UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Social Sciences

BENCHMARKING METHODS FOR REPEATED BUSINESS SURVEYS

by

Leonardo Trujillo

Thesis for the degree of Doctor of Philosophy

September, 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

SCHOOL OF SOCIAL SCIENCES

Doctor of Philosophy

BENCHMARKING METHODS FOR REPEATED BUSINESS SURVEYS

by Leonardo Trujillo

Benchmarking corresponds to a combination of two sources of information on a given variable. In many situations, the problem consists of combining a series of frequent data with a series of less frequent but more accurate data for producing more accurate estimates of the former series. For example, estimates of population characteristics are derived from the last census and researchers re-estimate the values for the time gap between two censuses using more regular information. In what follows we focus in the problem of benchmarking monthly data with annual estimates; then, the benchmarking consists of forcing the sum of the monthly signals to equal the signal of the benchmark. Alternative estimators have been proposed in the literature for benchmarking. When the adjusted series agrees exactly with these benchmarks, the benchmarking is called *binding*. The binding process is implemented by setting the variance of the annual survey errors to zero. However, it is necessary to account for the variance of the annual survey errors when computing the variances of the benchmarked estimators. In this thesis, we develop the theoretical expression of the correct variance as well as an expression for the excess in the variance due to the binding process. The results are extended to the most known benchmarking methods proposed in the literature. An application to business surveys used for official statistics in the UK is presented, illustrating some particular issues regarding the state space modelling. Finally, the problem of how to prepare tabular data classified by attributes as columns and points in time as rows is analyzed. This multivariate extension of the benchmarking problem distinguishes two basic type of problems: when only marginal totals are available (*contemporaneous disaggregation*) and when the aggregates do not correspond with the sum of the disaggregated values by year and/or by attributes (*reconciliation*). The scope of this thesis is based basically in a state space model approach.

Declaration of Authorship

I, Leonardo Trujillo, declare that the thesis entitled BENCHMARKING METHODS FOR REPEATED BUSINESS SURVEYS and the work presented in it are my own, I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- I have acknowledge all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed: [redacted]

Date: February 24th, 2008

Acknowledgements

There are so many people who have contributed in some or another way to this dissertation and may require special acknowledgment and appreciation. Firstly, I want to thank my Supervisors, *Chris Skinner* and *Danny Pfeffermann* because without their expertise, understanding and patience I would never have finished; I could never imagine having better advisors and mentors for my PhD. I also have to say a big 'thank-you' to *Fred Smith* for his valuable comments and contributions to the development of this thesis. I will extend these thanks to all the academic and support staff of the *S3RI* and the *Division of Social Statistics* at the *University of Southampton.* Much respect to my colleagues, and hopefully friends ever, *Solange Corrêa*, *Caroline Young* and *Hukum Chandra* for making the life in the office such an enjoyable experience. Thanks to *Marcel Vieira*, *Pedro* and *Denise Silva*, *Ebrahim* and *Leslie* for all those serious discussions and coffee times. I will thank also *Mac McDonald*, *James Brown*, *Fred Smith*, *Amos Channon* and *Gabrielle Durrant* for giving me some work that helped my finances during this long journey. A special recognition to Professor *Ray Chambers* when he was in the University of Southampton; *Michael Hidiroglou* and *Trevor Fenton* in the Office for National Statistics (ONS) for all their suggestions, comments and their help to obtain the appropriate databases finally used in the practical applications.

Special thanks to all my family; particularly my *Mum*, *Dad* and my siblings *Angélica* and *Germán* for all their constant support and enthusiasm. I am so happy for seeing all them again soon. Also, thanks to all the friends I have made in England during these years: *Laurence*, *Mark*, *Begoña*, *Karla*, *Pete*, *George*, *Eliana*, *Lee*, *Eloísa*, *Andy* and *Paula*; hopefully our lives will give us a chance to meet again sometime; I really hope so. Finally thanks to the *Department of Statistics* in the *National University of Colombia*; special thanks to *Fabio Nieto* and *Leonardo Bautista* who introduced me to the time series and the survey sampling areas, respectively; *Jorge Martínez* for his technical comments and also *Luis López*, *Nelcy Rodríguez*, *Germán Hernández*, *Luis F. Niño*, and *Luz M. González* for their constant motivation. This thesis was partially supported by the *Alβan Programme*, European Union Programme of High Level Scholarships for Latin America, identification number (E03D18720CO) and Colfuturo.

Contents

PAGE
NUMBERING
AS ORIGINAL

Part I

List of Tables

List of Tables

Part II

List of Figures

List of Figures

Part III

Notation

Notation

t : Index over high frequency time $t = 1, \cdots, n$

y_t : Survey estimate at time t

η_t : Population quantity of interest at time t

ℓ_t : Sampling error at time t

n : Total number of high frequency observations (months or quarters)

K : Number of high frequency periods per year (i.e. K=12 if months)

m : Total number of low frequency (annual) observations

$[x]$: Integer part function of x

$\boldsymbol{y}$: Vector n x 1 of monthly observations $\boldsymbol{y} = [y_1, \cdots, y_n]'$

$\boldsymbol{x}$: Vector m x 1 of annual observations $\boldsymbol{x} = [x_1, \cdots, x_m]'$

x_i : Annual survey estimate at year i

$\boldsymbol{\eta}$: Vector n x 1 of underlying values $= [\eta_1, \cdots, \eta_n]'$

$\hat{\boldsymbol{\eta}}$: Vector of benchmarked values $\hat{\boldsymbol{\eta}} = [\hat{\eta}_1, \hat{\eta}_2, \cdots, \hat{\eta}_n]'$

i : Index over years $i = 1, \cdots, m$

w_i : Annual BI ratios

$f(\hat{\eta}, y)$: Penalty function (function to minimize) in Denton's method

$\boldsymbol{L}$: Indicator matrix n x m converting monthly series into annual series

$\mathbf{1}_K$: K x 1 column vector of ones

$\mathbf{0}_K$: K x 1 column vector of zeroes

$\boldsymbol{M_1} \otimes \boldsymbol{M_2}$: Kronecker (tensor) product of matrices $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$

$\boldsymbol{A}$: Symmetric non-singular matrix - Denton method

$\boldsymbol{C}$: Matrix of linear combinations of the discrepancies - Denton method

$\boldsymbol{r}$: Vector of discrepancies $\boldsymbol{x} - \boldsymbol{L'y}$ - Denton method

I_n : Identity matrix of dimension n

$\triangle$: Backward difference operator $\triangle\eta_t = \eta_t - \eta_{t-1}$

d : Order of the backward difference operator

D : Matrix n x n of backward differences - Denton method

Y : Matrix n x n with the elements of y over its diagonal

A^* : Non-singular matrix in Proportional Denton Method

ν_t : Disturbance error term in a random walk process

a : Bias parameter

ℓ : Vector n x 1 of monthly errors $\ell = [\ell_1, \cdots, \ell_n]'$

e : Vector m x 1 of annual survey errors $e = [e_1, \cdots, e_m]'$

X : Model regression matrix

τ' : Vector 1 x $(n+m)$ of observations $\tau' = [yx]'$

β' : Vector 1 x $(n+1)$ of parameters $\beta' = [a\eta']$

u' : Vector 1 x $(n+m)$ of survey errors $u' = [\ell' e']$

Σ_u : General autocovariance matrix of survey errors $\Sigma_u = \text{diag}(\Sigma_\ell, \Sigma_e)$

Σ_ℓ : Autocovariance matrix of the monthly survey errors

Σ_e : Autocovariance matrix of the annual survey errors

$\sigma^2_{\hat{a}}$: Variance of $\hat{a}$

y^* : Bias-corrected observation vector $y^* = y - 1_n\hat{a}$

k_t : Weights representing heteroscedasticity of the survey errors

ℓ^*_t : Standardized survey errors $\ell^*_t = \ell_t / k_t$

Σ_{ℓ^*} : Autocovariance matrix of the standardized survey errors

W : Diagonal matrix $W = \text{diag}(k_t)$

B : Backshift operator $B(y_t) = y_{t-1}$

$\phi(B)$: Autoregressive (AR) operator

$\Phi(B)$: Seasonal autoregressive operator

$\theta(B)$: Moving average (MA) operator

$\Theta(B)$: Seasonal moving average operator

μ : Expected value of the process η_t

$\boldsymbol{\mu}$: Vector of means of the random vector $\boldsymbol{\eta}$; $\boldsymbol{\mu} = \mu \otimes \mathbf{1}_n$.

b_t, c_t : White noise processes Hillmer and Trabelsi method

$\hat{\boldsymbol{\eta}}^0$: Minimum mean squared error linear estimate de $\boldsymbol{\eta}$ given $\boldsymbol{y}$

η_c : Correction factor term Hillmer and Trabelsi method

$\boldsymbol{\Omega}$: Autocovariance matrix $Cov(\boldsymbol{\eta} \mid \boldsymbol{y})$

μ_t : Trend component in structural time series models

γ_t : Seasonal component in structural time series models

ϵ_t : Irregular component in structural time series models

σ_ϵ^2 : Variance of the irregular term

a_t : Random walk process

σ_ν^2 : Variance of the random walk disturbance

β_t : Local rate of change or slope in BSM

ξ_t : Disturbance for the local linear trend - BSM

ζ_t : Disturbance of the slope model - BSM

v : Index over high frequency time in a particular year, $v = 1, \cdots, K$

ω_t : Disturbance term for seasonality (dummy seasonality)

$\omega_{v,t}, \omega_{v,t}^*$: White noise processes (trigonometric seasonality)

κ_v : Seasonal frequency at instant v, $v = 1, \cdots, [K/2], \kappa_v = 2\pi v/K$

$\gamma_{vt}, \gamma_{vt}^*$: Cyclical components of the trigonometric seasonality term

τ_t : Trading day component in structural time series model

φ_t : Moving festival component in structural time series model

P : Dimension of y_t, $P = 1$ if y_t is an univariate series, $P > 1$ if multivariate

$\boldsymbol{y}_t$: Multivariate observed time series at instant t

$\boldsymbol{\alpha}_t$: Unobserved state vector with dimension r x 1

r : Dimension of the state vector α_t

Z_t : State matrix of dimension p x r

ϵ_t : Disturbance term of dimension p x 1 in the observation equation

T_t : Transition matrix of dimension r x r

ϑ_t : Disturbance term of dimension r x 1 in the transition equation

H_t : Covariance P x P matrix of the disturbances in the observation equation

Q_t : Covariance r x r matrix of the disturbances in the transition equation

α_0 : Initial vector state to the Kalman filter

$\hat{\alpha}_0$: Estimated initial vector state to the Kalman filter

P_0 : Covariance matrix of the initial vector state to the Kalman filter

α^*_t : New state vector after transformation $\alpha^*_t = B\alpha_t$

M : Non-singular r x r matrix to transform the original state vector

p : Autoregressive order

q : Moving average order

ϱ : Length of the ARMA model state vector, $\varrho = \max(p, q+1)$

χ_t : Disturbance term of the ARMA model

Y_t : Set of observed data until the instant t; $Y_t = (y_1, \cdots, y_t)$

$\hat{\alpha}_{t|s}$: Conditional mean of α_t based on Y_s

$P_{t|s}$: Conditional covariance matrix of α_t based on Y_s

$\hat{\alpha}_t$: Conditional mean of α_t based on Y_t, $\hat{\alpha}_t = \hat{\alpha}_{t|t}$

P_t : Conditional covariance matrix of α_t based on Y_t, $P_t = \hat{\alpha}_{t|t}$

v_t : Vector of innovations $v_t = y_t - Z_t\hat{\alpha}_{t|t-1}$

F_t : Covariance matrix of the innovations v_t

J_{t-1} : Matrix of dimension r x r in the backward recursions

K_t : Kalman gain matrix of dimension r x p, $K_t = P_{t|t-1}Z'_tF_t^{-1}$

L_t : Matrix of dimension r x r in the Kalman filter

r_t : Vector of dimension r x 1 in the backward smoothing recursions

N_t : Matrix of dimension r x r in the backward smoothing recursions

$V_{t|n}$: Covariance matrix of dimension r x r of the smoothed state vector

ψ_t : Vector of hyper-parameters

$L(y, \psi)$: Likelihood function for estimating hyper-parameters

a : Constant r x 1 vector in the model for α_0

b : Number of non-stationary components in the state vector

B : r x b selection matrix (non-stationary part)

R : r x $(r-b)$ selection matrix (stationary part)

δ : b x 1 vector of unknown quantities (non-stationary part)

λ : $(r-b)$ x 1 vector of unknown quantities (stationary part)

$\varkappa$: Arbitrary large number in a diffuse prior; $\text{Var}(\delta) = \varkappa I_b$

P_∞ : Matrix r x r for initialization of non-stationary part of the state vector

Q_0 : Covariance matrix of λ with dimension $(r-b)$ x $(r-b)$

P_* : Initial covariance matrix $(r-b)$ x $(r-b)$ stationary part of the state vector

$\tilde{v}_t$: Standardized innovations $\tilde{v}_t = v_t / \sqrt{F_t}$ (univariate case)

ι : Index over the number of structural components, $\iota = 1, \cdots, c$

c : Total number of components in the structural time series model

$Z_{\iota,t}$: Sub-observation matrix associated with the ι component

$\alpha_{\iota,t}$: Sub-state vector associated with the ι component

$\tilde{Z}_t$: Observation matrix - disturbances in the state vector

$\varpi_{s,t}$: Elements of the covariance matrix Ω

y^* : Single series incorporating both monthly and annual information

y_s : Elements of y^*

s : Index over the elements of y^*, $s = 1, \cdots, (n+m)$

$\hat{\eta}_B$: Vector of benchmarked estimators when binding

$\Sigma_{\hat{\eta}B}$: Covariance matrix of bound estimators

B^* : n x m matrix in the difference between the bound and the non-bound estimator

$\Sigma_{\hat{\eta}B}{}^c$: Corrected bound variance when $\Sigma_e \neq 0_m$

$|M|$: Determinant of matrix M

$M_1 \geq M_2$: $M_1 - M_2$ is a positive semidefinite matrix

$M_1 > M_2$: $M_1 - M_2$ is a positive definite matrix

$\hat{\eta}_{nb}$: Regression benchmarked estimator with zero bias

$\hat{t}_{y,R}$: Ratio estimator of the total of the variable y

H : Total number of strata

h : Indicator over the number of stratum $h = 1, \cdots, H$

$\hat{t}_{y,h,R}$: Ratio estimator of the total of the variable y in the stratum h

S_h : Selected sample in the stratum h

k : Index for units in S_h, $k = 1, \cdots, n_h$

n_h : Number of units in the stratum h

y_k : Value of the variable of study in the unit k

x_k : Value of the auxiliary variable in the unit k

$t_{x,h}$: Total of the auxiliary variable in the stratum h

$\hat{t}_{y,MP,t}$: Matched pairs estimator of the total of the variable y in the instant t

S_h^* : Matched sample between the instants t and $t-1$ for the stratum h,

$S_h^* = S_{h,t} \cap S_{h,t-1}$

$S_{h,t}$: Selected sample in the stratum h in the instant t

$\hat{t}_{y,h,R,t}$: Ratio estimator of the total of y in the stratum h at the instant t

$y_{k,t}$: Value of the variable of study in the unit k at the instant t

$t_{y,h,t-1}$: Total of the variable y in the stratum h in the instant $t-1$

R^2 : Coefficient of determination

S_t : Standard error of the estimate $\hat{\eta}$ at instant t

$\hat{S}_t$: Estimated standard error

k : Index over the subperiods (months, quarters) in each year $k = 1, \cdots, K$

η_{tj} : Unknown value of the signal at instant t for subsector j

$\eta_{t.}$: Sector (contemporaneous) total at instant t

$\eta_{.j(i)}$: Annual total at year i for the subsector j

η_j : Vectors of the signal in each subsector

$\eta_.$: Vector of contemporaneously aggregated signals

$\eta_{(i)}$: 1 x p stacked vectors of annual totals in a given year i

$\eta_{j(.)}$: m x 1 vectors of annually aggregated data in subsector j

$z_.$: Vector of dimension n x 1 of estimates of the monthly sector totals

$\in_.$: Vector of monthly survey errors for the total of the sector

$x_{(i)}$: Vector of annual total estimates in each subsector for the year i

$e_{(i)}$: Annual survey errors of the estimates of the annual total at year i

y^i : Joint vector of sector and annual totals per sector in year i

$y*$: Concatenated series of vectors $y^i \quad i = 1, \cdots, m$

y_{tj} : Survey estimate of the signal at instant t for subsector j

Glossary

ABI :	Annual Business Inquiry
AIC :	Akaike Information Criterion
AR :	Autoregressive
ARCH :	Autoregressive Conditional Heteroscedasticity
ARE :	Absolute Relative Error
ARIMA :	Autoregressive Integrated Moving Average
ARMA :	Autoregressive Moving Average
BI :	Benchmark to Indicator
BIC :	Bayes Information Criterion
BLS :	Bureau of Labor Statistics
BLUE :	Best Linear Unbiased Estimator
BLUP :	Best Linear Unbiased Predictor
BSM :	Basic Structural Model
CV :	Coefficient of Variation
DANE :	Departamento Administrativo Nacional de Estadística
eCV :	Estimated Coefficient of Variation
EEC :	European Economic Community
GARCH :	Generalized Autoregressive Conditional Heteroscedasticity
GDP :	Gross Domestic Product
GLM :	Generalized Linear Models
GLS :	Generalized Least Squares
GVF :	Generalized Variance Functions
IDBR :	Inter-Departmental Business Register
iidrv :	Independent Identically Distributed Random Variables

MA : Moving Average

MAPE : Mean Absolute Percentage Error

ML : Maximum Likelihood

MPE : Mean Percent Error

MPI : Monthly Production Inquiry

MSE : Mean Square Error

MPSE : Mean Square Percentage Error

NA : Not Available - Missing value

NID : Normally Independent Distributed

OLS : Ordinary Least Squares

ONS : Office for National Statistics

PAYE : Pay As You Earn

pps : Probability proportional to size

PRN : Permanent Random Numbers

RV : Relative Variance

RWN : Random Walk plus Noise Model

se : Standard error

SIC : Standard Industrial Classification

srs : Simple Random Stratified

SSM : State Space Models

TAE : Total Absolute Error

TSAB : Time Series Analysis Branch

UK : United Kingdom

VAT : Value-Added Tax

*** : Significance at 0.1%

** : Significance at 1%

* : Significance at 5%

Chapter 1

Introduction

1.1 The Benchmarking Problem

Repeated surveys are widely used in many statistical offices to obtain estimates for a set of variables at regular intervals of time and to follow their level through time. For instance, official business surveys are carried out to estimate monthly production, monthly sales or quarterly capital expenditure (National Statistics, 2004; DANE, 2006); labour force surveys are also conducted monthly to estimate the number of employed and the rate of unemployment (Holt and Skinner, 1998; National Statistics, 2001; Bureau of Labor Statistics, 2005) and, as an additional example, monthly surveys are conducted at regular intervals to measure vote preferences (Freeman, Houser, Kellstedt and Williams, 1998; Erikson and Wlezien, 1999; Chanley, Rudolph and Rahn, 2000; Yang, Goldstein and Heath, 2000).

These type of surveys are designed mainly to estimate finite population parameters such as totals and changes in totals and means over time. According to Särndal, Swensson and Wretman (1992), page 279, remark 9.9.2, the "design and estimation for such surveys may require special methods, for example, the use of time-series analysis combined with design or model based survey sampling tools.". The application of time series methods to repeated surveys was proposed with the aim of improving estimates in these surveys. Bell and Hillmer (1987*a*) and Binder and Hidiroglou (1988) make a distinction between two approaches: the *"classical sampling approach"* and the

"time series approach". Regarding the "classical sampling approach" (Tikkiwal, 1979; Wolter, 1979), the parameter of study is assumed to be an unknown constant and all the variability comes from the sampling. On the other hand, in the "time series approach", the parameter is assumed to be a random quantity produced by a stochastic process and this gives an additional "source of variability"; see for example Blight and Scott (1973); Scott and Smith (1974); Scott, Smith and Jones (1977); Jones (1980); Bell and Hillmer (1987*b*); Binder and Hidiroglou (1988); Duncan and Kalton (1988) and Pfeffermann (1991).

We will denote the value of the unobserved population true series (signal) at the time t as η_t. Using the time series approach, Smith (1978, page 208) justifies the approach thus; "... how strong is the assumption that (the parameter) η_t is an unknown constant. It implies that η_t cannot be predicted in any way from knowledge of the previous values η_{t-1}, η_{t-2}, etc. Surely in most repeated surveys the parameter would change only moderately with time, and hence knowledge of η_{t-1} would be very useful in predicting η_t. To ignore this information seems very wasteful".

Scott and Smith (1974) combined time series and sampling by considering the decomposition

$$y_t = \eta_t + \ell_t \quad t = 1, \cdots, n \tag{1.1.1}$$

where η_t is the signal at time t, ℓ_t is the sampling error associated with y_t representing the survey estimate of η_t at time t. Therefore, the equation above decomposes the observed series y_t into the signal η_t plus a noise ℓ_t with t denoting the repetition of the survey in n periods.

The estimate y_t based on the data at time t may be adjusted to increase the accuracy of the estimation of η_t. The most common adjustments made to the periodic observations are signal extraction (smoothing), interpolation, extrapolation and benchmarking (Dagum, Cholette and Chen, 1998). Signal extraction methods aim to improve the precision in the estimation of η_t (Bell and Hillmer, 1990; Pfeffermann and Bleuer, 1993; Binder, Bleuer and Dick, 1993) and interpolation (extrapolation) methods are commonly used if there are missing values within (outside) the period of

observation (Chow and Lin, 1971). In this thesis, we will focus specifically on benchmarking.

Benchmarking corresponds to the optimal combination of two sources of information on a variable (two different set of estimates, one of which, the benchmark, is more accurate than the other). In many situations, the problem consists in combining a series of high-frequency data (e.g. monthly data) and a series of less frequent data (e.g. annual data) to produce more accurate estimates of a time series for some specific flow variable. For instance, yearly estimates of population are derived from the last census and researchers re-estimate the flows for the time gap between two censuses using monthly or quarterly regional, subregional and inter-regional information (Dagum and Cholette (2006), page 3).

In the UK, as another example, results of the Annual Business Inquiry (ABI), produced by the Office for National Statistics (ONS) are normally used to improve monthly estimates from business surveys; although it is usually over a year, after the year in question, before the estimates become available. The monthly estimates are often assumed as biased due to coverage deficiencies in the sampling frame. Undercoverage is caused since new businesses are normally included in the frame with some delay. The improvement after benchmarking is achieved by assuming that the information contained in ABI is more accurate than the monthly data.

Figure 1.1 illustrates the benchmarking process using a fictitious example. The series in red corresponds to the series of original estimates coming from a monthly survey; vertical lines in black correspond to the exact date when benchmarking methods are applied using the new available information from an annual survey; and series in green corresponds to the adjusted series after benchmarking. Given this situation, it is necessary to combine the information in both series to obtain more precise estimates reflecting the true behaviour of the unknown original series (estimates are represented by the green series in Fig. 1.1). The most common aim is to improve high frequency series (e.g. monthly or quarterly), when there are low frequency (e.g. annual) benchmarks available from another more reliable survey. Typically, the low frequency series

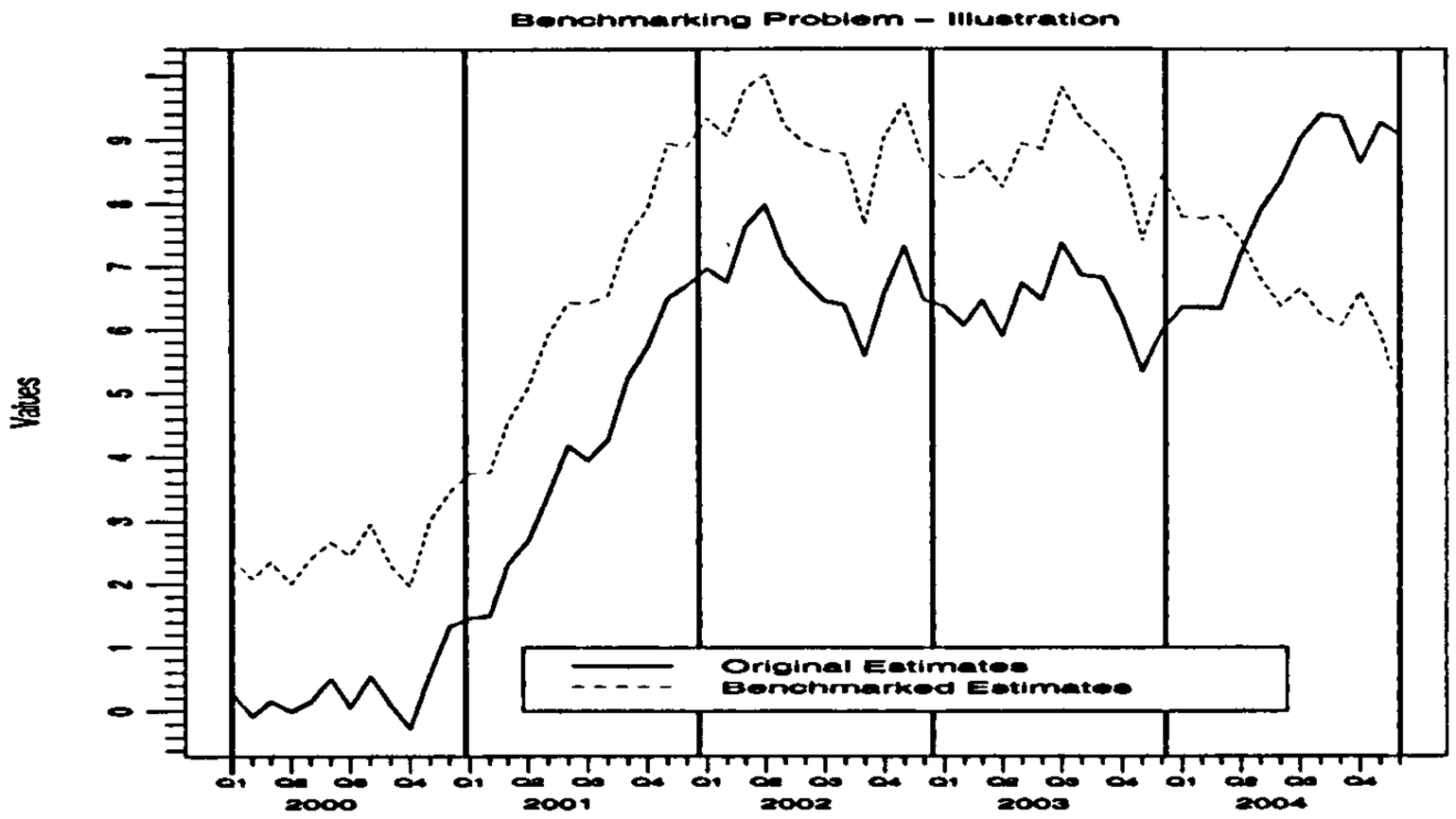


Figure 1.1. Graphical Description of the Benchmarking Problem

is more reliable than the high frequency series, because it originates from a larger sample or even a census. The more reliable measurements are then considered as benchmarks (Bloem, Dippelsman and Maehle, 2001). In most cases, the annual totals of the monthly estimates are not equal to the benchmarks. For this reason, benchmarking has been commonly considered as the process of adjusting the less reliable monthly series to make it consistent with the annual benchmarks. When the adjusted series agrees exactly with these benchmarks, the benchmarking is called *"binding"*. However, in the presence of annual survey errors, "benchmarking can be defined more broadly as the process of optimally combining two sources of measurements in order to achieve improved estimates of the signal under investigation" (Dagum et al., 1998). We will refer to the benchmarking estimation under the last definition as *non-binding estimation.*

A related problem, using the time series approach, has been named the *"disaggregation of univariate time series"* and studied by authors such as Chow and Lin (1971), Ginsburgh (1973), Fernandez (1981), Rossi (1982), Guerrero (1990), Wei and Stram (1990), Guerrero and Martinez (1995), Guerrero (2003) and Di Fonzo and Marini (2005), among others. The main purpose of this approach is to combine low frequency data from the series of study with high frequency data of auxiliary variables in order to obtain high frequency estimates of this series. For example, biennial census

of manufacturers are used to estimate annual estimates of national income in the US (Friedman, 1962). The problem of combining low frequency data from a given series with high frequency data coming from auxiliary variables is out of the scope of this thesis.

Benchmarking will be referred to in this thesis as the more general problem of improving subannual estimates derived from one source using annual estimates obtained via a second source, both including survey errors (Hillmer and Trabelsi, 1987; Cholette and Dagum, 1994; Laniel and Fyfe, 1990; Chen, Cholette and Dagum, 1997; Durbin and Quenneville, 1997). The benchmarking problem has also been called *ex-post estimation* from the point of view that the low frequency estimates are produced after (*post*) observing the benchmarks. Another different problem, called *ex-ante estimation* is how to do these adjustments before (*ante*) the most recent benchmark becomes available (Nieto, 1998; Nieto, 2007). In business surveys, for example, the annual total estimates are obtained only in the middle or the end of the following year. Then, the problem consists in how the estimation in the last few data points, in the partial year at the end of the series, can be improved before the benchmark estimate is obtained. Durbin and Quenneville (1997) have proposed an "online procedure", where it is not necessary to have the benchmark for the last year of study.

Another common problem, which can be seen as a multivariate extension of the benchmarking problem, is how to improve estimation in the cells of a table of data (for example, months by rows and industrial subsectors by column). The problem arises when the information is available in an aggregate form only (annually and/or by industrial sectors) or when the aggregates do not correspond with the sum of the disaggregated values because, for example, aggregate and disaggregate values come from different surveys or sources of information. For instance, aggregated estimates are available in National Accounts in two forms (Guerrero and Nieto, 1999): *temporally* (e.g. annually) and *contemporaneously* (e.g. by economic sector at a given time) but there is a necessity to get disaggregated estimates to carry out econometric modelling and for making decisions about some particular sectors or regions. Telser (1967), Zellner and Mornmarquette (1976), Abraham (1982), Lütkepohl (1984), Wei and Stram (1990),

among others, discuss the problems of drawing conclusions from econometric analyses at disaggregated levels when the data are temporally aggregated.

Two special cases of the multivariate problem may be distinguished :

1. *Reconciliation of time series* corresponds to the multivariate problem when a set of preliminary series obtained from a source subject to survey error is available for each subperiod and every subsector. Dagum and Cholette (2006), chapters 12 and 14, refer to this problem as *reconciling one-way or two-way classified systems of time series.* The aggregates by row and columns do not correspond with the sum of the preliminary series because the aggregate series could come from a different survey. The aim is to make this information consistent using both the auxiliary information contained in the history of the series and the information contained in the marginals.

2. *Contemporaneous and Temporal Disaggregation* corresponds to the multivariate problem when no prior auxiliary information about the subsectors or the subperiods is available; the only available information is the annual and sector totals. Dagum and Cholette (2006), chapter 13, refer to this problem as *reconciling marginal two-way systems* and set up the problem as a contingency table with for example "type of industry" in the rows vs "province" in the columns with available marginal totals but missing information in the inner cells. In this thesis, the same problem is studied but considering one of the dimensions as an index over "time";

In the first case, (Zaier and Trabelsi, 2007) have proposed a method to estimate the inner cells. However, this method does not provide any estimation of the standard error of the estimates and does not produce estimates for the first year of observation. Also, an adaptation of the Iterative Proportional Fitting (IPF) method (Deming and Stephan, 1940) is considered for this problem. In the second case, some alternatives (Di Fonzo, 1990; Guerrero and Nieto, 1999; Quenneville and Rancourt, 2005; Dagum and Cholette, 2006) have been recently studied but they either use auxiliary information from other highly correlated sources or make use of difficult assumptions such

as knowing the autocovariance matrices of the stochastic processes involved. In this thesis, new alternatives are proposed to obtain the estimates under the problem being described above through state space models in Chapters 5 and 6. The advantages and shortcomings of these proposed methods will also be discussed later on.

1.2 Aims of the Thesis

There are some desirable characteristics that a benchmarking method should have apart from the natural one as a solution to the consistency of high frequency series with low frequency benchmarks. The first one, which was mentioned above, is the capacity to deal with situations where the indicator series extends into a period for which there is no benchmark yet available; but also, preserving as much as possible the short term movements in the signal and ensuring that the sum of the sub-periods of the current year are as close as possible to the annual benchmarks. We will consider the problem of *benchmarking* as how to improve subannual estimates derived from one source by using annual estimates obtained from a second source, with both estimates (annual and subannual) subject to survey errors. In practice, it is also important to deal with specific problems such as incomplete or not available standard error of the survey estimates and specific issues under the state space model approach such as optimal specification of trends, seasonalities and ARMA modeling of the survey errors; maximum likelihood estimation of hyperparameters; goodness of fit tests and estimation of the variance of the estimators. Other problems not considered in this thesis and possible areas of further work are: missing data; multiplicative structure of the data when for example, the amplitude of the seasonal cycles increases or decreases jointly with the trend; specification of trading days/calendar effects and estimation of the survey bias.

Regarding the multivariate case, a first problem of estimating a set of monthly series for some specific subsectors of a whole industry is considered by using yearly totals for each subsector and monthly values of the total sector of industry. The estimated high frequency time series must fulfill temporal (by year or columns) and contemporaneous

(by sector or rows) aggregation constraints in the binding case; but also it is relevant to study the case when the totals are obtained from different sources subject to survey errors. Regarding the second problem in the multivariate case, Quenneville and Rancourt (2005, page 1) refer to this situation as "restoring the additivity of a system of time series, with the objective of balancing a table of seasonally adjusted series benchmarked to the corresponding annual totals from the raw series". When auxiliary disaggregated information exists, it is preferable to employ a disaggregation procedure that combines all available (aggregated and disaggregated) information rather than working only with aggregated data. The main aim is to restore the additivity to the table in order to keep the implied constraints by row and columns. Again, data can be considered from sample surveys and it is necessary to introduce the survey errors into the model.

1.3 Outline of the Thesis

The thesis is structured as follows. Chapter 2 reviews the available benchmarking methods that have been proposed in the literature. Specifically, some theoretical developments from three main benchmarking methods due to Denton (1971), Cholette and Dagum (1994) and Hillmer and Trabelsi (1987) are presented along with their main advantages and disadvantages being highlighted. Chapter 3 introduces basic concepts in structural time series and state space models and some additional sections referring to special issues about the Kalman filter, maximum likelihood estimation, initialization of the recursions and diagnostic checking. In addition to the other benchmarking methods presented in Chapter 2, two alternative state space model based methods due to Durbin and Quenneville (1997) are presented. The last sections of this chapter concentrate on the use of binding and non-binding estimators (they were introduced in page 4 above). It will be shown, at the end of the chapter, that the use of binding estimators, in the case of non-zero variance annual estimates, adds an additional component to the variance of the benchmarked values. The theoretical expression of the correct variance in this case is presented as well as an expression for the excess in the variance due to the binding. In particular for the two stage benchmarking model and under some

specific conditions, the estimates after binding could be even worse than the smoothed estimates without benchmarking. The results are extended to the other benchmarking methods presented in the previous chapters. Chapter 4 presents an overview of the main business surveys used in official statistics in the UK and describes the corresponding main parameters of study, the sampling designs and the need to apply benchmarking methods in this particular kind of surveys. At the end of the chapter, state space models are applied in order to benchmark the two main business surveys in the UK: the MPI, Monthly Production Inquiry and the ABI, Annual Business Inquiry. The information is available for most of the industrial sectors in the economy but some particular issues require to be solved before benchmarking. One particular problem is the non-availability of measures of precision for some periods of study in the survey and then generalised variance functions (GVF, Wolter (1985)) are used to overcome this problem. Other recommendations in terms of the specification of the trend and the seasonalities of the model are suggested and compared with those proposed by Durbin and Quenneville (1997) accordingly to the assumptions of the respective models. Additional issues such as initial values and constraints in the maximum likelihood estimation; initialisation methods for the filter; diagnostic tests over the innovations and the auxiliary residuals (Kohn and Ansley (1989), Durbin and Koopman (2001)) and Monte Carlo simulation of state space models are also considered.

Chapters 5 and 6 consider the multivariate extension of the benchmarking problem. The concern here is how to produce tabular data in a consistent and efficient way to get publishable values complying with both annual and contemporaneous restrictions. Two different situations are studied: the contemporaneous disaggregation with missing values case and the reconciliation case (as they were introduced in pages 5 and 6 above). The solutions for these two problems are presented using State Space Models (SSM); the reconciliation problem in Chapter 5 and the contemporaneous disaggregation case in Chapter 6. Additionally, a simulation has been carried out in both cases by an underlying model for the high frequency series that follows a random walk plus noise (RWN) process. In the reconciliation problem an AR(1) model is assumed to the survey errors. The results are presented for the binding case to check the consistency of the results. However, the method deals with both binding and non-binding cases. A

comparison of the estimates in both situations using the proposed methods and others proposed in the literature is presented in chapters 5 and 6. Chapter 7 presents some conclusions, a general overview of the final results of the thesis and possible directions for areas of future work.

Chapter 2

Benchmarking Methods

2.1 Preliminaries

This chapter presents some of the available benchmarking methods in the statistical literature and analyses their strengths and deficiencies. The chapter is structured as follows: firstly, some basic notation to be used throughout this dissertation will be given and then a review of the existing methods. The benchmarking problem, as described in the methods in this chapter, assumes the existence of two different series for the same variable but measured in different frequencies in time. The aim is the optimal combination of the information in the two series. Considering the most common case, the low frequency series will be considered as an annual series and it will be assumed that the high frequency series is observed over time periods of which there are K per year, (e.g. $K = 4$ or $K = 12$ depending on whether it is quarterly or monthly data respectively).

Let n be the length of the observed subannual series and m be the length of the series of annual benchmarks. When a benchmark is available for the last subannual observations, $n = mK$. Considering the most general case when the information for the last year is not necessarily complete, $m = [n/K]$ is the number of complete years with $[x]$ denoting the integer part of x. For instance, consider the situation when $n = 24$ quarterly observations ($K = 4$), this implies that $m = 6$ and $n = mK$; but

if, for example, $n = 26$ quarterly observations, m will denote the number of complete years being equal to [26/4], which is also equals to 6 years. The last year in this case is incomplete and the last two observations in the series do not have an available corresponding benchmark. In what follows we consider monthly and yearly estimates.

The values of the monthly estimates will be represented by the column vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]'$ and the values of the annual estimates will be represented by $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]'$. Let $\boldsymbol{\eta} = [\eta_1, \cdots, \eta_n]'$ denote the underlying signal (the true time series without survey errors). The annual series $\boldsymbol{x}$ is generally obtained from a different source (survey or administrative record) than the monthly series. Treating $\boldsymbol{y}$ as $\boldsymbol{\eta}$ subject to error; the benchmarking problem is how to adjust the vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]'$ to obtain a new more reliable vector of estimates $\hat{\boldsymbol{\eta}} = [\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_n]'$ using the information contained in $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]'$. In this chapter, we will consider the series $\boldsymbol{y}$ as the subannual values from a flow series. In other words, the yearly values of $\boldsymbol{x}$ should correspond to the yearly sums of the values in $\boldsymbol{y}$ for the corresponding year (we will refer to this kind of estimation where the annual sums of the monthly values will correspond exactly to the yearly values as *binding estimation*).

2.2 Quadratic Minimization Approach

Denton (1971) proposed a numerical benchmarking method according to the "principle of movement preservation" (Bloem et al., 2001, section 6.A1.6). This principle requires a benchmarking method with the following conditions:

(a) the variations in the subannual adjusted series will be close to those obtained in the subannual observed series, and

(b) the sums of the K subannual values by year are equal to the observed annual benchmarks in the corresponding year

Denton (1971) expressed mathematically condition (a) as the problem of minimising the differences between the adjusted subannual series ($\hat{\boldsymbol{\eta}}$) and the observed suban-

nual series (y). On the other hand, condition (b) can be expressed by the restriction $\sum_{(i-1)K+1}^{iK} \hat{\eta}_t = x_i$ with $i = 1, \cdots, m$.

One way to fulfill the latter condition is to equally distribute the difference between the annual value and the sum of the subannual values for the corresponding year among the subannual periods. Another possibility is to distribute the value of the annual benchmark across the subperiods as follows:

$$\hat{\eta}_t = y_t \times \frac{x_i}{\sum_{(i-1)K+1}^{iK} y_t} = y_t \times w_i \quad t = 1, \cdots, n \tag{2.2.1}$$

with i being the corresponding year for the observation t; $i = 1, \cdots, m$. The benchmarking procedure applying Equation 2.2.1 is known as "*prorata*". Notice that from Equation 2.2.1 is also possible to write $w_i \sim \frac{\hat{\eta}_t}{y_t}$. For this reason, the factors w_i's are commonly called the "Benchmark to Indicator (BI) ratios" (Maitland-Smith, 2002) and they could be used as a measure of "bias".

The prorata method is a good choice to benchmark the monthly series when it is possible to assume that the observed series and the target series have similar behaviours, i.e. close variations in the subperiods and similar seasonalities. This method is also acceptable when the BI ratio is approximately constant from year to year. If, however, BI ratios for consecutive years are very different and the prorata method is used, a discontinuity in the growth rate from the last subannual period in one year to the first in the next year will be introduced. This is known in the literature as "the step problem" (Bloem et al., 2001, section 6.16). A simple situation illustrating this problem will be illustrated in example 2.1 in the next subsection.

2.2.1 Denton's Method

Denton's (1971) method uses least squares optimization as a method to benchmark a monthly series according to annual totals for the same variable. The problem is formulated mathematically as minimizing a penalty function of the differences between the adjusted monthly series and the observed monthly series subject to the benchmark

constraints. Using the notation at the beginning of this chapter, and assuming $f(\hat{\eta}-y)$ is the function to be minimized; the problem consists of estimating $\hat{\eta}$ in such a way that $f(\hat{\eta}-y)$ is minimized subject to

$$\sum_{(i-1)K+1}^{iK} \hat{\eta}_t = x_i \quad i = 1, \cdots, m \tag{2.2.2}$$

The latter restriction can be written in a matrix form as $L'\hat{\eta} = x$ where

$$L_{nxm} = \begin{bmatrix} 1_K & 0_K & \cdots & 0_K \\ 0_K & 1_K & \cdots & 0_K \\ & \cdots\cdots\cdots & & \\ 0_K & 0_K & \cdots & 1_K \end{bmatrix} = I_m \otimes 1_K \tag{2.2.3}$$

where 1_K and 0_K are K x 1 column vectors in which each element is 1 or 0 respectively, L' is the transpose of L and $\otimes$ represents the Kronecker product. The elements of the Kronecker product $M_1 \otimes M_2$ of matrices M_1 and M_2 of dimension a x b and c x d, respectively, are given by $m_{1ij}M_2$ with m_{1ij} being the elements of matrix M_1. The resulting product has dimension ac x bd. The results of the Denton's method are summarized in Proposition 2.2.1.

Proposition 2.2.1. *Let $f(\hat{\eta}-y)$ be the quadratic form represented by $(\hat{\eta}-y)'A(\hat{\eta}-y)$ with A being a non-singular symmetric matrix of order n. The minimum of $f(\hat{\eta}-y)$ subject to the restriction 2.2.2 is obtained when*

$$\hat{\eta} = y + Cr \tag{2.2.4}$$

where $C = A^{-1}L(L'A^{-1}L)^{-1}$ and $r = x - L'y$.

Denton (1971) sets up a Lagrangian expression to achieve this result. All the mathematical details not included in Denton (1971) are shown in Appendix A.1.

The consideration of different matrices A produces different solutions. For instance, the choice of $A = I_n$ minimizes the differences between $\hat{\eta}$ and y according to the restriction 2.2.2, with I_n the identity matrix of dimension n, and then $C = (1/K)L$, which means the solution coincides with the method of equally distributing the discrepancies. Another alternative is, for example, to minimize the distance between the first or higher

order differences of the original and adjusted series. In that case, the penalty function can be expressed as (Denton, 1971)

$$f(\hat{\eta}, y) = \sum_{t=1}^{n} (\Delta^d \hat{\eta}_t - \Delta^d y_t)^2 = \sum_{t=1}^{n} [\Delta^d (\hat{\eta}_t - y_t)]^2 \tag{2.2.5}$$

where Δ is the backward difference operator $\Delta y_t = y_t - y_{t-1}$ and Δ^d denotes the application of this operator d times.

The vector of first backward differences may then be expressed as $D(\hat{\eta} - y)$, where D is the nxn matrix given by

$$D = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ & & \cdots\cdots\cdots & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \tag{2.2.6}$$

If $d = 1$, a restricted minimization is done over the distance of the first differences and the quadratic form to be minimized, subject to the annual constraints, is now $(\hat{\eta} - y)'D'D(\hat{\eta} - y)$ where $A = D'D$. In general, one could consider a more general quadratic form as for example $(\hat{\eta} - y)'D'MD(\hat{\eta} - y)$ with M an arbitrary matrix. It is also possible to specify the penalty function in terms of the distances between higher-order differences of the original and adjusted series making $A = \underbrace{D'D' \cdots D'}_{h \text{ times}} \underbrace{D \cdots DD}_{h \text{ times}}$. In all the cases, the benchmarked estimates are obtained by replacing the corresponding value of A in Equation 2.2.4.

Denton's method is considered as a pure numerical method. Even though the annual and monthly information could be obtained through periodic surveys, this method does not include any information about the survey errors. In other words, only binding estimators are considered. Later on, in this chapter, other benchmarking methods such as Hillmer and Trabelsi (1987), Cholette and Dagum (1994), Durbin and Quenneville (1997) will account for survey errors permitting the calculation of binding and non-binding estimates and their respective confidence intervals.

Regarding the "step problem"; i.e. the appearance of big discontinuities in the BI ratios from one year to the next, Denton (1971) proposed an alternative solution applying proportional differences. These alternative methods will be explained in section 2.2.1. Example 2.1 shows an example of the application of the prorata and Denton methods and highlights the presence of the step problem.

Example 2.1

Table 2.1 shows a fictitious example of series consisting of two years of values: 300 is the annual estimate for the first year and 500 for the second one. The example is similar to the one appearing in Denton (1971) but it has been constructed in a way which permits to appreciate more clearly some problems in the estimation. In this new example, the quarterly estimates for the first year in the third column do not add to 300.

Annual Totals	Quarter	Original y	Prorata	Equal Distr $A = I$	1st Diff $A = D'D$	2nd Diff $A = D'D'DD$
Year 1 (300)	1	80	48 *(0.60)*	30 *(0.38)*	45 *(0.56)*	57 *(0.71)*
	2	100	60 *(0.60)*	50 *(0.50)*	45 *(0.45)*	51 *(0.51)*
	3	190	114 *(0.60)*	140 *(0.74)*	130 *(0.68)*	125 *(0.66)*
	4	130	78 *(0.60)*	80 *(0.62)*	80 *(0.62)*	67 *(0.52)*
Year 2 (500)	1	80	80 *(1.00)*	80 *(1.00)*	57 *(0.71)*	34 *(0.42)*
	2	100	100 *(1.00)*	100 *(1.00)*	97 *(0.97)*	82 *(0.82)*
	3	190	190 *(1.00)*	190 *(1.00)*	200 *(1.05)*	205 *(1.08)*
	4	130	130 *(1.00)*	130 *(1.00)*	146 *(1.12)*	179 *(1.38)*

Table 2.1. Application of prorata and Denton's Method for three different penalty functions. BI ratios in parentheses.

The prorata and Denton methods were applied to benchmark the original series in the third column to the benchmarks of 300 and 500 using different penalty functions. In the fourth column, a prorata method was applied using Equation 2.2.1. It can be observed that the use of $A = I$ in the fifth column makes the difference 300-500 = -200 for the first year (the second year has not got any change in the third column) to be equally distributed in the corresponding four periods. The step problem becomes evident in the prorata (fourth) column as the BI ratios show a big discontinuity from the last period in the first year to the first value in the second year. In fact, they have approximately the same adjusted value, even though they come from very different original values. The same happened in the fifth column using $A = I$. The alternatives $A = D'D$ and $A = D'D'DD$ produce more smoothed BI ratios according to the plots in Figure 2.2.

2.2.2 Proportional Denton Method

Discrepancies from one year to the next could be smoothed using proportional differences as proposed in Denton (1971). This alternative is still a numerical method rather than a statistical one because it does not consider the survey errors, but is a good alternative to deal with the "step problem". In fact, according to Gubman and Burck (2005), this is the method most applied by statistical agencies around the world due to its simplicity.

The proportional Denton Method considers a penalty function in terms of proportionate differences between the adjusted and the observed series instead of arithmetic differences. The proportionate difference in period t is defined as $(\hat{\eta}_t - y_t)/y_t$. Defining the "subannual BI ratios" as $\hat{\eta}_t/y_t$, the idea of preserving the proportional changes in the series is equivalent to preserve the subannual BI ratios. This is because

$$\frac{\hat{\eta}_t - y_t}{y_t} - \frac{\hat{\eta}_{t-1} - y_{t-1}}{y_{t-1}} = \frac{\hat{\eta}_t}{y_t} - \frac{\hat{\eta}_{t-1}}{y_{t-1}} \tag{2.2.7}$$

Defining Y as the nxn diagonal matrix with the elements of the vector y in the diagonal, the function to minimise can be expressed in the form $(\hat{\eta} - y)'A(\hat{\eta} - y)$ with $A = Y^{-1}A^*Y^{-1}$ for some non-singular matrix A^*.

From Proposition 2.2.1., we follow that

$$\hat{\eta} = y + YA^{*-1}YL(L'YA^{*-1}YL)^{-1}r \tag{2.2.8}$$

is obtained. In 2.2.4 the adjustment was independent of the values in the observed series being adjusted, in 2.2.8 the adjustment depends on the original values. Bloem et al. (2001, Annex 6.1, section B2) propose some extensions to the proportional Denton technique.

Example 2.1 (continued)

Table 2.2 summarizes the results for the data in Example 2.1 using the Proportional Denton Method. The results in table 2.2, in this very particular case, suggest that the application of a second order proportional difference seems to ameliorate the step problem and make the variations in the adjusted series closer to those in the original one.

Annual Totals	Quarter	Original y	Prop. Diff $A = I$	1st Diff(Prop. Diff) $A = D'D$	2nd Diff(Prop. Diff) $A = D'D'DD$
Year 1 (300)	1	80	62 *(0.78)*	61 *(0.76)*	68 *(0.85)*
	2	100	71 *(0.71)*	61 *(0.61)*	65 *(0.65)*
	3	190	86 *(0.45)*	100 *(0.52)*	100 *(0.53)*
	4	130	81 *(0.62)*	78 *(0.60)*	67 *(0.52)*
Year 2 (500)	1	80	80 *(1.00)*	63 *(0.79)*	50 *(0.63)*
	2	100	100 *(1.00)*	94 *(0.94)*	82 *(0.82)*
	3	190	190 *(1.00)*	200 *(1.05)*	200 *(1.05)*
	4	130	130 *(1.00)*	143 *(1.10)*	108 *(0.83)*

Table 2.2. **Application of Proportional Denton Method for three different penalty functions. BI ratios in parentheses.**

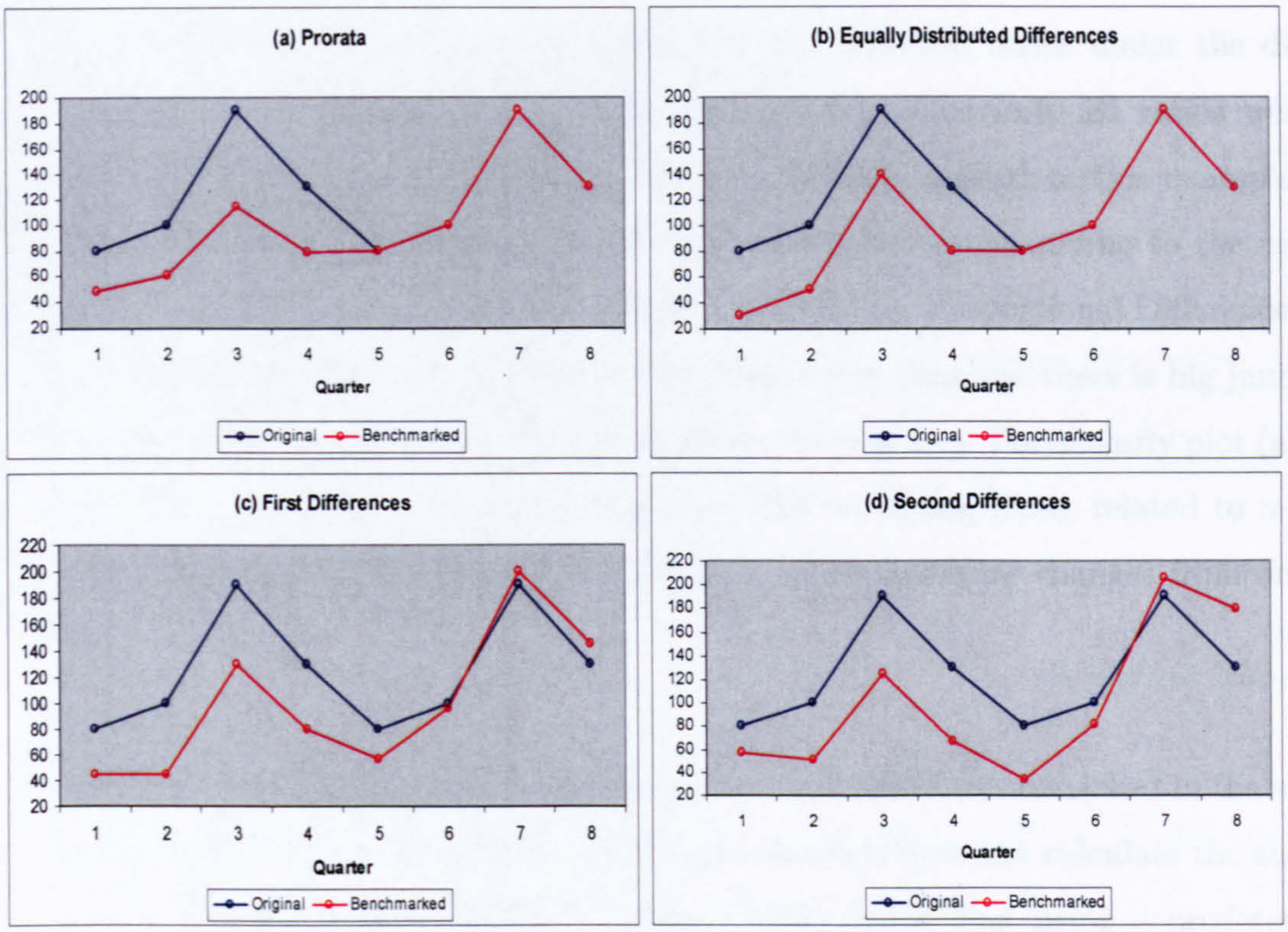


Figure 2.1. **Plots of Original and Benchmarked Data in Example 2.1 under the Denton Method**

Figure 2.1 shows a plot of the original and benchmarked series under the different variations of the Denton's method in Example 2.1. Quarterly BI ratios were also calculated and plotted in Figure 2.1 for each method applied in this example. The step problem becomes apparent for the BI ratio plots corresponding to the methods (a) Prorata, (b) Equally Distributed Differences and (e) Proportional Differences. The adjustments in these plots were done separately every year and there is big jump from the last quarter in year one and the first quarter in year two. Particularly plot (a) looks exactly as the mathematical step function. The remaining plots, related to methods using differences (absolute and proportional), show smoother changes from one year to the next.

However, some difficulties with the Denton method have been remarked in the statistical literature. The main difficulty is that this method does not calculate the standard error of the estimates. Besides, Cholette (1984) stated that using a predetermined value for the backward difference operator in Equation 2.2.5 could cause distortions to the benchmarked series. As an alternative, Bloem et al. (2001, Equation 6.3) considers the minimization of a function over the differences of the quarterly (or monthly) BI ratios from $t = 2$. Additionally, although the Denton method is a numerical procedure without any statistical criteria to be evaluated, Laniel and Fyfe (1990) have presented the proportional Denton method in statistical terms. They assume the model $\frac{\hat{\eta}_t}{y_t} = \frac{\hat{\eta}_{t-1}}{y_{t-1}} + \nu_t$, equivalent to

$$\frac{\hat{\eta}_t - y_t}{y_t} = \frac{\hat{\eta}_{t-1} - y_{t-1}}{y_{t-1}} + \nu_t \tag{2.2.9}$$

according to Equation 2.2.7 and ν_t being a sequence of uncorrelated and identically distributed errors with mean zero and constant variance and $\hat{\eta}_t$ values subject to the restriction in Equation 2.2.2. This formulation let them to conclude that assuming that a relative bias follows a random walk process and assuming that the subannual and annual data are observed without sampling errors is a condition unlikely to be satisfied by economic time series. Finally, the big difficulty with this method is that it is set up only in a binding scenario (see page 4) and one cannot calculate the variance of $\hat{\eta}$.

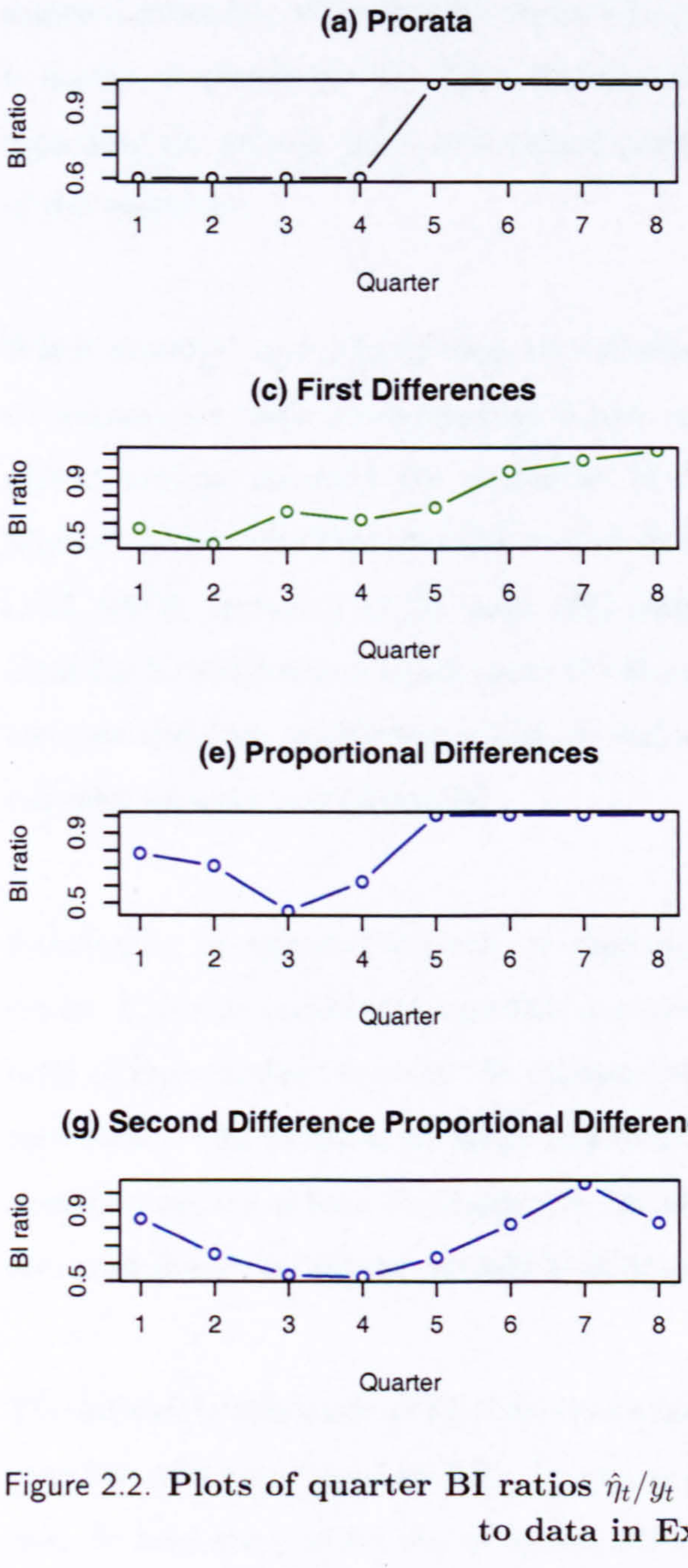


Figure 2.2. **Plots of quarter BI ratios $\hat{\eta}_t/y_t$ for the benchmarking methods applied to data in Example 2.1**

2.3 Regression Approach

Denton (1971) is more concerned that the variations (e.g. month-to-month changes and seasonal patterns) of the original series are not very badly affected after the adjustment is made. A problem with this method is that it does not include any information regarding the survey errors and cannot produce any estimates of the standard errors of the estimates.

When monthly and annual data are obtained from repeated surveys, it is desirable to account for their corresponding survey errors. Therefore, benchmarking methods should capture not only the properties of the survey errors associated with the estimated time series but also the special characteristics of the survey design. Bloem et al. (2001, section 6.A1.39, page 106) state that "...the survey design may provide identifiable information about parts of the stochastic properties of the series. Clearly, incorporating any such information, if available, in the estimation procedure may potentially improve the estimates".

Specifically, in repeated surveys, overlapping of the samples generates autocorrelated errors. It is also common in repeated surveys that the coefficients of variation (i.e. the ratio of the standard error of the estimator to its expected value) are nearly constant over time. This implies, in many cases, heteroscedastic survey errors. Other issues such as presence of bias, non-response, births and deaths in the sample frame must be also considered during the modelling (Cholette and Dagum, 1994).

The annual benchmarks are generally assumed more precise and less biased than the monthly estimates, because they are coming from censuses or surveys of bigger sample size. In business surveys, for example, the sub-annual estimates are often biased due to undercoverage in the sampling frame. This is caused by the delay in the inclusion of new businesses in the sampling frame monthly. This problem is less common in annual surveys (Laniel and Fyfe, 1990). Cholette and Dagum (1994) introduce a benchmarking

method which extends Denton's method. Their method not only takes into account the subannual and annual survey errors, not considered by Denton's method, but also considers special characteristics of the survey data such as the presence of bias in the original series and presence of autocorrelations and heteroscedasticity in the survey errors.

The Cholette and Dagum's method consists of the generalised regression model with autocorrelated survey errors given by

$$\begin{aligned} y_t &= a + \eta_t + \ell_t, \quad t = 1, \ldots, n \\ x_i &= \sum_{(i-1)K+1}^{iK} \eta_t + e_j, \quad i = 1, \ldots, m \end{aligned} \tag{2.3.1}$$

The first equation in 2.3.1 coincides with Equation 1.1.1 in the last chapter plus an additional constant term a to denote a bias parameter to be estimated. The estimates of η_t will correspond to the benchmarked series. The consideration of this bias parameter as a constant term will be discussed later on this section. Using the same notation from the first chapter, the ℓ_t's represent the monthly survey errors affecting the observations and they may have a general covariance structure resulting from the overlapping of the samples. Also, it is assumed that $E(\ell_t) = 0$ for all $t = 1, \cdots, n$. The second equation in 2.3.1 coincides with the Denton's restriction in Equation 2.2.2 with the addition of the term e_t denoting the corresponding annual survey error.

Using the notation in page 12 and denoting by $\boldsymbol{\ell}$ and $\boldsymbol{e}$ the subannual and annual survey error vectors, respectively; we also denote $\boldsymbol{\eta}$ the monthly vector of parameters (considered as fixed non-stochastic quantities). Then, the model in Equation 2.3.1 can be written in a matrix form as follows

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n & I_n \\ \mathbf{0}_m & L' \end{bmatrix} \begin{bmatrix} a \\ \boldsymbol{\eta} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\ell} \\ \boldsymbol{e} \end{bmatrix} = X \begin{bmatrix} a \\ \boldsymbol{\eta} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\ell} \\ \boldsymbol{e} \end{bmatrix} \tag{2.3.2}$$

$$E(\boldsymbol{\ell}) = \mathbf{0}, E(\boldsymbol{e}) = \mathbf{0}, E(\boldsymbol{\ell}\boldsymbol{\ell}') = \boldsymbol{\Sigma}_{\boldsymbol{\ell}}, E(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{\Sigma}_{\boldsymbol{e}}, E(\boldsymbol{e}\boldsymbol{\ell}') = \mathbf{0}$$

where $\mathbf{1}_n$ is a vector n-dimensional of ones, $\mathbf{0}_m$ is a vector m-dimensional of zeroes, $\boldsymbol{L}$ is the n x m design matrix in Equation 2.2.3, a is a bias parameter and $\boldsymbol{X}$ is a $(n+m)$ x $(n+1)$ matrix. In the last equation, the vectors $\boldsymbol{\ell}$ and $\boldsymbol{e}$ are assumed mutually

independent as the monthly and annual data come from two independent separate surveys. Cholette and Dagum (1994) expressed the model in Equation 2.3.2 as

$$\tau = X\beta + u, E(u) = 0, E(uu') = \Sigma_u \tag{2.3.3}$$

where $\tau' = [y', x']$, $\beta' = [a, \eta']$, $u' = [\ell', e']$, Σ_u is a block diagonal matrix with blocks Σ_ℓ and Σ_e, although Σ_ℓ and Σ_e are not necessarily diagonal matrices. The benchmarked estimator of this method is given in Proposition 2.3.1.

Proposition 2.3.1. *The BLUE estimators of the parameters a and η are respectively given by*

$$\hat{a} = -\sigma^2_{\hat{a}} 1' L(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y) \tag{2.3.4}$$

and

$$\hat{\eta} = y^* + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y^*), \quad y^* = y - 1_n\hat{a} \tag{2.3.5}$$

with respective variances

$$\sigma^2_{\hat{a}} = 1/[1'L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1] \tag{2.3.6}$$

and

$$\begin{aligned}\Sigma_{\hat{\eta}} =& [\Sigma_\ell - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell] \\ &+ [I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']1\sigma^2_{\hat{a}}1'[I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']'\end{aligned} \tag{2.3.7}$$

The proof is achieved using standard results for GLS and partitioned matrices and all the details are included in Appendix A.2.

As noted before the survey errors may be heteroscedastic. Cholette and Dagum (1994) dealt with this problem by expressing ℓ_t as

$$\ell_t = k_t \ell_t^* \tag{2.3.8}$$

where the k_t's are weights representing heteroscedasticity over time and it is assumed that the ℓ_t^*'s follow an ARMA model and they have the associated covariance matrix Σ_{ℓ^*} (McLeod, 1975). Then, the covariance matrix of ℓ_t can be expressed as

$$\Sigma_\ell = W\Sigma_{\ell^*}W \tag{2.3.9}$$

where W is a diagonal matrix of the weights k_t. A possibility is to consider Σ_{ℓ^*} as the autocorrelation matrix of the standardized survey errors and then the k_t's will be

equal to the standard deviations. We will get back to this representation in section 3.5 when introducing a state space approach for benchmarking.

All the equations used in this method assume knowledge of the autocovariance matrices of the annual and subannual survey errors. Additionally, considering the heteroscedastic case, it is necessary to consider an ARMA model of the standardized survey errors. This regression method is one of the core methods included in the software BENCH produced by Statistics Canada (Cholette, 1994; Bloem et al., 2001). The software makes the strong assumption that the survey errors follow an AR(1) model. The main difficulty in the application of the Cholette and Dagum method is that although statistical agencies sometimes produce reports with estimates of the variances of the survey errors, they rarely report either autocorrelations or the specification of the relevant ARIMA models (see Guerrero (1990, page 30)).

Repeated surveys usually use rotation sampling designs which can produce different expected values for estimates of the same characteristics from different rotation groups. The phenomenon has been called *rotation group bias* (Bailar, 1975). The regression method presented here includes a constant bias component in its formulation; however, response bias in the data can be at different magnitude over time, due to for example, conditioning of the respondent or familiarity with the survey after a long period (Ghangurde, 1982).

2.4 ARIMA Model Based Approach

Hillmer and Trabelsi (1987) formulate the benchmarking problem using time series analysis techniques. Their method provides a way to take into account the stochastic properties of the time series being benchmarked, the statistical properties of the sample survey from which the original estimates of the time series were derived and the properties of the errors of the benchmarks. The method was proposed in the context of improving subannual estimates using annual information and the stochastic properties of the subannual series itself.

This approach is naturally represented in terms of Equation 1.1.1 shown in Chapter 1, $y_t = \eta_t + \ell_t$. It is assumed that the two components η_t and ℓ_t in Equation 1.1.1 are mutually independent processes with known first and second order moments and also that η_t and ℓ_t follow the autoregressive integrated moving average (ARIMA) models

$$\begin{aligned} \phi_\eta(B)(\eta_t - \mu) &= \theta_\eta(B)b_t; \quad Var(b_t) = \sigma_b^2 \\ \phi_\ell(B)(\ell_t) &= \theta_\ell(B)c_t; \quad Var(c_t) = \sigma_c^2 \end{aligned} \tag{2.4.1}$$

where each of the pairs of polynomials $(\phi_\eta(B), \theta_\eta(B))$ and $(\phi_\ell(B), \theta_\ell(B))$ have no common zeros; $\phi_\eta(B)$ is a polynomial in the backshift operator B having its zeros lying on or outside the unit circle; $\theta_\eta(B)$, $\phi_\ell(B)$ and $\theta_\ell(B)$ are polynomials with all zeros outside the unit circle, μ is the mean of the process η_t and the processes b_t and c_t are uncorrelated white noise processes.

Using equation 1.1.1, the ARIMA model for y_t is

$$\phi(B)(y_t - \mu) = \theta(B)d_t; \quad Var(d_t) = \sigma_d^2 \tag{2.4.2}$$

where $\phi(B) = \phi_\eta(B).\phi_\ell(B)$ and $\theta(B)$ and σ_d^2 can be obtained using the results from Hillmer and Tiao (1982). Let $\boldsymbol{y} = (y_1, \cdots, y_n)'$, $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)'$ and $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_n)'$; it is assumed that the random vectors $\boldsymbol{\eta}$ and $\boldsymbol{\ell}$ have multivariate normal distributions: $\boldsymbol{\eta}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma_\eta})$ and $\boldsymbol{\ell}$ is $N(\boldsymbol{0}, \boldsymbol{\Sigma_\ell})$ where the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma_\eta}$ and $\boldsymbol{\Sigma_\ell}$ can be derived from the respective ARIMA models.

The aim of the method is to develop the appropriate modifications to the minimum mean squared error estimate when additional (annual) external information about η_t becomes available. From Equation 2.3.2 in the last section and considering no bias term,

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\ell} \tag{2.4.3}$$

$$\boldsymbol{x} = \boldsymbol{L'\eta} + \boldsymbol{e}$$

where $\boldsymbol{x}$ is an observed column vector of dimension m, $\boldsymbol{L}$ is an n x m matrix, and $\boldsymbol{e}$ is an error vector of dimension m x 1. It will be assumed that $\boldsymbol{e}$ has a multivariate normal distribution $\mathrm{N}(\boldsymbol{0}_m, \boldsymbol{\Sigma_e})$ and is independent of $\boldsymbol{\eta}$ and $\boldsymbol{\ell}$. Like the Cholette and Dagum method, Hillmer and Trabelsi (1987) consider the vector $\boldsymbol{\tau}' = (\boldsymbol{y}', \boldsymbol{x}')$ containing both

the monthly and annual observations; a vector of errors $\boldsymbol{u}' = (\boldsymbol{\ell}', \boldsymbol{e}')$ containing the corresponding monthly and annual errors. The vector of parameters consists of the annual benchmarked values to be estimated (without bias) and the matrix X in 2.3.3 takes now the form $X' = [I, L]$.

Proposition 2.4.1. *Consider the model*

$$\tau = X\eta + u$$

Assuming that η *has a* $N(\mu, \Sigma_\eta)$ *distribution and* u *has a* $N(0, \Sigma_u)$ *distribution, where* $\Sigma_u = diag(\Sigma_\ell, \Sigma_e)$*; the minimum mean squared error estimate of* η *given* τ *is*

$$\hat{\eta} = E(\eta \mid \tau) = \hat{\eta}^0 + \eta_c$$

where $\hat{\eta}^0$ *is the minimum mean squared error linear estimate of* η *given* y

$$\hat{\eta}^0 = E(\eta \mid y) = \begin{cases} (\Sigma_\ell{}^{-1} + \Sigma_\eta{}^{-1})^{-1} \times (\Sigma_\ell{}^{-1} y + \Sigma_\eta{}^{-1}\mu), & \text{if } \eta \text{ is stationary} \\ (\Sigma_\ell{}^{-1} + \Sigma_\eta^{-1})^{-1}\Sigma_\ell{}^{-1} y, & \text{if } \eta \text{ is non-stationary ARIMA} \end{cases}$$

and η_c *is the correction factor term*

$$\eta_c = \Omega L(L'\Omega L + \Sigma_e)^{-1}(x - L'\hat{\eta}^0)$$

where

$$\Omega = Cov(\eta \mid y) = (\Sigma_\ell{}^{-1} + \Sigma_\eta{}^{-1})^{-1}$$

and

$$\Sigma_{\hat{\eta}} = Cov(\eta \mid \tau) = \Omega - \Omega L(L'\Omega L + \Sigma_e)^{-1} L'\Omega$$

In the last proposition stationarity implies that all the zeros of the autoregressive part lie outside the unit circle, whereas non-stationarity ARIMA implies that all the zeros of the autoregressive part lie on or outside the unit circle. The proof of the last result is achieved by minimising the mean squared error. This technique is known in the time series literature as *signal extraction*. The proof of the property above appears in Hillmer and Trabelsi (1987). However, they make the strong assumption that a nonstationary series has zero mean according to results in Cleveland and Tiao (1976). Durbin and Quenneville (1997) suggest that this assumption is unnecessary and in some cases invalid. An alternative brief proof is included in this document in Appendix A.3.

Compared with the regression approach, the Hillmer and Trabelsi method takes into account the information about the stochastic properties of the series being benchmarked. Clearly, this information may improve the estimates but the main difficulty in the application of the method is that statistical agencies normally do not produce either autocorrelations or specification of ARIMA models of the survey errors. Bell and Hillmer (1987*b*) state that the estimation of these autocovariance matrices is the same as estimating sampling variances and in particular, Hillmer and Trabelsi (1987) propose the use of the random groups method using survey microdata (Wolter, 1985, chapter 2).

Scott et al. (1977) refer to the use of survey microdata to estimate the autocovariance of the sampling errors as a *primary analysis*. In spite of that, this is not always possible due to the need for confidentiality in the survey or the lack of a record linking data in repeated surveys. Scott et al. (1977) estimate the autocovariance matrix of the survey errors using only the published time series data and they called this alternative a *secondary analysis*. Results from Tiao and Hillmer (1978), Bell and Hillmer (1984) and Bell and Hillmer (1987*b*) establishes that there is a fundamental identification problem with this second alternative.

Chen et al. (1997) introduces a non-parametric solution to estimate the covariance matrix for the stationary part of the signal. Using relative mean squared errors as a measure of efficiency, they show that the ARIMA approach is more efficient than the regression approach and the non-parametric method gives very close values to the ARIMA method (Dagum and Cholette, 2006, page 208).

