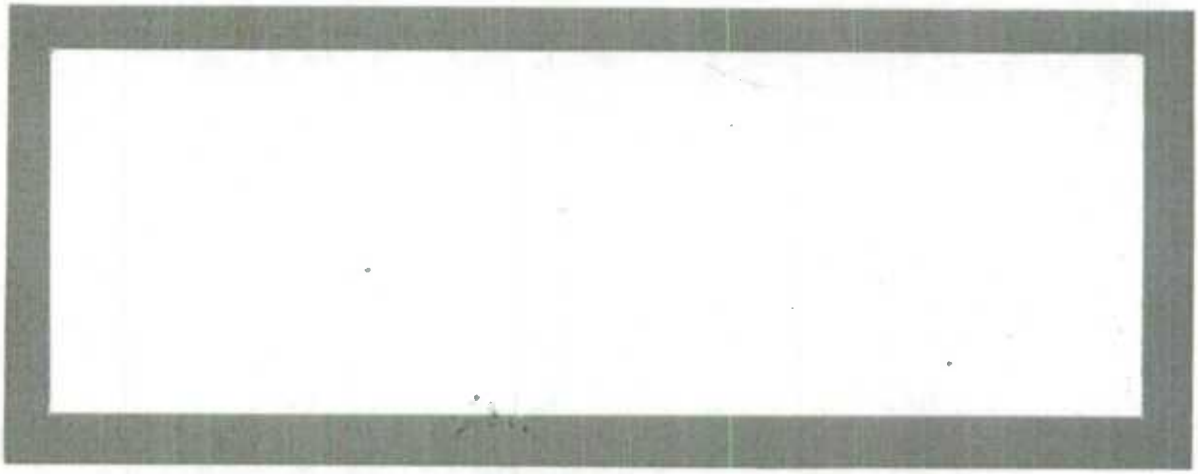
1-617
no. 97-04E
. 2

Statistics Canada / Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-97-004E

METHODOLOGY BRANCH

CAHIER DE TRAVAILNO. BSMD-97-004E

DIRECTION DE LA MÉTHODOLOGIE

STATISTICS CANADA STATISTIQUE CANADA

JUL 31 1997
JUIL

LIBRARY
BIBLIOTHÈQUE

A STATISTICAL APPROACH FOR DISAGGREGATING MIXED-FREQUENCY ECONOMIC TIME SERIES DATA

by

Wai-Sum Chan and Zhao-Guo Chen

July 1997

STATISTICS CANADA
LIBRARY
BIBLIOTHÈQUE

A Statistical Approach For Disaggregating Mixed-Frequency Economic Time Series Data

By

Wai-Sum Chan[1] and Zhao-Guo Chen[2]

July, 1997

Business Survey Methods Division

Statistics Canada

[1] *Department of Economics & Statistics, National University of Singapore, Republic of Singapore 119260*

[2] *Time Series Research & Analysis Centre, 3-H Coats Building, Statistics Canada Ottawa, ON, Canada K1A 0T6*

ABSTRACT

The problem of mixed-frequency time series data arises from changing the observation frequency. For example, we may have a time series with quarterly observations in the first portion and annual figures in the remainder. We shall call that quarter-year mixed-frequency data. In this paper we suggest a method to disaggregate the annual observations to quarterly values. The proposed method can easily be generalised to the year-quarter, quarter-month, year-month and other mixed-frequency situations; it may avoid difficulties of time series modelling and is easy to implement. A step-by-step algorithm of the method is given so that econometricians not expert in this area can still perform the procedure. The proposed method is illustrated through two real examples. We also conduct a small scale Monte Carlo experiment to compare the proposed procedure with two existing alternative methods. Finally, some concluding remarks are given.

RÉSUMÉ

Le problème suscité par la présence de fréquences mixtes dans les séries chronologiques apparaît lorsque la fréquence de la collecte des données change. Par exemple, les observations sont trimestrielles dans la première partie de la série et annulles par après. Ce sont des données de fréquences mixtes trimestrielles-annuelles. Dans cet article, nous suggérons une méthode de désagrégation des observations annuelles en données trimestrielles. La méthode proposée peut facilement être généralisée pour traiter les cas de données annuelles-trimestrielles, trimestrielles-mensuelles, annuelles-mensuelles et autres cas; de plus elle évite les difficultés reliées à la modélisation des séries chronologiques et sa mise en pratique est simple. L'algorithme de la méthode est conçu de façon à permettre aux économétriciens qui ne sont pas experts dans ce domaine de suivre chaque étape et d'exécuter la procédure. Deux exemples réels sont utilisés pour expliquer la méthode proposée. Nous avons également effectué une petite expérience à l'aide de la méthode de Monte-Carlo pour comparer la prédédure proposée à deux autres méthodes. Puis nous concluons.

1. INTRODUCTION

There are two types of mixed-frequency data problem in time series analysis. The first one occurs when we are analysing several time series jointly but their observation frequencies are not consistent. For example, in simultaneous equation modeling, quarterly data are available for all series but one, for which only annual data are observed. Instead of excluding that particular yearly time series in the model, we may disaggregate it to quarterly figures. Chow and Lin (1971), Fernandez (1981), Litterman (1983) and Guerrero (1990) have developed some prominent methods to perform such disaggregation.

The second type of mixed-frequency problem arises from changing the observation frequency of a time series. For example, a time series had been observed annually over several decades. Because of the increasing importance of the series, the reporting agency have been collecting and releasing quarterly figures for the variable in recent years. Now, we have a time series with annual observations in the first portion and quarterly figures in the remainder. We shall call this a *year-quarter* mixed-frequency time series. A well-known example is the US GNP series for which annual observations have been available since 1900 (cf. Vanderhoof *et al*, 1988, pp.582), but the quarterly figures of the series were reported only after 1954 (cf. Lutkepohl, 1991, pp.500-502). One of the solutions to the *year-quarter* mixed-frequency problem is to temporally aggregate the quarterly portion of the series to annual values. Wei and Stram (1990) proposed a model-based procedure to disaggregate all the available yearly totals to quarterly figures. Chan (1993) provided a comparative study for all existing alternative methods in the literature. However, temporal aggregation could lead to a considerable loss of information as discussed by Tiao (1972) and Rossana and Steater (1995). In this paper we propose a disaggregation procedure which will fully utilize all the information in the mixed-frequency series.

Most of the macroeconomic data are reported either quarterly or annually. Therefore we shall describe our proposed method for the *quarter-year* mixed-frequency situation. Assume that we have n quarterly observations $(y(1), \ldots, y(n))$ followed by N yearly observations $Z(T)$. It should be noted that $Z(T)$ is the observed annual temporal aggregates of the unobserved quarterly time series after time n, i.e.,

$$Z(T) = y(n + 4T - 3) + \ldots + y(n + 4T) \tag{1.1}$$

for $T = 1, \ldots, N$. Our primary goal is to disaggregate $Z(T)$ to $\hat{y}(n + t)$, for $t = 1, \ldots 4N$, such that the $\hat{y}(n + t)$ values satisfy the basic constraint in (1.1) and they are also as close to the unobserved $y(n + t)$ as possible. After the disaggregation we shall have a consistent quarterly time series with the length of $n + 4N$.

