UNIVERSITY OF SOUTHAMPTON

Modelling Compositional Time Series from Repeated Surveys

by

Denise Britz do Nascimento Silva

Thesis submitted for the degree of Doctor of Philosophy

Faculty of Mathematical Studies

December 1996

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty of Mathematical Studies

Doctor of Philosophy

MODELLING COMPOSITIONAL TIME SERIES FROM REPEATED SURVEYS

by Denise Britz do Nascimento Silva

A compositional time series is defined as a multiple time series in which each of the series has values bounded between zero and one and, moreover, the sum of the series equals one at each time point. Data with such characteristics are observed in repeated surveys when a survey variable has a multinomial response and the interest lies on the proportion of units classified in each of its categories. In this case the survey estimates are proportions of a whole subject to a unity-sum constraint.

This thesis proposes state-space models for improving estimation of compositional data from repeated surveys taking into account the sampling errors. The proposed modelling procedure provides bounded predictions and signal estimates for the compositions, satisfying the unity-sum constraint, while taking into account the sampling errors. This is accomplished by mapping the compositions from the Simplex onto the Real space using the additive logratio transformation, then modelling the transformed data via multivariate state-space models, and finally applying the additive logistic transformation to obtain estimates in the original scale. In addition it is shown that the modelling procedure is permutation invariant.

The method is applied to compositional data from the Brazilian Labour Force Survey. The model for the survey estimates is a combination of the multivariate models specified for the signal and noise processes. Estimates for the vector of proportions of labour market status and the unemployment rate are obtained. Estimates of seasonally adjusted series are also produced. The results of the empirical work lead to the conclusion that smoother trends are obtained with a model which explicitly accounts for the sampling errors, when compared with the results from other standard procedures for seasonal adjustment.

ACKNOWLEDGEMENTS

Research for this thesis was supported by a grant from Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and by my employers Fundação Instituto Brasileiro de Geografia e Estatística (IBGE).

I thank my supervisor Prof. T.M.F. Smith not only for his advice, continuous attention and support, but also for providing some fine opportunities for my development.

Many thanks to my colleagues Rory, Masoud, Tahzibi, Chris and Parviz who shared with me the highs and lows of this process and, who made my days in the office so pleasant. Special thanks go to Denise, Christine, Rahim and Mitra for their friendship and constant encouragement.

I am very grateful to Carvalho & Silvia, Roberto & Cristiane, Paulo & Graça, Iuri & Matildes, Ivanildo & Irani, Mário & Mônica, Eduardo & Sílvia, Márcia & Flávio, Ronald & Tatiana, Taís, Eduardo, Fernando, Paola and Carol for their invaluable support throughout these four years.

Many good friends in Brazil made their best to keep in touch through their constant messages of encouragement, in particular Lourdes, Magina, Cris, Teresa, Diva, Vermelho, Zélia and Sônia. Thanks are also due to my colleagues at the Escola Nacional de Ciências Estatísticas (ENCE).

I thank Prof. Danny Pfeffermann for the many helpful discussions and for providing software for fitting state-space models. I also thank Dr. David Binder for elucidating my queries regarding some of his earlier work.

Thanks are due to the staff of the Departamento de Emprego e Rendimento of IBGE for providing the Labour Force Survey data and corresponding documentation.

I am forever indebted to my parents, Golda & Oscar, to whom I owe much more than I can say. Their love and guidance brought me this far. I am also grateful to my sister, Marcia, for always being there for my parents while I was away.

Finally, I dedicate this thesis to my son, for his patience and understanding and to my husband Pedro, for his unconditional love and support and for being the best husband a research student could ever wish for.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

This thesis focuses on the estimation and analysis of repeated surveys. More precisely, it deals with the use of the state-space approach for modelling compositional time series from repeated surveys, while taking into account the sampling errors.

The term composition is used here to indicate a vector of non-negative elements representing proportions of a whole, subject to a unity-sum constraint (Aitchison,1986). A compositional time series is a multivariate time series comprising observations of compositions at each time point (Brunsdon, 1987).

There are two main approaches for estimation incorporating the past information available in repeated surveys. One is the **classical sampling approach** in which the parameters of interest are considered as fixed but unknown quantities. The other is the **time series approach** in which the targets for inference are regarded as random quantities which can be modelled by a time series process. Regarding the latter, there are two alternative ways of fitting time series models to repeated survey data. One applies Box & Jenkins(1970) type procedures for fitting ARIMA models and uses signal extraction theory (Whittle, 1983) to estimate the unknown population quantities. The other, very much in use nowadays, employs the state-space approach (Harvey, 1989) which can be used to fit both ARIMA and structural time series models.

Most of the work available in this area concentrates on modelling univariate series of survey data. The exceptions are mostly concerned with small area estimation, in which the same target variable is considered across different (small) domains. Although much has already been done regarding modelling univariate series of survey estimates, models for multivariate survey data are not readily available in the literature.

Therefore, a key feature of this thesis is the use of vector ARMA and multivariate structural models to improve estimation and analysis in repeated surveys, by allowing two or more survey variables to be modelled simultaneously, while taking the sampling errors into account.

The motivation for this work originated from the fact that statistical agencies around the world often investigate many variables in each of their surveys. In addition, some of these variables have multinomial response, and for these variables the interest lies on the estimation of the proportion of units classified in each of their categories. When is this case, the survey estimates are proportions of a whole subject to a unity-sum constraint (compositions).

Hence there is a need for a modelling procedure which, while taking into account the sampling errors, allows different survey variables to be modelled concurrently satisfying, when required, a unity-sum constraint.

Another topic of current interest which is addressed in the thesis regards the seasonal adjustment of vector time series of survey data. When two or more series subjected to a sum constraint need to be seasonally adjusted the analyst always faces the dual choice of either adjusting each of the series individually and then producing the seasonally adjusted aggregate series from the individually adjusted ones or to directly adjust the aggregate series depending on the prime objective of the analysis.

However, modelling the series simultaneously via multivariate structural models yields seasonally adjusted series which satisfy the underlying sum constraint. Moreover, because the time series models proposed in this thesis explicitly account for sampling errors (the noise process), the resulting seasonally adjusted figures and trend estimates are related to the underlying signal which represents the unobservable population quantities.

This thesis presents a framework to model multivariate data from repeated surveys, taking into account the sampling errors, with special emphasis on the compositional case.

1.2 Outline

Chapter 2 describes some basic ideas and concepts used in the repeated survey context. First, the key objectives of analysis and basic types of repeated surveys are examined. Then the **classical sampling** and the **traditional time series** approaches for estimation and analysis of repeated surveys are briefly reviewed. Chapter 3 defines state-space models and discusses the representation of ARMA and structural time series models in the state-space framework.

Chapter 4 reviews the use of state-space models for improving estimation in repeated surveys, including some references to the case in which the target quantities are proportions. The solutions available for this case consist in modelling the original series of estimated proportions using, the state-space approach, without applying any transformation. This approach is examined in detail in Chapter 5.

For simplicity, all the analysis carried out in Chapter 5 refers to the case of univariate time series. Despite this, the overall idea about the problem to be tackled here is developed in this chapter, which also provides evidence that the current state-space approach for modelling proportions does not guarantee that the predictions and signal estimates are bounded between zero and one.

A new approach for modelling compositional data from overlapping surveys is proposed in Chapter 6. It proposes a class of multivariate state-space models for series of compositional data which take into account the sampling errors and, which simultaneously attempt to guarantee predictions and signal estimates satisfying the underlying constraints imposed by compositions.

The complete specification of a time series model for survey data embraces the identification of a suitable model to represent the sampling error process. Chapter 7 addresses this issue, developing procedures for dealing with the multivariate and compositional case.

Finally, in Chapter 8, the methods are applied to compositional data from the Brazilian Labour Force Survey which comprises estimates of the vector of proportions of labour market status. The model for the survey estimates is a combination of the multivariate models specified for the signal and noise processes. Estimates of seasonally adjusted compositions and unemployment rate series are also produced. The results appear to show that the underlying structure obtained from this approach is simpler than that obtained using standard methods for seasonal adjustment such as the X-11 programme. Chapter 9 reviews the key conclusions of the thesis and presents some ideas for future work.

2 Analysis of Repeated Surveys

2.1 Introduction

Repeated surveys (also known as longitudinal surveys, surveys across time or surveys in successive occasions) are employed by many organizations, such as national statistical agencies, to provide information enabling a study of the evolution of the population in time.

Denote by θ_t a population quantity of interest at time t and assume that observations are made at equally spaced time intervals $t=1,2,\ldots,T$. Let y_t represent a survey-based estimate of θ_t . A repeated survey produce a time series $\{y_t\}$ comprising estimates of the unknown target series $\{\theta_t\}$. Examples of possible target quantities θ_t are the monthly proportion of unemployed people in a country or the quarterly total of industrial production. However, the target parameters may not be just simple means or totals.

Before proceeding further, some notation and terminology need to be specified. Adopting the usual convention, a statistic (a random variable) that can be used to estimate (or predict) some unknown quantity is called an *estimator*. The *value* associated with a particular realization of that statistic is called an *estimate*. The same notation will be used here to represent both an estimator and its realization, the estimate. That is, the same notation is used to denote an unknown but observable quantity and its observed value. The precise meaning will be made clear according to the context in which it is used. For example, y_t above represents an estimator for θ_t which becomes an estimate when the survey is effectively carried out.

Duncan & Kalton(1987) and Kalton & Citro(1993) identified several possible objectives for repeated surveys, including:

(i) the estimation of population parameters at distinct time points - $\{\theta_t\ ,\ t=1,2,\ldots\}$;

(ii) the estimation of population parameters averaged across time - $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t \ ,\ T=1,2,\ldots$;

(iii) the estimation of change - $\Delta_t = \theta_t - \theta_{t-1}$;

(iv) the cumulation of samples over time.

At this point, it is important to specify that this research is concerned with the first objective listed above. That is, the centre of interest throughout this thesis will be the estimation of population parameters such as means, totals or ratio means at each survey round. A variety of survey designs can be used to meet objectives *(i)* to *(iv)*, by collecting data at several points in time. The differences between these designs come mostly from the strategies for inclusion/exclusion of units in the sample on distinct survey rounds. One way to characterize these survey designs is by the level of sample overlap between occasions, based on which Duncan & Kalton(1987) distinguished four different types of element survey designs:

(a) series of cross-sectional surveys (non-overlapping surveys) -on each occasion a sample of the existing population is selected and no attempt is made to ensure that any unit is included on more than one occasion; instead it may be specified that units cannot be included in more than one survey round;

(b) repeated panel surveys (complete overlapping surveys) - on each occasion similar measurements are made on the same fixed sample, and usually the sampling units are kept in the panel during the whole course of the survey;

(c) rotating panel surveys or rotating sampling (partially overlapping surveys) - some units are retained on the sample from one occasion to the next, and some are replaced by newly selected ones; the set of sampling units that join and leave the survey at the same time is usually called a panel or a rotation group;

(d) split-panel surveys (partially overlapping surveys) - this design is a combination of a panel with a cross-sectional or rotating panel survey; in this case, one portion of the initially selected sample is maintained fixed for all occasions, and the other portion is partly (rotation) or wholly (cross-sectional) substituted at each survey round.

The choice of the survey design and the overlapping pattern depends on the survey's overall aims as well as on operational constraints (for a detailed discussion see Duncan & Kalton, 1987, Cochran,1977, pp.344-345 or Kish, 1987, Chapter 6). Although defined for element survey designs, the above categories also apply for multi-stage surveys. In the case of master samples, for example, the primary sampling units are maintained fixed for several occasions whereas other stage units can follow some rotation pattern.

2.2 Classical Sampling Approach

In the classical sampling approach, the sequence of population parameters $\{\theta_t\}$ is considered as a set of fixed yet unknown quantities. Early results about estimation of population parameters in time from repeated surveys, using the classical sampling approach, were reported by Jessen(1942), Yates(1949) and Patterson(1950). These papers provided a general theory for designing samples and estimating means, totals and change with partial replacement of the sampling units. It is interesting to note that both Yates(1949) and Patterson(1950) based their work on the assumption that the observations y_{ti} of the survey variable y for unit i at time t were random quantities, related to previous observations for the same unit $y_{t-h,i}$ for $h=1,2,...,t-1$, with some correlation structure, whereas the population parameters θ_t were treated as fixed and unknown quantities. Extensions of this early work were provided by Eckler(1955), Woodruff(1963), Rao & Graham(1964), Gurney & Daly(1965), Singh(1968), Wolter(1979) and Tikkiwal(1979).

A common feature of all the papers adopting the classical sampling approach for repeated surveys was the use of composite estimators. Gurney & Daly(1965) defined a composite estimator as a "weighted average of two or more *linear unbiased estimators* of a specific characteristic for a given time period, where the weights are selected in order to reduce the variance when compared with the variances of the original estimators". Generally, the linear unbiased estimators considered were based on the matched and unmatched sample portions from different survey rounds. The composite estimators were defined recursively and used a limited number of linear unbiased estimators combined with past composite estimators.

The papers listed above examined the formulation of composite estimators and the determination of the optimal proportion of sampling units to retain in the sample from one occasion to next, in order to minimize the variance of the composite estimator.

Gurney & Daly(1965) introduced the concept of an elementary estimate. An *elementary estimate* is based on data from a single time period and only includes values from a set of units that join and leave the survey at the same time. Therefore an elementary estimate only uses data from a single rotation group in a single survey round.

Let $y_t^{(k)}$ be the k^{th} elementary estimate at time t for the population parameter θ_t and let

$$y_t^{(k)} = \theta_t + e_t^{(k)} \tag{2.1}$$

where $e_t^{(k)}$ is the sampling error. Under the assumptions that θ_t is a fixed yet unknown quantity and that $y_t^{(k)}$ is design unbiased for θ_t , $\forall\ k=1,...,K$, it follows that $E_p(e_t^{(k)}) = 0$, $\forall\ k=1,...,K$. Here E_p denotes the expectation operator with respect to the randomization distribution and the index $k=1,2,...,K$ does not refer to an individual sampling unit, but to a rotation group (or panel). It is usually assumed that $\{e_t^{(k)}\}$ has a known correlation structure over time. The specification of the elementary estimates as well as their correlation structure over time depends on the survey's pattern of overlap.

Consider a rotating panel survey, such as the U.S. Current Population Survey, which has a 4-8-4 rotating system (Bureau of the Census, 1978). In this case, eight elementary estimates are available every month, because each of the eight rotation groups surveyed each month provides an (unbiased) estimate for the target population quantity.

Gurney & Daly(1965) obtained the minimum variance linear unbiased estimator (MVLUE) for θ_t within the class of estimators that are linear combinations of elementary estimators. They used multivariate analysis techniques in order to obtain the coefficients of the desired MVLUE. Later Smith(1978), Wolter(1979) and Jones(1980) provided simpler developments of their result, as follows.

Assume that K rotation groups were investigated at each survey round in a survey which has been carried out for T occasions. Hence KT elementary estimates would be available for the analyst after survey round T. In this case a generalization of equation (2.1) can be written in a matrix form as

$$y = X\,\boldsymbol{\theta} + e \quad , \tag{2.2}$$

where y is a $(KT \times 1)$ vector of elementary estimates for all occasions up to time T, X is a $(KT \times T)$ "design" matrix of 0's and 1's indicating which element of y is associated with each element of $\boldsymbol{\theta}$, a $(T\times 1)$ vector with components θ_t, and e is a $(KT \times 1)$ vector with $E_p(e) = \mathbf{0}$ and $E_p(e e') = \Sigma$, where Σ is the variance-covariance matrix of the elementary estimates assumed known. Note that e represents the vector of survey sampling errors regarding the elementary estimates.

Applying generalized least squares (Rao, 1973, p. 230), the MVLUE for $\boldsymbol{\theta}$ is obtained as

$$\hat{\boldsymbol{\theta}} = (X' \Sigma^{-1} X)^{-1}\ X' \Sigma^{-1}\ y \quad , \tag{2.3}$$

with $V(\hat{\theta}) = (X' \Sigma^{-1} X)^{-1}$. Furthermore, the MVLUE for any linear combination $L' \boldsymbol{\theta}$ is $L' \hat{\boldsymbol{\theta}}$, with variance $L' (X' \Sigma^{-1} X)^{-1} L$.

If the survey has no overlap, Σ becomes diagonal and each θ_t is estimated independently of θ_{t-h} $h = 1,2,\ldots,t-1$, that is, using only the information provided by $y_t^{(k)}$, $k=1,\ldots,K$.

Note that (2.3) makes use of all data available up to time T. Hence the coefficients used in the MVLUE would change on every occasion after collecting additional observations,which would imply the need to update all previous estimates. This is not a desirable property.

Wolter(1979) argued that it would be very simple to construct a MVLUE for any rotating scheme by setting up the appropriate vector y of elementary estimates and specifying the covariance matrix Σ and the design matrix X . However, he and other authors (Binder & Hidiroglou, 1988 , Binder & Dick, 1989) pointed out that some difficulties appear when computing MVLU estimates in practice. One of them is the inversion of the covariance matrix Σ , which can become large if many historical elementary estimates are considered. In fact, the order of $(X' \Sigma^{-1} X)$ increases as T increases.

In addition, as pointed out by Smith(1978), to employ an estimator $\hat{\theta}$ as in (2.3) the analyst must be able to obtain individual observations on each occasion and also to trace the same unit for all repetitions of the survey in order to evaluate the correlations between units at different times and to correctly form the rotation groups for computing the elementary estimates. Smith(1978) classified this procedure as primary analysis, and used the term secondary analysis to describe those other procedures based solely on the aggregate estimates from each survey round (such as y_t).

2.3 Traditional Time Series Approach

Focusing now on the unknown population quantity θ_t , it is quite natural to imagine that the knowledge of $\theta_1 ,\dots, \theta_{t-1}$ conveys useful information about θ_t . However this assumption does not imply that θ_t is perfectly predictable from $\theta_1 ,\dots, \theta_{t-1}$. One way of representing this situation is by considering θ_t to be a random variable which evolves stochastically in time following a certain time series model. Such a framework, first proposed by Blight & Scott(1973), Scott & Smith(1974) and Scott, Smith & Jones(1977) is considered in the this section.

Blight & Scott(1973) introduced an estimation procedure for partially overlapping repeated surveys in which the population mean was assumed to vary over time. They allowed the population means on different occasions to be correlated by assuming a first-order autoregressive model for $\{\theta_t\}$ (Box & Jenkins, 1970, p.56), denoted hereafter AR(1). An AR(1) model was also adopted for the individual population values $\{y_{ti}\}$.

A recursive relationship for the estimation of θ_t and for its estimator's variance was derived using the conditional probability distribution of θ_t given all the elementary estimates up to time t . Using Bayes' theorem, this probability function was factorized according to the information specified by the time series models for $\{\theta_t\}$ and $\{y_{ti}\}$. However, because the recursive formulae depended on the elementary estimates and on the correlation between individual values, this estimator required a form of primary analysis, as in the classical sampling approach.The problem is that the individual sample observations y_{ti} are not always available to the analyst.

Hence, Scott & Smith(1974) and later Scott, Smith & Jones(1977) proposed a time series approach based on secondary analysis. Using signal extraction results (see e.g. Whittle, 1983, pp.46-47 or Reinsel, 1993, pp. 218-220), estimators were provided for the population mean for both overlapping and non-overlapping surveys, considering the following decomposition:

$$y_t = \theta_t + e_t \tag{2.4}$$

where $\{\theta_t\}$, $\{y_t\}$ and $\{e_t\}$ are random processes with y_t being a design-unbiased estimate of the unknown population quantity θ_t based only on data from time t, and e_t denoting the sampling error such that $E(e_t\,|\,\theta_t) = 0$ and $V(e_t\,|\,\theta_t) = S_t^2$.

By analogy with the signal extraction approach, θ_t is the signal and e_t is the noise. In this context, one is interested in estimating the unobservable signal, θ_t , based on the past and current observations, $y_1, \dots, y_t$, in the presence of noise. In order to employ the signal extraction results to estimate θ_t , it is necessary to add the following assumptions to the model (2.4):

(i) $\{\theta_t\}$, or a suitable difference of it, is stationary;

(ii) $\{e_t\}$ is stationary;

(iii) $\{\theta_t\}$ and $\{e_t\}$ are uncorrelated time series.

Under this model, the classical signal extraction approach for the estimation of θ_t is to determine a linear filter:

$$\hat{\theta}_t = \sum_{j=0}^{\infty} a_j y_{t-j} \quad ,$$

such that $\hat{\theta}_t$ is close to θ_t in the sense that the mean squared estimation error (MSE)

$$E(\theta_t - \hat{\theta}_t)^2 = E(\theta_t - \sum_{j=0}^{\infty} a_j y_{t-j})^2 \quad ,$$

is a minimum among all possible linear filters.

In order to illustrate the use of the signal extraction procedure to estimate the current value of the population parameter consider the case of a non-overlapping repeated survey. Following Scott & Smith(1974), $\{\theta_t\}$ is assumed to follow an AR(1) process of the form:

$$\theta_t = \phi\theta_{t-1} + \eta_t \quad , \tag{2.5}$$

where η_t are white noise disturbances with mean zero and variance σ_η^2 , denoted hereafter by $\eta_t \sim WN(0,\sigma_\eta^2)$, and $|\phi| < 1$. For non-overlapping surveys with small sampling fractions, e_t and e_{t-j} are considered uncorrelated for $j = 1,2,\dots,t-1$, and the structure of the noise process $\{e_t\}$ is completely specified by the sampling variance of y_t as an estimator of θ_t , denoted here as σ_e^2 . The observed series $\{y_t\}$ is then assumed to be the sum of two random processes, an AR(1) and a white noise process, where $\{\theta_t\}$ and $\{e_t\}$ are uncorrelated time series. The main objective of the signal extraction procedure is to estimate the signal given the observed series. In this case, the sample estimate can be expressed as

$$y_t = \theta_t + e_t = \frac{\eta_t}{1 - \phi B} + e_t \quad , \tag{2.6}$$

with $\eta_t \sim WN(0,\sigma_\eta^2)$ and $e_t \sim WN(0,\sigma_e^2)$.

From (2.6), it follows that

$$(1 - \phi B)y_t = \eta_t + (1 - \phi B)e_t \quad . \tag{2.7}$$

Therefore $\{y_t\}$ follows an ARMA(1,1) model which can be alternatively represented by

$$(1 - \phi B)y_t = (1 - \delta B)\varepsilon_t \quad , \tag{2.8}$$

with $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$ having the same underlying autocovariance structure for $\{y_t\}$ as (2.7).

The model parameters σ_ε^2 and δ in (2.8) are determined by equating the covariances in (2.7) and (2.8). The signal estimator $\hat{\theta}_t$ based on the classical signal extraction approach is given by (for details see Scott, Smith & Jones, 1977 or Reinsel, 1993, p.221):

$$\hat{\theta}_t = \left[1 - \frac{\delta}{\phi}\right] \sum_{j=0}^{\infty} \delta^j y_{t-j} \quad . \tag{2.9}$$

This is exactly the estimator of θ_t proposed by Scott & Smith(1974).

Scott, Smith & Jones(1977) extended the results in Scott & Smith(1974) for complex survey designs. In addition to the non-overlapping case they examined single-stage and two-stage overlapping surveys. Since the autocovariance structure of the sampling errors depends on the pattern of overlap, the authors used different ARMA models for both $\{\theta_t\}$ and $\{e_t\}$. They also provided an interesting discussion about which ARMA models would be appropriate under different survey designs.

Regarding the sampling error process $\{e_t\}$, for example, they suggested that for completely overlapping surveys an AR(1) would be a reasonable model. For partially overlapping surveys in which the units are rotated out of the sample after q occasions in the survey (which implies that there are no common units in the samples of times t and $t-j$, for $j > q$, a moving-average process of order q (MA(q)) was recommended because the autocorrelation function for such models is zero for lags greater than q. A detailed discussion about modelling of the sampling error process is found in Chapter 7.

Jones(1980) derived a minimum mean squared linear estimator (MMSLE) for $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_T)'$ which encompasses all the previous results. Using the vector of unbiased estimates $y = (y_1, \dots, y_T)'$ such that

$$y = \boldsymbol{\theta} + \boldsymbol{e} \quad , \tag{2.10}$$

with $E(\boldsymbol{e} \mid \boldsymbol{\theta}) = \mathbf{0}$ and $E(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{\theta}) = \Sigma$, where Σ is assumed known, and allowing $\boldsymbol{\theta}$ itself to be a random variable with mean $\boldsymbol{\mu}$ and variance-covariance matrix V , the best linear estimator of $\boldsymbol{\theta}$ given y is obtained as (see Rao, 1973, p.234):

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu} + (\Sigma^{-1} + V^{-1})^{-1} \Sigma^{-1}(y - \boldsymbol{\mu}) \quad . \tag{2.11}$$

Noting that $(\Sigma^{-1} + V^{-1})^{-1} = V - V(V + \Sigma)^{-1} V$ (see result 2.9 from Rao, 1973, p.33), $\hat{\boldsymbol{\theta}}$ can be rewritten as:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu} + V(V + \Sigma)^{-1}(y - \boldsymbol{\mu}) \quad , \tag{2.12}$$

with variance $V - V(V + \Sigma)^{-1} V$.

Observing that:

$$\boldsymbol{\mu} = E(\boldsymbol{\theta}) = E(y) \quad , \quad V = COV(\boldsymbol{\theta}, y) = COV(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{e}) \quad , \tag{2.13a}$$

$$V(y) = E[V(y \mid \boldsymbol{\theta})] + V[E(y \mid \boldsymbol{\theta})] = (\Sigma + V) \quad , \tag{2.13b}$$

it becomes clear that, when assuming normality of $(\boldsymbol{\theta}, y)$, the estimator $\hat{\theta}$ given by (2.11) or (2.12) corresponds to the conditional expectation of $\boldsymbol{\theta}$ given y (see Rao, 1973, p.522). Note that the derivation in (2.13) assumes $\boldsymbol{\theta}$ and $\boldsymbol{e}$ uncorrelated.

If a primary analysis was to be carried out, instead of a secondary analysis, the vectors y and would have to include all the elementary estimates up to time T and the model (2.10) would have to be modified by including a matrix X , as in (2.2). In this case,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu} + (X' \Sigma^{-1} X + V^{-1})^{-1} X' \Sigma^{-1}(y - X\boldsymbol{\mu}) \quad . \tag{2.14}$$

When μ is completely unknown, Rao(1973, p.234) shows that (2.14) reduces to the estimator in (2.3) obtained under the assumption that θ is fixed rather than random.

As already discussed for the classical sampling approach, to employ an estimator such as (2.14) one needs to invert matrices with dimensionality equal to the number of elementary estimates available up to time T . Moreover, (2.14) depends on the covariance structure of the unobservable quantity θ . Assuming θ and e uncorrelated, V can be obtained as $V(y) - \Sigma$ which, in turn, can be estimated from the observed data. This constitutes an additional and very practical reason for requiring stationarity of $\{e_t\}$.

The autocovariance structure of $\{\theta_t\}$ can be obtained if one knows the autocovariance generating function of $\{y_t\}$ and $\{e_t\}$, since the sampling error and the signal processes are assumed to be uncorrelated. Then estimates of the autocovariance structure of the signal $\{\theta_t\}$ may be readily obtained if the sampling error autocovariances can be estimated using design-based methods. Moreover, assuming $\{e_t\}$ to be stationary, information about sampling covariances can be pooled over time to estimate $COV(e_t, e_{t-h})$, which in this case depend only on the lag h . Observe that these covariances need to be estimated from a single realization of the series.

One way of overcoming both problems (the estimation of the covariance matrix of θ and the manipulation of large matrices) is by formulating the signal extraction problem in terms of state-space models and the Kalman Filter (Anderson & Moore,1980, Harrison & Stevens,1976 and Harvey,1989). Once a model has been expressed in a state-space form, the Kalman Filter equations can be used to develop a recursive procedure for producing the MMSLE (or MMSE) of the unobservable population quantity of interest. Moreover, by expressing the model in state-space form, the likelihood function for y is directly obtained from the specified model and can be maximized to estimate the unknown model parameters.

Binder & Hidiroglou(1988), Binder & Dick(1989), Tiller(1989), and Pferffermann, Burck & Ben-Tuvia(1989) introduced the state-space models for estimation in repeated surveys. Some other applications of state-space models for repeated surveys can be found in

Binder & Dick(1990), Pfeffermann(1991), Tiller(1992), Pfeffermann & Bleuer(1993), Binder, Bleuer & Dick(1993) and Pfeffermann, Bell & Signorelli(1996).

Before reviewing the role that state-space models can play in developing an approach for improving estimation in repeated surveys, the next chapter introduces the state-space formulation with the Kalman Filter equations for the analysis of the time series.

3 State-Space Models

3.1 Introduction

The basic ideas of state-space models have their roots in control engineering and the physical sciences, where a system is a mathematical model for a real world process that accepts a number of inputs and gives rise to a number of outputs. One way of describing a system is by using the **State-Space** representation. State-space models consist of two equations. The first equation, *the observation (or measurement) equation*, represents the relationship between the observations and the current state of the unobservable model components. The second, *the transition (or system) equation*, describes how the unobservable components evolve stochastically in time.

Let $y_1, \dots, y_T$ be a sequence of vectors of M-dimensional observations (or measurements), namely $y_t = (y_{1t}, \dots, y_{Mt})'$, and let $\alpha_1, \dots, \alpha_T$ be a sequence of stochastic state-vectors. In a state-space framework the relationship between y_t and α_t is described by

the Observation Equation:

$$y_t = H_t \alpha_t + \varepsilon_t \qquad \textbf{(3.1a)}$$

where y_t is $(M \times 1)$, H_t is a $(M \times n)$ matrix, α_t is a $(n \times 1)$ vector called state-vector, and ε_t is a $(M \times 1)$ vector of serially uncorrelated normally distributed disturbances with mean zero and covariance matrix U_t .

In general, the elements of the state-vector are not observable. However, they are assumed to be generated by a first order Markov process that can be described by

the System Equation:

$$\alpha_t = T_t \alpha_{t-1} + G_t \eta_t \qquad \textbf{(3.1b)}$$

where T_t and G_t are $(n \times n)$ and $(n \times g)$ matrices, respectively, and η_t is a vector of serially uncorrelated normally distributed disturbances with mean zero and covariance matrix $\boldsymbol{Q}_t$. $\boldsymbol{T}_t$ is usually called transition or state matrix.

In general, the system matrices $\{\boldsymbol{H}_t , \boldsymbol{U}_t , \boldsymbol{T}_t , \boldsymbol{G}_t , \boldsymbol{Q}_t\}$ may depend on sets of unknown parameters, or hyperparameters, that must be estimated. If the system matrices do not change over time the model is said to be time-invariant. Most of the models considered throughout this thesis are of this type.

To complete the specification of the state-space system in (3.1) two additional assumptions are needed:

(i) the initial state-vector α_0 is normally distributed with mean and covariance matrix given by $E(\alpha_0) = \hat{\alpha}_{0|0}$; $V(\alpha_0) = \boldsymbol{P}_{0|0}$;

(ii) the disturbances ε_t and η_t are mutually uncorrelated over time and also uncorrelated with the initial state vector, namely

$$COV(\varepsilon_t , \eta_t) = \mathbf{0} , COV(\varepsilon_t , \alpha_0) = \mathbf{0} , COV(\eta_t , \alpha_0) = \mathbf{0} \quad \forall\, t \quad .$$

These assumptions are needed for a simpler derivation of the Kalman Filter recursion equations. However, the normality assumptions can be dropped and the transition and measurement disturbances can be correlated, allowing more general results to be obtained.

The main feature of this approach is the ability of providing filtered estimates of the unobservable state-vector and to predict future values of the observations. Anderson and Moore(1979, sect.2.1) emphasize the differences between *filtering*, *smoothing*, and *prediction*. For them, filtering means recovering at time t information about some unobservable quantity associated with a system using measurements or observations right up to time t (but not those available after time t). Smoothing is concerned with recovering information about unobservable system quantities using measurements obtained both before and after time t . Consequently, the recovery does not occur at time t (but after that). Finally, prediction forecasts the future system behaviour at time $t' > t$, given data up to time t .

One quantity of particular interest in the system is the state-vector. It is set up in such a way that it carries all the information about the system which is essential to determine its future behaviour. The current state is defined by Wei(1993, p.384) as the minimum set of information that, together with future inputs, is sufficient to describe the future system behaviour. Therefore, the state-space representation of a system is also called the Markovian representation, because given the present state, the future of a system is independent of its past.

Once a model has been put in a state-space form, one can use the Kalman Filter to establish a recursive procedure for making inference about the state-vector and the system measurements. The recursive procedure is carried out in two stages, the first one prior to observing y_t , which produces the prediction equations. The second is carried out after observing y_t , which produces the updating equations. The procedure was originally developed by R.E. Kalman(Kalman, 1960, Kalman & Bucy, 1961).

When using the Kalman Filter, an optimal estimator of the state-vector at time t can be computed based on the information available at that time. Further, assuming the disturbances and the initial state-vector to be normally distributed, the Kalman Filter provides the minimum mean square estimator (MMSE) of $\boldsymbol{\alpha}_t$, which is the conditional mean of $\boldsymbol{\alpha}_t$ given $y_1 , \dots , y_t$. When the normality assumptions are dropped, the Kalman Filter gives the minimum mean square linear estimator (MMSLE), since it minimizes the mean square error within the class of all linear estimators. In Section 3.3, the Kalman Filter recursion equations are presented. Before that, however, Section 3.2 recalls some results about estimation which are necessary for their derivation.

An important point to note is that a state-space representation of a linear system is not unique. In fact, if one considers an arbitrary non-singular $(n \times n)$ matrix $\boldsymbol{N}$ then, from equation (3.1b) a new state vector $\boldsymbol{\alpha}_t^* = \boldsymbol{N}\,\boldsymbol{\alpha}_t$ can be defined. Letting $\boldsymbol{T}_t^* = \boldsymbol{N}\,\boldsymbol{T}_t\,\boldsymbol{N}^{-1}$, $\boldsymbol{G}_t^* = \boldsymbol{N}\,\boldsymbol{G}_t$, $\boldsymbol{H}_t^* = \boldsymbol{H}_t\,\boldsymbol{N}^{-1}$ it follows that

$$\begin{cases} N\,\alpha_t = N\,T_t\,N^{-1}\,N\,\alpha_{t-1} + N\,G_t\,\eta_t \quad , \\ y_t = H_t\,N^{-1}\,N\,\alpha_t + \varepsilon_t \quad , \end{cases}$$

so that

$$\begin{cases} \alpha_t^* = T_t^*\ \alpha_{t-1}^* + G_t^*\ \eta_t \quad , \\ y_t = H_t^*\ \alpha_t^* + \varepsilon_t \quad , \end{cases}$$

which, in turn, is an alternative and equivalent representation for (3.1). The features of different state-space representations for a system are discussed in Harvey(1989, p.102) and Reinsel(1993, pp.193-215). The choice of the representation can depend, for example, on the objective of the analysis (see Section 3.5).

3.2 Estimation Criteria

As stated before, at any fixed time t, the filter aims to provide information about a random variable α_t given the (observed) values of another random variable, say $\boldsymbol{D}_t = (y_1', \ldots, y_t')'$ where $\underline{\ \boldsymbol{D}_T\ }$ represent all the available information up to time $\underline{T}$. The natural way of doing this is via the conditional probability function of α_t given $\boldsymbol{D}_t$. It is well known that the best estimator of α_t given $\boldsymbol{D}_t$ is the mean of the conditional distribution, that is $E(\alpha_t \mid \boldsymbol{D}_t)$. Here *best* is used in the sense of minimizing the *mean square error.*

Result 3.1 (see, e.g., Anderson & Moore, 1979, p.26)

Let α and $\boldsymbol{D}$ be two jointly distributed random vectors. The minimum mean square estimator(MMSE) $\hat{\alpha}$ of α in terms of $\boldsymbol{D}$ is given by $\hat{\alpha} = E(\alpha \mid \boldsymbol{D})$. Then $\hat{\alpha}$ is uniquely specified as the conditional mean of α given $\boldsymbol{D}$.

Another useful result regards the case in which α and D are jointly normally distributed. It can be found, for example, in Rao(1973, p.522).

Result 3.2

Let α and D be two jointly distributed normal vectors, with mean and covariance matrix given by

$$E\begin{bmatrix}\alpha\\ D\end{bmatrix} = \begin{bmatrix}m_\alpha\\ m_D\end{bmatrix} \quad V\begin{bmatrix}\alpha\\ D\end{bmatrix} = \begin{bmatrix}\Sigma_{\alpha\alpha} & \Sigma_{\alpha D}\\ \Sigma'_{\alpha D} & \Sigma_{DD}\end{bmatrix} \quad . \tag{3.2}$$

Then $\alpha|D$ is normally distributed with

$$E(\alpha|D) = m_\alpha + \Sigma_{\alpha D}\,\Sigma_{DD}^{-1}\,(D - m_D) \quad , \tag{3.3a}$$

$$V(\alpha|D) = \Sigma_{\alpha\alpha} - \Sigma_{\alpha D}\,\Sigma_{DD}^{-1}\,\Sigma'_{\alpha D} \quad . \tag{3.3b}$$

The next result gives the minimum mean square linear estimator of α based on the available information D .

Result 3.3 (Anderson & Moore,1979, sect. 5.2)

Let α and D be two jointly distributed random vectors with mean and covariance as in (3.2). A linear estimator of α given D is of the form $\hat{\alpha} = AD + b$, where A is a fixed matrix and b is a fixed vector. The minimum mean square linear estimator of α given D is the one for which A and b minimize

$$E\{\|\alpha - AD - b\|^2\} = E\{(\alpha - AD - b)'(\alpha - AD - b)\} \quad . \tag{3.4}$$

Then the minimum mean square linear estimator(MMSLE) corresponds to $A = \Sigma_{\alpha D}\Sigma_{DD}^{-1}$ and $b = m_\alpha - A\,m_D$, hence yielding the same estimator as in (3.3a), with covariance error matrix as in (3.3b).The value of corresponding minimum mean square error(MSE) is given by $Trace(\Sigma_{\alpha\alpha} - \Sigma_{\alpha D}\Sigma_{DD}^{-1}\Sigma'_{\alpha D})$.

Corollary 3.1

a) The MMSLE is unbiased, that is, $E(\alpha - \hat{\alpha}) = 0$.

b) If α and D are jointly normally distributed random vectors then the MMSLE is equal to the MMSE ($E(\alpha \mid D)$).

Using these results the Kalman Filter equations can be derived. Appendix A1 contains the derivation of the Kalman Filter equations under the assumption that the disturbances and the initial state-vector are normally distributed. When the normality assumptions are relaxed the expressions of the estimators remain unchanged based on Result 3.3. However, in this case, the filter provides the MMSLE of α_t given D_t , rather than the MMSE (see Harvey, 1989, pp.110-111).

Returning to model (3.1), note that the conditional probability function $P(\alpha_t | \; y_1, ..., y_{t-1})$ summarizes all the information that $y_1, ..., y_{t-1}$ contain about α_t . To solve the prediction and filtering problems it is necessary first to compute the sequence $E(\alpha_t | \; y_1 \;, \;... \;, \; y_{t-1})$ for $t = 1, 2, ...$.

Denote $\hat{\alpha}_{t|t-1} = E(\alpha_t | \; y_1, ..., y_{t-1}) = E(\alpha_t | D_{t-1})$, and by

$$P_{t|t-1} = E \; \{(\alpha_t - \hat{\alpha}_{t|t-1})(\alpha_t - \hat{\alpha}_{t|t-1})' \mid D_{t-1}\} = V(\alpha_t \mid D_{t-1}) \qquad ,$$

the error covariance matrix associated with $\hat{\alpha}_{t|t-1}$. Similarly, the filtered estimate $E(\alpha_t | D_{t-1}, y_t) = E(\alpha_t \mid D_t)$ is denoted by $\hat{\alpha}_{t|t}$ with associated error covariance matrix $P_{t|t}$.

3.3 The Kalman Filter Equations

3.3.1 The Prediction Equations

Suppose that the current time is $t-1$, so one has observed D_{t-1} but does not know any of the states $\alpha_1, \alpha_2, \dots, \alpha_{t-1}$. Suppose, however, that $\hat{\alpha}_{t-1|t-1}$, the *"best"* estimator of α_{t-1}, is available to the analyst. At time $t-1$, the knowledge about $\boldsymbol{\alpha}_{t-1}$ is expressed by its MMSE and the associated error covariance matrix $P_{t-1|t-1}$. Then, before observing y_t, an optimal estimator for α_t is $\hat{\alpha}_{t|t-1}$. Therefore the prediction equations for α_t and the corresponding covariance matrix are (details in Appendix A1):

$$\begin{aligned} \hat{\alpha}_{t|t-1} &= E(\alpha_t | D_{t-1}) = T_t \hat{\alpha}_{t-1|t-1} \quad , \\ P_{t|t-1} &= V(\alpha_t | D_{t-1}) = T_t P_{t-1|t-1} T_t' + G_t Q_t G_t' \; , \qquad \forall \; t=1,,\dots,T. \end{aligned} \tag{3.5a}$$

To predict y_t based on D_{t-1}, the one step ahead forecast is given by (details also in Appendix A1):

$$\begin{aligned} \hat{y}_{t|t-1} &= E(y_t | D_{t-1}) = H_t \hat{\alpha}_{t|t-1} \quad , \\ F_{t|t-1} &= V(y_t | D_{t-1}) = H_t P_{t|t-1} H_t' + U_t \; , \qquad \forall \; t=1,\dots,T \; . \end{aligned} \tag{3.5b}$$

3.3.2 The Updating Equations and the Steady-State

Just before observing y_t, the inference about the state-vector α_t relies on the distribution $P(\alpha_t | D_{t-1})$. After observing y_t the MMSE for α_t given y_t is $E(\alpha_t \,|\, D_t)$. This conditional distribution is directly obtained via the standard results for the multivariate normal distribution (see Result 3.2) using $P(\alpha_t, y_t \,|\, D_{t-1})$. Then, the updating equations (also in Appendix A1) are given by:

$$\begin{aligned} \hat{\alpha}_{t|t} &= E(\alpha_t \,|\, D_t) = \hat{\alpha}_{t|t-1} + P_{t|t-1} H_t' \, F_{t|t-1}^{-1}(y_t - \hat{y}_{t|t-1}) \quad , \\ P_{t|t} &= V(\alpha_t \,|\, D_t) = P_{t|t-1} - P_{t|t-1} H_t' \, F_{t|t-1}^{-1} H_t P_{t|t-1} \ , \quad \forall \ t=1,\dots,T. \end{aligned} \tag{3.5c}$$

The above equation for $\hat{\alpha}_{t|t}$ provides the filtered estimate for α_t as soon as the observation y_t becomes available to the analyst.

Equations (3.5a), (3.5b), (3.5c) are the so-called **Kalman Filter Equations.** Putting together (3.5a) and (3.5c), $P_{t+1|t}$ is obtained as:

$$P_{t+1|t} = T_{t+1}(P_{t|t-1} - P_{t|t-1} H_t' \, F_{t|t-1}^{-1} H_t P_{t|t-1}) T_{t+1}' + G_{t+1} \, Q_{t+1} \, G_{t+1}' \ . \tag{3.6}$$

In addition, letting

$$P_{t|t-1} H_t' \, F_{t|t-1}^{-1} = K_t \quad , \tag{3.7}$$

it follows that

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(y_t - \hat{y}_{t|t-1}) = K_t y_t + (I - K_t H_t)\hat{\alpha}_{t|t-1} \quad . \tag{3.8}$$

Note that $\hat{\alpha}_{t|t}$ is a weighted average of the previous estimate of α_t and the observation y_t , for cases in which $H_t = I$ .Denoting $(y_t - \hat{y}_{t|t-1}) = \varepsilon_{t|t-1}$, the MMSE for α_t is given by:

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t \, \varepsilon_{t|t-1} \quad . \tag{3.9}$$

The *prediction errors* $\varepsilon_{t|t-1}$ are often called *innovations*, because they sum up all the new information contained in y_t that was not available from the previous history of the system (summarized in $\hat{\alpha}_{t|t-1}$). It becomes clear from (3.9) that the filter has a recursive *prediction-correction* form since the estimate $\hat{\alpha}_{t|t}$ equals the prediction of α_t from observations up to time $t-1$ updated by a factor K_t times the innovation term $\varepsilon_{t|t-1}$. The matrix K_t is known as the Kalman gain.

Before proceeding further it is important to define the concept of steady-state filter of time-invariant models which will be used in Chapter 5. Following Harvey(1989, p.118), a Kalman filter is in a steady-state if the covariance matrix $P_{t+1|t}$ is time invariant, that is,

$$P_{t+1|t} = P_{t|t-1} = P \quad . \tag{3.10}$$

The Kalman Filter has a steady-state solution if there is a time invariant covariance matrix of the form (3.10) which satisfies equation (3.6). Therefore, if this matrix exists, it is the solution of the following equation:

$$P - T(P-PH' F^{-1} HP)T' - GQG' = 0 \quad , \tag{3.11a}$$

with

$$F = HPH' + U \quad . \tag{3.11b}$$

The state-space model and the Kalman Filter can also provide predicted values for y_{t+j} and α_{t+j} for any future time $t+j$ with $j > 1$, given the observations up to time t (for details, see Harvey,1989, pp.222-223). Another feature of the Kalman Filter is the ability to smooth the state-vector, as a retrospective estimator, when new information becomes available. In this case it is possible to revise the inferences about previous values of the state-vector based on recent data.

3.3.3 Smoothing

Harvey(1989, pp.149-155) discusses three alternative ways of computing smoothed estimates. The one adopted here is known as *fixed-interval smoothing*. The recursion starts at time T , with $\hat{\alpha}_{T|T}$ and $P_{T|T}$ obtained from (3.5c), and runs backwards producing smoothed estimates in the order $T, T-1, \ldots, 1$. In a Gaussian model, the fixed-interval smoothed estimator is defined as:

$$\begin{aligned} \hat{\alpha}_{t|T} &= \hat{\alpha}_{t|t} + P_t^* (\hat{\alpha}_{t+1|t} - T\hat{\alpha}_{t|t}) \quad , \\ P_{t|T} &= P_{t|t} + P_t^* (P_{t+1|T} - P_{t|t}) P_t^{*\prime} \quad , \end{aligned} \tag{3.12}$$

where $P_t^* = P_{t|t} T' P_{t+1|t}^{-1} \quad \forall \quad t = T-1, \ldots, 1$.

3.4 Decomposition of the Likelihood Function

As pointed out earlier, the system matrices may be unknown depending on hyperparameters $(\boldsymbol{\Omega})$ that must be estimated. In this case the use of the Kalman Filter enables the evaluation of the likelihood which is used to estimate any unknown parameter in the model. Harvey(1989, sect. 3.4), examines the application of maximum likelihood estimation to time series modelling, a situation in which the observations $y_1, y_2, \dots, y_t$ are not independently distributed. In this case, their joint distribution can be obtained using the conditional probability density function $p(y_t|\boldsymbol{D}_{t-1})$ as:

$$p(\boldsymbol{D}_T, \boldsymbol{\Omega}) = p(y_1, \dots, y_T, \boldsymbol{\Omega}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{D}_{t-1}) \quad .$$

Recalling the prediction equations in (3.5b) and working with normality assumptions, it follows that:

$$p(\boldsymbol{D}_t, \boldsymbol{\Omega}) = \prod_{t=1}^{T} \frac{1}{(2\pi)^{M/2} |\boldsymbol{F}_{t|t-1}|^{1/2}} \exp\{\frac{-1}{2}(y_t - \hat{y}_{t|t-1})' \boldsymbol{F}_{t|t-1}^{-1} (y_t - \hat{y}_{t|t-1})\} \ .$$

Then the log-likelihood $\mathcal{L}(\boldsymbol{\Omega}; \boldsymbol{D}_T)$ is given by:

$$\mathcal{L} = -\frac{MT}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \{\log |\boldsymbol{F}_{t|t-1}| + (y_t - \hat{y}_{t|t-1})' \boldsymbol{F}_{t|t-1}^{-1} (y_t - \hat{y}_{t|t-1})\}, \qquad \textbf{(3.13)}$$

where *log* denotes natural logarithm.

Since $(y_t - \hat{y}_{t|t-1}) = \varepsilon_{t|t-1}$ is a vector of prediction errors, (3.13) is known as the *prediction error decomposition* form of the log-likelihood. A complete description of the state-space modelling procedure embraces the computation of the log-likelihood function, its maximization with respect to the unknown parameters as well as the initialization of the Kalman Filter. These aspects are discussed in Harvey(1989, Chapters 3 and 4), Pferffermann(1991) and Binder, Bleuer & Dick(1993). They will, however, be addressed later in the thesis when describing the empirical work. For details about the use of state-space models and the Kalman Filter in a time series framework refer to Harvey(1989).

Using the definitions introduced in this and previous sections, the use of state-space models in a time series analysis context will be considered next. The state-space representation of univariate ARMA models is described in Section 3.5. Section 3.7.2 contains the state-space representation for the univariate structural time series models. As this thesis is mostly concerned with compositional data, which implies that multivariate models are needed to represent both the signal and sampling error processes, Sections 3.6 and 3.7.3 review typical VARMA and multivariate structural models, respectively. In Chapter 4, various examples of the use of state-space models for improving estimation in repeated surveys are given.

3.5 The State-Space Representation of Integrated Autoregressive Moving Average Models

Integrated Autoregressive Moving Average (ARIMA) models can be represented in a state-space form for both univariate and multiple time series (Harvey & Phillips, 1979, Wei, 1993, Chapter 15, Reinsel, 1993, Chapter 7). For details about the standard theory regarding ARMA/ARIMA models, the reader is referred to Box & Jenkins(1970, Chapters 3 and 4) or Wei(1993, Chapters 3 and 4). A stationary time series is represented via a time-invariant system and, as stated before, this representation is not unique. In a time-invariant state-space model the system matrices $\{\boldsymbol{H}_t , \boldsymbol{U}_t , \boldsymbol{T}_t , \boldsymbol{G}_t , \boldsymbol{Q}_t\}$ are all independent of time and can, therefore, be defined without the time subscript. Harvey(1989, Sect.3.3) and Anderson & Moore(1979, Chapter 4) examine the properties of time-invariant systems. The state-space formulation of the ARIMA models will be defined and illustrated considering the univariate case. Reinsel(1993, chapter 7) provides the results regarding the multivariate case which are considered in Section 3.6. Harvey & Phillips(1979) provided the state-space representation of seasonal ARIMA models. As a special case of it, they also provided an equivalent result for ARMA models. What follows next is based on their work.

Definition 3.1

Let B be the backshift operator, i.e. $By_t = y_{t-1}$, $B^2 y_t = y_{t-2}$, etc., and let s denote the seasonal period. Then, an integrated seasonal autoregressive moving average model, ARIMA(p,d,q)(P,D,Q)$_s$, for a univariate time series $\{y_t\}$ is given by:

$$\lambda(B^s)\,\phi(B)(1 - B)^d\,(1 - B^s)^D\, y_t = \gamma(B^s)\,\delta(B)\,\epsilon_t \quad , \tag{3.14a}$$

with ϵ_t independent $N(0,\sigma_\epsilon^2)$ and

$$\begin{aligned}
\lambda(B^s) &= 1 - \lambda_1 B^s - \lambda_2 B^{2s} - \ldots - \lambda_P B^{Ps} \quad , \\
\gamma(B^s) &= 1 - \gamma_1 B^s - \gamma_2 B^{2s} - \ldots - \gamma_Q B^{Qs} \quad , \\
\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \quad , \\
\delta(B) &= 1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_q B^q \quad .
\end{aligned}$$

An alternative representation is given by

$$\varphi(B)\, y_t = \vartheta(B)\, \epsilon_t \tag{3.14b}$$

where $\varphi(B) = \lambda(B^s)\,\phi(B)(1 - B)^d\,(1 - B^s)^D$ is a (p+d+sP+sD)-degree polynomial and $\vartheta(B) = \gamma(B^s)\,\delta(B)$ is a (q+sQ)-degree polynomial.

It follows as a special case from definition 3.1 that a zero mean stationary ARMA model can be written as:

$$\phi(B)\, y_t = \delta(B)\, \epsilon_t \tag{3.14c}$$

Result 3.5 (Harvey & Phillips,1979)

A seasonal ARIMA(p,d,q)(P,D,Q)$_s$ model, as in Definition 3.1 (see equation (3.14b)), can be represented by a state-space model, having the following observation and systems equations:

observation equation

$$y_t = \boldsymbol{H}\boldsymbol{\alpha}_t \quad ; \tag{3.15a}$$

system equation

$$\boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}\,\eta_t \quad , \tag{3.15b}$$

with:

$$\boldsymbol{T} = \left[\begin{array}{cccccc} \varphi_1 & \vdots & & & & \\ \vdots & \vdots & & I_{(r-1)\times(r-1)} & & \\ \vdots & \vdots & & & & \\ \cdots & \vdots & \cdots & & \cdots & \cdots \\ \varphi_r & \vdots & & \mathbf{0}_{1\times(r-1)} & & \end{array}\right] \quad ,$$

where $r = \max(p+d+sP+sD, q+sQ+1)$, $\boldsymbol{G} = [1, -\vartheta_1, -\vartheta_2, \ldots, -\vartheta_{r-1}]'$, $\boldsymbol{H} = [1,0,0,\ldots,0]$ and $\boldsymbol{\alpha}_t = (\alpha_{1t}, \alpha_{2t}, \ldots, \alpha_{rt})'$ is defined as

$$\begin{aligned} \alpha_{1t} &= y_t \quad , \\ \alpha_{it} &= \varphi_i y_{t-1} + \varphi_{i+1} y_{t-2} + \ldots + \varphi_r y_{t-(r-i+1)} \\ &\quad - \vartheta_{i-1}\epsilon_t - \vartheta_i \epsilon_{t-1} - \ldots - \vartheta_{r-1}\epsilon_{t-(r-i)} \quad , \end{aligned}$$

for $i = 2, \ldots, r$. Where necessary $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_{p+d+sP+sD})$ or $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{q+sQ})$ is augmented with zeros to have dimension r.

The representation of ARMA models in a state-space form is a special case of Result 3.5 in which $d=s=P=D=Q=0$. Consequently, $\varphi(B)$ and $\vartheta(B)$ reduce to $\phi(B)$ and $\delta(B)$, respectively, as in Definition 3.1. Appendix A2 contains an example of a state-space representation for an ARMA(2,2) model.

A state-space representation of an ARMA process is not unique. Wei(1993, Chapter 5) summarises Akaike's (1974,1975) Markovian representation which is valid either for univariate or for multiple time series. This is exactly the one used by the Statespace Procedure in the SAS\ETS software package (SAS,1988).

Focusing attention on the use of state-space models for survey estimation, ARMA models are expressed most conveniently by Harvey and Phillips' representation (as it will be shown in Chapter 4).

3.6 The State-Space Representation of Vector Time Series

Wei(1993, Chapter 14) presents the theory of vector (or multiple) time series whereas Reinsel(1993, chapter 7) provides a good review of different state-space formulations for vector time series. The vector ARMA model and its respective state-space representation are now introduced.

Definition 3.2 (Wei, 1993, pp.335-336)

A vector autoregressive moving average model (VARMA) for an M-dimensional multiple time series $\{y_t\}$ (with mean vector $E(y_t)$) is given by

$$\Phi(B)Y_t = \Theta(B)\epsilon_t \quad , \tag{3.16}$$

with $Y_t = y_t - E(y_t)$,

$$\Phi(B) = I - \Phi_1 B - \dots - \Phi_p B^p \quad ,$$

$$\Theta(B) = I - \Theta_1 B - \dots - \Theta_q B^q \quad ,$$

where $\Phi_1, \dots, \Phi_p, \Theta_1, \dots, \Theta_q$ are $M \times M$ coefficient matrices and ϵ_t is an M-dimensional white noise random vector with zero mean and covariance structure:

$$E(\epsilon_t \epsilon_{t-h}') = \begin{cases} \Sigma_\epsilon & h = 0 \ , \\ \mathbf{0} & h \neq 0 \ . \end{cases} \tag{3.17}$$

Result 3.6 (Reinsel, 1993, p.203-204)

An M-dimensional vector ARMA(p,q) model, as in Definition 3.2, can be represented by a state-space model, having the following observation and systems equations:

observation equation

$$\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{\alpha}_t \quad ; \tag{3.18a}$$

system equation

$$\boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}\,\boldsymbol{\eta}_t \quad , \tag{3.18b}$$

with:

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{I} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{\Phi}_2 & \boldsymbol{0} & \boldsymbol{I} & \cdots & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\Phi}_{r-1} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{I} \\ \boldsymbol{\Phi}_r & \boldsymbol{0} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \end{bmatrix} \quad ,$$

$$\boldsymbol{G} = [\boldsymbol{I}, -\boldsymbol{\Theta}_1, -\boldsymbol{\Theta}_2, \dots, -\boldsymbol{\Theta}_{r-1}]' \quad ,$$

$$\boldsymbol{\eta}_t = \boldsymbol{\epsilon}_t \quad ,$$

where $r = \max(p, q+1)$. In addition,

$\boldsymbol{H} = [\boldsymbol{I} \vdots \boldsymbol{0} \vdots \boldsymbol{0} \vdots \cdots \vdots \boldsymbol{0}]$ and $\boldsymbol{\alpha}_t = (\boldsymbol{\alpha}'_{1t}, \boldsymbol{\alpha}'_{2t}, \dots, \boldsymbol{\alpha}'_{rt})'$ is defined as

$$\boldsymbol{\alpha}'_{1t} = (y_{1t}, \dots, y_{Mt})' \quad ,$$

$$\begin{aligned} \boldsymbol{\alpha}_{it} = {} & \boldsymbol{\Phi}_i \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_{i+1} \boldsymbol{y}_{t-2} + \dots + \boldsymbol{\Phi}_r \boldsymbol{y}_{t-(r-i+1)} \\ & - \boldsymbol{\Theta}_{i-1} \boldsymbol{a}_t - \boldsymbol{\Theta}_i \boldsymbol{a}_{t-1} - \dots - \boldsymbol{\Theta}_{r-1} \boldsymbol{a}_{t-(r-i)} \quad , \quad i = 2, \dots, r \quad . \end{aligned}$$

Where necessary $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p)$ or $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q)$ is augmented with zero matrices.

Having introduced the state-space representation for ARMA and vector ARMA time series, it is important to consider another approach widely used for representing the signal process $\{ \theta_t \}$. It is concerned with Structural Time Series Models (Harvey, 1989), which is introduced in the next section.

