American Journal of Mathematics and Statistics 2014, 4(2): 51-57
DOI: 10.5923/j.ajms.20140402.01

Analysis of Compositional Time Series from Repeated Surveys

Etebong P. Clement

Department of Mathematics and Statistics University of Uyo, P.M.B.1017 Uyo, Uyo, Nigeria

Abstract A compositional time series is a multivariate time series in which each of the series has values bounded between zero and one and the sum of the series equals one at each time point. Data with such characteristics are observed in repeated surveys when a survey variable has a multinomial response but interest lies in the proportion of units classified in each of its categories. The main approach to analyzing Compositional Time Series data has been based on the application of an initial transform to break the unit sum constraint. Box-Cox transformation originally was envisioned as a panacea for simultaneously correcting normality, linearity and homoscedasticity. However, one thing is clear; that seldom does this transformation fulfill the basic assumptions as originally suggested. This paper aims at reviewing works relating to these transformations with some modifications and illustrative example as would be applicable to the analysis of compositional time series data.

Keywords Box-Cox transformation, Compositional time series, Multinomial response, Repeated surveys

1. Introduction

Repeated surveys produce time series $\{y_t\}$ comprising estimates of the unknown target series $\{\theta_t\}$. If a survey is repeated at times $t = 1, \dots, T$, then multinomial responses at each time $t, (r_t say)$, lead to compositions. A composition is a vector of non-negative components summing to a constant, usually a unity. Symbolically, a vector x such that: $x = (x_i, \dots, x_D)^\iota$; $x_i > 0$ $(T = 1, \dots D)$; $\sum_{i=1}^{D} x_i = 1$ is a composition. A time series of composition is referred to as a compositional time series (CTS).

A compositional time series (CTS) is defined as a multivariate time series in which each of the series has values bounded between zero and one and the sum of the series equals one at each time point. Data with such characteristics are observed in repeated surveys when a survey variable has a multinomial response but interest lies in the proportion of unit classified in each of its categories. Therefore, the survey estimates are proportions of a whole subject to a unity-sum constraint.

A repeated survey is a sample survey which is performed more than once with essentially the same questionnaire or schedule but not necessarily with the same sample units. Many repeated surveys are based on a rotating panel design in which K panels of sampling units are investigated at each survey round (time point) and panels are replaced in a systematic manner, according to the rotating pattern of the survey design. In these surveys, elementary design unbiased estimates $y_t^{(k)}, k = 1, \dots K$, for the population parameters θ_t, can be obtained from each rotation group. A rotation group is a set of sampling units that joins and leaves the sample at the same time [1].

A repeated survey enables estimation of changes for the population as well as cross-sectional estimate. Monitoring and detecting important changes will usually be a key reason for sampling in time. Common frequencies for repeated survey are monthly, quarterly and annual. However more frequent sampling may be adopted as in the opinion polls leading up to an election and monitoring Television or Radio rating [2].

Some examples of repeated surveys are monthly labour force surveys in Australia. Quarterly surveys include the labour force survey in U.K and Ireland and many business surveys. Annual surveys include the Annual Survey of Manufacturers of the U.S. Census Bureau enumerates a fixed panel of economic establishments for five survey years. Establishments are selected with probabilities proportionate to size using Poisson sampling. The June Enumerative Survey of the National Agricultural Statistics Service is a yearly survey of agricultural activities. The farm costs and returns survey, also of the National Agricultural Statistics Service, enumerates a stratified simple random sample of farms each year.

In a repeated survey there is not necessarily any overlap of the sample for different occasions. A rotating panel surveys also uses a sample that is followed over time, but the focus is on estimates at aggregate levels. When the emphasis is on

* Corresponding author:
epclement@yahoo.com (Etebong P. Clement)

Published online at http://journal.sapub.org/ajms

estimates for the population an independent sample may be used on each occasion, which is often the case when the interval between the surveys is quite large. An option is to use the same sample at each occasion, with additions so that the sample estimates refer to the current population. For monthly or quarterly surveys the sample is often designed with considerable overlap between successive surveys. The sample overlap will reduce the sampling variance of estimates of change and reduce costs. Many important surveys are conducted repeatedly to give estimates of the level or mean for several time periods.

Repeated surveys can provide estimates for each time periods y_t. A major value of repeated surveys is in their ability to provide estimates of change. The simplest analysis of change is the estimate of one period change $y_t - y_{t-1}$. In a monthly survey this corresponds to one month change. For a survey conducted annually this corresponds to annual change. In general, therefore the change $\boldsymbol{s}$ time periods apart can be estimated as the difference at lag s: $\Delta^{(s)} y_t = y_t - y_{t-s}$.

