, making this approach computationally intensive. For now, the question of how to best implement this approach remains open.

If y and C are assumed independent, then information in y and C can be combined via the joint likelihood function, $L(\alpha \mid y) L(\eta \mid C)$, where $L(\alpha \mid y)$ is the $N(X\beta, \Sigma(\psi) + V(\eta))$ likelihood, and $L(\eta \mid C)$ is the Wishart likelihood given in (2.4). From a classical perspective, inference might then be based on an assumed asymptotic normal distribution for the maximum likelihood estimate of α. This would recognize uncertainty about η, though I am unaware that such results have actually been established for this case, particularly including results that carry through to recognize uncertainty about η in making inferences about Y. (See the discussion in the next section.)

From a Bayesian perspective, if y and C are assumed independent, the combination of their information through the joint likelihood can be achieved by taking the posterior $p(\eta \mid C)$ from the sampling error modelling as a "prior" for η when analyzing the model (2.2). Coupled with a (possibly noninformative) prior for β and ψ, $p(\beta, \psi)$, these can be updated to a posterior for the full set of parameters α given both C and y, as follows:

$$p(\alpha \mid y, C) \propto p(y \mid \alpha) p(C \mid \alpha) p(\alpha) \tag{2.5}$$

$$\propto p(y \mid \alpha) p(\eta \mid C) p(\beta, \psi). \tag{2.6}$$

The posterior $p(\eta \mid C)$ reflects uncertainty about η from the data C. (2.6) then uses the data y to update the information about η obtained from C. Assuming values of α can be simulated from the distribution corresponding to the density (2.6), the Bayesian approach easily carries through to recognizing uncertainty about all the parameters when making inferences about Y, as discussed in the next section. Since (2.6) is readily calculated, such simulations can generally be obtained from specialized techniques (Gelfand and Smith 1990), or from asymptotic approximation by a normal distribution with mean given by the value $\hat{\alpha}$ that maximizes the posterior density (2.6), and with variance matrix given by the inverse Hessian (second derivative) matrix of $-\log[p(\alpha \mid y, C)]$ evaluated at $\hat{\alpha}$. Certain reparameterizations, such as variances to log-variances, can improve the normal approximation.

2.3 Inference About the True Population Quantities Y

Given the parameters α, the conditional distribution of Y given y is normal with mean and variance

$$E(Y \mid y, \alpha) = X\beta + \Sigma\, \Omega^{-1}(y - X\beta) \tag{2.7}$$

$$Var(Y \mid y, \alpha) = \Sigma - \Sigma \Omega^{-1} \Sigma, \tag{2.8}$$

where $\Omega = \Sigma(\psi) + V(\eta)$. These results are the foundation for inference, classical or Bayesian, about Y. We first discuss the Bayesian approach, explicitly considering the case where y and C are assumed independent. The treatment when y and C are not assumed independent is essentially the same, except that the parameters α are simulated differently. (See Section 2.2.)

In principle, one could consider adding C to the conditioning sets in (2.7) and (2.8). However, it is unclear how $E(Y \mid y, C, \alpha)$ would be computed in general, or that much would be gained from doing so. In what follows, it should be understood that when we write $p(Y \mid y, C)$ or $E(Y \mid y, C)$ these quantities depend on C only through the dependence of the posterior distribution of the parameters on C, that is, only through $p(\eta \mid C)$.

