Time series modeling of repeated survey data for estimation of finite population parameters

Danny Pfeffermann
University of Southampton, UK;
Hebrew University of Jerusalem, Israel.

Paper in honor of Alastair Scott and Fred Smith

Abstract

In the first part of the article, I review and discuss the pioneering contributions of the late Alastair Scott and T.M.F Smith to time series analysis of repeated survey data. In the second part, I review and discuss some of the extensive theoretical and applied developments in this area, emerging from their work over the ensuing 40 years or so. I conclude with a brief summary of Scott and Smith contributions and extensions, with some remarks on possible advances and challenges to time series analysis of repeated surveys.

Key words: Basic structural model; Benchmarking; Big data; Autocorrelated sampling errors; Design-based estimation; State-space models.

Acknowledgement: I am grateful to Jan Brakel van den from Statistics Netherlands and William Bell from the Census Bureau in the U.S. for sharing with me their important contributions related to the topic of this article.

1. Introduction

Most of the population and business surveys carried out by Statistical Bureaus all over the world are repeated over time at regular time intervals between them, such as every month, quarter, or annually. In some surveys a new sample is drawn every time the survey is carried out, but very often, the samples taken at the different times are partially overlapping, whether by design, or because the same sample is intended to be always surveyed, but some of the units drop out and others join the sample instead. The result of this process is a sequence of estimates, published regularly, forming a genuine time series. Familiar examples include employment and unemployment rates from Labor Force Surveys, mean income (or income inequality) estimates from Household Expenditure Surveys, and industrial production or trends of production from Business Surveys.

Classical survey sampling theory considers the unknown population values of a target variable as constants, basing the inference solely on the randomization distribution, implied by the random sample selection, as defined by the sampling design. See, e.g., Cochran (1977, Sections 12.10-12.13) and Binder and Dick (1989) for review of the main results of estimation of population means from repeated surveys under this theory. A fundamental paper of this approach is of Patterson (1950). The author considered the case of partially overlapping samples with exponentially decaying autocorrelations between observations relating to the same sampled units. With some additional assumptions, Patterson (1950) derived the Minimum Variance Linear Unbiased Estimators (MVLUE) of the population means for the current time point, previous time points and for the change between two successive time points.

The pioneering contribution of Alastair Scott and Fred Smith (hereafter S&S) to inference from repeated surveys was to consider the unknown target population parameters such as means, proportions, etc. as realizations of a stationary time series model, which evolves stochastically over time. Specifically, a model is assumed, which consists of two parts. The first part accounts for the statistical relationship between the sampling errors of concurrent estimators and past sampling errors, termed by S&S as *secondary analysis*, or between observations corresponding to the same unit at different points in time, termed by the authors as *primary analysis*. It is this part of the model that is used under the classical survey sampling inference approach described above. The second part of the model considers the target population parameters like

the unknown population means as an unobservable time series, although as shown in Section 2, the model underlying this time series need not be specified explicitly other than the assumption of stationarity, possibly after appropriate transformation such as differencing.

Scott, Smith, and their colleagues (students) published their work during the seventies of the previous century and since then, the use of two-part time series models for the estimation of population parameters has developed in all kinds of ways and directions and is now routinely used by many statistical bureaus for the production and publication of their official statistics. In Section 2, I highlight the main contributions of S&S. In the remaining sections, I attempt to review some of the more recent developments emerging from their fundamental contributions. Due to space and my own time limitations, I was unable to review many other important developments and in particular, I do not review the many related studies under the Bayesian framework, an extremely important area on its own.

2. Scott and Smith contributions to estimation from repeated surveys

To introduce the idea, suppose that it is required to estimate the population mean, θ_t of a variable Y at month t. Denote by y_{ti} the value linked to unit i at time t.

Blight and Scott (1973) consider the following explicit two-part model:

$$y_{ti} - \theta_t = \rho(y_{(t-1),i} - \theta_{t-1}) + \eta_{ti};\ E(\eta_{ti}) = 0,\ Var(\eta_{ti}) = \sigma_\eta^2,\ Cov(\eta_{ti}, \eta_{\tau i}) = 0,\ t \neq \tau \text{ , (2.1)}$$

$$\theta_t - \mu = \lambda(\theta_{t-1} - \mu) + \varepsilon_t;\ E(\varepsilon_t) = 0,\ Var(\varepsilon_t) = \sigma_\varepsilon^2,\ Cov(\varepsilon_t, \varepsilon_\tau) = 0,\ t \neq \tau \text{ .} \qquad \text{(2.2)}$$

It is also assumed that $Cov(\varepsilon_t, \eta_{\tau i}) = 0$ for all t and τ.

Equation (2.1) defines the relationship between successive values associated with the same unit. A similar assumption is made in many of the studies under the classical sampling approach mentioned above, sometimes implicitly. Equation (2.2) defines a simple model for the evolution of the unknown population means over time, the new fundamental contribution to estimation from repeated surveys. Both parts of the model are autoregressive models of order 1 [AR (1)].

Remark 1. Blight and Scott (1973) note the somewhat anomalous position under the classical sampling theory of assuming a time series relationship between the individual

observations y_{ti} at different times as in Equation (2.1), but not assuming a time series relationship between the population means of these observations.

Blight and Scott (1973) consider the estimation of the population mean θ_t and the change $\delta_t = (\theta_t - \theta_{t-1})$ under the model (2.1)-(2.2), with added normality assumptions, based on all the data observed until and including time t. Assuming that all the model parameters are known, the authors derive efficient recursive algorithms for the computation of the estimators $(\hat{\theta}_t, \hat{\delta}_t)$ and their variances, distinguishing between matched observations with the previous sample, (units observed in both samples), matched observations with the next sample, and unmatched observations with both samples. Two other problems considered are the optimal matching proportion between two successive samples and the estimation of the unknown model parameters.

Remark 2. Nowadays, the optimal estimators, their variances and variance estimators could be derived by writing the model holding for the matched and unmatched means in *state-space* form and applying the Kalman filter (Kalman, 1960), and the prediction error decomposition for maximum likelihood parameter estimation. See Binder and Dick (1989) and Section 3. The authors propose the use of moment estimators, stating that "there are no simple expressions for full maximum-likelihood or Bayes estimates." Smith (1978) mentions the use of the Kalman filter as a way for estimating the unknown population means recursively.

S&S (1974) follow similar ideas, but recognizing that the individual observations are often unknown to the person analyzing the data, they model the relationship between the direct design-based estimators, denoted hereafter by $\{Y_t\}$, which are assumed to be unbiased for $\{\theta_t\}$ with respect to the randomization (design-based) distribution over all possible sample selections, such that $Y_t = \theta_t + e_t;\ E(e_t) = 0, Var(e_t) = s_t^2$; where $e_t = (Y_t - \theta_t)$ is the sampling error.

An interesting observation made by the authors is that without modelling the relationship between the true population means, “if the samples are non-overlapping, the estimate of θ_t reduces to Y_t and the previous estimates cannot improve this estimate. This implies wasting valuable information, considering that under normal

circumstances, the population means are expected to only change slowly over time." As illustrated below and throughout the paper, by modelling the relationship between the true (unknown) population means, this limitation is removed and efficient estimates of the current mean or changes in the means are obtained, employing all the direct estimates for all the times.

Specifically, denoting $\boldsymbol{Y}_{(t)} = (Y_1,...,Y_t)'$, for non-overlapping surveys

$$f(\theta_t \mid \boldsymbol{Y}_{(t)}) = f(\theta_t, Y_t \mid \boldsymbol{Y}_{(t-1)}) / f(Y_t \mid \boldsymbol{Y}_{(t-1)}) \propto f(Y_t \mid \theta_t) f(\theta_t \mid \boldsymbol{Y}_{(t-1)}) \qquad (2.3)$$

and under normality assumptions, the best (optimal) predictor of θ_t and its variance are shown to be,

$$\hat{\theta}_t = (1-\pi_t)Y_t + \pi_t \hat{Y}_t,\ Mse(\hat{\theta}_t) = E(\hat{\theta}_t - \theta_t)^2 = (1-\pi_t)s_t^2, \qquad (2.4)$$

where $\pi_t = s_t^2 / Var(Y_t \mid \boldsymbol{Y}_{(t-1)})$ and $\hat{Y}_t = E(\theta_t \mid \boldsymbol{Y}_{(t-1)}) = E(Y_t \mid \boldsymbol{Y}_{(t-1)})$. Notice that the variance s_t^2 of the direct estimator Y_t is reduced by the factor $(1-\pi_t)$ by use of the estimator $\hat{\theta}_t$ and the reduction would normally increase as the sample size decreases. S&S (1974) propose to derive the estimates $\hat{Y}_t$ and $Var(Y_t \mid \boldsymbol{Y}_{(t-1)})$ by fitting standard time series models to the series $\boldsymbol{Y}_{(t)}$, such as $ARIMA(p,d,q)$ models, which permits also predicting future means θ_{t+l} at time t, without specifying an explicit model for $\{\theta_t\}$. No standard survey sampling theory exists when the surveys are not overlapping but the population means are random.

Remark 3. S&S (1974) already advocate the use of time series models for improving the estimates in small subgroups, nowadays referred to as *small area estimation*. I return to this topic in Section 4.

The estimator (2.4) is derived under the assumption of non-overlapping samples but in practice, repeated surveys are often partially overlapping, inducing correlations between the sampling errors $e_t = (Y_t - \theta_t)$. S&S (1974) consider the case where $\{\theta_t\}$ follow an AR (1) model (possibly after differencing), and $\{e_t\}$ follows an $ARIMA(1,0,1)$ model, and derive the optimal estimators of θ_t and $\delta_t = (\theta_t - \theta_{t-1})$ under the model for known model parameters. Simple estimates for the parameters are proposed, defining the corresponding empirical best predictors.