The focus is often on $s = 1$, but for a survey repeated on a monthly basis changes for $s = 2,3,12$ are also commonly examined [2]. Having sample overlap at lag s will usually lead to a positive correlation between the estimates. Since

$$Var\left(\Delta^{(s)}y_t\right) = Var(y_t) + Var(y_{t-s}) - 2\sqrt{Var(y_t)Var(y_{t-s})cov(y_t, y_{t-s})},$$

having sample overlap reduces the variance of $\Delta^{(s)}(y_t)$ compared with having no sample overlap. [3] considered the components of change in a repeated survey [4-6] give a general review of issues in the design and analysis of repeated surveys. [7] cover many of the important issues associated with panel surveys. [8-10] review estimation issues for repeated surveys.

The focus of this paper is on compositional data from repeated surveys. Data of this kind frequently arise in disciplines as disparate as biology, demography, ecology, economics, geology and politics. Examples are: the percentage 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 share 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 [11]. In this paper we give a detailed review of developments in the field of the statistical analysis of compositional time series (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. The inverse transformation is then used on the derived results to obtain results pertinent to the original sample space. That is, the inverse transformation is applied to obtain the equivalent inferential results for the original compositional time series (CTS).

This approach was first discussed by [12] in the context of analyzing CTS from repeated sample surveys. In [12-14], the authors first proved that such an approach is in variant to the choice of the component used as the common divisor in the additive log ratio (alr) transformation. Secondly, assuming normality for the distribution of y_t, they obtained forecasts for the original CTS x_t by calculating the mean of the corresponding additive logistic distribution numerically.

In this paper two methods of analyzing CTS is discussed: The direct modeling in the simplex, and transformation of the simplex. An attempt is made at reviewing the works relating to the transformation of the simplex with some modifications.

2. Compositional Time Series

Let

$$\theta_t = \left(\theta_{1t}, \dots, \theta_{D+1,t}\right) \quad (1)$$

be a vector of population quantities of interest at time t, and assume that observations are taken at equally spaced time intervals $t = 1,2, \dots, T$.

Let

$$y_t = \left(y_{1t}, \dots, y_{D+1,t}\right) \quad (2)$$

represent a survey-based estimate of θ_t based on data collected at time t.

Repeated surveys produce time series $\{y_t\}$ comprising estimates of the unknown target series $\{\theta_t\}$. According to [1] focusing on the unknown population vector θ_t, it is natural to imagine that knowledge of $\theta_1, \dots, \theta_{t-1}$ conveys useful information about θ_t but without implying that it is perfectly predictable from $\theta_1, \dots, \theta_{t-1}$.

One way of representing this situation is by considering θ_t being a random variable which evolves stochastically in time following a certain time series model, as was first proposed for univariate survey analysis by [15-17]

The survey estimates y_t and θ_t of (1) and (2) can then be expressed as

$$y_t = \theta_t + e_t \quad (3)$$

where $\{\theta_t\}$, $\{y_t\}$ and $\{e_t\}$ are random processes and $e_t = \left(e_{1t}, \dots, e_{D+1,t}\right)^1$ are the sampling errors such that $E(e_t|\theta_t) = 0$ and $Var(e_t|\theta_t) = \Sigma_\mathrm{t}$.

Many variables investigated by statistical agencies have a multinomial response and interest lies in the estimation of the proportion of units classified in each of the categories. If this is the case, the vector of proportion sums to one and forms what is known as a composition.

A composition is a vector of non-negative components summing to a constant, usually 1, or put symbolically, a vector x such that $x = (x_i, \dots, x_D)^1; x_i > 0$ $(i =$ *1,...D;i=1Dxi=1.*

A time series of compositions is referred to as a Compositional Time Series (CTS). A Compositional Time Series is a sequence of vectors $y_t = \left(y_{1t}, \dots, y_{D+1,t}\right)$ each belonging to the simplex S^D.

If a survey is repeated at time $t = 1, \dots, T$, then multinomial response at each time at $\underline{r_t}$ say constitute compositions.

$$\left\{\underline{U_t}: 0 < U_{it} < 1, i = 1, \dots d; \sum_{i=1}^{d} U_{it} < 1, t = 1, \dots, T\right\}$$

which forms a multivariate time series.

