ESTIMATION AND SEASONAL ADJUSTMENT OF POPULATION MEANS USING DATA FROM REPEATED SURVEYS

D. Pfeffermann
Hebrew University, Jerusalem 91905

Key Words: Kalman Filter, Panels, Rotation Bias, Smoothing, Structural Models

1. INTRODUCTION

The problem considered in this paper is the following: Having a sequence of surveys, carried out at regular time intervals on a given population, how can the data available to the analyst be combined in order to estimate population means and their unobservable components like trend levels and seasonal effects. As illustrated in subsequent sections, the answer to this question depends on three major factors.

1. The sampling design, and in particular, whether or not the samples are partially overlapping so that primary and/or ultimate sampling units are retained in the sample over more than one period,
2. The level of data availability: sometimes all past and current individual data is available with appropriate identification labels, but in other applications, the only available data are the aggregate estimates based on the samples selected in the corresponding time periods. These estimates may or may not include estimates for the sampling errors.
3. The relationships between individual observations at different points of time and the long term behavior of the population means and their components.

We adopt a time series approach by which the components of the population means are considered as random variates which evolve stochastically in time. The process underlying the evolution of the components is known up to a set of parameters which are estimated from the sample data. This approach is in contrast to the classical sampling approach for the analysis of repeated surveys which considers the population means as fixed parameters and hence uses the past data for the estimation of current means only when the surveys are partially overlapping and the distinct panel estimates are known.

The model assumed for the population means is known in the time series literature as the 'Basic Structural Model' and it has been shown to perform well in various empirical studies. A notable feature of this model is that it uses the traditional decomposition of the mean into a trend level component and a seasonal effect which has an immediate interpretation and is routinely used by government offices for the production of seasonally adjusted data. The model is extended to account for the correlations between the panel estimates and it can be applied both in the case of a "primary analysis" for which individual panel estimates are available and in the case of a "secondary analysis" where only the published aggregate estimates are known. The immediate implication of this property is that the extended model permits the estimation of the trend levels and the seasonal effects taking into account the correlations between individual data and employing the distinct panel estimates when available. Estimates for the mean square error of the estimated components are obtained as a by-product of the estimation process.

The plan of the paper is as follows: in section 2 we describe briefly the general form of state space models and their associated inferential method, the Kalman filter. Section 3 defines the basic structural model and discusses its application under various combinations of rotation patterns and data availability. The model is extended in section 4 to account for rotation group effects, a phenomenon known to sometimes affect estimates obtained from repeated surveys. Section 5 describes the method used for the initialization and estimation of the Kalman filter. Empirical results illustrating the main features of the proposed procedure and comparing its performance to the performance of other procedures are presented in section 6. The empirical study uses simulated data and two actual series collected as part of the Israel Labor Force Survey. Section 7 contains some concluding remarks.

Some key references to the classical sampling approach for the analysis of repeated survey data are the articles by Jessen (1942), Patterson (1950), Rao and Graham (1964), Gurney and Daly (1965) and Cochran (1977, sections 2.10-2.12). The time series approach has been explored in the articles by Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Jones (1979, 1980), Hausman and Watson (1985), Tam (1987) and Binder and Dick (1989). Smith (1979) and Binder and Hidiroglou (1988) review the above and other related articles discussing in detail the pros and cons of the two approaches.

2. STATE SPACE MODELS AND THE KALMAN FILTER

In this section we review briefly the basic structure of state space models and their accompanying Kalman filter equations (Kalman, 1960), focusing on aspects most germane to the anlaysis presented in subsequent sections.

State-Space models consist in general of two sets of linear equations which define how the observable and unobservable model components evolve stochastically in time. The following definitions and assumptions stem from the special structure of repeated survey data.

I. <u>Observations Equation:</u>

$$\underset{\sim}{Y}_t = Z_t \underset{\sim}{\alpha}_t + \underset{\sim}{\varepsilon}_t \tag{2.1}$$

where $\underset{\sim}{Y}_t$ is the vector of observations (estimators) at time t, Z_t is a known design matrix, $\underset{\sim}{\alpha}_t$ is a vector of unknown 'state components' (e.g. components comprising the population mean) which are allowed to vary in time and $\underset{\sim}{\varepsilon}_t$ is a vector of disturbances (estimation errors) satisfying the "wide sense" requirements,

$$E(\underset{\sim}{\varepsilon}_t) = 0; \; E(\underset{\sim}{\varepsilon}_t \, \underset{\sim}{\varepsilon}'_{t-k}) = V_t^{(k)}, \; k = 0, 1, \ldots \tag{2.2}$$

Notice that the error terms are allowed to be serially correlated. Serial correlations arise in repeated surveys when primary and/or ultimate sampling units are retained in the sample over several occasions.

II. <u>System Equation</u>

$$\underset{\sim}{\alpha}_t = T_t \underset{\sim}{\alpha}_{t-1} + \underset{\sim}{\eta}_t \tag{2.3}$$

where T_t is a transition matrix and $\underset{\sim}{\eta}_t$ is another vector of disturbances which is independent of the vectors $(\underset{\sim}{\varepsilon}_{t-k})$, k = 0, 1, ... and satisfies the conditions

$$E(\underset{\sim}{\eta}_t) = \underset{\sim}{0};\ E(\underset{\sim}{\eta}_t\underset{\sim}{\eta}_t') = Q_t;\ E(\underset{\sim}{\eta}_t\underset{\sim}{\eta}_{t-k}') = 0,\ k\geq 1 \quad (2.4)$$

Assuming that the V-C matrices $V_t^{(k)}$ and Q_t are known, the state vectors $\underset{\sim}{\alpha}_t$ can be estimated most conveniently by means of the Kalman filter. The filter consists of a set of recursive equations which can be used to update and smooth estimates of current and previous state vectors and to predict future vectors every time that new data become available. A good reference to the theory of the Kalman filter is the book by Anderson and Moore (1979). In the next section we show that the problem of serially correlated errors can be overcome by including the errors as part of the state vectors. Hence we present below the Kalman filter equations for the simpler case where all the error terms are serially independent.

Let $\hat{\underset{\sim}{\alpha}}_{t-1}$ be the best linear unbiased predictor (blup) of $\underset{\sim}{\alpha}_{t-1}$ based on the data observed up to time t-1. Since $\hat{\underset{\sim}{\alpha}}_{t-1}$ if blup for $\underset{\sim}{\alpha}_{t-1}$, $\hat{\underset{\sim}{\alpha}}_{t|t-1} = T_t\hat{\underset{\sim}{\alpha}}_{t-1}$ is the blup of $\underset{\sim}{\alpha}_t$ based on all the information up to time (t-1). Furthermore, if $P_{t-1} = E(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})'$ is the V-C matrix of the prediction errors at time t-1, $P_{t|t-1} = T_tP_{t-1}T_t' + Q_t$ is the V-C matrix of the prediction errors $(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)$. (Follows straightforwardly from equations 2.3 and 2.4).