Assuming we can simulate values of α from $p(\alpha \mid y, C)$ we can compute (2.7) and (2.8) for each simulated α, and then simulate Y from the corresponding multivariate normal distribution, thus simulating from $p(Y|y, C)$. Alternatively, if only the posterior means and variances of the elements Y_i of Y are desired, these are:

$$\begin{aligned} E(Y_i|y, C) &= E[E(Y_i|y, \alpha)|y, C] \\ &= \int E(Y_i|y, \alpha) p(\alpha|y, C)\, d\alpha \end{aligned} \tag{2.9}$$

$$\begin{aligned} Var(Y_i|y, C) &= Var[E(Y_i|y, \alpha)|y, C] \\ &\quad + E[Var(Y_i|y, \alpha)|y, C) \end{aligned} \tag{2.10}$$

We can approximate (2.9) for each $i = 1, ..., n$ via Monte Carlo integration by: (1) simulating α from the distribution corresponding to $p(\alpha|y, C)$; (2) computing $E(Y_i|y, \alpha)$ for each simulated α; and (3) taking the sample mean of the $E(Y_i|y,\alpha)$'s over the simulations. To approximate (2.10) we proceed similarly, evaluating both $E(Y_i|y,\alpha)$ and $Var(Y_i|y,\alpha)$ at step (2), then taking the sample variance and the sample mean over the simulations of these respective results, and then adding these together to get $Var(Y_i|y, C)$.

The above assumes that we can simulate from the posterior distribution with density $p(\alpha|y, C)$. As discussed in Section 2.2, this might be done using an asymptotic normal approximation, or using specialized Bayesian simulation techniques. If only the posterior means and variances (2.9) and (2.10) are desired, then importance sampling is an attractive approach that does not require simulation from the exact posterior -- see Geweke (1989).

The classical approach to inference about Y having estimated the model (2.2) is to plug the estimated parameters into (2.7) and (2.8). This assumes the model (2.2) is known exactly. An improvement results from making an addition to (2.8) to account for the effect of error in estimating the model parameters. Accounting for the additional variance due to estimating β is straightforward. As discussed in Bell and Otto (1993), the additional term follows from standard generalized least squares results. Accounting for additional variance due to estimating ψ and η is not straightforward. Methods of accounting for variance due to estimating ψ have been given by Prasad and Rao (1990) for small area estimation and by Binder, Bleuer, and Dick (1993) for time series applications. These approaches rely on somewhat complicated asymptotic approximations. I am unaware of any classical methods developed to account for variance due to estimating the sampling error model parameters η from C. In contrast, one of the primary advantages to the Bayesian approach is its conceptually straightforward treatment of both nonlinear sources of variation -- error due to not knowing ψ *and* η. The price paid in the Bayesian approach for this conceptual simplicity and potentially better results is an increase in computational complexity, which can be handled by simulation.

3. EXAMPLE: 5+ UNIT HOUSING STARTS, SOUTH REGION OF THE U.S.

This time series example uses data on 5 or more unit housing starts in the south region of the U. S. Figure 1 shows a plot of the time series of original monthly survey estimates (z_t) from January 1975 through November 1988 (167 observations). (The data were obtained from the Manufacturing and Construction Statistics Division of the Census Bureau. The scale is omitted from the plot to prevent identification of the actual numbers of this unpublished series.) This data was previously analyzed by Bell and Hillmer (1994), who estimated the following model for $y_t = \log(z_t)$:

$$\begin{gathered} y_t = Y_t + e_t \qquad t = 1, ..., 167 \\ (1-B)(1-B^{12})Y_t = (1-\theta_1 B)(1-\theta_{12}B^{12})a_t \quad a_t\, i.i.d.\, N(0,\sigma_a^2) \\ e_t\, i.i.d.\, N(0,\sigma_e^2) \end{gathered} \tag{3.1}$$

The parameter estimates were $\tilde{\theta}_1 = .53$, $\tilde{\theta}_{12} = .80$, $\tilde{\sigma}_a^2 = .0510$ and $\tilde{\sigma}_e^2 = .0162$. The estimate of σ_e^2 was obtained by averaging estimates of relative sampling error variances for each of the 60 months of the years 1982 through 1986. Averaging estimated sampling error autocorrelations for a given lag over the 60 months yielded small values, suggesting little autocorrelation is present in e_t and leading to its white noise model. The remaining parameters of (3.1) were then estimated by maximum likelihood using the time series data $y = (y_1, ..., y_{167})'$ but holding σ_e^2 fixed at its estimated value of .0162. Bell and Hillmer (1994) used this model for signal extraction estimation of Y_t assuming the parameters were known to be equal to their estimated values. Here, we shall obtain posterior means and variances of Y_t, allowing for uncertainty about all the parameters. Many of the results presented here are taken from a more detailed Bayesian analysis by Bell and Otto (1993), who investigated model-based seasonal adjustment of this series.