The transformation of the series produces a multivariate time series defined on $\mathbb{R}^d$ at each time point t which can be analysed using standard methods. In particular [13] examined the use of ARMA models on the transformed series defined by $\underline{\phi}(B)\underline{V_t} = \underline{\theta}(B)\underline{\varepsilon_t}$.

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

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

In the multivariate case, the ideas of [18] who give a very simple procedure for choosing, estimating and testing such models is always followed.

However, it is always necessary to consider if the choice of reference variable in any way influences the analysis. Consequently, [12] proves the following results.

(i) Let $\underline{V_t}^{(k)} = \underline{z}(k)\underline{V_t}$

$$= \underline{z}(k)\left(\underline{V_t} - \upsilon\right) = \underline{V_t}^{(k)} - \underline{\upsilon}^{(k)},$$
$$(t = 0, \underline{+}1, \dots), (k = 1, \dots, d)$$

where $\underline{z}(k)$ is given by

$$\underline{z}(k) = \left\{z_{ij}^{(k)}\right\}$$
$$z_{ij}^{(k)} = 1 \quad (i = j \neq k; i, j = 1, \dots, d)$$
$$= -1 \quad (j = k, i = 1, \dots, m)$$
$$= \text{elsewhere}$$

and $\underline{\mu} = E(\underline{V_t})$, then if $\{\underline{V_t}\}$ follows a multivariate ARMA (p, q) process of dimension m then $\left\{\underline{V_t}^{(k)}\right\}$ is also multivariate ARMA (p, q). The roots of the determinantal equations of both the AR and MA components from the two models are identical so that the stationarity and invertibility conditions remain consistent.

(ii) Consider the compositional time series $\{\underline{U_t}\}$ where $a_d^{(k)}(\underline{u_t})$,

$(k = 1, \dots, d+1)$ follows an ARMA (p, q) process. Then each ARMA model $(k = 1, \dots, d+1)$ 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 consequences of results (i) and (ii) is that any component of $\underline{u_t^f}$ may be selected as the reference variable without affecting the final results. In what follows, we assume that the reference variable is $u_{d+1,t}$. The application of compositional data to modelling and forecasting is straight forward when the argument of [19] is followed.

Let the series $\underline{U_t}$ be transformed to $\underline{V_t}$.

$$\underline{V_t} = a_d\left(\underline{u_t}\right)$$

$\{\underline{V_t}\}$ is then modeled by the vector ARMA (p, q), forecasts for $\underline{V_{t+l}}$ can be obtained. Let the l-step a head forecast $\underline{V_{t+l}}$ of $\underline{V_t}$ be denoted by $\underline{V_t}(l)$ and its covariance matrix $\sum_t(l)$, a "naïve" forecast for $\underline{u_{t+l}}$ as:

$$\underline{\upsilon_t}(l) = a_d^{-1}(\underline{V_t}(l))$$

Assuming normality for the distribution of V_t so that $(\underline{V_{t+l}}|\underline{V_{t-1}}, \dots) \sim N\left(\underline{V_t}(l), \underline{\Sigma}_{\mathrm{t}}(l)\right)$. The optimum forecast of $\underline{u_{t+l}}, \underline{u_t}(l)$ may be obtained numerically by calculating the mean of $L_d\left(\underline{V_t}(l), \underline{\Sigma}_{\mathrm{t}}(l)\right)$ or $\underline{u_t}(l)$ may be approximated. Also a confidence region for $\underline{U_{t+l}}$ may be obtained following standard multivariate theory, though the confidence region will not centered at $\underline{u_t}(l)$.

A 100 $(1 - \alpha)$% confidence region for $\underline{u_{t+l}}$ according to [13] can be formed from

$$\left[V_t(l) - ln\left\{\frac{\underline{U_{t+l}}}{U_{d+1,t+l}}\right\}^T \underline{\Sigma}_t^{-1}(l)\left[V_t(l) - ln\left\{\frac{\underline{U_{t\pm l}}}{U_{d+m,t+l}}\right\}\right]\right]$$
$$\leq \chi_{\alpha,d}^2$$

where $\chi_{\alpha,d}^2$ is the α% point of a $\chi_{(d)}^2$ distribution, by mapping points from $\mathbb{R}^d$ onto the simplex S^d.

