Time Series of Proportions: A Compositional Approach

C. Barceló-Vidal[1] and L. Aguilar[2]

[1] Dept. Informàtica i Matemàtica Aplicada, Campus de Montilivi, Univ. de Girona, E-17071 Girona, Spain – carles.barcelo@udg.edu
[2] Dept. de Matemáticas, Escuela Politécnica, Univ. de Extremadura, E-10071 Cáceres, Spain – luciaaz@unex.es

Abstract: Taking account of the compositional nature of proportions, the logistic function is the natural transformation to apply when analyzing and modelling time series of continuous proportions. From the metric structure on the simplex, an isomorphic structure is defined on the set of continuous proportions. This structure permits the translation of the standard analysis of univariate time series to the compositional analysis of proportions in which the logistic transformation arises naturally and not as a mere alternative to the transformation of the data.

Keywords: Compositional data; log odds ratio; logistic transformation; simplex; time series of proportions.

1 Introduction

Univariate time series (TS) of proportions, p_t, arise in a wide variety of applications. Authors often ignore the restricted range of variation of the p_t, namely $(0, 1)$, and use standard techniques to model TS of proportions (e.g., Box and Jenkins, 1976; Tiller, 1992). Such analyses can result in estimates of proportions erroneously lying outside the interval $(0, 1)$.
Wallis (1987) was the first author to propose the logistic transformation $y_t = \operatorname{logit} p_t = \log(p_t/(1 - p_t))$ as an appropriate transformation for TS of proportions. The arguments given by him for the use of the logit transformation are: i) the necessity to stabilize the variance and make the transformed data approximately normally distributed, and ii) to ensure that estimates and projections lie within $(0, 1)$. Other authors use the logarithmic transformation as an alternative transformation to model TS of proportions because it can be useful as a means of stabilizing the variance and normalizing transformed data. However, it is not a guarantee that estimates and projections will lie in $(0, 1)$. It would appear then that the analysis and modelling of TS of proportions simply requires, if really necessary, the identification of the appropriate transformation in each case. However, in our opinion, this transformation based approach ignores the compositional nature of proportions. Any proportion p_t inevitably has associated

with it the complementary proportion $1 - p_t$, and thus the modelling of a TS of proportions p_t should be based upon the time series of compositions $\mathbf{p}_t = (p_t, 1 - p_t)'$ in the simplex $\mathcal{S}^2$.
This paper presents the compositional approach to the modelling of TS of proportions based on the initial work of Barceló-Vidal *et al.* (2007) which uses the compositional data analysis methodology introduced by Aitchison (1986) and subsequently developed by Egozcue *et al.* (2006). In Section 2 we present the Euclidean vector space $(\mathcal{P}, \oplus, \odot)$ of proportions in $(0, 1)$ in correspondence with the simplex $(\mathcal{S}^2, \oplus, \odot)$. The algebraic structure of $\mathcal{P}$ serves as the basis of the development, in Section 3, of the compositional approach to the analysis of TS of proportions and to introduce the compositional ARIMA models. There it is shown that the logit transformation is the natural transformation which should be applied when attempting to analyze or model TS of proportions as it is the one that takes into account their compositional nature.

2 Continuous proportions as a compositional space

2.1 The metric space of continuous proportions

Let $\mathcal{P}$ be the set of continuous proportions $p \in (0, 1)$. We identify a proportion p with the 2-part $\mathbf{p} = (p, 1 - p)' \in \mathcal{S}^2$ and therefore we can easily translate to $\mathcal{P}$ the structure defined in $(\mathcal{S}^2, \oplus, \odot)$. The *perturbation* of p and p^* in $\mathcal{P}$ will be denoted as

$$p \oplus p^* = \frac{pp^*}{pp^* + (1-p)(1-p^*)} = \frac{\operatorname{odds} p \times \operatorname{odds} p^*}{1 + \operatorname{odds} p \times \operatorname{odds} p^*} \ ,$$