Jones (1980) considered the case under which the estimates at time *t* consist of a vector of elementary estimates, $\boldsymbol{Y}_t = (Y_{t1},...Y_{tJ})'$, (estimates based on individuals joining and leaving the sample at the same time). Denoting $\tilde{\boldsymbol{Y}}_{(t)} = (\boldsymbol{Y}_1',...,\boldsymbol{Y}_t')'$ and $\boldsymbol{\theta}_{(t)} = (\theta_1,...,\theta_t)'$, the author considers the general model,

$$\tilde{\boldsymbol{Y}}_{(t)} = X_t\boldsymbol{\theta}_{(t)} + \tilde{\boldsymbol{e}}_{(t)};\ \tilde{\boldsymbol{e}}_{(t)} \sim N(\boldsymbol{0}, E_t)\,, \tag{2.5}$$

where X_t is a fixed matrix of 0's and 1's linking the estimates and the parameters, $\boldsymbol{\theta}_{(t)}$ is assumed to be multivariate normal with zero mean and variance matrix $\Sigma_{\boldsymbol{\theta}_{(t)}}$ and $\tilde{\boldsymbol{e}}_{(t)}$ is the vector of survey errors. Using conditional arguments, it follows that

$$E(\boldsymbol{\theta}_{(t)} \mid \tilde{\boldsymbol{Y}}_{(t)}) = (\mathrm{X}_t' E_t^{-1} X_t + \Sigma_{\theta_{(t)}}^{-1})^{-1} X_t E_t^{-1} \tilde{\boldsymbol{Y}}_{(t)};\ Var(\boldsymbol{\theta}_t \mid \tilde{\boldsymbol{Y}}_{(t)}) = (\mathrm{X}_t' E_t^{-1} X_t + \Sigma_{\theta_{(t)}}^{-1})^{-1}. \tag{2.6}$$

The result (2.6) is quite general and it produces the current and smoothed estimators (of past means) in one run. However, it requires new computations at every time *t*, with inversion of matrices of increasing dimensions. As stated earlier, this problem is solved by use of state-space models with recursive filters. See Section 3.

Remark 4. S&S (1974) comment that even with non-overlapping samples, it is difficult to update estimates of past values, θ_{t-k} as new direct estimates become available but as shown below, Scott Smith and Jones (1977, hereafter SSJ) handle this problem.

SSJ consider a general formulation by which the series $\{Y_t\}$ and $\{\theta_t\}$ are uncorrelated stationary processes. By fitting an ARIMA model to the series $\{Y_t\}$, optimal estimators of the means $\{\theta_{t+l}, l = 0, \pm 1, ...\}$ (the *signals*) and their mean-squared error (*Mse*) are derived by signal extraction techniques, as found in Whittle (1963). The respective theoretical formulae when the model parameters and the covariances $c_e(j) = Cov(e_t, e_{t-j})$ are known are quite complicated, and can be found in SSJ. (In practice, the unknown model parameters and the sampling error covariances need to be replaced by sample estimates.) Denoting by $\hat{\theta}_t(l)$ the optimal predictor of θ_{t+l} at time *t*, the predictor has the attractive expression,

$$\hat{\theta}_t(l) = \hat{Y}_t(l) - \hat{e}_t(l)\,, \tag{2.7}$$

where $\hat{Y}_t(l)$ is the predictor of Y_{t+l} at time *t* as obtained from the ARIMA model, and $\hat{e}_t(l)$ is the forecasted sampling error (the *noise*), obtained from the signal extraction. For non-overlapping surveys, $\hat{\theta}_t(0)$ reduces to the estimator (2.4).

Noting that for non-overlapping surveys $\hat{e}_t(l) = 0$ for $l > 0$, by (2.5) $\hat{\theta}_t(l) = \hat{Y}_t(l)$ and $Mse[\hat{\theta}_t(l)] = Mse[\hat{y}_t(l)] - c_e(0)$. Perhaps more interesting is the case where it is desired to update the previous estimate, $\hat{\theta}_{t-1}(0)$, as a new estimate, Y_t, becomes available. The updated (*smoothed*) estimate and its *Mse* are,

$$\hat{\theta}_{t-1}(-1) = \hat{\theta}_{t-1}(0) - a_1[Y_t - \hat{\theta}_t(0)];\ Mse[\hat{\theta}_{t-1}(-1)] = (1-\pi)c_e(0) - a_1^2\pi c_e(0)\,, \qquad (2.8)$$

where a_1 is the coefficient of Y_{t-1} when representing the model for $\{Y_t\}$ in the general form $\sum_{j=0}^{\infty} a_j Y_{t-j} = \varepsilon_t$ and $\pi = c_e(0)/Var(\varepsilon_t)$. Recall that any stationary and invertible model can be represented in this form. Comparing (2.8) with (2.4) shows the further reduction in the *Mse* from using the new estimate Y_t at time *t*.

Remark 5. All the results above are derived by fitting a time series model to the estimates $\{Y_t\}$, without specifying explicitly the model for the true means $\{\theta_t\}$. As stated in the next section, modern applications of time series models to repeated surveys employ *state-space models*, which require specifying the model holding for the means $\{\theta_t\}$, but are more flexible and allow for further inference possibilities.

Samples overlap if the units at any stage of the sampling process appear in more than one survey. For example, in two-stage sampling it may happen that some or all of the primary sampling units (clusters) are drawn by design in more than one survey, but new secondary (ultimate) units are sampled in each survey. There are many variations of such designs but in all the cases, the effect of overlap is that the sampling errors are correlated. Estimation of these correlations becomes complicated under a secondary analysis for which the individual observations are not available, requiring additional strong assumptions.

SSJ (1977) consider the case where the sampled units are retained in the sample for at most $q+1$ occasions, such that irrespective of the pattern of the sampling design, the sampling errors can be represented by a moving average model of order q [

$MA(q)$]. This case covers in particular the multi-stage sampling designs mentioned above, yielding the following current, forecasted and smoothed estimators:

$$\hat{\theta}_t(0) = Y_t - \frac{1}{\sigma^2}\sum_{j=0}^{q}\sum_{k=0}^{q-j} a_k c_e(j+k)\varepsilon_{t-j},$$

$$\hat{\theta}_t(1) = \hat{Y}_t(1) - \hat{e}_t(1) = \hat{Y}_t(1) - \frac{1}{\sigma^2}\sum_{j=0}^{q-1}\sum_{k=0}^{q-1-j} a_k c_e(j+k+1)\varepsilon_{t-j}, \qquad (2.9)$$

$$\hat{\theta}_t(-1) = Y_{t-1} - \hat{e}_t(-1) = \hat{\theta}_{t-1}(0) - \frac{1}{\sigma^2}[\sum_{j=0}^{q+1} a_j c_e(j-1)]\varepsilon_t,$$

where, using previous notation, $c_e(l) = Cov(e_t, e_{t-l})$ and the coefficients $\{a_j\}$ are defined by the model representation $\sum_{j=0}^{\infty} a_j Y_{t-j} = \varepsilon_t$ with $\sigma^2 = Var(\varepsilon_t)$. Recall that the residuals ε_{t-j}, $j = 0,1,...$ are the one-step ahead prediction errors under the model fitted tor the observed series. For $q = 0$, the estimates (2.9) reduce to the estimates presented for non-overlapping surveys. SSJ (1977) discuss possible ways of estimating the covariances $c_e(l)$, depending on data availability.

Remark 6. The theoretical results derived in the article are illustrated empirically with detail, by applications to real sample data.

3. New developments following Scott and Smith contributions

Since the pioneering work of S&S described in Section 2, the use of time series models for finite population estimation became a common routine in many statistical bureaus throughout the world, mostly for estimation in small areas for which the sample sizes are not sufficient to base the inference on classical survey sampling theory. In this section, I describe some of the main developments following S&S, focusing on the models used and estimation procedures applied. Due to space limitations, I shall not elongate much on technical details, which can be found in the references provided.

3.1 Summary and extensions of SSJ results

I start with results of Bell and Hillmer (1987,1989, 1990, hereafter B&H). Denote by $(\boldsymbol{Y}_{(N)}, \boldsymbol{\theta}_{(N)}, \boldsymbol{e}_{(N)})$ the vectors of the observed estimates, the population means and the sampling errors for time $t = 1,...,N$, such that $\boldsymbol{Y}_{(N)} = \boldsymbol{\theta}_{(N)} + \boldsymbol{e}_{(N)}$. Assuming $E(\boldsymbol{e}_{(N)}) = \boldsymbol{0}$ and that $\boldsymbol{\theta}_{(N)}$ and $\boldsymbol{e}_{(N)}$ are stationary and independent,

$$E(\boldsymbol{Y}_{(N)}) = E(\boldsymbol{\theta}_{(N)}) = \boldsymbol{\mu}_{(N)} = (\mu_1,...,\mu_N)'; \Sigma_{\boldsymbol{Y}_{(N)}} = \Sigma_{\boldsymbol{\theta}_{(N)}} + E_N . \qquad (3.1)$$

Under (3.1), the minimum *Mse* linear predictor of $\theta_{(N)}$ and its variance matrix are,

$$\hat{\boldsymbol{\theta}}_{(N)} = \boldsymbol{\mu}_{(N)} + \Sigma_{\boldsymbol{\theta}_{(N)}} \Sigma^{-1}_{\boldsymbol{Y}_{(N)}} (\boldsymbol{Y}_{(N)} - \boldsymbol{\mu}_{(N)}); \; Var(\hat{\boldsymbol{\theta}}_{(N)} - \boldsymbol{\theta}_{(N)}) = E_{(N)} - E_{(N)} \Sigma^{-1}_{\boldsymbol{Y}_{(N)}} E_{(N)} . \qquad (3.2)$$

With added normality assumptions, (3.2) defines the conditional mean and variance matrix of $\hat{\boldsymbol{\theta}}_{(\mathrm{T})} \,|\, \boldsymbol{Y}_{(T)}$ and it is easily shown to coincide with the estimator (2.6) derived by Jones (1980) in the special case, $\boldsymbol{\mu}_{(N)} = \mathbf{0}$, $J = 1$, $X_t = I$ and $t = N$, by employing the formula for the inverse of the sum of two matrices. B&H extend the results to the case where the series $\boldsymbol{Y}_{(T)}$ requires differencing to achieve stationarity and show the optimality of the estimators obtained in this case. Similar results have been derived by Jones (1980) , but without the optimality properties. B&H (1990) establish some other properties of the optimal results in (3.2); in particular, that the estimator $\hat{\boldsymbol{\theta}}_{(N)}$ is consistent under the joint distribution of $\boldsymbol{Y}_{(N)}$ and $\hat{\boldsymbol{\theta}}_{(N)}$, but that it is biased under the randomization distribution, when $\hat{\boldsymbol{\theta}}_{(N)}$ is assumed to be constant.

