Journal of Official Statistics, Vol. 14, No. 3, 1998, pp. 237–253

The Time Series Analysis of Compositional Data

Teresa M. Brunsdon[1] *and T.M.F. Smith*[2]

The analysis of repeated surveys can be approached using model-based inference, utilising the methods of time series analysis. On a long run of repeated surveys it should then be possible to enhance the estimation of a survey parameter. However, many repeated surveys that are suited to this approach consist of variables that are proportions, and hence are bounded between 0 and 1. Furthermore interest is often in a multinomial vector of these proportions, that are sum-constrained to 1, i.e., a composition. A solution to using time series techniques on such data is to apply an additive logistic transformation to the data and then to model the resulting series using vector ARMA models. Here the additive logistic transformation is discussed which requires that one variable be selected as a reference variable. Its application to compositional time series is developed, which includes the result that the choice of reference variable will not affect any final results in this context. The discussion also includes the production of forecasts and confidence regions for these forecasts. The method is illustrated by application to the Australian Labour Force Survey.

Key words: Repeated surveys; additive logistic transformation; VARMA; dependence; labour force survey.

1. Introduction

The theory of sample surveys has mainly been concerned with univariate problems. Arguably this matters little for randomization inference since the only random variable is the indicator representing sample selection. For model-based inference, however, the multivariate nature of survey data must be taken into account. Scott and Smith (1974) developed a model-based theory for the analysis of repeated surveys which was essentially univariate. If y_t is a survey estimate of a parameter θ_t based on survey data at time t then we can express this in signal and noise form as

$$y_t = \theta_t + e_t, \quad t = 1, 2, \ldots, T \tag{1.1}$$

If the estimator is unbiased then the estimation error, e_t, will have mean zero and its covariance structure will be determined by the sample design. In randomization inference θ_t would be treated as an unknown constant with no relationship between θ_t and past values $\theta_{t-1}, \theta_{t-2}, \ldots$. Scott and Smith argued that θ_t would frequently change stochastically over time and could be represented by a time series model. The covariance structure of θ_t

[1] Sheffield Hallam University, Division of Applied Statistics, School of Computing and Management Sciences, Sheffield S1 1WB, U.K.
[2] University of Southampton, Faculty of Mathematical Studies, Highfield, Southampton SO9 5NH, U.K.
Acknowledgment: This research was supported by grants from the Economic and Social Research Council of the U.K., GR/E/39853 and H519345002.

could be inferred from the observed covariances of y_t, and the known covariance structure of e_t. They showed that time series predictors of θ_t could be more efficient than the classical randomization estimators.

Time series analysis requires a long run of data for efficient estimation. In addition, it is much easier to employ the covariance structure of e_t in a time series framework if the error structure remains constant over time. This implies a long run of surveys with the same design and sample size. One set of surveys which met these conditions were monthly public opinion polls of voting intentions. Scott, Smith and Jones (1977) and Smith (1978), fitted time series models to key variables such as, C_t, the proportion who would vote Conservative, L_t, the proportion who would vote Labour, and, $C_t - L_t$, the Conservative lead over Labour, which could be negative. The results demonstrated some of the potential gains of time series methods, but they also raised several additional problems. First, the proportions were bounded between 0 and 1 and yet the models fitted were not so constrained. Second, the true variable of interest was the complete vector of voting intentions, a multinomial vector, not the single variables, and the Labour and Conservative votes would be negatively correlated. The solution to these problems became clear when Aitchison (1982) read a paper on the statistical analysis of compositional data to the Royal Statistical Society. The multinomial vectors formed compositions and so the problem was that of the time series analysis of compositional data. Many series published by official statistical agencies satisfy the conditions of a compositional time series. In this article we analyze labour market data from the Australian Labour Force Survey. The multinomial vector of interest is the employment status of individuals categorized as employed (E_t), unemployed (U_t), and not in the labour force (N_t). Wallis (1987) analyzed a univariate time series of unemployment rates. Following the methods in Brunsdon (1987) we show how the analysis can be extended to the multivariate composition of all labour market states.

2. Compositional Data

Consider a multinomial response, $\underline{r}^T = (r_1, r_2, \ldots, r_{d+1})$, $\sum_{i=1}^{d+1} r_i = n$, which represents a d dimensional random variable. Let $x_i = r_i/n$, $\underline{x}^T = (x_1, \ldots, x_d)$, then $\underline{x}$ is a composition which lies in the simplex $S^d = \left\{\underline{x} : 0 < x_i < 1,\ i = 1, \ldots, d;\ \sum_{i=1}^{d} x_i < 1\right\}$. The value $x_D = 1 - \sum_{i=1}^{d} x_i$, where $D = d + 1$, is called the fill-up value, or FUV, and is determined by the d values $x_1, \ldots, x_d$. The problems of modelling and analyzing compositional data are discussed thoroughly in the monograph by Aitchison (1986). He demonstrates the difficulties of applying standard methods to the composition, $\underline{x}$, due to the constraints of the boundary of the simplex. Multivariate analyses based on null concepts such as independence are particularly difficult to handle. Aitchison's solution, which like all good ideas seems obvious when you hear it, is to map $\underline{x}$ from the simplex S^d onto $\mathbb{R}^d$ and then to examine the statistical properties within $\mathbb{R}^d$. He considers several transformations the most important of which is the additive-logistic or $a_d(x_i)$ transformation defined by:

$$y_i = a_d(x_i) = \log\left(\frac{x_i}{x_D}\right), \qquad (i = 1, \ldots, d) \tag{2.1}$$

where

$$x_D = 1 - \sum_{i=1}^{d} x_i$$

with inverse

$$\begin{aligned} x_i = a_d^{-1}(y_i) &= \frac{e^{y_i}}{1 + \sum_{j=1}^{d} e^{y_j}} \quad (i = 1, \ldots, d) \\ &= \frac{1}{1 + \sum_{j=1}^{d} e^{y_j}} \quad (i = D) \end{aligned} \tag{2.2}$$

where x_D is the FUV. Let $\underline{x}^f$ denote the $D \times 1$ vector, consisting of $\underline{x}$ augmented by x_D, so that

$$\left\{ \underline{x}^f : 0 < x_i^f < 1 (i = 1, \ldots, D); \quad \sum_{i=1}^{D} x_i^f = 1 \right\}$$

represents an alternative definition of a composition.