3.7 Structural Time Series Models in State-Space Form

3.7.1 Preliminaries

The principal feature of structural time series models is that they are formulated in terms of components which have a direct interpretation such as trend, seasonals and cycles. Moreover, structural time series models are set up in such a way that the unobservable components are stochastic. Since each component is influenced by a disturbance term, the structural time series can be represented in a state-space form, with the state of the system representing the various unobservable components. Then the Kalman Filter can be used to update the state as new observations become available. Hence, both ARIMA and structural time series models can be viewed as special classes of state-space models.

The idea of representing time series by unobservable components models is not a recent one. Nerlove et al.(1979) provide a good review of the historic use of this approach. They also present the unobservable components models as a possible extension of ARMA models. In this case each component is represented via an ARMA process and its optimal estimate (or smoothed value) is obtained as a solution to a signal extraction problem. Engle(1978) extended this framework for the case in which the unobservable components are assumed to follow ARIMA models. Later Bell(1984) established the results to extend Whittle's(1983) signal extraction formulation for the case of non-stationary unobservable components models.

Corresponding to each structural model there is a *reduced form model* that includes only observable variables. The reduced form of a structural model is an ARIMA model. For details on structural time series models the reader is referred to Harvey(1989), which is a book dedicated to the analysis of structural time series models and the Kalman Filter. Sections 3.7.2 and 3.7.3 below describes some univariate and multivariate structural time series models.

3.7.2 Univariate Structural Time Series Models

Consider the **local level** or **random walk plus noise** model defined by:

$$\begin{cases} y_t = \mu_t + \varepsilon_t \ , \\ \mu_t = \mu_{t-1} + \eta_t \ , \end{cases} \tag{3.19}$$

where μ_t is the trend or level, $\varepsilon_t \sim N(0,\sigma_\varepsilon^2)$, $\eta_t \sim N(0,\sigma_\eta^2)$ for $t=1,\dots,T$.

Note that the disturbances can also be considered pure white noise without normality assumptions. Examining the model in (3.19) in the light of the state-space formulation (3.1), it becomes clear that model (3.19) corresponds to a state-space model with one-dimensional state α_t equal to μ_t , and with transition matrix $\boldsymbol{T}$, as well as $\boldsymbol{H}$ and $\boldsymbol{G}$, now having a single element. The only parameters of the model which require estimation are σ_ε^2 and σ_η^2 . If the trend component of a structural model is not generated by a random walk, one can add a stochastic linear trend to the model. The **local linear trend model** is given by:

$$\begin{cases} y_t = \mu_t + \varepsilon_t \ , \\ \mu_t = \mu_{t-1} + \beta_{t-1} + \omega_t \ , \\ \beta_t = \beta_{t-1} + \xi_t \ , \end{cases} \tag{3.20a}$$

where ε_t , ω_t and ξ_t are mutually uncorrelated normally distributed disturbances with mean zero and variances σ_ε^2 , σ_ω^2 and σ_ξ^2 . Alternatively the disturbances could be white noise sequences.

The local linear trend model (3.20a) can be written in a state-space form as:

$$y_t = \boldsymbol{H}\boldsymbol{\alpha}_t + \varepsilon_t \quad , \tag{3.20b}$$

$$\boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \quad , \tag{3.20c}$$

where

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} , \boldsymbol{H} = [1\ 0] , \boldsymbol{T} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} , \boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} , \boldsymbol{\eta}_t = \begin{bmatrix} \omega_t \\ \xi_t \end{bmatrix} ,$$

with $V(\varepsilon_t) = \sigma^2_\varepsilon$ and $V(\eta_t) = \boldsymbol{Q}_t = \begin{bmatrix} \sigma^2_\omega & 0 \\ 0 & \sigma^2_\xi \end{bmatrix} = diag(\sigma^2_\omega, \sigma^2_\xi)$.

A widely used, yet more elaborate model is the **basic structural model (BSM)**, with observation equation given by:

$$y_t = \mu_t + \gamma_t + \varepsilon_t \quad ,$$

where μ_t is the trend or level, γ_t is the seasonal component and ε_t is the irregular component at time t . The system equations that describe the evolution of μ_t and γ_t are given by:

$$\begin{cases} \mu_t = \mu_{t-1} + \beta_{t-1} + \omega_t \ , \\ \beta_t = \beta_{t-1} + \xi_t \ , \\ \gamma_t = -\sum_{j=1}^{s-1} \gamma_{t-j} + \nu_t \ , \end{cases} \tag{3.21}$$

where s is the number of seasons and $\varepsilon_t, \omega_t, \xi_t, \nu_t$ are mutually uncorrelated normally distributed (or white noise) disturbances with zero mean and variances $\sigma^2_\varepsilon, \sigma^2_\omega, \sigma^2_\xi, \sigma^2_\nu$, respectively.

The use of the basic structural model is illustrated in Chapter 4. The next section introduces the multivariate structural time series models.

3.7.3 Multivariate Structural Time Series Models

Let $y_1, y_2, \ldots, y_T$ be a sequence of $M \times 1$ vectors. Following Harvey(1989, pp.429), the multivariate structural framework allows the unobservable components to be contemporaneously correlated. As in the univariate case, the simplest structural model is the **local level** or **random walk plus noise** model defined as:

$$\begin{cases} y_t = \mu_t + \varepsilon_t \ , \\ \mu_t = \mu_{t-1} + \eta_t \ , \end{cases} \qquad \textbf{(3.22)}$$

where μ_t is a $M \times 1$ vector of trend/level components and ε_t and η_t are $M \times 1$ vectors normally distributed with mean zero and covariance matrices Σ_ε and Σ_η for $t = 1, \ldots, T$. Note that the relation between the series arises via the off-diagonal elements of the disturbance covariance matrices. In a multivariate framework the disturbances can also be considered pure white noise without normality assumptions. Comparing model (3.19) with model (3.22) it becomes clear that the latter is a straightforward generalization of the univariate local level model.

The same is valid for the local linear trend model defined in (3.20), implying that the **multivariate local linear trend model** is given by:

$$\begin{cases} y_t = \mu_t + \varepsilon_t \ , \\ \mu_t = \mu_{t-1} + \beta_{t-1} + \omega_t \ , \\ \beta_t = \beta_{t-1} + \xi_t \ , \end{cases} \qquad \textbf{(3.23a)}$$

where ε_t , ω_t and ξ_t are $M \times 1$ mutually uncorrelated normally distributed vectors with mean zero and covariance matrices Σ_ε , Σ_ω and Σ_ξ .

The local linear trend model (3.23a) can be written in a state-space form as:

$$y_t = H\alpha_t + \varepsilon_t \quad , \tag{3.23b}$$

$$\alpha_t = T\alpha_{t-1} + \eta_t \quad , \tag{3.23c}$$

with

$$\alpha_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} \quad , \quad H = [10] \otimes I_M \quad , \quad T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \otimes I_M \quad ,$$

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes I_M \quad , \quad \eta_t = \begin{bmatrix} \omega_t \\ \xi_t \end{bmatrix} \quad ,$$

where I_M denotes a $M \times M$ identity matrix and $\otimes$ is the Kronecker product. In addition, it is assumed that:

$$V(\eta_t) = Q_t = \begin{bmatrix} \Sigma_\omega & 0 \\ 0 & \Sigma_\xi \end{bmatrix} \quad .$$

Finally, as expected, the **multivariate basic structural model** is a generalization of model (3.21). Its detailed specification is presented later in Sections 6.4 and 8.2.2 when multivariate basic structural models are employed to model compositional survey data.

The state-space formulation introduced in this chapter can be used to provide signal extraction results in the analysis of repeated surveys. In the presence of sampling error, the signal (the unknown population quantity) is not directly observable. Instead, an estimate or a vector of elementary estimates is observed which differs from unknown signal due to sampling errors. This situation can be modelled in the state-space approach by including any component required to describe the time series model assumed for the signal process in the state-vector α_t .

Moreover, the system equation can also be used to model the sampling error. Therefore, the state-vector would contain components of the time series models for both the signal and noise. In the case of non-overlapping surveys, with independent sampling errors, the disturbance in the observation equation could be used to represent the sampling error.

In general, when dealing with survey data, the signal process is either represented by an ARMA or a Structural time series model (Harvey, 1989) whereas the sampling error is assumed to follow an ARMA model(Box & Jenkins, 1970, sect.3.4). The following chapter reviews the use of state-space models in survey estimation.

4 State-Space Models for Survey Estimation

4.1 Introduction

Two different approaches for incorporating the information contained in repeated samples were introduced in Chapter 2, namely, the classical sampling approach and the traditional time series approach. In this and subsequent chapters, the state-space approach will be considered.

Binder & Hidiroglou(1988) introduced the use of state-space models in survey estimation to overcome the practical difficulties that arise when applying Jones' result (2.12 or 2.14). Their idea was to use the Kalman Filter equations to produce the MMSE of $\{\theta_t\}$ given all the available information $\boldsymbol{D}_t$. In order to outline their procedure consider a set-up where $\boldsymbol{y}_t$ is a $(k \times 1)$ vector of elementary estimates for time t , and $\boldsymbol{e}_t$ is a vector of sampling errors, θ_t is not directly observable, and

$$\boldsymbol{y}_t = \mathbf{1}_k \theta_t + \boldsymbol{e}_t \quad , \tag{4.1}$$

where $\mathbf{1}_k$ is a unit column vector.

The time series model for the survey estimate is the combination of two distinct models. One to describe the evolution of the unobservable population quantities $\{\theta_t\}$ over time and another to represent the time series relationship between the sampling errors $\{e_t\}$ of the sample estimates. As an example (from Binder and Hidiroglou,1988), consider the simple case where $\{\theta_t\}$ is an AR(1) process, $\{e_t\}$ is a MA(1), y_t and e_t are scalars, meaning that just one elementary estimate is available on each occasion. The model for y_t is then given by:

$$\begin{cases} y_t = \theta_t + e_t \\ \theta_t = \phi\,\theta_{t-1} + \xi_t \\ e_t = \gamma_t - \beta\,\gamma_{t-1} \end{cases} \quad \text{or} \quad \begin{cases} y_t = \theta_t + e_t \\ \theta_t = \phi\,\theta_{t-1} + \xi_t \\ e_t = b_{t-1} + \gamma_t \\ b_t = -\beta\,\gamma_t \end{cases} .$$

The corresponding state-space formulation is:

$$\begin{cases} y_t = H\alpha_t \quad , \\ \alpha_t = T\alpha_{t-1} + G\eta_t \quad , \end{cases} \tag{4.2a}$$

where:

$$\alpha_t = \begin{bmatrix} \theta_t \\ e_t \\ b_t \end{bmatrix}, H' = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, T = \begin{bmatrix} \phi & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -\beta \end{bmatrix}, \eta_t = \begin{bmatrix} \xi_t \\ \gamma_t \end{bmatrix} . \tag{4.2b}$$

It becomes clear from equations (4.2) that the complete model for $\{\theta_t, e_t, y_t\}$ can be formulated as a state-space model where the state-vector includes components from both the $\{\theta_t\}$ and $\{e_t\}$ processes.

When working with state-space models for repeated surveys, the disturbance term in the observation equation will not be used to model the sampling error, because the latter can be correlated over time due to the presence of overlapping units. Instead, the sampling errors must be included in the state-vector α_t. Hence, in general, state-space models for data from overlapping surveys have the form (4.2a), that is, have no disturbance term in the observation equation. Authors who worked with such models discovered that ignoring the serial correlation of the sampling errors may imply biased estimates for the model parameters (see, for example, Binder & Dick, 1989 and Binder, Bleuer and Dick, 1993).

Using the Kalman Filter equations in (3.5), an estimator for α_t given all available data at time t is given by:

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + P_{t|t-1} H' (H P_{t|t-1} H')^{-1} (y_t - H\hat{\alpha}_{t|t-1}) \quad , \tag{4.3a}$$

with error covariance matrix

$$P_{t|t} = P_{t|t-1} - P_{t|t-1} H' (H P_{t|t-1} H')^{-1} H \quad . \tag{4.3b}$$

From (4.2b) and (4.3a), it follows that $\hat{\theta}_{t|t} = L' \hat{\alpha}_{t|t}$ with $L' = (1\,0\,0)$ and associated error covariance matrix $L' P_{t|t} L$. Therefore:

$$\begin{aligned}
\hat{\theta}_{t|t} &= L' \hat{\alpha}_{t|t-1} + L' P_{t|t-1} H' (H P_{t|t-1} H')^{-1} (y_t - H \hat{\alpha}_{t|t-1}) \\
&= \hat{\theta}_{t|t-1} + L' P_{t|t-1} H' (H P_{t|t-1} H')^{-1} (y_t - \hat{y}_{t|t-1}) \qquad \textbf{(4.4)} \\
&= \hat{\theta}_{t|t-1} + L' K_t (y_t - \hat{y}_{t|t-1}) \quad ,
\end{aligned}$$

where $\hat{\theta}_{t|t-1}$ is the first component of $\hat{\alpha}_{t|t-1}$ and

$$K_{t|t-1} = P_{t|t-1} H' (H P_{t|t-1} H')^{-1} \quad .$$

Note that

$$L' P_{t|t-1} H' = COV(L'\alpha_t, y_t \,|\, D_{t-1}) = COV(\theta_t, y_t \,|\, D_{t-1}) \quad . \qquad \textbf{(4.5)}$$

Using the Kalman Filter equations, the observed sample estimate at time t can be separated into its signal and noise components as:

$$y_t = L' \hat{\alpha}_{t|t} + (H - L') \hat{\alpha}_{t|t} = \hat{\theta}_{t|t} + \hat{e}_{t|t} \quad ,$$

where $(H - L') = (0\ 1\ 0)$ is exactly the vector of coefficients that should be used to extract $\hat{e}_{t|t}$ from $\hat{\alpha}_{t|t}$. Both $\hat{\theta}_{t|t}$ and $\hat{e}_{t|t}$ can be interpreted as a composite-type estimator that allows the combination of an estimate based on past data with current sample information, as in (4.4), to obtain an improved estimate. Moreover, from (4.4), note that K_t can be interpret as a ratio between $COV(\theta_t, y_t \,|\, D_{t-1})$ and $V(y_t \,|\, D_{t-1})$. Then, rewriting (4.4) as

$$\hat{\theta}_{t|t} = L' K_t y_t + L' (1 - K_t H) \hat{\alpha}_{t|t} \quad ,$$

shows that, in both the time series and state-space approaches, the smaller the variance of the sample estimate, the greater its weight, and thus the closer $\hat{\theta}_{t|t}$ is to the current sample estimate y_t .

Before proceeding, it is important to note that the application of the classical signal extraction procedure (Section 2.3) to improve estimation in repeated surveys is based on the assumption that $\{\theta_t\}$ and $\{e_t\}$ are uncorrelated time series. To support the use of this approach, a key result was provided by Bell & Hillmer(1990), namely that, if y_t is a design unbiased estimator for θ_t then $\{\theta_t\}$ and $\{e_t\}$ are uncorrelated time series.

However, when using the state-space approach for modelling repeated surveys this assumption can be dropped. This is possible because both the signal θ_t and the sampling error e_t are placed in the state-vector α_t which can accommodate correlated components. Note that the basic assumption in the state-space approach is that the model disturbances should be mutually uncorrelated over time and also uncorrelated with the initial state-vector (Chapter 3, p.18). Moreover, a class of state-space models which are suitable for representing data from overlapping surveys have the form of (4.2a), with no disturbance term in the observation equation. Therefore the presence of correlated components in the state-vector when no disturbance term is added to the observation equation does not violate any model assumption. Indeed, the state-space representation of ARMA models (see Result 3.5) contains correlated components in the state-vector, since it is comprised of linear combinations of the observed series. Note, in addition, that there is no disturbance term in the observation equation. This is a quite interesting feature of the state-space approach since having both, signal and noise, depending on the same populations units it is not natural to assume that $\{\theta_t\}$ and $\{e_t\}$ are uncorrelated. For example when modelling series of estimated proportions, the variance of the sampling error is a function of the unobservable population proportion. If y_t is the unbiased estimator for a proportion θ_t under simple random sampling, then:

$$V(e_t \,|\, \theta_t) = V(y_t - \theta_t \,|\, \theta_t) = V(y_t \,|\, \theta_t) = \frac{(N_t - n_t)}{n_t(N_t - 1)} \, \theta_t(1-\theta_t) \quad ,$$

where N_t and n_t are the sizes of the finite population and the sample on time t .

Hence the assumption of uncorrelatedness of the signal and noise processes should not be considered when using state-space models for estimation in overlapping surveys .

The next section gives a brief review of how state-space models can be used to improve survey estimates. The papers by Binder & Hidiroglou(1988), Binder & Dick(1989,1990), Pfeffermann(1991), Binder, Bleuer & Dick(1993), Pfeffermann, Bell & Signorelli(1996) illustrate the usefulness of this approach.

4.2 Models for Single Time Series and for Series of Cross-Sectional Data

As mentioned before, when working with the time series or state-space approach, the model for the survey estimates is a combination of the models adopted for both $\{ \theta_t \}$ and $\{ e_t \}$.

As argued by Pfeffermann(1991), the time series models for $\{ \theta_t \}$ and $\{ e_t \}$ depend on the survey design and on the pattern of sample overlap (if any). In addition, they may vary according to the level of data available which determines when primary or secondary analysis is to take place. Moreover, the models are influenced by the presumed long term behaviour of the population means and their components. Models for single time series use either elementary (rotation group) estimates or the average of the rotation group estimates as input data.

Binder & Hidiroglou(1988), Binder & Dick(1989,1990), Pfeffermann(1991) and Tiller(1992) focused their work on the single time series case, estimating aggregate population means or totals. In all these papers, except Pfeffermann(1991), a secondary analysis was performed on the series of aggregate estimates. Pfeffermann(1991) used, instead, the elementary estimates as input data.

The papers by Pfeffermann, Burck & Ben-Tuvia(1989), Pfeffermann & Burck(1990), Pfeffermann & Bleuer(1993) and Pfeffermann, Bell & Signorelli(1996) propose models that use the rotation group estimates for each small area as the input data, in the context of small area estimation. The rest of this chapter examines these works, since they represent the state of the art regarding the use of state-space models for survey estimation.

Binder & Hidiroglou(1988) fitted an ARMA model to the Canadian Travel Survey assuming independent sampling errors. Binder & Dick(1989,1990) modelled the monthly number of unemployed people from the Canadian Labour Force Survey. In this rotating panel survey, each panel (containing one-sixth of the selected households) remains in the sample for six consecutive months. They assumed that the series of sample estimates could be decomposed as

$$y_t = x_t' \gamma + \theta_t + e_t \quad ,$$

where $x_t' \gamma$ was a deterministic regression term, $\{ \theta_t \}$ was represented by an ARIMA(1,1,0) model and $\{ e_t \}$ by an ARMA(3,6).

Pfeffermann(1991) fitted a basic structural model similar to (3.21) to the Israeli Labour Force Survey, which is a quarterly survey of households with four rotation groups in every quarter. On each survey occasion, one panel of households is fresh and other three have already been included in the sample on some occasion in the past. Every new panel is included in the survey for two quarters, is rotated out for two quarters and comes back for only two more survey rounds. Pfeffermann considered the elementary estimates as input data but also performed a secondary analysis, using the average of the rotation group estimates, and compared the two approaches. His model also assumed that observations y_{ti} belonging to the same household i (individual sampling units) follow an AR(1) model of the form

$$\begin{aligned} y_{ti} - \theta_t &= \rho \ (y_{t-1,i} - \theta_{t-1}) + \nu_{ti} \quad , \\ e_{ti} &= \rho \qquad e_{t-1,i} \qquad + \nu_{ti} \quad , \end{aligned}$$

where $|\rho| < 1$ and $\{\nu_{ti}\}$ $t = 2,3,...$ are white noise with mean zero and variance σ_ν^2 .

When using the individual rotation group estimates as input data, Pfeffermann(1991) also assumed that the four rotation groups, of equal size M, surveyed in a certain occasion were independent. So each rotation group was considered as an independent simple random sample of households. He also included rotation group bias as a linear regression term in the model. The model was fitted to simulated data as well as to a series of the number of hours worked and the number of weeks worked.

Once the model parameters have been estimated using the prediction error decomposition form of the likelihood (as in 3.13), the Kalman Filter equations can be applied to estimate the population mean and its components (such as seasonal effects) using linear combinations of the state vector as in (4.4). Pfeffermann points out that this approach enables the decomposition of the population mean using more information than is usually considered by the traditional procedures for seasonal adjustment like, for example, X11-ARIMA (Dagum, 1980).

Another example of the use of state-space models in survey estimation can be found in Tiller(1989). He fitted a state-space model to the unemployment rate series generated by the U.S. Current Population Survey (CPS) which has a 4-8-4 rotation scheme (see Bureau of the Census, 1978). The observed CPS estimate of the unemployment rate y_t is once again represented by the sum of two processes $\{\theta_t\}$ and $\{e_t\}$. The unobservable population mean process $\{\theta_t\}$ was modelled as a function of observable economic variables (assumed to be independent of the sampling error in the observed series) as:

$$\theta_t = X_t \beta_t + v_t \quad , \qquad (4.6)$$

where X_t is a $1 \times k$ vector of observed regressor variables, β_t is a $k \times 1$ vector of stochastic coefficients treated as varying according to an AR(1) process, written as

$$\beta_t = T_\beta \beta_{t-1} + \boldsymbol{\nu}_t \quad , \qquad (4.7a)$$

where T_β is a $k \times k$ matrix of fixed parameters and $\boldsymbol{\nu}_t$ is a $k \times 1$ vector of white noise disturbances with covariance matrix $V = DIAG(\sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_k})$.

The disturbance term v_t in (4.6) is the part of signal that is not accounted for by the regression term. Since v_t can be serially correlated, it is represented by the following ARMA(p,q) model

$$\phi_v(B)v_t = \delta_v(B)\epsilon_{vt} \quad , \tag{4.7b}$$

where ϵ_{vt} is white noise with variance $\sigma^2_{\epsilon_v}$.

The noise component e_t is used to incorporate features of the CPS sample design into the modelling process. It represents the sampling error and is modeled by:

$$e_t = \gamma_t e_t^* \quad , \tag{4.8a}$$

where γ_t accounts for changes in variance and e_t^* follows an ARMA(p*,q*) model:

$$\phi_e(B)e_t = \delta_e(B)\epsilon_{et} \quad , \tag{4.8b}$$

where ϵ_{et} is white noise with variance $\sigma^2_{\epsilon_e}$.

The model defined by equations (4.6), (4.7) and (4.8) can be written in a state-space form with the unobservable signal and the noise as state variables. For this specific problem the unobservable variables are β_t , v_t and e_t^* . Since v_t and e_t^* follow, respectively, ARMA(p,q) and ARMA(p*,q*) models, they are converted into vectors α_{vt} and α_{et} , using the procedure introduced in Result 3.5. Then, the state-space model is:

$$\left\{ \begin{aligned} y_t &= H_t \alpha_t \quad , \\ \alpha_t &= T\alpha_{t-1} + G\eta_t \quad , \end{aligned} \right. \tag{4.9}$$

with:

$$\alpha_t = (\beta_t', \alpha_{vt}', \alpha_{et}')' \quad ,$$

where α_{vt} and α_{et} are, respectively, $r \times 1$ and $r^* \times 1$ state vectors, as in Result 3.5, with $r = \max(p, q+1)$ and $r^* = \max(p^*, q^*+1)$. In addition,

$$H_t = (X_t, 1, \mathbf{0}_{r-1}, \gamma_t, \mathbf{0}_{r^*-1}) \quad ,$$

$$T = DIAG(T_\beta, T_\upsilon, T_e) \quad , \quad G = DIAG(I_{k\times k}, G_\upsilon, G_e) \quad ,$$

$$T_\upsilon = \begin{bmatrix} \phi_{\upsilon 1} & I_{(r-1)\times(r-1)} \\ \vdots & \\ \vdots & \\ \phi_{\upsilon r} & 0 \end{bmatrix} \quad , \quad T_e = \begin{bmatrix} \phi_{e1} & I_{(r^*-1)\times(r^*-1)} \\ \vdots & \\ \vdots & \\ \phi_{er^*} & 0 \end{bmatrix} \quad ,$$

$$G_\upsilon' = [1, -\delta_{\upsilon 1}, \ldots, -\delta_{\upsilon(r-1)}] \quad , \quad G_e' = [1, -\delta_{e1}, \ldots, -\delta_{e(r^*-1)}] \quad ,$$

$$\eta_t = (\upsilon_t', \epsilon_{\upsilon t}, \epsilon_{et})' \quad .$$

The random disturbances are assumed to be mutually uncorrelated, hence the variance-covariance matrix of η_t is (from 4.7 and 4.8):

$$Q = DIAG(V, \sigma_{\epsilon_\upsilon}, \sigma_{\epsilon_e}) \quad .$$

The signal component can be estimated using $\hat{\theta}_{t|t} = [X_t, 0, \mathbf{0}_{(r-1)+1+(r^*-1)}]\, \hat{\alpha}_{t|t}$.

Tiller(1992) extended the previous model by adding a trend, a seasonal and an irregular component to the signal decomposition. He reported encouraging results about the reduction in variance over the traditional survey estimation but also pointed out that, as expected, care is required when choosing the models for the signal and sampling error processes. Pfeffermann & Burck(1990) reported that the time series and state-space approaches are not routinely used by statistical agencies because the classical survey estimators of the aggregate means are usually as efficient as the model-based ones when the model holds and more robust when the model fails to hold. However, the use of structural time series models allows for the estimation of unobservable signal components such as trend, seasonals and cycles while taking into account the sampling errors.

Pfeffermann, Burck & Ben-Tuvia(1989), Pfeffermann & Burck(1990), Pfeffermann & Bleuer(1993), Pfeffermann, Bell & Signorelli(1996) developed state-space models for dealing with a time series of cross-sectional data. As stated before, the use of a model-based approach is suitable for cases in which the survey error component is large. This generally occurs in the context of small area (or domain) estimation due to small sample sizes. The authors considered a model to improve estimation in small domains that exploits both cross-sectional relationships between small areas and time series properties of the data.

Pfeffermann & Bleuer(1993) modelled the Canadian Labour Force Survey. A basic structural model, adapted to the case of a monthly survey (12 seasonals) with six rotation groups each month, was used to represent the signal of each small area. They also assumed separate autoregressive relationships for the six sampling error series, corresponding to the six monthly elementary estimates. For joint modelling of small areas, assuming that the sampling errors were independent between areas, just the model for the signal $\{ \theta_t \}$ was extended to account for the cross-sectional correlations. Then the model allowed for non-zero contemporary correlation between the error terms corresponding to the unobservable signal components. Working with this assumption means that if there is an increase in the trend level in one small area, similar increases are expected to occur in other areas. The final model was constructed by combining the models for all small areas.

To protect the procedure against model failures, monthly sample estimates in the aggregate level were forced to coincide with the model-based ones. This was done by adding linear constraints to the observation equation. Moreover, the model also accounted for changes in the variances of the sampling errors over time as well as for rotation group bias.

Pfeffermann & Bleuer examined the results of fitting this state-space model to the unemployment rate series. They compared the model-based estimates with the traditional design-based ones, reporting that the two sets of estimates behaved very similarly. However the model-based estimates (for the joint model) were more stable and had smaller standard error than the survey estimates. The estimates for the seasonal effects produced by the modelling procedure and by using X11-ARIMA were also compared. The seasonal effects produced by the two approaches were very close, the same happening for the trend estimate although in this case the X11 curve was smoother than the model curve.

Although some improvement in survey estimation can be achieved by modelling the time series relationships of the survey estimates, this approach is model dependent. Binder, Bleuer & Dick(1993) pointed out that a data set can match with a wide class of models. To examine this issue they fitted two different models to a series of the number of unemployed people from the Canadian Labour Force Survey.

One of the models used the log-transformed series as inputs, employing a secondary analysis in which just the aggregate estimate was considered. The unobservable population mean was assumed to follow a seasonal ARIMA(1,1,0)(1,1,0)$_{12}$ model. The model for the sampling errors was given by : $e_t = k_t \epsilon_t$, where k_t was the standard error estimated from the survey and ϵ_t assumed to follow an ARMA(3,6). The standard error for the log-transformed data was computed by using Taylor linearization.

The other model was the one described in Pfeffermann & Bleuer(1993) which uses the original monthly elementary estimates as the input data. They reported that both models appear to give estimates which are consistent with the data, concluding that models with entirely different structures can provide acceptable results.

In a recent work, Pfeffermann, Bell and Signorelli(1996) fitted state-space models to the Australian regional unemployment and labour force participation rate series. Their aim was to provide estimates for the trend of the series free from spurious cycles possibly induced by the autocorrelations of the sampling errors. The authors claimed that these cycles are not identified or removed when using standard procedures for seasonal adjustment such as the X-11 program. Using a basic structural model for the signal and an AR(2) model for the noise they produced estimates for the trend which were smoother than those obtained from the X-11.

This section has described the use of state-space models in survey estimation. It is interesting to note that some of these models were fitted to series of proportions. In fact, both Pfeffermann & Bleuer(1993) and Tiller(1992) implemented their procedures in order to improve estimation of unemployment rates. Hence, it is reasonable to ask whether the predictions for y_t and the filtered estimates for θ_t provided by those models are always bounded between zero and one.

Having reviewed the literature regarding the use of state-space models to improve estimation in repeated survey it becomes clear that no attempt has been made to model simultaneous more than one survey variable. Indeed, Binder, Bleuer and Dick(1993) reported that little work had been done on taking advantage of the correlations among variables in a repeated survey. Most of the work they reviewed dealt with the case of small area estimation where the same target variable is considered across different small domains. In fact, although much has already been discussed regarding modelling univariate series of survey estimates, there is little discussion of models for multivariate survey data in the literature. Chapter 6 and 7 introduce the use of multivariate time series models for survey estimation. Before that, however, the issue of modelling series of proportions will be examined; firstly in the context of time series analysis without taking into account the sampling error and then considering that the observed series are subject to sampling errors.

5 Modelling Time Series of Proportions

5.1 Introduction and Review of Existing Approaches

The problem of modelling both univariate and multivariate series of proportions was studied by Wallis(1987) and Brunsdon(1987). One of the difficulties in modelling time series of proportions arises because proportions are bounded between zero and one, but standard time series models are not similarly constrained. Thus there is no guarantee that the forecasts obtained when fitting standard time series models will belong to the interval $[0,1]$.

Wallis(1987) applied the logistic transformation, suggested by Aitchison & Shen(1980), to time series known to be bounded between zero and one. These series are typically measured as proportions or percentages, like unemployment rates. For a univariate time series $\{y_t\}$, such that $0 \leq y_t \leq 1 \;\; \forall \; t$, he recommended using the transformed series $\{y_t^*\}$ where

$$y_t^* = \log \left[\frac{y_t}{1 - y_t} \right] \quad ,$$

and to model the transformed series using conventional ARMA models.

Smith & Brunsdon(1986) pointed out, however, that if the proportions $\{y_t\}$ are all in the range $[0.2,0.8]$ they can be modelled in the conventional way with little risk of producing predictions lying outside $[0,1]$. But they agreed that if a series of proportions evolves close to any of the boundaries of the domain, the use of the logistic transformation is recommended.

Brunsdon(1987) considered the case of multiple time series in which each of the series has values bounded between zero and one and, moreover, the sum of the series equals one at each time point. Data with such characteristics (unity-sum constraint and bounded between 0 and 1) are known as compositional data. Formally, a compositional time series is defined (see Brunsdon, 1987, p.75) as a sequence of vectors $y_t = (y_{1t}, \dots, y_{M+1,t})'$ belonging to the

Simplex S^M , defined as

$$S^M = \{ y : 0 \leq y_m \leq 1 , m = 1,...,M+1 ; \sum_{i=1}^{M+1} y_m = 1 \} \quad .$$

Aitchison(1986) examined the difficulties of applying standard methods to modelling and analysing compositions. For instance, the interpretation of the covariance structure of a composition is different from the standard interpretation of covariances and correlation between components of an unrestricted vector. For example, considering $y_t = (y_{1t}, y_{2t}, y_{3t})'$ it follows from the properties of a composition that

$$COV(y_{1t} \, , \, y_{1t} + y_{2t} + y_{3t}) = COV(y_{1t}, 1) = 0 \quad .$$

But this also implies that

$$COV(y_{1t}, y_{1t}) + COV(y_{1t}, y_{2t}) + COV(y_{1t}, y_{3t}) = 0 \quad ,$$

and consequently

$$COV(y_{1t}, y_{2t}) + COV(y_{1t}, y_{3t}) = - VAR(y_{1t}) \quad ,$$

meaning that at least one of the covariances on the left-hand side must be negative. This feature of compositional data is known as the *negative bias difficulty*.

Another issue examined by Aitchison(1986) is the lack of parametric distributions defined on the appropriate Simplex. One class of distributions which was considered for describing compositional data is the Dirichlet. However, every Dirichlet composition can be considered as being formed from a basis of independent Gamma random variables(Aitchison, 1986, p.59). This independence property of the basis components does not match with the characteristics of many compositions found in practice. Therefore the Dirichlet class is not always suitable for the parametric modelling of compositions.

To tackle these problems Aitchison(1986) suggested the use of transformations to map compositions from the Simplex S^M onto $\mathbb{R}^M$. One such transformation is the *additive* *logratio transformation* (a_M) defined in Aitchison(1986, p.113) and first adopted for a time series context by Brunsdon(1987, p.75), given by $\boldsymbol{v}_t = a_M(\boldsymbol{y}_t) = (v_{1t}, \dots, v_{Mt})'$, with

$$v_{mt} = \log\left[\frac{y_{mt}}{y_{M+1,t}}\right] \quad , \quad m = 1,\dots,M \quad , \quad \forall\, t \quad , \tag{5.1}$$

where *log* denotes the natural logarithm. Note that $y_{M+1,t} = 1 - \sum_{m=1}^{M} y_{mt}$, usually called the fill-up value, is used as the reference variable or category.

The inverse transformation, known as the *additive logistic transformation* (a_M^{-1}), is given by $\boldsymbol{y}_t = a_M^{-1}(\boldsymbol{v}_t) = (y_{1t}, \dots, y_{M+1,t})'$ such that

$$y_{mt} = \begin{cases} \dfrac{\exp(v_{mt})}{1 + \sum_{j=1}^{M} \exp(v_{jt})} & m = 1,\dots,M \quad , \quad \forall\, t \quad , \\[2ex] \dfrac{1}{1 + \sum_{j=1}^{M} \exp(v_{jt})} & m = M+1 \quad , \quad \forall\, t \quad . \end{cases} \tag{5.2}$$

Regarding the choice of a parametric distribution defined on the Simplex S^M , Aitchison & Shen(1980) introduced the *additive logistic normal distribution* :

$$L^M(\boldsymbol{\mu}, \Sigma) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}} \prod_{m=1}^{M+1} y_{mt}} \exp\left\{ -\frac{1}{2}(\boldsymbol{v}_t - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{v}_t - \boldsymbol{\mu}) \right\} \quad , \tag{5.3}$$

where $\boldsymbol{v}_t = a_M(\boldsymbol{y}_t)$ is as defined above.

The authors showed that when the logratios $\boldsymbol{v}_t$ are normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ , i.e. $\boldsymbol{v}_t \sim N_M(\boldsymbol{\mu}, \Sigma)$, the M+1-part composition has an additive logistic normal distribution as defined in (5.3).

Working with compositional time series, Brunsdon(1987) recommended the use of transformations to map the series of compositions from the Simplex onto the Real space before using Vector ARMA models (Tiao & Box,1981) for estimation and analysis. Note that in the alternative, if applying standard univariate time series methods of analysis to each of the components of a composition, there is no guarantee either that the forecasts will be bounded between zero and one, or that their sum will be one. Although the application of a logistic transformation to each of the individual series of proportions would guarantee individual forecasts restricted to the interval $[0,1]$ it would still not satisfy automatically the vital unity-sum constraint. She also proved an important invariance property that when the vector of transformed series computed using the fill-up value as the reference variable (as in (5.1)) follows a VARMA(p,q) process of dimension M , then transformed vectors based on any other reference variable $(y_{mt}\ ,\ m = 1,...,M)$ also follow a VARMA(p,q), all of them representing the same model for the original composition. This implies that any vector component can be selected as the reference variable for the transformation without affecting the results of the modelling procedure. This approach for analysing compositional time series is quite interesting because, in general, the behaviour of the complete vector is of interest, not only of one of its components. Treating the set of series as a multiple time series enables the analyst to study the dynamics of relationships between the components.

The work of Wallis(1987) and Brunsdon(1987) provides useful insight about modelling time series of proportions. However, these authors did not consider the fact that the observed time series are often subject to sampling errors. They also did not make use of structural time series models or, more generally, state-space models. Hence this thesis extends their work by focusing on the development of a state-space approach for modelling proportions and compositions from repeated surveys taking into account the sampling errors. This raises the issue of whether it is necessary to transform the data in a state-space framework in order to guarantee that the signal estimates and observation predictions are always bounded between zero.

Hence, before proceeding further in the direction of modelling compositional survey data, it is important to take a decision about whether or not a transformation should be applied to the data. Although Wallis(1987) and Brunsdon(1987) recommended the use of transformations before fitting ARMA or VARMA models, other authors such as Tiller(1992), Pfeffermann & Bleuer(1993) and Pfeffermann, Bell & Signorelli(1996) used a state-space approach for modelling "raw" unemployment rate series. This could be an indication that, by using a specific state-space model, the problem of modelling "raw" proportions is under control. If so, what are the features of this formulation that lead to such results? To investigate this issue consider the following conjecture.

Conjecture 5.1

The use of the state-space approach for modelling proportions based on repeated surveys guarantees that the filtered estimates $\hat{\theta}_{t|t}$ are always bounded between zero and one.

If this conjecture was always true, it would imply that there was no need of a transformation. On the other hand, if the conjecture fails to hold, the models for series of proportions (and consequently for compositions) should be applied to the transformed data. To simplify the discussion, the use of structural time series models will initially be considered without taking into account the sampling errors. So, to begin with, the case of a local level model for a univariate time series of proportions will be analysed.

5.2 The Local Level Model for Proportions

Consider the local level model for a single series of estimated proportions $\{y_t\}$ given by,

$$\begin{cases} y_t = \theta_t + e_t \ , \\ \theta_t = \theta_{t-1} + \eta_t \ , \end{cases} \tag{5.4}$$

where θ_t is the unobservable level (or true unobservable proportion), e_t and η_t are mutually uncorrelated normally distributed disturbances, with mean zero and variances σ_e^2 and σ_η^2 , respectively.

The idea is to verify whether the use of the state-space approach for modelling the observed series of proportions $\{y_t\}$ guarantees that the signal estimates $\hat{\theta}_{t|t}$ and also the predictions $\hat{y}_{t+1|t}$ are bounded between zero and one. It is then necessary to examine the estimates obtained via the Kalman Filter, and to check whether these estimates lie inside $[0\ ,\ 1]$. Ideally, to carry out the analysis, these estimates should be expressed in terms of the observed series $\{y_t\}$.

Therefore, for each of the models examined in this chapter, explicit expressions for the filtered estimates are provided in terms of the observations. These expressions are obtained by using the steady-state filter (see Chapter 3, p.25) corresponding to each model. For the local level model considered in this section, the steady-state filtered estimate for θ_t has the form (details in Appendix B1):

$$\hat{\theta}_{t|t} = (1 - \lambda)\hat{\theta}_{t-1|t-1} + \lambda y_t \quad , \tag{5.5}$$

with

$$\lambda = \frac{P}{P + \sigma_\varepsilon^2} \quad , \tag{5.6}$$

and P is a solution to equation (3.11a). Then $\hat{\theta}_{t|t}$ has the form of an *exponentially weighted moving average (EWMA)* with smoothing constant λ given by (5.6) where $0 \leq \lambda \leq 1$ because P is strictly positive. By repeated substitution, it follows that

$$\hat{\theta}_{t|t} = \lambda \sum_{j=0}^{t-1} (1 - \lambda)^j y_{t-j} + (1 - \lambda)^t \theta_0 \quad . \tag{5.7}$$

Assuming that $\theta_0 = 0$ and that $\{y_t\}$ is a series of proportions, equation (5.7) expresses $\hat{\theta}_{t|t}$ as a linear combination of $y_1,...,y_t$ with $y_t \in [0,1]\ \forall\ t=1,...,t$. In order to guarantee that the estimate $\hat{\theta}_{t|t}$ is bounded between zero and one it is sufficient to have the sum of the weights in the linear combination (5.7) less than or equal to one, since each of the weights is also bounded between zero and one, i.e. since $0 \le \lambda(1-\lambda)^j \le 1 \quad \forall\, j$ because $0 \le \lambda \le 1$. Hence it suffices to show that

$$\lambda \sum_{j=0}^{t-1} (1 - \lambda)^j \le 1 \tag{5.8}$$

to prove that $\hat{\theta}_{t|t} \in [0,1]$. But this statement is true since the weights in the EWMA sum to unity. In fact, if t is large,

$$\begin{aligned}\lim_{t\to\infty} \left[\lambda \sum_{j=0}^{t-1} (1 - \lambda)^j\right] &= \lambda \lim_{t\to\infty} \sum_{j=0}^{t-1} (1 - \lambda)^j \\ &= \lambda \lim_{t\to\infty} \left[\frac{1 - (1 - \lambda)^t}{1 - (1 - \lambda)}\right] = 1 \quad .\end{aligned}$$

The one step ahead forecast for a local level model is given by $\hat{y}_{t|t-1} = \hat{\theta}_{t-1|t-1}$ so that both the model forecasts as well the estimates for the unobservable level are bounded between zero and one. Consequently, when dealing with a structural local level model there is no need to apply any transformation to the series in order to get predictions and estimates guaranteed to lie in the $[0,1]$ interval. Thus, Conjecture 5.1 is not contradicted for the local level model.

Although suitable for modelling non-overlapping surveys, the model in (5.4) does not allow for any time series structure regarding the sampling errors. Thus it seems reasonable to explore a model which includes some structure for the sampling errors and seems suitable for the case of overlapping surveys, in order to verify if Conjecture 5.1 stands for more complex models.

5.3 A Simple Model for Estimated Proportions from Completely Overlapping Surveys

Consider the case of a panel survey in which all the units are retained for all survey rounds. The effect of overlap is to introduce correlations between the sampling errors $e_t \, , \, e_{t-h} \;\forall\; h$ which must be estimated from the data. As suggested by Scott, Smith & Jones(1977), a first-order autoregressive model is commonly used for the sampling error in a complete overlapping survey. For the signal θ_t , which represents the true unobservable proportions, a structural model can be used. Here the simplest of the structural models is employed: the local level model. Then, the model for the survey estimate is given by:

$$\begin{cases} y_t = \theta_t + e_t \quad , \\ \theta_t = \theta_{t-1} + a_t \quad , \\ e_t = \beta\, e_{t-1} + b_t \quad . \end{cases} \tag{5.9}$$

where $\{a_t\}$ and $\{b_t\}$ are uncorrelated series of disturbances with $a_t \sim WN(0 \, , \, \sigma_a^2) \; , \; b_t \sim WN(0 \, , \, \sigma_b^2)$ and $|\beta| < 1$.

This model presents similar features to those introduced in Chapter 4, and when the sampling error is included in the state-vector there is no added disturbance term in the observation equation. Thus model (5.9) can be expressed in the state-space form as

$$\begin{cases} y_t = \boldsymbol{H}\boldsymbol{\alpha}_t \quad , \\ \boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}\boldsymbol{\eta}_t \quad . \end{cases} \tag{5.10a}$$

where

$$\alpha_t = \begin{bmatrix} \theta_t \\ e_t \end{bmatrix} , \; \boldsymbol{H} = [1 \;\; 1] , \; \boldsymbol{T} = \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix} , \; \boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} , \; \eta_t = \begin{bmatrix} a_t \\ b_t \end{bmatrix} . \qquad \textbf{(5.10b)}$$

Using the Kalman Filter equations (3.5), the filtered estimate for α_t is obtained as

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + \boldsymbol{P}_{t|t-1} \boldsymbol{H}' \boldsymbol{F}^{-1}_{t|t-1}(y_t - \hat{y}_{t|t-1}) \quad , \qquad \textbf{(5.11a)}$$

with

$$\begin{aligned} \hat{\alpha}_{t|t-1} &= \boldsymbol{T} \hat{\alpha}_{t-1|t-1} \quad , \\ \hat{y}_{t|t-1} &= \boldsymbol{H} \boldsymbol{T} \hat{\alpha}_{t-1|t-1} \quad . \end{aligned} \qquad \textbf{(5.11b)}$$

As t increases the filter can reach a steady-state. Once again, to obtain the steady-state filter, the error covariance matrix $\boldsymbol{P}$ must be the solution of the equation (3.11a) with

$$\boldsymbol{Q} = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix} .$$

Representing the steady-state covariance matrix $\boldsymbol{P}$ by

$$\boldsymbol{P} = \begin{bmatrix} p_1 & p_{12} \\ p_{12} & p_2 \end{bmatrix} , \qquad \textbf{(5.12)}$$

the steady-state filter is given by (details in appendix B2)

$$\begin{aligned} \hat{\theta}_{t|t} &= (1 - W)\hat{\theta}_{t-1|t-1} + W y_t - \beta W \hat{e}_{t-1|t-1} \quad , \\ \hat{e}_{t|t} &= \beta W \hat{e}_{t-1|t-1} + (1 - W) y_t - (1 - W)\hat{\theta}_{t-1|t-1} \quad , \end{aligned} \qquad \textbf{(5.13a)}$$

where

$$W = \frac{p_1 + p_{12}}{p_1 + 2p_{12} + p_2} \quad . \tag{5.13b}$$

Note that, by repeated substitution (details in Appendix B3):

$$\begin{aligned}\hat{\theta}_{t|t} = &\ (1 - W)[(1 - W) + \beta W]^{t-1}\theta_0 - \beta W[(1 - W) + \beta W]^{t-1}e_0 \\ &+ Wy_t + W(1 - W)(1 - \beta)\sum_{j=1}^{t-1}[(1 - W) + \beta W]^{j-1}y_{t-j} \ .\end{aligned} \tag{5.14}$$

Assuming that $\theta_0 = e_0 = 0$, it follows that the steady-state filtered estimate for θ_t is given by:

$$\hat{\theta}_{t|t} = Wy_t + W(1 - W)(1 - \beta)\sum_{j=1}^{t-1}[(1 - W) + \beta W]^{j-1}y_{t-j} \quad . \tag{5.15}$$

Using (5.15) it is possible to investigate whether the use of model (5.9) guarantees that the filtered signal estimates are bounded between zero and one, when dealing with a series of proportions. Before proceeding it is important to identify which restrictions are imposed on W . Note that W is a function of the elements of the steady-state error covariance matrix $\boldsymbol{P}$. The subsequent relations below follow directly from equation (3.11a):

$$p_1 - \frac{(p_2p_1 - p_{12}^2)}{p_1 + 2p_{12} + p_2} - \sigma_a^2 = 0 \quad , \tag{5.16a}$$

$$p_{12} + \frac{\beta(p_2p_1 - p_{12}^2)}{p_1 + 2p_{12} + p_2} = 0 \quad , \tag{5.16b}$$

$$p_2 - \frac{\beta^2(p_2p_1 - p_{12}^2)}{p_1 + 2p_{12} + p_2} - \sigma_b^2 = 0 \quad . \tag{5.16c}$$

Moreover, since $\boldsymbol{P}$ is a covariance matrix it must be nonnegative definite, that is:

$$p_2p_1 - p_{12}^2 \geq 0 \quad . \tag{5.17}$$

Also, it follows from equation (5.16a) that:

$$\frac{(p_1 + p_{12})^2}{p_1 + 2p_{12} + p_2} = \sigma_a^2 \quad ,$$

implying that

$$p_1 + 2p_{12} + p_2 > 0 \quad . \tag{5.18}$$

Substituting the solution for the system of equations (5.16) into (5.13b) it follows that W can be written as (see details in Appendix B4):

$$W = \frac{\tilde{p}_1 + \tilde{p}_{12}}{\tilde{p}_1 + 2\tilde{p}_{12} + \tilde{p}_2} \quad , \tag{5.19}$$

with

$$\tilde{p}_1 = \frac{(1 + 3\beta^2 - 4\beta) - (\beta - 1)\sqrt{\beta^2 + 2\beta + 4\tau + 1}}{2(\beta - 1)^2} \quad , \tag{5.20a}$$

$$\tilde{p}_{12} = -\beta\tilde{p}_1 + \beta \quad , \tag{5.20b}$$

$$\tilde{p}_2 = (1 - \beta)^2\, \tilde{p}_1^2 + (4\beta - 2\beta^2 - 1)\tilde{p}_1 + \beta^2 - 2\beta \quad , \tag{5.20c}$$

where $\tau = \sigma_b^2 / \sigma_a^2$.

Having obtained an expression for W in terms of β and τ , the next step is to analyse the use of model (5.9) for handling a series of proportions regarding the production of bounded estimates. That is, is time to check whether or not $0 \leq \hat{\theta}_{t|t} \leq 1 \;\; \forall\, t$. The conditions for getting the filter estimates in (5.15) bounded between zero and one are:

$$W + W(1-W)(1-\beta)\sum_{j=1}^{t-1}[(1-W) + \beta W]^{j-1} \leq 1 \ , \tag{5.21a}$$

$$0 \leq W \leq 1 \ , \tag{5.21b}$$

$$0 \leq W(1-W)(1-\beta)[(1-W) - \beta W]^{j-1} \leq 1 \quad j=1,\ldots,t-1 \ , \tag{5.21c}$$

Equations (5.21) state that the sum of the weights in the linear combination (5.15) must be less than or equal to one, and also that each of the weights must be bounded between zero and one.

For large t, equation (5.21a) is always satisfied and, in addition, W is bounded between zero and one. However, condition (5.21c) is not satisfied for all j. Figure 5.1 displays a contour plot for the function in (5.21c) when $j=2$, which is the coefficient of y_{t-2} in the linear combination (5.17). The range for β was chosen such that $|\beta| < 1$ and $\sigma_b^2 \leq \sigma_a^2$ (implying $0 \leq \tau \leq 1$). As σ_b^2 and σ_a^2 are the variances of the disturbances terms in the noise and signal model, there is no reason to assume $\sigma_b^2 > \sigma_a^2$ but, in any case, the result introduced next is also valid for $\tau > 1$.

Figure 5.1 shows that for values of $\beta \in [-1,0]$ the weight for y_{t-2} in $\hat{\theta}_{t|t}$ can be negative. Therefore, there is no guarantee that, when fitting model (5.9) to a series of estimated proportions, the signal estimates $\hat{\theta}_{t|t}$ will belong to $[0,1]$. Note, in addition, that although the model assumed for the signal θ_t was a local level with no added error in the observation equation, the filtered estimates are not necessarily bounded. The presence of possibly negative weights in (5.15) implies that the fact that $y_t \in [0,1] \ \forall \ t$ does *not* guarantee that $\hat{\theta}_{t|t} \in [0,1]$. When the analysis was carried out with $\tau > 1$, the same conclusions were achieved. Consequently, the model in (5.11) contradicts Conjecture 5.1.

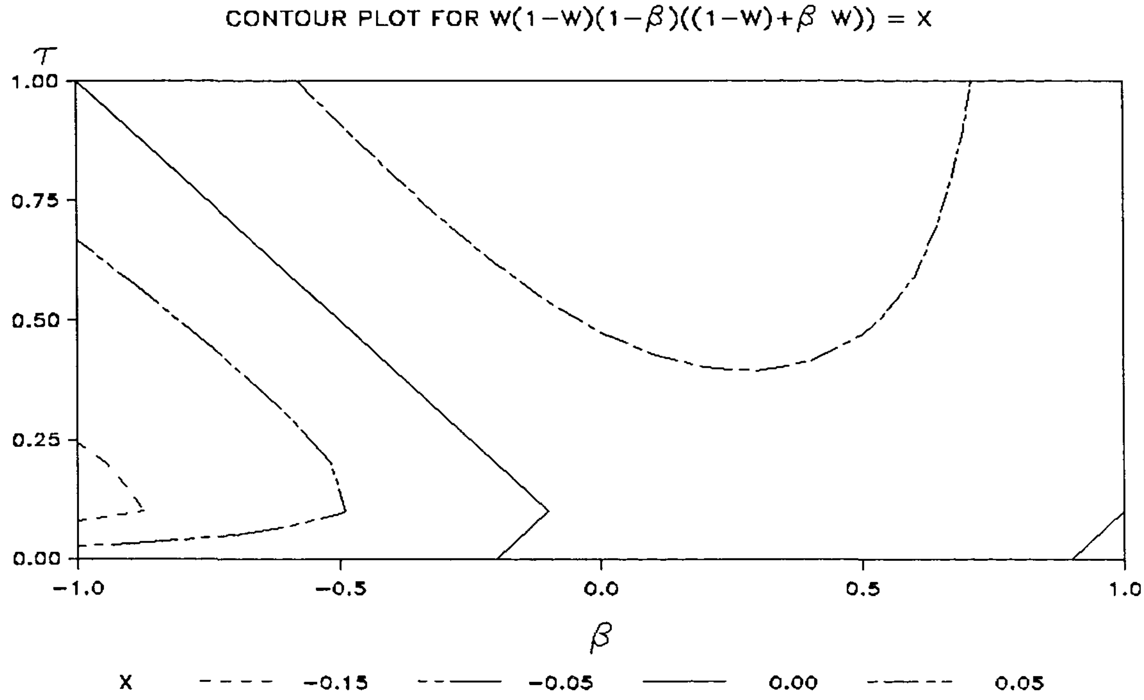


Figure 5.1

This indicates that model (5.9) may be of limited service for modelling series of proportions if the assurance of valid bounded filtered estimates is required. This can be used as an argument for the adoption of a logistic transformation, for example, prior to the use of the state-space approach for modelling proportions.

5.4 Conclusion

This chapter focused on the analysis of ordinary state-space models for series of estimated proportions. It was demonstrated, by disproving Conjecture 5.1, that it is not possible to ensure in general that signal estimates are always bounded between zero and one, if models are fitted to proportions directly without any transformation. Restricting the analysis to the case of modelling data from overlapping surveys, one procedure (see Chapter 4) is to fit a structural model for $\{ \theta_t \}$ assuming an ARMA model for the sampling errors

$\{e_t\}$. Model (5.9) is a simple but proper example of this sort of state-space model, widely used nowadays, to improve estimation in repeated surveys. The evidence provided in Section 5.3 leads to the conclusion that the direct use of the state-space approach for modelling proportions without any prior transformation in repeated surveys does not always guarantee bounded filtered estimates. Hence Conjecture 5.1 is false.

If the use of a state-space model does not itself guarantee bounded filtered estimates, it seems reasonable to require that the data to be transformed before modelling. In the case of compositional data, following the work of Brunsdon(1987), the data can be transformed using the additive logratio transformation in (5.1). The use of logit or logratio transformations of the estimates as inputs to the state-space model (or any other signal extraction approach) affects the way the sampling errors should be modelled. This issue will be discussed in Chapter 7. So far, all the analysis has been concerned with the univariate case. The multivariate/compositional case will be addressed in Chapter 6 where state-space models for handling compositional data from overlapping surveys are proposed.

6 State-Space Modelling of Compositional Data from Repeated Surveys

6.1 Introduction

This chapter considers the use of state-space models for improving estimation of compositional data in a repeated survey framework. Although compositional data were previously modelled using a state-space (structural time series) approach by Quintana & West(1988) and Shephard & Harvey(1989), these authors did not address the issue of modelling the autocovariance structure of the sampling errors when the observed compositions are obtained from repeated sample surveys.

Quintana and West(1988) employed a Bayesian approach and fitted a Dynamic Linear Model to a set of Mexican imports series, which were classified as: consumer, intermediate and capital. The data were first converted from the original scale to proportions in order to analyse the relative behaviour of the series. Then a log-ratio transformation (West & Harrison, 1989, p.641) was applied to the three series of proportions, with the transformed data as inputs when fitting a local linear model. The final analysis was presented in terms of the transformed compositions and no attempt was made to recover the trend estimates for the original series.

Shephard & Harvey(1989) fitted a local level model to the time series of shares of the vote for the three main parties at the British General Elections, from 1974 to 1987. Their approach was to fit this multivariate structural model to a vector of proportions of votes comprising all but one of the parties (or groups of parties) and to obtain the level of support of the remaining party (or group) by subtraction, since the sum of all the proportions is unity. The authors recommended this procedure, reporting that the technique of dropping one of the series is a standard practice in systems of regression equations, and that it leads to the same result as if a maximum-likelihood estimation procedure was applied to a full set of equations with constraints imposed on the covariance matrix of the disturbances. Although this procedure guarantees that the sum of the predictions equals one, it does not guarantee that the predictions for each of the series are bounded between zero and one.

Shephard & Harvey(1989) overcame this last problem by using a local level model which, as was demonstrated in Chapter 5, always produces bounded predictions when the inputs are bounded. They noted this fact and argued that the use of a local level model was appropriate because more general models could not guarantee predictions always belonging to $[0,1]$. Regarding the model disturbances, they assumed that the error of the observation equation was mainly the sampling error. They looked at the sequence of opinion polls as non-overlapping repeated surveys, and assumed that the auto-covariance of the sampling errors was zero for any lag greater than or equal to one, i.e.assumed that the observation errors were independent through time. However, the authors did not take into account that the polls are in general based on a master sample (see Smith, 1996) which implies an overlap between the primary sampling units. They defined the covariance matrix of the errors such as to represent the variance and covariance of the sampling errors on each occasion.

This review of the available literature in the area of modelling compositional data in repeated surveys reveals that little work was done to take the sampling errors into account, particularly for the case of overlapping surveys. Moreover, although suggested by Quintana & West(1988) and Harvey & Shephard(1993) that different structural models (rather than the local level model) could be fitted to compositional data, this has still not been tried yet and no other models were evaluated in practice.

The evidence provided in Chapter 5 for univariate series, together with the work of Brunsdon(1987), leads to the conclusion that the only way, in general, of guaranteeing bounded predictions and signal estimates for compositional data, when fitting either structural or a vector ARMA models, is by using a transformation such as the additive logratio transformation or one of the other transformations proposed by Aitchison(1986). However, one drawback of this approach is that, when applying structural models to the transformed data, great care is needed to define how (and which of) the unobservable components of the original series can be estimated from the model fitted to the transformed series.

In Section 6.2 two different models are proposed for compositional data from overlapping surveys which take the sampling error into account. Because the interest lies in survey data, the purpose of the modelling procedures is not only to provide predictions for the observed series but also to improve the estimation of the unobservable signal and its components. Both types of models are extensions to the work of Brunsdon(1987). The idea is to apply the additive logratio transformation to the data, which can then be modelled via either structural or vector ARMA models in a multivariate state-space formulation.

