Compositional Time Series: Past and Perspectives

Juan M.C. Larrosa

Universidad Nacional del Sur (UNS)

jlarrosa@uns.edu.ar

Resumen

Este trabajo revisa contribuciones académicas que se centran en el análisis de series dinámicas composicionales, un tema poco investigado a pesar de la amplia disponibilidad de datos en ciencias sociales. Explora las opciones disponibles y divide los artículos de investigación en dos enfoques principales de probabilidad, frecuentista y bayesiano, y enumera varias transformaciones y técnicas específicas de datos. Como conclusión, esta rama del análisis estadístico de la composición requiere una actualización profunda y, por esta misma razón, es un campo fértil para la investigación futura.

Abstract

This work reviews academic contributions that focus on the analysis of compositional dynamic series, a little investigated topic in spite of the wide data availability in social sciences. It explores the available options and divides research articles into two main probability approaches, frequentist and Bayesian, and enumerates several specific data transformations and techniques. As conclusion, this branch of the compositional statistical analysis requires a profound updating and, for this very same reason, is a fertile field for future research.

JEL Code: C22, C46

1. Introduction

Variations in compositions are difficult to observe in cases like geological processes where individuals under scrutiny (solid rocks, sand, sediments, and the like) can change their composition but only through a long period of time. However, in social Sciences processes these changes usually take shorter time and become a powerful dimension for explaining diverse social events: for instance, composition related to unemployment, portfolio investment, trade balance, and many others. When a broker manages a portfolio of assets she looks for components (assets) with different degree of risk for assembling the balanced portfolio. Once formed, valuations of components vary in time and how these valuations evolve is critical for the investor: decisions might be taken to change that composition. In another more macroeconomic scenario, the composition of the trade balance (by destination, type of products, or by another feature) would describe patterns for changing directions in international commerce. Household expenditure composition evolution might help to understand changes in preferences and needs in a target population. Many other examples describe the importance of compositions in economics.

This fact has been taken into account for non-constrained data and an enormous amount of literature has been written on Time Series Analysis (TSA), being Hamilton (1994) and Woodward et al. (2012) great contributions on the subject, but little has been said about Compositional Time Series (CTS). CTS represent multivariate time series of compositions, often characterized by a constant sum constraint representation, at each time point *t*. Thus a CT can be defined as the series of elements of the simplex S^D, the sample space of representations of compositional data to a chosen constant sum constraint. CTS are thus characterized by positive components with a constant sum at each time *t* (frequently the constant is taken as 1). This constraint forms a crucial problem when modeling compositional time series by standard multivariate time series methods. From the methodological point of view, the problem with a statistical analysis of CTS using standard methods is caused by the specific geometry of compositional data, the Aitchison geometry on the simplex, which accounts for inherent properties of compositional data.

Several approaches have been introduced to model CTS. The main strategy is based on the use of log-ratio transformations. This procedure consists of transforming CTS given in the coordinate space - the real vector space with Euclidean structure - to abandon Aitchison's geometry and, practically, to break the unit-sum constraint of the original time series. After transformations have been done, standard multivariate time series methods can be applied to transformed time series.

The goal of the paper is to describe four principal aspects on each quoted work: what transformation have been applied to raw data for avoiding *spurious* analysis? What statistical methodology has been used for analyzing transformed data? Has this methodology brought new insights into CTS analysis? And lastly, what CTS features, if any, remain unanswered? Most recent contributions such as Mills (2009a,b), Dawson et al. (2014), Kynclová et al (2015) show that by transforming data according to early suggestions as pointed out by Aitchison (1986) and Egozcue and Pawlowsky-Glahn (2006) statistical standard procedures can be successfully applied to data.

The paper follows with section 2 where we introduce compositional data and postulate initial definitions. We follow with section 3 where the approaches for CTS are divided in two subsections: one for the Bayesian approach and the second for the non-Bayesian or frequentist approach, ending with a summary. Section 4 discusses the survey and Section 5 states the conclusions.

2. The methodology of compositional data

Compositional data refers to proportions of a whole and because of it are subject to the constraint that the sum of its components is the unit or a constant. This restriction does not allow for an immediate interpretation of the covariance structure due to the presence of spurious correlation (Pearson, 1897). This has not been properly treated for long time by academic research across several disciplines. For instance, Brandt et al. (1999) describe the procedures commonly used by political scientists (among other social scientists) for avoiding this restriction: (1) ignoring the compositional nature of the data, i.e., by using independent equations for each component, (2) ignoring all but one component, i.e., any model of unemployment or political party vote share, or (3) converting a multipart composition into a two-part subcomposition and then employing (2). As they remarked all of these approaches ignore the deterministic structure of the correlation among components caused by the sum constraint; besides all approaches ignore the boundedness of the data and, finally, this subcompositional approach can mask (or create) substantively important variability in the data. Aitchison (1986) suggested a number of log-transformation in constrained data before applying any standard statistical procedure[1]. This section resumes the required concepts for understanding the findings of this work. It begins by defining what compositional data is.

Definition 1. *Compositional data* $x = (x_1, x_2, \ldots, x_D)'$ *with D parts, is a vector with strictly positive components, so the sum of all of the components equal a constant k. The sampling space is the simplex defined as* $S^D = \{(x_1, x_2, \ldots, x_D)' : x_j > 0; j = 1, 2, \ldots, D; x_1 + x_2 + \cdots + x_D = k\}$.

[1] Pawlowsky-Glahn, V. and A. Buccianti (2011) represents an enormous introduction to the topic of compositional data analysis.