One problem is that if the x_i's are permuted a different FUV is obtained and so a different version of a_d. In other words we may select any element of $\underline{x}^f$, x_k say, to be the reference variable and obtain:

$$y_i^{(k)} = a_d^{(k)}(x_i) = \log\left(\frac{x_i}{x_k}\right) \quad (i = 1, \ldots, D; i \neq k)$$

with inverse

$$\begin{aligned} a_d^{(k)-1}\left(y_i^{(k)}\right) &= \frac{e^{y_i^{(k)}}}{1 + \sum_{\substack{j=1 \\ j \neq k}} e^{y_j^{(k)}}} \quad (i = 1, \ldots, D; i \neq k) \\ &= \frac{1}{1 + \sum_{\substack{j=1 \\ j \neq k}} e^{y_i^{(k)}}} \quad (i = k) \end{aligned}$$

In using this transformation we must therefore establish whether subsequent analysis is invariant to the choice of reference variable. It is useful to note that $\underline{y}(k) = \underline{Z}(k)\underline{y}^D$, where $\underline{Z}(k) = \left\{z_{ij}(k)\right\}$
and

$$\begin{aligned} z_{ij}^{(k)} &= 1, \qquad (i = j \neq k; i, j = 1, \ldots, d) \\ &= -1, \qquad (j = k; i = 1, \ldots, d) \\ &= 0, \qquad \text{elsewhere} \end{aligned} \tag{2.3}$$

If we now assume that $\underline{y}^{(D)} \sim N_d(\underline{\mu}, \underline{\underline{\Sigma}})$ then $\underline{x} \sim L_d(\mu; \Sigma)$, the logistic-normal distribution, i.e.,

$$f\left(\underline{x}|\underline{\mu}, \underline{\underline{\sum}}\right) = \frac{1}{|2\pi\underline{\underline{\sum}}|^{1/2}\Pi_{i-1}^{D} x_i} \exp\left\{-1/2\left(\log\left(\frac{\underline{x}}{x_D}\right) - \underline{\mu}\right)^T \underline{\underline{\sum}}^{-1}\left(\log\left(\frac{\underline{x}}{x_D}\right) - \underline{\mu}\right)\right\} \tag{2.4}$$

Aitchison and Shen (1980) show that for $\underline{y}^{(k)} \sim N_d(\underline{\underline{Z}}(k)\underline{\mu}, \underline{\underline{Z}}(k)\underline{\underline{\sum}}\underline{\underline{Z}}^T(k))$, the distribution $L_d(\underline{\underline{Z}}(k)\underline{\mu}, \underline{\underline{Z}}(k)\underline{\underline{\sum}}\underline{\underline{Z}}^T(k))$ is simply the appropriate rotation of $L_d(\underline{\mu}, \underline{\underline{\sum}})$ i.e., it is the distribution of $\underline{x}^*$, where $\underline{x}^*$ is $\underline{x}$ but with x_k and x_D interchanged. Consequently any subsequent analysis is unaffected by the choice of reference variable. This invariance property may be extended to time series models and we examine this in Section 3.

When $d = 1$, a_d reduces to the univariate logistic transformation $\log(x/1-x)$ and $L_d(\underline{\mu}, \underline{\underline{\sum}})$ to $L_1(\mu, \sigma^2)$ which is equivalent to the S_B distribution of Johnson (1949) with parameters $\gamma = -\mu/\sigma$ and $\sigma = 1/\sigma$. Thus the a_d transformation and the L_d distribution provide a multivariate generalization of the approach suggested by Wallis (1987).

The moments of the $L_d(\underline{\mu}, \underline{\underline{\sum}})$ distribution, although finite, cannot be evaluated algebraically.

In the time series context the mean is employed to obtain minimum MSE forecasts, and Brunsdon (1987) shows how the mean may be evaluated using quadrature. Aitchison (1989) warns that the mean vector for compositional data may be a poor summary statistic when the distribution is multi-modal because it may lie outside the dense part of the distribution. For the data that we have analyzed this has not been a problem.

In many applications interest centres more naturally on the ratios x_j/x_k or their logarithms. From standard log-normal theory we have, for example,

$$E(x_j/x_k) = \exp\left\{\mu_j - \mu_k + 1/2(\sigma_{jj} - 2\sigma_{jk} + \sigma_{kk})\right\}$$

and

$$\mathrm{Cov}(x_j/x_k, x_i/x_\ell) = E(x_j/x_k)E(x_i/x_\ell)\left\{\exp(\sigma_{ij} + \sigma_{k\ell} - \sigma_{j\ell} - \sigma_{ik}) - 1\right\}$$

where

$$\underline{\underline{\sum}} = \left\{\sigma_{ij}\right\}$$

For further discussion of this see Aitchison and Shen (1980).

3. Compositional Time Series

If a survey is repeated at times $t = 1, \ldots, T$, then multinomial responses at each time t, $\underline{x}_t$ say, lead to compositions

$$\left\{\underline{x}_t : 0 < x_{it} < 1, i = 1, \ldots, d; \quad \sum_{i=1}^{d} x_{it} < 1; \quad t = 1, \ldots, T\right\}$$

which form a multivariate time series.

Transforming the series using the a_d transformation (2.1) produces a multivariate time series defined on $\mathbb{R}^d$ at each time point which can be analyzed using standard methods. In

particular we will examine the use of VARMA models on the transformed series defined by

$$\underline{\phi}(B)\underline{y}_t = \underline{\theta}(B)\underline{\epsilon}_t$$

where

$$\underline{\phi}(B) = \underline{I}_d + \underline{\phi}_1(B) + \ldots + \underline{\phi}_p B^p$$

and

$$\underline{\theta}(B) = \underline{I}_d + \underline{\theta}_1(B) + \ldots + \underline{\theta}_q B^q$$

In the multivariate case we follow the ideas of Tiao and Box (1981) who give a very simple procedure for choosing, estimating and testing such models.

As in the previous section it is necessary to consider if the choice of reference variable in any way influences the analysis. Brunsdon (1987) proves the following results.

Result 1.

Let $\underline{Y}_t = \underline{y}_t - \underline{v}$, where $\underline{v} = E(\underline{y}_t)$, then

$$\begin{aligned}\underline{Y}_t^{(k)} &= \underline{Z}(k)\underline{Y}_t \\ &= \underline{Z}(k)(\underline{y}_t - \underline{v}) = \underline{y}_t^{(k)} - \underline{v}^{(k)}, \quad (t = 0, \pm 1, \ldots), \quad (k = 1, \ldots, d)\end{aligned}$$

where $\underline{Z}(k)$ is given by (2.3). Then if $\left\{\underline{Y}_t\right\}$ follows a VARMA(p,q) process of dimension d, then $\left\{\underline{Y}_t^{(k)}\right\}$ is also VARMA(p,q). Further the roots of the determinantal equations of both the AR and the MA components from the two models are identical, so that the stationarity and invertibility conditions remain consistent.

Result 2.

Consider the compositional time series $\{\underline{y}_t\}$ where $a_d^{(k)}(\underline{x}_t)(k = 1, \ldots, D)$ follows a VARMA(p,q) process. Then each VARMA model $(k = 1, \ldots, D)$ represents the same model for $\underline{x}_t$, except that the elements of $\underline{x}_t^f$ and associated parameters have been permuted. That is, the model for $\underline{x}^f$ is totally invariant to the choice of reference variable.