At this point two alternative routes can be considered. One way is to model both signal and noise in the transformed scale via vector ARMA models (Tiao & Box, 1981; Wei, 1993, chapter 14). The final model is the superposition of these two multivariate models. Following West & Harrison(1989, p.182) the term *superposition* is used here to refer to a construction of a complex model from different, simpler, component models. This will be called the *General Multivariate Model.* An example of this formulation is presented in Section 6.3.

Another way of modelling the transformed data is by assuming that the signal of the transformed series (looking at each of them via an univariate perspective) can be modelled by similar univariate structural models, with model parameters being different across series. In addition, the noise process can be represented by a vector ARMA process. This type of model will be named hereafter as *Common Components Model,* as in Barbosa(1989). An example of this formulation is provided in Section 6.4.

It will be shown that both procedures produce predictions and signal estimates bounded between zero and one, satisfying the unity-sum constraint. The state-space formulation for compositional data is examined in more detail in the next section where, for simplicity, it is formally defined considering a composition lying in the Simplex S^2 . This case is of particular interest because the modelling procedures proposed here will be tested using data from the Brazilian Labour Force Survey which produces monthly estimates of the employed, the unemployed and of the people who are not in the labour force. The survey's target population comprises people who are 15 years old or more. Using this information it is possible to form, at each survey round, a vector of proportions using the total population aged 15 years or more as the reference variable. Each of these vectors represents a tri-dimensional random variable which lies in the Simplex S^2 .

6.2 A Framework for Modelling Compositional Data from Overlapping Surveys

Assume that the point estimates for a set of population characteristics obtained from a repeated overlapping survey form the basis of a composition. From Aitchison(1986, p.31), "a basis w of $M+1$ parts is a $(M+1)\times 1$ vector of positive components $(w_1,\dots,w_{M+1})$ all recorded on the same measurement scale". A basis completely determines its composition y , defined as the vector $y = w \;/\; \mathbf{1}'w$. The composition y belongs to the Simplex S^M , defined as

$$S^M = \{y: 0 \le y_m \le 1 \;,\; m=1,\dots,M+1; \sum_{m=1}^{M+1} y_m = 1\} \quad ,$$

In the case of a sample survey, let y_t be a vector of estimated proportions subject to a unity-sum constraint. Data with such characteristics are observed in repeated surveys when a survey variable has a multinomial response but the interest lies in the proportion of units classified in each of its categories. At this point, having defined the concept of a basis, one can argue that an alternative approach to deal with this problem is by directly modelling the components of the basis as a multivariate time series. Note, however, that the basis is not always available to the analyst as in the case of a survey in which individuals are requested to inform the composition of their expenditures as proportions of their income. And, even if they are asked to record their expenditures over a period of time, what is generally published is the overall pattern (or composition) of expenditures. In addition, if a component of the basis is small (recall that $w_m \in [0, \infty]$, $\forall m$), fitting a time series model to that components can lead to negative predictions.

In a labour force survey, for example, the estimates of the total number of people who are unemployed $(m=1)$, employed $(m=2)$ and not in the labour force $(m=3)$ form the basis $w_t = (w_{1t}, w_{2t}, w_{3t})'$ of a composition $y_t = (y_{1t}, y_{2t}, y_{3t})'$ where $y_{mt} = \dfrac{w_{mt}}{w_{1t}+w_{2t}+w_{3t}}$ for $m=1,2,3$.

The observed compositional time series $\{y_t\}$ can then be defined as the sequence of vectors

$$y_t = (y_{1t}, y_{2t}, y_{3t})' \quad , \tag{6.1}$$

where

y_{1t} is the estimated proportion of unemployed people at time *t*;

y_{2t} is the estimated proportion of employed people at time *t*;

y_{3t} is the estimated proportion of people who are not in the labour force at time *t*. It is easy to see that y_t lies in the Simplex S^2 .

Consider now that $y_t = (y_{1t}, \dots, y_{M+1,t})'$ is a vector of sample estimates belonging to the Simplex S^M . Since each of its components is subject to sampling errors, y_{mt} can be decomposed into signal and noise components as

$$y_{mt} = \theta_{mt} + e_{mt} \quad , \quad m=1,\dots,M+1 \quad , \tag{6.2}$$

where θ_{mt} is the unknown population proportion assumed to follow a time series model, and e_{mt} is the sampling error. Considering the $M+1$ series simultaneously, (6.2) can be written in vector form as:

$$\boldsymbol{y}_t = \boldsymbol{\theta}_t + \boldsymbol{e}_t \quad , \tag{6.3}$$

where $\boldsymbol{\theta}_t = (\theta_{1t}, \dots, \theta_{M+1,t})'$ and $\boldsymbol{e}_t = (e_{1t}, \dots, e_{M+1,t})'$. In addition, it is assumed that

$$\sum_{m=1}^{M+1} \theta_{mt} = \sum_{m=1}^{M+1} y_{mt} = 1 \quad \forall\, t \quad , \tag{6.4a}$$

which implies that

$$\sum_{m=1}^{M+1} e_{mt} = 0, \qquad \forall\, t \quad . \tag{6.4b}$$

Following Bell & Hillmer(1990), the model in (6.3) can be rewritten as

$$y_{mt} = \theta_{mt} \left[1 + \frac{e_{mt}}{\theta_{mt}} \right] = \theta_{mt} u_{mt} \quad , \tag{6.5a}$$

with

$$u_{mt} = \left[1 + \frac{e_{mt}}{\theta_{mt}} \right] = (1 + \tilde{u}_{mt}) \quad , \tag{6.5b}$$

where $\tilde{u}_{mt} = \frac{e_{mt}}{\theta_{mt}}$ represents the relative sampling error of the estimated proportion.

Applying the additive logratio transformation to the vector $\boldsymbol{y}_t$ with components given in (6.5a) produces a transformed vector $\boldsymbol{v}_t = a_M(\boldsymbol{y}_t) = (v_{1t}, \dots, v_{Mt})'$ defined on $\mathbb{R}^M$. If $y_{M+1,t}$ is used as the reference variable, the transformed vector has as its m^th component:

$$\begin{aligned} v_{mt} &= \log \left[\frac{y_{mt}}{y_{M+1,t}} \right] = \log \left[\frac{\theta_{mt} u_{mt}}{\theta_{M+1,t} u_{M+1,t}} \right] \\ &= \log \left[\frac{\theta_{mt}}{\theta_{M+1,t}} \right] + (\log u_{mt} - \log u_{M+1,t}) \quad , \quad m = 1, \dots, M \quad . \end{aligned} \tag{6.6}$$

From (6.6), a vector model for the transformed series can be written as:

$$\boldsymbol{v}_t = \boldsymbol{\theta}_t^* + \boldsymbol{e}_t^* \quad , \tag{6.7}$$

with $\boldsymbol{v}_t = (v_{1t}\ ,\dots,\ v_{Mt})'$, $\boldsymbol{\theta}_t^* = (\theta_{1t}^*\ ,\dots,\ \theta_{Mt}^*)'$ and $\boldsymbol{e}_t^* = (e_{1t}^*\ ,\dots,\ e_{Mt}^*)'$, where $v_{mt}=\log(y_{mt}/y_{M+1,t})$, $\theta_{mt}^* = \log\left(\theta_{mt}\ /\ \theta_{M+1,t}\right)$ and $e_{mt}^* = \log(u_{mt}/u_{M+1,t})$, for $m = 1,\dots,M$. Note that model (6.7) has the same form as model (6.3).

Before proceeding further, it is interesting to note that a Taylor linearization of $\log(u_{mt}) = \log(1 + \tilde{u}_{mt})$ yields (details in Appendix C1):

$$\log(u_{mt}) = \tilde{u}_{mt} + \mathrm{O}_p(n_t^{-1}) \quad . \tag{6.8}$$

Assuming n_t large, yields the approximation:

$$\log(u_{mt}) \approx \tilde{u}_{mt} \quad . \tag{6.9}$$

Substituting (6.9) in (6.7) results in,

$$v_{mt} = \log\left[\frac{y_{mt}}{y_{M+1,t}}\right] \approx \log\left[\frac{\theta_{mt}}{\theta_{M+1,t}}\right] + (\tilde{u}_{mt} - \tilde{u}_{M+1,t}) \quad . \tag{6.10}$$

Then $e_{mt}^* = (\tilde{u}_{mt} - \tilde{u}_{M+1,t})$ can be interpreted as a contrast of the relative sampling errors.

To describe the survey data model (6.7) must incorporate time series models for both $\{\boldsymbol{\theta}_t^*\}$ and $\{\boldsymbol{e}^*{}_t\}$. Hence a multivariate model for the transformed data is given by:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{\theta}_t^* + \boldsymbol{e}_t^* \quad ; \\ \boldsymbol{\theta}_t^* = \boldsymbol{T}^{(\theta)}\ \boldsymbol{\theta}_{t-1}^* + \boldsymbol{G}^{(\theta)}\ \boldsymbol{\eta}_t^{(\theta)} \quad ; \\ \boldsymbol{e}_t^* = \boldsymbol{T}^{(e)}\ \boldsymbol{e}_{t-1}^* + \boldsymbol{G}^{(e)}\ \boldsymbol{\eta}_t^{(e)} \quad . \end{cases} \tag{6.11}$$

The configuration of $\boldsymbol{T}^{(\theta)}$, $\boldsymbol{G}^{(\theta)}$, $\boldsymbol{\eta}^{(\theta)}$, $\boldsymbol{T}^{(e)}$, $\boldsymbol{G}^{(e)}$ and $\boldsymbol{\eta}^{(e)}$ will depend on the form of the time series models for $\{\boldsymbol{\theta}_t^*\}$ and $\{\boldsymbol{e}^*{}_t\}$. However, regardless of which model is chosen to represent the signal and noise processes, (6.11) can be expressed via a state-space formulation as:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{H}\boldsymbol{\alpha}_t \quad ; \\ \boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}\boldsymbol{\eta}_t \quad , \end{cases} \tag{6.12}$$

where

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \boldsymbol{\theta}_t^* \\ \boldsymbol{e}_t^* \end{bmatrix} \quad ,$$

or is any appropriate set of present and past information such that the future behaviour of the system can be completely described by the knowledge of the present state. For example, $\boldsymbol{\alpha}_t$ could comprise unobservable components representing the trend and seasonals of $\{\boldsymbol{\theta}_t^*\}$ and past values of $\{\boldsymbol{e}_t^*\}$. The other matrices $\boldsymbol{H}, \boldsymbol{T}, \boldsymbol{G}$ and $\boldsymbol{\eta}_t$ have the form:

$$\boldsymbol{H} = \left[\boldsymbol{H}^{(\theta)} \;\; \boldsymbol{H}^{(e)}\right] \;\; , \quad \boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}^{(\theta)} & \vdots & \boldsymbol{0} \\ \cdots & \cdots & \cdots \\ \boldsymbol{0} & \vdots & \boldsymbol{T}^{(e)} \end{bmatrix} \quad ,$$

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}^{(\theta)} & \vdots & \boldsymbol{0} \\ \cdots & \cdots & \cdots \\ \boldsymbol{0} & \vdots & \boldsymbol{G}^{(e)} \end{bmatrix} \quad , \quad \boldsymbol{\eta}_t = \begin{bmatrix} \boldsymbol{\eta}_t^{(\theta)} \\ \boldsymbol{\eta}_t^{(e)} \end{bmatrix} \quad .$$

If the vector $\boldsymbol{y}_t = (y_{1t}, \dots, y_{M+1,t})'$ is permuted, a different version of the additive logratio transformation is obtained. Let $\boldsymbol{y}_t^{(m)}$ denote a permutation of $\boldsymbol{y}_t$ with the elements y_{mt} and $y_{M+1,t}$ interchanged. In this case, the element $y_{mt} \neq y_{M+1,t}$ is used as the reference variable. Aitchison(1986, p.93) defined the *permutation matrix* $\mathcal{P}_m$ as an identity matrix of order $(M+1)$ with the columns m and $M+1$ interchanged. If the permutation $\mathcal{P}_m$ is applied to the composition $\boldsymbol{y}_t = (y_{1t}, \dots, y_{mt}, \dots, y_{M+1,t})'$ it yields:

$$\boldsymbol{y}_t^{(m)} = \mathcal{P}_m \boldsymbol{y}_t = (y_{1t}, \dots, y_{M+1,t}, \dots, y_{mt})' \quad . \tag{6.13}$$

The effect of the permutation $\mathcal{P}_m$ on the logratio vector $\boldsymbol{v}_t$ is such that (Aitchison, 1986, pp.93-94 and Brunsdon, 1987, pp.61-62):

$$v_t^{(m)} = Z_m v_t \quad , \tag{6.14}$$

with

$$\{Z_m\}_{ij} = \begin{cases} 1 & i=j \neq m \quad i,j=1,\ldots,M \quad , \\ -1 & j=m \quad i=1,\ldots,M \quad , \\ 0 & elsewhere \ . \end{cases} \tag{6.15a}$$

The matrix Z_m satisfies

$$Z_m^{-1} = Z_m \quad , \tag{6.15b}$$

and

$$|Z_m| = \pm 1 \quad . \tag{6.15c}$$

The additive logistic transformation is a one-to-one transformation from $v_t \in \mathbb{R}^M$ to $y_t \in S^M$ (Aitchison, 1986, p.113), meaning that $a_M^{-1}(v_t^{(m)}) = y_t^{(m)}$. Recall that the general state-space model for v_t defined in (6.7) is given by (6.12). Now, let $v_t^{(m)} = Z_m v_t$ for all $t=1,\ldots,T$. Then a general state-space model for $v_t^{(m)}$ is:

$$\begin{cases} Z_m v_t = Z_m H \alpha_t \quad ; \\ \alpha_t = T\alpha_{t-1} + G\eta_t \quad , \end{cases} \tag{6.16a}$$

or

$$\begin{cases} v_t^{(m)} = (Z_m H)\, \alpha_t \quad ; \\ \alpha_t = T\alpha_{t-1} + G\eta_t \quad . \end{cases} \tag{6.16b}$$

The issue here is how the permutations affect statistical procedures for compositional data. It is important to investigate if the modelling procedure proposed in this thesis is permutation invariant. That is, if the state-space modelling procedure for compositional time series is invariant to the choice of the reference variable. Theorem 6.1 settles this question stating that any permutation of y_t (and consequently any transformed v_t) can be taken as the input for the state-space model since they all yield the same results.

Theorem 6.1

The state-space approach is permutation invariant. That is, the state-space models in (6.12) and (6.16) represent the same model for y_t on the Simplex, except that y_t has been permuted.

Proof of Theorem 6.1:

A state-space modelling procedure consists of:

(i) the formulation of a state-space model;

(ii) the use of the Kalman Filter equations for prediction, updating and smoothing;

(iii) the estimation the unknown hyperparameters via maximum-likelihood.

Regarding items *(i)* and *(ii)*, it is evident that the permutation does not affect the state-vector or the system equation in (6.16). A permutation of the observations only affects the observation or measurement equation which represents the relationship between the observations and the current state components. Hence, although the relation between the observed transformed series and the state-vector was adjusted for the permutation, the state vector remained the same.

The prediction and smoothing equations for model (6.16) are given by (for details see Chapter 3, pp.23-25):

$$\begin{aligned} \hat{\alpha}_{t|t-1} &= E(\alpha_t | D_{t-1}) = T_t \hat{\alpha}_{t-1|t-1} \quad , \\ P_{t|t-1} &= V(\alpha_t | D_{t-1}) = TP_{t-1|t-1}T' + GQG' \quad , \end{aligned} \qquad \textbf{(6.17a)}$$

and

$$\begin{aligned} \hat{v}^{(m)}_{t|t-1} &= E(v^{(m)}_t \mid D_{t-1}) = Z_m H \hat{\alpha}_{t|t-1} = Z_m \hat{v}_{t|t-1} \quad , \\ F^{(m)}_{t|t-1} &= V(v^{(m)}_t \mid D_{t-1}) = Z_m H P_{t|t-1} (Z_m H)' = Z_m F_{t|t-1} Z'_m \quad , \end{aligned} \qquad \textbf{(6.17b)}$$

implying that

$$\begin{aligned} \hat{\alpha}_{t|t} = E(\alpha_t \mid D_t) &= \hat{\alpha}_{t|t-1} + P_{t|t-1} H' Z'_m F^{-1}_{p,t|t-1}(v^{(m)}_t - \hat{v}^{(m)}_{t|t-1}) \\ &= \hat{\alpha}_{t|t-1} + P_{t|t-1} H' Z'_m (Z_m F_{t|t-1} Z'_m)^{-1} (Z_m v_t - Z_m \hat{v}_{t|t-1}) \\ &= \hat{\alpha}_{t|t-1} + P_{t|t-1} H' F^{-1}_{t|t-1} (v_t - \hat{v}_{t|t-1}) \quad , \end{aligned} \qquad \textbf{(6.17c)}$$

$$P_{t|t} = V(\alpha_t \mid D_t) = P_{t|t-1} - P_{t|t-1} H' F^{-1}_{t|t-1} H P_{t|t-1} \quad ,$$

and

$$\begin{aligned} \hat{\alpha}_{t|T} &= \hat{\alpha}_{t|t} + P_{t|t} T P_{t+1|t} (\hat{a}_{t+1|t} - T \hat{a}_{t|t}) \quad , \\ P_{t|T} &= P_{t|t} + P_{t|t} T' P^{-1}_{t+1|t} (P_{t+1|T} - P_{t+1|t}) P^{-1'}_{t+1|t} T P'_{t|t} \quad . \end{aligned} \qquad \textbf{(6.17d)}$$

It becomes clear from equations (6.17a), (6.17c) and (6.17b) that the updating, filtering and smoothing equations remain unchanged.

Regarding the one-step ahead forecast in (6.17b), note that $v^{(m)}_{t|t-1} \sim N(Z_m \hat{v}_{t|t-1}, Z_m F_{t|t-1} Z'_m)$ whereas $v_{t|t-1} \sim N(\hat{v}_{t|t-1}, F_{t|t-1})$. Recall, from Chapter 5 (p.52), that when considering v normally distributed with mean μ and covariance matrix Σ , y is said to have an *additive logistic normal distribution* in the Simplex S^M denoted by $L^M(\mu, \Sigma)$. Hence, as

$$(v^{(m)}_t \mid v^{(m)}_1, \ldots, \ldots, v^{(m)}_{t-1}) \sim N(Z_m \hat{v}_{t|t-1}, Z_m F_{t|t-1} Z'_m) \quad ,$$

it follows that

$$(y^{(m)}_t \mid v^{(m)}_1, \ldots, \ldots, v^{(m)}_{t-1}) \sim L^M(Z_m \hat{v}_{t|t-1}, Z_m F_{t|t-1} Z'_m) \quad .$$

Because the additive logratio transformation is a one-to-one transformation, $(v^{(m)}_1, \ldots, \ldots, v^{(m)}_{t-1})'$ conveys the same information as $(y^{(m)}_1, \ldots, \ldots, y^{(m)}_{t-1})'$. Thus,

$$y^{(m)}_{t|t-1} = (y^{(m)}_t \,|\, y^{(m)}_1, \ldots, y^{(m)}_{t-1}) \sim L^M(Z_m \hat{v}_{t|t-1}, Z_m F_{t|t-1} Z'_m) \quad , \tag{6.18a}$$

whereas

$$y_{t|t-1} = (y_t \,|\, y_1, \ldots, y_{t-1}) \sim L^M(\hat{v}_{t|t-1}, F_{t|t-1}) \quad . \tag{6.18b}$$

However, Aitchison and Shen(1980) showed that $L^M(Z_m \mu, Z_m \Sigma Z'_m)$ is simply a rotation of $L^M(\mu, \Sigma)$.

In conclusion, $L^M(\hat{v}_{t|t-1}, F_{t|t-1})$ and $L^M(Z_m \hat{v}_{t|t-1}, Z_m F_{t|t-1} Z'_m)$ denote the same distribution but with a permutation in the Simplex S^M . Consequently, the same one-step ahead forecasts are obtained for the original compositional vector using either v_t or $v^{(m)}_t$ as inputs for the state-space models. Hence the prediction, updating and smoothing equations are permutation invariant.

Finally, as pointed out in Chapter 3, the system matrices can depend on hyperparameters (Ω) that must be estimated. The use of the Kalman Filter enables the evaluation of the likelihood which is used to estimate any unknown parameter in the model. For a situation in which the observations $v^{(m)}_1, v^{(m)}_2, \ldots, v^{(m)}_T$ are not independently distributed, their joint distribution can be obtained using the conditional probability density functions $p(v^{(m)}_t \,|\, D_{t-1})$ as:

$$p(v^{(m)}_1, \ldots, v^{(m)}_T, \Omega) = \prod_{t=1}^{T} p(v^{(m)}_t \,|\, D_{t-1}) \quad .$$

From the predictions equations in (6.17b) it follows that the likelihood is given by:

$$\mathcal{L}(\Omega, v^{(m)}_1, \ldots, v^{(m)}_T) = \prod_{t=1}^{T} \frac{1}{(2\pi)^{M/2} |Z_m F_{t|t-1} Z'_m|^{1/2}} \times$$
$$\exp\left\{ -\frac{1}{2}\left(Z_m v_t - Z_m \hat{v}_{t|t-1})' (Z_m F_{t|t-1} Z'_m)^{-1} (Z_m v_t - Z_m \hat{v}_{t|t-1}\right)\right\} \quad . \tag{6.19}$$

Note that

$$|Z_m\, F_{t|t-1}\, Z'_m| \;=\; |Z_m|\; |F_{t|t-1}|\; |Z'_m| \;=\; |F_{t|t-1}| \quad , \tag{6.20}$$

since $|Z_m| = \pm 1$.

In addition,

$$\begin{gathered}(Z_m v_t - Z_m \hat{v}_{t|t-1})'\,(Z_m F_{t|t-1} Z'_m)^{-1}(Z_m v_t - Z_m\; \hat{v}_{t|t-1}) \;= \\ (v_t - \hat{v}_{t|t-1})'\, F^{-1}_{t|t-1}(v_t - \hat{v}_{t|t-1}) \quad .\end{gathered} \tag{6.21}$$

Then, it becomes clear that $\mathscr{L}(\Omega, v_1^{(m)}, \ldots, v_T^{(m)}) \;=\; \mathscr{L}(\Omega, v_1, \ldots, v_T)$. Hence the likelihood function is also permutation invariant. Another way of reaching this conclusion is observing that the (modulus of) the Jacobian of this transformation $v_t^{(m)} = Z_m\, v_t$ equals one. Moreover, both $v_t^{(m)}$ and v_t represent the same composition. Therefore, the state-space modelling procedure is permutation invariant. In summary, it has been shown that whichever permutation is used to construct the time series of logratios, the same inferences are obtained when returning to the original Simplex.

In order to illustrate how to get filtered and smoothed estimates for $\{\theta_t\}$ after modelling the transformed series, the next two sections introduce examples of a General Multivariate Model and a Common Components Model, respectively. Later, in Chapter 8, a Common Components Model will be used to model compositional data from the Brazilian Labour Force Survey. It is interesting to note that the modelling procedure established in this thesis is directly applicable to unconstrained multivariate series.

6.3 The General Multivariate Model

The basic idea of the General Multivariate Model approach is to fit vector ARMA models (as in Definition 3.2, Chapter 3, p.30) to both $\{\theta_t^*\}$ and $\{e_t^*\}$. A detailed discussion about the specific problem of modelling the sampling error series in a multivariate and compositional framework is presented in Chapter 7.

Meanwhile, to illustrate the use of a General Multivariate Model in a compositional framework, consider the case of a completely overlapping survey estimating proportions subject to a unity-sum constraint belonging to the Simplex S^2 . By analogy with the univariate case (discussed in Scott, Smith & Jones, 1977), assume that both the signal and noise components (of the transformed survey data) follow a first-order autoregressive model. Hence, considering that $\boldsymbol{v}_t \in \mathbb{R}^2$, it will be assumed that both $\{\boldsymbol{\theta}_t^*\}$ and $\{\boldsymbol{e}_t^*\}$, as defined in (6.7), follow a VAR(1), given by:

$$\begin{bmatrix} \theta_{1t}^* \\ \theta_{2t}^* \end{bmatrix} = \begin{bmatrix} \phi_{11}^{(\theta)} & \phi_{12}^{(\theta)} \\ \phi_{21}^{(\theta)} & \phi_{22}^{(\theta)} \end{bmatrix} \begin{bmatrix} \theta_{1,t-1}^* \\ \theta_{2,t-1}^* \end{bmatrix} + \begin{bmatrix} \eta_{1t}^{(\theta)} \\ \eta_{2t}^{(\theta)} \end{bmatrix} , \tag{6.22a}$$

$$\begin{bmatrix} e_{1t}^* \\ e_{2t}^* \end{bmatrix} = \begin{bmatrix} \phi_{11}^{(e)} & \phi_{12}^{(e)} \\ \phi_{21}^{(e)} & \phi_{22}^{(e)} \end{bmatrix} \begin{bmatrix} e_{1,t-1}^* \\ e_{2,t-1}^* \end{bmatrix} + \begin{bmatrix} \eta_{1t}^{(e)} \\ \eta_{2t}^{(e)} \end{bmatrix} . \tag{6.22b}$$

The state-space representation of (6.22a) is:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{H}\boldsymbol{\alpha}_t , \\ \boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t , \end{cases} \tag{6.23}$$

with:

$$\boldsymbol{\alpha}_t = (\theta_{1t}^* , \theta_{2t}^* , e_{1t}^* , e_{2t}^*)' ,$$

$$\boldsymbol{\alpha}_{t-1} = (\theta_{1,t-1}^* , \theta_{1,t-1}^* , e_{1,t-1}^* , e_{2,t-1}^*)' ,$$

$$\boldsymbol{\eta}_t = (\eta_{1t}^{(\theta)} , \eta_{2t}^{(\theta)} , \eta_{1t}^{(e)} , \eta_{2t}^{(e)})' ,$$

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} ,$$

$$T = \left[\begin{array}{ccc:ccc} \phi_{11}^{(\theta)} & \phi_{12}^{(\theta)} & & & & \\ \phi_{21}^{(\theta)} & \phi_{22}^{(\theta)} & & & \mathbf{0} & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \mathbf{0} & & \phi_{11}^{(e)} & & \phi_{12}^{(e)} \\ & & & \phi_{21}^{(e)} & & \phi_{22}^{(e)} \end{array}\right] ,$$

where $\eta_t^{(\theta)}$ and $\eta_t^{(e)}$ are bivariate normal vectors, with mean zero and covariance matrices Σ_θ and Σ_e , respectively. Moreover, $\eta_t^{(\theta)}$ and $\eta_t^{(e)}$ are assumed to be mutually uncorrelated processes with no serial correlation.

Once the model has been expressed in a state-space formulation, the Kalman Filter equations can be used to provide filtered and smoothed estimates for α_t and predictions for $\boldsymbol{v}_t$. From equations (3.5) it follows that

$$\hat{\boldsymbol{v}}_{t|t-1} = \boldsymbol{H}\,\boldsymbol{T}\,\hat{\alpha}_{t-1|t-1} \quad , \tag{6.24a}$$

$$\hat{\alpha}_{t|t} = \boldsymbol{H}\,\hat{\alpha}_{t-1|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\,(\boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}')^{-1}(\boldsymbol{v}_t - \hat{\boldsymbol{v}}_{t|t-1}) \quad , \tag{6.24b}$$

$$\hat{\boldsymbol{\alpha}}_{t|T} = \hat{\boldsymbol{\alpha}}_{t|t} + \boldsymbol{P}_{t|t}\boldsymbol{T}'\boldsymbol{P}_{t+1|t}(\hat{\boldsymbol{a}}_{t+1|t} - \boldsymbol{T}\hat{\boldsymbol{a}}_{t|t}) \ . \tag{6.24c}$$

Hence the filtered estimates $\hat{\theta}^*_{t|t}$ for the transformed signal θ^*_t can be obtained from $\hat{\alpha}_{t|t}$ by using:

$$\hat{\theta}^*_{t|t} = [\,1\,1\,0\,0\,]\,\hat{\alpha}_{t|t} \quad . \tag{6.25a}$$

Similarly, smoothed estimates $\hat{\theta}^*_{t|T}$ for the transformed signal θ^*_t are computed via

$$\hat{\theta}^*_{t|T} = [\,1\,1\,0\,0\,]\,\hat{\alpha}_{t|T} \quad . \tag{6.25b}$$

Moreover, smoothed estimates $\hat{\theta}_{t|T}$ for the original signal θ_t can be retrieved from (6.25b) by applying the additive logistic transformation a_2^{-1} , as in equation (5.2), namely

From (6.26a) and (6.26b) it becomes clear that the use of the additive logistic transformation guarantees that $\sum_{m=1}^{3} \hat{\theta}_{m,t|T} = 1$, and $0 \le \hat{\theta}_{m,t|T} \le 1$, $m=1,2,3$,

$$\hat{\theta}_{m,t|T} = \frac{\exp(\hat{\theta}^{*}_{m,t|T})}{1 + \sum_{k=1}^{2} \exp(\hat{\theta}^{*}_{k,t|T})} \qquad m=1,2 \quad , \tag{6.26a}$$

$$\hat{\theta}_{3,t|T} = \frac{1}{1 + \sum_{k=1}^{2} \exp(\hat{\theta}^{*}_{k,t|T})} \quad . \tag{6.26b}$$

leading to bounded signal estimates that sum to one across the series at every time point t . The one-step ahead forecasts for the original series can also be obtained by applying the additive logistic transformation to $\hat{v}_{t|t-1}$. Thus

$$\hat{y}_{m,t|t-1} = \frac{\exp(\hat{v}_{m,t|t-1})}{1 + \sum_{k=1}^{2} \exp(\hat{v}_{k,t|t-1})} \qquad m=1,2 \quad , \tag{6.27a}$$

$$\hat{y}_{3,t|t-1} = \frac{1}{1 + \sum_{k=1}^{2} \exp(\hat{v}_{k,t|t-1})} \quad . \tag{6.27b}$$

An important point to note regarding equations (6.26) and (6.27) is that neither $\hat{\theta}_{t|T}$ nor $\hat{y}_{t|t-1}$ are minimum mean square estimators (MMSE). Recall from Results 3.1 and 3.2 that, for example, the MMSE for y_t in terms of $(y_1,\ldots,y_{t-1})$ is given by

$E[y_t \,|\, y_1,\ldots,y_{t-1}] = E[a_2^{-1}(v_t) \,|\, a_2^{-1}(v_1),\ldots,a_2^{-1}(v_{t-1})]$,

which differs from $\hat{y}_{t|t-1} = a_2^{-1}(E[v_t \,|\, v_1,\ldots,v_{t-1}])$. As mentioned before, $(y_t \,|\, y_1,\ldots,y_{t-1}) \sim L^2(\hat{v}_{t|t-1}, F_{t|t-1})$. Hence the MMSE for y_t in terms of $(y_1,\ldots,y_{t-1})$ is the mean of an additive logistic normal distribution with parameters $\hat{v}_{t|t-1}$ and $F_{t|t-1}$. The same occurs when estimating the signal in the original scale. A MMSE for θ_t is given by

$E[\theta_t \,|\, y_1,\ldots,y_T] = E[a_2^{-1}(\theta_t^{*}) \,|\, a_2^{-1}(v_1),\ldots,a_2^{-1}(v_T)]$,

whereas $\hat{\theta}_{t|T} = a_2^{-1}(E[\theta_t^{*} \,|\, v_1,\ldots,v_T])$. The MMSE for θ_t in terms of $(y_1,\ldots,y_T)$ is given by the mean of an additive logistic normal distribution with parameters $W\hat{\alpha}_{T|t}$ and

$WP_{t|T}W'$, where W is the a row vector as in equation 6.25.

Unfortunately closed forms for the moments of the additive logistic normal distribution are not available. As pointed out by Aitchison(1986, p.116), numerical approximations for the moments have to be computed by Hermitian integration. Brunsdon (1987, Chapter 5) investigated the problem of estimating the mean and variance of an additive logistic normal distribution and also considered other location parameters such as the mode. She derived numerical approximations for the mean and mode and compared the results with the so-called *naive* estimates (those obtained by applying the additive logistic transformation). Her findings suggest that when $F_{t|t-1}$ is small, such that the distribution is dense around one area, it is possible to use the inverse transformation as an approximation to the mean. She used the mean, mode and inverse transformation when predicting the proportion of votes for the three major parties in the British general elections via vector ARMA models and found that "the various predictors were almost indistinguishable" since all of the data points for the series were packed away from the extremes (Brunsdon, 1987, p.178). That is, there were no compositions y_t such that $y_{mt} = 1$, $y_{jt} = 0$ for all $j \neq m = 1,...,M+1$.

Following Brunsdon(1987) the additive logistic transformation will be used in this thesis to get filtered and smoothed estimates for the signal in the original scale as well as predictions for the survey estimates.

In summary, the use of a General Multivariate Model enables the analyst to model compositional time series data taking into account the sampling errors in repeated overlapping surveys. However no estimates of the structural components of the series such as trend and seasonals can be obtained within the vector ARMA formulation. The Common Components Model which can encompass structural time series models is considered in the next section as an alternative approach that might solve this problem.

6.4 The Common Components Model

For simplicity, consider the case of a completely overlapping survey that produces quarterly estimates of a composition which lies in the Simplex S^2 . In the Common Components Multivariate formulation, the models for each of the univariate signal process θ^*_{mt} must have the same form. Assume that each of θ^*_{mt} follows a basic structural model as follows:

$$\begin{cases} \theta^*_{mt} = L^*_{mt} + S^*_{mt} \quad , \qquad m=1,2 \ , \\ L^*_{mt} = L^*_{m,t-1} + R^*_{m,t-1} + \eta^{(l)}_{mt} \quad , \\ R^*_{mt} = R^*_{m,t-1} + \eta^{(r)}_{mt} \quad , \\ S^*_{mt} = -\sum_{j=1}^{3} S^*_{m,t-j} + \eta^{(s)}_{it} \quad , \end{cases} \tag{6.28}$$

where $\theta^*_{mt} = \log(\theta_{mt}/\theta_{3t})$, $e^*_{mt} = \log(u_{mt}/u_{3t})$, L^*_{mt} is the trend or level of the unobservable transformed signal θ^*_{mt} , R^*_{mt} is the corresponding change in the level, S^*_{mt} is the seasonal component of θ^*_{mt} . The disturbances $\eta^{(l)}_{mt}$, $\eta^{(r)}_{mt}$, $\eta^{(s)}_{mt}$ are assumed to be mutually uncorrelated normally distributed with mean zero and variances σ^2_{ml} , σ^2_{mr} , σ^2_{ms} , respectively. Note that an alternative model to the signal is one which decomposes θ^*_{mt} into trend, seasonal and irregular components. Here, no irregular term was included in the signal model (6.28) since it is assumed that, when modelling survey data, its variation is mainly the sampling variation which is in turn properly accounted for by the model for the noise process. In a practical situation one can fit a model with an added irregular term to check the validity of this assumption. Later, in Chapter 8, a model including an irregular component for the signal is also fitted to the data.

The multivariate model for $\{\theta_t^*\}$ has the following state-space formulation:

$$\left\{\begin{array}{lcl} \theta_t^* & = & H^{(\theta)}\,\alpha_t^{(\theta)} \quad ; \\ \alpha_t^{(\theta)} & = & T^{(\theta)}\,\alpha_{t-1}^{(\theta)} + G^{(\theta)}\,\eta_t^{(\theta)} \quad , \end{array}\right. \tag{6.29}$$

where

$H^{(\theta)} = [10100]\otimes I_2$, $\alpha_t^{(\theta)} = [\, L_{1t}^* \; L_{2t}^* \; R_{1t}^* \; R_{2t}^* \; S_{1t}^* \; S_{2t}^* \; \ldots \; S_{1,t-2}^* \; S_{2,t-2}^* \,]'$,

$\eta_t^{(\theta)} = (\, \eta_{1t}^{(l)} \; \eta_{2t}^{(l)} \; \eta_{1t}^{(r)} \; \eta_{2t}^{(r)} \; \eta_{1t}^{(s)} \; \eta_{2t}^{(s)} \,)'$,

$$G^{(\theta)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \otimes I_2$$

and

$$T^{(\theta)} = \left[\begin{array}{cc|ccc} 1 & 1 & & & \\ 0 & 1 & & \mathbf{0}_{2\times 3} & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & -1 & -1 & -1 \\ \mathbf{0}_{3\times 2} & & 1 & 0 & 0 \\ & & 0 & 1 & 0 \end{array}\right] \otimes I_2 \; .$$

The model equations in (6.29) are supplemented by the cross sectional assumption :

$$\Sigma_\theta = \begin{bmatrix} \Sigma_l & \mathbf{0} & \\ & \Sigma_r & \\ \mathbf{0} & & \Sigma_s \end{bmatrix} \quad ,$$

that is, the two series are linked via the off-diagonal elements of $\Sigma_l, \Sigma_r, \Sigma_s$.

Regarding the model for the sampling errors, assume that the (multivariate) noise process $\{e_t^*\}$ can be represented by a vector autoregressive model of first order as in (Section 6.3, 6.22b). Now, putting (6.29) and (6.22b) together, a common components model for the transformed survey estimates $\boldsymbol{v}_t$ is given by:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{H}\boldsymbol{\alpha}_t \quad , \\ \boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}\boldsymbol{\eta}_t \quad , \end{cases} \tag{6.30}$$

with

$$\boldsymbol{\alpha}_t = (\boldsymbol{L}_t^{*'}, \boldsymbol{R}_t^{*'}, \boldsymbol{S}_t^{*'}, \boldsymbol{S}_{t-1}^{*'}, \boldsymbol{S}_{t-2}^{*'}, \boldsymbol{e}_t^{*'})' = (L_{1t}^*, L_{2t}^*, \dots, S_{1,t-2}^*, S_{2,t-2}^*, e_{1t}^*, e_{2t}^*)' \quad ,$$

$$\boldsymbol{\alpha}_{t-1} = (\boldsymbol{L}_{t-1}^{*'}, \boldsymbol{R}_{t-1}^{*'}, \boldsymbol{S}_{t-1}^{*'}, \boldsymbol{S}_{t-2}^{*'}, \boldsymbol{S}_{t-3}^{*'}, \boldsymbol{e}_{t-1}^{*'})' \quad ,$$

$$\boldsymbol{\eta}_t = (\eta_{1t}^{(l)}, \eta_{2t}^{(l)}, \eta_{1t}^{(r)}, \dots, \eta_{2t}^{(e^*)})' \quad ,$$

$$\boldsymbol{H} = [\boldsymbol{H}^{(\theta)}, 1] \otimes \boldsymbol{I}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad ,$$

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}^{(\theta)} \otimes \boldsymbol{I}_2 & \vdots & \boldsymbol{0} & \\ \cdots\ \cdots\ \cdots & \vdots & \cdots & \cdots \\ & \vdots & \phi_{11} & \phi_{12} \\ \boldsymbol{0} & \vdots & \phi_{21} & \phi_{22} \end{bmatrix} \quad ,$$

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}^{(\theta)} & \vdots & \boldsymbol{0} \\ \cdots\ \cdots & \cdots & \cdots \\ \boldsymbol{0} & \vdots & \boldsymbol{I}_2 \end{bmatrix}$$

and

$$V(\boldsymbol{\eta}_t) = \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{\Sigma}_\theta & \vdots & \boldsymbol{0} \\ \cdots & \cdots & \cdots \\ \boldsymbol{0} & \vdots & \boldsymbol{\Sigma}_e \end{bmatrix} \quad .$$

The smoothed estimates $\hat{\alpha}_{t|T}$ for α_t in model (6.30) can be computed using the Kalman Filter equation in (3.12) and $\hat{\theta}^*_{t|T}$ can be obtained from $\hat{\alpha}_{t|T}$ using:

$$\hat{\theta}^*_{t|T} = [1 \;0\; 0\; 0\; 1\; 0 \;...\; 0]\,\hat{\alpha}_{t|T} = W\,\hat{\alpha}_{t|T} \quad , \tag{6.31a}$$

with covariance matrix given by

$$V(\hat{\theta}^*_{t|T}) = W\,P_{t|T}\,W' \quad , \tag{6.31b}$$

where $P_{t|T}$ is the covariance matrix of the smoothed state-vector $\hat{\alpha}_{t|T}$ .Note that (6.31) yields the signal extraction estimates for $\theta^*_{mt} = \log(\theta_{mt}\,/\,\theta_{3t})$, $m = 1, 2$. Estimates for L^*_t and S^*_t can also be obtained from $\hat{\alpha}_{t|T}$ simply by pre-multiplying the state-vector estimate by a suitable row vector as in (6.31a) to extract the component estimates. In addition

$$\hat{v}_{t+1|t} = H\,\hat{\alpha}_{t+1|t} = H\,T\,\hat{\alpha}_{t|t} \tag{6.32}$$

is the one-step ahead prediction for the transformed series.

As in Section 6.3, estimates for θ_{mt} and predictions for y_{mt} are computed by applying the a_2^{-1} transformation to $\hat{\theta}^*_{t|T}$ and $\hat{v}_{t+1|t}$. Once again the use of the additive logistic transformation guarantees that:

(i) $\sum_{m=1}^{3} \hat{y}_{m,t+1|t} = 1$ with $0 \leq \hat{y}_{m,t+1|t} \leq 1$;

(ii) $\sum_{m=1}^{3} \hat{\theta}_{m,t|t} = 1$ with $0 \leq \hat{\theta}_{m,t|t} \leq 1$.

Unfortunately, although $L^*_{t|T}$ and $S^*_{t|T}$ can be estimated from model (6.30), it is not straightforward to obtain estimates for the structural unobservable components of the original signal θ_t , such as $L_{t|T}$ and $S_{t|T}$. However, if a multiplicative model with no irregular component is assumed for $\{\theta_{mt}\}$, such that:

$$\begin{aligned} \theta_{1t} &= L_{1t}\, S_{1t} \quad , \\ \theta_{2t} &= L_{2t}\, S_{2t} \quad , \\ \theta_{3t} &= L_{3t}\, S_{3t} \quad , \end{aligned} \tag{6.33}$$

where L_{mt} and S_{mt} for $m=1,2,3$, represent the trend and seasonal components of each of the unobservable signal, then applying an additive logratio transformation to $\boldsymbol{\theta}_t$ results in:

$$\log(\theta_{mt}\ /\theta_{3t}) = \frac{\log(L_{mt}\ S_{mt}\ )}{\log(L_{3t}\ S_{3t}\ )} = \log\left[\frac{L_{mt}}{L_{3t}}\right] + \log\left[\frac{S_{mt}}{S_{3t}}\right] \ ,\ m=1,2\ . \tag{6.34}$$

But (6.34) can be rewritten as

$$\theta^*_{mt} = L^*_{mt} + S^*_{mt} \quad , \tag{6.35}$$

with $L^*_{mt} = \log(L_{mt}\ /L_{3t})$ and $S^*_{mt} = \log(S_{mt}\ /S_{3t})$ which can be estimated from model (6.30).

Hence the use of a basic structural model for $\{\boldsymbol{\theta}^*_t\}$ corresponds to the case in which the underlying model for $\{\boldsymbol{\theta}_t\}$ decomposes the original signal into its trend and seasonal components in a multiplicative way. Therefore, it is assumed hereafter that the relation between the original signal and its components is given by the equations in (6.33).

To derive estimates, either filtered or smoothed, for L_{mt} and S_{mt} note that:

$$\begin{cases} \exp(S^*_{1t}\ ) = S_{1t}\ /\ S_{3t} \quad , \\ \exp(S^*_{2t}\ ) = S_{2t}\ /\ S_{3t} \quad , \end{cases} \tag{6.36a}$$

and also that

$$\begin{cases} \exp(L_{1t}^{*}) = L_{1t} \,/\, L_{3t} \quad , \\ \exp(L_{2t}^{*}) = L_{2t} \,/\, L_{3t} \quad . \end{cases} \tag{6.36b}$$

To recover S_{1t}, S_{2t}, S_{3t} , in (6.36a), it is necessary to assume an explicit relationship between these unobservable components based on model (6.33). By doing this, a third equation can be added to each of the systems in (6.36) and an estimate of the original series components can be obtained. Note that the systems have three unknowns for just two equations. If, for example, it could be assumed that $\sum_{i=1}^{3} S_{it} = 1$ then S_{it} in (6.36a) would be obtained by using an expression equivalent to the additive logistic transformation. But this assumption does not appear natural when considering the multiplicative model in (6.33) which implies that

$$L_{1t}S_{1t} + L_{2t}S_{2t} + L_{3t}S_{3t} = 1 \quad ,$$

since, in a compositional framework, $\sum_{m=1}^{3} \theta_{mt} = 1$.

The same analysis is valid for the levels/trends L_{mt} . Although, in this case, it is quite natural to assume that they do sum to one across the series, being also bounded between zero and one. Hence, trend estimates for the original series can be obtained solving the system

$$\begin{cases} \exp(L_{1t}^{*}) = L_{1t} \,/\, L_{3t} \quad , \\ \exp(L_{2t}^{*}) = L_{2t} \,/\, L_{3t} \quad , \\ L_{1t} + L_{2t} + L_{3t} = 1 \quad , \end{cases} \tag{6.37a}$$

which results in

$$L_{mt} = \frac{\exp(L_{mt}^{*})}{1 + \sum_{k=1}^{2} \exp(L_{kt}^{*})} \qquad m = 1,2 \quad , \tag{6.37b}$$

$$L_{3t} = \frac{1}{1 + \sum_{k=1}^{2} \exp(L_{kt}^{*})} \quad . \tag{6.37c}$$

As there is no irregular component in model (6.33), and consequently in (6.35), the seasonally adjusted figures are given by the trend estimates in (6.37). Therefore, the smoothed estimates for the trend of the original series of proportions are obtained by applying the additive logistic transformation a_2^{-1} to $L_{t|T}^{*}$. Consequently, estimates for the seasonal components of the original proportions can be computed by taking

$$\hat{S}_{m,t|T} = \frac{\hat{\theta}_{m,t|T}}{\hat{L}_{m,t|T}} \quad , \quad m = 1, \dots, 3 \quad . \tag{6.38}$$

Alternatively, without assuming any relation between the trend or seasonals across the series what can be obtained from model (6.30) are estimates for the trend and seasonals of the series comprising ratios of the original proportions.

From (6.36), it follows that:

$$\frac{\exp(S_{2t}^{*})}{\exp(S_{1t}^{*})} = \frac{S_{2t}}{S_{1t}} \quad , \tag{6.39a}$$

$$\frac{\exp(L_{2t}^{*})}{\exp(L_{1t}^{*})} = \frac{L_{2t}}{L_{1t}} \quad . \tag{6.39b}$$

And, from the multiplicative model (6.33), one gets:

$$\frac{\theta_{2t}}{\theta_{1t}} = \frac{L_{2t}}{L_{1t}} \frac{S_{2t}}{S_{1t}} \quad . \tag{6.40}$$

For labour force surveys, an important issue is to estimate the unemployment rate series and also to produce seasonally adjusted figures. Recall from equations (6.1) and (6.2) that θ_{1t} and θ_{2t} represent the unknown population proportions of unemployed and employed people, respectively. Using these proportions, the unknown unemployment rate at time t is defined as

$$\gamma_t = \frac{\theta_{1t}}{\theta_{1t} + \theta_{2t}} = \frac{1}{(1 + \frac{\theta_{2t}}{\theta_{1t}})} = (\frac{\theta_{2t}}{\theta_{1t}} + 1)^{-1} \quad . \tag{6.41}$$

From Result 3.1, a MMSE for $\frac{\theta_{2t}}{\theta_{1t}}$ is given by $E\left[\frac{\theta_{2t}}{\theta_{1t}} \mid y_1, \ldots, y_T\right]$. For convenience of notation, drop the dependence of the moments on the available information. That, is denote the MMSE for $\frac{\theta_{2t}}{\theta_{1t}}$ simply as $E\left[\frac{\theta_{2t}}{\theta_{1t}} \mid D\right]$ where D represents the necessary information to produce either filtered or smoothed estimates. In order to get $E\left[\frac{\theta_{2t}}{\theta_{1t}} \mid D\right]$ note that:

$$\frac{\theta_{2t}}{\theta_{1t}} = \frac{\exp(\theta_{2t}^*)}{\exp(\theta_{1t}^*)} = \exp(\theta_{2t}^* - \theta_{1t}^*) \quad . \tag{6.42}$$

From the standard log-normal theory it folllows that (see, for example, Encyclopedia of Statistical Science,1985,vol.5, pp.134-136):

$$E\left[\frac{\theta_{2t}}{\theta_{1t}} | D\right] = \exp(E[\theta_{2t}^* - \theta_{1t}^* \mid D] + \frac{1}{2} V[\theta_{2t}^* - \theta_{1t}^* | D]) \quad , \tag{6.43a}$$

where

$$E[\theta_{2t}^* - \theta_{1t}^* \mid D] = E[\theta_{2t}^* \mid D] - E[\theta_{1t}^* \mid D] \quad , \tag{6.43b}$$

and

$$V[\theta_{2t}^* - \theta_{1t}^* \mid D] = V[\theta_{2t}^* \mid D] + V[\theta_{1t}^* | D] - 2COV[\theta_{2t}^*, \theta_{1t}^* \mid D] \quad . \tag{6.43c}$$

are obtained from the state smoothed (or filtered) estimates in (6.31) and respective covariance matrices. In addition,

$$V\left[\frac{\theta_{2t}}{\theta_{1t}}|D\right] = \{\exp(E[\theta_{2t}^{*} - \theta_{1t}^{*}\,|D]) - 1\} \quad \times \qquad \textbf{(6.44)}$$
$$\{\exp(2E[\theta_{2t}^{*} - \theta_{1t}^{*}\,|D) + V[\theta_{2t}^{*} - \theta_{1t}^{*}\,|D])\} \quad .$$

Employing an approximation for the first and second moments of functions of random variables based on Taylor expansion (details in Appendix C2) one gets the following approximation for the MMSE of γ_t :

$$E[\gamma_t\,|D] \approx \left\{E\left[\frac{\theta_{2t}}{\theta_{1t}}|D\right] + 1\right\}^{-1} + \frac{V\left[\frac{\theta_{2t}}{\theta_{1t}}|D\right]}{\left\{E\left[\frac{\theta_{2t}}{\theta_{1t}}\right] + 1\right\}^{3}} \quad , \qquad \textbf{(6.45a)}$$

with

$$V[\gamma_t\,|D] \approx \left\{E\left[\frac{\theta_{2t}}{\theta_{1t}}|D\right] + 1\right\}^{-4} V\left[\frac{\theta_{2t}}{\theta_{1t}}|D\right] \quad , \qquad \textbf{(6.45b)}$$

where $E\left[\frac{\theta_{2t}}{\theta_{1t}}|D\right]$ and $V\left[\frac{\theta_{2t}}{\theta_{1t}}\,|D\right]$ are obtained using expressions (6.43) and (6.44).

The expressions (6.45a) and (6.45b) will be employed in Chapter 8 to compute model dependent estimates for the unemployment rate series and corresponding standard errors. When analysing an unemployment rate series the analyst usually requires estimates for the underlying trend and/or seasonally adjusted figures. Based on model (6.30) seasonally adjusted figures can be obtained as:

$$\gamma_t^{sa} = \frac{L_{1t}}{L_{1t} + L_{2t}} = \frac{1}{(1 + \frac{L_{2t}}{L_{1t}})} = (\frac{L_{2t}}{L_{1t}} + 1)^{-1} \quad . \qquad \textbf{(6.46)}$$

Then the seasonally adjusted unemployment rate series can be estimated by using $E[\gamma_t^{sa} \mid D]$ which is computed following the same steps as described in (6.43) to (6.45).

One interesting point to note is that, because $E\left[\frac{\theta_{2t}}{\theta_{1t}} \mid D\right] \geq 0$, it follows that

$$\left\{ E\left[\frac{\theta_{2t}}{\theta_{1t}} \mid D\right] + 1 \right\}^{-1} \qquad \textbf{(6.47)}$$

is always bounded between zero and one. However, it is not clear whether or not the same is true regarding the expression in (6.45a). Note that, although (6.45a) is a better approximation than (6.47) for the MMSE of γ_t , the second term in (6.45) is usually negligible. Hence the choice of the estimator for the unemployment rate based on model (6.30) depends on whether or not the series evolve sufficiently close to boundaries, in which case the use of the estimator in (6.47) could be recommended.

In conclusion, a Common Components Model provides signal (and trend) estimates bounded between zero and one in accordance with the unity-sum constraint. Moreover, model (6.30), provides estimates for seasonal and trend components of series comprising ratios of the original proportions which is a quite useful feature of the proposed modelling procedure.

6.5 Summary

This chapter has introduced a method for modelling compositional time series from repeated surveys taking into account the sampling errors. It has been shown that the state-space modelling procedure is permutation invariant and that both the General Multivariate Model and the Common Components Model provide predictions and signal estimates which belong to the Simplex. In addition, the use of a Common Components Model enables the analyst to get trend and seasonally adjusted estimates for the original proportions and for series comprising ratios of the original proportions.

The application of the proposed state-space approach to survey data requires the estimation of the correlation structure of the sampling errors . When fitting a Common Components Model to the Brazilian Labour Force Survey data (or any other survey data), it will be necessary to estimate the autocovariance structure of $\{e_t^* \}$ in order to formulate an appropriate time series model for this "contrast" of sampling errors.

Chapter 7 outlines important issues, regarding the identification of time series models for the sampling error process that need to be addressed in order to complete the specification and implementation of the models proposed in this chapter.

7 Time Series Models for the Sampling Error Process

7.1 Introduction

Use of the State-Space approach for improving estimation in repeated surveys provides great flexibility in the specification of the time series models for both the signal process $\{\theta_t\}$ and the noise process $\{e_t\}$. This chapter deals with modelling of the sampling error process $\{e_t\}$. It focuses on the identification procedures for the sampling error model and discusses the links between the model formulation and the sampling design.

The model specification depends on the sampling design, particularly on the level of sample overlap between occasions, and also on the availability of data. In a panel survey, if the individual records y_{ti} are available on each occasion and can be linked throughout the survey period, a full *primary analysis* can be carried out. On the other hand, if only the published aggregate estimates y_t are available, then only a *secondary analysis* can be carried out. In rotating panel surveys, the analyst can obtain the elementary panel (or rotation group) estimates $y_t^{(k)}$ (as in Chapter 2, section 2.2). This case will be named hereafter as an *elementary analysis*, although some authors do not differentiate an elementary analysis from a primary analysis.

Many authors considered the problem of modelling the sampling error process, in a univariate framework, see for example, Pfeffermann(1989,1991), Binder & Dick(1990), Bell & Hillmer(1990), Tiller(1992), Pfeffermann & Bleuer(1993), Binder, Bleuer and Dick(1993) and Pfeffermann, Bell & Signorelli(1996). There are two general approaches for identifying the model for the sampling errors. One, which can be called the *quantitative* method of analysis, refers to procedures in which a direct estimate of the sampling error autocorrelation function is obtained from the survey data. The other, hereafter called the *qualitative* method, is usually adopted when the analyst is unable to obtain direct estimates of the sampling error autocorrelation function. In this case, assumptions are made regarding the pattern of such covariances (or correlations) according to the survey sampling design.

Models for univariate time series of sampling errors are reviewed first according to both qualitative and quantitative approaches and different cases of data availability. Then the multivariate case is considered, and the multivariate framework is adapted to model the sampling error process when dealing with compositional data from repeated surveys.

7.2 The Univariate Case

Recall from (2.4) the following decomposition for the survey-based estimate y_t :

$$y_t = \theta_t + e_t \quad ,$$

where θ_t and e_t represent the finite population parameter of interest at time t and the sampling error, respectively. Assume that y_t is a design-unbiased estimator for θ_t , i.e. $E(e_t \,|\, \theta_t) = 0$, and let $V(e_t \,|\, \theta_t) = S_t^2$ represent the design variance of y_t . With the time series approach $\{y_t\}$, $\{\theta_t\}$ and $\{e_t\}$ are treated as random quantities each, to be modelled by a time series process.

7.2.1 Quantitative Analysis via the Design-Based Approach

When data are available for individual sampling units and these can be identified throughout the duration of the survey (that is, when it is possible to link survey microdata over time), $\gamma_{e|\theta}(h) = COV(e_{t-h}, e_t \,|\, \theta_{t-h}, \theta_t)$ can be estimated via $\gamma_{y|\theta}(h)$ using design based methods because

$$\gamma_{e|\theta}(h) = COV(y_{t-h} - \theta_{t-h}, y_t - \theta_t \,|\, \theta_{t-h}, \theta_t) = COV(y_{t-h}, y_t \,|\, \theta_{t-h}, \theta_t) = \gamma_{y|\theta}(h).$$

Assuming $\{e_t\}$ stationary, $\gamma_{e|\theta}(h)$ depends on h but not on t . Therefore $\gamma_{e|\theta}(h)$ can be estimated by

$$\hat{\gamma}_{e|\theta}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} C\hat{O}V(y_{t-h}, y_t \,|\, \theta_{t-h}, \theta_t) \quad .$$

Using estimates of the sampling error autocovariances and variances, an estimated autocorrelation function can be obtained and the model for the sampling error identified.

Before proceeding further it is interesting to address the problem of modelling the noise process from a log-transformed series of survey estimates. Consider the case when a logarithmic transformation is to be applied to the survey estimates (as in Binder, Bleuer & Dick,1993). The model can be written as:

$$y_t = \theta_t + e_t = \theta_t(1 + \frac{e_t}{\theta_t}) \quad . \tag{7.1}$$

Applying the logarithmic transformation to both sides of equation (7.1) results in

$$\log(y_t) = \log(\theta_t) + \log(1 + \frac{e_t}{\theta_t}) \approx \log(\theta_t) + \frac{e_t}{\theta_t} \quad . \tag{7.2}$$

Hence, when modelling the log-transformed series of survey estimates, the noise component is approximately represented by the relative sampling errors of the original series of estimates. In addition, from (7.1) it follows that:

$$(i) \quad V\left[\frac{e_t}{\theta_t} \,|\, \theta_t\right] = \frac{V(e_t \,|\, \theta_t)}{\theta_t^2} = \frac{V(y_t \,|\, \theta_t)}{\theta_t^2} \; ;$$

$$(ii) \quad COV\left[\frac{e_{t-h}}{\theta_{t-h}} \, , \, \frac{e_t}{\theta_t} \,|\, \theta_{t-h}, \theta_t\right] = \frac{COV(y_{t-h} y_t \,|\, \theta_{t-h}, \theta_t)}{\theta_{t-h}\, \theta_t} \quad .$$

Putting *(i)* and *(ii)* together yields:

$$CORR\left[\frac{e_{t-h}}{\theta_{t-h}} \, , \, \frac{e_t}{\theta_t} \,|\, \theta_{t-h}, \theta_t\right] = CORR(y_{t-h} \, , \, y_t \,|\, \theta_{t-h}, \theta_t) \quad . \tag{7.3}$$

Therefore the model for the relative sampling errors can be identified from the autocorrelation function of the original series of estimates.