Laniel and Fyfe (1990) recommend the use of methods such as those of Cholette and Dagum or Hillmer and Trabelsi "for only a small number of very important economic indicators". Their reason is that "since ARIMA modelling is being used in this method, it would be costly to implement for large scale surveys dealing with hundreds of series" (pages 273-274). Another common problem with these two methods is the possible oversmoothing of the data due to a bad specification of the ARIMA models.

2.5 Conclusions and Further Issues

The main benchmarking methods available in the literature have been reviewed in this chapter. Denton's method is a good alternative when there is no additional information about the sampling design or the estimated standard errors in the survey and also particularly, when it is desirable to have binding estimators. Because this method does not include information about the survey errors, an implicit assumption of this method is that the variations in the monthly estimates are close to the real variations of the parameters and it is a pure numerical method without any statistical criteria to evaluate the precision of the estimates.

Methods which include the survey errors in the modelling such as those of Hillmer and Trabelsi or Cholette and Dagum may be more efficient because it is possible to take into consideration stochastic features of the time series structure such as the autocorrelation and heteroscedasticity of the survey errors and the presence of survey bias. An essential characteristic of these methods, making them preferable over the Denton method, is the possibility of calculating the variances of the benchmarked estimates.

Nonetheless, one disadvantage is that sometimes it is not possible to access detailed information generated from the sample survey such as autocovariance matrices or associated ARMA models of the survey errors. Regarding to these disadvantages, Guerrero (1990, page 30) states that: "these requirements are reasonable for a statistical agency in charge of publishing official statistics, but they might be very restrictive for a practitioner who occasionally wants to disaggregate a time series". Also, they require to obtain the annual benchmark before the end of a given year to adjust the subperiods in the year. That means, if subannual data is obtained before the end of the year, they cannot be adjusted until a benchmark is obtained.

The next chapter will present two additional alternatives for benchmarking using state space models proposed by Durbin and Quenneville (1997). These alternatives do not require the specification of the autocovariance matrices of the survey errors and in particular, one of the methods does not need to have an annual benchmark available

at the end of the year to update the estimates. Different components will make up the series (trends, seasonalities, cycles, calendar variations, effects of explanatory variables, interventions) and they are modeled separately. Also multivariate observations can be treated by straightforward extensions of the univariate state space form. A possible disadvantage of these models is the relative lack of software and the consideration of high dimensional vectors and matrices in the estimation. The methods are considered separately in the next chapter in order to introduce the basic theory of state space models and some new thoretical developments referred to the variance of the benchmarked estimates in the binding case which will be extended to all the benchmarking methods introduced in Chapter 2. There are other techniques not considered in this thesis that use auxiliary information for benchmarking. They are classified under the area of "disaggregation of time series" by authors such as Chow and Lin (1971), Fernandez (1981), Guerrero and Martinez (1995), Guerrero and Nieto (1999) or Di Fonzo and Marini (2003). They assume a set of auxiliary series highly correlated with the original one. Newly available software, ECOTRIM (Barcellan and Buono, 2002), is available for the implementation of these methods. Since the application to Business Surveys in the UK do not consider the use of auxiliary information, these methods were not considered in this thesis.

Chapter 3

State Space Models and Benchmarking

3.1 Preliminaries

The state space model approach provides a flexible approach to time series analysis. There are many references, including Durbin and Koopman (2001) who provide a recent treatment of the approach; also Janacek and Swift (1993), Harvey (1989), Tsay (2005) and Shumway and Stoffer (2006) are some textbooks with related chapters to the area.

The main idea in the use of state space models (SSM) in time series analysis of survey data is to extend the general theory of signal extraction by using the Kalman filter (Kalman, 1960). Some authors such as Tam (1987), Binder and Hidiroglou (1988), Binder and Dick (1989) and Pfeffermann (1991) introduced the idea for survey data. A parameter of interest is estimated in each individual survey and then the Kalman filter is applied on the series of estimates. The application of the Kalman filter under the correct model provides at least the "best linear unbiased prediction (BLUP)" estimates of the parameter in every instant (linear optimal). Specifically, when a Gaussian distribution is assumed, the estimator is even optimal in the sense of minimising the mean square error.

The use of state space models and the Kalman filter has also been proposed in the area of temporal disaggregation (Harvey and Pierse, 1984; Harvey, 1989, section 6.4.1). More recently, Durbin and Quenneville (1997), Harvey and Chung (2000) and Moauro and Savio (2002) have considered the temporal disaggregation problem including information about the survey errors. In particular, Durbin and Quenneville (1997) developed benchmarking methodology for the case where monthly estimates are constrained to add up to given annual estimates.

The advantages of the Durbin and Quenneville methods, with respect to the other benchmarking alternatives referred in the previous chapter, arise from the properties of the state space approach (Durbin, 2000; Durbin and Koopman, 2001, section 3.5). Under this approach, the original series is assumed to be decomposed into the unobserved components of trend, seasonality and irregular terms. A big difference of this method with the others considered above is that one of the methods presented below provides a solution to the problem of estimating subannual estimates when there is no a benchmark available in the horizon (*ex-ante estimation*). Another advantage of the state space approach is the possibility to introduce innovation terms and calendar effects into the model in an easy way. Innovation terms permit the consideration of outliers in the series and calendar effects arising due to variations in every specific year. For instance, when the activity of an industry varies according to the day of the week or when the exact days of a holiday change every year.

In this section, the state space model approach for benchmarking is presented as follows; firstly, a brief introduction about structural time series models, state space models and the Kalman filter is given and then, two solutions for the benchmarking problem are studied in subsections 3.5.1. and 3.5.2. Other issues such as the estimation of the survey bias (assumed as a constant parameter) and treatment of multiplicative time series data are presented in Durbin and Quenneville (1997) but not considered in this overview. Finally, the last section surveys two possible kinds of benchmarked estimators (binding and non-binding) and specifically develops the correct variance of binding

estimators when using temporal benchmarks that are subject to survey errors. The theory is presented considering each particular case under the benchmarking methods introduced in Chapters 2 and 3.

3.2 Structural Time Series Model

Structural time series models decompose the series of study into unobservable components which have a direct interpretation. The common decomposition is to consider the series as the sum of trend, seasonal and irregular terms (Harvey, 1989). Suppose we regard y_t as having the form:

$$y_t = \mu_t + \gamma_t + \epsilon_t, \quad t = 1, \cdots, n \tag{3.2.1}$$

where μ_t is a trend component, γ_t is a seasonal component and ϵ_t is the irregular or residual component. In the case of annual series, seasonal effects can be dropped.

Equation 3.2.1 is known as the *additive case*. Sometimes, the additive assumption may be unrealistic and it is preferable to assume the *multiplicative* decomposition given by:

$$y_t = \mu_t \cdot \gamma_t \cdot \epsilon_t, \quad t = 1, \cdots, n \tag{3.2.2}$$

This is a more suitable model when the amplitude of the seasonal cycles increases or decreases jointly with the trend. In an additive structure, the seasonal effects are independent of the evolution of the trend. One simple way to check the adequate decomposition for a time series is to overplot segments of the original series over the cycle. Seasonal adjustment software, such as X11 or X12ARIMA, include ANOVA and non parametric tests to decide what structure is more suitable to use for a particular time series (National Statistics, 2005*a*). It can be noticed that taking logarithms, model 3.2.2 reduces to the model 3.2.1.

Some common structural time series models are:

-*Random Walk plus Noise* (RWN, Muth (1960), Durbin and Koopman (2001, page 9)). Consider Equation 3.2.1 taking $\mu_t = a_t$ where a_t is a random walk; without the

seasonal component and all the disturbances are usually assumed to follow a normal distribution. These assumptions give the model:

$$\begin{aligned} y_t &= a_t + \epsilon_t, \epsilon_t \sim NID(0, \sigma_\epsilon^2) \\ a_t &= a_{t-1} + \nu_t, \nu_t \sim NID(0, \sigma_\nu^2) \end{aligned} \tag{3.2.3}$$

for all $t = 1, \cdots, n$. We assume that the irregular term ϵ_t has constant variance σ_ϵ^2. This model is also called the *local level model.* The notation $NID(0, \sigma^2)$ denotes a normally distributed, serially independent, random variable with mean zero and variance σ^2. When σ_ϵ^2 is zero, the series follow a random walk and the forecasts are equal to the last observation, y_n. On the other hand, if σ_ν^2 is zero, the trend is equal to a constant and the best forecast of future observations is the sample mean.

-*Basic Structural Model* (BSM, Harrison (1965), Harvey (1989, page 172)). A BSM follows a structural time series model given by Equation 3.2.1 with the following components. The trend component of a BSM consists of a local linear trend model given by

$$\begin{aligned} \mu_t &= \mu_{t-1} + \beta_{t-1} + \xi_t \\ \beta_t &= \beta_{t-1} + \zeta_t \end{aligned} \tag{3.2.4}$$

where μ_t is known as the adaptive level and β_t is a random walk known as the local rate of change or slope. The processes ξ_t and ζ_t correspond to uncorrelated white-noise terms with variances σ_ξ^2 and σ_ζ^2 respectively.

The seasonal component can be written in two ways, as a dummy variable type or a trigonometric type. In the dummy variable type, considering K subannual periods per year ($K = 12$ if monthly data, $K = 4$ if quarterly), it is assumed that the seasonal pattern is constant over time. Then, the seasonal values for the subannual periods can be modelled by constants $\gamma_t, \gamma_{t-1} \cdots, \gamma_{t-(K-1)})$ where

$$\sum_{v=0}^{K-1} \gamma_{t-v} = 0 \tag{3.2.5}$$

In practice, it is desirable to allow the seasonal effects to change over time. A simple way to achieve changing seasonality using the ideas before is by adding an error term

ω_t in Equation 3.2.5 and it follows that

$$\sum_{v=0}^{K-1} \gamma_{t-v} = \omega_t \quad \Rightarrow \quad \gamma_t = -\sum_{v=1}^{K-1} \gamma_{t-v} + \omega_t \tag{3.2.6}$$

with ω_t a disturbance term with mean zero. The zero expectation makes these effects sum to zero in the forecast function.

Another alternative to express seasonality is by using a trigonometric form proposed by Hannan, Terrell and Tuckwell (1970). Assuming constant seasonal, the seasonal effect at time t can be expressed as

$$\gamma_t = \sum_{v=1}^{[K/2]} (\gamma_v \cos \kappa_v t + \gamma_v^* \sin \kappa_v t) \tag{3.2.7}$$

where κ_v correspond to the seasonal frequencies, $\kappa_v = 2\pi v/K, v = 1, \cdots, [K/2]$ and $[x]$ denotes the integer part of x.

Equation 3.2.7 may be allowed to evolve over time using results in Harvey (1989, page 42), according to the model

$$\gamma_t = \sum_{v=1}^{[K/2]} \gamma_{vt} \tag{3.2.8}$$

where

$$\begin{aligned} \gamma_{vt} &= \gamma_{v,t-1} \cos \kappa_v + \gamma_{v,t-1}^* \sin \kappa_v + \omega_{vt} \\ \gamma_{vt}^* &= -\gamma_{v,t-1} \sin \kappa_v + \gamma_{v,t-1}^* \cos \kappa_v + \omega_{vt}^* \end{aligned} \tag{3.2.9}$$

with $v = 1, \cdots, [K/2]$ and with ω_{vt} and ω_{vt}^* white noise processes with mean zero and uncorrelated with each other.

Additional to the trend and the seasonal components (e.g. dummy variable or trigonometric type), the BSM accounts for an irregular component ϵ_t which is assumed to be a white noise. The BSM corresponds then to the sum of these three components.

Sometimes the BSM is extended to the more general form

$$\eta_t = \mu_t + \gamma_t + \tau_t + \varphi_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \quad t = 1, \cdots, n \tag{3.2.10}$$

where τ_t and φ_t represent trading day and moving festival components. *Trading day* refers to the modelling of any variation in the series which depends on how many

Mondays, Tuesdays, and so on, there are in the series. For example, if Fridays tend to have bigger sales, there is a necessity to account for the number of Fridays in the period of reference. On the other hand, certain holidays and religious festivals (most notably Easter, every Jewish festival, Chinese New Year or Ramadan) appear in different dates from year to year. Then, the addition of *moving festival components* to the model permit to take control over this situation because data from several time series (e.g. industrial production, retail sales and air traffic series) are affected by the date when these events are falling, Harvey (1989, page 335). Dagum, Quenneville and Sutradhar (1992) studied general models for trading day and its SSF, whereas Bell and Hillmer (1983) and Morris and Pfeffermann (1985) have suggested some models for the moving festivals. Also, Cleveland and Devlin (1980*a*) and Cleveland and Devlin (1980*b*) discussed some special methods to detect calendar effects.

3.3 State Space Form

The state space form provides a simple way to deal with structural time series models. The "state" of the system represents the unobserved structural components such as trends and seasonalities. The basic idea is to write a structural time series model in a special form, which will permit the Kalman filter to update the state when new observations become available. Additionally, a better estimate of the state is obtained using smoothing algorithms at any instant during the period of observation (Anderson and Moore, 1979; de Jong, 1988*a*; de Jong, 1989; Kohn and Ansley, 1989).

A state-space model (SSM) is a set of two equations related to an unobserved state vector $\boldsymbol{\alpha}_t$. The first equation states that the observations are linear combinations of $\boldsymbol{\alpha}_t$ and it is known as the *observation equation.* The second equation represents the evolution of $\boldsymbol{\alpha}_t$ over time and it is known as the *transition equation.* The set of two equations can be written in many ways Durbin and Koopman (2001, page 38) and particularly, in this thesis, the observation equation will take the form

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \quad \boldsymbol{\varepsilon}_t \sim iid(0_p, \boldsymbol{H}_t) \quad t = 1, \cdots, n \tag{3.3.1}$$

and the evolution of $\boldsymbol{\alpha}_t$ is given by the *transition equation* represented by the Markovian structure,

$$\boldsymbol{\alpha}_t = \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{\vartheta}_t \quad \boldsymbol{\vartheta}_t \sim iid(0_r, \boldsymbol{Q}_t) \quad t = 1, \cdots, n \tag{3.3.2}$$

where iid$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for "independent identically distributed" random variables with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In the pair of the equations above, $\boldsymbol{y}_t$ is the value of the observed time series at the instant t, which is a scalar if $\boldsymbol{y}$ is a univariate time series. Otherwise, $\boldsymbol{y}_t$ is considered as a P x 1 vector of observations at time t, where P represents the number of components in the multivariate time series. $\boldsymbol{\alpha}_t$ represents the r x 1 unobserved state vector, $\boldsymbol{Z}_t$ and $\boldsymbol{T}_t$ are deterministic matrices of dimension P x r and r x r respectively, and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\vartheta}$ are disturbance terms of dimension P x 1 and r x 1 respectively. Also, $\boldsymbol{H}_t$ and $\boldsymbol{Q}_t$ denote P x P and r x r known covariance matrices respectively.

It is also assumed in Equations 3.3.1 and 3.3.2 that the initial vector $\hat{\boldsymbol{\alpha}}_0 \sim N(\boldsymbol{\alpha}_0, \boldsymbol{P}_0)$; $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\vartheta}_t$ are serially uncorrelated and additionally; it will also be assumed that $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\vartheta}_t$ are mutually independent and uncorrelated with the initial vector $\boldsymbol{\alpha}_0$. In practice, however, there are some unknown elements in the system given by the observation and transition equations (e.g. in practice, $\boldsymbol{H}_t$ and $\boldsymbol{Q}_t$ are usually unknown). We will refer to these unknown parameters as *hyperparameters* and will discuss their estimation in section 3.4.2.

Having in mind the formulation in SSF of any structural time series model, Harvey (1989) highlights that $\boldsymbol{\alpha}_t$ must be determined by construction. Notice that, in particular, if a new state vector $\boldsymbol{\alpha}_t^*$ is obtained making $\boldsymbol{\alpha}_t^* = \boldsymbol{M}\boldsymbol{\alpha}_t$ where $\boldsymbol{M}$ is any nonsingular r x r matrix and $\boldsymbol{\alpha}_t$ is the corresponding state vector in equations 3.3.1 and 3.3.2; a new state space formulation is obtained given by

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{Z}_t^*\boldsymbol{\alpha}_t^* + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\alpha}_t^* &= \boldsymbol{T}_t^*\boldsymbol{\alpha}_{t-1}^* + \boldsymbol{\vartheta}_t^* \end{aligned} \tag{3.3.3}$$

where $\boldsymbol{Z}_t^* = \boldsymbol{Z}_t\boldsymbol{M}^{-1}, \boldsymbol{T}_t^* = \boldsymbol{M}\boldsymbol{T}_t\boldsymbol{M}^{-1}$ and $\boldsymbol{\vartheta}_t^* = \boldsymbol{M}\boldsymbol{\vartheta}_t$. Thus, there is no unique representation for any particular model by SSF.

The basic idea is to set up a state vector $\boldsymbol{\alpha}_t$ containing all the information of the system at the instant t with the smallest possible number of components. We will now present a state space formulation for the structural time series models presented in the last section.

- *Random Walk plus Noise.* The RWN model was introduced in Equation 3.2.3. The state space form (SSF) for this model is straightforward, using $\boldsymbol{\alpha}_t$ to be equal to the scalar a_t, $\boldsymbol{Z}_t$ and $\boldsymbol{T}_t$ are constants with value equal to 1, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\vartheta}_t$ are equal to the scalars ϵ_t and ν_t. The conditions $\epsilon_t \sim NID(0, \sigma_\epsilon^2)$ and $\nu_t \sim NID(0, \sigma_\nu^2)$ agree with those of the state space formulation.

Another possible state space formulation for a RWN model is obtained by defining $\boldsymbol{\alpha}_t = [a_t \quad \epsilon_t]'$, $\boldsymbol{Z}_t = [1 \quad 1]$, $\boldsymbol{T}_t = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $\boldsymbol{\varepsilon}_t = 0$ and $\boldsymbol{\vartheta}_t = [\nu_t \quad \epsilon_t]'$. Finally, in this second formulation, $\boldsymbol{\vartheta}_t \sim NID(\mathbf{0}_2, \boldsymbol{Q}_t)$ with $\boldsymbol{Q}_t = \text{diag}(\sigma_\nu^2, \sigma_\epsilon^2)$ and the final SSF is given by

$$y_t = [1 \quad 1] \cdot [a_t \quad \epsilon_t]'$$

$$\begin{bmatrix} a_t \\ \epsilon_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a_{t-1} \\ \epsilon_{t-1} \end{bmatrix} + \begin{bmatrix} \nu_t \\ \epsilon_t \end{bmatrix} \tag{3.3.4}$$

It should be noticed that in this second formulation there is no disturbance term in the observation equation and the transition matrix $\boldsymbol{T}_t$ is singular. Authors such as Godolphin and Stone (1980) and Kohn and Ansley (1983) have studied the implications of singular matrices in state space models.

-*Basic Structural Model.* The Basic Structural Model (BSM) was defined in Section 2.5.1 as a model composed of a local linear trend model defined in Equation 3.2.4; a dummy variable (Equation 3.2.6) or trigonometric model for the seasonal component (Equations 3.2.7 and 3.2.9) and a white noise irregular term. The transition equation in the SSF for the local linear trend can be formulated by setting the following vectors and matrices below. These elements conform with the vectors and matrices in Equation 3.2.4.

$$\alpha_{1,t} = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix}, \quad Z_{1,t} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad T_{1,t} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \vartheta_{1,t} = \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} \tag{3.3.5}$$

According to the Equation 3.2.6, $\gamma_t = -\sum_{\tau=1}^{3} \gamma_{t-\tau} + \omega_t$, the transition equation for a dummy variable seasonality for quarterly data can be formulated by defining

$$\alpha_{2,t} = \begin{bmatrix} \gamma_t \\ \gamma_{t-1} \\ \gamma_{t-2} \end{bmatrix}, \quad Z_{2,t} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad T_{2,t} = \begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \vartheta_{2,t} = \begin{bmatrix} \omega_t & 0 & 0 \end{bmatrix} . \tag{3.3.6}$$

These elements conform with the dummy variable seasonal model presented in the right side of Equation 3.2.6. Following Equations 3.2.7 and 3.2.9 with $v = 1, 2$ for quarterly data, the trigonometric seasonality can be formulated in SSM by defining

$$\alpha_{2,t} = \begin{bmatrix} \gamma_{1t} \\ \gamma_{1t}^* \\ \gamma_{2t} \end{bmatrix}, \quad T_{2,t} = \begin{bmatrix} \cos(\pi/2) & \sin(\pi/2) & 0 \\ -\sin(\pi/2) & \cos(\pi/2) & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{3.3.7}$$
$$\text{and} \quad Z_{2,t} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \vartheta_{2,t} = \begin{bmatrix} \omega_{1t} & \omega_{1t}^* & \omega_{2t} \end{bmatrix}$$

These are the vectors and matrices involved in the SSF of a trigonometric seasonal model corresponding to the model in Equation 3.2.9. Using Equations 3.3.5 - 3.3.7 and $\varepsilon_t = \epsilon_t$, the SSF for the BSM is as follows

$$y_t = [Z_{1,t} \quad Z_{2,t}] \cdot [\alpha_{1,t} \quad \alpha_{2,t}]' + \epsilon_t$$
$$\begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \end{bmatrix} = \begin{bmatrix} T_{1,t} & 0 \\ 0 & T_{2,t} \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{bmatrix} + \begin{bmatrix} \vartheta_{1,t} \\ \vartheta_{2,t} \end{bmatrix} \tag{3.3.8}$$

Another possible state space formulation for the BSM is obtained by including the irregular term into the state space vector as follows:

$$
\begin{aligned}
y_t &= [Z_{1,t} \quad Z_{2,t} \quad Z_{3,t}] \cdot [\alpha_{1,t} \quad \alpha_{2,t} \quad \alpha_{3,t}]' \\
&= [Z_{1,t} \quad Z_{2,t} \quad 1] \cdot [\alpha_{1,t} \quad \alpha_{2,t} \quad \epsilon_t]' \\
\begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \\ \alpha_{3,t} \end{bmatrix} &= \begin{bmatrix} T_{1,t} & 0 & 0 \\ 0 & T_{2,t} & 0 \\ 0 & 0 & T_{3,t} \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \alpha_{3,t-1} \end{bmatrix} + \begin{bmatrix} \vartheta_{1,t} \\ \vartheta_{2,t} \\ \vartheta_{3,t} \end{bmatrix} \\
&= \begin{bmatrix} T_{1,t} & 0 & 0 \\ 0 & T_{2,t} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \epsilon_{t-1} \end{bmatrix} + \begin{bmatrix} \vartheta_{1,t} \\ \vartheta_{2,t} \\ \epsilon_t \end{bmatrix}
\end{aligned} \tag{3.3.9}
$$

The last formulation does not have a disturbance term in the observation equation and the transition matrix T_t is singular. For the more general form involving calendar effects, Dagum et al. (1992) studied the state space formulation for trading day effects.

- *ARMA model* (Box and Jenkins, 1976; Box, Jenkins and Reinsel, 1994). A stationary ARMA(p, q) model is given by

$$\ell_t = \phi_1 \ell_{t-1} + \cdots + \phi_p \ell_{t-p} + \chi_t + \theta_1 \chi_{t-1} + \cdots + \theta_q \chi_{t-q} \tag{3.3.10}$$

where $\chi_t \sim \text{NID}(0, \sigma_\chi^2)$ and p and q are non-negative integers.

Assuming $\varrho = \max(p, q+1)$, a SSF representation for the ARMA model is achieved by setting the state space vector of length ϱ

$$
\alpha_{4,t} = \begin{bmatrix} \ell_t \\ \phi_2 \ell_{t-1} + \cdots + \phi_\varrho \ell_{t-\varrho+1} + \theta_1 \chi_t + \cdots + \theta_{\varrho-1} \chi_{t-\varrho+2} \\ \vdots \\ \phi_{\varrho-1} \ell_{t-1} + \phi_\varrho \ell_{t-2} + \theta_{\varrho-2} \chi_t + \theta_{\varrho-1} \chi_{t-1} \\ \phi_\varrho \ell_{t-1} + \theta_{\varrho-1} \chi_t \end{bmatrix} \tag{3.3.11}
$$

and the remaining vectors and matrices defining the SSF as follows

$$
Z_{4,t} = \begin{bmatrix} 1 \\ 0_{\varrho-1} \end{bmatrix}', \quad T_{4,t} = \begin{bmatrix} \phi_1 & \\ \vdots & I_{\varrho-1} \\ \phi_{\varrho-1} & \\ \phi_\varrho & 0'_{\varrho-1} \end{bmatrix}, \vartheta_{4,t} = \begin{bmatrix} \chi_t \\ \theta_1 \chi_t \\ \vdots \\ \theta_{\varrho-1} \chi_t \end{bmatrix} \text{ and } \quad \varepsilon_{4,t} = 0 \tag{3.3.12}
$$

Using this formulation some of the AR or MA coefficients will be equal to zero unless $p = q + 1$. There are many ways to transform such an ARMA model into a state space form and the interested reader could be referred to other options as presented in Akaike (1975) and Aoki (1987).

More sophisticated models could be considered in order to exhibit change in variance over time. Heteroscedasticity (volatility) models such as the *ARCH model* (Engle, 1982) and the *GARCH model* (Bollerslev, 1986) can be formulated into a state space representation with non-normally distributed disturbances but still providing minimum mean squared error linear estimators of the state and future observations (Harvey, Ruiz and Shephard, 1994).

3.4 Kalman Filtering and Smoothing

The Kalman Filter is a set of recursive equations for calculating optimal estimates of the state vector $\boldsymbol{\alpha}_t$ at time t, using the information available at time t (Kalman, 1960; Harvey, 1989; Durbin and Koopman, 2001). Once a model is set up into its SSF, it is possible to calculate the expectation and variance of $\boldsymbol{\alpha}_t$ conditional on the observed data $\boldsymbol{Y}_t = (y_1, \cdots, y_t)$. The "optimality" of the estimator of $\boldsymbol{\alpha}_t$ refers to the property of minimising the mean squared error (MSE). The application of the Kalman filter provides at least the "best linear unbiased prediction (BLUP)" estimates of the parameter in every instant (linear optimal); when a Gaussian distribution is assumed, the estimator is even optimal in the sense of minimising the mean square error.

3.4.1 Forward and Backward Equations

Let $\hat{\boldsymbol{\alpha}}_t = \hat{\boldsymbol{\alpha}}_{t|t}$ be the conditional mean of $\boldsymbol{\alpha}_t$ given the observed data $\boldsymbol{Y}_t = (y_1, \cdots, y_t)$ and let also $\boldsymbol{P}_t = \boldsymbol{P}_{t|t}$ be the r x r conditional covariance matrix $\boldsymbol{P}_t = \text{Cov}(\boldsymbol{\alpha}_t \mid \boldsymbol{Y}_t)$. Taking into account the assumptions in the SSF and assuming normality, the initial

vector $\hat{\boldsymbol{\alpha}}_0$ has a multivariate normal distribution $\mathrm{N}(\boldsymbol{\alpha}_0, P_0)$.

According to the transition equation

$$\boldsymbol{\alpha}_1 = \boldsymbol{T}_1\boldsymbol{\alpha}_0 + \boldsymbol{\vartheta}_1 \tag{3.4.1}$$

Then, $\boldsymbol{\alpha}_1$ is multivariate normal with conditional mean $\hat{\boldsymbol{\alpha}}_{1|0} = \boldsymbol{T}_1\boldsymbol{\alpha}_0$ and covariance matrix given by $P_{1|0} = \boldsymbol{T}_1 P_0 \boldsymbol{T}'_1 + Q_1$. The distribution of $\boldsymbol{\alpha}_1$ conditional on $\boldsymbol{y}_1$ is obtained by writing

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \hat{\boldsymbol{\alpha}}_{1|0} + (\boldsymbol{\alpha}_1 - \hat{\boldsymbol{\alpha}}_{1|0}) \\ \boldsymbol{y}_1 &= \boldsymbol{Z}_1\hat{\boldsymbol{\alpha}}_{1|0} + \boldsymbol{Z}_1(\boldsymbol{\alpha}_1 - \hat{\boldsymbol{\alpha}}_{1|0}) + \boldsymbol{\varepsilon}_1 \end{aligned} \tag{3.4.2}$$

The vector $[\boldsymbol{\alpha}'_1 \quad \boldsymbol{y}'_1]$ has also a multivariate normal distribution with mean $[\hat{\boldsymbol{\alpha}}_{1|0} \quad \boldsymbol{Z}_1\hat{\boldsymbol{\alpha}}_{1|0}]$ and covariance matrix

$$\begin{bmatrix} P_{1|0} & P_{1|0}\boldsymbol{Z}'_1 \\ \boldsymbol{Z}_1 P_{1|0} & \boldsymbol{Z}_1 P_{1|0}\boldsymbol{Z}'_1 + H_1 \end{bmatrix} \tag{3.4.3}$$

Using some properties of the multivariate normal distribution (Harvey, 1989, Appendix Chapter 3), the distribution of $\boldsymbol{\alpha}'_1$ conditional on $\boldsymbol{y}'_1$ is also multivariate normal with mean

$$\hat{\boldsymbol{\alpha}}_{1|1} = \hat{\boldsymbol{\alpha}}_{1|0} + P_{1|0}\boldsymbol{Z}'_1 F_1^{-1}\boldsymbol{v}_1 \tag{3.4.4}$$

and covariance matrix

$$P_1 = P_{1|0} - P_{1|0}\boldsymbol{Z}'_1 F_1^{-1}\boldsymbol{Z}_1 P_{1|0} \tag{3.4.5}$$

where $\boldsymbol{v}_1 = \boldsymbol{y}_1 - \boldsymbol{Z}_1\hat{\boldsymbol{\alpha}}_{1|0}$ are called the *innovations* or *one-step ahead prediction errors* and $F_1 = \boldsymbol{Z}_1 P_{1|0}\boldsymbol{Z}'_1 + H_1$ represents their covariance matrix.

Repeating this procedure for $t = 1, \cdots, n$ (Harvey, 1989; Durbin and Koopman, 2001) result in the *prediction equations*

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_{t|t-1} &= \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} \\ P_{t|t-1} &= \boldsymbol{T}_t P_{t-1}\boldsymbol{T}'_t + Q_t \end{aligned} \tag{3.4.6}$$

and the *updating equations*

$$\begin{aligned}
\hat{\alpha}_t &= \hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + P_{t|t-1}Z'_t F_t^{-1} v_t \\
P_t &= P_{t|t-1} - P_{t|t-1}Z'_t F_t^{-1} Z_t P_{t|t-1} \\
v_t &= y_t - Z_t \hat{\alpha}_{t|t-1} \\
F_t &= Z_t P_{t|t-1} Z'_t + H_t
\end{aligned} \tag{3.4.7}$$

where v_t denotes the innovations or one-step ahead prediction errors and F_t denote their variances. Equations 3.4.6 and 3.4.7 give the Kalman filter recursions for the filtered state, the innovations and their respective variances.

After applying the Kalman filter, it is possible to take account of the information made available after time t, that means updating the Kalman filter estimates using the information in the entire sample Y_n. This procedure is called *smoothing* and the corresponding estimator is called a *smoother*; since the smoother is based on more information than the filtered estimator, it will have a smaller MSE.

The *fixed-interval smoothing* algorithm is one of the possible smoothing algorithms, being a backward recursion which starts at time n after the Kalman filter is applied. Its derivation is obtained by authors such as Anderson and Moore (1979), Jazwinski (1970), and Ansley and Kohn (1982). The backward recursions for $t = n, \cdots, 1$ are given by

$$\begin{aligned}
J_{t-1} &= P_{t-1|t-1} T'_t P_{t|t-1}^{-1} \\
\hat{\alpha}_{t-1|n} &= \hat{\alpha}_{t-1|t-1} + J_{t-1}(\hat{\alpha}_{t|n} - T_t \hat{\alpha}_{t-1|t-1}) \\
P_{t-1|n} &= P_{t-1|t-1} - J_{t-1}(P_{t|n} - P_{t|t-1}) J'_{t-1}
\end{aligned} \tag{3.4.8}$$

Other forms of the same backward recursions have been proposed by de Jong (1988*a*), de Jong (1989), Kohn and Ansley (1989) considering the recursions

$$
\begin{aligned}
K_t &= P_{t|t-1} Z_t' F_t^{-1} \\
L_t &= T_t - K_t Z_t \\
r_{t-1} &= Z_{t-1}' F_t^{-1} v_t + L_t' r_t \\
\hat{\alpha}_{t|n} &= \hat{\alpha}_{t|t-1} + P_{t|t-1} r_{t-1} \\
N_{t-1} &= Z_{t-1}' F_t^{-1} Z_t + L_t' N_t L_t \\
V_{t|n} &= P_{t|t-1} - P_{t|t-1} N_{t-1} P_{t|t-1}
\end{aligned}
\tag{3.4.9}
$$

with $r_n = 0_r$ and $N_n = 0_{rxr}$ and $t = n, \cdots, 1$. This last set of recursions also permit to calculate the covariance matrices of the smoothed estimators at different times (de Jong and Mackinnon, 1988; de Jong, 1998) as follows

$$
Cov(\hat{\alpha}_{t|n}, \hat{\alpha}_{t^*|n}) = P_{t|t-1} L_t' L_{t+1}' \cdots L_{t^*-1}' (I - N_{j-1} P_{t^*|t^*-1}) \tag{3.4.10}
$$

for $t^* = t+1, \cdots, n$. These covariances will be used for benchmarking later on. Durbin and Koopman (2001, sections 4.3 and 4.5) show the rationale behind the filtering and smoothing equations and also the equivalence between the recursions in Equation 3.4.8 with the recursions in Equation 3.4.9.

Finally, the smoothed estimates of the response y and its variance can be obtained making

$$
\begin{aligned}
\hat{y}_t &= Z_t \hat{\alpha}_{t|n} \\
Var(\hat{y}_t) &= Z_t Var(\hat{\alpha}_{t|n}) Z_t' \\
Cov(\hat{y}_t, \hat{y}_{t^*}) &= Z_t Cov(\hat{\alpha}_{t|n}, \hat{\alpha}_{t^*|n}) Z_{t^*}'
\end{aligned}
\tag{3.4.11}
$$

for $t = n, \cdots, 1$ and $t^* = t+1, \cdots, n$.

An important case to take into consideration is when any of the matrices $P_{t|n}$ becomes singular. In that case, Harvey (1989), page 154 states that "if $P_{t|n}$ is singular for some t, it may be replaced by a generalised inverse as suggested by Kohn and Ansley (1983)". Particularly, this is the case when the RWN or the BSM are written in SSF putting the irregular term into the state vector. This formulation induces the transition matrix to have a row of zeros and then being a singular matrix.

3.4.2 Estimation of the hyperparameters

In the SSF the system matrices usually depend on unknown parameters known as hyperparameters. For instance, variance of disturbances or ARMA parameters need to be estimated before the Kalman filter is applied. These estimates can be obtained using maximum likelihood (ML) estimation. The Kalman filter is used to construct the likelihood function and then this function is maximised applying a suitable numerical procedure of optimization. However, the traditional assumption of a set of observations $y_1, \cdots, y_n$ being independent and identically distributed is not valid for a time series model. Denoting the vector of hyperparameters by ψ and assuming multivariate normality for the disturbances in the SSF, the joint density of the observations can be expressed by

$$L(y, \psi) = \prod_{t=1}^{n} p(y_t \mid Y_{t-1}) \tag{3.4.12}$$

where $p(y_t \mid Y_{t-1})$ denotes the distribution of y_t conditional in all the information available until time $t-1$. Assuming normality, $p(y_t \mid Y_{t-1})$ is also normal with mean $Z_t\hat{\alpha}_{t|t-1}$ and covariance matrix F_t. The likelihood function can be expressed as,

$$logL(\psi) = -(pn/2)log(2\pi) - (1/2)\sum_{t=1}^{n} log \mid F_t \mid -(1/2)\sum_{t=1}^{n} v_t' F_t^{-1} v_t \tag{3.4.13}$$

where v_t are the innovations and the matrices F_t are the prediction error variance-covariance matrices of the innovations as defined in Equation 3.4.7. Equation 3.4.13 is sometimes known as the *prediction error decomposition* form of the likelihood.

In order to estimate the set of hyperparameters, the likelihood function will be maximised with respect to the vector of unknown parameters ψ using the function nlminb in Splus 7.0. This is a function based on numerical maximisation algorithms (Durbin and Koopman, 2001, section 7.3.2) such as the Newton-Raphson's method of optimisation. The details and derivations of the Newton's method can be found in Harvey (1990). Other optimisation algorithms such as the functions ms and optim are available in R and Splus (Venables and Ripley, 2002). The algorithm used should be able to handle constraints, since estimates of variances must be non-negative while ARMA parameters must follow restrictions in order to get stationary and invertible processes.

Once the ML estimates are obtained, they are substituted for the unknown hyperparameters in the corresponding formulas of the state predictors and their variances. Pfeffermann and Tiller (2005) point out that this practice results in underestimation of the true prediction mean square errors (PMSE) due to ignoring the variability implied by the parameter estimation. They developed bootstrap procedures to get valid PMSE estimators when the state vector predictors use estimated hyperparameter values.

3.4.3 Initialization of the Kalman filter

In order to start the Kalman filter and smoother recursions, it is necessary to have initial values of the state $\boldsymbol{\alpha}_0$ and the covariance matrix $\boldsymbol{P}_0$. However, it can be shown that for models reaching a steady state; the state estimates for large t are not considerably affected by the choice of initial values, even if the model is non-stationary (Janacek and Swift, 1993).

If these values are not known, which is the common case, there are some alternatives to estimate them. If the state vector is stationary, the filter can be started using a zero mean and a covariance matrix, $\boldsymbol{P}_0$, representing the mean and covariance matrix of the unconditional distribution of the state vector provided that the unconditional mean is zero (Gardner, Harvey and Phillips, 1980).

Following the second line in Equation 3.4.6, $\boldsymbol{P}_0$ is considered as the solution of the equation

$$\boldsymbol{P}_0 = \boldsymbol{T}\boldsymbol{P}_0\boldsymbol{T}' + \boldsymbol{Q} \tag{3.4.14}$$

which is equivalent to the equation

$$\text{vec}(\boldsymbol{P}_0) - \text{vec}(\boldsymbol{T}\boldsymbol{P}_0\boldsymbol{T}') = \text{vec}(\boldsymbol{Q}) \tag{3.4.15}$$

and then using the property $\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A})vec(\boldsymbol{Q})$, it follows that

$$\text{vec}(\boldsymbol{P}_0) = [\boldsymbol{I}_r - \boldsymbol{T} \otimes \boldsymbol{T}]^{-1}\text{vec}(\boldsymbol{Q}) \tag{3.4.16}$$

where $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ and the $vec(.)$ operator transforms a matrix into a vector by stacking its columns one underneath the other.

On the other hand, if the state vector is generated by a non-stationary process, some different approaches have been considered to initialize the Kalman filter in this case. Considering a more general situation where the state vector contains elements in which there is no prior information available (say b non-stationary elements) and others with a known joint distribution ($m - b$ stationary elements); a general model for the initial state vector $\boldsymbol{\alpha}_0$ is

$$\boldsymbol{\alpha}_0 = \boldsymbol{a} + B\boldsymbol{\delta} + R\boldsymbol{\lambda} \tag{3.4.17}$$

with $\boldsymbol{a}$ being a known r x 1 vector, $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are b x 1 and $(r-b)$ x 1 vectors of unknown quantities, B and R being r x b and r x $(r-b)$ selection matrices, respectively. B and R constitute a set of columns of I_r and $B'R = 0_{b\times(r-b)}$. The aim is to separate $\boldsymbol{\alpha}_0$ into a constant part $\boldsymbol{a}$, a non-stationary part $B\boldsymbol{\delta}$ and a stationary part $R\boldsymbol{\lambda}$.

Example 2.2

Consider the decomposition of a time series in Equation 1.1.1. Using this expression, the observed sample time series is decomposed as the sum of the unobserved population true series plus the sampling error series. Assuming a BSM (Equation 3.2.4) for the true series and an AR(1) model for the sampling error series, the state vector will take the form of a column vector of dimension 7 given by:

$$\boldsymbol{\alpha}_t = [\mu_t', \beta_t, \gamma_t, \gamma_{t-1}, \gamma_{t-2}, \epsilon_t, \ell_t]' \tag{3.4.18}$$

and then, this vector has been formed using non-stationary and stationary components (the stationary component corresponds to the last element in the state vector). According to Equation 3.4.17, the state vector can be decomposed as

$$\boldsymbol{\alpha}_0 = B\boldsymbol{\delta} + R\boldsymbol{\lambda} \tag{3.4.19}$$

with

$$B_{7x6} = \begin{bmatrix} I_6 \\ 0'_6 \end{bmatrix}, \boldsymbol{\delta}_{6x1} = [\mu_t', \beta_t, \gamma_t, \gamma_{t-1}, \gamma_{t-2}, \epsilon_t]', R_{7x1} = \begin{bmatrix} 0_6 \\ 1 \end{bmatrix}, \boldsymbol{\lambda}_{1x1} = \ell_t \tag{3.4.20}$$

Two alternatives for initializing the Kalman Filter are described below. The first assumes that $\boldsymbol{\alpha}_0$ is random and nothing is known about the initial state. The second

one assumes that the initial state $\boldsymbol{\alpha}_0$ is fixed but unknown. Therefore, its elements must be estimated by treating them as unknown parameters in the model.

Diffuse Prior Initialization

Using some vocabulary of Bayesian inference, an informative prior corresponds to specific and definite information about a variable whereas a non-informative prior corresponds to vague information. Considering the decomposition of the state vector in Equation 3.4.17, we first need to specify the prior distribution of $\boldsymbol{\alpha}_0$ to let the Kalman filter update that distribution.

Assuming $\boldsymbol{\delta} \sim N(0, \kappa \boldsymbol{I}_b)$, the Kalman filter is started using as initial conditions $\boldsymbol{a}_0 = E(\boldsymbol{\alpha}_0) = \boldsymbol{a}$ and $\boldsymbol{P}_0 = Var(\boldsymbol{\alpha}_0)$ where

$$\boldsymbol{P}_0 = \kappa \boldsymbol{P}_\infty + \boldsymbol{P}_* \tag{3.4.21}$$

and $\kappa \to \infty$; $\boldsymbol{P}_\infty = \boldsymbol{B}\boldsymbol{B}'$ and $\boldsymbol{P}_* = \boldsymbol{R}\boldsymbol{Q}_0\boldsymbol{R}'$ with $\boldsymbol{Q}_0$ being the covariance matrix of $\boldsymbol{\lambda}$ in Equation 3.4.17. A simple approximate technique is to start the Kalman filter at $t = 0$ with $\boldsymbol{a}_0 = \boldsymbol{0}$ and replacing κ by an arbitrary large number (S+Finmetrics assumes $\kappa = 10^6$ x $\max\{1, \text{diag}(\boldsymbol{Q})\}$. The first b innovations and their associated variances are not considered in the prediction error decomposition in Equation 3.4.13. Using the notation above, if b represents the number of non-stationary components in the state vector and also if a diffuse prior is considered for $\boldsymbol{\alpha}_0$; the first b observations will permit to construct $\boldsymbol{a}_b$ and $\boldsymbol{P}_b$ as the starting values Harvey and Peters (1990, page 92). However, this approach could lead to large rounding errors and could complicate the numerical optimization if $\boldsymbol{\alpha}_0$ is very large (Harvey (1989, page 128); Durbin and Koopman (2001, page 101)).

More general ways of avoiding the "large κ" approximation include the methods due to Ansley and Kohn (1985) and de Jong (1988*b*). The Ansley and Kohn (1985) method propose a transformation which eliminates the dependence on initial conditions. A modified form of the Kalman filter is then constructed and this enables the likelihood function to be constructed via the prediction error decomposition (see Section

3.4.2). However, when smoothing, the usual backward recursions are applied from $t = n, \cdots, b+1$ and some modifications are required for the initial period $t = b, \cdots, 1$. Further developments of the Ansley and Kohn (1985) method were given by Koopman and Durbin (2003) making the collapse between $t = n, \cdots, b+1$ and $t = b, \cdots, 1$ to be automatic in the smoothing. Time series computational packages such as S+Finmetrics use Koopman and Durbin (2003) method for initializing the Kalman filter. The method is known as *exact diffuse prior* (Durbin and Koopman, 2001, chapter 5).

Alternatively, the de Jong (1988*b*) method is based on an extension of the Kalman filter augmenting the observed vector. However, in this approach, it is also necessary to modify the initial smoothing in the backward recursions. A recent solution to the problem appears in de Jong and Chu-Chun-Lin (2003). Other authors such as Bell and Hillmer (1991) and Snyder and Saligari (1996) considered the initialization problem for the Kalman filter recursion from a diffuse prior point of view. However, they did not consider the adaptation of the smoothing recursions under their initialization methods.

Fixed Initial State Vector

If it is assumed that all the elements of $\boldsymbol{\alpha}_0$ are fixed, this will imply that $\boldsymbol{P_0} = \mathbf{0}$. As a result, the problem of initialization concerns the estimation of $\boldsymbol{\alpha}_0$ only. Re-writing the observations in Equation 3.3.1 in terms of $\boldsymbol{\alpha}_0$ by repeated substitution of the Equation 3.3.2. It follows that

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{Z}_1\boldsymbol{T}_1\boldsymbol{\alpha}_0 + \boldsymbol{Z}_1\boldsymbol{\vartheta}_1 + \boldsymbol{\varepsilon}_1 \\ \boldsymbol{y}_2 &= \boldsymbol{Z}_2\boldsymbol{T}_2\boldsymbol{T}_1\boldsymbol{\alpha}_0 + \boldsymbol{Z}_2(\boldsymbol{T}_2\boldsymbol{\vartheta}_1 + \boldsymbol{\vartheta}_2) + \boldsymbol{\varepsilon}_2 \end{aligned} \tag{3.4.22}$$

and calling $\boldsymbol{X}_2 = \boldsymbol{Z}_2\boldsymbol{T}_2\boldsymbol{T}_1$ and $\boldsymbol{\alpha^*}_2 = \boldsymbol{T}_2\boldsymbol{\vartheta}_1 + \boldsymbol{\vartheta}_2$, it follows that, in general,

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\alpha}_0 + \boldsymbol{Z}_t\boldsymbol{\alpha^*}_t + \boldsymbol{\varepsilon}_t$$

with $\boldsymbol{X}_t = \boldsymbol{Z}_t \prod_{j=1}^{t} \boldsymbol{T}_j$ and $\boldsymbol{\alpha^*}_t = \boldsymbol{T}_t\boldsymbol{\alpha^*}_{t-1} + \boldsymbol{\vartheta}_t$

Equation 3.4.3 is considered as the measurement equation of a multivariate model and the GLS estimator of $\boldsymbol{\alpha}_0$ is obtained (Wecker and Ansley, 1983). A different method

by Rosenberg (1973) considers the estimation of α_0 by maximum likelihood. The method yields identical numerical results to Wecker and Ansley (1983) (see Harvey (1989, section 3.4.4.)).

Pfeffermann (1984) considered a more general problem of optimal prediction of vectors of coefficients considered as stochastic regression coefficients. The Kalman filter can be considered as a special case of the class of models treated in this paper and optimal estimators of the fixed starting state α_0 and "future" realizations α_t are obtained. Using a different approach, Shumway and Stoffer (1982) proposed the use of the EM algorithm (Dempster, Laird and Rubin, 1977) to derive a recursive procedure for estimating the parameters by maximum likelihood in time invariant state space models. Their method initialize the procedure by selecting starting values for the parameters $\boldsymbol{\alpha}_0$, $\boldsymbol{P}_0$, $\boldsymbol{T}$ and $\boldsymbol{Q}$ and then calculate the likelihood; then perform the E-step, running the traditional Kalman filter and obtaining smoothed values of $\boldsymbol{\alpha}_t$, $\boldsymbol{P}_{t|n}$ and $\boldsymbol{P}_{t,t-1|n}$; finally, perform the M-step updating the estimates $\boldsymbol{\alpha}_0$, $\boldsymbol{P}_0$, $\boldsymbol{T}_t$ and $\boldsymbol{Q}$. The same iterative procedure is repeated to convergence. Further details can be obtained in Shumway and Stoffer (2006, section 6.3). The method finally used for initialization in the state space model applied to business survey data in Chapter 4 was a diffuse prior as a better model in terms of fitting was achieved under this approach.

3.4.4 Diagnostic Checking and Goodness of Fit

In terms of the goodness of fit of the model, in a well-specified model, the standardized individual elements $\tilde{\boldsymbol{v}}_t$ (in the univariate case, $\boldsymbol{v}_t/\sqrt{\boldsymbol{F}_t}$ for $t = b+1, \cdots, n$) are serially uncorrelated and normally distributed with zero mean and constant variance, (Harvey, 1989, page 442). This can be checked by means of large-sample diagnostic tests and graphical procedures. QQ plots, histograms, tests of Shapiro and Wilk (1965) and Jarque and Bera (1980) are some ways to check normality. Additionally, autocorrelation plots and tests of serial correlation such as Ljung and Box (1978) (also called *Portmanteau test*) and Box and Pierce (1970) are useful. Harvey (1989, page

271) also proposes a F-diagnostic test for heteroscedasticity based on the standardized innovations $\tilde{\boldsymbol{v}}_t$ which has also been called *post-sample prediction test.*

Plots of the cumulative sum (CUSUM) and the cumulative sum of squares are also useful to detect stability problems in the parameters (Brown, Durbin and Evans, 1975). The CUSUM test is based on the cumulated sum of the standardized innovations

$$CUSUM_t = \sum_{j=b+1}^{t} \frac{\tilde{\boldsymbol{v}}_j}{\hat{\sigma}_{\tilde{\boldsymbol{v}}}} \tag{3.4.23}$$

Brown et al. (1975) show that $CUSUM_t$ has mean zero and variance proportional to $t-b-1$ and also that approximate 95% confidence intervals are given by the lines $\pm[0.948\sqrt{n-b}+1.896(t-b)/\sqrt{n-b}]$ for a significance level of 5%. If $CUSUM_t$ wanders outside these limits there is a failure in the stability assumption of the parameters. Harvey (1990, page 155) states that the CUSUM plot is also valuable for detecting structural breaks and includes an example of this use for the series of road accidents in Great Britain (Harvey, 1989, section 7.5.1). A set of statistics which also provide useful additional information are the estimates of the irregular disturbance term and the estimates of the disturbances in the transition equation, known in the literature as *auxiliary residuals.* For a general model in a SSF in Equations 3.3.1 and 3.3.2, these quantities are defined by (Durbin and Koopman, 2001)

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}_{t|n} &= \boldsymbol{y}_t - \boldsymbol{Z}_t\hat{\boldsymbol{\alpha}}_{t|n} \\ \hat{\boldsymbol{\vartheta}}_{t|n} &= \hat{\boldsymbol{\alpha}}_{t|n} - \boldsymbol{T}_t\hat{\boldsymbol{\alpha}}_{t-1|n} \end{aligned} \tag{3.4.24}$$

Kohn and Ansley (1989) and de Jong (1988*a*) developed the recursions to compute the disturbances in the observation equation directly during the Kalman filter and smoothing recursions without first calculating $\hat{\boldsymbol{\alpha}}_t$. Koopman (1993) developed the recursions for the disturbances in the transition equation. However, they are not serially independent (Kohn and Ansley, 1989) but they could be useful to detect outliers and structural breaks, respectively. In order to choose one from several candidate models, it is necessary to establish some comparison criterion. A possible way is to evaluate the value of the loglikelihood for each of the plausible models. In general, the larger the number of parameters the larger is the likelihood and then, information criteria such as the Akaike information criterion (AIC) and the Bayesian information

criterion (BIC) are used in order to penalise models with more parameters than others. These criteria are given by the formulae

$$\begin{aligned} \text{AIC} &= -2 * \text{log-likelihood} + 2 * \text{length}(\psi) \\ \text{BIC} &= -2 * \text{log-likelihood} + \log(n) * \text{length}(\psi) \end{aligned} \tag{3.4.25}$$

In general, big values of the loglikelihood are desirable and therefore, smaller values of AIC or BIC.

3.5 Benchmarking Based on State Space Methods

Considering the benchmarking problem presented in Chapter 1, Durbin and Quenneville (1997) propose two alternatives to produce benchmarked estimates using SSM that are cast in state space form. The first method (*two step method*) uses signal extraction to derive the smoothed estimators of the monthly signals without any benchmarking and then the smoothed series and the annual benchmarks are combined in order to compute the final adjusted estimates. In their paper, Durbin and Quenneville (1997) also consider the inclusion of trading days, treatment of multiplicative series and estimation of survey bias which will not be covered in this review. They also proposed a second method for benchmarking (*single step method*). The main difference from the first one is that instead of performing the estimation in two steps; the method incorporates into a single series both monthly and annual values and then arrange a suitable state space model for the combined series in order to obtain the benchmarked estimators.

This section describes the two methods proposed by Durbin and Quenneville (1997) and considers the special case when the adjusted series is forced to agree exactly with the benchmarks (*binding estimation*). The binding process is implemented by setting the variance of the annual survey errors to zero. However, it is necessary to account for the variance of the annual survey errors when computing the variances of the benchmarked estimators. We develop the theoretical expression of the correct

variance as well as an expression for the excess in the variance due to the binding process. As it turns out for the two step benchmarking method and under some specific conditions, the estimates in the second step after binding could actually be less accurate than the estimates obtained without benchmarking. The results are extended to the benchmarking methods based on regression and ARIMA model based approaches in the last chapter.

3.5.1 Two Step Benchmarking Method

Consider again the decomposition of univariate time series data into a signal η_t and a survey error ℓ_t in Equation 1.1.1. Using the Equation 2.3.8 to represent heteroscedastic survey errors, it follows that

$$y_t = \eta_t + \ell_t = \eta_t + k_t \ell_t^* \tag{3.5.1}$$

where ℓ_t^* is the standardized survey error and k_t is the standard deviation of the survey errors. The idea behind the term k_t is that since $\mathrm{Var}(\ell_t)=\mathrm{Var}(k_t\ell_t^*)$ then $\mathrm{Var}(\ell_t^*)=\mathrm{Var}(\ell_t/k_t)=\mathrm{Var}(\ell_t)/k_t^2 = 1$. Then, the term ℓ_t^* can be assumed as a unit-variance stationary ARMA(p,q) series and the model accounts for the heteroscedasticity of the survey errors (see Section 2.3). Consequently, it is necessary that all the values of k_t for $t = 1, \cdots, n$ and also the orders p and q in the survey errors ARMA model are known.

Following the same ideas as in other benchmarking methods it is assumed that there is a series of annual values x_i, $(i = 1, \cdots, m)$ available from another source and also considered as more accurate than the monthly values y_t's. We assume that errors in the benchmarks are independent of errors in the monthly observations. The x_i's are assumed to satisfy the benchmarking relations in Equation 2.4.3 introduced in section 2.4.

$$\boldsymbol{x} = \boldsymbol{L}'\boldsymbol{\eta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_e})$$

where $\boldsymbol{x}$ is an observed m x 1 vector of annual estimates, $\boldsymbol{L}$ is an n x m indicator matrix, and $\boldsymbol{e}$ is an m x 1 annual error vector having a multivariate normal distribution N($\mathbf{0}_m$, $\boldsymbol{\Sigma_e}$) and independent of $\boldsymbol{\eta}$ and $\boldsymbol{\ell}$. It is also assumed that $\boldsymbol{\Sigma_e}$ is known or can be estimated.

Suppose that the unobserved true series η_t follows a general structural time series model given by

$$\eta_t = \mu_t + \gamma_t + \epsilon_t \tag{3.5.2}$$

where μ_t, γ_t and ϵ_t are the trend, seasonal and irregular components respectively. It follows from Equation 3.5.1 that the observed series y_t follows the model

$$y_t = \eta_t + \ell_t = \underbrace{\mu_t + \gamma_t + \epsilon_t}_{\eta_t} + \underbrace{k_t \ell_t^*}_{\ell_t}, \quad t = 1, \cdots, n \tag{3.5.3}$$

with ℓ_t^* a unit variance ARMA(p,q) process and k_t denoting the standard deviation of the survey errors at time t as before. Following the ideas in subsection 2.5.2, the structural model in Equation 3.5.3 can be formulated into SSF and then using the Kalman filter one can get an estimate of η_t. Notice that we model y_t but the aim is to produce a preliminary estimate of η_t by signal extraction. We will denote the estimator of η_t in this first stage by $\tilde{\eta}_t$.

Proposition 3.5.1. *Consider the structural time series model for the observed series y_t given by Equation 3.5.3, writing this model into a SSF and then combining the results of the Kalman Filter and smoother with those in Hillmer and Trabelsi (1987), the BLUP estimator for $\boldsymbol{\eta}$ is given by*

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^0 + \hat{\boldsymbol{\eta}}_c \tag{3.5.4}$$

and its corresponding covariance matrix $\boldsymbol{\Sigma}_{\hat{\eta}}$ is given by

$$\boldsymbol{\Sigma}_{\hat{\eta}} = \boldsymbol{\Omega} - \boldsymbol{\Omega L}(\boldsymbol{L'\Omega L} + \boldsymbol{\Sigma}_e)^{-1}\boldsymbol{L'\Omega} \tag{3.5.5}$$

with $\hat{\boldsymbol{\eta}}^0$ the vector with single elements $\hat{\eta}_{0,t} = \boldsymbol{Z}_t\tilde{\boldsymbol{\alpha}}_t$; $\boldsymbol{\eta}_c = \boldsymbol{\Omega L}(\boldsymbol{L'\Omega L}+\boldsymbol{\Sigma}_e)^{-1}(\boldsymbol{x}-\boldsymbol{L'}\hat{\boldsymbol{\eta}}^0)$ and $\boldsymbol{\Omega} = [\varpi_{t,t^*}] = \boldsymbol{Z}_t Cov(\tilde{\boldsymbol{\alpha}}_t, \tilde{\boldsymbol{\alpha}}_{t^*})\boldsymbol{Z}'_{t^*}$ obtained after the application of the Kalman filter and smoother.

Proof. Let $\boldsymbol{\alpha}_{1,t}$ be the trend component in the state vector, $\boldsymbol{\alpha}_{2,t}$ the seasonal component, $\boldsymbol{\alpha}_{3,t}$ the irregular component and $\boldsymbol{\alpha}_{4,t}$ the survey error component as they were introduced in section 2.5.2. Setting $\boldsymbol{\alpha}_{3,t} = \epsilon_t$, implies that the irregular component of the observation equation is included in the state vector and it follows that

$$\boldsymbol{\alpha}_t = [\boldsymbol{\alpha}_{1,t} \quad \boldsymbol{\alpha}_{2,t} \quad \boldsymbol{\alpha}_{3,t} \quad \boldsymbol{\alpha}_{4,t}]' = [\boldsymbol{\alpha}_{1,t} \quad \boldsymbol{\alpha}_{2,t} \quad \epsilon_t \quad \boldsymbol{\alpha}_{4,t}]' \tag{3.5.6}$$

Let

$$\tilde{\boldsymbol{Z}}_t = [\boldsymbol{Z}_{1,t} \quad \boldsymbol{Z}_{2,t} \quad \boldsymbol{Z}_{3,t} \quad \boldsymbol{Z}^*_{4,t}] = [\boldsymbol{Z}_{1,t} \quad \boldsymbol{Z}_{2,t} \quad 1 \quad \boldsymbol{Z}^*_{4,t}] \tag{3.5.7}$$

where $\boldsymbol{Z}_{\iota,t}$ represents the corresponding matrix related to the element $\boldsymbol{\alpha}_{\iota,t}, \iota = 1, \cdots, 4$ in Equation 3.5.6 and $\boldsymbol{Z}^*_{4,t} = k_t \boldsymbol{Z}_{4,t}$. Equations 3.5.6 and 3.5.7 permit to write the observation equation in the form

$$y_t = \tilde{\boldsymbol{Z}}_t \boldsymbol{\alpha}_t = \mu_t + \gamma_t + \epsilon_t + k_t \ell^*_t \tag{3.5.8}$$

Moreover, using the corresponding elements $\boldsymbol{T}_{\iota,t}$ and $\boldsymbol{\vartheta}_{\iota,t}$ $(\iota = 1, \cdots, 3)$, it is possible to write the transition matrix and the disturbances in the transition equation by

$$\begin{aligned} \boldsymbol{T}_t &= \text{diag}(\boldsymbol{T}_{1,t}; \boldsymbol{T}_{2,t}; \boldsymbol{T}_{3,t}; \boldsymbol{T}_{4,t}) = \text{diag}(\boldsymbol{T}_{1,t}; \boldsymbol{T}_{2,t}; 0; \boldsymbol{T}_{4,t}) \\ \boldsymbol{\vartheta}_t &= [\boldsymbol{\vartheta}_{1,t} \quad \boldsymbol{\vartheta}_{2,t} \quad \boldsymbol{\vartheta}_{3,t} \quad \boldsymbol{\vartheta}_{4,t}] = [\boldsymbol{\vartheta}_{1,t} \quad \boldsymbol{\vartheta}_{2,t} \quad \epsilon_t \quad \boldsymbol{\vartheta}_{4,t}] \end{aligned} \tag{3.5.9}$$

to get the transition equation

$$\boldsymbol{\alpha}_t = \boldsymbol{T}_t \boldsymbol{\alpha}_{t-1} + \boldsymbol{\vartheta}_t \tag{3.5.10}$$

Once the model is formulated in SSF, it is possible to use the Kalman filter to get the estimate $\hat{\boldsymbol{\eta}}_0 = E(\boldsymbol{\eta} \mid \boldsymbol{y})$. Using the observation equation $y_t = \tilde{\boldsymbol{Z}}_t \boldsymbol{\alpha}_t$, then $\eta_t = \boldsymbol{Z}_t \boldsymbol{\alpha}_t$ where $\boldsymbol{Z}_t$ has the same form as $\tilde{\boldsymbol{Z}}_t$ replacing $\boldsymbol{Z}^*_{4,t}$ by a suitable vector of zeroes (this is because η_t does not contain the survey errors). Once $\tilde{\boldsymbol{\alpha}}_t = E(\alpha_t \mid \boldsymbol{y})$ is calculated

using the filtering and smoothing recursions analogously to Equations 3.4.11; it follows that a preliminary estimator of the signals η_t is $\hat{\eta}_{0,t} = \boldsymbol{Z}_t\tilde{\boldsymbol{\alpha}}_t$ with covariance matrix $\boldsymbol{\Omega} = [\varpi_{t,t^*}] = [\text{Cov}(\hat{\eta}_{0,t}; \hat{\eta}_{0,t^*})]$ for $t, t^* = 1, \cdots, n$. Also, according to the observation equation and the fact that the irregular term has been included in the state vector, it follows that $\varpi_{t,t^*} = \boldsymbol{Z}_t\text{Cov}(\tilde{\boldsymbol{\alpha}}_t, \tilde{\boldsymbol{\alpha}}_{t^*})\boldsymbol{Z}'_{t^*}$. The recursive expression for the last expression $\text{Cov}(\tilde{\boldsymbol{\alpha}}_t, \tilde{\boldsymbol{\alpha}}_{t^*})$ can be calculated from Equation 3.4.10.

The second stage uses the results in Hillmer and Trabelsi (1987) for which theoretical details are included in Appendix A.3. The final estimate incorporating the annual information and its corresponding variance are given by the expressions

$$\hat{\boldsymbol{\eta}} = E(\boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{y}) = \hat{\boldsymbol{\eta}}^0 + \boldsymbol{\eta}_c$$

where

$$\boldsymbol{\eta}_c = \boldsymbol{\Omega}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Omega}\boldsymbol{L} + \boldsymbol{\Sigma}_e)^{-1}(\boldsymbol{x} - \boldsymbol{L}'\hat{\boldsymbol{\eta}}^0)$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} = Cov(\boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Omega}\boldsymbol{L} + \boldsymbol{\Sigma}_e)^{-1}\boldsymbol{L}'\boldsymbol{\Omega}$$

□

An example of the application of this method to business surveys in the UK will be presented in Chapter 4 where some structural time series model will be considered with their corresponding formulation as SSF.

3.5.2 Single Step Benchmarking Method

Durbin and Quenneville (1997) proposed a second method of benchmarking. The difference from the first one is that instead of performing the estimation in two stages;

both series (monthly and annual) are combined into a single series with the filtering and smoothing procedures applied in a single stage.

The single series containing both the high and low frequency series is arranged as

$$\boldsymbol{y}^* = [y_1, \cdots, y_K, x_1, y_{K+1}, \cdots, y_{2K}, x_2, y_{2K+1}, \cdots, y_n, (x_m)] \tag{3.5.11}$$

where (x_m) signifies that the single series could represent complete years with the final element being x_m or could represent incomplete years with the final element being y_n representing the last available monthly data. Also K represents the number of subperiods per year as it was noted before and the elements $y_t, t = 1, \cdots, n$ and $x_i, i = 1, \cdots, m$ follow the equations below

$$\begin{aligned} y_t &= \eta_t + \ell_t = \eta_t + k_t \ell_t^* \\ x_i &= \sum_{(i-1)K+1}^{iK} \eta_t + e_i \end{aligned} \tag{3.5.12}$$

The total length of the series $\boldsymbol{y}^*$ will be $n + m$ independently of whether it ends with an element y_n or an element x_m and $s = 1, \cdots, n + m$ will denote the index of each position in the series $\boldsymbol{y}^*$. This characteristic allows the method to be an online procedure, i.e. it is not necessary to wait until the next benchmark is available to apply the procedure. All the benchmarking methods reviewed before require having available the benchmark for the last year before applying the corresponding method. This is not the case now and therefore, the single stage benchmarking method is a solution of the ex-ante estimation problem presented in pages 4 and 5.

Proposition 3.5.2. *Consider an array containing both the high and low frequency series in a single series. After writing the structural model for the new series in an appropriate SSF, the smoothed benchmarked estimates and their respective variances are given by*

$$\hat{\boldsymbol{\eta}}_s = \boldsymbol{Z}_s^* \boldsymbol{\alpha}_s^* + \boldsymbol{\epsilon}_s^* \tag{3.5.13a}$$

$$Var(\hat{\boldsymbol{\eta}}_s) = \boldsymbol{Z}_s^* \boldsymbol{P}_s^* \boldsymbol{Z}'^*_s + \sigma^2_{\epsilon^* s} \quad s = 1, \cdots, n + m \tag{3.5.13b}$$

with $n + m$ the length of the new series and the other components given by the SSF.

Proof. Let us assume that the values $y_t \quad t = 1, \cdots, n$ in $\boldsymbol{y}^*$ follow the same structural time series model used in the two stage method and formulated in Equation 3.5.3. It follows that

$$y_t = \mu_t + \gamma_t + \epsilon_t + k_t \ell_t^*, \quad t = 1, \cdots, n$$

Writing this structural time series model for the values y_t along with the equation $x_i = \sum_{(i-1)K+1}^{iK} \eta_t + e_i$ for the values $x_i, i = 1, \cdots, m$ into SSF, an estimate of η_t can be obtained using the Kalman filter. Again the modelling is done over $\boldsymbol{y}^*$ to produce an estimate of $\boldsymbol{\eta}$. Considering $\boldsymbol{\alpha}_{1,s}, \boldsymbol{\alpha}_{2,s}$, $\boldsymbol{\alpha}_{3,s}$ and $\boldsymbol{\alpha}_{4,s}$ the corresponding components of the trend, seasonality, irregular terms and the survey errors, respectively. Since every term x_i in the series $\boldsymbol{y}^*$ depends on the last K values, it will be necessary to consider the trend, seasonality and irregular components in the state vector each one with a subvector of length equal to K. That is not necessary for $\boldsymbol{\alpha}_{4,s}$ as η_t is not affected by the survey errors. Then, the length of the vector $\boldsymbol{\alpha}_{4,s}$ is $\varrho = \max(p, q)$ using the notation given in page 33 and p, q denoting the respective orders of the ARMA(p, q) model of the survey errors. The total length of the state vector will be equal to $3K + \varrho$ with K the number of high frequencies per year. In this way, the observation equation will refer to two kinds of values: $y_t, t = 1, \cdots, n$ and $x_i, i = 1, \cdots, m$ according to Equation 3.5.12. This is achieved by considering a state space vector given by

$$\begin{aligned} \boldsymbol{\alpha}^*{}_s &= [\boldsymbol{\alpha}_{1,s} \quad \boldsymbol{\alpha}_{2,s} \quad \boldsymbol{\alpha}_{3,s} \quad \boldsymbol{\alpha}_{4,s}] \\ &= [\mu_s, \cdots, \mu_{s-K+1} \mid \gamma_s, \cdots, \gamma_{s-K+1} \mid \epsilon_s, \cdots, \epsilon_{s-K+1} \mid \boldsymbol{\alpha}_{4,s}] \end{aligned} \tag{3.5.14}$$

with observation matrices as

$$\tilde{Z}^*{}_s = \begin{cases} [10_{K-1} \mid 10_{K-1} \mid 10_{K-1} \mid k_t 0_{\varrho-1}], & \text{if} \quad y_s^* = y_t \quad t = 1, \cdots, n \\ [1_K \mid 1_K \mid 1_K \mid 0_\varrho], & \text{if} \quad y_s^* = x_i \quad i = 1, \cdots, m \end{cases} \tag{3.5.15}$$

and disturbances

$$\boldsymbol{\varepsilon}^*{}_s = \begin{cases} 0, & y_s^* = y_t \quad t = 1, \cdots, n \\ e_i, & y_s^* = x_i \quad i = 1, \cdots, m \end{cases} \tag{3.5.16}$$

with variances given by

$$\sigma^2_{\varepsilon^* s} = \begin{cases} 0, & y^*_s = y_t \quad t = 1, \cdots, n \\ \sigma^2_{e_i}, & y^*_s = x_i \quad i = 1, \cdots, m \end{cases} \tag{3.5.17}$$

with $\sigma^2_{e_i}$ denoting the variance of the annual survey errors.

Durbin and Quenneville (1997, page 38) state that it is necessary to assume that Σ_e is a diagonal matrix; otherwise, the state vector becomes too large. It can also be noted that, more importantly, if this assumption is not made the disturbances in the observation equation will become autocorrelated. As an alternative, Pfeffermann and Tiller (2006) have developed a new filtering algorithm for state space models with correlated disturbances in the observation equation giving a possible solution to overcome this problem.

Equations 3.5.14 to 3.5.17 permit writing the observation equation in the form

$$y^*_s = \tilde{Z}^*{}_s \alpha^*{}_s + \varepsilon^*{}_s \quad \varepsilon^*{}_s \sim iid(0, \sigma^2_{\varepsilon^*}) \quad s = 1, \cdots, n+m \tag{3.5.18}$$

which is equivalent to the set of equations 3.5.12. In the binding case

$$y^*_s = \tilde{Z}^*{}_s \alpha^*{}_s \quad s = 1, \cdots, n+m \tag{3.5.19}$$

Also, using the corresponding elements $T_{\iota,s}$ and $\vartheta_{\iota,s}$, $\quad \iota = 1, \cdots, 4$, it is possible to write the transition equation. The transition equation defines the development of the system from one instant to the other. Going over the index $s = 1, \cdots, n+m$, the transition from one element y_t to the next is described by the original matrices $T_{\iota,s}$. At the same time, the transitions have to "jump" the elements x_i in order to keep the continuity in the series. Then, the transition from the last element in year i, denoted by $y_{iK}, i = 1, \cdots, m$ to the next element x_i must keep the same state vector in x_i as for y_{iK}. There is no transition from y_t to x_i and the identity matrix could be used as T^*_s. Now from x_i to y_{iK+1}, it is possible to use the same transition matrix from y_{iK} to the next y_{iK+1} as the state vector for x_i is the same as the state vector for y_{iK}.

Given all these technical considerations the transition matrices and the disturbances in the transition equation could be arranged as

$$T^*_s = \begin{cases} diag(T_{1,s}, T_{2,s}, T_{3,s}, T_{4,s}), & \text{if} \quad y^*_s = y_t \quad t = 1, \cdots, n \\ I_{3K+r_3}, & \text{if} \quad y^*_s = x_i \quad i = 1, \cdots, m \end{cases} \tag{3.5.20}$$

$$\vartheta^*_s = \begin{cases} [\vartheta_{1,s}, \vartheta_{2,s}, \vartheta_{3,s}, \vartheta_{4,s}], & \text{if} \quad y^*_s = y_t \quad t = 1, \cdots, n \\ 0_{3K+r_4}, & \text{if} \quad y^*_s = x_i \quad i = 1, \cdots, m \end{cases} \tag{3.5.21}$$

Durbin and Quenneville (1997, page 36) set up a different transition matrix for the application to the Canadian retail sales data leading to the same benchmarked estimates. After setting up this SSF, the procedure is implemented by applying the Kalman filter and smoother with the specifications before. Finally to obtain $\hat{\eta}$, it is necessary to replace $\tilde{Z}^*_s$ by

$$Z^*_s = \begin{cases} [10_{K-1} \mid 10_{K-1} \mid 10_{K-1} \mid 0_\varrho], & \text{if} \quad y^*_s = y_t \quad t = 1, \cdots, n \\ [1_K \mid 1_K \mid 1_K \mid 0_\varrho], & \text{if} \quad y^*_s = x_i \quad i = 1, \cdots, m \end{cases} \tag{3.5.22}$$

analogously to what was done in the two stage method above. The filtered and smoothed values are

$$\hat{\eta}_s = Z^*_s \alpha^*_s + \varepsilon^*_s \tag{3.5.23a}$$

$$Var(\hat{\eta}_s) = Z^*_s P^*_s Z'^*_s + \sigma^2_{\varepsilon^* s} \quad s = 1, \cdots, n+m \tag{3.5.23b}$$

with P^*_s obtained during the Kalman filter recursions for the single series y^* in Equation 3.5.11. Notice than after producing the values $\hat{\eta}_s$ with $s = 1, \cdots, n+m$ in Equation 3.5.23a; it is necessary to get the unstacked values $\hat{\eta}_t$ with $t = 1, \cdots, n$ corresponding to the monthly benchmarked values. □

This method has many advantages over the other benchmarking methods. Firstly, it is an online procedure, which means that one does not need to wait until the next benchmark is available. Secondly, it does not require any estimation of the autocovariance

matrices of the monthly and annual survey errors. This characteristic of the method makes easier its application. Nevertheless, it is still necessary to specify an ARMA model for the survey errors and estimate its corresponding parameters. Another disadvantage is that the dimensionality of the vectors and matrices involved in the state space formulation tends to increase very profoundly.

3.6 Binding and Non-Binding Estimation

A common practice of time series analysts is to adjust the monthly series such that it satisfies exactly the annual benchmarks. Alternative methods such as Denton (1971) considered this problem, although it was not possible to get an estimate of the variance of the estimates. The other benchmarking methods presented in the last chapter, consider the special case when the annual data come from a census or a complete enumeration making the covariance matrix of the annual survey errors to be equal to zero. The estimation, in this case, is called *binding estimation*. Sometimes the annual restrictions are imposed, even in the case where the annual benchmarks are obtained from a survey. In the latter case it is necessary, however, to account for the variability (sampling errors) of the benchmarks when computing the variances of the monthly benchmarked estimators. Using state space models, Pfeffermann and Tiller (2006) show how to obtain the variance of binding estimators (obtained from a binding procedure) when using contemporaneous benchmarks that are subject to survey errors.