Remark 7. Following S&S (1974) and SSJ (1977), B&H obtain the expressions in (3.2) by use of signal extraction, without specifying a time-series model for $\hat{\boldsymbol{\theta}}_{(N)}$. B&H (1989) discuss the use of models for $\hat{\boldsymbol{\theta}}_{(N)}$ and $\boldsymbol{e}_{(N)}$, putting them in state-space form and using the Kalman filter and smoother as an efficient way to obtain the predictors and their variances. See Section 3.2 for details. Pfeffermann and Tiller (2006) show that a model for $\boldsymbol{e}_{(N)}$ is not required and it is sufficient to account for the correlations of the sampling errors.

Remark 8. Equation (3.2) assumes known $\boldsymbol{\mu}_{(N)}$ and variance matrices. In practice, any unknown parameter needs to be estimated from the available data, in which case the variances of the prediction errors when substituting the model parameters in (3.2) by their sample estimators ignore the contribution to the variance of the prediction errors from the use of the estimates, thus underestimating the true variances. Pfeffermann and Tiller (2005) established a bootstrap procedure, which yields the variance of the prediction errors to correct order when using estimated parameters. See also Bollineni-Balabay et al. (2017).

3.2 *State-space models*

S&S fit ARIMA models to the survey estimates and obtain the estimates of the underlying population means by signal extraction. The big advantage of this approach is that the model holding for the means need not be specified. However, as discussed and illustrated by B&H (1987, 1989, 1990), simple models for the observed series cannot reflect non-stationarities in the sampling errors, such as sampling variances that change over time. Also, the use of a simple time series model for the observed series may imply an unreasonable model for the signal. In theory (although in extreme cases), the implied model for the signal could violate the requirement that the corresponding spectrum is non-negative at all frequencies. Signal extraction is often complicated and requires repeating the whole estimation process every time that new observations become available.

For these reasons, an alternative approach adopted by several statistical agencies, (the Bureau of Labor Statistics- BLS in the U.S., Statistics Netherland, the CBS in Israel, and currently experimented in other countries), is to specify a model for the observed data given the population means, and a model for the population means, combine the two models in state-space form and then apply the Kalman filter (Kalman, 1960) and a smoothing algorithm to obtain the required estimators and their *Mse*s. As detailed below, one of the main advantages of this approach is that it enables computing the desired estimators and their *Mse*s recursively, without having to refit the model every time that new data become available. In what follows I describe briefly the main steps of this procedure. The book by Harvey (1989) is an excellent reference for this modelling approach.

The basic (linear) state-space model is defined as follows:

***Observation (measurement) equation*,**

$$\mathbf{Y}_t = Z_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t;\ E(\boldsymbol{\varepsilon}_t) = \mathbf{0},\ E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \Sigma,\ E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t*}') = 0 \text{ for } t \neq t^*, \tag{3.3}$$

***Transition (state) equation*,**

$$\boldsymbol{\beta}_t = T\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t;\ E(\boldsymbol{\eta}_t) = \mathbf{0},\ E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = Q,\ E(\boldsymbol{\eta}_t\boldsymbol{\eta}_{t*}') = 0 \text{ for } t \neq t^*. \tag{3.4}$$

It is also assumed that $E(\boldsymbol{\varepsilon}_t\boldsymbol{\eta}_{t^*}') = 0$ for all (t,t*). In this formulation, $\mathbf{Y}_t, \boldsymbol{\beta}_t, \boldsymbol{\varepsilon}_t, \boldsymbol{\eta}_t$ are vectors, Z_t, T are "design" matrices, which may contain unknown parameters and $\mathbf{0}\,(0)$ is the null vector (matrix) of appropriate order. (The matrices Σ, T, Q can be

time-dependent but for convenience, I assume that they are fixed over time.) As noted before, the model defined by (2.1)-(2.2) is a simple special case of the model in (3.3)-(3.4). See Binder and Dick (1989) for an appropriate formulation. Notice that for $T = I, Q = 0$, $\boldsymbol{\beta}_t = \boldsymbol{\beta}$, the model (3.3)-(3.4) reduces to a standard regression model.

The Kalman filter is a recursive algorithm which updates the best linear unbiased predictor (BLUP) of the state vector $\boldsymbol{\beta}_t$ at time $(t-1)$ when new data $\boldsymbol{Y}_t$ become available. Denote by $\hat{\boldsymbol{\beta}}_{t-1}$ the BLUP of $\boldsymbol{\beta}_{t-1}$, based on the observations $\boldsymbol{Y}_{(t-1)}$, and by P_{t-1} the corresponding prediction variance matrix; $P_{t-1} = E[(\hat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1})(\hat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1})']$. Then, under the model the predictor of $\boldsymbol{\beta}_t$ at time $(t-1)$ is, $\hat{\boldsymbol{\beta}}_{t|t-1} = T\hat{\boldsymbol{\beta}}_{t-1}$ and $Var(\hat{\boldsymbol{\beta}}_{t|t-1}) = TP_{t-1}T' + Q = P_{t|t-1}$.

When the new data $\boldsymbol{Y}_t$ become available, the updated BLUP of $\boldsymbol{\beta}_t$ and its *Mse* are,

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t|t-1} + P_{t|t-1}Z_t'F_t^{-1}(\boldsymbol{Y}_t - Z_t\hat{\boldsymbol{\beta}}_{t|t-1});\ P_t = P_{t|t-1} - P_{t|t-1}Z_t'F_t^{-1}Z_tP_{t|t-1} \le P_{t|t-1}, \tag{3.5}$$

where $F_t = (Z_tP_{t|t-1}Z_t' + \Sigma) = Var(\hat{Y}_{t|t-1} - Y_t)$.

Past estimators of the state vectors can be updated by the following recursive smoothing algorithm, where we denote by $\hat{\boldsymbol{\beta}}_{t|N}$ the smoothed estimate of $\boldsymbol{\beta}_t$ based on all the data $\boldsymbol{Y}_{(N)}$, $t \le N$, and by $P_{t|N}$ the corresponding prediction error variance matrix:

$$\hat{\boldsymbol{\beta}}_{t-1|t} = \hat{\boldsymbol{\beta}}_{t-1} + P_{t-1}^*(\hat{\boldsymbol{\beta}}_t - T\hat{\boldsymbol{\beta}}_{t-1}),\ P_{t-1|t} = P_{t-1} - P_{t-1}^*(P_{t|t-1} - P_t)P_{t-1}^{*\prime} \le P_{t-1}, \tag{3.6}$$

$$P_{t-1}^* = P_{t-1}T'P_{t|t-1}^{-1};$$

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{t-1|N} = \hat{\boldsymbol{\beta}}_{t-1} + P_{t-1}^*(\hat{\boldsymbol{\beta}}_{t|N} - T\hat{\boldsymbol{\beta}}_{t-1}),\\
&P_{t-1|N} = E[(\hat{\boldsymbol{\beta}}_{t-1|N} - \boldsymbol{\beta}_{t-1})(\hat{\boldsymbol{\beta}}_{t-1|N} - \boldsymbol{\beta}_{t-1})'] = P_{t-1} - P_{t-1}^*(P_{t|t-1} - P_{t|N})P_{t-1}^{*\prime} \le P_{t-1}.
\end{aligned} \tag{3.7}$$

The algorithm starts with $\hat{\boldsymbol{\beta}}_{N|N} = \hat{\boldsymbol{\beta}}_N$, $P_{N|N} = P_N$.

The algorithms described so far assume known model parameters. In practice, they are replaced by sample estimates, yielding what is known as the empirical best linear predictors (EBLUP). See also Remark 8 above. Possible ways of estimating the unknown model parameters and the variances of the resulting prediction errors, which account for the variability of the parameter estimators, are mentioned later.

