Compositional Time Series Analysis: A Review

Aguilar Zuil, Lucía
Universidad de Extremadura, Dept. de Matemáticas, Escuela Politécnica
Avda. de la Universidad s/n
10071 Cáceres, Spain
E-mail: lucia@unex.es

Barceló-Vidal, Carles
Universitat de Girona, Dept. Informàtica i Matemàtica Aplicada
Campus de Montilivi
17071 Girona, Spain
E-mail: carles.barcelo@udg.es

Larrosa, Juan M.
CONICET, Universidad Nacional del Sur, Dept. de Economía
San Juan y 12 de Octubre, planta baja, gabinete 5
8000 Bahía Blanca (Buenos Aires), Argentina
E-mail: jlarrosa@criba.edu.ar

1. Introduction

Compositional data are inherently multivariate by nature but are characterized by the distinguishing feature that they are comprised of non-negative components which sum to a constant. Without loss of generality, it can be assumed that the constant in question is 1. More than twenty years have elapsed since Aitchinson's pioneering contributions to the field were published (see Aitchinson (1986) and references therein). These contributions to the literature were seminal in the sense that they were the first publications which developed statistical methods specifically designed for the analysis of compositional data and illustrated their use in the analysis of real compositional data. Nevertheless, a number of earlier studies, primarily involving the use of multivariate methods in the analysis of geological data, had discussed and criticized the application of standard multivariate techniques which ignored the non-negativity and unit-sum constraints inherent in compositional data. Indeed, Pearson, as far back as 1897, had warned of the consequences of using such techniques to explore the correlations between the components making up a composition. Moreover, it is undoubtedly the case that the seminal works of Aitchinson referred to above acted as the catalyst for subsequent developments in the field of the statistical analysis of compositional data. Whilst contributions from other authors were initially scarce, research in the field has, in recent years, exhibited renewed and increasing interest. Aitchinson & Egozcue (2005) document the major historical developments in the field and suggest potential directions for future research.

This paper focuses on a particular type of compositional data, namely multivariate time series comprised of compositional data. In what follows, we will use the abbreviation CTS when referring to a "compositional time series". Data of this kind frequently arise in disciplines as disparate as biology, demography, ecology, economics, geology and politics. Examples are: the percentages of different species of fish recorded in a lake at different instants in time; the composition of monthly immigration to a city according to the country of origin; the daily market shares at the end of trading; the breakdown of household monthly consumption by type of item in budget surveys; and the results of opinion polls conducted at different times during an election campaign. In Section 2 we review developments in the field of the statistical analysis of CTS. Historically, the main approach to analyzing CTS data has

been based on the application of an initial transform to break the unit sum constraint, followed by the use of standard time series techniques. Finally, the inverse transformation is used on the derived results so as to obtain results pertinent to the original sample space. Section 2 is structured around the various forms of transformation that have been proposed within the literature, although we also present the details of an alternative modelling approach in which the data are modelled directly in the space in which they were originally observed. The paper ends, in Section 3, with a brief summary and some concluding remarks.

2. Approaches to the Analysis of Compositional Time Series

2.1 The additive log-ratio transformation

Let $\mathbf{x}_t = (x_{t1}, \ldots, x_{tD})', t = 1, 2, , \ldots, n$ denote a k-dimensional CTS such that $x_{tj} > 0$ for $j = 1, \ldots, D$ and $\sum_{j=1}^{D} x_{tj} = 1$ at each time t. The additive log-ratio (alr) transformation of $\mathbf{x}_t$ produces the vector $\mathbf{y}_t$ in $\Re^d$ with components $y_{tj} = alr(x_{tj}) = \log(x_{tj}/x_{tD}), j = 1, \ldots, d, t = 1, 2, \ldots, n$, where $d = D - 1$. Thus, the alr transformation is a one-to-one transformation from the natural sample space, namely the simplex $S^D = \{(x_1, \ldots, x_D)' : x_1 > 0, \ldots, x_D > 0; x_1 + \ldots + x_D = 1\}$, to $\Re^d$. The inverse of the alr transformation is known as the additive logistic transformation. The original idea of applying the alr transformation in the analysis of compositional data is due to Aitchison (see Aitchison (1986)). In the context of time series, the use of the alr transformation has been common practice in the analysis of univariate time series of proportions. However, in such analyses, no reference is generally made to the compositional nature of the data. Wallis (1987) can be considered as the pioneer of this approach. In the remainder of our discussion we will concentrate on the multivariate case for which the associated research has been far less extensive.