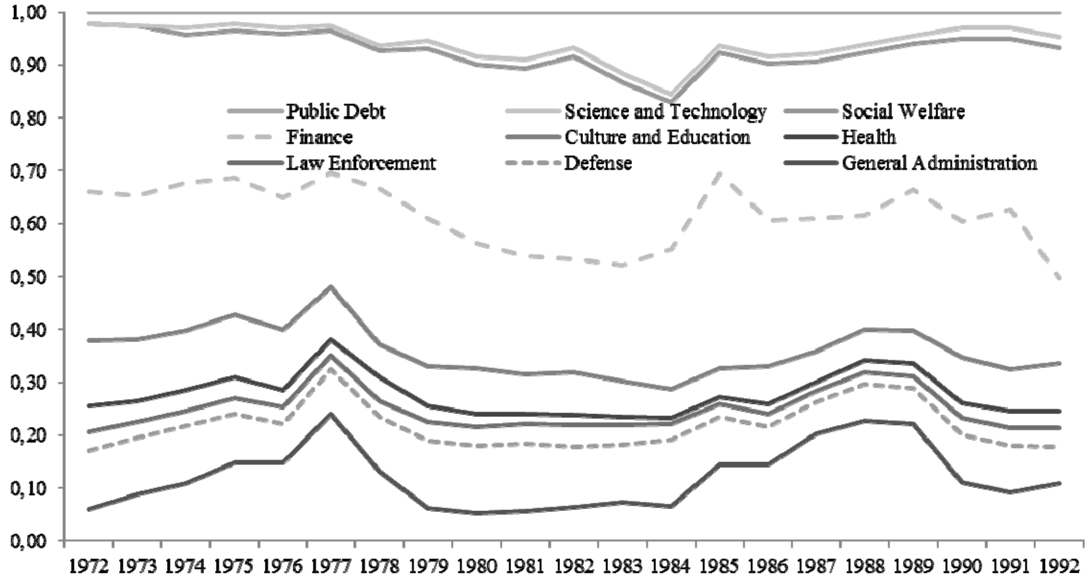

Figure 1. Nine chapter composition in the National Budget of Argentina (1972-1992)

Source: The author from data from Ministry of Economics (Argentina)

Figure 1 presents raw data from the composition of the National Budget of Argentina. We can always obtain compositional data on S^D if we have an initial vector of nonnegative components. We only require dividing each component by the sum of all components. Following, we define the first transformation:

Definition 2. *The additive logratio transformation (alr) of index j* $\left(j = 1, \ldots, D \right)$ *is the one-to-one transformation from* S^D *to* $\mathbb{R}^{D-1}$ *defined as*

$$alr_j x = \ln\left(\frac{x_{-j}}{x_j}\right),$$

where x_{-j} *denotes de vector x with the component* x_j *deleted.*

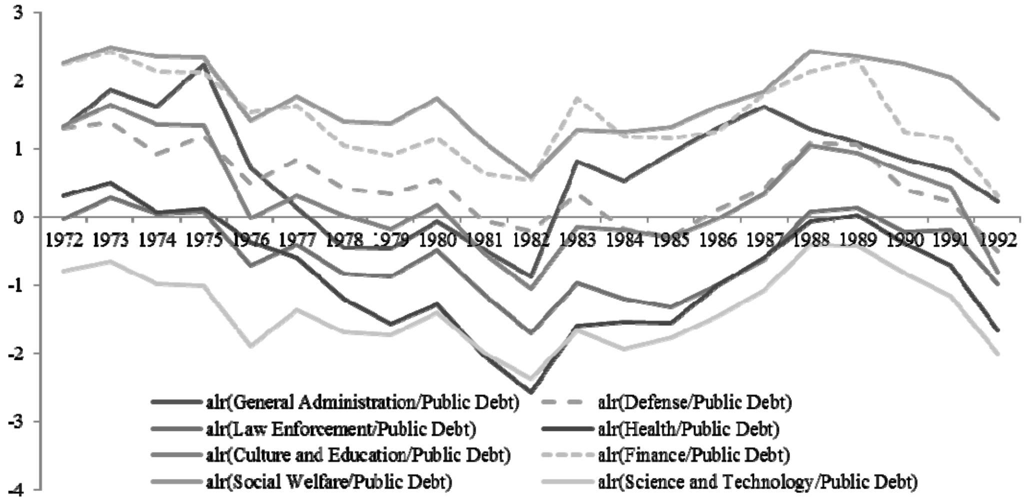

Figure 2. Data ALR-transformed from Figure 1. Public Debt chapter is used as fill-value.

Source: The author from data from Ministry of Economics (Argentina)

Figure 2 presents the data exposed in Figure 1 transformed by the *alr*-transformation by using Public Debt component as a fill-value. This is an asymmetric transformation. A case for symmetric transformation is the following one:

Definition 3. *The centered logratio transformation (clr) is a bijective application between* $x \in S^D$ *to* $c \in \mathbb{R}^D$ *defined by*

$$clr\left(x\right) = \ln\frac{x_j}{g\left(x\right)} = c_j,$$

with $g\left(x\right) = \left(\prod_{j=1}^{D} x_j\right)^{1/D}$ *as the geometric mean of the composition.*

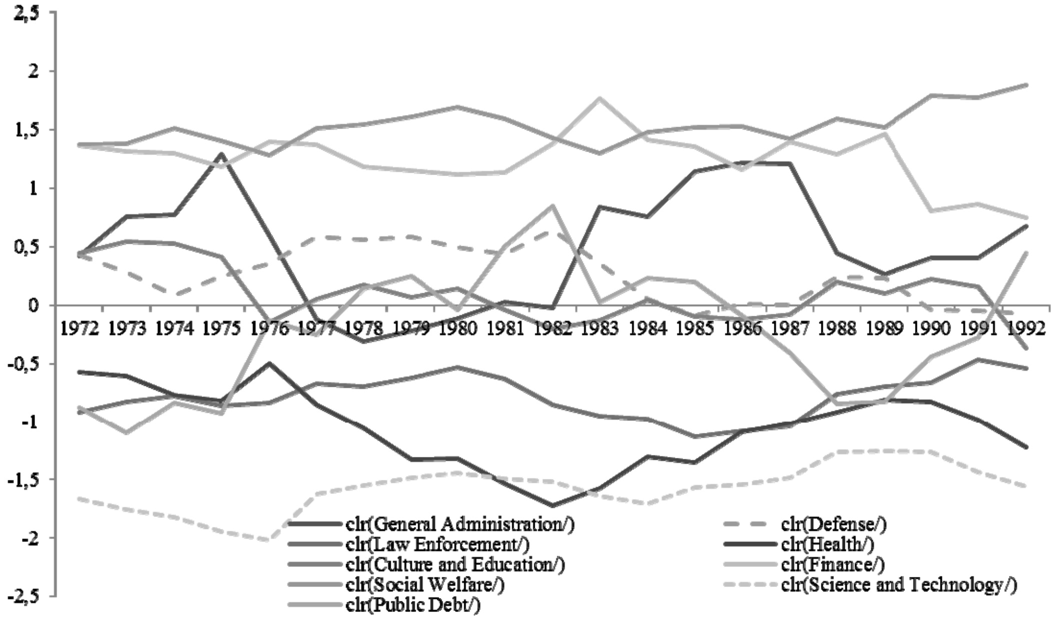

Figure 3. Data transformed by CLR-transformation.

