THE TIME SERIES ANALYSIS OF COMPOSITIONAL DATA

T.M. Fred Smith, University of Southampton, U.K.
Teresa M. Brunsdon, Sheffield City Polytechnic, U.K.

KEY WORDS: Repeated surveys, additive logistic transformation, ARMA, dependence.

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

$$y_t = \theta_t + e_t \ , \quad t = 1,2,..., T \ . \qquad (1.1)$$

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 $\theta_{t-1}, \theta_{t-2},...$. 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 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 if the covariance structure of e_t is to be employed in a time series framework then this is much easier if the error structure remains constant over time, implying 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), 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. The problem was given to Teresa Brunsdon, a new research student, and this paper reports some of the results contained in her thesis, Brunsdon (1987). For univariate problems Wallis (1987) provides a similar analysis.

2. Compositional Data

Consider a multinomial response,

$$\underline{r}^T = (r_1,r_2....,r_{m+1}), \quad \sum_{i+1}^{m+1} r_i = n \ , \ \text{which}$$

represents an m dimensional random variable. Let $u_i = r_i/n$, $\underline{u}^T = (u_1,..., u_m)$, then $\underline{u}$ is a composition which lies in the simplex

$$S^m = \left\{ \underline{u} : 0 < u_i < 1, i=1,...,m; \ \sum_{i=1}^{m} u_i < 1 \right\} .$$

The value $u_{m+1} = 1 - \sum_{i=1}^{m} u_i$ is called the fill-up value, or FUV, and is determined by the m values $u_1,...,u_m$. The problems of modelling and analysing compositional data are discussed thoroughly in the excellent monograph by Aitchison (1986). He demonstrates the difficulties of applying standard methods to the composition, $\underline{u}$, 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 first hear of it, is to map $\underline{u}$ from the simplex S^m onto $\mathbb{R}^m$ and then to examine the statistical properties within $\mathbb{R}^m$. He considers several transformations the most important of which is the additive-logistic or a_m transformation defined by:

$$v_i = a_m(u_i) = \log\left[\frac{u_i}{u_{m+1}}\right], \quad (i=1,\ldots,m), \qquad (2.1)$$

where $\quad u_{m+1} = 1 - \sum_{i=1}^{m} u_i$,

with inverse

$$u_i = a_m^{-1}(v_i) = \frac{e^{v_i}}{1 + \sum_{j=1}^{m} e^{v_j}}, (i=1,\ldots,m),$$

$$= \frac{1}{1 + \sum_{j=1}^{m} e^{v_j}} \quad (i=m+1), \qquad (2.2)$$

where u_{m+1} is the FUV. Let $\underline{u}^f$ denote the $(m+1) \times 1$ vector, consisting of $\underline{u}$ augmented by u_{m+1} , so that

$$\left\{\underline{u}^f : 0 < u_i^f < 1 (i=1,\ldots,m+1); \sum_{i=1}^{m+1} u_i^f = 1\right\}$$

represents an alternative definition of a composition.

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

$$a_m^{(k)} : v_i^{(k)} = \log\left[\frac{u_i}{u_k}\right] \qquad (i=1,\ldots,m+1; i \neq k);$$

with inverse

$$a_m^{(k)-1}: u_i = \frac{e^{v_i^{(k)}}}{1 + \sum_{\substack{j=1 \\ j\neq k}}^{m+1} e^{v_j^{(k)}}} \quad (i=1,\ldots,m+1; i \neq k),$$

$$= \frac{1}{1 + \sum_{\substack{j=1 \\ j\neq k}}^{m+1} e^{v_j^{(k)}}} \qquad (i=k) .$$

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,

$\underline{v}^{(k)} = \underline{Z}(k)\underline{v}^{(m+1)}$ where $\underline{Z}(k) = \{z_{ij}(k)\}$,