Also forecasts for either the ratios $U_{i,t+l}/U_{j,t+l}$ or generally the log-ratios may be obtained.

$$\left(U_i|U_j\right)_t(l) = exp\left\{V_{it}(l) - V_{jt}(l)\right.$$
$$\left.+\frac{1}{2}\left(\sigma_{iit}(l) - 2\sigma_{ijt}(l) + \sigma_{jjt}(l)\right)\right\}$$

where $\underline{\Sigma}_{\mathrm{t}}(l) = \{\sigma_{ijt} + (l)\}$

3. Analyzing Compositional Time Series

Two methods of analyzing compositional time series will be explored, namely: Direct method and transformation method. Under the transformation methods of analysis, we shall examine two techniques: Box and Cox transformation and the log-ratio transformation. Again, the log ratio transformation shall be viewed under: (i) additive log ratio (alr) transformation (ii) centered log ratio transformation (clr) and (iii) isometric log ratio transformation.

3.1. Direct Modeling in the Simplex

Around the same time as the publication of [12] and [20-21] introduced a different approach to analyzing CTS, which had also been inspired by some of the earlier ideas of Aitchison. There and in [22], 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. The state distribution was assumed to be Dirichlet conjugate. This was a new generalization of the Dirichlet distribution proposed by them in order to allow for dependence between the components.

A vector of continuous proportions consists of the proportions of some total accounted for by its constituent components (compositions). We consider the situations where time series data are available and where interest focuses on the proportions rather than the actual amounts. A state space model for time series of compositions conditionally on the unobserved state, the observation are assumed to follow the Dirichlet distribution, often considered to be the most natural distribution on the simplex. The state follows the Dirichlet conjugate distribution.

Let $y = (y_1, y_2, \dots, y_D)^1$ be a vector of continuous proportions, namely a vector with positive components such that $y^T u = 1$.

Where $U = (1,1,\dots,1)^1$ is a $(d+1)$ - vector of 1s.

Then y follows the Dirichlet distribution if it has the density

$$\mathrm{P}(y|\alpha) = D(\alpha)^{-1} \prod_{i=1}^{d+1} y_i^{\alpha_i - 1} \tag{4}$$

In density (4) $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{d+1})'$ where $\alpha_i > 0$ for $i = 1, \dots, d+1$ and

$$D(\alpha) = \Gamma(\alpha^T u)^{-1} \prod_{i=1}^{d+1} \Gamma\, \alpha_i$$

is the Dirichlet function, a $(d+1)$ - dimensional analogue of the beta function. We denote this situation by $y \sim D(\alpha)$.

The sample space is the d-dimensional simplex S^d ; $S^d = \{y \epsilon\, \mathbb{R}_t^D : y^T u = 1\}$

Expressing (4) in exponential family form, we have:

Let $V = log\, y$

$\tau = V^T u/(d+1)$ and

$Z = V - u\tau$

Z is the vector of symmetric log ratios (clr) and Z = clr (y)

Also let $\theta = \alpha/\lambda$

where $\lambda = \alpha^T u$ so that $y - D(\lambda\theta)$ .Then density (4) becomes:

$$\mathrm{P}(Z|\theta, \lambda) = \exp\{\lambda Z^T \theta + \lambda\tau - logD(\lambda\theta)\} \tag{5}$$

The sample space is $H^d = \{Z \epsilon \mathbb{R}^{d+1} : Z^T u = 0\}$ and the parameters space is $(\theta, \lambda)\ \epsilon\ S^d \times \mathbb{R}_t$. The purpose of this reparameterization according to [22] is to separate the effects of location θ and spread λ as far as possible.

3.2. Transformation Method

The sample space of a composition x is referred to as the simplex S^d. It has been known since the days of [23] that normal statistical methods are not applicable to element of the simplex (the compositions).

The major way, following the ideas of Aitchison of resolving these problems has been through transformation.

3.2.1. Box-Cox Transformation

[24] introduced the use of the well-known Box-Cox transformation as an alternative to the additive log ratio (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 known is that presented in [25]. These authors modeled 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, among others.

[25] used the same complex procedure as those proposed in [26] to carry out model selection and inference. They illustrated their approach using two CTS; the mortality data from Los Angeles (analyzed previously by [26] and a CTS on vehicle production which had been previously analyzed by [21].