3.1 Sampling Error Model

For this example the only parameter of the sampling error model is $Var(e_t) = \sigma_e^2$. We assume that the relative sampling error variance estimates, $(s_1^2, ..., s_m^2)$ are $m = 60$ independent, unbiased estimates of σ_e^2. Since the e_t are assumed independent over t, $V = \sigma_e^2 I$, and $C = \text{diag}(s_1^2, ..., s_m^2)$. (Note that since $m < n$ the sample variance estimates available do not span the full time series of y_t values.) The Wishart model (2.3) then reduces to assuming that $\nu s_i^2/\sigma_e^2 \sim \chi_\nu^2$, where ν is the degrees of freedom assigned to each s_i^2. Using the noninformative prior $p(\sigma_e^2) \propto 1/\sigma_e^2$, Bell and Otto (1993) show that, conditional on ν, the posterior of σ_e^2 is a scaled inverse $\chi_{m\nu}^2$ distribution (a $\chi_{m\nu}^2$ distribution for $\nu \sum_i s_i^2/\sigma_e^2$). This is a generalization of a standard Bayesian result for the case $m = 1$ (Box and Tiao 1973, Theorem 2.3.1).

The appropriate value for ν is unknown. Bell and Otto (1993) examined two ways of dealing with this problem. First, they used the result that $Var(s_i^2)/[E(s_i^2)]^2 = 2/\nu$, to estimate ν by $\hat{\nu} = 2(\bar{s}^2)^2/\hat{Var}(s_i^2)$ where $\bar{s}^2$ and $\hat{Var}(s_i^2)$ are the sample mean and variance of $(s_1^2, ..., s_m^2)$. This gave $\hat{\nu} = 5.1$. Then $p(\sigma_e^2|C \doteq p(\sigma_e^2|\hat{\nu}, C)$ is defined by $m\hat{\nu}\bar{s}^2/\sigma_e^2 = 4.96/\sigma_e^2 \sim \chi_{306}^2(m\hat{\nu} = 306)$. Alternatively, to recognize uncertainty about ν via a fully Bayesian approach, Bell and Otto multiplied $p(\sigma_e^2|\nu, C)$ by $p(\nu)$, a prior distribution for ν, and numerically integrated the product over ν to get the marginal posterior of σ_e^2. Essentially identical results were obtained using a flat prior ($p(\nu) \propto$ constant) or the prior obtained using Jeffreys' rule. Bell and Otto then compared $p(\sigma_e^2|C)$ resulting from exact (Jeffreys' prior) and approximate ($\nu = 5.1$) treatments of the chi-squared model, also comparing these results with $p(\sigma_e^2|C)$ obtained from exact and approximate treatments of a lognormal model for the s_i^2. All these versions of $p(\sigma_e^2|C)$ peak near .0162, the estimate of σ_e^2 obtained by Bell and Hillmer (1994), though there are some differences in the shapes of the various posteriors. Comparisons of maximized likelihood values slightly favored the lognormal model. It turns out, however, that the exact posterior for the lognormal model is very close to the approximate posterior with $\nu = 5.1$ under the chi-squared model. As in Bell and Otto (1993), I shall therefore proceed with the latter, which is easier to evaluate. This now becomes the "prior" for σ_e^2 for the analysis of the next section.