Source: The author from data from Ministry of Economics (Argentina)

We must note that *clr*-transformed data is constrained by a zero-sum constraint. Figure 3 shows data from Figure 1 transformed now by the clr-transformation. Finally, we use the following compositional time series definition:

Definition 4. *Let* $x_t = \left(x_{t1}, \dots, x_{tD} \right)', t = 0, \pm 1, \pm 2, \dots$ *be a compositional time series process (CTS process) defined on* S^D *for any t.*

A given CTS process $\left\{ x_t \right\}$ could be analyzed as a multivariate time series or sub-compositional process (by isolating two or more parts). The *alr* and *clr* transformations can be applied to any CTS process. We will call $\left\{ c_t \right\}$ to a *clr*-transformed CTS process defined on $\mathbb{R}^{D}$ then it is restricted to the hyperplane *V* because $z_t' 1_D = 0$. And $\left\{ a_t \right\}$ will be the notation for an *alr*-transformed CTS process defined on $\mathbb{R}^{D-1}$ that depends on the denominator used in the *alr*-transformation.

3. Approaches to CTS Analysis

While non-Bayesian approaches may be considered the mainstream for non-constrained time series analysis, one could argue that the opposite is the case for CTS analysis. Many of the works that will be recapitulated have been designed under the spirit of the Bayes theorem and this fact in some cases requires a quick review of these methods. Chronologically, earlier papers worked with transformed data on the log-normal distribution and latter papers introduce the Dirichlet distributions as assumption in parameters behavior. Original notations are homogenized for the sake of clarity. We begin next section describing Bayesian CTS models.

3.1 Bayesian Methods in CTS

Bayesian techniques require that researchers explicit their expectations on the distribution that actual data under analysis have (see Broemeling and Shaarawy (1986); Koop (2003: Ch. 8); Poirier and Tobias (2005); and especially Zellner (1984: part 3), for extensive references on Bayesian inference). A considerable part of the contributions is based on some form of state space models, which are hierarchical in nature. When applied to compositional time series this approach requires the definition of prior information on time series evolution. For example, Grunwald (1987) works with compositional time series by using state space modelling for non-Gaussian time series. He opts for the *clr* transformation for dealing with the constant-sum constraint. Then he applies a state space model by specifying initial observations and state distributions "which describe either diffuse or well-defined initial beliefs" (Grunwald, 1987: 16) for time series forecasting. This process is recursively done by the Kalman filter implemented on the filtering stage.

For those that are unrelated with state space models[2], it can briefly state that a time series $y_1, y_2, \dots$ could be thought as a (steady) model

$$\left(y_t \mid \theta_t, \tau_t\right) \sim \mathrm{Dir}\left(\tau_t \theta_t\right), \tag{2.1}$$

where, in the case of continuous proportions, they assume y_t follows a Dirichlet distribution. This is called the observation equation, that evolve conditional to a state θ_t with spread τ_t[3]. The state $\{\theta_t\}$ is assumed to evolve over time according to the steady state model, namely

$$p\left(\theta_{t+1} \mid \mathrm{D}_t\right) \propto \left\{p\left(\theta_t \mid \mathrm{D}_t\right)\right\}^{\gamma} \quad \text{with } 0 < \gamma < 1 \tag{2.2}$$

[2] For a general theory of state space models applied to time series analysis see Harvey (1989).

[3] τ_t is deliberated introduced by the author to cope with a forecasting problem. τ_t is updated separately from θ_t. See Grunwald (1987: Ch. 4) and Grunwald *et al.* (1993: 108-109) for details.

There, D$_t$ is defined recursively by $\mathrm{D}_t = \{I_t, \mathrm{D}_{t-1}\}$ where, for $t \geq 1$, $I_t = \{y_t,$ all other relevant information available at time *t* but not at *t* – 1} and D$_0$ are the externally determined estimated parameters and all available relevant information at *t* = 0.

Dirichlet distribution in (2.1) has the following form:

$$f(p) = \frac{\Gamma(\tau)}{\prod_{j=1}^{d+1}\Gamma(\beta_j)} p_j^{\beta_1 - 1} \ldots p_{d+1}^{\beta_{d+1}-1} \text{ with } \tau = \beta_1 + \ldots + \beta_{d+1} \qquad (2.3)$$

with sample space S^d and parameter space $\{(\beta_1, \ldots, \beta_{d+1}) : \beta_j > 0 \text{ for } j = 1, \ldots, d+1\}$.

As for any state space model, it must be defined:

(i) the assumptions underlying the state behavior,

(ii) the description of the (recursive) filtering process, as stated by Grumwald is described by:

Observation Distribution $$f(y_{t+1} \mid \theta_{t+1}) \qquad (2.4)$$

State Forecast Distribution $$f\left(\theta_{t+1} \mid y^{(t)}\right) = \int f\left(\theta_{t+1} \mid y^{(t)}\right) f\left(\theta_t \mid y^{(t)}\right) d\theta_t \qquad (2.5)$$

State Posterior Distribution $$f\left(\theta_{t+1} \mid y^{(t)}\right) = \frac{\int f(y_{t+1} \mid \theta_{t+1}) f\left(\theta_{t+1} \mid y^{(t)}\right)}{f\left(y_{t+1} \mid y^{(t)}\right)} \qquad (2.6)$$

(Note that the state posterior distribution is described by Bayes theorem.),

(iii) the forecasting stage (described in the denominator of the state distribution posterior), and

(iv) the smoothing stage (again, derived from state distribution posterior for $t \leq n$).