Since the transformed time series, $\mathbf{y}_t$, is an unconstrained multivariate time series in $\Re^d$, standard multivariate techniques can be used to analyze it. Thus, the possibility of vector ARMA (VARMA) modelling springs immediately to mind. Such an approach is based on the use of VARMA models to obtain estimates and forecasts for the transformed series, followed by the application of the additive logistic transformation to obtain the equivalent inferential results for the original CTS. This approach was first discussed by Brunsdon (1987) in the context of analyzing CTS from repeated sample surveys. In Brunsdon (1987), Smith & Brunsdon (1989) and Brunsdon & Smith (1998) the authors first proved that such an approach is invariant to the choice of the component used as the common divisor in the alr transformation. Secondly, assuming normality for the distribution of $\mathbf{y}_t$, they obtained forecasts for the original CTS, $\mathbf{x}_t$, by calculating the mean of the corresponding additive logistic distribution numerically. They also derived confidence regions for $\mathbf{x}_t$. The applications they considered involved compositional data from a Gallup poll conducted in the U.K. as well as data from the Australian Labour Force Survey. Also in the field of repeated sample surveys, Silva (1996) and Silva & Smith (2001) made use of the alr transformation too, but then employed a state space modelling approach for the transformed time series. They proved that their approach is also invariant to the choice of the component used in the common divisor of the alr transformation, and illustrated its use in the analysis of CTS data from the Brazilian Labour Force Survey.

Ravishanker, Dey & Iyengar (2001) generalised the approach of Brunsdon & Smith (1989) in the sense that they used an extension of VARMA models incorporating covariates. They assumed that the transformed alr time series, $\mathbf{y}_t$, follows a regression model with VARMA normal distributed errors. The procedures required to fit such models and carry out inference for them are really rather complex. The authors address the possible non-uniqueness of the fitted model using a Bayesian hierarchical approach to model selection. They also carried out a Monte Carlo experiment to estimate the expected proportions for the compositions, based on samples drawn from a simulated posterior density function. Their empirical results for CTS mortality data from Los Angeles showed the utility

of their approach in explaining the dependence of different categories of mortality on air quality.

Ratnaparkhi & Krishnamurthy (2002) extended yet further the potential application of the approach of Brunsdon & Smith (1989). They added another component to the regression model with VARMA normal errors which allows for the potential heteroscedasticity of the transformed *alr* time series. Specifically, they proposed fitting generalised autoregressive heteroscedastic (GARCH) models to the *alr* transformed series. They illustrated the use of this approach in the analysis of data on the micro-finance system employed in Maharashtra State (India). More precisely, they analysed a monthly CTS consisting of the proportions for four categories of loans made by self-help groups using a model incorporating a covariate that represented the savings of the different self-help groups. In their analysis they made use of cross-validation; fitting their model to part of the data and using the remaining of the data to evaluate the quality of the predicted proportions.

2.2 Box-Cox transformation

Aitchison (1986) introduced the use of the well-known Box-Cox transformation as an attractive alternative to the *alr* transformation. The Box-Cox transformation has the advantage of including the *alr* transformation as a special case. However, the only application of this approach that we are aware of is that presented in Bhaumik, Dey and Ravishanker (2003). These authors modelled the Box-Cox transformed data using dynamic linear models incorporating a rich class of distributions for the errors based on scale mixtures of multivariate normal distributions. This general class of distributions includes as special cases the multivariate normal, Student-t, logistic and stable distributions, amongst others. Bhaumik, Dey and Ravishanker (2003) used the same complex procedures as those proposed in Ravishanker, Dey & Iyengar (2001) to carry out model selection and inference. They illustrated their approach using two CTS; the mortality data from Los Angeles (analysed previously by Ravishanker, Dey & Iyengar (2001)), and a CTS on vehicle production which had been previously analysed by Grunwald (1987).

2.3 Centered log-ratio transformation

The centered log-ratio transformation (*clr*) was also proposed by Aitchinson (1986) as a means of transforming compositional data into data distributed throughout D-dimensional real space. The *clr* transformation of a CTS, $\mathbf{x}_t$, is defined as $\mathbf{y}_t = \log(\mathbf{x}_t/g(\mathbf{x}_t))$, where $g(.)$ denotes the geometric mean. Compared with the *alr* transformation, the *clr* transformation has the advantage of not requiring a reference component, but has the disadvantage of a singularity due to the fact that $\sum_{j=1}^{D} y_{tj} = 0$ for all t. Quintana & West (1988) were the first to use the *clr* transformation to analyse CTS data. They modelled *clr* transformed data on monthly Mexican imports using a type of dynamic regression model which allowed for subjective as well as exogenous interventions. Their approach assumes a multivariate logistic normal distribution for the underlying process. They resolved the singularity problem by transforming $\mathbf{y}_t$ to $\mathbf{y}_t'K$, where $K = I - D^{-1}11', 1 = (1, \ldots, 1)'$. Brandt, Monroe & Williams (1999) made use of the same solution to the singularity problem when they employed VAR normal models incorporating an unlagged covariate to model *clr* transformed CTS data from U.S. Gallup Polls.