When working with rotating panel surveys, instead of modelling the aggregate sampling error, the analyst has the choice of modelling independently the series for each rotation group. Let K be the number of rotation groups investigated on each survey occasion and $y_t^{(k)}$, $k=1,\ldots,K$ denote the elementary estimates at time t. The procedure described above could be applied for each of the series $\{e_t^{(k)}\}$ for $k = 1,\ldots,K$, where $e_t^{(k)} = y_t^{(k)} - \theta_t$. If there is no rotation bias and provided each $y_t^{(k)}$ is an unbiased estimator of θ_t , then the sample autocorrelation function of the elementary estimates can be used to identify a time series model for $\{e_t^{(k)}\}$.

Train, Cahoon & Makens(1978) reported the design-based autocovariance structure for some national level statistics of the U.S.Current Population Survey which Bell & Hillmer(1988) used to model national teenage unemployment. In addition, Bell & Hillmer(1990) developed sampling error models for the U.S. Retail Trade Survey analysing the design-based covariances obtained from a study using sample microdata. Lee(1990) presented design-based estimates for the correlations between rotation group estimates in the Canadian Labour Force Survey.

7.2.2 Quantitative Analysis Based on Pseudo Errors

In the case of a rotating panel survey, in which only the elementary estimates are available to the analyst, the model for the sampling errors can be identified (as suggested by Tiller,1992 , Pfeffermann & Bleuer,1993 and Pfeffermann, Bell & Signorelli,1996) by analysing the so-called pseudo errors given by:

$$\tilde{e}_t^{(k)} = y_t^{(k)} - y_t \quad , \qquad \textbf{(7.4)}$$

where $y_t = \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)}$. If there is no rotation bias, it follows that:

$$\begin{aligned} \tilde{e}_t^{(k)} &= y_t^{(k)} - y_t = y_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)} \\ &= (y_t^{(k)} - \theta_t) - \frac{1}{K}\sum_{k=1}^{K} (y_t^{(k)} - \theta_t) \qquad \textbf{(7.5)} \\ &= e_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)} = e_t^{(k)} - e_t \quad . \end{aligned}$$

Thus contrasts in $\tilde{e}_t^{(k)}$ are in fact functions of the rotation group sampling errors only. Assuming that the sampling errors are uncorrelated if the rotation groups do not overlap, and also that the autocorrelation structure of $\{e_t^{(k)}\}$ depends on the lag but not on the rotation group, it can be shown that $CORR(\tilde{e}_t^{(k)}, \tilde{e}_{t-h}^{(k)}) = CORR(e_t^{(k)}, e_{t-h}^{(k)})$ (see Appendix D1). In this case, the estimated covariances (or correlations) can be obtained without conditioning on θ_t and θ_{t-h} (note that θ_t is canceled out in expression (7.5)). Models for the rotation group sampling errors can be specified by applying simple model identification procedures to the various pseudo error series, $\{\tilde{e}_t^{(k)}\}$, $k=1,...,K$. Hence, after generating a pseudo error series, its autocorrelation function can be estimated using time series procedures from any standard statistical software. Note, however, that this procedure depends on the restrictive assumption that the autocorrelation structure of the sampling errors does not vary between rotation groups. To overcome this problem, Pfeffermann, Bell and Signorelli(1996) proposed a method which allows for different rotation group autocorrelation structures.

Consider a two-stage survey in which the rotation groups are composed of mutually exclusive primary sampling units which remain in the sample for all survey occasions. Consider, in addition, that the rotation pattern applies to panels of second stage units (for an example, refer to Chapter 8, section 8.1). In this case let $y_t^{(k)}$ denote an elementary estimate obtained from the k^{th} rotation group. Note that $y_t^{(k)}$, $t=1,2,...$, are based either on the same panel of second stage units or different panels selected from the same set of primary sampling units. Following Pfeffermann, Bell and Signorelli(1996) assume that:

(i) $COV(e^{(j)}_{t-h}, e^{(k)}_{t}) = 0$ if $k \neq j \quad \forall t,h$, that is, in the case of no overlap between rotation groups the sampling errors are uncorrelated;

(ii) $COV(e^{(k)}_{t-h}, e^{(k)}_{t}) = \gamma^{(k)}_{h} \quad \forall\, t$ for $k=1,\dots,K$, that is, the autocovariance depends on the lags and on the rotation groups but not on t.

These autocovariances refer to either the same panel of second stage units enumerated in different months or to different panels selected from the same set of primary sampling units (a new panel and its predecessors). Using (i) and (ii) one gets:

$$COV(e_{t-h}, e_{t}) = \gamma_{h} = \frac{1}{K^{2}} \sum_{k=1}^{K} \{COV(e^{(k)}_{t-h}, e^{(k)}_{t})\} = \frac{1}{K^{2}} \sum_{k=1}^{K} \gamma^{(k)}_{h} \quad . \tag{7.6}$$

From (7.4) and (7.5), it follows that $COV(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_{t}) = C^{(k)}_{h}$ is equal to

$$\begin{aligned} & COV(e^{(k)}_{t-h} - e_{t-h}, e^{(k)}_{t} - e_{t}) \\ &= COV(e^{(k)}_{t-h} - \frac{1}{K}\sum_{k=1}^{K} e^{(k)}_{t-h}, e^{(k)}_{t} - \frac{1}{K}\sum_{k=1}^{K} e^{(k)}_{t}) \\ &= COV(e^{(k)}_{t-h}, e^{(k)}_{t}) - \frac{1}{K} \sum_{j=1}^{K} COV(e^{(k)}_{t-h}, e^{(j)}_{t}) \\ &\quad - \frac{1}{K} \sum_{j=1}^{K} COV(e^{(j)}_{t-h}, e^{(k)}_{t}) + \frac{1}{K^{2}} \sum_{i=1}^{K}\sum_{j=1}^{K} COV(e^{(i)}_{t-h}, e^{(j)}_{t}) \quad . \end{aligned} \tag{7.7}$$

Using *(i)* and *(ii)*, Pfeffermann, Bell & Signorelli(1996) showed that (7.7) results in:

$$\begin{aligned} C^{(k)}_{h} &= \gamma^{(k)}_{h} - \frac{1}{K}\gamma^{(k)}_{h} - \frac{1}{K}\gamma^{(k)}_{h} + \frac{1}{K^{2}} \sum_{j=1}^{K} \gamma^{(j)}_{h} \\ &= \left[1 - \frac{2}{K}\right] \gamma^{(k)}_{h} + \frac{1}{K^{2}} \sum_{j=1}^{K} \gamma^{(j)}_{h} \\ &= \left[1 - \frac{2}{K} + \frac{1}{K^{2}}\right] \gamma^{(k)}_{h} + \frac{1}{K^{2}} \sum_{j\neq k}^{K} \gamma^{(j)}_{h} = \left[1 - \frac{1}{K}\right]^{2} \gamma^{(k)}_{h} + \frac{1}{K^{2}} \sum_{j \neq k} \gamma^{(j)}_{h} \quad . \end{aligned} \tag{7.8}$$

In addition, using (7.6) and (7.8), it follows that

$$
\begin{aligned}
\sum_{k=1}^{K} C_h^{(k)} &= \sum_{k=1}^{K} \left[1-\frac{1}{K}\right]^2 \gamma_h^{(k)} + \sum_{k=1}^{K} \frac{1}{K^2} \sum_{j\neq k}^{K} \gamma_h^{(j)} \\
&= \left[1-\frac{1}{K}\right]^2 \sum_{k=1}^{K} \gamma_h^{(k)} + \frac{(K-1)}{K^2} \sum_{k=1}^{K} \gamma_h^{(k)} \\
&= \left[1-\frac{2}{K}+\frac{1}{K^2}+\frac{K}{K^2}-\frac{1}{K^2}\right] \sum_{k=1}^{K} \gamma_h^{(k)} \\
&= \left[\frac{K^2-K}{K^2}\right] \sum_{k=1}^{K} \gamma_h^{(k)} = (K^2-K)\,\gamma_h = (K^2-K)\, COV(e_{t-h}, e_t) \quad .
\end{aligned}
\qquad \textbf{(7.9)}
$$

Hence, as in Pfeffermann, Bell & Signorelli(1996), the autocorrelation function ρ_h of the sampling errors can be obtained as:

$$
\begin{aligned}
\rho_h = \frac{\sum_{k=1}^{K} C_h^{(k)}}{\sum_{k=1}^{K} C_0^{(k)}} &= \frac{(K^2-K)\, COV(e_{t-h}, e_t)}{\sqrt{(K^2-K)\, COV(e_t, e_t)\, (K^2-K)\, COV(e_{t-h}, e_{t-h})}} \\
&= \frac{COV(e_{t-h}, e_t)}{COV(e_t, e_t)\, COV(e_{t-h}, e_{t-h})} \quad .
\end{aligned}
\qquad \textbf{(7.10)}
$$

In a practical situation, an estimate for the autocorrelation function in (7.10) is computed using the sample autocovariance function of the pseudo error series. In addition, using the Yule-Walker equations (see Wei,1993, p.135) and the estimated autocorrelation function, the analyst can obtain estimates for the partial autocorrelation function of the sampling error process as well as estimates for the parameters of time series model.

7.2.3 The Qualitative Analysis

For situations in which neither the individual observations nor the elementary estimates are available to the analyst, Scott, Smith and Jones(1977) proposed a time series approach based on a secondary analysis. They employed qualitative analysis to specify the model holding for the sampling errors. Following their approach, consider a single-stage overlapping survey and let $y_t = \sum_{i \in s_t} w_{ti} y_{ti}$ be a linear estimator for θ_t , where s_t denotes the set of units in the sample at time t. Assume that the conditional covariance between different units is zero, i.e. that

$$COV(y_{t-h,i}, y_{tj} \mid \theta_{t-h}, \theta_t) = 0 \quad , i \neq j \, , \forall \, h = 0, 1, \ldots \quad , \tag{7.11}$$

and denote by

$$COV(y_{t-h,i}, y_{ti} \mid \theta_{t-h}, \theta_t) = \gamma(h) \quad . \tag{7.12}$$

Then it follows that

$$\begin{aligned} COV(e_{t-h}, e_t \mid \theta_{t-h}, \theta_t) &= COV\left[\sum_{i \in s_{t-h}} w_{t-h,i} \; y_{t-h,i}, \sum_{j \in s_t} w_{tj} \; y_{tj} \mid \theta_{t-h}, \theta_t \right] \\ &= \sum_{i \in s_t \cap s_{t-h}} w_{t-h,i} w_{ti} \; COV(y_{t-h,i}, y_{ti} \mid \theta_{t-h}, \theta_t) \qquad \textbf{(7.13)} \\ &= \gamma(h) \sum_{i \in s_t \cap s_{t-h}} w_{t-h,i} w_{ti} \quad . \end{aligned}$$

In addition, $V(e_t \mid \theta_t)$ is given by

$$\begin{aligned} V(e_t \mid \theta_t) &= V\left[\sum_{i \in s_t} w_{ti} \; y_{ti} \mid \theta_t \right] \\ &= \sum_{i \in s_t} w_{ti}^2 \; V(y_{ti} \mid \theta_t) + 2 \sum_{i < j \in s_t} w_{ti} w_{tj} \, COV(y_{ti}, y_{tj} \mid \theta_t) \quad . \end{aligned} \tag{7.14}$$

Using (7.11) and (7.12) with $h = 0$ one gets

$$\begin{aligned} V(e_t \mid \theta_t) &= \sum_{i \in s_t} w_{ti}^2 \; V(y_{ti} \mid \theta_t) \\ &= \gamma(0) \sum_{i \in s_t} w_{ti}^2 \quad . \end{aligned} \tag{7.15}$$

Note that $V(e_t \mid \theta_t) = V(e_{t-h} \mid \theta_{t-h})$ since $\{e_t\}$ is assumed to be stationary. Putting (7.13) and (7.15) together results in:

$$CORR(e_{t-h}, e_t \mid \theta_{t-h}, \theta_t) \;=\; \rho_{e|\theta}(h) \;=\; \frac{\gamma(h) \sum\limits_{i \in s_t \cap s_{t-h}} w_{t-h,i}\, w_{ti}}{\gamma(0) \sum\limits_{i \in s_t} w_{ti}^2} \quad . \tag{7.16}$$

Assuming that y_t is an equally weighted estimator and that the weights are held fixed for all survey rounds, i.e. that $w_{ti} = w$,it follows that

$$COV(e_{t-h}, e_t \mid \theta_{t-h}, \theta_t) \;=\; \gamma(h) \sum_{i \in s_t \cap s_{t-h}} w_{t-h,i}\, w_{ti} \;=\; \gamma(h)\; n(h) w^2 \quad , \tag{7.17}$$

and also that

$$\begin{aligned} V(e_t \mid \theta_t) \;&=\; \sum_{i \in s_t} w_{ti}^2 \; V(y_{ti} \mid \theta_t) \\ &=\; \gamma(0) \sum_{i \in s_t} w_{ti}^2 \;=\; n\, w^2\, \gamma(0) \quad , \end{aligned} \tag{7.18}$$

where $n(h)$ is the number of common units in s_t and s_{t-h} .
Then putting (7.17) and (7.18) together leads to

$$\begin{aligned} CORR(e_{t-h}, e_t \mid \theta_{t-h}, \theta_t) \;&=\; \rho_{e|\theta}(h) \;=\; \frac{\gamma(h)\, n(h) w^2}{\gamma(0) n w^2} \\ &=\; \rho(h) \frac{n(h)}{n} \;=\; \pi_h \; \rho(h) \quad , \end{aligned} \tag{7.19}$$

where π_h is the proportion of overlap between the two occasions and $\rho(h) = \dfrac{\gamma(h)}{\gamma(0)}$.

Examining the above correlation structure in (7.19), Scott, Smith and Jones(1977) pointed out that if $\pi_h = 0$ for $h > q$ then $\rho_{e|\theta}(h) = 0$ for $h > q$, suggesting a MA(q) for the sampling errors. In addition, for panel surveys in which all units are retained in the sample for all survey rounds it follows that $\pi_h = 1$ and $\rho_{e|\theta}(h)$ has the general form $\rho(h)$. In this case the authors suggested adopting the simplifying assumption that $\rho(h) = \rho^h$, which implies an AR(1) model for the sampling errors.

When the proportion of overlap, π_t, is constant for all survey rounds it follows that $\rho_{e|\theta}(h) = \pi\,\rho(h)$. This can happen either in a split-panel or in a rotating panel survey. By assuming that the autocorrelation of the sampling errors decays exponentially over time, that is $\rho_{e|\theta}(h) \propto \pi\,\rho^h$, Scott, Smith & Jones(1977) suggested an ARMA(1,1) model to represent the sampling error process.

The authors also provided the same sort of analysis for multi-stage surveys pointing out that in this case the overlap can occur at any sampling stage. Consider for example a two-stage survey. The value y_{tij} for the j^{th} unit in the i^{th} psu at time t was modelled as:

$$y_{tij} = \theta_t + A_{ti} + B_{tij} \quad , \tag{7.20}$$

where A_{ti} and B_{tij} are cluster-level and unit-level random effects with zero means and variances σ_a^2 and σ_b^2, respectively. It was assumed that secondary stage units (ssu) within the same primary sampling unit (psu) are correlated whereas units in different primary sampling units are uncorrelated. Then, to complete the model specification consider:

$$(i)\ COV(A_{t-h,i}, B_{tij}) = 0 \quad \forall\, t,h,i,j \quad ,$$

$$(ii)\ COV(A_{t-h,i}, A_{ti'}) = \begin{cases} \gamma_a(h)\; i=i' \;\; \forall\, t,h \\ 0 \quad\; i \neq i' \;\; \forall\, t,h \end{cases} ,$$

$$(iii)\ COV(B_{t-h,ij}, B_{ti'j'}) = \begin{cases} \gamma_b(h)\; i=i'\; j=j' \;\; \forall\, t,h \\ 0 \quad\; otherwise \end{cases} ,$$

(iv) all the other covariances are zero.

Expressing y_t as a linear estimator in which the weights are attached to the secondary stage units yields $\sum\sum_{i,j \in s_t} w_{tij}\, y_{tij}$, where $\sum\sum_{i,j \in s_t} w_{tij} = 1$ is the unbiasedness condition. The sampling error is given by

$$\begin{aligned} e_t &= y_t - \theta_t = \sum\sum_{i,j\in s_t} w_{tij}\, y_{tij} - \theta_t &&= \sum\sum_{i,j\in s_t} w_{tij}\, y_{tij} - \theta_t \sum\sum_{i,j\in s_t} w_{tij} \\ &= \sum\sum_{i,j\in s_t} w_{tij}\, y_{tij} - \sum\sum_{i,j\in s_t} w_{tij}\, \theta_t &&= \sum\sum_{i,j\in s_t} w_{tij}\,(y_{tij} - \theta_t) \;. \end{aligned} \tag{7.21}$$

Using (7.20) and (7.21) it follows that

$$e_t = \sum\sum_{i,j\in s_t} w_{tij}\,(A_{ti} + B_{tij}) = \sum_{i\in s_t} w_{ti} A_{ti} + \sum\sum_{i,j\in s_t} w_{tij} B_{tij} \;,$$

with $w_{ti} = \sum_{j\in s_t} w_{tij}$.

Thus,

$$\begin{aligned} & COV(e_{t-h}, e_t \,|\, \theta_{t-h}, \theta_t) \\ &= COV\Big(\sum_{i\in s_{t-h}} w_{t-h,i} A_{t-h,i} + \sum\sum_{i,j\in s_{t-h}} w_{t-h,ij} B_{t-h,ij} \,,\, \sum_{i\in s_t} w_{ti} A_{ti} + \sum\sum_{i,j\in s_t} w_{tij} B_{tij}\Big) \\ &= COV\Big(\sum_{i\in s_{t-h}} w_{t-h,i} A_{t-h,i}, \sum_{i\in s_t} w_{ti} A_{ti}\Big) + COV\Big(\sum_{i\in s_{t-h}} w_{t-h,i} A_{t-h,i}, \sum\sum_{i,j\in s_t} w_{tij} B_{tij}\Big) \\ &+ COV\Big(\sum\sum_{i,j\in s_{t-h}} w_{t-h,ij} B_{t-h,ij}, \sum_{i\in s_t} w_{ti} A_{ti}\Big) + COV\Big(\sum\sum_{i,j\in s_{t-h}} w_{t-h,ij} B_{t-h,ij}, \sum\sum_{i,j\in s_t} w_{tij} B_{tij}\Big) \;. \end{aligned}$$

$$\begin{aligned} &= \sum_{i\in s_t \cup s_{t-h}} w_{t-h,i} w_{ti}\, COV(A_{t-h,i}, A_{ti}) + \sum_{i\in s_t \cup s_{t-h}} w_{t-h,i} \sum_{j\in s_t} w_{tij}\, COV(A_{t-h,i}, B_{tij}) \\ &+ \sum\sum_{i\in s_t \cup s_{t-h}} w_{ti} \sum_{j\in s_{t-h}} w_{t-h,ij} COV(B_{t-h,ij}, A_{ti}) + \sum\sum_{i,j\in s_t \cup s_{t-h}} w_{t-h,ij} w_{tij}\, COV(B_{t-h,ij}, B_{tij}) \;. \end{aligned}$$

Using the assumptions *(i)* to *(iv)* it follows that

$$COV(e_{t-h}, e_t \,|\, \theta_{t-h} \theta_t) = \gamma_a(h) \sum_{i\in s_t \cap s_{t-h}} w_{t-h,i}\, w_{ti} + \gamma_b(h) \sum\sum_{i,j\in s_t \cap s_{t-h}} w_{t-h,ij}\, w_{tij} \;.$$

Denoting

$$VAR(e_t \,|\, \theta_t) = \gamma_a(0) \sum_{i\in s_t} w_{ti}^2 + \gamma_b(0) \sum\sum_{i,j\in s_t} w_{tij}^2 = S^2 \quad \forall\ t,$$

yields

$$CORR(e_{t-h}, e_t \,|\, \theta_{t-h}, \theta_t) = \rho_{e|\theta}(h)$$

$$= \frac{\gamma_a(h)}{S^2} \sum_{i \in s_t \cap s_{t-h}} w_{t-h,i}\, w_{ti} + \frac{\gamma_b(h)}{S^2} \sum \sum_{i,j \in s_t \cap s_{t-h}} w_{t-h,ij}\, w_{tij} \qquad (7.22)$$

$$= \frac{\gamma_a(h)}{S^2}\, b_1(t,h) + \frac{\gamma_b(h)}{S^2}\, b_2(t,h) \quad ,$$

with $b_1(t,h) = \sum_{i \in s_t \cap s_{t-h}} w_{t-h,i}\, w_{ti}$ and $b_2(t,h) = \sum \sum_{i,j \in s_t \cap s_{t-h}} w_{t-h,ij}\, w_{tij}$.

Although the autocorrelation in (7.22) seems to represent a complex stochastic process, Scott, Smith & Jones(1977)showed that for some survey designs this function matches the autocorrelation function of low order ARMA processes. If the samples do not overlap for $h > q$ it follows that $s_t \cap s_{t-h} = \{\emptyset\} \;\; \forall\, h = q, q+1, \ldots$ then $b_1(h) = b_2(h) = 0 \;\; \forall\, h > q$ and the sampling error process can be represented by a MA(q) model. When independent samples of secondary sampling units are drawn from the selected primary sampling units on each occasion it follows that $b_2(t,h) = 0 \;\; \forall\, h > 0$. In this case,

$$\rho_{e|\theta}(h) = \frac{\gamma_a(h)}{S^2}\, b_1(t,h) \quad .$$

If the overlap between primary sampling units and the weights are constant over time, say $w_{tij} = w_{ij} \;\; \forall\, t$, then $b_1(t,h) = b$ and

$$\rho_{e|\theta}(h) = b\, \frac{\gamma_a(h)}{S_t^2} = b\, \rho_a(h) \quad .$$

Working with the assumption that $\rho_a(h)$ decays exponentially the authors suggested an ARMA(1,1) model for the sampling error process for any $0 < b \leq 1$.

The analysis presented above illustrates that even without an estimate for the autocorrelation function of the sampling errors the analyst can propose a reasonable time series model for the sampling error process. Note, however, that the qualitative approach requires rather strong assumptions about the underlying autocorrelation structure of the individual units. Before proceeding further it is important to note that the use of a qualitative analysis is not confined to cases in which the final model is fitted using aggregate estimates as inputs. Pfeffermann(1991) employed a qualitative analysis in order to define the model holding for the rotation group sampling errors in the Israeli Labour Force Survey. Following Blight & Scott(1973), the author assumed a relationship between individual values given by $y_{ti} - \theta_t = \rho\,(y_{t-h,i} - \theta_{t-h}) + \nu_{ti}$, which induces a first order autoregressive model for the rotation group sampling errors. Hence the model for the sampling error was again defined without the estimation of an autocorrelation function using sample data.

Having discussed the two different approaches to model the sampling error in a univariate framework, the next step is to introduce procedures to model multiple time series of sampling errors.

7.3 The Multivariate Case

Denote by $\boldsymbol{\theta}_t = (\theta_{1t},\dots,\theta_{Mt})'$ a vector of M finite population parameters of interest at time t. Note that here the indexes $1t,\dots,Mt$ are used to express that M variables are being modelled concurrently. In this case y_{mti} denotes the individual value of the characteristic m at time t for the unit i . Let $\boldsymbol{y}_t = (y_{1t},\dots,y_{Mt})'$ represent a vector of survey-based estimates of $\boldsymbol{\theta}_t$. If the vector of sampling errors is $\boldsymbol{e}_t = (e_{1t},\dots,e_{Mt})'$, it follows that:

$$\boldsymbol{y}_t = \boldsymbol{\theta}_t + \boldsymbol{e}_t \quad .$$

Assuming that $\boldsymbol{y}_t$ is design unbiased then $E(\boldsymbol{e}_t\,|\,\boldsymbol{\theta}_t) = \boldsymbol{0}$ and $V(\boldsymbol{e}_t\,|\,\boldsymbol{\theta}_t) = \Sigma_y$, where Σ_y contains the sampling variances and covariances of $\boldsymbol{y}_t$ as an estimator of $\boldsymbol{\theta}_t$.

In a multivariate framework the main objective is to model simultaneously a multiple time series of survey estimates in order to get signal estimates for $\boldsymbol{\theta}_t$. Following the ideas described for the univariate case, it seems natural that the analyst should now base the model identification procedure on the cross-correlation function (Wei,1993, p.333) of the sampling error series. Before proceeding further, recall from Definition 3.2 that a VARMA model for a M-dimensional multiple time series $\{\boldsymbol{e}_t\}$ (with mean vector $E(\boldsymbol{e}_t) = 0$) is defined as $\boldsymbol{\Phi}(B)\boldsymbol{e}_t = \boldsymbol{\delta}(B)\boldsymbol{a}_t$. The cross-covariance matrix function for the Vector ARMA process $\{\boldsymbol{e}_t\}$ (from Wei,1993, p.333) is given by:

$$\boldsymbol{\Gamma}_e(h) = COV(\boldsymbol{e}_{t-h}, \boldsymbol{e}_t) = E(\boldsymbol{e}_{t-h}\ \boldsymbol{e}_t^{\prime}) \quad ,$$

where $\{\boldsymbol{\Gamma}_e(h)\}_{ml} = \gamma_{eml}(h) = COV(e_{m,t-h}, e_{lt})$. The cross-correlation function for the vector process is defined as:

$$\mathbf{P}_e(h) = \boldsymbol{D}_e^{-1/2}\, \boldsymbol{\Gamma}_e(h)\, \boldsymbol{D}_e^{-1/2} \quad ,$$

where $\boldsymbol{D}_e$ is the diagonal matrix in which the mth diagonal element is the variance of the mth process. That is:

$$\boldsymbol{D}_e = diag(\gamma_{e11}(0), \ldots, \gamma_{eMM}(0)) \quad .$$

The sections that follow introduce the general framework for modelling multiple time series of sampling errors.

7.3.1 The Quantitative Analysis

When the individual sampling units are available and the survey microdata can be linked over time $\boldsymbol{\Gamma}_e(h)$ can be estimated via the conditional covariances $\boldsymbol{\Gamma}_{e|\theta}(h)$, given $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t-h}$, using design-based methods. Consider, for example, the bivariate case where $\boldsymbol{\Gamma}_{e|\theta}(h)$ has the form:

$$\boldsymbol{\Gamma}_{e|\theta}(h) = \begin{bmatrix} COV(e_{1,t-h}, e_{1t} | \theta_{t-h}, \theta_t) & COV(e_{1,t-h}, e_{2t} | \theta_{t-h}, \theta_t) \\ COV(e_{2,t-h}, e_{1t} | \theta_{t-h}, \theta_t) & COV(e_{2,t-h}, e_{2t} | \theta_{t-h}, \theta_t) \end{bmatrix}$$

$$= \begin{bmatrix} COV(y_{1,t-h}, y_{1t} | \theta_{t-h}, \theta_t) & COV(y_{1,t-h}, y_{2t)} | \theta_{t-h}, \theta_t) \\ COV(y_{2,t-h}, y_{1t} | \theta_{t-h}, \theta_t) & COV(y_{2,t-h}, y_{2t} | \theta_{t-h}, \theta_t) \end{bmatrix} \quad , \qquad \textbf{(7.23)}$$

and each matrix cell can be estimated as described in Section 7.2.1.

Note that, in practice, the estimates in (7.23) are computed for some set of time points t and lags h. The standard procedure (see for example, Bell & Hillmer, 1990) is to average them over t to obtain "improved" estimates of the covariances.

In principle, the identification for vector time series is similar to that for univariate time series. Hence, a suitable time series model for a multivariate process can be identified from the pattern of its cross-correlation and partial autoregression matrices as in Wei(1993, p.350-351). Based on the estimated cross-covariance matrices for the sampling errors, estimates for the cross-correlation and partial autoregression matrices and can be computed and a time series model to reproduce this correlation structure can be formulated.

In the case of a rotating panel survey, when only the K elementary estimates are available, K M-dimensional time series of pseudo errors can be constructed from deviations of the rotation group estimates about the overall mean. Following the same notation as in Section 7.2.1, a time series of pseudo errors for the k^{th} rotation group is defined as:

$$\tilde{e}_t^{(k)} = y_t^{(k)} - y_t = (\tilde{e}_{1t}^{(k)},\dots,\tilde{e}_{Mt}^{(k)})' = (y_{1t}^{(k)} - y_{1t},\dots,y_{Mt}^{(k)} - y_{Mt})' \quad , \tag{7.24a}$$

where $y_t = \dfrac{1}{K}\sum_{k=1}^{K} y_t^{(k)}$.

Note that if there is no rotation bias and if y_{mt} is an unbiased estimator for θ_{mt} , then $e_{mt}^{(k)} = y_{mt}^{(k)} - \theta_{mt}$ for $m=1,\dots,M$ and

$$\begin{aligned}\tilde{e}_t^{(k)} &= y_t^{(k)} - y_t = y_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)} \\ &= (y_t^{(k)} - \theta_t) - \frac{1}{K}\sum_{k=1}^{K}(y_t^{(k)} - \theta_t) \\ &= e_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)} = e_t^{(k)} - e_t \quad .\end{aligned} \tag{7.24b}$$

Assuming that the correlations between sampling errors do not depend on the rotation groups (although they may depend on the lags and may vary between characteristics) and assuming, in addition, that if the rotation groups do not overlap the corresponding sampling errors are uncorrelated, it can be shown that $\mathbf{P}_e^{(k)}(h) = \mathbf{P}_{\tilde{e}}^{(k)}(h)$ (see Appendix D2 for details). In this case, the model holding for the rotation group sampling error series $\{e_t^{(k)}\}$ can be specified by applying model identification procedures to each of the (multiple) pseudo error series $\{\tilde{e}_t^{(k)}\}$. The sample cross-correlation functions can be estimated using any statistical software capable of handling multiple time series (such as the SCA, 1986, for example).

However, as in the univariate case (Section 7.2.1), the assumption that the cross-correlations do not depend on the rotation group is not always appropriate. To circumvent this problem, the method introduced by Pfeffermann, Bell & Signorelli(1996), described in Section 7.2.1, is adapted for a multivariate framework as follows.

Consider, as in Section 7.2.2, a two-stage survey in which the rotation groups are composed of mutually exclusive primary sampling units which remain in the sample for all survey occasions. Recall that the rotation pattern applies to panels of second stage units. Denote by $e_{mt}^{(k)}$ the error term for the m^{th} characteristic at time t refering to the k^{th} rotation group. Note that $e_{mt}^{(k)}$, $t = 1, 2, \ldots$, refer either to the same panel of second stage units or different panels selected from the same set of primary sampling units.

Let

$$CORR(e_{t-h}, e_t) = P_e(h) = D_e^{-1/2}\, \Gamma_e(h) D_e^{-1/2} \quad , \tag{7.25}$$

where the element in row m and column l of $\Gamma_e(h)$, denoted $\gamma_{ml}(h)$ is given by

$$\gamma_{ml}(h) = COV(e_{m,t-h}, e_{lt}) = COV\left[\frac{1}{K}\sum_{j=1}^{K} e_{m,t-h}^{(j)}, \frac{1}{K}\sum_{k=1}^{K} e_{lt}^{(k)}\right] \quad . \tag{7.26}$$

and

$$D_e = diag(\gamma_{11}(0), \ldots, \gamma_{MM}(0)) \quad . \tag{7.27}$$

In addition, assume that

(i) $COV(e^{(j)}_{m,t-h}, e^{(k)}_{lt}) = 0$ if $j \neq k \quad \forall t,h,m,l$, that is, in the case of no overlap between rotation groups the sampling errors are uncorrelated;

(ii) the sampling error autocovariances vary between characteristics, depend on the lags and on the rotation groups, but not on t . That is, $COV(e^{(k)}_{m,t-h}, e^{(k)}_{lt}) = \gamma^{(k)}_{ml}(h)$ for $k=1,\dots,K$ and $m,l=1,\dots,M$.

Using (i) and (ii) it follows that

$$\begin{aligned}\gamma_{ml}(h) &= COV(e_{m,t-h}, e_{lt}) &= COV\left[\frac{1}{K}\sum_{j=1}^{K} e^{(j)}_{m,t-h}, \frac{1}{K}\sum_{k=1}^{K} e^{(k)}_{lt}\right] \\ &= \frac{1}{K^2}\sum_{k=1}^{K} \gamma^{(k)}_{ml}(h) \quad .\end{aligned} \tag{7.28}$$

Now denote the cross-correlation matrix of the series of pseudo errors by $\mathbf{P}_{\tilde{e}}(h)$, where:

$$\mathbf{P}_{\tilde{e}}(h) = CORR(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_{t}) = (D^{(k)}_{\tilde{e}})^{-1/2}\ \Gamma^{(k)}_{\tilde{e}}(h)\,(D^{(k)}_{\tilde{e}})^{-1/2} \quad , \tag{7.29}$$

with

$$\{\Gamma^{(k)}_{\tilde{e}}(h)\}_{ml} = COV(\tilde{e}^{(k)}_{m,t-h}, \tilde{e}^{(k)}_{lt}) \quad . \tag{7.30}$$

From (7.24), it follows that

$$COV(\tilde{e}^{(k)}_{m,t-h}, \tilde{e}^{(k)}_{lt}) = COV(e^{(k)}_{m,t-h} - e_{m,t-h}, e^{(k)}_{lt} - e_{lt}) \quad , \tag{7.31}$$

with

$$
\begin{aligned}
& COV(e^{(k)}_{m,t-h} - e_{m,t-h}, e^{(k)}_{lt} - e_{lt}) \\
&= COV(e^{(k)}_{m,t-h} - \frac{1}{K}\sum_{k=1}^{K} e^{(k)}_{m,t-h},\, e^{(k)}_{lt} - \frac{1}{K}\sum_{k=1}^{K} e^{(k)}_{lt}) \\
&= COV(e^{(k)}_{m,t-h}, e^{(k)}_{lt}) \;-\; \frac{1}{K}\sum_{j=1}^{K} COV(e^{(k)}_{m,t-h}, e^{(j)}_{lt}) \\
&- \frac{1}{K}\sum_{j=1}^{K} COV(e^{(j)}_{m,t-h}, e^{(k)}_{lt}) \;+\; \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} COV(e^{(i)}_{m,t-h}, e^{(j)}_{lt}) \quad .
\end{aligned}
\tag{7.32}
$$

Using *(i)* and *(ii)*, (7.32) results in:

$$
\begin{aligned}
\{\,\Gamma^{(k)}_{\tilde{e}}(h)\,\}_{ml} &= \gamma^{(k)}_{ml}(h) \;-\; \frac{1}{K}\,\gamma^{(k)}_{ml}(h) \;-\; \frac{1}{K}\,\gamma^{(k)}_{ml}(h) \;+\; \frac{1}{K^2}\sum_{j=1}^{K}\gamma^{(j)}_{ml}(h) \\
&= \left[1 - \frac{2}{K}\right]\gamma^{(k)}_{ml}(h) \;+\; \frac{1}{K^2}\sum_{j=1}^{K}\gamma^{(j)}_{ml}(h) \\
&= \left[1 - \frac{2}{K} + \frac{1}{K^2}\right]\gamma^{(k)}_{ml}(h) \;+\; \frac{1}{K^2}\sum_{j\neq k}\gamma^{(j)}_{ml}(h) \\
&= \left[1 - \frac{1}{K}\right]^2\gamma^{(k)}_{ml}(h) \;+\; \frac{1}{K^2}\sum_{j\neq k}\gamma^{(j)}_{ml}(h) \quad .
\end{aligned}
\tag{7.33}
$$

To obtain the cross-correlation matrices of the sampling errors note that using (7.28) one gets

$$
\begin{aligned}
\left\{\sum_{k=1}^{K}\Gamma^{(k)}_{\tilde{e}}(h)\right\}_{ml} &= \sum_{k=1}^{K}\left[(1-1/K)^2\,\gamma^{(k)}_{ml}(h) \;+\; \frac{1}{K^2}\sum_{j\neq k}\gamma^{(j)}_{ml}(h)\right] \\
&= \frac{K^2-K}{K^2}\sum_{k=1}^{K}\gamma^{(k)}_{ml}(h) \;=\; (K^2-K)\gamma_{ml}(h) \quad .
\end{aligned}
\tag{7.34}
$$

From (7.34) it follows that

$$
(K^2-K)\Gamma_{\tilde{e}}(h) \;=\; \sum_{k=1}^{K}\Gamma^{(k)}_{\tilde{e}}(h) \quad .
\tag{7.35}
$$

The same sort of argument implies that

$$(K^2-K)D_e = \sum_{k=1}^{K} D_{\tilde{e}}^{(k)} \quad . \tag{7.36}$$

Now, taking (7.35) and (7.37) into the expression (7.25) for the sampling error cross-correlation matrix, leads to

$$P_{\tilde{e}}(h) = \left[\sum_{k=1}^{K} (D_{\tilde{e}}^{(k)})\right]^{-1/2} \left[\sum_{k=1}^{K} \Gamma_{\tilde{e}}^{(k)}(h)\right] \left[\sum_{k=1}^{K} (D_{\tilde{e}}^{(k)})\right]^{-1/2} \quad . \tag{7.37}$$

Therefore the cross-covariance matrices and the cross-correlation function of the sampling error process can be obtained by averaging the pseudo error cross-covariance matrices. Consequently, a multivariate time series model for the sampling errors can be identified and estimates of the parameter matrices can be computed, provided the series of pseudo-errors are available.

7.3.2 The Qualitative Analysis

This section introduces a framework for modelling multiple time series of sampling errors based on a qualitative analysis. This is, in fact, a generalization of the ideas presented by Scott, Smith & Jones(1977).

Consider the case of a single-stage survey and let

$$y_{mt} = \sum_{i \in s_t} w_{mti}\, y_{mti} \tag{7.38}$$

be a linear estimator of θ_{mt} for $m=1,\dots,M$. Assume that, given $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t-h}$:

(i) $COV(y_{m,t-h,i}, y_{ltj}) = 0$ for $i \neq j$, $\forall\ m,l,t,h$;

(ii) $COV(y_{m,t-h,i}, y_{lti}) = \gamma_{ml}(h)$, $\forall\ m,l,t,h,i$.

Since

$$\begin{aligned}\Gamma_{e|\theta}(h) &= COV(\mathbf{e}_{t-h},\mathbf{e}_t \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) = COV(\mathbf{y}_{t-h}-\boldsymbol{\theta}_{t-h},\mathbf{y}_t-\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) \\ &= COV(\mathbf{y}_{t-h},\mathbf{y}_t \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) = \Gamma_{y|\theta}(h) \quad ,\end{aligned}$$

then

$$\{\boldsymbol{\Gamma}_{e|\theta}(h)\}_{ml} = COV(e_{m,t-h},e_{lt} \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) = COV(y_{m,t-h},y_{lt} \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) \ . \tag{7.39}$$

Substituting (7.38) in (7.39) and using *(i)* and *(ii)* yields

$$\begin{aligned}COV(e_{m,t-h},e_{lt} \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) &= COV\left[\sum_{i\in s_{t-h}} w_{m,t-h,i}\; y_{m,t-h,i}, \sum_{j\in s_t} w_{ltj}\; y_{ltj} \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t\right] \\ &= \sum_{i\in s_t\cap s_{t-h}} w_{m,t-h,i}\, w_{lti}\; COV(y_{m,t-h,i},y_{lti} \,|\, \boldsymbol{\theta}_{t-h},\boldsymbol{\theta}_t) \qquad (7.40) \\ &= \gamma_{ml}(h) \sum_{i\in s_t\cap s_{t-h}} w_{m,t-h,i}\, w_{lti} \quad .\end{aligned}$$

As a special case, consider that y_{mt} is an equally weighted estimator and that the same weights are used for all M characteristics of interest. If the weights are held fixed for all survey rounds ($w_{mti}=w \quad , \quad \forall \ m,t,i$) then equation (7.40) can be simplified to:

$$COV(e_{m,t-h},e_{lt} \,|\, \theta_{t-h},\theta_t) = \gamma_{ml}(h) \sum_{i\in s_t\cap s_{t-h}} w_{m,t-h,i}\, w_{lti} = \gamma_{ml}(h)\ n(h)\, w^2 \ , \tag{7.41}$$

for $m,l = 1,\dots,M$, where $n(h)$ is the number of common units in s_t and s_{t-h} . Then it follows that

$$\boldsymbol{\Gamma}_{e|\theta}(h) = n(h)\, w^2\ \boldsymbol{\Gamma}(h) \quad , \tag{7.42}$$

where $\{\boldsymbol{\Gamma}(h)\}_{ml} = \gamma_{ml}(h)$, as defined in *(ii)*. That is, $\Gamma(h)$ is a matrix in which the elements represent covariances between individual units. Note in addition that

UNIVERSITY OF SOUTHAMPTON LIBRARY

$$
\begin{aligned}
COV(e_{mt}, e_{mt} \mid \theta_{t-h}, \theta_t) &= COV\left[\sum_{i \in s_t} w_{mti}\; y_{mti}, \sum_{j \in s_t} w_{mtj}\; y_{mtj} \mid \theta_{t-h}, \theta_t\right] \\
&= \sum_{i \in s_t} w_{mti}\, w_{mti}\;\; COV(y_{mti}, y_{mti} \mid \theta_{t-h}, \theta_t) \qquad \textbf{(7.43)} \\
&= \gamma_{mm}(0) \sum_{i \in s_t} w_{mti}^2 = \gamma_{mm}(0)\;\, n w^2 \quad .
\end{aligned}
$$

where n is the total sample size on each survey round. From equation (7.43) it follows that, for the variances,

$$
D_{e|\theta} = n\, w^2 D = n\, w^2\, diag(\gamma_{11}(0)\;,\; \dots\;,\; \gamma_{MM}(0)) \quad . \qquad \textbf{(7.44)}
$$

Putting equations (7.42) and (7.44) together results in:

$$
\begin{aligned}
\mathbf{P}_{e|\theta}(h) &= D_{e|\theta}^{-1/2}\;\, \Gamma_{e|\theta}(h)\;\, D_{e|\theta}^{-1/2} \\
&= (n\, w^2 D)^{-1/2}\; n(h)\, w^2\, \Gamma(h)\;\; (n\, w^2 D)^{-1/2} \qquad \textbf{(7.45)} \\
&= \frac{n(h)}{n}\;\, \mathbf{P}(h) = \pi_h\; \mathbf{P}(h) \quad ,
\end{aligned}
$$

where $\mathbf{P}(h) = D^{-1/2}\, \Gamma(h)\, D^{-1/2}$ is the cross-correlation matrix of the individual units.

Similarly to the univariate case, if there is no overlap between the sampling units ($\pi_h = 0$) for $h > q$ then $\mathbf{P}_{e|\theta}(h) = \mathbf{0} \;\; \forall \;\; h > q$. The cross-correlation matrix for a vector moving-average process of order q is zero for lags greater than q (see appendix D3 for details). Thus for single-stage short-term overlapping surveys, in which units are retained in the sample for q consecutive occasions, a vector MA(q) model should be appropriate for the sampling error process. The same sort of argument can be used to model the sampling error process according to different patterns of overlap and the above results can be readily extended for the case of two-stage surveys.

For a single-stage completely overlapping survey it follows that $\pi_h = 1$ and $P_{e|\theta}(h)$ in (47) has the general form $\mathbf{P}(h)$. Assuming that $P(h) \propto P^h$, a vector AR(1) model may be suggested for the sampling error process $\{e_t\}$. Note that the cross-correlation matrix for a vector AR(1) has the form (see Appendix D3):

$$P(h) = P(0)\,(\Phi_1^*)^h \tag{7.46}$$

with $\Phi_1^* = D^{-1/2}\Phi_1 D^{1/2}$. The expression in (7.46) indicates that $P(h)$ has a die-out pattern as in the autocorrelation function of a univariate AR(1) process.

A VARMA(1,1) model could also be used to represent the sampling error process, since its cross-correlation function can be expressed (details in Appendix D3) as

$$P(h) = P(1)(\Phi_1^*)^{h-1} \quad h \geq 2 \quad .$$

It is interesting to note that this dual choice is not available when modelling the sampling error process from completely overlapping surveys in a univariate framework. Recall that the autocorrelation function of an AR(1) model is given by

$$\rho(h) \approx \rho^h \quad h \geq 1 \quad , \tag{7.47}$$

with $\rho(0) = 1$, and that the autocorrelation function of an ARMA(1,1) has the form

$$\rho(h) \approx \rho^h c \quad h \geq 2 \quad . \tag{7.48}$$

Comparing the functions in (7.47) and (7.48) it becomes clear that the AR(1) model is a better choice for completely overlapping surveys whereas an ARMA(1,1) is a more suitable model for partially overlapping surveys with constant proportion of overlap.

In a multivariate framework there are reasons to favour the VAR(1) model. Foremost, the univariate model structure implied by a Vector AR(1) model is such that the individual series follow ARMA(M,M-1) models (see Reinsel, 1993, p.29, Maravall & Mathis,1994 or Chan & Wallis,1978). Note that, M and M-1 are the maximum orders for the individual ARMA models. Particularly, if $\{\Phi_1\}_{ij} = 0$ for $i \neq j = 1,\dots,M$, each $\{e_{mt}\}$ series ($m = 1,\dots,M$) would follow an AR(1) model (for an example see Appendix D4). In fact, this special case mirrors a situation in which the analyst is modelling each series separately since the off-diagonal elements in the parameter matrix Φ_1 represent exactly the lack of influence of each series on the others. Therefore a VAR(1) model for the multivariate process $\{e_t\}$

matches quite well with the existing theory for the univariate case.

In addition, two VARMA(p,q) models (with $p,q > 0$) can give rise to the same covariance matrix structure $\Gamma(h)$. In other words the models can be observationally equivalent (see Reinsel, 1993, pp.36-39). In this case certain constraints need to be imposed on the matrix operators $\Phi(B)$ and $\delta(h)$ to select uniquely one parameter set from a class of equivalent structures. Hence some analysts prefer to restrict attention to Vector Autoregressive models.

In a partially overlapping survey with constant proportion of overlap, the sampling error process can also be modelled by a VAR(1) model. In this case $\pi_h = \pi$, which implies $P_{e|\theta}(h) = \pi P(h)$. Assuming $P(h) \propto P^h$ yields $P_{e|\theta}(h) \propto \pi P^h$. Consequently, once again a VAR(1) can be used to model the sampling errors.

Consider now the case of a two-stage survey. As in section 7.2.2, it will be assumed that units within the same psu are correlated whereas units in different psu's are uncorrelated. Let y_{mtij} be the value of the j^{th} unit within the i^{th} primary sampling unit for characteristic m at time t . Based on Scott, Smith and Jones(1977), given θ_{mt} , y_{mtij} can be modelled by the following random effects model:

$$y_{mtij} = \theta_{mt} + A_{mti} + B_{mtij} \quad , \tag{7.49}$$

where A_{mti} is a cluster-level random term with zero mean and variance σ^2_{ma} and B_{mtij} is a unit-level random term with zero mean and variance σ^2_{mb} . To complete the model specification it is assumed that:

$$(i)\; COV(A_{m,t-h,i}, B_{lti'j}) = 0 \quad \forall\, m,l,t,h,i,i',j \quad , \tag{7.50a}$$

$$(ii)\; COV(A_{m,t-h,i}, A_{lti'}) = \begin{cases} \gamma_{ma}(h) & i=i'\; m=l\; \forall\, t,h \\ \gamma_{mla}(h) & i=i'\; m\neq l\; \forall\, t,h \\ 0 & i\neq i'\; \forall\, m,l,t,h \end{cases} \quad , \tag{7.50b}$$

$$(iii)\ COV(B_{m,t-h,ij}, B_{lti'j'}) = \begin{cases} \gamma_{mb}(h)\ i=i'\ j=j'\ m=l\ \forall\, t,h \\ \gamma_{mlb}(h)\ i=i'\ j=j'\ m\neq l\ \forall\, t,h \\ 0 \qquad otherwise \end{cases} \quad . \tag{7.50c}$$

Expressing y_{mt} as a linear estimator in which the weights are attached to the secondary stage units yields $y_{mt} = \sum\sum_{i,j\in s_t} w_{mtij}\, y_{mtij}$ where $\sum\sum_{i,j\in s_t} w_{mtij} = 1$ is the unbiasedness condition. Then, the sampling error is given by

$$\begin{aligned} e_t = y_{mt} - \theta_{mt} &= \sum\sum_{i,j\in s_t} w_{mtij}\, y_{mtij} - \theta_{mt} \\ &= \sum\sum_{i,j\in s_t} w_{mtij}\, y_{mtij} - \sum\sum_{i,j\in s_t} w_{mtij}\, \theta_{mt} \\ &= \sum\sum_{i,j\in s_t} w_{mtij}\, (y_{mtij} - \theta_{mt}) \quad . \end{aligned} \tag{7.51}$$

Using (7.49) and (7.51) it follows that

$$e_{mt} = \sum\sum_{i,j\in s_t} w_{mtij}\,(A_{mti} + B_{mtij}) = \sum_{i\in s_t} w_{mti} A_{mti} + \sum\sum_{i,j\in s_t} w_{mtij} B_{mtij} \ ,$$

with $w_{mti} = \sum_{j\in s_t} w_{mtij}$. Hence,

$$\begin{aligned} & COV(e_{m,t-h}, e_{lt} \,|\, \theta_{t-h}, \theta_t) \\ &= COV\left[\sum_{i\in s_{t-h}} w_{m,t-h,i} A_{m,t-h,i} + \sum\sum_{i,j\in s_{t-h}} w_{m,t-h,ij} B_{m,t-h,ij}, \sum_{i\in s_t} w_{lti} A_{lti} + \sum\sum_{i,j\in s_t} w_{ltij} B_{ltij} \right] \\ &= COV\Big(\sum_{i\in s_{t-h}} w_{m,t-h,i} A_{m,t-h,i}, \sum_{i\in s_t} w_{lti} A_{lti}\Big) \\ &+ COV\Big(\sum_{i\in s_{t-h}} w_{m,t-h,i} A_{m,t-h,i}, \sum\sum_{i,j\in s_t} w_{ltij} B_{ltij}\Big) \\ &+ COV\Big(\sum\sum_{i,j\in s_{t-h}} w_{m,t-h,ij} B_{m,t-h,ij}, \sum_{i\in s_t} w_{lti} A_{lti}\Big) \\ &+ COV\Big(\sum\sum_{i,j\in s_{t-h}} w_{m,t-h,ij} B_{m,t-h,ij}, \sum\sum_{i,j\in s_t} w_{ltij} B_{ltij}\Big). \end{aligned}$$

Then

$$COV(e_{m,t-h}, e_{lt} \mid \theta_{t-h}, \theta_t)$$

$$\begin{aligned}
&= \sum_{i \in s_t \bigcup s_{t-h}} w_{m,t-h,i} w_{lti}\, COV(A_{m,t-h,i}, A_{lti}) \\
&+ \sum_{i \in s_t \bigcup s_{t-h}} w_{m,t-h,i} \sum_{j \in s_t} w_{ltij}\, COV(A_{m,t-h,i}, B_{ltij}) \\
&+ \sum_{i \in s_t \bigcup s_{t-h}} \sum w_{lti} \sum_{j \in s_{t-h}} w_{m,t-h,ij}\, COV(B_{m,t-h,ij}, A_{lti}) \\
&+ \sum_{i,j \in s_t \bigcup s_{t-h}} \sum w_{m,t-h,ij} w_{ltij}\, COV(B_{m,t-h,ij}, B_{ltij}) \ .
\end{aligned}$$

Using the assumptions (7.50) it follows that

$$\begin{aligned}
&COV(e_{m,t-h}, e_{lt)} \mid \theta_{t-h}, \theta_t) = \{\boldsymbol{\Gamma}_{e|\theta}(h)\}_{ml} = \\
&\gamma_{mla}(h) \sum_{i \in s_t \bigcap s_{t-h}} w_{m,t-h,i}\, w_{lti} + \gamma_{mlb}(h) \sum_{i,j \in s_t \bigcap s_{t-h}} \sum w_{m,t-h,ij} w_{ltij} \quad .
\end{aligned} \tag{7.52}$$

Let

$$V(e_{mt} \mid \theta_t) = \gamma_{ma}(0) \sum_{i \in s_t} w_{mti}^2 + \gamma_{mb}(0) \sum_{i,j \in s_t} \sum w_{mtij}^2 = S_m^2 \quad \forall\ t.$$

and

$$\boldsymbol{D}_{e|\theta} = diag(S_1^2, S_2^2, \ldots, S_M^2) \quad . \tag{7.53}$$

The cross-correlation matrix of the sampling error process given $\theta_t, \theta_{t-1}, \theta_{t-2}, \ldots$ is defined as

$$\mathbf{P}_{e|\theta}(h) = \boldsymbol{D}_{e|\theta}^{-1/2}\, \boldsymbol{\Gamma}_{e|\theta}(h)\, \boldsymbol{D}_{e|\theta}^{-1/2} \quad . \tag{7.54}$$

Substituting (7.52) and (7.53) into (7.54) results in:

$$\{\mathbf{P}_{e|\theta}(h)\}_{ml} = \frac{\gamma_{mla}(h) \sum_{i \in s_t \bigcap s_{t-h}} w_{m,t-h,i}\, w_{lti} + \gamma_{mlb}(h) \sum_{i,j \in s_t \bigcap s_{t-h}} \sum w_{m,t-h,ij} w_{ltij}}{\sqrt{S_m^2 S_l^2}} \quad . \tag{7.55}$$

From (7.55) it becomes clear that if the samples do not overlap for $h > q$, then $s_t \cap s_{t-h} = \{\varnothing\} \;\; \forall h = q, q+1, \ldots$ and $\mathbf{P}_{e|\theta}(h) = \mathbf{0}$. Therefore, as it was suggested for single-stage surveys, the sampling error process can be modelled by a vector MA(q).

When independent samples of units are drawn from the selected psu's on each occasion one gets

$$\sum\sum_{i,j \in s_t \cap s_{t-h}} w_{m,t-h,ij}\, w_{ltij} = 0 \quad \forall h > 0 \quad ,$$

and

$$\{\mathbf{P}_{e|\theta}(h)\}_{ml} = \frac{\gamma_{mla}(h) \sum_{i \in s_t \cap s_{t-h}} w_{m,t-hi}\, w_{lti}}{\sqrt{S_m^2 S_l^2}} \quad . \qquad \textbf{(7.56)}$$

If the same weights are used for all M characteristics, say $w_{mti} = w_{lti} = w_{ti} \quad \forall m, l = 1, \ldots, M$, equation (7.56) becomes

$$\mathbf{P}_{e|\theta}(h) = \sum\sum_{i,j \in s_t \cap s_{t-h}} w_{t-h,i}\, w_{ti}\, \mathbf{P}(h) \quad , \qquad \textbf{(7.57)}$$

with

$$\{\mathbf{P}(h)\}_{ml} = \frac{\gamma_{mla}(h)}{\sqrt{S_m^2 S_l^2}} \quad .$$

If, in addition, the proportion of overlap as well as the weights are held fixed for all survey rounds the cross-correlation matrix in (7.57) takes the form

$$\mathbf{P}_{e|\theta}(h) \propto C\, \mathbf{P}(h) \quad .$$

Assuming

$$\mathbf{P}(h) \propto \mathbf{P}^h \tag{7.58}$$

one gets

$$\mathbf{P}_{e|\theta}(h) \propto C\,\mathbf{P}^h \quad .$$

Based on the above simplifying assumptions a VAR(1) can be used to represent the sampling error process in a two-stage survey with no overlap between ssu and constant (partial or complete) overlap between psu's.

Although the assumption in (7.58) is a bit restrictive, it relates quite well to the theory introduced by Scott, Smith and Jones(1977) for the univariate case. Regarding the weights, the above assumptions require an estimation procedure in which no adjustment for non-response or any sort of calibration is employed.

7.4 The Compositional Case

Let $\boldsymbol{y}_t = (y_{1t}, \dots, y_{M+1,t})'$ be a vector of sample estimates belonging to the Simplex S^M as defined in Section 6.2. Since each of its components is subject to sampling errors recall that y_{mt} can be decomposed into signal and noise as

$$y_{mt} = \theta_{mt} + e_{mt} \quad , \quad m = 1, \dots, M+1 \quad , \tag{7.59}$$

where θ_{mt} is the unknown population proportion assumed to follow a time series model, and e_{mt} is the sampling error. Considering the $M+1$ series simultaneously, (7.59) can be written in vector form as:

$$\boldsymbol{y}_t = \boldsymbol{\theta}_t + \boldsymbol{e}_t \quad , \tag{7.60}$$

where $\boldsymbol{\theta}_t = (\theta_{1t}, \dots, \theta_{M+1,t})'$ and $\boldsymbol{e}_t = (e_{1t}, \dots, e_{M+1,t})'$. In addition, it is assumed that

$$\sum_{m=1}^{M+1} \theta_{mt} = \sum_{m=1}^{M+1} y_{mt} \quad , \tag{7.61}$$

which implies that

$$\sum_{m=1}^{M+1} e_{mt} = 0 \quad , \quad \forall\, t \quad . \tag{7.61}$$

The model in (7.59) can be rewritten as

$$y_{mt} = \theta_{mt} \left[1 + \frac{e_{mt}}{\theta_{mt}} \right] = \theta_{mt} u_{mt} \quad , \quad m = 1, \ldots, M+1 \quad . \tag{7.62}$$

Applying the additive logratio transformation to the vector y_t with components given in (7.62) produces a transformed series $v_t = a_M(y_t) = (v_{1t}, \ldots, v_{Mt})'$ defined on $\mathbb{R}^M$ which has as its mth component:

$$\begin{aligned} v_{mt} &= \log\left[\frac{y_{mt}}{y_{M+1,t}} \right] = \log\left[\frac{\theta_{mt} u_{mt}}{\theta_{M+1,t} u_{M+1,t}} \right] \\ &= \log\left[\frac{\theta_{mt}}{\theta_{M+1,t}} \right] + \log\left[\frac{u_{mt}}{u_{M+1,t}} \right] \quad , \quad m = 1, \ldots, M \quad . \end{aligned} \tag{7.63}$$

The sequence of vectors $\{v_t\}$, $t = 1, 2, \ldots$ is a multivariate time series in $\mathbb{R}^M$. From (7.63), a vector model for the transformed series $\{v_t\}$ can be written as:

$$v_t = \theta_t^* + e_t^* \quad , \tag{7.64}$$

with $\theta_t^* = (\theta_{1t}^*, \ldots, \theta_{Mt}^*)'$, $e_t^* = (e_{1t}^*, \ldots, e_{Mt}^*)'$, where $\theta_{mt}^* = \log\left(\theta_{mt} \,/\, \theta_{M+1,t}\right)$ and $e_{mt}^* = \log(u_{mt}/u_{M+1,t})$, for $m = 1, \ldots, M$.