Finally, a crucial item is the likelihood function that can be used for estimating parameters outside the internal updating procedure. This function is usually maximized through numerical methods. It is assumed that the observation distribution $f_\phi(y_t \mid \theta_t)$ and the state forecasting mechanism $f_\phi(\theta_{t+1} \mid \theta_t)$ are known in form but they depend on an unknown parameter ϕ. The log-likelihood for ϕ is

$$L(\phi) = \sum_{i=1}^{t} \log f_\phi\left(y_{t+1} \mid y^{(t)}\right) \qquad (2.7)$$

Finally, Grunwald uses US Federal Government data (on tax revenues and external trade) for testing his model and applies the Dirichlet distribution in the updating and forecasting procedures obtaining acceptable good fitting and forecasted values.

Another case is Quintana and West (1988) who work with Mexican import time series by using Aitchison's additive log transformation (*alr*). They model series as a class of dynamic multivariate regression (DMR), closely related to state space modelling. This technique allows for modeling multiple variate time series by using a basic structure that assume the existence of an observation equation (observed values), evolution equation (state equation) and prior information (assumptions on state equation probability distribution). In a similar fashion but using a matrix notation they present the following model:

$$\text{Observation Equation} \quad y_t' = x_t'\Theta_t + e_t' \qquad e_t \sim N\left(0, v_t\Sigma\right) \qquad (2.8)$$

$$\text{Evolution Equation} \quad \Theta_t = G_t\Theta_{t-1} + F_t \qquad F_t \sim N\left(0, W_t, \Sigma\right) \qquad (2.9)$$

$$\text{Prior Information} \quad \left(\Theta_t \mid \Sigma\right) \sim N\left(M_{t-1}, C_{t-1}, \Sigma\right) \quad \Sigma \sim W^{-1}\left(S_{t-1}, d_{t-1}\right) \qquad (2.10)$$

In the above equations, y_t is a $\left(q \times 1\right)$ vector of observations made at time *t*, x_t is a $\left(p \times 1\right)$ vector of independent variables, Θ_t is an unknown $\left(p \times q\right)$ matrix of system (regression) parameters, e_t is a $\left(q \times 1\right)$ observation error vector, v_t is a scalar variance associated with e_t and Σ is an unknown $\left(q \times q\right)$ system scale variance matrix providing cross sectional correlation structure for the components of y_t. *N* (*M*, *C*, Σ) and W^{-1} (*S*, *d*) denote the general matrix normal and inverted-Wishart distributions (this are derived in the appendix of the original paper).

The nature of the model component series can be seen as follows. For $j = 1, \ldots, q$ let y_{ti} be the observation on the *j*th series, simply the *j*th element of y_t; e_{tj} the corresponding element of e_t; θ_t the *j*th column of Θ_t; f_{tj} the *j*th column of F_t; m_{tj} the *j*th column of M_t; and σ_t^2 the *j*th diagonal element of Σ. Then, y_{tj} marginally follows the DRM:

$$\text{Observation Equation} \quad y_{tj} = x_t'\theta_{tj} + e_{tj} \qquad e_{tj} \sim N\left(0, v_t\sigma_j^2\right) \qquad (2.11)$$

$$\text{Evolution Equation} \quad \theta_{tj} = G_t\theta_{t-1} + f_{tj} \qquad f_{tj} \sim N\left(0, W_t\sigma_j^2\right) \qquad (2.12)$$

$$\text{Prior Information} \quad \left(\theta_{t-1,j} \mid \Sigma\right) \sim N\left(m_{t-1,j}, C_t\sigma_j^2\right) \quad \Sigma \sim W^{-1}\left(S_{t-1}, d_{t-1}\right) \qquad (2.13)$$

The joint structure comes in via the covariance, conditional upon Σ:

$$Cov\left(e_{ti}, e_{tj}\right) = v_t\sigma_{ij}, \qquad (2.14)$$

$$Cov\left(f_{ti}, f_{tj}\right) = W_t\sigma_{ij}, \qquad (2.15)$$

$$Cov\left(\theta_{ti}, \theta_{tj}\right) = C_{t-1}\sigma_{ij}, \qquad (2.16)$$

for $i \neq j$, where σ_{ij} is the *ij* off-diagonal element of Σ.

They use CTS process $y_t = \left(y_{t1}, \ldots, y_{tq}\right)$, $t = 1, 2, \ldots$, multivariate time series such that $y_{ti} > 0$ for all *i* and *t*. They are concerned only in the proportions $p_t = \left(\sum_{i=1}^{q} y_{ti}\right)^{-1} y_t$. Later they apply the *clr* transformation as in Definition 3:

$$c_{tj} = \log\left(\frac{p_{tj}}{g\left(p_{tj}\right)}\right) = \log p_{tj} - \log g\left(p_{tj}\right),\ j = 1, \ldots, q, \qquad (2.17)$$

where $g(p_{tj})$ is the geometric mean of the p_{ij}. Modeling c_t with the DRM previously introduced derives in a conditional multivariate normal structure. Thus the observational distribution of the proportions p_{ti} is the multivariate logistic-normal distribution as defined in Aitchison and Shen (1980).

A difference between state space models and DRM approach is that DRM include *discount factors* to adapt W_t to subjective or exogenously given interventions. Thus, for a given discount factor δ, such that $0 < \delta \leq 1$, we have that:

$$W_t = \left(\delta^{-1} - 1\right) G_t C_{t-1} G_t' \tag{2.18}$$

When δ = 1, W_t = 0 and then Θ_t will evolve purely deterministic (also called static model), but for smaller values δ they can model greater variation in Θ_t. This is used, for example, for taking into account shocks or trends that could modify W_t evolution. Notice that state space modelling approach simply add covariates (for instance, dummies that represents such shocks or trends) explicitly and then their statistical significance can be measured.