In the next subsections, we develop the correct variance of binding estimators when using temporal annual benchmarks that are subject to survey error. The theory is first presented for the two-step method using state space models, and then extended to the ARIMA and Regression approaches introduced in Chapter 2.

3.6.1 Two Step Benchmarking Method

There are two special situations to be considered when the two stage benchmarking method is applied to benchmark monthly data to annual totals. In the first situation, after the monthly smoothed values $\tilde{\eta}$ are obtained in a first stage, they are benchmarked to annual estimates, considering the latter as restrictions (*binding estimation*). In a second situation, the annual estimates are considered as auxiliary information subject to survey errors in order to get more precise estimators, but they do not necessarily satisfy the annual restrictions (*non binding estimation*).

In the second stage, after the filtering and smoothing processes are completed, the benchmarked estimates could be obtained using the Hillmer and Trabelsi estimator given in Proposition 2.4.1. We will refer to this estimator as the *non-binding estimator.* This estimator corresponds to the n x 1 vector

$$\hat{\eta} = E(\eta \mid x, y) = \tilde{\eta} + \Omega L[L'\Omega L + \Sigma_e]^{-1}(x - L'\tilde{\eta}) \tag{3.6.1}$$

and variance given by the matrix

$$\Sigma_{\hat{\eta}} = Var(\eta \mid x, y) = \Omega - \Omega L[L'\Omega L + \Sigma_e]^{-1}L'\Omega \tag{3.6.2}$$

with dimension n x n.

Another possibility to consider is the use of the *binding estimator* corresponding to

$$\hat{\eta}_B = \tilde{\eta} + \Omega L[L'\Omega L]^{-1}(x - L'\tilde{\eta}) \tag{3.6.3}$$

assuming that the autocovariance of the annual estimates is equal to a zero matrix. This assumption implies a complete fulfilment of the benchmark restrictions since

$$L'\hat{\eta}_B = L'\tilde{\eta} + x - L'\tilde{\eta} = x \tag{3.6.4}$$

Using the estimator in Equation 3.6.3, the sum of the monthly estimates per year will be exactly the corresponding annual estimates.

However, the practice of making the variance of the last estimator to be equal to the analogue of Equation 3.6.2, replacing Σ_e as a zero matrix and using the formula

$$\Sigma_{\hat{\eta}B} = \Omega - \Omega L[L'\Omega L]^{-1}L'\Omega \tag{3.6.5}$$

is only valid if it is possible to guarantee that the annual values come from a census or a complete enumeration. This is because if the sum of the binding estimates is considered, then

$$Var(x) = Var(L'\hat{\eta}_B) = L'\Sigma_{\hat{\eta}B}L = L'\Omega L - L'\Omega L = 0_{mxm} \tag{3.6.6}$$

As it was explained above, binding estimation requires to assume that the variance of the annual estimates is zero. This is sometimes assumed even when the annual benchmarks are obtained from a survey instead from a census. Then, it is necessary to add a new term in the expression of the variance of the benchmarked estimates in order to account for the variability of the benchmarks when computing the variances of the monthly benchmarked estimators.

In order to get the correct expression for the variance of the binding estimator subject to annual survey errors; the binding estimator will be expressed as a function of the non binding estimator writing

$$\begin{aligned}
\hat{\eta}_B &= \hat{\eta} + (\hat{\eta}_B - \hat{\eta}) \\
&= \hat{\eta} + \Omega L[L'\Omega L]^{-1}(x - L'\tilde{\eta}) - \Omega L[L'\Omega L + \Sigma_e]^{-1}(x - L'\tilde{\eta}) \\
&= \hat{\eta} + \Omega L[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}](x - L'\tilde{\eta}) \\
&= \hat{\eta} + B^*(x - L'\tilde{\eta})
\end{aligned} \tag{3.6.7}$$

where B^* is a n x m matrix equals to $\Omega L[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}]$. Since $\Sigma_e \neq 0$, Equations 3.6.5 and 3.6.6 do not hold. Denoting the variance matrix of the binding estimator under the presence of annual survey errors as $\Sigma^c_{\hat{\eta}B}$, with the superindex c indicating that the binding variance has been corrected due to $\Sigma_e \neq 0$; the next proposition calculates the variance for $\hat{\eta}_B$.

Proposition 3.6.1. *The variance $\Sigma^c_{\hat{\eta}B}$ can be decomposed in terms of the variance of the non-binding estimator plus an additional term as*

$$\Sigma^c_{\hat{\eta}B} = \Sigma_{\hat{\eta}} + \Omega L(L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}]L'\Omega \tag{3.6.8}$$

with Ω being the autocovariance matrix of the smoothed values $\tilde{\eta}$ in the first step.

Proof. Using the decomposition of the binding estimator in terms of the non-binding estimator in Equation 3.6.7, multivariate properties of the variance and covariance operators and some matrix algebra, it follows that

$$\begin{aligned}
\Sigma^c_{\hat{\eta}B} &= \mathrm{Var}(\hat{\eta}) + \mathrm{Cov}(\hat{\eta}, B^*(x - L'\tilde{\eta})) + \mathrm{Cov}(B^*(x - L'\tilde{\eta}), \hat{\eta}) + \mathrm{Var}(B^*(x - L'\tilde{\eta})) \\
&= \mathrm{Var}(\hat{\eta}) + \mathrm{Cov}(\hat{\eta}, x)B^{*'} - \mathrm{Cov}(\hat{\eta}, L'\tilde{\eta})B^{*'} + B^*\mathrm{Cov}(x, \hat{\eta}) - B^*\mathrm{Cov}(L'\tilde{\eta}, \hat{\eta}) \\
&\quad + B^*(\Sigma_e + L'\Omega L)B^{*'} \\
&= \mathrm{Var}(\hat{\eta}) + \mathrm{Cov}(\tilde{\eta}, x)B^{*'} + \mathrm{Cov}(\Omega L[L'\Omega L + \Sigma_e]^{-1}(x - L'\tilde{\eta}), x)B'^* \\
&\quad - \mathrm{Cov}(\tilde{\eta}, L'\tilde{\eta})B^{*'} - \mathrm{Cov}(\Omega L[L'\Omega L + \Sigma_e]^{-1}(x - L'\tilde{\eta}), L'\tilde{\eta})B^{*'} + B^*\mathrm{Cov}(x, \tilde{\eta}) \\
&\quad + B^*\mathrm{Cov}(x, \Omega L[L'\Omega L + \Sigma_e]^{-1}(x - L'\eta)) - B^*\mathrm{Cov}(L'\tilde{\eta}, \tilde{\eta}) \\
&\quad - B^*\mathrm{Cov}(L'\tilde{\eta}, \Omega L[L'\Omega L + \Sigma_e]^{-1}(x - L'\tilde{\eta})) + B^*(\Sigma_e + L'\Omega L)B^{*'} \\
&= \mathrm{Var}(\hat{\eta}) + \Omega L[L'\Omega L + \Sigma_e]^{-1}[\Sigma_e + L'\Omega L]B^{*'} - \Omega L B^{*'} \\
&\quad + B^*[\Sigma_e + L'\Omega L][L'\Omega L + \Sigma_e]^{-1}L'\Omega + B^*(\Sigma_e + L'\Omega L)B^{*'} - B^*L'\Omega \\
&= \mathrm{Var}(\hat{\eta}) \\
&\quad + \Omega L[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}](\Sigma_e + L'\Omega L)[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}]L'\Omega \\
&= \mathrm{Var}(\hat{\eta}) \\
&\quad + [\Omega L(L'\Omega L)^{-1}\Sigma_e + \Omega L(L'\Omega L)^{-1}(L'\Omega L) - \Omega L] \cdot [(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}]L'\Omega
\end{aligned}$$

and then, finally,

$$\boxed{\Sigma^c_{\hat{\eta}B} = \Sigma_{\hat{\eta}} + \Omega L(L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}]L'\Omega}$$

as it was required. □

The last expression permits to decompose the variance of the binding estimator into the variance of the non-binding estimator plus an extra term. In the particular case when $\Sigma_e = 0$, $\Sigma^c_{\hat{\eta}B} = \Sigma_{\hat{\eta}}$. One of the main purposes of benchmarking is to obtain better estimates than those used before to the benchmarking process.

Denoting the expression $M_1 \geq M_2$ to represent that the matrix $M_1 - M_2$ is positive semidefinite and, in the same way, considering $M_1 > M_2$ to denote that the matrix

$M_1 - M_2$ is positive definite. The following two propositions address the issue of efficiency of binding and non-binding estimators. Firstly, the optimality of the non-binding estimator $\hat{\eta}$ against the binding estimator $\hat{\eta}_B$ subject to annual survey errors is studied. Then, some conditions are determined in order to decide if the binding estimator in the second stage $\hat{\eta}_B$ is optimal than the smoothed estimates in the first stage $\tilde{\eta}$ or the other way round.

Proposition 3.6.2. *The binding estimator under annual survey errors is less efficient than the non-binding estimator. In other words,* $\Sigma^c_{\hat{\eta}B} > \Sigma_{\hat{\eta}}$. *The excess in the variance of* $\hat{\eta}_B$ *with respect to the variance of* $\hat{\eta}$ *is given by* $\Omega L(L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}L'\Omega$ *representing the excess due to the presence of a non-zero annual survey error variance* Σ_e.

Proof. In order to prove that $\Sigma^c_{\hat{\eta}B} > \Sigma_{\hat{\eta}}$, it is necessary to guarantee that the matrix $\Omega L(L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}]L'\Omega$ in the second term of Equation 3.6.8 is always positive definite. This is possible to assure if the matrix

$$(L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}] \tag{3.6.9}$$

is always positive definite (see Harville (1997, theorem 14.2.9)). Using the Woodbury matrix identity (Golub and van Loan, 1996, page 50)

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B(D^{-1} + B'A^{-1}B)^{-1}B'A^{-1} \tag{3.6.10}$$

and taking $A = L'\Omega L$, $B = I$ and $D = \Sigma_e$, the matrix in Equation 3.6.9 can be expressed as

$$\begin{aligned}
&(L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1} - (L'\Omega L + \Sigma_e)^{-1}] \\
&= (L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1} - (L'\Omega L)^{-1} + (L'\Omega L)^{-1}({\Sigma_e}^{-1} + (L'\Omega L)^{-1})^{-1}(L'\Omega L)^{-1}] \\
&= (L'\Omega L)^{-1}\Sigma_e[(L'\Omega L)^{-1}({\Sigma_e}^{-1} + (L'\Omega L)^{-1})^{-1}(L'\Omega L)^{-1}]
\end{aligned} \tag{3.6.11}$$

Now using that $(A+B)^{-1} = A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}$, it follows that the Equation 3.6.11 can be expressed as

$$\begin{aligned}
&(L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}[\Sigma_e{}^{-1} + (L'\Omega L)^{-1}]^{-1}(L'\Omega L)^{-1} \\
&= (L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}(L'\Omega L)(L'\Omega L + \Sigma_e)^{-1}\Sigma_e(L'\Omega L)^{-1} \qquad (3.6.12)\\
&= (L'\Omega L)^{-1}\Sigma_e(L'\Omega L + \Sigma_e)^{-1}\Sigma_e(L'\Omega L)^{-1}
\end{aligned}$$

Then,

$$\Sigma^c_{\hat{\eta}B} = \Sigma_{\hat{\eta}} + \Omega L(L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}L'\Omega \qquad (3.6.13)$$

where the second term is a positive definite matrix as $(\Sigma_e + L'\Omega L)$ is the sum of two positive definite matrices and its inverse is also a positive definite matrix. Then, assuming $\Sigma_e \neq 0$, $\Sigma^c_{\hat{\eta}B}$ is always greater than $\Sigma_{\hat{\eta}}$ and the second term in Equation 3.6.13 shows how much you lose in variance when considering binding estimators. □

Finally, we found conditions under which the binding estimator in the second step $\hat{\eta}_B$ could have a bigger variance than the smoothed estimator in the first step, $\tilde{\eta}$.

Proposition 3.6.3. *The variance of the binding estimator in the case of annual survey errors is greater than the variance of the monthly smoothed estimator in the first stage, if the generalized total variance of Σ_e (the covariance matrix of the annual survey errors) is greater than the generalized total variance of $L'\Omega L$ (the covariance matrix of the sum of the smoothed monthly estimates obtained in the first step). In matrix notation, if $|L'\Omega L| \leq |\Sigma_e|$ then $\Sigma^c_{\hat{\eta}B} \geq \Omega$.*

Proof. If the $\Sigma^c_{\hat{\eta}B}$ is written not in terms of $\Sigma_{\hat{\eta}}$ but in terms of Ω, which represents the variance of the monthly smoothed estimates in the first stage, it follows that

$$\begin{aligned}
\Sigma^c_{\hat{\eta}B} &= \Omega - \Omega L[L'\Omega L + \Sigma_e]^{-1}L'\Omega + \Omega L(L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}L'\Omega \\
&= \Omega - \Omega L\{[L'\Omega L + \Sigma_e]^{-1} - (L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}\}L'\Omega
\end{aligned} \qquad (3.6.14)$$

Because the negative sign in Equation 3.6.14, if $\Sigma^{c}_{\hat{\eta}B} < \Omega$ then the matrix

$$[L'\Omega L + \Sigma_e]^{-1} - (L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1} \tag{3.6.15}$$

must be positive definite.

Thus, *if* $\Sigma^{c}_{\hat{\eta}B} < \Omega$ *then it can be guaranteed that* $[L'\Omega L + \Sigma_e]^{-1} - (L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}$ *is positive definite* (1). It can also be proved that if $A - B$ is positive definite, then $|A| > |B|$ by using theorem 25 in Magnus and Neudecker (1988, page 22). According to this last proposition, it follows that if $[L'\Omega L + \Sigma_e]^{-1} - (L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}$ is positive definite then

$$\begin{aligned} &|(L'\Omega L + \Sigma_e)^{-1}| > |(L'\Omega L)^{-1}\Sigma_e(\Sigma_e + L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}| \\ &\Rightarrow \frac{1}{|L'\Omega L + \Sigma_e|} > \frac{|\Sigma_e|^2}{|L'\Omega L|^2|\Sigma_e + L'\Omega L|} \\ &\Rightarrow |L'\Omega L|^2 > |\Sigma_e|^2 \\ &\Rightarrow |L'\Omega L| > |\Sigma_e| \end{aligned} \tag{3.6.16}$$

The last result can be summarized as *if* $[L'\Omega L+\Sigma_e]^{-1}-(L'\Omega L)^{-1}\Sigma_e(\Sigma_e+L'\Omega L)^{-1}\Sigma_e(L'\Omega L)^{-1}$ *is positive definite then the generalised total variance of* $L'\Omega L$ *is greater than the generalised total variance of* Σ_e (2). From conclusions (1) and (2), it can be concluded that *if* $\Sigma^{c}_{\hat{\eta}B} \leq \Omega$ *then the generalised total variance of* $L'\Omega L$ *is greater than the generalised total variance of* Σ_e (3). Now using the negation of (3), it follows that *if* $|L'\Omega L| \leq |\Sigma_e|$ *then* $\Sigma^{c}_{\hat{\eta}B} \geq \Omega$ as appears in the proposition. □

The last result implies that the actual variance of the binding estimates in the second stage could be greater than the variance of the smoothed estimates in the first stage. Clearly, in this case, there is no reason for the benchmarking except for consistency in publication. All the results in this section can be extended to the ARIMA and Regression approaches as it will be shown in the next subsections. This extension is done by considering both approaches to be analogous to a "two-step" method type.

3.6.2 ARIMA Approach Method

The ARIMA method is also a "two step method". The first step uses the stochastic structure of the monthly data to produce an estimator of the signal and its corresponding autocovariance matrix $\boldsymbol{\Omega}_1 = Cov(\boldsymbol{\eta} \mid \boldsymbol{y})$. In the second step, the annual information is incorporated in exactly the same way as the two-step benchmarking method. Regarding the variance of the benchmarked estimates, it follows that $\boldsymbol{\Omega}_1 = (\boldsymbol{\Sigma}_{\boldsymbol{\ell}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1})^{-1}$ with $\boldsymbol{\Sigma}_{\boldsymbol{\ell}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ representing the covariance matrices of the monthly survey errors and the underlying true time series, respectively. These covariance matrices are obtained from the corresponding ARIMA models in the way that was explained in section 2.4. The same results in the last subsection above are obtained with $\boldsymbol{\Omega}$ replaced by $\boldsymbol{\Omega}_1$. Then, it can also be concluded that *if* $|\boldsymbol{L}'(\boldsymbol{\Sigma}_{\boldsymbol{\ell}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1})^{-1}\boldsymbol{L}| \leq |\boldsymbol{\Sigma}_{\boldsymbol{e}}|$ *then* $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\eta}}B}^{c} \geq \boldsymbol{\Omega}_1$ and the incorporation of the annual restrictions would not make any improvement in the estimation when this condition is satisfied.

3.6.3 Regression Approach Method

The regression method can also be considered as a "two-step method". An estimator of the monthly signal is obtained subtracting the estimated bias in a first step. Then, in a second stage, the annual information is incorporated into the estimation in the form of a generalized regression model. Hillmer and Trabelsi (1987) method could be considered as a particular case of the Cholette and Dagum (1994) method replacing the matrix $\boldsymbol{\Omega}$ for the covariance monthly survey error matrix $\boldsymbol{\Sigma}_{\boldsymbol{\ell}}$, the vector $\tilde{\boldsymbol{\eta}}$ for the vector $\boldsymbol{y}$ and considering no bias parameter. Therefore, two different cases will be considered according to the presence of bias in the estimation.

Zero bias case

If it is assumed that the bias is zero, analogous results to those obtained for the two stage method are produced for the binding estimator. The equations below summarize

the main equations for the non-binding and binding estimator under the regression approach assuming no survey bias. $\hat{\eta}_{nb}$ denotes the estimator under the no bias assumption.

Non-binding Estimator

$$\hat{\eta}_{nb} = y + \Sigma_\ell L[L'\Sigma_\ell L + \Sigma_e]^{-1}(x - L'y) \tag{3.6.17a}$$

$$\Sigma_{\hat{\eta},nb} = \Sigma_\ell - \Sigma_\ell L[L'\Sigma_\ell L + \Sigma_e]^{-1}L'\Sigma_\ell \tag{3.6.17b}$$

Binding Estimator

$$\hat{\eta}_{B,nb} = y + \Sigma_\ell L[L'\Sigma_\ell L]^{-1}(x - L'y) \tag{3.6.18a}$$

$$\Sigma^c_{\hat{\eta}B,nb} = \Sigma_{\hat{\eta},nb} + \Sigma_\ell L(L'\Sigma_\ell L)^{-1}\Sigma_e(\Sigma_e + L'\Sigma_\ell L)^{-1}\Sigma_e(L'\Sigma_\ell L)^{-1}L'\Sigma_\ell \tag{3.6.18b}$$

Equation 3.6.17a gives the expression for the non-binding estimator; Equation 3.6.17b the corresponding variance; Equation 3.6.18a the corresponding formula for the binding estimator and Equation 3.6.18b the corrected variance for the binding estimator with benchmarks being subjected to survey errors. Following the results for binding estimation in the two stage method, it is possible to conclude that *if* $|L'\Sigma_\ell L| \leq |\Sigma_e|$ *then* $\Sigma^c_{\hat{\eta}B} \geq \Sigma_\ell$ and the variance of the binding estimator could be greater than the variance of the original monthly estimates.

Non-zero bias case

According to the results in Proposition 2.3.1., the bias and the benchmarked estimator with their respective variances can be obtained using

$$\hat{a} = -\sigma^2_{\hat{a}} 1'L(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y) \tag{3.6.19a}$$

$$\hat{\eta} = y^* + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y^*), \quad y^* = y - 1_n\hat{a} \tag{3.6.19b}$$

$$\sigma^2_{\hat{a}} = 1/[1'_n L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n] \tag{3.6.19c}$$

$$\begin{aligned}\Sigma_{\hat{\eta}} &= [\Sigma_\ell - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell] \\ &\quad + [I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']1_n\sigma^2_{\hat{a}}1'_n[I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']'\end{aligned} \tag{3.6.19d}$$

When considering a survey bias, Cholette and Dagum (1994) method could be consider as an analogous case of Hillmer and Trabelsi (1987) replacing the matrix $\boldsymbol{\Omega}$ for the covariance matrix $\boldsymbol{\Sigma_\ell}$ and the vector $\tilde{\boldsymbol{\eta}}$ for the vector $\boldsymbol{y} - \mathbf{1}_n\hat{a}$. Now, analogously to the other benchmarking methods, the binding estimator could be defined as

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{B}} = \boldsymbol{y}^* + \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}(\boldsymbol{x} - \boldsymbol{L}'\boldsymbol{y}^*), \quad \boldsymbol{y}^* = \boldsymbol{y} - \mathbf{1}_n\hat{a} \tag{3.6.20}$$

assuming $\boldsymbol{\Sigma_e} = \mathbf{0}$.

Now, considering the difference between the estimator under non-zero bias in Equation 3.6.19b with the estimator under zero bias in Equation 3.6.17a. It follows that

$$\begin{aligned} \hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}_{nb} &= \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L} + \boldsymbol{\Sigma_e})^{-1}\boldsymbol{L}'\mathbf{1}_n\hat{a} - \mathbf{1}_n\hat{a} \\ &= [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L} + \boldsymbol{\Sigma_e})^{-1}\boldsymbol{L}']\mathbf{1}_n\hat{a} \end{aligned} \tag{3.6.21}$$

Also, considering the differences in the corresponding variances for the two estimators in Equations 3.6.19d and 3.6.17b, respectively. It follows that

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}},nb} = [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L} + \boldsymbol{\Sigma_e})^{-1}\boldsymbol{L}']\mathbf{1}\sigma^2_{\hat{a}}\mathbf{1}'[\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L} + \boldsymbol{\Sigma_e})^{-1}\boldsymbol{L}']' \tag{3.6.22}$$

That means, when considering binding with $\boldsymbol{\Sigma_e} = \mathbf{0}$, the benchmarked estimator could be expressed as

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{B}} = \hat{\boldsymbol{\eta}}_{nb,B} + [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']\mathbf{1}_n\hat{a} \tag{3.6.23}$$

with variance given by

$$\begin{aligned} \boldsymbol{\Sigma}^{\boldsymbol{c}}_{\hat{\boldsymbol{\eta}}\boldsymbol{B}} &= \boldsymbol{\Sigma}^{\boldsymbol{c}}_{\hat{\boldsymbol{\eta}},\boldsymbol{nb}} + Cov(\hat{\boldsymbol{\eta}}_{nb,B}, [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']\mathbf{1}_n\hat{a}) + Cov([\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']\mathbf{1}_n\hat{a}, \hat{\boldsymbol{\eta}}_{nb,B}) \\ &+ [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']\mathbf{1}_n\sigma^2_{\hat{a}}\mathbf{1}'_n[\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']' \\ =&\boldsymbol{\Sigma}^{\boldsymbol{c}}_{\hat{\boldsymbol{\eta}},\boldsymbol{nb}} + Cov(\hat{\boldsymbol{\eta}}_{nb,B}, \hat{a})\mathbf{1}'_n[\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']' + [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']\mathbf{1}_n Cov(\hat{a}, \hat{\boldsymbol{\eta}}_{nb,B}) \\ &+ [\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']\mathbf{1}_n\sigma^2_{\hat{a}}\mathbf{1}'_n[\boldsymbol{I} - \boldsymbol{\Sigma_\ell} \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{\Sigma_\ell}\boldsymbol{L})^{-1}\boldsymbol{L}']' \end{aligned}$$

where $Cov(\hat{\boldsymbol{\eta}}_{nb,B}, \hat{a})$ refers to the cross-covariance between two vectors with dimensions $k \times 1$ and $l \times 1$ according to the definition given by Bickel and Doksum (2001, page 504, Equation B.5.8) and specific values of $k = n$ and $l = 1$.

Now since,

$$
\begin{aligned}
&Cov(\hat{\eta}_{nb,B}, \hat{a}) = Cov(\hat{a}, \hat{\eta}_{nb,B})' \\
&= Cov(y + \Sigma_\ell L[L'\Sigma_\ell L + \Sigma_e]^{-1}(x - L'y), -\sigma^2_{\hat{a}} 1' L[L'\Sigma_\ell L + \Sigma_e]^{-1}(x - L'y)) \\
&= -Cov(y, -L'y)[L'\Sigma_\ell L + \Sigma_e]^{-1} L' 1 \sigma^2_{\hat{a}} \\
&\quad - \Sigma_\ell L[L'\Sigma_\ell L + \Sigma_e]^{-1} Var(x - L'y)[L'\Sigma_\ell L + \Sigma_e]^{-1} L' 1 \sigma^2_{\hat{a}} \\
&= \Sigma_\ell L[L'\Sigma_\ell L + \Sigma_e]^{-1} L' 1 \sigma^2_{\hat{a}} - \Sigma_\ell L[L'\Sigma_\ell L + \Sigma_e]^{-1} L' 1 \sigma^2_{\hat{a}} = 0_{n \times 1}
\end{aligned}
$$

it follows that

$$
\Sigma^c_{\hat{\eta}B} = \Sigma^c_{\hat{\eta},nb} + [I - \Sigma_\ell L(L'\Sigma_\ell L)^{-1} L'] 1_n \sigma^2_{\hat{a}} 1'_n [I - \Sigma_\ell L(L'\Sigma_\ell L)^{-1} L']' \tag{3.6.24}
$$

Replacing Equation 3.6.17d in Equation 3.6.25, it follows that

$$
\begin{aligned}
\Sigma^c_{\hat{\eta}B} = \Sigma_{\hat{\eta},nb} + \underbrace{\Sigma_\ell L(L'\Sigma_\ell L)^{-1} \Sigma_e (\Sigma_e + L'\Sigma_\ell L)^{-1} \Sigma_e (L'\Sigma_\ell L)^{-1} L'\Sigma_\ell}_{\text{Binding Excess}} \\
+ \underbrace{[I - \Sigma_\ell L(L'\Sigma_\ell L)^{-1} L'] 1_n \sigma^2_{\hat{a}} 1'_n [I - \Sigma_\ell L(L'\Sigma_\ell L)^{-1} L']'}_{\text{Bias Estimation Excess}}
\end{aligned} \tag{3.6.25}
$$

with the two excess terms being positive definite matrices. This permits us to conclude that, also for the regression method, the binding estimator subject to survey errors is less efficient than the non-binding estimator.

3.7 Conclusions and Further Issues

State space models have been introduced in this chapter. The key advantage of these models is that it permits analysing time series in a structural form, that is the decomposition of the original estimates into different components as trends, seasonality and calendar variations Godolphin and Johnson (2003). This structural form covers a wide range of time series with ARIMA or stochastic volatility models as particular cases. Some important aspects in the implementation of these models were covered, specifically the initialization of the Kalman filter and maximum likelihood estimation of the hyperparameters.

Regarding the benchmarking problem, it is possible to formulate the problem in this context. Two alternative methods have been proposed by Durbin and Quenneville (1997) using state space models and they were presented using a more general formulation than the one presented in their paper. Durbin and Quenneville (1997) considered a very specific model to be applied to the Canadian Retail Trade series and this model has been extended here in a more general form in order to consider any kind of structural time series model. One advantage of the Durbin and Quenneville (1997) methods is that one of the methods is an online procedure which means it is no longer necessary to have the last annual benchmark available for producing high frequency estimates in the last year.

One criticism of the approach is that "the models are very complex compared to those used in the regression methods. Moreover, the smoothing part of the state space approach requires the storage of a large number of covariance matrices" (Dagum and Cholette, 2006, page 205). In our humble opinion, this is a problem that is not impossible to handle by using any of the modern software tools available nowadays.

Furthermore, a possible way to reduce the dimensionality of the vectors and matrices involved is by not including the observation errors in the state vector but under the cost of producing an autocorrelated measurement model. Pfeffermann and Tiller (2006) proposed a modification to the Kalman filter theory which is able to deal with this problem. They proposed a new filter that coincides with the traditional Kalman filter when the disturbances in the observation equation are uncorrelated. Another possibility in order to reduce the dimensionality is to drop some of the higher frequencies in the seasonal term. This is because seasonal patterns change relatively smoothly over the year. Abraham and Box (1978) and Anderson (1971, page 106) give an example of a model using only one and two frequencies respectively (instead of six in the complete model). The first frequency, which corresponds to a period of twelve months, is known as the *fundamental frequency* while the remaining frequencies are *harmonics.* (In the Anderson example, the second frequency correspond to a period of six months).

We have also considered two cases of estimators for benchmarking. When the adjusted series agrees exactly with the benchmarks, the benchmarking has been called *binding*. This type of estimation is considered even in cases when the annual benchmarks have been obtained from a survey rather than a complete enumeration or census. In this case, with the benchmarks being subject to annual survey errors, it is necessary to account for the variability of the benchmarks when computing the variances of the monthly benchmarked estimators. In Section 3.6.1., the correct variance of the binding estimators when using temporal annual benchmarks that are subject to survey error has been developed. Additionally, some conditions under which the variance of binding estimators could be higher than the variance of the obtained estimators after signal extraction without the benchmarks have been established. In this latter case, there is no reason for binding apart from consistency in publication. The theory has been presented for the two-step method (Durbin and Quenneville, 1997) and also extended to the ARIMA and Regression approaches outlined in Chapter 2 with similar conclusions than those obtained for the benchmarking method using state space models.

Chapter 4

Benchmarking Methods Applied to Business Surveys in the UK

In this chapter, the benchmarking theory will be applied to real data obtained from business surveys in the UK. We will describe the data sources used. Smith, Pont and Jones (2003) give an overview of the business surveys carried out in the UK. Approximately 100 different business surveys are carried out by the ONS and most of them adopt different methodologies although some coordination among them has been proposed. The main characteristics (sampling frame, sampling design, parameters and estimators) of the two more important surveys will be explained in the next sections. This corresponds to the MPI (Monthly Production Inquiry) and the ABI (Annual Business Inquiry). Finally, the advantages of benchmarking these surveys are highlighted through the application in a particular industrial sector. The precision of the estimates of benchmarking methods under the state space models approach is studied using the theoretical developments in the last chapters.

4.1 Preliminaries

The concept of repeated surveys was introduced in Chapter 1. Repeated surveys differ from longitudinal surveys in that the sampling units need not be the same over time. Instead, there could be any degree of overlap between the units in two adjacent

periods. The idea of this partial overlap is to reduce respondent burden, especially in smaller businesses and on the other hand to get better estimates of change that take into account the correlations that are present (see Wolter (1979) for a more detailed explanation) .

Business surveys are a particular case of repeated surveys with different possible frequencies: monthly, quarterly or annual surveys. These surveys are designed with the aim of producing estimates of totals, averages, ratios and change between two periods in measures of economic activity (Srinath, 1987; Hidiroglou and Srinath, 1993). Additionally, results for business surveys are used to construct other official statistics such as national accounts. National accounts show the major transactions occurring during an economic period. The main user of this information is the government, which uses the results to measure how the nation is performing, to propose new policies and to evaluate and control the implementation of them (Lewington, 1995). Business surveys receive a special consideration in the literature as their design and application are different from social surveys. Cox and Chinnappa (1995) distinguish between social and economic statistics, where "social" refers to "people and their activities as individuals" and "economic" refers to "organizational entities" and their economic activities. The information is generally collected through "household surveys" and "business surveys" respectively.

Riviere (2002) and Smith et al. (2003) highlight the main conceptual differences between household and business surveys. In particular, the obligation to respond (in the case of business surveys); the necessity of interviewers and multi-stage sampling techniques (in the case of households); among other differences. However, the main difference is perhaps the heterogeneity of businesses. Size variables such as turnover or number of employees normally have highly skewed distributions. This is because, there is a small number of large businesses with a huge contribution to the economy (according to the ONS, in 2000, there were around 9000 businesses with more than 250 employees in the IDBR covering a total of 14 million employees) but there are also a very large number of small businesses (there were around 1655000 businesses with less than 20 employees covering 5.5 million of employees in the same period). An-

other difference valid for business surveys is the availability of alternative data sources (administrative records or marketing data) which can be used to validate survey estimates or for imputation (Cox and Chinnappa, 1995) and for the sampling design and estimation.

Two of the most important business surveys carried out by the ONS in the UK are:

1. *Monthly Production Inquiry (MPI).* The MPI is a survey covering the manufacturing industry all over the UK. It is the main source of the monthly Index of Production (IoP), it is also used to estimate the change in the total number of employees and contributes to the "income measure" and the "output measure" of the Gross Domestic Product (GDP). Since 1948, the survey collects monthly information from manufacturing industries about turnover and since 1996 about employment variables. Currently it covers around 9300 out of a total around 160000 businesses monthly (National Statistics, 2005*b*)

2. *Annual Business Inquiry (ABI).* This is the main annual survey of businesses, covering the same MPI variables (turnover and employment) but also covering purchases, inventories and capital expenditure. The survey covers around 75000 businesses since 1998 (Jones, 2000; Partington, 2001; National Statistics, 2004).

The main characteristics of these two surveys (sampling frame, sampling design, parameters and estimators) will be presented in the next sections and a summary of the key facts is presented in the Appendix C.

4.2 Sampling Frame

Konschnik, Monsour and Detlefsen (1985) and Hidiroglou and Srinath (1993) point out that maintaining a frame of businesses is complicated due to the rapid rate of change of the frame. Hidiroglou and Srinath (1993) state that "mergers, acquisitions, changes in ownership, reorganizations, and so forth, require setting up rules for handling changes".

The sampling frames in business surveys are mainly list frames although it is not unusual to combine area and list frames. In particular, in the UK, the Inter-Departmental Business Register (IDBR) is a list established in 1995 consisting about 1.8 million businesses undertaking activities in diverse sectors such as agriculture, mining, catering, transport, banking, public administration, among others (Perry, 1995).

The sampling frame is made up of businesses called *units*. Having in mind the construction of the sampling frame some definitional questions arise (e.g. what is a business?). Following the definitions in Smith et al. (2003), the basic statistical unit on the frame of businesses is the *enterprise* (or business) defined as: "the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit" (Smith et al., 2003; European Legislation Council, 1993, Section III-A).

In the same context, when regional data is required, the *local unit* is defined as "an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place. At or from this place economic activity is carried out for which - save for certain exceptions - one or more persons work (even if only part-time) for one and the same enterprise" (Smith et al., 2003; European Legislation Council, 1993, Section III-B).

In order to classify businesses according to their industrial activity and according to the goods and services they produce, the standard industrial classification (SIC03) is used. This classification corresponds to a hierarchical five digit code system, useful to determine the corresponding strata and is relevant in the processes of editing, imputation and estimation (National Statistics, 2003). The 17 main sections in the SIC92 are denoted by a single capital letter from A to Q and they are summarized in Appendix D. Some sections are, in turn, divided into subsections (each denoted by the addition of a second letter). The letters of the sections or subsections can be uniquely defined by the next breakdown, the divisions (denoted by two digits). There are 17 sections,

16 subsections, 62 divisions, 225 groups, 517 classes and 285 subclasses. A new industrial classification (SIC2007) is planned to be adopted in January 2008 (National Statistics, 2006).

The IDBR includes information about the SIC classification, location and names of the businesses and the register is maintained and actualized by the ONS. The IDBR is updated using information from business surveys, registers of VAT (value-added tax) and the Department of Employment. Normally there are some delays in registering births and deaths of businesses in the IDBR. Hedlin, Pont and Fenton (2001) examined lags in recording births and deaths of businesses on the IDBR. The average lag in recording births was shorter than recording deaths implying over-coverage in the frame. Smith et al. (2003) state that these enterprises opening and closing are smaller than the average and thus, the impact over the economy is small.

There are two variables commonly collected from MPI and ABI (Smith et al., 2003):

- Turnover: Defined as the amount receivable by the business for services provided or goods sold during the period covered by the form and
- Employment: Measured as the total number of employees at a certain date.

4.3 Sampling Design

The sampling design for business surveys is usually the same in all the statistical agencies around the world: simple random stratified (srs) sampling and sometimes sampling by probabilities proportional to size (pps). Multistage samples are common for household surveys but not for business surveys (Riviere, 2002). In the UK, systematic samples were used and also pps relative to the size of the enterprise; but these techniques were abandoned due to the difficulties in implementing a new permanent random number system for the rotation of the samples (Smith et al., 2003). The two surveys considered in this chapter (MPI and ABI) use a stratified srs sampling but they

Size Bands	Employment Range	Sampling Fraction
1	0-9 employees	1%
2	10-49 employees	5%
3	50-149 employees	19%
4/7	150 or more employees (50 if a Band 7)	100%

Table 4.1. Employment Size Bands for Stratification in MPI. Source: ONS data

still have some small differences in the stratification and the rotation of the samples.

It is well known in the sampling literature that in order to estimate change most efficiently from one period to another, it is better to retain the same sample throughout all occasions (Cochran, 1977). For estimates of both level and change, partial overlap is optimal and this optimum matching proportion depends on the value of the correlation between one occasion and the previous one (Finkner and Nisselson, 1978). Additional considerations enter into the decision of what matching proportion should be used. For instance, due to respondent burden, it may be advisable to replace units in the sample more frequently (Hidiroglou and Srinath, 1993). Nevertheless, the cost of introducing new units into the sample makes it more attractive to replace units less frequently (Finkner and Nisselson, 1978). In the next subsections, sampling designs and rotation will be explained in more detailed for MPI and ABI respectively.

4.3.1 Sampling Design for MPI

The MPI uses simple random stratified sampling (National Statistics, 2005*b*). The population is stratified by the 4 digit-SIC code and employment bands within industry. In the MPI, the number of employment size bands (strata) is three or four depending on the total number of businesses in the respective industry. Table 4.1 shows the size bands used in MPI with their respective sampling fractions.

Some industries with small populations are chosen (around 40 of the 4 digit industries) to be stratified into only three strata with all businesses with more than 50 employees

being sampled. In this case, the strata are coded as "1-2-7". In the other cases, the stratification follows the structure in Table 4.1 with the natural codification "1-2-3-4". The sampling fractions were chosen with consideration that if an industry is dominated by few very large companies, they are included in each period. This corresponds to the forced inclusion band 4/7.

A rotation scheme is done for strata 1, 2 and 3 (1 and 2 only if the population is small). The rotation is done in the following way: businesses in the sample from strata 2 and 3 are selected for 27 consecutive months minimum whereas businesses in stratum 1 are selected for a total of 15 consecutive months. Samples are generated from the IDBR using a system of Permanent Random Numbers (PRN, Ohlsson (1995)). Firstly, each unit on the sampling frame is assigned a random number between 0 and 1. These PRN create a new order in the sampling frame (ordering the units by PRN). If the desired size of the sample is n_s, the sample is selected choosing a random starting point, noted by κ, and selecting the first n_s elements with PRN higher than κ. This is the sample for the first period. For the following periods, the rotation of successive samples is forced by moving the starting point in a way that overlaps in the required proportion (for more details, see Ohlsson (1995)). Notice that these numbers are "permanent" as there is not a new assignation of random numbers for the next periods. However, if a new business appears in the frame a new PRN has to be chosen for this unit.

The sample is allocated to strata using the Neyman optimum allocation (Neyman, 1934). Under Neyman's formula, the allocation minimizes the variance of total turnover over all the strata. The sample is allocated using as weights the product of the stratum size and the stratum standard deviation in each stratum.

4.3.2 Sampling Design for ABI

Jones (2000, Annex, pages 55-57) and Partington (2001, Technical note, pages 5-7) describe the sampling design for ABI. The ABI questionnaire is divided into two forms: ABI/1 is used to collect employment data and ABI/2 is used to collect accounting data.

The reason why this division is done is because the time of collection of the data is different. Employment data is available from businesses at the end of march while accounting data is usually available around six months later.

ABI/2 is a sub-sample of ABI/1 with some industry sectors not covered by ABI/2 (Jones, 2000). The sample sizes are around 78500 businesses for ABI/1 and 75000 for ABI/2. A stratified srs sample is selected from ABI/1 and the ABI/2 sample is obtained by excluding the sectors that are not covered. The population is stratified by six employment size bands (1-9, 10-19, 20-49, 50-99, 100-249, 250+); three regions (England and Wales combined, Scotland and Northern Ireland) and an hybrid 2/3/4 digit SIC depending on the region. For instance, for Northern Ireland a 2 digit SIC is used whereas for England and Wales a 4-digit SIC is used.

In a similar way to the MPI, a Neyman optimum allocation is used, with the result that in some cases some strata of under 250 employees will be also completely enumerated. The rotation scheme is as follows (Jones, 2000): businesses in the first stratum of small businesses are completely replaced every year; businesses with number of employees between 10 and under the cutting point for forced inclusion have a rotation rate of 50 per cent (i.e. half are replaced each year). Other issues as post-stratification, estimation, scaling and synthetic estimation are discussed in Partington (2001).

4.4 Parameters and Estimators

Estimation of totals corresponding to the variables turnover and employment is usually carried out at the ONS using ratio estimators (Cochran, 1977; Särndal et al., 1992, section 7.3). Generally, turnover is highly correlated with employment size and the MPI uses Pay as You Earn's (PAYE) employment as an auxiliary variable for the estimation. PAYE is an administrative source from a tax deduction system, introduced in the UK in 1944, which takes a certain amount of money from an employee's income when paid by the employer. Having in mind the stratified srs sampling design in both surveys; a separate ratio estimator (Cochran (1977, page 164), Särndal et al. (1992, page 270)) is

used to estimate employment and turnover (Smith et al., 2003). This is a special case of a ratio estimator given by

$$\hat{t}_{y,R} = \sum_h \hat{t}_{y,h,R} = \sum_h \frac{\sum_{k \in S_h} y_k}{\sum_{k \in S_h} x_k} \cdot t_{x,h} \tag{4.4.1}$$

where y_k denotes the value of the variable of interest for sample unit k and x_k the corresponding value of the deducted tax. S_h is the selected sample in stratum h and $t_{x,h}$ is the population total of the auxiliary variable (PAYE) in the stratum $h = 1, \cdots, H$. According to Smith et al. (2003) such estimates are calibrated to the stratum totals of the auxiliary variable and the calibration takes place by stratum or groups of strata (combined ratio estimation). This last situation is what actually happens most of the time in the ABI where the size bands are combined if the sample size is too small.

Another very important parameter to estimate is the change from one period to the next one. This is useful to detect turning-points in time. Business surveys in the UK use the method of matched pairs to measure levels in a particular month using the estimated level in the month before as auxiliary information by using units that are common in consecutive months. Then, an estimate of change is produced by taking differences of the estimated levels.

The calculation of level estimates is done over those common units from one period to the next by the formula

$$\hat{t}_{y,MP,t} = \sum_h \frac{\sum_{k \in S_h^*} y_{k,t}}{\sum_{k \in S_h^*} y_{k,t-1}} \cdot \hat{t}_{y,h,R,t-1} \tag{4.4.2}$$

where $S_h^* = S_{h,t} \bigcap S_{h,t-1}$ and $S_{h,t}$ denotes the selected sample in the stratum h at the instant t. It can be noticed that Equation 4.4.2 is analogous to the Equation 4.4.1. Using the estimated total in the last period as the auxiliary variable, Equation 4.4.1 can be written analogously as follows

$$\hat{t}_{y,R,t} = \sum_h \hat{t}_{y,h,R,t} = \sum_h \frac{\sum_{k \in S_h^*} y_{k,t}}{\sum_{k \in S_h^*} y_{k,t-1}} \cdot t_{y,h,t-1} \tag{4.4.3}$$

Note that because the internal sums are over S_h^*, both the numerator and the denominator inside the sum in Equation 4.4.3 are available. However, because the term $t_{y,h,t-1}$ is unknown, it has to be approximated as in Equation 4.4.2. Once a level estimate is obtained, differences between successive levels are calculated producing an estimate of change. Smith et al. (2003, section 3.7, page 269) points out that measuring successive levels and taking differences is not a good alternative as survey errors tend to distort the actual difference.

4.5 Reasons for Benchmarking

Apart from the natural advantage of combining information obtained from a monthly survey with more precise annual information for the same variables; other reasons arising from the estimation procedures using benchmarking methods for business surveys in the UK are as follows.

According to results of Kokic and Jones (1998) and Smith et al. (2003), matched pairs is a more reliable measure of change in business surveys. However, because the method only considers common responses from one period to the next, it cannot take into account births and deaths of businesses. Kokic and Jones (1998) and Smith et al. (2003) explain the effect of benchmarking on the quality of matched pairs estimates. Without benchmarking, the advantages of the matched pairs method are weakened. Using simulated data, Kokic and Jones (1998) showed that the variances of the matched pairs estimates without benchmarking increase over time and affect the variance of the estimate of change (Smith et al., 2003, page 270). In fact, the quality of the estimation gets worse "as the distance from the starting benchmark increases". Smith et al. (2003), page 270, stated that "because of the delay in introducing any benchmark data, the advantages of the matched pairs method cannot be realized in real time". The next section will present an application of state space benchmarking methods, which will permit to overcome the problems that have been referred to in this section for the main business surveys in the UK.

4.6 Benchmarking MPI to ABI

The main aim of benchmarking is to combine the monthly information (estimates and their respective standard errors) with auxiliary annual information in order to get better monthly estimates which will add up to a new improved annual estimate. In this section we will present an application of benchmarking methods using state space models applied to business surveys in the UK. Data of turnover in the period of January 1998 to December 2003 is obtained from two different sources: the Monthly Production Inquiry (MPI) and the Annual Business Inquiry (ABI) survey in the Office for National Statistics (ONS). This information was available from the ONS for 215 industrial sub-sectors in the manufacturing industry in the UK. The application will permit to study some practical issues about the specification of structural time series models, the initialization of the Kalman filter and the maximum likelihood estimation of the hyper-parameters. The goodness of fit of the various models to be considered will be evaluated according to sample diagnostic tests and graphical displays of the innovations, by checking for outliers and structural breaks based on an analysis of the auxiliary residuals (as they were defined in Section 3.4.4.) and also through the use of Monte Carlo experiments.

The application requires having information in both surveys for the same period of observation. 95 sub-sectors, out of the 215, were discarded as they had information available from one survey but not the other (monthly but no annual or the inverse situation), or because it was not possible to get any standard error information from the surveys. In the remaining group of 120 sub-sectors, 73 have got standard errors in a higher level (e.g. sector level) but not in the sub-sector level. In total, only 45 sub-sectors have all the necessary information available (95+73+45=213). The remaining two sub-sectors (with SIC codes 37.1 and 37.2) are a special case conforming to the high level sector 37. For these sub-sectors, the information corresponding to the monthly estimates and the monthly standard errors is available only at the sector level. The problem of how to get disaggregated estimates at the sub-sector level corresponds to the contemporaneous disaggregation problem studied in the next chapter.

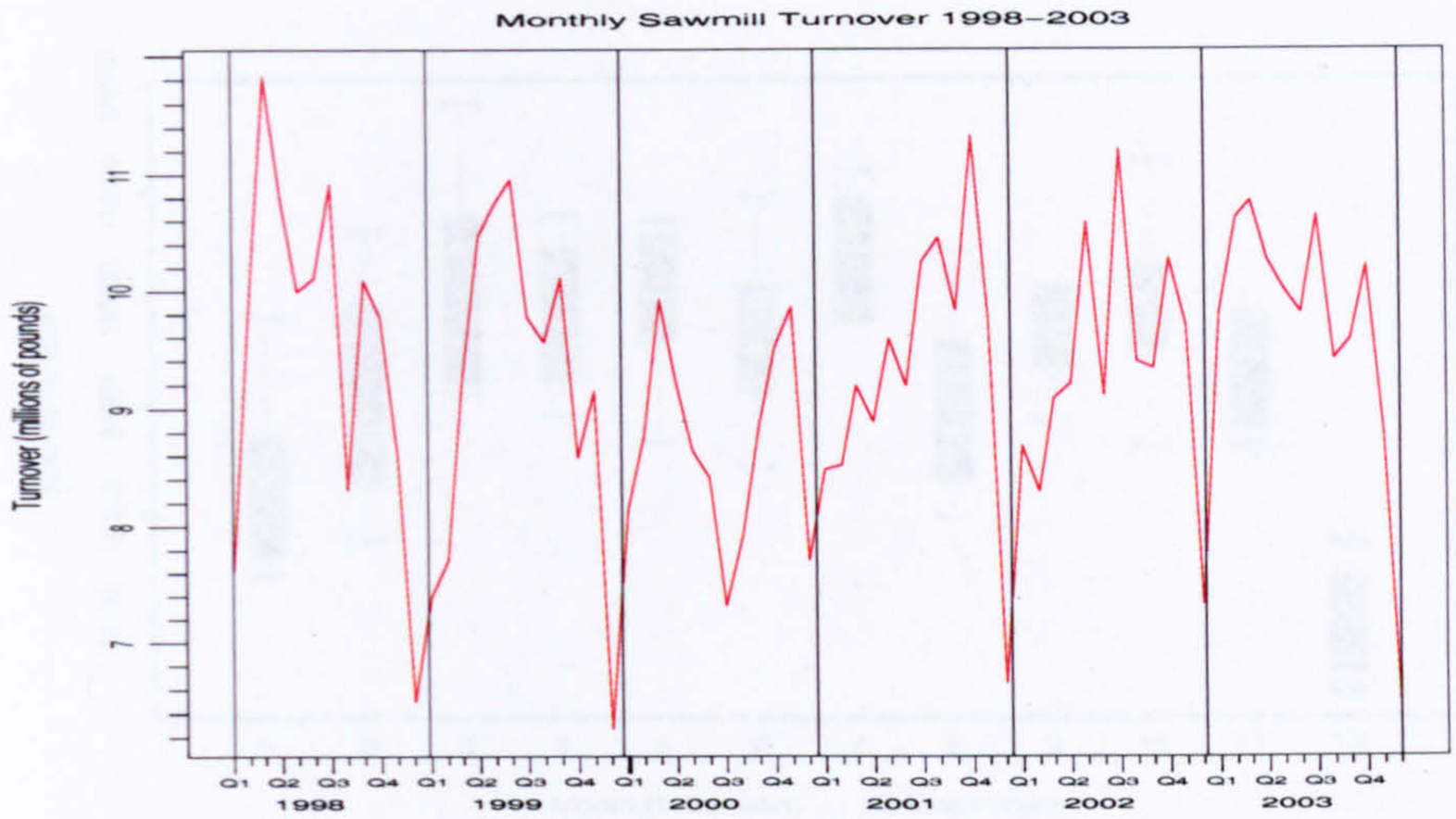


Figure 4.1. **Monthly Estimates for Turnover of Sawmills**

The industrial sector of wood manufacturing was selected for illustration. The plot in Figure 4.1 shows the monthly values of turnover obtained from MPI, during the period of reference, for the industrial sub-sector of sawmills. This corresponds to the SIC code 20.1 in the subsection DD - Manufacture of wood and wood products. The length of the monthly series is $n = 72$ months whereas the length of the annual series is $m = 6$ years.

It can be noticed in Figure 4.1 that the monthly series is seasonal and there is a significant drop of the turnover values for this industrial sub-sector at the end of each year. This can be confirmed in the plot in the Figure 4.2, where the values were grouped by month and values corresponding to the group "December" (month 12) show lower values than the other months. Notice also the presence of an outlier in the value of July (month 7), corresponding to the big drop in July 2000. In Figure 4.1 the value of Turnover for December 2000 is even higher than the value in July 2000, making one ask if there were untypical external influences in the year 2000. This will be reflected in the benchmarking models used for this application and presented later on.

Using the Scott and Smith (1974) decomposition in Equation 1.1.1, $y_t = \eta_t + \ell_t$ with η_t the value of the unobserved population true series and ℓ_t the sampling error associated with y_t, the survey estimate of η_t at time t. Considering the series in this application,

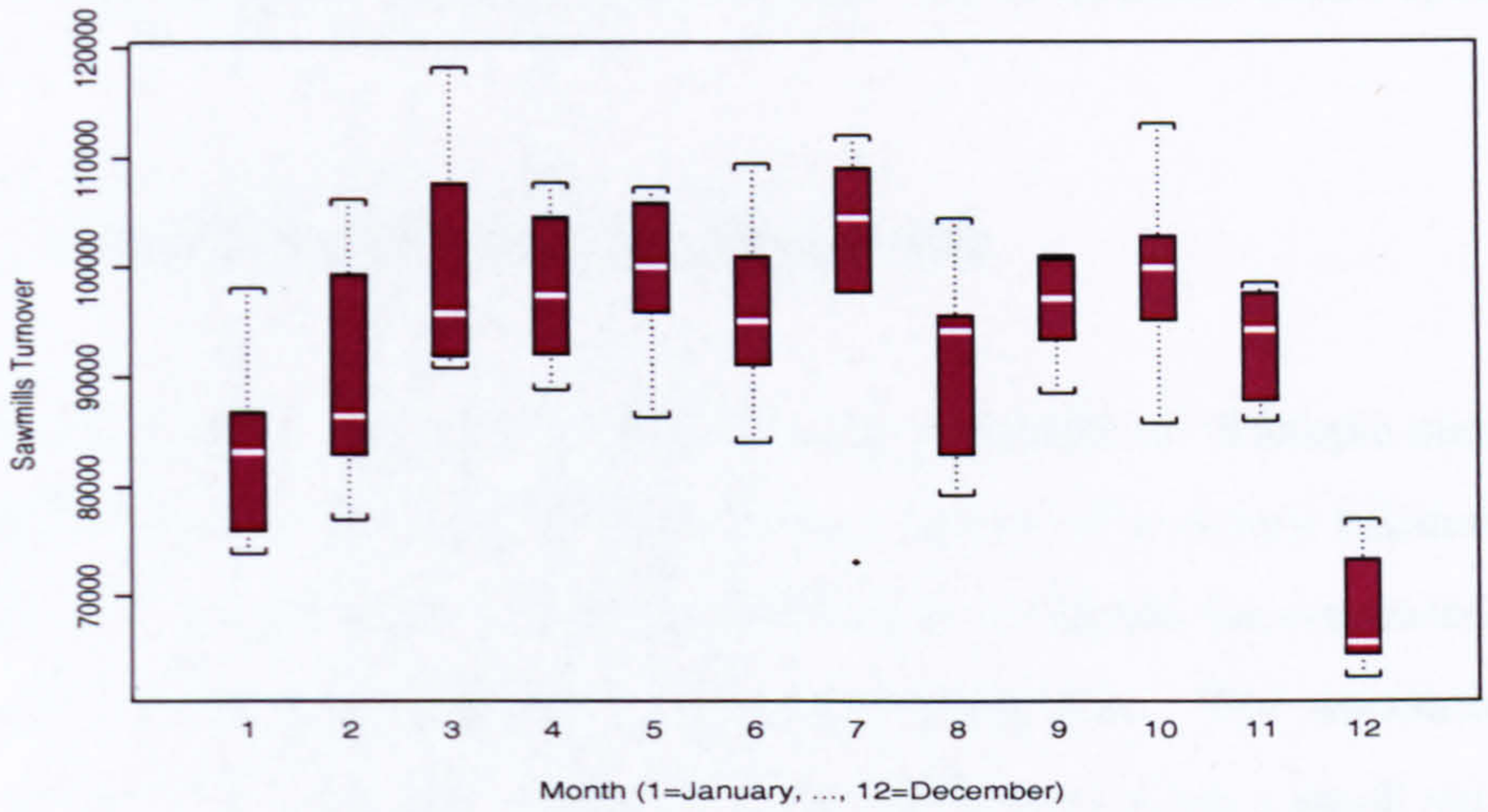


Figure 4.2. **Monthly Estimates for Turnover of Sawmills Grouped by Month**

the monthly series of Turnover for sawmills, y_t can be modeled using the signal plus noise model, $y_t = \underbrace{\mu_t + \gamma_t + \epsilon_t}_{\eta_t} + \underbrace{k_t \ell_t^*}_{\ell_t}$. In this model, the monthly signal series η_t has been decomposed into a trend, a seasonal, an irregular component and the survey error ℓ_t. In particular, ℓ_t has been expressed as $\ell_t = k_t \ell_t^*$ following the Cholette and Dagum (1994) formula in Equation 2.3.8.

The term k_t in the survey error component is the standard deviation of the survey errors ($k_t = \text{s.d.}(y_t) = \text{s.d.}(\eta_t + \ell_t) = \text{s.d.}(\ell_t)$) assuming the η_t's are constant for each instant. The value of k_t is taken from the standard error information obtained from the MPI survey in the ONS and is incorporated into the model to represent heteroscedasticity in the survey errors.

Although estimates of turnover for the sawmills sub-sector are available with their respective annual standard errors (and coefficients of variation) for ABI from 1998 to 2003, the monthly standard errors are available only in the period from January 2002 to December 2003. Before the application of the benchmarking methods in section 2.5, it is necessary therefore to estimate these standard error values in order to allow heteroscedasticity in the survey errors. A generalized variance functions (GVF) approach (Wolter, 1985; Johnson and King, 1987; Valliant, 1987) is proposed in the next section;

this is a methodology to estimate variances based upon regression model specifications.

4.6.1 Generalized Variance Functions

Wolter (1985, Chapter 5, page 201) discusses the possibility of "a simple mathematical relationship between the variance or relative variance of a survey estimator to the expectation of the estimator". Then, a possibility is to regress the estimated variances of an estimator to the estimated values of the parameter. The estimation of this model can be done using data from surveys in the past or from a small subset of the elements in the current survey. Variance estimates can be obtained by evaluating the model at the survey estimates, rather than by direct computations. This method was called the method of *generalized variance functions* (GVF) (Wolter, 1985; Johnson and King, 1987; Valliant, 1987).

Let $\hat{\theta}$ denote an estimator of the parameter θ and let $\theta = E(\hat{\theta})$ denote its expectation. GVF models are fitted to predict the relative variance (RV) of the estimator defined by

$$RV = Var(\hat{\theta})/\theta^2 = \sigma^2_{\hat{\theta}}/\theta^2 = CV^2(\hat{\theta}) \tag{4.6.1}$$

where $\sigma^2_{\hat{\theta}}$ denote the variance of $\hat{\theta}$ and $CV(\hat{\theta})$ its corresponding coefficient of variation.

Some alternative models have been proposed in the literature in order to achieve the best fit to relate RV to θ. Apart of the case when θ is a proportion, there is no rigorous theoretical justification for any of these models and so, optimum estimators of the model parameters are difficult to construct. Wolter (1985, page 206) states that "discussions of optimality would require an exact model and an exact statement of the error structure of the estimators $(\hat{RV})$ and $(\hat{\theta})$. In the absence of a completely specified model, we shall simply seek to achieve a good empirical fit to the data $(\hat{\theta}, \hat{RV})$". In the specific case, when $\hat{\theta}$ denotes an estimator of a total for a binary variable, some justification has been established. In this case, when $\hat{\theta}$ approaches N, the size of the population of study, the variance of $\hat{\theta}$ approaches zero and then, the relative variance RV is a decreasing function of the magnitude of the expectation θ. Wolter (1985),

pages 203-205 presents some justification in this specific case in terms of the concept of design effects, the application of clustered simple random sampling and considers also the specific case of the estimation of proportions. Valliant (1987) studies the application of GVF for estimators of totals that are linear combinations of sample cluster means from stratified two-stage cluster samples.

Considering the case of repeated surveys, a subindex t will be added to represent the values of the estimator of the signal at different points of time

$$RV_t = Var(\hat{\eta}_t)/\eta_t^2 = \sigma_{\hat{\theta}_t}^2/\eta_t^2 = \sigma_{y_t}^2/\eta_t^2 = CV^2(y_t) \tag{4.6.2}$$

Some alternative models relating RV_t to η_t are given by the expressions:

$$RV_t = \beta_0 + \beta_1/\eta_t \tag{4.6.3}$$

$$RV_t = \beta_0 + \beta_1/\eta_t + \beta_2/\eta_t^2 \tag{4.6.4}$$

$$RV_t = (\beta_0 + \beta_1\eta_t)^{-1} \tag{4.6.5}$$

$$RV_t = (\beta_0 + \beta_1\eta_t + \beta_2\eta_t^2)^{-1} \tag{4.6.6}$$

$$log(RV_t) = \beta_0 + \beta_1 log(\eta_t) \tag{4.6.7}$$

The parameters β_0, β_1 and β_2 are unknown and need to be estimated from all the available pair of values $(y_t,\ \hat{RV}_t)$. After that, variance estimates can be obtained by evaluating the model at the survey estimates as was mentioned above. This is the methodology used by the Current Population Survey (CPS) at the US Census Bureau under the assumption that the sample design has not changed during the period of reference (Bureau of Labor Statistics, 2002). CPS uses generalized variances for estimates of month-to-month changes as well as for estimates of monthly levels since 1947 (Hansen, Hurwitz and Madow, 1953; Bureau of Labor Statistics, 2002). In a different context, Johnson and King (1987) applied GVF in a US survey of reading ability among young adults using a multistage stratified probability sample.

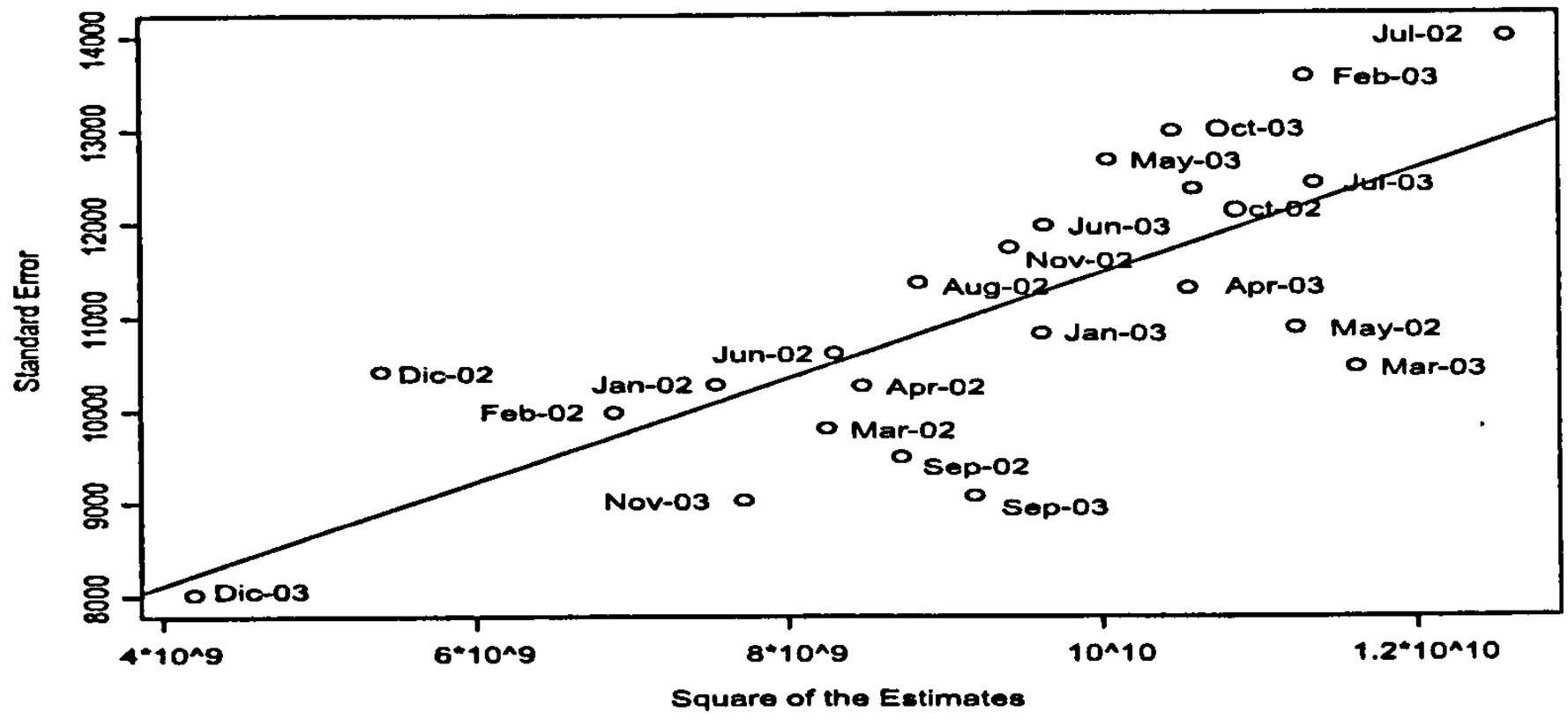


Figure 4.3. Scatterplot Standard Error vs. Square of the Estimates. Regression Line: Final Model.

MPI Example

In this application, data on turnover from the manufacturing sectors in MPI are available from January 1998 to December 2003. However, monthly standard error information is only available for the last two years (2002 and 2003) at the ONS. The sub-sector 20.1 corresponding to the manufacturing industries of wood (sawing and milling) will be considered as an example for the application of GVF to estimate the missing standard error information.

A set of models, including Models 4.6.3-4.6.7, were fitted to the 24 available pairs of values $(\hat{\eta}, \hat{RV})$ from January 2002 to December 2003. However, none of them provided a good fit when considering their corresponding residuals. Therefore, models for standard errors were studied instead. The model with best fitting to the data (s.e. stands for "standard error") was

$$k_t = s.e.(y_t) = \beta_0 + \beta_1 y_t + \beta_2 y_t^2 \qquad (4.6.8)$$

The scatterplot appearing in Figure 4.3 shows a moderate positive association between the square of the estimates and their standard deviation with a correlation of 0.74. One outlier (August 2003) was detected and it was not considered for the fitting. All the

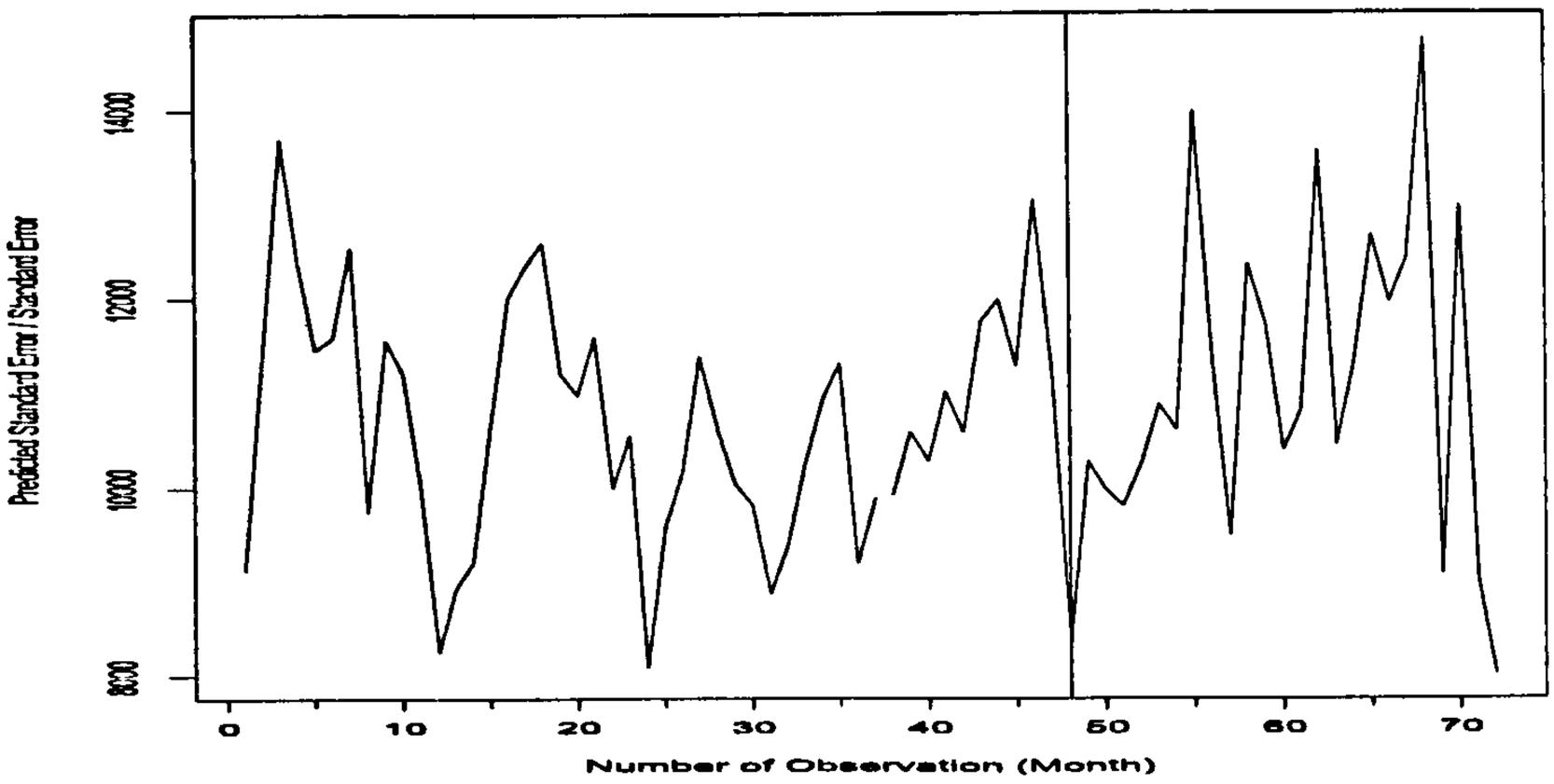


Figure 4.4. Series of Predicted Standard Errors. Last 24 Observations are Original Values from MPI.

detailed information about the regression modeling is presented in Appendix E.

The model used to predict the standard errors before 2003 after considering only the significant terms is given by the equation

$$k_t = s.e.(y_t) = 5878.68 + (5.59 \times 10^{-7})y_t^2 + \epsilon_t \qquad (4.6.9)$$

taking one outlier out of the sample. The R^2 value for the linear model was 0.55. Examination of significance tests and significance of the parameters gives no reason to question the adequacy of the model apart from the presence of some influential points as presented in the diagnostic plots in Figures E.7 - E.10 and test diagnostics in Table E.6 in Appendix E. The first plot corresponds to a histogram of the studentized residuals, then the qqplot shows the standardized residuals against the quantiles of the appropriate t distribution, with the plot being approximately linear. The third plot of studentized residuals against fitted values does not suggest any non-linear relationships, non-constant variances or outliers. Also, a plot of the autocorrelation function of the residuals does not show significant serial autocorrelation among them. Same conclusions are obtained after examining Table E.6 under the column "Model(-1obs)".

Standard errors for the months before January 2002 were obtained by calculating the predicted values according to Equation 4.6.9 and they are plotted in Figure 4.4 with

the original estimated values from the survey from January 2002 to December 2003. However they originate from different models as the part before January 2002 does not include the modeling error. In other words, the values of the standard errors before January 2002 obey the model $\hat{k}_t$=$s\hat{e}.(y_t) = 5878.68 + (5.59 \times 10^{-7})y_t^2$ and after January 2002 the model $\hat{k}_t$=$s.e.(y_t) = 5878.68 + (5.59 \times 10^{-7})y_t^2 + \epsilon$.

In order to have compatible sets of values, the mean and the variance of the residuals were calculated and then 48 values from a normal distribution with this mean and variance were randomly chosen and assigned for each one of the values from 1998 - 2003. The final set of standard errors of the estimates to be used for benchmarking are plotted in Figure 4.5. It should be noticed that the standard error value for the observation which was an outlier was also recalculated according to the final estimated model.

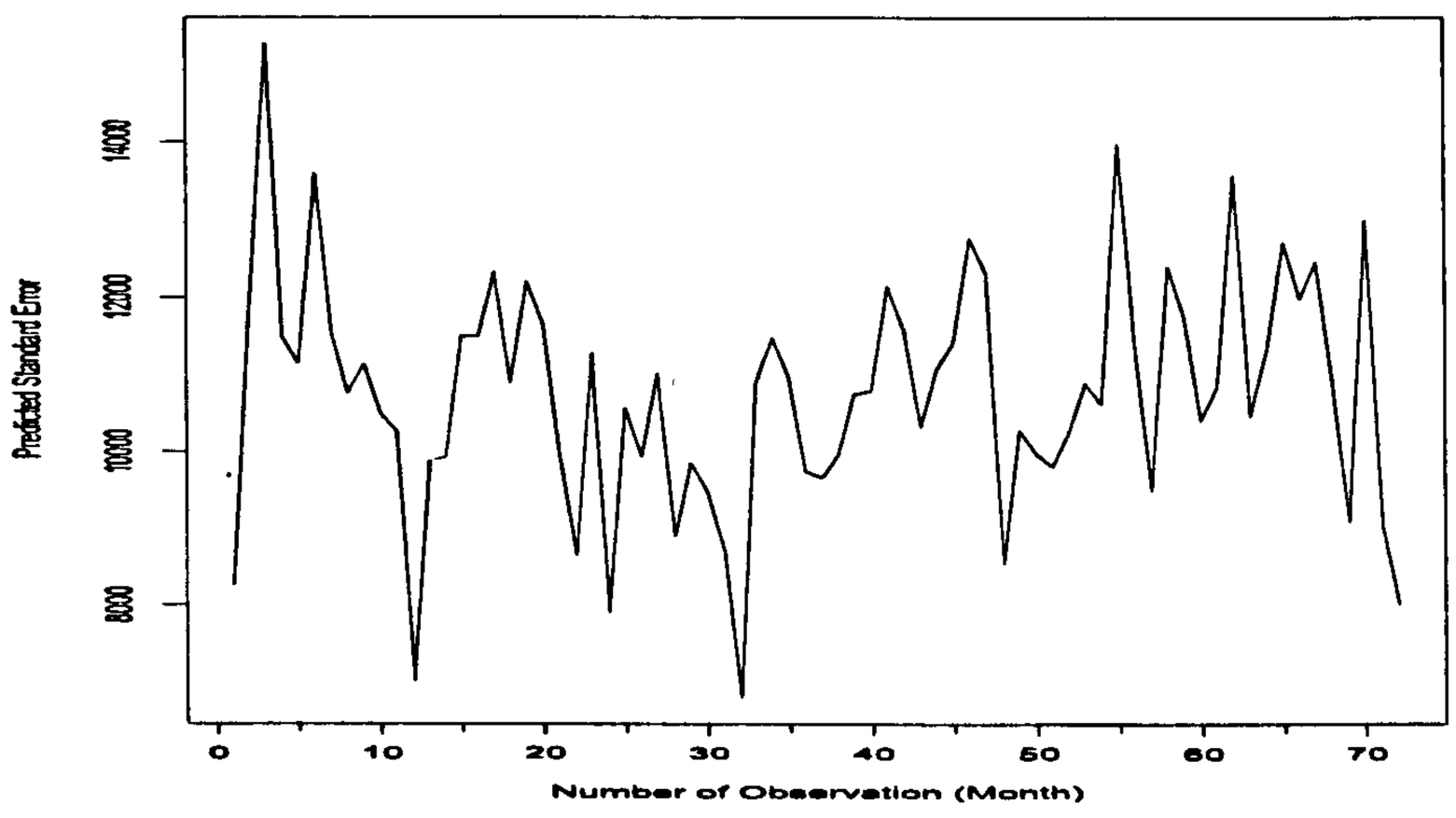


Figure 4.5. Series of Predicted Standard Errors. First 48 Observations Were Obtained Plus a Random Error

4.6.2 State Space Modelling of MPI

We will consider the same data set in the last subsection corresponding to the Turnover of sawmills in the UK for the period January 1998 to December 2003. Once all the stan-

dard errors are available; all the required information (annual and monthly estimates and the annual and monthly standard errors) is available and ready for benchmarking. Under the state space model approach, both the two step and the single step method, introduced in section 3.5, will require a suitable structural model for the monthly observations. The aim is to get a good model in terms of fitting, with small dimensions in the vectors and matrices used in the state space form.