The consequence of the above two results is that any component of $\underline{x}_t^f$ may be selected as the reference variable without affecting the final results. For the rest of this section, we will assume, without loss of generality, that the reference variable is $x_{D,t}$.

The application of Section 2 to modelling and forecasting is now straightforward and follows the same argument as Wallis (1987). The series $\underline{x}_t$ is transformed to $\underline{y}_t$:

$$\underline{y}_t = a_d(\underline{x}_t)$$

$\left\{\underline{y}_t\right\}$ is then modelled by a VARMA(p,q). It is then a relatively simple matter to obtain forecasts for $\underline{y}_{t+\ell}$. If the ℓ-step ahead forecast of $\underline{y}_{t+\ell}$ is denoted by $\underline{y}_t(\ell)$ and its covariance matrix $\sum_t(\ell)$ then we may obtain the ''naive'' forecast for $\underline{x}_{t+\ell}$ as

$$\underline{x}_t(\ell) = a_d^{-1}\left(\underline{y}_t(\ell)\right)$$

Assuming normality for the distribution of y_t, so that

$$\left(\underline{y}_{t+\ell}|\underline{y}_t, \underline{y}_{t-1}, \ldots\right) \sim N\left(\underline{y}_t(\ell), \underline{\underline{\sum}}_t(\ell)\right)$$

the optimum forecast of $\underline{x}_{t+\ell}, \underline{x}_t(\ell)$ may be found numerically by calculating the mean of $L_d(\underline{y}_t(\ell), \underline{\sum}_t(\ell))$.

From standard multivariate theory a confidence region for $\underline{x}_{t+\ell}$ may also be obtained, although it will not be centred at $\underline{x}_t(\ell)$. A $100(1-\alpha)\%$ confidence region for $\underline{x}_{t+\ell}$ can be formed from

$$\left[\underline{y}_t(\ell) - \log\left\{\frac{\underline{x}_{t+\ell}}{x_{D,t+\ell}}\right\}\right]^T \underline{\underline{\sum}}_t^{-1}(\ell)\left[\underline{y}_t(\ell) - \log\left\{\frac{\underline{x}_{t+\ell}}{x_{D,t+\ell}}\right\}\right] \leq \chi^2_{a;d}$$

where $\chi^2_{a;d}$ is the $\alpha\%$ point of a $\chi^2_{(d)}$ distribution, by mapping points from $\mathbb{R}^d$ onto the simplex S_d, see Figure 1.

Finally, forecasts for either the ratios $x_{i,t+\ell}/x_{j,t+\ell}$, or the log-ratios, may be found. For example

$$(x_i/x_j)_t(\ell) = \exp\left\{y_{it}(\ell) - y_{jt}(\ell) + 1/2(\sigma_{iit}(\ell) - 2\sigma_{ijt}(\ell) + \sigma_{jjt}(\ell))\right\}$$

where $\underline{\sum}_t(\ell) = \left\{\sigma_{ijt}(\ell)\right\}$.

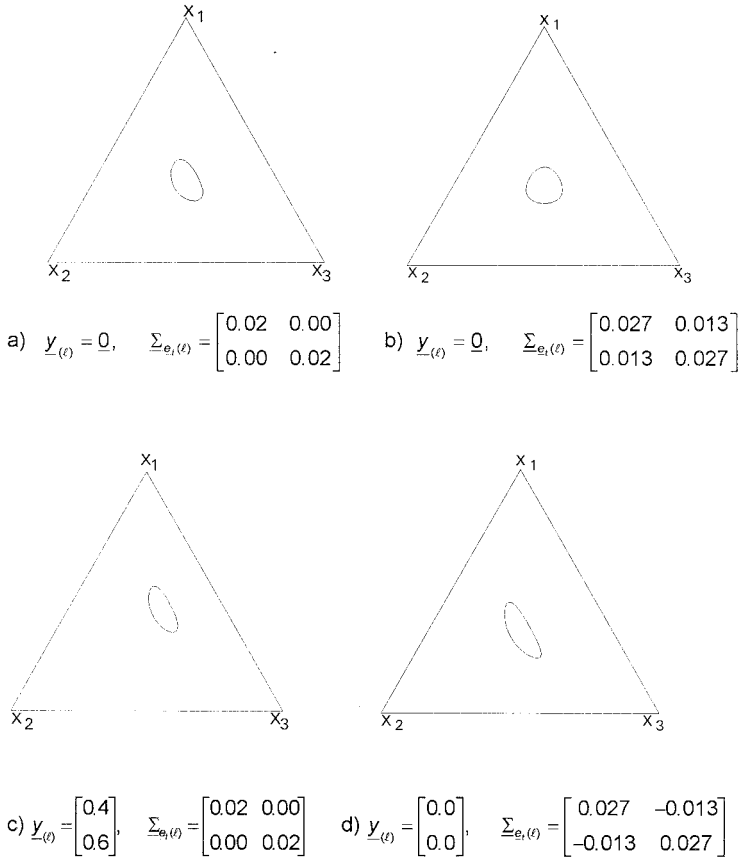


Fig. 1. Confidence regions for $\underline{x}_t$

4. An Application to Labour Market Data

Data from the Australian Labour Force Survey (LFS), provided by the Australian Bureau of Statistics, were available monthly for the period February 1978 to July 1991, a total of 162 observations. The vector of interest is the triple (E_t, U_t, N_t) which defines the numbers in each employment state. For time series analysis it is easier to work with the deseasonalized series. The series for E_t and U_t were available after seasonal adjustment using the X11 routine. The seasonally adjusted series for N_t was deduced by using the constraint that the sum

$$C_t = E_t + U_t + N_t$$

gives the total population which should not be subject to seasonal variation. We denote the seasonally adjusted values by the vector

$$(E_t^*, U_t^*, N_t^*)$$

Our interest centres on the bivariate composition of the relative proportions of employed and unemployed, that is, E_t^*/C_t and U_t^*/C_t, or E_t/C_t and U_t/C_t.

The two compositional time series are plotted in Figures 2 and 3 respectively.

The annual cycle is clearly evident in the raw series, whilst the seasonally adjusted series contains the underlying trend with the seasonal component apparently removed. Thus the formation of N_t^* appears to have been validated. The next step is to form the transformed series. In this case we form

$$\underline{Y}_t = \log \begin{pmatrix} E_t/N_t \\ U_t/N_t \end{pmatrix}$$

and,

$$\underline{Y}_t^* = \log \begin{pmatrix} E_t^*/N_t^* \\ U_t^*/N_t^* \end{pmatrix}$$

The last six observations were removed from the series to enable a comparison of forecasts. Thus the series from February 1978 to January 1991 (156 observations) will be used to forecast February 1991–July 1991. The multiple time series package SCA, Liu et al. (1986), was used to identify and fit a VARIMA model to each of these two series. Summaries of sample cross correlation, and partial autocorrelation matrices, are given in Figures 4 and 5 for each series respectively. Note that there is more than one definition of these matrices; we use the one in the SCA manual, Liu et al. (1986). We also use the conventional ''+'', ''−'', ''.'' notation.