where $\operatorname{odds} p = \frac{p}{1-p}$. The proportion $1/2$ is the neutral element of the group $(\mathcal{P}, \oplus)$, the *inverse* of p in $(\mathcal{P}, \oplus)$ is

$$\frac{\operatorname{odds} p}{1 + \operatorname{odds} p},$$

and the *compositional difference* between $p, p^* \in (0, 1)$ will be given by

$$p \ominus p^* = \frac{\frac{p}{p^*}}{\frac{p}{p^*} + \frac{1-p}{1-p^*}} = \frac{\frac{\operatorname{odds} p}{\operatorname{odds} p^*}}{1 + \frac{\operatorname{odds} p}{\operatorname{odds} p^*}} \ .$$

The *power transformation* of $p \in (0, 1)$ and $\alpha \in \mathbb{R}$ will be defined by

$$\alpha \odot p = \frac{p^\alpha}{p^\alpha + (1-p)^\alpha} = \frac{(\operatorname{odds} p)^\alpha}{1 + (\operatorname{odds} p)^\alpha} \ .$$

In this manner, $(\mathcal{P}, \oplus, \odot)$ becomes a one-dimensional real vector space. It is important to note that the algebraic structure of $\mathcal{P}$ is based on the odds

of the proportions and thus not only takes into account of p but also its complement $1-p$.
The *additive logratio* transformation (alr) on $\mathcal{S}^2$ corresponds to the logit transformation on $\mathcal{P}$, and the *centered* (or *symmetric*) *logratio* transformation (clr) corresponds to the $\frac{1}{2}$ logit transformation on $\mathcal{P}$. As they are linear transformations from the vector space $(\mathcal{P}, \oplus, \odot)$ to $\mathbb{R}$ it holds that

$$\text{logit}\,((\alpha \odot p) \;\oplus\; (\alpha^* \odot p^*)) = \alpha \,\text{logit}\, p \;+\; \alpha^* \,\text{logit}\, p^*,$$

$$\text{logit}^{-1}\,(\alpha\, y \;+\; \alpha^*\, y^*) = (\alpha \odot \text{logit}^{-1} y) \;\oplus\; (\alpha^* \odot \text{logit}^{-1} y^*),$$

for any $p, p^* \in \mathcal{P}$, and any $\alpha, \alpha^*, y, y^* \in \mathbb{R}$. Recall that the inverse of the logistic transformation can be expressed as $\text{logit}^{-1} y = \exp y / (1 + \exp y)$. It also holds that

$$\text{odds}\,((\alpha \odot p) \;\oplus\; (\alpha^* \odot p^*)) = (\text{odds}\, p)^{\alpha} \times (\text{odds}\, p^*)^{\alpha *}.$$

The $\mathcal{C}$-norm of a proportion $p \in (0,1)$ is given by

$$\|p\|_{\mathcal{C}} = \frac{1}{\sqrt{2}} \left|\text{logit}\, p\right|,$$

and the $\mathcal{C}$-distance between two proportions p and p^* in $(0,1)$ by

$$d_{\mathcal{C}}(p, p^*) = \|p \ominus p^*\|_{\mathcal{C}} = \frac{1}{\sqrt{2}} \left|\text{logit}\, p - \text{logit}\, p^*\right|.$$

The $\mathcal{C}$-norm converts the vector space $(\mathcal{P}, \oplus, \odot)$ into a metric space, and the $\frac{1}{2}$ logit transformation can be viewed as an isometry between $\mathcal{P}$ and $\mathbb{R}$.