3.3. *Applications to estimation from repeated surveys*

I start by defining what is known as the Basic Structural Model (BSM), which when combined with a model for the sampling errors, forms the basis for many applications of analysing repeated survey data. See Harvey (1989) for the theoretical properties of the BSM. For convenience, I assume a monthly series.

$$
\begin{aligned}
&Y_t = L_t + S_t + \varepsilon_t \,;\; \varepsilon_t \sim N(0,\sigma_\varepsilon^2) \\
&L_t = L_{t-1} + R_{t-1} + \eta_{Lt} \;,\; \eta_{Lt} \sim N(0,\sigma_L^2) \;;\; R_t = R_{t-1} + \eta_{Rt} \;,\; \eta_{Rt} \sim N(0,\sigma_R^2), \\
&S_t = \sum\nolimits_{j=1}^{6} S_{j,t} \,; \qquad\qquad (3.8) \\
&S_{j,t} = \cos\omega_j S_{j,t-1} + \sin\omega_j S^*_{j,t-1} + \nu_{j,t} \;,\quad \nu_{j,t} \sim N(0,\sigma_S^2) \\
&S^*_{j,t} = -\sin\omega_j S_{j,t-1} + \cos\omega_j S^*_{j,t-1} + \nu^*_{j,t} \;,\quad \nu^*_{j,t} \sim N(0,\sigma_S^2) \;;\; \omega_j = 2\pi j/12;\; j = 1,...,6 \,.
\end{aligned}
$$

In this model, Y_t is the (univariate) direct estimate at time t, L_t is the trend level, R_t is the slope of the trend, S_t is the seasonal effect and ε_t is the irregular term. The disturbances $\varepsilon_t, \eta_{L_t}, \eta_{R_t}, \nu_{j,t}, \nu^*_{j,t}$ are independent white noise series. The model for the trend approximates a local linear trend, while the model for the seasonal effects uses the traditional decomposition of the seasonal component into 6 cyclical components corresponding to the seasonal frequencies. The added noise term permits the seasonal effects to evolve stochastically over time, but imposing that the expectation of the sum of 12 successive seasonal effects is 0. The BSM is easily formulated in state-space form, with Z_t being a row vector and ε_t a scalar.

Remark 9. The BSM can be extended to allow also for the effect of moving festivals and trading days effects. See, e.g., Morris and Pfeffermann (1984) and Bell and Hillmer (1990).

The BSM does not account for autocorrelations between the sampling errors of the estimators Y_t in the case of repeated surveys. For this, the model has to be extended by adding to the observation equation (or the state equations, see later) an additional component e_t, representing the sampling error. In what follows, I describe several extensions of the BSM that account for the sampling errors of the survey estimates and their correlation structure. I refer to such models as extended BSM (EBSM).

Pfeffermann (1991) fitted an EBSM to estimates obtained from the Labor Force Survey in Israel (ILFS). At that time, the ILFS was a quarterly survey consisting of 4 panels,

with 3 of them included in previous surveys and one panel surveyed for the first time. Specifically, every new panel was included in the sample for two successive quarters, left out of the survey for the next two quarters and then included again for two more quarters. This rotation pattern produced 50% sample overlaps between two successive quarters and between the same quarter in two successive years. Denote by y_t^{t-j} the survey estimate of the population mean at time *t*, based on the panel joining the sample for the first time at time $t-j, j=0,1,4,5$. The separate panel estimates at any given time are independent by design, but correlated over time. Following Scott and Blight (1973), Pfeffermann assumed an AR(1) model for the individual observations y_{ti} (Eq., 2.1) with mean (signal) $\theta_t = L_t + S_t$, implying the following model for the sampling errors of the panel estimators:

$$e_t^{t-1} = \rho e_{t-1}^{t-1} + \nu_t^{t-1}, e_t^{t-4} = \rho^3 e_{t-3}^{t-4} + \nu_t^{t-4},\ e_t^{t-5} = \rho e_{t-1}^{t-5} + \nu_t^{t-5}, Corr(e_t^t, e_t^{t-k}) = 0,\ k > 0\ . \quad (3.9)$$

Denoting $\mathbf{Y}_t = (y_t^t, y_t^{t-1}, y_t^{t-4}, y_t^{t-5})'$, the model holding for $\mathbf{Y}_t$ and the corresponding sampling errors is easily formulated in the state-space form (3.3)-(3.4), with the models holding for L_t and S_t as under the BSM (see Remark 10 below), and the model holding for the sampling errors defined by (3.9). Notice that the sampling errors are part of the state vector. See Section (3.4) for an alternative approach.

Remark 10. Pfeffermann (1991) actually assumed that the seasonal effects evolve according to the model $\sum_{k=0}^{3} S_{t-k} = \eta_{S_t}$, instead of the trigonometric model in (3.8).

Pfeffermann (1991) extended the model considered so far by accounting for what is known as *rotation group bias* (RGB), under which sampled units tend to provide different information on different rounds of interview, because of “interview fatigue”, the use of different modes of data collection in different rounds of interview (for example, face to face interview Vs. telephone interview, which is common in LFS), etc. The bias of the panel estimates may also result from different patterns of nonresponse across the panels. To account for the possible bias of the separate panel estimates, Pfeffermann (1991) added to the observation equation a constant term γ_j for the panel joining the sample for the first time at time $(t-j)$, such that $E(y_t^{t-j}) = L_t + S_t + \gamma_j;\ j = 0,1,4,5$.

By assuming normality of the disturbance terms in the model, the unknown model variances, the RGB constants and the AR(1) coefficients have been estimated by maximum likelihood (MLE), with the log likelihood obtained from the prediction error decomposition. Denoting by φ the set of unknown parameters and using the notation in Section (3.2), the log-likelihood is,

$$l(\varphi) = const - \frac{1}{2}\sum_{t=1}^{N}[\log|F_t| + (\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})' F_t^{-1} (\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})]. \qquad (3.10)$$

Remark 11. Statistical bureaus all over the world produce routinely seasonally adjusted and trend estimates of all their core time series, which constitute a major part of the published official statistics. The BSM and its extension (EBSM), yield such estimates, although as far as I can tell, this family of models is not routinely used for this purpose. The procedure in common use is X13ARIMA. See the reference list. Maravall (1985) compares the BSM with the conventional filters used by the X-11 seasonal adjustment procedure, preceding but forming the basis for X13ARIMA. See also Pfeffermann et al. (1998) for empirical comparisons.

Brakel and Krieg (2009) fitted a similar model to Pfeffermann (1991) to the LFS survey estimates in the Netherlands. The Dutch LFS is also based on a rotating panel design by which a new sample enters the survey every month. The new sample is observed 5 times with an interval of 3 months between successive observations, and then it is replaced by a new sample. Thus, using previous notation with *t* defining months, the observed series at time *t* consists in this case of the vector $\boldsymbol{Y}_t = (y_t^t, y_t^{t-3}, y_t^{t-6}, y_t^{t-9}, y_t^{t-12})'$. The model accounts for RGB by assuming that there is no bias in the sample observed for the first time [$E(y_t^t) = L_t + S_t$], and modelling the other RGBs as random walk. The model accounts also for “shocks” in the series (e.g., change of the sampling design), by adding a level shift; a dummy variable taking the value 1 when the shock starts.

Remark 12. The model is used by Statistics Netherlands since 2010 for the production of the official monthly Labor Force figures at the national level, and for six domains defined by age and gender.

Brakel and Krieg (2016) combined the separate models for the six domains into a single thirty-dimensional model, thus accounting also for the cross-sectional

correlations between the trend disturbance terms of the domain models. This way, the model *borrows strength over time and space*. Another extension considered by the authors, proposed originally by Harvey and Chung (2000), is to add to the observation equation the series of "claimant counts" (people claiming unemployment benefits), as auxiliary variables "explaining" the variation in the labor force estimates. Notice that by adding auxiliary variables to the observation equation of the EBSM, the seasonal and trend components in the model account for the "residual" trend and seasonal effects, not explained by the auxiliary variables.

3.4. *Filtering algorithm for state-space models with correlated sampling errors*

As mentioned before, a common practice to account for the correlation structure of the sampling errors in overlapping surveys is to assume an ARMA model for them, and add the model to the state equations of the state-space model. This paradigm has been used in the studies reviewed so far. However, the ARMA model may include many terms and when modelling jointly many time series, as for example for *benchmarking* (Section 4), the resulting state vector is of very high dimension and application of the Kalman filter and smoother runs into all kinds of problems, even with modern high power computers. To deal with this problem, Pfeffermann and Tiller (2006) developed a filtering algorithm for state-space models with autocorrelated measurement errors in the observation equation, which does not require modelling the sampling errors. The filter coincides with the Kalman filter when the measurement errors are uncorrelated over time.

Consider the following multivariate state-space model,

$$\boldsymbol{Y}_t = Z_t\boldsymbol{\beta}_t + \boldsymbol{e}_t\,;\;\; E(\boldsymbol{e}_t) = \boldsymbol{0},\;\; E(\boldsymbol{e}_\tau \boldsymbol{e}_t') = \Sigma_{\tau t} = Diag[\sigma_{1,\tau t},...,\sigma_{D,\tau t}]\,, \tag{3.11}$$

$$\boldsymbol{\beta}_t = T\boldsymbol{\beta}_{t\text{-}1} + \boldsymbol{\eta}_t\,;\; E(\boldsymbol{\eta}_t) = \boldsymbol{0},\; E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = Q,\; E(\boldsymbol{\eta}_t\boldsymbol{\eta}_{t^*}') = 0 \text{ for } t \neq t^*. \tag{3.12}$$

In this model, $\boldsymbol{Y}_t = (Y_{1t},...,Y_{Dt})'$ is a vector of independent survey estimators measured for $D \geq 1$ series at time *t*, but each series d contains autocorrelated sampling errors e_{dt}. Accordingly, $\boldsymbol{e}_t = (e_{1t},...,e_{Dt})'$, $\boldsymbol{\eta}_t = (\eta_{1t}',...,\eta_{Dt}')'$, $Z_t = I_D \oplus \boldsymbol{z}_{\boldsymbol{dt}}'$, a block diagonal matrix with $\boldsymbol{z}_{dt}'$ as the dth block, $Q = I_D \oplus Q_t$ and $T = [T_1',...,T_D']'$. (I_D defines the identity matrix of order D). It is assumed that all the state vectors $\boldsymbol{\beta}_{(t)}$ are of fixed dimension q, which defines the dimension of all the other vectors and matrices above.