Following Quintana and West (1988), they also notice the first *complication* on the transformed data. As we suppose that y_t in (2.1) follows (2.11) then it emerges the singularity of the model due to the zero-sum constraint, where $y_t'1 = 0$, for all t, where $1' = (1, \ldots, 1)$. This follows from the definition and leads to singularity of the matrices Σ, V_t, V_t^*, etc. The way they deal with this problem is by transforming y_t using $y_t'K$ where:

$$K = I - q^{-1}11',\ 1 = [1, \ldots, 1]' \tag{2.19}$$

Now it has to retransform (2.11) by including (2.19), so we get:

Observation Equation $\quad y_t'K = x_t'\Psi_t + \left(Ke_t\right)', \qquad Ke_t \sim N\left(0, v_t \Xi\right), \quad$ (2.20)

Evolution Equation $\quad \Psi_t = G_t \Psi_{t-1} + F_t K, \qquad F_t K \sim N\left(0, W_t, \Xi\right) \quad$ (2.21)

Prior Information $\quad \Psi_{t-1} \sim N\left(M_{t-1}K, C_{t-1}, \Xi\right), \quad \Xi \sim W^{-1}\left(K' S_{t-1} K, d_{t-1}\right) \quad$ (2.22)

Where $\Psi_t = \Theta_t K$ and $\Xi = K'\Sigma K$. By these linear transformations, quantities x_t' , v_t, G_t, W_t, and C_t remain unaffected by the transformation. This way, the constrained data follows now a DMR. Quintana and West end the paper with an application to Mexican import composition with very good results.

Grunwald et al. (1993) review Grunwald (1987)'s thesis. They specify symmetric logratios (*clr*) as transformation for the raw data and delineate more concisely the Dirichlet state space modeling approach. They describe it as based on the idea that dynamic proportions are constructed of an unobserved random walk component and a noise component. Then they apply their stylized model on world car production composition forecasting.

The next paper that uses Bayesian approach is Cargnoni et al. (1997). They use as the motivating case of study the forecasting of the number of high school students in Italy. They divide students as (i) students that repeat the same grade in consecutive years, (ii) students that proceed to the following grade and do not leave the school, and (iii) students that leave the school. They don't clearly specify the transformation to apply, but they put in the options of transformation those of Aitchison (1986)'s. As previous investigations, they rely on a kind of state space time series model. By assuming that there exists cross-sectional conditional independence of the series (independence among individuals) they derive a class of conditionally Gaussian dynamic models, a bit more complex than Quintana and West's.

In another more complex approach, Ravishanker et al. (2004) study the relationship between air pollution and mortality proportions in the Los Angeles area by using a Hierarchical Bayesian modeling framework. They first transform raw data by the additive logratio (alr) transformation. Then they use linear regression with vector autoregressive moving average (VARMA) errors. Inference is derived from Bayesian framework using Markov chain Monte Carlo algorithm in order to simultaneously generate samples from the posterior distributions of the parameters.
The framework can be briefly described as follows: Let y_t denote a *g*-dimensional composition at time *t*; i.e. a vector of quantities Y_{tj}, *j* = 1, …, *G* such that $\sum_{j=1}^{G} Y_{tj} = 1$, *t* = 1, …., *T*. Let a_t denote the vector resulting from the *alr* transformation of y_t, i.e.,

$$A_{tj} = \ln\left(\frac{Y_{tj}}{Y_{tG}}\right), \text{ with } j = 1,\ldots,g\ ,\ t = 1,\ldots,T \tag{2.23}$$

Let z_t be a *t*-dimensional vector of covariates at time *t*. A normal linear regression model with VARMA errors for the *g*-dimensional vector time series a_t is given by:

$$a_t = \gamma + \eta' z_t + w_t, \tag{2.24}$$

$$\Phi(B) w_t = \Theta(B) h_t \tag{2.25}$$

where *γ* is a *g*-dimensional intercept term, η is a $t \times g$ matrix of regression coefficients, w_t denotes the *g*-dimensional vector of regression errors, h_t are *g*-variate iid *N* (0,Σ) variates with unknown positive definite covariance matrix Σ. It is assumed that w_t = ($W_{1,t}$, …, $W_{g,t}$) are generated by a zero mean VARMA (*p*, *q*) process. Once this model is estimated arises the problem that solution may be non-unique, so the authors apply a Bayesian selection mechanism among best solution candidates. So, they maximize a Gaussian likelihood function, then they specify a prior density function and, using Bayes theorem, they also specify the posterior density. As this last posterior density is analytically intractable they must rely on numerical simulations. They use Monte Carlo simulations for the expected composition proportions based on the samples from the (simulated) posterior density function.

All of this enormous and complex simulated process makes difficult for direct interpretation of the steps of the estimation procedure. As final result, they obtain twelve possible models from where they choose, by selecting those with lower Bayesian Information Criterion (BIC).

Finally, Bhaumik et al. (2003) combine *alr* and Box-Cox transformations for dealing with the same data previously used by Grunwald (1993) and Ravishanker et al. (2004). These compositions use the following transformation:

$$BCA_{tj} = \begin{cases} \dfrac{\left(\frac{Y_{tj}}{Y_{tG}}\right)^{\lambda_{tj}-1}}{\lambda_{tj}} & \text{if } \lambda_{tj} \neq 0 \\ \ln\left(\frac{Y_{tj}}{Y_{tG}}\right) & \text{if } \lambda_{tj} = 0 \end{cases} \text{, with } j = 1,\ldots,g \text{ , } t = 1,\ldots,T \qquad (2.26)$$

where $\lambda_{tj} \in \mathbb{R}$ is an unknown parameter named as the Box-Cox parameter and $g = G-1$. The transformation is defined across the paper as a $BCA_t = BC(Y_t, \lambda_t)$, a special case that it gets *alr* transformation when $\lambda_t = 0$, for all i,t (as suggested by Aitchison, 1986).

The dynamical linear model is defined as $BCA_t \middle| \lambda_t = \gamma_t + Y_t\beta_t + e_t$, where α_t and β_t are g-dimensional vector of unknown parameters and e_t is the random error. By using the scale mixture of multivariate normal (SM-MVN) error distribution they develop a complex procedure for estimating and selecting alternative models as in Ravishanker et al. (2004).

3.2 Frequentist Approach to CTS

While Bayesian approaches rely on the researcher's specifications about the a priori data distribution and then update the estimation with the observed values, non-Bayesian or frequentist procedures, as linear regressions, assume some known (usually Gaussian) probability distribution of the stochastic part of the model.