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

If we now assume that

$$\underline{v}^{(m+1)} \sim N_m(\underline{\mu}, \underline{\Sigma})$$

then $\underline{u} \sim L_m(\mu. \Sigma)$, the logistic-normal distribution

i.e. $$f(\underline{u}\,|\,\underline{\mu}, \underline{\Sigma}) = \frac{1}{|2\pi\Sigma|^{1/2} \prod_{i=1}^{m+1} u_i}$$

$$\exp\left\{-\tfrac{1}{2}\left[\ln\left[\frac{\underline{u}}{u_{m+1}}\right] - \underline{\mu}\right]^T \underline{\Sigma}^{-1} \left[\ln\left[\frac{\underline{u}}{u_{m+1}}\right] - \underline{\mu}\right]\right\} \qquad (2.4)$$

Aitchison and Shen (1980) show that for $\underline{v}^{(k)} \sim N_m(\underline{Z}(k)\underline{\mu}, \underline{Z}(k)\underline{\Sigma}\ \underline{Z}^T(k))$ the distribution $L_m(\underline{Z}(k)\underline{\mu}\ , \underline{Z}(k)\underline{\Sigma}\ \underline{Z}^T(k))$ is simply the appropriate rotation of $L_m(\underline{\mu}, \underline{\Sigma})$ i.e. it is the distribution of $\underline{u}^*$, where $\underline{u}^*$ is $\underline{u}$ but with u_k and u_{m+1} 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 $m = 1$, a_m reduces to the univariate logistic transformation $\log(u/1-u)$ and $L_m(\underline{\mu},\underline{\Sigma})$ 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_m transformation and the L_m distribution provide a multivariate generalization of the approach suggested by Wallis (1987).

The moments of the $L_m(\underline{\mu},\underline{\Sigma})$ distribution, although finite, cannot be evaluated algebraically. However, they may be evaluated numerically by employing a suitable quadrature technique. As an alternative Brunsdon (1987) considers two approximations to the mean, but the preferred solution is numerical evaluation of the mean by quadrature.

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

$$E(u_j/u_k) = \exp\{\mu_j-\mu_k+\tfrac{1}{2}(\sigma_{jj}-2\sigma_{jk}+\sigma_{kk})\}$$

and

$$\mathrm{Cov}(u_j/u_k,\ u_i/u_\ell) = E(u_j/u_k)E(u_i/u_\ell)\{\exp(\sigma_{ij}+\sigma_{k\ell}-\sigma_{j\ell}-\sigma_{ik}) - 1\}$$

where

$$\underline{\Sigma} = \{\sigma_{ij}\}\ .$$

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

3. Compositional time series

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

$$\{\underline{u}_t: 0<u_{it}<1, i=1,\dots,m; \sum_{i=1}^{m} u_{it}<1;\ t=1,\dots,t\}$$

which form a multivariate time series. Transforming the series using the a_m transformation (2.1) produces a multivariate time series defined on $\mathbb{R}^m$ at each time point t which can be analysed using standard methods. In particular we will examine the use of ARMA models on the transformed series defined by

$$\underline{\Phi}(B)\underline{v}_t = \underline{\theta}(B)\underline{\varepsilon}_t\ ,$$

where $\underline{\Phi}(B) = \underline{I}_m + \underline{\Phi}_1 B + \dots + \underline{\Phi}_p B^p$,

and $\underline{\theta}(B) = \underline{I}_m + \underline{\theta}_1 B + \dots + \underline{\theta}_q B^q$.

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

$$\underline{V}_t^{(k)} = \underline{Z}(k)\underline{V}_t$$
$$= \underline{Z}(k)(\underline{v}_t-\underline{\nu}) = \underline{v}_t^{(k)}-\underline{\nu}^{(k)},\quad (t=0,\pm1,\dots),\quad (k=1,\dots,m)\ ,$$

where $\underline{Z}(k)$ is given by (2.3) and $\underline{\mu} = E(\underline{v}_t)$, then if $\{\underline{V}_t\}$ follows a multivariate ARMA(p,q) process of dimension m then, $\{\underline{V}_t^{(k)}\}$ is also multivariate ARMA(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 $\{\underline{u}_t\}$ where $a_m^{(k)}(\underline{u}_t)(k=1,\dots,m+1)$ follows an ARMA(p,q) process. Then each ARMA model $(k=1,\dots,m+1)$ represents the same model for $\underline{u}_t$, except that the elements of $\underline{u}_t^f$ and associated parameters have been permuted. That is, the model for $\underline{u}^f$ is totally invariant to the choice of reference variable.