Using previous notation, let $\hat{\boldsymbol{\beta}}_{t|t-1} = T\hat{\boldsymbol{\beta}}_{t-1}$ define the predictor of $\boldsymbol{\beta}_t$ at time $(t-1)$, with variance matrix $P_{t|t-1} = TP_{t-1}T' + Q$. Consider the following generalized regression model (GLS) with random coefficients,

$$\begin{pmatrix} T\hat{\boldsymbol{\beta}}_{t-1} \\ \boldsymbol{Y}_t \end{pmatrix} = \begin{pmatrix} I_q \\ Z_t \end{pmatrix}\boldsymbol{\beta}_t + \begin{pmatrix} \boldsymbol{u}_{t|t-1} \\ \boldsymbol{e}_t \end{pmatrix};\ \boldsymbol{u}_{t|t-1} = T\hat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_t, \tag{3.13}$$

and define $V_t = Var\begin{pmatrix} \boldsymbol{u}_{t|t-1} \\ \boldsymbol{e}_t \end{pmatrix} = \begin{bmatrix} P_{t|t-1} & , & C_t \\ C_t' & , & \Sigma_t \end{bmatrix}$. The covariance matrix $C_t = Cov[\boldsymbol{u}_{t|t-1}, \boldsymbol{e}_t]$ is computed as follows: Let $[I_q, Z_j']V_j^{-1} = [B_{j1}, B_{j2}]$, where B_{j1} contains the first q columns and B_{j2} the remaining columns. Define $A_j = TP_jB_{j1}$, $\tilde{A}_j = TP_jB_{j2};\ j = 2,\ldots,(t-1)$ and $\tilde{A}_1 = TK_1$, where $K_1 = P_{1|0}Z_1'F_1^{-1}$ is the 'Kalman gain' with $P_{1|0} = TP_0T' + Q$ and $F_1 = Z_1P_{1|0}Z_1' + \Sigma_1$. Then,

$$C_t = A_{t-1}A_{t-2}\times\ldots\times A_2\tilde{A}_1\Sigma_{1t} + A_{t-1}A_{t-2}\times\ldots\times A_3\tilde{A}_2\Sigma_{2t} + \ldots + A_{t-1}\tilde{A}_{t-2}\Sigma_{t-2,t} + \tilde{A}_{t-1}\Sigma_{t-1,t}. \tag{3.14}$$

The (GLS) predictor of $\boldsymbol{\beta}_t$ and the respective prediction error variance matrix are,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_t &= \left[(I_q, Z_t')V_t^{-1}\begin{pmatrix} I_q \\ Z_t \end{pmatrix}\right]^{-1}(I_q, Z_t')V_t^{-1}\begin{pmatrix} T\hat{\boldsymbol{\beta}}_{t-1} \\ \boldsymbol{Y}_t \end{pmatrix} \\ P_t &= E[(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t)(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t)'] = \left[(I_q, Z_t')V_t^{-1}\begin{pmatrix} I_q \\ Z_t \end{pmatrix}\right]^{-1}. \end{aligned} \tag{3.15}$$

Pfeffermann and Tiller (2006) show that the predictor $\hat{\boldsymbol{\beta}}_t$ is the BLUP of $\boldsymbol{\beta}_t$ based on $T\hat{\boldsymbol{\beta}}_{t-1}$ and $\boldsymbol{Y}_t$. It is not the BLUP based on all the observations $\boldsymbol{Y}_{(t)} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_t')'$, but simulation results show that the loss from not including all the observations for the prediction of $\boldsymbol{\beta}_t$ (which is not practical in a production environment), is very mild. As stated before, when the sampling errors are uncorrelated, the GLS predictor (3.15) coincides with the optimal, Kalman filter predictor.

4. Benchmarking in Small Area Estimation

4.1. *Introduction*

The use of time series models for finite population estimation from repeated surveys becomes essential in small area estimation (SAE), because the sample sizes in at least some of the areas are usually too small to allow using design-based estimators.

Furthermore, in many real applications, no samples are available for many of the areas and there exists no design-based theory for estimation in areas with no samples.

On the other hand, at the national level or for large areas, the sample sizes are often sufficiently large and statistical bureaus tend to use design-based estimators for the large areas, thus avoiding the use of models. This, however, may result in "publication inconsistency", in the sense that the model-based estimators in the small areas do not conform to the design-based estimator in the large area to which they belong. For example, the sum of the model-based estimators in the small areas may differ from the design-based estimator in the large area. While this may indicate model misspecification or breakdown, in practice it is often the case that the model cannot be modified easily in a production environment, and it may take many time points before the change in the model can be detected and accounted for properly.

To deal this problem, it is common practice to benchmark the model-based small area predictors, such that they conform to the corresponding design-based estimator in the large area. Denote by $\hat{\theta}_{dt,m}$ the model dependent predictor at time *t* in area *d*, belonging to a large group *L* of *D* areas for which the design-based estimator is sufficiently accurate, and by Y_{dt} the corresponding design-based estimator. Benchmarking means imposing in every time *t* the constraint,

$$\sum_{\mathrm{d}=1}^{D} b_{dt}\hat{\theta}_{dt,m} = \sum_{\mathrm{d}=1}^{D} b_{dt}Y_{dt};\ t = 1, 2\ldots, \tag{4.1}$$

with constant weights $\{b_{dt}\}$. For example, $b_{dt} = 1$ when estimating totals, $b_{dt} = N_{dt} / \sum_{d=1}^{D} N_{dt}$ when estimating means or proportions, where N_{dt} is the size of the target population in area *d*.

Other than guaranteeing publication consistency, the use of benchmarking has two other important properties. First, if all the design-based estimates in the group *L* jointly increase or decrease due to some external effects that are not accounted for by the model, the benchmarked estimators will reflect this change much faster than the model dependent estimators. This happens because time series models adapt to changes in the behaviour of the observed series much slower. This property is illustrated very strikingly in Pfeffermann and Tiller (2006), using data from the U.S Current Population Survey (CPS). Second, by incorporating the constraints, the benchmarked estimators

borrow information from both past data and cross-sectionally, unlike the model dependent estimators, which only borrow information from past data.

Remark 13. A notable feature of the benchmark equations (4.1) is that the model-dependent estimates are benchmarked to a weighted mean of the design-based estimates, which are the input data for the model, known as *internal benchmarking*. This is different from *external benchmarking*, under which the model-dependent estimates are benchmarked to external (independent) data sources. Benchmarks of this kind are not frequently available, and in what follows I restrict to internal benchmarking. See, e.g., Hillmer and Trabelsi (1987) and Durbin and Quenneville (1997) for external benchmarking in state-space modelling.

Remark 14. External and internal benchmarking have been investigated extensively in cross-sectional SAE under the frequentist and Bayesian paradigms. See, e.g., Bell et al. (2012) and Pfeffermann (2013). Pfeffermann et al. (2014) compare cross-sectional and time series benchmarking procedures using simulated series.

A simple way in common use of enforcing the constraint (4.1), known as ratio or pro-rata benchmarking is,

$$\hat{\theta}_{dt}^{bmk} = \hat{\theta}_{dt,m} \sum_{\mathrm{d}=1}^{D} b_{dt} Y_{dt} \,/\, \sum_{\mathrm{d}=1}^{D} b_{dt} \hat{\theta}_{dt,m}; \; t = 1, 2.... \tag{4.2}$$

However, the use of (4.2) has three important limitations. (i)- it multiplies all the model-based predictors by the same ratio, irrespective of their precision, (ii)- the benchmarked estimator in a given area does not converge to the true population value when only increasing the sample size in that area, (iii)- the use of (4.2) does not lend itself to simple variance estimation and hence, correct filtering in state-space models.

4.2 Internal b*enchmarking in state-space models with autocorrelated measurement errors*

Pfeffermann and Burck (1990) developed a benchmarking procedure for SAE in the context of a state-space model, but assumed uncorrelated sampling errors. Pfeffermann and Tiller (2006, hereafter PT) extended their work to account for correlated sampling errors, with reference to the U.S. Current Population Survey (CPS, the U.S. LFS), and in what follows I review this procedure.

The CPS is a monthly survey of households (HH) with a rather complicated rotation pattern under which every sampled HH is retained in the survey for 4 successive

months, it is dropped from the survey in the ensuing 8 months and then it is included again in the following 4 successive months. This rotation pattern induces highly correlated sampling errors. The same sampling design is used in Brazil and Israel.

Tiller (1982) fitted the BSM (3.8) separately for each of the 51 States of the U.S, but with the measurement errors replaced by sampling errors which have been modelled by an AR(15) model and added to the state equations. The resulting state vector consists of 29 elements. Benchmarking the monthly estimates within the filtering and smoothing algorithms, requires fitting the State models jointly, imposing each month the benchmark constraint. This implies that even with only 10 States, the state vector would consist of 290 elements, with the prediction variance matrix P_t which needs to be inverted as part of the filter (see Section 3.2), being of dimension 290.

To deal with this problem, PT combined the state-space models holding for the areas under consideration into a single joint model, such that $\boldsymbol{Y}_t = (Y_{1t},...,Y_{Dt})'$ is the vector of the survey estimates of the area means $\boldsymbol{\theta}_t = (\theta_{1t},...,\theta_{Dt})'$ at time *t*, and added the constraints (4.1) to the measurement equations of the combined model.