[27] introduced a family of power transformation such that the transformed values are a monotonic function of the observations over some admissible range and indexed by

$$y_i^{(\lambda)} = \begin{cases} y_i^{(\lambda)} & \lambda \neq 0 \\ \log y_i & \lambda = 0 \end{cases} \tag{6}$$

for $y_i > 0$. However, this family has been modified by [28] to take account of the discontinuity at $\lambda = 0$, such that

$$y_i^{(\lambda)} = \begin{cases} \left(y_i^{(\lambda)} - 1\right)/\lambda & \lambda \neq 0 \\ \log y_i & \lambda = 0 \end{cases} \tag{7}$$

and that for unknown λ

$$y^{(\lambda)} = \left(y_1^{(\lambda)}, y_2^{(\lambda)}, y_n^{(\lambda)}\right)' = \mathrm{X}\theta + e$$

where X is a matrix of known constants, θ is a vector of unknown parameters associated with the transformed values and $e \sim MVN(0, \sigma^2 \mathrm{I_n})$ is a vector of random errors. The transformation in equation (7) is valid only for $y_i > 0$ and, therefore, modifications have had to be made for negative observations. [28] proposed the shifted power transformation with the form

$$y_i^{(\lambda)} = \begin{cases} \{(y_i + \lambda_2)^{\lambda_1} - 1\}/\lambda_1 & \lambda_1 \neq 0 \\ \log(y_i + \lambda_2) & \lambda_1 = 0 \end{cases} \tag{8}$$

where λ_1 is the transformation parameter and λ_2 is chosen such that $y_i > -\lambda_2$.

[29] introduced the so-called modulus transformation which is considered to normalize distributions already possessing some measure of approximate symmetry and carries the form

$$y_i^{(\lambda)} = \begin{cases} sign\ (y_i)\{(|y_i| + 1)^{\lambda} - 1\}/\lambda & \lambda \neq 0 \\ sign\ (y_i)\ \{log\ (|y_i| + 1)\} & \lambda = 0 \end{cases} \tag{9}$$

[30] suggested another alternative which can be used with negative observations and which is claimed to be effective at turning skew unimodal distributions into nearly symmetric normal-like distributions and is of the form:

$$y_i^{(\lambda)} = \begin{cases} (\exp(\lambda y_i) - 1)/\lambda & \lambda \neq 0 \\ y_i & \lambda = 0 \end{cases} \tag{10}$$

[31] suggested another modification so that distributions of $y_i^{(\lambda)}$ with unbounded support such as the normal distribution can be included. For $\lambda > 0$, the extension is:

$$y_i^{(\lambda)} = \{|y_i|^{\lambda} sign(y_i) - 1\}/\lambda \tag{11}$$

It is important to note that the ranged of $y_i^{(\lambda)}$ in equations (6) – (9) is restricted according to whether λ is positive or negative. This implies that the transformed values do not cover the entire range $(-\infty, +\infty)$ and, hence, their distributions are of bounded support. Consequently, only approximate normality is to be expected.

It is also remarked that since [28] transformation, other modifications of the transformation for special applications and circumstances had been made, but for most researchers, the original Box-Cox transformation of equation (7) suffices and is preferable due to computational simplicity.

3.2.2. Log Ratio Transformation

Let $\mathcal{A}_{\mathrm{DxD}}$ denote the family of all real D x D matrices such that $\mathrm{AI_D} = \mathrm{A'I_D} = 0_\mathrm{D}$

Let $\mathrm{X} \in \mathrm{S^D}$ and $\mathrm{A} \in \mathcal{A}_{\mathrm{DxD}}$. We defined the product A⊙X as:

$$\mathrm{A} \odot \mathrm{X} = \mathrm{C}\left(\prod_{\mathrm{i}=1}^{\mathrm{D}} x_i^{a_{1i}}, \dots, \prod_{\mathrm{i}=1}^{\mathrm{D}} x_i^{a_{\mathrm{D}i}}\right)'$$

The function $\mathrm{X} \longrightarrow \mathrm{A} \odot \mathrm{X}$ is an endomorphism of the vector space $(\mathrm{S^D}, \oplus, \odot)$. Moreover, any endomorphism of $\mathrm{S^D}$ can be written in this form. The matrix associated to identity endomorphism is the well-known centering matrix $\mathrm{G_D} = \mathrm{I_D} - \mathrm{D^{-1}J_D}$ of order D X D.