The consequence of the above two results is that any component of $\underline{u}_t^f$ 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 $u_{m+1,t}$.

The application of section 2 to modelling and forecasting is now straight forward and follows the same argument as Wallis (1987). The series $\underline{u}_t$ is transformed to $\underline{v}_t$:-

$$\underline{v}_t = a_m(\underline{u}_t) \ .$$

$\{\underline{V}_T\}$ is then modelled by the (multivariate) ARMA(p,q). It is then a straight forward matter to obtain forecasts for $\underline{v}_{t+\ell}$. If the ℓ-step ahead forecast $\underline{v}_{t+\ell}$ of $\underline{v}_t$ is denoted by $\underline{v}_t(\ell)$ and its covariance matrix $\Sigma_t(\ell)$ then we may obtain the "naive" forecast for $\underline{u}_{t+\ell}$ as

$$\underline{\nu}_t(\ell) = a_m^{-1}(\underline{v}_t(\ell)) \ .$$

Assuming normality for the distribution of v_t , so that

$$(\underline{v}_{t+\ell}/\underline{v}_{t-1},\ldots) \sim N(\underline{v}_t(\ell) \ , \ \underline{\Sigma}_t(\ell)) \ .$$

the optimum forecast of $\underline{u}_{t+\ell}$, $\underline{u}_t(\ell)$ may be found numerically by calculating the mean of $L_m(\underline{v}_t(\ell), \_\Sigma_t(\ell))$ or $\underline{u}_t(\ell)$ may be approximated.

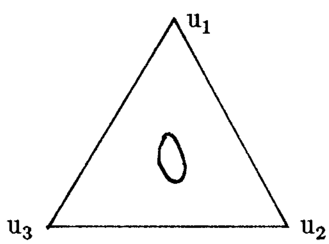
From standard multivariate theory a confidence region for $\underline{u}_{t+\ell}$ may also be obtained, although it will not be centred at $\underline{u}_t(\ell)$. A $100(1-\alpha)\%$ confidence region for $\underline{u}_{t+\ell}$ can be formed from

$$\left[\underline{v}_t(\ell)-\ln\left\{\frac{\underline{u}_{t+\ell}}{u_{m+1,\,t+\ell}}\right\}\right]^T \underline{\Sigma}_t^{-1}(\ell) \left[\underline{v}_t(\ell)-\ln\left\{\frac{\underline{u}_{t\pm\ell}}{u_{m+1,\,t+\ell}}\right\}\right] \leq \chi^2_{\alpha;m}$$

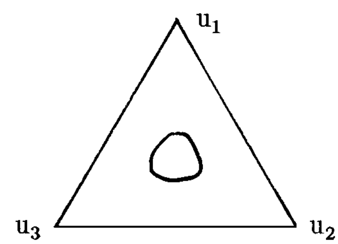
where $\chi^2_{\alpha;m}$ is the $\alpha\%$ point of a $\chi^2_{(m)}$ distribution, by mapping points from $\mathbb{R}^m$ onto the simplex S_m , see Figure 1.

Figure 1.Confidence regions for $\underline{u}_t$

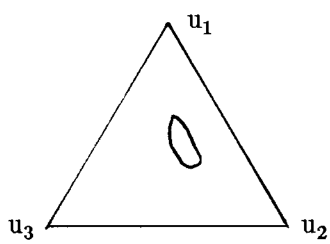
a) $\underline{v}_{t(\ell)} = \underline{0}$, and $\underline{\Sigma}_{\underline{e}_{t(\ell)}} = \begin{bmatrix} 0.02 & 0.00 \\ 0.00 & 0.02 \end{bmatrix}$



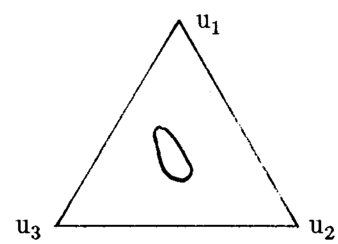
b) $\underline{v}_{t(\ell)} = \underline{0}$, and $\underline{\Sigma}_{\underline{e}_{t(\ell)}} = \begin{bmatrix} 0.026 & 0.013 \\ 0.013 & 0.026 \end{bmatrix}$



c) $\underline{v}_{t(\ell)} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$, and $\underline{\Sigma}_{\underline{e}_{t(\ell)}} = \begin{bmatrix} 0.02 & 0.00 \\ 0.00 & 0.02 \end{bmatrix}$



d) $\underline{v}_{t(\ell)} = \underline{0}$, and $\underline{\Sigma}_{\underline{e}_{t(\ell)}} = \begin{bmatrix} 0.026 & 0.013 \\ 0.013 & 0.026 \end{bmatrix}$



