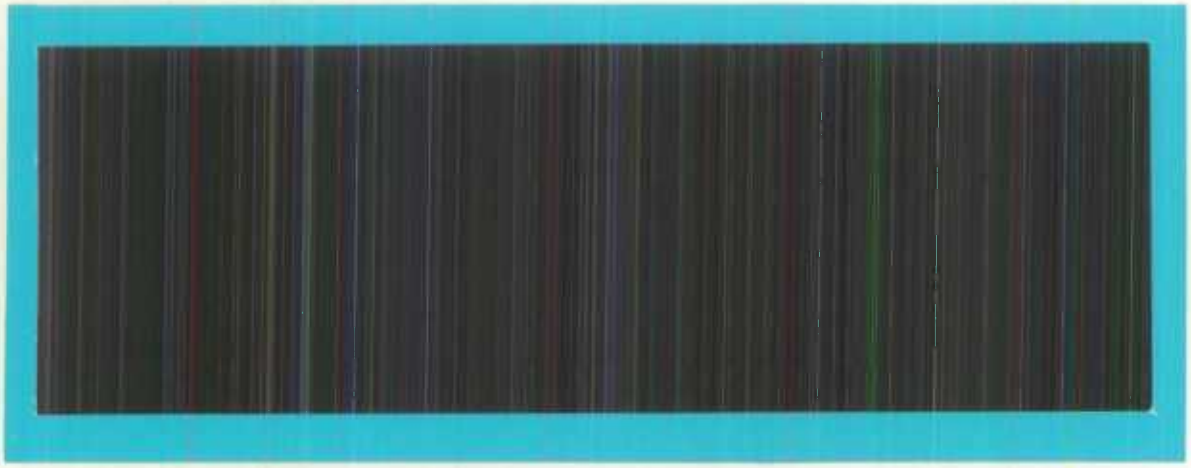
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

11-613
no. 92-02
c. 2

Canada

WORKING PAPER NO. SSMD 92-002 E

METHODOLOGY BRANCH

C.2

ROBUST JOINT MODELLING OF CANADIAN LABOUR FORCE SERIES OF SMALL AREAS

D. PFEFFERMANN AND S.R. BLEUER

STATISTICS CANADA STATISTIQUE CANADA

APR 1 1992

LIBRARY BIBLIOTHÈQUE

January 1992

STATISTICS CANADA
LIBRARY
BIBLIOTHÈQUE

ROBUST JOINT MODELLING OF CANADIAN LABOUR FORCE SERIES OF SMALL AREAS

D. Pfeffermann and S.R. Bleuer[1]

ABSTRACT

In this article we report the results of fitting a state-space model to Canadian unemployment rates. The model assumes an additive decomposition of the population values into a trend, seasonal and irregular component and separate autoregressive relationships for the six survey error series corresponding to the six monthly panel estimators. The model includes rotation group effects and permits the design variances of the survey errors to change over time. The model is fitted at the small area level but it accounts for correlations between the component series of different areas. The robustness of estimators obtained under the model is achieved by imposing the constraint that the monthly aggregate model based estimators in a group of small areas for which the total sample size is sufficiently large coincide with the corresponding direct survey estimators. The performance of the model when fitted to the Atlantic provinces is assessed by a variety of diagnostic statistics and residual plots and by comparisons with estimators in current use.

KEY WORDS: Design variance; Kalman filter; Panel survey; Rotation bias; State-space model.

RÉSUMÉ

Dans cet article, nous présentons les résultats de l'ajustement de modèle d'espace d'états aux taux de chômage du Canada. Le modèle suppose une décomposition additive des valeurs de la population en une composante tendancielle, saisonnière et irrégulière et des relations autorégressives distinctes pour les six séries d'erreurs d'enquête correspondant aux six estimateurs de panel mensuels. Le modèle inclut les effets des groupes de rotation et permet aux variances du plan de sondage des erreurs d'enquête de varier dans le temps. Le modèle est ajusté au niveau de la région, mais il prend en compte les corrélations entre les séries composantes de différentes régions. On obtient des estimateurs robustes en posant la contrainte selon laquelle les estimateurs de modèle agrégés mensuels dans un groupe de régions dont la taille d'échantillon est suffisamment importante coïncident avec les estimateurs d'enquête directs. Les résultats du modèle, ajusté pour les provinces de l'Atlantique, sont évalués grâce à une série de statistiques de diagnostic et des tracés résiduels ainsi que par des comparaisons avec des estimateurs couramment utilisés.

MOTS-CLÉS: variance selon le plan, filtre de Kalman, enquête avec panel, biai dû à la rotation, modèle d'espace d'états,

[1] D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905, S.R. Bleuer, Social Surveys Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

1. INTRODUCTION

A time series model for survey data is the combination of two distinct models. The "census model" describing the evolution of the finite population values over time and the survey errors model representing the time series relationships between the survey errors of the survey estimators. There are at least four main reasons for wishing to model the raw survey estimators:

A- The model based estimators of the population values resulting from the modelling process have in general smaller variances than the survey estimators, particularly in small areas where the sample sizes are small.

B- The model we employ yields estimators for the seasonal effects and for the variances of these estimators as a by-product of the estimation process.

C- The model can be used to forecast the population values, the trend and the seasonal components for time periods beyond the sample time period for which the direct survey estimators are available. Such forecasts are important when assessing the performance of the model and for policy decision making.

D- The model can be used to detect turning points in the level of the series and assess their significance. (Work on this problem will be addressed in a separate article).

The methodology described in this article integrates the methodologies presented in Pfeffermann and Burck (1990) and Pfeffermann (1991) with some new modifications and extensions. The main features of the model are as follows:

1) The model decomposes the population values into the unobservable components of trend, seasonality and irregular terms. Smoothed predictors of the three components (and hence of the population values) based on all the available data, and standard errors of the prediction errors are obtained straightforwardly by application of the Kalman filter. The standard errors are modified to account for the extra variation induced by the use of estimated parameter values.

2) The model uses the distinct monthly panel estimators as input data. The use of the panel estimators has two important advantages over the use of the mean estimators: i) It identifies better the time series model holding for the survey errors by analyzing contrasts between the panel estimators, ii) It yields more efficient estimators for the model parameters and hence better predictors for the unobservable model components.

3) The model accounts for changes in the variances of the survey errors over time and for possible rotation group effects.

4) The model can be applied simultaneously to the panel estimators in separate small areas. The census model is extended in this case to account for the cross-correlations between the unobservable components of the population values operating in these areas.

5) A modification to ensure the robustness of the small area estimators against possible model breakdowns is incorporated into the model equations. The modification consists of constraining the model based estimators of aggregates of the population values over a group of small areas for which the total sample size is sufficiently large to coincide with the corresponding aggregate survey estimators. As a result, sudden changes in the level of the series are reflected in the model based estimators with no time lag.

The model and the robustness modifications are described in more detail in section 2. Empirical results obtained when fitting the model to the four Atlantic provinces of Canada are presented in section 3. Section 4 contains a short summary with suggestions for extension of the analysis.

2. A STATE-SPACE MODEL FOR CANADA UNEMPLOYMENT SERIES

2.1 The Canadian Labour Force Survey

Data on unemployment are collected as part of the labour force survey (LFS) carried out by Statistics Canada. The Canadian LFS is a rotating monthly panel survey by which every new sampled panel of households is retained in the sample for six successive months before being replaced by another panel from the same PSU or stratum. The PSU's are defined by geographic locations (city blocks or urban centers in the urban regions and enumeration areas in the rural regions). The strata are homogeneous groups of clusters with respect to geographic location and social economic variables. Every PSU is represented in only one rotation group. The separate panel estimators are assumed to be independent for variance estimation purposes, a property validated in past studies. For a recent report describing the design of the LFS and the construction of the direct survey estimators, the reader is referred to Singh *et al.* (1990).

2.2 The Census Model

In what follows we consider a single small area. In section 2.4 we consider joint modelling of the panel estimates in a group of small areas. The model postulated for the population values is the Basic Structural Model (BSM) which consists of the following sets of equations.

$$Y_t = L_t + S_t + \epsilon_t;\ L_t = L_{t-1} + R_{t-1} + \eta_{Lt};\ R_t = R_{t-1} + \eta_{Rt};\ \sum_{j=0}^{11} S_{t-j} = \eta_{St} \qquad (2.1)$$

In (2.1) Y_t is the population value ("true" unemployment rate) at time t, L_t is the trend level, R_t is the increment, S_t the seasonal effect and ϵ_t the irregular term assumed to be white noise with zero mean and variance σ_ϵ^2. Thus, the equation in the left-hand side of (2.1) postulates the classical decomposition of a time series into a trend, seasonal and irregular components. This decomposition is inherent in the commonly used procedures for seasonal adjustment, see e.g. Dagum (1980). Notice however that in the present case the series $\{Y_t\}$ is itself unobservable. The series $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ are independent white noise disturbances with mean zero and variances σ_L^2, σ_R^2 and σ_S^2 respectively. Hence, the second and third equations of (2.1) define a local approximation to a linear trend whereas the equation in the right-hand side models the evolution of the seasonal effect such that the sum of every 12 successive effects fluctuates around zero.

The theoretical properties of the BSM in relation to other models are discussed in Harrison and Stevens (1976), Harvey (1984) and Maravall (1985). Empirical results illustrating the performance of the model are shown in Harvey and Todd (1983), Morris and Pfeffermann (1984) and Pfeffermann (1991). Although more restricted than the family of ARIMA models, the BSM is now recognised as being flexible enough to approximate the behaviour of many diverse time series.

2.3 The survey errors model

The model holding for the survey errors was identified initially by analyzing separately the pseudo errors series $e_{tp}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, t=1...N, where $y_t^{(j)}$ is the estimator of Y_t based on the j-t panel, j = 1...6, (the panel surveyed for the j-th successive month) and $\bar{y}_t = \sum_{j=1}^{6} y_t^{(j)}/6$ is the mean estimator. Notice that $(y_t^{(j)} - \bar{y}_t) = (e_t^{(j)} - \sum_{j=1}^{6} e_t^{(j)}/6)$, where $e_t^{(j)} = (y_t^{(j)} - Y_t)$ are the true survey errors. Thus, the notable feature of the contrasts $(y_t^{(j)} - \bar{y}_t)$ is that they are functions of only the survey errors irrespective of the model holding for the population values.