In this section, a brief review of the modeling for the signal and the sampling error is presented. Firstly, a preliminary model considering a Basic Structural Linear Model (BSM) (pages 26 and 31) for the monthly signal and an autoregressive model for the survey error is implemented. The BSM will permit to formulate the monthly series in terms of unobserved components such as trend, seasonality and irregular terms. It is possible to consider this model as an extension of a multiple regression approach, letting the regression parameters to change stochastically over time (Harvey, 1990, page 31). The corresponding results and fitting tests under this basic model are shown. Then, different modifications were applied to the model trying to get a better fit to the data. The goodness of fit is evaluated in terms of the standardized innovations of the model and through Monte Carlo experiments as it will be explained in the next subsection. Finally, the results for a model with a good fit are presented.

BSM Signal plus Noise Model - Application

According to Equation 1.1.1, the observed turnover for the sawmill industries in the UK, y_t, can be represented as the sum of two independent processes; the true population or signal η_t and the sampling error or noise ℓ_t using Equation 1.1.1. The signal η_t will be represented as a structural time series model in order to obtain a decomposition in terms of a trend, a seasonal and an irregular component. Additionally, we consider a multiplicative form (Bell and Hillmer, 1990) of the survey error by assuming $\ell_t = k_t \ell_t^*$ with ℓ_t^* reflecting the auto-covariance structure of an unit variance ARMA process and k_t representing the standard errors of the survey in order to represent the heteroscedastic structure of the sampling errors. Non-sampling errors were not considered in this

application.

Modeling an ARMA process for the standardized survey error ℓ_t^* will require to get information about the auto-covariance matrix structure of the survey errors. A first approach is to estimate this structure directly from the survey microdata using the sample design information (called *primary analysis* in page 26). In this particular case, this approach is not possible not only because the variance estimation would involve complex computations on huge microdata but also because of confidentiality reasons that prevent getting the required information at a microdata level. Another alternative, modeling the error directly from the aggregate data (*secondary analysis* in page 26) was not considered because as some authors discuss there is a fundamental identification problem (e.g. Bell and Hillmer (1987*a*, page 86)). We will proceed using a different approach as follows here. In order to get a model as simple as possible in terms of the dimension of the state vector, we consider an AR(1) process for ℓ_t^* first. The incorporation of this model will add an extra term in the state vector as explained in page 38 (in order to add an ARMA(p,q) model, the number of extra terms in the state vector is equal to $\max(p, q+1)$). If the fitting is not satisfactory in terms of the significance of the parameters and the diagnostic tests, we will proceed to increase both the size of the state vector and the number of hyper-parameters to be estimated in a sequential way. We will consider the sequence of models: AR(1) (an extra term in the state vector, one additional hyper-parameter); then MA(1) (two extra terms in the state vector, one hyper-parameter), AR(2), ARMA(1,1) (two extra terms and two hyper-parameters), ARMA(2,1) and so on. We will stop this search once we get a good fitting to the data after considering different tests of misspecification such as diagnostic plots and traditional tests of autocorrelation, normality and heteroscedasticity for the standardized innovations. This search is extended not only to the most appropriate ARMA model for the survey errors, also for the remaining components in the structural time series model. For instance, in a different context, Durbin and Quenneville (1997) considered a benchmarking model for the Canadian retail trade sales using a differentiated trend, no slope, seasonal dummy variables and a seasonal SARMA(1,0) $\times(1,0)_{12}$ survey error.

We will first consider a BSM signal plus AR(1) standardized survey errors model for the turnover of sawmills industries. The assumption considering an AR(1) model for the noise is a standard assumption followed by authors such as Blight and Scott (1973) and Pfeffermann (1991). It implies that the autocorrelations of the sampling error decay geometrically as time passes. This is also the model assumed by the software BENCH in Statistics Canada under the Cholette and Dagum (1994) method. The BSM model was introduced for quarterly data in Equation 3.2.10 and its state space form was presented in page 36. Setting a BSM for the signal and also considering an AR(1) model for the standardized survey errors; the following model for the monthly turnover of sawmills in the UK was considered

$$y_t = \eta_t + \ell_t = \underbrace{\mu_t + \gamma_t + \epsilon_t}_{\eta_t} + \underbrace{k_t \ell_t^*}_{\ell_t} \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

where each component is given by

$$\text{Trend}: \begin{cases} \mu_t = \mu_{t-1} + \beta_{t-1} + \xi_t; & \xi_t \sim N(0, \sigma_\xi^2) \\ \beta_t = \beta_{t-1} + \zeta_t; & N(0, \sigma_\zeta^2) \end{cases}$$

$$\text{Seasonality}: \begin{cases} \gamma_t = \sum_{v=1}^{6} \gamma_{vt} & \\ \gamma_{vt} = \gamma_{v,t-1} \cos \kappa_v + \gamma_{v,t-1}^* \sin \kappa_v + \omega_{vt}; & \omega_{vt} \sim N(0, \sigma_\omega^2) \\ \gamma_{vt}^* = -\gamma_{v,t-1} \sin \kappa_v + \gamma_{v,t-1}^* \cos \kappa_v + \omega_{vt}^*; & \omega_{vt}^* \sim N(0, \sigma_\omega^2) \end{cases} \tag{4.6.10}$$

with $t = 1, \cdots, n$, $\kappa_v = 2\pi v/12 = v\pi/6$, $v = 1, \cdots, 6$ and standardized survey errors ℓ_t^* given by $\ell_t^* = \phi \ell_{t-1}^* + \chi_t; \qquad \chi_t \sim N(0, 1 - \phi^2)$

The signal η_t is modeled as a structural time series model decomposed in the trend μ_t, the seasonal component γ_t and the irregular white noise component ϵ_t. The processes ξ_t and ζ_t correspond to uncorrelated white-noise terms with variances σ_ξ^2 and σ_ζ^2 respectively. Also the disturbances ω_{vt} and ω_{vt}^* are white noise uncorrelated processes with mean zero and common variance σ_ω^2. As part of the modeling, the variances of the disturbances σ_ξ^2, σ_ζ^2 and σ_ω^2 need to be estimated and then they will constitute the hyperparameters of the signal. Then, they will be estimated using their maximum likelihood estimators. A positive variance for a component implies a stochastic process, whereas a zero variance implies deterministic behavior. In particular, the trend com-

ponent corresponds to a local linear trend model (see page 32) with the trend shifted by σ_ξ^2 and its first difference shifted by σ_ζ^2. In the special case when $\beta_t = 0$, the trend corresponds to a simple random walk. Regarding the seasonal component, this corresponds to the sum of six trigonometric terms associated with a fundamental frequency. Because we are considering monthly data, the fundamental frequency corresponds to a period of twelve months and its five harmonics (the first harmonic is associated with a period of six months and the second one to three months). It is assumed here that each trigonometric term has the same variance σ_ω^2. Finally, the irregular term in the signal is a residual not explained by the structural time series components. In order to estimate the hyperparameters and obtain estimates of the unknown signal we now cast the components of the signal plus noise model into state space form. The state vector has a dimension equal to 15 (one component for the trend, one for the slope, eleven components for the seasonal term, one irregular and one for the AR(1) standardized survey error), and it is represented by

$$\boldsymbol{\alpha}_t = [\mu_t, \beta_t, \gamma_{1t}, \gamma_{1t}^*, \gamma_{2t}, \gamma_{2t}^*, \gamma_{3t}, \gamma_{3t}^*, \gamma_{4t}, \gamma_{4t}^*, \gamma_{5t}, \gamma_{5t}^*, \gamma_{6t}, \epsilon_t, \ell_t^*] \tag{4.6.11}$$

with all the fourteen first components considered as non-stationary processes. The model in Equation 4.6.10 can be cast into a state space form using the matrices and vectors in Equations 3.3.5 - 3.3.12 for monthly data.

The hyper-parameters corresponding to the variances of the white noise disturbances in the signal and the coefficient in the AR(1) noise model are estimated by maximum likelihood (see section 3.4.2). The maximum likelihood (ML) estimation of the hyper-parameters was implemented by use of the nlminb routine available in R and S+ using a quasi-Newton approach. In order to reduce the possibility of obtaining sub-optimal local maxima after the numerical optimization procedure, different sets of starting values in a grid of positive values for the variances and the range [-1,1] for the autoregressive term was used. Because the problem is constrained to positive variances and autoregressive terms in a specific range, different transformations were tested to convert the constrained problem into one of unconstrained maximization (see Durbin and Koopman (2001), pages 143-144). After transformation, the standard error of the estimates were obtained from the associated information matrix by using

the *delta method* (Oehlert, 1992). Considering the model in Equation 4.6.10 and after an extensive search on a grid of about 3000 different sets of initial values for the five hyperparameters; the corresponding estimated values (standard errors in brackets) under this initial model were $\hat{\phi} = 0.788(0.086)$, $\hat{\sigma}^2_{\xi}$= 403.01908 (2.785), $\hat{\sigma}^2_{\zeta}$= 3580.29101 (9336.656), $\hat{\sigma}^2_{\omega}$= 127037.06409 (57885.94), $\hat{\sigma}^2_{\epsilon}$= 20.08723 (468.6856). The estimated variance of the disturbance in the AR(1) model was equal to $\hat{\sigma}^2_{\chi}$= 1-$\hat{\phi}^2$= 0.38 (see Appendix B2). The hyperparameters were estimated by maximum likelihood using the SsfPack algorithms implemented in the S-Plus module S+Finmetrics version 2.0 and the function nlminb (Zivot, Wang and Koopman (2004)). The log-likelihood using this model was -639.7366. The innovations (one-step-ahead prediction errors) and their variances, appearing in the likelihood, were calculated by use of the Kalman filter, initiated by a diffuse prior for the first 14 elements (non-stationary components) in the initial state vector $\boldsymbol{\alpha}_0$.

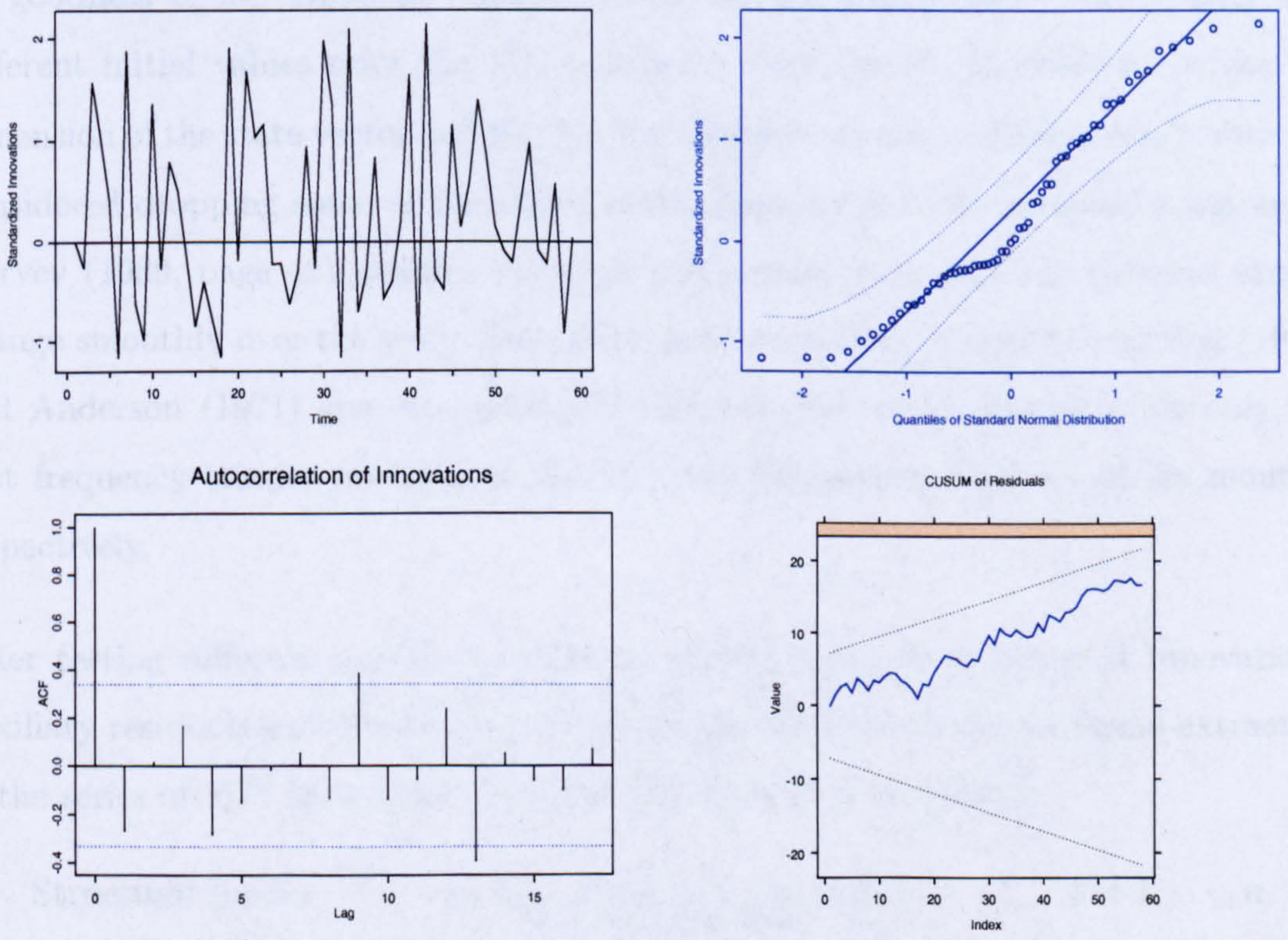


Figure 4.6. **Diagnostic Plots of the Innovations. Initial Model.**

Figure 4.6 shows some diagnostic plots for the innovations. The upper left time series

plot of the standardized innovations does not show any trace of outliers; the quantile comparison plot shows all the observed points being within the bootstrapped pointwise 99% confidence envelope around the normal line (Fox, 2002, page 29) but with long tails at the beginning and at the end of the curve of quantiles; the lower left autocorrelation plot show significant autocorrelations at lags 9 and 13 outside the 99% confidence interval (limits are calculated as $\pm 2.57/\sqrt{57} \sim 0.34$). The last plot in the right lower corner corresponds to a CUSUM plot (see section 3.4.4) showing no stability problems with this model. However, there is a definite downward movement after the observation in July 2000 (the same observation which appeared to be an outlier in Figure 4.2).

Final Model

Different alternative models were tested before we chose a final model according to its goodness of fit. Different ARMA models for the standardized survey error and different initial values prior the ML estimation were tested. In order to reduce the dimension of the state vector and the number of non-stationary components, it was also considered dropping some of the higher order frequencies in the seasonal component. Harvey (1989, page 42) justifies the approach arguing that seasonal patterns should change smoothly over the year. Also, some authors such as Abraham and Box (1978) and Anderson (1971) give examples of structural time series models using only the first frequency (twelve months) or the first two frequencies (twelve and six months) respectively.

After testing different models in order to achieve good fit in terms of innovations, auxiliary residuals and Monte Carlo experiments; the final model for signal extraction of the series of MPI turnover in the sawmills industry is as follows:

$$\text{Structural Model}: \quad y_t = \underbrace{\mu_t + \gamma_t + \epsilon_t + \lambda w_t}_{\eta_t} + \underbrace{k_t \ell_t^*}_{\ell_t} \qquad t = 1, \cdots, n$$

$$\text{Trend}: \quad \begin{cases} \mu_t = \mu_{t-1} + \beta_{t-1} + \xi_t; & \xi_t \sim N(0, \sigma_\xi^2) \\ \beta_t = \beta_{t-1} + \zeta_t; & \zeta \sim N(0, \sigma_\zeta^2) \end{cases}$$

$$\text{Seasonality}: \begin{cases} \gamma_t = \sum_{v=1}^{6} \gamma_{vt} & \\ \gamma_{vt} = \gamma_{v,t-1} \cos \kappa_v + \gamma^*_{v,t-1} \sin \kappa_v + \omega_{vt}; & \omega_{vt} \sim N(0, \sigma^2_\omega) \\ \gamma^*_{vt} = -\gamma_{v,t-1} \sin \kappa_v + \gamma^*_{v,t-1} \cos \kappa_v + \omega^*_{vt}; & \omega^*_{vt} \sim N(0, \sigma^2_\omega) \end{cases} \tag{4.6.12}$$

$$\text{Survey Error}: \quad \ell^*_t = \phi_1 \ell^*_{t-1} + \chi_t; \qquad \chi_t \sim N(0, \sigma^2_\chi)$$

The extra term λw_t in the signal accounts for the intervention effect due to the outlier detected in Figure 4.2 with w_t being a pulse variable of the form

$$w_t = \begin{cases} 0, & t \neq 31 \quad \text{(July 2000)} \\ 1, & t = 31 \quad \text{(July 2000)} \end{cases} \tag{4.6.13}$$

Harvey (1989, page 399) considers the state space form and the estimation under this model with intervention. The model depends on five hyper-parameters corresponding to the variances $\sigma^2_\xi, \sigma^2_\zeta, \sigma^2_\omega$ and σ^2_ϵ of the disturbances associated with the trend, slope, seasonal and irregular terms respectively, and the AR(1) coefficient ϕ_1 in the standardized survey error model. Considering the model in Equation 4.6.12. and after an extensive search on a grid of about 3000 different sets of initial values for the five hyperparameters; the corresponding estimated values (standard errors in brackets) under this final model were $\hat{\phi} = 0.738(0.079)$, $\hat{\sigma}^2_\xi$= 36315.503 (543615.5), $\hat{\sigma}^2_\zeta$= 1.00 (14.833), $\hat{\sigma}^2_\omega$= 22026.466 (53880.03), $\hat{\sigma}^2_\epsilon$= 36315.503 (880569.20). The estimated variance of the disturbance in the AR(1) model was equal to $\hat{\sigma}^2_\chi$= 1-$\hat{\phi}^2$= 0.455 (see Appendix B2). The hyperparameters were estimated by maximum likelihood using the SsfPack algorithms implemented in the S-Plus module S+Finmetrics version 2.0 and the function nlminb (Zivot et al. (2004)). The log-likelihood using this model was -637.0514.

We are aware that these variance estimators are highly insignificant. Pfeffermann, Feder and Signorelli (1998, pages 344-345) discuss how under series of short length (72 data points in this case, of which 15 were used for initialization of the Kalman filter) the asymptotic standard errors obtained from the inverse of the information (Hessian) matrix overestimate the true variances even though the point estimators could perform satisfactorily in terms of unbiasedness. Through simulations, they shown that the estimates of the hyperparameters become significant only with series of about 500 observations. A simulation study was conducted here in order to check the stability

of the maximum likelihood estimates under the particular set of initial values being chosen. Different series (300 in total) with the same length (72 observations) and under the final model in Equation 4.6.12. were simulated and a set of maximum likelihood estimates were obtained for the four hyper-parameters in every simulated series in order to check the stability and the adequacy of the initial values before the optimization. The maximum likelihood estimates using the series of the sawmills industry and the mean, median, and standard error (se) of the maximum likelihood estimates in the 300 simulated series appear in Table 4.2. The results show good approximation to the estimates obtained with the original series of study as they fall in the respective confidence intervals of the simulated values.

Statistic	$\hat{\sigma}^2_\xi$	$\hat{\sigma}^2_\zeta$	$\hat{\sigma}^2_\omega$	$\hat{\sigma}^2_\epsilon$	$\hat{\phi}_1$
MLE Estimates	36315.503	1.000	22026.466	36315.503	0.738
MLE se	540627.5	14.833	53880.03	880569.20	0.079
Simulated MLE Mean	45686.604	11.567	46394.223	637304.46	0.690
Simulated MLE Median	38520.984	10.602	40564.000	523886.720	0.739
Simulated MLE se	357554.500	8.479	31786.260	518792.800	0.102

Table 4.2. **Original MLE estimates, standard erors (se) and summary of statistics of simulated MLE estimates in 300 series under the final model.**

The results shows the presence of a very small value for the variance of the slope. Given the magnitude of the standard errors for the estimates of the hyperparameters, we used the likelihood ratio test (LRT) to evaluate the adequacy of nested models. We tried to estimate a simpler model fixing the variance of the slope to zero and then we used the LRT test to compare between the two models, one with a random and the other one with a fixed slope, respectively. The LRT is a statistical test of the goodness of fit between two models used to compare a more complex model to a simpler nested one, in order to see if the more extensive model is required (see Mood, Graybill and Boes (1974, section 5.1) and Harvey (1989, section 5.1.1)). The LRT statistic approximately follows a chi square distribution with degrees of freedom equal to the number of additional parameters in the more complex model. Using this information and because the p-value

of the test was not close to zero in this particular case; we decided to use a fixed slope in the structural component for the trend. Further LRT checks sequentially fixing the other hyperparameters in the model to zero showed significant differences and hence, the other components were kept as random components.

Table 4.3 (at the end of this chapter) summarizes the main results for the models in Equations 4.6.10 and 4.6.2, the latter with a fixed slope. Maximum likelihood estimates and their respective standard errors; test statistics of autocorrelation (Ljung-Box and Box-Pierce); normality (Shapiro-Wilks and Jarque-Bera) and heteroscedasticity with their respective p-values are included in this table. The results for the last column correspond to the model after fixing the variance to the slope component to zero.

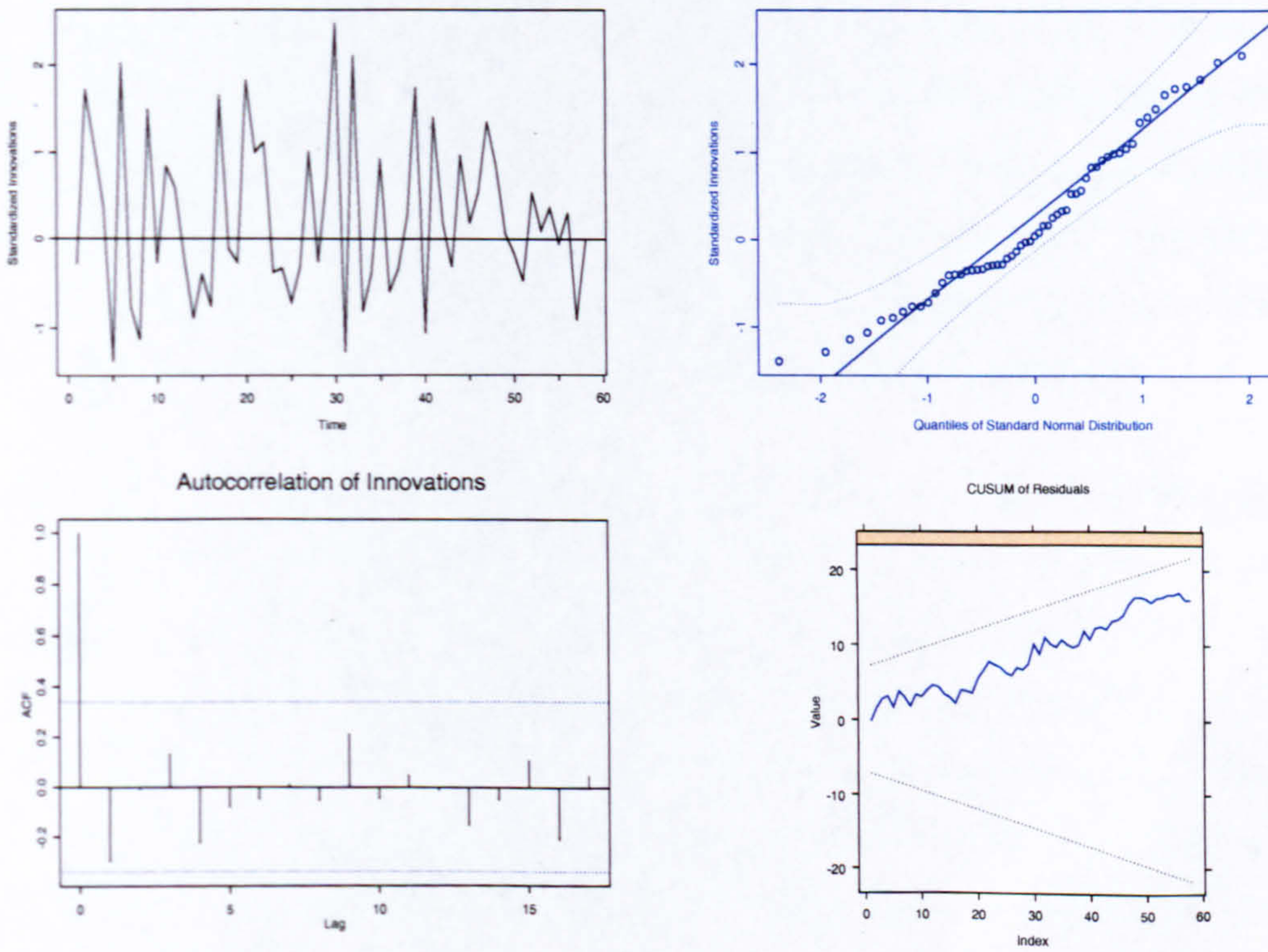


Figure 4.7. **Diagnostic Plots of the Innovations. Final Model.**

Diagnostic testing performed on the innovations (one-step-ahead predictions) generated from the Kalman filter under both models should lead to conclude that they are approximately normal distributed white noise variables. Examination of the test

results gives no reason to question the adequacy of the final model whereas they indicate the presence of serial correlation, non-normality and heteroscedasticity in the innovations for the initial model in page 92. Test procedures based on plots of the innovations in Figure 4.7 show no trace of outliers and constant level in the time series plot; all points within the bootstrapped pointwise 99% confidence envelope in the quantile comparison plot without the long tails at the beginning and at the end of the plot; the autocorrelation plot does not show significant autocorrelations and the CUSUM plot does not show any stability problems with this model.

Figure 4.8 shows the smoothed estimates for each of the components of the state vector. The first graph shows the filtered estimates of the trend component (series in blue) plotted against the original values of turnover in this industrial sector. The trend estimates look as an approximate constant trend in the middle of the series of interest (series in black). The remaining plots correspond to the estimated seasonal effects (γ coefficients) (six in total), then , one component for the irregular term and finally the AR(1) standardized survey errors. The sum of these nine processes coincides exactly with the values in the original monthly series as the sampling error was included in the state vector and there is no disturbance in the observation equation.

Finally, in order to test the goodness of fit of the model, we consider the final model for 3000 simulated series under the same model. Then, for each time point, we calculate the α% and the $(1-\alpha)$% percentile and we overplot the series of turnover with the series of percentiles. The model was considered to give a good fit in the model if the original series under consideration is approximately between the 5th (1st) and the 95th (99th) percentile of the simulated values. Figure 4.9 shows good fitting of the model with few points outside the 95% confidence interval, most of them corresponding to the lower values in December each year (due to the high seasonality of the series).

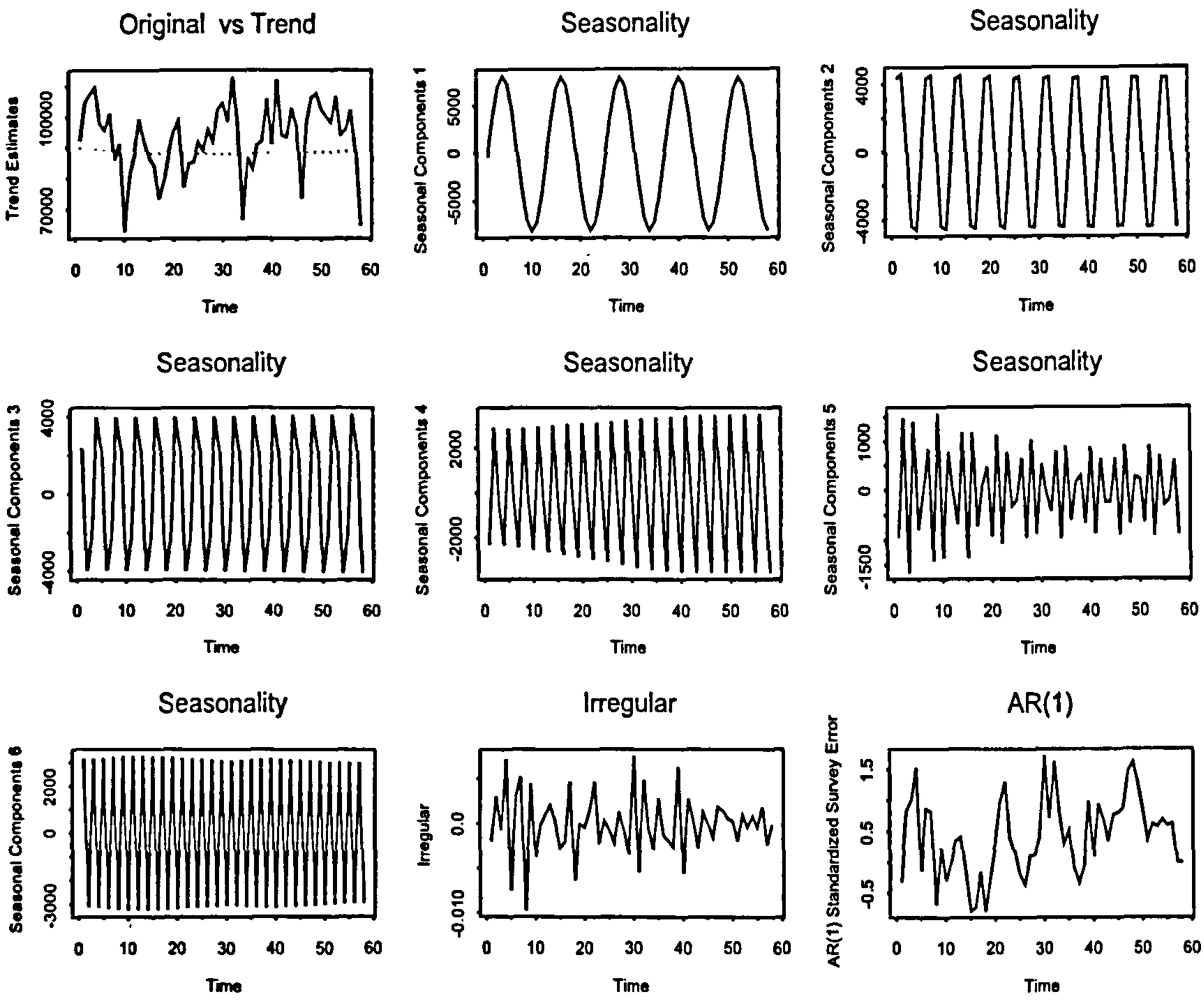


Figure 4.8. **Smoothed Structural Values. Final Model.**

4.6.3 Benchmarked Estimators

After obtaining the values of the smoothed monthly series with the final model, the annual information is used now to benchmark the information. In the second step, monthly benchmarked values will be obtained after incorporating the information contained in the annual total estimates from ABI. There are two possible alternatives to obtain the benchmarked values: *Binding estimation* will refer to the situation when the sums of the monthly series per year equal the annual benchmarks exactly and the *non-binding case* is when the estimation takes account for the annual sampling errors. In both cases, the corresponding reductions in the standard errors and coefficients of variation from the MPI monthly estimates are shown. Figure 4.10 shows the benchmarked

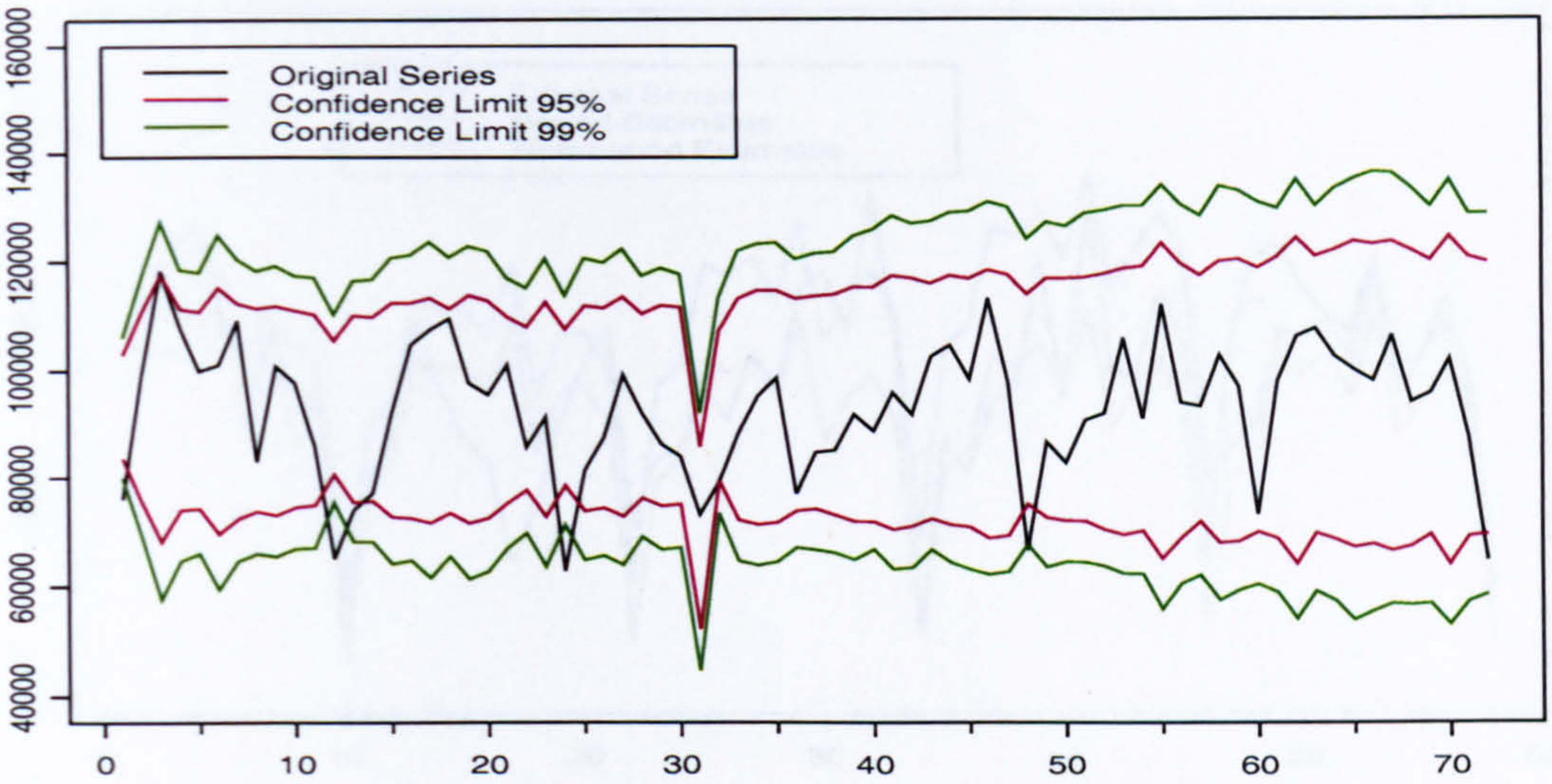


Figure 4.9. **Simulation of 3000 Series Under the Final Model. Series in Both Extremes are 95 and 99% Confidence Limits.**

values by the annual totals for the final model for the data of Turnover in the industrial subsector of wood manufacturing (sawing and milling) in Equation 4.6.2 considering both binding and non-binding estimation. It should be noticed that, for example, the annual estimate for the last year in the series was 1120000 whereas the sum of the original monthly estimates for this specific year is 1164650. The idea of benchmarking is to make these two totals more consistent. Then with binding totals, the sum of the benchmarked estimates is exactly 1120000 and with non binding totals it is 1071779 under the final model. However, in the non binding case and depending on a good specification of the model, it is expected to obtain a better set of monthly estimators with the corresponding monthly CV's being lower after benchmarking according to the results obtained in section 3.6.

Figure 4.11 shows a comparison between the original standard errors with those obtained after benchmarking using the final model in Equation 4.6.2. Series in black shows the standard errors k_t of the original monthly series, the dotted blue series shows the standard errors for the binding estimates ($\hat{\boldsymbol{\eta}}_B$; Equation 3.6.3) being even bigger than the standard errors of the smoothed estimates ($\hat{\boldsymbol{\eta}}_0$; page 53) in the first

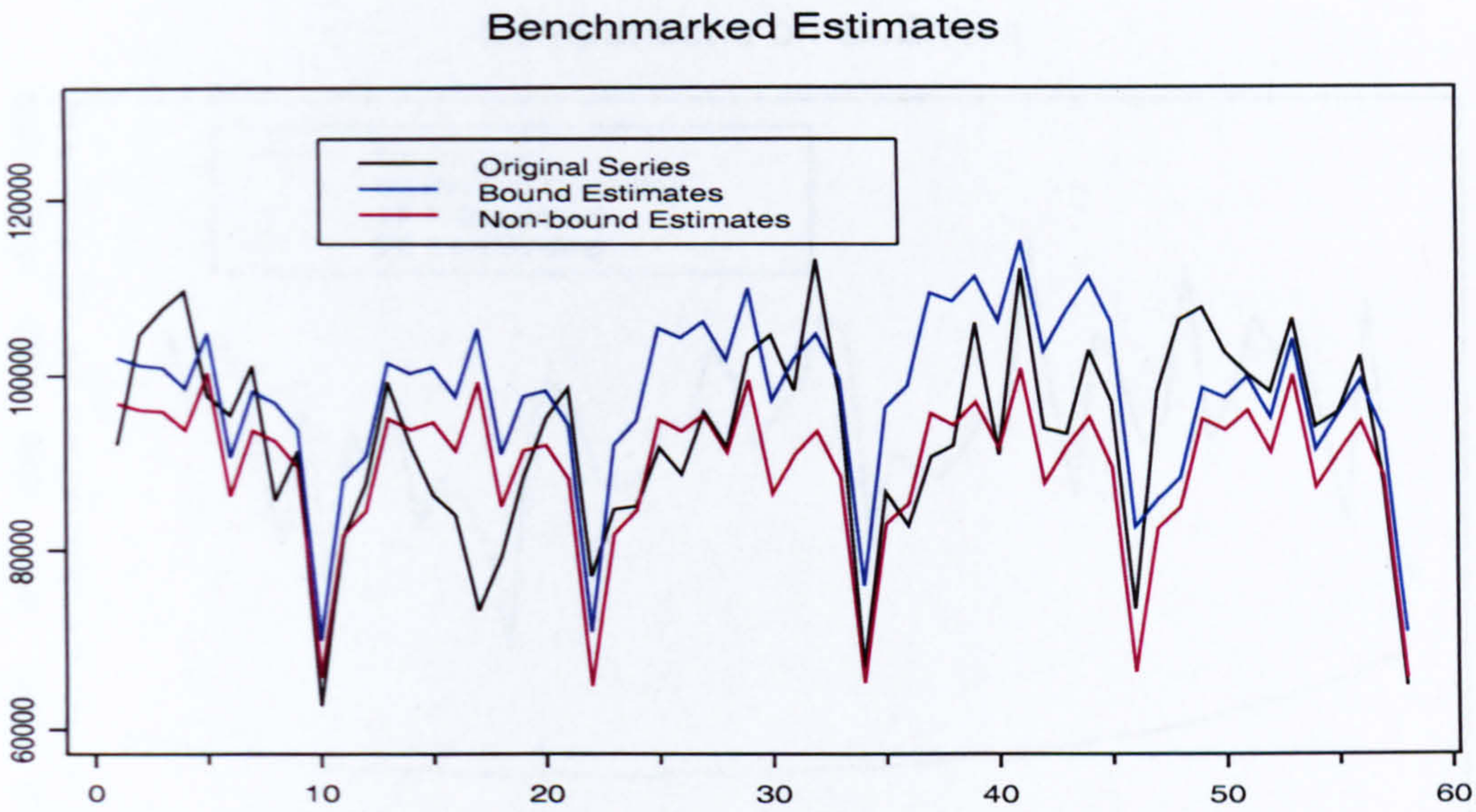


Figure 4.10. **Binding and Non-binding Estimates. Final Model**

step, plotted in red. Both smoothed and non-binding estimates (Equation 3.6.1., plotted in green) look very close. The same conclusions are obtained after examination of the estimated coefficients of variation (CVs) for the original series and the alternative estimators under consideration. In conclusion, there is a good reduction in the variability of the estimates after benchmarking, the reduction being bigger by the use of non-binding estimators. It is also noticed that the standard errors and CVs for binding estimators could be bigger than those from the smoothed values in the first step, confirming the results obtained in the last chapter. The fact that the variability of the binding estimators is bigger than the variability of the smoothed values indicates that a second step would not be necessary in order to improve the estimation. Table 4.4 shows a comparison of the corresponding standard errors for each estimator in the last six months in the series.

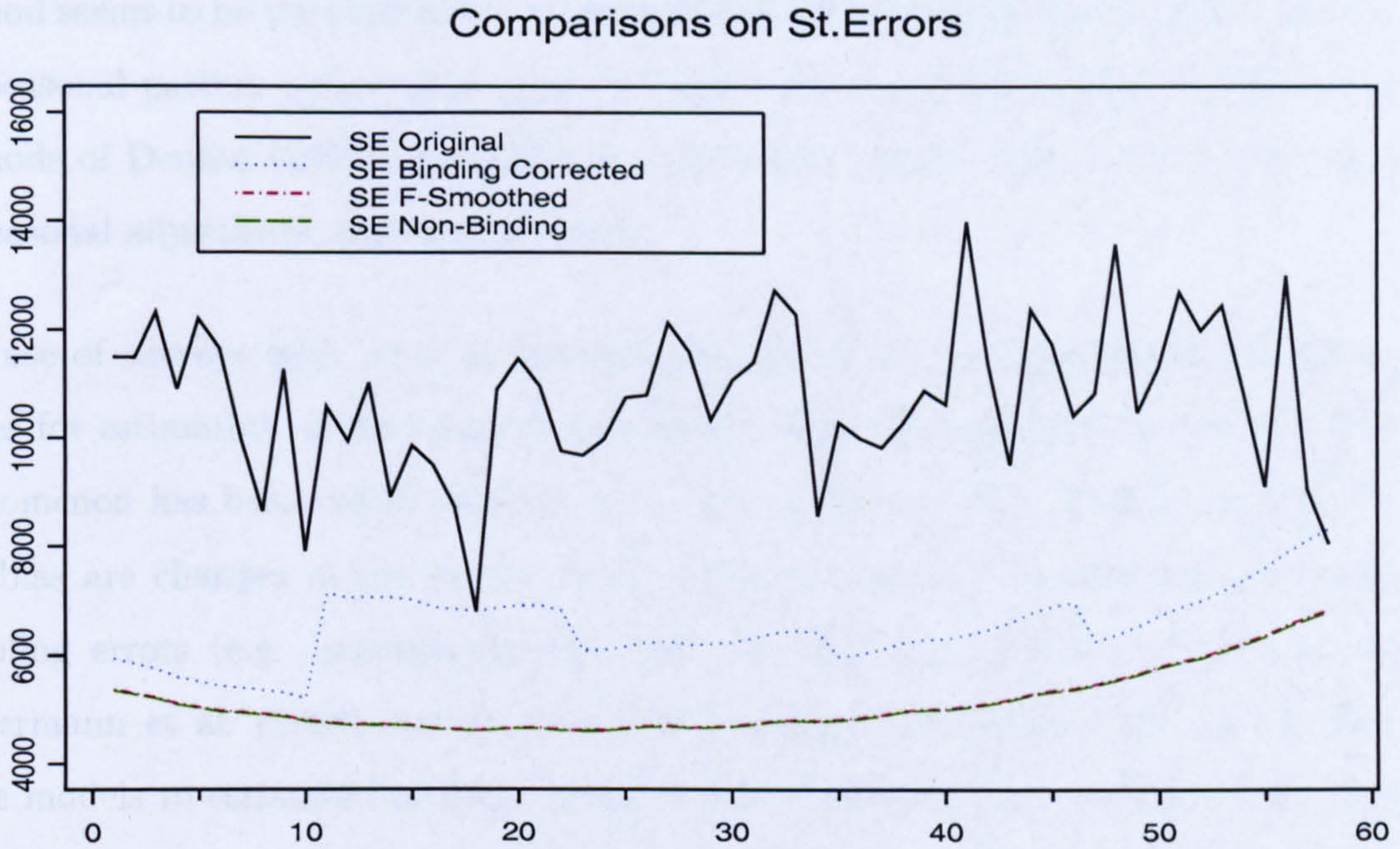


Figure 4.11. **Comparison of Standard Errors. Final Model.**

4.7 Conclusions and Further Issues

After reviewing the literature on business surveys and studying the particularities of this kind of surveys in the UK, some key points must be highlighted: the sampling frame is the same for all business surveys (IDBR) with the statistical units (enterprises) being classified according to the SIC92 code; the sampling design is also basically the same for all business surveys (a stratified srs sampling) with slight differences in the stratification and the rotation schemes. In particular, ABI and MPI use both stratified random sampling surveys according to different employment sizebands. Because of the use of the matched pairs estimator for measuring change between two periods, it is necessary to benchmark the levels according to results of Kokic and Jones (1998).

One important characteristic was not mentioned here: the estimates that are published in business surveys are normally seasonally adjusted. One important question is: when should the benchmarking be applied, before the seasonal adjustment or afterwards? Gubman and Burck (2005, section 7) state that "the Denton proportional

method seems to be the preferable one among the binding benchmarking methods from the seasonal pattern preservation point of view", when comparing the benchmarking methods of Denton (1971) and Cholette and Dagum (1994) with a special emphasis on seasonal adjustment and quality issues.

The use of surveys with rotating sampling designs could produce different expected values for estimators of the same characteristics from different rotation groups. The phenomenon has been called *rotation group bias* (Bailar, 1975). Possible reasons for this bias are changes in the survey methodologies from time to time and also non-sampling errors (e.g. missclassification and non-response). Some authors such as Pfeffermann et al. (1998) and Pitta and Silva (2004) have studied the use of state space models to estimate this bias. However due to the fact that the survey rotation schemes for ABI and MPI are different according to the strata in each survey and the stratification is different comparing the two surveys; the rotation scheme was not considered into the final state space model. Another aspect to take into consideration is what information is completely available from the repeated surveys being analysed. For instance, in some cases, not all the standard error information is available and there is a high rate of missing values and outliers. Also, some data and/or their standard errors may not be available at lower levels of disaggregation. The outlier problem causes instability in the standard errors making necessary to include heteroscedasticity factors and intervention terms in the benchmarking model.

All the examples in the last sections were considered under the application of the two step benchmarking method. The final model which has been proposed in the application was obtained in an empirical way, trying different modifications to standard, well known models and by checking the assumptions over the innovations, standardised smoothing residuals, statistics of goodness of fit, MLE estimates of the hyperparameters and the reduction in the standard errors and coefficients of variation in comparison to the original survey estimates. It is not easy to postulate a model for the trend and seasonality, particularly since the ideal situation is to have a short state vector as possible. One possibility to reduce the dimensionality of the state vector is to exclude the observation errors from the state equation, which will produce autocorrelated errors

in the observation equation. Pfeffermann and Tiller (2005) proposed a modification to the Kalman filter, which can deal with this problem. The GLS filter proposed by them could be used in order to compare the two approaches but this approach was not considered here.

Most of the series in Business Surveys behave multiplicatively instead of additively. It would be necessary to formulate adequate models to deal with this case. The common practice using logarithms is not possible in these situations because although $y_t = \eta_t \cdot \ell_t$, the benchmarking relations $x = L\eta + e_t$ are still linear and expressed in terms of the values η_t and not in terms of $\log(\eta_t)$. Durbin and Quenneville (1997) proposed some alternatives using what they called posterior mode estimates (Farmeir (1992) and Durbin and Cordero (1993)). Multiplicative models are out of the scope of this thesis but they can be considered as an area of further work, specially in the multivariate extension of the next chapter.

This chapter has also proposed to model the monthly ARMA survey errors in an empirical way trying different models and checking both the assumptions and some empirical results through Monte Carlo experiments. The initial values are an important decision before the maximum likelihood estimation. When there is no knowledge about the initial values of the parameters, it is recommended to perform an extensive search considering restrictions on the parameters such as positive variances and conditions of stationarity and invertibility in the ARMA parameters. In this application, some transformations were done in order to get the maximum likelihood estimates and the delta method was used in order to get the standard errors of the estimates.

Additionally, in order to reduce the dimensionality of the state vector, some alternatives to the seasonal component were considered. Harvey (1989, page 42) states that because seasonal patterns change relatively smoothly over the year, it may be sometimes reasonable to drop some of the higher-order frequencies in the trigonometric form of the seasonality. The first frequency, which corresponds to a period of twelve months is known as the fundamental frequency while the remaining are harmonics. Since we are benchmarking monthly data, the seasonal term is equal to the sum of the fundamen-

tal frequency plus five harmonics. Different models were considered by using different number of harmonics. In short, different initializations, different starting values for the ML estimation, different number of frequencies in the seasonal component and different ARMA models for the standardized survey errors were considered before getting a final model. Traditional diagnostics in the innovations and the auxiliary residuals were considered to evaluate the fitting and also some Monte Carlo experiments.

	Initial Model Equation 4.6.10	**Final Model** Equation 4.6.2
I. Maximum Likelihood Estimation		
Log-Likelihood	−639.737	−637.051
AR(1) Coefficient(ϕ)	0.788	0.738
Standard Error	*(0.086***)*	*(0.081***)*
Trend Variance($\hat{\sigma}^2_\xi$)	403.019	36315.503
Standard Error	*(2.785)*	*(543615.5)*
Slope Variance($\hat{\sigma}^2_\zeta$)	3580.291	(*Fixed*)
Standard Error	*(9336.656)*	(*Fixed*)
Seasonal Variance($\hat{\sigma}^2_\omega$)	127037.1	22026.466
Standard Error	*(57885.94)*	*(51650.31)*
Irregular Variance($\hat{\sigma}^2_\epsilon$)	20.087	36315.503
Standard Error	*(468.686)*	*(44282.56)*
II. Diagnostic Tests		
Ljung-Box Statistic	52.502	20.2992
p-Value (H_0: No autocorrelation)	*(0.016**)*	*(0.2592)*
Box-Pierce Statistic	42.873	24..4308
p-Value (H_0: No autocorrelation)	*(0.002***)*	*(0.1082)*
Shapiro-Wilks Statistic	0.947	0.9670
p-Value (H_0: Normality)	*(0.049*)*	*(0.1155)*
Jarque-Bera Statistic	3.571	2.6447
p-Value (H_0: Normality)	*(0.205)*	*(0.2665)*
Heteroscedasticity Test	0.683	0.4996
p-Value (H_0: Homoscedasticity)	*(0.04*)*	*(0.1056)*

Table 4.3. Summary of maximum likelihood estimates and diagnostic tests - Initial and Final Model

Statistic	Jan	Feb	Mar	Apr	May	Jun
Original MPI series	10826.72	13570.44	10472.30	11309.91	12681.12	11975.08
Binding estimates	6289.458	6414.147	6530.867	6629.079	6799.957	7014.555
Smoothed est. (step 1)	5493.261	5580.918	5662.305	5730.294	5847.532	5992.906
Non-bind. estimates	5465.060	5551.342	5631.413	5698.274	5813.503	5956.269
Statistic	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Original MPI series	12432.81	10844.88	9092.28	12989.12	9039.86	8454.807
Binding estimates	7198.613	7394.963	7589.441	7767.592	8025.776	8325.419
Smoothed est. (step 1)	6115.981	6245.678	6372.553	6487.426	6651.676	6839.101
Non-bind. estimates	6077.036	6204.196	6328.486	6440.927	6601.546	6784.600

Table 4.4. Standard Errors of Estimators for the Months Corresponding to the Last Year (2003) Using the Final Model.

Chapter 5

Benchmarking and Contemporaneous Disaggregation

In the previous chapters, we were interested in how to adjust for discrepancies between a high frequency series and aggregated low frequency series (benchmarking problem). Another common problem in the analysis of business surveys, and related to benchmarking, is how to prepare tabular data classified by attributes (e.g in a contingency table with attributes as columns and points in time as rows) when auxiliary information for the variable of study is available not only with annual frequency but also in aggregates such as the whole industrial sector.

Aggregated data corresponds to information obtained during the data collection process which is not accessible at smaller levels (including the micro-level). For instance, National Accounts are usually aggregated both in time (say annually) and contemporaneously (say by economic sector or geographical region). In practice, however, prediction and economical planning is often required at sub-annual frequency and also disaggregated at sub-sectorial or sub-regional levels. In particular, carrying out econometric analysis with temporally aggregated data is far from optimal (Zellner and Mornmarquette, 1976; Guerrero and Nieto, 1999). On the other hand, as aggregated data normally come from administrative or larger sources, they are particularly useful as auxiliary information to improve estimation at disaggregated levels.

Quenneville and Rancourt (2005), Fortier and Quenneville (2006) and Quenneville, Fortier, Chen and Latendresse (2006) discuss how many surveys publish seasonally adjusted (SA) series (following seasonal adjustment by specialized software as X11 or X12-ARIMA) that must fulfill various aggregation constraints. In particular, the Monthly Retail Trade Survey (MRTS) in Statistics Canada publishes SA series by industry and region. The SA national total is obtained as the sum of the 19 industries; however, this total is not necessarily equal to the sum of the 13 regions. Quenneville and Rancourt (2005, page 1) state that:

"'An alternative (...) is to present the monthly discrepancies openly; however, showing explicit discrepancies usually causes confusion among users and criticism or embarrassment to the publishers".

Two particular problems will be considered in this chapter: firstly, when the aggregates (marginal totals) do not correspond with the sum of the disaggregated values because, for example, aggregate and disaggregate values were estimated from different surveys (Dagum and Cholette, 2006, chapter 12) and secondly, when data is available in an aggregate form only (i.e. only the marginal totals by rows and columns are available) and it is desirable to estimate the disaggregated high-frequency data (Dagum and Cholette, 2006, chapter 13). These two situations are explained with more detail in Table 5.1. Being the most common case, we will assume for this chapter that the low-frequency series is observed annually.

5.1 Preliminaries

The general structure of the disaggregation problem is presented in Table 5.1. This table provides a general overview of the final table of parameters to be estimated as presented in Di Fonzo and Marini (2005) with slight changes in notation. In this general form, every cell is an unknown parameter to be estimated. Later on, we will refer to two particular cases of this table.

Letting $i = 1, \cdots, m$ be an index denoting year and $j = 1, \cdots, P$ an index denoting subsector; we wish to estimate $\eta_1, \cdots, \eta_j, \cdots, \eta_P$, P unknown column vectors each of dimension n satisfying both contemporaneous row totals and temporal aggregation constraints every year. Using the same notation as in the chapters before, n will denote the length of the high frequency series (i.e. total number of quarters or months); K denotes the number of high-frequency periods per year (i.e. $K = 4$ if quarters, $K = 12$ if months) and m will denote the number of years ($m = [n]_K$, with $[x]_b$ denoting the integer part of x/b). In the case of monthly data, the problem can be considered as $m + 1$ temporal contingency tables with the first m tables having dimension 12 x P and their corresponding row and column totals. Throughout this chapter, we will refer to the P columns as subsectors adding to the total sector (last column in Table 5.1).

In the UK, for example, a Standard Industrial Classification (SIC) was first introduced in 1948 in order to classify business, establishments and other statistical units according to their type of economic activity. For instance, considering the sector 15.3 " "Processing of Fruits and Vegetables", this sector is subdivided into three subsectors:

- *15.31* Processing and Preserving of Potatoes
- *15.32* Manufacture of Fruit and Vegetable Juice and
- *15.33* Processing and Preserving of Fruits and Vegetables

The sector 15.3 is itself a subsector of the more general sector 15 (Manufacture of Food Products and Beverages), with subsectors 15.1 to 15.9. An example of a problem, related to the one studied in this chapter, would be how to use monthly information collected from the complete sector 15 to obtain monthly disaggregated estimates for its P=9 subsectors. After obtaining new or more accurate estimates for the subsectors, the new estimates for the subsector 15.3 could be disaggregated into its P=3 subsectors and so on. Suppose the information about businesses in sector 15 is collected through a monthly survey, so that K=12. If, for example, information for n=70 months is available, there are m=$[70]_{12}$= 5 complete years. The total sector at time t

Year	Period	Subsectors η_1	$\cdots$	η_j	$\cdots$	η_P	Total $\eta_.$
	1	η_{11}	$\cdots$	η_{1j}	$\cdots$	η_{1P}	$\eta_{1.}$
	2	η_{21}	$\cdots$	η_{2j}	$\cdots$	η_{2P}	$\eta_{2.}$
1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	K	η_{K1}	$\cdots$	η_{Kj}	$\cdots$	η_{KP}	$\eta_{K.}$
Total $\eta_{(1)}$		$\eta_{.1(1)}$	$\cdots$	$\eta_{.j(1)}$	$\cdots$	$\eta_{.P(1)}$	$\eta_{..(1)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	K(i-1)+1	$\eta_{K(i-1)+1,1}$	$\cdots$	$\eta_{K(i-1)+1,j}$	$\cdots$	$\eta_{K(i-1)+1,P}$	$\eta_{K(i-1)+1.}$
	K(i-1)+2	$\eta_{K(i-1)+2,1}$	$\cdots$	$\eta_{K(i-1)+2,j}$	$\cdots$	$\eta_{K(i-1)+2,P}$	$\eta_{K(i-1)+2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	t	η_{t1}	$\vdots$	η_{tj}	$\vdots$	η_{tP}	$\eta_{t.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Ki	$\eta_{Ki,1}$	$\cdots$	$\eta_{Ki,j}$	$\cdots$	$\eta_{Ki,P}$	$\eta_{Ki.}$
Total $\eta_{(i)}$		$\eta_{.1(i)}$	$\cdots$	$\eta_{.j(i)}$	$\cdots$	$\eta_{.P(i)}$	$\eta_{..(i)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	K(m-1)+1	$\eta_{K(m-1)+1,1}$	$\cdots$	$\eta_{K(m-1)+1,j}$	$\cdots$	$\eta_{K(m-1)+1,P}$	$\eta_{K(m-1)+1.}$
	K(m-1)+2	$\eta_{K(m-1)+2,1}$	$\cdots$	$\eta_{K(m-1)+2,j}$	$\cdots$	$\eta_{K(m-1)+2,P}$	$\eta_{K(m-1)+2.}$
m	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Km	$\eta_{Km,1}$	$\cdots$	$\eta_{Km,j}$	$\cdots$	$\eta_{Km,P}$	$\eta_{Km.}$
Total $\eta_{(m)}$		$\eta_{.1(m)}$	$\cdots$	$\eta_{.j(m)}$	$\cdots$	$\eta_{.P(m)}$	$\eta_{..(m)}$
	Km+1	$\eta_{Km+1,1}$	$\cdots$	$\eta_{Km+1,j}$	$\cdots$	$\eta_{Km+1,P}$	$\eta_{Km+1.}$
	Km+2	$\eta_{Km+2,1}$	$\cdots$	$\eta_{Km+2,j}$	$\cdots$	$\eta_{Km+2,P}$	$\eta_{Km+2.}$
$(m+1)$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	η_{n1}	$\cdots$	η_{nj}	$\cdots$	η_{nP}	$\eta_{n.}$

Table 5.1. Parameters to be Estimated under Multiple Disaggregation

is represented as $\eta_{t.}$ being equal to $\eta_{t.} = \sum_{j=1}^{9} \eta_{tj}$ and these values correspond to the last column in table 5.1. In this case, t=1, $\cdots$, 70.

The auxiliary information to be combined with the monthly data will consist of an annual survey (considered as the sum of the monthly values and containing estimates for all the subsectors) and also monthly estimates for the sector total. Normally large sample surveys are conducted annually, instead of a higher time frequency, because of their high costs in the collection of the information. Then, annual surveys will permit better sector disaggregation than monthly surveys as they can be prepared with more anticipation and they are more precise. However, business and economic data are often required in a disaggregated and subannual form and their estimation is the purpose of this chapter.

The problem is displayed in Table 5.1 with the subsector annual totals (subtotals by columns) represented by $\eta_{.j(i)}$ where j corresponds to the index of the subsector and i the index of the year of observation; $i = 1, \cdots, m$ and $j = 1, \cdots, P$. Throughout all this chapter, we will assume rows are months and columns are subsectors although, of course, the problem can be extended into many other applications.

Using the notation in Table 5.1, in the next paragraphs we will consider the following vectors:

- η_j, $j = 1, \cdots, P$ are the n x 1 vectors consisting of the disaggregated signal, $\eta_j = [\eta_{1j}, \cdots, \eta_{nj}]'$. In table 5.1, these vectors correspond to the columns (without the stacked total values each year);

- $\eta_{.}$ (n x 1) is the vector of contemporaneously aggregated signals, $\eta_{.} = [\eta_{1.}, \cdots, \eta_{n.}]'$ (total sector by month). In table 5.1, this vector correspond to the last column (without the stacked total values each year);

- $\eta_{(i)}$, i=1, $\cdots$, m will denote the 1 x P vectors of annual aggregated data in a given year i (stacked vectors in Table 5.1), $\eta_{(i)} = [\eta_{.1(i)}, \cdots, \eta_{.j(i)}, \cdots, \eta_{.P(i)}]$;

- $\eta_{j(.)}$; $j = 1, \cdots, P$; are mx1 vectors of annually aggregated data (or

another period of reference) by columns, $\eta_{j(.)} = [\eta_{.j(1)}, \cdots, \eta_{.j(i)}, \cdots, \eta_{.j(m)}]'$

Then, the following accounting constraints must hold in order to impose additivity to the rows and columns in table 5.1:

$$\sum_{j=1}^{P} \eta_j \quad = \quad \eta_. \tag{5.1.1a}$$

$$L'\eta_j \quad = \quad \eta_{j(.)}, \quad j = 1, \cdots, P \tag{5.1.1b}$$

where L is the n x m aggregation matrix converting high-frequency to low frequency data in Equation 2.2.3. Each element of $\eta_{j(.)}$ can be considered as a non overlapping linear combination of η_j, with coefficients given by the K x 1 vector $\mathbf{1}$, K being the temporal aggregation order. Thus, in general, the matrix L' is equal to $L' = [I_m \otimes \mathbf{1}' : \mathbf{0}]$, where $\mathbf{0}$ is a null m x $(n-Km)P$ matrix added in the matrix L' to consider observations without an available benchmark in the horizon (ex-ante estimation) and the symbol $\otimes$ represents the Kronecker product between matrices.

Two particular problems will be addressed in the next chapters:

Case 1. *Reconciliation Problem.* We consider the case where in addition to the annually and sectorial aggregated information, p preliminary vectors of survey estimates of the vectors η_j are available, $j = 1, \cdots, P$. Representing these vectors as y_j of dimension n; the problem now is that $\sum_{j=1}^{P} y_j \neq \eta_.$ and y_j does not comply with $\eta_{(i)}$. Then, it is necessary to adjust these survey estimates in order to arrive at useful, consistent and publishable values that fulfill the constraints by rows and columns. Dagum and Cholette (2006, chapter 12) also name this case as *one-way classified systems.* The reconciliation problem is applied in the National Accounts context under the name of *balancing* and also, in a different context, is used for small area estimates which are not consistent with values obtained from corresponding larger areas (Pfeffermann and Bleuer, 1993; Rao, 2003).

Case 2. *Contemporaneous and Temporal Disaggregation.* We consider the case where the only available information are the vectors of marginal totals $\eta_.$ and $\eta_{(i)}$. In the example before, if only the annual totals and the total for the whole sector are available,

then the high frequency information inside the tables is missing and it is necessary to obtain estimates of the elements of the vectors η_j, $j = 1, \cdots, P$ using the stochastic properties of these series. Another case, related with the last one is when the available information consists of only the survey estimates of $\eta_.$ and $\eta_{(i)}$ and possibly also, measures of precision such as their estimated standard errors or coefficients of variation. The aim is how to combine monthly and annual information to get estimates of the missing values and more precise annual and sector totals. The situation is similar to the problem presented in Dagum and Cholette (2006, chapter 13) where they considered information from the Canadian Retail and Wholesale Trade Series which are classified by Province and Trade Group and where only the marginal totals are sufficiently reliable for publication. They called this two-way classified system as *marginal two-way systems*. The difference with the approach in this chapter is that one of the attributes of classification is time (by rows) in order to use the stochastic properties of the series of study. If there is more than one attribute to classify the series, all the possible combinations between the categories in different attributes could be included as columns.

The two problems presented above have been considered in the literature. Specifically for the first problem; Almon (1988) proposes an univariate polynomial method to convert annual series to quarterly figures by interpolation. Zaier and Trabelsi (2007) proposed a procedure which extends the polynomial method to the multivariate case when only the marginal totals are known. However, this procedure does not provide estimates for the subperiods in the first year of observation and also not standard deviations of the estimators. Dagum and Cholette (2006, section 13.2.3) propose a general analytical solution of the marginal two-way reconciliation model using generalised linear models. However, both procedures (Zaier and Trabelsi, 2007; Dagum and Cholette, 2006) do not take into account the stochastic properties of the series of study and they fail to define an online procedure. That means, they do not produce estimates for the months in a year if there is not a benchmark yet available.

On the other hand, the reconciliation case could be considered as a special case of balancing contingency tables, with time as one of the attributes. This problem has been widely studied in the literature; Deming and Stephan (1940) proposed the Iterative Proportional Fitting approach, also known as "raking". However, the use of this method could affect the original movements of the series (Dagum and Cholette, 2006, page 266). Regarding the reconciliation case, Di Fonzo and Marini (2003) present a multivariate extension of Denton's benchmarking procedure, according to which the temporal dynamics of the reconciled series should be as close as possible to those of the preliminary figures. This method does not account for survey errors. Also the dimensions of the matrices involved in the calculations can be considerable in practical situations, possibly giving rise to computational burden. Guerrero and Nieto (1999) developed a benchmarking method which exploits the autoregressive features of the preliminary series to determine the unobserved values of multiple time series whose temporal and contemporaneous aggregates are known. This method however, does not account for survey errors. Guerrero (2005) proposes a discrepancy measure to validate empirically the "compatibility" between the benchmarked estimates and the preliminary information.

In another context, Quenneville, Huot, Cholette, Chiu and Di Fonzo (2003), Quenneville and Rancourt (2005), Fortier and Quenneville (2006) and Quenneville et al. (2006) studied the problem of reconciling series after the application of seasonal adjustment procedures. The objective is to reconcile these series in order to satisfy some aggregation constraints, making sure that the annual totals constraints after seasonal adjustment remain satisfied. They use special regression models to perform the numerical computations for prorating. However, the consistency with the annual totals from the raw series is achieved at the expense of the quality of the seasonal adjustment. For series with significant calendar-related effects of moving seasonality effects, the annual totals of a seasonally adjusted series should differ from the unadjusted series. All the methods above fail to be online procedures, which is the requirement to produce benchmarked values even when, in the horizon, there is no a benchmark yet available. In the next sections, a benchmarking method is proposed for the two situations (contemporaneous disaggregation with missing values and reconciliation), which

is an online procedure and takes into consideration the survey errors and the stochastic structure of the series of study. A simulation has been carried out to illustrate the two procedures and evaluate the standard errors of the estimates produced by them.

The methods proposed in the next chapters require modelling the available information by state space structural time series models. We give solutions both for the case when there is no preliminary information and the high frequency values are missing for the subsectors, and also for the reconciliation case. In the first case, when it is not possible to get any preliminary information, the marginal totals (vectors $\boldsymbol{\eta}_{\cdot}$ and $\boldsymbol{\eta}_{(i)}$) are arranged into a special single series and then, Kalman filtering and smoothing is applied. In the second case, when there is some preliminary information , Kalman filtering and smoothing is applied in a multivariate state space models context.

5.2 Benchmarking and Contemporaneous Disaggregation

5.2.1 Preliminaries and Proposed Method

A solution for the second multivariate problem in Chapter 1 is presented here. This has been called Contemporaneous and Temporal Disaggregation and corresponds to the case when the vector $\boldsymbol{\eta}_{\cdot} = [\eta_{1\cdot}, \cdots, \eta_{n\cdot}]'$ (last column in Table 5.1 and the m stacked vectors $\boldsymbol{\eta}_{(i)}$ $(i = 1, \cdots, m)$ are all the available information. In many cases, this information is not even obtainable at the population level and the only available data are survey estimates of these vectors. We shall assume that the vector $\boldsymbol{z}$, of dimension n, contains the estimates of the monthly sector totals; and the vectors $\boldsymbol{x}_{(i)}$, of dimension P, contain the estimates of the annual subsector totals, $\boldsymbol{x}_{(i)} = [x_{1(i)}, \cdots, x_{P(i)}]$.

The initial configuration is similar to the structure presented in Table 5.1, but replacing the values η_{tj} $(t = 1, \cdots, n;\ j = 1, \cdots, P)$ in the inner cells by missing values (NA=

Year	Period	Subsectors η_1	$\cdots$	η_j	$\cdots$	η_P	z
	1	NA	$\cdots$	NA	$\cdots$	NA	$z_{K(i-1)+1}$
	2	NA	$\cdots$	NA	$\cdots$	NA	$z_{K(i-1)+2}$
i	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	K	NA	$\cdots$	NA	$\cdots$	NA	z_{Ki}
Total $x_{(i)}$		$x_{1(i)}$	$\cdots$	$x_{j(i)}$	$\cdots$	$x_{P(i)}$	

Table 5.2. Tabular representation of the contemporaneous disaggregation problem for year i, $i = 1, \cdots, m$.

"Not Available") and the row totals and column totals by their survey estimates $\boldsymbol{x}_{(i)}$ and $\boldsymbol{z}$. The aim is to predict the unobserved cells η_{tj} $t = 1, \cdots, n$; $j = 1, \cdots, P$ taking into account both the monthly and the annual survey errors and according to the available marginal totals per sector and per year. Table 5.2 shows the initial configuration for the year i, $i = 1, \cdots, m$.

Considering the row annual total vectors $\boldsymbol{\eta}_{(i)}$; $i = 1, \cdots, m$ as stacked vectors of annually aggregated data of dimension P, it follows that $\boldsymbol{x}_{(i)} = \boldsymbol{\eta}_{(i)} + \boldsymbol{e}_{(i)}$, with the vectors $\boldsymbol{x}_{(i)}$ containing the estimates of the annual subsector totals; $\boldsymbol{e}_{(i)}$ are the non-observed vectors of annual survey errors by rows in Table 5.1; $\boldsymbol{e}_{(i)} = [e_{1(i)}, \cdots, e_{P(i)}]$. Also, $\in_{\cdot} = [\in_{1\cdot}, \cdots, \in_{n\cdot}]'$ will denote the subannual survey errors associated to the sector totals. Then, these vectors of survey estimates can be decomposed in the form

$$\boldsymbol{z} = \boldsymbol{\eta}_{\cdot} + \in_{\cdot} = \sum_{j=1}^{P} \eta_j + \in_{\cdot}$$

$$\boldsymbol{x}_{(i)} = \boldsymbol{\eta}_{(i)} + \boldsymbol{e}_{(i)} \tag{5.2.1}$$

The last line in the former equation implies that $x_{\cdot j(i)} = \eta_{\cdot j(i)} + e_{j(i)}$ for each $j = 1, \cdots, P$ and then $x_{\cdot j(i)} = \sum_{t \in i} \eta_{tj} + e_{j(i)}$. It is assumed, as in the chapters before, that the vector of monthly survey errors $\in_{\cdot}$ is independent of the vector of annual survey errors $\boldsymbol{e}_{(i)}$ as the information contained in them comes from different independent sources. In the binding case, when the inner cells have to add exactly to the marginal

totals by row and columns, we will assume that $\in$. and $e_{(i)}$ are null vectors.

The proposed method consists of building a single series of values for each year as follows. In a given year i,

$$y^i_{CD} = (z_{K(i-1)+1}, \cdots, z_{Ki}, x_{(i)}) =: (z_{K(i-1)+1}, \cdots, z_{Ki}, x_{1(i)}, \cdots, x_{P(i)}) \quad i = 1, ..., m \tag{5.2.2}$$

putting the sector totals for each period and the stacked annual totals together in each year and the subindex CD indicating "Contemporaneous Disaggregation" to differentiate with the series to be used in the next chapter. In the particular case when the totals are population values (binding case), the sampling errors are zero and the last vector is equivalent to

$$y^i_{CD} = (\eta_{K(i-1)+1,.}, \cdots, \eta_{Ki,.}, x_{(i)}) = (\eta_{K(i-1)+1,.}, \cdots, \eta_{Ki,.}, \eta_{.1(i)}, \cdots, \eta_{.P(i)}) \quad i = 1, ..., m \tag{5.2.3}$$

After the series y^i is built and all the series y^i are concatenated, we produce the total series of observed values

$$y^*_{CD} = \begin{cases} (y^1_{CD}; \cdots; y^m_{CD}), & \text{if n is a multiple of m} \\ (y^1_{CD}; \cdots; y^m_{CD}; z_{Km+1}; \cdots; z_n), & \text{otherwise} \end{cases} \tag{5.2.4}$$

The length of this new series y^*_{CD} is equal to $n + Pm$ with n the number of observed high-frequency periods, P the number of vectors to be estimated and m the number of complete years ($m = [n]_K$ with K the number of high-frequency periods per year). After arranging the available information in this way, standard Kalman filter and smoothing is applied to the new series y^*. This series has implicitly all the contemporaneous and temporal constraints by row and sub-columns in Table 5.1. The application of the Kalman filter requires to express each single element y^*_s in its state space form introduced in Equations 3.3.1 and 3.3.2 given by

$$\begin{aligned} y^*_{s,CD} &= Z^*_{s,CD}\alpha^*_{s,CD} + \varepsilon^*_{s,CD} \quad \varepsilon^*_{s,CD} \sim N(0, h_{s,CD}) \\ \alpha^*_{s,CD} &= T^*_{s,CD}\alpha^*_{s-1,CD} + \vartheta^*_{s,CD} \quad \vartheta^*_{s,CD} \sim N(0_r, Q_{s,CD}) \end{aligned} \tag{5.2.5}$$

with r being the dimension of the state vector. The use of the method requires specifying suitable models for the trend, the seasonal effects, trading days effects and the

survey errors. A brief illustration of the state space form is presented below for the case when $P = 2$ and $K = 4$ and then, the general case for arbitrary values of P and K is presented. We have used the subindexes s,CD in order to reduce the formulas in the next chapter and to differentiate the SSF for Contemporaneous Disaggregation (CD) from the SSF for Reconciliation.

5.2.2 Binding Estimation for Quarterly Data - Bivariate Case

In this subsection, we present the proposed method for contemporaneous disaggregation in a simple scenario considering binding totals (without survey errors, i.e. $\epsilon_{\cdot} = \mathbf{0}_{n\times 1}$, $\boldsymbol{e}_{(i)} = \mathbf{0}_{1\times P}$ for $i = 1, \cdots, m$). We will consider the general case in the next subsection. The state space form for this simplified case is presented as an illustration.