3.2 Posterior Distribution for the Parameters

The time series data y can now be used to develop a posterior distribution for the model parameters $\alpha = (\theta_1, \theta_{12}, \sigma_a^2, \sigma_e^2)'$. From (2.6) the posterior density is

$$p(\alpha|y, C) \propto p(y|\alpha)\, p(\sigma_e^2|C)\, p(\theta_1, \theta_{12}, \sigma_a^2)$$

where $p(\sigma_e^2|C)$ is defined by inverse chi-squared distribution developed above, and $p(y|\alpha)$ is a multivariate normal density that can be computed by the suitably initialized Kalman filter as discussed by Bell and Otto. We use the noninformative prior $p(\theta_1, \theta_{12}, \sigma_a^2) \propto 1/\sigma_a^2$ truncated to $|\theta_1| \leq 1$, $|\theta_{12}| \leq 1$, $\sigma_a^2 > 0$. A multivariate normal approximation to the posterior for $(\theta_1, \theta_{12}, \log(\sigma_a^2), \log(\sigma_e^2))'$ (restricted to $|\theta_1| \leq 1$, $|\theta_{12}| \leq 1$) has mean vector given by the maximum posterior density estimates, $(\hat{\theta}_1, \hat{\theta}_{12}, \log(\hat{\sigma}_a^2), \log(\hat{\sigma}_e^2))'$, and covariance matrix given by the inverse Hessian (second derivative) matrix of $-\log(p(\theta_1, \theta_{12}, \log(\sigma_a^2), \log(\sigma_e^2)|y, C))$, evaluated at $(\hat{\theta}_1, \hat{\theta}_{12}, \log(\hat{\sigma}_a^2), \log(\hat{\sigma}_e^2))'$. Bell and Otto report the following estimates, with standard errors from the inverse Hessian in parentheses:

$$\hat{\theta}_1 = .52 \quad (.083) \qquad \hat{\theta}_{12} = .79 \quad (.11) \qquad (3.2)$$

$$\log(\hat{\sigma}_a^2) = -2.98 \quad (.17) \quad \Rightarrow \quad E(\sigma_a^2|y, C) = .0516.$$

$$\log(\hat{\sigma}_e^2) = -4.12 \quad (.081) \quad \Rightarrow \quad E(\sigma_e^2|y, C) = .0162.$$

The Hessian matrix obtained using numerical second derivatives also produced the following posterior correlation matrix of the parameter vector $(\theta_1, \theta_{12}, \log(\sigma_a^2), \log(\sigma_e^2))'$:

$$\begin{matrix} 1.00 & & & \\ .01 & 1.00 & & \\ .29 & -.12 & 1.00 & \\ -.11 & -.03 & -.22 & 1.00 \end{matrix} \qquad (3.3)$$

Note that the posterior mode estimates are very close to the parameter estimates obtained by Bell and Hillmer (1994), and that (3.3) shows little posterior correlation between the parameters.

3.3. Posterior Means and Variances of Y_t

The posterior means and variances of the true series Y_t were computed by simulation as discussed in Section 2.3. The simulations from the exact posterior distribution of α obtained in Bell and Otto (1993) were used. Figure 1 shows both the time series of the original direct estimates z_t along with exponentiated posterior means, $\exp(E[Y_t|y, C])$, which shall be called the "Bayes estimates". The smoothing that results from computing the posterior means is evident. Figure 2 shows the corresponding estimates of (multiplicative) sampling error, $\exp(E[e_t|y, C])$, which are the ratios of the z_t to the $\exp(E[Y_t|y, C])$. This shows that many of the Bayes estimates differ from the direct estimates by more than 5 percent, and several by more than 10 percent. Figure 3 shows the ratios of the Bayes estimates to the classical estimates Bell and Hillmer (1994) obtained by exponentiating the result from plugging estimated parameters into (2.7) (actually, into an analogous expression appropriate for nonstationary time series models). We see the differences between the classical and Bayesian point estimates are small: only a few exceed one half of one percent in magnitude. Thus, uncertainty about the model parameters has very little effect on point estimates for this example.

