# A simultaneous spatial autoregressive model for compositional data

Thi Huong An Nguyen[13], Christine Thomas-Agnan[1], Thibault Laurent[2] & Anne Ruiz-Gazen[1]

[1] *Toulouse School of Economics, University of Toulouse Capitole, France*
[2] *Toulouse School of Economics, CNRS, University of Toulouse Capitole, France*
[3] *Danang University of Architecture, Vietnam*

**Abstract.** In an election, the vote shares by party on a given subdivision of a territory form a vector with positive components adding up to 1 called a composition. Using a conventional multiple linear regression model to explain this vector by some factors is not adapted for at least two reasons: the existence of the constraint on the sum of the components and the assumption of statistical independence across territorial units questionable due to potential spatial autocorrelation. We develop a simultaneous spatial autoregressive model for compositional data which allows for both spatial correlation and correlations across equations. We propose an estimation method based on two-stage and three-stage least squares. We illustrate the method with simulations and with a data set from the 2015 French departmental election.

**Keywords.** multivariate spatial autocorrelation, spatial weight matrix, three-stage least squares, two-stage least squares, simplex, electoral data, CoDa.

## 1 Introduction

Some data present simultaneously the characteristics of compositional data (vectors with positive components conveying relative information) as well as the characteristics of spatial data (presence of spatial heterogeneity and spatial dependence). For example, land cover data contain information about different land use shares and the statistical unit is a subdivision of a territory; among the many papers that treat this type of data (see Leininger et al., 2013; Overmars et al., 2003; Yoshida and Tsutsumi, 2018; Pirzamanbein et al., 2018). Another instance is in geochemistry where data consist of composition of mineral deposits into chemical elements at different locations in geographical space, see for example Rubio et al. (2016) who study sediments in an artic lake or Filzmoser et al. (2010) who examine the Kola moss layer composition data from the R package StatDA (Filzmoser, 2020). This is also the case in political economy for electoral data containing the vote shares by party in a multiparty election for a list of administrative subdivisions of a territory as in Katz and King (1999) or for data about turnout rates as in Borghesi and Bouchaud (2010). Other examples include the distribution of temperature data at weather stations as in Salazar et al. (2015), the distribution of benthic macroinvertebrates at sampling stations in the Delaware Bay in Billheimer et al. (1997).

The challenge for modelling such data is to accomodate at the same time their compositional and spatial nature. For spatial data with a continuous domain, Pawlowsky and Burger (1992); Pawlowsky-Glahn and Egozcue (2016); Rubio et al. (2016); Martins et al. (2016) adopt a geostatistical approach. On the other hand, conditional autoregressive models have been developed for the multivariate regression framework (MCAR), e.g. Mardia (1988) and Gelfand and Vounatsou (2003). An application of MCAR to compositional data can be found in Billheimer et al. (1998) where vectors of counts are modelled by multinomial distributions whose parameters follow a prior Gaussian Markov random field. These hierarchical Bayesian models need Markov Chain Monte Carlo procedures for fitting. As in our application, Pirzamanbein et al. (2018) and Leininger et al. (2013) deal with areal data. Pirzamanbein et al. (2018) combine a Dirichlet distribution with a Gaussian Markov random field prior for the alr transformed vector of Dirichlet parameters. Leininger et al. (2013) propose a power transformation combined with an MCAR model in order to address the problem of observed zero proportions. We choose to focus rather

on a combination of ilr transformations with a spatial econometrics model which has the advantage of a very easy implementation involving only ordinary least squares steps. The compositional vector being the dependent variable, we will need spatial econometrics models for multivariate dependent variable as in Kelejian and Prucha (2004). We develop a simultaneous spatial autoregressive model for compositional data (CoDa) which allows for both spatial correlation and correlations across equations. We propose an estimation method based on two-stage (S2SLS) and three-stage (S3SLS) least squares.

In Section 2, we first recall some classical facts adapted to work with CoDa. We then introduce a new operation which will be necessary later to write our model in a simplex fashion and study its properties. In Section 3, we recall facts about the definition and estimation of simultaneous autoregressive models for multivariate output spatial data and combine with the tools of Section 2 to define our model for spatio-compositional data. Section 4 presents some simulations to evaluate the quality of the S2SLS and S3SLS methods in the multivariate case. Section 5 presents an application to election results with the question of the impact of socio-economic variables on parties vote shares with a data set from the 2015 French departmental election. Section 6 concludes.