Let $\boldsymbol{\beta}_{dt}$ be the sub-vector of $\boldsymbol{\beta}_t$ in the model (3.11)-(3.12) corresponding to area d, such that $\boldsymbol{\beta}_t = (\boldsymbol{\beta}'_{1t},...,\boldsymbol{\beta}'_{Dt})'$. With this notation, the benchmark constraints (4.1) are,

$$\sum_{d=1}^{D} b_{dt}\boldsymbol{z}'_{dt}\hat{\boldsymbol{\beta}}_{dt}^{bmk} = \sum_{d=1}^{D} b_{dt}Y_{dt},\ t = 1,2,... \qquad (4.3)$$

However, under the model, $\sum_{d=1}^{D} b_{dt}Y_{dt} = \sum_{d=1}^{D} b_{dt}\boldsymbol{z}'_{dt}\boldsymbol{\beta}_{dt} + \sum_{d=1}^{D} b_{dt}e_{dt}$. In order to impose the benchmarks, PT added the last equation to the measurement equations in (3.11), such that the new measurement equations are,

$$\tilde{\boldsymbol{Y}}_t = \tilde{Z}_t\boldsymbol{\beta}_t + \tilde{\boldsymbol{e}}_t;\ \tilde{\boldsymbol{Y}}_t = \left(\boldsymbol{Y}'_t, \sum_{d=1}^{D} b_{dt}Y_{dt}\right)',\ \tilde{Z}_t = \begin{bmatrix} Z_t \\ b_{1t}\boldsymbol{z}'_{1t},\ldots,b_{Dt}\boldsymbol{z}'_{Dt} \end{bmatrix},\ \tilde{\boldsymbol{e}}_t = \left(\boldsymbol{e}'_{t,} \sum_{d=1}^{D} b_{dt}e_{dt}\right)'. \quad (4.4)$$

The benchmarked predictors are computed as follow: First, use the random coefficients regression model representation (3.13) for the augmented measurement equations, using $\hat{\boldsymbol{\beta}}_{t|t-1}^{bmk} = T\hat{\boldsymbol{\beta}}_{t-1}^{bmk}$ as the state vector predictor;

$$\begin{pmatrix} T\hat{\boldsymbol{\beta}}_{t\text{-}1}^{bmk} \\ \tilde{\boldsymbol{Y}}_t \end{pmatrix} = \begin{pmatrix} I_q \\ \tilde{Z}_t \end{pmatrix}\boldsymbol{\beta}_t + \begin{pmatrix} \boldsymbol{u}_{t|t\text{-}1}^{bmk} \\ \tilde{\boldsymbol{e}}_t \end{pmatrix};\ \boldsymbol{u}_{t|t\text{-}1}^{bmk} = (T\hat{\boldsymbol{\beta}}_{t\text{-}1}^{bmk} - \boldsymbol{\beta}_t),\ Var\begin{pmatrix} \boldsymbol{u}_{t|t\text{-}1}^{bmk} \\ \tilde{\boldsymbol{e}}_t \end{pmatrix} = \begin{bmatrix} P_{t|t-1}^{bmk} & C_t^{bmk} \\ C_t^{bmk\prime} & \tilde{\Sigma}_{tt} \end{bmatrix} = \tilde{V}_t \quad (4.5)$$

$P_{t|t-1}^{bmk} = E[(T\hat{\boldsymbol{\beta}}_{t-1}^{bmk} - \boldsymbol{\beta}_t)(T\hat{\boldsymbol{\beta}}_{t-1}^{bmk} - \boldsymbol{\beta}_t)'] = TP_{t-1}^{bmk}T' + Q$;

$P_{t-1}^{bmk} = E[(\hat{\boldsymbol{\beta}}_{t-1}^{bmk} - \boldsymbol{\beta}_{t-1})(\hat{\boldsymbol{\beta}}_{t-1}^{bmk} - \boldsymbol{\beta}_{t-1})']$, $C_t^{bmk} = E(\boldsymbol{u}_{t|t-1}^{bmk}, \tilde{\boldsymbol{e}}_t')$, $\tilde{\Sigma}_{\boldsymbol{tt}} = E(\tilde{\boldsymbol{e}}_t\tilde{\boldsymbol{e}}_t') = \begin{bmatrix} \Sigma_{tt} & \mathbf{h}_{tt} \\ \mathbf{h}_{tt}' & v_{tt} \end{bmatrix}$,

$v_{tt} = Var(\sum_{d=1}^{D} b_{dt}e_{dt}) = \sum_{d=1}^{D} b_{dt}^2 Var(e_{dt})$ and $\mathbf{h}_{tt} = Cov(\boldsymbol{e}_t, \sum_{d=1}^{D} b_{dt}e_{dt})$

$= [b_{1t}Var(e_{1t}),...,b_{Dt}Var(e_{Dt})]'$.

Second, compute the benchmarked predictor for time t by imposing $\sum_{d=1}^{D} b_{dt}e_{dt} = 0$ ($\sum_{d=1}^{D} b_{dt}y_{dt} = \sum_{d=1}^{D} b_{dt}\boldsymbol{z}_{dt}'\boldsymbol{\beta}_{dt}$) in the last equation of (4.4). This is done by enforcing $Var(\sum_{d=1}^{D} b_{dt}e_{dt}) = 0$, $Cov(\sum_{d=1}^{D} b_{dt}e_{dt}, e_{rt}) = 0$ for $r = 1,...,D$. Defining $\tilde{\boldsymbol{e}}_{t,0} = (\boldsymbol{e}_t', 0)'$, $\tilde{\Sigma}_{tt,0} = E(\tilde{\boldsymbol{e}}_{t,0}\tilde{\boldsymbol{e}}_{t,0}')$, $C_{t,0}^{bmk} = E(\boldsymbol{u}_{t|t-1}^{bmk}\tilde{\boldsymbol{e}}_{t,0}')$, it follows from (3.15) after some additional algebra that the benchmarked predictor $\hat{\boldsymbol{\beta}}_t^{bmk}$ of $\boldsymbol{\beta}_t$ is,

$$\hat{\boldsymbol{\beta}}_t^{bmk} = T\hat{\boldsymbol{\beta}}_{t-1}^{bmk} + (P_{t|t-1}^{bmk}\tilde{Z}_t' - C_{t,0}^{bmk})(\tilde{Z}_tP_{t|t-1}^{bmk}\tilde{Z}_t' - \tilde{Z}_tC_{t,0}^{bmk} - C_{t,0}^{bmk\prime}\tilde{Z}_t' + \tilde{\Sigma}_{tt,0})^{-1}(\tilde{Y}_t - \tilde{Z}_tT\hat{\boldsymbol{\beta}}_{t-1}^{bmk}). \quad (4.6)$$

The benchmarked predictors of the small area means and the corresponding variance matrix of the prediction errors are,

$$\hat{\boldsymbol{\theta}}_t^{bmk} = Z_t\hat{\boldsymbol{\beta}}_t^{bmk},\ E[(\hat{\boldsymbol{\theta}}_t^{bmk} - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_t^{bmk} - \boldsymbol{\theta}_t)'] = Z_tP_t^{bmk}Z_t'. \quad (4.7)$$

See Appendix D in PT for the computation of the matrices $C_t^{bmk} = E(\boldsymbol{u}_{t|t-1}^{bmk}, \tilde{\boldsymbol{e}}_t')$ and $P_t^{bmk} = E[(\hat{\boldsymbol{\beta}}_t^{bmk} - \boldsymbol{\beta}_t)(\hat{\boldsymbol{\beta}}_t^{bmk} - \boldsymbol{\beta}_t)']$.

Notice that the enforcement $\sum_{d=1}^{D} b_{dt}e_{dt} = 0$ is only used for computing the benchmarked predictor, but not when computing the variance matrices of the prediction errors. Thus, the matrix P_{t-1}^{bmk} and hence $P_{t|t-1}^{bmk} = TP_{t-1}^{bmk}T' + Q$ appearing in (4.6) are the correct prediction variance matrices.

Remark 15. The variance matrix in (4.7) accounts for the variances and autocovariances of the sampling errors, the variances and autocovariances of the benchmark errors $\sum_{d=1}^{D} b_{dt}e_{dt} = \sum_{d=1}^{D} b_{dt}Y_{dt} - \sum_{d=1}^{D} b_{dt}\boldsymbol{z}_{dt}'\boldsymbol{\beta}_{dt}$ and their covariances with the area sampling errors, and the variances of the model errors.

Two other important properties of the benchmarked predictors are:

(a)- *Unbiasedness*: if $E(\hat{\boldsymbol{\beta}}_{t\text{-}1}^{bmk} - \boldsymbol{\beta}_{t\text{-}1}) = \mathbf{0}$, then under the model, $E(T\hat{\boldsymbol{\beta}}_{t-1}^{bmk} - \boldsymbol{\beta}_t) = \mathbf{0}$ and hence $E(\hat{\boldsymbol{\beta}}_t^{bmk} - \boldsymbol{\beta}_t) = \mathbf{0}$ by (4.6). Thus, to ensure unbiasedness at all time points, it is only required to initialize the filtering process with an unbiased predictor.

(b)- *Consistency*: unlike with pro-rata benchmarking (Eq. 4.2), as the sample size n_{dt} in area d increases, the benchmarked predictor $\hat{\theta}_{dt}^{bmk}$ is consistent for the true area mean θ_{dt}. See Pfeffermann et al. (2014) for discussion of this property.

Remark 16. Pfeffermann and Tiller also developed a corresponding benchmarked smoothing algorithm. Interested readers can contact the authors for details.

4.3 *Two-stage benchmarking*

Survey data are often structured hierarchically, in which case it might be necessary to benchmark the model-dependent predictors at each level of the hierarchy. For example, in the U.S. CPS the States are classified into 9 Census Divisions (CD), and estimates are produced and published for each CD and State. Thus, it is necessary to first benchmark the model-based CD estimates to agree with the reliable design-based national estimate, and then benchmark the model-dependent State estimates within each CD to agree with the CD benchmarked estimate obtained in the first stage. Application of this two-stage benchmarking procedure guarantees publication consistency at each level of the hierarchy.