When a new vector of observations $\underset{\sim}{Y}_t$ becomes available, the predictor of $\underset{\sim}{\alpha}_t$ and the V-C matrix P_t are updated according to the formulae

$$\hat{\underset{\sim}{\alpha}}_t = \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1}Z_t'F_t^{-1}(\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1}) \quad (2.5)$$

and

$$P_t = P_{t|t-1} - P_{t|t-1}Z_t'F_t^{-1}Z_tP_{t|t-1} \quad (2.6)$$

where $\hat{\underset{\sim}{Y}}_{t|t-1} = Z_t\hat{\underset{\sim}{\alpha}}_{t|t-1}$ is the blup of $\underset{\sim}{Y}_t$ at time (t-1) so that $(\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1})$ is the vector of innovations (prediction errors) with V-C matrix $F_t = (Z_tP_{t|t-1}Z_t' + V_t)$.

An important use of the Kalman filter is the updating (smoothing) of past state estimates as new, more recent data become available, e.g. smoothing the estimates of the seasonal effects for previous periods. Denoting by T^* the most recent period for which observations are available, the smoothing is carried out using the equation

$$\hat{\underset{\sim}{\alpha}}_{t|T^*} = \hat{\underset{\sim}{\alpha}}_t + P_tT_{t+1}'P_{t+1|t}^{-1}(\hat{\underset{\sim}{\alpha}}_{t+1|T^*} - T_{t+1}\hat{\underset{\sim}{\alpha}}_t),\quad t = 2, 3, \ldots, T^* \quad (2.7)$$

where $P_{t|T^*} = E(\hat{\underset{\sim}{\alpha}}_{t|T^*} - \underset{\sim}{\alpha}_t)(\hat{\underset{\sim}{\alpha}}_{t|T^*} - \underset{\sim}{\alpha}_t)'$ satisfies the equation

$$P_{t|T^*} = P_t + P_tT_{t+1}'P_{t+1|t}^{-1}(P_{t+1|T^*} - P_{t+1|t})P_{t+1|t}^{-1}T_{t+1}P_t,\quad t = 2, \ldots, T^* \quad (2.8)$$

Notice from (2.7) and (2.8) that $\underset{\sim}{\alpha}_{T^*|T^*} = \underset{\sim}{\alpha}_{T^*}$ and $P_{T^*|T^*} = P_{T^*}$ which defines the starting values for the smoothing equations.

The actual application of the Kalman filter requires the estimation of the unknown V-C matrices V_t and Q_t, the initial state vector α_0 and the initial V-C matric P_0. We address these issues in section 5.

3. BASIC STRUCTURAL MODELS FOR REPEATED SURVEYS

3.1 System Equations for the Components of the Population Mean

The model considered in this study consists of the following system equations describing the evolution of the population mean and its components over time. For convenience of presentation we assume that the data are collected on a quarterly basis.

$$\theta_t = L_t + S_t$$

$$L_t = L_{t-1} + R_{t-1} + \eta_{Lt};\quad R_t = R_{t-1} + \eta_{Rt}$$

$$\sum_{j=0}^{3} S_{t-j} = \eta_{St} \quad (3.1)$$

where $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ are three independent white noise processes with mean zero and variances σ_L^2, σ_R^2 and σ_S^2, respectively. The first equation postulates an additive decomposition of the population mean θ_t into a trend level component L_t and a seasonal effect S_t. As noted earlier, such a decomposition is inherent in the seasonal adjustment procedures in common use like, for example, X-11 ARIMA (Dagum, 1980). Other components like moving festivals and trading days effects can likewise be incorporated in the decomposition equation - Morris and Pfeffermann (1984), Dagum and Quenneville (1988).

The second and third equations approximate a local linear trend (the case of a constant level is a special case by which $\sigma_L^2=\sigma_R^2=0$ and $R_0=0$) where as the last equation models the variation of the seasonal effects. As can be seen, the model permits changes in the seasonal pattern but imposes the condition that the expectation of the sum of the seasonal effects over a given span Δ (four quarters in our case) is zero. (Constant seasonality is obtained when $\sigma_S^2=0$).

The model defined by (3.1) is known in the statistical literature as the "Basic Structural Model". The theoretical properties of this model in relation to other models are discussed in Harrison and Stevens (1976), Harvey and Todd (1983), Harvey (1984) and Maravall (1985). Although this model is more restricted compared to the family of ARIMA models, it is flexible enough to approximate the behaviour of many different time series as illustrated empirically by Harvey and Todd (1983), Morris and Pfeffermann (1984), Dagum and Quenneville (1988) and Quenneville and Dagum (1988). Important features of the model pertaining to the present problem are discussed in subsequent sections.

The model defined by the last three equations of (3.1) can be written alternatively as

$$\underset{\sim}{\alpha}_t^{(1)} = T_{11}\underset{\sim}{\alpha}_{t-1}^{(1)} + \underset{\sim}{\eta}_t^{(1)} \quad (3.1')$$

where $\underset{\sim}{\alpha}_t^{(1)\prime} = (L_t, R_t, S_t, S_{t-1}, S_{t-2})$ is the state vector at time t,

$$T_{11} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$ is a time invariant transition matrix and

$\underset{\sim}{\eta}_t^{(1)}$ is the corresponding error vector with mean zero and V-C matrix $Q_{11} = \mathrm{Diag}(\sigma_L^2, \sigma_R^2, \sigma_S^2, 0, 0)$. We use the representation (3.1') in subsequent sections.

3.2 Observations Equation for the Survey Estimators

The model equations for the survey estimators depend on the sampling design, the rotation pattern and the covariances between individual observations. In the present study we follow Blight and Scott (1973) and assume that observations $\{Y_{ti}\}$ pertaining to the same unit i follow a first order autoregressive model, i.e.

$$Y_{ti} - \theta_t = \rho(Y_{(t-1),i} - \theta_{t-1}) + v_{ti} \quad (3.2)$$

where the errors $\{v_{ti};\ t = 2, 3, \ldots\}$ are white noise with mean zero and variance σ_v^2 and $|\rho| < 1$. This is a standard assumption made (sometimes implicitly) in essentially all the articles mentioned in section 1. It implies that correlations between individual observations decay geometrically as time passes. It is assumed also that the sampling design is ignorable (Sugden and Smith, 1984) and that observations pertaining to different individuals are independent. The model can be extended to the case of a two stage sampling design by adding random components λ_{tk} to represent random cluster effects so that $Y_{tkj} = \theta_t + \lambda_{tk} + \varepsilon_{tkj}$. Assuming that the cluster effects follow a separate autoregressive relationship, the model can be analysed similarly to the present case. The model accounts then for contemporary and serial correlations between observations pertaining to different ultimate units belonging to the same cluster (cf. Scott, Smith and Jones, 1977).