Figure 4 shows posterior variances for Y_t along with variances obtained by plugging the parameter estimates of Bell and Hillmer into (2.8) (actually, into the appropriate analog for nonstationary models). The latter is very close to $E[Var(Y_t|y,\alpha)|y,C]$ the second term on the right hand side of (2.10). Figure 4 shows that the posterior variances nearly always exceed the variances conditional on parameter estimates, but the amount of the difference varies erratically over time. Also note that both the variance curves tend to increase toward the end of the time series. This is a standard result that reflects the fact that data close to the time point for which Y_t is being estimated provide more information than data further away, so that variances are lowest in the center of the series, where there is more of this highly relevant information. Figure 5 shows the ratios of $Var[E(Y_t|y,\alpha)|y,C]$, the first term in (2.10), to the posterior variances $Var(Y_t|y,C)$. This provides a measure of the contribution of parameter uncertainty to posterior variance, and quantifies how different the two curves are in Figure 4. We see that only a few of the ratios in Figure 5 exceed 4 percent, though two of them approach 8 and 10 percent. This shows that for this example parameter uncertainty has a generally small effect on posterior variances, though the magnitude of this effect varies erratically over time.

Disclaimer. This paper reports the general results of research undertaken by Census Bureau staff. The views are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

Bell, William R., and Hillmer, Steven C. (1990). "The time series approach to estimation for repeated surveys," *Survey Methodology*, 16, 195-215.

Bell, William R., and Hillmer, Steven C. (1994), "Applying Time Series Models in Survey Estimation," submitted for publication.

Bell, William R., and Otto, Mark C. (1993). "Bayesian assessment of uncertainty in seasonal adjustment with sampling error present," submitted for publication.

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

Binder, David A., and Dick, J. Peter (1990). "A method for the analysis of seasonal ARIMA models," *Survey Methodology*, 16, 239-253.

Binder, David A., Bleuer, Susana R., and Dick, J. Peter (1993). "Time series methods applied to survey data," paper presented at the meeting of the International Statistical Institute.

Box, G.E.P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Reading, Massachusetts: Addison-Wesley.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*, New York: McGraw-Hill.

Dempster, A. P., and Hwang, J. S. (1993). "Component models and Bayesian technology for estimation of state employment and unemployment rates," *Proceedings of the 1993 Annual Research Conference*, U. S. Bureau of the Census, 571-581.

Dick, Peter (1995). "Modelling net undercoverage in the 1991 Canadian Census," *Survey Methodology*, 21, 45-54.

Gelfand, A. E., and Smith, A. F. M. (1990). "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.

Geweke, J. (1989). "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, 57, 1317-1339.

Ghosh, M., and Nangia, N. (1993). "Estimation of median income of four-person families: a Bayesian time series approach," *Proceedings of the 1993 Annual Research Conference*, U. S. Bureau of the Census, 555-570.

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

Isaki, C.T., Huang, E.T., and Tsay, J.H. (1991). "Smoothing adjustment factors from the 1990 Post Enumeration Survey," *American Statistical Association, Proceedings of the Social Statistics Section*, 338-343.

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

Prasad, N.G.N., and Rao, J.N.K. (1990). "The estimation of mean squared errors of small area estimators," *Journal of the American Statistical Association*, 78, 47-59.

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

Scott, A. J., Smith, T.M.F., and Jones, R. G. (1977). "The application of time series methods to the analysis of repeated surveys," *International Statistical Review*, 45, 13-28.

Tiller, Richard B. (1992). "Time series modeling of sample survey data from the U.S. Current Population Survey," *Journal of Official Statistics*, 8, 149-166.