Similar models were identified for both series but it was found necessary to further difference the series. This is evident in Figures 4 and 5 by the pattern of the ACFs which do not appear to decay rapidly. After taking first differences of the $\underline{Y}_t$ series, it was apparent that the seasonal lags likewise indicated non-stationarity. Seasonal differences were also taken and the new ACF/PACF type functions are also given in Figure 4.

For $\underline{Y}_t$ we found $\text{VARIMA}_2(3,0,0)(0,1,1)_{12}$, $\text{VARIMA}_2(0,1,1)(0,1,1)_{12}$, $\text{VARIMA}_2(2,1,0)(0,1,1)_{12}$ and $\text{VARIMA}_2(1,1,1)(0,1,1)_{12}$ to be possible contenders. The seasonal MA component is the obvious choice because of the truncation of the ACF at lag 12, whilst the PACF decays at the seasonal lags. The remaining components depend on the exact interpretation of the early lags of the ACF and PACFs. Each of the above models was fitted and the residual ACF/PACFs etc. examined. All the models fitted reasonably well but

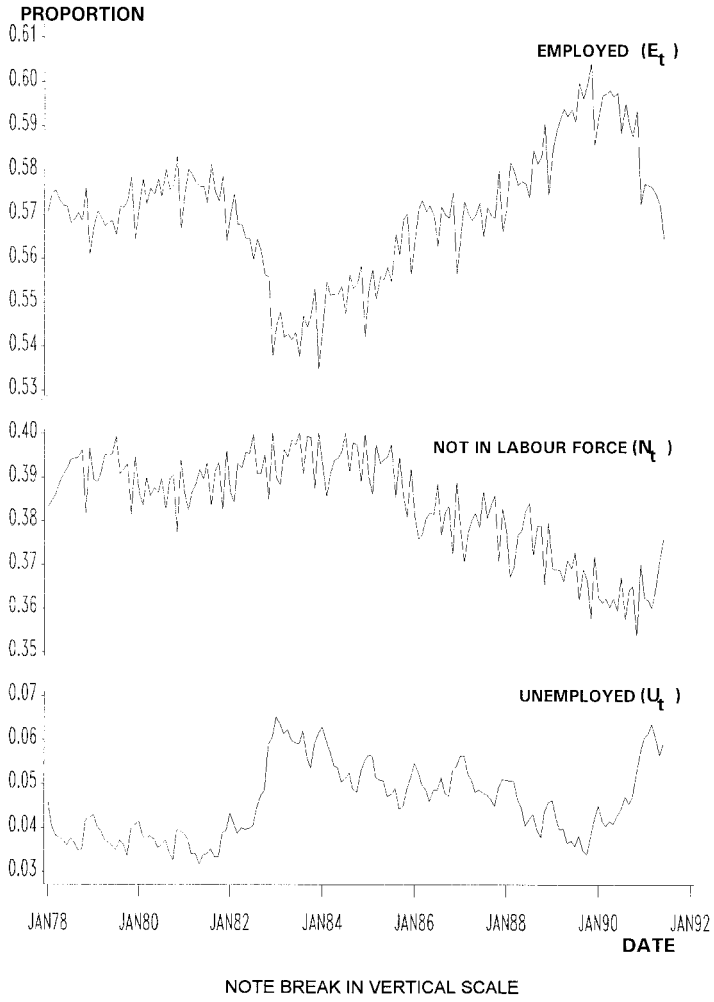


Fig. 2. Australian Labour Force Data expressed as proportions

the worst model was the VARIMA$_2$(0,1,1)(0,1,1)$_{12}$. Marginally the best model was the VARIMA$_2$(1,1,1)(0,1,1)$_{12}$, and so this was selected for forecasting. This model was re-fitted fixing any near zero parameter values to zero, in order to reduce the number of parameters.

The seasonally adjusted series, $\underline{Y}_t^*$, only required ordinary differences, as one might expect, and the resulting ACF/PACF type functions are reported in Figure 5. There is a slight indication of a significant lags in the ACF at lags 12 and 24, whilst the PACF again seems to indicate significant results at all three seasonal lags computed (12, 24 and 36). If we ignore this we can tentatively identify VARIMA$_2$(3,0,0), VARIMA$_2$(0,1,1), VARIMA$_2$(2,1,0) and, VARIMA$_2$(1,1,1) to be possible contenders. When these were fitted it became evident from the residual ACF/PACFs that a seasonal component still remained. Again a seasonal MA(1) seemed the best choice. The VARIMA$_2$(1,1,1)(0,0,1)$_{12}$ was selected as the most appropriate model, and apart from the seasonal difference this is identical to that for $\underline{Y}_t$.

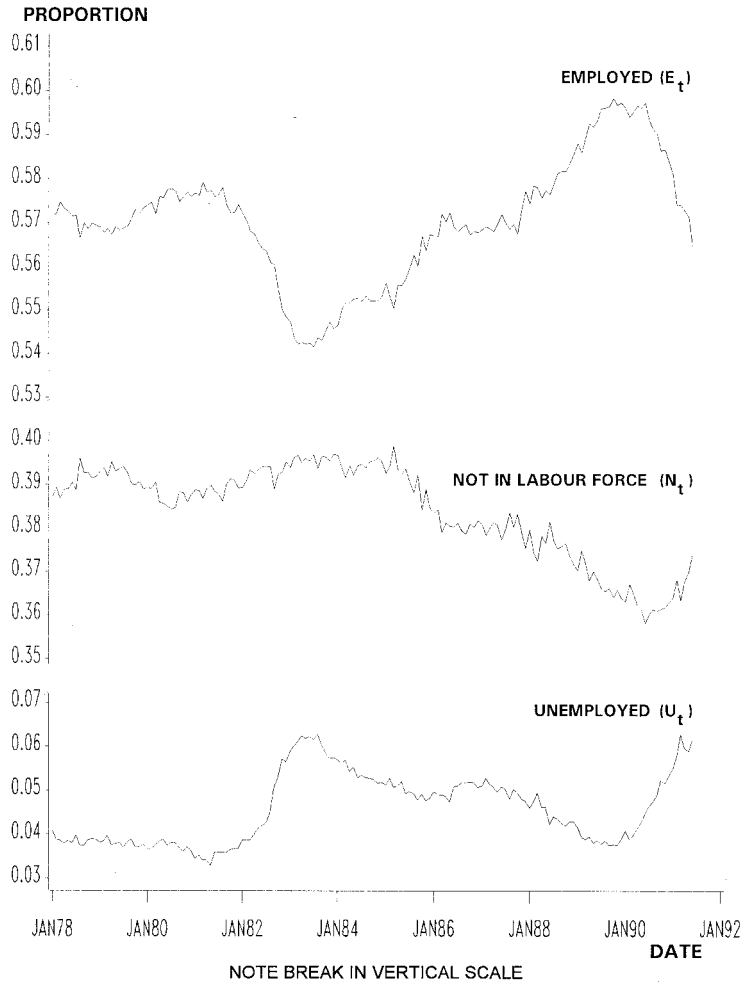


Fig. 3. Australian Labour Force Data (seasonally adjusted) expressed as proportions