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

$$(u_i/u_j)_t(\ell) = \exp\left\{ v_{it}(\ell) - v_{jt}(\ell) + \tfrac{1}{2}(\sigma_{iit}(\ell) - 2\sigma_{ijt}(\ell) + \sigma_{jjt}(\ell)) \right\}$$

where $\underline{\Sigma}_t(\ell) = \{\sigma_{ijt}(\ell)\}$.

4. Dependence for compositional time series

Aitchison (1986) contains a good discussion of the ideas of dependence and independence for cross-sectional compositional data. The sum constraint on compositional data induces an automatic dependence and if we wish to understand the inter-relationships between compositional time series it is necessary to develop new forms of dependence.

In Section 2 we developed the idea of a composition from multinomial data $\underline{r}^T = (\underline{r}_1,\ldots,r_{m+1})$, $\sum_{i=1}^{m+1} r_i = n$. More generally we can consider any positive variable $\underline{w}^T = (\underline{w}_1,\ldots,w_{m+1})$, $w_i > 0$, $\tau = \sum_{i=1}^{m+1} w_i$, and form the composition

$$C(\underline{w}) = \underline{u} \ ,$$

where $u_i = w_i/\tau$, $i = 1,\ldots, m$, with FUV, $u_{m+1} = w_{m+1}/\tau$. For example, in a family expenditure survey w_i would represent the amount spent on some group of commodities and u_i would be the proportion of expenditure on the group. The variable $\underline{w}$ is called the basis of the composition. If the basis is available then analyses involving the basis are called extrinsic analyses.

One of the earliest notions of extrinsic independence was to consider independence in the basis $\underline{w}$. If

(i) $\coprod \underline{w}$, **(ii)** $\underline{u} = C(\underline{w})$,

then $\underline{u}$ has basis independence. For time series data this independence property must hold at each time point and also across time. If $\underline{u}_t$ has basis independence then the autocovariance function $\underline{\Gamma}(\underline{k})$ of $\underline{v}_t$ has the form

$$\underline{\Gamma}(k) = dg(\gamma_1^{(k)},\ldots,\gamma_m^{(k)}) + \gamma_{m+1}^{(k)} \underline{U}_m, \qquad \gamma_i^{(k)} > 0, \qquad (4.1)$$

for every lag k , which is a very strong condition as might be expected.

In many problems the basis may not be available, only the composition $\underline{u}$ is known, in which case the analysis is termed intrinsic. For example, in geology the composition of a soil or rock sample is all that is known. An intrinsic analysis examines amalgamations and subcompositions. For the composition, $\underline{u}_t$, we have the partition

$$\underline{u}_t \equiv (\underline{u}_{t(c)}, \underline{u}_t^{(c)}, \tau_{t(c)}), \qquad (4.2)$$

where

$$\underline{u}_{t(c)} = (u_{t,1},\ldots, u_{t,c}) \in S^c,$$

$$\underline{u}_t^{(c)} = (u_{t,c+1},\ldots,u_{t,m+1}) \in S^{m-c+1},$$

and

$$\tau_{t(c)} = \sum_{j=c+1}^{m+1} u_{t,j} \ .$$

Hence $\tau_{t(c)}$ is the FUV for $\underline{u}_{t(c)}$, and is an amalgamation of some of the elements of $\underline{u}_t$. Interest centres on whether one subcompositon is independent of another, for example, whether in a public opinion poll we can study the vote for the major parties independent of the votes for minor parties. If the subcompositions formed from all partitions are independent then the composition has complete subcompositional independence. Aitchison (1986) shows that the Dirichlet distribution possesses this and other stronger properties which is why it is unsuitable for modelling compositional structures.