Fig. 1 South 5+ Unit Housing Starts

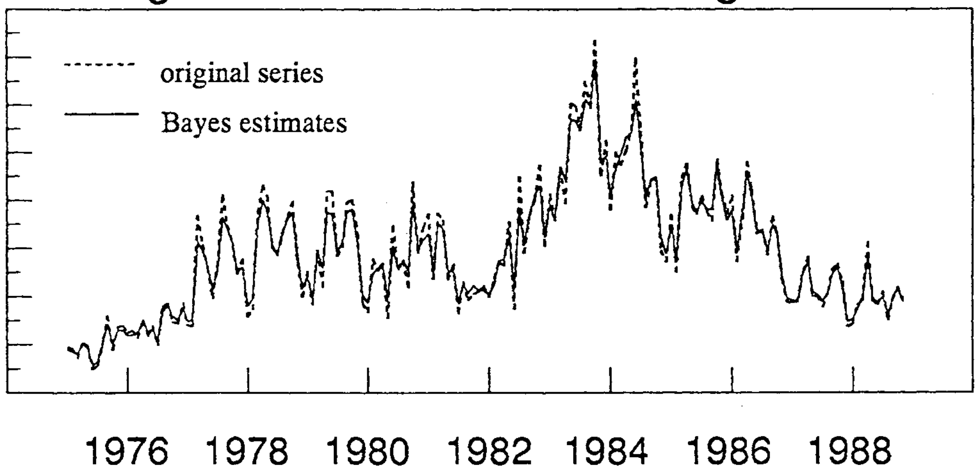

Fig. 2 exp(Posterior Means of e_t)