The resulting two models fitted were:

$$\left(\underline{I} - \begin{bmatrix} 0.386 & -0.092 \\ (0.097) & (0.017) \\ -1.183 & 0.601 \\ (0.283) & (0.125) \end{bmatrix} B\right)(\underline{I} - B^{12})(\underline{I} - B)\underline{Y}_t = \left(\underline{I} - \begin{bmatrix} 0.684 & 0.000 \\ (0.83) & (\textit{fixed}) \\ 0.000 & 0.629 \\ (\textit{fixed}) & (0.128) \end{bmatrix} B\right)$$

$$\times \left(\underline{I} - \begin{bmatrix} 0.787 & 0.000 \\ (0.049) & (\textit{fixed}) \\ 0.000 & 0.834 \\ (\textit{fixed}) & (0.049) \end{bmatrix} B^{12}\right)\underline{\epsilon}_t$$

$$\text{with } \operatorname{cov}(\underline{\epsilon}_t) = \begin{bmatrix} 0.000106 & 0.000187 \\ 0.000187 & 0.001517 \end{bmatrix}$$

and

$$\left(\underline{I} - \begin{bmatrix} 0.000 & -0.097 \\ (\textit{fixed}) & (0.020) \\ -1.271 & 0.631 \\ (0.288) & (0.107) \end{bmatrix} B\right)(\underline{I} - B)\underline{Y}_t^* = \left(\underline{I} - \begin{bmatrix} 0.240 & 0.000 \\ (0.081) & (\textit{fixed}) \\ 0.000 & 0.582 \\ (\textit{fixed}) & (0.117) \end{bmatrix} B\right)$$

(continued on next page)

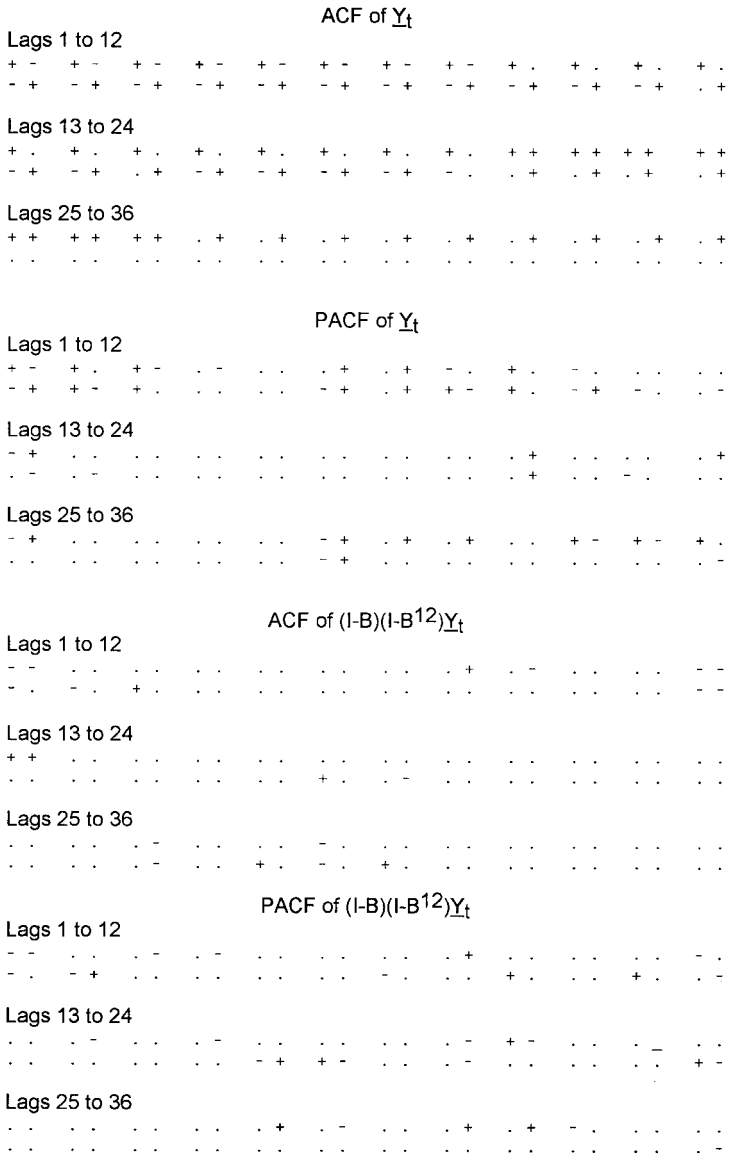


Fig. 4. Cross-correlation and partial autocorrelation functions of $\underline{Y}_t$

$$\times \left(\underline{I} - \begin{bmatrix} 0.305 & 0.000 \\ (0.072) & (\textit{fixed}) \\ 0.000 & 0.317 \\ (\textit{fixed}) & (0.073) \end{bmatrix} B^{12} \right) \underline{\epsilon}_t$$

$$\text{with } \operatorname{cov}(\underline{\epsilon}_t) = \begin{bmatrix} 0.000076 & 0.000117 \\ 0.000117 & 0.001085 \end{bmatrix}$$