Considering the bivariate case ($P = 2$) and data collected by quarters ($K = 4$); the table for the first year takes the form appearing in Table 5.2. The single series $\boldsymbol{y}^*_{CD}$ is built by arranging all the available information as follows

$$\boldsymbol{y}^*_{CD} = (y^*_{s,CD}) = (z_1, z_2, z_3, z_4, x_{1(1)}, x_{2(1)}, z_5, z_6, \cdots) \tag{5.2.6}$$

Year	Period	Subsectors η_1	Subsectors η_2	z
1	1	NA	NA	z_1
	2	NA	NA	z_2
	3	NA	NA	z_3
	4	NA	NA	z_4
Total $x_{(1)}$		$x_{1(1)}$	$x_{2(1)}$	

Table 5.3. Tabular representation of the contemporaneous disaggregation problem for the first year, P =2 and $K = 4$.

The length of the new series $\boldsymbol{y}^*_{CD}$ is $n + 2m$ with n the total number of observations

in the vector $\boldsymbol{z}$ and m the number of complete years. Considering the respective state space models for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$,

$$\begin{cases} \eta_{t1} &= \boldsymbol{Z}_{t1}\,\boldsymbol{\alpha}_{t1} + \varepsilon_{t1} \\ \boldsymbol{\alpha}_{t1} &= \boldsymbol{T}_{t1}\,\boldsymbol{\alpha}_{t-1,1} + \boldsymbol{\vartheta}_{t1} \\ \varepsilon_{t1} &\sim N(0, \sigma^2_{t1}) \\ \boldsymbol{\vartheta}_{t1} &\sim N(0, \boldsymbol{Q}_{t1}) \end{cases} \quad \text{and} \quad \begin{cases} \eta_{t2} &= \boldsymbol{Z}_{t2}\,\boldsymbol{\alpha}_{t2} + \varepsilon_{t2} \\ \boldsymbol{\alpha}_{t2} &= \boldsymbol{T}_{t2}\,\boldsymbol{\alpha}_{t-1,2} + \boldsymbol{\vartheta}_{t2} \\ \varepsilon_{t2} &\sim N(0, \sigma^2_{t2}) \\ \boldsymbol{\vartheta}_{t2} &\sim N(0, \boldsymbol{Q}_{t2}) \end{cases} \tag{5.2.7}$$

Using Equations 5.2.1 and 5.2.7, the values for the single series $\boldsymbol{y}^*_{CD}$ in the first year are given by

$$\begin{aligned}
y^*_{1,CD} = z_1 = \eta_{1.} &= \eta_{11} + \eta_{12} \\
&= \boldsymbol{Z}_{11}\boldsymbol{\alpha}_{11} + \varepsilon_{11} + \boldsymbol{Z}_{12}\boldsymbol{\alpha}_{12} + \varepsilon_{12} = \{\boldsymbol{Z}_{11}\boldsymbol{\alpha}_{11} + \boldsymbol{Z}_{12}\boldsymbol{\alpha}_{12}\} + \varepsilon^*_1 \\
y^*_{2,CD} = z_2 = \eta_{2.} &= \eta_{21} + \eta_{22} \\
&= \boldsymbol{Z}_{21}\boldsymbol{\alpha}_{21} + \varepsilon_{11} + \boldsymbol{Z}_{22}\boldsymbol{\alpha}_{22} + \varepsilon_{22} = \{\boldsymbol{Z}_{21}\boldsymbol{\alpha}_{21} + \boldsymbol{Z}_{22}\boldsymbol{\alpha}_{22}\} + \varepsilon^*_2 \\
y^*_{3,CD} = z_3 = \eta_{3.} &= \eta_{31} + \eta_{32} \\
&= \boldsymbol{Z}_{31}\boldsymbol{\alpha}_{31} + \varepsilon_{31} + \boldsymbol{Z}_{32}\boldsymbol{\alpha}_{32} + \varepsilon_{32} = \{\boldsymbol{Z}_{31}\boldsymbol{\alpha}_{31} + \boldsymbol{Z}_{32}\boldsymbol{\alpha}_{32}\} + \varepsilon^*_3 \\
y^*_{4,CD} = z_4 = \eta_{4.} &= \eta_{41} + \eta_{42} \\
&= \boldsymbol{Z}_{41}\boldsymbol{\alpha}_{41} + \varepsilon_{41} + \boldsymbol{Z}_{42}\boldsymbol{\alpha}_{42} + \varepsilon_{42} = \{\boldsymbol{Z}_{41}\boldsymbol{\alpha}_{41} + \boldsymbol{Z}_{42}\boldsymbol{\alpha}_{42}\} + \varepsilon^*_4 \\
y^*_{5,CD} = x_{.11} &= \eta_{11} + \eta_{21} + \eta_{31} + \eta_{41} \\
&= \boldsymbol{Z}_{11}\boldsymbol{\alpha}_{11} + \varepsilon_{11} + \boldsymbol{Z}_{21}\boldsymbol{\alpha}_{21} + \varepsilon_{21} + \boldsymbol{Z}_{31}\boldsymbol{\alpha}_{31} + \varepsilon_{31} + \boldsymbol{Z}_{41}\boldsymbol{\alpha}_{41} + \varepsilon_{41} \\
&= \{\boldsymbol{Z}_{11}\boldsymbol{\alpha}_{11} + \boldsymbol{Z}_{21}\boldsymbol{\alpha}_{21} + \boldsymbol{Z}_{31}\boldsymbol{\alpha}_{31} + \boldsymbol{Z}_{41}\boldsymbol{\alpha}_{41}\} + \varepsilon^*_5 \\
y^*_{6,CD} = x_{.21} &= \eta_{12} + \eta_{22} + \eta_{32} + \eta_{42} \\
&= \boldsymbol{Z}_{12}\boldsymbol{\alpha}_{12} + \varepsilon_{12} + \boldsymbol{Z}_{22}\boldsymbol{\alpha}_{22} + \varepsilon_{22} + \boldsymbol{Z}_{32}\boldsymbol{\alpha}_{32} + \varepsilon_{32} + \boldsymbol{Z}_{42}\boldsymbol{\alpha}_{42} + \varepsilon_{42} \\
&= \{\boldsymbol{Z}_{12}\boldsymbol{\alpha}_{12} + \boldsymbol{Z}_{22}\boldsymbol{\alpha}_{22} + \boldsymbol{Z}_{32}\boldsymbol{\alpha}_{32} + \boldsymbol{Z}_{42}\boldsymbol{\alpha}_{42}\} + \varepsilon^*_6
\end{aligned} \tag{5.2.8}$$

and so on for the following years. Now, the idea is to obtain filtered and smoothed values of the single series $\boldsymbol{y}^*_{CD}$. Then, it is necessary to write the last equations into a state space form given by Equation 5.2.5. However, considering the disturbances in Equation 5.2.8, the vector $\boldsymbol{\varepsilon}^*_s = [\varepsilon^*_1, \cdots, \varepsilon^*_n]$, is not a vector of serially uncorrelated disturbances. This is because, for example, $\text{Cov}(\varepsilon^*_1, \varepsilon^*_5) = E(\varepsilon^*_1\varepsilon^*_5) = \sigma^2_1 \neq 0$ and the Kalman filter would not yield the conditional mean of the state vector.

Instead, considering the analogous state space models in Equation 5.2.7 but including the disturbances into the state vector

$$\begin{cases} \eta_{t1} &= Z_{t1}\,\alpha_{t1} \\ \alpha_{t1} &= T_{t1}\,\alpha_{t-1,1} + \vartheta_{t1} \\ \vartheta_{t1} &\sim N(0, Q_{t1}) \end{cases} \quad \text{and} \quad \begin{cases} \eta_{t2} &= Z_{t2}\,\alpha_{t2} \\ \alpha_{t2} &= T_{t2}\,\alpha_{t-1,2} + \vartheta_{t2} \\ \vartheta_{t2} &\sim N(0, Q_{t2}) \end{cases} \tag{5.2.9}$$

in the way it was done for the RWN model in Equation 3.3.4 and for the BSM model in Equation 3.3.9, the observation equation can be expressed as $y^*_{s,CD} = Z^*_{s,CD}\alpha^*_{s,CD}$ by building the state vector $\alpha^*_{s,CD}$ and the observation matrix $Z^*_{s,CD}$ in the following specific way.

Analogous to the notation in Chapter 3, we will denote the number of components of the state vector α_{tj} by r_j with $j = 1, 2$ and we will let $\alpha^s_j = [\alpha_{sj}, \cdots, \alpha_{s-3,j}]$ to be a vector of dimension $4r_j$. The state vector $\alpha^*_{s,CD}$ in the observation and transition equations is expressed as the concatenation of the individual vectors associated with each column. In this particular case, the dimension of this new state vector $\alpha^*_{s,CD}$ is equal to $4r$ x 1 with $r = r_1 + r_2$.

$$\alpha^*_{s,CD} = [\alpha^s_1; \alpha^s_2]' = [\alpha_{s1}; \cdots; \alpha_{s-3,1}; \alpha_{s2}; \cdots; \alpha_{s-3,2}]' \tag{5.2.10}$$

The observation 1 x $4r$ system matrices $Z^*_{s,CD}$ for the first year will be equal to

$$\begin{aligned} Z^*_{1,CD} &= [Z_{11}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_{12}; 0_{r_2}; 0_{r_2}; 0_{r_2}] \\ Z^*_{2,CD} &= [Z_{21}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_{22}; 0_{r_2}; 0_{r_2}; 0_{r_2}] \\ Z^*_{3,CD} &= [Z_{31}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_{32}; 0_{r_2}; 0_{r_2}; 0_{r_2}] \\ Z^*_{4,CD} &= [Z_{41}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_{42}; 0_{r_2}; 0_{r_2}; 0_{r_2}] \\ Z^*_{5,CD} &= [Z_{41}; Z_{31}; Z_{21}; Z_{11}; 0_{r_2}; 0_{r_2}; 0_{r_2}; 0_{r_2}] \\ Z^*_{6,CD} &= [0_{r_1}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_{42}; Z_{32}; Z_{22}; Z_{12}] \end{aligned} \tag{5.2.11}$$

and at the beginning of the second year, $Z^*_{7,CD} = [Z_{51}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_{52}; 0_{r_2}; 0_{r_2}; 0_{r_2}]$ and the same equations are repeated every six periods (four quarters and two additional instants for the number of variables). Notice that the product $Z^*_{s,CD}\alpha^*_{s,CD}$ produces the

values $y^*_{s,CD}$. The formula $y^*_{s,CD} = Z^*_{s,CD}\alpha^*_{s,CD}$ corresponds to the observation equation for y^*_{CD}.

Considering now the transition equation, this takes the form $\alpha^*_{s,CD} = T^*_{s,CD}\alpha^*_{s-1,CD} + \vartheta^*_{s,CD}$. This set of equations can be obtained by skipping the Kalman filter updating equations for the SSF of y^*_{CD} at the points related to the values $x_{1(i)}$ and $x_{2(i)}$ with $i = 1, \cdots, m$. This is done to keep the stochastic nature of the vector of sector totals z. In this way, the value $y^*_{7,CD}$ is related with the value $y^*_{4,CD}$ but not with the values $y^*_{5,CD}$ and $y^*_{6,CD}$. Values $y^*_{5,CD}$ and $y^*_{6,CD}$ are used to update the annual totals during the iterative procedure.

Considering time variant state space models with different transition matrices for the values z_t $(t = 1, \cdots, n)$ and the values $x_{j(i)}$ $(i = 1, \cdots, m;\ j = 1, 2)$. The transition matrices for the first year can be written as follows

$$T^*_{s,CD} = \text{diag}\left(\begin{bmatrix} T_{s1} & 0_{r_1 x r_1} & 0_{r_1 x r_1} & 0_{r_1 x r_1} \\ I_{r_1} & 0_{r_1 x r_1} & 0_{r_1 x r_1} & 0_{r_1 x r_1} \\ 0_{r_1 x r_1} & I_{r_1} & 0_{r_1 x r_1} & 0_{r_1 x r_1} \\ 0_{r_1 x r_1} & 0_{r_1 x r_1} & I_{r_1} & 0_{r_1 x r_1} \end{bmatrix}, \begin{bmatrix} T_{s2} & 0_{r_2 x r_2} & 0_{r_2 x r_2} & 0_{r_2 x r_2} \\ I_{r_2} & 0_{r_2 x r_2} & 0_{r_2 x r_2} & 0_{r_2 x r_2} \\ 0_{r_2 x r_2} & I_{r_2} & 0_{r_2 x r_2} & 0_{r_2 x r_2} \\ 0_{r_2 x r_2} & 0_{r_2 x r_2} & I_{r_2} & 0_{r_2 x r_2} \end{bmatrix}\right)_{4r \times 4r} \tag{5.2.12}$$

for $s = 1, \cdots, 4$ and $T^*_{s,CD} = I_{4r}$ for $s = 5, 6$ and the same kind of matrices are repeated every six periods (four quarters and two additional instants for the number of variables). For instance, $T^*_{7,CD}$ is obtained using T_{51} and T_{52} replacing T_{s1} and T_{s2} in Equation 5.2.12. The formulation of the transition equation in the first year is complete by setting

$$\vartheta^*_{s,CD} = \begin{cases} [\vartheta_{s1}; 0_{r_1}; 0_{r_1}; 0_{r_1}; \vartheta_{s2}; 0_{r_2}; 0_{r_2}; 0_{r_2}], & s = 1, \cdots, 4 \\ 0_{4r}, & s = 5, 6 \end{cases} \tag{5.2.13}$$

and this pattern is repeated in the same way every six indices s (at the beginning of the next year, $\vartheta^*_{7,CD} = [\vartheta_{51}; 0_{r_1}; 0_{r_1}; 0_{r_1}; \vartheta_{52}; 0_{r_2}; 0_{r_2}; 0_{r_2}]$). Also, the covariance matrix

$Q^*_{s,CD}$ will be equal to a diagonal block matrix

$$Q^*_{s,CD} = \begin{cases} \text{diag}(Q_{s1}, 0_{r_1 x r_1}, 0_{r_1 x r_1}, 0_{r_1 x r_1}, Q_{s2}, 0_{r_2 x r_2}, 0_{r_2 x r_2}, 0_{r_2 x r_2}), & s = 1, \cdots, 4 \\ 0_{4r x 4r}, & s = 5, 6 \end{cases} \tag{5.2.14}$$

Using the last specifications, it is noticed that $\alpha^*_{s,CD} = T^*_{s,CD}\alpha^*_{s-1,CD} + \vartheta^*_{s,CD}$ as required in the transition equation in Equation 5.2.5. It is also noticed that starting from the state vector at time 4, $\alpha^*_{4,CD} = \alpha^*_{5,CD} = \alpha^*_{6,CD}$ because the transition matrix has been defined as $T_s = I_{4r}$ and the vector $\alpha^*_{7,CD}$=z_5 is directly related to $\alpha^*_{4,CD}$=z_4 through the matrix $T^*_{s,CD}$ in Equation 5.2.12. Using Kalman filter and smoothing, the state vector $\alpha^*_{s,CD} = [\alpha^s_1; \alpha^s_2]'$ can be estimated and using the appropriate components in this vector, disaggregated values for the subsectors can be obtained. For instance, in this particular case, $\eta_{s1} = Z^1\alpha^*_{s,CD}$ and $\eta_{s2} = Z^2\alpha^*_{s,CD}$ with $Z^1 = [Z_1 \otimes [1, 0_3]; 0_{4r_2}] = [Z_1; 0_{r_1}; 0_{r_1}; 0_{r_1}; 0_{r_2}; 0_{r_2}; 0_{r_2}; 0_{r_2}]$ and $Z^2 = [0_{4r_1}; Z_2 \otimes [1, 0_3]] = [0_{r_1}; 0_{r_1}; 0_{r_1}; 0_{r_1}; Z_2; 0_{r_2}; 0_{r_2}; 0_{r_2}]$. Also, $\text{Var}(\hat{\eta}_{s1}) = Z^1 P^*_{s,CD}(Z^1)'$ and $\text{Var}(\hat{\eta}_{s2}) = Z^2 P^*_{s,CD}(Z^2)'$ with $P^*_{s,CD}$ obtained during the Kalman filter and smoother recursions. Then, the values with indexes s that are multiples of 5 or 6 are discarded in order to recover the indexation over t in the original series.

5.2.3 General Case

In this subsection we will develop the expressions of the vectors and matrices in the state space form for binding or non-binding estimation, for arbitrary number of variables and subperiods per year. In the same way as in section 5.2.2, $y^*_{s,CD}$ will denote the sth element in the series y^*_{CD} with $s = 1, \cdots, n + mP$. Equation 5.2.15 shows the relationship between the indexes s in $y^*_{s,CD}$ with the indexes t in z_t and also with the indexes j in $x_{.j(i)}$.

$$y^*_{s,CD} = \begin{cases} z_{s-P(i-1)}, & mod(s, K+P) = 1, \cdots, K \\ x_{.j(i)}, & mod(s, K+P) = K + j; \quad j = 1, \cdots, P \end{cases} \tag{5.2.15}$$

with $i = [s-1]_{K+P} + 1$, $j = mod(s, K+P) - K$ and $mod(a,b)$ denoting the number $a - b[a-1]_b$. On the other hand, $z_t = y^*_{t+Pi,CD}$ and $x_{.j(i)} = y^*_{(j-P)+(K+P)i,CD}$.

The idea behind Equation 5.2.15 is the arrangement of the contemporaneous totals as the K first values in the series y^*_{CD} for a given year and also, the annual totals as the P last values in the same year. This equation also permits to determine which kind of total (contemporaneous or annual) is connected with a particular index s, $s = 1, \cdots, n + mP$ and viceversa.

Now considering the P state space models with the irregular terms included in the state vector

$$\begin{cases} \eta_{t1} &= Z_{t1}\,\alpha_{t1} \\ \alpha_{t1} &= T_{t1}\,\alpha_{t-1,1} + \vartheta_{t1} \\ \vartheta_{t1} &\sim N(0, Q_{t1}) \end{cases} \quad \begin{cases} \eta_{t2} &= Z_{t2}\,\alpha_{t2} \\ \alpha_{t2} &= T_{t2}\,\alpha_{t-1,2} + \vartheta_{t2} \\ \vartheta_{t2} &\sim N(0, Q_{t2}) \end{cases} \quad \cdots \quad \begin{cases} \eta_{tP} &= Z_{tP}\,\alpha_{tP} \\ \alpha_{tP} &= T_{tP}\,\alpha_{t-1,P} + \vartheta_{tP} \\ \vartheta_{tP} &\sim N(0, Q_{tP}) \end{cases} \tag{5.2.16}$$

and if a non-binding total sector is considered, $z = \eta_{.} + \in_{.}$. Then, a state space model will be considered for the monthly survey error model in the total sector given by

$$\begin{cases} \in_{t.} &= Z_{t.}\,\alpha_{t.} \\ \alpha_{t.} &= T_{t.}\,\alpha_{t-1,.} + \vartheta_{t.} \\ \vartheta_{t.} &\sim N(0, Q_{t.}) \end{cases} \tag{5.2.17}$$

A formulation of the general state space model for the new series y^*_{CD} is given by the system of equations in Equation 5.2.5 by defining the system vectors and matrices below. Firstly, the state vector is built concatenating the P state vectors $\alpha^s_j = [\alpha_{js}; \cdots; \alpha_{j,s-K+1}]$ plus an extra term to account for the monthly survey errors in the total.

$$\alpha^*_{s,CD} = [\alpha^s_1; \alpha^s_2; \cdots; \alpha^s_P; \alpha_{s.}]' \tag{5.2.18}$$

Then, the dimension of the state vector is $(rK + \varrho_{.})$ with $r = \sum_{j=1}^{P} r_j$, r_j the dimension of the single state vector α^s_j and $\varrho_{.}$ being the dimension of the state vector $\alpha_{t.}$ in the model for survey error (for instance, $\varrho_{.} = \max(p, q+1)$ if an ARMA(p,q) model is considered

for $\in$.). In the following equations, we use the mathematical relationship between the indexes t from the original series to the new ones s built for contemporaneous disaggregation in Equation 5.2.15, $t = s - P(i-1)$, $i = [s-1]_{K+P} + 1$ and $j = mod(s, K+P) - K$. Then, the observation matrix can be expressed in a partitioned form as

$$Z^*_{s,CD} = \begin{cases} \left[[Z_{s-P(i-1),1}; \cdots ; Z_{s-P(i-1),P}] \otimes [1; 0'_{K-1}]; Z_{s-P(i-1),.}\right], \\ \qquad\qquad 1 \leq mod(s, K+P) \leq K, \\ \left[\delta_j \otimes [Z_{s-j,1}; \cdots ; Z_{s-j,P}]; 0_{\varrho.}\right], \\ \qquad\qquad mod(s, K+P) = K + j, \end{cases} \tag{5.2.19}$$

with δ_j being a 1 x P vector with 1 in the j-th position and zeros elsewhere. δ_j is called the Kronecker j-th delta vector with single elements $\delta_{jj'}$ given by

$$\delta_{jj'} = \begin{cases} 1 & j = j' \\ 0 & j \neq j' \end{cases} \quad j' = 1, \cdots, P \tag{5.2.20}$$

The dimension of the observation matrix Z^*_s is 1 x $(rK + \varrho.)$.

In order to include the annual survey errors in the annual subsector totals, we will consider the unobserved vector

$$\epsilon^*_{s,CD} = \begin{cases} 0, & 1 \leq mod(s, K+P) \leq K \\ e_{j(i)}, & mod(s, K+P) = K + j \end{cases} \tag{5.2.21}$$

Consequently, the covariance matrix of the disturbances $\epsilon^*_{s,CD}$ is given by the scalars (which are assumed as known from the annual survey in each subsector)

$$h^*_{s,CD} = \begin{cases} 0, & 1 \leq mod(s, K+P) \leq K \\ \sigma^2_{e_{j(i)}}, & mod(s, K+P) = K + j \end{cases} \tag{5.2.22}$$

In the same way it was assumed by Durbin and Quenneville (1997, page 38), we will consider $\Sigma_{e_{(i)}}$ as a diagonal matrix since otherwise the state vector becomes too large.

The transition matrix is a $rK + \varrho_.$ diagonal matrix of $P + 1$ matrices given by

$$T^*_{s,CD} = \begin{cases} \text{diag}\left(T^*_{s1,CD}, \cdots, T^*_{sP,CD}, T_{s-P(i-1),.}\right), & 1 \leq mod(s, K+P) \leq K \\ I_{rK}, & mod(s, K+P) > K \end{cases} \tag{5.2.23}$$

with

$$T^*_{sj,CD} = \begin{pmatrix} T_{s-P(i-1),j} & 0_{r_j x r_j} & 0_{r_j x r_j} & 0_{r_j x r_j} \\ I_{r_j} & 0_{r_j x r_j} & 0_{r_j x r_j} & 0_{r_j x r_j} \\ 0_{r_j x r_j} & I_{r_j} & 0_{r_j x r_j} & 0_{r_j x r_j} \\ 0_{r_j x r_j} & 0_{r_j x r_j} & I_{r_j} & 0_{r_j x r_j} \end{pmatrix} \quad j = 1, \cdots, P \tag{5.2.24}$$

and $T_{s-P(i-1),.}$ is the transition matrix associated with the series of sector totals. Finally, the formulation of the transition equation is complete by arranging the $(rK + \varrho_.)$ dimensional vector and its corresponding variance matrix given by

$$\vartheta^*_{s,CD} = \begin{cases} \left[[\vartheta_{s-P(i-1),1} \cdots \vartheta_{s-P(i-1),P}] \otimes [1; 0'_{(K-1)}]; \vartheta_{t.}\right], & 1 \leq mod(s, K+P) \leq K \\ 0_{rK}, & mod(s, K+P) > K \end{cases} \tag{5.2.25}$$

and

$$Q^*_{s,CD} = \begin{cases} \text{diag}\left(Q_{s-P(i-1),1}, 0_{Kr_1 x Kr_1}, \cdots, Q_{s-P(i-1),P}, 0_{Kr_1 x Kr_1}\right), & 1 \leq mod(s, K+P) \leq K \\ 0_{4rx4r}, & mod(s, K+P) > K \end{cases} \tag{5.2.26}$$

Equations 5.2.18 - 5.2.26 form the SSF required in Equation 5.2.5. The Kalman filter produces the estimates $\hat{y}^*_{s,CD}$ and $\hat{\alpha}^*_{s,CD}$ for every $s = 1, \cdots, n + mP$. Values of $\hat{\alpha}^*_{s,CD}$ are expected to coincide exactly with the observed values in the series $y^*_{s,CD}$ when $1 \leq mod(s, K+P) \leq K$. This is because for these time points, the observation equation was written with $\epsilon_s = 0$. Now, because the main interest is to estimate the disaggregated values η_j $(j = 1, \cdots, P)$, the procedure ends by obtaining the estimates $\hat{\eta}_{sj,CD} = Z^j \hat{\alpha}^*_{j,CD}$ $(j = 1, \cdots, P)$ with $Z^j = [0_{Kr_1}; 0_{Kr_2}; \cdots; Z_j \otimes [10_3]; \cdots; 0_{Kr_P}]$ and $\text{Var}(\hat{\eta}_{sj,CD}) = Z^j P^*_{s,CD} (Z^j)'$ with the matrix $P^*_{s,CD}$ obtained during the Kalman filter and smoothing recursions.

5.3 Simulation 1

In our first application, a random walk plus noise model (RWN) will be assumed for the series $\eta_j, j = 1, \cdots, P$ with no consideration of survey errors. The concept of a RWN model was explained in pages 31 and 36. This section reports the results of a simulation study in a very simple case, assuming $P = 2$ and $K = 4$. The experiment consists on generating one pair of series η_1 and η_2 from the RWN model defined by Equation 3.2.3

$$\begin{cases} \eta_{tj} = \mu_j + a_{tj} + \epsilon_{tj}, & \epsilon_{tj} \sim NID(0, \sigma_\epsilon^2) \\ a_{tj} = a_{t-1,j} + \nu_{tj}, & \nu_{tj} \sim NID(0, \sigma_\nu^2) \end{cases} \quad t = 1, \cdots, 250 \qquad (5.3.1)$$

with parameters $\mu_1 = 30$, $\mu_2 = 70$, $\sigma_\epsilon^2 = 3$ and $\sigma_\nu^2 = 0.5$. The graphs of the simulated processes in a single iteration appear in Figure 5.1. The corresponding vector of totals

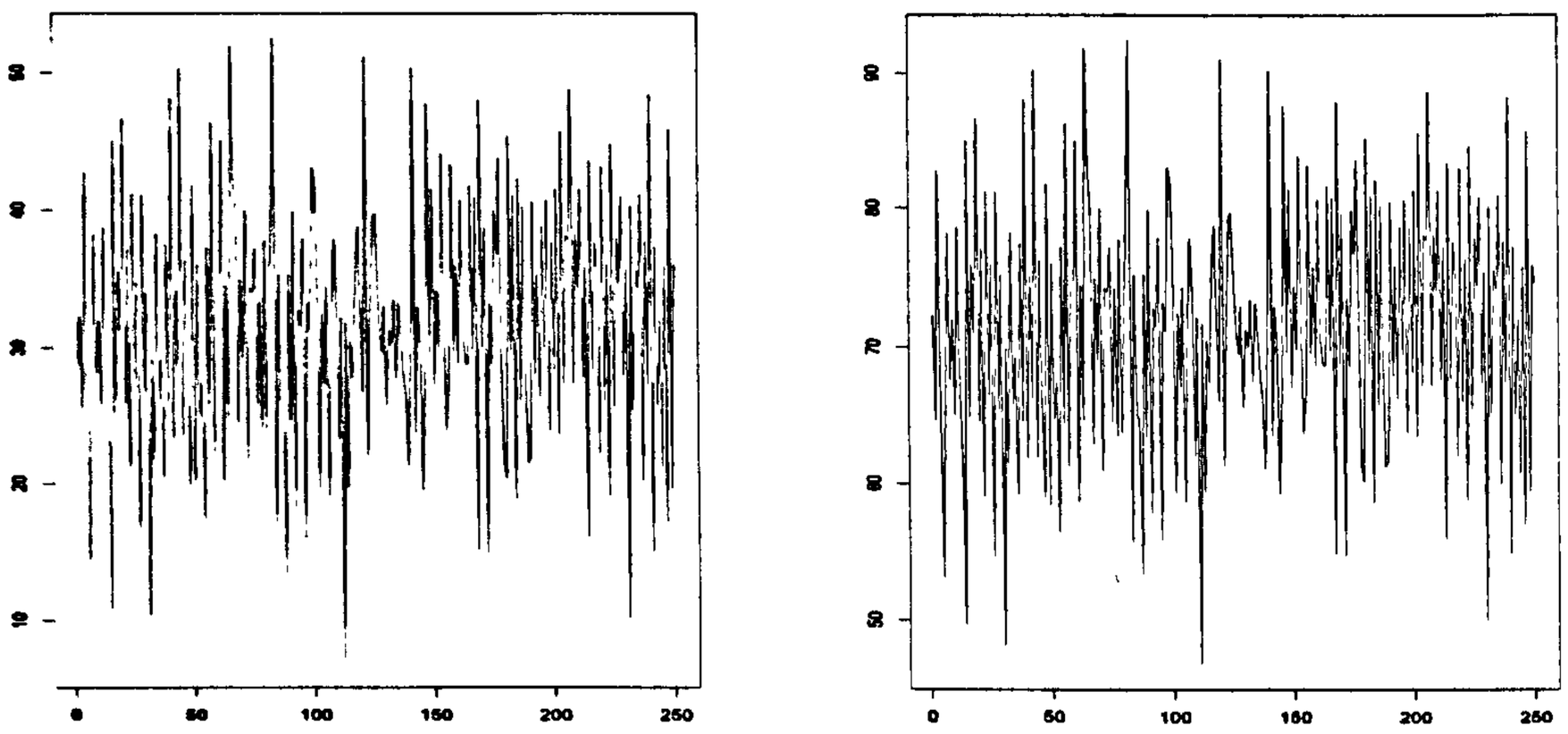


Figure 5.1. Pair of simulated RWN processes.

z of dimension 250 and the vectors x_1 and x_2 of dimension 62, $62 = [250]_4$ are built. Table 5.4 is the analoge of Table 5.1 for this special case in the first year. The idea is to the estimate the missing values in Table 5.5. In this case, we are considering no survey errors in the row and column subtotals. Using the state space form for a RWN in Equation 2.5.14, page 30; the proposed method was applied by building the series $y^*_{s,CD}$ $(s = 1, \cdots, n + mP = 374)$ and writing the observation and transition equations

Year	Period	Subsectors η_1	Subsectors η_2	$\eta_.$
1	1	26.87582	64.29220	91.16802
	2	24.33702	58.95520	83.29222
	3	19.00001	58.39089	77.39090
	4	18.43571	70.80347	89.23918
Total $x_{(i)}$		88.64856	252.44176	341.09032

Table 5.4. Simulated RWN values for the first year

Year	Period	Subsectors η_1	Subsectors η_2	$\eta_.$
1	1	NA	NA	91.16802
	2	NA	NA	83.29222
	3	NA	NA	77.39090
	4	NA	NA	89.23918
Total $x_{(i)}$		88.64856	252.44176	341.09032

Table 5.5. Initial values simulation RWN

using the vectors and matrices below.

$$\alpha^*_{s,CD} = [a_{s1}, \epsilon_{s1}; \cdots ; a_{s-3,1}, \epsilon_{s-3,1}; ; a_{s2}, \epsilon_{s2}; \cdots ; a_{s-3,2}, \epsilon_{s-3,2}] \tag{5.3.2}$$

$$Z^*_{s,CD} = \begin{cases} [11;00;00;00;11;00;00;00], & 1 \leq mod(s,6) \leq 4 \\ [11;11;11;11;00;00;00;00], & mod(s,6) = 5 \\ [00;00;00;00;11;11;11;11], & mod(s,6) = 6 \end{cases} \tag{5.3.3}$$

and

$$T^*_{s,CD} = \text{diag} \left(\begin{bmatrix} T & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ I_2 & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ 0_{2x2} & I_2 & 0_{2x2} & 0_{2x2} \\ 0_{2x2} & 0_{2x2} & I_2 & 0_{2x2} \end{bmatrix} , \begin{bmatrix} T & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ I_2 & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ 0_{2x2} & I_2 & 0_{2x2} & 0_{2x2} \\ 0_{2x2} & 0_{2x2} & I_2 & 0_{2x2} \end{bmatrix} \right)_{16 \text{ x } 16} \tag{5.3.4}$$

for $1 \leq mod(s,6) \leq 4$ where

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{5.3.5}$$

for $j = 1, 2$. $T^*_{s,CD} = I_{16}$ for $5 \leq mod(s,6) \leq 6$. The vector of disturbances ϑ^*_s is equal to

$$\vartheta^*_{s,CD} = \begin{cases} [\vartheta_{t1} 0_6 | \vartheta_{t2} 0_6]' = [\nu_{t1} \epsilon_{t1} 0_6 | \nu_{t2} \epsilon_{t2} 0_6]', & 1 \leq mod(s,6) \leq 4 \\ 0'_{16}, & 5 \leq mod(s,6) \leq 6 \end{cases} \tag{5.3.6}$$

with covariance matrix

$$Q^*_{s,CD} = \begin{cases} \text{diag}(Q_{t1}, 0_{6x6}, Q_{t2}, 0_{6x6}), & s = 1, \cdots, 4 \\ 0_{16x16}, & s = 5, 6 \end{cases} \tag{5.3.7}$$

with $t = s - P(i-1)$ and

$$Q_{t1} = Q_{t2} = \begin{bmatrix} 0.5 & 0 \\ 0 & 3 \end{bmatrix} \tag{5.3.8}$$

Equations 5.3.2. - 5.3.8. form the SSF required in Equation 5.2.5. The Kalman Filter was initialised using a diffuse prior. The Kalman filter produces the estimates $\hat{y}^*_{s,CD}$ and $\hat{\alpha}^*_{s,CD}$ for every $s = 1, ..., 374$. Values of $\hat{y}^*_{s,CD}$ were expected to coincide exactly with the values in the series $y^*_{s,CD}$ due to the fact that the observation equation was written with $\epsilon^*_s = 0$ in all cases. Now, because the main interest is to estimate the disaggregated values η_1 and η_2; these values can be obtained from the respective components for each subsector in $\hat{\alpha}^*_{s,CD}$ as follows

$$\hat{\eta}_{s1} = Z^1 \hat{\alpha}^*_{s,CD} = a_{s1} + \epsilon_{s1} \qquad \hat{\eta}_{s2} = Z^2 \hat{\alpha}^*_{s,CD} = a_{s2} + \epsilon_{s2} \tag{5.3.9}$$

with $Z^1 = [1,1] \otimes [1,0] \otimes [1, 0'_3] = [11|00|00|00||00|00|00|00]$ and $Z^2 = [1,1] \otimes [0,1] \otimes [1, 0'_3] = [00|00|00|00||11|00|00|00]$. The six first smoothed values for the first year of $\hat{\eta}^s_1$ and $\hat{\eta}^s_2$ were, in this simulation, $\hat{\eta}^s_1$ =[25.11428 21.18519 18.21688 24.13221 24.13221 24.13221] and $\hat{\eta}^s_2$=[66.05374 62.10703 59.17402 65.10697 65.10697 65.10697]. It can be noticed than, in the six points associated to the first year, the last two values are equal to the fourth one. This is because we are skipping the recursion in these two points corresponding to the annual totals using the matrix I_{16} as the transition matrix. Then, eliminating these values for each year in the series, we finally get the vector of disaggregated estimates $\hat{\eta}_1$ and $\hat{\eta}_2$. The information has been disaggregated

both contemporaneously and temporally, and the estimates have been obtained with their respective standard errors (see) and coefficients of variation (cv). The variances of the estimates were calculated by

$$Var(\hat{\eta}_{si}) = Z^i P^*_{s,CD} (Z^i)' \quad i = 1, 2 \tag{5.3.10}$$

where $P^*_{s,CD}$ is the covariance matrix obtained during the Kalman filter and smoother recursions. The simulation was carried out in different steps. Firstly, 1000 series from a RWN model were generated and secondly, the proposed method was applied in order to produce 1000 series of estimated values with total length of 250 values. Also, two alternative methods (IPF (Deming and Stephan, 1940) and (Zaier and Trabelsi, 2007)) were applied in order to be compared with the proposed method. Several statistics have been suggested in the literature for evaluating the performance of methods of adjusting contingency tables to known marginal totals. Upton (1985) and Wong (1992) have proposed what they called the absolute relative error (ARE) to compare matrices across different matrix and sample sizes. For a given year i ($i = 1, \cdots, m$); ARE_i corresponds to the total absolute error (TAE_i) divided by the number of cells in the corresponding year and TAE_i given by

$$TAE_i = \sum_{t \in i} \sum_{j=1}^{P} ; p_{tj} - q_{tj} ; \tag{5.3.11}$$

where p_{tj} and q_{tj} are the corresponding elements of the t-th row and the j-th column in the original matrix P_i and the estimated matrix Q_i. Then, TAE_i corresponds to the total deviation of the estimated matrix from the population matrix in a given year i. The ARE statistic is preferable when comparing matrices of different sizes (different number of instants or sectors) and they were the statistics finally considered to evaluate the methods used in this simulation. Notice that ARE is just an average of TAE values, which means the term "relative" in its acronym is used under this definition. Also, in this simulation, each table per year has eight inner cells in total with exception of the last year which has got four cells and it is used to illustrate the problem of ex-ante estimation.

5.3.1 Iterative Proportional Fitting - Results

The method known as iterative proportional fitting (IPF) or raking was proposed by Deming and Stephan (1940) as an iterative procedure to estimate the cells in a contingency table subject to some known marginal constraints. The main idea is to adjust a matrix of any dimension until the totals by rows and columns converge to some pre-defined values (Fienberg, 1970).

IPF is an iterative procedure whereby the original table values are gradually adjusted in several iterations to fit the row and column constraints. The final estimated contingency table after the iterations converge corresponds to the maximum likelihood estimates obtained when the values in the cells are convergent within an acceptable pre-defined limit (Bishop, Fienberg and Holland, 1975, pages 82-101). Given known row and column constraints, IPF can also be used to compute the maximum likelihood estimates of a two-dimensional matrix where the values are not known.

The iterations start by considering an a priori initial table. We will consider two possible initial tables in the contemporaneous disaggregation case with missing values. One possibility, assuming that the initial table values are constant (i.e. every cell is just the grand total divided by the number of cells) (Bishop et.al., 1975, denoted in the tables below as Raking 1) or assuming an initial table with statistical independence by row and columns (*prorata method*, denoted in the tables below as Raking 2); by this approach the estimates for every cell η_{tj} with $t = 1, \cdots, n$ and $j = 1, \cdots, P$ are computed as

$$\hat{\eta}_{tj} = \frac{\eta_{.j(i)}\eta_{t.}}{\eta_{..(i)}} \tag{5.3.12}$$

In summary, we have adapted the Deming and Stephan (1940) method to the contemporaneous disaggregation case with missing values as follows:

1. Consider a multidimensional table with initial values $\eta_{tj}^{(0)} = \frac{\eta_{..(i)}}{k*p}$ for every cell $tj; t = 1, \cdots, n; j = 1, \cdots, P$ in the contingency table associated to every year.

2. Adjust the initial values to the first marginal sub-totals (e.g. totals by columns) in each year to derive an estimate: $\eta_{tj}^{(1)r} = \eta_{tj}^{(0)} * \left(\frac{\eta_{.j(i)}}{\sum_{j=1}^{P} \eta_{tj}^{(0)}} \right)$ for every cell $tj(t = 1, \cdots, n; j = 1, \cdots, P$ in the contingency table associated to every year with the superindex r indicating the r-th adjustment by rows.

3. Repeat the adjustment to the marginal sub-total in the other dimension (e.g. totals by rows) to complete one cycle of $s = 1, \cdots, S$ steps $\eta_{tj}^{(1)} = \eta_{tj}^{(1)c} = \eta_{tj}^{(1)r} * \left(\frac{\eta_{t.}}{\sum_{K*(i-1)+1}^{K_i} \eta_{tj}^{(0)}} \right)$ with the superindex c indicating adjustment by columns.

4. In general at the s-th step, we have $\eta_{tj}^{(s)} = \eta_{tj}^{(s-1)} * \left(\frac{\eta_{.j(i)}}{\sum_{j=1}^{P} \eta_{tj}^{(0)}} \right) * \left(\frac{\eta_{t.}}{\sum_{K*(i-1)+1}^{K_i} \eta_{tj}^{(0)}} \right)$ for $s = 1, \cdots, S$.

5. Steps (1)-(4) are repeated until the factors $\left(\frac{\eta_{.j(i)}}{\sum_{j=1}^{P} \eta_{tj}^{(0)}} \right) \approx \left(\frac{\eta_{t.}}{\sum_{K*(i-1)+1}^{K_i} \eta_{tj}^{(0)}} \right) \approx 1$ under some convergence criterion. Since the procedure is proven to converge when the marginal sub-totals are consistent with each other (for example add to the same overall total), the choice of convergence criterion only affects the number of cycles that will be needed before the criterion is met.

Both alternatives Raking 1 and Raking 2 gave the same results. In other words, starting with a constant table in the inner cells makes the iterative procedure converge to a table with statistical independence by row and columns. Each cell is the product of the corresponding marginal totals divided by the grand total of the table. In this particular case, with 250 observations, it is impossible to get estimates of the inner cells for the last incomplete year as raking requires the availability of marginal totals by both row and columns. In conclusion, this alternative does not give a solution to the ex-ante estimation problem in Chapter 1.

Tables 5.6 (one iteration) and 5.8 (1000 iterations) show the actual simulated values and the corresponding estimates under this method with complete fulfilment of the restrictions by row and columns. The application of this method does not permit

to produce either estimates for the incomplete last year at the end of the series or standard error of the estimates.

The two first plots in Figure 5.2 show the estimated levels for each subsector through time. The graphs show how close are the means of the estimated (raking) values after 1000 iterations to the means of the simulated (original) values and also how the estimated values preserve the behaviour of the original series. Also, the difference between the actual and the estimated value for a given pair of month/subsector were calculated and plotted in the plot at the bottom showing values closer to zero for each subsector. The plot of the differences shows corresponding mirror images showing the strong dependence present in the disaggregated data.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	23.69435	67.47367
	2	24.33702	58.95520	83.29222	21.64745	61.64477
	3	19.00001	58.39089	77.39090	20.11371	57.27719
	4	18.43571	70.80347	89.23918	23.19305	66.04613
Total		88.64856	252.44176	341.09032	88.64856	252.44176
2	1	30.84828	64.79764	95.64592	29.41833	66.22759
	2	24.84245	70.78847	95.63092	29.41372	66.21720
	3	30.83328	77.11675	107.95003	33.20277	74.74726
	4	37.16156	65.74247	102.90403	31.65075	71.25328
Total		123.68557	278.44533	402.13090	123.68557	278.44533
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	24.47317	60.57627
	2	19.57317	79.39541	98.96858	28.47843	70.49015
	3	39.44022	75.16697	114.60719	32.97848	81.62811
	4	35.21179	82.30659	117.51838	33.81618	83.70220
Total		119.74626	296.39733	416.14359	119.74626	296.39733
63	1	42.35141	64.84352	107.19493	NA	NA
	2	24.88833	66.51744	91.40577	NA	NA

Table 5.6. Results of Contemporaneous Disaggregation with Missing Values. Raking Estimates. Single Iteration.

Statistic	Year 1	Year 2	Year 3	···	Year 62	Year 63 (incomplete)
TAE	23.48415	27.76307	16.62095	···	35.62101	NA
ARE	2.93552	3.47038	2.07762	···	4.45263	NA

Table 5.7. TAE and ARE. Raking Estimates. Contemporaneous Disaggregation with Missing Values. Single Iteration. Mean(ARE) = 5.36283.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.10970	70.57724	100.63694	30.12170	70.51524
	2	30.11886	70.53026	100.64912	30.12354	70.52558
	3	30.12508	70.51454	100.63962	30.10587	70.53375
	4	30.12646	70.53522	100.66168	30.12899	70.53269
Total		120.48010	282.10726	402.58736	120.48010	282.10726
2	1	30.14526	70.54183	100.68709	30.15291	70.53418
	2	30.14947	70.53614	100.68561	30.14891	70.53670
	3	30.15734	70.54471	100.70205	30.14236	70.55969
	4	30.15508	70.55313	100.70821	30.16297	70.54524
Total		120.60715	282.17581	402.78296	120.60715	282.17581
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.49690	71.67515
	2	30.49117	71.68520	102.17637	30.48423	71.69214
	3	30.49767	71.68915	102.18682	30.47857	71.70825
	4	30.49078	71.70624	102.19702	30.50204	71.69498
Total		121.96173	286.7053	408.73226	121.96173	286.7053
63	1	30.48295	71.72075	102.20370	NA	NA
	2	30.46696	71.72257	102.18953	NA	NA

Table 5.8. Results of Contemporaneous Disaggregation with Missing Values. Averages of Raking Estimates. 1000 Iterations.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	39.47833	39.52109	39.50098	⋯	39.45207	NA
ARE	4.93479	4.94014	4.93762	⋯	4.93151	NA

Table 5.9. TAE and ARE. Raking Estimates. Contemporaneous Disaggregation with Missing Values. 1000 Iterations. Mean(ARE)= 4.95142.

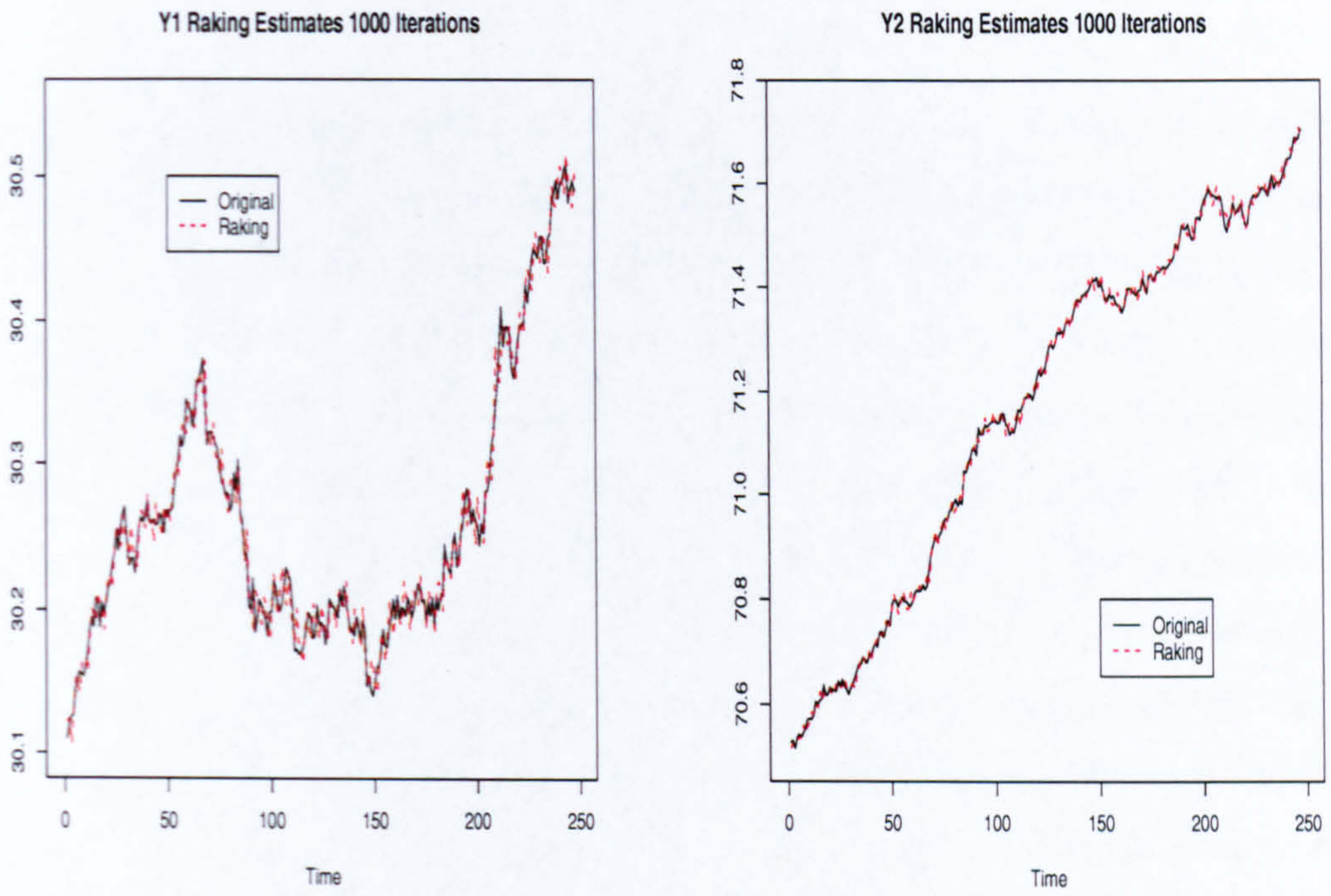


Figure 5.2. **Mean of the Disaggregated Raking Estimates for a RWN model in 1000 Iterations. Contemporaneous Disaggregation.**

5.3.2 Polynomial Interpolation Method - Results

Zaier and Trabelsi (2007) propose a procedure which extends the polynomial method for disaggregation of time series (Almon, 1988) to the multivariate case. The Almon's method provides a univariate polynomial approach to convert annual series to sub-annual figures by polynomial interpolation and it is implemented in the econometric package G. The method was proposed only for stock data. Then, the method was extended to the multivariate case for flow series by Zaier and Trabelsi (2007). They also show how to transform the flow data into stock data before applying the method. The basic idea of disaggregating the series is to obtain a smoothed curve between two consecutive years in each of the series by sector. This implies that, with this method, it is impossible to get estimates for the first year as it will require to have an initial and a final point for interpolation. Almon (1988) assumes a cubic polynomial to be fitted to each pair of successive years, Additionally, this method is only applicable when n is a multiple of m (e.g. $n = km$), which means that there is still a problem of ex-ante estimation and it is required to wait until the end of the year when the next vector of benchmarks is available to produce the estimates. The method does not produce standard errors of the estimated values either.

Zaier and Trabelsi (2007) method distinguishes three different cases and solutions:

1. $m = \frac{2P}{K-1} + 1$, which gives a unique solution.
2. $m < \frac{2P}{K-1} + 1$, which is not considered in their paper as it coincides with the method proposed by Almon (1988), considering arbitrary assumptions on the form of the interpolation curve and
3. $m > \frac{2P}{K-1} + 1$, being the most common case. In this particular case, the polynomial method fulfills exactly the temporal aggregation but the contemporaneous aggregation constraints are just approximated. In this case, it is necessary to adjust the polynomial estimates in a further step, using the Hillmer and Trabelsi (1987) method which was extended to the multivariate case by Trabelsi and Hillmer (1990).

Tables 5.10 shows the results for the first two and the last two years when applying the polynomial method in order to estimate the inner cells of the simulated contingency tables according to the procedure described in the last subsection. The two columns η_1 and η_2 correspond to the actual values generated from the RWN model and the column $\eta_.$ corresponds to the row totals. The estimated values correspond to the columns $\hat{\eta}_1$ and $\hat{\eta}_2$. It is seen that the estimated values add perfectly to the row totals. The annual totals only coincide after using the multivariate extension of the Hillmer and Trabelsi method (Trabelsi and Hillmer (1990)) according to the third solution presented above. One big disadvantage of this method in this simulation experiment is that it did not produce estimates for the inner cells corresponding to the first and the last year. The estimated values in Table 5.10 coincide perfectly in both, rows and columns per year, with complete fulfilment of the restrictions (shaded columns and rows). Because the impossibility of producing standard error of the estimates, we will consider ARE and TAE statistics to compare with other available methods and the proposed method in this chapter.

The same process was repeated in 1000 iterations. Table 5.12 shows the average of these results after repeating the process 1000 times. The simulated values do not converge to the mean of the process (30 and 70 respectively) at the beginning of the series but they do converge at the end of the series. The application of this method does not permit to produce either estimates for the initial or the estimates for the incomplete last year at the end of the series nor standard errors of the estimates. Figure 5.3 shows how far are the means of the 1000 estimated values at the beginning of the series from the original simulated series (with values outside the bounds of the graph) and how they stochastically converge to the simulated values after around seven years of observations. The difference between the actual and the estimated value for a given pair of month/subsector were calculated and plotted in the plot at the bottom showing values closer to zero for each subsector. Differences between actual and estimated values are close to zero but again being closer only at the end of the series. The plot of the differences shows corresponding mirror images showing the strong dependence present in the disaggregated data.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	NA	NA
	2	24.33702	58.95520	83.29222	NA	NA
	3	19.00001	58.39089	77.39090	NA	NA
	4	18.43571	70.80347	89.23918	NA	NA
Total		88.64856	252.44176	341.09032	NA	NA
2	1	30.84828	64.79764	95.64592	36.90419	58.74173
	2	24.84245	70.78847	95.63092	24.73131	70.89961
	3	30.83328	77.11675	107.95003	28.54735	79.40268
	4	37.16156	65.74247	102.90403	33.50272	69.40131
Total		123.68557	278.44533	402.13090	123.68557	278.44533
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	21.14424	63.90520
	2	19.57317	79.39541	98.96858	27.55509	71.41349
	3	39.44022	75.16697	114.60719	34.94785	79.65934
	4	35.21179	82.30659	117.51838	36.09908	81.41930
Total		119.74626	296.39733	416.14359	119.74626	296.39733
63	1	42.35141	64.84352	107.19493	NA	NA
	2	24.88833	66.51744	91.40577	NA	NA

Table 5.10. Results of Contemporaneous Disaggregation with Missing Values. Polynomial Estimates. Single Iteration.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	NA	24.22366	74.17148	⋯	35.47681	NA
ARE	NA	3.027957	9.27144	⋯	4.434601	NA

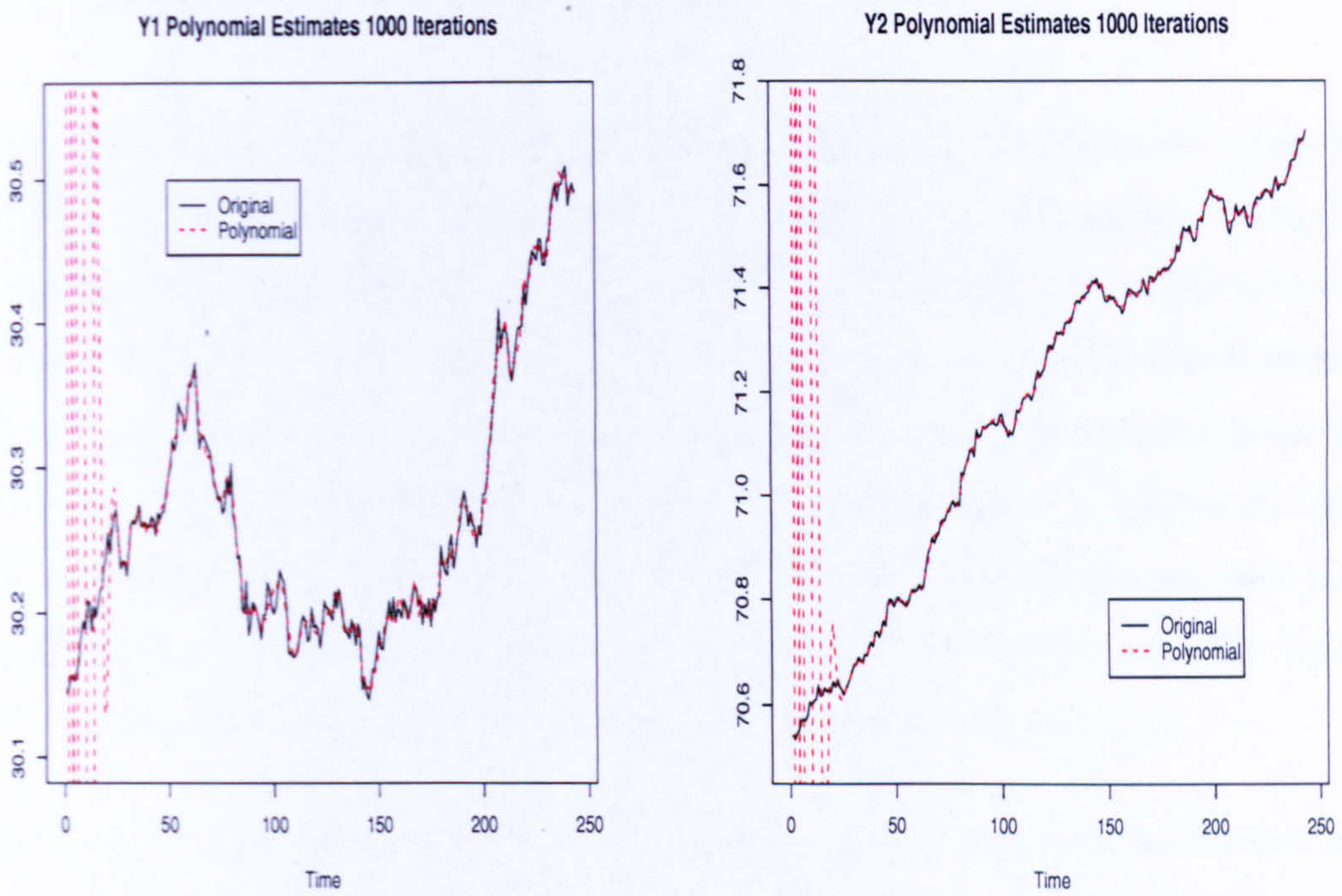
Table 5.11. TAE and ARE. Polynomial Estimates. Contemporaneous Disaggregation with Missing Values. Single Iteration. Mean(ARE) = 5.46884.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.10970	70.52724	100.63694	NA	NA
	2	30.11886	70.53026	100.64912	NA	NA
	3	30.12508	70.51454	100.63962	NA	NA
	4	30.12646	70.53522	100.66168	NA	NA
Total		120.48010	282.10726	402.58736	120.57014	282.1072
2	1	30.14526	70.54183	100.68709	40.05056	60.63653
	2	30.14947	70.53614	100.68561	25.94233	74.74328
	3	30.15734	70.54471	100.70205	23.10120	77.60085
	4	30.15508	70.55313	100.70821	31.51306	69.19515
Total		120.60715	282.17581	402.78296	120.60715	282.17581
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.49248	71.67957
	2	30.49117	71.68520	102.17637	30.48856	71.68780
	3	30.49767	71.68915	102.18682	30.48928	71.69755
	4	30.49078	71.70624	102.19702	30.49142	71.70560
Total		121.96173	286.7053	408.73226	121.96173	286.7053
63	1	30.48295	71.72075	102.20370	NA	NA
	2	30.46696	71.72257	102.18953	NA	NA

Table 5.12. Results of Contemporaneous Disaggregation with Missing Values. Polynomial Estimates. Average in 1000 Iterations.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	NA	57.70716	70.30178	⋯	38.42159	NA
ARE	NA	7.21340	8.78772	⋯	4.80270	NA

Table 5.13. TAE and ARE. Polynomial Estimates. Contemporaneous Disaggregation with Missing Values. 1000 Iterations. Mean(ARE) = 4.88312.



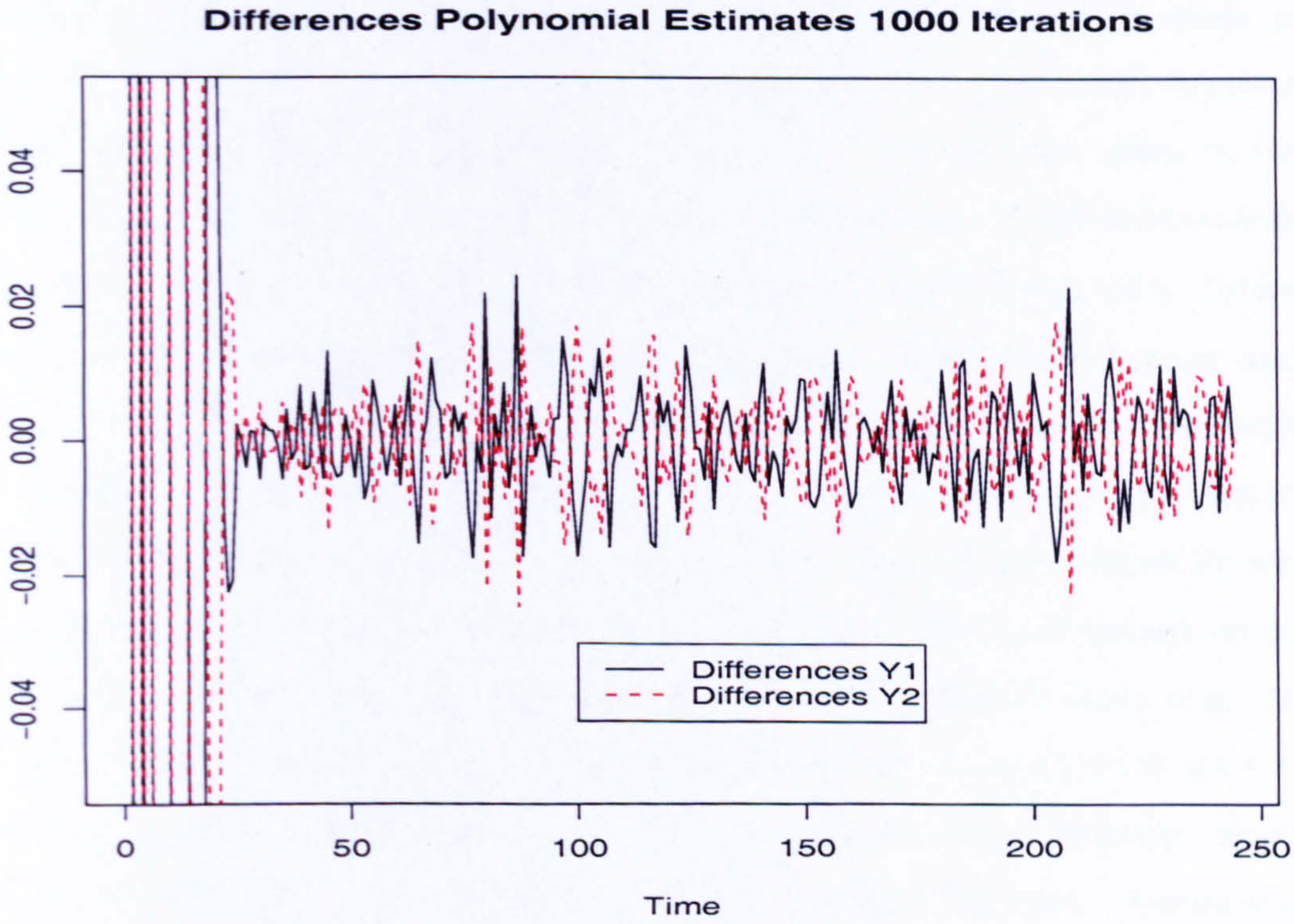


Figure 5.3. **Mean of the Disaggregated Polynomial Estimates for a RWN model in 1000 Iterations (First and Last Year are not Considered). Contemporaneous Disaggregation.**

5.3.3 Proposed Method - Results

Tables 5.14 and 5.17 show the results of one single iteration for the two first and the two last years (the 63th year was simulated as an incomplete year on purpose). Table 5.14 corresponds to the application of the Kalman filter and Table 5.17 correspond to the smoothed estimates in Equation 3.4.11 for the state space model defined in Equations 5.3.2 to 5.3.8. In both tables, the columns η_1 and η_2 are the actual values generated according to the RWN model and the column $\eta_.$ is the row totals. The estimated values are represented by the columns $\hat{\eta}_1$ and $\hat{\eta}_2$. From the results, the estimated values add perfectly to the row totals in both the filtered and the smoothed cases. The annual totals only coincide with the smoothed ones but not in the filtering case.

Additionally, in the filtered case for the first year (Table 5.14), both subsectors have the same estimate. The sum by rows in both sides of the table (actual values and estimates) is equal to the shaded column in the middle. The estimated values in Table 5.17 coincide perfectly in both rows and columns every year, with complete fulfilment of the restrictions (shaded columns and rows). The last two values in the series correspond to the two first quarters in year 63, where the annual benchmarks have not been observed yet and the filtered and smoothed values are the same. Tables 5.15 and 5.18 show reduction in terms of the magnitude of the standard errors and coefficients of variation when they are compared from one year to the next one with the exception of the last two years. According to the magnitude of the coefficients of variation, filtered values are not useful in the first year of observation (possibly not useful since the third one). On the other hand, the values for the standard errors are relatively small and stable in the Table 5.18 for all the periods of study with the exception perhaps of the last incomplete year. Standard errors are exactly the same in both subsectors given that the values were simulated with the same variances. Since the means are different, the cvs for $\hat{\eta}_2$ are smaller than $\hat{\eta}_1$. The same process was repeated 1000 times. Tables 5.20 and 5.22 show the average of these results over all the iterations. The means of the simulated values converge to the means of the process (30 and 70 respectively) and we obtain the same conclusions as obtained for the single iteration case.

Figures 5.4 and 5.5 show how close are the means of the estimated values after 1000 iterations to the means of the simulated values and also how the estimated values preserve the behaviour of the original series. Figure 5.4 show the results for the filtered estimates and Figure 5.5 for the smoothed ones. The difference between the actual and the estimated value for a given pair of month/subsector were calculated and plotted in the plot at the bottom showing values closer to zero for each subsector. Differences between actual and estimated values are close to zero in both figures, with smaller differences for the smoothed estimates. Again, the plot of the differences shows corresponding mirror images showing the strong dependence present in the disaggregated data. The proposed method has some advantages over the previous methods above. The main advantage is the possibility to obtain estimates of the initial and the final year (ex-ante estimation) with their corresponding standard errors. However, the method makes a big assumption. The application of the method requires prior knowledge about the stochastic structure of each of the subsectors.

5.3.4 Comparison of the Methods

Table 5.24 and Figure 5.6 compare the ARE values used to evaluate the performance of the different methods applied in this simulation. It was mentioned before that the ARE values measure the average distance between the original table and the estimated one. Figure 5.6 shows the results for different number of simulated series (1, 250 and 1000). In these three cases (polynomial method in red, raking in blue and the proposed method in black), the polynomial method appears to be the less precise, particularly at the beginning of the series. Values for the first and the last year are missing; the same happens for the last year with the raking method. There is an increase in the series in black (estimates under the proposed method) in the last year as this year is incomplete, affecting the quality of the estimation. However, the two other methods do not calculate any estimates for this last year and the proposed method looks as the most efficient one based on the ARE measure.

Year	Period	Actual Values η_1	η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	$\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	45.58401	45.58401
	2	24.33702	58.95520	83.29222	41.64611	41.64611
	3	19.00001	58.39089	77.39090	38.69545	38.69545
	4	18.43571	70.80347	89.23918	44.61958	44.61958
Total		88.64856	252.44176	341.09032	170.54416	170.54416
2	1	30.84828	64.79764	81.63651	27.34885	68.29706
	2	24.84245	70.78847	95.63092	27.34135	68.28956
	3	30.83328	77.11675	107.95003	33.50091	74.44912
	4	37.16156	65.74247	102.90403	30.97791	71.92612
Total		123.68557	278.44533	402.13090	119.16903	282.96187
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	22.64258	62.40686
	2	19.57317	79.39541	98.96858	29.60216	69.36643
	3	39.44022	75.16697	114.60719	37.42146	77.18573
	4	35.21179	82.30659	117.51838	38.87705	78.64132
Total		119.74626	296.39733	416.14359	128.54326	287.60033
63	1	42.35141	64.84352	107.19493	32.89890	74.29603
	2	24.88833	66.51744	91.40577	25.00432	66.40145

Table 5.14. Results of Contemporaneous Disaggregation with Missing Values. Filtered Estimates. Single Iteration.

Year	Period	se $(\hat{\eta}_1)$	se $(\hat{\eta}_2)$	cv $(\hat{\eta}_1)$	cv $(\hat{\eta}_2)$
	1	707.1101	707.1101	14.0137	14.0137
	2	707.1101	707.1101	14.5806	14.5806
1	3	707.1102	707.1102	13.1620	13.1620
	4	707.1103	707.1103	12.7075	12.7075
	1	2.4077	2.4077	0.1154	0.0396
	2	2.4335	2.4335	0.1259	0.0411
2	3	2.4590	2.4590	0.0771	0.0342
	4	2.4843	2.4843	0.0753	0.0341
⋮	⋮	⋮	⋮	⋮	⋮
	1	2.2538	2.2538	0.0793	0.0330
	2	2.2814	2.2814	0.0861	0.0344
62	3	2.3086	2.3086	0.0734	0.0324
	4	2.3355	2.3355	0.0715	0.0322
63	1	2.2538	2.2538	0.0810	0.0333
	2	2.2814	2.2814	0.0645	0.0303

Table 5.15. Standard Errors and Coefficients of Variation. Filtered Estimates. Contemporaneous Disaggregation with Missing Values. Single Iteration

Statistic	Year 1	Year 2	Year 3	···	Year 62	Year 63 (incomplete)
TAE	163.79320	29.69924	16.81079	···	37.18299	19.13699
ARE	20.47415	3.71240	2.10135	···	4.64787	4.78425

Table 5.16. TAE and ARE. Filtered Estimates. Contemporaneous Disaggregation with Missing Values. Single Iteration. Mean(ARE) = 5.57298.

Year	Period	Actual Values η_1	η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	$\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	25.11428	66.05374
	2	24.33702	58.95520	83.29222	21.18519	62.10703
	3	19.00001	58.39089	77.39090	18.21688	59.17402
	4	18.43571	70.80347	89.23918	24.13221	65.10697
Total		88.64856	252.44176	341.09032	88.64856	252.44176
2	1	30.84828	64.79764	95.64592	28.44144	67.20448
	2	24.84245	70.78847	95.63092	28.44903	67.18189
	3	30.83328	77.11675	107.95003	34.64896	73.80107
	4	37.16156	65.74247	102.90403	32.14614	70.75789
Total		123.68557	278.44533	402.13090	123.68557	278.44533
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	20.57594	64.47350
	2	19.57317	79.39541	98.96858	27.42185	71.54673
	3	39.44022	75.16697	114.60719	35.16538	79.44181
	4	35.21179	82.30659	117.51838	36.58309	80.93529
Total		119.74626	296.39733	416.14359	119.74626	296.39733
63	1	42.35141	64.84352	107.19493	32.89890	74.29603
	2	24.88833	66.51744	91.40577	25.00432	66.40145

Table 5.17. Results of Contemporaneous Disaggregation with Missing Values. Smoothed Estimates. Single Iteration.