There are two major considerations in the choice of a model for the survey errors

a) The model should account for possible rotation group biases or more generally, allow for different survey error means in different panels

b) The model should account for changes in the variances of the survey errors over time.

Rotation group biases may arise from providing different information on different rounds of interview, depending on the length of time that respondents are included in the sample, or on the method of data collection, say, whether by telephone or by home interview. (In the Canadian LFS, the first panel is interviewed by home visits, the other panels are interviewed by telephone). Another possible reason for differences between the panel survey error means is differences in the nonresponse patterns across the panels. See Pfeffermann (1991) for further discussion with references to earlier studies on this problem.

Changes in the variances of the survey errors over time occur when the variances are function of the level of the series. Indeed, as revealed by figure 1 in section 3, the estimates of the standard deviations of the survey errors are subject to seasonal effects with a seasonal pattern that follows the seasonal pattern of the population values. Another possible explanation for changes in the variances of the survey errors is changes in the sampling design. For example, the overall sample size of the Canadian LFS was reduced in 1985-86 from 55.000 households to 48.000 households. This reduction in the sample size was associated with other changes in the design. See Singh *et al.* (1990) for details.

Application of simple model estimation and diagnostic procedures to the pseudo survey errors suggest a 3rd order autoregressive (AR) model for the standardized survey errors $\tilde{e}_t^{(j)} = (e_t^{(j)} - \beta_j) / SD(e_t^{(j)})$, i.e.

$$\tilde{e}_t^{(j)} = \phi_{j1}\, \tilde{e}_{t-1}^{(j-1)} + \phi_{j2}\, \tilde{e}_{t-2}^{(j-2)} + \phi_{j3}\, \tilde{e}_{t-3}^{(j-3)} + u_t^{(j)},\ j = 1 \ldots 6 \qquad (2.2)$$

where $\beta_j = E(e_t^{(j)})$ are the rotation group biases, $SD(e_t^{(j)})$ are the design standard deviations and $u_t^{(j)}$ are independent white noise series with mean zero and variances σ_j^2. It is assumed that $\sum_{j=1}^{6} \beta_j = 0$ which implies that the mean survey estimator, $\bar{y}_t$, is unbiased. See Pfeffermann (1991) for discussion on the need to constraint the bias coefficients. Subsequent analysis when fitting the combined model defined by (2.1) and (2.2) (see section 2.4) validates this model with the further observation that the coefficients $(\phi_{j1}, \phi_{j2}, \phi_{j3})$ can be assumed to be equal for $j = 4,5,6$. Furthermore, for the first panel an AR(1) model already gives

a good fit whereas for the second and third panel an AR(2) model is appropriate although with different coefficients. These relationships hold for each of the four Atlantic provinces.

The model defined by (2.2) satisfies the two prior considerations discussed above. In particular, $\mathrm{VAR}(u_t^{(j)}) = \sigma_j^2 \mathrm{VAR}(e_t^{(j)})$. The actual application of the model raises however two questions:

1. For the first three panels there is not enough history to permit the fitting of an AR(3) model. For example, the survey error $e_t^{(1)}$ corresponds to the panel which is in the sample for the first month. In order to overcome this problem, we replace the missing survey errors by the survey errors corresponding to the panels previously selected from the same cluster or stratum. For example, the AR(2) model fitted to $e_t^{(2)}$ is

$$\tilde{e}_t^{(2)} = \phi_{21}\, \tilde{e}_{t-1}^{(1)} + \phi_{22}\, \tilde{e}_{t-2}^{(6)} + u_t^{(2)} \tag{2.3}$$

   Notice that the panel surveyed for the second time at month t replaces at time $(t-1)$ the panel observed for the sixth time at month $(t-2)$ so that both panels represent the same cluster or stratum. The use of surrogate survey errors in the case of the first three panels may explain the different models identified for these panels as compared to the model identified for the other three panels.

2. The true standard deviations of the survey errors are unknown whereas the survey estimates of the standard deviations are themselves subject to sampling errors. To overcome this problem, we use smoothed values of the estimated standard deviations, obtained by fitting the relationship

$$(\tilde{SD})_t = \hat{\gamma}\,(\hat{SD})_{t-1} + \hat{\gamma}_0 t + \sum_{i=1}^{11} \hat{\gamma}_i D_{it} \tag{2.4}$$

   with the γ-coefficients estimated by ordinary least squares. The notation $(\hat{SD})_t$ defines the raw, unsmoothed design standard deviation of the aggregate survey estimator at month t and $\{D_{it}\}$ are dummy variables accounting for monthly seasonal effects so that $D_{it} = 1$ when $t=12k+i$, $k=0,1,\ldots$ and $D_{it}=0$ otherwise. The smoothed standard deviations of the panel survey errors are obtained as $\tilde{SD}(e_t^{(j)}) = \sqrt{6}(\tilde{SD})_t$. The latter estimates are used as surrogates for the true, unknown, standard deviations.

2.4 State-space representation and estimation of the model holding for the survey estimators

It follows from (2.1) that the panel estimators can be modeled as

$$y_t^{(j)} = L_t + S_t + \epsilon_t + e_t^{(j)}, \qquad j=1...6 \tag{2.5}$$

where

$$L_t + L_{t-1} + R_{t-1} + \eta_{Lt};\ R_t = R_{t-1} + \eta_{Rt};\ \sum_{j=0}^{11} S_{t-j} = \eta_{St} \tag{2.6}$$

with $\{\epsilon_t\}$, $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ defined as in (2.1). The separate models defined by (2.5), (2.6) and (2.2) can be cast into a compact state-space representation with $\underline{y}_t' = (y_t^{(1)}...y_t^{(6)})$ as the input data, similarly to the representation in Pfeffermann (1991). Following that representation, the survey errors (and in the present study also the census irregular term) are included as part of the state vector so that there are no residual terms in the observation equation defined by (2.5). Unlike in Pfeffermann (1991), however, the transition matrix and the Variance-Covariance (V-C) matrix of the state error terms are not fixed in time since they depend on the design variances of the survey errors which, as explained in section 2.3, change over time.

The state-space representation of the model permits to update, smooth or predict the state vectors and hence the seasonal, trend and population values at any given time t by means of the Kalman filter. Denote by $\underline{\alpha}_t$ the state vector corresponding to time t. The state vector comprises the trend level, increment and seasonal effects, the rotation group effects and the survey errors. See Pfeffermann (1991) for details. By "updating" we mean estimation of $\underline{\alpha}_t$ at time t based on all the data until and including time t. "Smoothing" refers to the estimation of $\underline{\alpha}_t$ based on all the available data for all the months before and after time t. Smoothing is required for improving past estimates as, for example, when estimating the seasonal effects or when estimating changes in the population values or the trend levels. "Prediction" of state vectors corresponding to postsample months is important for policy making. Predictions within the sample period allow to assess the performance of the model, e.g. by comparing the forecasted panel estimates as derived from the predicted state vectors with the actual estimates. See section 3 for details. The theory of state-space models and the Kalman filter is developed in numerous publications, see Pfeffermann (1991) for the filtering and smoothing equations with references. Notice that the filtering and the smoothing equations not only yield the three sets of estimators at any given time t but also the V-C matrices of the corresponding estimation errors.

The actual application of the Kalman filter requires the estimation of the unknown model parameters and the initialization of the filter, that is, the estimation of the initial state vector $\underset{\sim}{\alpha}_0$ and the corresponding V-C matrix of the estimation errors. For a single small area, the unknown model parameters are the four variances of the error terms in the census model (2.1) and the eight autoregression coefficients and six residual variances in the panel survey errors models (2.2). (The rotation group means are included in the state vectors as fixed, time invariant coefficients). In order to reduce the number of free parameters in the combined state-space model, we assume $\sigma_j^2 = \sigma^2 \times \tilde{\sigma}_j^2$ j=1...6, where $\{\sigma_j^2\}$ are the residual variances in (2.2) and $\tilde{\sigma}_j^2$ are the estimates of the residual variances obtained by fitting the autoregression equations to the pseudo survey errors $e_{tp}^{(i)}$ defined in section 2.3. This assumption reduces the number of unknown parameters from 18 to 13. (The estimates $\tilde{\sigma}_j^2$ are very close for j=4,5,6 and have been set equal).

Assuming that the error terms in the census and survey error models have a normal distribution, the unknown model parameters can be estimated by maximization of the likelihood. See Pfeffermann and Burck (1991) for a brief description of the method of scoring maximization algorithm and for the initialization of the filter. This article includes references to more rigorous discussions.

2.5 Adjustments to account for the use of estimated parameter values

Once the unknown model parameters have been estimated, the Kalman filter equations can be applied with the true parameter values replaced by the parameter estimates. As noted in section 2.4, the Kalman filter not only produces estimates for the state vectors but also the V-C matrices of the corresponding estimation errors. A possible problem arising from the use of these V-C matrices, however, is that they ignore the extra variation implied by parameter estimation, thus resulting in underestimation of the true variances.