Suppose that the quarterly time series, $y(t)$, follows

$$\nabla^d \nabla_4^{d'} y(t) = x(t) \tag{1.2}$$

where d and d' are non-negative integers, $\{x(t)\}$ is a stationary time series, $\nabla = (1 - B)$ and $\nabla_4 = (1 - B^4)$ are the ordinary and the seasonal difference operator respectively, and B is the backshift operator, i.e., $B^s y(t) = y(t - s)$. We shall only consider the cases of $0 \leq d + d' \leq 2$ and $d' \leq 1$, which are the most commonly encountered situations in economic

time series. Let $d_0 = d + 4d'$. A general formula to express the unobserved $y(n+t)$ in (1.2) through its past observations $\mathbf{y}^* = (y(n-d_0+1), \ldots, y(n))'$ and the unknown values of $\{x(t)\}$ is given by

$$y(n+t) = f(t, \mathbf{y}^*) + \sum_{j=1}^{t} a(t,j)\, x(n+j), \quad \text{for } t \geq 1 \tag{1.3}$$

where f is a linear function of t and $\mathbf{y}^*$. Chen and Anderson (1994) and Chen and Liu (1997) derived the forms of $f(t, \mathbf{y}^*)$ and $a(t,j)$ for many combinations of d and d'. For $0 \leq d + d' \leq 2$ and $d' \leq 1$, the results are summarised in Table 1.

Table 1 about here

So far, our discussion and notations concentrate on the *quarter-year* mixed-frequency problem. However, *year-quarter* mixed-frequency series are more often encountered in economic analysis (i.e., annual data are available in the earlier period and quarterly figures are observed later). In those cases, we can still perform our proposed procedure in the next section by reversing the time axis. The *year-quarter* series will therefore become *quarter-year* data. However, it should be cautious that if $\mu(t) = E[y(t)]$ is, say, a quadratic function, then the extrapolation of $\mu(t)$ from one end is very different from the other end. There will be no such problem if we assume that

$$\mu(t) = \alpha_0 + \alpha_1 t + \sum_{i=1}^{4} \beta_i \delta_{\{t,i\}} \tag{1.4}$$

where α_0 and α_1 are constants, and

$$\beta_i \delta_{\{t,i\}} = \begin{cases} \beta_i, & \text{if } t = i \\ 0, & \text{otherwise} \end{cases}$$

represents the seasonality. Fortunately, if we assume that $E[x(t)]$ is in the form of $\mu(t)$ in (1.4) for $d + d' = 2$, and $E[x(t)]$ equals any constant μ for $d + d' = 1$, then the condition in (1.4) is automatically met. It can be easily seen from the equation (1.3) and the results in Table 1. Finally, the method can be generalised without any difficulty to *month-year*, *month-quarter* and other mixed-frequency situations.

2. THE DISAGGREGATION FOMULA

Let

$$\mathbf{y_1} = \begin{pmatrix} y(1) \\ \vdots \\ y(n) \end{pmatrix}, \qquad \mathbf{y_2} = \begin{pmatrix} y(n+1) \\ \vdots \\ y(n+4N) \end{pmatrix},$$

$$\mathbf{x_1} = \begin{pmatrix} x(d_0+1) \\ \vdots \\ x(n) \end{pmatrix}, \qquad \mathbf{x_2} = \begin{pmatrix} x(n+1) \\ \vdots \\ x(n+4N) \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} Z(1) \\ \vdots \\ Z(N) \end{pmatrix}. \tag{2.1}$$

It should be noted that (a) $\mathbf{y_1}$ and $\mathbf{Z}$ are quarterly and yearly vector of observations, respectively; (b) $\mathbf{x_1}$ is obtained from $\mathbf{y_1}$ through the difference equation in (1.2); (c) $\mathbf{y_2}$ and $\mathbf{x_2}$ are vectors of unknown future values connected by the relation in (1.3); and (d) $\mathbf{y_2}$ and $\mathbf{Z}$ are restricted by the constraint in (1.1). Our main objective in this section is to find a formula to estimate $\mathbf{y_2}$ such that the basic requirement in (1.1) is satisfied and the estimate is also as close to the true $\mathbf{y_2}$ as possible.

Denote the best linear forecast of $\mathbf{x_2}$ from the observed vector $\mathbf{x_1}$ by $\hat{\mathbf{x}}_\mathbf{2} = (\hat{x}(n+1), \ldots, \hat{x}(n+4N))'$. In other words, $\hat{\mathbf{x}}_\mathbf{2}$ is the projection of $\mathbf{x_2}$ on the linear space spanned by $\mathbf{x_1}$. We shall have more discussion on the forecasting methods for $\mathbf{x_2}$ in the next section. The corresponding best linear forecast of $\mathbf{y_2}$ from the observed vector $\mathbf{y_1}$ is

$$\hat{\mathbf{y}}_\mathbf{2}^{(1)} = (\hat{y}^{(1)}(n+1), \ldots, \hat{y}^{(1)}(n+4N))' \tag{2.2}$$

where

$$\hat{y}^{(1)}(n+t) = f(t, \mathbf{y}^*) + \sum_{j=1}^{t} a(t,j)\ \hat{x}(n+j), \quad \text{for } t = 1, \ldots, 4N \tag{2.3}$$

with $f(t, \mathbf{y}^*)$ and $a(t,j)$ functions defined in Table 1. The covariance matrix of the forecast error is given by $\mathbf{V}_{\mathbf{e_y}}^{(1)}$ with elements:

$$\begin{aligned} \mathbf{V}_{\mathbf{e_y}}^{(1)}(t_1, t_2) &= \mathrm{Cov}[e_y^{(1)}(t_1), e_y^{(1)}(t_2)] \\ &= \sum_{j_1=1}^{t_1} \sum_{j_2=1}^{t_2} a(t_1, j_1)\ \mathbf{V}_{\mathbf{e_x}}(j_1, j_2)\ a(t_2, j_2) \end{aligned} \tag{2.4}$$

where

$$e_y^{(1)}(t) = \hat{y}^{(1)}(n+t) - y(n+t), \quad \text{for } t = 1, \ldots, 4N \tag{2.5}$$

$$e_x(t) = \hat{x}(n+t) - x(n+t), \quad \text{for } t = 1, \ldots, 4N \tag{2.6}$$

and

$$\mathbf{V}_{\mathbf{e_x}}(j_1, j_2) = \mathrm{Cov}[e_x(j_1), e_x(j_2)]. \tag{2.7}$$