Year	Period	se $(\hat{\eta}_1)$	se $(\hat{\eta}_2)$	cv $(\hat{\eta}_1)$	cv $(\hat{\eta}_2)$
1	1	1.8667	1.8667	0.0612	0.0265
	2	1.8496	1.8496	0.0648	0.0270
	3	1.8500	1.8500	0.0548	0.0251
	4	1.8664	1.8664	0.0523	0.0247
2	1	1.8651	1.8651	0.0890	0.0307
	2	1.8495	1.8495	0.0952	0.0312
	3	1.8496	1.8496	0.0578	0.0258
	4	1.8660	1.8660	0.0563	0.0256
⋮	⋮	⋮	⋮	⋮	⋮
62	1	1.8649	1.8649	0.0653	0.0274
	2	1.8498	1.8498	0.0694	0.0279
	3	1.8495	1.8495	0.0585	0.0260
	4	1.8657	1.8657	0.0568	0.0258
63	1	2.2538	2.2538	0.0810	0.0333
	2	2.2814	2.2814	0.0645	0.0333

Table 5.18. Standard Errors and Coefficients of Variation. Smoothed Estimates. Contemporaneous Disaggregation with Missing Values. Single Iteration

Statistic	Year 1	Year 2	Year 3	···	Year 62	Year 63 (incomplete)
TAE	22.78600	29.68905	15.09291	···	36.87990	19.13699
ARE	2.84825	3.71113	1.88661	···	4.60999	4.78425

Table 5.19. TAE and ARE. Smoothed Estimates. Contemporaneous Disaggregation with Missing Values. Single Iteration. Mean(ARE) = 5.14010.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.10970	70.52724	100.63694	50.31847	50.31487
	2	30.11886	70.53026	100.64912	50.32456	50.32456
	3	30.12508	70.51454	100.63962	50.31981	50.31981
	4	30.12646	70.53522	100.66168	50.33084	50.33084
Total		120.48011	282.1072	402.58736	201.29368	201.29368
2	1	30.14526	70.54183	100.68709	30.14020	70.54689
	2	30.14947	70.53614	100.68561	30.13946	70.54615
	3	30.15734	70.54471	100.70205	30.14768	70.55437
	4	30.15508	70.55313	100.70821	30.15076	70.55745
Total		120.60715	282.17581	402.78296	120.57810	282.20486
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.51350	71.65855
	2	30.49117	71.68520	102.17637	30.51566	71.66071
	3	30.49767	71.68915	102.18682	30.52089	71.66594
	4	30.49078	71.70624	102.19702	30.52599	71.67103
Total		121.96173	286.7053	408.73226	122.07604	286.65623
63	1	30.48295	71.72075	102.20370	30.51872	71.68498
	2	30.46696	71.72257	102.18953	30.51163	71.67790

Table 5.20. Results of Contemporaneous Disaggregation with Missing Values. Filtered Estimates. Average Values in 1000 Iterations.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	161.62714	47.37955	44.70552	⋯	43.57636	21.81447
ARE	20.20339	5.92244	5.58819	⋯	5.44704	5.45362

Table 5.21. TAE and ARE. Filtered Estimates. Contemporaneous Disaggregation with Missing Values. Average Values in 1000 Iterations. Mean(ARE) = 5.68202.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.10970	70.52724	100.63694	30.11510	70.52184
	2	30.11886	70.53026	100.64912	30.12120	70.52792
	3	30.12508	70.51454	100.63962	30.11640	70.52322
	4	30.12646	70.53522	100.66168	30.12740	70.53428
Total		120.48010	282.10726	402.58736	120.48010	282.10726
2	1	30.14526	70.54183	100.68709	30.14730	70.53979
	2	30.14947	70.53614	100.68561	30.14650	70.53911
	3	30.15734	70.54471	100.70205	30.15505	70.54700
	4	30.15508	70.55313	100.70821	30.15830	70.54991
Total		120.60715	282.17581	402.78296	120.60715	282.17581
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.48665	71.68540
	2	30.49117	71.68520	102.17637	30.48733	71.68904
	3	30.49767	71.68915	102.18682	30.49157	71.69525
	4	30.49078	71.70624	102.19702	30.49618	71.70084
Total		121.96173	286.77053	408.73226	121.96173	286.77053
63	1	30.48295	71.72075	102.20370	30.51872	71.68498
	2	30.46696	71.72257	102.18953	30.51163	71.67790

Table 5.22. **Results of Contemporaneous Disaggregation with Missing Values. Smoothed Estimates. Average Values in 1000 Iterations.**

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	37.03708	37.10982	37.13806	⋯	37.34501	21.81447
ARE	4.62696	4.63873	4.64226	⋯	4.66813	5.45362

Table 5.23. **TAE and ARE. Smoothed Estimates. Contemporaneous Disaggregation. Average Values in 1000 Iterations. Mean(ARE) = 4.66956.**

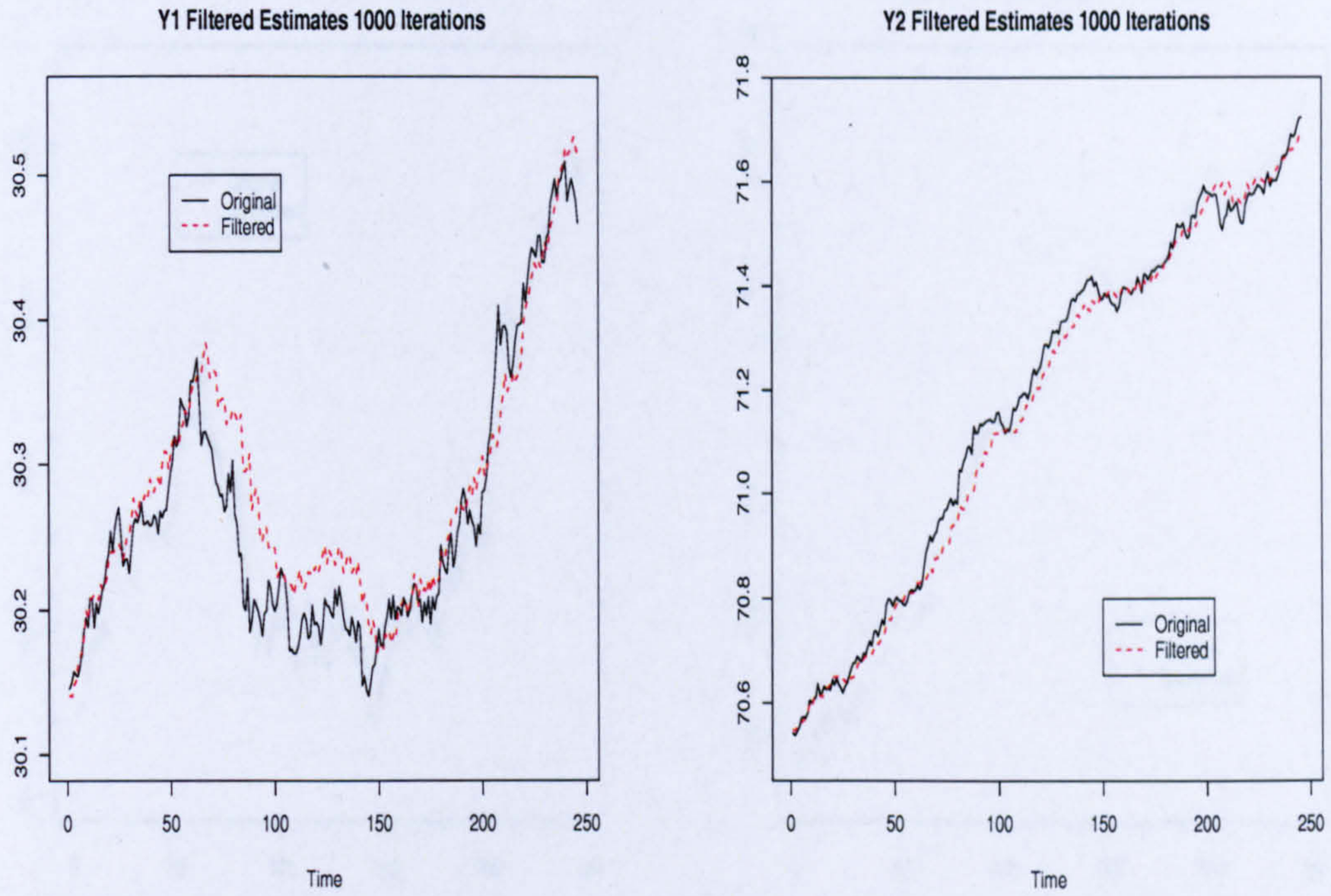


Figure 5.4. **Mean of the Disaggregated Filtered Estimates for a RWN model in 1000 Iterations (First Year was not Considered). Contemporaneous Disaggregation.**

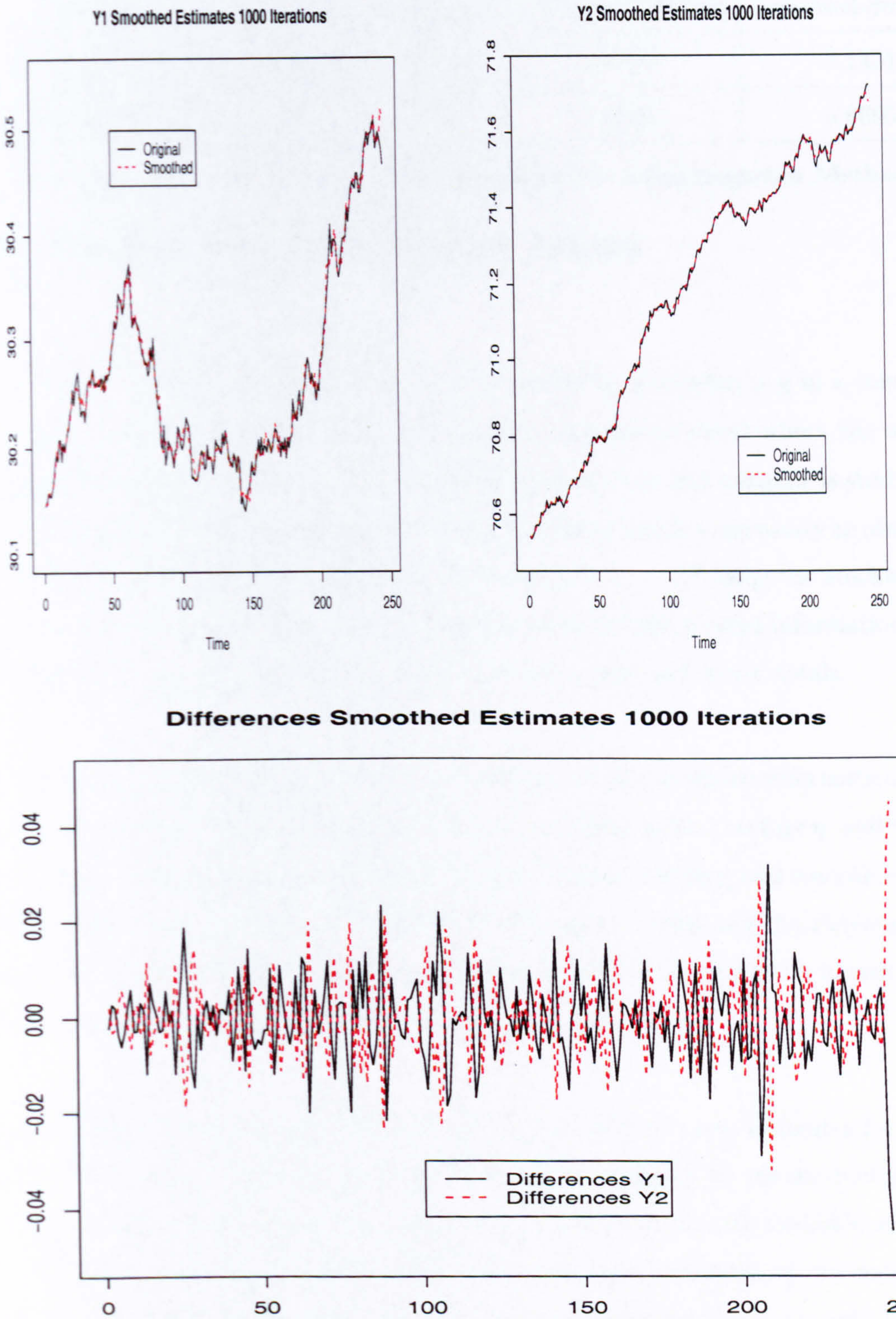


Figure 5.5. **Mean of the Disaggregated Smoothed Estimates for a RWN model in 1000 Iterations. Contemporaneous Disaggregation with Missing Values.**

Method	Raking	Polynomial	Proposed-Filtered	Proposed-Smoothed
ARE - Single Iteration	5.36283	5.46884	5.57298	5.14010
ARE - 1000 Iterations	4.95142	4.88312	5.68202	4.66956

Table 5.24. Average of ARE Values. Contemporaneous Disaggregation Methods.

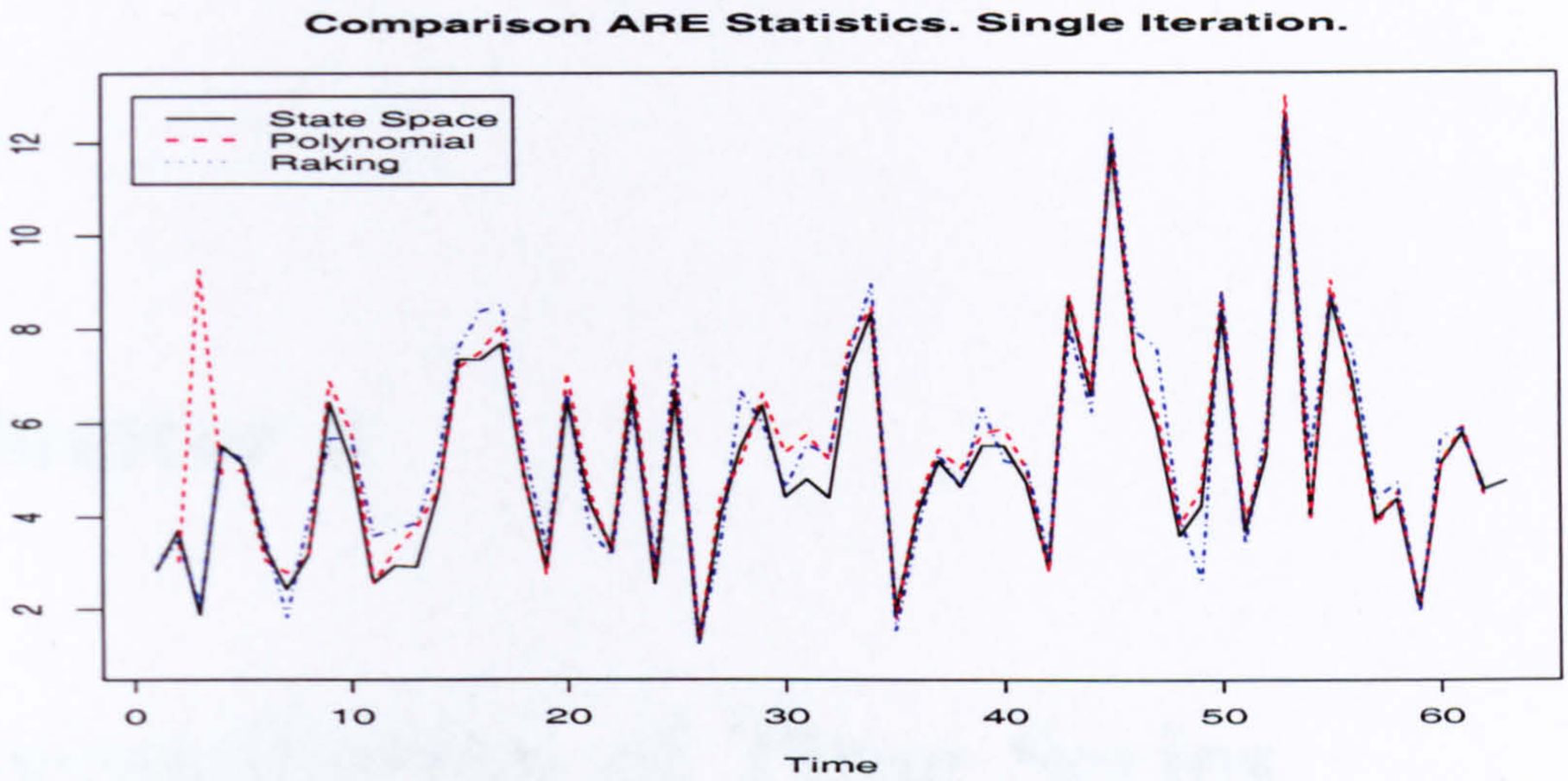
5.4 Conclusions and Further Issues

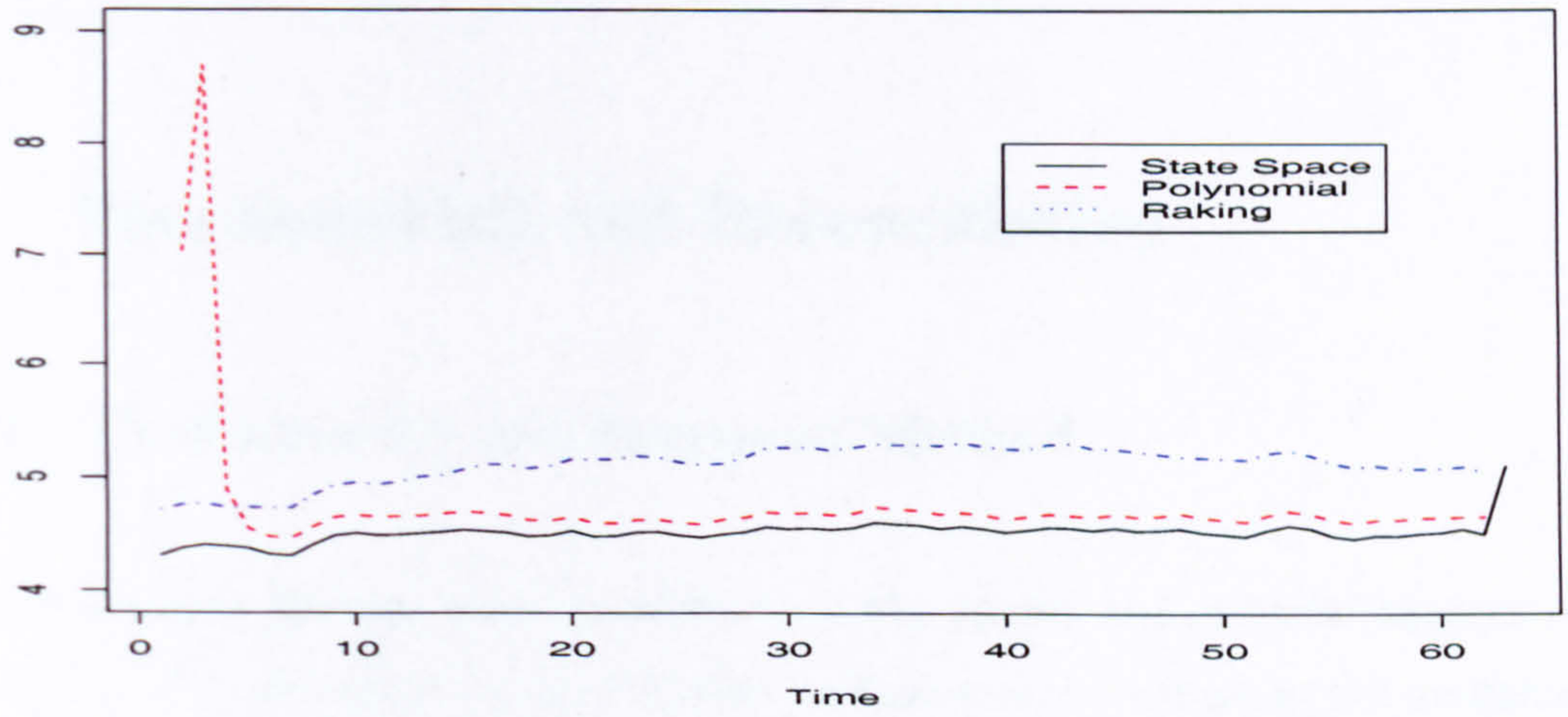
The problem of how to produce tabular data classified by attributes (e.g in a contingency table with attributes as columns and points in time as rows) where the only available information are the vectors of marginal totals by row and columns is studied. The high frequency information inside the tables is missing and it is necessary to obtain estimates of the elements in the separate vectors η_j, $j = 1, \cdots, P$ using the stochastic properties of these series. The aim is to combine monthly and annual information to get estimates of the missing values and more precise annual and sector totals.

The method proposed corresponds to the modelling of the available information by state space structural time series models. All the marginal totals (vectors $\eta_{.}$ and $\eta_{(i)}$) are arranged into a special single series and then, Kalman filtering and smoothing is applied. Other two methods: IPF (Deming and Stephan, 1940) and the Polynomial method (Zaier and Trabelsi, 2007) were applied to a RWN simulated series in order to compare these methods with the proposed method in this chapter.

The advantages of the Proposed Method are: a) the method gives estimates for the initial year (this is not possible under the Polynomial method); b) the method gives a solution to the ex-ante estimation (when there is not a benchmark available at the end of the series, which is a problem under the other two methods); c) the method produces estimates of the standard errors of the estimates (which is not possible under the other two alternatives). However, it assumes knowledge of the stochastic structure of the series involved.

Since it is not possible to calculate standard errors in all the methods considered, we quantified the quality of the estimation under two possible measures of precision: TAE and ARE (Equation 5.3.11). When comparing the averages over 1000 simulated series, the polynomial method appears to be less precise, particularly at the beginning of the series. On the other hand, the proposed method seems to be the most efficient based on the ARE measure. Other measures of precision as those considered in Equations E.2.1 to E.2.4 (Fair, 1984; Pindyck and Rubinfeld, 1991) in the Appendix E could be considered for further experiments.





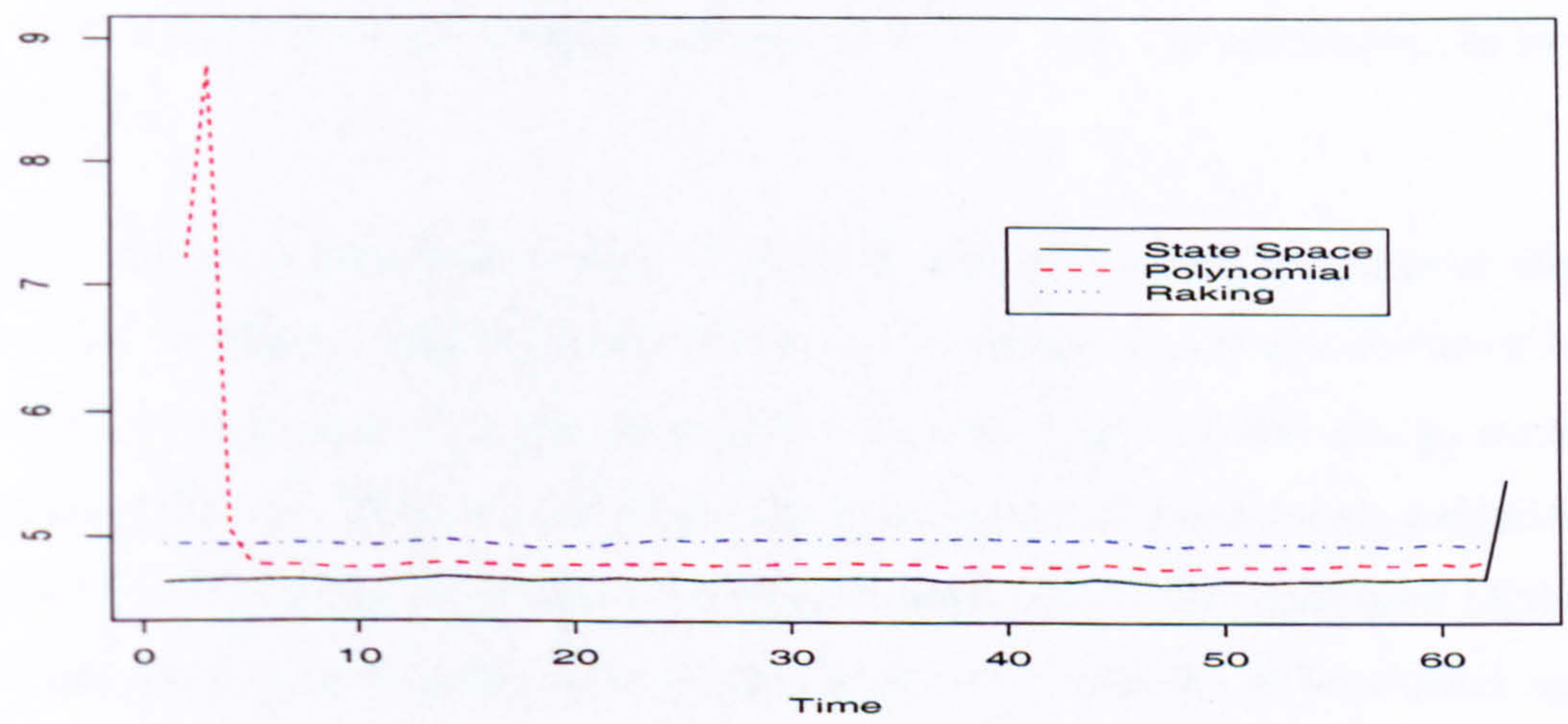


Figure 5.6. **Plots of ARE Values. Contemporaneous Disaggregation.**

Chapter 6

Reconciliation of Time Series

6.1 Benchmarking and Reconciliation

6.1.1 Preliminaries and Proposed Method

We now consider the case when in addition to the annual and sectorial aggregated information, P preliminary vectors of survey estimates of the vectors $\boldsymbol{\eta}_j$ are available, $j = 1, \cdots, P$. Denoting these vectors by $\boldsymbol{y}_j$ of dimension n; the problem is that $\sum_{j=1}^{P} \boldsymbol{y}_j \neq \boldsymbol{\eta}_.$ and $\boldsymbol{y}_j$ does not comply with $\boldsymbol{\eta}_{(i)} \quad i = 1, \cdots, m$, the vectors of stacked annual totals.

In many cases, the population vector of totals $\boldsymbol{\eta}_.$ and $\boldsymbol{\eta}_{(i)}$ cannot be observed and only survey estimates $\boldsymbol{z}$ and $\boldsymbol{x}_{(i)}$, respectively, can be obtained as it was discussed in equation 5.2.1; but again it is not necessarily true that $\sum_{j=1}^{P} \boldsymbol{y}_j = \boldsymbol{z}$ and also $\boldsymbol{y}_j$ could not comply with $\boldsymbol{x}_{(i)}$. Then, it is necessary adjust or correct all these survey estimates in order to arrive at useful, consistent and publishable values with fulfilment of the constraints by row and columns. Also, using a more broad definition of benchmarking; we can use the information contained in the more reliable totals by row and columns to get more precise disaggregated estimates in the inner cells.

Year	Period	Subsectors y_1	$\cdots$	y_j	$\cdots$	y_P	z
	1	$y_{K(i-1)+1,1}$	$\cdots$	$y_{K(i-1)+1,j}$	$\cdots$	$y_{K(i-1)+1,P}$	$z_{K(i-1)+1}$
	2	$y_{K(i-1)+1,1}$	$\cdots$	$y_{K(i-1)+1,j}$	$\cdots$	$y_{K(i-1)+1,P}$	$z_{K(i-1)+2}$
i	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	K	$y_{K(i-1)+1,1}$	$\cdots$	$y_{K(i-1)+1,j}$	$\cdots$	$y_{K(i-1)+1,P}$	z_{Ki}
Total $x_{(i)}$		$x_{1(i)}$	$\cdots$	$x_{j(i)}$	$\cdots$	$x_{P(i)}$	

Table 6.1. **Reconciliation problem for year i, i=1, $\cdots$, m.**

The initial configuration is similar to Table 5.1 with the difference that instead of missing values, we now have survey estimates without fulfilment of the row and column totals. Table 6.1 shows the initial configuration for the year i ($i = 1, \cdots, m$). The aim is to restore the additivity for each one of these tables with more precise estimates, combining all the information available from the monthly and annual estimates, and estimate the respective monthly and annual standard errors (or any measure of precision in these surveys).

According to Equation 5.2.1, it follows that

$$\begin{aligned} z &= \eta_{.} + \epsilon \\ x_{(i)} &= \eta_{(i)} + e_{(i)} \\ y_j &= \eta_j + \ell_j \end{aligned} \tag{6.1.1}$$

for $i = 1, \cdots, P$ and $j = 1, \cdots, m$ with y_j a vector of dimension n x 1, containing the preliminary estimates for each cell for each variable in every month and ℓ_j denoting the corresponding vectors of monthly survey errors. Notice how, compared with the equations 5.2.1 for the contemporaneous disaggregation case, now it is necessary to include a third equation to capture the preliminary information.

Without consideration of the stacked values, all the available information can be organized in a matrix of dimension n x $(P+1)$, superposing tables that are similar to Table 6.1 for the other years. If for each instant $t = 1, \cdots, n$, we consider the vector $y_t = (y_{t1}, \cdots, y_{tP}, z_t)'$ of dimension $(P+1)$ and also consider the vector of annual

totals $\boldsymbol{x}_{.(i)} = (x_{.1(i)}, \cdots, x_{.P(i)})'$ of dimension P for every year $i = 1, \cdots, m$; the series $\boldsymbol{y}^i = (y_{K(i-1)+1}; \cdots; y_{Ki}; \boldsymbol{x}_{.(i)})$ is constructed juxtaposing the individual series with monthly indexes related to the year $i = 1, \cdots, m$.

$$\boldsymbol{y}^i = \left(\begin{pmatrix} y_{K(i-1)+1,1} \\ \cdots \\ y_{K(i-1)+1,P} \\ z_{K(i-1)+1} \end{pmatrix}, \begin{pmatrix} y_{K(i-1)+2,1} \\ \cdots \\ y_{K(i-1)+2,P} \\ z_{K(i-1)+2} \end{pmatrix}, \cdots \begin{pmatrix} y_{Ki,1} \\ \cdots \\ y_{Ki,P} \\ z_{K} \end{pmatrix}, \begin{pmatrix} x_{.1(i)} \\ \cdots \\ x_{.P(i)} \end{pmatrix} \right) \tag{6.1.2}$$

Compared with the disaggregation case in the last chapter, now the series to be studied is no longer univariate. Now each instant will represent one row in Table 6.1. considered as a column vector of dimension $(P+1)$ x 1 with exception of the last component which has dimension P x 1. This difference in dimensions of the observations in some instants can be handled with the use of appropriate transition matrices. The length of the this multivariate series is $K + 1$. Finally, we produce the whole series

$$\boldsymbol{y}^* = \begin{cases} [\boldsymbol{y}^1; \cdots; \boldsymbol{y}^m], & \text{if n is a multiple of m} \\ [\boldsymbol{y}^1; \cdots; \boldsymbol{y}^m; y_{Km+1}, \cdots, y_n], & \text{otherwise} \end{cases} \tag{6.1.3}$$

of length $n+m$, with each single element being a vector of dimension $(P+1)$ or P.The proposed method consists in applying standard Kalman filtering and smoothing to the new series $\boldsymbol{y}^*$. This series has all the contemporaneous and temporal estimates by row and columns per year in Table 5.1. It is necessary to express $\boldsymbol{y}^*$ in the state space form $y_s^* = \boldsymbol{Z}_s^* \boldsymbol{\alpha}_s^* + \varepsilon_s^*$, $\boldsymbol{\alpha}_s^* = \boldsymbol{T}_s^* \boldsymbol{\alpha}_{s-1}^* + \boldsymbol{\vartheta}_s^*$. A brief illustration of the state space model for reconciliation is presented below for the case of binding totals with $P = 2$ and $K = 4$.

6.1.2 Binding Estimation for Quarterly Data - Bivariate Case

The state space form for a simplified case is presented in this subsection as an illustration. It consists of a system or contingency table with just two variables ($P = 2$) and data collected by quarters ($K = 4$). The table for the first year appears in Table 6.2. We will consider a binding case here where the marginal totals are not subject to survey error. It means that the only processes including survey errors are the auxiliary

information for the subsectors as obtained from a quarterly survey. In this particular case, the single series y^* is built with the available information as follows

$$y^* = [y_s^*] = \left(\begin{bmatrix} y_{11} \\ y_{12} \\ z_1 \end{bmatrix} \quad \begin{bmatrix} y_{21} \\ y_{22} \\ z_2 \end{bmatrix} \quad \begin{bmatrix} y_{31} \\ y_{32} \\ z_3 \end{bmatrix} \quad \begin{bmatrix} y_{41} \\ y_{42} \\ z_4 \end{bmatrix} \quad \begin{bmatrix} x_{.1(1)} \\ x_{.2(1)} \end{bmatrix} \quad \begin{bmatrix} y_{51} \\ y_{52} \\ z_5 \end{bmatrix} \quad , \cdots \right) \tag{6.1.4}$$

Year	Period	y_1	y_2	z
1	1	y_{11}	y_{12}	z_1
	2	y_{21}	y_{22}	z_2
	3	y_{31}	y_{32}	z_3
	4	y_{41}	y_{42}	z_4
Total $x_{(1)}$		$x_{.1(1)}$	$x_{.2(1)}$	$x_{..(1)}$

Table 6.2. Tabular representation of the reconciliation data for the first year, P =2 and K=4.

The length of the new series y^* is $n + m$ with n defining the total number of observations in the vector z and m the number of complete years. Every component of y^* has dimension 3 x 1 with exception of the components in multiples of 5 which have dimension 2 x 1. Another alternative is to consider the vector $[x_{.1(1)}, x_{.2(1)}, x_{..(1)}]$, instead of just $[x_{.1(1)}, x_{.2(1)}]$, in the positions that are multiple of 5. However, this alternative gives correlated disturbances in the observation equation with the last element being redundant. Now, we will proceed differently than we did in the contemporaneous disaggregation case by considering the state space models for y_1 and y_2 considered as auxiliary information for η_1 and η_2. As usual, $y_1 = \eta_1 + \ell_1$ and $y_2 = \eta_2 + \ell_2$. Then, we will consider different state space models for η_1, η_2, ℓ_1 and ℓ_2, with all the models including the irregular terms in the state vector.

$$\begin{cases} \eta_{t1} & = Z_{t1}\,\alpha_{t1} \\ \alpha_{t1} & = T_{t1}\,\alpha_{t-1,1} + \vartheta_{t,1} \\ \vartheta_{t1} & \sim N(0, Q_{t1}) \end{cases} \qquad \begin{cases} \ell_{t1} & = \tilde{Z}_{t1}\,\tilde{\alpha}_{t1} \\ \tilde{\alpha}_{t1} & = \tilde{T}_{t1}\,\tilde{\alpha}_{t-1,1} + \tilde{\vartheta}_{t1} \\ \tilde{\vartheta_{t1}} & \sim N(0, \tilde{Q}_{t1}) \end{cases}$$

$$\begin{cases} \eta_{t2} &= Z_{t2}\,\alpha_{t2} \\ \alpha_{t2} &= T_{t2}\,\alpha_{t-1,2} + \vartheta_{t,2} \\ \vartheta_{t2} &\sim N(0, Q_{t2}) \end{cases} \qquad \begin{cases} \ell_{t2} &= \tilde{Z}_{t2}\,\tilde{\alpha}_{t2} \\ \tilde{\alpha}_{t2} &= \tilde{T}_{t2}\,\tilde{\alpha}_{t-1,2} + \tilde{\vartheta}_{t2} \\ \tilde{\vartheta}_{t2} &\sim N(0, \tilde{Q}_{t2}) \end{cases} \tag{6.1.5}$$

Although the terms ℓ_{t1} and ℓ_{t2} could be included in the state vector in the models for η_{t1} and η_{t2}, it is better to consider their models separately in order to reduce the size of the state vector for the reconciliation model, as will be explained later on. The values for the single series y^* in the first year are given by

$$\begin{aligned} y_1^* = \begin{bmatrix} y_{11} \\ y_{12} \\ z_1 \end{bmatrix} &= \begin{bmatrix} Z_{11}\alpha_{11} + \tilde{Z}_{11}\tilde{\alpha}_{11} \\ Z_{12}\alpha_{12} + \tilde{Z}_{12}\tilde{\alpha_{12}} \\ Z_{11}\alpha_{11} + Z_{12}\alpha_{12} \end{bmatrix} \\ y_2^* = \begin{bmatrix} y_{21} \\ y_{22} \\ z_2 \end{bmatrix} &= \begin{bmatrix} Z_{21}\alpha_{21} + \tilde{Z}_{21}\tilde{\alpha}_{21} \\ Z_{22}\alpha_{22} + \tilde{Z}_{22}\tilde{\alpha}_{22} \\ Z_{21}\alpha_{21} + Z_{22}\alpha_{22} \end{bmatrix} \\ y_3^* = \begin{bmatrix} y_{31} \\ y_{32} \\ z_3 \end{bmatrix} &= \begin{bmatrix} Z_{31}\alpha_{31} + \tilde{Z}_{31}\tilde{\alpha}_{31} \\ Z_{32}\alpha_{32} + \tilde{Z}_{32}\tilde{\alpha}_{32} \\ Z_{31}\alpha_{31} + Z_{32}\alpha_{32} \end{bmatrix} \\ y_4^* = \begin{bmatrix} y_{41} \\ y_{42} \\ z_4 \end{bmatrix} &= \begin{bmatrix} Z_{41}\alpha_{41} + \tilde{Z}_{41}\tilde{\alpha}_{41} \\ Z_{42}\alpha_{42} + \tilde{Z}_{42}\tilde{\alpha}_{42} \\ Z_{41}\alpha_{41} + Z_{42}\alpha_{42} \end{bmatrix} \\ y_5^* = \begin{bmatrix} x_{.1(i)} \\ x_{.2(i)} \end{bmatrix} &= \begin{bmatrix} Z_{11}\alpha_{11} + Z_{21}\alpha_{21} + Z_{31}\alpha_{31} + Z_{41}\alpha_{41} \\ Z_{12}\alpha_{12} + Z_{22}\alpha_{22} + Z_{32}\alpha_{32} + Z_{42}\alpha_{42} \end{bmatrix} \end{aligned} \tag{6.1.6}$$

and so on for the following years.

As before, let r_j denotes the number of components of the state vector α_{sj} with $s = 1, \cdots, n+m$ and $j = 1, 2$. Let also $\alpha_j^s = [\alpha_{sj}, \cdots, \alpha_{s-3,j}]$ be a vector of dimension $4r_j$; the state vector α_s^* in the observation and transition equations can be written as the concatenation of the individual vectors associated with each column. In this particular case, the dimension of this new state vector α_s^* is equal to $(4r_1 + 4r_2 + \varrho_1 + \varrho_2)$ x 1 or

$(4r + \varrho)$ x 1 with $r = r_1 + r_2$ and $\varrho = \varrho_1 + \varrho_2$.

$$\boldsymbol{\alpha}_s^* = [\boldsymbol{\alpha}_1^s; \boldsymbol{\alpha}_2^s; \tilde{\boldsymbol{\alpha}}_{s1}; \tilde{\boldsymbol{\alpha}}_{s2}]' \tag{6.1.7}$$

$$= [\boldsymbol{\alpha}_{s1}; \cdots; \boldsymbol{\alpha}_{s-3,1}; \boldsymbol{\alpha}_{s2}; \cdots; \boldsymbol{\alpha}_{s-3,2}; \tilde{\boldsymbol{\alpha}}_{s1}; \tilde{\boldsymbol{\alpha}}_{s2}]' \tag{6.1.8}$$

$$= [\boldsymbol{\alpha}_{s,CD}^*; \tilde{\boldsymbol{\alpha}}_{s1}; \tilde{\boldsymbol{\alpha}}_{s2}]' \tag{6.1.9}$$

Notice that this is the same state vector used in the contemporaneous disaggregation case in the last section, with the addition of the survey error terms in each subsector. Another option would be to include the survey errors ℓ_{sj} in the state vector $\boldsymbol{\alpha}_{sj}$ but this would make $\boldsymbol{\alpha}_s^*$ to have a bigger dimension with some unnecessary terms.

The observation or system matrices Z_s^* for the first year will be equal to

$$Z_s^* = \begin{cases} \begin{bmatrix} Z_{s1} & 0_{r_1} & 0_{r_1} & 0_{r_1} & 0_{r_2} & 0_{r_2} & 0_{r_2} & 0_{r_2} & \tilde{Z}_{s1} & 0_{\varrho 2} \\ 0_{r_1} & 0_{r_1} & 0_{r_1} & 0_{r_1} & Z_{s2} & 0_{r_2} & 0_{r_2} & 0_{r_2} & 0_{\varrho 1} & \tilde{Z}_{s2} \\ Z_{s1} & 0_{r_1} & 0_{r_1} & 0_{r_1} & Z_{s2} & 0_{r_2} & 0_{r_2} & 0_{r_2} & 0_{\varrho 1} & 0_{\varrho 2} \end{bmatrix}, \\ \qquad\qquad mod(s,5) \leq 4, \\ \begin{bmatrix} Z_{s-1,1} & Z_{s-2,1} & Z_{s-3,1} & Z_{s-4,1} & 0_{r_2} & 0_{r_2} & 0_{r_2} & 0_{r_2} & 0_{\varrho 1} & 0_{\varrho 2} \\ 0_{r_1} & 0_{r_1} & 0_{r_1} & 0_{r_1} & Z_{s-1,2} & Z_{s-2,2} & Z_{s-3,2} & Z_{s-4,2} & 0_{\varrho 1} & 0_{\varrho 2} \end{bmatrix}, \\ \qquad\qquad mod(s,5) = 5, \end{cases} \tag{6.1.10}$$

Considering now the transition equation, this takes the form $\boldsymbol{\alpha}_s^* = T_s^* \boldsymbol{\alpha}_{s-1}^* + \boldsymbol{\vartheta}_s^*$. In the same way as it was done for the contemporaneous disaggregation case, this set of equations can be obtained by skipping the Kalman filter updating equations for the SSF of $\boldsymbol{y}^*$ at the points related to the values $x_{.1(i)}$ and $x_{.2(i)}$ with $i = 1, \cdots, m$. This is done in order to preserve the stochastic nature of the vector of sector totals $\boldsymbol{z}$. In this way, the state vector of $\boldsymbol{y}_6^*$ is related to the state vector of $\boldsymbol{y}_4^*$ but not with the state vector of $\boldsymbol{y}_5^*$. Then, for the first year of observations, it follows that

$$T_s^* = \begin{cases} \text{diag}(T_{s,CD}^*, \tilde{T}_{s1}, \tilde{T}_{s2}), & 1 \leq mod(s,5) \leq 4 \\ I_{4r+\varrho}, & mod(s,5) = 5 \end{cases} \tag{6.1.11}$$

Also,

$$\vartheta_s^* = \begin{cases} [\vartheta_{s,CD}^*, \tilde{\vartheta}_{s1}, \tilde{\vartheta}_{s2}], & 1 \leq mod(s,5) \leq 4 \\ 0_{4r+\varrho}, & mod(s,5) = 5 \end{cases} \tag{6.1.12}$$

and

$$Q_s^* = \begin{cases} \text{diag}(Q_{s,CD}^*, \tilde{Q}_{s1}, \tilde{Q}_{s2}), & 1 \leq mod(s,5) \leq 4 \\ 0_{(4r+\varrho)\text{x}(4r+\varrho)}, & mod(s,5) = 5 \end{cases} \tag{6.1.13}$$

6.1.3 General Case

In this subsection we develop the expressions for the vectors and matrices in the state space form for binding and non-binding estimation, any number of variables and sub-periods of observation per year. In the same way was done in section 2.5.5. and 5.3.1, y_s^* will denote each single element in the series y^* with $s = 1, \cdots, n+m$. This time, in a different way that was done for the contemporaneous disaggregation case, y_s^* will be a vector of observations with dimension $(P+1)$ x 1 or P x 1 according to

$$y_s^* = \begin{cases} [y_{t1}; \cdots; y_{tP}; z_t]', & mod(s, K+1) = 1, \cdots, K \\ [x_{.1(i)}; \cdots; x_{.P(i)}]', & mod(s, K+1) = K+1; \end{cases} \tag{6.1.14}$$

with $i = [s-1]_{K+1} + 1$ and $mod(a,b)$ denoting the number $a - b[a-1]_b$. On the other hand,

$$[y_{t1}; \cdots; y_{tP}; z_t]' = y_{t+(i-1)}^* \text{ and } [x_{.1(i)}; \cdots; x_{.P(i)}]' = y_{i(K+1)}^* \tag{6.1.15}$$

with $t = s - (i-1)$. The idea behind Equation 6.1.14 is to consider the vectors of annual totals as stacked values at the end of each year.

Now considering the P state space models with the irregular terms included in the state vector

$$\begin{cases} \eta_{t1} = Z_{t1}\,\alpha_{t1} \\ \alpha_{t1} = T_{t1}\,\alpha_{t-1,1} + \vartheta_{t1} \\ \vartheta_{t1} \sim N(0, Q_{t1}) \end{cases} \begin{cases} \eta_{t2} = Z_{t2}\,\alpha_{t2} \\ \alpha_{t2} = T_{t2}\,\alpha_{t-1,2} + \vartheta_{t2} \\ \vartheta_{t2} \sim N(0, Q_{t2}) \end{cases} \cdots \begin{cases} \eta_{tP} = Z_{tP}\,\alpha_{tP} \\ \alpha_{tP} = T_{tP}\,\alpha_{t-1,P} + \vartheta_{tP} \\ \vartheta_{tP} \sim N(0, Q_{tP}) \end{cases} \tag{6.1.16}$$

and the corresponding P state space models for their associated survey errors,

$$\begin{cases} \ell_{t1} &= \tilde{Z}_{t1}\,\tilde{\alpha}_{t1} \\ \tilde{\alpha}_{t1} &= \tilde{T}_{t1}\,\tilde{\alpha}_{t-1,1} + \tilde{\vartheta}_{t1} \\ \tilde{\vartheta}_{t1} &\sim N(0, \tilde{Q}_{t1}) \end{cases} \begin{cases} \ell_{t2} &= \tilde{Z}_{t2}\,\tilde{\alpha}_{t2} \\ \tilde{\alpha}_{t2} &= \tilde{T}_{t2}\,\tilde{\alpha}_{t-1,2} + \tilde{\vartheta}_{t2} \\ \tilde{\vartheta}_{t2} &\sim N(0, \tilde{Q}_{t2}) \end{cases} \cdots \begin{cases} \ell_{tP} &= \tilde{Z}_{tP}\,\tilde{\alpha}_{tP} \\ \tilde{\alpha}_{tP} &= \tilde{T}_{tP}\,\tilde{\alpha}_{t-1,P} + \tilde{\vartheta}_{tP} \\ \tilde{\vartheta}_{tP} &\sim N(0, \tilde{Q}_{tP}) \end{cases} \tag{6.1.17}$$

If a non-binding total sector is considered, $z = \eta_{\cdot} + \epsilon_{\cdot}$, a state space model will be considered for the monthly survey error model in the total sector given by

$$\begin{cases} \epsilon_{t\cdot} &= Z_{t\cdot}\,\alpha_{t\cdot} \\ \alpha_{t\cdot} &= T_{t\cdot}\,\alpha_{t-1,\cdot} + \vartheta_{t\cdot} \\ \vartheta_{t\cdot} &\sim N(0, Q_{t\cdot}) \end{cases} \tag{6.1.18}$$

A formulation of the general state space model for the new series y^* is given by the system of equations in Equation 5.2.5 by defining the system vectors and matrices below. Firstly, the state vector is built concatenating the P state vectors $\alpha_j^s = [\alpha_{js}; \cdots; \alpha_{j,s-K+1}]$ plus some extra terms to consider the monthly survey errors in each of the subsectors and the total sector.

$$\alpha_s^* = [\alpha_1^s; \alpha_2^s; \cdots; \alpha_P^s; \alpha_{s\cdot}; \tilde{\alpha}_{s1}; \tilde{\alpha}_{s2}; \cdots; \tilde{\alpha}_{sP}]' \tag{6.1.19}$$

$$= [\alpha_{s,CD}^*; \tilde{\alpha}_{s1}; \tilde{\alpha}_{s2}; \cdots; \tilde{\alpha}_{sP}]' \tag{6.1.20}$$

Then, the length of the state vector is $rK + (\varrho + \varrho_{\cdot})$ with $r = \sum_{j=1}^{P} r_j$, r_j the dimension of the single state vector α_j^s, $\varrho = \sum_{j=1}^{P} \varrho_j$, ϱ_j being the dimension of the state vector $\tilde{\alpha}_j^s$ and $\varrho_{\cdot}$ being the dimension of the state vector $\alpha_{t\cdot}$ in the monthly survey error model.

In the following equations, we use the mathematical relationship between the indexes t from the original series to the new ones s built for contemporaneous disaggregation in Equation 5.2.15, $t = s-(i-1)$ and $i = [s-1]_{K+1}+1$. Then, the observation matrix

can be expressed by

$$Z_s^* = \begin{cases} \begin{bmatrix} Z_{t1} & 0 & \cdots & 0 & 0 & 0 & \tilde{Z}_{t1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & Z_{tP} & 0 & 0 & 0 & \cdots & \tilde{Z}_{tP} \\ Z_{t1} & 0 & \cdots & Z_{tP} & 0 & Z_{t.} & 0 & \cdots & 0 \end{bmatrix}, \\ \qquad mod(s, K=1) \leq K, \\ \begin{bmatrix} Z_{s-1,1} & Z_{s-2,1} & Z_{s-3,1} & Z_{s-4,1} & \cdots & 0 & 0 & 0 & 0 & 0_{(\varrho.+\varrho)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & Z_{s-1,P} & Z_{s-2,P} & Z_{s-3,P} & Z_{s-4,P} & 0_{(\varrho.+\varrho)} \end{bmatrix}, \\ \qquad mod(s, K+1) = K+1, \end{cases} \tag{6.1.21}$$

The dimension of the observation matrix Z_s^* is $P+1$ x $(Kr+\varrho+\varrho.)$ if $1 \leq mod(s, K+1) \leq K$ or P x $(Kr+\varrho+\varrho.)$ if $mod(s, K+1) = K+1$.

In order to include the annual survey errors in the annual subsector totals, we will consider the unobserved vector

$$\epsilon_s^* = \begin{cases} 0, & 1 \leq mod(s, K+1) \leq K \\ [e_{1(i)}; \cdots; e_{P(i)}]', & mod(s, K+1) = K+1 \end{cases} \tag{6.1.22}$$

with $i = [s-1]_{K+P} + 1$ and $j = mod(s, K+P) - K$. Consequently, the covariance matrix of the disturbances ϵ_s^* is given by the scalars (which are assumed as known from the annual survey in each subsector)

$$h_s^* = \begin{cases} 0, & 1 \leq mod(s, K+1) \leq K \\ \Sigma_{e_{(i)}}, & mod(s, K+1) = K+1 \end{cases} \tag{6.1.23}$$

with $i = [s-1]_{K+P}+1$ and $j = mod(s, K+P)-K$ according to the index equivalences in Equation 5.2.15. In the same way it was assumed by Durbin and Quenneville (1997, page 38), we will consider $\Sigma_{e_{(i)}}$ as a diagonal matrix since otherwise the state vector becomes too large.

The transition matrix is a diagonal matrix of $P+1$ matrices given by

$$T_s^* = \begin{cases} \text{diag}(T^*_{s,CD}; \tilde{T}_{s-P(i-1),1}; \cdots; \tilde{T}_{s-P(i-1),P}), & 1 \leq mod(s, K+1) \leq K \\ I_{rK}, & mod(s, K+1) = K+1 \end{cases} \tag{6.1.24}$$

Finally, the formulation of the transition equation is complete by arranging the vector and variance matrix of the disturbances

$$\vartheta_s^* = \begin{cases} [\vartheta^*_{s,CD}, \tilde{\vartheta}_{s-P(i-1),1}, \cdots, \tilde{\vartheta}_{s-P(i-1),P}], & 1 \leq mod(s, K+1) \leq K \\ 0_{4r+\varrho}, & mod(s, K+1) = K+1 \end{cases} \tag{6.1.25}$$

and

$$Q_s^* = \begin{cases} \text{diag}(Q^*_{s,CD}, \tilde{Q}_{s-P(i-1),1}, \cdots, \tilde{Q}_{s-P(i-1),P}), & 1 \leq mod(s, K+1) \leq K \\ 0_{(4r+\varrho)x(4r+\varrho)}, & mod(s, K+1) = K+1 \end{cases} \tag{6.1.26}$$

6.2 Simulation 2

The theoretical developments of suitable models to predict the actual values in Table 6.1 are going to be illustrated for a random walk plus noise with binding totals; it means when the row and column totals are observed without any survey errors assuming $\in_{.} = e_{.j} = 0$.

The RWN model was presented in Equation 3.2.3. Assuming a RWN model for the underlying processes η_{tj}. A new modification is considered for the model of the auxiliary information y_{tj}. Adding a survey error term in the RWN model, that gives the state space model

$$\begin{aligned} y_{tj} &= \eta_{tj} + \ell_{tj} = \underbrace{a_{tj} + \epsilon_{tj}}_{\eta_{tj}} + \ell_{tj}, \epsilon_{tj} \sim NID(0, \sigma^2_{\epsilon_j}) \\ a_{tj} &= a_{t-1,j} + \nu_{tj}, \nu_{tj} \sim NID(0, \sigma^2_{\nu_j}) \end{aligned} \tag{6.2.1}$$

for $t = 1, \cdots, n$ and $j = 1, \cdots, P$; ϵ_{tj}'s and ν_{tj}'s all mutually independent and independent of a_0; and ℓ_{tj} representing the distortion to the RWN model corresponding

to the non-observed vector of survey errors for the subsector j at instant t. The underlying level of the process is assumed to be generated by the random walk, a_t, but like the irregular term, ϵ_t and the survey errors ℓ_{tj}, they are not directly observable. Additionally, assuming the process ℓ_{tj} is an AR(1) process, ℓ_{tj} can be expressed as

$$\ell_{tj} = \phi_j \ell_{t-1,j} + \chi_{tj}, \chi_{tj} \sim NID(0, \sigma^2_{\chi_j}) \tag{6.2.2}$$

for $j = 1, \cdots, P$ and $t = 1, \cdots, n$. This section performs a simulation study in a very simple case, assuming $P = 2$ and $K = 4$. The same seed used to generate the values in the contemporaneous disaggregation with missing values was used in this case to produce comparable results. The experiment consisted on generating one pair of series $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ from the RWN model with initial values $a_{01} = 30$, $a_{02} = 70$, $\phi_1 = \phi_2 = 0.7$, $\sigma^2_{\epsilon} = 3$ and $\sigma^2_{v} = 0.5$. Their corresponding values will be the same as in Figure 5.1.

Year	Period	$\boldsymbol{\eta_1}$	$\boldsymbol{\eta_2}$	$\boldsymbol{z}$
1	1	28.65669	72.26048	100.91717
	2	32.33554	64.65780	96.99334
	3	24.73287	82.71407	107.44694
	4	42.78914	68.50138	111.29052
Total		128.51424	288.13373	416.64797

Table 6.3. Simulated values for reconciliation. RWN model. First year.

The corresponding vector of totals $\boldsymbol{z}$ with dimension 250 x 1 and the vectors $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ with dimension 62 x 1, $62 = [250]_4$ are built. Table 6.3 presented above is the analogous to the Table 5.1 for this special case in the first year.

Generating and adding two series of AR(1) survey errors with coefficient 0.7 in both cases to the quarter subsector values inside the table, now we produced Table 6.4 which does not comply with the row and column totals. Assuming there are no survey errors in these totals (both, by rows and columns), the idea is to disaggregate these totals using the information contained in the preliminary survey estimates series.

The proposed method was applied by building the series $\boldsymbol{y^*_s}$ $(s = 1, \cdots, n + m = 312)$ and writing the observation and transition equations in the form $y^*_s = Z^*_s \alpha^*_s + \varepsilon^*_s$,

Year	Period	y_1	y_2	z
1	1	27.05718	72.42865	100.91717
	2	30.61360	65.47342	96.99334
	3	26.20821	82.85157	107.44694
	4	45.35098	68.46961	111.29052
Total		128.51424	288.13373	416.64797

Table 6.4. **Simulated values for reconciliation. RWN with AR(1) survey errors. First year.**

$\boldsymbol{\alpha}_s^* = T_s^* \boldsymbol{\alpha}_{s-1}^* + \vartheta_s^*$ with

$$\boldsymbol{\alpha}_s^* = [a_{s1}, \epsilon_{s1}; \cdots; a_{s-3,1}, \epsilon_{s-3,1}; a_{s2}, \epsilon_{s2}; \cdots; a_{s-3,2}\epsilon_{s-3,2}; \ell_{s1}; \ell_{s2}] \tag{6.2.3}$$

$$Z_s^* = \begin{cases} \begin{bmatrix} 1_2' & 0_6' & 0_2' & 0_6' & 1 & 0 \\ 0_2' & 0_6' & 1_2' & 0_6' & 0 & 1 \\ 1_2' & 0_6' & 1_2' & 0_6' & 0 & 0 \end{bmatrix}, & 1 \leq mod(s,5) \leq 4 \\ \begin{bmatrix} 1_8' & 0_8' & 0_2' \\ 0_8' & 1_8' & 0_2' \end{bmatrix}, & mod(s,5) = 5 \end{cases} \tag{6.2.4}$$

and

$$T_s^* = \text{diag}\left[\left(\begin{array}{c|c|c|c} T & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ \hline I_2 & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ \hline 0_{2x2} & I_2 & 0_{2x2} & 0_{2x2} \\ \hline 0_{2x2} & 0_{2x2} & I_2 & 0_{2x2} \end{array}\right), \left(\begin{array}{c|c|c|c} T & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ \hline I_2 & 0_{2x2} & 0_{2x2} & 0_{2x2} \\ \hline 0_{2x2} & I_2 & 0_{2x2} & 0_{2x2} \\ \hline 0_{2x2} & 0_{2x2} & I_2 & 0_{2x2} \end{array}\right), \; 0.7, \; 0.7\right]_{18 \times 18} \tag{6.2.5}$$

for $1 \leq mod(s,5) \leq 4$ where

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{6.2.6}$$

for $j = 1, 2$. $T_s^* = I_{18}$ for $mod(s,6) = 5$

The vector of disturbances $\boldsymbol{v}_s^*$ is equal to

$$\begin{cases} \boldsymbol{\vartheta}_s^* = [\nu_{s1}\epsilon_{1s}; 0_6; \nu_{s2}\epsilon_{s2}; 0_6; \chi_{s1}; \chi_{s2}]', & 1 \leq mod(s,5) \leq 4 \\ 0'_{18}, & mod(s,5) = 5 \end{cases} \tag{6.2.7}$$

and

$$Q_s^* = \begin{cases} \text{diag}(Q_{s-P(i-1),1}, 0_{6x6}, Q_{s-P(i-1),2}, 0_{6x6}, \tilde{Q}_{s-P(i-1),1}, \tilde{Q}_{s-P(i-1),2}), & 1 \leq mod(s,5) \leq 4 \\ 0_{18x18}, & mod(s,5) = 5 \end{cases} \tag{6.2.8}$$

The Kalman Filter was initialised using a diffuse prior. Again, the Moore-Penrose inverse was used in the Kalman filter equations obtaining the estimates $\hat{\boldsymbol{y}}_s^*$ and $\hat{\boldsymbol{\alpha}}_s^*$ for every $s = 1, \cdots, 312$.

Then, using the vectors

$$Z_1 = ([1,1] \otimes [10'_3] \otimes [1,0]; 0'_2) \tag{6.2.9}$$

$$Z_2 = ([1,1] \otimes [10'_3] \otimes [0,1]; 0'_2) \tag{6.2.10}$$

the estimated values are calculating using

$$\hat{\eta}_{s1} = Z^1 \hat{\alpha}_s^* \quad \hat{\eta}_{s2} = Z^2 \hat{\alpha}_s^* \tag{6.2.11}$$

and their respective variances

$$Var(\hat{\eta}_{sj}) = Z^j P_s (Z^j)' \quad j = 1,2 \tag{6.2.12}$$

where $\boldsymbol{P_s}$ is the covariance matrix obtained by the Kalman Filter.

6.2.1 Iterative Proportional Fitting - Results

The iterative proportional fitting (IPF) or raking method was introduced in the last chapter. The main idea is to adjust a matrix of any dimension until the totals by row and columns converge to some pre-defined values. The original table values are gradually adjusted in several iterations to fit the row and column constraints. The final

estimated contingency table after iteration corresponds to the maximum likelihood estimates obtained when the probabilities are convergent within an acceptable predefined limit (Bishop et al., 1975, pages 82-101).

In summary, the method is implemented as follows:

1. Take the multidimensional table with initial values $\eta_{tj}^{(0)} = \frac{\eta_{..(i)}}{k*p}$ for every cell $tj; t = 1, \cdots, n; j = 1, \cdots, P$ in the contingency table associated to every year but without adding the marginal totals by row and columns.

2. Scale the initial values to the first marginal sub-totals (e.g. totals by columns) in each year to derive an estimate: $\eta_{tj}^{(1)r} = \eta_{tj}^{(0)} * \left(\frac{\eta_{.j(i)}}{\sum_{j=1}^{P} \eta_{tj}^{(0)}} \right)$ for every cell $tj(t = 1, \cdots, n; j = 1, \cdots, P$ in the contingency table associated to every year with the superindex r indicating the r-th adjustment by rows.

3. Repeat the scaling to the marginal sub-total in the other dimension (e.g. totals by rows) to complete one cycle of $s = 1, \cdots, S$ steps $\eta_{tj}^{(1)} = \eta_{tj}^{(1)c} = \eta_{tj}^{(1)r} * \left(\frac{\eta_{t.}}{\sum_{K*(i-1)+1}^{K_i} \eta_{tj}^{(0)}} \right)$ with the superindex c indicating adjustment by columns.

4. In general at the s-th step, we have $\eta_{tj}^{(s)} = \eta_{tj}^{(s-1)} * \left(\frac{\eta_{.j(i)}}{\sum_{j=1}^{P} \eta_{tj}^{(0)}} \right) * \left(\frac{\eta_{t.}}{\sum_{K*(i-1)+1}^{K_i} \eta_{tj}^{(0)}} \right)$ for $s = 1, \cdots, S$.

5. Steps (1)-(4) are repeated until the factors $\left(\frac{\eta_{.j(i)}}{\sum_{j=1}^{P} \eta_{tj}^{(0)}} \right) \approx \left(\frac{\eta_{t.}}{\sum_{K*(i-1)+1}^{K_i} \eta_{tj}^{(0)}} \right) \approx 1$ under some convergence criterion. Since the procedure is proven to converge when the marginal sub-totals are consistent with each other (for example add to the same overall total), the choice of convergence criterion only affects the number of cycles that will be needed before the criterion is met.

Tables 6.5 (one iteration) and 6.7 (1000 iterations) show the actual simulated values and the corresponding estimates under this method with complete fulfilment of the restrictions by row and columns. The application of this method does not permit to

produce neither estimates for the incomplete last year at the end of the series nor standard error of the estimates.

The two first plots in Figure 6.1 show the estimated levels for each subsector through time. The graphs show how close are the means of the estimated (raking) values after 1000 iterations to the means of the simulated (original) values and also how the estimated values preserve the behaviour of the original series. The application of this method in the reconciliation case shows better performance than when it was applied for contemporaneous disaggregation. Also, the difference between the actual and the estimated value for a given pair of month/subsector were calculated and plotted in the plot at the bottom showing values closer to zero for each subsector. The plot of the differences shows corresponding mirror images showing the strong dependence present in the disaggregated data and values very close to zero.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	26.13927	65.02875
	2	24.33702	58.95520	83.29222	24.54552	58.74670
	3	19.00001	58.39089	77.39090	19.27451	58.11639
	4	18.43571	70.80347	89.23918	18.68926	70.54991
Total		88.64856	252.44176	341.09032	88.64856	252.44176
2	1	30.84828	64.79764	95.64592	31.35793	64.28799
	2	24.84245	70.78847	95.63092	25.99778	69.63314
	3	30.83328	77.11675	107.95003	29.91166	78.03837
	4	37.16156	65.74247	102.90403	36.41822	66.48581
Total		123.68557	278.44533	402.13090	123.68557	278.44533
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	26.60833	58.44111
	2	19.57317	79.39541	98.96858	19.35426	79.61433
	3	39.44022	75.16697	114.60719	39.01559	75.59160
	4	35.21179	82.30659	117.51838	34.76809	82.75029
Total		119.74626	296.39733	416.14359	119.74626	296.39733
63	1	42.35141	64.84352	107.19493	NA	NA
	2	24.88833	66.51744	91.40577	NA	NA

Table 6.5. Results Reconciliation Case. Raking Estimates. Single Iteration.

Statistic	Year 1	Year 2	Year 3	···	Year 62	Year 63 (incomplete)
TAE	2.94620	6.65988	1.20531	···	4.34901	NA
ARE	0.36828	0.83249	0.15066	···	0.54363	NA

Table 6.6. TAE and ARE. Raking Estimates. Reconciliation Case. Single Iteration. Mean(ARE) = 0.56569.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.10970	70.57724	100.63694	30.12782	70.50912
	2	30.11886	70.53026	100.64912	30.12351	70.52560
	3	30.12508	70.51454	100.63962	30.13597	70.50365
	4	30.12646	70.53522	100.66168	30.09280	70.56888
Total		120.48010	282.10726	402.58736	120.48010	282.10725
2	1	30.14526	70.54183	100.68709	30.13573	70.55136
	2	30.14947	70.53614	100.68561	30.15971	70.52599
	3	30.15734	70.54471	100.70205	30.14422	70.55783
	4	30.15508	70.55313	100.70821	30.16750	70.54071
Total		120.60715	282.17581	402.78296	120.60715	282.17581
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.47008	71.70197
	2	30.49117	71.68520	102.17637	30.52220	71.65417
	3	30.49767	71.68915	102.18682	30.46884	71.71799
	4	30.49078	71.70624	102.19702	30.50063	71.69639
Total		121.96173	286.7053	408.73226	121.96173	286.7053
63	1	30.48295	71.72075	102.20370	NA	NA
	2	30.46696	71.72257	102.18953	NA	NA

Table 6.7. Results Reconciliation Case. Raking Estimates. Average 1000 Iterations.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	4.18103	4.10984	4.08860	⋯	4.28053	NA
ARE	0.52263	0.51373	0.51108	⋯	0.53507	NA

Table 6.8. TAE and ARE. Raking Estimates. Reconciliation Case. 1000 Iterations.

Mean(ARE)= 0.52053.

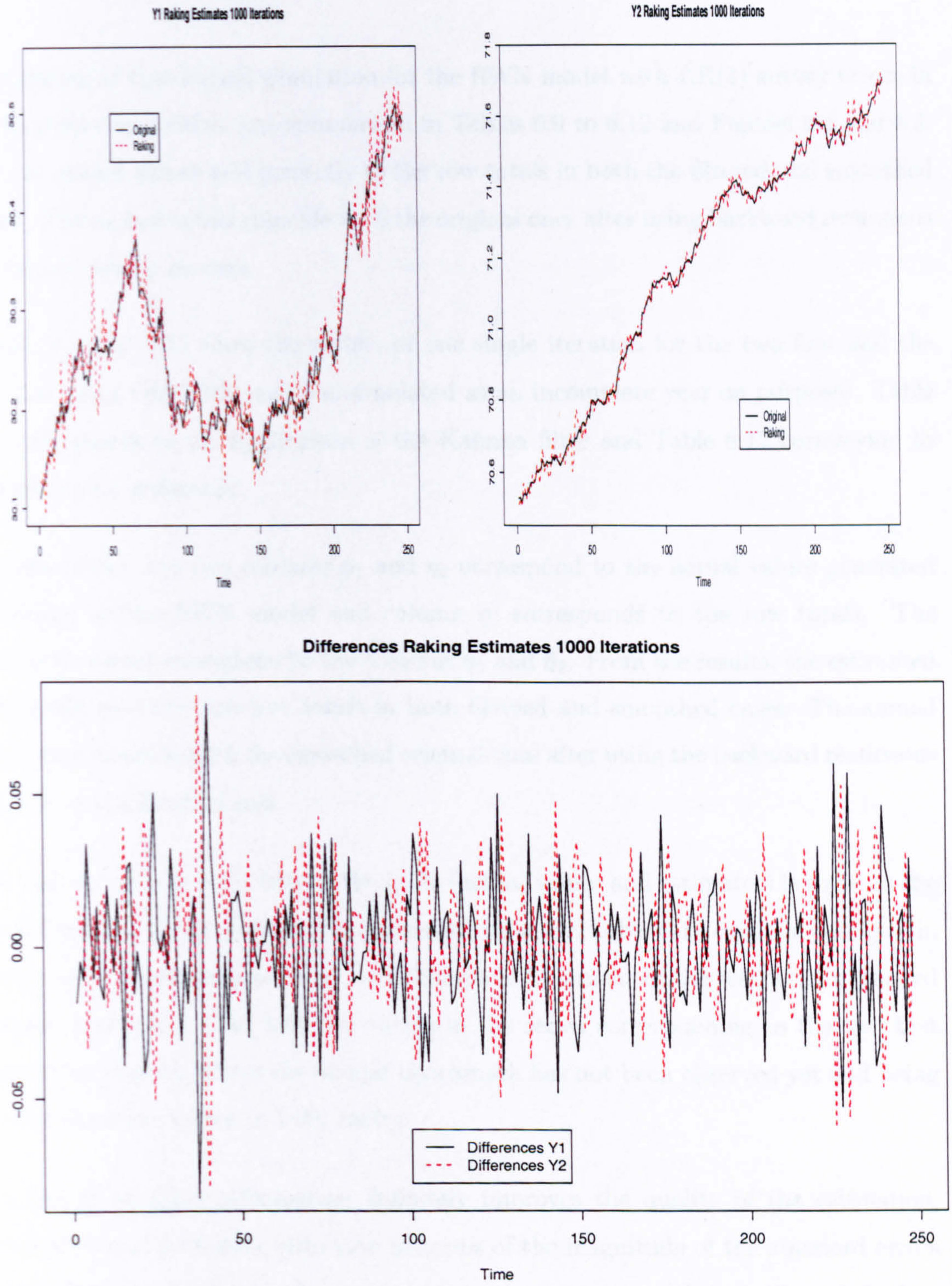


Figure 6.1. **Mean of the Disaggregated Raking Estimates for a RWN model in 1000 Iterations (First Year was not Considered). Reconciliation Case.**

6.2.2 Proposed Method - Results

The results of this second simulation for the RWN model with AR(1) survey errors in a reconciliation context are summarised in Tables 6.9 to 6.12 and Figures 6.2 and 6.3. The estimated values add perfectly to the row totals in both the filtered and smoothed cases. The annual totals coincide with the original ones after using backward recursions in the smoothing process.

Tables 6.9 and 6.12 show the results of one single iteration for the two first and the two last years (the 63th year was simulated as an incomplete year on purpose). Table 6.9 corresponds to the application of the Kalman filter and Table 6.12 correspond to the smoothed estimates.

In both tables, the two columns η_1 and η_2 correspond to the actual values generated according to the RWN model and column $\eta_.$ corresponds to the row totals. The estimated values correspond to the columns $\hat{\eta}_1$ and $\hat{\eta}_2$. From the results, the estimated values add perfectly the row totals in both filtered and smoothed cases. The annual totals only coincide with the smoothed original ones after using the backward recursions but not in the filtering case.

The sum by rows in both sides of the table (actual values and estimates) is equal to the shaded column in the middle. The estimated values in Table 6.12 coincide perfectly in both, rows and columns per year, with complete fulfilment of the restrictions (shaded columns and rows). The last two values in the series corresponding to the two first quarters in year 63, where the annual benchmark has not been observed yet and being exactly the same values in both tables.

The use of auxiliary information definitely improves the quality of the estimation. Tables 6.10 and 6.13 show reduction in terms of the magnitude of the standard errors and coefficients of variation from one year to the next one with exception of the last two years. According to the magnitude of the coefficients of variation, filtered values are not useful in the first year of observation (possibly not useful until the third one). On the other hand, the values for the standard errors are relatively small and stable in

the Table 6.10 for all the periods of study with exception perhaps of the last incomplete year. Standard errors are exactly the same in both subsectors given that the values were simulated with the same variances. Since the means are different, the cvs for $\hat{\eta}_2$ are smaller than $\hat{\eta}_1$. Comparing tables 6.10 and 6.13, there is a big reduction in terms of the magnitude of the standard errors and coefficients of variation. These values are bigger in earlier years than in the later ones. Filtered values in the first year are now better estimated but not enough to use the estimates for the first three quarters in the first year. The standard errors and coefficients of variation are relatively stable in both tables without considering these three first values.

The same process was repeated in 1000 iterations. Tables 6.15 and 6.17 show the average of these results after repeating the process 1000 times. The simulated values converge to the mean of the process (30 and 70 respectively) and practically follow the same conclusions obtained in the single iteration case.

Figures 6.2 and 6.3 confirm that the results for the reconciliation case are better than those in the first case of contemporaneous and temporal disaggregation. The two series filtered and FS are very close to the series of original values. These figures show how close are the means of the estimated values after 1000 iterations with respect to the means of the simulated actual values and also how the estimated values preserve the behaviour of each one of the original series. Figure 6.2 show the results for the filtered estimates and Figure 6.3 for the smoothed ones. Differences between actual and estimated values are close to zero in both figures with less dispersion for the smoothed estimates. The range of values of the differences are smaller than those of the differences in the first simulation; that means they are closer to zero.

The proposed method has some advantages over the previous methods above. As happened in the contemporaneous disaggregation case and compared with the raking method, the main advantage of the proposed method is the possibility to obtain estimates of the final year (ex-ante estimation) with their corresponding standard errors.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	27.14925	64.01877
	2	24.33702	58.95520	83.29222	24.92325	58.36897
	3	19.00001	58.39089	77.39090	19.40400	57.98690
	4	18.43571	70.80347	89.23918	18.97550	70.26368
Total		88.64856	252.44176	341.09032	90.45200	250.63832
2	1	30.84828	64.79764	81.63651	30.96340	64.68251
	2	24.84245	70.78847	95.63092	26.03952	69.59139
	3	30.83328	77.11675	107.95003	30.49359	77.45644
	4	37.16156	65.74247	102.90403	36.37559	66.52844
Total		123.68557	278.44533	402.13090	123.87210	278.25878
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	25.73565	59.31379
	2	19.57317	79.39541	98.96858	19.62171	79.34688
	3	39.44022	75.16697	114.60719	39.42992	75.17727
	4	35.21179	82.30659	117.51838	35.72890	81.78948
Total		119.74626	296.39733	416.14359	120.51618	295.62742
63	1	42.35141	64.84352	107.19493	42.11710	65.07783
	2	24.88833	66.51744	91.40577	25.03634	66.36943

Table 6.9. Results Reconciliation Case. Filtered Estimates. Single Iteration.

Year	Period	se $(\hat{\eta}_1)$	se $(\hat{\eta}_2)$	cv $(\hat{\eta}_1)$	cv $(\hat{\eta}_2)$
1	1	0.70712	0.70712	0.02631	0.01010
	2	0.70542	0.70542	0.02899	0.01197
	3	0.70233	0.70233	0.03696	0.01203
	4	0.69857	0.69857	0.03789	0.00987
2	1	0.57412	0.57412	0.01861	0.00886
	2	0.62834	0.62834	0.02529	0.00888
	3	0.65283	0.65283	0.02117	0.00847
	4	0.66450	0.66450	0.01788	0.01011
⋮	⋮	⋮	⋮	⋮	⋮
62	1	0.56881	0.56881	0.02229	0.01007
	2	0.62342	0.62342	0.03185	0.00956
	3	0.64778	0.64778	0.01642	0.00785
	4	0.65965	0.65965	0.01873	0.00862
63	1	0.56881	0.56881	0.01343	0.00877
	2	0.62342	0.62342	0.02505	0.00937

Table 6.10. Standard Errors and Coefficients of Variation. Filtered Estimates. Reconciliation Case. Single Iteration

Statistic	Year 1	Year 2	Year 3	···	Year 62	Year 63 (incomplete)
TAE	3.60689	4.87573	4.19792	···	1.58104	0.76465
ARE	0.45086	0.60947	0.52474	···	0.19763	0.19116

Table 6.11. TAE and ARE. Filtered Estimates. Reconciliation Case. Single Iteration. Mean(ARE) = 0.70164.

Year	Period	Actual Values η_1	η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	$\hat{\eta}_2$
1	1	26.87852	64.29220	91.16802	26.64019	64.52783
	2	24.33702	58.95520	83.29222	24.39748	58.89474
	3	19.00001	58.39089	77.39090	19.01801	58.37289
	4	18.43571	70.80347	89.23918	18.59288	70.64630
Total		88.64856	252.44176	341.09032	88.64856	252.44176
2	1	30.84828	64.79764	95.64592	31.00080	64.64511
	2	24.84245	70.78847	95.63092	26.02879	69.60212
	3	30.83328	77.11675	107.95003	30.28763	77.66240
	4	37.16156	65.74247	102.90403	36.36834	66.53569
Total		123.68557	278.44533	402.13090	123.68557	278.44533
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	25.52108	59.52836	85.04944	25.92837	59.12107
	2	19.57317	79.39541	98.96858	19.35450	79.61409
	3	39.44022	75.16697	114.60719	39.21180	75.39539
	4	35.21179	82.30659	117.51838	35.25160	82.26678
Total		119.74626	296.39733	416.14359	119.74626	296.39733
63	1	42.35141	64.84352	107.19493	42.15395	65.04098
	2	24.88833	66.51744	91.40577	25.03634	66.36943

Table 6.12. Results Reconciliation Case. Smoothed Estimates. Single Iteration.