Pfeffermann et al. (2014) developed a two-stage benchmarking procedure, which is similar to the single-stage procedure in Section (4.2), but with the following main changes. Consider a first-level hierarchy *d*, consisting of S small areas. The benchmark equation is now,

$$\sum_{s=1}^{S} b_{ds,t}\boldsymbol{z}'_{ds,t}\hat{\boldsymbol{\alpha}}_{ds,t} = \boldsymbol{z}'_{dt}\hat{\boldsymbol{\beta}}_{dt}^{bmk} = \hat{\theta}_{dt}^{bmk};\ \ t = 1,2,\ldots, \tag{4.8}$$

where $\hat{\boldsymbol{\beta}}_{dt}^{bmk}$ is the benchmarked predictor in level d. Let $r_{dt}^{bmk} = (\boldsymbol{z}'_{dt}\hat{\boldsymbol{\beta}}_{dt}^{bmk} - \boldsymbol{z}'_{dt}\boldsymbol{\beta}_{dt})$ define the benchmark error at the higher hierarchy. Assuming $\boldsymbol{z}_{ds,t} = \boldsymbol{z}_{dt}\ \forall s$, $\boldsymbol{\beta}_{dt} = \sum_{s=1}^{S} b_{ds,t}\boldsymbol{\beta}_{ds,t}$ and hence, $r_{dt}^{bmk} = (\boldsymbol{z}'_{dt}\hat{\boldsymbol{\beta}}_{dt}^{bmk} - \boldsymbol{z}'_{dt}\sum_{s=1}^{S} b_{ds,t}\boldsymbol{\beta}_{ds,t})$ and $E(r_{dt}^{bmk}) = 0$.

The observed values in the measurement equations are now $\tilde{\boldsymbol{Y}}_t^d = (Y_{d1,t},...,Y_{dS,t},\hat{\theta}_{dt}^{bmk})'$ with $Y_{ds,t}$ denoting the small area survey estimates in the lower hierarchy, and the corresponding vector of sampling and benchmark errors is $\tilde{\boldsymbol{e}}_t^d = (e_{d1,t},...,e_{dS,t},r_{dt}^{bmk})'$; compare with (4.4). The rest of the computations are similar (but more complicated) to the computations in Section (4.2). See Pfeffermann et al. (2014) for details.

5. Recent advances in time-series analysis of repeated surveys

A "hot topic" for more than a decade, not only in time series modelling of survey data, is how to use "big data" information to improve survey estimates, or even replace them. Brakel et al. (2017) apply a bivariate time series model that combines a time series of monthly survey estimates of consumer confidence with a related index series, derived from messages left in social media platforms. The latter series is timelier than the consumer confidence series. The models assumed for each series are similar to the BSM (Section 3), with the sampling errors of the survey estimates assumed to be independent and absorbed in the population irregular term, but allowing the error terms of the slopes in the two series to be correlated. The variances of the error terms of the trends are set to zero to avoid numerical identification problems, which is a common practice in many studies. As shown by the authors, the use of the bivariate model improves the estimates of the population means and permits nowcasting the mean in a concurrent month, when the social media figure is already known, while the survey estimate is still unknown (only available a month later).

It may happen that many big data series related to the survey estimates are available. Schiavoni et al. (2021) propose a dynamic state-space factor model to nowcast monthly unemployment figures, with $n \approx 100$ Google trends. Denote by $\boldsymbol{X}_t$ the n google trend series for time t. In the first step, common factors are computed by applying principal components analysis, using the model,

$$\boldsymbol{X}_t = \Lambda \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t,\ Var(\boldsymbol{\varepsilon}_t) = \Psi;\quad \boldsymbol{f}_t = \boldsymbol{f}_{t-1} + \boldsymbol{u}_t, Var(\boldsymbol{u}_t) = I_r, \tag{5.1}$$

where $\boldsymbol{f}_t$ is a *r*-dimensional vector of common factors with $r \ll n$, Λ is the $n \times r$ matrix of factor loadings and Ψ is diagonal.

In the second step, a joint state-space model is fitted to the LFS series $\boldsymbol{Y}_t = (y_t^t, y_t^{t-3}, y_t^{t-6}, y_t^{t-9}, y_t^{t-12})'$ defined before and the series $\boldsymbol{X}_t$,

$$\begin{pmatrix} \boldsymbol{Y}_t \\ \boldsymbol{X}_t \end{pmatrix} = \begin{pmatrix} \mathbf{1}_5(L_t + S_t) \\ \hat{\Lambda}\boldsymbol{f}_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ \mathbf{0}_n \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_t \\ \boldsymbol{\varepsilon}_t \end{pmatrix}, \tag{5.2}$$

where the model for $\boldsymbol{Y}_t$ is defined as in Brakel and Krieg (2015) mentioned in Section 3.3 with $\boldsymbol{\lambda}_t$ representing the random walk RGB, $\mathbf{1}_5$ is the unit vector of length 5 and $\mathbf{0}_n$ the null vector of order n. The common factors $\boldsymbol{f}_t$ are re-estimated under the combined model.

Another major topic occupying many statisticians throughout the world for more than two years is how to model, or account for the effects of *COVID-19*. In the context of time series analysis of repeated surveys, Brakel et al. (2021) modified the model of Brakel and Krieg (2015) to account for the effects of the pandemic by increasing the variance of the slope disturbance during that period. This way, the trend estimates assign more weight to the current survey estimates and less to past data. See also Harrison and Stevens (1976). (As noted to me by Brakel van den in private communication, if the overall sample size is large enough to publish monthly direct estimates at the national level, one could use instead the benchmark procedure in Section 4.2.) One of the effects of COVID-19 in many countries is that no face-to-face (CAPI) interviews were possible at certain time periods. To deal with this effect, the authors fit a separate time series model with no CAPI responses, and used this model to identify the coefficient of a level shift during the the pandemic. (As mentioned in Section 3.3, Brakel and Krieg, 2009 already account for “shocks” in the series by adding level shifts.) Gonçalves et al. (2021) compare different structural time series models aimed for producing monthly labor force estimates in Brazil during COVID-19.

I conclude this section by considering some recent advances in SAE. The results presented in Section 4 assume a fixed number of areas with observations, implying that the benchmarking procedures considered relate only to these areas. In practice, it is often the case that survey estimates are only available for some of the areas, and the sampled areas with survey estimates, as well as their number may change from one time to another.

Braverman (2022, in final prep), considers this situation. Suppose that the population consists of a fixed number of *M* areas, but at each time *t*, samples $S_{it}, i = 1,...,m_t$ are

available for only m_t out of the *M* areas. The author assumes the following general linear mixed model (LMM) for the sampled area survey estimates:

$$\boldsymbol{Y}_t = X_t\boldsymbol{\beta}_t + Z_t\boldsymbol{v}_t + \boldsymbol{e}_t,\ t = 1,2,... \tag{5.3}$$

where $\boldsymbol{Y}_t$ is now the $m_t \times 1$ vector of survey estimates, $\boldsymbol{\beta}_t$ is a vector of regression coefficients that can vary across areas and over time and $\boldsymbol{v}_t, \boldsymbol{e}_t$ are vectors of random effects and sampling errors, satisfying $E(\boldsymbol{v}_t) = \boldsymbol{0}$, $E(\boldsymbol{v}_\tau \boldsymbol{v}_t') = \Sigma_{\tau t}^{v}$; $E(\boldsymbol{e}_t) = \boldsymbol{0}$, $E(\boldsymbol{e}_\tau \boldsymbol{e}_t') = \Sigma_{\tau t}^{e}$; $E(\boldsymbol{v}_\tau \boldsymbol{e}_t') = 0$ for all τ, t.

The model (5.3) is very general and includes the SAE models reviewed in Section 4 as special cases. It also extends the models considered by Pfeffermann and Burck (1990) and by Rao and Yu (1994), which account for cross-sectional area random effects that do not change over time. However, the model is defined for only the m_t survey estimates.

To include the non-sampled areas in the benchmarking and prediction processes, Braverman (2022) writes the model holding for all the *M* areas. Let $X_t^*, Z_t^*, \boldsymbol{\beta}_t^*, \boldsymbol{v}_t^*$ define the equivalent matrices and vectors of $X_t, Z_t, \boldsymbol{\beta}_t, \boldsymbol{v}_t$ when applied to the *M* areas, such that X_t^* and Z_t^* have now *M* rows, with similar extensions of the dimensions of the other vectors and matrices. Denote by $\boldsymbol{i}_i$ the column vector of length M, with 1 in position i and zeroes elsewhere, such that $[\boldsymbol{i}_1,...,\boldsymbol{i}_M] = \mathrm{I}_M$. Let $S_t = \cup_{i=1}^{m_t} S_{it}$ and define the $(M \times m_t)$ matrix operator,

$$\Delta_t^* = \mathrm{col}_{i \in S_t}\left(\boldsymbol{i}_i'\right)',\ t = 1, 2, ... \tag{5.4}$$

For example, if at time t the first m_t areas are sampled, $\Delta_t^* = [\mathrm{I}_{m_t} \vdots 0_{m_t, M-m_t}]'$. ($0_{a,b}$ defines the zero matrix of dimension (a,b).) It follows that for all t and τ, $\Delta_t^{*\prime} X_t^* \boldsymbol{\beta}_t^* = X_t\boldsymbol{\beta}_t$, $\Delta_t^{*\prime} Z_t^* \boldsymbol{v}_t^* = Z_t\boldsymbol{v}_t$ and $\Delta_\tau^{*\prime}\Sigma_{\tau t}^{v*}\Delta_t^* = \Sigma_{\tau t}^{v}$. Consequently, the model (5.3) can be written in terms of all the *M* areas as,

$$\boldsymbol{Y}_t = \Delta_t^{*\prime} X_t^* \boldsymbol{\beta}_t^* + \Delta_t^{*\prime} Z_t^* \boldsymbol{v}_t^* + \Delta_t^{*\prime} \boldsymbol{e}_t^*, \tag{5.5}$$

with the variance matrices $\Sigma_{\tau t}^{v*}$, $\Sigma_{\tau t}^{e*}$ defined similarly to (5.3), but with respect to all the M areas. (Notice, however, that no sampling errors exist for the nonsampled areas.)