From the variance formula of the projection (e.g., see Whittle, 1963, section 4.1 and Brillinger, 1981, section 8.2), we have

$$\mathbf{V}_{\mathbf{e_x}} = \mathbf{\Gamma}_{22}^{\mathbf{x}} - \mathbf{\Gamma}_{21}^{\mathbf{x}} (\mathbf{\Gamma}_{11}^{\mathbf{x}})^{-1} \mathbf{\Gamma}_{12}^{\mathbf{x}} \tag{2.8}$$

where $\mathbf{\Gamma}_{\mathbf{ij}}^{\mathbf{x}} = \mathrm{Cov}[\mathbf{x_i}, \mathbf{x_j}]$. Define a lower triangular matrix $\mathbf{A}$ with dimension $4N \times 4N$,

$$\mathbf{A} = \begin{pmatrix} a(1,1) & 0 & \ldots & 0 \\ a(2,1) & a(2,2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a(4N,1) & a(4N,2) & \ldots & a(4N,4N) \end{pmatrix}. \tag{2.9}$$

We can simplify the equation (2.4) to

$$\mathbf{V}_{\mathbf{e_y}}^{(1)} = \mathbf{A}\ \mathbf{V}_{\mathbf{e_x}}\ \mathbf{A}'. \tag{2.10}$$

The superscript (1) on $\hat{\mathbf{y}}_2^{(1)}$ in (2.2) means that this estimate of $\mathbf{y_2}$ is only from the "first source" of data, $\mathbf{y_1}$. The "second source" or additional information from the vector observation $\mathbf{Z}$ has not been used. The basic constraint in (1.1) could be violated by $\hat{\mathbf{y}}_2^{(1)}$ and hence, it is not our final estimate.

The requirement in (1.1) can be re-written in matrix form:

$$\mathbf{Z} = \mathbf{L}\ \mathbf{y_2}, \qquad \mathbf{L} = \mathbf{I_N} \otimes (1\ 1\ 1\ 1) \tag{2.11}$$

where $\otimes$ denotes the matrix Kronecker product and $\mathbf{I_N}$ is an Identity matrix with dimension $N \times N$. The final estimate $\hat{\mathbf{y}}_2^{(2)}$, which combines information from both the "first source" and the "second source" of data, is given by

$$\hat{\mathbf{y}}_2^{(2)} = \hat{\mathbf{y}}_2^{(1)} + \mathbf{V}_{\mathbf{e_y}}^{(1)}\mathbf{L}'(\mathbf{L}\mathbf{V}_{\mathbf{e_y}}^{(1)}\mathbf{L}')^{-1}(\mathbf{Z} - \mathbf{L}\hat{\mathbf{y}}_2^{(1)}). \tag{2.12}$$

The covariance matrix of the final estimation error is

$$\mathbf{V}_{\mathbf{e_y}}^{(2)} = \mathbf{V}_{\mathbf{e_y}}^{(1)} - \mathbf{V}_{\mathbf{e_y}}^{(1)}\mathbf{L}'(\mathbf{L}\mathbf{V}_{\mathbf{e_y}}^{(1)}\mathbf{L}')^{-1}\mathbf{L}\mathbf{V}_{\mathbf{e_y}}^{(1)}. \tag{2.13}$$

From (2.12), we note that the final estimate, $\hat{\mathbf{y}}_2^{(2)}$, is simply the preliminary estimate $\hat{\mathbf{y}}_2^{(1)}$ plus an adjustment factor. It is also obvious that $\mathbf{L}\hat{\mathbf{y}}_2^{(2)} = \mathbf{Z}$ and hence $\hat{\mathbf{y}}_2^{(2)}$ is a proper disaggregated vector for $\mathbf{Z}$. Moreover, the final estimate $\hat{\mathbf{y}}_2^{(2)}$ is always an improvement of the initial estimate $\hat{\mathbf{y}}_2^{(1)}$. It is clear from (2.13) that

$$\mathbf{V}_{\mathbf{e_y}}^{(2)} \leq \mathbf{V}_{\mathbf{e_y}}^{(1)}. \tag{2.14}$$

Formulae (2.12) and (2.13) have been discussed in Trabelsi and Hillmer (1989) and Guerrero (1989) under the assumption of normality and in terms of $\hat{\mathbf{y}}_2^{(1)} = E[\mathbf{y_2}|\mathbf{y_1}]$ and $\hat{\mathbf{y}}_2^{(2)} = E[\mathbf{y_2}|\mathbf{y_1}, \mathbf{Z}]$. This problem is often referred to as constrained forecast in the time series literature. In fact it has the same form as those of benchmarking time series (e.g., see Hillmer and Trabelsi, 1987; Chen, Cholette and Dagum, 1997) and the Gaussian assumption can be removed if the conditional expection (the best estimate) is replaced by the projection (the best linear estimate).

A vital assumption of the methods proposed by Trabelsi and Hillmer (1989) and Guerrero (1989) is that $y(t)$ follows an ARIMA model. In our proposed procedure, the equation (2.3) enables us to convert *any* prediction of $\mathbf{x_2}$ (not necessary ARMA forecast) to $\hat{\mathbf{y}}_2^{(1)}$ and (2.4) converts the variance of prediction error of $\mathbf{x_2}$ to those of $\mathbf{y_2}$. The only requirement of the proposed method is $\nabla^d\nabla_4^{d'}y(t)$ being stationary. It offers greater flexibility by relieving the ARIMA model assumption for $y(t)$. We often encounter difficulties with specifying a suitable ARMA model for $x(t)$ and obtaining efficient parameter estimates when the sample size is small. Furthermore, the proposed method is able to handle non-linear time series which is not uncommon in economic data. Therefore, if we can find a better (compared to ARMA forecast) estimate of $\mathbf{x_2}$, we should be able to obtain a more accurate disaggregated vector $\hat{\mathbf{y}}_2^{(2)}$. It can be shown that a better estimate of $\mathbf{x_2}$ will lead to smaller values in $\mathbf{V}_{\mathbf{e_x}}$ and hence lower the error variance of $\hat{\mathbf{y}}_2^{(2)}$.

3. PRACTICAL CONSIDERATIONS

A step-by-step algorithm is given in this section so that econometricians not expert in this field can still perform the proposed disaggregation procedure. Furthermore, we shall discuss in detail of the practical issues involved in each step.

Step[1]: *Determine the non-negative integer values of d and d′ in (1.2).*

In model (1.2), we assume that $x(t) = \nabla^d \nabla_4^{d'} y(t)$ is stationary. There are many identification techniques available in the literature to specify the values of d and d' from the observed vector $\mathbf{y_1}$. We may employ, among many others, (a) the autocorrelation method by Box and Jenkins (1976, pp.174); (b) unit root tests suggested by various authors, ranging from the classical Dickey and Fuller (1981) test, the specialized seasonal unit root tests by Osborn *et al* (1988) and Hylleberg *et al* (1990) to the most recent method by Chan and Tsay (1996); or (c) graphical techniques developed by Cressie (1988) and Chen and Anderson (1998).