# 2 Definitions and notations in compositional data analysis

A $D$-composition $\mathbf{u}$ is a vector of $D$ parts of some whole which carries relative information and therefore can be represented in the so-called simplex space $\mathbf{S}^D$ defined by Aitchison (1986):

$$\mathbf{S}^D = \left\{ \mathbf{u} = (u_1, \ldots, u_D)^T : u_m > 0, m = 1, ..., D; \sum_{m=1}^{D} u_m = 1 \right\},$$

where $^T$ is the transposition operator. For any vector $\mathbf{w} \in \mathbb{R}^{+D}$, the closure operation is defined by

$$\mathcal{C}(\mathbf{w}) = \left( \frac{w_1}{\sum_{m=1}^{D} w_m}, \cdots, \frac{w_D}{\sum_{m=1}^{D} w_m} \right).$$

Let us recall the usual operations used to define a vector structure on the simplex space.

1. $\oplus$ is the perturbation operation, corresponding to the addition in $\mathbb{R}^D$:

$$\mathbf{u} \oplus \mathbf{v} = \mathcal{C}(u_1 v_1, \ldots, u_D v_D), \ \mathbf{u}, \mathbf{v} \in \mathbf{S}^D$$

2. $\odot$ is the power operation, corresponding to the scalar multiplication in $\mathbb{R}^D$:

$$\lambda \odot \mathbf{u} = \mathcal{C}(u_1^\lambda, \ldots, u_D^\lambda), \ \lambda \text{ is a scalar}, \mathbf{u} \in \mathbf{S}^D$$

Moreover, the compositional product of a matrix by a vector denoted by $\boxdot$ is defined as follows

$$\mathbf{B} \boxdot \mathbf{u} = \mathcal{C} \left( \prod_{m=1}^{D} u_m^{b_{1m}}, \cdots, \prod_{m=1}^{D} u_m^{b_{Lm}} \right)^T$$

where $\mathbf{u} \in \mathbf{S}^D$, $\mathbf{B} = (b_{lm})$ with $l = 1, \ldots L$, $m = 1, \ldots, D$ is a $L \times D$ matrix.

The simplex $\mathbf{S}^D$ can also be equipped with the compositional/Aitchison inner product (see Aitchison, 1985; Pawlowsky-Glahn et al., 2015) in order to define distances.

The analysis of compositional data makes use of log-ratio transformations which map the simplex $\mathbf{S}^D$ to $\mathbb{R}^q$ (where most often $q = D - 1$) because of their degree 0 homogeneneity (scale invariance). The

classical ones are the additive log-ratio (alr), the centered log-ratio (clr) and the isometric log-ratio (ilr) transformations. In this paper, we will mainly use some ilr transformations. Because it is needed to define the ilr, let us first recall the definition of the clr transformation of a vector $\mathbf{u} \in \mathbf{S}^D$

$$\mathrm{clr}(\mathbf{u}) = \left( \ln \frac{u_m}{g(\mathbf{u})} \right)_{m=1,\dots,D},$$

where $g(\mathbf{u}) = \sqrt[D]{u_1 \cdot u_2 \cdots u_D}$ is the geometric mean of the components. Let $\mathbf{V}_D$ be a $D \times (D-1)$ contrast matrix (e.g. Pawlowsky-Glahn et al., 2015) associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \mathrm{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1}),$$

where clr is understood columnwise. For each such matrix $\mathbf{V}_D$, an isometric log-ratio transformation (ilr) is then defined by:

$$\mathbf{u}^* = \mathrm{ilr}(\mathbf{u}) = \mathbf{V}_D^T \ln(\mathbf{u})$$

where the logarithm of $\mathbf{u} \in \mathbf{S}^D$ is understood componentwise. The inverse transformation is given by:

$$\mathbf{u} = \mathrm{ilr}^{-1}(\mathbf{u}^*) = \mathcal{C}(\exp(\mathbf{V}_D \mathbf{u}^*)).$$

Since our data is made of samples of composition vectors, we store them in a $n \times D$ matrix $\mathbf{Y} = (\mathbf{Y}_{il})$, $i = 1, \dots, n,\ l = 1, \dots, D$. Each row of this matrix, denoted by $\mathbf{Y}_{i.}$, is a compositional column vector of $\mathbf{S}^D$. $\mathbf{Y}_{.l}$, $l = 1, \dots, D$ denotes the $l^{th}$ column of $\mathbf{Y}$ and we have

$$\mathbf{Y} = [\mathbf{Y}_{1.} \dots \mathbf{Y}_{n.}]^T = [\mathbf{Y}_{.1} \dots \mathbf{Y}_{.D}] \tag{1}$$

Let us define an extension of the ilr transformation of a matrix $\mathbf{Y}$ by

$$\mathrm{ilr}(\mathbf{Y}) = \ln(\mathbf{Y})\mathbf{V}_D = \begin{bmatrix} \mathrm{ilr}(\mathbf{Y}_{1.})^T \\ \vdots \\ \mathrm{ilr}(\mathbf{Y}_{n.})^T \end{bmatrix}$$

Note that $\mathrm{ilr}(\mathbf{Y})$ is a $n \times (D-1)$ matrix.