The other factor determining the observations equation for the survey estimators is the rotation pattern. Consider first the special case of a non-overlapping survey. Assuming that the samples selected at different time periods can be considered as independent, the observation equation is

$$\bar{Y}_t = \theta_t + \bar{v}_t;\quad E(\bar{v}_t) = 0,\ E(\bar{v}_t\bar{v}_{t-k}) = \begin{cases} \sigma_v^2/n_t & k=0 \\ 0 & \text{otherwise} \end{cases} \quad (3.3)$$

where $\bar{Y}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} Y_{ti}$ is the aggregate survey estimator at time t and $\bar{v}_t$ is the corresponding survey error. The model defined by (3.1') and (3.3) specifies the basic structural model to be used in the case of a nonoverlapping survey.

Next we consider the case of overlapping surveys and to illustrate the ideas we focus for convenience on the Israel Labour Force Survey (ILFS) which provides the data used in the empirical study of section 6. Other rotation patterns can be handled in a similar way. The ILFS is a quarterly survey of households carried out by the Central Bureau of Statistics (CBS) to provide information on employment and other important demographic and socio-economic characteristics of the labour force in Israel. Every quarter the CBS surveys four panels each composed of approximately 3000 households so that three panels have been included in some past surveys and one panel is new. Every new panel is included in the survey for two quarters, left out of the survey for the next two quarters and then included again for two more quarters. This rotation pattern produces a 50% overlap between two successive quarters and a 50% overlap between quarters representing the same months in two successive years. For a brief description of the sampling design used for the ILFS, see Nathan and Eliav (1988). As discussed there, the four concurrent panels can be considered as independent simple random samples of households.

In what follows we define $\bar{Y}_t^{t-j} = \frac{1}{m}\sum_{i=1}^{m} y_{ti}^{t-j}$ to be the mean observed at time t for the panel joining the survey for the first time at time t-j, j=0, 1, 4, 5. It is assumed for convenience that the panels are of fixed size m. The aggregate survey estimate at time t will be denoted as before by $\bar{Y}_t = \frac{1}{4}(\bar{Y}_t^t + \bar{Y}_t^{t-1} + \bar{Y}_t^{t-4} + \bar{Y}_t^{t-5})$. We distinguish between the case where the panel estimates are known and the case where the only available data at time t is the aggregate estimate $\bar{Y}_t$. For the first case we have

$$\underset{\sim}{\bar{Y}}_t = \underset{\sim}{1}_4\theta_t + \underset{\sim}{\bar{\varepsilon}}_t \quad (3.4)$$

where $\underset{\sim}{1}_4$ is the unit vector of length 4, $\underset{\sim}{\bar{Y}}_t' = (\bar{Y}_t^t, \bar{Y}_t^{t-1}, \bar{Y}_t^{t-4}, \bar{Y}_t^{t-5})$ is the row vector of panel estimators at time t and $\underset{\sim}{\bar{\varepsilon}}_t' = (\bar{\varepsilon}_t^t, \bar{\varepsilon}_t^{t-1}, \bar{\varepsilon}_t^{t-4}, \bar{\varepsilon}_t^{t-5})$ is the corresponding vector of survey errors satisfying the transitive relationships

$$\underset{\sim}{\alpha}_t^{(2)} = \begin{bmatrix} \bar{\varepsilon}_t^t \\ \bar{\varepsilon}_t^{t-1} \\ \bar{\varepsilon}_t^{t-4} \\ \bar{\varepsilon}_t^{t-5} \\ \bar{\varepsilon}_{t-2}^{t-3} \\ \bar{\varepsilon}_{t-1}^{t-2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \rho & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho^3 & 0 \\ 0 & 0 & \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{\varepsilon}_{t-1}^{t-1} \\ \bar{\varepsilon}_{t-1}^{t-2} \\ \bar{\varepsilon}_{t-1}^{t-5} \\ \bar{\varepsilon}_{t-1}^{t-6} \\ \bar{\varepsilon}_{t-3}^{t-4} \\ \bar{\varepsilon}_{t-2}^{t-3} \end{bmatrix} + \underset{\sim}{\eta}_t^{(2)} = T_{22}\underset{\sim}{\alpha}_{t-1}^{(2)} + \underset{\sim}{\eta}_t^{(2)} \quad (3.5)$$

where $\underset{\sim}{\eta}_t^{(2)\prime} = (\bar{\varepsilon}_t^t, \bar{v}_t^{t-1}, \rho^2\bar{v}_{t-2}^{t-4} + \rho\bar{v}_{t-1}^{t-4} + \bar{v}_t^{t-4}, \bar{v}_t^{t-5}, 0, 0)$ is a vector of independent disturbances which is uncorrelated with the vectors $\{\underset{\sim}{\alpha}_{t-j}^{(2)}\}$ and $\{\underset{\sim}{\eta}_{t-j}^{(2)}\}$, $j \geq 1$ and has mean zero and V-C matrix

$$V(\underset{\sim}{\eta}_t^{(2)}) = (\tfrac{1}{m}\sigma^2 v)\ \mathrm{Diag}[(1-\rho^2)^{-1}, 1, (\rho^4 + \rho^2 + 1), 1, 0, 0] = Q_{22} \quad (3.6)$$

Equations (3.5) and (3.6) follow directly from the autoregressive assumption (3.2) so that $\bar{v}_{t-k}^{t-j} = \frac{1}{m} \sum_{i=1}^{m} v_{t-k,i}^{t-j}$ is the mean of the white noise disturbances at time (t-k) for the panel joining the sample for the first time at time (t-j), $j \geq k$. We included the survey errors $\bar{\varepsilon}_{t-1}^{t-2}$ and $\bar{\varepsilon}_{t-1}^{t-6}$ in the vector $\alpha_{t-1}^{(2)}$ and the errors $\bar{\varepsilon}_t^t$ and $\bar{\varepsilon}_{t-2}^{t-3}$ in the vector $\alpha_t^{(2)}$ in order that $\alpha_{t-1}^{(2)}$ will contain the same components as $\alpha_t^{(2)}$ with a time shift of 1. Since $\bar{\varepsilon}_{t-2}^{t-3}$ had to be added to $\alpha_t^{(2)}$ we added it also to $\alpha_{t-1}^{(2)}$, insuring that way that the error vector $\eta_t^{(2)}$ will be independent of past state vectors. This in turn required that $\bar{\varepsilon}_{t-1}^{t-2}$ will be added to $\alpha_t^{(2)}$.
Obviously when the same components are included in both the vectors, the corresponding residual variance is set to zero. This strategy can be applied for general rotation patterns.

For the case of a secondary analysis we have

$$\bar{Y}_t = \theta_t + \frac{1}{4}(\bar{\varepsilon}_t + \bar{\varepsilon}_t^{t-1} + \bar{\varepsilon}_t^{t-4} + \bar{\varepsilon}_t^{t-5}) = \theta_t + \frac{1}{4} 1'\bar{\varepsilon}_t \quad (3.7)$$

with (3.5) and (3.6) remaining unchanged.