2.2 Compositional random continuous proportions

If p is a random continuous proportion in $(0,1)$, the *compositional* expected value ($\mathcal{C}$-mean) of p will be given by

$$\text{E}_{\mathcal{C}}\{p\} = \text{logit}^{-1}\left(\text{E}\left\{\text{logit}\, p\right\}\right).$$

In agreement with the concept of variance of a random variable and the $\mathcal{C}$-distance between two proportions, the *compositional* variance ($\mathcal{C}$-variance) of p will be defined as

$$\text{var}_{\mathcal{C}}\{p\} = \text{E}\left\{d^2_{\mathcal{C}}\left(p, \text{E}_{\mathcal{C}}\{p\}\right)\right\} = \text{E}\left\{\frac{1}{2}\left(\text{logit}\, p - \text{logit}\, \text{E}_{\mathcal{C}}\{p\}\right)^2\right\},$$

and, therefore, $\text{var}_{\mathcal{C}}\{p\} = \frac{1}{2}\text{var}\left\{\text{logit}\, p\right\}$.
Similarly, if (p, p^*) is a bivariate random proportion defined in $(0,1) \times (0,1)$, the *compositional* covariance ($\mathcal{C}$-covariance) and the *compositional* correlation ($\mathcal{C}$-correlation) of p and p^* will be defined as

$$\text{cov}_{\mathcal{C}}\{p, p^*\} = \tfrac{1}{2}\text{cov}\left\{\text{logit}\, p, \text{logit}\, p^*\right\},$$

$$\text{corr}_{\mathcal{C}}\{p, p^*\} = \text{corr}\left\{\text{logit}\, p, \text{logit}\, p^*\right\}.$$

The $\mathcal{C}$-mean and $\mathcal{C}$-variance of a random proportion p in $(0,1)$ are compatible with the algebraic structure of $(\mathcal{P}, \oplus, \odot)$ by which it holds that:

(i) $\mathrm{E}_{\mathcal{C}}\{p \oplus p^*\} = \mathrm{E}_{\mathcal{C}}\{p\} \oplus \mathrm{E}_{\mathcal{C}}\{p^*\}$;

(ii) $\mathrm{E}_{\mathcal{C}}\{\alpha \odot p\} = \alpha \odot \mathrm{E}_{\mathcal{C}}\{p\}$;

(iii) $\mathrm{var}_{\mathcal{C}}\{p \oplus p^*\} = \mathrm{var}_{\mathcal{C}}\{p\} + \mathrm{var}_{\mathcal{C}}\{p^*\} + 2\,\mathrm{cov}_{\mathcal{C}}\{p, p^*\}$;

(iv) $\mathrm{var}_{\mathcal{C}}\{\alpha \odot p\} = \alpha\,\mathrm{var}_{\mathcal{C}}\{p\}$,

for any $p, p^* \in \mathcal{P}$ and any $\alpha \in \mathbb{R}$.
Finally, the *compositional* normality of a random continuous proportion, p, will be associated with the normality of $\mathrm{logit}\, p$. Therefore, we will say that p is $\mathcal{C}$-normally distributed if $\mathrm{logit}\, p$ is normally distributed.
It would thus appear obvious that the compositional structure of the random continuous proportion p is based on that of the transformed proportion $\mathrm{logit}\, p$ and that the latter is compatible with the algebraic structure of $\mathcal{P}$ defined by the operators $\oplus$ and $\odot$.

3 Compositional approach to time series of proportions

From a *compositional* point of view, the time series analysis of continuous proportions p_t is based on the standard analysis of the series $\mathrm{logit}\, p_t$ and the fact that the algebraic operators on $\mathcal{P}$ that are compatible with this compositional approach are the perturbation operator $\oplus$ and the power transformation operator $\odot$, instead of the sum and multiplication by a scalar within in $\mathbb{R}$.