Step[2]: *Given d and d′ in Step [1], determine the corresponding* $f(t, \mathbf{y}^*)$ *and* $a(t, j)$ *functions from Table 1.*

Step[3]: *Find the "best" forecast of* $\mathbf{x_2}$ *from the observed vector* $\mathbf{x_1}$.

There are many methods for forecasting economic time series. An excellent reference book is given by Granger and Newbold (1986). A popular choice is the linear ARMA model forecast described in Box and Jenkins (1976, Chapter 5). An alternative approach to obtain the forecast is directly applying variance-covariance based projection formula. Although most economic time series models considered in the literature are linear, there is increasing amount of evidence that some economic time series are chaotic or non-linear. We may consider the non-linear forecasting methods discussed in Tong (1990). Finally, if one or more auxiliary series are available for $\mathbf{x_1}$, we may also employ multiple time series forecasting models. Examples of these models include transfer function models, vector ARMA models, and simultaneous equation models. We should find the "best" forecast of $\mathbf{x_2}$, using our own judgement, from one of the above mentioned forecasting models.

Step[4]: *Convert* $\hat{\mathbf{x}}_\mathbf{2}$ *from step [3] to* $\hat{\mathbf{y}}_\mathbf{2}^{(\mathbf{1})}$ *using formula (2.3).*

Step[5]: *Estimate* $\mathbf{V_{e_x}}$ *in (2.7).*

Let

$$\mathbf{\Gamma^x} = \begin{pmatrix} \mathbf{\Gamma_{11}^x} & \mathbf{\Gamma_{12}^x} \\ \mathbf{\Gamma_{21}^x} & \mathbf{\Gamma_{22}^x} \end{pmatrix} \quad \text{with dimension } (n - d_0 + 4N) \times (n - d_0 + 4N). \tag{3.1}$$

In order to obtain $\hat{\mathbf{V}}_{\mathbf{e_x}}$, we first need to estimate $\mathbf{\Gamma^x}$ in (3.1), then $\hat{\mathbf{V}}_{\mathbf{e_x}}$ follows from $\hat{\mathbf{\Gamma}}^{\mathbf{x}}$ using equation (2.8). There are three different available estimation methods of $\mathbf{\Gamma^x}$ for us to choose:

(a) linear ARMA model-based method;
(b) non-linear model-based method; and
(c) data-based sample autocovariance function method.

Suppose that we forecast $\mathbf{x_2}$ in Step [3] using a fitted linear ARMA(p, q) model

$$\phi_p(B)\tilde{x}(t) = \theta_q(B)a_t \tag{3.2}$$

where $\tilde{x}(t)$ is the mean-corrected series such that $\tilde{x}(t) = x(t) - \bar{x}$ for $t = (d_0 + 1), \ldots, n$ and $\bar{x} = (n - d_0)^{-1} \sum_{t=d_0+1}^{n} x(t)$. An estimate of the autocovariance function $\gamma^x(k)$ can be obtained from the fitted parameters in (3.2). McLeod (1975) provided an efficient algorithm to compute theoretical $\gamma^x(k)$ from an ARMA model. If a computer algebra system, such as MATHEMATICA or MAPLE, is available, specific formulae are easily derived. Therefore, we may use the fitted model in (3.2) to obtain $\tilde{\gamma}^x(k)$. Then the estimate of the Toeplitz matrix $\mathbf{\Gamma^x}$ in (3.1) is formed by setting its $(i, j)th$ element equal to $\tilde{\gamma}^x(|i - j|)$.

On the other hand, if we employ a non-linear time series model to forecast $\mathbf{x_2}$ in Step [3], we may perform a spectral analysis to get $\hat{\mathbf{\Gamma}}^{\mathbf{x}}$ (e.g., see Priestley, 1981, Chapter 11).

Finally, we can consider to use the data-based sample autocovariance function to estimate $\mathbf{\Gamma^x}$. This approach is preferred when the sample size is small or moderate (in such cases ARMA modeling is not very stable), or there is a possibility of non-linearity in the series. Define

$$\hat{\gamma}^x(k) = \frac{1}{n - d_0 - k} \sum_{t=d_0+1}^{n-k} \tilde{x}(t + k)\tilde{x}(t), \quad \text{for } 0 \le k \le (n - d_0 - 1) \tag{3.3}$$

as the sample autocovariance function for $\{x(t)\}$. Brillinger (1981, section 5.6) considered a tapered version of (3.3),

$$\hat{\gamma}^x_M(k) = \begin{cases} \hat{\gamma}^x(k)\ \mathcal{W}(k/M), & \text{if } k = 0, \ldots, M \\ 0, & \text{if } k > M \end{cases} \tag{3.4}$$

such that its corresponding (Fourier transform) spectral estimate is consistent. The $\mathcal{W}$ function in (3.4) is a lag window and M is the truncation point that depends on the sample size. For example, we may employ the Parzen lag window

$$\mathcal{W}(u) = \begin{cases} 1 - 6u^2 + 6|u|^3, & |u| \le 1/2 \\ 2(1 - |u|)^3, & 1/2 \le |u| \le 1 \\ 0, & |u| \ge 1. \end{cases} \tag{3.5}$$

Other commonly used lag windows are described in Priestley (1981, section 7.2) and Brillinger (1981, pp.55). The choice of the truncation point M, theoretically, is of $O(n^\alpha)$ for $0 < \alpha < 1$. For large n, there is a theoretical guideline for optimally choosing the value of α which depends on the property of the lag window. In practice, we choose $M \approx n/2$ and $M \approx n/3$ for small and moderate sample size, respectively. It should be noted that the ordinary sample autocovariance function $\hat{\gamma}^x(k)$ in (3.3) with the restriction $k \le (n - d_0 - 1)$ is not sufficient to form the Toeplitz matrix $\hat{\mathbf{\Gamma}}^{\mathbf{x}}$ with larger dimension $(n - d_0 + 4N) \times (n - d_0 + 4N)$. On the other hand, the autocovariance function in (3.4)

is tapered and there is no upper bound restriction on its k. Therefore, we should employ $\hat{\gamma}_M^x(k)$ to form $\hat{\mathbf{\Gamma}}^{\mathbf{x}}$ such that

$$\hat{\mathbf{\Gamma}}^{\mathbf{x}}(i,j) = \hat{\gamma}_M^x(|i-j|) \tag{3.6}$$

for $i = 1, \ldots, (n - d_0 + 4N)$ and $j = 1, \ldots, (n - d_0 + 4N)$. Once the $\mathbf{\Gamma}^{\mathbf{x}}$ is estimated, we can obtain $\hat{\mathbf{V}}_{\mathbf{e_x}}$ through equation (2.8).