For time series of compositions the cross sectional ideas of dependence must be extended to embrace time series ideas of dependence or causality. Brunsdon (1987) adopts the Wiener-Granger-Geweke causality framework of linear predictability, see Geweke (1982, 1984) Granger (1969). If X_t is the present value, X_{tp} the set of past values, X_{tpp} the set of past and present values and Ω the universe of information then for two series X_t, Y_t, we have:

(a) Causality : $Y_t \rightarrow X_t$.

If $\mathrm{Var}(X_t|\Omega_{tp}) < \mathrm{Var}(X_t|\Omega - Y)_{tp})$ then Y causes X.

(b) Feedback : $Y_t \leftrightarrow X_t$

If $Y_t \rightarrow X_t$ and $X_t \rightarrow Y_t$ then we have feedback between Y and X.

(c) Instantaneous Causality : $Y_t \cdot X_t$.

If $\mathrm{Var}(X_t | \Omega_{tp}, Y_{tpp}) < \mathrm{Var}(X_t | \Omega_{tp})$ then Y is instantaneously causing X and vice versa.

The two extreme cases are linear independence $Y_t \amalg X_t$, and complete dependence $Y_t \Longleftrightarrow X_t$, the latter holding if (a), (b), (c) all hold.

Following Geweke (1984) Brunsdon (1987) derives a sequence of tests of sub-compositional dependence over time for a specific model. The number of possibilities is very large and details will be given in a subsequent paper. The basis of the tests is to map the partition of $\underline{u}_t \in S^m$ onto $\underline{v}_t \in \mathbb{R}^m$ and to assume that $\underline{v}_t$ can be approximated by an $AR_m(p)$ process for some value of p . For the partition in (4.2) we have

$$\underline{v}_t \equiv (\underline{v}_{t(c)}, \underline{v}_t^{(c)}, \overset{*}{\tau}_{t(c)}), \ \overset{*}{\tau}_{t(c)} = \sum_{i=c+1}^{m} v_{ti} \quad ,$$

where $\underline{v}_{t(c)} = a_c(\underline{u}_{t(c)})$ etc., and the $AR_m(p)$ model is of the form

$$\begin{bmatrix} \underline{\Phi}_{11}(B) & \underline{\Phi}_{12}(B) & \underline{\Phi}_{13}(B) \\ \underline{\Phi}_{21}(B) & \underline{\Phi}_{22}(B) & \underline{\Phi}_{23}(B) \\ \underline{\Phi}_{31}(B) & \underline{\Phi}_{32}(B) & \underline{\Phi}_{33}(B) \end{bmatrix} \begin{bmatrix} \underline{v}_{t(c)} \\ \underline{v}_t^{(c)} \\ \overset{*}{\tau}_{t(c)} \end{bmatrix} = \begin{bmatrix} \underline{a}_{t(c)} \\ \underline{a}_t^{(c)} \\ \overset{*}{a_t} \end{bmatrix},$$

where $\underline{a}_t$ is a white noise process with

$$\mathrm{Var}(\underline{a}_t) = \underline{\Sigma} = \begin{bmatrix} \underline{\Sigma}_{11} & \underline{\Sigma}_{12} & \underline{\Sigma}_{13} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} & \underline{\Sigma}_{23} \\ \underline{\Sigma}_{31} & \underline{\Sigma}_{32} & \underline{\Sigma}_{33} \end{bmatrix} .$$

The various forms of dependence can be modelled by making assumptions about the coefficients $\underline{\Phi}_{ij}$ and the covariances $\underline{\Sigma}_{ij}$. Tests are based on natural logarithms of ratios of the determinants of suitably chosen estimates of the residual covariance matrices.

5. An application to public opinion polls.

Gallup poll data in the U.K. were available monthly for the nine years from January 1965 to December 1973 with no major change in design. The best fitting multivariate ARMA model was an ARMA (1,1), which is consistent with the univariate analysis carried out by Scott, Smith and Jones (1977) which took into account the rotating survey design. For the purposes of testing dependence this was approximated by an AR(2) model, which fitted almost as well. To justify analysing the data for the three main parties independent of votes for minor parties we tested for the independence of $\underline{u}_{t(c)}$ = (CON, LAB, LIB) from $\underline{u}_t^{(c)}$ = OTHER, DON'T KNOW) . The tests supported the hypothesis that $\underline{u}_{t(c)}$ may be modelled independently of $\underline{u}_t^{(c)}$. However the partition into (CON, LAB) alone, namely the two major parties, was not independent of the LIB party.