Assuming n_t large, recall from Chapter 6 the approximation:

$$\log(u_{mt}) = \log\left[1 + \frac{e_t}{\theta_t} \right] \approx \tilde{u}_{mt} \quad , \tag{7.65}$$

where $\tilde{u}_{mt} = \dfrac{e_{mt}}{\theta_{mt}}$, which are the relative sampling errors.

Hence $e^{*}_{mt} = \log(u_{mt}) - \log(u_{M+1,t}) = \tilde{u}_{mt} - \tilde{u}_{M+1,t}$ can be interpreted as a contrast of the relative sampling errors. Substituting (7.65) into (7.63) results in,

$$v_{mt} = \log\left[\frac{y_{mt}}{y_{M+1,t}}\right] \approx \log\left[\frac{\theta_{mt}}{\theta_{M+1,t}}\right] + (\tilde{u}_{mt} - \tilde{u}_{M+1,t}) \quad . \tag{7.66}$$

A time series model to describe the transformed survey data $\{v_t\}$ must incorporate time series models for both $\{\theta^{*}_{t}\}$ and $\{e^{*}{}_{t}\}$. Hence, the analyst faces once again the problem of modelling the sampling error process. The issue here is how to estimate the cross-correlation matrix function of $\{e^{*}_{t}\}$ condition or not on $\{\theta^{*}_{t}\}$ (if one is working with the quantitative approach).

It is important to note that expression (7.64) for the transformed series $\{v_t\}$ is of the same form as expression (7.60) for the original series $\{y_t\}$. Consequently a design-based estimate for $\Gamma_{e\cdot}(h)$ is denoted as

$$\Gamma_{e\cdot|\theta\cdot}(h) = \begin{bmatrix} COV(e^{*}_{1,t-h}, e^{*}_{1t} \,|\, \theta^{*}_{t-h}, \theta^{*}_{t}) & \cdots\cdots & COV(e^{*}_{1,t-h}, e^{*}_{Mt} \,|\, \theta^{*}_{t-h}, \theta^{*}_{t}) \\ \vdots & \vdots\;\vdots & \vdots \\ \vdots & \vdots\;\vdots & \vdots \\ COV(e^{*}_{M,t-h}, e^{*}_{1t} \,|\, \theta^{*}_{t-h}, \theta^{*}_{t}) & \cdots\cdots & COV(e^{*}_{M,t-h}, e^{*}_{Mt} \,|\, \theta^{*}_{t-h}, \theta^{*}_{t}) \end{bmatrix} ,$$

which can be estimated via $\Gamma_{v|\theta\cdot}(h)$, the sampling variances and covariances of the observed transformed series.

Although, in principle, the framework introduced in Sections 7.3.1 and 7.3.2 should also apply to the transformed model (7.64), the solution is not so straightforward. First and foremost, the transformed variables v_{mt} $(m = 1, \ldots, M)$ are not defined at the individual level. Hence $\Gamma_{e\cdot|\theta\cdot}(h)$ cannot be estimated via $\Gamma_{v|\theta\cdot}(h)$ using microdata.

The problems regarding the estimation of the covariance (correlation) structure of the sampling errors from compositional time series are discussed in the following sections. Section 7.4.1 presents a quantitative design-based approach for evaluating $\mathbf{P}_{e\cdot}(h)$, whereas in Section 7.4.2 a method based on pseudo errors is introduced.

7.4.1 Quantitative Design-Based Analysis for the Compositional Case

As introduced in Sections 7.2.1 and 7.3.1, one way of estimating the autocorrelation structure of the sampling errors is via design-based methods. In the compositional case, design-based estimates for the cross-correlation matrix function $\mathbf{P}_{e^{\bullet}}(h) = D_{e^{\bullet}}^{-1/2}\,\mathbf{\Gamma}_{e^{\bullet}}(h)\,D_{e^{\bullet}}^{-1/2}$ can be obtained via $\mathbf{P}_{e^{\bullet}|\theta^{\bullet}}(h) = D_{e^{\bullet}|\theta^{\bullet}}^{-1/2}\,\mathbf{\Gamma}_{e^{\bullet}|\theta^{\bullet}}(h)\,D_{e^{\bullet}|\theta^{\bullet}}^{-1/2}$, where

$$\{\mathbf{\Gamma}_{e^{\bullet}|\theta^{\bullet}}(h)\}_{ml} = COV(e^{*}_{m,t-h}, e^{*}_{lt)}\,|\,\theta^{*}_{t-h}, \theta^{*}_{t}\,) = COV(e^{*}_{m,t-h}, e^{*}_{lt}\,|\,\theta_{t-h}, \theta_{t})$$

and

$$D_{e^{\bullet}|\theta^{\bullet}} = diag[\,V(e^{*}_{1t}\,|\,\theta_t), \dots, V(e^{*}_{Mt}\,|\,\theta_t)] \quad .$$

Using (7.66) it follows that $COV(e^{*}_{m,t-h}, e^{*}_{lt}\,|\,\theta_{t-h}, \theta_t)$ is given by

$$COV\left(\frac{e_{m,t-h}}{\theta_{m,t-h}} - \frac{e_{M+1,t-h}}{\theta_{M+1,t-h}}\,,\,\frac{e_{lt}}{\theta_{lt}} - \frac{e_{M+1,t}}{\theta_{M+1,t}}\,\middle|\,\theta_{t-h}, \theta_t\right)$$

$$\begin{aligned}
&= \frac{1}{\theta_{m,t-h}\,\theta_{lt}}\;COV(e_{m,t-h}, e_{lt}|\theta_{t-h}, \theta_t)\\
&\quad - \frac{1}{\theta_{m,t-h}\,\theta_{M+1,t}}\;COV(e_{m,t-h}, e_{M+1,t}|\theta_{t-h}, \theta_t)\\
&\quad - \frac{1}{\theta_{M+1,t-h}\,\theta_{lt}}\;COV(e_{M+1,t-h}, e_{lt}|\theta_{t-h}, \theta_t)\\
&\quad + \frac{1}{\theta_{M+1,t-h}\,\theta_{M+1,t}}\;COV(e_{M+1,t-h}, e_{M+1,t}|\theta_{t-h}, \theta_t)\;.
\end{aligned} \qquad \mathbf{(7.67)}$$

Using the model formulation in (7.60) one could be tempted to estimate the covariance components $COV(e_{m,t-h}, e_{lt}\,|\,\theta_{t-h}, \theta_t)$, for $m, l = 1, \dots, M \;\; \forall\; t, h$ in (7.67) via $COV(y_{m,t-h}, y_{lt}\,|\,\theta_{t-h}, \theta_t)$. However, because $\{y_t\}$ is a time series of compositions these covariances, known as "crude covariances" (Aitchinson 1986, p.52), give rise to values which are difficult to interpret. This happens not only because of the unity-sum constraint which is inherent of any composition but also because its components are ratios with common denominators and such that, each ratio has common elements in its numerator and denominator.

As pointed out by Aitchison(1986, p.64), "All the difficulties arising in the traditional approach to covariances and correlations between components of compositions come from a lack of appreciation that to carry over ideas which are highly successful for one particular space, such as $\mathbb{R}^M$, into another very different sample space, namely S^M , may be completely inappropriate. ... the adoption of a crude covariance structure causes more confusion about the nature of compositional variability than it removes". For a detailed discussion regarding the difficulties of interpreting the "crude covariance structure" of compositional data refer to Aitchison(1986, p.52-58).

Note that, for the compositional case, it follows from (7.61) that sampling errors e_{mt} are subject to a zero-sum constraint. Hence, the covariances in (7.67) are also subject to some of the interpretation difficulties pointed out by Aitchison(1986) for compositions. For example, since

$$COV(e_{1t}, e_{1t} + ... + e_{M+1,t}) = 0 \quad ,$$

then

$$COV(e_{1t}, e_{2t}) + ... + COV(e_{1t}, e_{M+1,t}) = -V(e_{1t}) \quad .$$

Therefore the approximation in (7.66), which is generally used when modelling series of log-transformed survey data (as in Section 7.2.1), is not very useful in the compositional case. In fact, it forces the analysis back to the original Simplex space which incorporates all the unwanted constraints. Hence, the standard procedure of evaluating the autocovariance structure of the sampling errors using the design-based covariances of the observed (compositional) series of survey estimates is not recommended in this case due to the well known pitfalls of applying standard statistical procedures to compositional data.

Furthermore, any attempt to estimate $\boldsymbol{\Gamma}_{e^{\cdot}|\theta^{\cdot}}(h)$ via

$$\{\boldsymbol{\Gamma}_{v|\theta^{\cdot}}\}_{ml} = COV(\log y_{m,t-h} - \log y_{M+1,t-h}, \log y_{lt} - \log y_{M+1,t} | \theta^{*}_{t-h}, \theta^{*}_{t}) \tag{7.68}$$

would also be quite difficult to interpret because the estimation of such covariances by using Taylor series methods (see Wolter, 1985, Chapter 6) would also require the computation of $COV(y_{m,t-h}, y_{lt} | \theta_{t-h}, \theta_{t})$ for $m=1,...,M$ $\forall\ t,h$. It is important to emphasize that this thesis deals with survey estimates, so that the basis $\{w_t\}$ and the respective compositions $\{y_t\}$ are defined at the "aggregate" level, instead of at the "unit" level as in Aitchison(1986).

Consider the case of a two-stage labour force survey in which enumeration areas are the primary sampling units, households are the secondary sampling units and the residents are the analysis units. In this case w_{mt} is an estimate of a total in a subdomain. If w_{mt} can be written as a linear estimator in which the weights are attached to the secondary sampling units and if the weights are the same for all the characteristics, it follows that

$$w_{mt} = \sum\sum_{i,j \in s_t} a_{tij}\, w^{*}_{mtij} \quad , \tag{7.69}$$

with $\sum\sum_{i,j \in s_t} a_{tij} = 1$, where w^{*}_{mtij} is the total number of residents, in the sample of time t , in the j^th^ household of the i^th^ enumeration area in the m^th^ domain. Accordingly,

$$y_{mt} = \frac{\sum\sum_{i,j \in s_t} a_{tij}\, w^{*}_{mtij}}{\sum\sum_{i,j \in s_t} a_{tij} \sum_{m=1}^{3} w^{*}_{mtij}} = \frac{w_{mt}}{w_{1t} + w_{2t} + w_{3t}} \quad , \quad m=1,2,3 \quad . \tag{7.70}$$

Therefore, for the scope of this thesis, the basis $\{w_t\}$ are vectors of estimated totals. As a consequence, some of the solutions proposed by Aitchison(1986) to overcome the "covariance difficulties" of compositional data are not directly applicable in this sample survey environment.

One case of particular interest is the concept of correlation based on the covariances of logratios introduced by Aitchison(1986, p.77). He suggested the following logratio covariance matrix to determine the covariance structure of a composition.

Definition 7.1

For a $M+1-$ part composition x , the $M\times M$ matrix

$$\{\Sigma\}_{ml} = COV[\log(x_m/x_{M+1}) \ , \ \log(x_l/x_{M+1})] \quad , \tag{7.71}$$

where $m,l=1,...,M$, $x_m>0 \ \forall \ m$ and $\sum_{m=1}^{M+1} x_M = 1$, is termed *the logratio covariance matrix of a composition.*

Although the expression in (7.68) is also a logratio covariance, y_{mt} (in 7.70) is not defined at the individual level and, moreover,

$$\log w_{mt} = \log \sum_{i,j\in s_t}\sum a_{tij}\, w^*_{mtij} \neq \sum_{i,j\in s_t}\sum a_{tij}\log w^*_{mtij} \quad .$$

However, Aitchison(1986, p.86) also provided the following definition for the covariance matrix of a basis.

Definition 7.2

The covariance structure of a $M+1-$ part basis w is the $(M+1)\times(M+1)$ covariance matrix Ω of the vector $\ln w$, such that

$$\{\Omega\}_{ml} = COV(\log w_m, \log w_l) \quad m,l=1,...,M+1 \quad . \tag{7.72}$$

Aitchison(1986, p.86) pointed out that the above covariance structure is perfectly compatible with the covariance structure for compositions defined in (7.71). In fact $\Sigma = \mathbf{A}\,\Omega\,\mathbf{A}'$, where $\mathbf{A} = [I_{M\times M} \vdots \mathbf{1}_{M\times 1}]$.

Note that v_{mt} (in 7.63) can be alternatively expressed as

$$v_{mt} = \log(y_{mt}/y_{M+1,t}) = \log(w_{mt}/w_{M+1,t}) = \log w_{mt} - \log w_{M+1,t} \quad .$$

Hence the covariance in (7.68) can be expressed as a basis logratio covariance

$$\{\Gamma_{v|\theta^*}\}_{ml} = COV(\log w_{m,t-h} - \log w_{M+1,t-h}, \log w_{lt} - \log w_{M+1,t} | \,\theta^*_{t-h}, \theta^*_t) \ . \tag{7.73}$$

In a survey framework, the estimated total in each of the $M+1$ subdomains are random variables. In addition, the population total on each occasion is, in general, estimated by the survey (it does not come as fixed value from an external source), thus $\sum_{m=1}^{M+1} w_{mt} \;\; \forall\, t$ are also random variables. Therefore, there are no particular constraints regarding the components of $w_t = (w_{1t}, \ldots, w_{M+1,t})'$ or w_t itself. Consequently, the covariance in (7.68) can be estimated via (7.73) using Taylor linearization.

From (7.73) we get that

$$\begin{aligned}
\{\boldsymbol{\Gamma}_{v|\theta^*}\}_{ml} &= COV(v_{m,t-h}, v_{lt} \,|\, \theta^*_{t-h}, \theta^*_t) \\
&= COV(\log w_{m,t-h} - \log w_{M+1,t-h}, \log w_{lt} - \log w_{M+1,t} \,|\, \theta^*_{t-h}, \theta^*_t) \\
&= COV(\log w_{m,t-h}, \log w_{lt} \,|\, \theta^*_{t-h}, \theta^*_t) \\
&\quad - COV(\log w_{m,t-h}, \log w_{M+1,t} \,|\, \theta^*_{t-h}, \theta^*_t) \\
&\quad - COV(\log w_{M+1,t-h}, \log w_{lt} \,|\, \theta^*_{t-h}, \theta^*_t) \\
&\quad + COV(\log w_{M+1,t-h}, \log w_{M+1,t} \,|\, \theta^*_{t-h}, \theta^*_t) \;.
\end{aligned} \tag{7.74}$$

In addition, from (7.64) it follows that $\boldsymbol{\Gamma}_{e^*|\theta^*}(h) = \boldsymbol{\Gamma}_{v|\theta^*}(h) = \boldsymbol{\Gamma}_{v|\theta}(h)$ and consequently

$$\mathbf{P}_{e^*|\theta^*}(h) = \mathbf{P}_{v|\theta}(h) = \boldsymbol{D}^{-1/2}_{v|\theta} \, \boldsymbol{\Gamma}_{v|\theta}(h) \, \boldsymbol{D}^{-1/2}_{v|\theta} \quad ,$$

with

$$\boldsymbol{D}_{v|\theta} = diag[\, V(\log w_{1t} - \log w_{M+1,t} \,|\, \theta_t), \ldots, V(\log w_{Mt} - \log w_{M+1,t} \,|\, \theta_t)] \quad .$$

Thus an estimate of the cross-correlation matrix function of $\{e^*\}$ can be computed using

$$COV(\log w_{m,t-h}, \log w_{lt} \,|\, \theta_{t-h}, \theta_t) \quad \forall \quad \begin{cases} t\;; \\ h=0,1,2,\ldots\;; \\ m,l=1,\ldots,M+1\;. \end{cases} \tag{7.75}$$

The covariances in (7.75) can be obtained via design-based methods using Taylor linearization (see Wolter, 1985, p.225-227) as follows.

Let $W^{(t,h)} = (W_{1,t-h}, \dots, W_{M+1,t-h}, W_{1t}, \dots, W_{M+1,t})'$ be a vector of population totals and let $w_t = (w_{1,t-h}, \dots, w_{M+1,t-h}, w_{1t}, \dots, w_{M+1,t})'$ denote the corresponding vector of estimators based on the sample. Define

$$\begin{aligned} G(W^{(t,h)}\ ) &= (\log W_{1,t-h}, \dots, \log W_{M+1,t-h}, \log W_{1t}, \dots, \log W_{M+1,t})' \\ &= [g_1(W^{(t,h)}), g_2(W^{(t,h)}), \dots, g_{2M+2}(W^{(t,h)})]' \quad , \end{aligned}$$

as the parameter of interest which can be estimated by

$$\begin{aligned} G(w^{(t,h)}\ ) &= (\log w_{1,t-h}, \dots, \log w_{M+1,t-h}, \log w_{1t}, \dots, \log w_{M+1,t})' \\ &= [g_1(w^{(t,h)}), g_2(w^{(t,h)}), \dots, g_{2M+2}(w^{(t,h)})]' \quad . \end{aligned}$$

The matrix of mean squared errors is given approximately by (Wolter, 1985, p.226):

$$E\{[G(w^{(t,h)}) - G(W^{(t,h)})][G(w^{(t,h)}) - G(W^{(t,h)})]'\} \approx \Lambda_t \Gamma_{w|\theta}^{(t,h)} \Lambda_t' \ , \tag{7.76}$$

where $\{\Gamma_{w|\theta}^{(t,h)}\}_{ij} = COV(w_i, w_j \,|\, \theta_{t-h}, \theta_t)$, $i,j = (1,t-h), \dots, (M+1,t)$ and the matrix Λ_t is the Jacobian of the transformation with $\{\Lambda_t\}_{i^*j} = \dfrac{\partial g_{i^*}(W^{(t,h)})}{\partial w_j}$, $i^* = 1, \dots, 2M+2$. An estimator of (7.76) is given by:

$$\hat{V}(G(w^{(t,h)})) = \hat{\Lambda}_t \ \hat{\Gamma}_{w|\theta}^{(t,h)} \ \hat{\Lambda}_t' \quad , \tag{7.77}$$

where

$$\begin{aligned} \{\hat{\Lambda}_t\}_{i^*j} &= \frac{\partial g_{i^*}(w^{(t,h)})}{\partial w_j} \\ &= diag[1/w_{1,t-h}, \dots, 1/w_{M+1,t-h}, \dots, 1/w_{M,t}, 1/w_{M+1,t}] \ , \end{aligned} \tag{7.78}$$

and $\{\hat{\Gamma}_{w|\theta}^{(t,h)}\}_{ij} = C\hat{O}V(w_i, w_j \,|\, \theta_{-h}, \theta_t)$. Substituting the covariances in (7.74) by the values obtained in (7.78) yields an estimate of the cross-covariance matrix of the transformed sampling errors. Note that, in practice, an estimate is computed for each t and h . However, assuming $\{e_t^*\}$ a jointly stationary process, $\Gamma_{e^*}(h)$ depends on the lag h but not on t . For this reason we can average over t the estimated cross-covariances matrices of lag h to obtain an estimate of $\Gamma_{e^*}(h)$ (and also $\mathbf{P}_{e^*}(h)$).

Hence, by representing the transformed variables v_{mt} as logratios of the basis components, a design-based procedure is available for estimating the correlation structure of the sampling errors. In addition, based on the estimated cross-covariance matrix of the sampling error process, we can also compute estimates of the partial autoregression matrices (as in Section 7.3.1). Finally, using the estimates of the cross-correlation and partial autoregression function of the sampling error process, a multivariate time series model for the $\{e_t^*\}$ process can be identified.

One drawback of this method is the amount of computation it requires. Also, in practice, it may be difficult to link survey microdata over time to estimate sampling covariances directly. Fortunately, when elementary estimates are available, the cross-correlation matrices of the (transformed) noise process can be obtained using (transformed) pseudo errors. In fact, the method introduced in Section 7.3.1 can be adapted to the compositional case as follows.

7.4.2 Estimation of the Noise Cross-Correlation Matrices Using Pseudo Errors in a Compositional Case

Consider the case of a rotating panel survey in which K elementary estimates are available. Let y_t be a vector of estimated proportions subject to a unity-sum constraint. Assume also that the same constraint holds for each vector of elementary estimates $y_t^{(k)}$. In this case, each of its components can be decomposed as

$$y_{mt}^{(k)} = \theta_{mt} + e_{mt}^{(k)} \quad , \quad m = 1, ..., M+1 \quad , \qquad (7.79)$$

where $e_{mt}^{(k)}$ is the k^{th} rotation group sampling error. Considering the series simultaneously, (7.79) can be written in vector form as:

$$y_t^{(k)} = \theta_t + e_t^{(k)} \quad , \qquad (7.80)$$

where $e_t^{(k)} = (e_{1t}^{(k)}, ..., e_{(M+1)t}^{(k)})'$, with

$$\sum_{m=1}^{M+1} y_{mt}^{(k)} = \sum_{m=1}^{M+1} \theta_{mt} \qquad \forall\ t,k \quad ,$$

which implies that

$$\sum_{m=1}^{M+1} e_{mt}^{(k)} = 0, \quad \forall\ t,k \quad .$$

The model in (7.79) can be rewritten as

$$y_{mt}^{(k)} = \theta_{mt} \left[1 + \frac{e_{mt}^{(k)}}{\theta_{mt}} \right] = \theta_{mt} u_{mt}^{(k)} \quad . \tag{7.81}$$

Applying the additive logratio transformation to the vector $y_t^{(k)}$ in (7.79) produces the transformed series $v_t^{(k)} = a_m(y_t^{(k)}) = (v_{1t}^{(k)}, ..., v_{Mt}^{(k)})'$, defined on $\mathbb{R}^M$, which has as its mth component $(m=1,...,M)$:

$$\begin{aligned} v_{mt}^{(k)} &= \log \left[\frac{y_{mt}^{(k)}}{y_{M+1,t}^{(k)}} \right] = \log \left[\frac{\theta_{mt} u_{mt}^{(k)}}{\theta_{M+1,t} u_{M+1,t}} \right] \\ &= \log \left[\frac{\theta_{mt}}{\theta_{M+1,t}} \right] + \log \left[\frac{u_{mt}^{(k)}}{u_{M+1,t}^{(k)}} \right] \quad . \end{aligned} \tag{7.82}$$

From (7.82), a vector model for the k^{th} series of transformed elementary estimates can be written as:

$$v_t^{(k)} = \theta_t^* + e_t^{*(k)} \quad , \tag{7.83}$$

with $e_t^{*(k)} = (e_{1t}^{*(k)}, ..., e_{Mt}^{*(k)})'$ and $e_{mt}^{*(k)} = \log(u_{mt}^{(k)} / u_{M+1,t}^{(k)})$, for $m = 1,...,M$.

From (7.83) it becomes clear that K M-dimensional time series of transformed pseudo errors can be constructed from deviations of the transformed rotation group estimates about their overall mean. Following the same notation as in Section 7.3.1, the transformed pseudo errors for the k^{th} rotation group are defined as:

$$\tilde{\mathbf{e}}_t^{*(k)} = (\tilde{e}_{1t}^{*(k)},\dots,\tilde{e}_{Mt}^{*(k)})' = \mathbf{v}_t^{(k)} - \mathbf{v}_t = (v_{1t}^{(k)} - v_{1t},\dots, v_{Mt}^{(k)} - v_{Mt})' \quad , \tag{7.84}$$

where $\mathbf{v}_t = \frac{1}{K}\sum_{k=1}^{K} \mathbf{v}_t^{(k)}$. Note, in addition, that

$$\begin{aligned}\tilde{\mathbf{e}}_t^{*(k)} &= \mathbf{v}_t^{(k)} - \mathbf{v}_t = \mathbf{v}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} \mathbf{v}_t^{(k)} \\ &= (\mathbf{v}_t^{(k)} - \boldsymbol{\theta}_t^*) - \frac{1}{K}\sum_{k=1}^{K} (\mathbf{v}_t^{(k)} - \boldsymbol{\theta}_t^*) \\ &= \mathbf{e}_t^{*(k)} - \frac{1}{K}\sum_{k=1}^{K} \mathbf{e}_t^{*(k)} = \mathbf{e}_t^{*(k)} - \mathbf{e}_t^* \quad .\end{aligned} \tag{7.85}$$

From (7.83), together with (7.84) and (7.85), it becomes clear that the framework introduced in Section 7.3.1 can also be applied to the transformed model. That is, the cross-correlation matrices of the transformed sampling errors can be obtained by computing the cross-covariances matrices of the transformed pseudo errors using

$$(K^2 - K)\boldsymbol{\Gamma}_e(h) = \sum_{k=1}^{K} \boldsymbol{\Gamma}_{\tilde{e}}^{(k)}(h) \quad , \tag{7.86a}$$

and

$$(K^2 - K)\boldsymbol{D}_{e\cdot} = \sum_{k=1}^{K} \boldsymbol{D}_{\tilde{e}\cdot}^{(k)} \quad . \tag{7.86b}$$

Consequently,

$$\boldsymbol{P}_e(h) = \left[\sum_{k=1}^{K} \boldsymbol{D}_{\tilde{e}\cdot}^{(k)}\right]^{-1/2} \left[\sum_{k=1}^{K} \boldsymbol{\Gamma}_{\tilde{e}}^{(k)}(h)\right] \left[\sum_{k=1}^{K} \boldsymbol{D}_{\tilde{e}\cdot}^{(k)}\right]^{-1/2} \quad . \tag{7.87}$$

It is important to note that the assumptions (items *(i)* and *(ii)* from Section 7.3.1) on which these results are based are still reasonable for the compositional case. This is because it seems sensible to assume that if the rotation groups do not overlap the transformed sampling errors are uncorrelated. It also seems appropriate to assume that the correlation between transformed sampling errors depends on the rotation groups and on the lags, but not on t .

Before proceeding further a final issue must be raised. Note that in (7.84) and (7.85) v_t is defined as $v_t = \frac{1}{K}\sum_{k=1}^{K} v_t^{(k)}$ with $v_{mt} = \frac{1}{K}\sum_{k=1}^{K} v_{mt}^{(k)}$. Using this above definition and expressing v_{mt} in terms of the original proportions, leads to

$$
\begin{aligned}
v_{mt} &= \frac{1}{K}\sum_{k=1}^{K} v_{mt}^{(k)} = \frac{1}{K}\sum_{k=1}^{K}\log\left[\frac{y_{mt}^{(k)}}{y_{M+1,t}^{(k)}}\right] \\
&= \log\prod_{k=1}^{K}\left[\frac{y_{mt}^{(k)}}{y_{M+1,t}^{(k)}}\right]^{\frac{1}{K}} = \log\left[\frac{\left[\prod_{k=1}^{K} y_{mt}^{(k)}\right]^{\frac{1}{K}}}{\left[\prod_{k=1}^{K} y_{M+1,t}^{(k)}\right]^{\frac{1}{K}}}\right] .
\end{aligned}
\tag{7.88}
$$

Hence, from this perspective, the transformed series v_{mt} is a function of the geometric means of the original rotation group estimates of the target proportions.

On the other hand, by assuming that the survey estimator for the original proportions is obtained by averaging the original rotation group estimates (as in Section 7.2.2 and Section 7.3.1), it would follow that

$$
v_{mt} = \log\left[\frac{y_{mt}}{y_{M+1,t}}\right] = \log\left[\frac{\frac{1}{K}\sum_{k=1}^{K} y_{mt}^{(k)}}{\frac{1}{K}\sum_{k=1}^{K} y_{M+1,t}^{(k)}}\right] , \tag{7.89}
$$

which indicates that v_{mt} is a function of the arithmetic mean of the original rotation group estimates of the target proportions.

The differences between (7.88) and (7.89) are expected to be negligible since it is known that the geometric mean is approximately equal to the arithmetic mean when the dispersion is relatively small. If there is no rotation bias, it is reasonable to expect that the differences between the rotation group estimates are modest.

Having addressed the problems of how to model the survey estimates in a compositional framework (see Chapter 6) and how to identify the time series model for the sampling errors, the next step is to test the proposed modelling procedure in practice. Chapter 8 presents the results of an empirical study using compositional data from the Brazilian Labour Force Survey.

8 Modelling Compositional Time Series in the Brazilian Labour Force Survey

This chapter presents the empirical results obtained when fitting a Common Components Multivariate Model to the Brazilian Labour Force Survey data. The overall aim here is to illustrate the usefulness of this modelling procedure in a genuine survey situation.

The computer programs to implement the modelling procedures were written using the Interactive Matrix Language of the SAS System (IML/SAS) and are available in Appendix E.

8.1 The Brazilian Labour Force Survey and the Data Set for The Empirical Work

The Brazilian Labour Force Survey (BLFS) collects monthly information about employment, hours of work, education and wages together with some demographic information. It classifies the survey respondents aged 15 and over into three main groups - in employment, unemployed or economically inactive (not in the labour force) - according to their circumstances in the week prior to the interview, following the International Labour Organization (ILO) definitions.

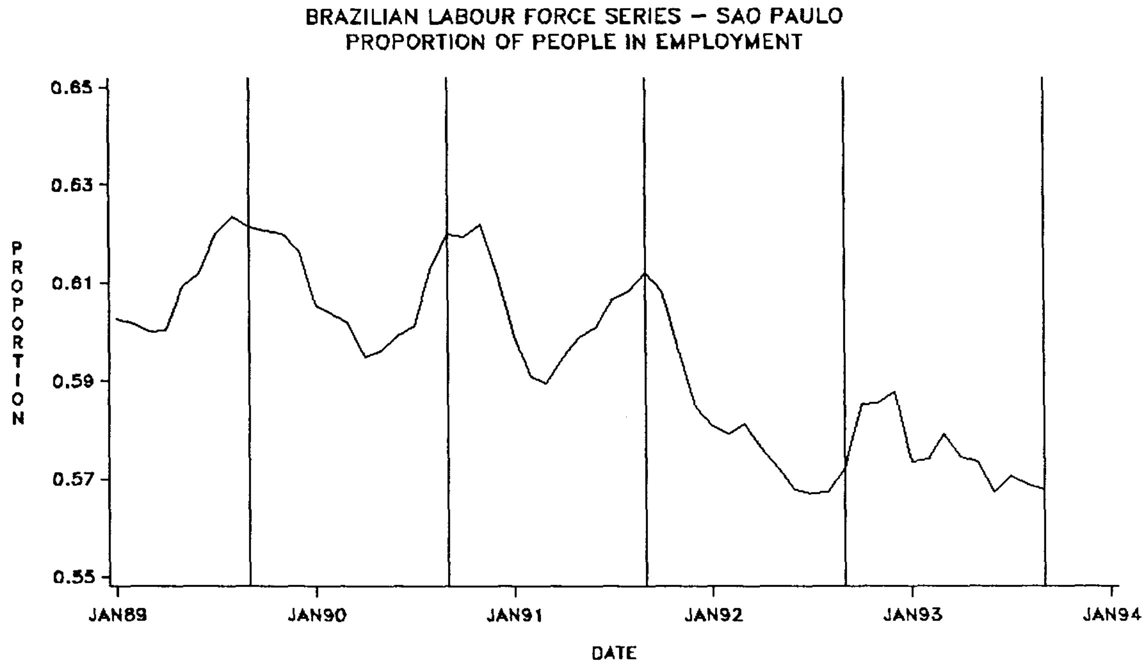
The survey targets the population aged 15 or over, living at one of the six major metropolitan areas in the country. The BLFS is a two-stage sample survey in which the primary sampling units (psu) are the census' enumeration areas and the second-stage units (ssu) are the households, which are completely enumerated. Within each area the primary sampling units are selected with probabilities proportional to their sizes and then a fixed number of households is selected from each psu via systematic sampling. The primary sampling units remain the same for a period of roughly 10 years (as in a Master sample). New primary sampling units are selected when information from a new population census becomes available.

In addition, the BLFS is a rotating panel survey. For any given month the sample is composed of four rotation groups of mutually exclusive sets of primary sampling units. The rotation pattern applies to panels of second-stage units (households). Within each rotation group a panel of households stays in the sample for four consecutive months, is rotated out for the following 8 months and then returns for another four consecutive months. It is then dropped from the survey. Each month one panel is rotated out of the sample. The substitute panel can be a completely new panel or one that is returning after eight months of absence. Note that the 4-8-4 rotation pattern induces a complex correlation structure for the sampling errors over time. The rotation pattern is illustrated in Figure 8.1 where, within a rotation group, common letters denote the same panel of households whereas different letters denote different panels of households selected from the same set of primary sampling units. Note that in any given month 75% of the households are in common with the previous month.

The empirical work was carried out using data from one metropolitan area, São Paulo, covering the period from January 1989 to September 1993 (57 observations). Approximately 6000 households are visited every month in this area. Data for the years before 1989 were not included due to changes in the sample design (the sample size was reduced at the end of 1988). Also, from November 1993 onwards a new sample was implemented based on the results of the 1991 population census. The quantities of interest are the proportions of people classified as employed, unemployed and inactive (as defined in Section 6.2, p.68), and also the unemployment rate (as in (6.41), p.87). Using the monthly individual observations, the series of sample estimates and their respective estimated standard errors were computed using the survey's standard estimators which are based on data of each specific survey round at a time. For each month two sets of estimates were obtained. The direct sample estimates, derived from the complete set of data collected at a given month, and the four elementary estimates, each of them based on data from a single rotation group. The panel estimates will be used to help identify the time series model for the sampling errors. Figures 8.2 to 8.5 display the series of sample estimates.

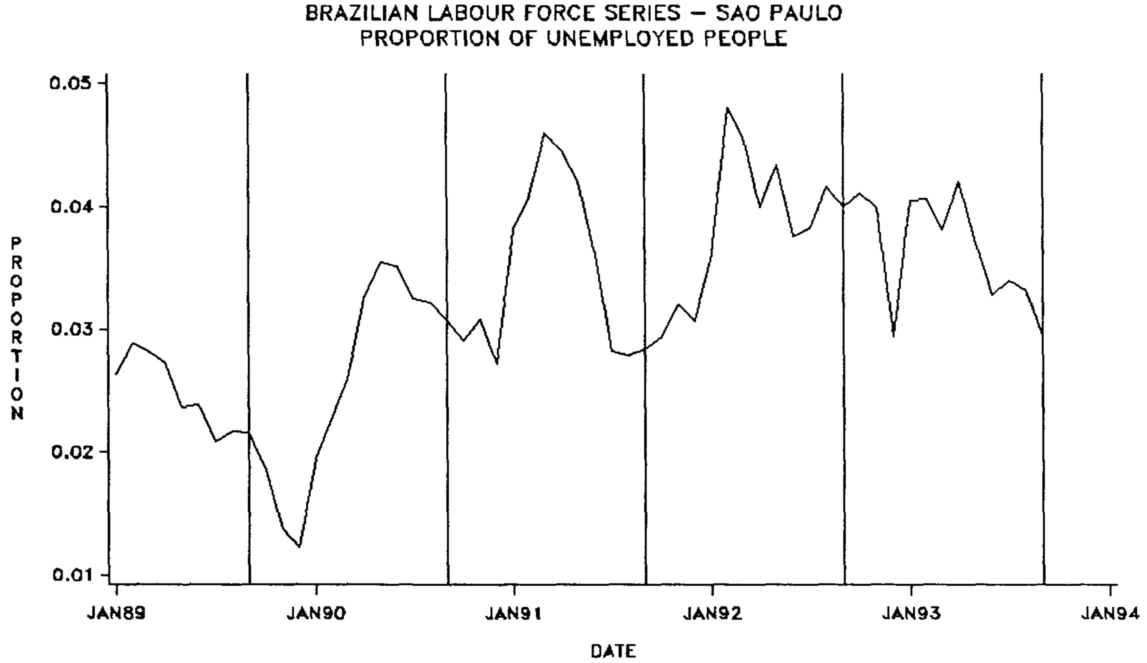
Figure 8.1 - Rotation Pattern of the Brazilian Labour Force Survey

Time/ Month	Rotation Groups			
	1	2	3	4
1	a	b	c	d
2	e	b	c	d
3	e	f	c	d
4	e	f	g	d
5	e	f	g	h
6	i	f	g	h
7	i	j	g	h
8	i	j	k	h
9	i	j	k	l
10	a	j	k	l
11	a	b	k	l
12	a	b	c	l
13	a	b	c	d
14	e	b	c	d
15	e	f	c	d
16	e	f	g	d
17	e	f	g	h
18	i	f	g	h
19	i	j	g	h
20	i	j	k	h
21	i	j	k	l
22	m	j	k	l
23	m	n	k	l
24	m	n	o	l
25	m	n	o	p
26	q	n	o	p
27	q	r	o	p
28	q	r	s	p



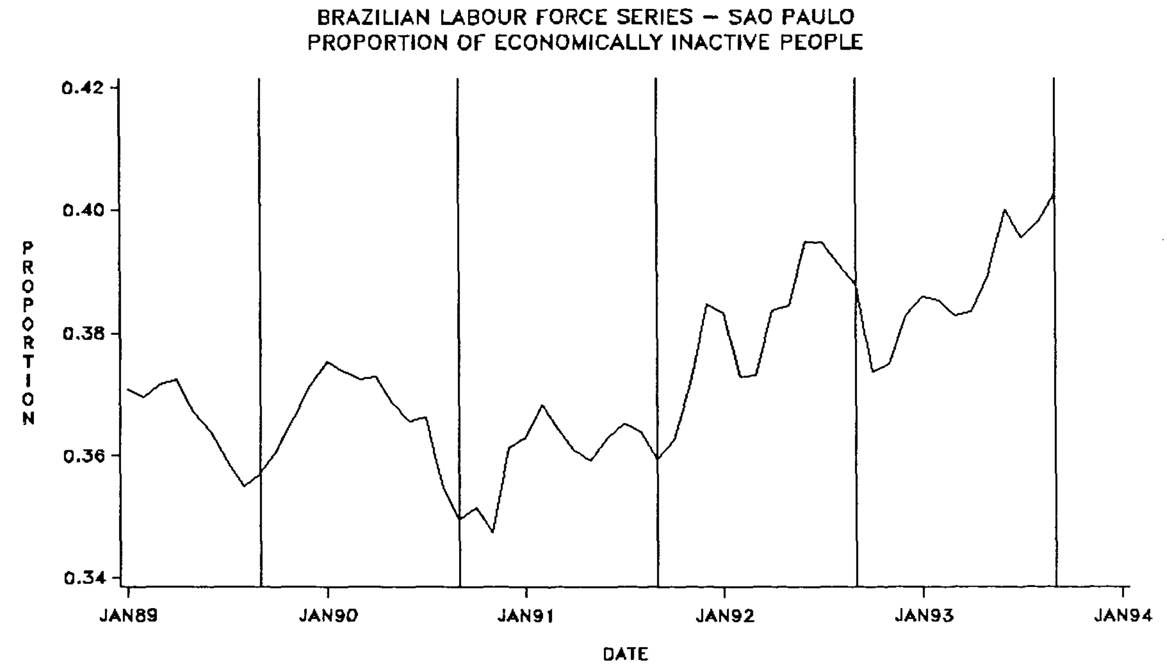
VERTICAL LINES = SEPTEMBER 89 – 93

Figure 8.2



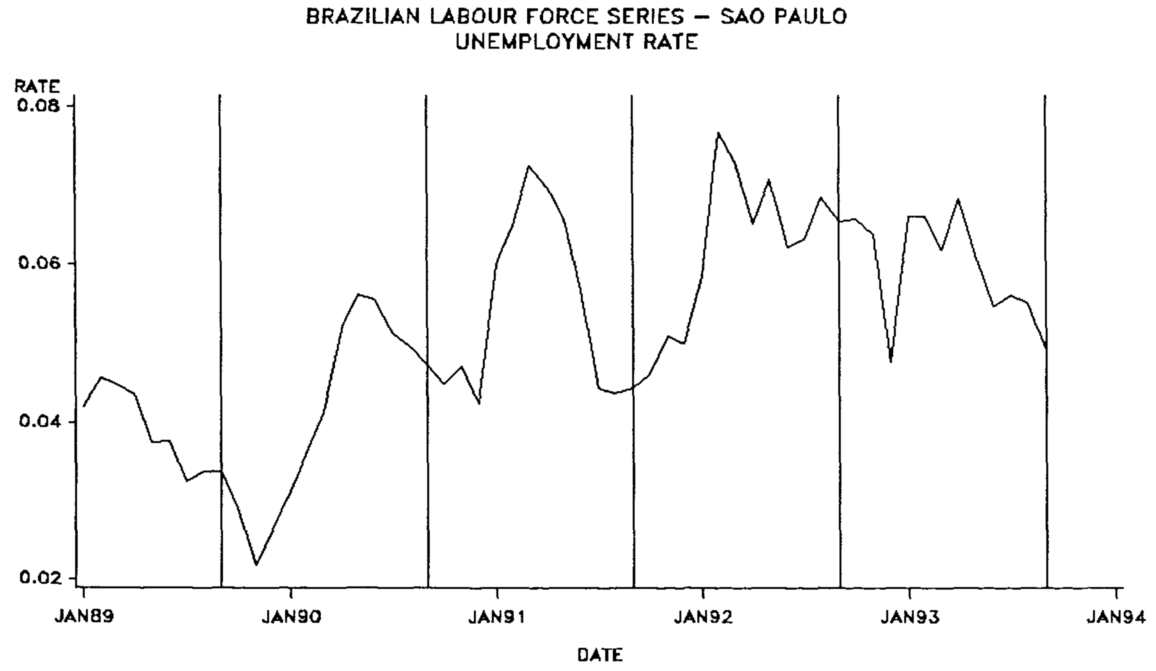
VERTICAL LINES = SEPTEMBER 89 – 93

Figure 8.3



VERTICAL LINES = SEPTEMBER 89 – 93

Figure 8.4



VERTICAL LINES = SEPTEMBER 89 – 93

Figure 8.5

In order to understand the behaviour of the series one has to consider the Brazilian political and economical environment during 1989-1993. In November/December 1989 a presidential election took place. This was the first president elected by direct vote since 1960. At the end of 1989 the inflation rate reached around 80% per month. In March 1990 the new president took over and carried out an economic reform. He stayed in the job for less than three years, being impeached in 1992. In October 1992 the vice-president assumed the presidency. During 1989-1993 different economic reform plans were implemented by different governments and the country's currency changed twice. A comprehensive analysis of the labour market in Brazil is out of the scope of this thesis and the issues raised above were just to illustrate some of the underlying factors affecting the behaviour of the series. In the following section a time series model for compositional data from the Brazilian Labour Force is introduced.

In this study the observed compositional time series is defined as the sequence of vectors

$$y_t = (y_{1t}, y_{2t}, y_{3t})' \quad , \tag{8.1}$$

where:

y_{1t} is the estimated proportion of unemployed people in month t ;

y_{2t} is the estimated proportion of employed people in month t ;

y_{3t} is the estimated proportion of economically inactive people in month t .

The estimated unemployment rate for month t is defined as:

$$r_t = \frac{y_{1t}}{y_{1t} + y_{2t}} \quad . \tag{8.2}$$

8.2 The Modelling Procedure

In a survey environment the principal aim of the modelling procedure is to improve estimation of the unobservable signal and its components. Note that when the time series is obtained from a sample survey it is subject to sampling error. Moreover, in a rotating panel survey such as the BLFS, the sampling errors are autocorrelated. As pointed out by Tiller(1992) and Pfeffermann, Bell & Signorelli(1996), these autocorrelations can induce spurious trends which get confounded with the underlying signal trend when the latter is the one of real interest. When sampling errors are not taken into account, their autocorrelation structure may be absorbed into the irregular term or even into the seasonal or trend components, possibly affecting the inferences about the model.

The model for the BLFS must take into account the special features of the data. First, it is a compositional time series belonging to the Simplex S^2 at each time t . Second, the time series are subject to sampling errors. Following the results provided in Chapter 6, the idea is to map the composition onto $\mathbb{R}^2$ using the additive logratio transformation. Then, the transformed composition is modelled using a multivariate state-space model taking into account the autocorrelation between the sampling errors. That is, the multivariate model for the sample estimates is a combination of the multivariate models for the signal and noise processes. Finally, the model based estimates are transformed back to the original space. Before proceeding further, recall from Chapter 6 the following notation and definitions.

As y_t in (8.1) is a vector of sample estimates, each of its components is subject to sampling errors. Hence, it can be modelled as

$$y_{mt} = \theta_{mt} + e_{mt} \quad , \quad m=1,2,3 \quad , \tag{8.3}$$

where θ_{mt} is the target population quantity and e_{mt} is the sampling error.

Model (8.3) can be expressed in a vector form as:

$$\boldsymbol{y}_t = \boldsymbol{\theta}_t + \boldsymbol{e}_t \quad ,$$

where $\boldsymbol{\theta}_t = (\theta_{1t}, \theta_{2t}, \theta_{3t})'$ and $\boldsymbol{e}_t = (e_{1t}, e_{2t}, e_{3t})'$. Rewriting (8.3) as

$$y_{mt} = \theta_{mt}\left[1 + \frac{e_{mt}}{\theta_{mt}}\right] = \theta_{mt}u_{mt} \quad , \quad m=1,2,3 \quad , \tag{8.4}$$

and applying the additive logratio transformation to the vector $\boldsymbol{y}_t$ in (8.4) results in

$$\begin{aligned} v_{mt} &= \log\left[\frac{y_{mt}}{y_{3t}}\right] = \log\left[\frac{\theta_{mt}u_{mt}}{\theta_{3t}u_{3t}}\right] \\ &= \log\left[\frac{\theta_{mt}}{\theta_{3t}}\right] + \log\left[\frac{u_{mt}}{u_{3t}}\right] \quad , \quad m=1,2 \quad . \end{aligned} \tag{8.5}$$

The vector model for the transformed series is:

$$\boldsymbol{v}_t = \boldsymbol{\theta}_t^* + \boldsymbol{e}_t^* \quad , \tag{8.6}$$

with $\boldsymbol{v}_t = (v_{1t} \ , \ v_{2t})'$, $\boldsymbol{\theta}_t^* = (\theta_{1t}^* \ , \ \theta_{2t}^*)'$, $\boldsymbol{e}_t^* = (e_{1t}^* \ , \ e_{2t}^*)'$, $\theta_{mt}^* = \log(\theta_{mt} \ / \ \theta_{3t})$ and $e_{mt}^* = \log(u_{mt}/u_{3t})$, for $m = 1,2$.

Hence the model for the transformed sample estimates, $\boldsymbol{v}_t$, is composed of a multivariate model for the transformed signal $\boldsymbol{\theta}_t^*$, describing how the transformed population quantities evolve in time, and a multivariate model representing the time series relationship between transformed sampling errors $\boldsymbol{e}_t^*$. Sections 8.2.2 and 8.2.3 present the models for the signal and the noise processes, respectively. Figure 8.6 below displays the series of transformed compositions.

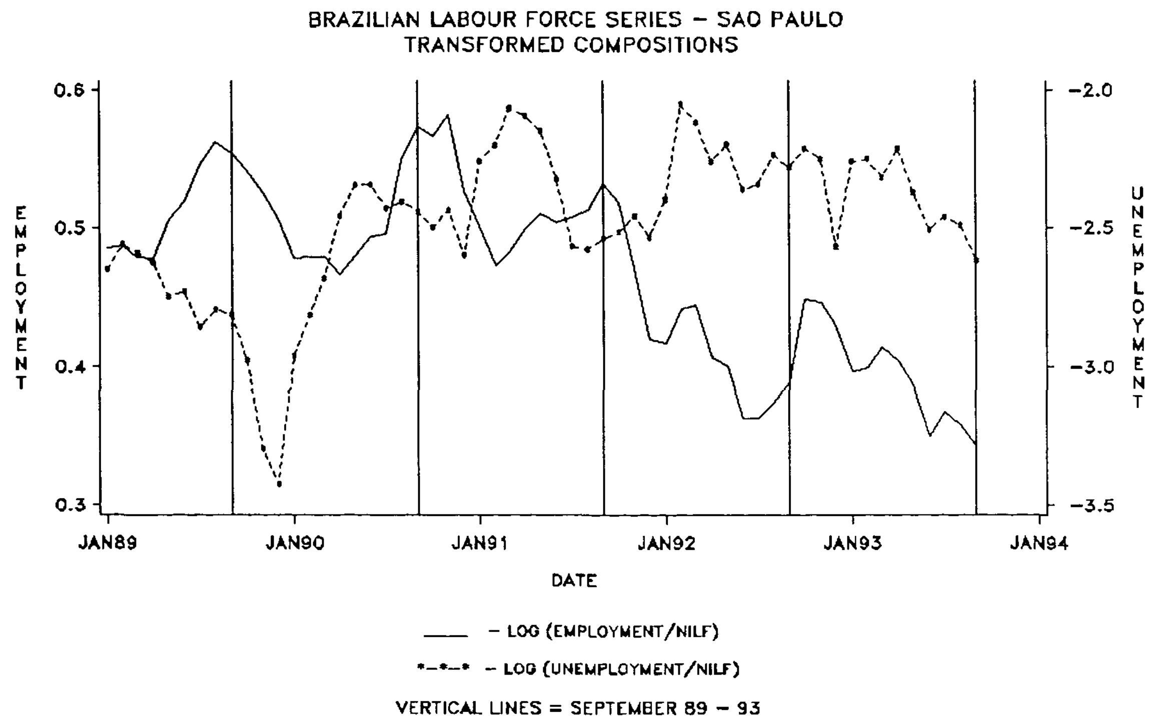


Figure 8.6

8.2.2 The Model for the Transformed Signal θ_t^*

The transformed signal process $\{\theta_t^*\}$ is assumed to follow a multivariate basic structural model (BSM), in which each of $\{\theta_{mt}^*\}$ follows a basic structural time series model, with the model parameters being possibly different across the series. The cross-sectional relationship between the series comes in via the correlation structure of the system disturbances. The model for $\{\theta_{mt}^*\}$, $m=1,2$ is given by

$$\left\{\begin{array}{lcl} \theta_{mt}^* & = & L_{mt}^* + S_{mt}^* \quad , \\ L_{mt}^* & = & L_{m,t-1}^* + R_{m,t-1}^* + \eta_{mt}^{(l)} \quad , \\ R_{mt}^* & = & R_{m,t-1}^* + \eta_{mt}^{(r)} \quad , \\ S_{mt}^* & = & -\sum_{j=1}^{11} S_{m,t-1}^* + \eta_{mt}^{(s)} \quad , \end{array}\right. \tag{8.7}$$

where L^*_{mt} is the trend/level component of the unobservable transformed signal θ^*_{mt} , R^*_{mt} is the corresponding change in the level,and S^*_{mt} is the seasonal component of θ^*_{mt} . The disturbances $\eta^{(l)}_{m}$, $\eta^{(r)}_{mt}$, $\eta^{(s)}_{mt}$ are assumed to be mutually uncorrelated normally distributed with mean zero and variances $\sigma^2_{m_l}$, $\sigma^2_{m_r}$, $\sigma^2_{m_s}$, respectively.

The multivariate model (8.7) for $\{\theta^*_t\}$ has the following state-space formulation:

$$\left\{ \begin{array}{lcl} \theta^*_t & = & H^{(\theta)} \alpha^{(\theta)}_t \quad ; \\ \alpha^{(\theta)}_t & = & T^{(\theta)} \alpha^{(\theta)}_{t-1} + G^{(\theta)} \eta^{(\theta)}_t \quad , \end{array} \right. \tag{8.8}$$

where

$H^{(\theta)} = [1010000000000] \otimes I_2$, $\alpha^{(\theta)}_t = [\, L^*_{1t}\ L^*_{2t}\ R^*_{1t}\ R^*_{2t}\ S^*_{1t}\ S^*_{2t}\ \ldots\ S^*_{1,t-10}\ S^*_{2,t-10}\,]'$,

$\eta^{(\theta)}_t = (\ \eta^{(l)}_{1t}\ \eta^{(l)}_{2t}\ \eta^{(r)}_{1t}\ \eta^{(r)}_{2t}\ \eta^{(s)}_{1t}\ \eta^{(s)}_{2t}\ )'$,

$$G^{(\theta)} = \begin{bmatrix} I_3 \\ \ldots\ \ldots \\ \mathbf{0}_{10\times 3} \end{bmatrix} \otimes I_2$$

and

$$T^{(\theta)} = \left[\begin{array}{cccccccc} 1 & 1 & \vdots & & & \mathbf{0}_{2\times 11} & & \\ 0 & 1 & \vdots & & & & & \\ \ldots & \ldots & \vdots & \ldots & \ldots & \ldots & \ldots & \ldots \\ & & \vdots & -1 & -1 & \ldots & -1 & -1 \\ & & \vdots & 1 & 0 & \ldots & 0 & 0 \\ \mathbf{0}_{11\times 2} & & \vdots & 0 & 1 & \ldots & 0 & 0 \\ & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \vdots & 0 & 0 & \ldots & 1 & 0 \end{array} \right] \otimes I_2 \ .$$

The model equations in (8.8) are supplemented by the cross sectional assumption:

$$\Sigma_\theta = \begin{bmatrix} \Sigma_l & & \mathbf{0} \\ & \Sigma_r & \\ \mathbf{0} & & \Sigma_s \end{bmatrix} ,$$

that is, the two series are linked via the off-diagonal elements of $\Sigma_l, \Sigma_r, \Sigma_s$.

8.2.3 The Model for the Noise Process $\{e_t^*\}$

The use of a vector time series model to represent the sampling error series is a key feature of this modelling procedure. The model identification for multiple time series is usually based on the cross-correlation matrices and partial lag correlation matrices (for details, see Wei,1993, p.356).

The correlation structure of the sampling errors, which are unobservable components, can be estimated using pseudo errors as suggested in Chapter 7. Recall from section 7.4.2 that, when dealing with compositional time series from rotating panel surveys, the cross-covariance matrices for the noise process $\{e_t^*\}$ can be obtained from the cross-covariance matrices of the transformed pseudo errors.

For the Brazilian Labour Force Survey, the estimates for cross-covariances of the transformed sampling errors were obtained using

$$\hat{\Gamma}_{e^*}(h) = \frac{\left[\sum_{k=1}^{K} \hat{\Gamma}_{\tilde{e}^*}^{(k)}(h)\right]}{K^2-K} = \frac{\left[\sum_{k=1}^{4} \hat{\Gamma}_{\tilde{e}^*}^{(k)}(h)\right]}{12} , \qquad \textbf{(8.9)}$$

where $\hat{\Gamma}_{\tilde{e}^*}^{(k)}(h)$, $k=1,\ldots,4$ is the estimated cross-covariance of lag h for the k^{th} transformed pseudo error series.

Estimates of the cross-correlation function and partial lag correlation matrices for $\{e_t^*\}$ were computed, based on the estimates in (8.9), using a recursive algorithm provided in Wei(1993, pp.359-362). This algorithm is a vector generalization of Durbin's (1960) recursive computational procedure for univariate partial autocorrelations.

A program in SAS-IML was developed to compute the estimates and provide the corresponding schematic representations (as in Tiao & Box, 1981). In addition, a statistical test to help establish the order of a suitable vector autoregressive process to represent the noise series was also carried out. Using the results in Wei(1993,p.362), if $\hat{P}_{ij}$ are the elements of the sample partial lag correlation matrices, then under the null hypothesis that $\{e_t^*\}$ is a vector autoregressive process of order $s-1$, $[M \,.\, \hat{P}_{ij}(s)]^2$ is asymptotically distributed as a χ^2 with one degree of freedom. Consequently $X(s) = M \sum_{i=1}^{M} \sum_{j=1}^{M} [\hat{P}_{ij}(s)]^2 \sim \chi^2_{M^2}$. Figures 8.7 and 8.8 provide the estimates for the cross-correlation and partial lag correlation matrices and their respective schematic representation together with the p-values for the statistical test.

The form of the correlation matrices and the results for the statistical test indicate that a VAR(1) can be used to represent the transformed sampling error process. Two other models were considered to represent the noise process, a VAR(2) and a VARMA(1,1).

Care must be taken when using the results of the schematic representations and the χ^2 test statistic $X(s)$ above. The asymptotic expression for the variance of the sample autocorrelations (which is used to construct the schematic representation) is related to the standard estimator of the cross-correlation matrix, namely the sample cross-correlation matrix (see Wei,1993, pp.350), which cannot be computed because the sampling error series is itself unobservable. The estimates used here were obtained from a different estimator, namely that defined in (8.9). Although aware of these potential problems, the estimates obtained here using the transformed pseudo errors provided the only clues for identifying a suitable model for the error process.

Figure 8.7

```
                CROSS-CORRELATION MATRICES

   LAG       1      LAG       2      LAG       3      LAG       4
  0.422 -0.035   0.268 -0.071   0.134  0.098 -0.092 -0.049
  0.044  0.060   0.103 -0.021   0.179  0.040  0.081  0.019


  LAG      5      LAG      6      LAG      7      LAG      8
 -0.138  0.004 -0.063  0.160 -0.054 -0.024 -0.053 -0.072
  0.059 -0.097  0.040 -0.029 -0.043 -0.116 -0.015  0.100


  LAG      9      LAG     10      LAG     11      LAG     12
  0.055 -0.007  0.090 -0.076  0.103 -0.050  0.069  0.055
 -0.053  0.110  0.011 -0.091  0.014  0.058  0.096  0.167


  LAG     13      LAG     14      LAG     15      LAG     16
 -0.017 -0.092 -0.070  0.006 -0.092  0.006 -0.205 -0.047
 -0.064 -0.140  0.064 -0.169  0.020 -0.071  0.031 -0.118


   LAG     17      LAG     18      LAG     19      LAG     20
 -0.171 -0.041 -0.145  0.030 -0.159 -0.050 -0.116 -0.103
 -0.049 -0.023  0.038  0.012 -0.028  0.022 -0.037  0.046


       SCHEMATIC REPRESENTATION OF CROSS-CORRELATIONS

           LAG   1   LAG   2   LAG   3   LAG   4
            +    .    +    .    .    .    .    .
            .    .    .    .    .    .    .    .

           LAG   5   LAG   6   LAG   7   LAG   8
            .    .    .    .    .    .    .    .
            .    .    .    .    .    .    .    .

           LAG   9   LAG  10   LAG  11   LAG  12
            .    .    .    .    .    .    .    .
            .    .    .    .    .    .    .    .

            LAG  13   LAG  14   LAG  15   LAG  16
            .    .    .    .    .    .    .    .
            .    .    .    .    .    .    .    .

            LAG  17   LAG  18   LAG  19   LAG  20
            .    .    .    .    .    .    .    .
            .    .    .    .    .    .    .    .
```

Figure 8.8

```
            PARTIAL LAG CORRELATION MATRICES


 LAG     1     LAG     2     LAG     3     LAG     4
 0.422 -0.035  0.094 -0.050 -0.013  0.152 -0.168 -0.102
 0.044  0.060  0.099 -0.023  0.144  0.048 -0.004  0.032


 LAG     5     LAG     6     LAG     7     LAG     8
-0.054  0.015  0.084  0.164  0.020 -0.134 -0.095 -0.097
 0.012 -0.101  0.013 -0.025 -0.035 -0.103  0.025  0.100


 LAG     9     LAG    10     LAG    11     LAG    12
 0.062  0.045  0.092 -0.015  0.033  0.006 -0.019  0.081
-0.033  0.081  0.079 -0.102 -0.007  0.089  0.088  0.157

SCHEMATIC REPRESENTATION OF THE PARTIAL LAG CORRELATIONS

          LAG  1   LAG  2   LAG  3   LAG  4
           +   .    .   .    .   .    .   .
           .   .    .   .    .   .    .   .

          LAG  5   LAG  6   LAG  7   LAG  8
           .   .    .   .    .   .    .   .
           .   .    .   .    .   .    .   .

          LAG  9   LAG 10   LAG 11   LAG 12
           .   .    .   .    .   .    .   .
           .   .    .   .    .   .    .   .
```

X(S) TO BE COMPARED WITH A χ^2 WITH M^2 DEGREES OF FREEDOM
(M=2 IS THE DIMENSION OF THE SERIES)

```
LAG     X(S)    p_value
 1     10.52    0.0325
 2      1.23    0.8731
 3      2.64    0.6198
 4      2.26    0.6881
 5      0.77    0.9424
 6      1.99    0.7376
 7      1.72    0.7871
 8      1.66    0.7980
 9      0.77    0.9424
10      1.43    0.8390
11      0.51    0.9725
12      2.25    0.6899
```

The parameter estimates for the VAR(1) and VAR(2) models were computed via the same algorithm used to compute the partial lag correlation. For vector moving average models, the model parameters can be obtained using the corresponding cross-covariance function. A simple method for solving the resulting system of equations can be found in Jenkins & Alavi(1981). The parameter estimates for the VARMA(1,1) were computed from the relation between the estimated cross-covariance function and the parameter matrices as in Wei(1993,pp.346-347).