Formally, let $\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})$ define the estimator of $\underset{\sim}{\alpha}_t$ at time t based on all the data available until some given time n, where $\hat{\underset{\sim}{\lambda}}$ represents the estimators of the unknown model parameters. The estimation error can be decomposed as

$$[\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \underline{\alpha}_t] = [\hat{\underline{\alpha}}_t(\underline{\lambda}) - \underline{\alpha}_t] + [\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \hat{\underline{\alpha}}_t(\underline{\lambda})] \tag{2.7}$$

which is the sum of the error if $\underline{\lambda}$ were known plus the error due to estimation of $\underline{\lambda}$. The two terms in the right-hand side of (2.7) are independent. A simple way to verify this property is by noting that $\hat{\underline{\alpha}}_t(\underline{\lambda}) = E(\underline{\alpha}_t | Y, \underline{\lambda})$ where Y represents all the available data. By conditioning on Y and $\underline{\lambda}$, $[\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \hat{\underline{\alpha}}_t(\underline{\lambda})]$ is nonstochastic whereas $E\{[\hat{\underline{\alpha}}_t(\underline{\lambda}) - \underline{\alpha}_t] | Y\} = \underline{0}$. It follows therefore from (2.7) that

$$\begin{aligned} Q_t &= E\{[\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \underline{\alpha}_t][\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \underline{\alpha}_t]'\} = \\ &= E\{[\hat{\underline{\alpha}}_t(\underline{\lambda}) - \underline{\alpha}_t][\hat{\underline{\alpha}}_t(\underline{\lambda}) - \underline{\alpha}_t]'\} + E\{[\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \hat{\underline{\alpha}}_t(\underline{\lambda})][\hat{\underline{\alpha}}_t(\hat{\underline{\lambda}}) - \hat{\underline{\alpha}}_t(\underline{\lambda})]'\} = P_t + R_t \end{aligned} \tag{2.8}$$

The V-C matrix P_t is the matrix produced by the Kalman filter for a given vector parameter $\underline{\lambda}$ and it can be estimated by replacing $\underline{\lambda}$ by $\hat{\underline{\lambda}}$. (This is the V-C matrix obtained by ignoring the variation resulting from parameter estimation). The V-C matrix R_t measures the variation due to parameter uncertainty.

In order to estimate R_t, we follow the approach proposed by Hamilton (1986). By this approach, realizations $\underline{\lambda}_{(k)}$, $k = 1 \ldots K$ are generated from the asymptotic normal posterior distribution of $\underline{\lambda}$, that is, from a $N(\hat{\underline{\lambda}}, \hat{\Lambda})$ distribution where $\hat{\underline{\lambda}}$ is the maximum likelihood estimator of $\underline{\lambda}$ and $\hat{\Lambda}$ is the asymptotic V-C matrix of $\hat{\underline{\lambda}}$. (Both $\hat{\underline{\lambda}}$ and $\hat{\Lambda}$ are obtained by the method of scoring). The Kalman filter is then applied with each of these realizations yielding estimates $\hat{\underline{\alpha}}_t(\underline{\lambda}_{(k)})$. The matrix R_t is estimated as

$$\hat{R}_t = \frac{1}{K} \sum_{k=1}^{K} [\hat{\underline{\alpha}}_t(\underline{\lambda}_{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \hat{\underline{\alpha}}_t(\underline{\lambda}_{(k)})][\hat{\underline{\alpha}}_t(\underline{\lambda}_{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \hat{\underline{\alpha}}_t(\underline{\lambda}_{(k)})]' \tag{2.9}$$

Ansley and Kohn (1986) propose an estimator for R_t based on first order Taylor series approximation. The use of their estimator is computationally less intensive but the procedure proposed by Hamilton is somewhat more flexible in terms of the assumptions involved and it enables a better insight into the sensitivity of the Kalman filter output to errors in the parameter estimators.

2.6 Joint modelling in several small areas

The model considered so far refers to a single area. When the sample sizes in the various areas are small, more efficient estimators can often be derived by modelling in addition the cross-sectional

relationships between the area population values. Clearly, the increase in efficiency resulting from such joint modelling depends on the sample sizes within the small areas and the closeness of the behaviours of the area population values over time.

The survey errors are independent between the areas so that any joint modelling of the survey estimators applies only to the census model. For modelling the unemployment rates in the four Atlantic provinces, we follow Pfeffermann and Burck (1990) and allow for nonzero contemporary correlations between corresponding error terms of the census models operating in these provinces. Thus, if $\underset{\sim}{\nu}'_{t,a} = (\epsilon_t^{(a)}, \eta_{Lt}^{(a)}, \eta_{At}^{(a)}, \eta_{St}^{(a)})$ denotes the vector of error terms at time t associated with the census model operating in area a, it is assumed that $C_{a,b} = E(\underset{\sim}{\nu}_{ta}\underset{\sim}{\nu}'_{tb})$ is diagonal but with possibly non zero covariances on the main diagonal. The actual implication of this assumption is that if, for example, there is a significant increase in the trend level in one province, similar increases can be expected to occur in other provinces.

The resulting joint model holding for the four provinces (or more generally for a group of areas) can again be casted into a state-space form, see equations (2.7) and (2.8) in Pfeffermann and Burck (1990). A major problem with fitting this model, however, is the joint estimation of all the unknown parameters which is computationally too intensive in terms of computer time and storage space. (The computer program written for this study uses numerical derivatives so that each iteration of the method of scoring used for maximization of the likelihood requires a separate sweep through all the data. Each sweep involves the computation of the Kalman filter equations for each month included in the sample period.)

To deal with this problem, we first fitted the models defined by (2.5), (2.6) and (2.2) separately for each of the provinces. We also postulated equal correlations across the provinces between the corresponding error terms of the separate census models so that

$$\Phi_{a,b} = C_{a,a}^{-\frac{1}{2}} C_{a,b} C_{b,b}^{-\frac{1}{2}} = \Phi \quad 1 \le a,b \le 4 \tag{2.10}$$

where $C_{a,a} = E(\underset{\sim}{\nu}_{ta}\underset{\sim}{\nu}'_{ta})$. The four correlations maximizing the likelihood of the joint model were determined by a grid search procedure with the other model parameters held fixed at their previously estimated values.

The assumption of equal correlations reduces the number of unknown parameters considerably. It can be justified also by the small number of areas considered for this study implying that no other pre-imposed structure on these correlations can be safely detected. More substantively, a simple breakdown of the labour force by industry (table 1 of Section 3) shows very similar relative frequencies in the four provinces suggesting a high degree of homogeneity in their economies.

2.7 Modifications to protect against model failures

The use of a model for the production of official statistics raises the question of how to protect against possible model failures. Testing the model every time that new data become available is not practical, requiring instead the development of a built-in mechanism to ensure the robustness of the estimators when the model fails to hold.

For modelling the labour force series in small areas we employed the modification proposed by Pfeffermann and Burck (1990). By this modification, the updated state vector estimates at any given time t, are constraint to satisfy the condition

$$\sum_{a=1}^{A} w_{ta}\hat{Y}_{ta} = \sum_{a=1}^{A} w_{ta}\bar{y}_{ta} \qquad t=1,2,... \tag{2.11}$$

where $\hat{Y}_{ta}$ is the model based estimator of the population value Y_{ta} in area a, $\bar{y}_{ta} = \frac{1}{6}\sum_{j=1}^{6} y_{ta}^{(j)}$ is the corresponding survey estimator and $w_{ta} = M_{ta} / M_t$ is the relative size of the labour force in that area so that $M_t = \sum_{a=1}^{4} M_{ta}$ and $\sum_{a=1}^{4} w_{ta} = 1$. Notice that $\sum_{a=1}^{A} w_{ta}\hat{Y}_{ta}$ and $\sum_{a=1}^{A} w_{ta}\bar{y}_t$ are correspondingly the model based and direct survey estimators of the aggregate population value in the group of areas considered.

The rationale behind the modification is simple. It assumes that the total sample size in all the areas is sufficiently large and hence that the aggregate survey estimators can be trusted. (This condition in fact dictates the level of aggregation required). By constraining the aggregate model based estimators to coincide with the aggregate survey estimators, the analyst ensures that any abrupt change in the population values reflected in the survey estimators will be likewise reflected in the model based estimators. Notice that without constraining the estimators, sudden changes in the level of the series, for example, will be reflected in the model based estimators only after several months because these estimators depend not only on current data but also on past data. On the other hand, if no substantial changes occur, the model based estimators can be expected to satisfy approximately the constraints even without imposing them explicitly. Thus, the constrainted estimators perform almost as well as the unconstrainted estimators in regular time periods.

The use of constraints of the form (2.11) was previously considered by Battese, Harter and Fuller (1988) and by Pfeffermann and Barnard (1991) for analyzing cross-sectional surveys. Pfeffermann and Burck (1990) show how to incorporate the constraints in the Kalman filter equations. Empirical results presented by the authors illustrate the good performance of the modified estimators in abnormal time periods. See also section 3.

3. FITTING THE MODEL TO THE ATLANTIC PROVINCES, EMPIRICAL RESULTS

The model defined by (2.2), (2.5) (2.6) and (2.10) was fitted to the monthly panel estimators in the four Atlantic provinces in two stages. In the first stage the model defined by (2.2), (2.5) and (2.6) was fitted to each of the provinces separately. In the second stage, the correlations defining the matrix ϕ of (2.10) were estimated using a grid search procedure. (See section 2.6). The estimators obtained are, $Diag(\phi) = (0.5, 0.25, 0.80, 0.0)$. The data used for estimation of the model cover the years 1982-1988. Data for 1989 were used for model diagnostics by comparing the results within and outside the sample period.

3.1 Prior considerations

Table 1 shows a breakdown of the labour force in the four provinces by industry. The figures in the table refer to March 1991. The (expected) sample sizes of the LFS are also shown. As can be seen, the

percentage breakdowns in the four provinces are very similar justifying the assumption of equal correlations across the provinces between the error terms of the census models. The similarity of the percentage breakdowns suggests also possible improvements in the efficiency of the model based estimators derived from the joint model over estimators which ignore the cross-sectional correlations between the province population values.