3.1 Some definitions

Let $p_t,\ t = 0, \pm 1, \pm 2, \ldots$ be a random process of continuous proportions in $(0,1)$. According to the compositional approach we define the $\mathcal{C}$-mean and the $\mathcal{C}$-variance of the process at time t as

$$\tilde{\mu}_t = \mathrm{E}_{\mathcal{C}}\{p_t\} = \mathrm{logit}^{-1}\left(\mathrm{E}\{\mathrm{logit}\, p_t\}\right); \quad \tilde{\sigma}_t^2 = \mathrm{var}_{\mathcal{C}}\{p_t\} = \tfrac{1}{2}\mathrm{var}\{\mathrm{logit}\, p_t\}.$$

Similarly, the $\mathcal{C}$-covariance and $\mathcal{C}$-correlation between p_{t_1} and p_{t_2} as

$$\tilde{\gamma}_{t_1,t_2} = \mathrm{cov}_{\mathcal{C}}\{p_{t_1}, p_{t_2}\} = \tfrac{1}{2}\mathrm{cov}\{\mathrm{logit}\, p_{t_1}, \mathrm{logit}\, p_{t_2}\},$$
$$\tilde{\varrho}_{t_1,t_2} = \varrho_{\mathcal{C}}\{p_{t_1}, p_{t_2}\} = \varrho\{\mathrm{logit}\, p_{t_1}, \mathrm{logit}\, p_{t_2}\}.$$

3.2 $\mathcal{C}$-stationarity and $\mathcal{C}$-white noise

A process of proportions p_t is called (weakly) $\mathcal{C}$-stationary if the following conditions are satisfied for all values of t:

$$\mathrm{E}_{\mathcal{C}}\{p_t\} = \tilde{\mu} = \text{constant}; \qquad \mathrm{cov}_{\mathcal{C}}\{p_t, p_{t+\tau}\} = \tilde{\gamma}(\tau),\ \tau = 0, \pm 1, \pm 2, \ldots$$

From a compositional perspective, a random process of proportions p_t is considered to be *$\mathcal{C}$-white noise* if

$$\mathrm{E}_{\mathcal{C}}\{p_t\} = 1/2, \quad \mathrm{var}_{\mathcal{C}}\{p_t\} = \tilde{\sigma}^2 \quad \text{and} \quad \mathrm{cov}_{\mathcal{C}}\{p_t, p_{t+\tau}\} = 0,$$

for $t = 0, \pm 1, \pm 2, \ldots$, and $\tau = \pm 1, \pm 2, \ldots$ Equivalently $\mathrm{logit}\, p_t$ should be white noise in the usual sense of the term, with variance $2\tilde{\sigma}^2$. We use the symbol ϵ_t to denote $\mathcal{C}$-white noise and represent by $\tilde{\sigma}^2_\epsilon$ the constant $\mathcal{C}$-variance of ϵ_t. If ϵ_t is $\mathcal{C}$-normally distributed, using the well known properties of the lognormal distribution, it is easy to prove that

$$\mathrm{E}\{\mathrm{odds}\, \epsilon_t\} = \exp\left(\tilde{\sigma}^2_\epsilon\right); \quad \mathrm{var}\{\mathrm{odds}\, \epsilon_t\} = \left(\exp(2\tilde{\sigma}^2_\epsilon) - 1\right)\exp(2\tilde{\sigma}^2_\epsilon).$$

3.3 The $\mathcal{C}$-difference operator

The $\mathcal{C}$-first difference operator $\nabla_{\mathcal{C}}$ is given by

$$\nabla_{\mathcal{C}} p_t = p_t \ominus p_{t-1} = (1 - L_{\mathcal{C}})p_t,$$

where $L_{\mathcal{C}}$ is the usual backshift operator. When the operator $L_{\mathcal{C}}$ is applied to a time series of proportions in a compositional context we have to take account of the algebraic structure of $(\mathcal{P}, \oplus, \odot)$. Thus, for example,

$$(1 - 2L_{\mathcal{C}} + L^2_{\mathcal{C}})p_t = p_t \ominus (2 \odot p_{t-1}) \oplus p_{t-2}.$$