(i) Additive Log ratio Transformation (alr)

The alr transformation of index i $(i = 1, \dots, D)$ denoted by alr(x) is the one-to-one transformation from $\mathrm{S^D}$ to $\mathbb{R}^\mathrm{D}$ define as:

$$\mathrm{X} \longrightarrow \mathrm{y} = \mathrm{alr(x)} = \mathrm{y}$$

$$y = \left[ln\frac{x_1}{x_D}, ln\frac{x_2}{x_D}, \dots, ln\frac{x_{D-1}}{x_D}\right]$$

$$y = \ln(x)\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & \cdots & 1 \\ -1 & -1 & \cdots & -1 \end{pmatrix}$$

where $y \epsilon\, \mathbb{R}^\mathrm{D}$ and $d = D - 1$

The inverse denoted $alr^{-1}(y)$ or (gal) is defined as:

$$alr^{-1}(y) = C[exp([y; 0])]$$

where gal means generalized additive logistic transformation

$$\mathrm{gal(y)} = \left[\frac{\exp(\mathrm{y_1})}{\sum_{i=1}^{\mathrm{D}-1}(\exp(\mathrm{y}_i)) + 1}, \dots, \frac{\exp(\mathrm{y_{D-1}})}{\sum_{i=1}^{\mathrm{D}-1}(\exp(\mathrm{y}_i) + 1)}, 1 - \mathrm{x_1} - \cdots - \mathrm{x_{D-1}}\right]$$

The additive log ratio transformation is asymmetric in the parts of the compositions.

(ii) Centered Log Ratio Transformation (clr)

The centered (or symmetric) log ratio transformation denoted by clr is the function from the compositional space $\mathrm{S^D}$ to $\mathbb{R}^\mathrm{D}$, defined by: $\mathrm{X} \longrightarrow \mathrm{Z} = \mathrm{clr(x)} = \mathrm{Z}$

$$Z = \left[ln\frac{x_1}{g(x)}, \dots, ln\frac{x_D}{g(x)}\right]$$

$$= \frac{\ln(x)}{\mathrm{D}}\begin{pmatrix} D-1 & -1 & \cdots & -1 \\ -1 & D-1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & \cdots & D-1 \end{pmatrix}$$

where $Z \epsilon \mathbb{R}^\mathrm{D}$ and $g_D(\mathrm{x})$ is the geometric mean $\left(\prod_{\mathrm{k}=1}^{\mathrm{D}} \mathrm{x_k}\right)^{1/\mathrm{D}}$ of x.

The inverse denoted by $(\mathrm{clr})^{-1}(\mathrm{Z})$ is defined by

$$(\mathrm{clr})^{-1}(\mathrm{Z}) = \mathrm{C}[\exp(\mathrm{Z})]$$

$$= \frac{\exp(\mathrm{y_1})}{\sum_{i=1}^{\mathrm{D}} \exp(\mathrm{y}_i)}, \dots, \frac{\exp(\mathrm{y_D})}{\sum_{i=1}^{\mathrm{D}} \exp(\mathrm{y}_i)}$$

This transformation is symmetric in the parts of the composition. The transformation maps $\mathrm{S^D}$ in the subspace $V = \left\{Z \epsilon \mathbb{R}^\mathrm{D} : \sum_{i=1}^{\mathrm{D}} \mathrm{Z} = 0\right\}$ of $\mathbb{R}^\mathrm{D}$, which can be seen to be a hyperplane through the origin of $\mathbb{R}^\mathrm{D}$, orthogonal to I_D (vector of units). This subspace has dimension $D - 1$. If $V_1, \dots, V_{D-1}$ be any orthonormal basis of V and if V be the D x (D − 1) matrix $[V_1, : \dots : V_{D-1}]$.

(iii) Isometric Log Ratio Transformation (ilr)

The isometric log ratio transformation denoted by $ilr_v = clr(x) \cdot V = \ln(X) \cdot V$.

For a given matrix V of D rows and (D-1) columns such that $\mathrm{V} \cdot \mathrm{V}' = \mathrm{I_{D-1}}$ (identity matrix of D − 1 elements) and $\mathrm{V} \cdot \mathrm{V}' = \mathrm{I_D} + \boldsymbol{a}\mathbf{1}$ where $\boldsymbol{a}$ may be any value and **1** is a matrix full of ones.

Alternatively, $ilr(x) = (\mathrm{y_1}, \dots, \mathrm{y_{D-1}}) \epsilon \mathbb{R}^\mathrm{d}$ where d=D-1

where $y_k = \frac{1}{\sqrt{k(k+1)}} ln\left(\frac{\prod_{i=1}^{k} \chi_i}{(x_{k+1})^k}\right)$ $(\mathrm{k} = 1, \dots, \mathrm{D} - 1)$