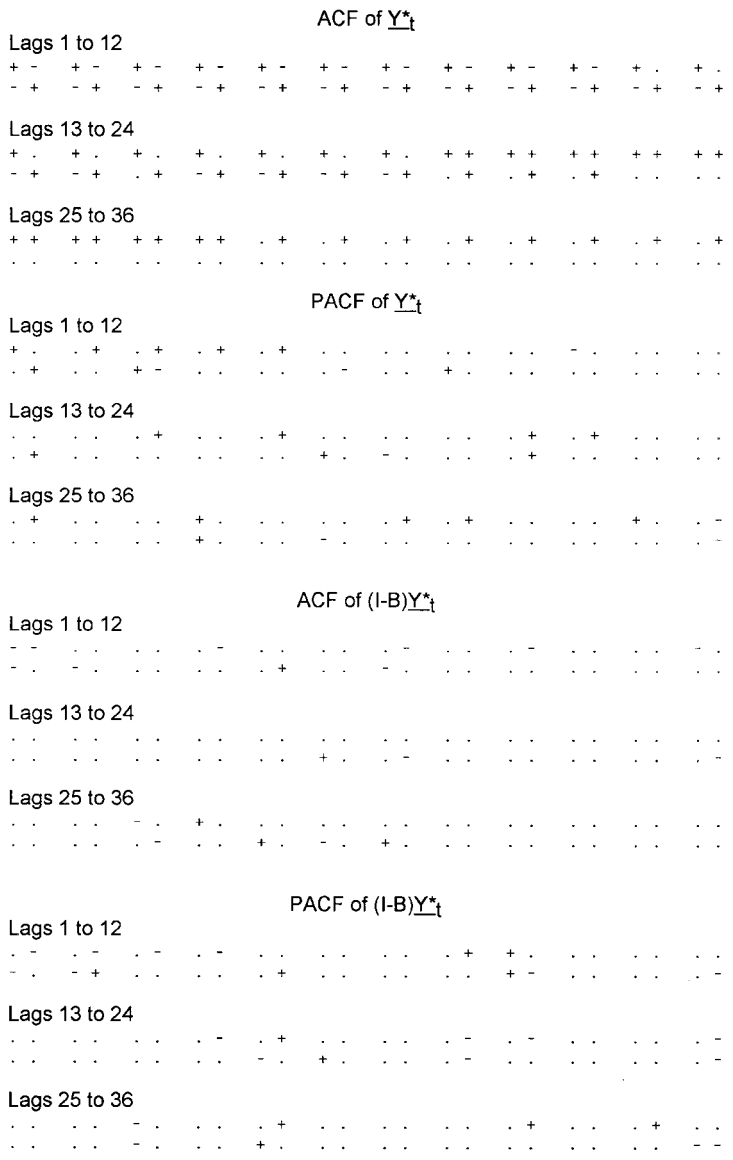


Fig. 5. Cross-correlation and partial autocorrelation functions of $\underline{Y}_t^*$

Table 1. Comparison of actual and forecasted values

Date	Actual proportions	Forecasted proportions
Raw series		
Feb 91	(0.57695, 0.06055, 0.36250)	(0.57902, 0.06054, 0.36044)
Mar 91	(0.57666, 0.06138, 0.36196)	(0.58471, 0.05851, 0.36196)
Apr 91	(0.57619, 0.06366, 0.36015)	(0.58269, 0.05675, 0.36056)
May 91	(0.57457, 0.06030, 0.36513)	(0.58192, 0.05656, 0.36152)
Jun 91	(0.57213, 0.05666, 0.37118)	(0.58107, 0.05503, 0.36390)
Jul 91	(0.56459, 0.05938, 0.37603)*	(0.58135, 0.05495, 0.36371)
Seasonally adjusted series		
Feb 91	(0.58090, 0.05500, 0.36400)	(0.58330, 0.05323, 0.36347)
Mar 91	(0.57390, 0.05780, 0.36820)	(0.58317, 0.05490, 0.36193)
Apr 91	(0.57400, 0.06270, 0.36330)	(0.58194, 0.05506, 0.36299)
May 91	(0.57280, 0.05970, 0.36750)	(0.58095, 0.05485, 0.36419)
Jun 91	(0.57140, 0.05880, 0.36990)	(0.58102, 0.05475, 0.36423)
Jul 91	(0.56450, 0.06160, 0.37391)*	(0.58049, 0.05416, 0.36536)

*No actual values of proportions were available for July 1991, values given are the official predicted values

Using these two models forecasts of $\underline{Y}_t$ and $\underline{Y}_t^*$ were produced for the next six months. These were then converted back to forecasts of the proportions using the inverse transformation

$$\begin{pmatrix} E_t \\ U_t \\ N_t \end{pmatrix} = \begin{pmatrix} e^{Y_{1t}}/(1 + e^{Y_{1t}} + e^{Y_{2t}}) \\ e^{Y_{1t}}/(1 + e^{Y_{1t}} + e^{Y_{2t}}) \\ 1/(1 + e^{Y_{1t}} + e^{Y_{2t}}) \end{pmatrix}$$

A comparison of the resulting forecasts is given in Table 1. It can be seen that the forecasts are reasonably close to the actual values for both series.

We now compute the confidence regions for the forecasts. This may be done using the formula in Section 3. The matrix $\underline{\sum}_t(l)$ may be easily calculated using the standard formula for an l-step ahead forecast in multivariate VARIMA models, namely

$$\underline{\sum}_t(l) = \underline{\sum} + \underline{\Psi}_1\underline{\sum}\underline{\Psi}_1' + \underline{\Psi}_2\underline{\sum}\underline{\Psi}_2' + \ldots + \underline{\Psi}_{(l-1)}\underline{\sum}\underline{\Psi}_{(l-1)}'$$

where the $\underline{\Psi}_i$'s are the usual MA weights. These may be easily output from SCA. The resulting confidence region for the 1-step ahead and the 6-step ahead forecasts are given in Figures 6 and 8. We observe that the confidence regions are very small. A small area has been enlarged and the resulting regions plotted in Figures 7 and 9 for each of the 1 to 6-step ahead forecasts. These use the triangular co-ordinates, so some reference points are also given in the first grid. The actual values are also plotted for comparison.

The small confidence regions in this example are due to the relatively small values of the error covariance matrix. Because of small confidence regions the use of a sophisticated method to estimate the MMSE forecast rather than these ''raw'' forecasts would add an insignificant amount of accuracy to the analyses.

The enlarged confidence regions help to illustrate the success of this approach: all bar the 2-step ahead and 6-step ahead confidence region contain the actual value. The seasonally adjusted series does slightly worse, with the actual value being further from the predicted region of confidence. Also the 3-step ahead forecast is only just inside the

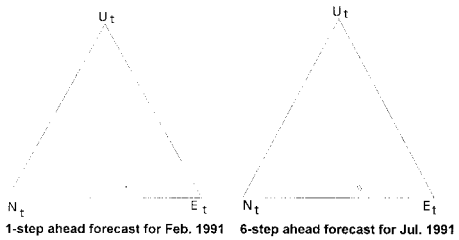
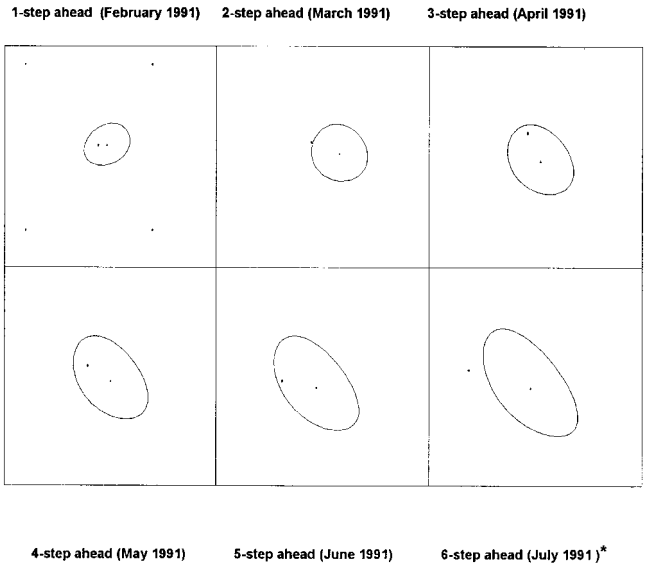