Although the data constitute a multivariate time series, ARIMA techniques based on multivariate autoregressive integrated moving average are usable thanks to Aitchison’s transformations. Brundson (1987), Brundson and Smith (1988) and Smith and Brundson (1989) use the additive logistic transformation for modelling time series as autoregressive processes. On the second paper, they review main Aitchison’s findings on compositional data and adapt them into a time series framework. Finally, they try to test subcompositional independence on time series by applying their methodology to UK vote-intention’s poll time series data, in the first and third paper, and try to forecast unemployment rate in Australian labor force in the second one.

So, they transform data by applying *alr* transformation as in Definition 2:

$$a_m\left(p_i\right)=\log\left[\frac{p_i}{p_{m+1}}\right],\ \left(i=1,\ldots,m\right) \qquad (2.26)$$

where $p_{m+1}=1-\sum_{i=1}^{m}p_i$. Brunsdon (1987) and Smith and Brundson (1989) are first attempts to test whether subcompositions in CTS data might be independently studied. Once a positive answer emerges, they define a Granger causality test (Granger 1969) with data from UK Gallup poll test. They verify independence between vote intentions on main political and other kinds of responses in political survey questions but that there was no independence within vote intentions on main political parties.
Later, in Brundson and Smith (1988), they apply Box-Jenkins methodology directly to *alr*-transformed data. This is by far the most common technique taught in time series courses. The goal is to predict labor force components in the Australia. They model transformed data as a VARMA process and helped with autocorrelograms and partial autocorrelograms they identify the order of the time series. Forecasted proportions were reasonably close to actual data.

A number of recent contributions emphasize that the approach to compositional time series need not be remote from many traditional time series techniques. Thus, Mills (2009a, b) represent two interesting contributions in this regard. Mills (2009a) focuses on predicting trends in obesity in the UK. For this it transforms data of the population percentage with overweight with the log-quotient additive alr and then applies an ARIMA prediction process. The predictive advantage is limited and the exercise of application is fruitful in predicting values significantly. Mills (2009b) broadens the scope of application to other sectors by repeating obesity figures and adding percentages of national income and age ratios for cricket players. For this the car extensively uses an *alr* transformation.
The prediction of time series has also been the point of analysis of Koehler et al. (2010) who model compositional time series with alr transformations. The specific case of analysis is the composition of adjustable rate loans. However, the authors add an exponential smoothing vector model that improves prediction indicators. Bergman and Holmquist (2014) more recently uses *clr*-transformed election polls data from Sweden. The smooth compositional data provided by different consultants by using compositional weighted least squares estimations and provided a clearer picture of the trend in voting preferences. van der Braker and Roels (2010) deals with discontinuities in sample variables. The contribution uses centered and additive logratio transformations on data for estimating and simulating data in series that was redesigned potentially harming the temporal comparability in data.

Finally, a more complex analysis is carry out in Brandt et al. (1999). They implement a vector autoregression (VAR) representation for dealing with compositional time series. The VAR was originally proposed by Sims (1980) for non-constrained data. They try to elucidate how the evolution of economic and political indexes affects vote intentions in the USA. As VAR models assume that we can best explain the current values of the endogenous variables (both compositions and non-compositions) using a sequence of predetermined past values. Formally, they write a system of compositions in reduced form for each observation as:

$$Y_t=\gamma Z_t+\sum_{j=1}^{P}\beta_j Y_{t-j}+\varepsilon_t \qquad (2.27)$$

where Y_t is an $M \times 1 = (Q + S) \times 1$ vector. Z_t is a matrix of exogenous variables (including an intercept) and Y_{t-j} is the j^{th} lag of Y_t. If we assume that the $M \times 1$ error term $\varepsilon_t \sim N(0,\Sigma)$ then we have a time series model for the symmetric (clr-transformed) log-ratios of the components. Assuming that the series Y_t are multivariate log-normal is a sufficient condition for the proportions to have a logistic-normal distribution (Aitchison 1986, Quintana and West 1988). They called this system a Compositional VAR or CVAR.

As noted by Quintana and West (1988), there is singularity into this VAR model due to the zero-sum constraint of the transformed values of the dependent variables. A traditional solution implemented in economic literature has been to drop one of the variables (usually the last variable) as Theil (1971: 326-356) suggested. So, they adopt Quintana and West proposal and create a matrix K defined as:

$$K = I - \frac{1}{q}hh', \tag{2.28}$$

where, again, q is the number of components, I is a $q \times q$ identity matrix and h is a $q \times 1$ vector of ones. The matrix performs an elementary row operation that maps the logarithms of the proportions to the symmetric logratio space. By using K they impose a constraint in the VAR system represented by (2.27) which is modified by (2.28) in the following way:

$$KY_t = \delta Z_t + \sum_{j=1}^{p} \theta_j Y_{t-j} + K\varepsilon_t \tag{2.29}$$

where $\delta = K\gamma$, and $\theta_j = K\beta_j$. This way, as in Quintana and West (1988), the transformation leaves the lagged and exogenous right-hand side variables unaffected. Kynčlová et al. (2015) presents VAR analysis applied to raw compositional data and explains the misleading results and then transforms the data and reapply the procedures and compares to former analysis.

Final estimation requires the usual procedure for VAR estimation (i.e., to estimate the q equations one by one or the system simultaneously), and then used a numerically extensive work for compute bootstrap samples and Monte Carlo integration for computing the moments of the posterior distribution. They apply the model to estimate the incidence of socioeconomic and political variables to voters' partisanship in the USA.

3.3 Summary

The Table 1 summarizes the previous reviews. There it can be noticed the respective paper reference, the transformation applied to raw data, the statistical technique, specific comments of the reviewer (if any), and the authors application field. As observed, *alr* and *clr* transformations were both applied in the different papers, the predominant statistical method is (variations of) state space model and most of the cases of study are from the social sciences area.