The VAR(1) model fitted to $\{e_t^*\}$ is given by:

$$\begin{bmatrix} e_{1t}^* \\ e_{2t}^* \end{bmatrix} = \begin{bmatrix} 0.4497 & -0.0187 \\ -0.2867 & 0.0773 \end{bmatrix} \begin{bmatrix} e_{1,t-1}^* \\ e_{2,t-1}^* \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} , \qquad \textbf{(8.10a)}$$

with

$$\hat{\Sigma}_a = \begin{bmatrix} 0.0001736 & \mathbf{0.32} \\ 0.0003051 & 0.0052033 \end{bmatrix} , \qquad \textbf{(8.10b)}$$

where the element in bold is the estimated correlation based on the variances and covariance in the lower triangular matrix.

The parameter estimates for the VAR(2) are:

$$\begin{bmatrix} e_{1t}^* \\ e_{2t}^* \end{bmatrix} = \begin{bmatrix} 0.4497 & -0.0187 \\ -0.2867 & 0.0773 \end{bmatrix} \begin{bmatrix} e_{1,t-1}^* \\ e_{2,t-1}^* \end{bmatrix} + \begin{bmatrix} 0.06957 & 0.0140 \\ -0.2617 & -0.0072 \end{bmatrix} \begin{bmatrix} e_{1,t-2}^* \\ e_{2,t-2}^* \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} , \qquad \textbf{(8.11a)}$$

with

$$\hat{\Sigma}_a = \begin{bmatrix} 0.0001739 & \mathbf{0.31} \\ 0.0003001 & 0.0052281 \end{bmatrix} . \qquad \textbf{(8.11b)}$$

Finally, the VARMA(1,1) has the form:

$$\begin{bmatrix} e^*_{1t} \\ e^*_{2t} \end{bmatrix} = \begin{bmatrix} 0.7347 & 0.2414 \\ -0.9224 & -0.2072 \end{bmatrix} \begin{bmatrix} e^*_{1,t-1} \\ e^*_{2,t-1} \end{bmatrix} - \begin{bmatrix} 0.3162 & 0.2590 \\ -0.7666 & -0.2749 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} , \quad \mathbf{(8.12a)}$$

with

$$\hat{\Sigma}_a = \begin{bmatrix} 0.0001723 & \mathbf{0.37} \\ 0.0003476 & 0.0051660 \end{bmatrix} . \quad \mathbf{(8.12b)}$$

Having defined the candidate models for the noise process, the next step is to obtain the model for the survey estimates.

8.2.4 The Model for the Survey Estimates

The model for the transformed survey estimates is a superposition of the models proposed in Sections 8.2.2 and 8.2.3. The general state-space representation of the multivariate model for $\boldsymbol{v}_t$ is given by:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{H}\boldsymbol{\alpha}_t , \\ \boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}\boldsymbol{\eta}_t . \end{cases} \quad \mathbf{(8.13)}$$

The configuration of the system matrices together with the state and disturbance vectors are defined according to the model chosen for the noise process, as presented in Table 8.1. For details on the state-space representation of VARMA models refer to Reinsel(1993, Section 7.2).

Table 8.1 - State-Space Models for the Transformed Estimates

System Matrices	BSM+VAR(1)	BSM+VAR(2)	BSM+VARMA(1,1)
α_t	$(\alpha_t^{(\theta)\prime}, e_t^{*\prime})'$	$(\alpha_t^{(\theta)\prime}, e_t^{*\prime}, e_{t-1}^{*\prime})'$	$(\alpha_t^{(\theta)\prime}, e_t^{*\prime}, (-\Theta a_t)')'$
η_t	$(\eta^{(\theta)\prime}, a_t')'$	$(\eta^{(\theta)\prime}, a_t')'$	$(\eta^{(\theta)\prime}, a_t')'$
H	$[H^{(\theta)}, 1] \otimes I_2$	$[H^{(\theta)}, 1, 0] \otimes I_2$	$[H^{(\theta)}, 1, 0] \otimes I_2$
T	$\left[\begin{array}{ccc} T^{(\theta)} & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \Phi \end{array}\right]$	$\left[\begin{array}{ccccc} T^{(\theta)} & \vdots & \mathbf{0} & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \Phi_1 & \vdots & \Phi_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & I_2 & \vdots & \mathbf{0} \end{array}\right]$	$\left[\begin{array}{ccccc} T^{(\theta)} & \vdots & \mathbf{0} & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \Phi & \vdots & I_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \mathbf{0} & \vdots & \mathbf{0} \end{array}\right]$
G	$\left[\begin{array}{ccc} G^{(\theta)} & \vdots & \mathbf{0} \\ \cdots\cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & I_2 \end{array}\right]$	$\left[\begin{array}{ccc} G^{(\theta)} & \vdots & \mathbf{0} \\ \cdots\cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & I_2 \end{array}\right]$	$\left[\begin{array}{ccc} G^{(\theta)} & \vdots & \mathbf{0} \\ \cdots\cdots & \cdots & \cdots \\ & \vdots & I_2 \\ \mathbf{0} & & \\ & \vdots & -\Theta \end{array}\right]$

For all the above models the covariance matrix of the model disturbances is given by:

$$Q = V(\eta_t) = \left[\begin{array}{ccc} \Sigma_\theta & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \Sigma_a \end{array}\right] .$$

Once the model for the survey estimates has been put in the state-space form, the Kalman Filter equations (from Chapter 3, Section 3.3) are used to get filtered and smoothed estimates for the unobservable components. The application of the Kalman Filter requires the estimation of the unknown hyperparameters (the covariances $\Sigma_l, \Sigma_r, \Sigma_s$) and the estimation of the initial state vector and respective covariance matrix.

8.2.5 The Estimation of the Model and Initialization of the Kalman Filter

Assuming that the disturbances η_t are normally distributed the log-likelihood function of the (transformed) observations can be expressed via the prediction error decomposition (see Chapter 3, Section 3.4 for details). Estimates for the model covariances were obtained by maximum likelihood, applying a quasi-Newton optimization technique. Following Fernandez and Harvey(1990), the constraint that the estimates for the unknown covariance matrices $\mathbf{\Sigma}_*$ are positive semi-definite was implemented by defining lower triangular matrices $\mathbf{A}$ such that $\mathbf{A}\,\mathbf{A}' = \mathbf{\Sigma}_*$ (see Graybill, 1983, pp.208-209) and maximizing the likelihood with respect to the elements of $\mathbf{A}$. From the theory of maximum likelihood estimation it follows that estimates for the original parameters (φ^*) can be readily obtained from the estimates of the derived parameters (φ) applying the inverse transformation to the maximum likelihood estimates (for details see Cramer,1989, pp.31-33). In this case, the numerically evaluated Hessian matrix provides variance estimates for the transformed hyperparameters (the elements of the triangular matrices $\mathbf{A}$). The covariance matrix for the original parameter set was obtained via:

$$V(\hat{\varphi}^*) = J\ V(\hat{\varphi})J' \quad ,$$

where J is the Jacobian of the transformation.

A computer program to implement the maximization procedure was developed using the optimization routine NLPQN from SAS-IML. The Hessian was evaluated numerically using the NLPFDD routine from SAS-IML which computes finite difference approximations for first- and second-order derivatives. The program for Kalman Filter updating, smoothing and prediction is an adaptation of a SAS-IML program kindly provided by Prof. D.Pfeffermann.

The initialization of the Kalman filter was carried out using a diffuse prior. By this approach the non-stationary components $(\alpha^{(\theta)})'$ of the state-vector were initialized with very large error variances and the respective components of the initial state vector were taken as zero. The stationary components $(e_{1t}^{*}\ e_{2t}^{*})'$ were initialized by the corresponding unconditional mean and variance. Following Mittnik(1991), the unconditional state covariance matrix for a VAR(1) is simply its autocovariance matrix $\boldsymbol{P}^{e^*}_{1|0} = \boldsymbol{\Gamma}_{e^*}(0)$ and in the case of a VAR(2) it is given by

$$\boldsymbol{P}^{e^*}_{1|0} = \begin{bmatrix} \boldsymbol{\Gamma}_{e^*}(0) & \boldsymbol{\Gamma}_{e^*}(1) \\ \boldsymbol{\Gamma}'_{e^*}(1) & \boldsymbol{\Gamma}_{e^*}(0) \end{bmatrix} \quad .$$

The unconditional variance for the VARMA(1,1) process was obtained solving the equation $vec\ (\boldsymbol{P}^{e^*}_{1|0}) = [\boldsymbol{I} - \boldsymbol{T}\otimes\boldsymbol{T}]^{-1}\ vec(\ \boldsymbol{G}\boldsymbol{\Sigma}_a\boldsymbol{G}'\ )$, from Harvey(1989,p.121), where

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{I}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{\Theta} \end{bmatrix} \ .$$

Hence the initialization of the Kalman Filter was carried out with:

$$\hat{\alpha}_{1|0} = \mathbf{0}_{30\times 1} \quad , \quad \boldsymbol{P}_{1|0} = \begin{bmatrix} c\boldsymbol{I}_{26} & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \boldsymbol{P}^{e^*}_{1|0} \end{bmatrix} \quad ,$$

where $c = 10^5$ and $\boldsymbol{P}^{e^*}_{1|0}$ was computed using the sample cross-correlation and parameters matrices from (8.10), (8.11) and (8.12). The use of a diffuse prior implies that the first 13 innovations and their associated variances are not included in the computation of the likelihood. Note that $d = 13$ is the number of non-stationary elements in each of the univariate models (for details see Harvey,1989, Chapters 3,4 and 8). This procedure has the advantage of being computationally simple. Different approaches for initializing the Kalman Filter are discussed in Harvey & Peters(1984), Harvey(1989, Sections 3.4.3 and 4.2.2) and De Jong(1988,1989,1991). The estimation procedure was tested by fitting structural models (as in Chapter 3) to some of the tutorial data sets available in STAMP5.0(Koopman et al,1995). The results from STAMP5.0 were used as benchmarks, since it is well-known software designed to analyse time series using univariate and multivariate structural time series models. Note, however, that it does not allow for sampling variation.

The following section presents the results obtained when fitting the candidate models defined in Table 8.1 to the Brazilian Labour Force Survey data.

8.3 Parameter Estimation

The parameter estimates and respective asymptotic standard errors (displayed in parenthesis) for the different state-space models are presented in Table 8.2. The values in bold are the estimated correlations. Note that the estimation of the model for the sampling errors was implemented outside the Kalman Filter as described in Section 8.2.3. As the seasonal and slope covariance matrices are consistently very small they can be omitted from the signal model. This implies that the seasonals are assumed to be deterministic and the slope is assumed to be fixed, giving rise to a local level model with drift and non-stochastic seasonals for the signal. Indeed, as pointed out by Koopman et al(1995, p.39), when the number of years considered in the analysis is small it seems reasonable to fix the seasonals since there is not enough data to allow the estimation of a changing pattern. Besides, the fact that fixed seasonals are the outcome of estimation is a satisfactory feature of the modelling procedure. Models containing irregular terms (for the signal) were also tested. As expected, in the presence of an explicit model for the sampling error, there was no need to include irregular components in the model for the signal. This followed because all the elements of their estimated covariance matrix were small and, consequently, were taken as zero. Regarding the model for the noise process, the choice between a pure autoregressive or a VARMA(1,1) model has little effect on the parameter estimates and goodness-of-fit measures (Table 8.4). Recall from Section 8.2.3 that based on the identification procedure a VAR(1) seemed adequate to represent the noise process. However, it will be seen later in this chapter that a basic structural model (with fixed slope and seasonals) plus a VARMA(1,1) yields smaller standard errors for the unemployment rate estimates when compared with those from a VAR(1) or VAR(2).

Table 8.2 - Estimates for the Hyperparameters and Respective Standard Errors

Model (log-lik)[1]	$\hat{\Sigma}_{l} x\, 10^{-4}$	$\hat{\Sigma}_{r} x\, 10^{-7}$	$\hat{\Sigma}_{s} x\, 10^{-10}$
BSM+AR(1) (200.40)	$\begin{bmatrix} 2.64 & \mathbf{0.10} \\ (0.94) & \\ 1.48 & 85.4 \\ (4.26) & (30.92) \end{bmatrix}$	$\begin{bmatrix} 7.08 & \mathbf{1.00} \\ (19) & \\ 25.75 & 93.61 \\ (73) & (388) \end{bmatrix}$	$\begin{bmatrix} 0.29 & \mathbf{-0.70} \\ (253) & \\ -0.36 & 0.90 \\ (750) & (2000) \end{bmatrix}$
⇒ BSM + AR(1)[2] (200.24)	$\begin{bmatrix} 2.77 & \mathbf{0.13} \\ (0.93) & \\ 2.05 & 87.8 \\ (3.56) & (27.11) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$
BSM+AR(2) (200.28)	$\begin{bmatrix} 2.69 & \mathbf{0.10} \\ (0.90) & \\ 1.55 & 84.0 \\ (3.18) & (27.18) \end{bmatrix}$	$\begin{bmatrix} 6.72 & \mathbf{1.00} \\ (18) & \\ 24.39 & 88.44 \\ (70) & (375) \end{bmatrix}$	$\begin{bmatrix} < 10^{-11} & \\ (\ ) & \\ 0.03 & 0.99 \\ (236) & (2000) \end{bmatrix}$
BSM + VARMA(1,1) (200.51)	$\begin{bmatrix} 2.64 & \mathbf{0.09} \\ (0.84) & \\ 1.35 & 84.5 \\ (3.67) & (28.56) \end{bmatrix}$	$\begin{bmatrix} 6.80 & \mathbf{1.00} \\ (18) & \\ 24.59 & 88.92 \\ (70) & (376) \end{bmatrix}$	$\begin{bmatrix} < 10^{-12} & \\ (\ ) & \\ < 10^{-12} & 0.08 \\ (\ ) & (722) \end{bmatrix}$
⇒BSM +[2] VARMA(1,1) (200.35)	$\begin{bmatrix} 2.78 & \mathbf{0.12} \\ (0.91) & \\ 1.95 & 87.0 \\ (3.55) & (27.10) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$

(1) Maximum value of the Log-Likelihood

(2) Local Level Model with Drift and Fixed Seasonals for the Signal

In addition to the models introduced in Section 8.2.4 an alternative basic structural model (as in Chapter 3, Section 3.7.3) that did not explicitly take into account the sampling errors was also fitted to the data (see the parameter estimates in Table 8.3).

Table 8.3 Parameter Estimates When Ignoring the Sampling Errors

Hyperparameters	BSM Stochastic Level Slope , Seasonals	⇒ Local Level + Drift + Fixed Seasonals
$\hat{\Sigma}_{l} x\, 10^{-4}$	$\begin{bmatrix} 4.21 & \mathbf{0.19} \\ (0.91) & \\ 4.39 & 129.71 \\ (3.74) & (26.86) \end{bmatrix}$	$\begin{bmatrix} 4.27 & \mathbf{0.21} \\ (0.90) & \\ 4.59 & 113.18 \\ (3.79) & (28.37) \end{bmatrix}$
$\hat{\Sigma}_{r} x\, 10^{-7}$	$\begin{bmatrix} 3.38 & \mathbf{1.00} \\ (18) & \\ 10.42 & 32.10 \\ (60) & (260) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$
$\hat{\Sigma}_{s} x\, 10^{-8}$	$\begin{bmatrix} 0.05 & \mathbf{0.90} \\ (20) & \\ 0.30 & 2.29 \\ (460) & (1050) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$
$\hat{\Sigma}_{\epsilon} x\, 10^{-6}$	$\begin{bmatrix} 0.51 & \mathbf{1.00} \\ (3.0) & \\ 5.23 & 53.95 \\ (5.68) & (771.1) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$
Log-Lik.[1]	204.48	204.47

[1] Maximum value of the Log-Likelihood

The irregular term was removed from the model because the elements of its covariance matrix were quite small when compared with the respective standard errors, although it has in some way captured the autocorrelation structure of the sampling errors. The same happened to the disturbances of the slope and seasonal equations. It is not straightforward to test whether the variances are zero or not in a structural time series framework, as discussed in Harvey(1989, Chapter 5). Problems arise when, under the null hypothesis, the parameters lie on the boundary of the parameter space (which is exactly the case of testing for null variances). However Koopman et al(1995) pointed out that it is not unusual to find a variance going to zero. Indeed, when using STAMP5.0 to fit a basic structural model to the BLFS data, the resulting covariance matrices for the slope and seasonal disturbances and for the irregular term were estimated as being zero, confirming the above results. Hence, the models that will be considered here for further analysis are those with fixed slope and seasonals, i.e. the models marked with $(\Rightarrow)$ on Tables 8.2 and 8.3.

8.4 Model Performance and Diagnostic Tests

The empirical distributions of the standardized residuals (as defined in Koopman,1995,p.203) were compared with a standard normal distribution to verify the assumption that the one-step ahead forecasting errors, i.e. the innovations $(\nu_t - \nu_{t|t-1})$, are normal deviates. The Shapiro-Wilk statistic was computed using the UNIVARIATE procedure from SAS which, also produces normal probability plots. Examination of the residuals and their autocorrelations revealed no evidence to reject the hypothesis of normality and suggested no major inadequacies regarding the fitted models. In addition, three measures of goodness of fit were computed as follows:

(i) the mean bias in predicting the transformed survey estimates

$$MB = \sum_{t=d+1}^{T} \frac{v_t - v_{t|t-1}}{t-d} \quad ,$$

(ii) the mean absolute bias

$$MAB = \sum_{t=d+1}^{T} \frac{| v_t - v_{t|t-1} |}{t-d} \quad ,$$

(iii) the square root of the mean squared relative prediction error

$$SQRE = \sum_{t=d+1}^{T} \left[\frac{(v_t - v_{t|t-1})}{v_t} \right]^2 / (t-d) \quad .$$

Table 8.4, which contains a summary of the goodness of fit measures together with the prediction error covariance matrix(PEV), shows that the models performed in a similar way. However, care must be taken when analysing the PEV since for a BSM with dummy seasonals it does not converge exponentially fast to a steady-state. Note that all three models have the same number of unknown parameters since the model parameters for the noise process are computed separately, and consequently are assumed to be known. The errors when predicting the transformed survey estimates are quite small indicating that all the models fitted well. Notice, however, that although the model which ignores the sampling variation presents a slightly better performance it has the clear drawback of not being able to produce separate estimates for the noise and signal components when the latter is the one of real interest.

Tiller(1992) compared univariate state-space models that do and do not account for sampling errors and reported that, when the sampling error was not explicitly modelled, its autocorrelation ended up in the irregular term leading to a well fitted model. He also reported that in the presence of a model accounting for the sampling variation, there was no need to include an irregular term in the signal model. Interesting enough, a model to represent the BLFS survey estimates (without accounting for sampling errors) does not require an irregular term. This can be due to the fact that two series are modelled simultaneously, so that the series variation is satisfactorily represented through the relationship between their underlying trends/levels.

Table 8.4 Measures of Goodness of Fit Based on the Innovations

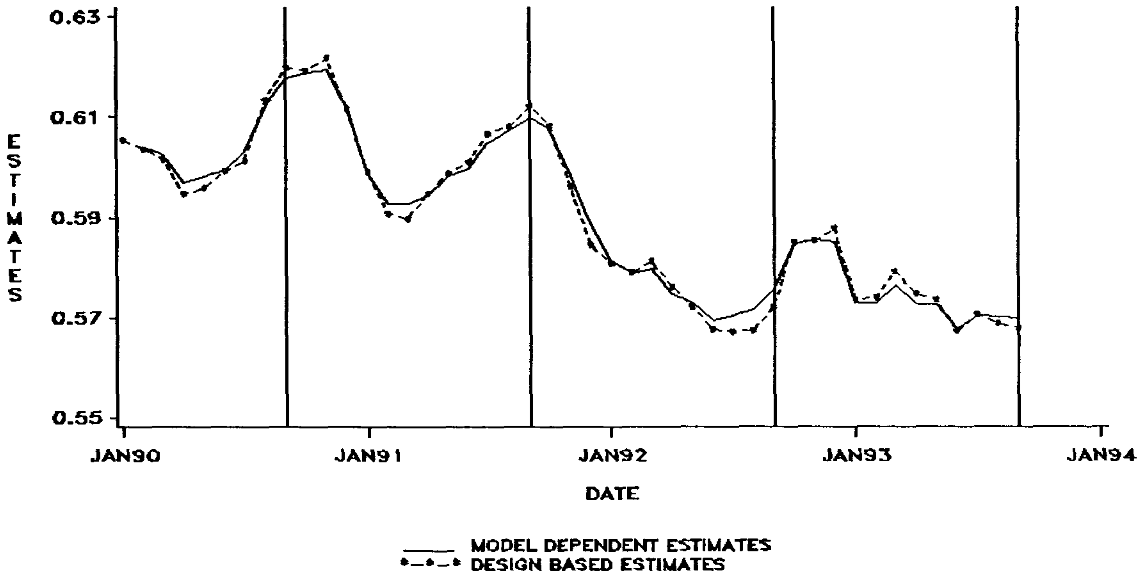
Model[1]	MB	MAB	SQRE(%)	PEV $x\,10^{-4}$
BSM+AR(1)	-0.0041[3] 0.0093[4]	0.0209 0.1264	3.58 4.07	$\begin{bmatrix} 6.34 & \mathbf{0.24} \\ 8.84 & 216.12 \end{bmatrix}$
BSM+ VARMA(1,1)	-0.0040 0.0091	0.0211 0.1260	3.59 4.06	$\begin{bmatrix} 6.33 & \mathbf{0.24} \\ 9.02 & 215.04 \end{bmatrix}$
BSM[2]	-0.0035 0.0073	0.0202 0.1192	3.26 3.67	$\begin{bmatrix} 5.32 & \mathbf{0.20} \\ 5.74 & 163.98 \end{bmatrix}$

(1) Here the model components are a local level with drift and fixed seasonals.
(2) Model without accounting for sampling variation.
(3) Results obtained when estimating log(emp/nilf).
(4) Results obtained when estimating log(une/nilf).

8.5 Results and Discussion

As mentioned before, the main objective of this modelling procedure was to improve estimation in repeated surveys dealing with compositional data. Recall from Chapter 4 (p.40), that the filtered estimate (and consequently the smoothed estimate) for the signal can be interpreted as a composite-type estimate obtained from the combination of an estimate based on past data $(\theta^*_{t|t-1})$ with the current sample information (v_t). Recall, in addition, that the smaller the variance of the design based estimate, the closer the filtered estimate is to the current sample estimate. Figures 8.9 to 8.12 display the design based estimates and the model dependent estimates for the vector of proportions of labour market status and the unemployment rate.

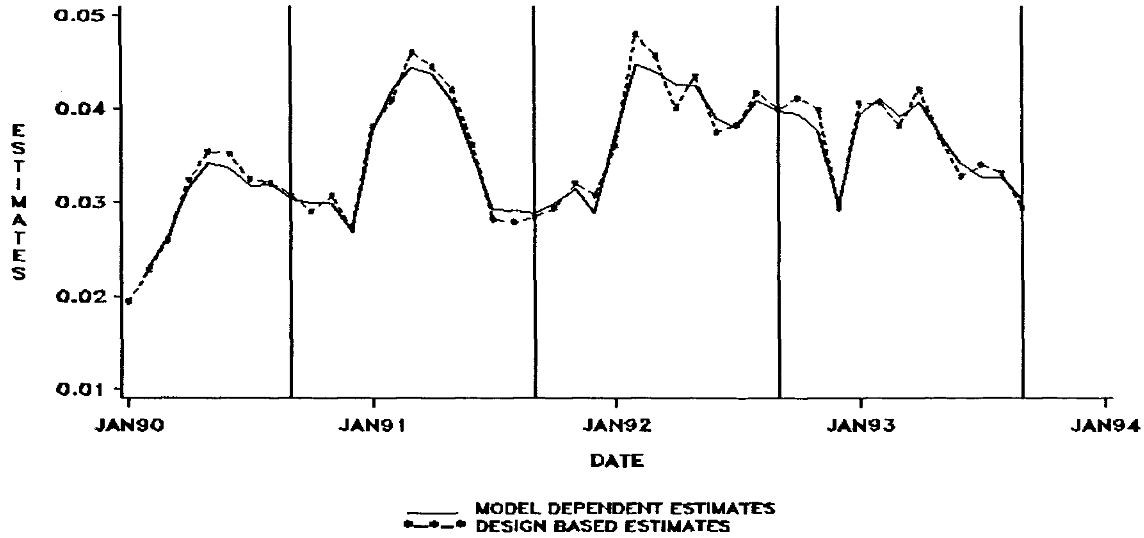
BRAZILIAN LABOUR FORCE SERIES – SAO PAULO
DESIGN BASED AND MODEL DEPENDENT ESTIMATES
PROPORTION OF PEOPLE IN EMPLOYMENT



VERTICAL LINES = SEPTEMBER 90 – 93

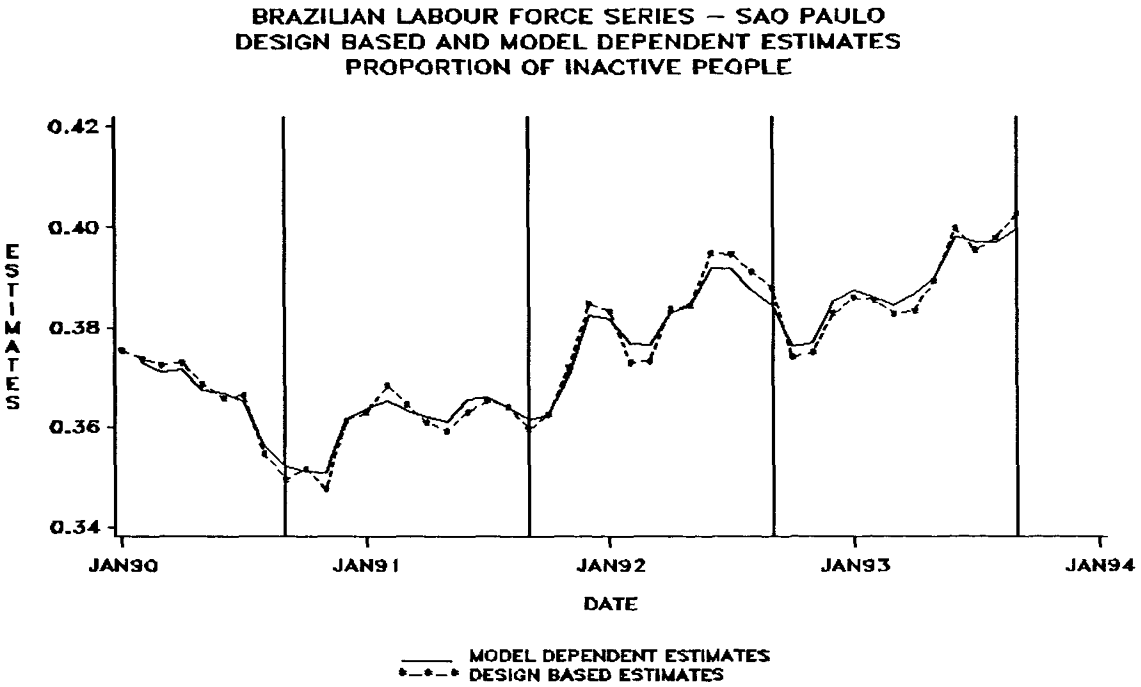
Figure 8.9

BRAZILIAN LABOUR FORCE SERIES – SAO PAULO
DESIGN BASED AND MODEL DEPENDENT ESTIMATES
PROPORTION OF UNEMPLOYED PEOPLE



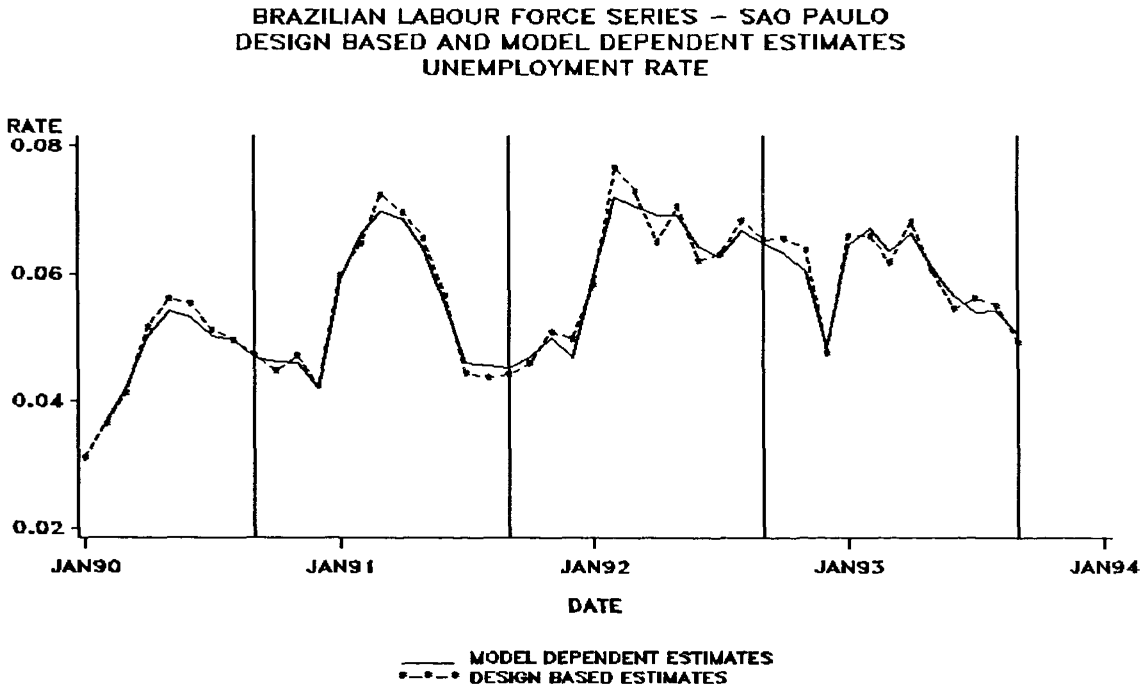
VERTICAL LINES = SEPTEMBER 90 – 93

Figure 8.10



VERTICAL LINES = SEPTEMBER 90 – 93

Figure 8.11



VERTICAL LINES = SEPTEMBER 90 – 93

Figure 8.12

The model dependent estimates are the smoothed estimates, which are based on data from the whole sample period, obtained when fitting a basic structural model for the signal and a VARMA(1,1) model to the noise. The expressions for mapping the transformed estimates back to the original Simplex were introduced in Chapter 6, Sections 6.3 and 6.4. For all four target quantities the signal estimates behave similarly to the design based estimates although some of the turning points of the latter were smoothed out. Figure 8.13 displays the estimated relative errors (defined as the ratio between the estimated standard errors and the corresponding point estimates). The estimated relative errors for the model dependent estimates show much less variability and are, in general, lower than those obtained for the survey estimates.

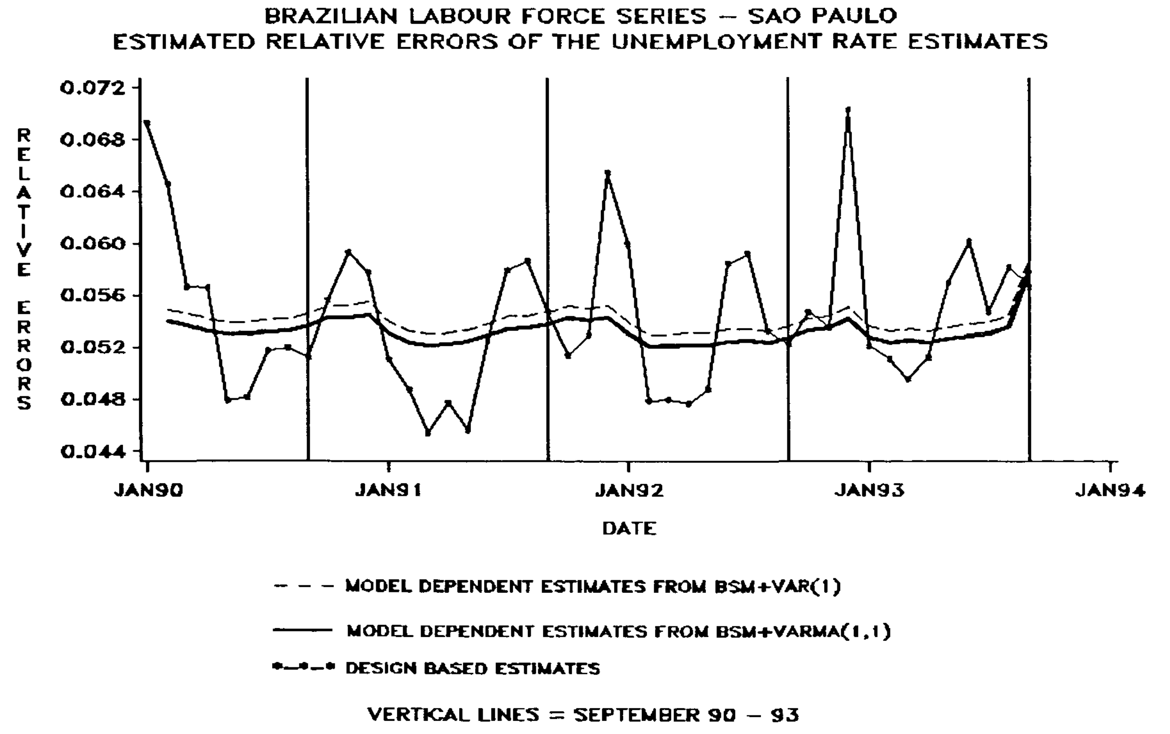

BRAZILIAN LABOUR FORCE SERIES – SAO PAULO
ESTIMATED RELATIVE ERRORS OF THE UNEMPLOYMENT RATE ESTIMATES

Figure 8.13

Seasonally adjusted estimates were obtained based on the basic structural model defined for the signal process. The seasonally adjusted estimates for the unemployment rate series obtained under the model were compared with those produced by the X11 procedure from SAS-ETS which is an adaptation of the Bureau of the Census X-11 seasonal adjustment program.

Figure 8.14 displays the seasonally adjusted series produced by both methods. Note that the model proposed here produces a smoother curve than the X11. This can be explained by the fact that the model dependent seasonally adjusted values represent mostly the underlying trend of the unobservable signal, since there was no irregular component in the signal model due to the inclusion of a specific model for the sampling error component. On the other hand, the X11 estimates were obtained based on a multiplicative decomposition for the observed unemployment rate series. In this case, the seasonally adjusted values comprise the trend/level and irregular components.

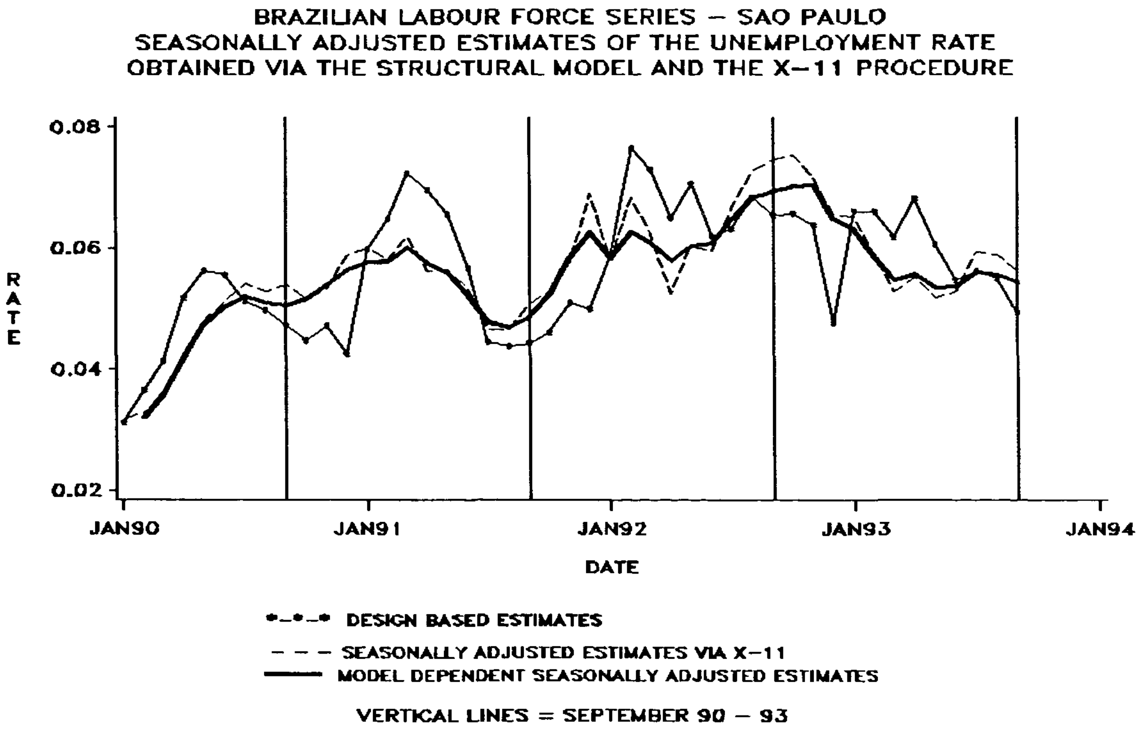


Figure 8.14

In addition, Figure 8.15 shows that the X11 and model dependent estimates of the seasonal effects behave quite similarly. This leads to the conclusion that, in this study, the X11 trend and irregular components have incorporated features from the sampling error process. Pfeffermann, Bell & Signorelli(1996) had reached a similar conclusion when analysing the Australian unemployment rate series.

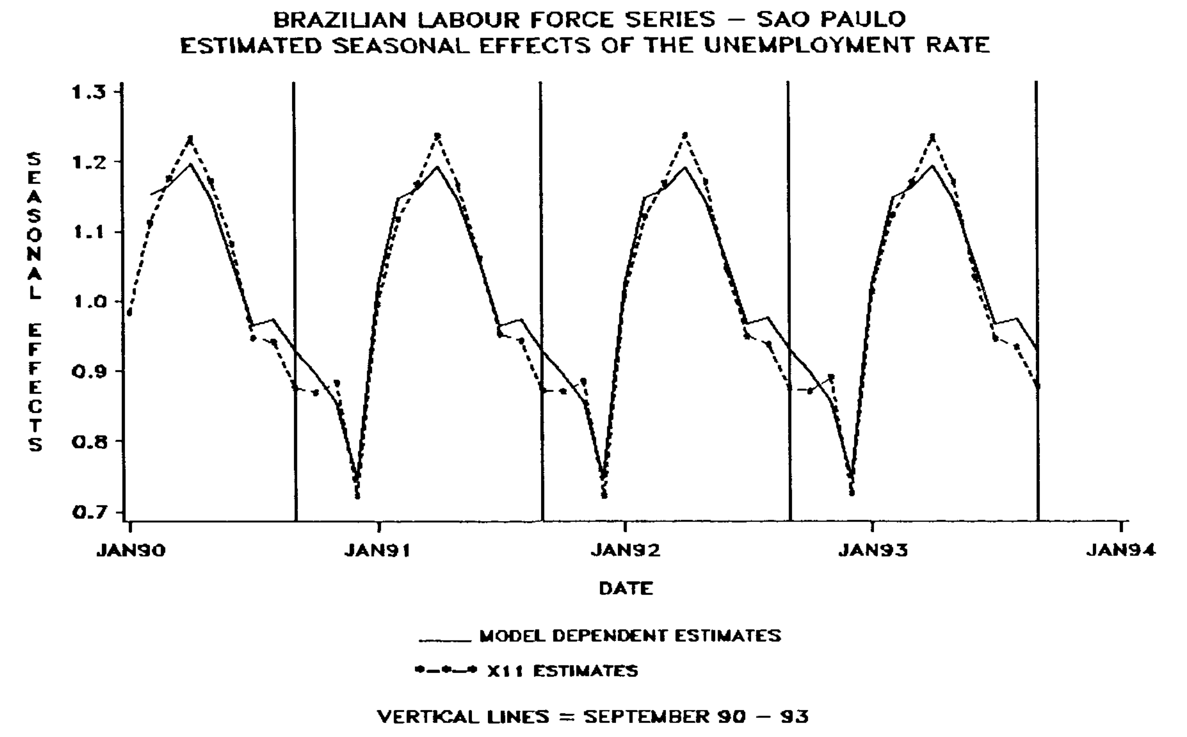


Figure 8.15

Figure 8.16 compares the trend estimates obtained when ignoring the sampling variation with those derived when their autocorrelation is modelled via a VARMA(1,1). The trend produced by the model which takes into account the sampling error is smoother, suggesting that the model succeeded in removing the underlying fluctuations induced by the correlation structure of the sampling errors. Finally, Figure 8.17 displays the model dependent estimates for the seasonal effects of the original compositions. The important point to note here is that these estimates were obtained concurrently from a multivariate model which took into account two very important features of the data, namely the compositional constraints and the presence of sampling errors.

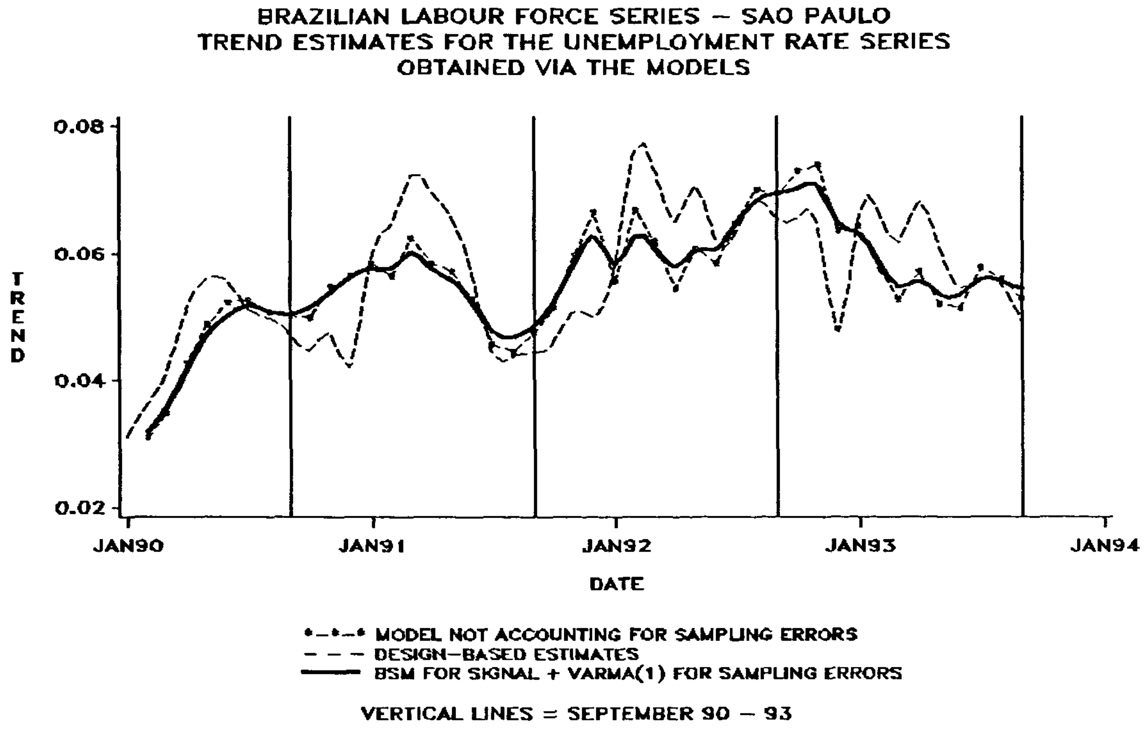


Figure 8.16

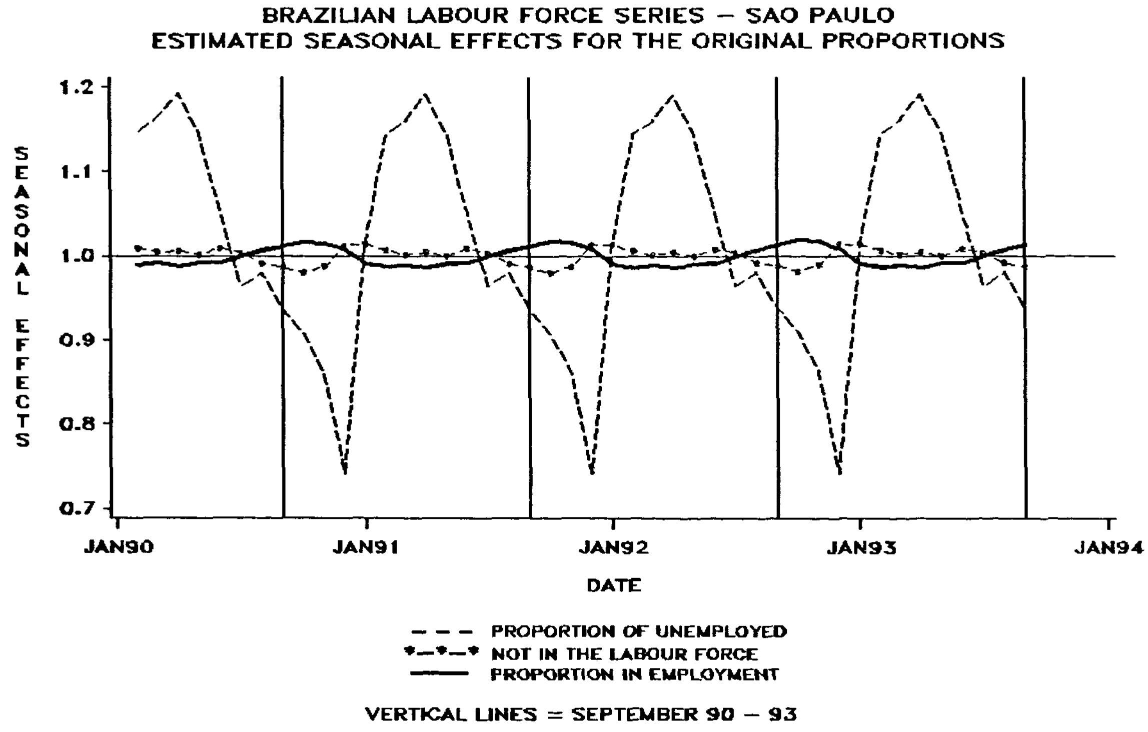


Figure 8.17

The present study using compositional data from the Brazilian Labour Force Survey has illustrated the usefulness of the proposed modelling procedure, yielding signal estimates satisfying the unity-sum constraint. The results appear to show that the underlying trend obtained from this approach is smoother than that obtained from standard methods for seasonal adjustment, suggesting that when the sampling errors are not properly accounted for they end up incorporated into the trend. In addition, the estimated relative errors of model dependent unemployment rate estimates are in general lower than those regarding the design based estimates.

9 Conclusions and Further Research

9.1 Conclusions

This thesis proposed a state-space approach for modelling compositional time series from repeated surveys while taking the sampling errors into account. The idea was to combine the existing theory for the analysis of compositional time series with the state-space formulation of time series models, to improve estimation of population parameters using data from repeated sample surveys. The most important feature of the modelling procedure is that it provides bounded predictions and signal estimates of the parameters in a composition, while satisfying the unity-sum constraint and taking into account the sampling errors.

This was accomplished by mapping the compositions from the Simplex onto the Real space using the additive logratio transformation, then modelling the transformed data via multivariate state-space models, and finally applying the additive logistic transformation to obtain estimates in the original scale. Previous work regarding compositional time series did not address the problem that, in a survey situation, the series are subject to sampling errors. On the other hand, the state-space approach for improving estimation in repeated surveys has never before been applied to model compositional data from overlapping surveys.

Most of the previous work in this area was concerned with improving estimation of univariate series of proportions like, for example, unemployment rate series. The procedure usually adopted for such cases was to fit the time series models directly to the original series of estimated proportions. However, the analysis in Chapter 5 led to the conclusion that it is not possible to ensure that the signal estimates are always bounded between zero and one, if state-space models are fitted directly to series of proportions without any transformation. Because this reasoning extends naturally to series of compositions, a decision in favour of transforming the original compositional data before modelling was taken here.

The use of the additive logratio transformation to map the compositions from the Simplex onto the Real space gave rise to the inevitable question of whether or not the choice of the reference variable would affect the modelling procedure and corresponding results. This issue was addressed in Section 6.2, where it was shown that the proposed modelling procedure is permutation invariant, i.e. whichever permutation of y is used to construct the time series of logratios, the same inferences are obtained when returning to the original Simplex. Although the emphasis of this thesis was on the compositional case, the framework introduced in Sections 6.3 and 6.4 is directly applicable for modelling unconstrained multivariate data from repeated surveys, allowing two or more survey variables to be modelled simultaneously while taking into account the sampling errors.

Section 7.3 established a comprehensive framework for specifying time series models for the unobservable sampling error process in the multivariate case, extending previous work for the univariate case by Pfeffermann & Bleuer(1993), Pfeffermann, Bell & Signorelli(1996) and Scott, Smith & Jones(1977). In addition, Section 7.4 provided the theory for modelling time series of transformed sampling errors for the compositional case.

The empirical work using the Brazilian Labour Force Survey data, described in Chapter 8, demonstrated the usefulness of this modelling procedure in a genuine survey situation. It proved that the proposed approach could be routinely employed to obtain improved estimates in large scale surveys, such as labour force surveys. The results of the empirical work also lead to the conclusion that smoother trends are obtained with a model which explicitly accounts for the sampling errors, when compared with the results from other standard procedures. This suggests that methods which do not model the sampling errors properly may end up propagating their influence into the trend and seasonal components. In addition, because the model dependent estimators can be viewed as composite-type estimators, combining past and current survey data, the estimated relative errors for the model dependent estimates of the unemployment rate were in general lower than those corresponding to the design based estimates.

One drawback of the modelling procedure proposed here is that although confidence regions for the original compositional vector can be constructed based on the model dependent estimates using the additive logistic normal distribution, confidence intervals for the individual proportions are not readily available. These intervals could be obtained from marginal distributions of the additive logistic normal distribution, which can only be evaluated by integrating out some of the elements of the compositional vector. However, as pointed out by Brunsdon(1987, p.135), this produces intractable expressions.

9.2 Recommendations for Further Research

Under a state-space formulation, a wide variety of models are available to represent the multivariate signal and noise processes, which favours a broader use of this modelling procedure. Therefore a potential extension of this work would be the application of the modelling procedure developed here to different data sets. Further empirical research could consider situations where the composition lies on a Simplex with a dimension higher than two and/or with compositions evolving closer to the boundaries of the interval [0,1]. Another point regards the Kalman Filter routine employed for the empirical work in Chapter 8, which might probably be improved by using the diffuse Kalman Filter as recommended in recent literature.

In addition, better insight on the performance of the modelling procedure might be gained via its application to simulated data, for which the "true" underlying models are known. The models considered here can also be extended to incorporate rotation group bias effects and explanatory variables.

In the view of a more general survey framework, note that the quantitative procedure proposed for estimating the cross-correlation function of the sampling errors was mostly based on the pseudo-errors obtained from panel estimates, with the panels pre-defined according to the sampling scheme. Further research is now needed to extend this procedure for surveys in which the elementary estimates and corresponding pseudo-errors are not those pre-defined by the sample selection scheme. For example, the elementary estimates could be obtained from disjoint sub-samples selected from the original sample, similarly to the idea of random groups for variance estimation (see Wolter, Chapter 2).

Another interesting issue regards the seasonal adjustment of multivariate time series. This thesis provided a method for seasonally adjusting compositional data taking into account the unity-sum constraint and the presence of sampling errors. However further research is needed to model sum-constrained series in general, to handle situations in which the interest lies on the analysis of the basis $\boldsymbol{w}$ instead of the compositions $\boldsymbol{y}$. For example, statistical agencies often need to produce seasonally adjusted estimates of industrial production per type of industry in accordance with the seasonally adjusted figures for the total industrial production. The problem is similar to that of benchmarking series of cross-sectional data.

One theoretical issue which also deserves further attention is the provision of confidence intervals for the individual proportions which, as mentioned before, has not been covered in this thesis. Another theoretical issue deserving further research is that only the additive logratio transformation and its inverse, the additive logistic transformation, were considered here, although other permutation invariant transformations that map compositions from the Simplex onto the Real space could also be considered.

10 Bibliography

Aitchison,J.(1986). The Statistical Analysis of Compositional Data. Chapman and Hall. New York.

Aitchison,J. & Shen,S.M.(1980). Logistic-Normal distributions: some properties and uses. Biometrika 67, 261-272.

Anderson,B.D.O. & Moore,J.B.(1979). Optimal Filtering. Prentice-Hall.

Barbosa,E.(1989). Dynamic Bayesian Models for Vector Time Series Analysis and Forecasting. Unpublised Ph.D. Thesis. University of Warwick.

Bell,W.R.(1984). Signal extraction for nonstationary time series. Annals of Statistics 13, 646-664.

Bell,W.R. & Hillmer,S.C.(1990). The time series approach to estimation for repeated surveys. Survey Methodology 19, No.2, 195-215.

Binder,D.A. & Hidiroglou,M.A.(1988). Sampling in time. In Handbook of Statistics, 6, 187-211. Eds, P.R.Krishnaiah and C.R.Rao. Elsevier Science.

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

Binder,D.A., Bleuer,S.R. & Dick,J.P.(1993). Time series methods applied to survey data. Paper presented at the 49th ISI Meeting, Florence, Italy.

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

Box,G.E.P. & Jenkins,G.M.(1970). Time Series Analysis Forecasting and Control. Holden-Day. San Francisco.

Brunsdon,T.M.(1987). Time Series Analysis of Compositional Data. Unpublished Ph.D. Thesis. University of Southampton.

Bureau of the Census(1978). The Current Population Survey: design and methodology. Technical Paper 40.

Chan,W.Y.T. & Wallis,K.F.(1978). Multiple time series modelling: another look at the mink-muskrat interactions. Applied Statistics, 27, 168-175.

Cramer,J.S.(1986). Econometric Applications of the Maximum Likelihood Methods. Cambridge University Press. Cambridge.

Cochran,W.G.(1977). Sampling Techniques. John Wiley & Sons.

Dagum,E.B.(1980). The X11-ARIMA seasonal adjustment method (catalog No.12-564E). Statistics Canada.

Dagum,E.B. & Quenneville,B.(1993). Dynamic linear models for time series components. Journal of Econometrics 55, 333-351.

De Jong,P.(1988). 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 Society, 84, 1085-1088.

De Jong,P.(1991). The diffuse Kalman filter. The Annals of Statistics, 19, 1073-1083.

Duncan,G.J. & Kalton,G.(1987). Issues of design and analysis of surveys across time. International Statistical Review 55, 1, 97-11.

Encyclopedia of Statistical Science(1985). Volume 5. Eds. Kotz,S. & Johnson,N.L. John Wiley & Sons.

Engle,R.F.(1978). Estimating structural models of seasonality. In Seasonal Analysis of Economic Time Series, 281-309. A. Zellner(Ed.). Bureau of the Census.

Eckler,A.R.(1955). Rotation sampling. Annals of Mathematical Statistics 26, 664-685.

Fernandez,F.J.M.(1986). Estimation and testing of multivariate time series models. Unpublished Ph.D. Thesis. London School of Economics, University of London.

Fernandez,F.J.M & Harvey,A.C.(1990). Seemingly unrelated time series equations and a test for homogeneity. Journal of Business and Economic Statistics, vol.8, no.1, 71-81.

Graybill,F.A.(1983). Matrices with Applications in Statistics. Second Edition. Wadsworth International Group. Belmont, California.

Granger,C.W.J. & Newbold,P.(1976). Forecasting transformed series. Journal of the Royal Statistical Society B, 38, 189-203.

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

Hamilton,J.D.(1994). Time Series Analysis. Princeton University Press.

Harrison,P.J. & Stevens,C.F.(1976). Bayesian forecasting. Journal of the Royal Statistical Society, Series B 38, 205-47.

Harvey,A.C.(1986). Analysis and generalisation of a multivariate exponential smoothing model. Management Science, vol.32, no.3, 374-380.

Harvey,A.C.(1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press. Cambridge.

Harvey,A.C.(1993). Time Series Models.Second Edition. Harvester Wheatsheaf. London.

Harvey,A.C. & Peters,S.(1984). Estimation procedures for structural time series models. London School of Economics. Mimeo.

Harvey,A.C. & Phillips,G.D.A.(1979). Maximum likelihood estimation of regression models with autoregressive moving-average disturbances. Biometrika 66, 49-58.

Harvey,A.C. & Shephard,N.(1993). Structural time series models. In Handbook of Statistics, vol.11, 261-302. Eds. S.Maddala, C.R.Rao and H.D.Vinod. Elsevier Science Publishers.

Janacek,G. & Swift,L.(1993). Time Series - forecasting, simulation, applications. Ellis Horwood. West Sussex.

Jenkins,G.M. & Alavi,S.A.(1981). Some aspects of modelling and forecasting multivariate time series. Journal of Time Series Analysis, vol.2, no.1, 1-47.