Year	Period	se $(\hat{\eta}_1)$	se $(\hat{\eta}_2)$	cv $(\hat{\eta}_1)$	cv $(\hat{\eta}_2)$
1	1	0.18115	0.18113	0.01584	0.00662
	2	0.10448	0.10448	0.01328	0.00548
	3	0.11017	0.11017	0.01747	0.00568
	4	0.16286	0.16286	0.02189	0.00570
2	1	0.15649	0.15649	0.01282	0.00610
	2	0.10447	0.10447	0.01301	0.00457
	3	0.10469	0.10469	0.01049	0.00420
	4	0.15578	0.15578	0.01062	0.00600
⋮	⋮	⋮	⋮	⋮	⋮
62	1	0.16205	0.16205	0.01577	0.00676
	2	0.10950	0.10950	0.01691	0.00417
	3	0.10448	0.10448	0.00819	0.00430
	4	0.17804	0.17804	0.01198	0.00513
63	1	0.31832	0.31832	0.01332	0.00870
	2	0.38865	0.38865	0.02505	0.00937

Table 6.13. Standard Errors and Coefficients of Variation. Smoothed Estimates. Reconciliation Case. Single Iteration

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	0.94257	5.35547	3.46367	⋯	1.78839	0.69095
ARE	0.11782	0.66943	0.43296	⋯	0.22355	0.17274

Table 6.14. TAE and ARE. Smoothed Estimates. Reconciliation Case. Single Iteration. Mean(ARE) = 0.41943.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.10970	70.52724	100.63694	30.12339	70.51355
	2	30.11886	70.53026	100.64912	30.13438	70.51473
	3	30.12508	70.51454	100.63962	30.12091	70.51872
	4	30.12646	70.53522	100.66168	30.11316	70.54852
Total		120.48011	282.1072	402.58736	120.49184	282.09552
2	1	30.14526	70.54183	100.68709	30.12864	70.55845
	2	30.14947	70.53614	100.68561	30.15983	70.52578
	3	30.15734	70.54471	100.70205	30.15594	70.54611
	4	30.15508	70.55313	100.70821	30.15645	70.55176
Total		120.60715	282.17581	402.78296	120.60086	282.18210
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.44919	71.72286
	2	30.49117	71.68520	102.17637	30.50529	71.67108
	3	30.49767	71.68915	102.18682	30.47420	71.71262
	4	30.49078	71.70624	102.19702	30.46804	71.72898
Total		121.96173	286.7053	408.73226	121.89672	286.83554
63	1	30.48295	71.72075	102.20370	30.47929	71.72441
	2	30.46696	71.72257	102.18953	30.48089	71.70864

Table 6.15. Results Reconciliation Case. Filtered Estimates. Average Values in 1000 Iterations.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	6.51140	6.24243	6.29449	⋯	6.34553	3.11674
ARE	0.81393	0.78030	0.78681	⋯	0.79319	0.77918

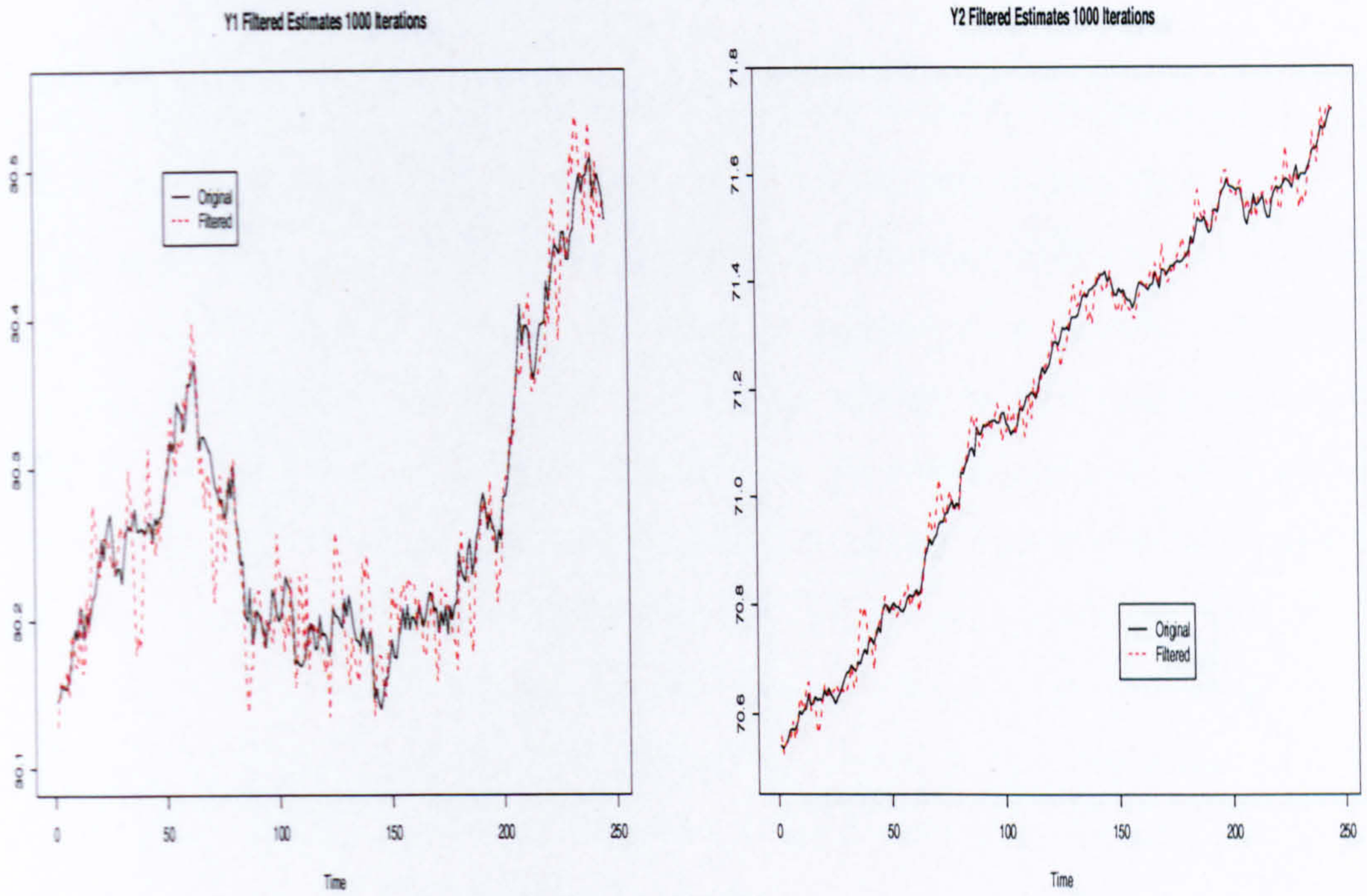
Table 6.16. TAE and ARE. Filtered Estimates. Reconciliation Case. Average Values in 1000 Iterations. Mean(ARE) = 0.78249.

Year	Period	Actual Values η_1	Actual Values η_2	Total $\eta_.$	Estimates $\hat{\eta}_1$	Estimates $\hat{\eta}_2$
1	1	30.19970	70.52724	100.63694	30.11953	70.51741
	2	30.11886	70.53025	100.64912	30.13049	70.51862
	3	30.12509	70.51454	100.63962	30.11816	70.52147
	4	30.12646	70.53522	100.66168	30.11196	70.54972
Total		120.57014	282.10722	402.58736	120.57014	282.10722
2	1	30.14526	70.54183	100.68709	30.13107	70.55603
	2	30.14947	70.53614	100.68561	30.16298	70.52263
	3	30.15734	70.54471	100.70205	30.15787	70.54418
	4	30.15508	70.55313	100.70821	30.15524	70.55297
Total		120.60715	282.17581	402.78296	120.60715	282.17581
⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	1	30.48211	71.68994	102.17205	30.46147	71.71058
	2	30.49117	71.68520	102.17637	30.52275	71.65362
	3	30.49767	71.68915	102.18682	30.49284	71.69399
	4	30.49078	71.70624	102.19702	30.48468	71.71234
Total		121.96173	286.77053	408.73226	121.96173	286.77053
63	1	30.48295	71.72075	102.20370	30.48002	71.72367
	2	30.46696	71.72257	102.18953	30.48089	71.70864

Table 6.17. Results Reconciliation Case. Smoothed Estimates. Average Values in 1000 Iterations.

Statistic	Year 1	Year 2	Year 3	⋯	Year 62	Year 63 (incomplete)
TAE	3.64349	3.45483	3.45785	⋯	3.70694	3.12199
ARE	0.45544	0.43185	0.43223	⋯	0.46337	0.78049

Table 6.18. TAE and ARE. Smoothed Estimates. Reconciliation Case. Average Values in 1000 Iterations. Mean(ARE) = 0.43560.



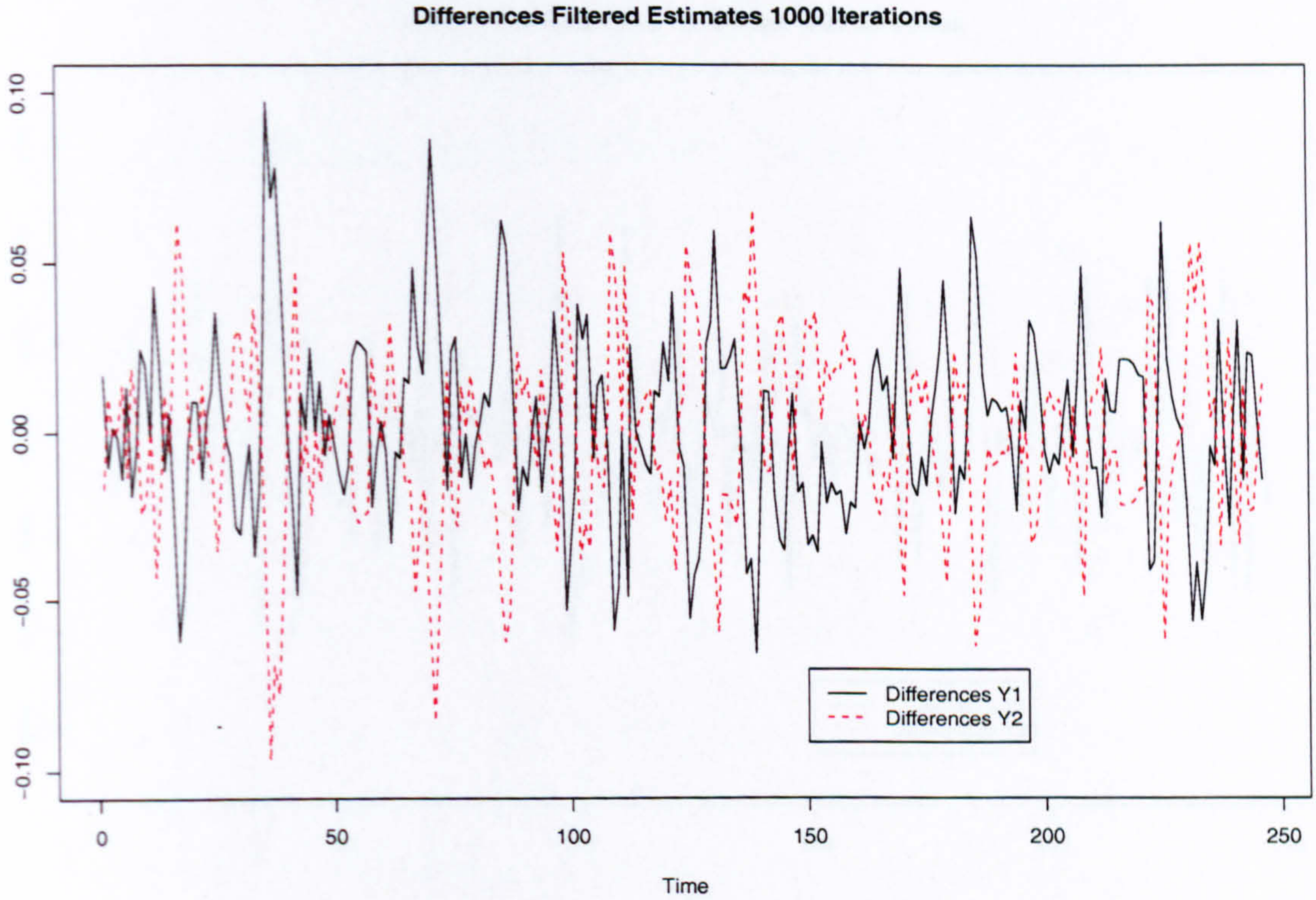


Figure 6.2. **Mean of the Disaggregated Filtered Estimates for a RWN model in 1000 Iterations (First Year was not Considered). Reconciliation Case.**

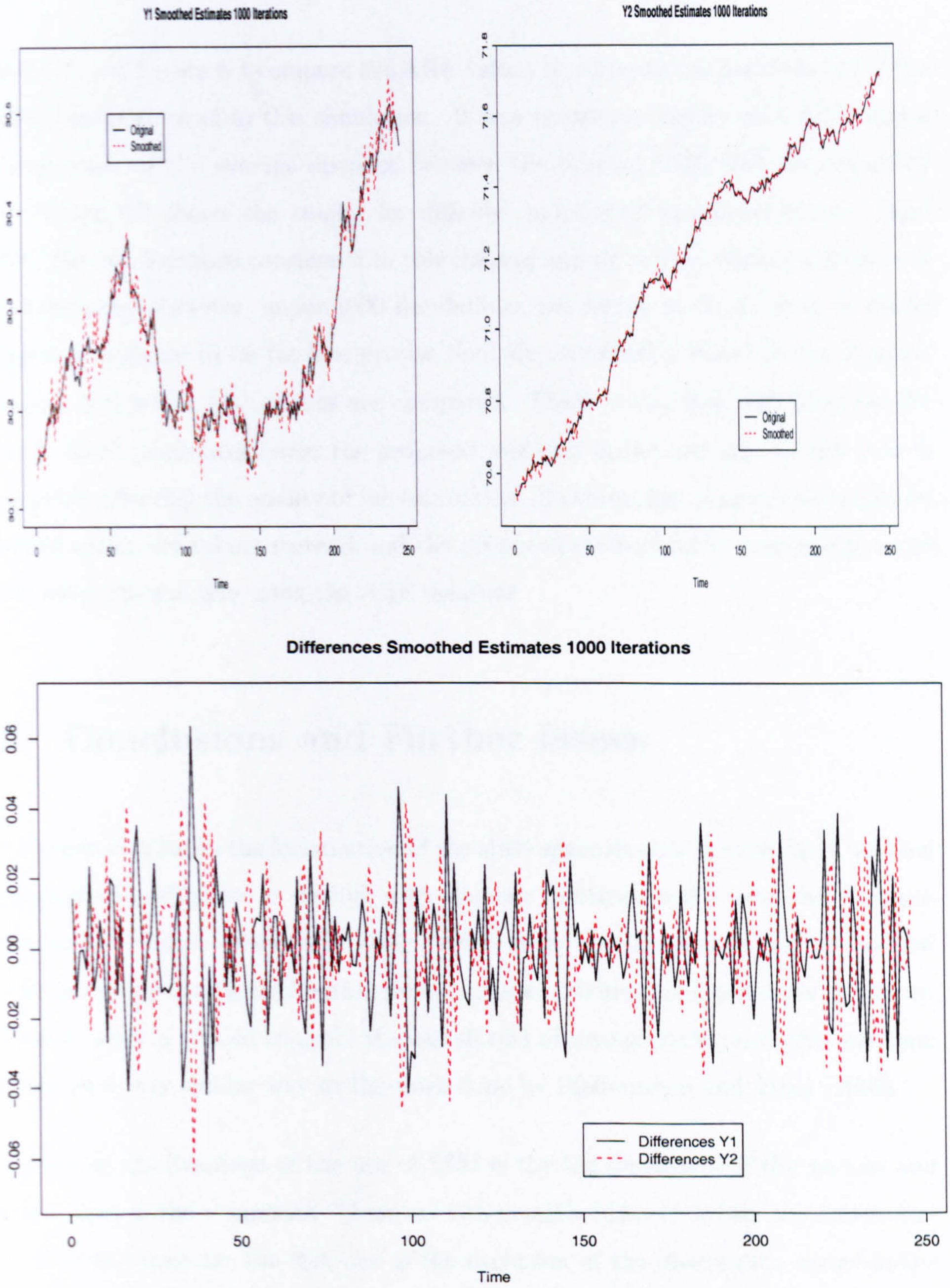


Figure 6.3. **Mean of the Disaggregated Smoothed Estimates for a RWN model in 1000 Iterations. Reconciliation Case.**

6.2.3 Comparison of the Methods

Table 6.19 and Figure 6.4 compare the ARE values to evaluate the performance of the different methods used in this simulation. It was mentioned before that ARE values are a measure of the average distance between the original table and the estimated one. Figure 6.4 shows the results for different number of simulated series (1 and 1000). The two methods considered in this chapter appear to have similar precision in their estimates. However, under 1000 simulations, the raking method (series in red in Figure 6.4) appears to be the less precise than the proposed method (series in black in Figure 6.4) when ARE values are compared. There is an ascendant drop for the series in black (estimates under the proposed method) in the last year as this year is incomplete, affecting the quality of the estimation. However, last year values cannot be obtained under the raking method and the proposed method (after smoothing) looks as the most efficient one using the ARE measure.

6.3 Conclusions and Further Issues

One important issue in the formulation of the state space models in both cases, missing values and reconciliation, is dealing with singular matrices in the recursion formulas of the Kalman filter. Kohn and Asley (1993) show how the use of any generalised inverse, including the Moore Penrose pseudoinverse, produce good estimates. However, additional work is needed to avoid the calculation of inverse matrices in the recursion formulas in a very similar way to the work done by Pfeffermann and Tiller (2005).

One possible disadvantage of the use of SSM is the big dimension of the vectors and matrices used in the recursions. There are two possible ideas to reduce the dimensionality of these elements: the first one is the exclusion of the observation errors in the state vector, which will produce autocorrelated errors in the formulation. Pfeffermann and Tiller (2005) have proposed a modification to the Kalman filter which could deal with this problem.

Method	Raking	Proposed-Filtered	Proposed-Smoothed
ARE - Single Iteration	0.56569	0.70164	0.41943
ARE - 1000 Iterations	0.52053	0.78249	0.43560

Table 6.19. **Average of ARE Values. Contemporaneous Disaggregation Methods.**

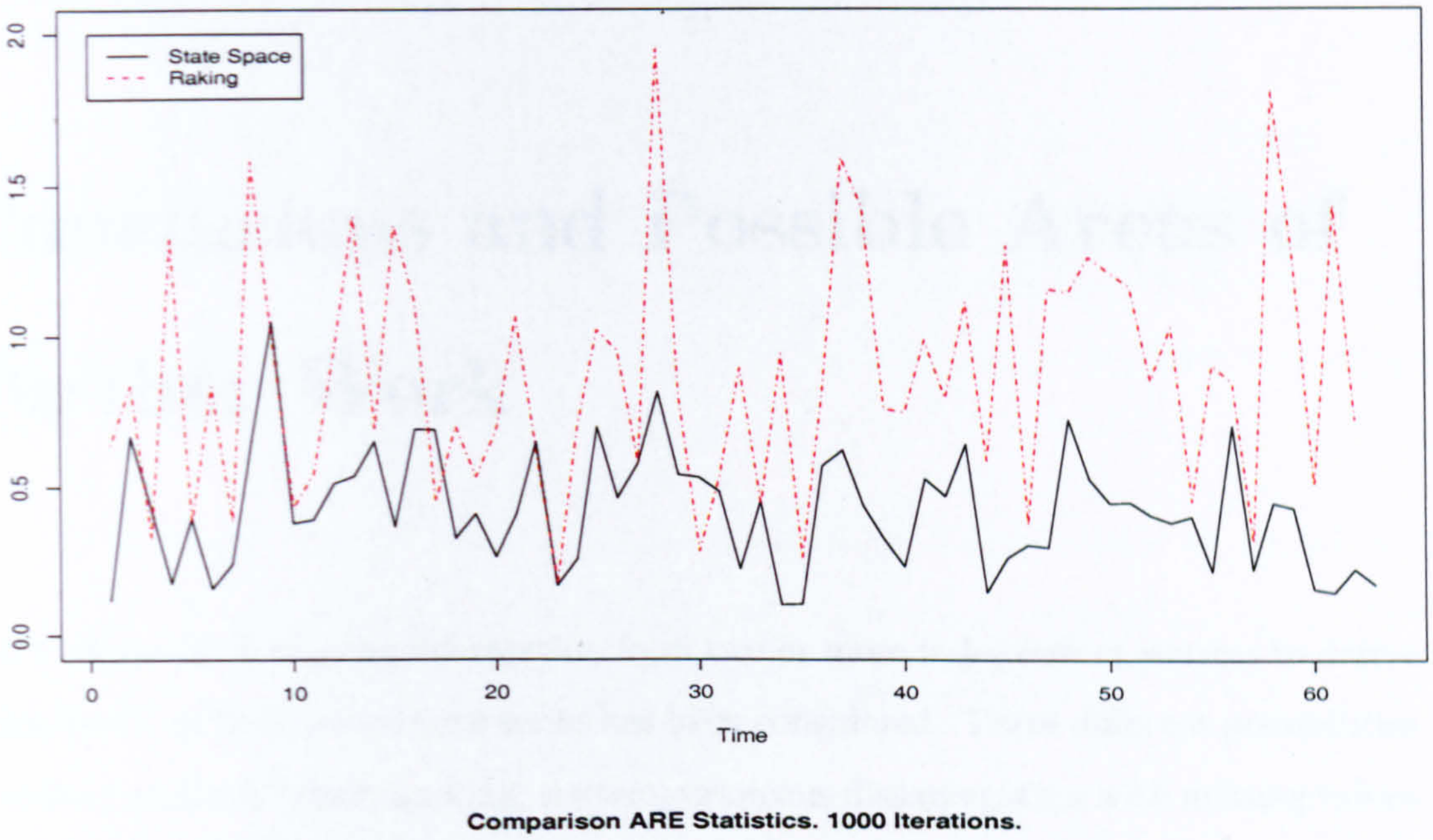


Figure 6.4. **Plots of ARE for each of the methods considered in the simulation. Reconciliation Case.**

Chapter 7

Conclusions and Possible Areas of Further Work

The problem of combining information from two or more independent sources to derive estimates of an unobserved time series has been considered. Three different possibilities have been studied: benchmarking, contemporaneous dissagregation with missing values and reconciliation of time series data.

The univariate case was studied trying to incorporate the stochastic properties of the series being estimated into the procedure. The aim is to produce benchmark estimates having smaller mean squared error than the prebenchmarked estimates. A state space model approach has been used as this kind of model covers a very wide range of models including the ARIMA ones. An application to Business Survey Methods has shown how important is a correct specification of the components: trends, seasonalities, trading days and survey errors. Aspects such as the sampling design, rotation in the samples, the fact that statistical agencies keep the same coefficient of variation during the repeated surveys introduce additional characteristics to take into account in the model as, for example, rotation could generate autocorrelated errors; constant coefficients of variation could generate heteroscedasticity; not to mention other problems such as non-response, frame deterioration through time, etc.

Benchmarking provides better estimates when the annual data provide the most reliable information on the overall level, while the high frequency source provide the only available explicit information about the short-term movements in the series. Denton Method is a good alternative when there is no additional information about the survey or the standard errors of the data. However, signal extraction methods, and particularly those associated with state space models show better properties as for example, they provide standard errors of the estimates.

Additionally, the use of Generalised Variance Functions to complete the missing standard error information, the structural time series modelling of the monthly data and the benchmarking process were illustrated for a particular series from Business Surveys in the UK. Some warning was done about the considerations to be taken to produce standard errors of binding estimates.

When state space models are used for benchmarking, it is not easy to postulate a model for the trend, the seasonalities and the survey errors as they are unobserved components of the observed series. Particularly, the ideal situation is choosing models which produce short length state vectors as much as it possible and they keep the valid assumptions in the innovations and standardised smoothing residuals in the model.

Many series have a structure where the seasonal effects change proportionately with the trend. If the trend increases, so do the seasonal effects and if the trend decreases the seasonal effects diminish too. This is a characteristic of most of the economic series, particularly those referred to as Business Surveys. This structure is known as a multiplicative structure, different than the additive one, where the seasonal effects remain more or less the same no matter which direction the trend is moving. Then, it is necessary to study how the methods applied here must be adapted to treat this specific structure of data.

Regarding the multivariate case, the use of aggregated data permits to eliminate or soften some of the weak points of the data collection. When only annual and sector totals are available and it is necessary to obtain dissagregated values in months or subsectors, a state space formulation presented here shows how to do the estimation

of the missing values. On the other hand, when additional to the annual and sectorial aggregated information, there is a highly correlated information with the missing values by sector and month; or even when survey estimates of the same variable disaggregated by sector and month are available, a new methodology again using state space models was presented. The idea in the second case is to recover the additivity in the tables in order to produce consistent and publishable values complying with both row and column totals.

Aspects as the use of generalised inverses in the Kalman filter, addition of survey error models and specification of trends and seasonalities are points of focus later on. It is also important to consider more challenging structural time series models in the simulations for the multivariate problems in Chapter 5 and 6. Under the consideration of a more general structural model, it is very useful to know how the different components of a sectorial time series could be disaggregated for their corresponding subsectors.

PAGE
NUMBERING
AS ORIGINAL

Part IV

Appendices

Appendix A

Mathematical background - Chapter 2

Some of the main details to obtain the benchmarked estimates and their variances for each one of the methods presented in Chapter 2 are included here. The proofs include details not presented in the original papers.

A.1 Quadratic Minimization Approach

Proposition A.1.1. *Let $f(\hat{\eta} - y)$ be a quadratic form $(\hat{\eta} - y)'A(\hat{\eta} - y)$ with A being a non-singular symmetric $n \times n$ matrix. The minimum of $f(\hat{\eta} - y)$ subject to the restriction 2.2.2 is obtained when*

$$\hat{\eta} = y + Cr$$

where $C = A^{-1}L(L'A^{-1}L)^{-1}$ and $r = x - L'y$.

Proof. The problem of minimising $f(\hat{\eta} - y) = (\hat{\eta} - y)'A(\hat{\eta} - y)$ under the restriction in Equation 2.2.2 is a constrained minimisation problem. The "Lagrangian expression" to be optimized is

$$U(\hat{\eta}, \lambda) = (\hat{\eta} - y)'A(\hat{\eta} - y) - 2\lambda'(x - L'\hat{\eta})$$

where λ is the vector of Langrange multipliers $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_m]'$. If the constrained function is optimized, then the final term above will always be equal to zero. Taking partial derivatives of f with respect to the elements of $\hat{\eta}$ and λ, equating them to zero, and using that

$$\frac{\partial(\hat{\eta}'A\hat{\eta})}{\partial\hat{\eta}} = 2A\hat{\eta}; \qquad \frac{\partial L'\hat{\eta}}{\partial\hat{\eta}} = L$$

It follows that

$$\begin{cases} \frac{\partial U}{\partial \hat{\eta}} = 2A(\hat{\eta} - y) + 2L\lambda = 0 \\ \frac{\partial U}{\partial \lambda} = -2(x - L'\hat{\eta}) = 0 \end{cases} \tag{A.1.1}$$

If $r = x - L'y$ is the vector of discrepancies between the two sets of annual totals, the linear system of equations in Equation A.1.1 can be expressed as

$$\begin{cases} A\hat{\eta} + L\lambda = Ay, \\ L'\hat{\eta} = r + L'y, \end{cases}$$

and its solution as

$$\begin{bmatrix} \hat{\eta} \\ \lambda \end{bmatrix} = \begin{bmatrix} A & L \\ L' & 0 \end{bmatrix}^{-1} \begin{bmatrix} A & 0 \\ L' & I \end{bmatrix} \begin{bmatrix} y \\ r \end{bmatrix} \tag{A.1.2}$$

where I is the mxm identity matrix. Obtaining the inverse of this special partitioned matrix (Faliva and Zoia (2002)),

$$\begin{bmatrix} A & L \\ L' & 0 \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} - A^{-1}L(L'A^{-1}L)^{-1}L'A^{-1} & A^{-1}L(L'A^{-1}L)^{-1} \\ (L'A^{-1}L)^{-1}L'A^{-1} & -(L'A^{-1}L)^{-1} \end{bmatrix}$$

and replacing this matrix in Equation A.1.2, it follows that

$$\begin{bmatrix} \hat{\eta} \\ \lambda \end{bmatrix} = \begin{bmatrix} I - A^{-1}L(L'A^{-1}L)^{-1}L' + A^{-1}L(L'A^{-1}L)^{-1}L' & A^{-1}L(L'A^{-1}L)^{-1} \\ (L'A^{-1}L)^{-1}L' - (L'A^{-1}L)^{-1}L' & -(L'A^{-1}L)^{-1} \end{bmatrix} \begin{bmatrix} y \\ r \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \hat{\eta} \\ \lambda \end{bmatrix} = \begin{bmatrix} I & A^{-1}L(L'A^{-1}L)^{-1} \\ 0 & -(L'A^{-1}L)^{-1} \end{bmatrix} \begin{bmatrix} y \\ r \end{bmatrix}$$

Then, the benchmarked estimates can be expressed as

$$\hat{\eta} = y + Cr$$

where $C = A^{-1}L(L'A^{-1}L)^{-1}$ as it was required. □

A.2 GLM Regression Approach

Proposition A.2.1. *The final BLUE estimators of the parameters a and η are respectively given by*

$$\hat{a} = -\sigma^2_{\hat{a}} 1' L(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y)$$

and

$$\hat{\eta} = y^* + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y^*), \quad y^* = y - 1_n\hat{a}$$

with respective variances

$$\sigma^2_{\hat{a}} = 1/[1'L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1]$$

and

$$\begin{aligned}\Sigma_{\hat{\eta}} =&[\Sigma_\ell - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell] \\ &+ [I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']1\sigma^2_{\hat{a}}1'[I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']'\end{aligned}$$

Proof. Using the Aitken estimator (Aitken (1935)), this is the BLUE for the model in Equation 2.3.3 and can be expressed as

$$\hat{\beta} = (X'\Sigma_u{}^{-1}X)^{-1}X'\Sigma_u{}^{-1}\tau \tag{A.2.1}$$

with covariance matrix given by

$$Cov\ (\hat{\beta}) = (X'\Sigma_u{}^{-1}X)^{-1} \tag{A.2.2}$$

where Σ_u denotes the true assumed known covariance matrix of the disturbances u.

Replacing the corresponding values in A.2.1, it follows that

$$\begin{bmatrix}\hat{a}\\ \hat{\eta}\end{bmatrix} = \left(\begin{bmatrix}1'_n & 0'_m\\ I_n & L\end{bmatrix}\begin{bmatrix}\Sigma_\ell^{-1} & 0\\ 0 & \Sigma_e^{-1}\end{bmatrix}\begin{bmatrix}1_n & I_n\\ 0_m & L'\end{bmatrix}\right)^{-1}\begin{bmatrix}1'_n & 0'_m\\ I_n & L\end{bmatrix}\begin{bmatrix}\Sigma_\ell^{-1} & 0\\ 0 & \Sigma_e^{-1}\end{bmatrix}\begin{bmatrix}y\\ x\end{bmatrix}$$

Thus,

$$\begin{bmatrix}\hat{a}\\ \hat{\eta}\end{bmatrix} = \begin{bmatrix}1'_n\Sigma_\ell^{-1}1_n & 1'_n\Sigma_\ell^{-1}\\ \Sigma_\ell^{-1}1_n & \Sigma_\ell^{-1}+L\Sigma_e^{-1}L'\end{bmatrix}^{-1}\begin{bmatrix}1'_n\Sigma_\ell^{-1}y\\ \Sigma_\ell^{-1}y+L\Sigma_e^{-1}x\end{bmatrix} = \begin{bmatrix}A & B\\ C & D\end{bmatrix}\begin{bmatrix}1'_n\Sigma_\ell^{-1}y\\ \Sigma_\ell^{-1}y+L\Sigma_e^{-1}x\end{bmatrix} \tag{A.2.3}$$

with $A = 1'_n\Sigma_e^{-1}1_n, B = 1'_n\Sigma_\ell^{-1}, C = \Sigma_\ell^{-1}1_n$ and $D = \Sigma_\ell^{-1} + L\Sigma_e^{-1}L'$, respectively.

Also, using Equation A.2.2 it follows that

$$Cov\ (\hat{\beta}) = \begin{bmatrix}\sigma^2_{\hat{a}} & \Sigma_{\hat{a}\hat{\eta}}\\ \Sigma_{\hat{\eta}\hat{a}} & \Sigma_{\hat{\eta}}\end{bmatrix} = \begin{bmatrix}1'_n\Sigma_\ell^{-1}1_n & 1'_n\Sigma_\ell^{-1}\\ \Sigma_\ell^{-1}1_n & \Sigma_\ell^{-1}+L\Sigma_e^{-1}L'\end{bmatrix}^{-1} = \begin{bmatrix}A & B\\ C & D\end{bmatrix}^{-1} \tag{A.2.4}$$

It is easily verified that the inverse of a symmetric partitioned matrix (Anderson (1984), Zwillinger and Kokoska (2000)) can be written as

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}^{-1} = \begin{bmatrix}A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1}\\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1}\end{bmatrix} \tag{A.2.5}$$

Combining the results in Equations A.2.4 and A.2.5 it follows that

$$\begin{aligned}\sigma^2_{\hat{a}} &= A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} \\ &= A^{-1} - A^{-1}B[I + D^{-1}C(-A)^{-1}B]^{-1}D^{-1}C(-A)^{-1}\end{aligned}$$

Using the matrix identity in Jazwinski (1970), Appendix 7B, Identity 3

$$\begin{aligned}\sigma^2_{\hat{a}} &= A^{-1} - A^{-1}BD^{-1}C(BD^{-1}C - A)^{-1} \\ &= A^{-1}(BD^{-1}C - A - BD^{-1}C)(BD^{-1}C - A)^{-1} \\ &= -(BD^{-1}C - A)^{-1}\end{aligned}$$

It can be noticed that

$$D^{-1} = (\Sigma_\ell^{-1} + L\Sigma_e^{-1}L')^{-1} = [I_n + \Sigma_\ell L\Sigma_e^{-1}L']^{-1}\Sigma_\ell$$

and using Jazwinski (1970), Appendix 7B, Identity 2

$$D^{-1} = \Sigma_\ell - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell \tag{A.2.6}$$

Replacing the corresponding values, it becomes that

$$\sigma^2_{\hat{a}} = \frac{-1}{1_n'\Sigma_\ell^{-1}[\Sigma_\ell - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell]\Sigma_\ell^{-1}1_n - 1_n'\Sigma_\ell^{-1}1_n}$$

and finally after some small algebra

$$\boxed{\sigma^2_{\hat{a}} = \frac{1}{1_n'L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n}}$$

Now to calculate the variance of the benchmarked estimates, we have

$$\Sigma_{\hat{\eta}} = (D - CA^{-1}B)^{-1} = [I - D^{-1}CA^{-1}B]^{-1}D^{-1} = [I + D^{-1}C(-A)^{-1}B]^{-1}D^{-1} \tag{A.2.7}$$

Again, using Jazwinski (1970), Appendix 7B, Identity 2

$$\Sigma_{\hat{\eta}} = D^{-1} - D^{-1}C\underbrace{(BD^{-1}C - A)^{-1}}_{-\sigma^2_{\hat{a}}}BD^{-1} \tag{A.2.8}$$

replacing Equation A.2.6 in Equation A.2.8, it follows that

$$\boxed{\begin{aligned}\Sigma_{\hat{\eta}} = {} & [\Sigma_\ell - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell] \\ & + [I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']1_n\sigma^2_{\hat{a}}1_n'[I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']'\end{aligned}}$$

It can be shown that the two blocks not on the diagonal of the matrix in Equation A.2.4. are followed by

$$\Sigma'_{\hat{a}\hat{\eta}} = -[A^{-1}B(D - CA^{-1}B)^{-1}]' = -(D - CA^{-1}B)^{-1}CA^{-1} = \Sigma_{\hat{\eta}\hat{a}}$$

Then, using Equation A.2.7

$$\Sigma_{\hat{\eta}\hat{a}} = [I + D^{-1}C(-A)^{-1}B]^{-1}D^{-1}C(-A)^{-1}$$

Using the matrix identity in Jazwinski (1970), Appendix 7B, Identity 3

$$\Sigma_{\hat{\eta}\hat{a}} = -D^{-1}C\sigma^2_{\hat{a}}$$

Finally replacing the appropriate values

$$\boxed{\Sigma_{\hat{\eta}\hat{a}} = -1_n\sigma^2_{\hat{a}} + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\sigma^2_{\hat{a}}}$$

and

$$\boxed{\Sigma_{\hat{a}\hat{\eta}} = -\sigma^2_{\hat{a}}1'_n + \sigma^2_{\hat{a}}1'_nL(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell}$$

The estimated bias and the benchmarked estimates are obtained replacing all the boxed formulas in Equation A.2.3 to produce

$$\begin{bmatrix}\hat{a}\\ \hat{\eta}\end{bmatrix} = \begin{bmatrix}\sigma^2_{\hat{a}} & -\sigma^2_{\hat{a}}1'_n + \sigma^2_{\hat{a}}1'_nL(L'\Sigma_\ell L + \Sigma_e)^{-1}L\Sigma_\ell \\ -1_n\sigma^2_{\hat{a}} + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\sigma^2_{\hat{a}} & \Sigma_{\hat{\eta}}\end{bmatrix} \times \begin{bmatrix}1'_n\Sigma_\ell^{-1}y \\ \Sigma_\ell^{-1}y + L\Sigma_e^{-1}x\end{bmatrix}$$

then

$$\hat{a} = \sigma^2_{\hat{a}}1'_n\Sigma_\ell^{-1}y + [-\sigma^2_{\hat{a}}1'_n + \sigma^2_{\hat{a}}1'_nL(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell][\Sigma_\ell^{-1}y + L\Sigma_e{}^{-1}x]$$

and after some algebra

$$\boxed{\hat{a} = -\sigma^2_{\hat{a}}1'_nL(L'\Sigma_\ell L + \Sigma_e)^{-1}(x - L'y)}$$

Now

$$\begin{aligned}\hat{\eta} = &- 1_n\sigma^2_{\hat{a}}1_n\Sigma_\ell^{-1}y + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\sigma^2_{\hat{a}}1'_n\Sigma_\ell{}^{-1}y + [y - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y] \\ &+ [\Sigma_\ell L\Sigma_e{}^{-1}x - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell L\Sigma_e{}^{-1}x] \\ &+ [I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']1_n\Sigma_\ell{}^{-1}1'_n[I - L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell]\Sigma_\ell{}^{-1}y \\ &+ [I - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L']1_n\Sigma_\ell{}^{-1}1'_n[I - L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell]L\Sigma_e{}^{-1}x\end{aligned}$$

after some algebra

$$\begin{aligned}\hat{\eta} =& y - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y - 1_n\sigma_{\hat{a}}^2 1_n' L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y \\ &+ \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\sigma_{\hat{a}}^2 1_n' L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y \\ &+ 1_n\sigma_{\hat{a}}^2 1_n' L(\Sigma_e{}^{-1} - (L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell L\Sigma_e{}^{-1})x \\ &+ \Sigma_\ell L(\Sigma_e{}^{-1} - (L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell L\Sigma_e{}^{-1})x \\ &- \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\Sigma_\ell{}^{-1}1_n' L[\Sigma_e{}^{-1} - (L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell L\Sigma_e{}^{-1}]x\end{aligned}$$

but analogously to Equation A.2.6.

$$\Sigma_e{}^{-1} - (L'\Sigma_\ell L + \Sigma_e)^{-1}L'\Sigma_\ell L\Sigma_e{}^{-1} = (L'\Sigma_\ell L + \Sigma_e)^{-1}$$

then

$$\begin{aligned}\hat{\eta} =& y - \Sigma_\ell L(L\Sigma_\ell L + \Sigma_e)^{-1}L'y - 1_n\sigma_{\hat{a}}^2 1_n' L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y \\ &+ \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\sigma_{\hat{a}}^2 1_n' L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y \\ &+ 1_n\sigma_{\hat{a}}^2 1_n' L(L'\Sigma_\ell L + \Sigma_e)^{-1}x + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}x \\ &- \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'1_n\Sigma_\ell{}^{-1}1_n' L(L'\Sigma_\ell L + \Sigma_e)^{-1}x\end{aligned}$$

and after more algebra

$$\hat{\eta} = (y - 1_n\hat{a}) + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}x - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'(y - 1_n\hat{a})$$

to produce

$$\boxed{\hat{\eta} = y^* + \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}x - \Sigma_\ell L(L'\Sigma_\ell L + \Sigma_e)^{-1}L'y^*}$$

with $y^* = y - 1_n\hat{a}$ □

A.3 ARIMA Model Based Approach

Proposition A.3.1. *Consider the model*

$$\tau = X\eta + u$$

Assume η *has a* $N(\mu, \Sigma_\eta)$ *distribution and* u *has a* $N(0, \Sigma_u)$ *distribution, where* $\Sigma_u = diag(\Sigma_\ell, \Sigma_e)$. *The minimum mean squared error estimate of* η *given* τ *is*

$$\hat{\eta} = E(\eta \mid \tau) = \hat{\eta}^0 + \eta_c$$

where $\hat{\eta}^0$ *is the minimum mean squared error linear estimate of* η *given* y

$$\hat{\eta}^0 = E(\eta \mid y) = \begin{cases} (\Sigma_\ell{}^{-1} + \Sigma_\eta{}^{-1})^{-1} \times (\Sigma_\ell{}^{-1} y + \Sigma_\eta{}^{-1}\mu), & \text{if } \eta \text{ is stationary} \\ (\Sigma_\ell{}^{-1} + \Sigma_\eta{}^{-1})^{-1}\Sigma_\ell{}^{-1} y & \text{if } \eta \text{ is not stationary} \end{cases}$$

and η_c *is the correction factor term*

$$\eta_c = \Omega L'(L\Omega L' + \Sigma_e)^{-1}(x - L\hat{\eta}^0)$$

where

$$\Omega = Cov(\eta \mid y) = (\Sigma_\ell{}^{-1} + \Sigma_\eta{}^{-1})^{-1}$$

and

$$\Sigma_{\hat{\eta}} = Cov(\eta \mid \tau) = \Omega - \Omega L'(L\Omega L' + \Sigma_e)^{-1} L\Omega$$

Proof. Cleveland and Tiao (1976) (Appendix A.1) showed that the problem to find the minimum squared error linear estimate (MMSE) of η given y could distinguish two different situations. In the first one, when η_t is stationary (all the zeros of $\phi_\eta(B)$ lie outside the unit circle), $\mu = (\mu, \mu, \cdots, \mu)' = 1_n\mu$ and the MMSE estimate (Cleveland and Tiao (1976), Equation A.4) is

$$\begin{aligned} E(\eta \mid y) = \hat{\eta}^0 &= (I + \Sigma_\ell \Sigma_\eta^{-1})^{-1}(y + \Sigma_\ell \Sigma_\eta{}^{-1}\mu) \\ &= (I + \Sigma_\ell \Sigma_\eta^{-1})^{-1}(\Sigma_\ell \Sigma_\ell{}^{-1} y + \Sigma_\ell \Sigma_\eta{}^{-1}\mu) \\ &= (I + \Sigma_\ell \Sigma_\eta^{-1})^{-1}\Sigma_\ell(\Sigma_\ell{}^{-1} y + \Sigma_\eta{}^{-1}\mu) \\ &= (\Sigma_\ell{}^{-1} + \Sigma_\eta^{-1})^{-1}(\Sigma_\ell{}^{-1} y + \Sigma_\eta{}^{-1}\mu) \\ &= (\Sigma_\ell{}^{-1} + \Sigma_\eta^{-1})^{-1}(\Sigma_\ell{}^{-1} y + \Sigma_\eta{}^{-1} 1_n \mu) \end{aligned}$$

On the other hand, they also proved (Cleveland and Tiao (1976), Appendix A.1, Situation 2) that if η_t follows a non-stationary ARIMA model (all the zeros of $\phi_\eta(B)$ lying on or outside the unit circle) the MMSE estimate (Cleveland and Tiao (1976), Equation A.8) is

$$\begin{aligned} E(\eta \mid y) = \hat{\eta}^0 &= (I + \Sigma_\ell \Sigma_\eta^{-1})^{-1} y \\ &= (\Sigma_\ell{}^{-1} + \Sigma_\eta^{-1})^{-1}\Sigma_\ell^{-1} y \end{aligned}$$

In both cases (Cleveland and Tiao, 1976, Equation A.4 and A.8)

$$\begin{aligned} cov(\eta \mid y) = \Omega &= (I + \Sigma_\ell \Sigma_\eta{}^{-1})^{-1}\Sigma_\ell \\ &= (\Sigma_\ell{}^{-1} + \Sigma_\eta{}^{-1})^{-1} \end{aligned}$$

The equations above depend on values of μ and any two of Σ_y, Σ_η and Σ_ℓ (the third one can be obtained using the relation $\Sigma_y = \Sigma_\eta + \Sigma_\ell$). μ and Σ_y can be estimated

from the ARIMA modelling on y_t and Σ_ℓ could be estimated according to the survey experts, the information about the sampling design or using survey microdata. Finally, Hillmer and Trabelsi (1987) give an adaptive form to write the benchmarked estimate $\hat{\eta}$, in terms of the prebenchmark estimate $\hat{\eta}^0$ and a correction term η_c.

A briefer proof than the one presented in Hillmer and Trabelsi (1987) is presented here. Since $y = \eta + \ell$ and $x = L\eta + e$ then,

$$E(x \mid y) = E(L\eta + e \mid y) = LE(\eta \mid y) = L\hat{\eta}^0$$

Also

$$\hat{\eta} = E(\eta \mid x, y) = E(\eta \mid x - L\hat{\eta}^0, y) \tag{A.3.1}$$

Applying a lemma in multivariate normal regression theory (Durbin and Koopman (2001), Appendix 2.13)

$$\begin{aligned} E(x \mid y, z) &= E(x \mid y) + \Sigma_{xz}\Sigma_{zz}^{-1}z \\ Cov(x \mid y, z) &= Cov(x \mid y) + \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}' \end{aligned} \tag{A.3.2}$$

where $\Sigma_{v1,v2}$ is the covariance matrix $\Sigma_{v1,v2} = E[(v_1 - \mu_{v1})(v_2 - \mu_{v2})']$. Replacing $x = \eta$ and $z = x - L\hat{\eta}^0$ in Equations A.3.2 and A.3.1

$$\begin{aligned} E(\eta \mid x - L\hat{\eta}^0, y) &= E(\eta \mid y) + Cov(\eta, x - L\hat{\eta}^0)[Var(x - L\hat{\eta}^0)]^{-1}(x - L\hat{\eta}^0) \\ Cov(\eta \mid x - L\hat{\eta}^0, y) &= Cov(\eta \mid y) - Cov(\eta, x - L\hat{\eta}^0)[Var(x - L\hat{\eta}^0)]^{-1}(Cov(\eta, x - L\hat{\eta}^0))' \end{aligned} \tag{A.3.3}$$

Given that,

$$\begin{aligned} &* E(\eta \mid y) = \hat{\eta}^0 \\ &* Cov(\eta \mid y) = \Omega \\ &* Var(x - L\hat{\eta}^0) = Var[L(\eta - \hat{\eta}^0) + e] = L\Omega L' + \Sigma_e \\ &* Cov(\eta, x - L\hat{\eta}^0) = E[\eta(x - L\hat{\eta}^0)'] - E[\eta]E[(x - L\hat{\eta}^0)'] \\ &\quad = E[\eta[L(\eta - \hat{\eta}^0)]'] = \Omega L' \end{aligned} \tag{A.3.4}$$

and replacing the corresponding expressions of Equations A.3.4 in Equations A.3.3, it follows that

$$\boxed{\hat{\eta} = E(\eta \mid x - L\hat{\eta}^0, y) = \hat{\eta}^0 + \Omega L'(L\Omega L' + \Sigma_e)^{-1}(x - L\hat{\eta}^0) = \hat{\eta}^0 + \eta_c}$$

with $\eta_c = \Omega L'(L\Omega L' + \Sigma_e)^{-1}(x - L\hat{\eta}^0)$ and

$$\boxed{\Sigma_{\hat{\eta}} = Cov(\eta \mid x - L\hat{\eta}^0, y) = \Omega - \Omega L'(L\Omega L' + \Sigma_e)^{-1}L\Omega}$$

□

Appendix B

Review of ARMA Model Survey Error Variances

It was discussed in page 73 that $\text{Var}(u_t) = 1$. Then, once a model has been chosen for the standardised survey errors u_t, the discrepancies of this model must follow this restriction. In this appendix, the basic ARMA models are discussed with the respective variances of the disturbance term χ_t in the model

$$u_t = \phi_1 u_{t-1} + \cdots + \phi_p u_{t-p} + \chi_t + \theta_1 \chi_{t-1} + \cdots + \theta_q \chi_{t-q} \tag{B.0.1}$$

B.1 MA(q) Model

The MA(q) model given by

$$u_t = \chi_t + \theta_1 \chi_{t-1} + \cdots + \theta_q \chi_{t-q} \tag{B.1.1}$$

is the simplest of the ARMA models. The expectation of u_t is zero with variance given by

$$\gamma(0) = E(u_t^2) = (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma_\chi^2 \tag{B.1.2}$$

Since $\text{Var}(u_t) = 1$, it follows that

$$\sigma_\chi^2 = \frac{1}{1 + \theta_1^2 + \cdots + \theta_q^2} \tag{B.1.3}$$

The values $\theta_1, \cdots, \theta_q$ are hyperparameters being estimated by maximum likelihood prior to the filtering and smoothing of the series of study.

If $q = 1$, then using Equation 3.4.16 and

$$T = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, Q = \begin{bmatrix} 1 & \theta_1 \\ \theta_1 & \theta_1^2 \end{bmatrix} \sigma_x^2 \tag{B.1.4}$$

it follows that

$$\text{vec}(P_0) = [I_r - T \otimes T]^{-1} \text{vec}(Q) \tag{B.1.5}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \theta \\ \theta \\ \theta^2 \end{bmatrix} \sigma_x^2 \tag{B.1.6}$$

$$= \begin{bmatrix} 1+\theta^2 \\ \theta \\ \theta \\ \theta^2 \end{bmatrix} \sigma_x^2 \tag{B.1.7}$$

and then

$$P_0 = \begin{bmatrix} 1 & \frac{\theta}{1+\theta^2} \\ \frac{\theta}{1+\theta^2} & \frac{\theta^2}{1+\theta^2} \end{bmatrix} \tag{B.1.8}$$

B.2 AR(1) Model

The AR(1) process is

$$u_t = \phi u_{t-1} + \chi_t; \qquad t = 1, \cdots, n; \quad -1 < \phi < 1 \tag{B.2.1}$$

The expectation of u_t is zero for all t, while its variance is given by

$$\gamma(0) = E(u_t^2) = E(\sum_{j=0}^{\infty} \phi^j \chi_{t-j})^2 = \sum_{j=0}^{\infty} \phi^{2j} E(\chi_{t-j}^2) = \sigma_\chi^2 \sum_{j=0}^{\infty} \phi^{2j} = \sigma_\chi^2/(1-\phi^2) = 1 \tag{B.2.2}$$

Then because it is assumed that $\text{Var}(u_t) = 1$,

$$\sigma_\chi^2 = 1 - \phi^2 \tag{B.2.3}$$

The value ϕ is a hyperparameter and is estimated using maximum likelihood before to start the Kalman filter and the parameter σ_χ^2 is a function of ϕ. Using Equation 3.4.16 and $T = \phi$ and $Q = \sigma_\chi^2$ it follows that $P_0 = \frac{1}{1-\phi^2}\sigma_\chi^2 = 1$

B.3 AR(2) Model

The next model is an AR(2) given by

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \chi_t \qquad t = 1, \cdots, n \tag{B.3.1}$$

Now, the expectation of u_t is zero for all t, while

$$\begin{aligned} \gamma(0) = E(u_t^2) &= \phi_1 E(u_t u_{t-1}) + \phi_2 E(u_t u_{t-2}) + E(\chi_t u_t) \\ &= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \sigma_\chi^2 \end{aligned} \tag{B.3.2}$$

Also,

$$\begin{aligned} \gamma(1) = E(u_t u_{t-1}) &= \phi_1 E(u_{t-1} u_{t-1}) + \phi_2 E(u_{t-1} u_{t-2}) + E(u_{t-1} \chi_t) \\ &= \phi_1 \gamma(0) + \phi_2 \gamma(1) \end{aligned} \tag{B.3.3}$$

and

$$\begin{aligned} \gamma(2) = E(u_t u_{t-2}) &= \phi_1 E(u_{t-1} u_{t-2}) + \phi_2 E(u_{t-2}^2) + E(u_{t-2} \chi_t) \\ &= \phi_1 \gamma(1) + \phi_2 \gamma(0) \end{aligned} \tag{B.3.4}$$

Dividing Equations B.3.2 - B.3.4 by $\gamma(0)$, a well known system of equations known as the Yule-Walker equations is derived with

$$1 = \frac{\sigma_\chi^2}{1 - \phi_1 \rho_1 - \phi_2 \rho_2} \tag{B.3.5}$$

and the values of ρ_1 and ρ_2 being the solution of the system equations

$$\begin{cases} \rho_1 &= \phi_1 + \phi_2\rho_1 \\ \rho_2 &= \phi_1\rho_1 + \phi_2 \end{cases} \Rightarrow \begin{cases} (1-\phi_2)\rho_1 &= \phi_1 \\ -\phi_1\rho_1 + \rho_2 &= \phi_2 \end{cases} \tag{B.3.6}$$

Solving this system of equations, it follows that $\rho_1 = \frac{\phi_1}{1-\phi_2}$ and $\rho_2 = \frac{\phi_1^2+(1-\phi_2)\phi_2}{1-\phi_2}$. Then since $\mathrm{Var}(u_t) = 1$, it follows that

$$1 = \frac{(1-\phi_2)\sigma_x^2}{(1+\phi_2)((1-\phi_2)^2-\phi_1^2)} \tag{B.3.7}$$

and then,

$$\sigma_x^2 = \frac{(1+\phi_2)((1-\phi_2)^2-\phi_1^2)}{(1-\phi_2)} \tag{B.3.8}$$

and σ_x^2 is a function of the hyperparameters ϕ_1 and ϕ_2 following the stationary restrictions (see for example Shumway and Stoffer (2006), page 97)

$$\phi_1 + \phi_2 < 1 \tag{B.3.9}$$

$$\phi_2 - \phi_1 < 1 \tag{B.3.10}$$

$$|\phi_2| < 1 \tag{B.3.11}$$

B.4 AR(p) Model

In general, considering the AR(p) model given by

$$u_t = \phi_1 u_{t-1} + \cdots + \phi_p u_{t-p} + \chi_t \quad t = 1, \cdots, n \tag{B.4.1}$$

The variance of u_t is given by

$$\gamma(0) = \frac{\sigma_x^2}{1-\phi_1\rho_1-\phi_2\rho_2-\cdots\phi_p\rho_p} \tag{B.4.2}$$

where the values $\rho_1, \cdots, \rho_p$ are given as the solution of the Yule-Walker equations

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1} \\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-2} \\ &\cdots\cdots \\ \rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p \end{aligned} \tag{B.4.3}$$

Making $\text{Var}(u_t) = \gamma(0) = 1$, then it is possible to get the corresponding value of σ_x^2. This value will be written as a function of the hyperparameters $\phi_1, \cdots, \phi_p$.

Example

Considering the final model in Chapter 4, and in particular the SARMA$(1,0)$x$(1,0)_{12}$ model for the survey errors; this model can be seen as an special case of an AR(13) with $\phi_1 = \phi$, $\phi_{12} = \Phi$ and $\phi_{13} = -\phi\Phi$ and $\phi_i = 0$ for $i = 2, \cdots, 11$ (Harvey, 1993, page 136).

Then, the variance of u_t is given by

$$\gamma(0) = \frac{\sigma_x^2}{1 - \phi\rho_1 - \Phi\rho_{12} - \phi\Phi\rho_{13}} \tag{B.4.4}$$

where the values ρ_1, ρ_{12} and ρ_{13} come from the solution of the corresponding system of Yule-Walker equations given by

$$\begin{aligned} \rho_1 &= \phi + \Phi\rho_{11} - \phi\Phi\rho_{12} \\ \rho_2 &= \phi\rho_1 + \Phi\rho_{10} - \phi\Phi\rho_{11} \\ &\cdots\cdots \\ \rho_{13} &= \phi\rho_{12} + \Phi\rho_1 - \phi\Phi \end{aligned} \tag{B.4.5}$$

The solution of the system provides the values

$$\begin{aligned} \rho_1 &= \frac{\phi(1 + \phi^{10}\Phi)}{1 + \phi^{12}\Phi} \\ \rho_{12} &= \frac{\phi^{12} + \Phi}{1 + \phi^{12}\Phi} \\ \rho_{13} &= \frac{\Phi - \phi^{12}\Phi^2 + \phi^{10}\Phi^2 + \phi^{12}}{1 + \phi^{12}\Phi} \end{aligned} \tag{B.4.6}$$

and finally

$$\sigma_x^2 = \frac{-(\Phi - 1)(\Phi + 1)(\phi - 1)(\phi + 1)(\phi^{12}\Phi - 1)}{1 + \phi^{12}\Phi} \tag{B.4.7}$$

B.5 ARMA(1,1) Model

Considering the ARMA(1,1) model given by

$$u_t = \phi_1 u_{t-1} + \chi_t + \theta_1 \chi_{t-1} \tag{B.5.1}$$

Then,

$$\begin{aligned} u_t u_{t-k} &= \phi_1 u_{t-1} u_{t-k} + \chi_t u_{t-k} + \theta_1 \chi_{t-1} u_{t-k} \\ \gamma(k) &= \phi_1 \gamma(k-1) + \gamma_{u\chi}(k) + \theta_1 \gamma_{u\chi}(k-1) \end{aligned} \tag{B.5.2}$$

It must be noticed that

$$\gamma_{u\chi}(k) = 0 \quad \text{if} \quad k > 0 \tag{B.5.3}$$

and following the last line in Equation B.5.2, the next system of equations is obtained

$$\begin{aligned} \gamma(0) &= \phi_1 \gamma(1) + \sigma_\chi^2 + \theta_1 \gamma_{u\chi}(-1) \\ \gamma(1) &= \phi_1 \gamma(0) + \theta_1 \sigma_\chi^2 \end{aligned} \tag{B.5.4}$$

and in particular, since

$$\begin{aligned} \gamma_{u\chi}(-1) = E(u_t \chi_{t-1}) &= \phi_1 E(u_{t-1}\chi_{t-1}) + E(\chi_t \chi_{t-1}) + \theta_1 E(\chi_{t-1}^2) \\ &= (\phi_1 + \theta_1)\sigma_\chi^2 \end{aligned} \tag{B.5.5}$$

then the system of equations can be written as

$$\begin{aligned} \gamma(0) - \phi\gamma(1) &= \sigma_\chi^2(1 + \theta_1(\phi_1 + \theta_1)) \\ \phi_1\gamma(0) + \gamma(1) &= \theta_1 \sigma_\chi^2 \end{aligned} \tag{B.5.6}$$

In the last system of equation, the solution for $\gamma(0)$ is

$$\gamma(0) = 1 = \frac{(1 + \theta_1^2 + 2\phi_1\theta_1)\sigma_\chi^2}{1 - \phi_1^2} \tag{B.5.7}$$

and then, it follows that

$$\sigma_\chi^2 = \frac{1 - \phi_1^2}{1 + \theta_1^2 + 2\phi_1\theta_1} \tag{B.5.8}$$

Making $\text{Var}(u_t) = \gamma(0) = 1$, then it is possible to get the corresponding value of σ_χ^2. This value will be written as a function of the hyperparameters $\phi_1, \cdots, \phi_p$.

Since $p = 1, q = 1$, then using Equation 3.4.16 and

$$T = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix}, Q = \begin{bmatrix} 1 & \theta_1 \\ \theta_1 & \theta_1^2 \end{bmatrix} \sigma_\chi^2 \tag{B.5.9}$$

it follows that

$$\text{vec}(P_0) = [I_r - T \otimes T]^{-1}\text{vec}(Q) \tag{B.5.10}$$

and then (see (Durbin and Koopman, 2001), page 112)

$$P_0 = \begin{bmatrix} 1 & \frac{(1-\phi^2)\theta}{1+\theta^2+2\phi\theta} \\ \frac{(1-\phi^2)\theta}{1+\theta^2+2\phi\theta} & \frac{(1-\phi^2)\theta^2}{1+\theta^2+2\phi\theta} \end{bmatrix} \tag{B.5.11}$$

B.6 ARMA(2,1) Model

Now considering the model ARMA(2,1) model given by

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \chi_t + \theta_1 \chi_{t-1} \tag{B.6.1}$$

Then,

$$\begin{aligned} u_t u_{t-k} &= \phi_1 u_{t-1} u_{t-k} + \phi_2 u_{t-2} u_{t-k} + \chi_t u_{t-k} + \theta_1 \chi_{t-1} u_{t-k} \chi_{t-1} u_{t-k} \\ \gamma(k) &= \phi_1 \gamma(k-1) + \phi_2 \gamma(k-2) + \gamma_{u\chi}(k) + \theta_1 \gamma_{u\chi}(k-1) \end{aligned} \tag{B.6.2}$$

Again, it must be noticed that

$$\gamma_{u\chi}(k) = 0 \quad \text{if} \quad k > 0 \tag{B.6.3}$$

and following the last line in Equation B.6.2, the next system of equations is obtained

$$\begin{aligned} \gamma(0) &= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \sigma_\chi^2 - \theta_1 \gamma_{u\chi}(-1) \\ \gamma(1) &= \phi_1 \gamma(0) + \phi_2 \gamma(1) + \theta_1 \sigma_\chi^2 \\ \gamma(2) &= \phi_1 \gamma(1) + \phi_2 \gamma(0) \end{aligned} \tag{B.6.4}$$

and in particular, since

$$\begin{aligned}\gamma_{ux}(-1) &= \phi_1\sigma_x^2 + \phi_2\gamma ux(-1) + \theta_1\sigma_x^2 \\ &= \frac{\phi_1+\theta_1}{1-\phi_2}\sigma_x^2\end{aligned} \tag{B.6.5}$$

then the system of equations can be written as

$$\gamma(0) - \phi_1\gamma(1) - \phi_2\gamma(2) = \sigma_x^2\left(\frac{1-\phi_2+\theta_1\phi_1+\theta_1^2}{1-\phi_2}\right) \tag{B.6.6}$$

$$\phi_1\gamma(0) + (\phi_2-1)\gamma(1) = -\theta_1\sigma_x^2$$

$$\phi_2\gamma(0) + \phi_1\gamma(1) - \gamma(2) = 0$$

In the last system of equation, the solution for $\gamma(0)$ is

$$\gamma(0) = \frac{(\sigma_x^2(\phi_2 - 2\theta_1\phi_1 - \theta_1\phi_1\phi_2 - \theta_1^2 - 1)}{(\phi_1 - 1 + \phi_2)(\phi_1 + 1 - \phi_2)(1+\phi_2)} \tag{B.6.7}$$

and then, it follows that

$$\sigma_x^2 = \frac{(\phi_1 - 1 + \phi_2)(\phi_1 + 1 - \phi_2)(1+\phi_2)}{(\phi_2 - 2\theta_1\phi_1 - \theta_1\phi_1\phi_2 - \theta_1^2 - 1)} \tag{B.6.8}$$

Making $\text{Var}(u_t) = \gamma(0) = 1$, then it is possible to get the corresponding value of σ_x^2. This value will be written as a function of the hyperparameters $\phi_1, \cdots, \phi_p$.

Appendix C

MPI and ABI - Key Facts

Fact	MPI	ABI
Date Commenced	1958	First held in 1998. It replaced several sector specific annual inquiries.
Statutory/Voluntary	Statutory	Statutory
Frequency	Monthly	Annual
Main Information Collected	* Total Turnover * Number of Employees * Export Turnover * Full-time/part-time and male/female employees (quarterly) * Orders on hand (engineering industries) * Export orders on hand (engineering industries) * New order (engineering industries) * New export orders (engineering industries)	* Total Turnover * Number of Employees * Employment Costs * Purchases of goods and services * Taxes and levies * Stocks * Capital expenditure

Fact	MPI	ABI
Respondents	Businesses in the production sector SIC03	Businesses in the production, construction motor trades, wholesale, retail , catering property financial and services trades sectors.
Sample size	9000 each month	74000 each year
Frame	IDBR	IDBR
Coverage	63% by employment	50% by employment
Method	srs with complete coverage of businesses with employment above a threshold of 150 (or 50 in some industries)	srs with 100% coverage of businesses with employment above a threshold of 250
Target Response	80%	85%
Users of Results	* IoP * UK National Accounts * ONS Labour Market Division	* IoP * UK National Accounts

Source: National Statistics (2006)

Appendix D

Standard Industrial Classification

Main Division	Description
A	Agriculture, Hunting and Forestry
B	Fishing
C	Mining and Quarrying
D	Manufacturing
E	Electricity, Gas and Water Supply
F	Construction
G	Wholesale and Retail Trade: Repair of Motor Vehicles, and Personal Household Goods
H	Hotels and Restaurants
I	Transport, Storage and Communication
J	Financial Intermediation
K	Real Estate, Renting and Business Activities
L	Public Administration and Defence: Compulsory Social Security
M	Education
N	Health and Social Work
O	Other Community, Social and Personal Service Activities
P	Private Households with Employed Persons
Q	Extra-Territorial Organisations and Bodies

Appendix E

GVF models - Example 4.6.1.

The modelling to obtain the standard deviation values of the estimates prior to January 2002 in the MPI example is presented in this Appendix. The final model was obtained using Generalized Variance Functions (Wolter, 1985). Equations 4.6.3 - 4.6.7, which assume a relationship between the relative variance and the estimates were used but also linear and quadratic relationships between standard deviations and variances with the estimates obtained in each period of observation. The model with the best fitting was that one relating the estimators with their corresponding standard deviations.

E.1 Initial GVF Model

According to Equation E.1.1 and using the model

$$\hat{se}.(y_t) = \beta_0 + \beta_1 y_t + \beta_2 y_t^2 + \epsilon \tag{E.1.1}$$

The scatterplot between the estimated standard deviation of the estimates and the estimates shows a positive association in Figure E.1. The correlation between the standard deviation of the estimates and the estimates is equal to 0.6400143.

Using R, version 2.0, the corresponding output for the model 6 is as follows.

Residual standard error: 1334 on 21 degrees of freedom Multiple R-Squared: 0.4157. Adjusted R-squared: 0.3601 F-statistic: 7.471 on 2 and 21 DF. p-value: 0.003544. Null

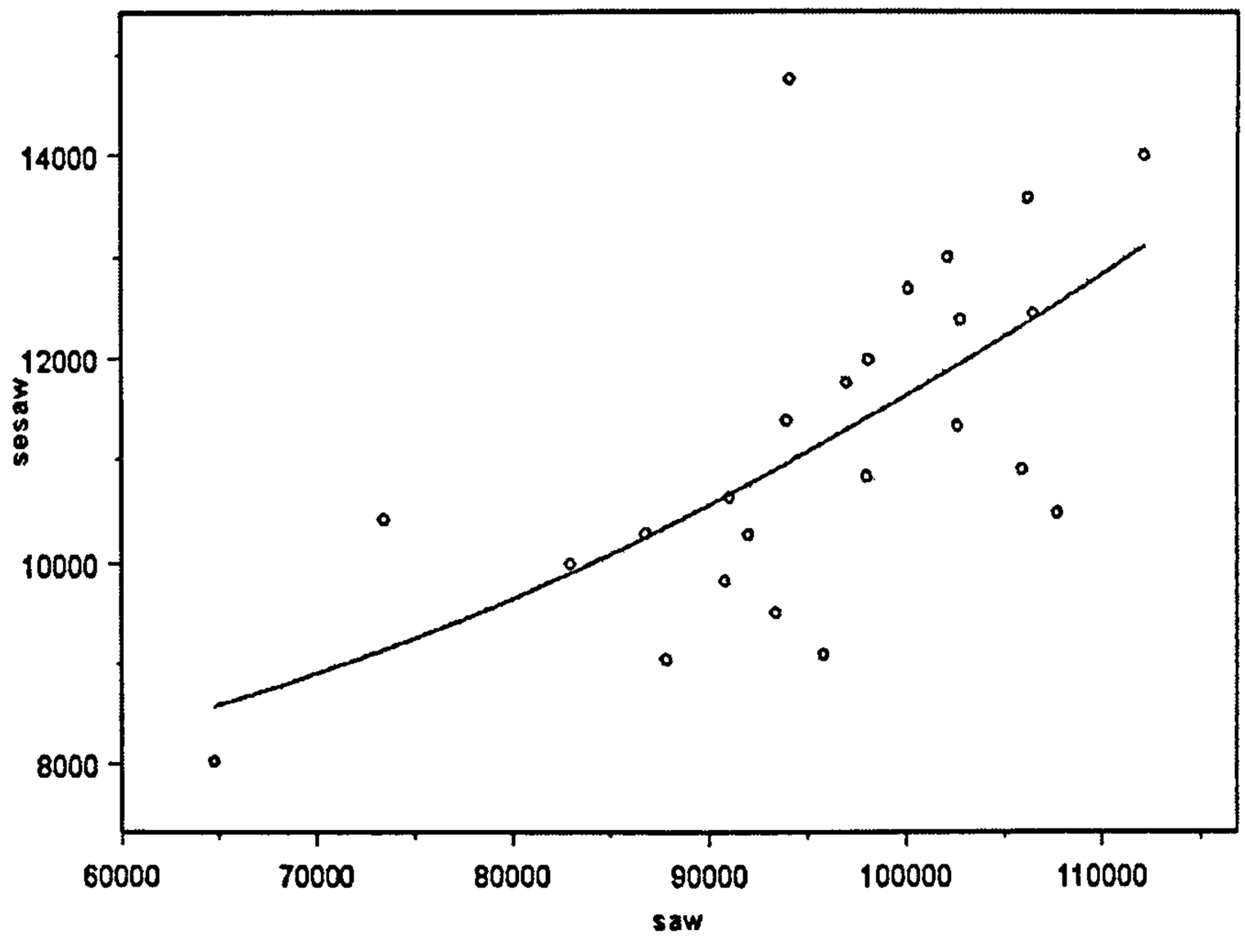


Figure E.1. **Scatterplot estimated standard deviation of the estimates vs. estimates**

Variable	Coefficient	Std.Error	t value	P-value
Intercept	8048.00	13400.00	0.601	0.555
Estimates	-0.04224	0.2998	-0.141	0.889
Estimates^2	0.0000007	0.000001	0.468	0.644

Table E.1. **Coefficients of the initial regression model.**

deviance: 64007455 on 23 degrees of freedom. Residual deviance: 37397973 on 21 degrees of freedom. AIC: 418.33

This is a significant model but the coefficients are not significant. Using a backward procedure, the next model to evaluate corresponds to

$$\hat{se.}(\hat{t}) = \beta_0 + \beta_2 y_t^2 + \epsilon \tag{E.1.2}$$

Variable	Coefficient	Std.Error	t value	P-value
Intercept	6169.00	1297.00	4.757	0.0000951***
Estimates^2	0.0000005	0.0000001	3.952	0.000678***

Table E.2. **Coefficients of the initial quadratic regression model.**

Residual standard error: 1304 on 22 degrees of freedom. Multiple R-Squared: 0.4152. Adjusted R-squared: 0.3886. F-statistic: 15.62 on 1 and 22 DF, p-value: 0.000678. Null

deviance: 64007455 on 23 degrees of freedom. Residual deviance: 37433338 on 22 degrees of freedom. AIC: 416.35

This model looks better as it is significant, all the parameters are significant, and the AIC is smaller than the model before. There was a small lost in terms of R^2 but now all the parameters are significant.

This alternative provides the model given by

$$\hat{s.e.}(y_t) = 6169 + 0.0000005y_t^2 + \epsilon$$

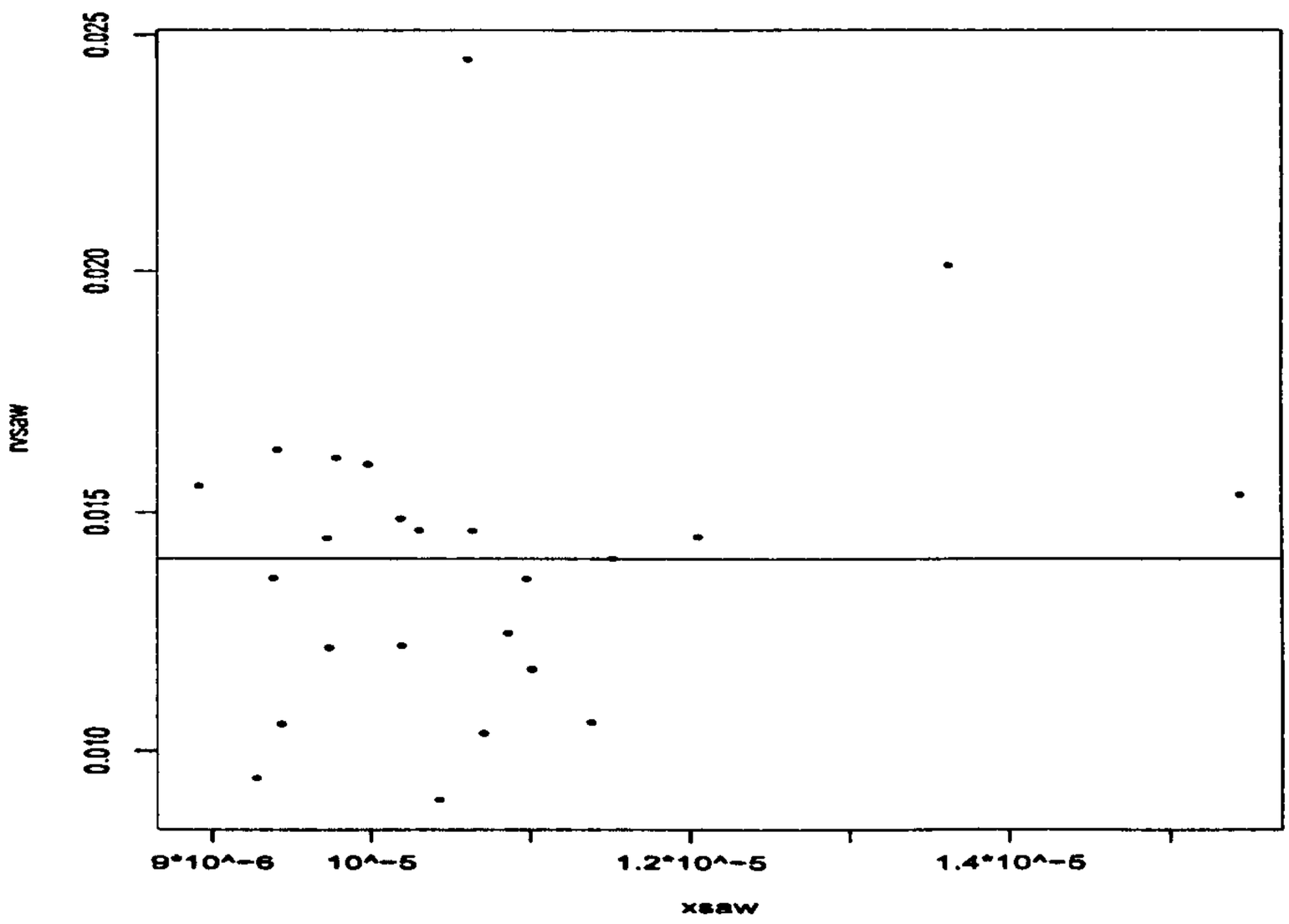


Figure E.2. **Scatterplot relative variance vs. inverse of the estimates**

Since there was a negative association between the relative variance of the estimates and the estimates in this example, the scatterplot between relative variance and the inverse of the estimates has a positive association as it is shown in Figure E.2. The correlation between relative variance and inverse of the estimates is equal to 0.2267701.

Using R, version 2.0, the corresponding output for the model 1 is as follows.