Equations (3.4) and (3.7) define the observations equation for the case of overlapping surveys. However, unlike the case of independent samples, the survey errors are now correlated. A simple way to overcome this problem in our case is by including the survey errors as part of the state vector and setting the residual variances of the observations equation to zero. The resulting model is specified in the next section.

3.3 A Compact Model Representation

The model defined by (3.1'), (3.4), (3.5) and (3.6) corresponding to the case of a primary analysis can be written compactly as

$$\bar{Y}_t = [1_4, 0_4, 1_4, 0_4, 0_4 \,|\, I_4, 0_4, 0_4] \begin{bmatrix} \alpha_t^{(1)} \\ \alpha_t^{(2)} \end{bmatrix} = Z\alpha_t \quad (3.8)$$

where 0_4 is a vector of zeros of length 4, I_4 is the identity matrix of order 4 and $\alpha_t' = (\alpha_t^{(1)\prime}, \alpha_t^{(2)\prime})$ is the augmented state vector satisfying the transition equation

$$\alpha_t = \begin{bmatrix} T_{11}, & {}_5 0_6 \\ {}_6 0_5, & T_{22} \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} \eta_t^{(1)} \\ \eta_t^{(2)} \end{bmatrix} = T\alpha_{t-1} + \eta_t \quad (3.9)$$

In (3.9) T_{11} is the transition matrix of the state vector $\alpha_t^{(1)}$ defining the evolution of the population mean components (equation 3.1'), T_{22} is the transition matrix of the survey errors (equation 3.5) and ${}_\ell 0_k$ defines a zero matrix of order $(\ell \times k)$. Notice that the elements of η_t are independent so that $Q = V(\eta_t)$ is diagonal with $Q_{11} = V(\eta_t^{(1)})$ and $Q_{22} = V(\eta_t^{(2)})$ comprising the diagonal elements.

For the case of a secondary analysis (equations 3.1', 3.5, 3.6 and 3.7) the matrix Z of equation (4.8) is replaced by the row vector

$$z' = (1, 0, 1, 0, 0, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, 0, 0) \quad (3.10)$$

so that $\bar{Y}_t = z'\alpha_t$ (compare with 3.7). However, the system equations (3.9) remain unchanged. Thus, the model preserves the intrinsic relationships (3.5) among the separate panel estimators even though the only available data are the aggregate estimators $\{\bar{Y}_t\}$. An interesting consequence of this formulation is that one can actually predict the original panel estimates $\{\bar{Y}_t^{t-j}\}$ using the relationship $\bar{Y}_t^{t-j} = L_t + S_t + \bar{\varepsilon}_t^{t-j} = c'\alpha_t$, say. (Equation 3.7 guarantees that the average of the four predictors equals the aggregate estimate $\bar{Y}_t$).
Such an analysis might be useful for model diagnostic, e.g. by comparing the prediction bias and mean square error (MSE) of the distinct panel estimates as obtained under primary and secondary analyses: see table 3 of section 6.2.

The common approach to the modelling of the behaviour of the survey errors in the case of a secondary analysis is to postulate a moving average (M.A.) process for the errors $\bar{e}_t = \bar{Y}_t - \theta_t$ as induced by the fact that the errors are uncorrelated after a certain lag determined by the rotation pattern. This formulation does not allow the prediction of the separate panel estimates. Notice also that by utilizing the relationship (3.2), the model holding for the survey errors includes in our case only two unknown parameters compared to five parameters if a general M.A. process of order 5 is used. (As easily seen, for the model defined by 3.10 and 3.9, $\mathrm{Cov}(\bar{Y}_t, \bar{Y}_{t-j}) \neq 0$ for j=0, 1, 3, 4, 5. One could argue on the other hand that postulating a general moving average process is more robust. In particular, the M.A. formulation does not require that the panel estimators corresponding to the same time period are independent.) The use of M.A. models for the survey errors is less obvious in the case of a primary analysis because of the different time gaps in which the panels are not observed.

The model defined by (3.8) (or 3.10) and (3.9) conforms to the general state-space formulation presented in section 2. Hence, once the unknown variances and the initial state components have been estimated, (the analysis of Maravall, 1985 illustrates that the model is uniquely identified), the Kalman filter equations can be applied to estimate the population means or changes in the means using the relationship, $\theta_t = (L_t + S_t)$. Moreover, the use of the present model permits the extraction of the seasonal effects in a straightforward manner taking into account the correlations between the survey estimation errors and using the distinct panel estimates in the case of a primary analysis. Thus, the approach outlined in this article enables to decompose the means into a trend level component and seasonal effects using more information than is commonly used by the traditional procedures for seasonal adjustments. These advantages are illustrated in the simulation study described in section 6.1. As mentioned in the introduction and becomes evident from the Kalman filter equations, the model provides estimates for the mean square errors of the estimated components at any given time period. (Quenneville and Dagum, 1988, propose to estimate the variances of the X-11 ARIMA estimates by fitting basic structural models which approximate the behaviour of the X-11 ARIMA components.) Obviously, the price paid for this flexibility is that the analysis is more model dependent compared, for example, to the use of the X-11 ARIMA procedure.

4. ACCOUNTING FOR ROTATION GROUP BIAS

The problem of rotation group bias (RGB) is that some of the panel estimators may be biased. In its classical use, RGB refers to a phenomenon by which respondents provide different information on different rounds of interview, depending on the length of time that they have been included in the sample. However, the phenomenon of RGB or at least its magnitude could be related to the method of data collection (e.g. home interview in some rounds and telephone interview in other rounds) or even result from differential nonresponse. Here and in section 6.2 we refer to RGB in this broader context.

Bailar (1975) found clear evidence for rotation bias in some of the labour force data collected at the U.S. Current Population Survey. Kumar and Lee (1983) found similar evidence in the Canadian Labour Force Survey. A review of these and other studies on rotation bias can be found in Binder and Hidiroglou (1988).

Using previous notation, rotation bias implies that $E(\bar{y}_t^{t-j} - \theta_t) = \beta_{jt} \neq 0$ for some j. Bailar (1975) and Kumar and Lee (1983) assume that the bias factors are time invariant which implies in our case that $\bar{y}_t^{t-j} = \theta_t + \bar{\varepsilon}_t^{t-j} + \beta_j$ or that

$$\bar{Y}_t = 1_4 (L_t + S_t) + I_4 \beta + \bar{\varepsilon}_t \qquad (4.1)$$

where $\beta' = (\beta_0, \beta_1, \beta_4, \beta_5)$ is a vector of constants. Equation (4.1), combined with (3.9) defines a model for incorporating constant RGB effects. However, the equations (4.1) and (3.9) alone are not sufficient for estimating the group effects and securing unbiased predictors for the population means because of the confounding effects of the trend level and a fixed shift in the bias coefficients. Thus, one needs to augment the model by a linear constraint of the form

$$\sum_j w_j \beta_j = w_0 \;., \; \sum_j w_j \neq 0 \qquad (4.2)$$

with known coefficients w_j in order to secure the identifiability of all the model components.