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

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

Kalman,R.E.(1960). A new approach to filtering and prediction problems. Journal of Basic Engineering Transactions ASME, Series D 82, 35-45.

Kalman,R.E. & Bucy,R.S.(1961). New results in linear filtering and prediction theory. Journal of Basic Engineering Transactions ASME, Series D 83, 95-108.

Kalton,G. & Citro,C.F.(1993). Panel surveys: adding the fourth dimension. Survey Methodology, vol.19, 2, 205-216.

Kish,L.(1987). Statistical Design for Research. John Wiley & Sons.

Koopman,S.J., Harvey,A.C., Doornik,J.A. & Shephard,N.(1995) Stamp 5.0 - Structural Time Series Analyser, Modeller and Predictor. Chapman & Hall. London.

Lee,H.(1990). Estimation of panel correlations for the Canadian Labour Force Survey. Survey Methodology, 16, 283-292.

Lindley,D.V.(1965). Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 1. Cambridge University Press. Cambridge.

Liu,L.(1986). Multivariate time series analysis using vector ARMA models. Scientific Computing Associates. Delkab, Illinois.

Liu,L. & Hudak,G.(1986). The SCA Statistical System. Reference Manual for Forecasting and Time Series Analysis. Scientific Computing Associates. Dekalb, Illinois.

Maravall,A. & Mathis,A.(1994). Encompassing univariate models is multivariate time series. Journal of Econometrics, 61, 197-233.

IBGE(1980). Metodologia da Pesquisa Mensal de Emprego 1980. Relatórios Metodológicos. Fundação Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro.

Morettin,P.A. & Toloi,C.M.C.(1985). Previsão de Séries Temporais. 2ª Edição. Atual Editora. São Paulo.

Mittnik,S.(1991). Derivation of the unconditional state-covariance matrix for exact-likelihood estimation of ARMA models. Journal of Economic Dynamics and Control, 15, 731-740.

Nerlove,M., Grether,D.M. & Carvalho,J.L.(1995) Analysis of Economic Time Series. Revised Edition. Academic Press Limited. London.

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

Pfeffermann,D.(1991). Estimation and seasonal adjustment of population means using data from repeated surveys. Journal of Business & Economic Statistics 9, 163-177.

Pfeffermann,D., Burck,L. & Ben-Tuvia,S.(1989). A time series model for estimating housing price indexes adjusted for changes in quality. Proceedings of the Statistics Canada Symposium on Analysis of Data in Time.

Pfeffermann,D. & Burck,L.(1990). Robust small area estimation combining time series and cross-sectional data. Survey Methodology. Vol. 16, No. 2, 217-237.

Pfeffermann,D. & Bleuer,S.R.(1993). Robust joint modelling of labour force series of small areas. Survey Methodology, vol.19, 2, 149-164.

Pfeffermann,D., Bell,P. & Signorelli,D.(1996). Labour force trend estimation in small areas. Paper presented at the Bureau of Census Annual Research Conference.

Quintana,J.M. & West,M.(1988). Time series analysis of compositional data. In Bayesian Statistics 3. Bernardo,J.M., DeGroot,M.A., Lindley,D.V., Smith,A.F.M. (Eds.). Oxford University Press.

Rao,C.R.(1973). Linear Statistical Inference and its Applications. John Wiley & Sons.

Rao,J.N.K. & Graham,J.E.(1964). Rotation designs for sampling on repeated occasions. Journal of the American Statistical Association 50, 492-509.

Reinsel,G.C.(1993). Elements of Multivariate Time Series Analysis. Springer-Verlag.

SAS Institute Inc.(1988). SAS/ETS User's Guide, Version 6. SAS Institute Inc. Cary, NC.

SAS Institute Inc.(1989). SAS/IML Software: Usage & Reference, Version 6, First Edition. SAS Institute Inc. Cary, NC.

SAS Institute Inc.(1995). SAS/IML Software: Changes and Enhancements through Release 6.11. SAS institute Inc. Cary, NC.

SAS Institute Inc.(1985). SAS User's Guide: Basics, Version 5 Edition. SAS Institute Inc. Cary, NC.

Scott,A.J. & Smith,T.M.F.(1974). Analysis of repeated surveys using time series methods. Journal of the American Statistical Association 69, 674-678.

Scott,A.J., Smith,T.M.F. & Jones,R.G.(1977). The application of time series methods to the analysis of repeated surveys. International Statistics Review 45, 13-28.

Shephard,N.G. & Harvey,A.C.(1989). Tracking the level of support for the parties during general election campaigns. Mimeo. Dept. of Statistics, London School of Economics.

Singh,D.(1968). Estimates in successive sampling using multi-stage design. Journal of the American Statistical Association 63, 99-112.

Smith,T.M.F.(1978). Principles and Problems in the analysis of repeated surveys. In Survey Sampling and Measurement, 201-216. Eds, N.K.Namboodini. Academic Press.

Smith, T.M.F.(1996). Public opinion polls: the UK general election, 1992. Journal of the Royal Statistical Society A, 159, part 3, 535-545.

Smith,T.M.F. & Brunsdon,T.M.(1986). Time series methods for small areas. Unpublished Report. University of Southampton.

Tiao,G.C. & Box,G.E.P.(1981). Modelling multiple time series with applications. Journal of the American Statistical Association 76, 802-816.

Tikkiwal, B.D.(1979). Successive sampling - a review. Proceedings of the 42nd Session of the International Statistical Institute held in Manila. Book 2, 367-84.

Tiller,R.B.(1989). A Kalman filter approach to labor force estimation using survey data. In Proceedings of the Survey Research Methods Section, American Statistical Association, 16-25.

Tiller,R.B.(1992). Time series modelling of sample data from the U.S. Current Population Survey. Journal of Official Statistics, vol.8, 2, 149-166.

Train,G., Cahoon,L. & Makens,P.(1978). The Current Population Survey variances, interrelationships, and design-effects. In Proceedings of the Survey Research Methods Section. American Statistical Association, 443-448.

Wallis,F.(1987). Time series analysis of bounded economic variables. Journal of Time Series Analysis 8, 115-23.

Wei,W.W.S.(1993). Time Series Analysis - univariate and multivariate methods. Addison-Wesley.

West,M. & Harrison,J.(1989). Bayesian Forecasting and Dynamic Models. Springer-Verlag.

Whittle,P.(1983). Prediction and Regulation, 2nd edn (revised). Blackwell.

Wolter,K.M.(1979). Composite estimation in finite populations. Journal of the American Statistical Association 74, 604-613.

Wolter,K.M.(1985). Introduction to Variance Estimation. Springer-Verlag. New York.

Woodruff,R.S.(1963). The use of rotating samples in the Census Bureau's monthly surveys. Journal of the American Statistical Association.

Yates,F.(1949). Sampling Methods for Censuses and Surveys. Griffin & Co.

Appendix A1 - The Kalman Filter Equations

Consider the system:

$$y_t = H_t \alpha_t + \varepsilon_t \quad , \tag{A1.1a}$$

$$\alpha_t = T_t \alpha_{t-1} + G_t \eta_t \quad . \tag{A1.1b}$$

By convention define, for $t=0$, $\alpha_0 \sim N(\hat{\alpha}_{0|0}, P_{0|0})$. $\{\varepsilon_t\}$ and $\{\eta_t\}$ are mutually uncorrelated normally distributed disturbances with mean zero and covariance matrices U_t and Q_t , respectively. The vectors of disturbances are also uncorrelated with the initial state-vector α_0 .

Considering $\hat{\alpha}_{0|0} = E(\alpha_0 \,|\, y_0)$ and $P_{0|0} = V(\alpha_0 \,|\, y_0)$ it follows that:

$$\hat{\alpha}_{1|0} = E(\alpha_1 \,|\, y_0) = E(T_1 \alpha_0 + G_1 \eta_1 \,|\, y_0) = E(T_1 \alpha_0 \,|\, y_0) = T_1 \hat{\alpha}_{0|0} \ , \tag{A1.2a}$$

$$\begin{aligned} P_{1|0} &= E[(\alpha_1 - \hat{\alpha}_{1|0})\,(\alpha_1 - \hat{\alpha}_{1|0})' \,|\, y_0] \\ &= E[(T_1 \alpha_0 + G_1 \eta_1 - T_1 \hat{\alpha}_{0|0})\,(T_1 \alpha_0 + G_1 \eta_1 - T_1 \hat{\alpha}_{0|0})' \,|\, y_0) \\ &= E[T_1 (\alpha_0 - \hat{\alpha}_{0|0})(\alpha_0 - \hat{\alpha}_{0|0})' \, T_1' \,|\, y_0] + G_1 E(\eta_1 \eta_1' \,|\, y_0)\, G_1' \\ &= T_1 P_{0|0} T_1' + G_1 Q_1 G_1' \quad . \end{aligned} \tag{A1.2b}$$

In addition, from (A1.2), it follows that

$$\hat{y}_{1|0} = E(y_1 \,|\, y_0) = E(H_1 \alpha_1 + \varepsilon_1 \,|\, y_0) = H_1 \hat{\alpha}_{1|0} \ ,$$

$$\begin{aligned} F_{1|0} &= E[(y_1 - \hat{y}_{1|0})\,(y_1 - \hat{y}_{1|0})' \,|\, y_0] \\ &= E[(H_1 \alpha_1 + \varepsilon_1 - H_1 \hat{\alpha}_{1|0})\,(H_1 \alpha_1 + \varepsilon_1 - H_1 \hat{\alpha}_{1|0})' \,|\, y_0) \\ &= E[H_1 (\alpha_1 - \hat{\alpha}_{1|0})(\alpha_1 - \hat{\alpha}_{1|0})' \, H_1' \,|\, y_0] + E(\varepsilon_1 \varepsilon_1' \,|\, y_0) \\ &= H_1 P_{1|0} H_1' + U_1 \quad . \end{aligned}$$

$$\begin{aligned} COV(\alpha_1, y_1 \mid y_0) &= E[(\alpha_1 - \hat{\alpha}_{1\mid 0})\,(H_1\,\alpha_1 + \varepsilon_1 - H_1\,\hat{\alpha}_{1\mid 0})' \mid y_0] \\ &= E[(\alpha_1 - \hat{\alpha}_{1\mid 0})(\alpha_1 - \hat{\alpha}_{1\mid 0})'\,H_1' \mid y_0] \\ &= P_{1\mid 0}H_1' \quad . \end{aligned}$$

Therefore,

$$\left[\begin{matrix} \boldsymbol{\alpha}_1 \\ y_1 \end{matrix} \,\middle|\, y_0 \right] \sim N \left\{ \left[\begin{matrix} \hat{\boldsymbol{\alpha}}_{1\mid 0} \\ H\,\hat{\boldsymbol{\alpha}}_{1\mid 0} \end{matrix}\right] , \left[\begin{matrix} P_{1\mid 0} & P_{1\mid 0}H_1' \\ H_1\,P_{1\mid 0} & H_1\,P_{1\mid 0}H_1' + U_1 \end{matrix}\right] \right\} .$$

Using Result 3.2, yields:

$$\hat{\alpha}_{1\mid 1} = E(\alpha_1 \mid y_0, y_1) = \hat{\alpha}_{1\mid 0} + P_{1\mid 0}H_1'\,(H_1\,P_{1\mid 0}H_1' + U_1)^{-1}(y_1 - H_1\,\hat{\alpha}_{1\mid 0}) \ ,$$

$$P_{1\mid 1} = V(\alpha_1 \mid y_0, y_1) = P_{1\mid 0} - P_{1\mid 0}H_1'\,(H_1\,P_{1\mid 0}H_1' + U_1)^{-1}\,H_1\,P_{1\mid 0} \quad .$$

Repeating these steps for $t = 2, 3, \ldots$ one gets the general recursion equations in (3.5).

Appendix A2 - State-Space Representation of an ARMA(2,2) Model

Consider the following ARMA(2,2) model:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t - \delta_1 \epsilon_{t-1} + \delta_2 \epsilon_{t-2} \quad . \tag{A2.1}$$

Using Result 3.5, the state-space representation for (A2.1) is given by:

$$y_t = (1\,0\,0)\,\alpha_t \quad ,$$

$$\alpha_t = \begin{bmatrix} \phi_1 & 1 & 0 \\ \phi_2 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} 1 \\ -\delta_1 \\ -\delta_2 \end{bmatrix} \epsilon_t \quad ,$$

with

$$\alpha_t = \begin{bmatrix} y_t \\ \phi_2 y_{t-1} - \delta_1 \epsilon_t - \delta_2 \epsilon_{t-1} \\ -\delta_2 \epsilon_t \end{bmatrix} \quad ,$$

and

$$\alpha_{t-1} = \begin{bmatrix} y_{t-1} \\ \phi_2 y_{t-2} - \delta_1 \epsilon_{t-1} - \delta_2 \epsilon_{t-2} \\ -\delta_2 \epsilon_{t-1} \end{bmatrix} \quad .$$

Appendix B1 - Equation (5.5)

From the Kalman Filter equations in (3.5), the prediction and updating equations for the local level model are given by:

$$\begin{aligned}
\hat{\theta}_{t|t-1} &= \hat{\theta}_{t-1|t-1} \quad , \\
P_{t|t-1} &= P_{t-1|t-1} + \sigma_\eta^2 \quad , \\
\\
\hat{y}_{t|t-1} &= \hat{\theta}_{t-1|t-1} \quad , \\
F_{t|t-1} &= P_{t-1|t-1} + \sigma_\eta^2 + \sigma_e^2 = P_{t|t-1} + \sigma_e^2 \quad , \\
\\
\hat{\theta}_{t|t} &= \hat{\theta}_{t|t-1} + P_{t|t-1}(y_t - \hat{y}_{t|t-1})/F_{t|t-1} \quad , \\
P_{t|t} &= P_{t|t-1} - P^2_{t|t-1}/F_{t|t-1} \quad .
\end{aligned} \tag{B1.1}$$

The Kalman Filter has a steady-state solution if there exist a time-invariant error covariance matrix that satisfies equation (3.11a) which, in this case, is given by $P = P - PF^{-1}P + \sigma_\eta^2$, where $F^{-1} = P + \sigma_e^2$.

Substituting P and F for $P_{t-1|t}$ and $F_{t-1|t}$ in (B1.1), it follows that:

$$\begin{aligned}
\hat{\theta}_{t|t} &= \hat{\theta}_{t|t-1} + \frac{P(y_t - \hat{y}_{t|t-1})}{P + \sigma_e^2} \\
&= \hat{\theta}_{t-1|t-1} + \frac{P(y_t - \hat{\theta}_{t-1|t-1})}{P + \sigma_e^2} \\
&= \left[1 - \frac{P}{P + \sigma_e^2} \right] \hat{\theta}_{t-1|t-1} + \left[\frac{P}{P + \sigma_e^2} \right] y_t \quad .
\end{aligned}$$

Appendix B2 - Equation (5.13)

Using the Kalman Filter equations (3.5), the filtered estimate for α_t is:

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\,F^{-1}_{t|t-1}(y_t - \hat{y}_{t|t-1}) \quad , \tag{B2.1a}$$

with

$$\begin{aligned} \hat{\alpha}_{t|t-1} &= \boldsymbol{T}\hat{\alpha}_{t-1|t-1} \quad , \\ \hat{y}_{t|t-1} &= \boldsymbol{H}\boldsymbol{T}\hat{\alpha}_{t-1|t-1} \quad . \end{aligned} \tag{B2.1b}$$

The steady-state covariance matrix $\boldsymbol{P}$ is the solution of the equation (as in (3.11a)):

$$\boldsymbol{P} - \boldsymbol{T}\left[\boldsymbol{P} - \boldsymbol{P}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}')^{-1}\boldsymbol{H}\boldsymbol{P}\right]\boldsymbol{T}' - \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G} = 0' \quad , \tag{B2.2}$$

where,

$$\boldsymbol{Q} = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix} \quad .$$

Representing the steady-state covariance matrix as:

$$\boldsymbol{P} = \begin{bmatrix} p_1 & p_{12} \\ p_{12} & p_2 \end{bmatrix} \quad , \tag{B2.3}$$

implies that:

$$\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}' = p_1 + 2p_{12} + p_2 \quad , \tag{B2.4}$$

Substituting (B2.3) and (B2.4) into (B2.2) it follows that the steady-state filter is given by:

$$\hat{\alpha}_{t|t} = \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix} \hat{\alpha}_{t-1|t-1} + \frac{1}{p_1 + 2p_{12} + p_2} \begin{bmatrix} p_1 & p_{12} \\ p_{12} & p_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left\{ y_t - [1 \;\; 1] \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix} \hat{\alpha}_{t-1|t-1} \right\}. \quad \textbf{(B2.5)}$$

Consequently:

$$\begin{bmatrix} \hat{\theta}_{t|t} \\ \hat{e}_{t|t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} \hat{\theta}_{t-1|t-1} \\ \hat{e}_{t-1|t-1} \end{bmatrix} + \frac{1}{p_1 + 2p_{12} + p_2} \begin{bmatrix} p_1 + p_{12} \\ p_{12} + p_2 \end{bmatrix} [y_t - \hat{\theta}_{t-1|t-1} - \beta \hat{e}_{t-1|t-1}].$$

Hence:

$$\begin{aligned} \hat{\theta}_{t|t} &= \hat{\theta}_{t-1|t-1} + \frac{p_1 + p_{12}}{p_1 + 2p_{12} + p_2} [y_t - \hat{\theta}_{t-1|t-1} - \beta \hat{e}_{t-1|t-1}] \quad , \\ & \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \textbf{(B2.6)} \\ \hat{e}_{t|t} &= \beta \hat{e}_{t-1|t-1} + \frac{p_2 + p_{12}}{p_1 + 2p_{12} + p_2} [y_t - \hat{\theta}_{t-1|t-1} - \beta \hat{e}_{t-1|t-1}] \quad . \end{aligned}$$

Because,

$$\frac{p_2 + p_{12}}{p_1 + 2p_{12} + p_2} = 1 - \frac{p_1 + p_{12}}{p_1 + 2p_{12} + p_2} \quad ,$$

denote $W = \dfrac{p_1 + p_{12}}{p_1 + 2p_{12} + p_2}$ and $1 - W = \dfrac{p_2 + p_{12}}{p_1 + 2p_{12} + p_2}$.

Therefore,

$$\begin{aligned} \hat{\theta}_{t|t} &= (1 - W)\hat{\theta}_{t-1|t-1} + Wy_t - \beta W \hat{e}_{t-1|t-1} \quad , \\ \hat{e}_{t|t} &= \beta W \hat{e}_{t-1|t-1} + (1 - W)y_t - (1 - W)\hat{\theta}_{t-1|t-1} \quad . \end{aligned}$$

Appendix B3 - Equation (5.14)

From equations (5.13) it follows that:

$$\begin{aligned}
\hat{\theta}_{t|t} &= (1 - W)[(1 - W)\hat{\theta}_{t-2|t-2} + W y_{t-1} - \beta W \hat{e}_{t-2|t-2}] + W y_t \\
&\quad -\beta W [\beta W \hat{e}_{t-2|t-2} + (1 - W) y_{t-1} - (1 - W)\hat{\theta}_{t-2|t-2}] \\[2ex]
&= (1 - W)[(1 - W) + \beta W]\hat{\theta}_{t-2|t-2} - \beta W[(1 - W) + \beta W]\hat{e}_{t-2|t-2} \\
&\quad + W y_t + W(1 - W)(1 - \beta) y_{t-1} \\[2ex]
&= (1 - W)[(1 - W) + \beta W]\{(1 - W)\hat{\theta}_{t-3|t-3} + W y_{t-2} - \beta W \hat{e}_{t-3|t-3}\} \\
&\quad -\beta W[(1 - W) + \beta W]\{\beta W \hat{e}_{t-3|t-3} + (1 - W) y_{t-2} - (1 - w)\hat{\theta}_{t-3|t-3}\} \\
&\quad + W y_t + W(1 - W)(1 - \beta) y_{t-1} \\[2ex]
&= [(1 - W)^3 + (1 - W)^2 \beta W]\hat{\theta}_{t-3|t-3} + [(1 - W)^2 + \beta W(1 - W)] W\, y_{t-2} \\
&\quad - \beta W[(1 - W)^2 + \beta W(1 - W)]\hat{e}_{t-3|t-3} - \beta^2 W^2 [(1 - W) + \beta W]\hat{e}_{t-3|t-3} \\
&\quad -\beta W[(1 - W) + \beta W](1 - W) y_{t-2} - \beta W(1 - W)[(1 - W) + \beta W]\hat{\theta}_{t-3|t-3} \\
&\quad + W y_t + W(1 - W)(1 - \beta) y_{t-1} \\[2ex]
&= [(1 - W)^3 + 2(1 - W)^2 \beta W + \beta^2 W^2 (1 - W)]\hat{\theta}_{t-3|t-3} \\
&\quad [-\beta W(1 - W)^2 - 2\beta^2 W^2 (1 - W) - \beta^3 W^3]\hat{e}_{t-3|t-3} \\
&\quad +[(1 - W)^2 W + \beta W^2 (1 - W) - \beta W(1 - W)^2 - \beta^2 W^2] y_{t-2} \\
&\quad + W y_t + W(1 - W)(1 - \beta) y_{t-1} \\[2ex]
&= (1 - W)[(1 - W)^2 + 2\beta W(1 - W) + \beta^2 W^2]\hat{\theta}_{t-3|t-3} \\
&\quad -\beta W[(1 - W)^2 + 2\beta W(1 - W) + \beta^2 W^2]\hat{e}_{t-3|t-3} \\
&\quad + W(1 - W)(1 - \beta)[1 - W(1 - \beta)] y_{t-2} + W(1 - W)(1 - \beta) y_{t-1} + W y_t \\[2ex]
&= (1 - W)[(1 - W) + \beta W]^2 \hat{\theta}_{t-3|t-3} - \beta W[(1 - W) + \beta W]^2 \hat{e}_{t-3|t-3} \\
&\quad + W(1 - W)(1 - \beta)[(1 - W) + \beta W] y_{t-2} + W(1 - W)(1 - \beta) y_{t-1} + W y_t \\
&= \quad \dots \dots \\
&= (1 - W)[(1 - W) + \beta W]^{t-1}\theta_0 - \beta W[(1 - W) + \beta W]^{t-1} e_0 \\
&\quad + W y_t + W(1 - W)(1 - \beta)\sum_{j=1}^{t-1} [(1 - W) + \beta W]^{j-1} y_{t-j}.
\end{aligned}$$

Appendix B4 - Final Expressions for $\tilde{p}_{12}$, $\tilde{p}_2$ and W

Using relations (5.17) and (5.18), the solution for the system of equations (5.16) is given by:

$$p_1 = \frac{(\sigma_a^2 + 3\beta^2\sigma_a^2 - 4\beta\sigma_a^2) \pm (\beta - 1)\sqrt{\beta^2(\sigma_a^2)^2 + 2\beta(\sigma_a^2)^2 + 4\sigma_b^2\sigma_a^2 + (\sigma_a^2)^2}}{2(\beta - 1)^2}, \tag{B4.1a}$$

$$p_{12} = -\beta p_1 + \beta\sigma_a^2 \quad , \tag{B4.1b}$$

$$p_2 = \frac{p_1^2 - 2\beta p_1^2 + 4p_1\beta\sigma_a^2 + \beta^2 p_1^2 - 2p_1\beta^2\sigma_a^2 + \beta^2(\sigma_a^2)^2 - \sigma_a^2 p_1 - 2\beta(\sigma_a^2)^2}{\sigma_a^2}. \tag{B4.1c}$$

Letting $\sigma_b^2 = \tau\sigma_a^2$, allows p_1 to be expressed as:

$$p_1 = \sigma_a^2 \left[\frac{(1 + 3\beta^2 - 4\beta) - (\beta - 1)\sqrt{\beta^2 + 2\beta + 4\tau + 1}}{2(\beta - 1)^2} \right] \quad , \tag{B4.2}$$

because the other expression in (B4.1a) would imply $p_1 p_2 - p_{12}^2 < 0$, for $|\beta| < 1$ and $\tau > 0$, in disagreement with (5.17).

Defining $\tilde{p}_1$ as:

$$\tilde{p}_1 = \left[\frac{(1 + 3\beta^2 - 4\beta) - (\beta - 1)\sqrt{\beta^2 + 2\beta + 4\tau + 1}}{2(\beta - 1)^2} \right] \quad , \tag{B4.3}$$

it follows that p_1 , p_{12} and p_2 can be expressed as

$$p_1 = \sigma_a^2\,\tilde{p}_1 \quad , \tag{B4.3a}$$

$$p_{12} = -\beta\,\sigma_a^2\,\tilde{p}_1 + \beta\,\sigma_a^2 = \sigma_a^2(-\beta\,\tilde{p}_1 + \beta) = \sigma_a^2\,\tilde{p}_{12} \quad , \qquad \textbf{(B4.3b)}$$

$$p_2 = \frac{1}{\sigma_a^2}\Big[(\sigma_a^2)^2\tilde{p}^2 - 2\beta(\sigma_a^2)^2\tilde{p}_1^2 + 4\beta(\sigma_a^2)^2\tilde{p}_1 + \beta^2(\sigma_a^2)^2\tilde{p}_1^2$$

$$-2\beta^2(\sigma_a^2)^2\tilde{p}_1 + \beta^2(\sigma_a^2)^2 - (\sigma_a^2)^2\tilde{p}_1 - 2\beta(\sigma_a^2)^2\Big] \qquad \textbf{(B4.3c)}$$

$$= \sigma_a^2[\ \tilde{p}_1^2 - 2\beta\,\tilde{p}_1^2 + 4\beta\,\tilde{p}_1 + \beta^2\tilde{p}_1^2 - 2\beta^2\tilde{p}_1 + \beta^2 - \tilde{p}_1 - 2\beta\,]$$

$$= \sigma_a^2\,\tilde{p}_2 \quad .$$

In addition, it becomes clear that W can be written as:

$$W = \frac{p_1 + p_{12}}{p_1 + 2p_{12} + p_2} = \frac{\sigma_a^2\tilde{p}_1 + \sigma_a^2\tilde{p}_{12}}{\sigma_a^2\tilde{p}_1 + 2\sigma_a^2\tilde{p}_{12} + \sigma_a^2\tilde{p}_2} = \frac{\tilde{p}_1 + \tilde{p}_{12}}{\tilde{p}_1 + 2\tilde{p}_{12} + \tilde{p}_2} \quad . \qquad \textbf{(B4.4))}$$

Hence by depending on $\tilde{p}_1$, $\tilde{p}_{12}$ and $\tilde{p}_2$, W is in turn a function of β and τ .

Appendix C1 - Equation (6.8)

$$\log(u_{mt}) = \tilde{u}_{mt} + O_p(n_t^{-1})$$

The following proof is based on Bell & Hillmer(1990).

Note that $\tilde{u}_{mt} = \frac{e_{mt}}{\theta_{mt}} = \frac{y_{mt} - \theta_{mt}}{\theta_{mt}}$ is the relative sampling error of the estimator y_{mt} . Therefore

$$E(\tilde{u}_{mt} | \theta_{mt}) = E\left(\frac{y_{mt} - \theta_{mt}}{\theta_{mt}} \middle| \theta_{mt} \right) = 0 \quad , \tag{C1.1}$$

if y_{mt} is design-unbiased for θ_{mt} , and

$$V(\tilde{u}_{mt} | \theta_{mt}) = V\left(\frac{y_{mt} - \theta_{mt}}{\theta_{mt}} \middle| \theta_{mt} \right) = \frac{V(y_{mt} - \theta_{mt} | \theta_{mt})}{\theta_{mt}^2} = \frac{V(y_{mt} | \theta_{mt})}{\theta_{mt}^2} \; . \tag{C1.2}$$

When estimating a population mean, it is often true

$$V(y_{mt} | \theta_{mt}) \leq \frac{K}{n_t} \quad , \tag{C1.3}$$

where n_t is the sample size at time t and K is some constant.
Equation (C1.3) can be expressed alternatively as

$$V(y_{mt} | \theta_{mt}) = O(n_t^{-1}) \tag{C1.4}$$

meaning that $V(y_{mt} | \theta_{mt})$ is of roughly the same order of magnitude as n_t^{-1} . This result is based on the theory of simple random sampling, although this relation can often be accepted as valid for other sample designs.

Note that (C1.2) together with (C1.3) imply that:

$$V(\tilde{u}_{mt} | \theta_{mt}) = \frac{V(y_{mt} | \theta_{mt})}{\theta_{mt}^2} \leq \frac{K / n_t}{\theta_{mt}^2} \leq \frac{K}{\theta_{mt}^2 \; n_t} \quad , \tag{C1.5}$$

then

$$V(\tilde{u}_{mt} | \theta_{mt}) = O(n_t^{-1}) \quad , \qquad \textbf{(C1.6)}$$

provided θ_{mt} is bounded away from zero.

Putting (C1.6) and (C1.1) together, results in:

$$V(\tilde{u}_{mt} | \theta_{mt}) = E(\tilde{u}_{mt}^2 | \theta_{mt}) = O(n_t^{-1}) \quad .$$

Now, using Theorem 6.2.1 from Wolter(1985, p.222), it follows that if

$$E(\tilde{u}_{mt}^2 | \theta_{mt}) = O(n_t^{-1}) \qquad \textbf{(C1.7a)}$$

then

$$\tilde{u}_{mt} = O_p(n_t^{-1/2}) \quad . \qquad \textbf{(C1.7b)}$$

Therefore,

$$u_{mt} = 1 + \tilde{u}_{mt} = 1 + O_p(n_t^{-1/2}) \quad . \qquad \textbf{(C1.8)}$$

Wolter(1985, p.223) provides the Taylor approximation for a continuous function $g(.)$ of a random variable u around a as follows:

$$g(u) = g(a) + g'(a)(u-a) + R(u,a) \quad , \qquad \textbf{(C1.9)}$$

where $g'(a)$ is the first derivative of $g(.)$ evaluated at a and $R(u,a)$ is the remainder series.

In addition, theorem 6.2.2 from Wolter(1985, p.223) states that, if

$$u = a + O_p(b_n) \quad , \qquad \textbf{(C1.10a)}$$

then $g(u)$ can be expressed using (C1.9) with

$$R(u,a) = O_p(b_n^2) \quad . \qquad \textbf{(C1.10b)}$$

Therefore, using (C1.9) and (C1.10) for the case in which

$$g(u) = \log(u_{mt}) = \log(1 + \tilde{u}_{mt}) = \log(1 + \mathrm{O}_p(n_t^{-1/2})) \quad ,$$

with $a = 1$ and $\mathrm{O}_p(b_n) = \mathrm{O}_p(n_t^{-1/2})$, it follows that

$$\begin{aligned} \log(u_{mt}) &= \log(1) + \left.\frac{1}{u_{mt}}\right|_{u_{mt}=1} (u_{mt} - 1) + \mathrm{O}_p(n_t^{-1}) \\ &= 0 + u_{mt} - 1 + \mathrm{O}_p(n_t^{-1}) \\ &= 1 + \tilde{u}_{mt} - 1 + \mathrm{O}_p(n_t^{-1}) \\ &= \tilde{u}_{mt} + \mathrm{O}_p(n_t^{-1}) \quad , \end{aligned}$$

Which proves expression (6.8).

Appendix C2 - Equations (6.45a) and (6.45b) $E[\gamma_t|D]$ and $V[\gamma_t|D]$

Following Lindley(1965, pp.134-135), consider a function $z = g(x)$ of a random variable x such that $E(x) = \mu$. Assume that the function $g(.)$ is differentiable up to a second-order and assume in addition that $V(x)$ is finite. The Taylor approximation for $E[z] = E[g(x)]$ is given by

$$E[z] = E[g(x)] \approx g(\mu) + 0.5\, g''(\mu)\, V(x) \quad , \tag{C2.1}$$

and an approximation for the variance has the form

$$V[z] = V[g(x)] \approx [g'(\mu)]^2 \; V(x) \quad . \tag{C2.2}$$

Let

$$x_t = \left\{ \frac{\theta_{2t}}{\theta_{1t}} + 1 \right\} \quad . \tag{C2.3}$$

Then

$$E[x_t\,|\,D] = E\left[\frac{\theta_{2t}}{\theta_{1t}} |D \right] + 1 \quad , \tag{C2.4}$$

$$V[x_t|D] = V\left[\frac{\theta_{2t}}{\theta_{1t}} |D \right] \quad . \tag{C2.5}$$

Now noting that $\gamma_t = g(x_t) = x_t^{-1}$ and also that $g'(x_t) = -x_t^{-2}$ and $g''(x_t) = 2x_t^{-3}$, it follows, by putting (C2.1) to (C2.5) together,that

$$
\begin{aligned}
E[\gamma_t | D] &\approx E^{-1}[x_t | D] + \frac{V[x_t | D]}{E^3[x_t | D]} \\
&= \left\{ E\left[\frac{\theta_{2t}}{\theta_{1t}} | D\right] + 1 \right\}^{-1} + \frac{V\left[\frac{\theta_{2t}}{\theta_{1t}} | D\right]}{\left\{ E\left[\frac{\theta_{2t}}{\theta_{1t}} | D\right] + 1 \right\}^{3}} \quad ,
\end{aligned}
\qquad \textbf{(C2.6)}
$$

and

$$
\begin{aligned}
V[\gamma_t | D] &\approx (-E[x_t | D]^{-2})^2 \, V[x_t | D] \\
&= \left\{ E\left[\frac{\theta_{2t}}{\theta_{1t}} | D\right] + 1 \right\}^{-4} V\left[\frac{\theta_{2t}}{\theta_{1t}} | D\right] \quad ,
\end{aligned}
\qquad \textbf{(C2.7)}
$$

as in Chapter 6, Equations (6.45a) and (6.45b).

Appendix D1 - Section 7.2.2

$$CORR(\tilde{e}_{t-h}^{(k)}, \tilde{e}_t^{(k)}) = CORR(e_{t-h}^{(k)}, e_t^{(k)})$$

A pseudo error is defined as

$$\tilde{e}_t^{(k)} = y_t^{(k)} - y_t \quad , \tag{D1.1}$$

where $y_t = \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)}$. If there is no rotation bias, it follows that:

$$\begin{aligned} \tilde{e}_t^{(k)} &= y_t^{(k)} - y_t = y_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)} \\ &= (y_t^{(k)} - \theta_t) - \frac{1}{K}\sum_{k=1}^{K} (y_t^{(k)} - \theta_t) \\ &= e_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)} = e_t^{(k)} - e_t \quad . \end{aligned} \tag{D1.2}$$

From (D1.2), $COV(\tilde{e}_{t-h}^{(k)}, \tilde{e}_t^{(k)})$ is equal to

$$\begin{aligned} COV(e_{t-h}^{(k)} - e_{t-h}, e_t^{(k)} - e_t) &= COV(e_{t-h}^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_{t-h}^{(k)}, e_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)}) \\ = COV(e_{t-h}^{(k)}, e_t^{(k)}) &- \frac{1}{K}\sum_{j=1}^{K} COV(e_{t-h}^{(k)}, e_t^{(j)}) \\ - \frac{1}{K}\sum_{j=1}^{K} COV(e_{t-h}^{(j)}, e_t^{(k)}) &+ \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} COV(e_{t-h}^{(i)}, e_t^{(j)}) \quad . \end{aligned} \tag{D1.3}$$

It is assumed that

(i) $COV(e_{t-h}^{(i)}, e_t^{(k)}) = 0$ if $i \neq k \quad \forall t,h$, that is, in the case of no overlap between rotation groups the sampling errors are uncorrelated;

(ii) $COV(e_{t-h}^{(k)}, e_t^{(k)}) = COV(e_{t-h}^{(i)}, e_t^{(i)}) \quad \forall\ t,h$ for $i,k = 1,\ldots,K$, that is, the autocovariance depend on the lags but not on the rotation groups.

Using *(i)* and *(ii)*, results in:

$$COV(e_{t-h}^{(k)} - e_{t-h}, e_t^{(k)} - e_t) = \left[1 - \frac{1}{K} \right] COV(e_{t-h}^{(k)}, e_t^{(k)}) \quad . \tag{D1.4}$$

Also, assuming $\{e_t^{(k)}\}$ $\forall\, k$ stationary,

$$\begin{aligned} COV(e_t^{(k)} - e_t, e_t^{(k)} - e_t) &= COV(e_{t-h}^{(k)} - e_{t-h}, e_{t-h}^{(k)} - e_{t-h}) \\ &= \left[1 - \frac{1}{K} \right] COV(e_t^{(k)}, e_t^{(k)}) \quad . \end{aligned} \tag{D1.5}$$

Since

$$CORR(\tilde{e}_{t-h}^{(k)}, \tilde{e}_t^{(k)}) = \frac{COV(\tilde{e}_{t-h}^{(k)}, \tilde{e}_t^{(k)})}{COV(\tilde{e}_{t-h}^{(k)}, \tilde{e}_{t-h}^{(k)})^{\frac{1}{2}}\, COV(\tilde{e}_t^{(k)}, \tilde{e}_t^{(k)})^{\frac{1}{2}}} \quad ,$$

using (D1.4) and (D1.5), it follows that

$$CORR(\tilde{e}_{t-h}^{(k)}, \tilde{e}_t^{(k)}) = CORR(e_{t-h}^{(k)}, e_t^{(k)}) \quad .$$

Appendix D2 - Section 7.3.1

$$CORR(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_{t}) = CORR(e^{(k)}_{t-h}, e^{(k)}_{t})$$

$$CORR(\tilde{e}^{(h)}_{t-h}, \tilde{e}^{(k)}_{t}) \;=\; P^{(k)}_{\tilde{e}}(h) \;=\; (D^{(k)}_{\tilde{e}})^{-1/2}\;\Gamma^{(k)}_{\tilde{e}}(h)(D^{(k)}_{\tilde{e}})^{-1/2} \quad , \tag{D2.1}$$

with

$$\{\Gamma^{(k)}_{\tilde{e}}(h)\}_{ml} \;=\; COV(\tilde{e}^{(k)}_{m,t-h}, \tilde{e}^{(k)}_{lt}) \quad ,$$

and

$$D^{(k)}_{\tilde{e}} \;=\; diag(\,VAR(\tilde{e}^{(k)}_{1t}), \ldots, VAR(\tilde{e}^{(k)}_{Mt})\,) \quad .$$

Note that, if there is no rotation bias and if y_{mt} is an unbiased estimator for θ_{mt} , then $e^{(k)}_{mt} \;=\; y^{(k)}_{mt} \;-\; \theta_{mt}$ for $m = 1, \ldots, M$ and

$$\begin{aligned}
\tilde{e}^{(k)}_{t} \;&=\; y^{(k)}_{t} - y_{t} \;=\; y^{(k)}_{t} \;-\; \frac{1}{K}\sum_{k=1}^{K} y^{(k)}_{t} \\
&=\; (y^{(k)}_{t} \;-\; \theta_{t}) \;-\; \frac{1}{K}\sum_{k=1}^{K}(y^{(k)}_{t} \;-\; \theta_{t}) \\
&=\; e^{(k)}_{t} - \frac{1}{K}\sum_{k=1}^{K} e^{(k)}_{t} \;=\; e^{(k)}_{t} \;-\; e_{t} \quad .
\end{aligned} \tag{D2.2}$$

Then (D2.2) implies that

$$COV(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_{t}) \;=\; COV(\,e^{(k)}_{t-h} - e_{t-h}, e^{(k)}_{t} - e_{t}) \quad . \tag{D2.3}$$

Using the properties of covariance matrices of random vectors (see Reinsel,1993, p.13) the covariance in (2.3) becomes:

$$\begin{aligned}
COV(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_{t}) \;=\; \quad & COV(e^{(k)}_{t-h}, e^{(k)}_{t}) \;-\; COV(e^{(k)}_{t-h}, e_{t}) \\
& -\; COV(e_{t-h}, e^{(k)}_{t}) \;+\; COV(e_{t-h}, e_{t}) .
\end{aligned} \tag{D2.4}$$

It is assumed that

(i) $COV(e^{(i)}_{t-h}, e^{(k)}_{t}) \;=\; \mathbf{0}$ if $i \neq k \quad \forall\, t, h$, that is, in the case of no overlap between rotation groups the sampling errors are uncorrelated;

(ii) the sampling error autocovariances vary between characteristics and depend on the lags but not on the rotation groups, that is, $COV(e^{(k)}_{t-h}, e^{(k)}_t) = COV(e^{(i)}_{t-h}, e^{(i)}_t) = \mathbf{\Gamma}^{(k)}_e(h) \quad \forall\ t,h$ for $i,k = 1,\dots,K$.

Using *(i)*, *(ii)* and properties of the covariance of random vectors (see Reinsel,1993, p.13) one gets:

$$COV(e^{(k)}_{t-h}, e^{(k)}_t) = \mathbf{\Gamma}^{(k)}_e(h) \quad ; \tag{D2.5}$$

$$\begin{aligned} COV(e^{(k)}_{t-h}, e_t) &= COV(e^{(k)}_{t-h}, \frac{1}{K}\sum_{j=1}^{K} e^{(k)}_t) = \frac{1}{K}\sum_{j=1}^{K} COV(e^{(k)}_{t-h}, e^{(j)}_t) \\ &= \frac{1}{K}\mathbf{\Gamma}^{(k)}_e(h) \quad ; \end{aligned} \tag{D2.6}$$

$$\begin{aligned} COV(e_{t-h}, e_t) &= COV(\frac{1}{K}\sum_{i=k}^{K} e^{(i)}_{t-h}, \frac{1}{K}\sum_{j=1}^{K} e^{(k)}_t) \\ &= \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} COV(e^{(k)}_{t-h}, e^{(j)}_t) \quad = \frac{K}{K^2}\,\mathbf{\Gamma}^{(k)}_e(h) \quad . \end{aligned} \tag{D2.7}$$

Substituting (D2.5)-(D2.7) into (D2.4) results in

$$\mathbf{\Gamma}^{(k)}_{\tilde{e}}(h) = COV(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_t) = \left[1 - \frac{1}{K}\right] \mathbf{\Gamma}^{(k)}_e(h) \quad , \tag{D2.8}$$

In addition, $\mathbf{D}^{(k)}_{\tilde{e}} = diag(VAR(\tilde{e}^{(k)}_{1t}), \dots, VAR(\tilde{e}^{(k)}_{Mt}))$ with

$$VAR(\tilde{e}^{(k)}_{mt}) = (1 - 1/K)\, VAR(e^{(k)}_{mt}) \quad . \tag{D2.9}$$

Then, from (D2.8) and (D2.9), it follows that

$$\begin{aligned} \mathbf{\Gamma}^{(k)}_{\tilde{e}}(h) &= CORR(\tilde{e}^{(k)}_{t-h}, \tilde{e}^{(k)}_t) \\ &= \left[1 - \frac{1}{K}\right]^{-1/2} (\mathbf{D}^{(k)}_e)^{-1/2} \left[1 - \frac{1}{K}\right] \mathbf{\Gamma}^{(k)}_e(h) \left[1 - \frac{1}{K}\right]^{-1/2} (\mathbf{D}^{(k)}_e)^{-1/2} \\ &= (\mathbf{D}^{(k)}_e)^{-1/2}\ \mathbf{\Gamma}^{(k)}_e(h)\ (\mathbf{D}^{(k)}_e)^{-1/2} = CORR(e^{(k)}_{t-h}, e^{(k)}_t) \quad . \end{aligned}$$

Appendix D3 - Cross-Covariance and Cross-Correlation Matrices of Vector ARMA Models

The following results can be found in Reinsel(1993, Chapter 2), Wei(1993, Chapter 14) and Liu(1986, Chapter 5).

Let $z_t = (z_{1t}, z_{2t}, \dots, z_{Mt})'$ with $t = \pm 1, \pm 2, \dots$ be a M-dimensional jointly stationary real valued process so that the mean $E(z_{mt}) = E(z_m)$ is constant for $m = 1, \dots, M$, and the cross-covariance between $z_{m,t-h}$ and z_{lt} for all $m, l = 1, \dots, M$ are functions only of the lag h . The cross-covariance matrix function of $\{Z_t\}$ is given by:

$$\Gamma_z(h) = COV(z_{t-h}, z_t) = E(Z_{t-h}\ Z_t') \quad ,$$

where $Z_t = z_t - E(z_t)$ and $\{\Gamma_z(h)\}_{ml} = \gamma_{z\,ml}(h) = COV(z_{m,t-h}, z_{lt})$.

The cross-correlation matrix function for a vector process is defined by:

$$\mathbf{P}_z(h) = D_z^{-1/2}\, \Gamma_z(h)\, D_z^{-1/2} \quad ,$$

where D_z is a diagonal matrix in which the mth diagonal element is the variance of the mth process. That is

$$D_z = diag(\gamma_{z\,11}(0), \dots, \gamma_{z\,MM}(0)) \quad ,$$

and $\{P_z(h)\}_{ml} = \rho_{z\,ml}(h) = \dfrac{\gamma_{z\,ml}(h)}{[\gamma_{z\,mm}(0)\ \gamma_{z\,ll}(0)]^{1/2}}$.

Note that, since

$$\begin{aligned} \gamma_{z\,ml}(h) &= E[(z_{m,t-h} - E(z_m))(z_{lt)} - E(z_l))] \\ &= E[(z_{lt} - E(z_l)) - (z_{m,t-h} - E(z_m))] \\ &= \gamma_{z\,lm}(-h) \quad . \end{aligned}$$

it follows that

$$\begin{cases} \boldsymbol{\Gamma}_z(h) = \boldsymbol{\Gamma}_z'(-h) \;, \\ \boldsymbol{P}_z(h) = \boldsymbol{P}_z'(-h) \;. \end{cases}$$

A vector autoregressive moving average model of orders p and q (VARMA(p,q)) for a M-dimensional multiple time series $\{z_t\}$ (with mean vector $E(z_t)$) is given by

$\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\delta}(B)\boldsymbol{a}_t$, where $\boldsymbol{Z}_t = z_t - E(z_t)$,

$\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \dots - \boldsymbol{\Phi}_p B^p$,

$\boldsymbol{\delta}(B) = \boldsymbol{I} - \boldsymbol{\delta}_1 B - \dots - \boldsymbol{\delta}_q B^q$ and $\boldsymbol{a}_t$ is a M-dimensional white noise random vector with zero mean and covariance structure:

$$E(\boldsymbol{a}_{t-h}\boldsymbol{a}_t') = \begin{cases} \boldsymbol{\Sigma}_a & h = 0 \;; \\ \boldsymbol{0} & h \neq 0 \;. \end{cases}$$

For a **Vector MA(q)** process

$$\boldsymbol{Z}_t = (\boldsymbol{I} - \boldsymbol{\delta}_1 B - \dots - \boldsymbol{\delta}_q B^q)\boldsymbol{a}_t \quad , \qquad \textbf{(D3.1)}$$

it can be shown that

$$\boldsymbol{\Gamma}_z(h) = \begin{cases} \sum_{j=0}^{q-h} \boldsymbol{\delta}_j \; \boldsymbol{\Sigma}_a \; \boldsymbol{\delta}_{j+h}' & h = 0, \dots, q \;, \\ 0 & h > q \;, \end{cases} \qquad \textbf{(D3.2)}$$

where $\boldsymbol{\delta}_0 = -\boldsymbol{I}$. Note that the cross-covariance cuts off after lag q. Therefore the cross-correlation matrix for the Vector MA(q) process in (D3.1) is given by:

$$\boldsymbol{P}_z(h) = \boldsymbol{0} \quad \forall\, h > q \quad .$$

For a **Vector MA(1)** process one gets

$$\boldsymbol{\Gamma}_z(h) = \begin{cases} \sum_{j=0}^{1-h} \boldsymbol{\delta}_j \, \boldsymbol{\Sigma}_a \, \boldsymbol{\delta}_{j+h}' & h=0,1 \ , \\ \\ 0 & h>1 \ , \end{cases} \qquad \textbf{(D3.3)}$$

with

$$\begin{cases} \boldsymbol{\Gamma}_z(0) = \boldsymbol{\Sigma}_a + \boldsymbol{\delta}_1 \, \boldsymbol{\Sigma}_a \, \boldsymbol{\delta}_1' \ , \\ \\ \boldsymbol{\Gamma}_z(1) = -\boldsymbol{\Sigma}_a \boldsymbol{\delta}_1' \ . \end{cases} \qquad \textbf{(D3.4)}$$

Then it becomes clear that the cross-covariance and the cross-correlation matrix functions for a Vector MA(1) process cut off after lag 1.

The general **Vector AR(p)** process is given by

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 B - \dots - \boldsymbol{\Phi}_p B^p)\boldsymbol{Z}_t = \boldsymbol{a}_t \quad . \qquad \textbf{(D3.5)}$$

The cross-covariance matrix of (D3.5) is

$$\boldsymbol{\Gamma}_z(h) = \begin{cases} \sum_{j=1}^{p} \boldsymbol{\Gamma}_z(-j)\boldsymbol{\Phi}_j' + \boldsymbol{\Sigma}_a & h=0 \ ; \\ \\ \sum_{j=1}^{p} \boldsymbol{\Gamma}_z(h-j)\boldsymbol{\Phi}_j' & h=1,2,\dots \ . \end{cases} \qquad \textbf{(D3.6)}$$

For a stationary **Vector AR(1)** process

$$\boldsymbol{\Gamma}_z(h) = \begin{cases} \boldsymbol{\Gamma}_z(-1)\,\boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}_a & h=0 \ , \\ \boldsymbol{\Gamma}_z(h-1)\,\boldsymbol{\Phi}_1' & h \geq 1 \ , \end{cases} \qquad \textbf{(D3.7)}$$

yielding

$$\left\{\begin{array}{l} \boldsymbol{\Gamma}_z(0) = \boldsymbol{\Gamma}_z(1)' \, \boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}_a = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}_z(0) \, \boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}_a \,, \\ \boldsymbol{\Gamma}_z(1) = \boldsymbol{\Gamma}_z(0) \, \boldsymbol{\phi}_1' \,, \\ \boldsymbol{\Gamma}_z(2) = \boldsymbol{\Gamma}_z(1) \, \boldsymbol{\Phi}_1' = \boldsymbol{\Gamma}_z(0) \boldsymbol{\Phi}_1'^{2} \,, \\ \dots \; \dots \; \dots \; \dots \\ \boldsymbol{\Gamma}_z(h) = \boldsymbol{\Gamma}_z(0) \, \boldsymbol{\Phi}_1'^{\,h} \,, \end{array}\right. \tag{D3.8}$$

which implies

$$\begin{aligned} \mathbf{P}_z(h) &= D_z^{-1/2} \boldsymbol{\Gamma}_z(h) \, D_z^{-1/2} = D_z^{-1/2} \, \boldsymbol{\Gamma}_z(0) \, \boldsymbol{\Phi}_1'^{\,h} D_z^{-1/2} \\ &= D_z^{-1/2} \, \boldsymbol{\Gamma}_z(0) \, D_z^{-1/2} D_z^{1/2} \, \boldsymbol{\Phi}_1'^{\,h} D_z^{-1/2} \\ &= \mathbf{P}_z(0) \, (\boldsymbol{\Phi}_1^{*})^{h} \quad h \geq 1 \;\;, \end{aligned} \tag{D3.9}$$

with $(\boldsymbol{\Phi}_1^{*})^{h} = D_z^{1/2} \, \boldsymbol{\Phi}_1'^{\,h} D_z^{-1/2}$.

The cross-covariance matrix of a **Vector ARMA(p,q)** process is given by

$$\boldsymbol{\Gamma}_z(h) = \left\{\begin{array}{ll} \displaystyle\sum_{j=1}^{p} \boldsymbol{\Gamma}_z(h-j) \, \boldsymbol{\Phi}_j' - \sum_{j=h}^{q} \psi_{j-h} \boldsymbol{\Sigma}_a \, \delta_j' & h = 0, \dots, q \;, \\ \displaystyle\sum_{j=1}^{p} \boldsymbol{\Gamma}_z(h-j) \, \boldsymbol{\Phi}_j' & h > q \;, \end{array}\right. \tag{D3.10}$$

with $\psi(B) = \boldsymbol{\Phi}(B)^{-1} \, \delta(B)$.

For a stationary **Vector ARMA(1,1)** process

$$\boldsymbol{\Gamma}_z(h) = \left\{\begin{array}{ll} \displaystyle\boldsymbol{\Gamma}_z(h-1) \, \boldsymbol{\Phi}_1' - \sum_{j=0}^{1} \psi_{j-h} \boldsymbol{\Sigma}_a \, \delta_j' & h = 0, 1 \;, \\ \boldsymbol{\Gamma}_z(h-1) \, \boldsymbol{\Phi}_1' & h \geq 1 \;, \end{array}\right. \tag{D3.11}$$

yielding

$$\left\{ \begin{array}{l} \mathbf{\Gamma}_z(0) = \mathbf{\Gamma}_z(1)' \, \mathbf{\Phi}_1' + \mathbf{\Sigma}_a - \psi_1 \mathbf{\Sigma}_a \delta_1' \\ \mathbf{\Gamma}_z(1) = \mathbf{\Gamma}_z(0) \, \phi_1' - \mathbf{\Sigma}_a \delta_1' \, , \\ \mathbf{\Gamma}_z(2) = \mathbf{\Gamma}_z(1) \, \mathbf{\Phi}_1' \, , \\ \cdots \; \cdots \; \cdots \; \cdots \\ \mathbf{\Gamma}_z(h) = \mathbf{\Gamma}_z(h-1) \, \mathbf{\Phi}_1' = \mathbf{\Gamma}_z(1) \, \mathbf{\Phi}_1'^{h-1} \, , \end{array} \right. \tag{D3.12}$$

which implies

$$\begin{aligned} \mathbf{P}_z(h) &= \boldsymbol{D}_z^{-1/2} \mathbf{\Gamma}_z(h) \, \boldsymbol{D}_z^{-1/2} = \boldsymbol{D}_z^{-1/2} \, \mathbf{\Gamma}_z(1) \, \mathbf{\Phi}_1'^{h-1} \, \boldsymbol{D}_z^{-1/2} \\ &= \boldsymbol{D}_z^{-1/2} \, \mathbf{\Gamma}_z(1) \, \boldsymbol{D}_z^{-1/2} \, \boldsymbol{D}_z^{1/2} \, \mathbf{\Phi}_1'^{h-1} \, \boldsymbol{D}_z^{-1/2} \\ &= \mathbf{P}_z(1) \, (\mathbf{\Phi}_1^*)^{h-1} \quad , \quad h \geq 2 \quad , \end{aligned} \tag{3.15}$$

with $(\mathbf{\Phi}_1^*)^{h-1} = \boldsymbol{D}_z^{1/2} \, \mathbf{\Phi}_1'^{h-1} \, \boldsymbol{D}_z^{-1/2}$.

Appendix D4 - Univariate Models Implied by a Vector Autoregressive Model

Let $z_t = (z_{1t}, z_{2t}, \dots, z_{Mt})'$ with $t = \pm 1, \pm 2, \dots$ be a M-dimensional jointly stationary real valued process.

The general **Vector AR(p)** process for a M-dimensional multiple time series $\{z_t\}$ (with mean vector $E(z_t)$) is given by

$$\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{a}_t \qquad \textbf{(D4.1)}$$

where $Z_t = z_t - E(z_t)$, $\boldsymbol{\Phi}(B) = I - \boldsymbol{\Phi}_1 B - \dots - \boldsymbol{\Phi}_p B^p$ and $\boldsymbol{a}_t$ is a M-dimensional white noise random vector.

To obtain the univariate representation of Z_{mt} $(m = 1, \dots, M)$ express (D4.1) as

$$\boldsymbol{Z}_t = [\boldsymbol{\Phi}(B)]^{-1} \boldsymbol{a}_t = |\boldsymbol{\Phi}(B)|^{-1} \boldsymbol{\Phi}^*(B)\, \boldsymbol{a}_t \quad , \qquad \textbf{(D4.2)}$$

where $|\bullet|$ denotes determinant of a matrix and $\boldsymbol{\Phi}^*(B)$ is the adjoint matrix of $\boldsymbol{\Phi}(B)$ (see Reinsel, 1993, p.29, Maravall & Mathis, 1994 and Chan & Wallis, 1978).

Multiplying both sides of (D4.2) by $|\boldsymbol{\Phi}(B)|$ results in

$$|\boldsymbol{\Phi}(B)|\, \boldsymbol{Z}_t = \boldsymbol{\Phi}^*(B)\, \boldsymbol{a}_t \quad . \qquad \textbf{(D4.3)}$$

Writing the relation (D4.3) as

$$\mathbf{Y}(B)\, \boldsymbol{Z}_t = \boldsymbol{\Xi}(B)\, \boldsymbol{a}_t \quad , \qquad \textbf{(D4.4)}$$

it becomes clear that (D4.4) is an alternative Vector ARMA representation in which the AR coefficient matrix is diagonal.

As an example (from Liu, 1986), consider the simple stationary bivariate AR(1) model $(I - \boldsymbol{\Phi}_1 B)\boldsymbol{Z}_t = \boldsymbol{a}_t$ with

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \quad .$$

Note that:

$$(i)\ |\Phi(B)| = |(I - \Phi_1(B))| = \begin{vmatrix} 1-\phi_{11}B & -\phi_{12}B \\ -\phi_{21}B & 1-\phi_{22}B \end{vmatrix}$$

$$= (1-\phi_{11}B)(1-\phi_{22}B) - \phi_{12}\phi_{21}B^2 \quad ;$$

$$(ii)\ \Phi^*(B) = \begin{vmatrix} 1-\phi_{22}B & \phi_{12}B \\ \phi_{21}B & 1-\phi_{11}B \end{vmatrix} \quad .$$

Substituting *(i)* and *(ii)* in (D4.3) yields

$$[(1-\phi_{11}B)(1-\phi_{22}B) - \phi_{12}\phi_{21}B^2] \begin{bmatrix} Z_{1t} \\ Z_{2t} \end{bmatrix} = \begin{bmatrix} 1-\phi_{11}B & -\phi_{12}B \\ -\phi_{21}B & 1-\phi_{22}B \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} . \qquad \textbf{(D4.5)}$$

Thus, each series of a bivariate AR(1) model follows an univariate ARMA(2,1) model. Note that if $\phi_{12} = \phi_{21} = 0$, (D4.5) becomes

$$(1-\phi_{11}B)(1-\phi_{22}B)\,Z_{1t} = (1-\phi_{22}B)\,a_{1t} \quad ,$$
$$(1-\phi_{11}B)(1-\phi_{22}B)\,Z_{2t} = (1-\phi_{11}B)\,a_{2t} \quad ,$$

resulting in

$$(1-\phi_{11}B)\,Z_{1t} = a_{1t} \quad ,$$
$$(1-\phi_{22}B)\,Z_{2t} = a_{2t} \quad .$$

Hence, if $\phi_{ij}\ (i \neq j)$, each series would follow an AR(1) model.