3.4 The $\mathcal{C}$-ARIMA model of proportions

A process of continuous proportions $p_t,\ t = 0, \pm 1, \pm 2, \ldots$, is a $\mathcal{C}$-ARMA(p, q) process if for every t,

$$p_t = (\phi_1 \odot p_{t-1}) \oplus \ldots \oplus (\phi_p \odot p_{t-p}) \oplus \epsilon_t \ominus (\theta_1 \odot \epsilon_{t-1}) \ominus \ldots \ominus (\theta_q \odot \epsilon_{t-q}), \quad (1)$$

where ϵ_t is $\mathcal{C}$-white noise $\mathcal{C}$-normally distributed with $\mathcal{C}$-variance $\tilde{\sigma}^2_\epsilon$. This equation can be written symbolically in the more compact form

$$\phi(L_{\mathcal{C}})(p_t) = \theta(L_{\mathcal{C}})\epsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where ϕ and θ are p^{th} and q^{th} degree polynomials in the $L_{\mathcal{C}}$ operator

$$\phi(L_{\mathcal{C}}) = 1 - \phi_1 L_{\mathcal{C}} - \ldots - \phi_p L^p_{\mathcal{C}}; \quad \theta(L_{\mathcal{C}}) = 1 - \theta_1 L_{\mathcal{C}} - \ldots - \theta_q L^q_{\mathcal{C}}.$$

Finally, a process of continuous proportions p_t is a $\mathcal{C}$-ARIMA(p,d,q) process if $(1-L_{\mathcal{C}})^d p_t$ is a $\mathcal{C}$-ARMA(p,q) process. It is clear that p_t is a $\mathcal{C}$-ARIMA(p,d,q) process if and only if $\operatorname{logit} p_t$ is a ARIMA(p,d,q) process. Therefore, in practice, the estimation of the parameters of a $\mathcal{C}$-ARIMA(p,d,q) process p_t reduces to the estimation of the parameters of the the transformed $\operatorname{logit} p_t$ process. $\mathcal{C}$-ARIMA(p,d,q) models can be represented in *logit* or *odds* formats. Thus, for example, equation (1) of a $\mathcal{C}$-ARMA(p,q) model can be expreseed in *logit* format as

$$\begin{aligned}\operatorname{logit} p_t = \;& \phi_1 \operatorname{logit} p_{t-1} + \ldots + \phi_p \operatorname{logit} p_{t-p} \\ & + \operatorname{logit} \epsilon_t - \theta_1 \operatorname{logit} \epsilon_{t-1} - \ldots - \theta_q \operatorname{logit} \epsilon_{t-q},\end{aligned}$$

where $\operatorname{logit} \epsilon_t \sim \mathrm{N}(0, 2\tilde{\sigma}_\epsilon^2)$; and in *odds* format as

$$\operatorname{odds} p_t = (\operatorname{odds} p_{t-1})^{\phi_1} \times \ldots \times (\operatorname{odds} p_{t-p})^{\phi_p} \times \omega_t \times (\omega_{t-1})^{-\theta_1} \times \ldots \times (\omega_{t-q})^{-\theta_q},$$

where ω_t is log-normally distributed, i.e., $\omega_t \sim \Lambda(0, 2\tilde{\sigma}_\epsilon^2)$.

Acknowledgments: This research has been supported by the Spanish Ministry of Science and Innovation under the projects "CODA-RSS" (Ref. MTM2009-13272) and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya (Ref. 2009SGR424).

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, New York: Chapman & Hall. Reprinted in 2003 by Blackburn Press.

Barceló-Vidal, C., Aguilar, L., and Martín-Fernández, J.A. (2007). Compositional time series: a first approach. In: *Proceedings of the 22nd International Workshop of Statistical Modelling*, Barcelona, Spain, pp. 81–86.

Box, G.E.P., and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Egozcue, J.J., and Pawlowsky-Glahn, V. (2006). Simplicial goemetry for compositional data. In: *Compositional Data Analysis: from Theory to Practice*, The Geological Society, London, UK, pp. 145-159.

Tiller, R.B. (1992). Time series modelling of sample data from the U.S. Current Population Survey. *Journal of Official Statistics*, **8**, 149–166.

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