Table 1: Labour Force by Industry in the Atlantic Provinces, March 1991

Sample size	Nova Scotia		New Brunswick		Newfoundland		Prince-Edward Island	
	4409		3843		2970		1421	
	Thousands	Percentages	Thousands	Percentages	Thousands	Percentages	Thousands	Percentages
Agriculture	7	1.7	7	2.3	.5	.2	6	9.8
Other primary ind	18	4.4	13	4.2	18	7.7	4	6.6
Manufacturing	44	10.7	37	11.9	23	9.9	6	9.8
Construction	24	5.9	21	6.8	18	7.7	4	6.6
Transp. & commun.	35	8.6	30	9.6	20	8.6	5	8.3
Trade & Commerce	81	19.8	61	19.6	41	17.6	10	16.4
Finance	20	4.9	12	3.9	6	2.6	.5	.8
Services	143	35.0	107	34.4	83	35.6	19	31.1
Public Admin.	36	8.8	22	7.0	23	9.9	6	9.8
Unclassified	1	.2	1	.3	.5	.2	.5	.8
Total	409	100	311	100	233	100	61	100

Two other prior considerations mentioned in section 2.3 are that the model should account for possible rotation group effects and for changes in the variances of the survey errors over time. In order to obtain initial estimates for the rotation group effects, we averaged the pseudo survey errors, $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, $j=1, \ldots, 6$ over all the months in the sample period. We then divided the averages by the conventional estimates of the standard errors. (The errors $e_{t,p}^{(j)}$ are correlated over time but the correlations are small because except for lags 7, 13 etc. the estimators $y_t^{(j)}$ refer to different psu's or strata. See section 2.1). Notice that in the absence of rotation group effects, $E(e_{t,p}^{(j)}) = 0$ for all j and t irrespective of the model postulated for the population values.

This preliminary (model free) analysis yields similar results to the results obtained under the full model, presented in table 2 of section 3.3.

Next consider the variances of the survey errors.

Figure 1 plots the seasonal effects of the aggregate survey estimators in the four provinces along with the seasonal effects of the standard errors of these estimators (multiplied by 100). The seasonal effects were estimated by application of the additive mode of X-11. Denote as before by w_{ta} the relative labour force size in province a at time t. The aggregate survey estimator is defined as $y_t^* = \Sigma_{a=1}^{4} w_{ta} \bar{y}_{ta}$ (Equation 2.11). The standard error of y_t^* is $(SD^*)_t = \left[\sum_{a=1}^{4} w_{ta}^2 (\hat{SD})_{ta}^2\right]^{1/2}$.

Figure 1 reveals that the standard errors are influenced by seasonal variations with a seasonal pattern that follows closely the seasonal pattern of the survey estimators and hence of the corresponding population values.

As discussed in section 2.3, ratter than using the original estimates of the design standard errors in the models fitted to the panel survey errors we use smoothed values, thus reducing the effect of the sampling errors on the former estimators. Figure 2 plots the two sets of estimators for Prince Edward Island (PEI) province which is the smallest province in the Atlantic region and hence has the smallest sample sizes. As can be seen, the effect of the smoothing is to trim the extreme raw estimates but otherwise the smoothed values behave similarly to the raw estimates. The plots for the other provinces show a similar pattern but the differences between the raw and the smoothed estimates are smaller because of the larger sample sizes in these provinces.

3.2 Modification to the original model

After fitting the separate models and computing simple 12-terms moving averages of the estimated seasonal effects, it became evident that unlike the assumption in (2.1), the variance of the sums of the seasonal effects decreases with time. The models have been modified accordingly and re-fitted to the data.

3.3 Results

3.3.1 Rotation Group Biases

Table 2 shows the rotation group Biases (RGB) and their estimated standard errors (SE) in the four provinces as obtained under the full model defined by (2.3), (2.5), (2.6) and (2.10).

Table 2: Rotation Group Effects and Standard Errors in the Four Provinces (X100)

	Nova Scotia		New Brunswick		Newfoundland		Prince Edward Island	
Panels	RGB	SE	RGB	SE	RGB	SE	RGB	SE
1	-0.20	0.10	-0.02	0.11	-0.47	0.13	0.32	0.17
2	0.18	0.09	0.40	0.10	0.42	0.12	0.18	0.15
3	0.32	0.08	0.24	0.09	0.47	0.12	0.31	0.15
4	0.06	0.07	0.01	0.09	0.18	0.12	0.03	0.15
5	-0.03	0.08	-0.15	0.10	-0.10	0.13	-0.25	0.16
6	-0.34	0.08	-0.50	0.11	-0.50	0.14	-0.60	0.16

The RGB behave very consistently across the provinces. Thus, the biases for the 3rd and 6th panel are all highly significant using the conventional t-statistic, having a positive sign for the 3rd panel and a negative sign for the 6th panel. The biases for the 4th and 5th panels have again the same sign in all the provinces and they are all nonsignificant.

For the 2nd panel all the biases are positive but the bias in P.E.I. is not significant. (P.E.I. is the province with the smallest sample size). It is also in P.E.I. that the sign of the bias for the 1st panel is different from the signs in the other provinces.

As discussed in section 2.3, there is more than one possible reason for the existence of RGB but the results emerging from the table provide a strong indication that whatever the reason is, the biases found for some of the panels are real and not just the outcome of sampling errors. A drawback of the present analysis, however, is that the RGB are assumed to be fixed over time. Section 4 proposes a more flexible model.

3.3.2 Goodness of Fit

A. TESTING FOR NORMALITY

Let $I_{ta}^{(j)} = \left(y_{ta}^{(j)} - y_{ta|(t-1)}^{(j)}\right)$ define the innovation when predicting the $j-th$ panel estimator one month ahead and denote $I_{ta}' = \left(I_{ta}^{(1)} \ldots I_{ta}^{(6)}\right)$. The use of maximum likelihood estimation in this study assumes that the vectors I_{ta} are normal deviates (see section 2.4). To test this assumption, we computed the empirical distribution of the standardized innovations $\left\{(SI)_{ta}^{(j)} = \left[I_{ta}^{(j)} / \hat{SD}\left(I_{ta}^{(j)}\right)\right], t = (k+1) \ldots N\right\}$ and compared it to the standard normal distribution using the Kolmogorov-Smirnov test statistic. This test statistic was computed for each of the six panels in the four provinces yielding p-values larger than 0.15 in 21 out of the 24 cases.

Applying the same test procedure to the standardized innovations $\{(SI)_{ta} = [I_{ta} / \hat{SD}(I_{ta})], t = (k+1) \ldots N\}$ where $I_{ta} = \left[\sum_{j=1}^{6} I_{ta}^{(j)} / 6\right]$ yields p-values larger than 0.15 in all four provinces.

The estimators of the standard deviations of the innovations used for the tests are those produced by the Kalman filter, without accounting for the variance component resulting from parameter estimation (see section 2.5). The latter component is negligible even in P.E.I. which has the smallest samples sizes among the four provinces. We come back to this finding in section 3.4.

B

COMPARISON OF THEORETICAL AND EMPIRICAL VARIANCES

The appropriateness of the model can be assessed also by comparing the empirical means of the squares of the innovations with their theoretical variances under the model. Figure 3 shows the square roots of the weighted sums $(wI)_t = \sum_{a=1}^{4} w_{ta}\left[\sum_{j=1}^{6} (I_{ta}^{(j)})^2 / 6\right]$ along the square roots of the weighted variances $V_t = \sum_{a=1}^{4} w_{ta}\left[\sum_{j=1}^{6} \hat{Var}\left(I_{ta}^{(j)}\right) / 6\right]$ where w_{ta} denotes as before the relative labour force size in province a and $\hat{Var}\left(I_{ta}^{(j)}\right)$ is the estimated variance of $\left(I_{ta}^{(j)}\right)$ under the model. Notice that the last 12 pairs of values refer to the year 1989 which data were not used for model estimation.

The picture revealed from Figure 3 is that the discrepancy between the square roots of the model dependent variance estimators and the empirical root mean squared errors (RMSE) is in most cases less than 20 percent. Notice that the innovations defining the sum $(w\ell)_t$ are correlated so that the effective number of innovations in each month is less than 24. This fact could explain the occasional large discrepancies. It is seen also that the model dependent estimators of the variances of the innovations are unbiased. The average of $[(w\ell)_t]^{1/2}$ over all the years considered is 0.0158, the corresponding average of $[V_t]^{1/2}$ is 0.0159.

C. PREDICTION ERRORS WITH DIFFERENT PREDICTORS

Table 3 contains summary statistics comparing the behaviour of the prediction errors (innovations) in the four provinces as obtained for three different sets of estimators of the state vector: 1) The estimators obtained under the separate models (SM) defined by (2.2), (2.5) and 2.6, 2) the estimators obtained under the joint model (JM) defined by (2.2), (2.5), (2.6) and (2.10) and 3) the estimators obtained by imposing the robustness constraints (2.11) on the joint model (ROB). Below we define the summary statistics using as before the notation $I_{ta}^{(j)} = \left(y_{ta}^{(j)} - \hat{y}_{ta|(t-1)}^{(j)}\right)$ for the prediction error when predicting the j-th panel estimator one month ahead.

$MB_a = \sum_{t=k+1}^{N} \left(\sum_{j=1}^{6} I_{ta}^{(j)} / 6 \right) / (N-k)$ - mean bias in predicting the mean survey estimator $\bar{y}_{ta} = \sum_{j=1}^{6} y_{ta}^{(j)} / 6$.

$MAB_a = \sum_{j=1}^{6} \left| \sum_{t=k+1}^{N} I_{ta}^{(j)} / (N-k) \right| / 6$ - mean absolute bias in predicting the panel estimators.

$SQRE_a = \left\{ \sum_{t=k+1}^{N} \left[\frac{1}{6} \sum_{j=1}^{6} (I_{ta}^{(j)} / \bar{y}_{ta})^2 \right] / (N-k) \right\}^{1/2}$ - square root of mean square relative prediction error

in predicting the mean survey estimator.

The above summary statistics are shown separately for the sample period of July 83 - December 88 and for the postsample period of January 89 - December 89.