Step[6]: *Set up the matrices* $\mathbf{A}$ *and* $\mathbf{L}$ *which are defined in (2.9) and (2.11) respectively.*

Step[7]: *Calculate* $\hat{\mathbf{V}}_{\mathbf{e_y}}^{(1)}$ *from equation (2.10) using* $\hat{\mathbf{V}}_{\mathbf{e_x}}$ *to replace* $\mathbf{V}_{\mathbf{e_x}}$.

Step[8]: *Compute* $\hat{\mathbf{y}}_{\mathbf{2}}^{(2)}$ *from the disaggregation formula (2.12) using* $\hat{\mathbf{V}}_{\mathbf{e_y}}^{(1)}$ *to replace* $\mathbf{V}_{\mathbf{e_y}}^{(1)}$.

4. ILLUSTRATIVE EXAMPLES

Example 1

We consider the quarterly beer production in the United States between the first quarter of 1967 to the fourth quarter of 1982. The time series is measured in terms of millions of barrels, and the data can be obtained from Nazem (1988, pp.330) and Wei (1990, pp.451). In order to illustrate our proposed disaggregation procedure, we assume that quarterly figures are observed from 1967.Q1 to 1978.Q4 and only yearly values are available after 1979. Therefore we have a *quarter-year* mixed-frequency time series with $n = 48$ and $N = 4$.

A time series plot of the first 48 quarterly observations is given in Figure 1. The graph displays an upward trend with a strong seasonal pattern. As the first step in our analysis, we calculate the sample autocorrelations of the series $y(t)$. The results are plotted in Figure 2(a). It clearly indicates the need for a seasonal differencing. Another sample autocorrelation plot is given in Figure 2(b) for the differenced series $x(t) = \nabla_4\, y(t)$. It suggests stationarity and a pure seasonal MA(1) or AR(1) model with period $s = 4$ for $x(t)$. However, the seasonal AR(1) model is marginally better fitted than the MA(1) process. The estimation results of the specified process for $x(t)$ are

$$(1 + 0.5595B^4)\, \tilde{x}(t) = a(t) \qquad \hat{\sigma}_a^2 = 2.0394 \tag{4.1}$$

with 0.1311 as the standard error of $\hat{\phi}_4$. The adequacy of the model is further confirmed by the residual analysis.

Figure 1 & Figure 2 about here

Now, we have $d = 0$ and $d' = 1$ which implies that $d_0 = 4$. Sixteen forecasts are obtained from the fitted model in (4.1) as $\hat{\mathbf{x}}_{\mathbf{2}}$. We convert $\hat{\mathbf{x}}_{\mathbf{2}}$ to $\hat{\mathbf{y}}_{\mathbf{2}}^{(1)}$ using equation (2.3) with the $f(t, \mathbf{y}^*)$ and $a(t,j)$ functions are employed according to the case of $d = 0$ and $d' = 1$ in Table 1. Since the effective sample size $(n - d_0 - 1 = 43)$ in (4.1) is small, we decide to use the tapered autocovariance function $\hat{\gamma}_M^x(k)$ in (3.4) to estimate $\mathbf{\Gamma}^{\mathbf{x}}$. The Parzen lag window in (3.5) with $M = 22$ is chosen. The remaining steps listed in the last section are carried out and the final disaggregation results are summarised in Table 2.

Table 2 about here

Define

$$\text{RMSE}(i) = \sqrt{\frac{1}{4N}\sum_{t=1}^{4N}\left[\hat{y}^{(i)}(n+t) - y(n+t)\right]^2}. \tag{4.2}$$

It should be noted that $y(n+t)$ for $t = 1, \ldots, 4N$ in this example are known. We may compute the RMSE values for $\hat{\mathbf{y}}_2^{(1)}$ and $\hat{\mathbf{y}}_2^{(2)}$ from Table 2. The results are: RMSE(1)=1.4959 and RMSE(2)=1.2424. There is a 16.95% reduction in RMSE of $\hat{\mathbf{y}}_2^{(2)}$ over $\hat{\mathbf{y}}_2^{(1)}$. It shows that the proposed disaggregation formula is able to capture the important information in $\mathbf{Z}$ and achieve significant improvement over the initial estimate.

An alternative method to handle *quarter-year* mixed-frequency problem is to aggregate the quarterly portion of the series to annual values. In this example, we have 16 yearly totals (1967-1982) after aggregation. Almon (1988) provided a method to convert annual series to quarterly figures by interpolation. Wei and Stram (1990) proposed a model-based procedure for such disaggregation. Both methods are briefly described in Chan (1993). The interpolated quarterly values by the Almon's (1988) method, $\mathbf{y}_2^{(\mathbf{A})}$, are also reported in Table 2. Unfortunately, the Wei and Stram's (1990) procedure is not applicable to our example due to the negative AR parameter as discussed in Chan (1993, pp.686). From Table 2, we can see that the information concerning the seasonal pattern of the series are eliminated during annual aggregation. It is totally unrecoverable by Almon's (1988) or Wei and Stram's (1990) methods. The RMSE(A) = 4.7376, a 381% higher than the RMSE(2).

Example 2

The real US GNP series (seasonally adjusted and in 1982 constant dollars) is selected for the demonstration. The yearly observations from 1900 to 1953 are available in Vanderhoof *et al* (1988) and the quarterly figures for 1954-87 are listed in Lutkepohl (1991, pp.500-502). We have a *year-quarter* mixed-frequency series.

In the first part of this example, we shall illustrate our proposed method to estimate the unobserved quarterly US GNP figures for the years from 1949-1953. In order to fit the setup in the proposed disaggregation formula in (2.12), we have to reverse the time axis. Therefore, our *year-quarter* US GNP series will be transformed to a *quarter-year* mixed-frequency time series. We simply set the quarterly observation of 1987.Q4 as $y(1)$, the observation of 1987.Q3 as $y(2), \ldots$, and the observation of 1954.Q1 as $y(136)$. Furthermore, we also set the annual observation of 1953 as $Z(1), \ldots$, and the observation of 1949 as $Z(5)$.

Routine model specification procedure is applied to the series $y(t)$. It concludes an AR(1) process for $x(t)$ with $(d, d') = (1, 0)$. The corresponding estimation results are

$$(1 - 0.3089B)\,\tilde{x}(t) = a(t) \qquad \hat{\sigma}_a^2 = 39.4849 \tag{4.3}$$

with 0.0822 as the standard error of $\hat{\phi}$. A residual analysis is performed and it indicates model adequacy in (4.3).

The implied autocovariance function $\tilde{\gamma}^x(k)$ from the fitted model in (4.3) is

$$\tilde{\gamma}^x(0) = 43.65 \quad \text{and} \quad \tilde{\gamma}^x(k) = 0.3089\,\tilde{\gamma}^x(k-1) \quad \text{for } k \geq 1. \tag{4.4}$$

The estimated Toeplitz matrix $\mathbf{\Gamma}^{\mathbf{x}}$ in (3.1) can be obtained from (4.4). We continue the proposed procedure by performing all the remaining steps mentioned in the last section. Finally, the time axis is reversed back and the disaggregation results are listed in Table 3. In order to be consistent and comparable to the table provided by Lutkepohl (1991, pp.500-502), the reported values in Table 3 are annualised figures (i.e., $4 \times \hat{\mathbf{y}}_2^{(2)}$).