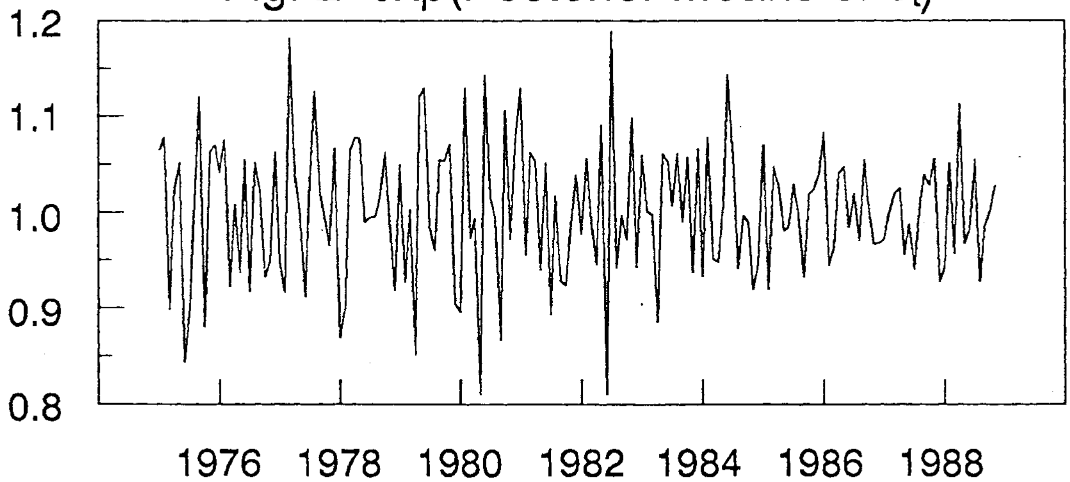

Fig. 3 Bayes Estimates/Plug-in Estimates

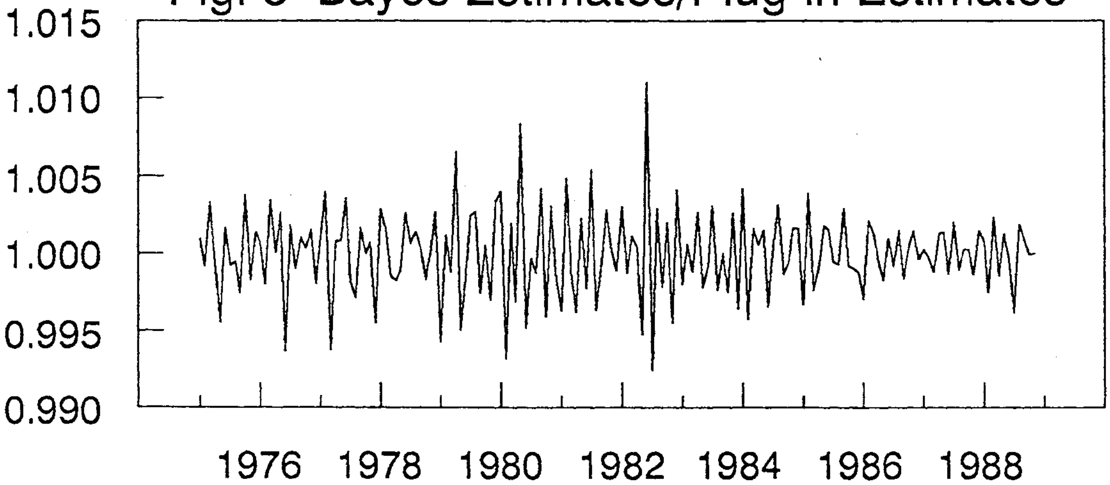

Figure 4.

Posterior Variances and Variances Conditional on the Parameters

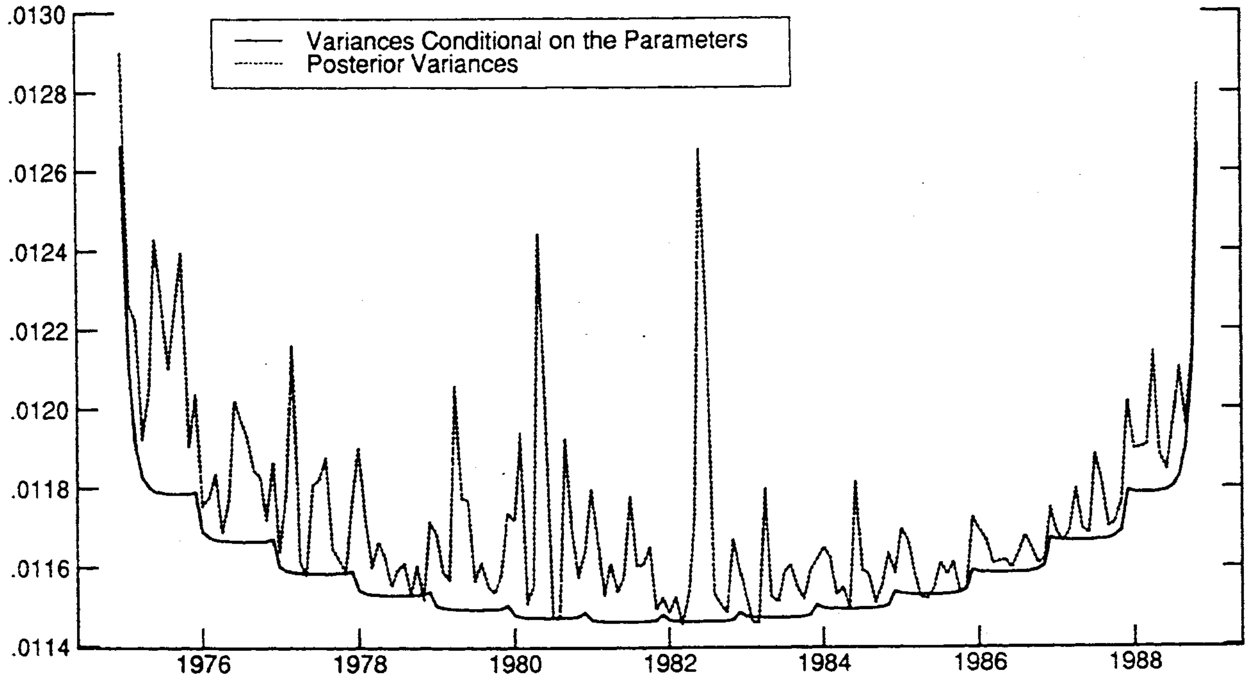

Figure 5.

Proportion of Posterior Variances that is due to Parameter Uncertainty