TABLE 3: PREDICTION ERRORS IN THE FOUR PROVINCES, SUMMARY STATISTICS (X100)

	Nova Scotia			New Brunswick			NewFoundland			Prince Ed. Island		
	SM	JM	ROB	SM	JM	ROB	SM	JM	ROB	SM	JM	ROB
7.83 - 12.88												
MB	-.11	-.07	-.06	-.12	-.09	-.06	-.25	-.18	-.08	.06	.14	.15
MAB	.12	.11	.10	.14	.12	.11	.29	.24	.20	.20	.23	.23
SQRE	5.76	5.62	5.70	5.48	5.47	5.47	7.03	6.91	6.96	9.34	9.13	9.17
1.89 - 12.89												
MB	.14	.11	.04	.47	.47	.46	.36	.33	.17	.84	.85	.86
MAB	.32	.32	.30	.51	.51	.50	.39	.37	.29	.84	.85	.86
SQRE	6.39	6.27	6.82	6.25	6.25	6.32	5.92	5.90	5.61	9.45	9.26	9.30

The main conclusions from table 3 are as follows:

1) The results obtained for the three sets of predictors are in general very similar, indicating that for the data analyzed the use of the joint model improves only slightly over the use of the separate modes and that there are no abrupt changes in the level of the series in the years considered.

2) The errors when predicting the survey estimators are small both within and outside the sample period, suggesting a good fit of the model. Notice that except in P.E.I., the relative prediction errors as measured by the statistics $SQRE_a$ are all less than 7%.

3) The biases of the prediction errors in the postsample period are larger than in the sample period with relatively large differences in New Brunswick and P.E.I. This outcome by itself could suggest some model failure in the year 1989. Inspection of the individual panel prediction errors in the four provinces for this year, (not shown in the paper), indicates however that although the errors are in general mostly positive, the relatively large biases are mainly the result of one or two extreme errors which, with only 12 data points, has a large effect on the average summary statistics. It should be noted also that the estimated unemployment rates in the four provinces in the year 1989 are between 0.11 and 0.18 so that a prediction bias of .005 or even .009 as obtained for P.E.I. is not high. Clearly, the model can be

modified to account for these biases if they persist with additional data. On the other hand, notice that the discussion above refers only to the bias of the prediction errors since the bias of the model based estimators of the concurrent population values is controlled by the robustness constraints (2.11).

In view of the very similar results obtained for the three sets of predictors considered and in order to highlight the performance of the robustness constraints, we deliberately deflated the unemployment rates in the period March 85 to March 87 by 33%, deflated the rates in the period April 87 - November 88 by 25% and inflated the rates in the period December 88 - December 89 by 33%. The effect of these operations is to introduce sudden drifts in the data in the months t=39, t=64 and t=84. Figure 4 displays the aggregate, one step ahead prediction errors (APE), $I_t^a = 6\sum_{a=1}^{4} w_{ta}\left[\sum_{j=1}^{6}\left(y_{ta}^{(j)} - \hat{y}_{ta|(t-1)}^{(j)}\right)/6\right]$ as obtained for the joint model with and without the robustness constraints, and also for the separate models.

The clear conclusion from Figure 4 is that by imposing the constraints, the APE in the periods following the three months with sudden drifts are smaller than the APE obtained without the constraints. Thus, in March 85 for example, (t=39), the APE are very large in absolute value both with and without the constraints which is obvious since the predictors use only the data until February 85. The APE corresponding to the robust predictors however, return to their normal level much faster than the APE of the nonrobust predictors. A similar behaviour is seen to hold in the other two periods. Another notable result featured in the graph is that in the periods following the months with the sudden drifts, the joint model performs better than the separate models even without imposing the robustness constraints. Thus, by borrowing information from one province to the other, the joint model adapts itself more rapidly to the new level of the series. For more illustrations of the performance of the robustness constraints see Pfeffermann and Burck (1990).

D. COMPARISONS WITH ESTIMATORS PRODUCED BY X-11

As a final assessment of the appropriateness of the model, we compare the estimates of the seasonal effects and the trend levels as obtained under the model with the estimates produced by the X-11 procedure (Dagum, 1980). The latter is known to be more flexible and less dependent on specific model assumptions. This procedure is the commonly used method for seasonal adjustment throughout the world. Figure 5 displays the average seasonal effects for the four provinces as obtained by X-11 and under the model. Figure 6 displays the corresponding trend level estimates. The averages are computed using the weights

(w_{tb}) employed in previous analyses. The model based estimates shown in the two figures are the smoothed estimates which, like X-11, employ all the data in the sample period.

As can be seen, the seasonal effects produced by the two approaches are very close. The trend level estimates are also close but the X-11 trend curve is smoother than the model curve. Similar close correspondence between X-11 and the model is obtained for each of the four provinces separately, including, in particular, P.E.I. with its relatively small sample sizes.

3.4 Comparison of Design Based and Model Dependent Estimators

We mention in the introduction that one of the major reasons for wishing to model the raw survey estimators is that the model produces estimates for the population values which, at least in small areas, are more accurate (when the model holds) than the survey estimators. Figure 7 displays the two sets of estimates of P.E.I. unemployment rates. The model dependent estimates are the smoothed values of the joint model which use all the data in all the months. As can be seen, the estimates produced by the two approaches behave very similar but the design based estimators have in general higher peaks and lower troughs. Figure 8 displays the corresponding standard errors (S.E.) as computed under the design, (smoothed values, see figure 2), and under the joint model. Also shown are the S.E. when fitting the separate model defined by (2.2), (2.5) and (2.6) and the corresponding S.E. after accounting for the use of parameter estimates instead of the unknown parameter values. See section 2.5 for details. (The latter have been computed only for the separate model to save in computing time).

There are three notable features emerging from the graphs:

1) The S.E. of the model dependent estimators under the joint model are only mildly smaller than the S.E. obtained for the separate model but considerably smaller than the S.E. of the survey estimators.

2) The S.E. of the model dependent estimators behave similarly to the S.E. of the survey estimators, a direct consequence of accounting for the changes in the variances of the survey errors over time in the model. See section 2.3 for details.

3) Accounting for the use of estimated parameter values in the computation of the S.E. of the model dependent estimators has only a marginal effect on the computed S.E. Recall that P.E.I. is the province with the smallest sample sizes. The effect of accounting for the use of parameter estimates in the other provinces is even smaller.

4. SUMMARY

This article illustrates that even data collected by a complex sampling design, consisting of several stages of selection with rotating panels, can be successfully modelled by a relatively simple model. The model consists of two parts: the census model holding for the population values and the survey errors model describing the time series relationship between the survey errors. The use of the model yields more accurate estimators for the population values and their components like trend and seasonality and it permits estimating the S.E. of these estimators in a rather simple way. The model equations can be modified to secure the robustness of the model-dependent estimators against possible model failures.

The model used in this article can be refined in various directions. Two refinements of particular relevance are to relax the assumption of constant variance for the error term ϵ_t in the census model and to let the rotation group biases to change over time.

The first refinement is suggested by the observation made in section 3.1 that the variances of the survey errors are subject to seasonal effects, with a seasonal pattern that is similar to the seasonal pattern of the raw estimates. Fitting the equations (2.4) in the four provinces indicates also the existence of a mild trend in the variances which again behaves similarly to the trend of the raw survey estimates. Thus, the variances of the survey errors seem to depend on the magnitude of the survey estimators which suggests that the variances $\sigma_t^2 = V(\epsilon_t)$ change with the level of the population values. As a first approximation one could assume that σ_t^2 is proportional to the corresponding variance of the survey error.

Letting the rotation group biases change over time is a natural extension of the model, considering that the population values means are time dependent. Modelling the evolution of the group biases can however be problematic because of possible identifiability problems with the models holding for the trend and the seasonal effects. See the discussion in Pfeffermann (1991).

The two refinements mentioned above are important and should be explored but based on our experience with the unemployment data, we expect that they will affect the model estimators very mildly.

ACKNOWLEDGEMENT

Work on this study was carried out while the first author was staying at Statistics Canada under its Research Fellowship Program.

REFERENCES

DAGUM, E.B. (1980). The X-11 ARIMA seasonal adjustment method. Catalog No. 12-564E, Statistics Canada, Ottawa, Ontario. K1A 0T6.

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

HARVEY, A.C. (1984). A unified view of statistical forecasting procedures (with discussion). *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.

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*, 5, 255-268.

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

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

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). Methodology of the Canadian labour force survey. Catalogue No. 71-526, Statistics Canada, Ottawa, Ontario, K1A 0T6.

Figure 1: Seasonal Effects of Aggregate Survey Estimators and of Standard Errors of Aggregate Survey Estimators (x100)