This problem is obviously not unique to the present model. If all the panel estimators are biased, one cannot hope for an unbiased estimator of the population mean without some information on the magnitude and relationship of the bias factors. Bailar (1975) assesses the bias by examining alternative data sources. Kumar and Lee (1983) assume that the bias coefficients add to zero in their analysis. We make a similar assumption in section 6.2 based on preliminary analysis of the data. In the absence of external information, this is a reasonable condition since it permits to test for the existence of group effects conditional on the assumption that the aggregate estimates are unbiased.

The model defined by (4.1) and (4.2) can be extended to the case of time changing group effects by permitting the bias factors to vary stochastically over time, similarly to the other model components, taking into account the possible correlations between them. Such an extension is not considered in the present study.

5. ESTIMATION AND INITIALIZATION OF THE KALMAN FILTER

The actual application of the Kalman filter requires the estimation of the autocorrelation coefficient ρ and the unknown elements of the V-C matrix Q as well as the initialization of the filter, that is, the estimation of the state vector α_0 and the V-C matrix P_0. In this section we describe the estimation methods used in the present study.

Assuming that the disturbances $\eta_t' = (\eta_t^{(1)'}, \eta_t^{(2)'})$ are normally distributed, the log likelihood function for the observations can be written as

$$L(\delta) = \text{constant } (-1/2) \sum_{t=1}^{T} (\log |F_t| + e_t' F_t^{-1} e_t) \qquad (5.1)$$

where $e_t = Y_t - \hat{Y}_{t|t-1}$ is the vector of innovations (prediction errors) and $\delta' = (\sigma_L^2, \sigma_R^2, \sigma_S^2, \sigma_V^2, \rho)$ is the vector of unknown model parameters.

Let $\delta_{(0)}$, $\hat{\alpha}_0$ and $\hat{p}_0$ define the initial estimates of δ, α_0 and p_0. A simple way to maximize the likelihood function (5.1) is by application of the method of scoring which consists of solving iteratively the equation

$$\delta_{(i)} = \delta_{(i-1)} + \lambda_i \{I[\delta_{(i-1)}]\}^{-1} G[\delta_{(i-1)}] \qquad (5.2)$$

In (5.2), $\delta_{(i-1)}$ is the estimator of δ as obtained in the (i-1)th iteration, $I[\delta_{(i-1)}]$ is the information matrix evaluated at $\delta_{(i-1)}$ and $G[\delta_{(i-1)}]$ is the gradient of the log likelihood again evaluated at $\delta_{(i-1)}$. The coefficient λ_i is a variable step length determined by a grid search procedure and introduced to guarantee that $L[\delta_{(i)}] \geq L[\delta_{(i-1)}]$ at each iteration. The formulae for the k-th element of the gradient vector and the kl-th element of the information matrix are given in Watson and Engle (1983).

The iterative solution of (5.2) may converge to negative variance estimators or become unstable if an iteration produces negative estimates. A similar instability may occur if ρ is estimated by a value outside the unit circle. In order to avoid this possibility, we transformed the vector δ' to the vector $\delta^{*\prime} = [\sigma_L, \sigma_R, \sigma_S, \sigma_V, \psi]$ where $\rho = \psi/(1+|\psi|)$ so that the likelihood function has been maximized with respect to the elements of δ^* rather than the elements of δ. We used two convergence criterions: $|\{L[\delta_{(i)}] - L[\delta_{(i-1)}]\}/L[\delta_{(i-1)}]| < 10^{-9}$ and $\max_j \{|\hat{\delta}_{(i)}(j) - \hat{\delta}_{(i-1)}(j)|/|\delta_{(i-1)}(j)|\} \leq .01$ where $\delta(j)$ stands for any one of the parameters. The algorithm has been stopped once one of the criterions was fulfilled (usually in less than 20 iterations).

Initialization of the Kalman filter was carried out following the approach proposed by Harvey and Peters (1984). By this approach, the nonstationary components of the state vector are initialized with very large error variances (which amounts to postulating a diffuse prior) so that the corresponding state estimates can conveniently be taken as zeros. The stationary components are initialized by the corresponding unconditional means and variances. For the model defined by (3.9), the stationary components are the six survey errors comprising the sub-vector $\alpha_0^{(2)}$, having zero mean and variance $(1/m)\sigma_v^2(1 - \rho^2)^{-1}$. In view of the use of large error variances for the nonstationary state components, the likelihood is estimated based on the last T-d observations where d is the number of the nonstationary components.

The use of this procedure has the clear advantage of being computationally very simple. Other approaches to initializing the Kalman filter that could be applied to the models considered in the present paper are discussed in Ansley and Kohn (1985) and DeJong (1988).

A computer program which implements the methods described in this section for the updating, smoothing and prediction of the state vectors of the models proposed in sections 3 and 4 has been written using the procedure PROC-IML of the SAS system. The program is a modification of the software DLM developed at Statistics Canada by Quenneville (1988).

6. SIMULATION AND EMPIRICAL RESULTS

In this section we describe the results of an empirical study aimed to illustrate some of the major features of the models presented in section 3. The study consists of two parts. In the first part discussed in section 6.1, we use simulated series thus enabling us to control the values of the model parameters. In the second part we use two actual series collected as part of the Israel Labour Force Survey. In this part we extend the models of section 3 by permitting for rotation group effects as described in section 4. The results are discussed in section 6.2.

6.1 Simulation Results

We generated several data sets, each containing 15 independent series of panel estimates with four panels for every time period. The panel estimates were generated so that they obey the model and rotation pattern defined by (3.8) and (3.9). The difference between the various sets is in the values of the model parameters and in the length of the series.

Here we focus mainly on the results obtained for two groups of data, each consisting of 2 separate data sets, one composed of series of length $T^*=100$ and the other restricted to series of length $T^*=36$. The latter is the length of the labour force series analysed in section 6.2. For the first group we used very small values of σ_L^2, σ_R^2, σ_S^2 implying an almost perfect linear trend with constant seasonality. We increased the values of the variances for the other group imposing that way more rapid changes in the components of the population mean. We used a relatively high value of $\rho=0.7$ for both series thereby emphasizing the effect of the rotation pattern. The variances σ_v^2 used are such that the survey errors account for about 20 percent of the MSE of the quarter to quarter difference in the aggregate means in the first group and for about 6 percent in the second group. In the discussion below we mention briefly the results obtained for other values of ρ and σ_v^2.