Table 1. Summary of papers

Author/s	Transformation Applied on Raw Data	Statistical Technique	Comments	Applied Case of Study
Brunsdon (1986)	Additive logratio	Log-Normal based autoregressive integrate moving average (ARIMA) model.		UK poll data on vote intentions
Grunwald (1987)	Centered logratio	Dirichlet conjugate state space model.	Several other time series approaches are presented.	Tax revenues compositions and world car production composition.
Brunsdon and Smith (1988)	Additive logratio	Log-Normal based vector autoregressive moving average (VARMA) model	They use (a more traditional) Box-Jenkins methodology.	Forecasting of Australian labor force composition
Quintana and West (1988)	Centered logratio	Log-normal state space model (Dynamic linear model)	They must introduce transformations on the regressand for avoiding singularity emergence on the variance and covariance matrix.	Mexican imports and exports composition
Smith and Brunsdon (1989)	Additive logratio	Log-Normal based ARIMA model.		UK poll data on vote intentions
Grunwald, Raftery, and Guttorp (1993)	Centered logratio	Dirichlet conjugate state space model		World car production composition.
Cargnoni, Müller, and West (1997)	Logratio (not declared explicitly)	Conditionally Gaussian dynamic model		Forecasting of number and composition of secondary school

				students in Italy.
Brandt, Monroe, and Williams (1999)	Centered logratio	Compositional Vector Autoregression (CVAR) system.	They deal with the same problem that Quintana and West (1988) and introduce analogous transformations on regressands.	Socioeconomic and political determinants of Partisanship composition.
Bhaumik, Dey and Ravishanker (2003)	Box-Cox with Alr-transformations	Linear regression with hierarchical priors		Use the same data as Ravishanker et al (2004) and Grunwald et al (1993)
Ravishanker, Dey, and Iyengar (2004)	Additive logistic ratios	Linear regression with (VARMA) errors and Hierarchical Bayesian selection model.		Los Angeles mortality composition.
Mills (2009a)	Additive logratio	ARMA model base on Log-Normal		Bugdet composition, obesity trends, cricket results in UK
Mills (2009b)	Additive logratio	ARMA based on Log-Normal		Obesity trends in UK
Koehler et al. (2010)	Additive logratio	Vector of Exponential smoothing Model		Adjusted loan rates, election win chances
van der Braker and Roels (2010)	Centered logratio	Seemingly Unrelated Structural Time Series and Restricted Multivariate Model		Netherlands's Permanent Survey on Living Conditions (PSLC).
Bergman and Holmquist (2014)	Centered logratio	Compositional Weighted Least Squares (C-WLS)	They smooth political party preferences obtained by different sources	Poll Party Vote Preference Data in Sweden

Dawson et al. (2014)	Additive logratio	ARMA based on Log-Normal	Sentiment analysis in Taking Part Survey (TPS)
Kynčlová et al (2015)	Additive and isometric logratio	VAR Model	Paper production shares in Europa.

4. Discussion

Since their sample space is the simplex rather than the real space with the usual Euclidean geometry, they need to be expressed in appropriate (preferably orthonormal) coordinates with respect to Aitchison geometry before any statistical analysis are performed. As observed along the survey, different approaches have been used to understand the effect of time in compositions. It is proposed an exploration by distinguishing two main pathways: first, contributions that are based in the most commonly used frequentist approach to time series analysis, and secondly, a Bayesian approach to CTS. However, it is hard to state which of the two has been shown to be the most efficient way to deal with CTS. This aspect will depend, of course, of the specific requirement of a particular research. In any case, transformation on raw data will be presents given that constrained nature of data. In general terms, compositional data should not be treated in a raw scale, but only after a log-ratio transformation (Aitchison, 1986). This is so because the information inherent to a compositional data is relative, each component depends on the value of other components. The principle of working in coordinates allows applying any sort of multivariate analysis to a log-ratio transformed composition, as long as this transformation is invertible. This principle is then of full applicability to time series analysis.

One aspect should be pointed out: Economic theory seldom explains models relying on the composition of relevant variables. This way, TSA has been focus mainly on non-compositional data. But as exposed, examples of CTS are present all across the economic analysis and in relevant and present topics, such as unemployment, voting processes, portfolio composition, national accounts, and budgetary decisions, among others.

5. Conclusions

It´s was found that academic literature is scant and scattered and it seems to be no clear mainstream. Several authors freely use two of the most known Aitchison's transformations and ad-hoc statistical model and sometimes these infrequent modeling approaches seem to be the center of the investigation rather than the compositional nature of data. Throughout this brief review three main aspects have been observed: the transformations, the statistical models, and the cases of study. First, the additive and centered logratios have been equaled used in the scant literature. However, none of the papers have compared the efficiency or appropriateness of each of the transformations for the specific case of study or statistical modelling. We know that *alr* transformation is not isometric and the *clr* transformation is isometric but constrained[4]. As a good remark has to be noted that Quintana and West (1988) and Brandt et al. (1999) have dealt with the problems of *clr*-tranformation zero-sum constraint by exogenously modifying the regressands in the linear regression equation. Further studies are required, again, for the appropriateness of this ad-hoc solution.

Second, diverse statistical techniques have been summarized. Such diversity remarks the lack of a mainstream methodology for dealing with CTS. Traditional TSA has a stock of available techniques that has not been applied using transformations from compositional data analysis, for instance, error-correction models, panel data analysis (Baltagi, 1995), dynamic panel data (Arellano and Bond, 1991), among others. While contributions that make use of VAR and ARIMA modelling procedures have been quoted, most of the literature relies on state space model variants that diverse degree of success have shown in dealing with constrained data. But for most social scientists this specific model (and more generally, Bayesian econometrics) usually is not studied in regular courses on Statistics or Econometrics.

Finally, the majority of the motivational cases of study of these papers come from social sciences problems. This is again paradoxical with the finding that only some of these statistical techniques are widely available for an average social scientist. We could say the same in terms of the required transformation for dealing with the constant-sum constraint.

Dynamic compositional problems are of substantive interest for social sciences. Examples like the evolution of federal budgets components, tax revenues compositions, income distribution, savings and investment composition during periods of crisis, among others represent interesting issues for future analysis. It is lacking the application of well known transformations into also well known least-squares-based methods for widening the knowledge and understanding of compositional time series.

[4] Besides, none of the works have used the isometric logratio that possesses such nice mathematical proprieties (Egozcué et al. 2003).

References

AITCHISON, J. *The Statistical Analysis of Compositional Data*. London, New York: Chapman and Hall. 1986, 417 p.