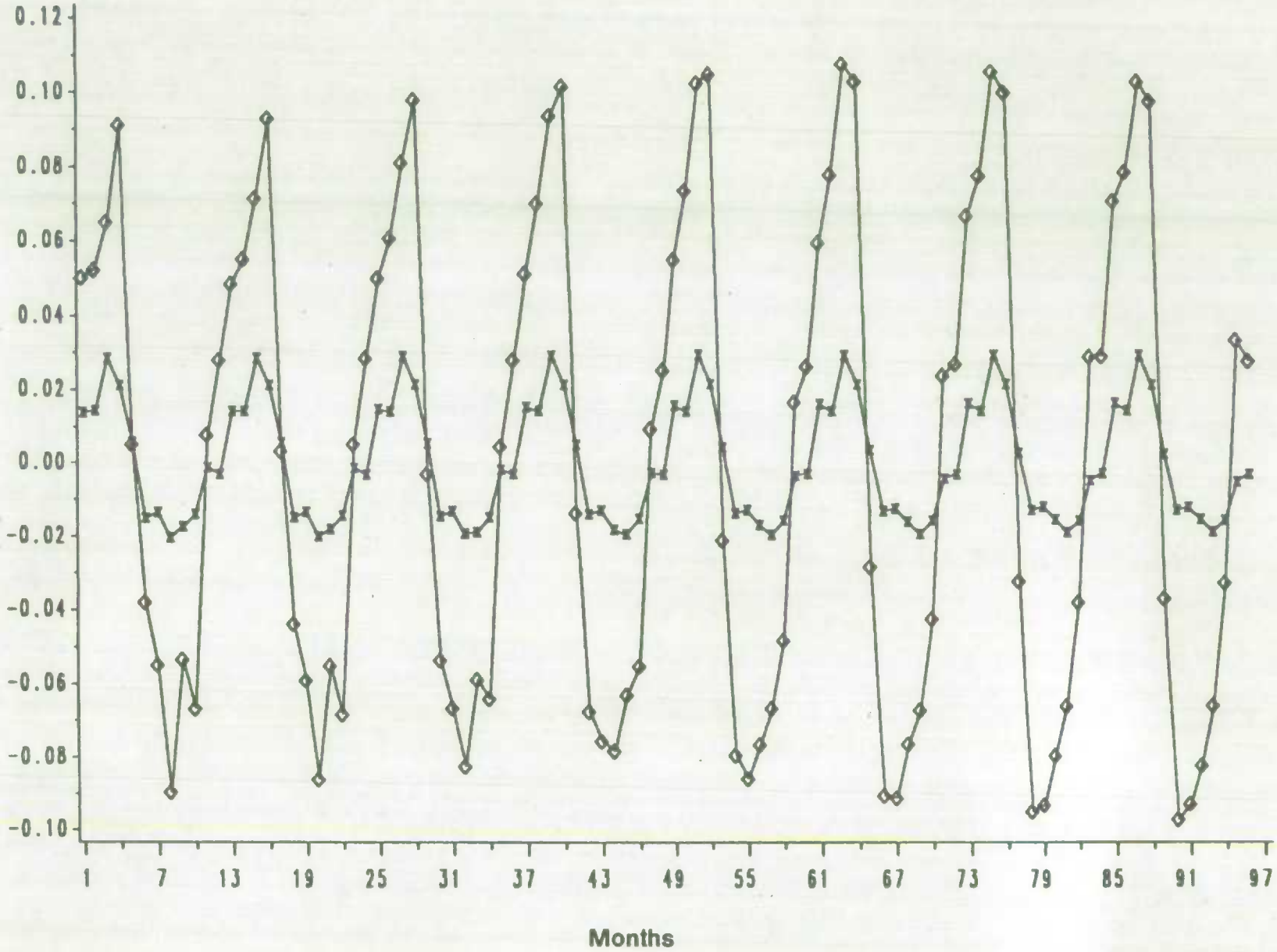


Figure 2: Original and smoothed standard errors of survey estimators (x100), P.E.I. Province

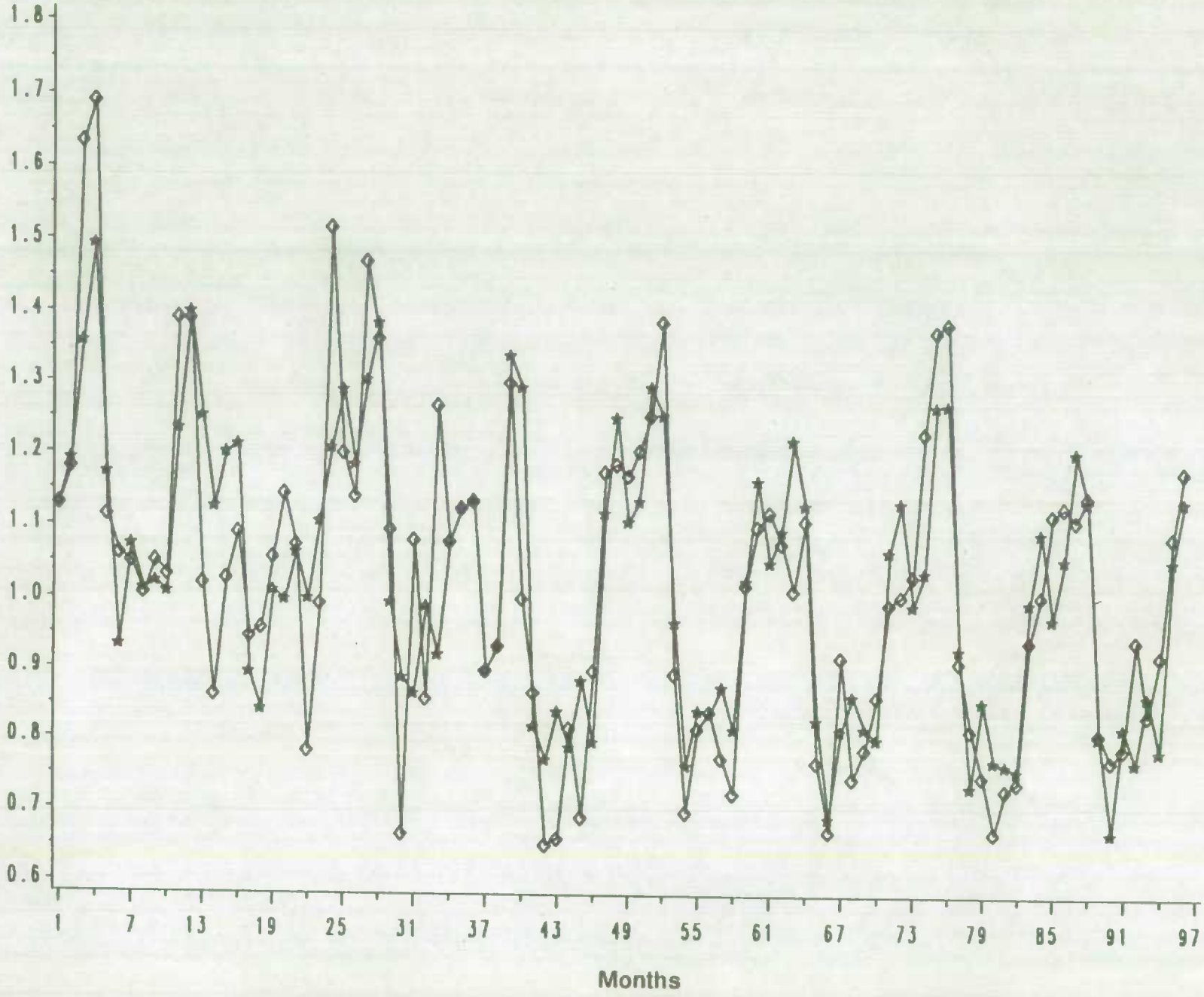


Figure 3: Empirical RMSE $[(wI)_c]^{1/2}$ and Square Roots of Estimators of Theoretical Variances $[V_c]^{1/2}$

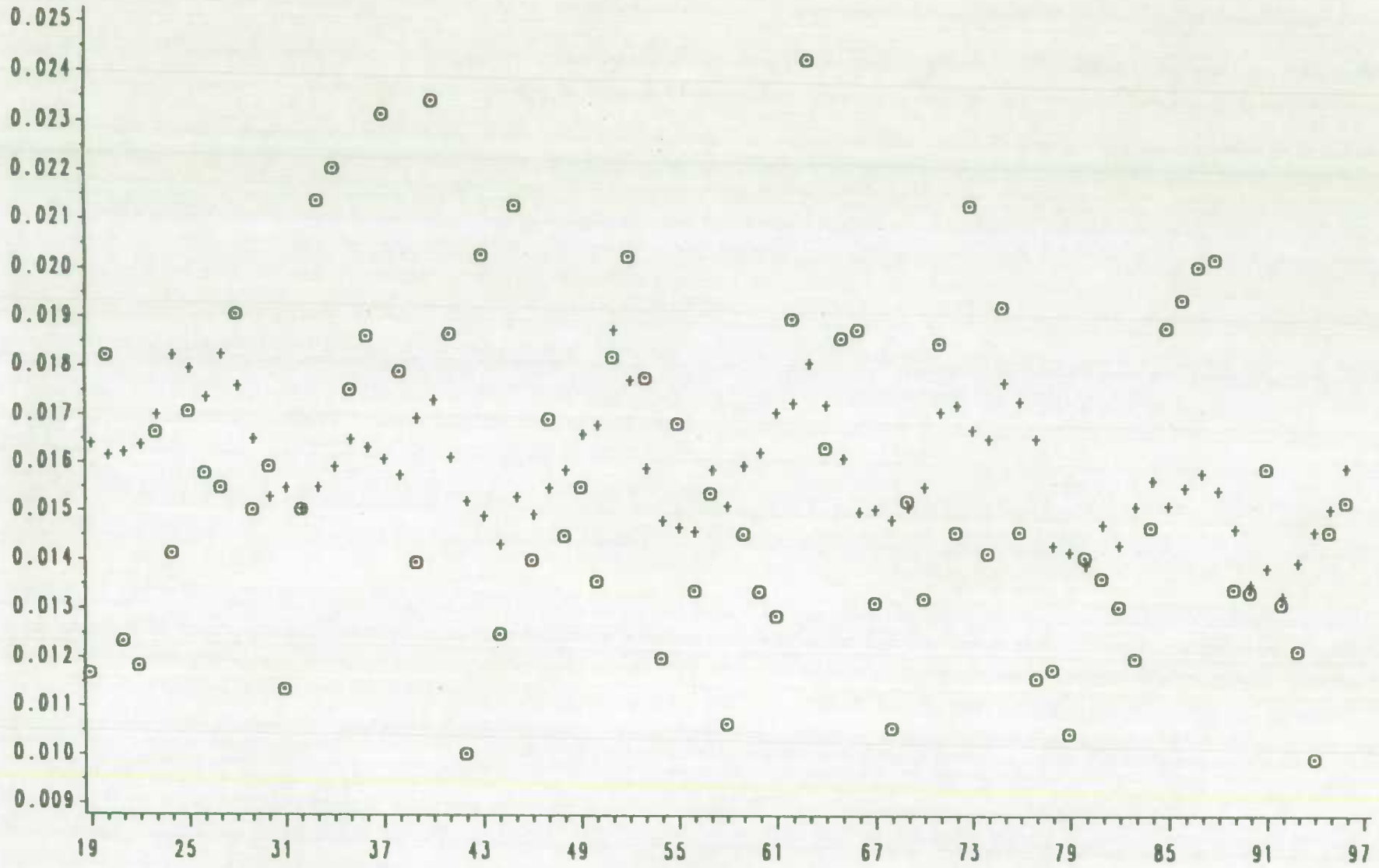


Figure 4: Aggregate One-Step Ahead Prediction Errors of Three Sets of Predictors (x100). Contaminated Data