In spatial econometrics models, spatial weight matrices are used to specify the neighborhood structure. For $n$ spatial locations, the elements $w_{ij}$ of the $n \times n$ matrix $\mathbf{W}$ are measures of proximity between locations $i$ and $j$ (see for instance Bivand et al., 2008, for different specifications). These matrices determine a covariance model for the data vector and play a role similar to the spatial variogram in geostatistics. For such a matrix and a given data vector $\mathbf{Z}$ of size $n$, the lagged vector $\mathbf{WZ}$ contains averages of the values of the variable $\mathbf{Z}$ in neighboring locations when $\mathbf{W}$ is row normalized. In our case, we need to apply such an operation to each column of the data matrix $\mathbf{Y}$ and we wish that the application of this process to each column of $\mathbf{Y}$ results in a matrix in the same space as the original one $(\mathbf{S}^D)^n$. As usual in CoDa, we use the principle of working in log-ratio coordinates (Mateu-Figueras et al., 2011) and expressing the results in the simplex. We thus define the following operation.

**Definition 1.** Let $\mathbf{W}$ be a $n \times n$ matrix. The operation $\triangle$ is a map from the cartesian product of simplex spaces $(\mathbf{S}^D)^n$ to itself defined by

$$\mathbf{W} \triangle \mathbf{Y} = \mathrm{ilr}^{-1}(\mathbf{W}\mathrm{ilr}(\mathbf{Y})) = \mathrm{ilr}^{-1}(\mathbf{W} \ln(\mathbf{Y})\mathbf{V}_D) \tag{2}$$

where $\mathbf{V}_D$ is a $D \times (D-1)$ contrast matrix.

Note that $(\mathbf{W}\triangle\mathbf{Y}) \in (\mathbf{S}^D)^n$ and $\mathbf{W}\text{ilr}(\mathbf{Y}) \in (\mathbb{R}^{(D-1)})^n$.

This operation satisfies the following properties:

**Proposition 1.** Let $\mathbf{Y}$ be a $n \times D$ matrix such that each row, denoted by $\mathbf{Y}_{i.}$, $i = 1, \ldots, n$ is a compositional vector in $\mathbf{S}^D$. Let $\mathbf{W} = (W_{ij})$, $(i, j = 1, \ldots, n)$, a $n \times n$ matrix and $\alpha \in \mathbb{R}$. We have

1. $\mathbf{W}\triangle(\alpha \odot \mathbf{Y}) = \alpha \odot (\mathbf{W}\triangle\mathbf{Y})$.
2. $\text{ilr}(\mathbf{W}\triangle(\alpha \odot \mathbf{Y})) = \alpha\mathbf{W}\text{ilr}(\mathbf{Y}) = \alpha\ \text{ilr}(\mathbf{W}\triangle\mathbf{Y})$.
3. $(\mathbf{W}\triangle\mathbf{Y})_{i.} = \mathcal{C}\left(\prod_{j=1}^n Y_{j1}^{W_{ij}}; \prod_{j=1}^n Y_{j2}^{W_{ij}}; \ldots; \prod_{j=1}^n Y_{jD}^{W_{ij}}\right)$, for $i = 1, \ldots, n$, where $(\mathbf{W}\triangle\mathbf{Y})_{i.}$ denotes the $i^{th}$ row of $\mathbf{W}\triangle\mathbf{Y}$.
4. Let $\mathbf{Y}_1, \mathbf{Y}_2 \in (\mathbf{S}^D)^n$, then $\mathbf{W}\triangle(\mathbf{Y}_1 \oplus \mathbf{Y}_2) = (\mathbf{W}\triangle\mathbf{Y}_1) \oplus (\mathbf{W}\triangle\mathbf{Y}_2)$.

The proof of Proposition 1 is in the appendix. Note that property 3. implies that for each individual, each component of $\mathbf{W}\triangle\mathbf{Y}$ is a weighted geometric mean of its neighboring values, weighted by the neighborhood weights.