Remark 16. Braverman (2022) assumes an ARMA model for the autocorrelated sampling errors, which as mentioned and illustrated before, is the common practice in many studies. In Section 3.4, I describe and discuss an alternative approach, which only requires defining the correlation structure of the sampling errors.

To fit the model, Braverman (2022) formulates it in state-space form and extends the GLS algorithm developed by Pfeffermann and Tiller (2006). In this case, the observation equation has the form,

$$\begin{pmatrix} \boldsymbol{Y}_t \\ \mathbf{0}_M \end{pmatrix} = \begin{bmatrix} \Delta_{\mathrm{t}}^{*'}\mathrm{X}_{\mathrm{t}}^{*} & \Delta_{\mathrm{t}}^{*'}Z_t^{*} \\ 0_M & -\mathrm{I}_{\mathrm{M}} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_{\mathrm{t}}^{*} \\ \mathbf{v}_t^{*} \end{pmatrix} + \begin{bmatrix} \Delta_{\mathrm{t}}^{*'} & 0_{m_t,M} \\ 0_M & \mathrm{I}_{\mathrm{M}} \end{bmatrix} \begin{pmatrix} \boldsymbol{e}_{\mathrm{t}}^{*} \\ \mathbf{v}_t^{*} \end{pmatrix}, \tag{5.6}$$

with the corresponding state equation defined by the model. As noted in Remark 16, the sampling errors are part of the state vector. Notice the extension of the observed data by a vector of zeros, which permits estimating the area means of non-sampled areas. See Pfeffermann (1984) for optimal properties of this model formulation.

Below is a brief summary of the other new developments in Braverman (2022):

1- Development of a second filtering algorithm, which consists of two separate state equations, one for $\boldsymbol{\beta}_t^{*}$ and the other for $\mathbf{v}_t^{*}$ and $\boldsymbol{e}_{\mathrm{t}}^{*}$. The use of this algorithm has computational advantages compared to the Kalman filter. Like the Kalman filter but unlike the GLS, the filter produces the best predictors (BP) of the area means at time t, based on all the observations until that time point under normality of the error terms, (BLUP otherwise). See the end of Section 3.4 for properties of the GLS filter.

2- Derivation of predictors and prediction variances for areas with no samples, as an integral part of the estimation process,

3- Benchmarking of all the area estimates and not just the areas with direct survey estimates, with corresponding prediction variances of the benchmarked predictors,

4- Development of smoothing algorithms under the LMM (5.3) with the benchmarks.

Braverman (2022) studies the performance of his new developments by an extensive simulation study.

6. Summary

This article is a tribute for Alastair Scott and Fred Smith, highlighting their outstanding contributions to inference on population parameters from repeated surveys. What distinguishes their work is that the ideas were all new, with no direct preceding references. As noted to me by Doctor Bill Bell from the Census Bureau in the U.S., S&S ideas were *"ahead of their time"*.

The pioneering contribution of S&S to inference from repeated surveys was to consider the unknown target population parameters such as means, proportions, etc. as realizations of a stationary time series, which evolves stochastically over time. This was in contrast to the classical survey sampling inference approach used until the publication of their work, under which the true target population parameters are considered as fixed constants, thus accounting only for the randomization distribution over all possible sample selections in the inference process. As shown theoretically and illustrated empirically under different scenarios of the sample overlap over time, assuming that the population parameters evolve stochastically results in much more accurate estimates, notably with small sample sizes. This is true even without specifying explicitly the time series model underlying their evolvement and holds also for the case of new independent samples at each time point. As illustrated in the present article, the approach also formed the basis for SAE from repeated surveys and in away, also for SAE based on cross-sectional surveys where again, a model is assumed for the target population parameters. Many other developments and applications emerging from S&S ideas are reviewed and discussed.

In recent years, more and more external data sources, such as administrative files and big data become available, and with the advancement of data science, there is an obvious desire to use these data sources for the production of official statistics. There are even opinions that in the long run, external data sources should replace the use of surveys. I personally think that at least in the foreseen future, surveys will continue to be needed, for the simple reason that I don't believe that administrative files and big data will be available for all the thousands of questions asked in surveys, and because based on my own experience and unlike what is often claimed, results from a survey are often much faster than from external sources. For example, information about income is available at the end of a survey, whereas by tax office regulations, this

information only becomes available after a year from the end of the reference year (up to two years for businesses).

Irrespectively, I clearly hope to see more and more studies combining survey data with external data sources for enhancing official statistics produced from repeated surveys. The work of Schiavoni et al. (2021) reviewed in Section 5 is a nice example for this kind of inference.

All the developments in this article assume probability samples, under which every unit in the target population has a positive probability to be included in the sample, with known probabilities for the sampled units. However, in recent years there is a growing tendency to use nonprobability samples, although not yet in official statistical bureaus, except for isolated cases. Inference from repeated nonprobability samples is another big challenge for the future.

To conclude, the present article contains many important theoretical developments and applications for inference from repeated surveys, starting with the pioneering contributions of Alastair Scott and Fred Smith, and I do hope that it will prompt new research in this very important area of statistics.

7. References

Bell, W. R. and Hillmer, S.C. (1987). Time series methods for survey estimation. Research report number 87/20, Statistical Research Division, U.S. Census Bureau. Available at, http://www.census.gov/srd/papers/pdf/rr87-20.pdf.

Bell, W.R. and Hillmer, S.C. (1989). Modeling time series subject to sampling error. Research report number 89/01, Statistical Research Division, U.S. Census Bureau. Available at https://www.census.gov/srd/papers/pdf/rr89-01.pdf.

Bell. W.R., Datta, G.S and Ghosh, M. (2012). Benchmarking small area estimators. *Biometrika*, **100**, 189–202.

Bell, W.R. and Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, **16**, 195—215.

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

Blight, B.J.N. and Scott, A.J. (1973). A stochastic model for repeated surveys. *J. royal statist. Soc.,* B, **35**, 61-66.

Bollineni-Balabay, O., J.A. van den Brakel, and Franz Palm (2017). State space time series modelling of the Dutch Labour Force Survey: Model selection and mean squared error estimation. *Survey Methodology*, **43**, 41-67.

Brakel, J.A. van den and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, **35**, 177-190.

Brakel, J.A. van den and Krieg, S. (2016). Small area estimation with state-space common factor models for rotating panels. *J. royal statist. Soc.,* A, **179**, 763-791.

Brakel, J.A. van den., E. Söhler., P. Daas and B. Buelens (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology*, **43**, 183-210.

Brakel, J.A. van den., Souren,M., and Krieg, S. (2021). Estimating monthly labor force figures during the COVID-19 pandemic in the Netherlands. (Submitted to the special issue of *JRSS-A*).

Braverman A. (2022). New Filtering and Smoothing algorithms with benchmarks for small area estimation. (In final preparation).

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. John Wiley & Sons, New York.

Durbin, J. and Quenneville, B. (1997). Benchmarking by State Space Models. *International Statistical Review*, **65**, 23-48.

Gonçalves, C., L. Hidalgo, D. Silva and J.A. van den Brakel (2021). Model-based single-month unemployment rate estimates for the Brazilian Labour Force Survey. (Submitted to the special issue of *JRSS-A*).

Harrison, P.J. and Stevens, C.F. (1976). Bayesian forecasting. *J. royal statist. Soc.,* B, **38**, 205-228.

Harvey, A.C. (1989). *Forecasting Structural Time Series with the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *J. royal statist. Soc.,* A, **163**, 303-339.

Hillmer, S.C. and Trabelsi, A. (1987). Benchmarking of Economic Time Series. *J. Amer. Statist. Ass.* **82**, 1064-1071.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *J. royal statist. Soc.,* B, **42**, 221-226.

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng*. **82** (1), 35-45.

Maravall, A. (1985). On structural time series models and characterization of Components. *J. Bus. Econ. Statist*., **3**, 350-355.

Morris, N.D. and Pfeffermann, D. (1984). A Kalman filter approach to the forecasting of monthly time series affected by moving festivals. *Journal of time series*, **5**, 255-268.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *J. royal statist. Soc.,* B*,* **12**, 241-255.

Pfeffermann, D. (1984). On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients. *J. royal statist. Soc.,* B, **46**, 139-148.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 40-68.

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

Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *J. Bus. Econ. Statist*., **16,** 339-348.

Pfeffermann, D. and Tiller, R. (2005). Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters. *Journal of Time Series Analysis,* **26**, 893-916.

Pfeffermann, D. and Tiller, R. (2006). Small Area Estimation With State-Space Models Subject to Benchmark Constraints. *J. Amer. Statist. Ass.,* **101**, 1387-1397.

Pfefermann, D., Sikov, A. And Tiller, M. (2014). Single- and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test*, **23**, 631–666.

Rao, J.N.K and Yu, M. (1994). Small-Area Estimation by Combining Time-series and Cross-Sectional Data. *The Canadian Journal of Statistics*, **22**, 511-528.

Schiavoni, C., F. Palm, S. Smeekes and Brakel J.A. van den (2021). A dynamic factor model approach to incorporate big data in state space models for official statistics. *J. royal statist. Soc., A*, **184**, 324-353.

Scott, A.J. and Smith, T.M.F. (1974). Analysis of repeated surveys using time series models. *J. Amer. Statist. Ass.,* **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*. Inter. Statist. Rev.*, 45, 13-28.

Smith, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. In: *Survey sampling and measurement*, ed. N.K. Nawboodivi, New York: Academic Press, pp. 201-216.

Tiller, R. B. (1982). Time series modelling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, **8**, 149-166.

Whittle, P. (1963). *Prediction and regulation by linear least-square methods*. London: English Universities Press.

X-13ARIMA-SEATS spec files seasonal adjustment program. Reference Manual Version 1.1 (last revision 2022), Time Series Research Staff, Statistical Research Division U.S. Census Bureau Washington, DC 20233. https://www.census.gov/data/software/x13as/x13sam.html