Table 3 about here

In the second part of this example, we compare the performances of our proposed method (CC method) with other alternatives. They are the A (Almon, 1988) method and the WS (Wei and Stram, 1990) procedure. We assume that quarterly data are only available from 1964.Q1 to 1987.Q4. The disaggregation methods are applied to the years 1954-1963. The resulting quarterly values are compared with the known observed figures. Table 4 gives the RMSE of the 4 quarterly estimates for each year by various methods. The proposed method is superior to the other alternatives when the estimates are close to the forecast origin. Both the ratios RMSE(A)/RMSE(CC) and RMSE(WS)/RMSE(CC) are all greater than one for the first three years. It should be noted that the time axis has been reversed. In conclusion, the overall performance of the proposed procedure is reassuring in this example.

Table 4 about here

5. MONTE CARLO RESULTS

To further compare the performances of the CC, A and WS disaggregation methods in handling mixed-frequency time series, a small scale simulation experiment is carried out. It is clear from the beer production example in the last section that the A and WS methods are not suitable for seasonal time series. Therefore, we shall focus on non-seasonal mixed-frequency time series in this section.

Time series $y(t)$ with the length of $(100+n+4N)$ are generated from an ARIMA(1,1,0) process

$$y(t) = (1+\phi)\, y(t-1) - \phi\, y(t-2) + a(t). \tag{5.1}$$

We use the IMSL subroutine RNNOA to produce the white noise, $a(t)$, with mean zero and variance one. The first 100 observations are discarded to reduce the effects of starting values. The last $4N$ observations are aggregated to form N yearly totals. Therefore, a *quarter-year* mixed-frequency time series is generated with n quarterly values in the first portion and N annual figures in the remainder. We consider the parameter $\phi = 0.8, 0.2, -0.2$ and -0.8. For each model, the cases of $n = 80$ with $N = 1, 2, 5, 10$, and 20 are studied. The RMSE of the estimates from each method is computed over the 200 replications. The ratios of the RMSE of the two alternative methods (A and WS) to the proposed method (CC) are reported in Table 5.

Table 5 about here

Most ratios in Table 5 are greater than one. It indicates that the performances of the proposed method are usually better than the alternatives. There are remarkable improvements of using the proposed procedure over the WS and A methods when $x(t) = \nabla y(t)$ is highly correlated. The proposed method is also in general superior to the other alternatives when N is small. In other words, the CC method performs better when the estimates are close to the forecast origin (or backcast origin if the time axis has been reversed). The results are consistent with the findings in Table 4 of the US GNP example.

6. CONCLUDING REMARKS

In this paper, we apply the constrained forecast techniques in time series analysis to handle the mixed-frequency data problem. We have demonstrated that the proposed method has advantages over the other alternatives. It can handle seasonal mixed-frequency time series. Furthermore, the overall performance of the proposed procedure is stable and promising in our simulation study. The proposed method also offers greater flexibility by relieving the ARIMA model assumption which is required by the existing constrained forecast methods. The only requirement of the proposed procedure is $\nabla^d \nabla_4^{d'} y(t)$ being stationary. We may choose any forecasting method for $\mathbf{x_2}$. Robust time series forecasting techniques can also be employed when there are spurious observations in the series. If we choose a $\hat{\mathbf{x}}_\mathbf{2}$ which is a best linear unbiased estimator (BLUE) for $\mathbf{x_2}$, it can be shown that both $\hat{\mathbf{y}}_\mathbf{2}^{(1)}$ and $\hat{\mathbf{y}}_\mathbf{2}^{(2)}$ are also BLUE for $\mathbf{y_2}$.

The proposed procedure can easily be generalised to *month-year, month-quarter* and other mixed-frequency situations. Another possible extension of the method is to handle mixed-frequency problem of a stock variable. The value of a stock variable such as the price of a given commodity exists at every time point. However, in practice, it is rarely the case that every possible value of a stock variable is observed and recorded. Instead, the observed series is often only a subseries obtained by a scheme of systematic sampling. For example, in the share market, the often-reported series is the daily closing price. Therefore, the stock variable cannot be temporally aggregated. The constraint in (1.1) for a stock series may now become $Z(T) = y(n + 4T)$, if we assume that only year-end values are observed after time n. In such cases, we can still perform the procedure by changing the $\mathbf{L}$ in (2.11) to $\mathbf{L} = \mathbf{I_N} \otimes (0\ 0\ 0\ 1)$.

ACKNOWLEDGEMENTS

Most of the computations and simulations for this research was done when the first author was visiting the Department of Statistics & Actuarial Science, University of Waterloo, Canada. He is grateful to Professor R. L. Brown and the University of Waterloo for providing financial and computing support.

REFERENCES

Almon, C. (1988) *The Craft of Economic Modeling.* Boston: Ginn Press.

Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control.* Second Edition, San Francisco: Holden-Day.

Brillinger, D.R. (1981) *Time Series: Data Analysis and Theory.* San Francisco: Holden-Day.

Chan, W.S. (1993) Disaggregation of annual time-series data to quarterly figures: A comparative study. *Journal of Forecasting*, **12**, 677-688.

Chan, N.H. and Tsay, R.S. (1996) Asymptotic inference for non-invertible moving average time series. *Journal of Time Series Analysis*, **17**, 1-17.

Chen, Z.G. and Anderson, O.D. (1994) The representation and decomposition of integrated stationary time series. *Advances in Applied Probability*, **26**, 799-819.

Chen, Z.G. and Anderson, O.D. (1998) Increment-vector methodology: transforming nonstationary series to stationary series. To appear in *Journal of Applied Probability*, June, 1998.

Chen, Z.G., Cholette, P.A. and Dagum, E.B. (1997) A nonparametric time series approach for estimating and benchmarking repeated survey data. To appear in *Journal of the American Statistical Association*, December, 1997.

Chen, Z.G. and Liu, J. (1997) A note on forecasting and interpolating seasonal integrated time series. *mimeo.*

Chow, G.C. and Lin, A. (1971) Best linear unbiased interpolation, distribution and extrapolation of time series by related series. *Review of Economics and Statistics*, **53**, 372-375.

Cressie, N. (1988) A graphical procedure for determining nonstationarity in time series. *Journal of the American Statistical Association*, **83**, 1108-1116. Correction, **85**, 272 (1990).

Dickey, D.A. and Fuller, W.A. (1981) Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, **49**, 1057-1072.

Fernandez, R.B. (1981) A methodological note on estimation of time series. *Review of Economics and Statistics*, **63**, 471-476.