Fig. 6. Confidence regions for forecasts of Australian Labour Force series using the model

$$\left(\mathbf{I}-\begin{bmatrix}0.386 & -0.092\\ -1.183 & 0.601\end{bmatrix}B\right)(\mathbf{I}-B^{12})(\mathbf{I}-B)\mathbf{y}_t=\left(\mathbf{I}-\begin{bmatrix}0.684 & 0.000\\ 0.000 & 0.629\end{bmatrix}B\right)\left(\mathbf{I}-\begin{bmatrix}0.787 & 0.000\\ 0.000 & 0.834\end{bmatrix}B^{12}\right)\boldsymbol{\varepsilon}_t$$

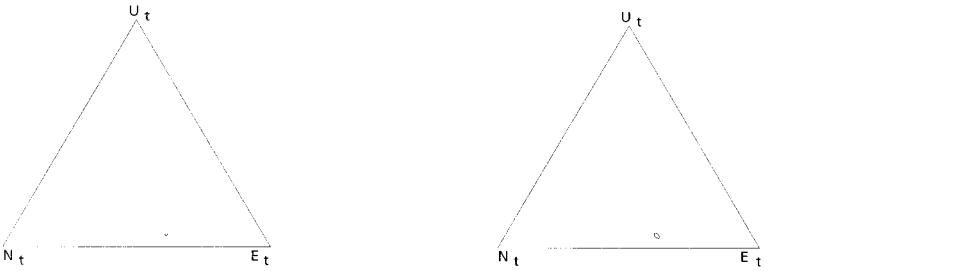
with $\quad \mathrm{cov}(\boldsymbol{\varepsilon}_t)=\begin{bmatrix}0.000106 & 0.000187\\ 0.000187 & 0.001517\end{bmatrix}$



The four points in the corners of the 1-step ahead cell represent some reference points to give an indication of scale. Reading from left to right, and down, the points are, (0.55 0.08 0.37), (0.58 0.08 0.34), (0.57 0.04 0.39) and (0.60 0.04 0.36). The central point in each plot is the forecasted value. The second point is the actual value *(except for July 1991, which is the official predicted value).

Fig. 7. Enlarged 95% confidence regions for the first six forecasts using the original data series

confidence region. Similarly for the non-adjusted series, the 2-step ahead is nearly inside the confidence region. The so-called actual values for July 1991 (6-step ahead) are both outside the regions that our models predict. It transpired that these themselves were predicted values, not actual values, and that the method has picked up this anomaly. The failure to forecast the March 1991 (2-step ahead) actual values could be due either to an unusual occurrence in that month, or to a failure in our model of some kind. However, all-in-all our model does well.



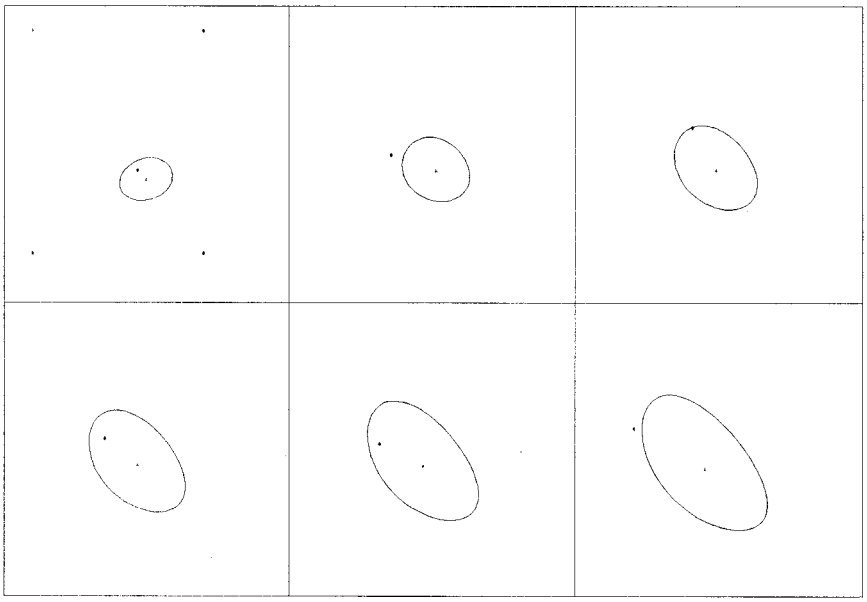
1-step ahead forecast for Feb. 1991 (seasonally adjusted) 6-step ahead forecast for Jul. 1991 (seasonally adjusted)

Fig. 8. Confidence regions for forecasts of the seasonally adjusted Australian Labour Force series using the model

$$\left(\underline{I}-\begin{bmatrix}0.000 & -0.097\\ -1.271 & 0.631\end{bmatrix}B\right)(\underline{I}-B)\underline{y}_t=\left(\underline{I}-\begin{bmatrix}0.240 & 0.000\\ 0.000 & 0.582\end{bmatrix}B\right)\left(\underline{I}-\begin{bmatrix}0.305 & 0.000\\ 0.000 & 0.317\end{bmatrix}B^{12}\right)\underline{\varepsilon}_t$$

$$\text{with}\qquad \text{cov}(\underline{\varepsilon}_t)=\begin{bmatrix}0.000076 & 0.000117\\ 0.000117 & 0.001085\end{bmatrix}$$

1-step ahead (February 1991) 2-step ahead (March 1991) 3-step ahead (April 1991)



4-step ahead (May 1991) 5-step ahead (June 1991) 6-step ahead (July 1991)*

The four points in the corners of the 1-step ahead cell represent some reference points to give an indication of scale. Reading from left to right, and down, the points are, (0.55 0.08 0.37), (0.58 0.08 0.34), (0.57 0.04 0.39) and (0.60 0.04 0.36). The central point in each plot is the forecasted value. The second point is the actual value *(except for July 1991, which is the official predicted value).

Fig. 9. Enlarged 95% confidence regions for the first six forecasts using the seasonally adjusted data series

5. Further Discussion of Results

We initially set out to illustrate the procedure outlined above for the raw series. Software problems led us also to consider the seasonally adjusted series in the belief that no seasonal component would be required in our model. However, we have found that such a component is required, and consequently we have decided to report on both sets of data. The fact that a seasonal component still remains in the seasonally adjusted series indicates a potential problem with the seasonal adjustment procedures employed.