The inverse denoted by (ilr)$^{-1}$ is defined as:

$$ilr_v^{-1}(x) = \mathrm{C}[exp(x.v')]$$

$$= \left[\left(1 + \sum_{i=0, i\neq 1}^{D} f(i)\right)^{-1}, \dots, \left(1 + \sum_{i=0, i\neq D}^{D} f(i)\right)^{-1}\right],$$

where $f(i) = \left(\frac{1}{f(i-1)} exp\left(\sqrt{i(i+1)\mathrm{y_i}}\right)\right)^{\frac{-1}{i}}$ and $f(0) = 1$

Let evaluation the log ratio transformations when D=3 and 4.
For $\mathrm{D} = 3 : \mathrm{x} = (x_1, x_2, x_3)'$

(i) $alr_i(x) = [y_1; y_2] = ln\left[\frac{x_1}{x_3}; ln\frac{x_2}{x_3}\right]$

where $x = \left[\frac{\exp(y_1); \exp(y_2); 1}{exp(y_1) + \exp(y_2) + 1}\right]$

(ii) $clr_i(x) = \mathrm{Z_i} = ln\frac{x_\mathrm{i}}{\sqrt[3]{x_1 x_2 x_3}}$

where $x_i = \frac{\exp(Z_\mathrm{i})}{exp(Z_1) + \exp(Z_2) + \exp(Z_3)}$

$$= \frac{\exp(Z_\mathrm{i})}{\sum_{i=1}^{3} exp(Z_\mathrm{i})}$$

(iii) $ilr_v(x) = \left[\frac{1}{\sqrt{2}} ln\frac{x_2}{x_3}; \frac{1}{\sqrt{6}} ln\frac{x_1^2}{x_1 x_3}\right]$

$$\text{where V} = \begin{pmatrix} 0 & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \end{pmatrix}$$

Again if D=4, that is, $x = (x_1, x_2, x_3, x_4)'$ then the resulting vectors of the different transformations are the following:

$$alr\ (x) = \left(ln\frac{x_1}{x_4}, ln\frac{x_2}{x_4}, ln\frac{x_3}{x_4}\right)'$$

$$clr\ (x) = \left(ln\frac{x_1}{g(x)}, ln\frac{x_2}{g(x)}, ln\frac{x_3}{g(x)}, ln\frac{x_4}{g(x)}\right)'$$

$$ilr\ (x) = \left(\frac{1}{\sqrt{2}} ln\frac{x_1}{x_2}, \frac{1}{\sqrt{6}} ln\frac{x_1 x_2}{x_3^2}, \frac{1}{\sqrt{12}}\ ln\frac{x_1 x_2 x_3}{x_4^3}\right)'$$

where g(x) is the geometric mean as defined earlier.

It is very important to emphasize that all these transformations - $alr(x)$, $clr(x)$, $ilr(x)$ and its inverses are one-to-one linear transformations between the compositional vector space $(\mathrm{S}^{\mathrm{D}}, \oplus, \odot)$ and the real vector space $(\mathbb{R}^{D-1}, +, \cdot)(or\ V\epsilon\mathbb{R}^{D})$ with the natural structure. Vectors $\boldsymbol{u} = ilr_v\boldsymbol{x}, \boldsymbol{y} = alr_D\boldsymbol{x}$ and $\boldsymbol{z} = clr\ \boldsymbol{x}$ associated with the same composition $x\epsilon\mathrm{S}^{\mathrm{D}}$ are related by the following linear relationship expressed in matrix form.

1. $\mathrm{u} = (\mathrm{FV})^{-1}\mathrm{y}$ and $\mathrm{u} = (\mathrm{FV})^{-1}\mathrm{Fz}$.
2. $y = FVu$ and $y = Fz$
3. $z = ((FV)^{-1}F)'u$ and $z = F'H^{-1}y$ where H is the $(\mathrm{D}-1)$ x $(\mathrm{D}-1)$ matrix $FF' = \mathrm{I}_{\mathrm{D}-1} + \mathrm{J}_{\mathrm{D}-1}$, with $\mathrm{J}_{\mathrm{D}-1} = \mathrm{I}_{\mathrm{D}-1}\mathrm{I}_{\mathrm{D}-1}'$.

4. Conclusions

The Box-Cox transformation has been widely used since it was first proposed. It has inspired a large amount of research on its applicability as well as on the drawbacks arising from its usage. However, one thing is clear; that seldom does this transformation fulfill the basic assumptions of linearity, normality and homoscedasticity simultaneously as originally suggested by [28]. A review of alternatives approaches is presented with modifications and illustrations useful to the analysis of compositional time series data.

REFERENCES

[1] Silva, D.B.N. and Smith, T.M.F. 2001, Modeling compositional time series from repeated surveys. Survey Methodology, 27,205-215.