Granger, C.W.J. and Newbold, P. (1986) *Forecasting Economic Time Series.* Second Edition, New York: Academic Press.

Guerrero, V.M. (1989) Optimal conditional ARIMA forecasts. *Journal of Forecasting*, **8**, 215-229.

Guerrero, V.M. (1990) Temporal disaggregation of time series: an ARIMA-based approach. *International Statistical Review*, **58**,29-46.

Hillmer, S.C. and Trabelsi, A. (1987) Benchmarking of Economic time series. *Journal of the American Statistical Association*, **82**, 1064-1071.

Hylleberg, S., Engle, R.F., Granger, C.W.J. and Yoo, B.S. (1990) Seasonal integration and cointegration. *Journal of Econometrics*, **44**, 215-238.

Litterman, R.B. (1983) A random walk Markov model for the distribution of time series. *Journal of Business and Economic Statistics*, **1**, 169-173.

Lutkepohl, H. (1991) *Introduction to Multiple Time Series Analysis.* Berlin: Springer-Verlag.

McLeod, I. (1975) Derivation of the theoretical autocovariance function of autoregressive moving average time series. *Applied Statistics*, **24**, 255-256.

Nazem, S.M. (1988) *Applied Time Series Analysis for Business and Economic Forecasting.* New York: Marcel Dekker.

Osborn, D.R., Chui, A.P.L., Smith, J.P. and Birchenhall, C.R. (1988) Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, **50**, 361-377.

Priestley, M.B. (1981) *Spectral Analysis and Time Series.* New York: Academic Press.

Rossana, R.J. and Steater, J.J. (1995) Temporal aggregation and economic time series. *Journal of Business and Economic Statistics*, **13**, 441-451.

Tiao, G.C. (1972) Asymptotic behavior of temporal aggregates of time series. *Biometrika*, **59**, 525-531.

Tong, H. (1990) *Non-linear Time Series: A Dynamical System Approach.* Oxford: Clarendon Press.

Trabelsi, A. and Hillmer, S.C. (1989) A benchmarking approach to forecast combination. *Journal of Business and Economic Statistics*, **7**, 353-362.

Vanderhoof, I.T., Faye, A., Tenenbein, A. and Verin, R. (1988) The risk of asset default: Report of the Society of Actuaries C1 Risk Task Force of the Committee on Valuation and Related Areas. *Transactions of the Society of Actuaries*, **XLI**, 547-591.

Wei, W.W.S. (1990) *Time Series Analysis: Univariate and Multivariate Methods.* New York: Addison-Wesley.

Wei, W.W.S. and Stram, D.O. (1990) Disaggregation of time series models. *Journal of the Royal Statistical Society*, Series B, **52**, 453-467.

Whittle, P. (1963) *Prediction and Regulation by Linear Least-Square Methods.* Minneapolis: University of Minnesota Press.

Figure 1. Quarterly beer production in the United States, 1967-78

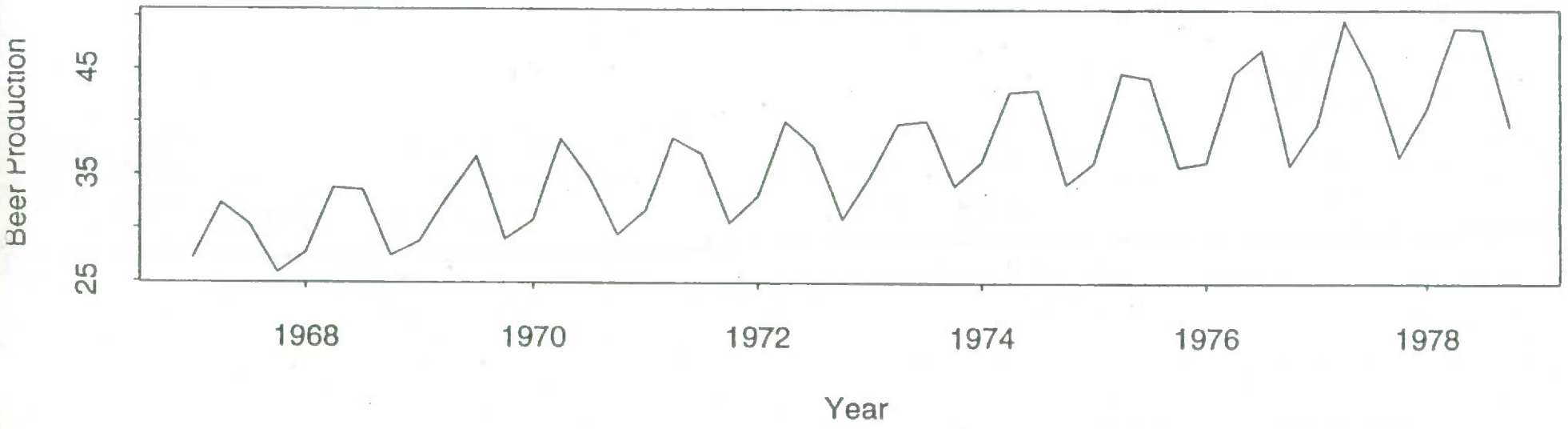
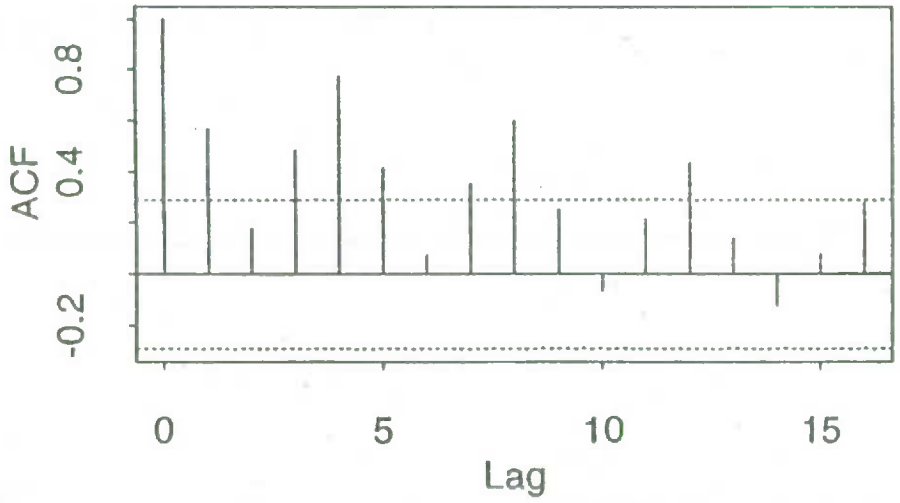


Figure 2. Model specification plots for the beer production data

(a) ACF plot for y(t)



(b) ACF plot for x(t)