Residual standard error: 0.003381 on 22 degrees of freedom. Multiple R-Squared: 0.05142 Adjusted R-squared: 0.008308 F-statistic: 1.193 on 1 and 22 degrees of freedom, the p-

Variable	Coefficient	Std.Error	t value	P-value
Intercept	0.0084	0.0052	1.6020	0.1234
Inverse(Estimates)	531.3052	486.5000	1.0921	0.2866

Table E.3. **Coefficients of the regression model. Complete model 1**

value is 0.2866. Correlation of Coefficients: Cor(Intercept, Inverse(Estimates))=-0.9912. Null deviance: 0.00026505 on 23 degrees of freedom. Residual deviance: 0.00025142 on 22 degrees of freedom. AIC: -201.09

In conclusion, this is not a good model as it is not significant, neither the intercept nor the slope are significant. Using backward regression, the output above suggest the use of a constant model.

$$V^2 = \beta_0 + \epsilon \tag{E.1.3}$$

Variable	Coefficient	Std.Error	t value	P-value
Intercept	0.0140	0.0007	20.2494	0.0000

Table E.4. **Coefficients of the regression model. Constant model**

Residual standard error: 0.003395 on 23 degrees of freedom. Multiple R-Squared: 2.271e-032 F-statistic: Inf on 0 and 23 degrees of freedom, the p-value is NA. Null deviance: 0.00026505 on 23 degrees of freedom. Residual deviance: 0.00026505 on 23 degrees of freedom. AIC=-201.82

The last model is clearly not a very good model based on the significance of the model or the R^2 value. A model without intercept in the equation above was implemented. However, following (Draper and Smith, 1998, page 27): "The omission of β_0 (the intercept) from a model implies that the response is zero when all the predictors are zero. This is a very strong assumption which is usually unjustified...". They also claims that R^2 does not make any sense in this case because the denominator in the definition of R^2 has a null model with an intercept in mind. Then, the R^2 should not be compared to those models with an intercept. The Akaike information criterion (AIC) is a measure of fit taking into account the parsimony of the model by penalizing for the number of parameters in the model and it will be the measure to be used to compare the fitting of the models.

Under the assumption that $1/\hat{t} = 0$ implies $\hat{V}^2 = 0$, a new alternative is

$$\hat{V}^2 = \beta_1/y_t + \epsilon \tag{E.1.4}$$

which is equivalent to the model

$$\hat{Var}(y_t) = \beta_1 y_t + \epsilon^* \tag{E.1.5}$$

The output for this model is

Variable	Coefficient	Std.Error	t value	P-value
Inverse(estimates)	1303.8599	66.3800	19.6424	0.0000

Table E.5. **Coefficients of the regression model. Model without intercept**

Residual standard error: 0.003494 on 23 degrees of freedom. Multiple R-Squared: 0.9437. Adjusted R-Squared: 0.9413. F-statistic: 385.8 on 1 and 23 degrees of freedom, the p-value is 6.661e-016. Null deviance: 0.00026505 on 23 degrees of freedom. Residual deviance: 0.00026505 on 23 degrees of freedom. AIC=-200.44

The R-Squared is not interpretable here. The AIC value is not smaller than the constant model. The best alternative under this model (although not a very good one) is given by

$$\begin{array}{rlll} \hat{V}^2 = & 0.0140 & + & \epsilon \\ & (66.38 * **) & & \\ \text{Multiple R-Squared:} & 2.271e-032 & \text{AIC} = -201.82 & \end{array}$$

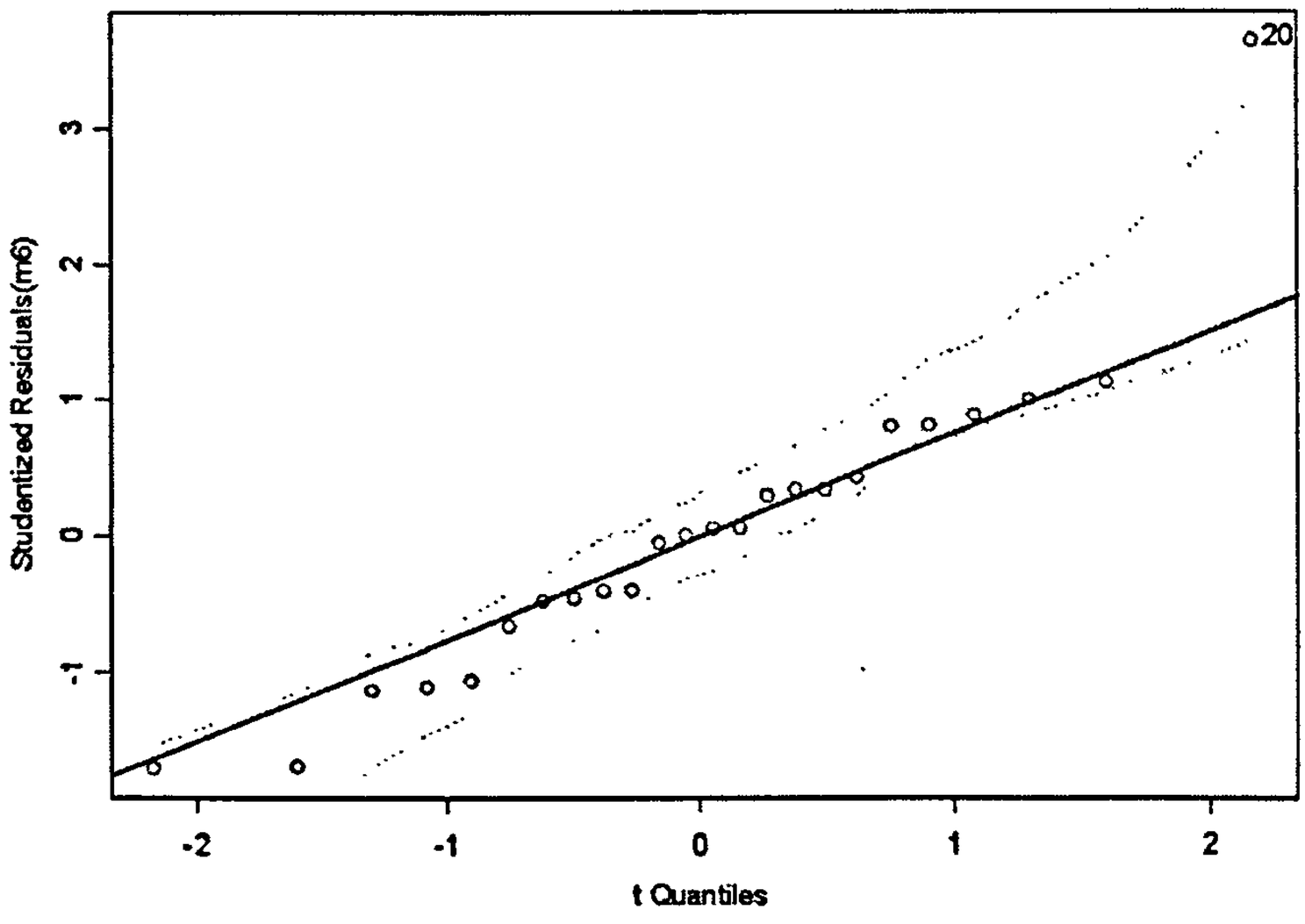


Figure E.3. Residual Diagnostic Plots. QQ Plot. Final Model.

E.2 Final GVF Model

Among the final three possible models, using the R^2 criterion the model finally chosen is given by

$$\begin{array}{|lcccccl|} \hline \hat{se}.(y_t) & = & 6168.73 & + & 5.45e-07y_t^2 & + & \epsilon \\ & & (1296.88 **) & & (1.38e-07 **) & & \\ \hline \end{array}$$

with its corresponding graph in Figure 4.3 in the main document.

Figures E.4 - E.6 summarizes the main diagnostic plots for the residuals of the final model. Other diagnostic tests were presented in Table E.6 in the main document. All the plots does not show evidence of non-normality, autocorrelation or heteroscedasticity. However, Figures E.3 and E.6 show the presence of an outlier corresponding to the observation 20 (August 2003). This observation was not considered into the analysis and a new model was obtained in the next section.

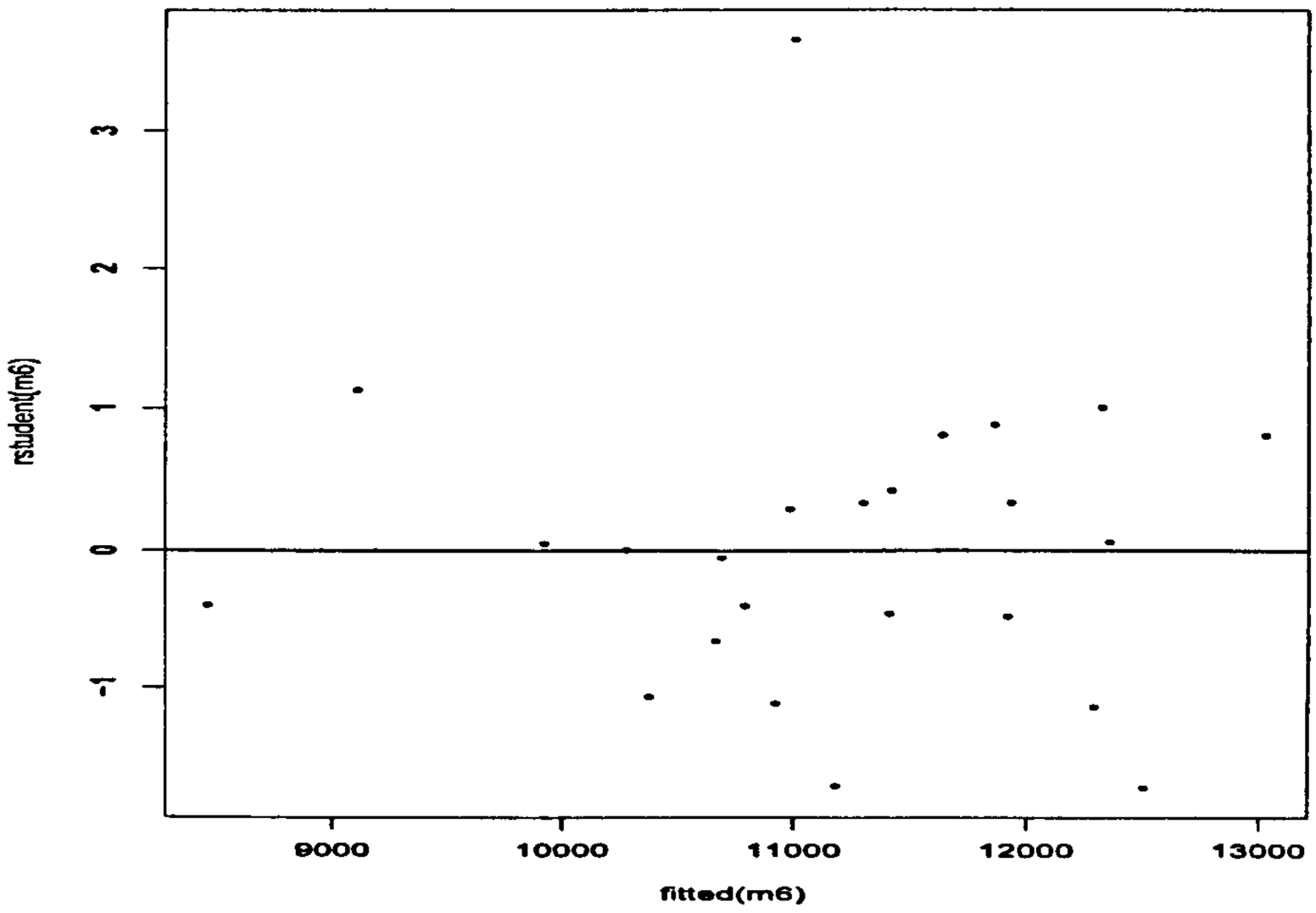


Figure E.4. Standardised Residuals vs Fitted Values Plot. Final Model.

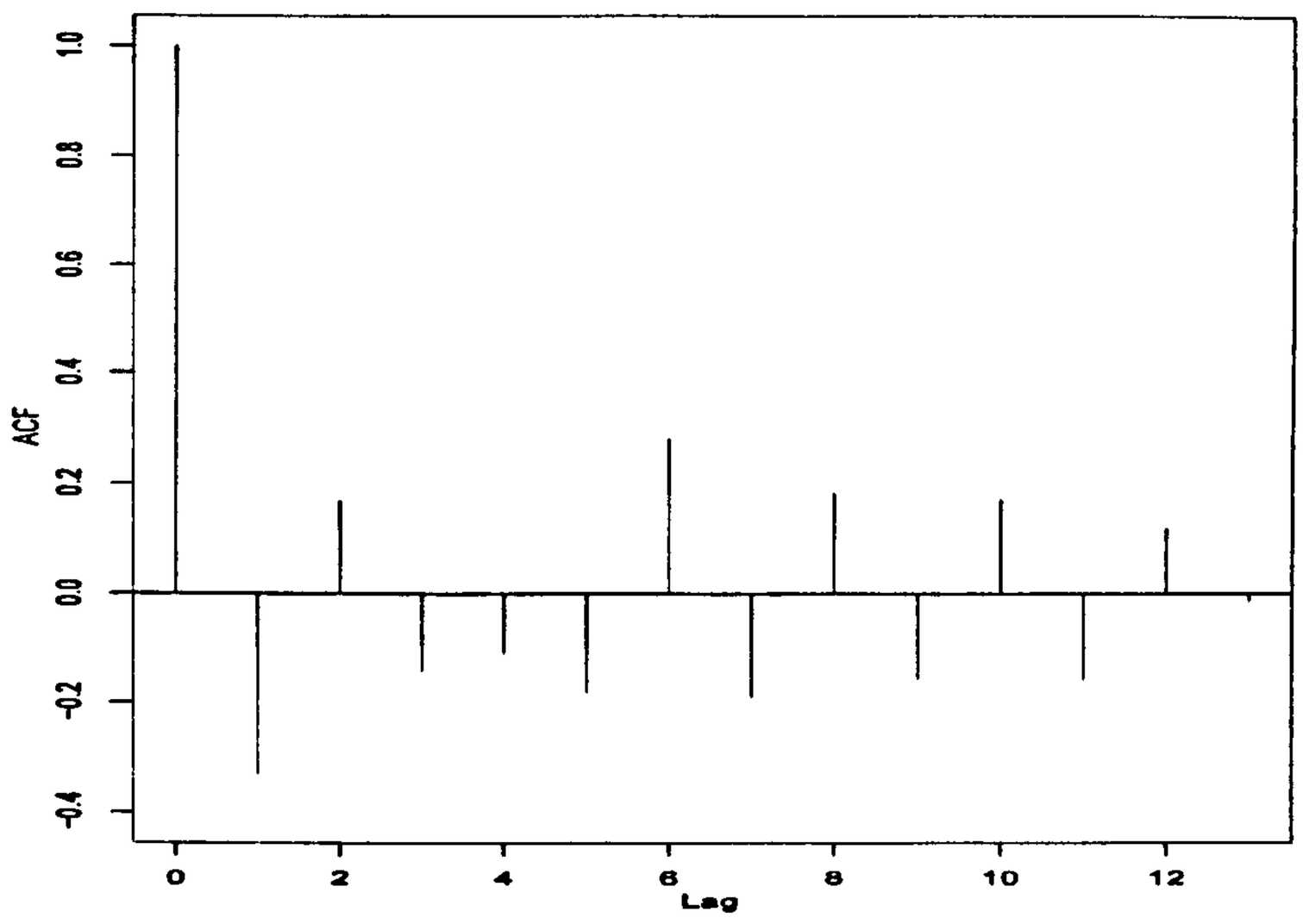


Figure E.5. Autocorrelation Function of Residuals. Final Model.

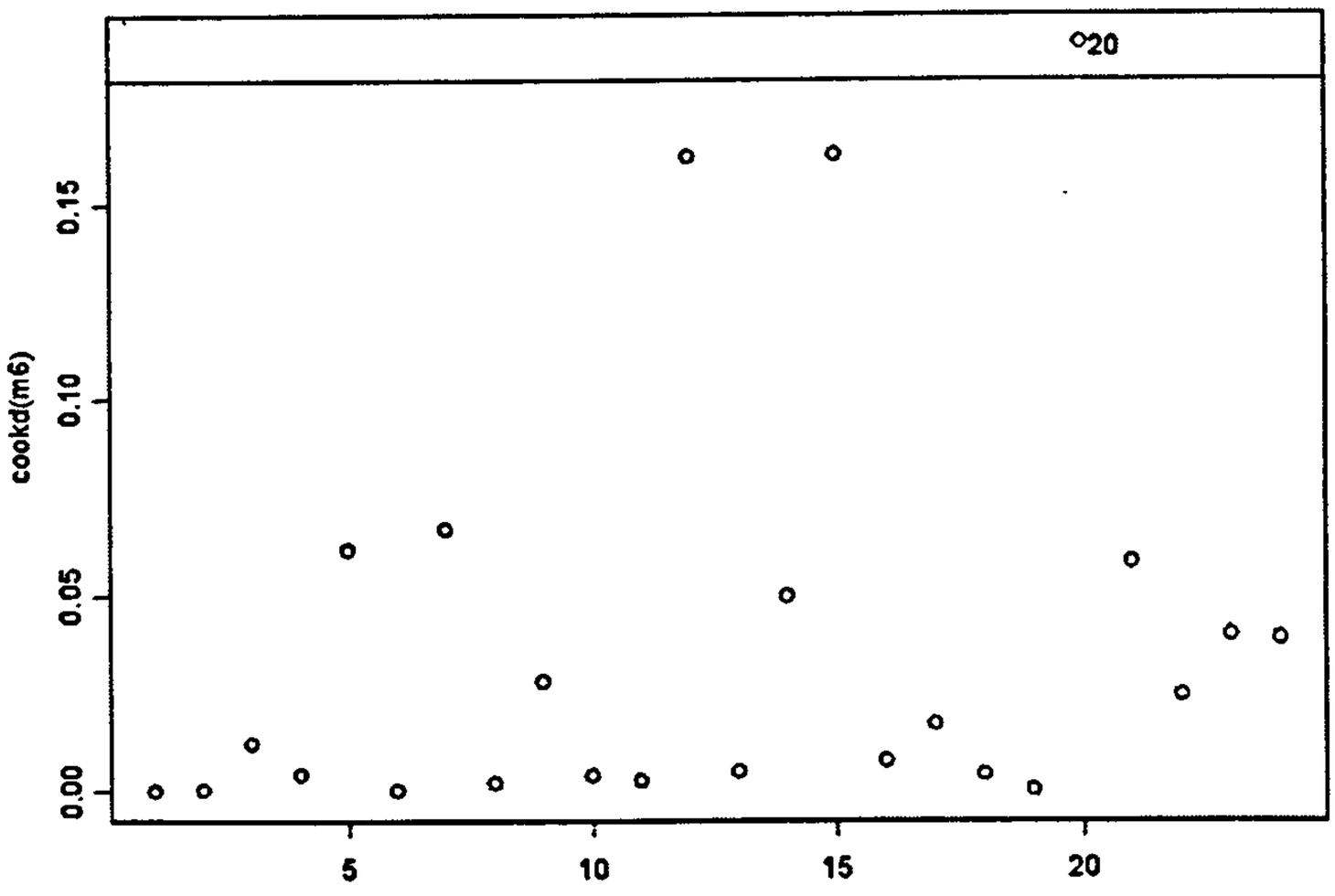


Figure E.6. Cook's Distances Plot. Final Model.

E.2.1 Final Model without Outlier

After deleting the observation 20, which was an outlier according to the diagnostic plots in the last section, the final model follows the equation given by

$$\begin{array}{|lcccccl|} \hline \hat{se.}(y_t) & = & 5878.68 & + & 5.59e-07y_t^2 & + & \epsilon \\ & & (1040.77**) & & (1.10e-07**) & & \\ \hline \end{array}$$

Figures E.7 - E.10 summarizes the main diagnostic plots for the residuals of the final model. Other diagnostic tests were presented in Table E.6 in the main document. All the plots does not show evidence of non-normality, autocorrelation or heteroscedasticity. However Plot E.10 show the presence of two influential observations 12 and 15 (December 2002 and March 2003). Keeping taking out influential observations, finally we got a model without five influential observations 5, 12, 15, 20 and 24 which is presented in the next section.

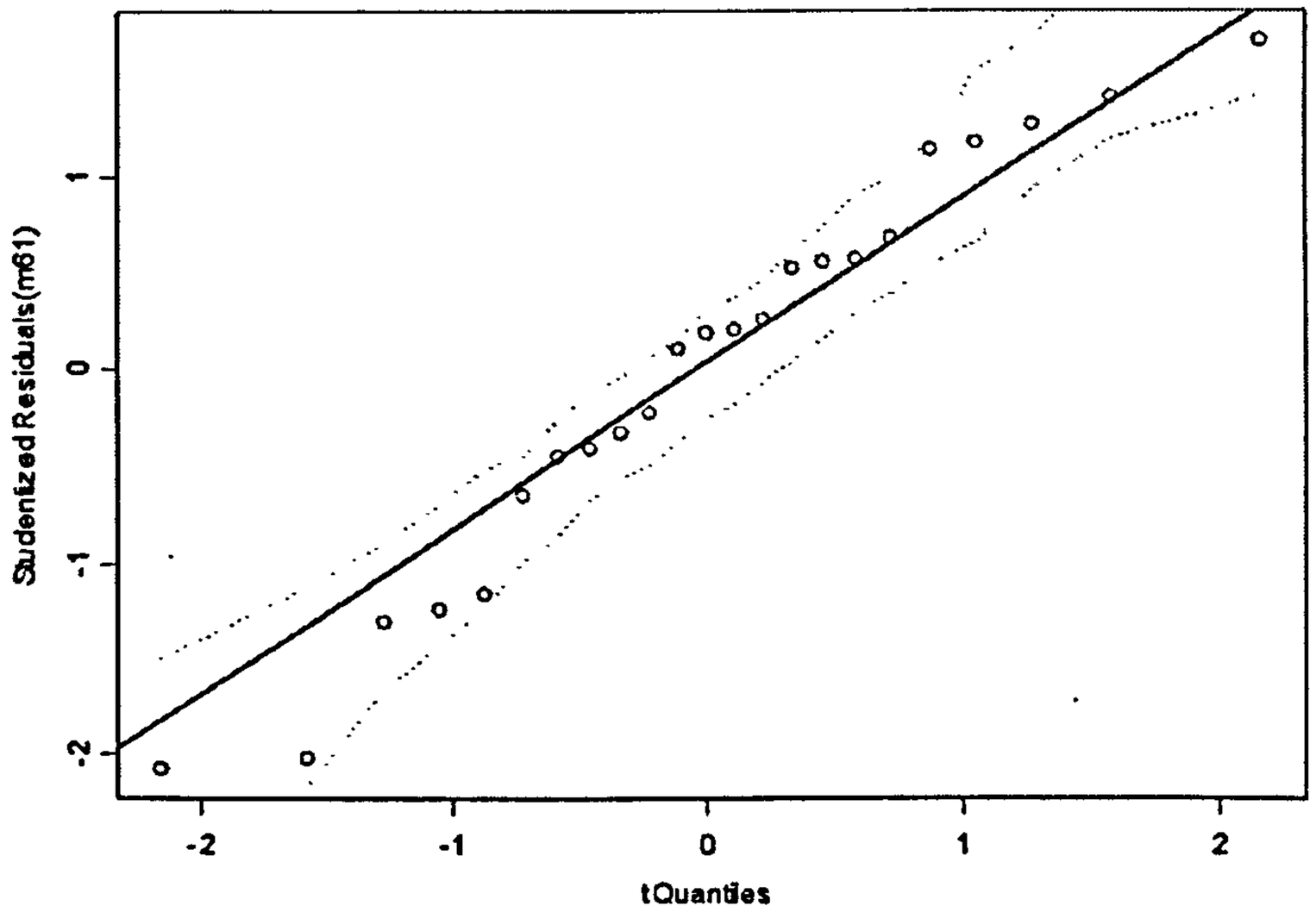


Figure E.7. Residual Diagnostic Plots. QQ Plot. Model without Outlier.

E.2.2 Final Model without Influential Observations

After deleting five influential observations which were detected through the Cook's distance plot, the final model takes the form

$$\begin{array}{|lcccc|} \hline \hat{se}.(y_t) & = & 3159.47 & + & 8.57e-07y_t^2 \quad + \quad \epsilon \\ & & (1357.22**) & & (1.34e-07**) \\ \hline \end{array}$$

Figures E.11 - ?? summarizes the main diagnostic plots for the residuals of the final model. Other diagnostic tests were presented in Table E.6 in the main document. All the plots does not show evidence of non-normality, autocorrelation or heteroscedasticity. There is no presence of outliers or influential points in this case either.

Even though, the model without influential observations has a higher value of R^2 and good performance in the diagnostic plots, it was estimated under 19 observations only. The gain in the adequacy of the model is obtained under the loss of the number of observations and in this particular case, this loss is more than the 20% (5/24) of the

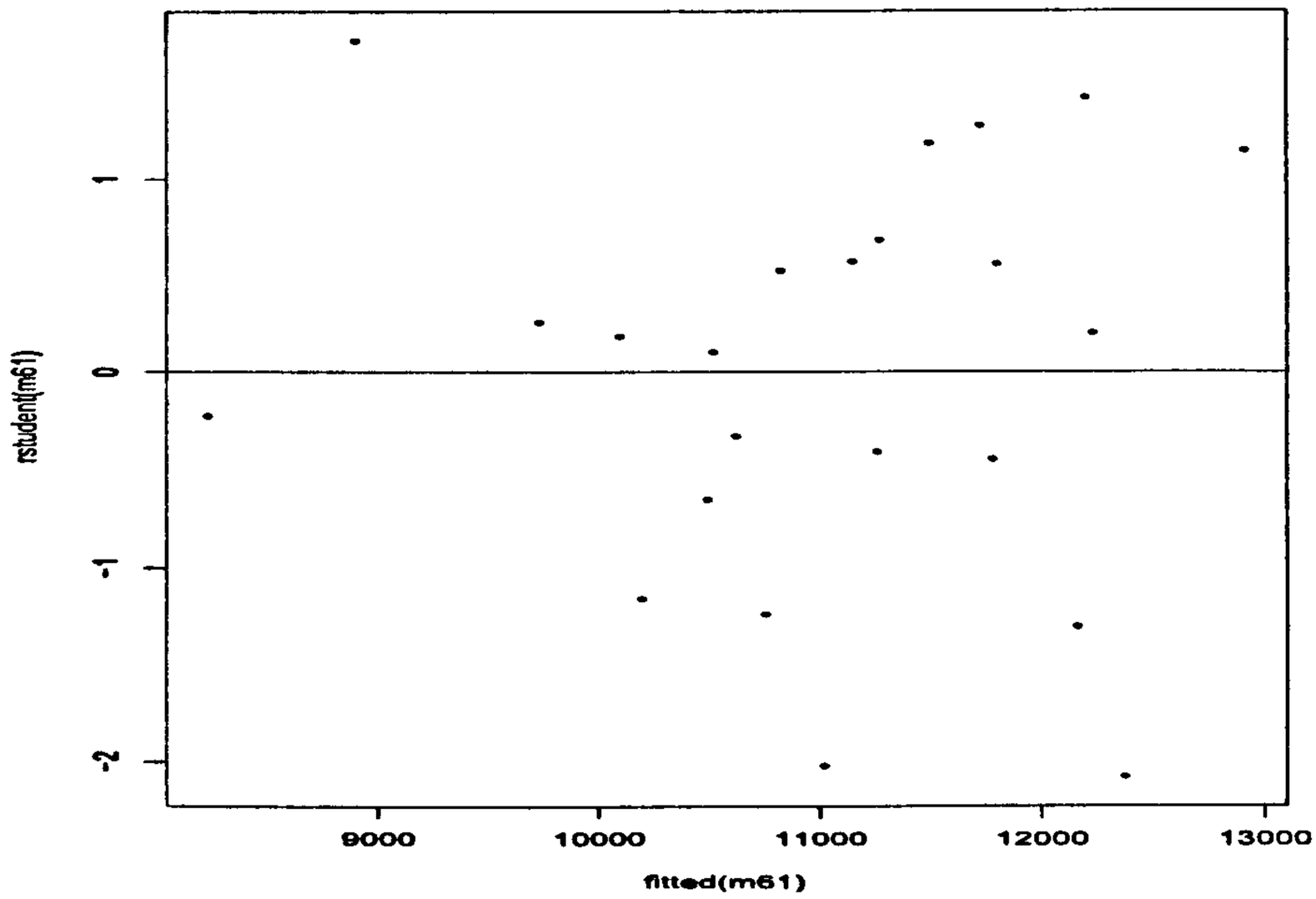


Figure E.8. **Standardised Residuals vs Fitted Values. Model without Outlier.**

original number of observations. To evaluate if the loss of information compensates the gain in the fitness of the model other criteria were used to evaluate the two final possible models. It is therefore desirable to have some quantitative measures to evaluate the performance of both models, one without the outlier and the other without the outlier and the four influential observations. If the improvement is too little, it would be better to keep the influential observations in the dataset as the loss of information is quite considerable. Four different measures of performance were considered with n denoting the number of observations with values 24, 23 or 19 in the respective models:

(a) *Mean Square Error.*

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{E.2.1}$$

(b) *Mean Percent Error.*

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{y_t - \hat{y}_t}{y_t} \tag{E.2.2}$$

(c) *Mean Absolute Percentage Error.*

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{E.2.3}$$

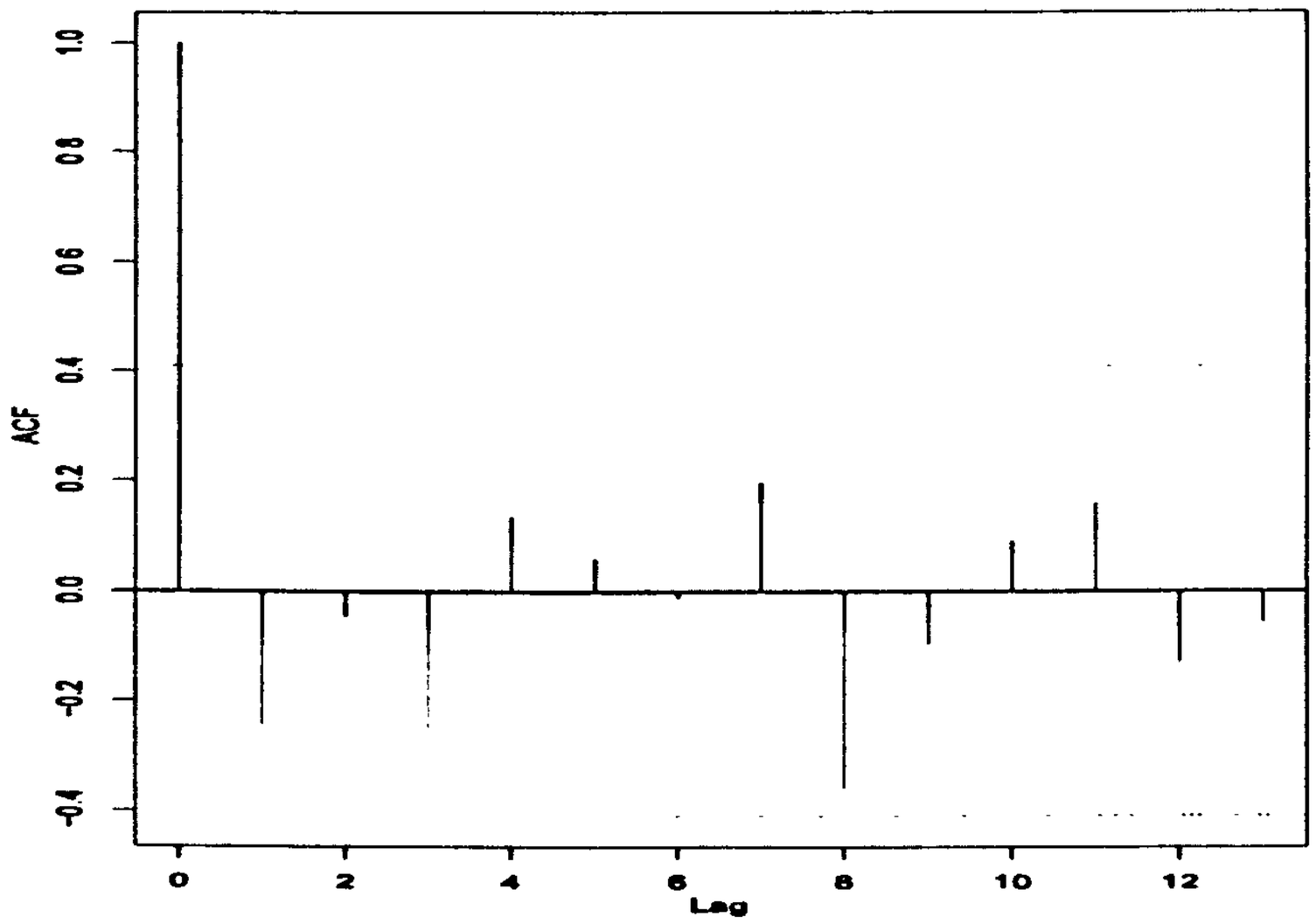


Figure E.9. **Autocorrelation Function of Residuals. Model without Outlier.**

(d) *Mean Square Percentage Error.*

$$MSPE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{y_t - \hat{y}_t}{y_t}\right)^2 \qquad \text{(E.2.4)}$$

For the model under consideration in Equation 4.6.8, $y_t = s.e.(\hat{\theta})$. The first quantity calculated in Equation E.2.1 is a quantitative measure of how closely the estimated values under a particular model follow the actual values. This is a measure of dispersion and its magnitude can only be evaluated by comparing it with the average size of the variable of study. The last three measures in Equations E.2.2 - E.2.4 make this comparison, in the last two the absolute value and the square of the value are calculated to avoid the problem of positive and negative errors canceling and penalizing large individual errors more heavily (Fair, 1984; Pindyck and Rubinfeld, 1991). According to the results in Table E.7, all the last three relative measures are small (values lower than 10%) and quite similar. Since the results are very close for the two optional models, we will work with the model losing less amount of information that is the model with the outlier in the second column of Table E.7 and given by the Equation in page 210.

A. Regression Estimates			
Coefficients (p-value)	Model	Model (-1 obs)	Model (-5 obs)
Intercept	6168.73***(0.00)	5878.68***(0.00)	3159.47*(0.02)
Estimates^2	5.45e-7***(0.00)	5.59e-7***(0.00)	8.57e-7***(0.00)
B. Diagnostic Tests of the Residuals			
Test Statistics (p-value)			
Shapiro-Wilks	0.94(0.19)	0.95(0.37)	0.92(0.12)
Jarque-Bera	3.72(0.15)	1.13(0.57)	1.80(0.41)
Ljung-Box	15.48(0.27)	7.48(0.82)	13.53(0.41)
Box-Pierce	10.70(0.63)	8.74(0.79)	4.66(0.97)
Durbin-Watson, lag=1	2.66(0.12)	2.49(0.25)	2.58(0.18)
Durbin-Watson, lag=12	0.77(0.32)	1.15(0.67)	0.97(0.56)
Homoscedasticity	0.12(0.72)	0.73(0.39)	0.27(0.60)
C. Fit Indexes			
Index			
R-squared	0.42	0.55	0.71
Adjusted R-squared	0.39	0.53	0.69
AIC	416.35	388.91	313.14

Table E.6. Estimates and Test Diagnostics GVFs.

Statistic	Model (-1 obs)	Model (-5 obs)
MSE	1586731	1911588
MPE	0.0032 (0.3%)	0.0064 (0.6%)
MAPE	0.0847 (8.4%)	0.0954 (9.5%)
MSPE	0.0114 (1.1%)	0.0153 (1.5%)

Table E.7. Mean Square Error Related Measures for Comparison Between Models without One and without Five Observations Respectively.

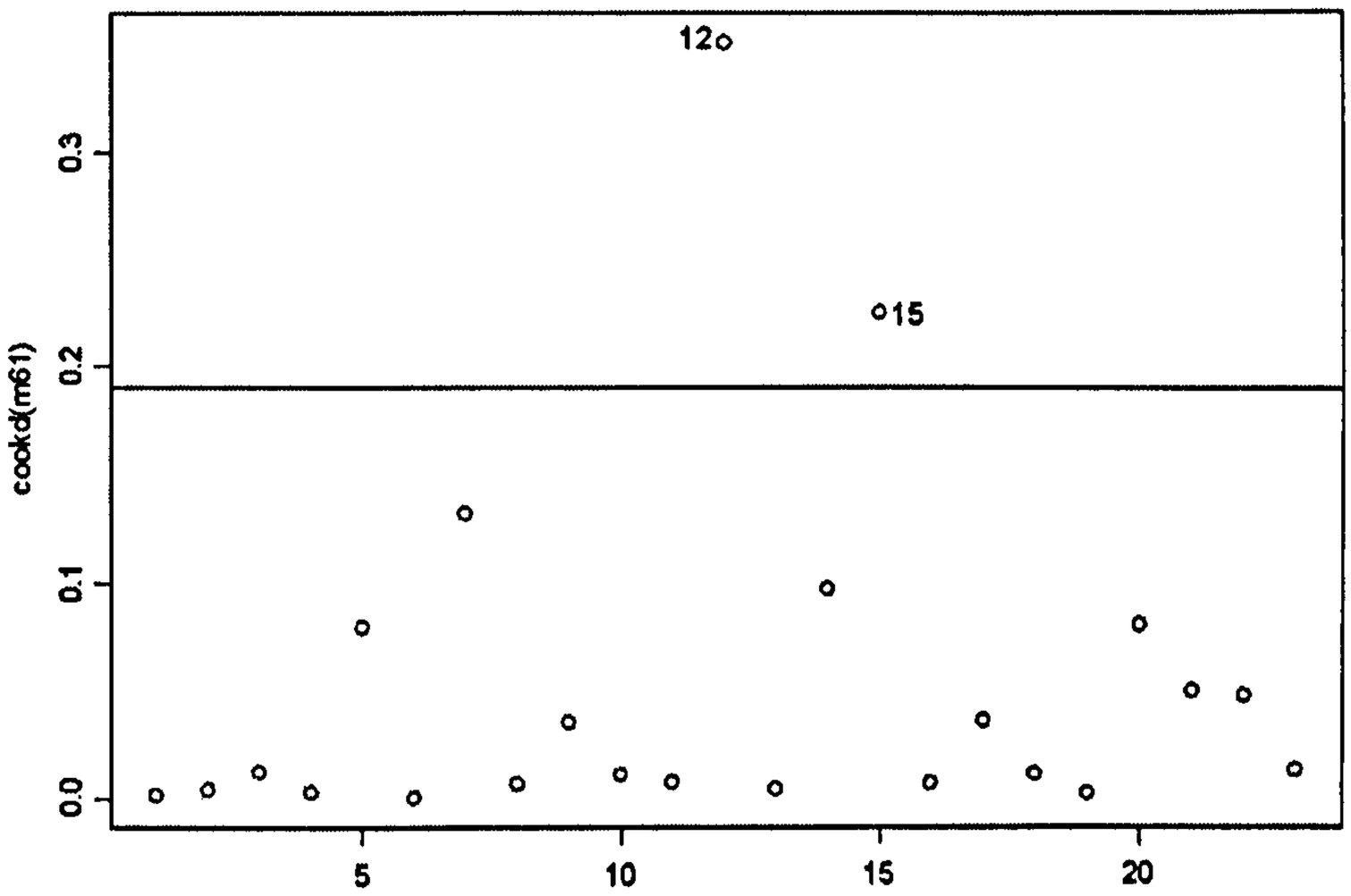


Figure E.10. **Cook's Distances Plot. Model without Outlier.**

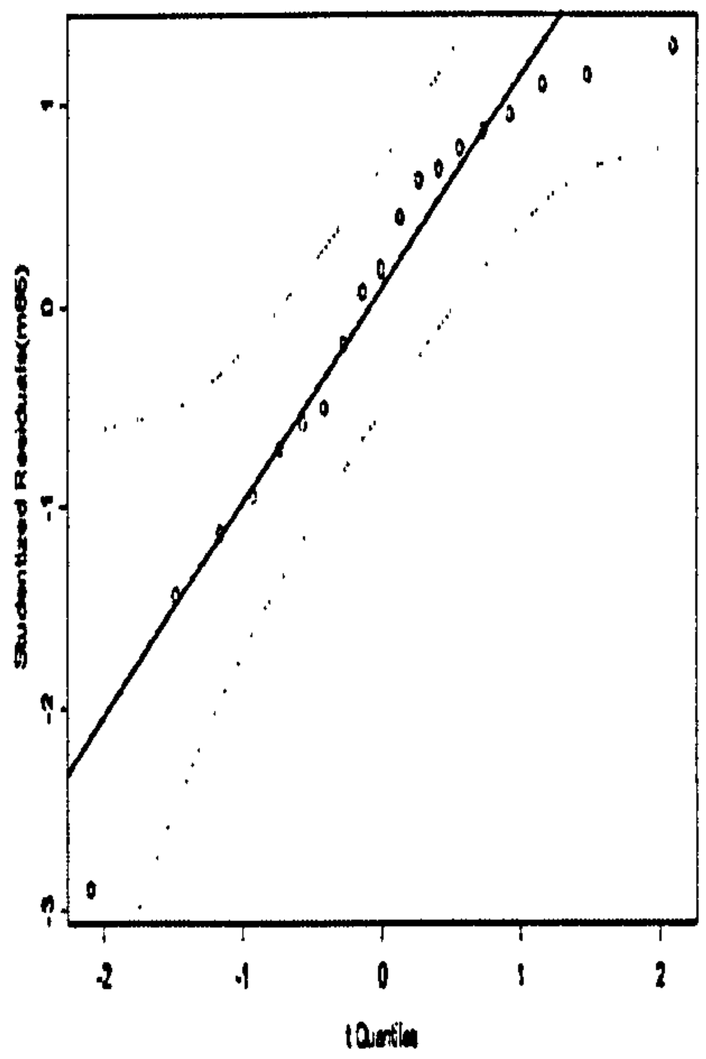


Figure E.11. **QQ Plot and Residuals vs Fitted Values. Model without Influential Points.**

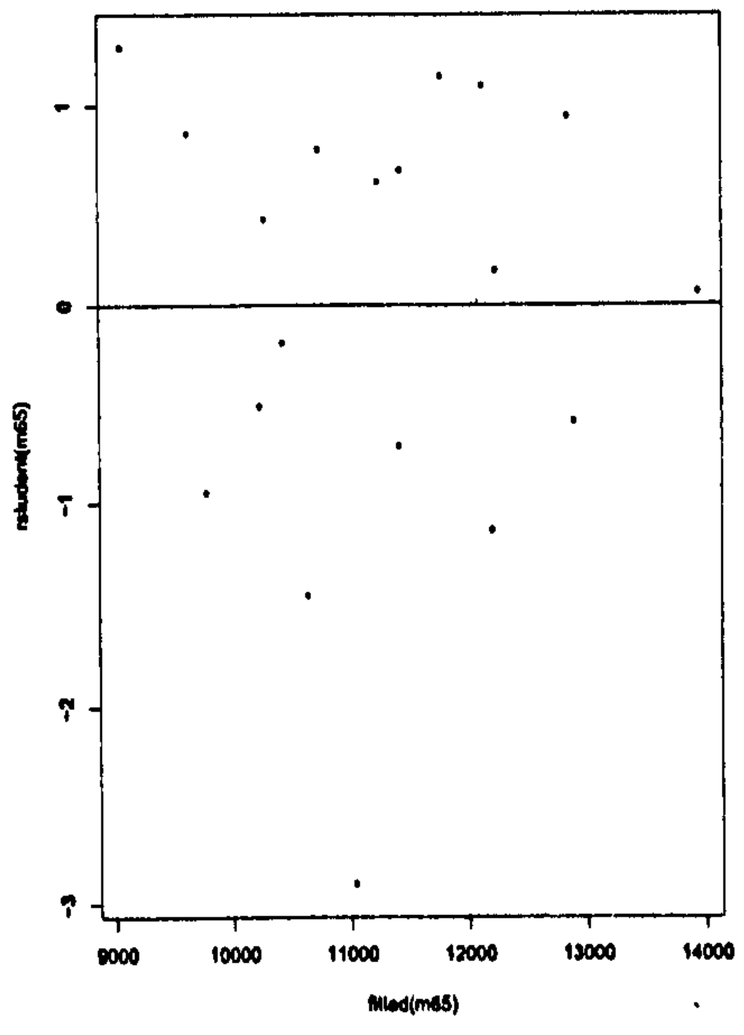


Figure E.12. Standardised Residuals vs Fitted values. Model without Influential Points

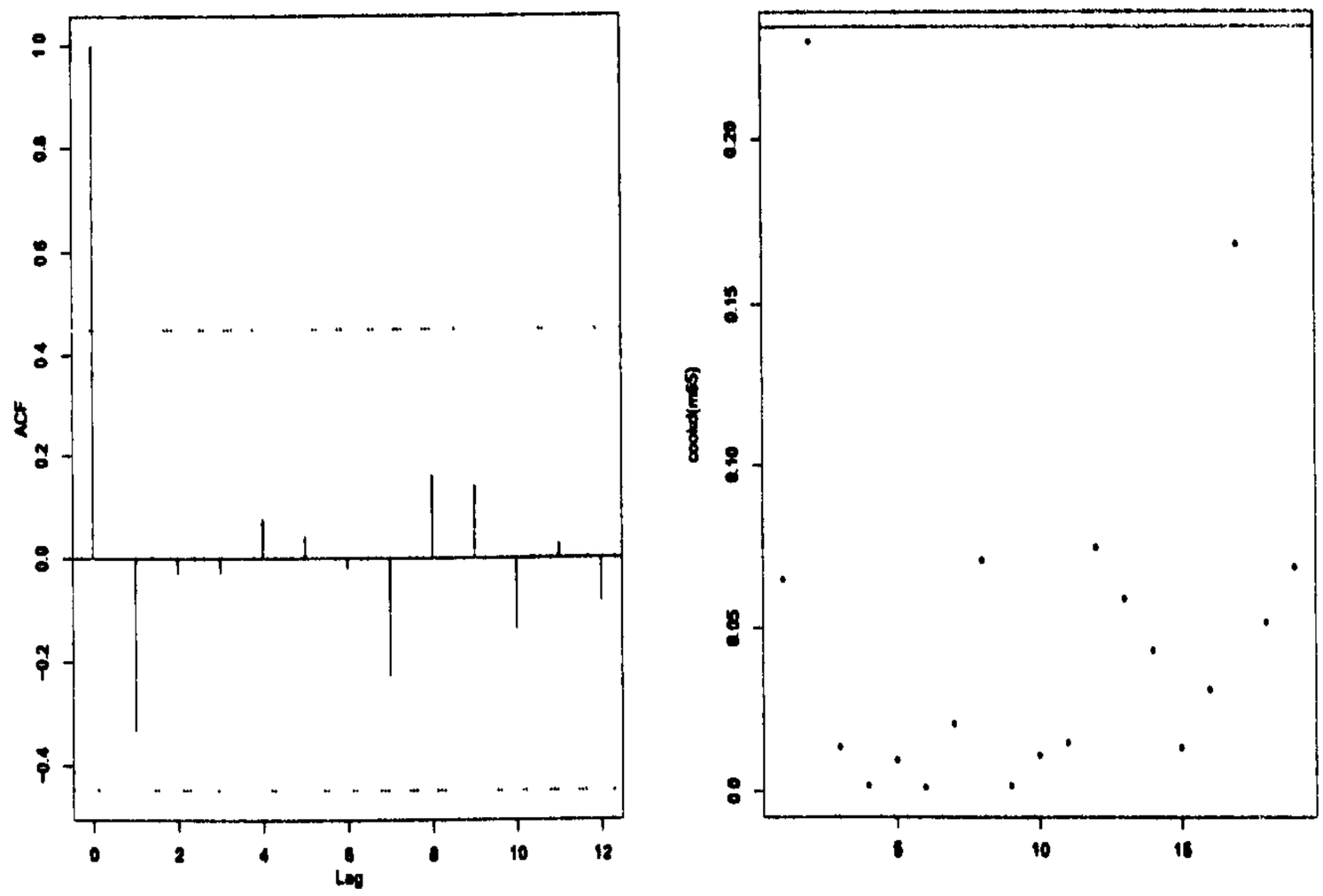


Figure E.13. Autocorrelation Function of Residuals and Cook's Distances. Model without Influential Points.

Bibliography

Abraham, B. (1982), 'Temporal aggregation and time series', *International Statistical Review,* **50**, 285–291.

Abraham, B. and Box, G. (1978), 'Deterministic and forecast-adaptive time-dependent models', *Applied Statistics* **27**, 120–130.

Aitken, A. (1935), On least squares and linear combinations of observations, *in* 'Proceedings of the Royal Society of Edinburgh, series A, 55', pp. 42–48.

Akaike, H. (1975), 'Markovian representation of stochastic processes by canonical variables', *SIAM Journal of Control,* **13**, 162–173.

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

Anderson, B. and Moore, J. (1979), *Optimal Filtering*, Prentice Hall.

Anderson, T. (1971), *The Statistical Analysis of Time Series*, Wiley, New York.

Anderson, T. (1984), *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Stanford.

Ansley, C. and Kohn, R. (1982), 'A geometrical derivation of the fixed interval smoothing algorithm', *Biometrika,* **69**, 486–487.

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

Aoki, M. (1987), *State Space Modelling of Time Series*, Springer-Verlag, New York.

Bailar, B. (1975), 'The effects of rotation group bias on estimates from panel surveys', *Journal of the American Statistical Association,* **70**, 23–30.

Barcellan, R. and Buono, D. (2002), *ECOTRIM Interface. Temporal Disaggregation Techniques.*, Eurostat.

Bell, W. and Hillmer, S. (1983), 'Modelling time series with calendar variation', *Journal of the American Statistical Association,* **78**, 526–534.

Bell, W. and Hillmer, S. (1984), 'Issues involved with the seasonal adjustment of economic time series', *Journal of Business and Economic Statistics,* **2**, 291–320.

Bell, W. and Hillmer, S. (1987*a*), Time series methods for survey estimation, *in* 'Proceedings of the American Statistical Association, Survey Research Methods Section', Washington, USA, pp. 83–92.

Bell, W. and Hillmer, S. (1987*b*), Time series methods for survey estimation, Technical Report CENSUS/SRD/RR-87/20, Bureau of the Census.

Bell, W. and Hillmer, S. (1990), 'The time series approach to estimation for repeated surveys', *Survey Methodology,* **16**(2), 195–216.

Bell, W. and Hillmer, S. (1991), 'Initializing the Kalman filter for nonstationary time series models', *Journal of the American Statistical Association,* **12**, 283–300.

Bickel, P. and Doksum, K. (2001), *Mathematical Statistics. Basic Ideas and Selected Topics*, Vol. 1, Prentice Hall, New Jersey.

Binder, D., Bleuer, S. and Dick, P. (1993), 'Time series methods applied to survey data', *Bulletin of the International Statistical Institute,* **49(1)**, 327–344.

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

Binder, D. and Hidiroglou, M. (1988), Sampling in time, *in* P. K. et.al., ed., 'Handbook of Statistics', Vol. 6, North-Holland Publishing Co, Amsterdam, pp. 187–211.

Bishop, Y., Fienberg, S. and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Massachussetts.

Blight, B. and Scott, A. (1973), 'A stochastic model for repeated surveys', *Journal of the Royal Statistical Society. Series B*, **35**, 61–68.

Bloem, A., Dippelsman, R. and Maehle, N. (2001), *Quarterly National Accounts Manual - Concepts, Data Sources and Compilation*, International Monetary Fund.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics*, **31**, 307–327.

Box, G. E. P. and Pierce, D. A. (1970), 'Distribution of residual correlations in autoregressive-integrated moving average time series models', *Journal of the American Statistical Association*, **65**, 1509–1526.

Box, G. and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, CA.

Box, G., Jenkins, G. and Reinsel, G. (1994), *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, CA.

Brown, R., Durbin, J. and Evans, J. (1975), 'Techniques of testing the constancy of regression relationships over time', *Journal of the Royal Statistical Society. Series B*, **37**, 141–192.

Bureau of Labor Statistics (2002), Current population survey, Technical paper g3rv. design and methodology, US Department of Labor.

Bureau of Labor Statistics (2005), BLS handbook of methods, Technical report, Department of Labor, US. www.bls.gov/opub/hom/pdf/homch4.pdf.

Chanley, V., Rudolph, T. and Rahn, W. (2000), 'The origins and consequences of public trust in government: A time series analysis', *Public Opinion Quarterly*, 64, 239–256.

Chen, Z., Cholette, P. and Dagum, E. (1997), 'A nonparametric method for benchmarking survey data via signal extraction', *Journal of the American Statistical Association,* **92**(440), 1563–1571.

Cholette, P. (1984), 'Adjusting sub-annual series to yearly benchmarks', *Survey Methodology,* **10**, 35–49.

Cholette, P. (1994), User's manual of programme BENCH to benchmark, interpolate and calendarize time series data, Technical Report TSRA-90-008, Statistics Canada, Ottawa.

Cholette, P. and Dagum, E. (1994), 'Benchmarking time series with autocorrelated survey errors', *International Statistical Review,* **62**, 365–377.

Chow, G. and Lin, A. (1971), 'Best linear unbiased interpolation, distribution and extrapolation of time series by related series', *Journal of the American Statistical Association,* **71**(355), 719–721.

Cleveland, W. and Devlin, S. (1980*a*), 'Calendar effects in monthly time series: Detection by spectrum analysis and graphical methods', *Journal of the American Statistical Association,* **75**, 487–496.

Cleveland, W. and Devlin, S. (1980*b*), 'Calendar effects in monthly time series: Modelling and adjustment', *Journal of the American Statistical Association,* **77**, 520–528.

Cleveland, W. and Tiao, G. (1976), 'Decomposition of seasonal time series: A model for the census X-11 program', *Journal of the American Statistical Association,* **71**, 581–587.

Cochran, W. (1977), *Sampling Techniques*, Wiley, New York.

Cox, B. and Chinnappa, B. (1995), Unique features of business surveys, *in* B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge and P. Kott, eds, 'Business Survey Methods', John Wiley and Sons, New York, pp. 153–169.

Dagum, E. and Cholette, P. (2006), *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series Data*, Springer-Verlag, New York.

Dagum, E., Cholette, P. and Chen, Z. (1998), 'A unified view of signal extraction, benchmarking, interpolation and extrapolation of time series', *International Statistical Review,* **66**(3), 245–269.

Dagum, E., Quenneville, B. and Sutradhar, B. (1992), 'Trading day variations multiple regression models with random parameters', *International Statistical Review,* **60**, 57–73.

DANE (2006), Encuesta anual manufacturera, Technical report, Departamento Administrativo Nacional de Estadistica, Bogota, Colombia. www.dane.gov.co/files/investigaciones/fichas/industria/ficha-eam.pdf.

de Jong, P. (1988*a*), 'A cross-validation filter for time series models', *Biometrika,* **75**, 594–600.

de Jong, P. (1988*b*), 'The likelihood for a state space model', *Biometrika,* **75**, 165–169.

de Jong, P. (1989), 'Smoothing and interpolation with the state space model', *Journal of the American Statistical Association,* **84**, 1085–1088.

de Jong, P. (1998), Fixed interval smoothing, Discussion paper, London School of Economics.

de Jong, P. and Chu-Chun-Lin, S. (2003), 'Smoothing with an unknown initial condition', *Journal of Time Series Analysis,* **24**, 141–148.

de Jong, P. and Mackinnon, M. (1988), 'Covariances for smoothed estimates in state space models', *Biometrika,* **75**(3), 601–602.

Deming, W. and Stephan, D. (1940), 'On the least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals of Mathematical Statistics,* **11**, 427–444.

Dempster, A., Laird, N. and Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B*, **39**, 1–38.

Denton, F. (1971), 'Adjustment of monthly and quarterly series to annual totals: An approach based on quadratic minimization', *Journal of the American Statistical Association*, **66**(333), 99–102.

Di Fonzo, T. (1990), 'The estimation of m disaggregated time series when contemporaneous and temporal aggregates are known', *The Review of Economics and Statistics*, **72**(1), 178–182.

Di Fonzo, T. and Marini, M. (2003), Benchmarking systems of seasonally adjusted time series according to Denton's movement preservation principle, Working Paper 2003.09, Universitá degli Studi di Padova. Dipartamento di Scienze Statistiche.

Di Fonzo, T. and Marini, M. (2005), Benchmarking a system of time series: Denton's movement preservation principle vs. data based procedure, *in* 'Workshop on Frontiers in Benchmarking Techniques and Their Application to Official Statistics', Luxembourg.

Draper, N. and Smith, H. (1998), *Applied Regression Analysis*, Wiley Series in Probability and Mathematical Statistics.

Duncan, G. and Kalton, G. (1988), 'Issues of design and analysis of surveys across time', *International Statistical Review*, **55**, 97–117.

Durbin, J. (2000), 'The foreman lecture: the state space approach to time series analysis and its potential for official statistics', *Australian and New Zealand Journal of Statistics*, **42**, 1–23.

Durbin, J. and Cordero, M. (1993), Handling structural shifts, outliers and heavy-tailed distributions in state space time series models, Working paper. statistics research division, US Bureau of the Census.

Durbin, J. and Koopman, S. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford, UK.

Durbin, J. and Quenneville, B. (1997), 'Benchmarking by state space models', *International Statistical Review,* **65c**, 23–48.

Engle, R. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations', *Econometrica,* **50**, 987–1007.

Erikson, R. and Wlezien, C. (1999), 'Presidential polls as a time series. The case of 1996', *Public Opinion Quarterly,* **63**, 163–177.

European Legislation Council (1993), Council regulation EEC 696/93, Technical report, Official Journal pp. 1-11. European Legislation Council Regulation http://forum.europa.eu.int/irc/dsis/bmethods/info/data/new/696-93en.htm.

Fair, R. (1984), *Specification, Estimation and Analysis of Macroeconometric Models*, Harvard University Press, Cambridge, USA.

Faliva, M. and Zoia, M. (2002), 'On a partitioned inversion formula having useful applications in econometrics', *Econometric Theory,* **18**, 525–530.

Farmeir, L. (1992), 'Posterior mode estimation by extended kalman filtering for multivariate dynamic generalised linear models', *Journal of the American Statistical Association,* **87**, 501–509.

Fernandez, R. (1981), 'A methodological note on the estimation of time series', *The Review of Economics and Statistics,* **63**, 471–478.

Fienberg, S. (1970), 'An iterative procedure for estimation in contingency tables', *The Annals of Mathematical Statistics* **41**(3), 907–917.

Finkner, A. and Nisselson, H. (1978), Some statistical problems associated with continuing cross-sectional surveys, *in* N. Namboodiri, ed., 'Survey Sampling and Measurement', Academic Press, New York, pp. 201–216.

Fortier, S. and Quenneville, B. (2006), Statistics Canada experience towards new standards for time series processing, *in* 'Conference on Seasonality, Seasonal Adjustment and their Implications for Short-Term Analysis and Forecasting', Luxembourg.

Fox, J. (2002), *An R and S-plus Companion to Applied Regression*, Sage Publications.

Freeman, J., Houser, D., Kellstedt, P. and Williams, J. (1998), 'Memoired processes, unit roots and casual inference in political science', *American Journal of Political Science*, **42**, 1289–1327.

Friedman, M. (1962), 'The interpolation of time series by related series', *Journal of the American Statistical Association*, pp. 729–757.

Gardner, G., Harvey, A. and Phillips, G. (1980), 'An algorithm for exact maximum likelihood estimation of autoregressive moving average models by means of Kalman filtering', *Applied Statistics*, **29**, 311–322.

Ghangurde, P. (1982), Rotation group bias in the LFS estimates, *in* 'Proceedings of the American Statistical Association, Survey Research Methods Section', Washington, USA, pp. 421–426.

Ginsburgh, V. (1973), 'A further note on the derivation of quarterly figures consistent with annual data', *Applied Statistics*, **22**, 368–374.

Godolphin, E. and Johnson, S. (2003), 'Decomposition of time series dynamic linear models', *Journal of Time Series Analysis*, **24**, 513–527.

Godolphin, E. and Stone, J. (1980), 'On the structural representation for polynomial-projecting predictor models based on the kalman filter', *Journal of the Royal Statistical Society. Series B*, **42**, 35–45.

Golub, G. and van Loan, C. (1996), *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland.

Gubman, Y. and Burck, L. (2005), Benchmarking of Israeli economic time series and seasonal adjustment, *in* 'Workshop on Frontiers in Benchmarking Techniques and Their Application to Official Statistics', Luxembourg.

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

Guerrero, V. (2003), 'Monthly disaggregation of a quarterly time series and forecasts of its unobservable monthly values', *Journal of Official Statistics,* **19**, 215–235.

Guerrero, V. (2005), Validating a preliminary estimate when temporally disaggregating a time series, *in* 'Workshop on Frontiers in Benchmarking Techniques and Their Application to Official Statistics', Luxembourg.

Guerrero, V. and Martinez, J. (1995), 'A recursive ARIMA-Based procedure for disaggregating a time series variable using concurrent data', *TEST,* **2**, 359–376.

Guerrero, V. and Nieto, F. (1999), 'Temporal and contemporaneous disaggregation of multiple economic time series', *TEST,* 8(2), 459–489.

Hannan, E., Terrell, R. and Tuckwell, N. (1970), 'The seasonal adjustment of economic time series', *International Economic Review,* **11**(1), 24–52.

Hansen, M., Hurwitz, W. and Madow, W. (1953), *Sample Survey Methods and Theory (vol. 1)*, Wiley, New York.

Harrison, P. (1965), 'Short term sales forecasting', *Applied Statistics* **14**, 102–139.

Harvey, A. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, UK.

Harvey, A. (1990), *The Econometric Analysis of Time Series*, LSE Handbooks in Economics, Hertfordshire, UK.

Harvey, A. (1993), *Time Series Models*, Harvester Wheatsheaf, Hertfordshire, UK.

Harvey, A. and Chung, C. (2000), 'Estimating the underlying change in unemployment', *Journal of the Royal Statistical Society,* **163**(3), 303–339.

Harvey, A. and Peters, S. (1990), 'Estimation procedures for structural time series models', *Journal of Forecasting* **9**, 89–108.

Harvey, A. and Pierse, G. (1984), 'Estimating missing observations in economic time series', *Journal of the American Statistical Association,* **79**(385), 125–131.

Harvey, A., Ruiz, E. and Shephard, N. (1994), 'Multivariate stochastic variance models', *Review of Economic Studies,* **61**, 247–264.

Harville, D. (1997), *Matrix Algebra from a Statistician's Perspective*, Springer-Verlag, New York.

Hedlin, D., Pont, M. and Fenton, T. (2001), Estimating the effects of birth and death lags on business register, *in* 'Proceedings of The Second International Conference on Establishment Surveys ICES II', ASA, Alexandria, USA, pp. 1099–1104.

Hidiroglou, M. and Srinath, K. (1993), 'Problems associated with designing subannual business surveys', *Journal of Business and Economic Statistics,* **11**, 397–405.

Hillmer, S. and Tiao, G. (1982), 'An ARIMA model based approach to seasonal adjustment', *Journal of the American Statistical Association,* **77**(377), 63–70.

Hillmer, S. and Trabelsi, A. (1987), 'Benchmarking of economic time series', *Journal of the American Statistical Association,* **82**(400), 1064–1071.

Holt, D. and Skinner, C. (1998), The transition from annual to quarterly labour force surveys, *in* 'The Community Labour Force Survey in the 1990s', Statistical Office of the European Communities, Luxembourg, pp. 345–376.

Janacek, G. and Swift, L. (1993), *Time Series: Forecasting, Simulation, Applications*, Ellis Horwood Series in Mathematics and its Applications, Chichester, UK.

Jarque, C. and Bera, A. (1980), 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals', *Economic Letters* **6**, 255–259.

Jazwinski, A. (1970), *Stochastic Processes and Filtering Theory*, Academic Press.

Johnson, E. and King, B. (1987), 'Generalized variance functions for a complex sample survey', *Journal of Official Statistics,* **3**, 235–250.

Jones, G. (2000), 'The development of the annual business inquiry', *Economic Trends,* **564**, 49–57.

Jones, R. (1980), 'Best linear unbiased estimators for repeated surveys', *Journal of the Royal Statistical Society. Series B,* **42**, 221–226.

Kalman, R. (1960), 'A new approach to linear filtering and prediction problems', *Journal of Basic Engineering, Transactions ASME, Series D* **82**, 35–45.

Kohn, R. and Ansley, C. (1983), 'Fixed interval estimation in state space models when some of the data are missing or aggregated', *Biometrika,* **70**(3), 683–688.

Kohn, R. and Ansley, C. (1989), 'A fast algorithm for signal extraction, influence and cross-validation', *Biometrika,* **76**, 65–79.

Kokic, P. and Jones, T. (1998), Comparison of matched pairs and ratio estimators, *in* 'Proceedings of the Symposia New Directions in Surveys and Censuses', Statistics Canada, Ottawa, pp. 269–272.

Konschnik, C., Monsour, N. and Detlefsen, R. (1985), Constructing and mantaining frames and samples for business surveys, *in* 'Proceedings of the American Statistical Association, Survey Research Methods Section', Washington, USA, pp. 113–122.

Koopman, S. (1993), 'Disturbance smoother for state space models', *Biometrika,* 80(1), 117–126.

Koopman, S. and Durbin, J. (2003), 'Filtering and smoothing of state vector for diffuse state', *Journal of Time Series Analysis,* **24,1**, 85–98.

Laniel, N. and Fyfe, K. (1990), 'Benchmarking of economic time series', *Survey Methodology,* **16**, 271–277.

Lewington, R. (1995), The role of national accounts and their impact on business surveys, *in* B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge and P. Kott, eds, 'Business Survey Methods', John Wiley and Sons, New York, pp. 153–169.

Ljung, G. and Box, G. (1978), 'On a measure of lack of fit in time series models', *Biometrika,* **66**, 67–72.

Lütkepohl, H. (1984), 'Linear transformations of vector ARMA processes', *Journal of Econometrics*, **26**, 283–293.

Magnus, J. and Neudecker, H. (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley Series in Probability and Mathematical Statistics.

Maitland-Smith, A. (2002), Use of benchmark data to align or derive quarterly / monthly estimates, *in* 'Meeting of the OECD Short - Term Economic Statistics Expert Group', Paris, France.

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

Moauro, F. and Savio, G. (2002), 'Temporal disaggregation using multivariate structural time series', *The Econometrics Journal*, **8**, 214–227.

Mood, A., Graybill, F. and Boes, D. (1974), *Introduction to the Theory of Statistics*, Mc-Graw Hill Series in Probability and Statistics, Fort Collins, Colorado.

Morris, N. and Pfeffermann, D. (1985), 'A Kalman filter approach to the forecasting of monthly time series affected by moving festivals', *Journal of Time Series Analysis*, **5**, 225–268.

Muth, J. (1960), 'Optimal properties of exponentially weighted forecasts', *Journal of the American Statistical Association*, **163**, 49–62.

National Statistics (2001), What exactly is the labour force survey?, Technical report, Office for National Statistics, Newport, UK. http://www.statistics.gov.uk/downloads/theme labour/What exactly is LFS1.pdf.

National Statistics (2003), UK standard industrial classification of economic activities SIC(2003), Technical report, Office for National Statistics, London, UK. http://www.statistics.gov.uk/methods quality/sic/downloads/UK SIC Vol1(2003).pdf.

National Statistics (2004), Report on the quinquennial review of the Annual Business Inquiry, Technical report, Office for National Statistics, Newport, UK. www.nationalstatistics.org.uk/downloads/theme-commerce/ABI-Quinn-rev3.pdf.

National Statistics (2005*a*), Guide to seasonal adjustment with X12ARIMA, Technical report, ONS Methodology and Statistical Development. Time Series Analysis Branch.

National Statistics (2005*b*), Report on the full triennal review of the MPI, Technical report, Office for National Statistics, Newport,UK. www.nationalstatistics.org.uk/downloads/reviews/MPItriennialReport2005.pdf.

National Statistics (2006), Operation 2007. the 2007 revision of the UK standard industrial classification, Technical report, Office for National Statistics, Newport, UK. www.statistics.gov.uk/methods-quality/sic/operation2007.asp.

Neyman, J. (1934), 'On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection', *Journal of the Royal Statistical Society,* **97**, 558–606.

Nieto, F. (1998), 'Ex-post and ex-ante prediction of unobserved economic time series: A case study', *Journal of Forecasting,* **17**, 35–58.

Nieto, F. (2007), 'Ex-post and ex-ante prediction of unobserved multivariate time series: a structural model based approach', *Journal of Forecasting* **26**, 53–76.

Oehlert, G. W. (1992), 'A note on the delta method', *The American Statistician* **46**(1), 27–29.

Ohlsson, E. (1995), Coordination of samples using permanent random numbers, *in* B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge and P. Kott, eds, 'Business Survey Methods', Wiley, New York, pp. 153–169.

Partington, J. (2001), The launch of the annual business inquiry, Technical report, ONS, UK. Labour Market Trends, www.nomisweb.co.uk/articles/ref/ABI-lmt-may2000.pdf.

Perry, J. (1995), 'The inter-departmental business register', *Economic Trends,* **505**, 27–30.

Pfeffermann, D. (1984), 'On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients', *Journal of the Royal Statistical Society. Series B,* **46**(1), 139–148.

Pfeffermann, D. (1991), 'Estimation and seasonal adjustment of population means using data from repeated surveys', *Journal of Business and Economic Statistics,* **9**(2), 163–177.

Pfeffermann, D. and Bleuer, S. (1993), 'Robust joint modelling of labour force series of small areas', *Survey Methodology,* **19**(2), 149–164.

Pfeffermann, D., Feder, M. and Signorelli, D. (1998), 'Estimation of autocorrelations of survey errors with application to trend estimation in small areas', *Journal of Business and Economic Statistics,* **16**(3), 339–348.

Pfeffermann, D. and Tiller, R. (2005), 'Bootstrap approximation to prediction MSE for state-space models with estimated parameters', *Journal of Time Series Analysis,* **26**(6), 893–916.

Pfeffermann, D. and Tiller, R. (2006), 'Small area estimation with state space models subject to benchmark constraints', *Journal of the American Statistical Association,* **101**, 1387–1397.

Pindyck, R. and Rubinfeld, D. (1991), *Econometric Models and Economic Forecasts,* McGraw Hill. Economic Series.

Pitta, M. and Silva, D. (2004), 'Uso de modelos de espaços de estados para a estimação do efeito de vício de grupos de rotação na PME/IBGE', *Revista Brasileira de Estatistica,* **65**(224), 89–121.

Quenneville, B., Fortier, S., Chen, Z. and Latendresse, E. (2006), Recent developments in benchmarking to annual totals in X-12 ARIMA and at Statistics Canada, *in* 'Conference on Seasonality, Seasonal Adjustment and their Implications for Short-Term Analysis and Forecasting', Luxembourg.

Quenneville, B., Huot, G., Cholette, P., Chiu, K. and Di Fonzo, T. (2003), Adjustment of seasonally adjusted series to annual totals, *in* 'Proceedings of the Statistics Canada Symposium'.

Quenneville, B. and Rancourt, E. (2005), Simple methods to restore the additivity of a system of time series, *in* 'Workshop on Frontiers in Benchmarking Techniques and Their Application to Official Statistics', Luxembourg.

Rao, J. (2003), *Small Area Estimation*, John Wiley, New York.

Riviere, P. (2002), 'What make business statistics special?', *International Statistical Review,* **70**, 145–159.

Rosenberg, B. (1973), 'Random coefficients model: the analysis of a cross-section of time series by stochastically convergent parameter regression', *Annals of Economic and Social Measurement,* **2**, 399–428.

Rossi, N. (1982), 'A note on the estimation of disaggregate time series when the aggregate is known', *The Review of Economics and Statistics,* **64**, 695–696.

Särndal, C., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer Verlag, New York.

Scott, A. and Smith, T. (1974), 'Analysis of repeated surveys using time series methods', *Journal of the American Statistical Association,* **69**, 674–678.

Scott, A., Smith, T. and Jones, R. (1977), 'The application of time series methods to the analysis of repeated surveys', *International Statistical Review,* **45**, 13–28.

Shapiro, S. S. and Wilk, M. B. (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika,* **52**, 591–611.

Shumway, R. and Stoffer, D. (1982), 'An approach to time series smoothing and forecasting using the EM algorithm', *Journal of Time Series Analysis,* **3**(4), 253–263.

Shumway, R. and Stoffer, D. (2006), *Time Series Analysis and its Applications- with R Examples*, Springer-Verlag, New York.

Smith, P., Pont, M. and Jones, T. (2003), 'Developments in business survey methodology in the Office for National Statistics', *Journal of the Royal Statistical Society. Series D*, **52**(3), 257–295.

Smith, T. (1978), Principles and problems in the analysis of repeated surveys, *in* N. Namboodiri, ed., 'Survey Sampling and Measurement', Academic Press, New York, pp. 201–216.

Snyder, R. and Saligari, G. (1996), 'Initialization of the Kalman filter with partially diffuse initial conditions', *Journal of Time Series Analysis*, **17**, 409–424.

Srinath, K. (1987), 'Methodological problems in designing continous business surveys: Some Canadian experiences', *Journal of Official Statistics*, **3**, 283–288.

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

Telser, L. (1967), 'Discrete samples and moving sums in stationary stochastic processes', *Journal of the American Statistical Association*, **62**, 484–499.

Tiao, G. and Hillmer, S. (1978), 'Some consideration of decomposition of a time series', *biometrika* **65**, 497–502.

Tikkiwal, B. (1979), Succesive sampling - a review, *in* 'Proceedings of the 42nd Session of the International Statistical Institute', Manila, Philippines, pp. 367–384.

Trabelsi, A. and Hillmer, S. (1990), 'Benchmarking time series with reliable benchmarks', *Applied Statistics*, **39**(3), 367–379.

Tsay, R. (2005), *Analysis of Financial Time Series*, Wiley Series in Probability and Statistics, New Jersey.

Valliant, R. (1987), 'Generalized variance functions in stratified two-stage sampling', *Journal of the American Statistical Association*, **82**, 499–508.

Venables, W. and Ripley, B. (2002), *Modern Applied Statistics with S*, Springer Verlag, New York.

Wecker, W. and Ansley, C. (1983), 'The signal extraction approach to nonlinear regression and spline smoothing', *Journal of the American Statistical Association,* **78**, 81–89.

Wei, W. and Stram, D. (1990), 'Disaggregation of time series models', *Journal of the Royal Statistical Society,* **52**, 453–467.

Wolter, K. (1979), 'Composite estimation in finite populations', *Journal of the American Statistical Association,* **74**, 604–613.

Wolter, K. (1985), *Introduction to Variance Estimation*, Springer Verlag, Washington, USA.

Yang, M., Goldstein, H. and Heath, A. (2000), 'Multilevel models for repeated binary outcomes: Attitudes and votes over the electoral cycle', *Journal of the Royal Statistical Society. Series A,* **163**, 49–62.

Zaier, L. and Trabelsi, A. (2007), 'A polynomial method for temporal disaggregation of multivariate time series', *Communications in Statistics - Simulation and Computation* **36**(3), 741–759.

Zellner, A. and Mornmarquette, G. (1976), 'A study of some aspects of temporal aggregation problems in econometric analyses', *The Review of Economics and Statistics,* **58**, 335–342.

Zivot, E., Wang, J. and Koopman, S. (2004), State space modelling in macroeconomics and finance using SsfPack in S+Finmetrics, *in* A. Harvey, S. Koopman and N. Shepard, eds, 'State Space and Unobserved Component Models: Theory and Applications', Cambridge University Press.

Zwillinger, D. and Kokoska, S. (2000), *Standard Probability and Statistical Tables and Formulae*, Chapman and Hall, Boca Raton.