[2] Steel, D.G. and McLaren, C. 2008, Design and analysis of repeated surveys. Centre for Statistical and Survey Methodology. University of Wollongong, Working Paper Series, 11-08, 2008,13p.http://ro.uow.edu.au/cssmwp/10.

[3] Hott, D. and Skinner, C. J. 1983, Component of change in repeaters surveys. International Statistical Review 57, 1-18.

[4] Duncan, G.J. and Kalton,G. 1987, Issues of design and analysis of surveys across time. International Statistical Review, 55, 97-117.

[5] Kalton, G. and Citro, C.F. 1993, Panel surveys:adding the fourth dimension. Survey Methodology,19,205-215.

[6] Steel, D.G., 2004, Sampling in time. Encyclopaedia of social measurement. Academic Press, 823-832.

[7] Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. 1989, Panel surveys. John Willey and Sons, New York.

[8] Smith, T.M.F. 1978, Principal and problems in the analysis of repeated surveys. Survey Sampling and Measurement. Ed. N. K. Namboodiri, Academic Press, New York.

[9] Binder, D.A. and Hidiroglou,M,A. (1988). Sampling in time: in handbook of Statistics, (Eds., P.R. Krishnaiah and C.R. Rao). Elsevier Science 6,187-211.

[10] Fuller, W. A. 1990, Analysis of repeated surveys. Survey Methodology, 16,167-180.

[11] Aguilar Zuil,L., Barcelo-Vidal, C. and Larrosa, J.M. 2007, Compositional time series analysis: a review in proceedings of the 56th session of the ISI (ISI 2007), Lisbon, August 22-29.

[12] Brunsdon, T.M. 1987, The time series analysis of compositional data. Ph.D. Thesis, university of Southampton, U.K.

[13] Smith, T.M.F., and Brunsdon, T.M. 1989, The time series analysis of compositional data. Proceedings of American Statistical Association, 26-32.

[14] Brunsdon, T.M. and Smith, T.M.F. 1998, The time series analysis of compositional data. Journal of Official Statistics 14 (3), 237-252.

[15] Blight, B. J. N. and Scott, A. J. 1973, A stochastic model for repeated surveys. Journal of the Royal Statistical Society B: Methodological 35, 61-68.

[16] Scott,A.J. and Smith, T.M.F. 1974, Analysis of repeated surveys using time series methods. Journal of American Statistical Association 69,674-678.

[17] Scott, A. J., Smith, T.M.F. and Jones, R.G. 1977, The application of time series methods to the analysis of repeated surveys. International Statistical Review, 43, 13-28.

[18] Tiao,G.C. AND Box, G.E.P. 1981, Modelling multiple time series with applications. Journal of American Statistical Association, 76, 802-816.

[19] Wallis, K.P. 1987, Time series analysis of bounded economic variables. Journal of Time Series Analysis, 8,115-123.

[20] Quintana, J. M. and West, M. 1988, The time series analysis of compositional data. Journal of Bayesian Statistics, 3,747-756.

[21] Grunwald, G.K. 1987, Time series models for continuous proportions. Ph.D. Thesis, University of Washington.

[22] Grunwald, G.K. Raftery, A.E. and Guttorp, P. 1993, Time series models for continuous proportions. Journal of Royal Statistical Society B, 55,103-116.

[23] 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. Proceedings of the Royal Society of London, LX, 489-498.

[24] Aitchison, J. 1986, The statistical analysis of compositional data. Chapman and Hall, London.

[25] Bhaumik,A., Dey, D.K and Ravishanker, N. 2003, A dynamic linear model approach for compositional time series analysis. Technical Report. University of Connecticut.

[26] Ravishanker, N., Dey, D.K. and Iyengar, M. 2001, Compositional time series analysis of mortality proportions. Communication in Statistics Theory Methodology, 30(11),

2281-2291.

[27] Turkey, J.W. 1957, The comparative anatomy of transformation. Annals of Mathematical Statistics, 28,602-632.

[28] Box,G.E.P. and Cox,D.R. 1964, An analysis of transformation. Journal of the Royal Statistical Society, Series B, 26,211-252.

[29] John, J. A and Drapper, N.R. 1980, An alternative family of transformation. Applied Statistics, 29,190-197.

[30] Manly, B. F. 1976, Exponential data transformation. The Statistician, 25, 37-42.

[31] Bickel, P.J. and Doksum, K. A. 1981, An analysis of transformation revisited. Journal of the American Statistical Association, 76,296-311.