As benchmark comparisons with the model results we have estimated the seasonal effects using the X-11 procedure and computed the sampling variance of Patterson's (1950) estimator of the population mean for the case of "sampling on more than two occasions". The variance is specified in formula 12.84 of Cochran (1977). Notice that the variance of this estimator is minimized in the case of a 50 percent sample overlap between two successive surveys which is the case in the present study.

The results obtained for the two groups are exhibited in tables 1 and 2 as averages over the 15 series considered in each case. We distinguish between the case where Q is known (denoted "Corr.Q") and the case where Q (and hence also the transition matrix T) is estimated, and between the results obtained for a primary analysis which uses the distinct panel estimates and the results obtained for a secondary analysis using only the aggregate estimates. Another distinction made is between the prediction one step ahead of the aggregate sample estimates (using either the correct model or the model which ignores the correlations between the survey errors) and the prediction of the distinct panel estimates (presented as average over the 4 panels). The prediction errors are presented mainly for comparison between primary and secondary analysis.

The main results emerging from the tables can be summarized as follows:

1) The use of primary analysis dominates the use of secondary analysis in almost every aspect studied (notice in particular the estimation of the survey error parameters in the case of the short series). The better performance in the case of a primary analysis is seen to hold also in the estimation of the seasonal effects (see table 2) despite the use of the smoothed estimators, an issue not investigated so far.

2) Estimating the unknown model parameters by the method of scoring yields satisfactory results in the case of a primary analysis. The results are less encouraging, however, in the case of a secondary analysis with respect to the estimation of the survey error parameters. This outcome is explained by the fact that these parameters index the relationship between the panel estimators whereas the panel estimators are not observable in the case of a secondary analysis. Nevertheless and as emphasized in section 4, the survey parameters can be estimated consistently even in the case of a secondary analysis which is illustrated very clearly by comparing the results obtained for the long and the short series.

   Another notable result is the estimation of the seasonal effects in the case of the first group of data. The smoothed estimates of the seasonal effects using the correct Q matrix outperform in this case the smoothed empirical estimates obtained by using the estimated variances which could be expected considering the very small value of σ_S^2.

   Still, the empirical estimates perform well even in this case and interesting enough, the use of the estimated variances is reflected also in the estimates of the MSE's of the smoothed estimators so that for the long series, the latter estimates again perform relatively well. For the short series, the MSE's estimates underestimate the true MSE's, a well known phenomenon in other applications

Table 1: Simulation Results, Data Set I

Initial Components : $L_0 = 100,\ R_0 = 5,\ \underset{\sim}{S}_0' = (4, 1, -3, -2)$

Residual State Variances: $\sigma_L^2 = 0.1,\ \sigma_R^2 = 0.1,\ \sigma_S^2 = 10^{-4}$

Survey Error Parameters: $\sigma_e^2 = 8$[1] ., $\rho = 0.7$

		100 Time Points				36 Time Points			
		Primary		Secondary		Primary		Secondary	
		Corr.Q	Est.Q	Corr.Q	Est.Q	Corr.Q	Est.Q	Corr.Q	Est.Q
Predic. Bias	panels	0.35	0.65	0.36	0.38	0.63	0.67	0.69	0.60
	aggregate	-0.09	0.25	-0.16	-0.14	-0.05	0.06	-0.05	0.09
Predic. MSE	panels	26.00	26.20	32.70	32.90	25.80	25.40	32.40	32.60
	aggregate	8.25	8.40	9.67	9.76	8.42	7.74	9.58	8.62
	$\rho=0$[2]	10.75	-	10.75	-	11.32	-	11.32	-
Est. of survey error parameters	σ_e^2 Bias	-	-0.24	-	-0.66	-	-0.72	-	-1.20
	σ_e^2 RMSE	-	0.56	-	3.20	-	1.44	-	6.20
	ρ Bias	-	-0.01	-	-0.05	-	0.00	-	-0.27
	ρ RMSE	-	0.03	-	0.10	-	0.07	-	0.26
Var. of Est.[3] of population means	model	3.90	4.22	4.13	4.56	3.55	3.97	3.76	3.35
	realized	4.10	4.45	4.35	4.85	3.65	4.20	3.84	5.53
	patterson	6.50	6.50	-	-	6.60	6.60	-	-
MSE of Est.[4] of Seasonal Effects	model	0.003	0.25	0.003	0.32	0.001	0.25	0.001	0.26
	realized	0.003	0.21	0.003	0.38	0.001	0.47	0.001	0.88
	$\rho=0$[2]	0.003	-	0.003	-	0.001	-	0.001	-
	X-11	-	-	-	0.79	-	-	-	0.93

1 $\sigma_e^2 = E(\bar{Y}_t - \theta_t)^2$ is the variance of the aggregate mean.

2 Results obtained when ignoring the serial correlations between the panel estimators.

3 Estimators are based on past and current data.

4 Estimators are "smoothed" using all the data.

Table 2 : Simulation Results, Data Set II

Initial Components : $L_0 = 100,\ R_0 = 5,\ \underset{\sim}{S}_0' = (4, 1, -3, -2)$

Residual State Variances: $\sigma_L^2 = 0.8,\ \sigma_R^2 = 1,\ \sigma_S^2 = 0.4$

Survey Error Parameters: $\sigma_e^2 = 4$[1] ., $\rho = 0.7$

		100 Time Points				36 Time Points			
		Primary		Secondary		Primary		Secondary	
		Corr.Q	Est.Q	Corr.Q	Est.Q	Corr.Q	Est.Q	Corr.Q	Est.Q
Predic. Bias	panels	0.38	0.38	0.43	0.39	0.58	0.58	0.63	0.59
	aggregate	-0.29	-0.28	-0.34	-0.37	-0.08	-0.03	-0.06	0.00
Predic. MSE	panels	20.85	20.45	25.30	24.70	20.80	19.34	25.40	22.97
	aggregate	12.16	11.70	13.80	13.10	11.90	10.24	13.82	10.66
	$\rho=0$[2]	14.20	-	14.20	-	14.20	-	14.20	-
Est. of survey error parameters	σ_e^2 Bias	-	-0.12	-	-0.72	-	-0.06	-	-0.32
	σ_e^2 RMSE	-	0.36	-	2.80	-	0.76	-	4.88
	ρ Bias	-	-0.01	-	-0.24	-	-0.01	-	-0.26
	ρ RMSE	-	0.03	-	0.36	-	0.08	-	0.45
Var. of Est.[3] of population means	model	2.90	2.87	3.31	2.40	2.81	2.84	3.20	1.95
	realized	2.90	2.94	3.27	3.42	3.12	3.09	3.56	3.68
	patterson	3.16	3.16	-	-	3.30	3.30	-	-
MSE of Est.[4] of Seasonal Effects	model	0.52	0.51	0.69	0.68	0.54	0.45	0.72	0.55
	realized	0.53	0.63	0.61	0.75	0.51	0.65	0.67	0.87
	$\rho=0$[2]	0.70	-	0.70	-	0.71	-	0.71	-
	X-11	-	-	-	0.96	-	-	-	1.00

1 $\sigma_e^2 = E(\bar{Y}_t - \theta_t)^2$ is the variance of the aggregate mean.