# 3 Multivariate LAG regression model

The principle of compositional regression models is to use a transformation to send the data from the simplex to some coordinate space and to postulate a gaussian regression model in the coordinate space as in Egozcue et al. (2012). The model can then be transferred back to the simplex by the inverse transformation. In our case, the model in coordinate space must be a multivariate regression model because we have several response variables. For simplicity, we concentrate on the so-called LAG model which includes endogenous lagged variables on the right hand side of the model equations. An extension to a Durbin model would be immediate (LeSage and Pace, 2009). Since our model will be postulated in the coordinate space we choose to star all variables and parameters in Subsection 3.1. Note that the model we describe in Subsection 3.1 is not specific to CoDa.

## 3.1 Model in the log-ratio coordinates space

We consider a sample of size $n$ and assume that we have $M$ endogenous variables, hence $M$ linear regression equations ($M$ will be $D-1$ in Section 3.2). For a $n \times M$ matrix $\mathbf{A}$, we will use the same notation as in Section 2 with $\mathbf{A}_{.l}$ the $l^{th}$ column, and $\mathbf{A}_{i.}$ the $i^{th}$ row of $\mathbf{A}$ as a column vector.

Let $\mathbf{Y}^*$ be a $n \times M$ matrix of dependent variables and $\mathbf{X}$ be a $n \times K$ matrix of explanatory variables. We will allow for using a different set of explanatory variables in each equation. For this reason, we denote by $S_l^{\mathbf{Y}^*}$, $S_l^{\mathbf{X}}$, $S_l^{\mathbf{WY}^*}$ the sets of indices of the variables which appear in the $l^{th}$ equation for $\mathbf{Y}^*$, $\mathbf{X}$, $\mathbf{WY}^*$ respectively. Accordingly $\mathbf{Y}^*_{S_l^{\mathbf{Y}^*}}$, $\mathbf{X}_{S_l^{\mathbf{X}}}$, $\mathbf{Y}^*_{S_l^{\mathbf{WY}^*}}$ will denote the columns of $\mathbf{Y}^*$, $\mathbf{X}$, $\mathbf{WY}^*$ which appear in the $l^{th}$ equation.

Let $\mathbf{\Gamma}^* = (\Gamma^*_{ml})$ and $\mathbf{R}^* = (R^*_{ml}), (\ m,\ l = 1, \ldots, M)$ be $M \times M$ matrices of parameters. $\mathbf{R}^*$ contains the parameters associated to the lagged endogenous variables on the right hand side of the model equation. As in the simultaneous equations literature in econometrics, each endogenous variable may also appear in each model equation so that $\mathbf{\Gamma}^*$ contains the corresponding parameters.

Finally $\boldsymbol{\beta}^*$ is a $K \times M$ matrix of parameters for the explanatory variables. $\boldsymbol{\epsilon}^*$ denotes a $n \times M$ error matrix.

As in Kelejian and Prucha (2004), we consider the following model

$$\mathbf{Y}^*_{.l} = \sum_{m \in S_l^{\mathbf{Y}^*}} \Gamma^*_{ml}\mathbf{Y}^*_{.m} + \mathbf{X}_{S_l^{\mathbf{X}}}\boldsymbol{\beta}^*_{S_l^{\mathbf{X}}} + \sum_{m \in S_l^{\mathbf{WY}^*}} R^*_{ml}\mathbf{WY}^*_{.m} + \boldsymbol{\epsilon}^*_{.l} \tag{3}$$

Note that model (3) is written for each column of $\mathbf{Y}^*$ i.e. for each component of the composition dependent vector but the $M$ equations are linked by the covariance structure of the errors. Indeed, we assume that the errors are centered $\mathbb{E}(\boldsymbol{\epsilon}^*) = \mathbf{0}_M$ and that $\mathbb{E}(\boldsymbol{\epsilon}^*_{i.}\boldsymbol{\epsilon}^*_{j.}) = \boldsymbol{\Sigma}^*$ if $i = j$ and $\mathbf{0}$ if $i \neq j$ where $\boldsymbol{\Sigma}^*$ is a $(D-1)\times(D-1)$ covariance matrix. This means that individuals are independent but that components of a given individual have a covariance structure $\boldsymbol{\Sigma}^*$.