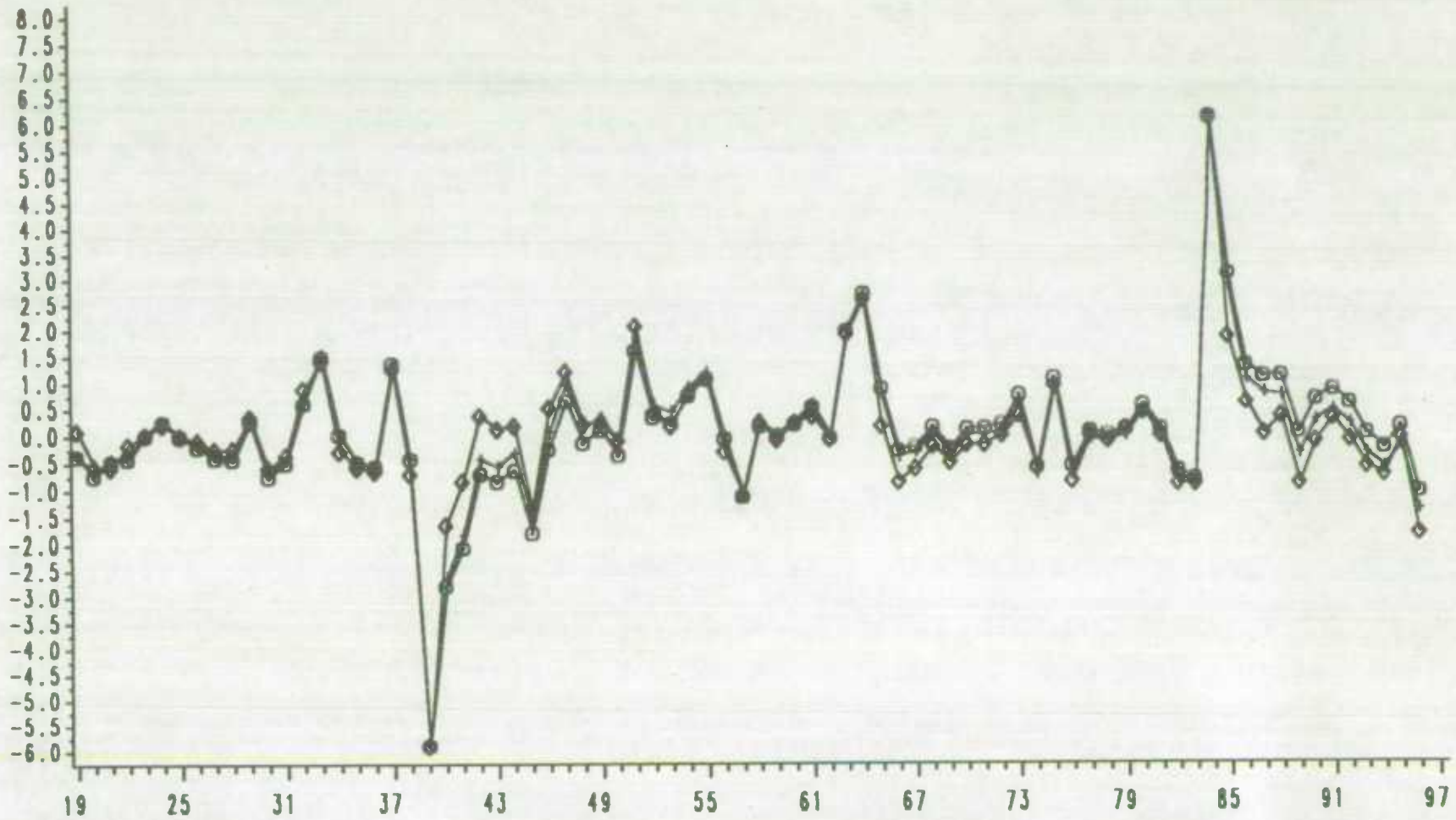


Figure 5: Weighted Averages of Seasonal Effects as Obtained by X-11 and Under the Model (x100)

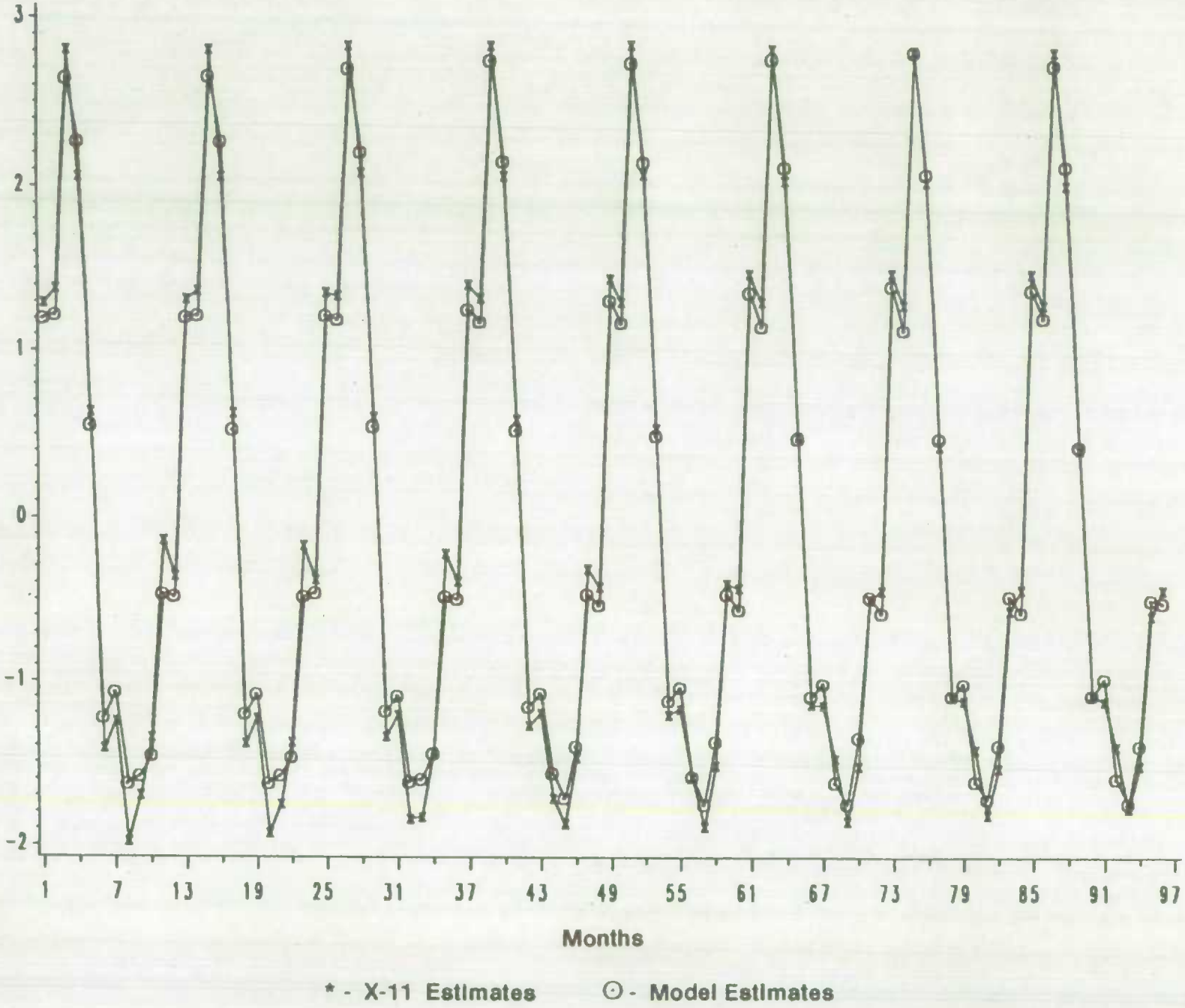


Figure 6: Weighted Averages of Trend Levels as Obtained by X-11 and Under the Model