2 Results obtained when ignoring the serial correlations between the panel estimators.

3 Estimators are based on past and current data.

4 Estimators are "smoothed" using all the data.

resulting from ignoring the extra variation due to estimating the model variances. A similar phenomenon can be observed with the variances of the estimators of the population means. Ansley and Kohn (1986) propose a correction factor of order $1/T^*$ to account for this extra variation in state space modelling.

3) The use of the full model taking into account the intrinsic relationships between the survey errors decreases the prediction errors very significantly compared to the case where these relationships are ignored (i.e. setting $\rho=0$ in the transition matrix). The same holds for the second group of data with respect to the estimation of the seasonal effects when using a primary analysis.

Although not shown in the table, we generated different data sets using smaller values of ρ but increasing each time the value of σ_v^2 so that the unconditional variance $\sigma_e^2 = E(\bar{Y}_t-\theta_t)^2$ remained fixed. Evidently, the smaller the value of ρ, the larger are the prediction errors under both a primary and a secondary analysis (although the differences between the two analyses are diminished) and the smaller is the impact of setting $\rho=0$ in the analysis. Decreasing the value of ρ increases also the MSE's of the estimates of the seasonal effects under both a primary and a secondary analysis using either the correct Q or its sample estimate. The increase is again very evident. Thus, the major factor determining the efficiency of the model is the residual variance, σ_v^2, and not the unconditional variance of the survey errors which is quite intuitive considering the imposed autoregressive relationships between the survey errors.

4) The use of the model yields more accurate estimates for the population means than does the classical sampling approach. (For the first group, the model performs better even under a secondary analysis.) The major factor affecting the performance of the classical estimator is the variance of the survey errors and since it is twice as large for the first group of data as for the second group, the variance of the estimator is likewise doubled. Under the model, the variance of the estimators depends also on the residual variances of the population mean components and since they are much smaller for the first group than for the second, the overall increase in the variance of the estimators is only in the magnitude of about 30%. These results indicate very clearly the possible advantages of modelling the evolution of the population means over time.

5) The use of the model yields in general better predictors for the seasonal effects than does the X-11 procedure. The much better performance of the model in the case of a primary analysis may not surprise but it is another indication for the possible gains in using the separate panel estimates taking the design features into account. The superiority of the model in the case of a secondary analysis is less obvious despite the exploitment of the model assumptions since almost all the series analyzed were found to be in the acceptance regions of the "Monitoring and Quality Assessment Statistics" calculated by the X-11 program. It should be noted however that the two procedures yield closer estimates once the survey error variance is decreased.

6.2 Empirical Results Using Labor Force Data

We present the results obtained for two series:

Series 1: Number of hours worked in the week preceding the survey

Series 2: Number of weeks worked in the year preceding the survey

In order not to burden the computations, we restricted the analysis to households in the city of Tel-Aviv. The time period covered was 1979-1987 so that each series consists of 36 x 4 panel estimates. We didn't include data for the years before 1979 because of changes in the sampling design and the questionnaire introduced in 1978. Data for 1988 was not available to us at the time of the analysis. Interested readers may obtain the data of the two series from the author.

The trend levels of these two series are almost constant. The seasonal effects account for about 50 percent of the MSE of the quarter to quarter differences in the aggregate means in the case of the first series and for about 30 percent in the case of the second series

The results obtained when fitting the models of sections 3 and 4 to the series are exhibited in table 3. The column headed "primrot" gives the results obtained when accounting for rotation group bias. We first analysed the series without including the bias factors (other columns of the table) and found that for both series the predictor one step ahead of the first panel estimator is essentially unbiased. This result suggests that $\beta_1 \doteq (\beta_2+\beta_4+\beta_5)/3$ but since there is no apparent reason for a bias associated with the first panel (see the discussion below) we presupposed also that $\beta_1 = 0$ which, together with the previous relationship implies that $\Sigma_j\beta_j = 0$. Nonetheless, when estimating the bias coefficients we only imposed the milder condition $\Sigma_j \beta_j = 0$. (Starting values for the coefficients were set to zero with correspondingly large error variances).

Comparing the panel prediction biases with and without the accounting for the group effects reveals relatively large biases for the other panels in the latter case. For the first series the smoothed estimators of the bias coefficients and the corresponding standard deviations (in brackets) are: $\hat{\beta}_1 = -.027(.17)$, $\hat{\beta}_2 = .26(.17)$, $\hat{\beta}_4 = .13(.17)$, $\hat{\beta}_5 = .-37(.17)$. Thus $\hat{\beta}_5$ is significant at the .03 level and likewise with respect to the difference $(\hat{\beta}_5 - \hat{\beta}_2)$ (the S.D. of the difference is .29). For the second series the corresponding values are $\hat{\beta}_1 = -.10(.16)$, $\hat{\beta}_2 = .20(.16)$, $\hat{\beta}_4 = .09(.16)$, $\hat{\beta}_5 = -.19(.16)$ so that none of the coefficients is significant even though they exhibit a similar pattern to that observed for the first series.

The fact that only one group effect came out significant may result from the short length and the relatively large error variances of the two seires. Notwithstanding, it is likewise not clear that the observed prediction biases reflect real rotation group effects. It was suggested to us that the negative effect observed for the forth panel could result from the fact

Table 3: Empirical Results, Labour Force Survey, Israel, 1979-1987

Series I : Hours Worked in the Week Preceding the Survey
Series II: Weeks Worked in the Year Preceding the Survey

	Series I			Series II		
	Primary	Primrot[1]	Secondary	Primary	Primrot[1]	Secondary
Prediction Bias						
panel 1	0.07	0.16	0.05	0.02	0.16	0.02
panel 2	0.36	0.08	0.34	0.27	-0.02	0.23
panel 3	0.27	0.10	0.26	0.17	0.07	0.18
panel 4	-0.31	0.11	-0.21	-0.20	0.04	-0.19
aggregate	0.10	0.10	0.11	0.07	0.06	0.06
Prediction MSE						
panels (average)	1.63	1.56	1.71	1.16	1.14	1.18
aggregate	0.51	0.50	0.48	0.21	0.21	0.24
$\rho=0$[2]	1.21	1.36	1.35	0.51	0.52	0.50
$\hat{Y}_{t\mid t-1}=Y_{t-1}$	1.47	1.47	1.47	0.41	0.41	0.41
Est. Survey. Param.						
σ^2_e	0.42	0.38	0.69	0.38	0.36	0.37
ρ	0.39	0.42	0.07	0.36	0.38	0.46
Var. Est. Pop. Means						
Model	0.15	0.14	0.20	0.10	0.11	0.10
Patterson	0.39	0.39	-	0.31	0.31	-

[1] Results obtained when accounting for rotation group effects.