Appendix E1 - SAS/IML program for Identification of Multiple Time Series Models for the Sampling Error Process

```
/* program errormul.sas ---------------------------------------

   computes the autocovariances, the crosscorrelations,
   the autoregression and partial lag correlation matrices of a
   multiple time series process of sampling errors. It also
   computes preliminary parameter estimates for Vector AR(p)

-------------------------------------------------------------*/
options linesize=72 ps=58 nodate;
options mprint symbolgen;

/* the Brazilian Labour Force Survey is a rotating panel survey
in which 4 panels are enumerated per month. Each of the following
data sets contains the series of transformed pseudo-errors for
a specific panel */

filename in1      'c:/denise/data/pseudo1.dat';
filename in2      'c:/denise/data/pseudo2.dat';
filename in3      'c:/denise/data/pseudo3.dat';
filename in4      'c:/denise/data/pseudo4.dat';

/* files with the subroutines for model identification
   and parameter estimation */

filename identmul 'c:/denise/sas/identmul.mac';
filename estimvar 'c:/denise/sas/estimvar.mac';


/* Reads the transformed pseudo-errors for a given
   rotation group. The values are multiplied by 1,000 to
   increase the number of significant digits when caculating
   the autocovariances  */

%MACRO READ;
 %do rg=1 %to 4;
    data pseudo&rg;
       infile in&rg;
         input year month panel $ peemp peune;
    peemp = peemp * 1000;
    peune = peune * 1000;
 %end;
%MEND;
%READ;

TITLE1 'BRAZILIAN LABOUR FORCE SURVEY';
TITLE2 'IDENTIFICATION OF THE TIME SERIES MODEL';
TITLE3 'FOR THE TRANSFORMED SAMPLING ERRORS';
run;

PROC IML;
 * RESET LOG; /* use this option if you want to have the results
   printed in the .LOG . in this case no .LST wil be created */

%MACRO MATRIX;

%do rg=1 %to 4;

/* Inputs the bivariate series of the pseudo errors
   (per rotation group) into IML */

START READDATA;
  USE PSEUDO&rg;
   READ ALL INTO RG&rg  VAR {PEEMP PEUNE};
  CLOSE PSEUDO&rg;
FINISH READDATA;
RUN READDATA;

/**************************************************************
* The autocovariances are computed for each rotation group and*
*then averaged to produce the autocovariances of the sampling *
*error series. Note that the autocovariance of the sampling   *
*errors is equal the sum of the autocovariances of the pseudo *
*errors divided by (k^2 - k), where k in the number of panels *
**************************************************************/
/* calculates autocovariances  matrices of lag = 0 to 24,
   the autocovariance is defined as
   E[(y_t - ymean_t)(y_t+h - ymean_t+h)'] */

START AUTOCOV;
TT= NROW(RG&rg); /* length of the series */
M = NCOL(RG&rg);/*dimension of the series, in this example M=2*/
```

```
GAMMA&rg = J(M*24,M,0); /* vertical concatenation of gamma1 to
                            gamma24*/
GAMMAT&rg = GAMMA&rg;
MEANRG= RG&rg[:,];
VECMEAN= REPEAT(MEANRG,TT,1);
CENTER = RG&rg - VECMEAN;

DO H=0 TO 24  BY 1;

CENTER1= CENTER[1:TT-H,];
CENTER2= CENTER[H+1:TT,];

  IF H = 0 THEN DO; /* note that peemp and peune have been
                       multiplied by 1000 */
    GAMMA0&rg =  (CENTER1` * CENTER2)/(1000000#TT);
  END;
  ELSE DO;
     GAMMA&rg[M*(H-1)+1:M*(H-1)+M,1:M] =
                     (CENTER1`*CENTER2)/(1000000#TT);
     GAMMAT&rg[M*(H-1)+1:M*(H-1)+M,1:M] =
                     ((CENTER1`*CENTER2)/(1000000#TT))`;
  END;
END;
FREE MEANRG VECMEAN CENTER CENTER1 CENTER2;
FINISH AUTOCOV;
RUN AUTOCOV;

%end;

%MEND;

%MATRIX;

/* computes de average of the pseudo-errors autocovariance
   matrices  note that k^2 - k = 12 for the Brazilian Labour
   Force Survey  (K=4)*/

START MEANCOV;
GAMMA0 = ( GAMMA01 + GAMMA02 + GAMMA03 + GAMMA04) / 12;
GAMMA  = ( GAMMA1  + GAMMA2  + GAMMA3  + GAMMA4 ) / 12;
GAMMAT = ( GAMMAT1 + GAMMAT2 + GAMMAT3 + GAMMAT4) / 12;
```

```
FREE GAMMA01 GAMMA02 GAMMA03 GAMMA04 GAMMA1 GAMMA2 GAMMA3 GAMMA4
     GAMMAT1 GAMMAT2 GAMMAT3 GAMMAT4;

GAMMA1 = GAMMA[1:M,1:M];
GAMMA2 = GAMMA[M+1:M+M,1:M];

print gamma0  GAMMA1 GAMMA2;
FINISH MEANCOV;
RUN MEANCOV;

%include identmul; /* calls routine which computes the cross-
                      correlations and partial lag correlation*/
%identmul;

%include estimvar; /* call the routine which computes the
                      preliminary estimates*/

%estimvar(2); /* p=2 is the order of the VAR(p) */

QUIT;


/* macro identul.mac -----------------------------------------

 Computes the crosscorrelations, autoregression and
 partial lag correlation matrices of a multiple time series .
 It uses the covariances computed with the subroutines autocov
 and meancov.

   ==> REFERENCE ===>
        WEI(1993), pages 359-361

   ==> TO EXECUTE: %INCLUDE IDENTMUL;
                   %IDENTMUL
 -------------------------------------------------------- */
%MACRO IDENTMUL;
START IDENTMUL;

 /* computes inverse of gamma0 */
    IGAMMA0 = INV(GAMMA0);

 /* Inverse of square root of diagonal of GAMMA0 */
    D = INV(SQRT(DIAG(GAMMA0)));
```

```
/* crosscorrelation matrix function */
CORR=J(M,M*24,0);
RHO    = GAMMA;

DO H = 1 TO 24;

   RHO[M*(H-1)+1:M*(H-1)+M,] = D*GAMMA[M*(H-1)+1:M*(H-1)+M,]* D;

   CORR[,M*(H-1)+1:M*(H-1)+M] =
              ROUND(RHO[M*(H-1)+1:M*(H-1)+M,],0.001);

END;

/* prints the cross-correlation matrices */

CORR1_4    = CORR[1:M,1:4*M];
CORR5_8    = CORR[1:M,4*M+1:8*M];
COR9_12   = CORR[1:M,8*M+1:12*M];
COR13_16   = CORR[1:M,12*M+1:16*M];
COR17_20   = CORR[1:M,16*M+1:20*M];

LAG = J(1,20*M,'    ');

DO  J = 0 TO 19;
         LAG[1,J*M+1]  = ' LAG';
         NO = J+1;
         LAG[1,J*M+2] = CHAR(NO,3);

END;

LAG1  = LAG[1,1:4*M];
LAG2  = LAG[1,4*M+1:8*M];
LAG3  = LAG[1,8*M+1:12*M];
LAG4  = LAG[1,12*M+1:16*M];
LAG5  = LAG[1,16*M+1:20*M];

CORR1_4    = LAG1 // CHAR(CORR1_4 ,6,3);
CORR5_8   =  LAG2 // CHAR(CORR5_8,6,3);
COR9_12   = LAG3 // CHAR(COR9_12,6,3);
COR13_16   = LAG4 // CHAR(COR13_16,6,3);
COR17_20   = LAG5 // CHAR(COR17_20,6,3);
```

```
PRINT 'CROSS-CORRELATION MATRICES',
CORR1_4   , CORR5_8 ,
COR9_12 , COR13_16 ,
COR17_20 /;

FREE CORR CORR1_4 CORR5_8 COR9_12 COR13_16 COR17_20;


/*schematic representation of the cross-correlation matrices*/

GRAPH = J(20*M,M,'   .');

DO I = 1 TO 20*M;

   DO J = 1 TO M;

         IF RHO[I,J] <  -(2/SQRT(TT)) THEN GRAPH[I,J] = '   -';
    ELSE IF RHO[I,J] >  ( 2/SQRT(TT)) THEN GRAPH[I,J] = '   +';

   END;
END;

CORR=J(M,20*M,'   .');

DO H = 1 TO 20;

CORR[,M*(H-1)+1:M*(H-1)+M] = GRAPH[M*(H-1)+1:M*(H-1)+M,];

END;

CORR1_4    = LAG1 // CORR[1:M,1:4*M];
CORR5_8    = LAG2 // CORR[1:M,4*M+1:8*M];
COR9_12    = LAG3 // CORR[1:M,8*M+1:12*M];
COR13_16   = LAG4 // CORR[1:M,12*M+1:16*M];
COR17_20   = LAG5 // CORR[1:M,16*M+1:20*M];

PRINT 'SCHEMATIC REPRESENTATION OF CROSS-CORRELATIONS',
      CORR1_4 , CORR5_8 , COR9_12 , COR13_16 , COR17_20 /;

FREE GRAPH CORR CORR1_4 CORR5_8 COR9_12 COR13_16 COR17_20;
```

```
/*********************************************
 Calculates partial autoregression matrices and
 partial lag correlation matrix function
 using algorithm in Wei(1993, p.359-361).
*********************************************/

DO H=1 TO 15;

 IF H=1 THEN DO;

    VU        = GAMMA0;
    VV        = GAMMA0;
    VVU       = GAMMA[1:M,];


    ALPHA = GAMMAT[1:M,] * IGAMMA0;
    BETA  = GAMMA[1:M,] * IGAMMA0;
    ALPHAT = ALPHA;

    DU = INV(SQRT(DIAG(VU)));
    DV = INV(SQRT(DIAG(VV)));

    LAGCORR = DV * VVU * DU;
    AUTOREG = (VVU)`* INV(VV);

 END;
 ELSE DO;

    VU = GAMMA0 - ALPHA * GAMMA[1:M*(H-1),];
    VV = GAMMA0 - BETA   * GAMMAT[1:M*(H-1),];
    VVU = GAMMA[M*(H-1)+1:M*(H-1)+M,] -
                           (ALPHAT * GAMMAT[1:M*(H-1),])` ;

    ALPHAHH = VVU` * INV(VV);
    BETAHH = VVU * INV(VU);


    IF H=2 THEN DO;
     ALPHAHK = ALPHA - ALPHAHH * BETA;
     ALPHAN = ALPHAHK || ALPHAHH;
     ALPHANT = ALPHAHH || ALPHAHK;
     BETAHK = BETA - BETAHH * ALPHA;
     BETAN = BETAHK || BETAHH;
    END;

    ELSE DO;

     ALPHAN = ALPHAHH;
     ALPHANT = ALPHAHH;
     BETAN = BETAHH;

     DO K = H-1 TO 1 BY -1;
      ALPHAHK = ALPHA[,M*K-1:M*K] -
                           ALPHAHH * BETA[,M*(H-K)-1:M*(H-K)];
      ALPHAN  = ALPHAHK || ALPHAN;
      ALPHANT = ALPHANT || ALPHAHK;
      BETAHK  = BETA[,M*K-1:M*K] -
                           BETAHH * ALPHA[,M*(H-K)-1:M*(H-K)];
      BETAN   = BETAHK || BETAN;
     END;

    END;

    DU = INV(SQRT(DIAG(VU)));
    DV = INV(SQRT(DIAG(VV)));

    LAGCORR = LAGCORR // ( DV * VVU * DU );
    AUTOREG = AUTOREG // ( (VVU)` * INV(VV) );

    ALPHA = ALPHAN;
    ALPHAT = ALPHANT;
    BETA   = BETAN;
 END;
END;

/* prints the autoregression and
   partial lag correlation matrices */

AUTO = J(M,M*15,0);
LAGC = J(M,M*15,0);

DO H = 1 TO 15;

AUTO[,M*(H-1)+1:M*(H-1)+M] =
ROUND(AUTOREG[M*(H-1)+1:M*(H-1)+M,],.001);
LAGC[,M*(H-1)+1:M*(H-1)+M] =
ROUND(LAGCORR[M*(H-1)+1:M*(H-1)+M,],.001);

END;

AR1_4    = LAG1 // CHAR(AUTO[1:M,1:4*M],6,3);
AR5_8    = LAG2 // CHAR(AUTO[1:M,4*M+1:8*M],6,3);
AR9_12   = LAG3 // CHAR(AUTO[1:M,8*M+1:12*M],6,3);


PAR1_4    = LAG1 // CHAR(LAGC[1:M,1:4*M],6,3);
PAR5_8    = LAG2 // CHAR(LAGC[1:M,4*M+1:8*M],6,3);
PAR9_12   = LAG3 // CHAR(LAGC[1:M,8*M+1:12*M],6,3);


PRINT 'AUTOREGRESSION MATRICES',
AR1_4    , AR5_8   ,  AR9_12 ;


PRINT 'PARTIAL LAG CORRELATION MATRICES',
PAR1_4   , PAR5_8   , PAR9_12 /;


FREE AUTO LAGC AR1_4  AR5_8  AR9_12 PAR1_4 PAR5_8 PAR9_12;


/*schematic representation of the partial lag correlation
  matrices*/

GRAPH = J(15*M,M,'   .');

DO I = 1 TO 15*M;

   DO J = 1 TO M;

        IF LAGCORR[I,J] <  -(2/SQRT(TT)) THEN GRAPH[I,J] = '  -';
   ELSE IF LAGCORR[I,J] >  ( 2/SQRT(TT)) THEN GRAPH[I,J] = '  +';
   END;
END;

CORR=J(M,15*M,'   .');
DO H = 1 TO 15;
  CORR[,M*(H-1)+1:M*(H-1)+M] = GRAPH[M*(H-1)+1:M*(H-1)+M,];
END;
```

```
CORR1_4   = LAG1 // CORR[1:M,1:4*M];
CORR5_8   = LAG2 // CORR[1:M,4*M+1:8*M];
COR9_12   = LAG3 // CORR[1:M,8*M+1:12*M];

PRINT 'SCHEMATIC REPRESENTATION OF THE PARTIAL LAG CORRELATIONS',
      CORR1_4  , CORR5_8   , COR9_12 /;

FREE GRAPH CORR CORR1_4 CORR5_8 COR9_12 LAG LAG1 LAG2 LAG3 LAG3
LAG4 LAG5;

/* computes the statistic sum(i and j)_lagcorr(h)^2
   to help identifying the order of an VAR,
   see Wei, 1993, page 362 */

X_H   = J(15,1,0);
LAG   = J(15,1,0);
P_VALUE = J(15,1,0);

DO H = 1 TO 15;

X_H[H,] = ROUND(TT * SSQ(LAGCORR[M*(H-1)+1:M*(H-1)+M,]),0.01);
LAG[H,] = H;
P_VALUE[H,] = ROUND((1 - PROBCHI(X_H[H,],M**2)),.0001);

END;

PRINT 'X(H) TO BE COMPARED WITH A CHI-SQUARED WITH M^2 DEGREES
OF FREEDOM','(M IS THE DIMENSION OF THE SERIES)',
LAG  X_H  P_VALUE ;

FINISH IDENTMUL;

RUN IDENTMUL;

%MEND IDENTMUL;
```

```
/* macro estimvar.mac ------------------------------------------

  computes the Yule-Walker estimates for Vector autoregressive
  model. It uses the covariances computed with the subroutines
  autocov and meancov.

   ==> REFERENCE ===>
        WEI(1993), pages 359-361

   ==> TO EXECUTE: %INCLUDE ESTIMVAR;
                    %ESTIMVAR(PROGRAM PARAMETERS DESCRIBED BELOW)

   ==> PARAMETERS NEEDED TO CALL THE ESTIMVAR MACRO

    &ORDER    ---> ORDER OF THE AUTOREGRESSIVE MODEL
 --------------------------------------------------------- */

%MACRO ESTIMVAR(ORDER);

START ESTIMVAR;

/********************************************
 Calculates the Yule-Walker estimates using
 the algorithm in Wei(1993, p.359-361).
*******************************************/

DO H=1 TO &order;

 IF H=1 THEN DO;

    VU        = GAMMA0;
    VV        = GAMMA0;
    VVU       = GAMMA[1:M,];

    PHI    = GAMMAT[1:M,] * IGAMMA0;
    BETA   = GAMMA[1:M,] * IGAMMA0;
    PHIT   = PHI;

    %IF &order = 1 %THEN %DO;
        PHITT = PHI`;
```

```
      SIGMA = GAMMA0 - (GAMMAT[1:M*&order,])`*PHITT;
      DO I=1 TO M;
         DO J=1 TO M;
            SIGMA[I,J] = SIGMA[J,I];
         END;
      END;

      PRINT 'PARAMETER MATRICES OF A VAR(' "&order" ') MODEL',
                                                          PHI;
      PRINT 'NOISE COVARIANCE MATRIX', SIGMA;
   %END;

END;
ELSE DO;

   VU = GAMMA0 - PHI * GAMMA[1:M*(H-1),];
   VV = GAMMA0 - BETA  * GAMMAT[1:M*(H-1),];
   VVU = GAMMA[M*(H-1)+1:M*(H-1)+M,] -
                     (PHIT * GAMMAT[1:M*(H-1),])` ;

   PHIHH  = VVU` * INV(VV);
   BETAHH = VVU * INV(VU);

   IF H=2 THEN DO;
    PHIHK  = PHI - PHIHH * BETA;
    PHIN   = PHIHK  || PHIHH;
    PHINT  = PHIHH  || PHIHK;
    PHINTT = PHIHK` // PHIHH`;
    BETAHK = BETA - BETAHH * PHI;
    BETAN = BETAHK || BETAHH;
   END;
   ELSE DO;

    PHIN   = PHIHH;
    PHINT  = PHIHH;
    PHINTT = PHIHH`;
    BETAN = BETAHH;

    DO K = H-1 TO 1 BY -1;
     PHIHK = PHI[,M*K-1:M*K] - PHIHH * BETA[,M*(H-K)-1:M*(H-K)];
     PHIN  = PHIHK || PHIN;
     PHINT = PHINT || PHIHK;
     PHINTT = PHIHK` // PHINTT;
```

```
      BETAHK  = BETA[,M*K-1:M*K] -
                     BETAHH * PHI[,M*(H-K)-1:M*(H-K)];
      BETAN   = BETAHK ¦¦ BETAN;
     END;

    END;

    PHI   = PHIN;
    PHIT  = PHINT;
    PHITT = PHINTT;

    BETA  = BETAN;

    SIGMA = GAMMA0 - (GAMMAT[1:M*&order,])`*PHITT;
    DO I=1 TO M;
       DO J=1 TO M;
          SIGMA[I,J] = SIGMA[J,I];
       END;
    END;

    PRINT ’PARAMETER MATRICES OF A VAR(’ "&order" ’) MODEL’ ,
                                                          PHI;
    PRINT ’NOISE COVARIANCE MATRIX’, SIGMA;
 END;
END;

FREE VU VV VVU GAMMA0 IGAMMA0 GAMMA GAMMAT BETA BETAHK BETAHH
BETANH PHIT PHIN PHINT PHINTT PHITT PHIHH PHIHK DU DV;

FINISH ESTIMVAR;
RUN ESTIMVAR;

%MEND ESTIMVAR;
```

Appendix E2 - SAS/IML programs to fit State-Space Models

```
/** BLFS.SAS ***********************************************
** THIS PROGRAM FITS A STATE SPACE MODEL TO THE BRAZILIAN
** LABOUR FORCE SURVEY DATA. THE MODEL FOR THE SIGNAL IS
** IS A COMMOM COMPONENT BASIC STRUCTURAL MODEL. THE MODEL
** FOR THE SAMPLING ERRORS IS A VAR(1) MODEL.
** THE INPUT DATA ARE THE AVERAGE OF THE TRANSFORMED PANEL
   ESTIMATES
************************************************************/

options linesize=72 ps=58 nodate;
options mprint symbolgen;


/* files with the transformed sample estimates    v_t^(k)

  per rotation group */

filename in1      'c:/denise/data/estim1.dat';
filename in2      'c:/denise/data/estim2.dat';
filename in3      'c:/denise/data/estim3.dat';
filename in4      'c:/denise/data/estim4.dat';

/* routines for computing smoothed estimates and
   for estimating the signal of the original compositions */

filename kalsmt   'c:/denise/sas/kalsmt.mac';
filename estim    'c:/denise/sas/estim.mac';

/* Reads the transformed panel estimates from Brazilian Labour
   Force for a given rotation group */

/* file with the estimates in the original scale */

filename out 'c:/denise/data/results.dat';
```

```
%MACRO READ;

 %do rg=1 %to 4;
    data estim&rg;
       infile in&rg;
         input year month panel $ pemp pune pnilf v1 v2;
   run;
 %end;
%MEND;

%READ;

TITLE1 'BRAZILIAN LABOUR FORCE SURVEY';

TITLE2
'(STOCHASTIC TREND + STOCHASTIC SLOPE + DUMMY SEASONAL)+ VAR(1)';

data estim;
     set estim1 estim2 estim3 estim4;

proc sort data=estim;
     by year month;
```

/* the input data is $\frac{1}{4}\sum_{K=1}^{4} \mathbf{v}_t^{(k)}$ */

```
proc means data = estim mean noprint;
     by year month;
     var v1 v2;
     output out= estim   mean=v1 v2;

run;

PROC IML;

/* Inputs the bivariate series into IML */

START READDATA;
  USE ESTIM;

   READ ALL INTO V VAR {v1 v2};
  CLOSE ESTIM;
FINISH READDATA;
RUN READDATA;

PRINT'ESTIMATION OF THE UNKNOWN PARAMETERS';
PRINT'USING QUASI-NEWTON METHOD FROM SAS_IML';

/****  state-space model defined as **************************

   v_t = Z x alpha_t + epsilon_t          V(epislon) = U

   alpha_t = T x alpha_t-1 + G x eta_t    V(eta ) = Q
**************************************************************/

COL1={'V1' 'V2'};
PRINT / 'INPUT DATA', V (¦COLNAME=COL1¦);

VPRIME = V`; /* it will be used as input to LIKL   */

/* observation matrix */
Z = I(2)¦¦J(2,2,0)¦¦I(2)¦¦J(2,20,0)¦¦I(2);

ZPRIME = Z`;
PRINT 'TRANSPOSE OF THE OBSERVATION MATRIX', ZPRIME;
FREE ZPRIME;

/* initial state vector alpha0 and variance P0 */
ALPHA0 = J(28,1,0);

K=100000; /* kI = P0  the difuse prior */

SIGNALP0 = I(26) * K;

/* stationary components of
   the state vector errorP0 = Gamma0 */

ERRORP0 = { 0.000215  0.00032 ,
            0.00032  0.0052142 };

P0 = BLOCK(SIGNALP0,ERRORP0);

PRINT 'INITIAL STATE-VECTOR AND INITIAL STATE COVARIANCE MATRIX',
    ' ALPHA1¦0  ' ALPHA0 , 'P1¦0   '   P0;
```

```
 /* transition matrix */

  T1  = { 1 1  0  0  0  0  0  0  0  0  0  0  0,
          0 1  0  0  0  0  0  0  0  0  0  0  0,
          0 0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1,
          0 0  1  0  0  0  0  0  0  0  0  0  0,
          0 0  0  1  0  0  0  0  0  0  0  0  0,
          0 0  0  0  1  0  0  0  0  0  0  0  0,
          0 0  0  0  0  1  0  0  0  0  0  0  0,
          0 0  0  0  0  0  1  0  0  0  0  0  0,
          0 0  0  0  0  0  0  1  0  0  0  0  0,
          0 0  0  0  0  0  0  0  1  0  0  0  0,
          0 0  0  0  0  0  0  0  0  1  0  0  0,
          0 0  0  0  0  0  0  0  0  0  1  0  0,
          0 0  0  0  0  0  0  0  0  0  0  1  0 };

T11 = T1 @ I(2);

/* parameter matrix of the VAR(1) for the
   transformed sampling errors as in Chapter 8 */

T22 = {  .4496519  -.018747 ,
        -.286699    .0772841 };

TRANS = (T11 || J(26,2,0)) // (J(2,26,0) || T22) ;

FREE  T1 T11 T22;

 /* G matrix from alpha_t = T x alpha_t-1 + G x eta_t */
  G1  = { 1 0 0 0,
          0 1 0 0,
          0 0 1 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 0,
          0 0 0 1};
```

```
G = G1 @ I(2);
FREE G1;

/* number of observations */
  TT=NCOL(VPRIME);
/* dimension of the state vector */
DIM1=NROW(ALPHA0) ;
/* dimension of the observation vector */
  DIM2=NROW(VPRIME);
/* first observation for computation of the likelihood */
/* I am using difuse prior so the estimates for the
   non-stationary components of the state-vector will be computed
   automatically based on the d=13 first observations */
ZZ=14;
PRINT 'FIRST OBSERVATION FOR COMPUTATION OF THE LIKELIHOOD', ZZ;

***************************************************************
* SUBROUTINE TO DO THE KALMAN FILTER AND CALCULATE THE        *
* INNOVATIONS, THEIR VARIANCES(F) AND THE KALMAN GAIN(K).     *
* THE STATE VECTOR ALPHA AND ITS COVARIANCE MATRIX P ARE      *
* NOT PRODUCED AS PART OF THE OUTPUT.                         *
*                                                             *
* INPUT: OBS:     TRANSPOSED MATRIX OF THE OBSERVATIONS       *
*                 DIMENSION OF THE MULTIPLE TIME SERIES x TT  *
*        ALPHA0:  INITIAL STATE VECTOR (COLUMN VECTOR)        *
*        P0:      COVARIANCE MATRIS OF ALPHA0                 *
*        Z:       MEASUREMENT EQUATION MATRIX (OR VECTOR)     *
*        TRANS:   TRANSITION MATRIX                           *
*        G:       SYSTEM NOISE MATRIX                         *
*        Q:       COVARIANCE MATRIX OF THE SYSTEM ERRORS      *
*                                                             *
* OUTPUT: INNOV:  INNOVATION  MATRIX                          *
*         F:      VARIANCE OF THE INNOVATION                  *
*         K:      KALMAN GAIN MATRIX                          *
* NOTE:  EACH COLUMN OF OBS,  INNOV, F AND K                  *
*        REPRESENTS A TIME POINT.                             *
*        THE  SUB-MATRIX OF K LOCATED IN                      *
*        COLUMN DIM2*(T-1)+1 TO DIM2*T (WHERE DIM2 IS THE     *
*        NUMBER OF ROW OF OBS) IS THE KALMAN GAIN MATRIX      *
*        OF THE VECTOR ALPHA IN THE T-TH COLUMN OF ALPHA.     *
*                                                             *
* This routine was kindly provided by Prof. D.Pfeffermann     *
***************************************************************;
```

```
START KALFIL(INNOV,F,ALPHA,P,APRE,PPRE,DIM1,DIM2,TT,
             ZZ,VPRIME,ALPHA0,P0, Z,TRANS,G,Q,U);

/* matrices initialisation and definitions */
/* APRE:   FORECAST OF ALPHA_t¦t-1
   PPRE:   COVARIANCE MATRIX OF APRE
   ALPHA:  ALPHA_t¦t
   P:      COVARIANCE MATRIX OF ALPHA */

INNOV = J(DIM2,TT,0);
K = J(DIM1,DIM2*TT,0);
F = J(DIM2,DIM2*TT,0);

DO T=1 TO TT BY 1;
    IF T=1 THEN DO;
       APRE = ALPHA0;
       PPRE = P0;
   END;
   ELSE DO;
      APRE = TRANS*ALPHA;
      PPRE = TRANS*P*TRANS` + G*Q*G`;
   END;

   F[,DIM2*(T-1)+1:DIM2*T] = Z*PPRE*Z` + U;
   K[,DIM2*(T-1)+1:DIM2*T] = PPRE*Z`*
                                INV(F[,DIM2*(T-1)+1:DIM2*T]);
   INNOV[,T] = VPRIME[,T] - Z*APRE;
   ALPHA = APRE + K[,DIM2*(T-1)+1:DIM2*T]*INNOV[,T];
   P=(I(DIM1)-K[,DIM2*(T-1)+1:DIM2*T]*Z)*PPRE;
END;

FREE K;

FINISH KALFIL;
```

```
*************************************************************
*                                                           *
* SUBROUTINE TO EVALUATE THE LIKELIHOOD                     *
* FUNCTION AT DIFFERENT ESTIMATED VALUES OF THE COVARIANCES *
* MATRICES OF THE MODEL.                                    *
*                                                           *
*INPUT: ZZ:  FIRST OBSERVATION FOR COMPUTATION OF LIKELIHOOD*
*       OBS: OBSERVATIONS                                   *
*                                                           *
* R: PARAMETER ESTIMATES                                    *
*                                                           *
* OUTPUT:   F:VALUES OF THE LOG-LIK. EVALUATED AT R         *
*           FINALR : FINAL PARAMETER ESTIMATES              *
*************************************************************;

START LIKL(R) GLOBAL(INNOV,F,ALPHA,P,APRE,PPRE,DIM1,DIM2,TT,
                     ZZ,VPRIME,ALPHA0,P0,Z,TRANS,G,Q,U);

/* from Fernandez & Harvey, JBES,1990, vol 8, pp.71-81
   The unknown parameters in the BSM are the elements
   of Sigma_L , Sigma_R and Sigma_S. The constraint that
   these covariances matrices are positive semi-definite is

   implemented by defining lower triangular matrices  $\Sigma^{\frac{1}{2}}$

   such that  $\Sigma^{\frac{1}{2}}\Sigma^{\frac{1'}{2}}=\Sigma$  and by  maximizing the likelihood with

   respect to the elements  $\Sigma^{\frac{1}{2}}$   */


/* system noise covariance matrix */
SQRT_QL = (R[,1] || {0}) // (R[,2] || R[,3]);
QL      = SQRT_QL * SQRT_QL`;

SQRT_QR = (R[,4] || {0}) // (R[,5] || R[,6]);
QR      = SQRT_QR * SQRT_QR`;

SQRT_QS = (R[,7] || {0}) // (R[,8] || R[,9]);
QS      = SQRT_QS * SQRT_QS`;
```

```
/* covariance matrix of the white noise disturbances
   of the VAR(1) process for the transformed sampling
   errors as in Chapter 8 */

SIGMAE = { .0001736  .0003051 ,
           .0003051  .0052033 };

Q=BLOCK(QL,QR,QS,SIGMAE);

FREE SQRT_QL SQRT_QR SQRT_QS QL QR QS;

/* observation noise covariance matrix */
U=J(2,2,0);

/*obtains the innovations and their variances
  with R as parameter */

RUN KALFIL(INNOV,F,ALPHA,P,APRE,PPRE,DIM1,DIM2,TT,
           ZZ,VPRIME,ALPHA0,P0,Z,TRANS,G,Q,U);

/* calculates the log-likelihood function */
LIKT = J(1,TT,0);

 DO T=1 TO TT BY 1;
     LIKT[,T]=LOG(DET(F[,DIM2*(T-1)+1:DIM2*T]))+
              INNOV[,T]`*INV(F[,DIM2*(T-1)+1:DIM2*T])*INNOV[,T];
 END;

NIKT=LIKT[,ZZ:NCOL(LIKT)];
F=-.5#NIKT[,+];

FREE LIKT NIKT;
RETURN(F);
FINISH LIKL;

/* use the quasi-Newton optimization routine to maximize the
   likelihood and get the parameters estimates */

/* initial values of the unknown parameter vector */

R = {0.367880  3.6789E-8 0.367880  0.223130 2.2323E-8 0.223130
     0.1353353 1.3534E-8 0.1353353 };
```

```
/* initial system noise covariance matrix */
SQRT_QL0 = (R[,1] || {0}) // (R[,2] || R[,3]);

QL0      = SQRT_QL0 * SQRT_QL0`;

SQRT_QR0 = (R[,4] || {0}) // (R[,5] || R[,6]);
QR0      = SQRT_QR0 * SQRT_QR0`;

SQRT_QS0= (R[,7] || {0}) // (R[,8] || R[,9]);
QS0      = SQRT_QS0 * SQRT_QS0`;

SIGMAE = { .0001736  .0003051 ,
           .0003051  .0052033 };

PRINT 'INITIAL SYSTEM NOISE COVARIANCE MATRIX' , QL0,
        QR0,QS0,SIGMAE;

FREE SQRT_QL0 SQRT_QR0 SQRT_QS0 QL0 QR0 QS0 ;

/* sas-iml routine to maximize the log-likelihood */
OPTN = { 1 3 . 3 };
TC = { . . . 0 0 0 1e-9. . . 1E-9. . };

CALL NLPQN(RC,FINALR,"LIKL",R,OPTN,,TC);

/* define a new Q matrix */
/* system noise covariance matrix */

SQRT_QL = (FINALR[,1] || {0}) // (FINALR[,2] || FINALR[,3]);
QL      = SQRT_QL * SQRT_QL`;

SQRT_QR=(FINALR[,4] || {0}) // (FINALR[,5] || FINALR[,6]);
QR      = SQRT_QR * SQRT_QR`;

SQRT_QS = (FINALR[,7] || {0}) // (FINALR[,8] || FINALR[,9]);
QS      = SQRT_QS * SQRT_QS`;

SIGMAE = { .0001736  .0003051 ,
           .0003051  .0052033 };

Q=BLOCK(QL,QR,QS,SIGMAE);
```

```
PRINT 'ESTIMATE OF Q VIA QUASI-NEWTON OPTIMIZATION',
       QL [FORMAT = 19.12],  QR [FORMAT = 19.12] ,
       QS [FORMAT = 19.12];

FREE SQRT_QL SQRT_QR SQRT_QS QL QR QS;

/* sas-iml routine to evaluate the second derivatives (hessian
   matrix) of the log-likelihood function */

CALL NLPFDD(MAXLIKL,GRAD,HESS,"LIKL",FINALR);

PRINT 'maximum value of the log-likelihood' MAXLIKL;

/* estimates of the covariance matrix of the hyperparameters*/
/* V = J x (abs(hess)) x  J`
where J=Jacobian of the Cholesky transformation. Note that
```

$$\begin{bmatrix} a & 0 \\ b & c \end{bmatrix} * \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} = \begin{bmatrix} a^2 & ab \\ ab & b^2+c^2 \end{bmatrix}$$

```
with, for example, FINALR[,1]=a, FINALR[,2]=b and FINAL[,3]=c */

/* Jacobian of the transformation */

JL = ( (2*FINALR[,1]) || J(1,2,0) ) //
     ( FINALR[,2] || FINALR[,1] || 0 ) //
     (  0 || (2*FINALR[,2]) || (2*FINALR[,3]) );

JR = ( (2*FINALR[,4]) || J(1,2,0) ) //
     ( FINALR[,5] || FINALR[,4] || 0 ) //
     (  0 || (2*FINALR[,5]) || (2*FINALR[,6]) );


JS = ( (2*FINALR[,7]) || J(1,2,0) ) //
     ( FINALR[,8] || FINALR[,7] || 0 ) //
     (  0 || (2*FINALR[,8]) || (2*FINALR[,9]) );

COVLT =-(INV(HESS[1:3,1:3]));
COVL = JL * COVLT * JL`;

COVRT = -(INV(HESS[4:6,4:6]));
COVR = JR * COVRT * JR`;
```

```
COVST = -(INV(HESS[7:9,7:9]));
COVS = JS * COVST * JS`;

PRINT 'ESTIMATED VARIANCE OF THE PARAMETER COMPONENTS',
       COVL, COVR, COVS;

/* run the filter with the final parameter estimates
   to get estimates of the smoothed values */

%INCLUDE ESTIM; /* macro with fixed-lag smoothing algorithm
                   to compute estimates of the state in the
                    original scale. */


/* creates data sets with the final estimates
   in the original scale*/
/* matrix with: filtered, smoothed and seasonally adjusted signal
   estimates, plus the one-step ahead forecasts for the observed
   series, the innovations, estimates for the undemployment rate,
   and the seasonally adjusted unemployment rate with respective
   standard errors */

RESULTS = OALPHA`|| OASMT`  || SAOASMT` || OSTEP`|| INNOV` ||
          UNEMP` || STDUNE` || SAUNE`   || STDSAUNE` ;

COL={'EMPFIL'   'UNEFIL'   'NILFFIL'
     'EMPSMT'   'UNESMT'   'NILFSMT'
     'EMPSA'    'UNESA'    'NILFSA'
     'EMPSTEP'  'UNESTEP'  'NILFSTEP'
     'VEMPINOV' 'VUNEINOV'
     'RATE'     'STDRATE'  'SARATE'   'STDSARAT'};


CREATE SAIDA FROM RESULTS [COLNAME=COL] ;
APPEND FROM RESULTS;
CLOSE SAIDA;

QUIT;

PROC PRINT DATA=SAIDA;
```

```
DATA _NULL_;
    SET SAIDA;
    FILE OUT;
    PUT EMPFIL   UNEFIL   NILFFIL  EMPSMT  UNESMT  NILFSMT
        EMPSA    UNESA    NILFSA   EMPSTEP UNESTEP NILFSTEP
        VEMPINOV VUNEINOV
        RATE     STDRATE  SARATE   STDSARAT;

RUN;

***************************************************************
* ESTIM.MAC                                                   *
* SUBROUTINE TO EVALUATE MODEL BASED ESTIMATES (PREDICTED,    *
* FILTERED, SMOOTHED AND SEASONALLY ADJUSTED) WITH THEIR      *
* RESPECTIVE CONFIDENCE INTERVALS AND/OR VARIANCES            *
*                                                             *
* INPUT: OBS:      MATRIX OF THE OBSERVATIONS                 *
*        ALPHA:    ESTIMATE OF THE CURRENT STATE              *
*        P:        COVARIANCE MATRIX OF ALPHA                 *
*        APRE:     PREDICTED STATE VECTOR                     *
*        PPRE:     COVARIANCE MATRIX OF APRE                  *
*        ONESTEP:  ONE STEP AHED FORECAST OF TRANSFORMED      *
*                  INPUT SERIES (V)                           *
*        F:        COVARIANCE MATRIX OF ONESTEP               *
*                                                             *
*        ASMT:     SMOOTHED STATE VECTOR                      *
*        PSMT:     COVARIANCE MATRIX OF ASMT                  *
*                                                             *
*                                                             *
*      VPRIME:     TRANSFORMED INPUT SERIES                   *
*                  TRANSPOSED MATRIX OF THE OBSERVATIONS      *
*                  DIMENSION OF THE MULTIPLE TIME SERIES x TT *
*        Z:        MESUREMENT EQUATION MATRIX (OR VECTOR)     *
*        TRANS:    TRANSITION MATRIX                          *
*        U:        COVARIANCE MATRIX OF THE OBSERVATION ERROR *
*        Q:        COVARIANCE MATRIX OF THE SYSTEM ERRORS     *
*        G:        SYSTEM NOISE MATRIX                        *
*        DIM1:     DIMENSION OF THE STATE-VECTOR              *
*        DIM2:     DIMENSION OF THE OBSERVATION VECTOR        *
*        TT:       LENGTH OF THE SERIES                       *
*                                                             *
*                                                             *
*                                                             *
* OUTPUT: OALPHA: ESTIMATE OF CURRENT STATE IN THE ORIGINAL   *
*                 SCALE                                       *
*         OSTEP:  ONE STEP AHEAD FORECAST OF THE              *
*                 ORIGINAL SERIES                             *
*         OASMT:  SMOOOTHED ESTIMATE OF THE STATE             *
*                 IN THE ORIGINAL SCALE                       *
*         SAOASMT:SEASONALLY ADJUSTED SMOOOTHED ESTIMATES     *
*                 IN THE ORIGINAL SCALE                       *
*        UNEMP:   SMOOTHED ESTIMATES OF THE UNEMPLOYMENT RATE *
*       STDUNE:   ESTIMATED STANDARD ERROR OF THE ESTIMATES   *
*       SAUNE :   SEASONALLY ADJUSTED SMOOTHED ESTIMATES OF   *
*                 THE UNEMPLOYMENT RATE                       *
*     STDSAUNE:   ESTIMATE STANDARD ERROR OF SAUNE            *
*                                                             *
* NOTE:  EACH COLUMN OF OBS, ALPHA, APRE , OALPHA, OSTEP, ETC.*
*        REPRESENTS A TIME POINT.                             *
*        THE  SUB-MATRIX OF P LOCATED IN                      *
*        COLUMN DIM1*(T-1)+1 TO DIM1*T (WHERE DIM1 IS THE     *
*        NUMBER OF ROWS OF ALPHA) IS THE COVATIANCE MATRIX    *
*        OF THE VECTOR ALPHA IN THE T-TH COLUMN OF ALPHA.     *
*        THE  SUB-MATRIX OF F LOCATED IN                      *
*        COLUMN DIM2*(T-1)+1 TO DIM2*T (WHERE DIM2 IS THE     *
*        NUMBER OF COLS OF OBS) IS THE COVARIANCE  MATRIX     *
*        OF THE VECTOR OBS IN THE T-TH COL OF OBS (M * TT)    *
*                                                             *
*                                                             *
**************************************************************;

%MACRO ESTIM;

START ESTIM(ALPHA,P,APRE,PPRE,ONESTEP,F,ASMT,PSMT,VPRIME,Z,ZZ,
            TRANS,G,U,Q,DIM1,DIM2,TT,ALPHA0,P0,INNOV,K,SMOOTH,
            OALPHA,OSTEP,OASMT,SAOASMT,UNEMP,STDUNE,
            SAUNE, STDSAUNE);

/* run the filter with the final parameter estimates
   to get estimates of the smoothed values */

SMOOTH=1;
/*macro with fixed-interval smoothing algorithm*/
%INCLUDE KALSMT;
%KALSMT;
FREE PRINTPRE PSTAR K ;
```

```
/* filtered estimates of the signal theta_t = L_t + S_t  */
/* filtered estimates of theta in the original scale  */
/* t = 14 depends on  firstobs(zz) */

W = { 1 0 1 0 0 0 0 0 0 0 0 0 0 0} @ I(2) ;
VARTHE  = J(2,DIM2*NCOL(ALPHA),0);
THETA   = J(2,NCOL(ALPHA),0);
OALPHA  = J(3,NCOL(ALPHA),0);

DO T=14 TO NCOL(ALPHA) BY 1;

  THETA[,T] = W * ALPHA[,T];

  VARTHE[,DIM2#(T-1)+1:DIM2#T] =
           W * P[,DIM1#(T-1)+1:DIM1#T] * W`;

/* filtered estimates for the signal in the original scale */

 OALPHA[3,T] = 1 / ( 1 + EXP(THETA[1,T]) + EXP(THETA[2,T]));

  DO K=1 TO 2 BY 1;

   OALPHA[K,T] = EXP(THETA[K,T]) / ( 1 + EXP(THETA[1,T]) +
                                        EXP(THETA[2,T]));
  END;
END;

/* smoothed estimates of the signal theta_t = L_t + S_t */
/* smoothed estimates of theta in the original scale  */
W = { 1 0 1 0 0 0 0 0 0 0 0 0 0 0} @ I(2) ;

THETA    = J(2,NCOL(ALPHA),0);
OASMT    = J(3,NCOL(ALPHA),0);
THETA1_2 = J(1,NCOL(ALPHA),0);
UNEMP     = J(1,NCOL(ALPHA),0);
VTHETA1_2= J(1,NCOL(ALPHA),0);
STDUNE    = J(1,NCOL(ALPHA),0);
EXP1_2    = J(1,NCOL(ALPHA),0);
VEXP1_2   = J(1,NCOL(ALPHA),0);

DO T=14 TO NCOL(ASMT) BY 1;

  THETA[,T] = W * ASMT[,T];
  VARTHE[,DIM2#(T-1)+1:DIM2#T] =
                W * PSMT[,DIM1#(T-1)+1:DIM1#T] * W`;

/* mean and variance of THETA_STAR_1 - THETA_STAR_1  */
/* where theta_star_1 = log (theta_emp / theta_nilf) */
/* and theta_star_2   = log (theta_une / theta_nilf) */

  THETA1_2[,T]  = THETA[1,T] - THETA[2,T];

  VTHETA1_2[,T] = VARTHE[1,DIM2#(T-1)+1] + VARTHE[2,DIM2#T]
                  - 2 * VARTHE[1,DIM2#T];

 /* mean and variance of exp(theta_star_1 - theta_star_2) */

  EXP1_2[,T]  =  EXP(THETA1_2[,T] + 0.5 * VTHETA1_2[,T]);

  VEXP1_2[,T] =  (EXP(VTHETA1_2[,T]) - 1) *
                        (EXP(2*THETA1_2[,T] + VTHETA1_2[,T]));

 /* approximate MMSE and standard error of unemployment rate */

  UNEMP[,T]  =  INV(EXP1_2[,T] + 1) +
                    VEXP1_2[,T] / (EXP1_2[,T] + 1)**3;

  STDUNE[,T] =  SQRT((INV(EXP1_2[,T] + 1))**4 *  VEXP1_2[,T]) ;

/* Smoothed estimates for the signal in the original scale */

 OASMT[3,T] = 1 / ( 1 + EXP(THETA[1,T]) + EXP(THETA[2,T]));

  DO K=1 TO 2 BY 1;

   OASMT[K,T] = EXP(THETA[K,T]) / ( 1 + EXP(THETA[1,T]) +
                                             EXP(THETA[2,T]));

  END;
END;
```

```
/* seasonally adjusted smoothed estimates of the signal
   theta_t = L_t , note that there is no irregular term*/
/* seasonally adjusted smoothed estimates of theta in the
   original scale*/

W = { 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 } @ I(2) ;

VARTHE     = J(2,DIM2*NCOL(ALPHA),0);
ADJTHETA   = J(2,NCOL(ALPHA),0);
SAOASMT    = J(3,NCOL(ALPHA),0);

THETA1_2 = J(1,NCOL(ALPHA),0);
SAUNE    = J(1,NCOL(ALPHA),0);
VTHETA1_2= J(1,NCOL(ALPHA),0);
STDSAUNE = J(1,NCOL(ALPHA),0);
EXP1_2   = J(1,NCOL(ALPHA),0);
VEXP1_2  = J(1,NCOL(ALPHA),0);

DO T=14 TO NCOL(ALPHA) BY 1;

  ADJTHETA[,T] = W * ASMT[,T];

  VARTHE[,DIM2#(T-1)+1:DIM2#T]=W*PSMT[,DIM1#(T-1)+1:DIM1#T]*W`;

/* mean and variance of THETA_STAR_1 - THETA_STAR_2 */

  THETA1_2[,T]  = ADJTHETA[1,T] - ADJTHETA[2,T];
  VTHETA1_2[,T] = VARTHE[1,DIM2#(T-1)+1] + VARTHE[2,DIM2#T]
                                        - 2 * VARTHE[1,DIM2#T];

 /* mean and variance of exp(theta_star_2 - theta_star_1) */

  EXP1_2[,T]   =  EXP(THETA1_2[,T] + 0.5 * VTHETA1_2[,T]);

  VEXP1_2[,T] =  (EXP(VTHETA1_2[,T]) - 1) *
                       (EXP(2*THETA1_2[,T] + VTHETA1_2[,T]));

 /* approximate MMSE and standard error of seasonally adjusted
          unemployment rate*/

  SAUNE[,T]     =  INV(EXP1_2[,T] + 1) +
                    VEXP1_2[,T] / (EXP1_2[,T] + 1)**3;
  STDSAUNE[,T] =  SQRT((INV(EXP1_2[,T] + 1))**4 *  VEXP1_2[,T]);

/* seasonally adjusted smoothed estimates for the signal
   in the original scale */

  SAOASMT[3,T] = 1 / ( 1 + EXP(ADJTHETA[1,T]) +
                              EXP(ADJTHETA[2,T]));

  DO K=1 TO 2 BY 1;

   SAOASMT[K,T] =
       EXP(ADJTHETA[K,T]) / ( 1 + EXP(ADJTHETA[1,T]) +
                                       EXP(ADJTHETA[2,T]));
  END;
END;

/*estimates of the one-step ahead forecast y_t¦t-1 (in the
  original scale)*/

OSTEP = J(3,NCOL(ONESTEP),0);

DO T=14 TO NCOL(ONESTEP) BY 1;

   OSTEP[3,T] = 1 / ( 1 + EXP(ONESTEP[1,T]) + EXP(ONESTEP[2,T]));

     DO K=1 TO 2 BY 1;

         OSTEP[K,T] = EXP(ONESTEP[K,T]) /
                       (1 + EXP(ONESTEP[1,T]) +
                                    EXP(ONESTEP[2,T]));
     END;
END;

FREE THETA VARTHE ADJTHETA VTHETA1_2 THETA1_2 EXP1_2 VEXP1_2;

FINISH ESTIM;

RUN ESTIM(ALPHA,P,APRE,PPRE,ONESTEP,F,ASMT,PSMT,VPRIME,Z,ZZ,
          TRANS,G,U,Q,DIM1,DIM2,TT,ALPHA0,P0,INNOV,K,SMOOTH,
          OALPHA,OSTEP,OASMT,SAOASMT,UNEMP,STDUNE,
          SAUNE, STDSAUNE);

%MEND ESTIM;
```

```
***************************************************************
* KALSMT.MAC                                                  *
* SUBROUTINE TO DO THE KALMAN FILTER WITH TIME INVARIANT      *
* MATRICES Z,TRANS, G, U AND Q.                               *
* INPUT: OBS:     MATRIX OF THE OBSERVATIONS                  *
*        ALPHA0:  INITIAL STATE VECTOR (COLUM VECTOR)         *
*        P0:      COVARIANCE MATRIS OF ALPHA0                 *
*        Z:       MESUREMENT EQUATION MATRIX (OR VECTOR)      *
*        TRANS:   TRANSITION MATRIX                           *
*        U:       COVARIANCE MATRIX OF THE OBSERVATION ERROR  *
*        Q:       COVARIANCE MATRIX OF THE SYSTEM ERRORS      *
*        G:       SYSTEM NOISE MATRIX                         *
*        SMOOTH:  1=SMOOTHING, 0=NO SMOOTHING                 *
*                                                             *
* OUTPUT: APRE:   STATE VECTOR MATRIX FROM THE PREDICTION     *
*         PPRE:   COVARIANCE MATRIX OF APRE                   *
*         ALPHA:  STATE VECTOR MATRIX                         *
*         P:      COVARIANCE MATRICES OF THE STATE VECTORS    *
*         INNOV:  INNOVATION  MATRIX                          *
*         F:      VARIANCE OF THE INNOVATION                  *
*         K:      KALMAN GAIN MATRIX                          *
*         ASMT:   STATE VECTOR MATRIX SMOOTHED                *
*         PSMT0:   COVARIANCE MATRIX OF AMST AT TIME 0        *
* NOTE:  EACH COLUMN OF OBS, ALPHA, INNOV, F AND K            *
*        REPRESENTS A TIME POINT.                             *
*        THE  SUB-MATRIX OF P LOCATED IN                      *
*        COLUMN DIM1*(T-1)+1 TO DIM1*T (WHERE DIM1 IS THE     *
*        NUMBER OF ROWS OF ALPHA) IS THE COVATIANCE MATRIX    *
*        OF THE VECTOR ALPHA IN THE T-TH COLUMN OF ALPHA.     *
*        THE  SUB-MATRIX OF K LOCATED IN                      *
*        COLUMN DIM2*(T-1)+1 TO DIM2*T (WHERE DIM2 IS THE     *
*        NUMBER OF COLS OF OBS) IS THE KALMAN GAIN MATRIX     *
*        OF THE VECTOR ALPHA IN THE T-TH COLUMN OF ALPHA.     *
*                                                             *
*The core of this routine was kindly provided by:             *
* Prof D.Pfeffermann                                          *
**************************************************************;

%MACRO KALSMT;

START KALSMT(APRE,PPRE,ALPHA,P,INNOV,ONESTEP,K,F,ASMT,PSMT,
             DIM1,DIM2,TT,VPRIME,ZZ,
             ALPHA0,P0,Z,TRANS,G,U,Q,SMOOTH);
```

```
/* dimension of the state vector */

DIM1=NROW(ALPHA0);

/* dimension of the observation vector */

DIM2=NROW(VPRIME);

/* number of observations */

TT=NCOL(VPRIME);

/* matrices initialisation and definitions */
/* APRE : forecast of alpha = alpha t¦t-1 */

APRE=J(DIM1,TT,0);
ALPHA=J(DIM1,TT,0);
ASMT=J(DIM1,TT,0);
INNOV=J(DIM2,TT,0);
ONESTEP=J(DIM2,TT,0);
K=J(DIM1,DIM2*TT,0);
F=J(DIM2,DIM2*TT,0);
P=J(DIM1,DIM1*TT,0);
PPRE=J(DIM1,DIM1*TT,0);
PSMT=J(DIM1,DIM1*TT,0);

/* standardized  and relative innovations */
INNOVRES = J(DIM2,NCOL(INNOV),0);
FSQRT    = J(DIM2,DIM2*TT,0);
RELAT    = J(DIM2,NCOL(INNOV),0);

DO T=1 TO TT BY 1;
   IF T=1 THEN DO;

      APRE[,T] =ALPHA0;
      PPRE[,DIM1*(T-1)+1:DIM1*T] = P0;

      F[,DIM2*(T-1)+1:DIM2*T] =
                   Z*PPRE[,DIM1*(T-1)+1:DIM1*T]*Z` + U;
      K[,DIM2*(T-1)+1:DIM2*T] = PPRE[,DIM1*(T-1)+1:DIM1*T]*Z`*
                                 INV(F[,DIM2*(T-1)+1:DIM2*T]);

      ONESTEP[,T]= Z*APRE[,T];
```

```
     INNOV[,T] = VPRIME[,T] - ONESTEP[,T];
     ALPHA[,T] = APRE[,T] + K[,DIM2*(T-1)+1:DIM2*T]*INNOV[,T];
     P[,DIM1*(T-1)+1:DIM1*T] =
                (I(DIM1)-K[,DIM2*(T-1)+1:DIM2*T]*Z)
                                *PPRE[,DIM1*(T-1)+1:DIM1*T];

   END;
   ELSE DO;

     APRE[,T]=TRANS*ALPHA[,T-1];
     F[,DIM2*(T-1)+1:DIM2*T] =
      Z*(TRANS*P[,DIM1*(T-2)+1:DIM1*(T-1)]*TRANS`+G*Q*G`)*Z`+U;

     PPRE[,DIM1*(T-1)+1:DIM1*T] =
       TRANS*P[,DIM1*(T-2)+1:DIM1*(T-1)]*TRANS` + G * Q * G` ;

     K[,DIM2*(T-1)+1:DIM2*T] =
      (TRANS*P[,DIM1*(T-2)+1:DIM1*(T-1)]*TRANS`+ G*Q*G`) * Z`*
                                INV(F[,DIM2*(T-1)+1:DIM2*T]);
     ONESTEP[,T]= Z*APRE[,T];
     INNOV[,T] = VPRIME[,T] - Z*APRE[,T];

     ALPHA[,T] = APRE[,T] + K[,DIM2*(T-1)+1:DIM2*T]*INNOV[,T];


     P[,DIM1*(T-1)+1:DIM1*T] =
       (I(DIM1)-K[,DIM2*(T-1)+1:DIM2*T]*Z)*
           (TRANS*P[,DIM1*(T-2)+1:DIM1*(T-1)]*TRANS` + G*Q*G`);

/* standardized innovations */

    SQRTF  = SQRT(VECDIAG(F[,DIM2*(T-1)+1:DIM2*T]));

    INNOVRES[,T] = INNOV[,T] / SQRTF;

/* RELATIVE  INNOVATIONS */

   RELAT[,T] = INNOV[,T] / VPRIME[,T];

  END;
END;

FREE K   ;
```

```
/* PREDICTION ERROR VARIANCE FOR THE T=TT-2, TT-1, TT */

PEV = F[,DIM2*(TT-2-1)+1:DIM2*TT];

/* MEASURES OF GOODNESS OF FIT BASED ON THE ONE-STEP AHEAD
   PREDICTIONS ERRORS ( THE INNOVATIONS ) */

ERRORS= J(DIM2,TT-ZZ+1,0);

ERRORS = INNOV[,ZZ:TT];
REL    = RELAT[,ZZ:TT];

MB = ERRORS[,+] / (TT-ZZ+1);

ABSERROR = ABS ( ERRORS);

MAB = ABSERROR[,+] / (TT-ZZ+1);

SQRE = REL[,##] / (TT-ZZ+1);

/* putting the standardized innov in the vector innov */

DO T = 1 TO TT BY 1;

  IF T < ZZ THEN DO;

     INNOV[,T] = 0;

  END;
  ELSE DO;

     INNOV[,T] = INNOVRES[,T];
  END;
END;

PRINT MB MAB SQRE, PEV;

FREE SQRTF INNOVRES ERRORS RELAT REL ABSERROR PEV;
```

```
/* this part does the smoothing */
/* fixed-interval smoothing - Harvey,1989,pp154-155 */

IF SMOOTH=1 THEN DO;

    /* initialisation for time t = tt */

    ASMT[,TT] = ALPHA[,TT];
    ASMTFI=ASMT[,TT]`;
   * PRINT'ESTIMATE OF THE CURRENT STATE VECTOR -ALPHA_t¦t',
                                                       ASMTFI;
    PSMT[,DIM1#(TT-1)+1:DIM1#TT] = P[,DIM1#(TT-1)+1:DIM1#TT];
    *PRINT 'V-C MATRIX OF FILTERED STATE ESTIMATORS AT TIME TT',
                                                          PSMT;
     /* smoothing for t = tt-1,...,2 */
    /* if MT is singular for some t , INV(MT) can be replaced
       by its generalised inverse as suggested by Kohn and
       Ansley(1983) - from Harvey,1989,p.154   */

    DO T= TT-1 TO 1 BY -1;

      MT = TRANS*P[,DIM1*(T-1)+1:DIM1*T]*TRANS` + G * Q * G`;
      PSTAR = P[,DIM1#(T-1)+1:DIM1#T]*TRANS`*GINV(MT);

      ASMT[,T]= ALPHA[,T] + PSTAR * (ASMT[,T+1]-TRANS*ALPHA[,T]);

      PSMT[,DIM1#(T-1)+1:DIM1#T] =
                  P[,DIM1#(T-1)+1:DIM1#T]+
                 PSTAR*(PSMT[,DIM1#(T)+1:DIM1#(T+1)]-MT)*PSTAR`;

    END;
END;

FINISH KALSMT;

RUN KALSMT(APRE,PPRE,ALPHA,P,INNOV,ONESTEP,K,F,ASMT,PSMT,
           DIM1,DIM2,TT,VPRIME,ZZ,
           ALPHA0,P0,Z,TRANS,G,U,Q,SMOOTH);

%MEND KALSMT;
```