The subcomposition formed from $\underline{u}_{t(c)}$ was modelled by an $ARMA_2(1,1)$ process and the results were compared with those of univariate modelling. As so often happens in multivariate time series analysis there were no obvious gains from fitting multivariate models over univariate models apart from the built in consistency of the results. The main benefits appeared at the earlier stage when justifying the analysis of the subcomposition.

6. CONCLUSION

In this paper a model for compositional time series has been proposed which could be used for modelling data from repeated sample surveys. The approach is to apply an instantaneous transformation which will map the data from the positive simplex S^m to the m-dimensional real space $\mathbb{R}^m$. In particular we have suggested the use of the multivariate additive-logistic transformation, a_m , because of its wide application. This 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. Forecasts may be obtained by finding the mean of the appropriate additive-logistic-normal distribution. A numerical integration routine may be used for this purpose.

It is possible to generalise the procedure we have explored by considering any transformation f (say) which maps S^m to $\mathbb{R}^m$. A transformation may be selected so that, for example, the transformed variables are not only on $\mathbb{R}^m$, but also 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. An important example is that a whole range of transformed-normal distributions become available to describe $\underline{u} \in S^m$. Previously the only distribution available was the Dirichlet and generalizations 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 $\underline{u} \in S^m$ (other than the linear 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 the components of $\underline{u}_t \in S^m$ ($t = 0, \pm 1, \ldots$). In such a context these relationships may be directional as well as instantaneous.

A particular difficulty with the additive logistic transformation is that of zero values, see Aitchison (1986, Ch.11). If any elements of $\underline{u}_t$ are zero the resulting transformed series will take values of $\pm\infty$. Two possibilities are obvious, the first is to find an alternative transformation. However, many transformations which map S^m to $\mathbb{R}^m$ yield the same results unless the transformed series $\underline{v}_t$ (say) is bounded above and below. The second possibility is to re-code zero as some sufficiently small number. For example, if the data is recorded to the nearest decimal place then any value in the range $0 \leq u < 0.05$ 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 set a lower and upper bound on the $\underline{v}_t$ series. Thus the two solutions are virtually identical.

In summary the use of the additive-logistic transformation provides a practical approach to modelling compositional time series. This may be expanded to include other transformations and investigate a variety of properties such as the causal relationships between variables.

Acknowledgement: This research was supported by a gramt from the ESRC.

REFERENCES

AITCHISON J. (1982) The statistical analysis of compositional data. *J. Roy. Statist. Soc.* **B**, 38, 189-203.

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

BRUNSDON T.M. (1987) *The time series analysis of compositional data*, PhD thesis, University of Southampton, U.K.

CONNOR R.J. and MOSIMANN J.E. (1969) Concepts for proportions with a generalization of the dirichlet distribution. *J. Amer. Statist. Assoc.* 64 194-206.

GEWEKE J. (1982) Measurement of linear dependence and feed back between multiple time series. *J. Amer. Statist. Assoc.* 77 304-324.

GEWEKE J. (1984) Measure of conditional linear dependence and feedback between time series. *J. Amer. Statist. Assoc.* 79 907-915.

GRANGER C.W.J. (1969) Investigating causal relationships by econometric models and cross-specral models. *Econometrica*, **37**, 424-438.

SCOTT A.J. and SMITH T.M.F. (1974) Analysis of repeated surveys using time series methods. *J. Amer. Statist. Assoc.* 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. *Inter. Statist. Review.* 43 13-28.

SMITH T.M.F. (1978) Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement.* Ed. N.K. Namboodiri, Academic Press, New York.

TIAO G.C. and BOX G.E.P. (1981) Modelling multiple time series with applications. *J. Amer. Statist. Assoc.* 76 802-816.

WALLIS K.F. (1987) Time series analysis of bounded economic variables. *J. Time Series Analysis*, 8, 115-123.