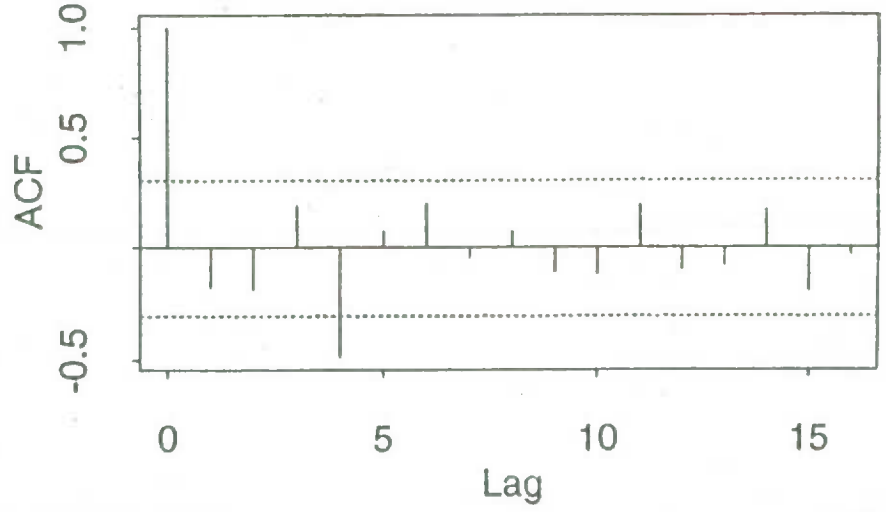


Table 1. Some commonly used cases of the general formula in (1.3)

$d = 0, d' = 0$	For $t \geq 1;\ j = 1, \ldots, t$	$f(t, \mathbf{y}^*) = 0$	$a(t,j) = \begin{cases} 1, & j = t; \\ 0, & j \neq t \end{cases}$
$d = 1, d' = 0$	For $t \geq 1;\ j = 1, \ldots, t$	$f(t, \mathbf{y}^*) = y(n)$	$a(t,j) = 1$
$d = 2, d' = 0$	For $t \geq 1;\ j = 1, \ldots, t$	$f(t, \mathbf{y}^*) = y(n) + t\{y(n) - y(n-1)\}$	$a(t,j) = 1 + t - j$
$d = 0, d' = 1$	For $t \geq 1;\ j = 1, \ldots, t$ let $h = \text{int}[(t+3)/4]$, and $i = t - 4(h-1)$	$f(t, \mathbf{y}^*) = y(n - 4 + i)$	$a(t,j) = \begin{cases} 1, & j = 4k + i - 4, \text{ for } k = 1, \ldots, h; \\ 0, & \text{otherwise} \end{cases}$
$d = 1, d' = 1$	For $t \geq 1;\ j = 1, \ldots, t$ let $h = \text{int}[(t+3)/4]$, and $i = t - 4(h-1)$	$f(t, \mathbf{y}^*) = h\{y(n) - y(n-4)\} + y(n - 4 + i)$	$a(t,j) = 1 + \text{int}[(t-j)/4]$

Note: $\text{int}[w]$ denotes the integer part of w.

Table 2. Disaggregation results for the beer production series

Year	Quarter	$\hat{y}_2^{(1)}$	$\hat{y}_2^{(2)}$	$y_2^{(A)}$	y_2
1979	1	41.89	43.44	45.35	44.29
	2	50.88	52.01	45.75	50.09
	3	47.93	48.46	46.25	48.42
	4	39.38	40.28	46.84	41.39
	Total	180.08	184.19	184.19	184.19
1980	1	43.89	45.48	48.23	46.11
	2	52.12	53.12	48.71	53.44
	3	50.76	52.47	49.01	53.00
	4	41.75	44.00	49.12	42.52
	Total	188.52	195.07	195.07	195.07
1981	1	45.02	44.46	48.39	44.61
	2	53.67	53.10	48.37	55.18
	3	51.43	52.45	48.41	52.24
	4	42.68	43.68	48.52	41.66
	Total	192.80	193.69	193.69	193.69
1982	1	46.64	45.60	48.69	47.84
	2	55.05	54.38	48.93	54.27
	3	53.31	52.98	49.23	52.31
	4	44.41	43.49	49.60	42.03
	Total	199.41	196.45	196.45	196.45

Note: (1) $\hat{y}_2^{(1)}$ is the preliminary estimate from equation (2.2)
(2) $\hat{y}_2^{(2)}$ is the final estimate from equation (2.12)
(3) $\hat{y}_2^{(A)}$ is the estimate obtained from Almon's (1988) method

Table 3. Disaggregated quarterly US GNP figures (1949-1953)

Year	Quarter				Yearly Total
	1	2	3	4	
1949	1101.0	1100.0	1108.4	1126.5	1109.0
1950	1153.5	1185.0	1219.8	1256.6	1203.7
1951	1293.2	1322.6	1342.8	1354.2	1328.2
1952	1359.6	1368.9	1384.7	1406.8	1380.0
1953	1431.1	1443.6	1441.1	1425.4	1435.3

Note: (1) All figures are annualised
(2) Observed yearly totals can be found in Vanderhoof *et al* (1988)

Table 4. Performances of various methods for the US GNP data

Year	RMSE(CC)	RMSE(A)	RMSE(WS)	RMSE(A)/ RMSE(CC)	RMSE(WS)/ RMSE(CC)
1963	1.1524	1.4471	1.3027	1.2557	1.1304
1962	1.4064	2.8997	2.0045	2.0618	1.4253
1961	2.1318	3.1116	2.5313	1.4596	1.1874
1960	4.9113	4.3006	4.4418	0.8757	0.9044
1959	2.0739	2.3360	1.9956	1.1264	0.9622
1958	5.9259	5.1903	3.2515	0.8759	0.5487
1957	1.5172	2.2930	1.6094	1.5113	1.0608
1956	0.9061	0.9065	0.7511	1.0004	0.8289
1955	0.8222	0.9148	1.2609	1.1126	1.5336
1954	4.9663	4.5312	3.3313	0.9124	0.6708

Table 5. Ratios of the RMSE of alternative methods to the CC method

Sample size		Disaggregation method	
n	N	A	WS
		(a) $\phi = 0.8$	
80	1	1.3173	1.2058
	2	1.1014	1.0963
	5	1.0024	0.9997
	10	1.0003	1.0101
	20	1.0089	0.9633
		(b) $\phi = 0.2$	
80	1	1.0371	1.0217
	2	1.0333	0.9712
	5	1.0224	0.9809
	10	0.9844	0.9897
	20	0.9711	0.9786
		(c) $\phi = -0.2$	
80	1	1.0431	NA
	2	1.0415	NA
	5	0.9823	NA
	10	1.0311	NA
	20	0.9945	NA
		(d) $\phi = -0.8$	
80	1	1.4043	NA
	2	1.3211	NA
	5	1.0985	NA
	10	1.0041	NA
	20	0.9970	NA

Note: NA means that the WS method is not applicable to these cases (see Chan, 1993, pp.686).

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010197435

- 005