2.4 Direct modelling in the simplex

Around the same time as the publication of Brunsdon (1987) and Quintana & West (1988), Grunwald (1987) introduced a rather different approach to analysing CTS, which had also been inspired by some of the earlier ideas of Aitchinson. There, and in Grunwald, Raftery & Guttorp (1993), the authors developed space state models which could be used to model CTS data directly in the simplex. The distribution of the CTS conditioned on the unobserved state was assumed to be Dirichlet (see Aitchinson (1986)). The state distribution was assumed to be Dirichlet conjugate. This was a new

generalisation of the Dirichlet distribution proposed by them in order to allow for dependence between the components. They illustrated the application of their model using CTS data from the U.S. Federal Government as well as data on global motor vehicle production.

3. Concluding remarks

We have considered two general approaches to analysing compositional time series data; one involving transformation from the simplex to real space and a second which models the data directly in the simplex. The former approach makes use of the *alr*, Box-Cox or *clr* transformations as well as numerous established techniques for analysing multivariate time series. However, there are potentially other transformations as well as different time series modelling approaches that one might consider using in the future. The second approach is based on the use of the Dirichlet model and extensions thereof. In a similar vein, the possibility of using other models defined on the simplex to model the data directly suggests itself as a potential line of future research.

In the majority of the publications in the field, the authors make use of the transformations proposed by Aitchinson and concentrate more on the time series analysis of the transformed data than on the issues related to their original compositional nature. One important issue of this kind is the problem of dealing with 0's and 1's which may occur at any instant t. Another is the problem of the singularity associated with the *clr* transformation. A further issue which certainly deserves consideration is the objective comparison of the different approaches and an evaluation of the quality of forecasts produced by them.

REFERENCES

Aitchinson, J. (1986) *The Statistical Analysis of Compositional Data*. Chapman& Hall, London.

Aitchinson, J. & Egozcue, J. J. (2005) Compositional data analysis: where are we and where should we be heading? *Math. Geol.*, **37**, 7, 829–850.

Bhaumik, A., Dey, D. K. & Ravishanker, N. (2003) A dynamic linear model approach for compositional time series analysis. Tech. Report, Univ. Connecticut.

Brandt, T. B., Monroe, B. L. & Williams, J. T. (1999) Time series models for compositional data. Tech. Report, Indiana Univ.

Brunsdon, T. M. (1987) The time series analysis of compositional data. Ph.D. Thesis, Univ. Southampton.

Brunsdon, T. M. & Smith, T. M. F. (1998) The time series analysis of compositional data. *J. Off. Statist.*, **14**, 237–253.

Grunwald, G. K. (1987) Time series models for continuous proportions. Ph.D. Thesis, Univ. Washington.

Grunwald, G. K. Raftery, A. E. & Guttorp, P. (1993) Time series of continuous proportions. *J. Roy. Statist. Soc. B*, **55**, 103–116.

Pearson, K. (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. Roy. Soc. London*, **VX**, 489–502.

Quintana, J. M. & West, M. (1988) The time series analysis of compositional data. *Bayesian Statist.*, **3**, 747–756.

Ratnaparkhi, M. V. & Krishnamurthy, R. (2002) Compositional multivariate time series analysis of the savings and loans by the micro finance institutions in Maharashtra State (India). Proc. IAOS Conf. London.

Ravishanker, N., Dey, D. K. & Iyengar, M. (2001) Compositional time series analysis of mortality proportions. *Commun. Statist.-Theory Meth.*, **30**(11), 2281–2291.

Silva, D. B. N. (1996) Modelling compositional time series from repeated surveys. Ph.D. Thesis, Univ. Southamptom.

Silva, D. B. M. & Smith, T. M. F. (2001) Modelling compositional time series from repeated surveys. *Survey Meth.*, **27**, 205–215.

Smith, T. M. F. & Brunsdon, T. M. (1989) The time series analysis of compositional data. *Proc. Amer. Statist. Assoc.*, 26–32.

Wallis, K. F. (1987) Time series analysis of bounded economic variables. *J. Time Series Anal.*, **8**, 115–123.