The series E_t^*, N_t^*, U_t^* show slight residual seasonality which seems to be accentuated by the logistic transformation. We could speculate that it was something to do with our method of calculating N_t^*, but this does not explain why N_t^* did not appear to contain a seasonal component. A similar problem has been found when considering the Brazilian Labour Force Survey. The cause of the problem needs further investigation.

When we compare the two models it is also interesting to note the similarity between the parameter estimates. For the non-seasonal parameter estimates they are roughly identical except for the first diagonal terms. The seasonal components also exhibit some similarity; they are both diagonal, with the two parameters being similar in each model. For the seasonal series after seasonal differencing the seasonal MA parameter is about $0.8\underline{\mathrm{I}}$, whilst for the seasonally adjusted series it is about $0.3\underline{\mathrm{I}}$, where $\underline{\mathrm{I}}$ is the identity matrix. Some seasonal adjustment has therefore taken place as expected.

In producing our forecasts we have used the raw transformation of the series. However, because of the small values of the covariance matrix in our model there will be little difference between these and the true MMSE forecast. As seen in Section 3, for time series the MMSE is found from the expected value of the forecast, given that the transformed forecast is (Multivariate)-Normal. For our transformation this may be evaluated numerically but not algebraically. Aitchison (1989) argues that one is often interested in the ratios of one category to another; the forecasts for these can be easily computed, we then have only the log transformation to take account of. Again the methodology was given in Section 3. In the time series context, we may be more interested in the raw series such as the number unemployed. In this case, whenever the components have a covariance matrix of a fairly small order the inverse transformation should suffice.

Another word of warning is that the underlying distribution for the proportions can be multimodal. In such an instance the highest density of probability is not at the mean, and the MMSE forecast may not be what is required. It might be better to quote the modes. From studies in Brunsdon (1987) the problem of multi-modality only occurs for what are relatively high values of the covariance matrix of the parent normal distribution (i.e., $\sum_t(l)$ in this case). Such a series would tend to fluctuate rapidly from one category near 1 to another near 1 and so on, and is rare in this context. Should it occur these methods may be inappropriate.

6. Conclusion

We have illustrated how the problem of forecasting proportions may be dealt with, not only for a univariate series, but also a multivariate series. These techniques may be easily applied to the analysis of repeated surveys, and our example is of one such survey, the Australian Labour Force Survey. The technique is based on applying an instantaneous

transformation which will map the data from the positive simplex S^d to the d-dimensional real space $\mathbb{R}^d$. In particular we suggested the use of the multivariate additive-logistic transformation, a_d, because of its wide application. The transformation requires that one of the compositional variables be used as a reference variable. We have demonstrated that our approach is invariant to the choice of reference variable. Clearly having applied the appropriate transformation, any range of forecasting techniques could be used, provided one takes care in calculating the inverse transformation of the forecasts. In most cases this will be the straight inverse. Not only can other time series models be used, e.g., the structural models of Harvey (1989), but other transformations are available which map S^d to $\mathbb{R}^d$. A transformation may be selected so that, for example, the transformed variables have some further property, e.g., normality or stationarity. The advantages of this general approach for static compositional data have been well investigated and are summarised in Aitchison (1986). Many of these advantages will carry over into this time series context. Previously the only distribution available was the Dirichlet and generalisations of it (e.g., Connor and Mosimann (1969)). These distributions impose a strong independence structure on the data such as neutrality or 'independence except for the constraint.' The 'f-normal' distribution overcomes this problem and allows dependence between the variables $\underline{x}_t \in S^d$ (other than the constraint). The additive-logistic-normal distribution was used by Aitchison (1982) for just this purpose. Applied to compositional time series it is similarly possible to look for relationships between components of $\underline{x}_t \in S^d (t = 0, \pm 1, \ldots)$. In such a context these relationships may be directional as well as instantaneous. Part of the motivation for the use of other transformations could be partly due to the wish to investigate such interactions between each of the series; again Aitchison (1986) summarises many possibilities for the static case. A further area of investigation would be to use this technique to enhance the estimates of a repeated survey, as outlined in the Introduction using a multivariate version of Scott, Smith, and Jones (1977). This should be a fairly straightforward extension, with a little thought needed to transpose the cross-correlation structure of the properties to the transformed series. Again Aitchison (1986) has tackled this for the static situation.

A particular difficulty with the additive logistic transformation is that of zero values, see Aitchison (1986, Ch.11). If any elements of $\underline{x}_t$ are zero the resulting transformed series will take values of $\pm\infty$. One possible solution is to find an alternative transformation. However, many transformations which map S^d to $\mathbb{R}^d$ yield the same results unless the transformed series $\underline{y}_t$ (say) is bounded above and below. The second possibility is to recode zero as some sufficiently small number. For example, if the data is recorded to the nearest decimal place then any value in the range $0 < x < 0.05$ would have been rounded down and recorded as zero. Thus one might re-code zero as 0.025, the mid-point of this range. The success of this technique needs further investigation, but is likely to be adequate for most situations where the data does not contain too many zeros. The effect of recoding zeros will be to set a lower and upper bound on the $\underline{y}_t$ series. Thus the two solutions are virtually identical.

In general the method outlined above for forecasting multivariate time series with a sum-constraint seems to work well. There is scope for further methodological development and for combining with other fields, such as survey analysis. Finally there are new problems to investigate, such as ''seasonal *non* adjustment.''

7. References

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society, Series B, 38, 189–203.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman and Hall, New York.

Aitchison, J. (1989). Measures of Location of Compositional Data Sets. Math. Geol. 21, 787–790.

Aitchison, J. and Shen, S.M. (1980). Logistic-Normal Distributions: Some Properties and Uses. Biometrika, 67, 261–272.

Brunsdon, T.M. (1987). The Time Series Analysis of Compositional Data. Ph.D. Thesis, University of Southampton, U.K.

Connor, R.J. and Mosimann, J.E. (1969). Concepts for Proportions with a Generalization of the Dirichlet Distribution. Journal of the American Statistical Association, 64, 194–206.

Harvey, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Johnson, N.L. (1949). Systems of Frequency Curves Generated by Methods of Translation. Biometrika, 36, 149–176.

Liu, L., Hudak, G.B., Box, G.E.P., Muller, M.E., and Tiao, G.C. (1986). The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis. SCA, Illinois, U.S.A.

Scott, A.J. and 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., and Jones, R.G. (1977). The Application of Time Series Methods to the Analysis of Repeated Surveys. International Statistical Review, 43, 13–28.

Smith, T.M.F. (1978). Principles and Problems in the Analysis of Repeated Surveys. In Survey Sampling and Measurement. N.K. Namboodiri (ed.). Academic Press, New York.

Tiao, G.C. and Box, G.E.P. (1981). Modelling Multiple Time Series with Applications. Journal of the American Statistical Association, 76, 802–816.

Wallis, K.F. (1987). Time Series Analysis of Bounded Economic Variables. Journal of Time Series Analysis, 8, 115–123.

Received April 1991
Revised July 1997