AITCHISON, J. and SHEN, S.M. "Logistic-normal distribution: some properties and uses". *Biometrika* 1980, vol. **2,** n**.** 67, p. 261-272.

ANDERSON, T.W. *The Statistical Analysis of Time Series*. New York: John Wiley & Sons, 1994.

ARELLANO, M. and BOND, S. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". *Review of Economic Studies,* 1991, vol. 58, 277-294.

BHAUMIK, A., DEY and D. K., RAVISHANKER, N. "A dynamic linear model approach for compositional time series analysis". Working Paper, University of Connecticut. 2003.

BALTAGI, B.H. *Econometric Analysis of Panel Data*. New York: John Wiley and Sons, 1995.

BERGMAN, J. and B. HOLMQUIST, "Poll of Polls: A Compositional Loess Model", *Scandinavian Journal of Statistics*, Vol. 41, p. 301–310, 2014. (DOI: 10.1111/sjos.12023).

BRAKEL, VAN DEN, J. and J. ROELS, "Intervention Analysis with State-Space Models to Estimate Discontinuities Due to a Survey Redesign", *The Annals of Applied Statistics*, vol. 4, no. 2 (June 2010), p. 1105-1138. (DOI: 10.1214/09-AOAS305).

BRANDT, P.T., MONROE, B.L. and WILLIAMS, J.T. "Time Series Models for Compositional Data". *Proceedings of the Meeting of the American Political Science Association*, Atlanta, 1999.

BROEMELING, L.D. and SHAARAWY, S. "A Bayesian Analysis of Time Series". In Goel, P., Zellner, A. (eds.). *Bayesian Inference and Decision Techniques*. Elsevier Science Publishers B.V., 1986.

BRUNSDON, T.M. *Times series of compositional data*. Ph.D. Thesis Dissertation. University of Southampton, 1986.

BRUNSDON, T.M. and SMITH, T.M.F. "The Time Series Analysis of Compositional Data". *Journal of Official Statistics,* 1988, vol. 14 n. 3, p. 237-253.

CARGNONI, C., MÜLLER, P. and WEST, M. “Bayesian Forecasting of Multinomial Time Series through Conditionally Gaussian Dynamic Models”, *Journal of the American Statistical Association* 1997, vol. 92, p. 640-647.

DAWSON, P. P. DOWNWARD, and T.C. MILLS (2014), “Olympic news and attitudes towards the Olympics: a compositional time-series analysis of how sentiment is affected by events”, *Journal of Applied Statistics* vol. 41 n. 6, p. 1307-1314, DOI: 10.1080/02664763.2013.868417.

EGOZCUE, J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. and BARCELÓ-VIDAL, C. “Isometric Logratio Transformations for Compositional Data Analysis”. *Mathematical Geology* 2003, vol. 35 n. 3, p. 279-300.

EGOZCUE, J., and PAWLOWSKY-GLAHN, V. “Simplicial geometry for compositional data”. In *Compositional Data Analysis in the Geosciences: From Theory to Practice*, Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds). Geological Society: London; 2006, p. 145–160.

ENDERS, W. *Applied Econometric Time Series*. 1995, Toronto: John Wiley & Sons. 433 p.

GRUNWALD, G.K. “*Time Series Models for Continuous Proportions*.” Ph.D. Thesis Dissertation. Department of Statistics. University of Washington. 1987, 104 p.

GRUNWALD, G.K., RAFTERY, A.E. and GUTTORP, P. “Time Series of Continuous Proportions”. *Journal of the Royal Statistical Society*, 1993 Series B 55**,** n. 1, p. 103-116.

HAMILTON, W. *Time Series Analysis*. Princeton: Princeton University Press, 1994, 816p.

HARVEY, A.C. *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press, 1989, 572 p.

KYNCLOVÁ, P., P. FILZMOSER and K. HRON, “Modeling Compositional Time Series with Vector Autoregressive Models”, *Journal of Forecasting*, 2015, vol. 34, 303–314.

KOEHLER, A.B., R.D. SNYDER, J. K. ORD and A. BEAUMONT, “Forecasting Compositional Time Series with Exponential Smoothing Methods”, Monash University Working Paper 20/10, 2010.

KOOP, G. *Bayesian Econometrics.* West Sussex: John Wiley & Sons, 2003, 376 p.

MILLS, T.C. “Forecasting obesity trends in England”. *Journal of The Royal Statistical Society Series* A - *Statistics in Society,* 2009a, vol. 172 n. 1, p. 107-117.

MILLS, T.C. “Forecasting compositional time series”, 2009b, *Quality & Quantity* vol. 44, n.4, p. 673-690.

PAWLOWSKY-GLAHN, V. and A. BUCCIANTI, *Compositional Data Analysis: Theory and Applications*, Wiley & Sons, 2011, 400 p.

PEARSON, K. “Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurements of organs.” *Proceedings of the Royal Society,* 1897, vol. 60, p. 489-498.

PORIER, D.J. and TOBIAS, J.L. “Bayesian Econometrics”. *Staff General Research Papers 12428*, 2005, Iowa State University, Department of Economics.

QUINTANA, J.M. and WEST, M. “Time Series Analysis of Compositional Data”. In Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.). *Bayesian Statistics,* 1988, vol. 3, p. 747-756.

RAVISHANKER, N., DEY, D.K. and IYENGAR, N. #Compositional Time Series of Mortality Proportions”. *Communications in Statistics - Theory and Methods* 2001, vol. 30, n. 11, p. 2281- 2291.

SIMS, C. “Macroeconomics and Reality”. *Econometrica,* 1980, vol. 48, n.1, p. 1-48.

SMITH, T.M.F., and BRUNSDON, T.M., “The Time Series Analysis of Compositional Data.” *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1989, p. 26-32.

THEIL, H. *Principles of Econometrics*. New York: John Wiley & Sons, 1971, 768 p.

WOODWARD, W. A., GRAY, H. L. and ELLIOT, A. C., *Applied Time Series Analysis*, CRC Press, 2012, 564 p.

ZELLNER, A. *Basic Issues in Econometrics*. Chicago: The Chicago University Press, 1984, 360 p.