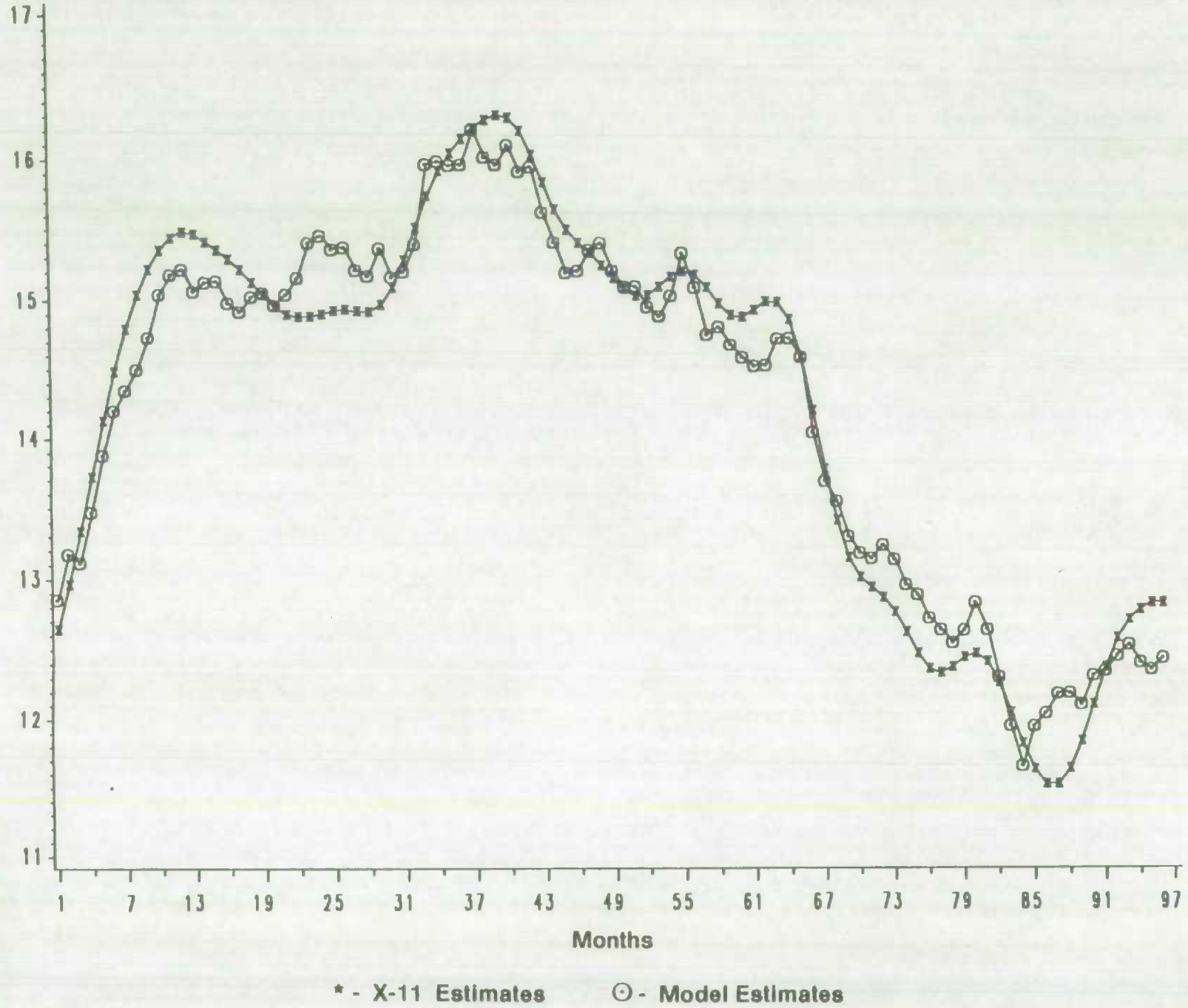


Figure 7: Design Based and Model Dependent Estimates of P.E.I. Unemployment Rates (x100)

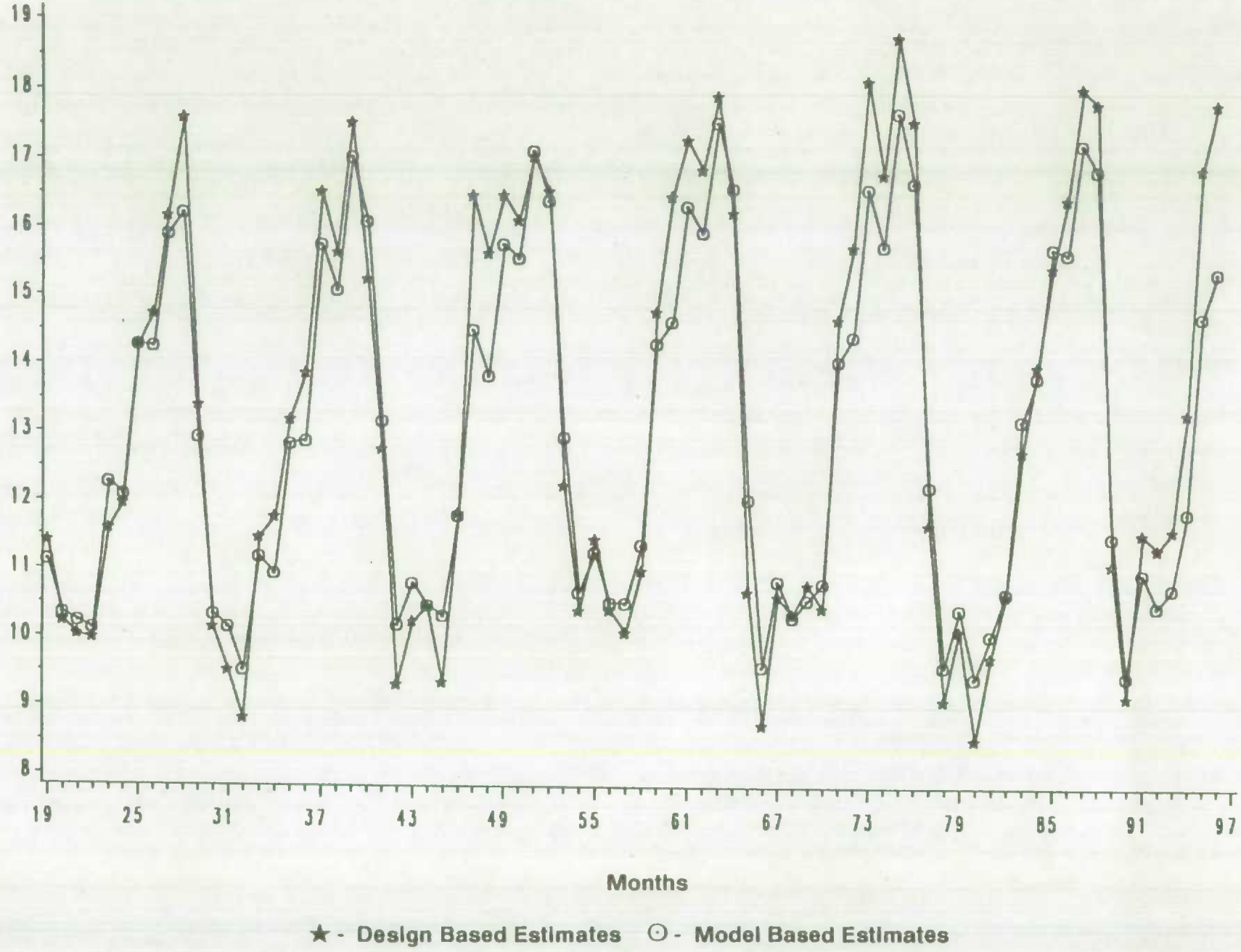
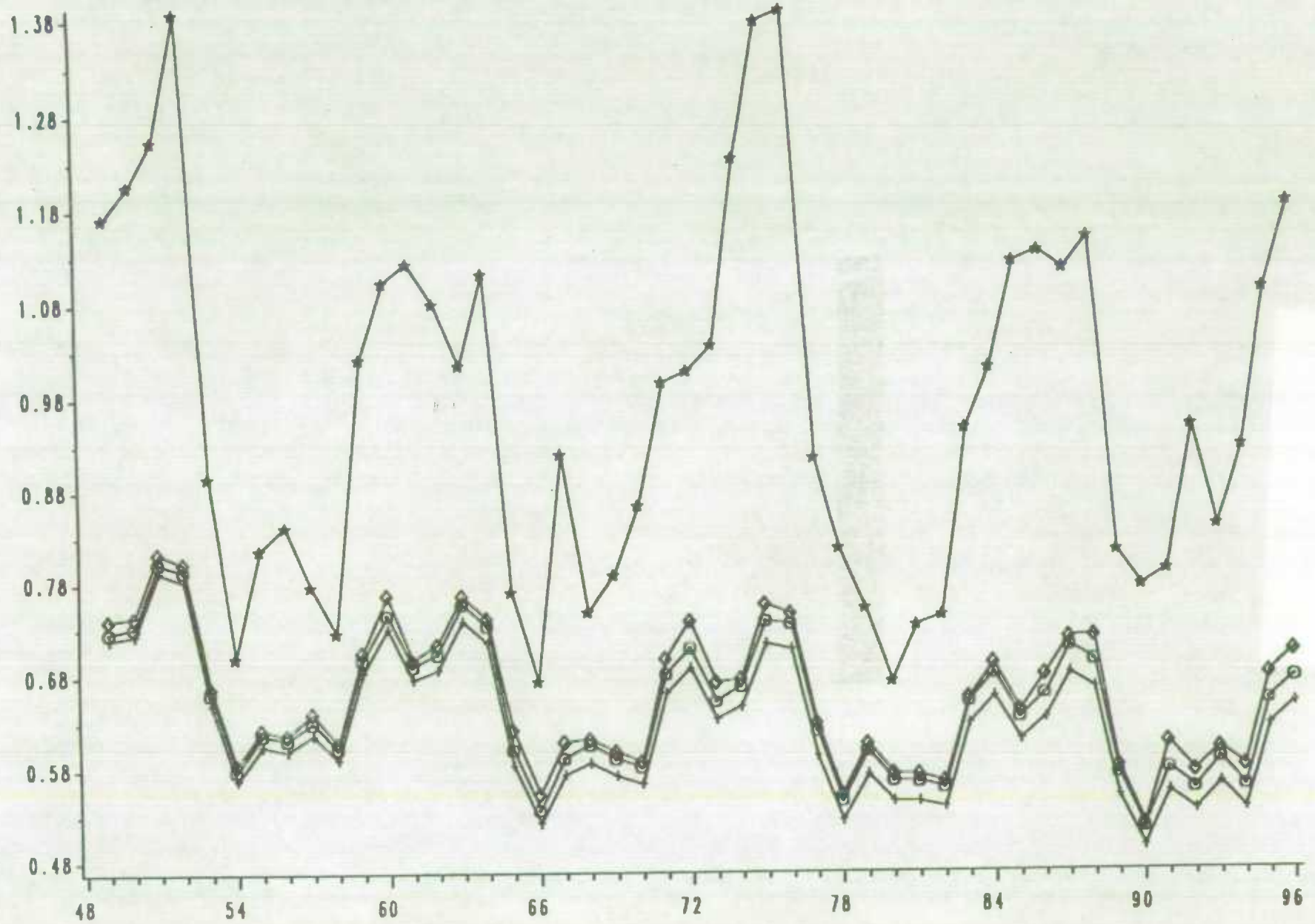


Figure 8: S.E. of Survey Estimators and of Model Dependent Estimators With and Without Accounting for Parameter Estimation (x100). P.E.I. Province



Ca 005

Statistics Canada Library
Bibliothèque Statistique Canada

1010089477