[2] Results obtained when ignoring the serial correlations between the panel estimators.

that this is the only panel surveyed also on income. Another noteworthy feature of the ILFS is that about half of the interviews of the second and third panels are carried out by telephone. Thus, although the results of our analysis are inconclusive at this stage they are indicative enough to justify a more rigorous and comprehensive study using more series and if possible more detailed data.

Two other notable outcomes in talbe 3 are: i) The use of primary and secondary analysis gives consistent results. This is true in particular for the second series but holds also for the first series except for the estimation of the survey error parameters and hence the estimation of the variance of the population mean estimator ii) The estimates of the variance of the model dependent estimators of the population means are substantially smaller than the estimated variances of the sampling (Patterson) estimator. One needs to be cautious in comparing the two sets of variances since the former are model dependent and employ estimates for all the unknown model parameters. However the substantial reduction in the prediction MSE's under the model as compared to the case where the correlations between the survey errors are ignored or when predicting the aggregate means by the means observed in the previous periods as well as the other results discussed above make us believe that the difference in the variances is real and not just the result of model misspecification or sampling errors. (see also the concluding remarks). What seems to make the difference between the model and the sampling estimator is the fact that these series exhibit almost constant trend levels and very stable seasonal components which of course is ideal for the use of model based predictors. (See the discussion in point 4 of section 6.1).

7. CONCLUDING REMARKS

The results obtained in the empirical study illustrate the possible advantages of using a primary analysis as compared to the use of a secondary analysis. First, the use of a primary analysis yields more accurate estimates for the model parameters and in particular the survey error parameters and second, it produces better predictors for the population means and the seasonal effects. Practitioners in the survey sampling area often prefer the use of a secondary analysis because of rotation group effects but as we have illustrated, these effects can be incorporated in the model.

The use of the model for seasonal adjustment problems is an important aspect which should be further explored. Our study indicates the potential advantages of using a primary analysis taking into account the correlations between the observed estimates. It is unfortunate that despite the increasing use of repeated surveys by statistical bureaus and the widely recognized need for producing accurate estimates for the seasonal effects, the special structure of the data is generally ignored when producing such estimates. In this respect, the procedure proposed in the present article is a first attempt to use more data and take the design features into account when estimating the seasonal effects which hopefully will encourage further research.

ACKNOWLEDGEMENT

Work on this paper was completed while I was staying at Statistics Canada under a Research Fellowship Program. Particular thanks are due to D. Friedman for his assistance in an early version of this paper and to B. Quenneville for helping with the computations. I would like to thank also the associate editor and two referees for very helpful comments.

REFERENCES

Anderson, B.D.O. and Moore, J.B. (1979) Optimal Filtering. Prentice-Hall, Englewood Cliffs, N.J.

Ansley, C.F. and Kohn, R. (1985) Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. The Annals of Statistics, 13, 1286-1316.

Ansley, C.F. and Kohn, R. (1986) Prediction mean squared error for state space models with estimated parameters. Biometrika, 73, 467-473.

Bailar, B.A. (1975) The effects of rotation group bias on estimates from panel surveys. Journal of the American Statistical Association, 70, 23-29.

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

Binder, D.A. and Hidiroglou, M.A. (1988) Sampling in Time. Handbook of Statistics, Vol. 6 (P.R. Krishnaiah and C.R. Rao, eds.), Amsterdam: Elsevier Science, pp. 187-211.

Blight, B.J.N. and Scott, A.J. (1973) A stochastic model for repeated surveys. Journal of the Royal Statistical Society, B, 35, 61-68.

Cochran, W.G. (1977) Sampling Techniques (Third Edition), New York: Wiley.

Dagum, E.B. (1980) The X-11 ARIMA seasonal adjustment method. Catalogue 12-564E, Statistics Canada, Ottawa, Ont, K1A 0T6.

Dagum, E.B. and Quenneville, B. (1988) Deterministic and stochastic models for the estimation of trading-day variations. Working Paper No. TSRAD-88-003E, Statistics Canada, Ottawa, Ont, K1A 0T6.

DeJong, P. (1988) The likelihood of a state-space model. Biometrika, 75, 165-169.

Gurney, M. and Daly, J.F. (1965) A multivariate approach to estimation in periodic sample surveys. Proceedings of the Social Statistics Section, American Statistical Association, 242-257.

Harrison, P.J. and Stevens, C.F. (1976) Bayesian forecasting (with discussion). Journal of the Royal Statistical Society, B, 38, 205-247.

Harvey, A.C. (1984) A unified view of statistical forecasting procedures. Journal of Forecasting 3, 245-275.

Harvey, A.C. and Todd, P.H.J. (1983) Forecasting economic time series with structural and Box-Jenkins Models (with discussion), Journal of Business and Economic Statistics, 1, 299-315.

Harvey, A.C. and Peters, S. (1984) Estimation procedures for structural time series models. Dememic Econometrics Programmes, Discussion Paper A.44, London School of Economics, London WC2A 2AE.

Hausman, J.A. and Watson, M.W. (1985) Errors in variables and seasonal adjustment procedures. Journal of the American Statistical Association, 80, 531-540.

Jessen, R.J. (1942) Statistical Investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experimental Station, Research Bulletin 304, 54-59.

Jones, R.G. (1979) The efficiency of time series estimators for repeated survey. Australian Journal of Statistics, 21, 1-12.

Jones, R.G. (1980) Best linear unbiased estimators for repeated survey. Journal of the Royal Statistical Society, B, 42, 221-226.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82 D, 33-45.

Kumar, S. and Lee, H. (1983). Evaluation of composite estimation for the Canadian labour foce survey. Survey Methodology, 9, 1-24.

Maravall, A. (1985) On structural time series models and the characterization of components. Journal of Business and Economic Statistics, 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, 255-268.

Nathan, G. and Elihav, T. (1988) Comparison of Measurement Errors in Telephone Interviewing and Home Visits by Misclassification Models. Journal of Official Statistics, 4, 363-374.

Patterson, H.D. (1950) Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society, B, 12, 241-255.

Quenneville, B. (1988) DLM: Software for dynamic linear models. Time Series Research and Analysis Division, Statistics Canada, Ottawa, Ont, K1A 0T6.

Quenneville, B. and Dagum, E.G. (1988) Variance of X-11 ARIMA estimates - A structural approach. Working Paper TSRAD-88-00, Statistics Canada, Ottawa, Ont, K1A 0T6.

Scott, A.J. and Smith, T.M.F. (1974) Analysis of repeated surveys using time series models. 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.

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

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

Tam, S.M. (1987) Analysis of repeated surveys using a dynamic linear model. International Statistical Review, 55, 63-73.

Watson, M.W. and Engle, R.F. (1983) Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. Journal of Econometrics, 23, 385-400.