Kelejian and Prucha (2004) suggest and study the properties of a Spatial Two Stage Least Square (S2SLS) procedure as well as a Spatial Three Stage Least Square (S3SLS) procedure to estimate model (3). Following their suggestion, we consider $\mathbf{H}$ a subset of linearly independent columns of the $n \times 3K$ matrix $(\mathbf{X}, \mathbf{WX}, \mathbf{W^2X})$. Let $\mathbf{P}_H = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ denote the projection matrix onto the space generated by the columns of $\mathbf{H}$. For the $l^{th}$ equation, we group the variables and parameters of the right hand side into a matrix $\mathbf{Z}_l$ of variables and a vector $\boldsymbol{\delta}^*_l$ of parameters:

$$\mathbf{Z}_l = \left[\mathbf{Y}^*_{S_l^{\mathbf{Y}^*}} \quad \mathbf{X}_{S_l^{\mathbf{X}}} \quad \mathbf{WY}^*_{S_l^{\mathbf{WY}^*}}\right]; \boldsymbol{\delta}^*_l = \left[\boldsymbol{\Gamma}^{*T}_{.l} \; \boldsymbol{\beta}^{*T}_l \; \mathbf{R}^{*T}_{.l}\right]^T.$$

The S2SLS estimation method for this model then proceeds as follows for each equation (i.e. each component) separately:

1. Perform a univariate regression of each column of $\mathbf{Z}$ on $\mathbf{H}$ and compute the fitted values $\tilde{\mathbf{Z}}_l$:
$$\tilde{\mathbf{Z}}_l = \mathbf{P}_H\mathbf{Z}_l = \left[\mathbf{P}_H\mathbf{Y}^*_{S_l^{\mathbf{Y}^*}} \quad \mathbf{X}_{S_l^{\mathbf{X}}} \quad \mathbf{P}_H\mathbf{WY}^*_{S_l^{\mathbf{WY}^*}}\right].$$
2. Perform a univariate regression of $\mathbf{Y}^*_{.l}$ on $\tilde{\mathbf{Z}}_l$:
$$\tilde{\boldsymbol{\delta}}^*_l = (\tilde{\mathbf{Z}}_l^T\tilde{\mathbf{Z}}_l)^{-1}\tilde{\mathbf{Z}}_l^T\mathbf{Y}^*_{.l}.$$

At the end of step 2, we can calculate the residuals by

$$\tilde{\boldsymbol{\epsilon}}^*_{.l} = \mathbf{Y}^*_{.l} - \hat{\mathbf{Y}}^*_{.l} = \mathbf{Y}^*_{.l} - \mathbf{Z}_l\tilde{\boldsymbol{\delta}}^*_l,$$

and get an estimate of the covariance matrix $\boldsymbol{\Sigma}^*$ with

$$\tilde{\boldsymbol{\Sigma}}^*_{ml} = \frac{\tilde{\boldsymbol{\epsilon}}^{*T}_{.m}\tilde{\boldsymbol{\epsilon}}^*_{.l}}{n}.$$

Until now, the covariance structure between equations has not been taken into account and the Three Stage Least Square (S3SLS) method is supposed to correct for this. To write the expression of the S3SLS estimator, we need to vectorize $\mathbf{Y}^*$ (by stacking the columns of $\mathbf{Y}$) resulting in $\mathbf{y}^* = \text{vec}(\mathbf{Y}^*)$ and to write the explanatory matrices as follows

$$\mathcal{Z} = \begin{bmatrix} \mathbf{Z}_1 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \mathbf{Z}_M \end{bmatrix}, \; \tilde{\mathcal{Z}} = \begin{bmatrix} \mathbf{P}_H\mathbf{Z}_1 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \mathbf{P}_H\mathbf{Z}_M \end{bmatrix}.$$

We then get a corrected estimator $\hat{\boldsymbol{\delta}}^*$ of $\boldsymbol{\delta}^*$

$$\hat{\boldsymbol{\delta}}^* = (\tilde{\mathcal{Z}}^T(\tilde{\boldsymbol{\Sigma}}^{*-1} \otimes \mathbf{I}_n)\mathcal{Z})^{-1}\tilde{\mathcal{Z}}^T(\tilde{\boldsymbol{\Sigma}}^{*-1} \otimes \mathbf{I}_n)\mathbf{y}^* \tag{4}$$

It is known that if the matrix $\mathbf{R}^*$ is not diagonal, the S2SLS $\tilde{\boldsymbol{\delta}}^*$ and S3SLS $\hat{\boldsymbol{\delta}}^*$ estimators are identical (see Greene, 2003, p. 488).

In the application of Section 5, we consider a slightly more general model in which we include compositional variables among the explanatory (see for example Filzmoser et al., 2018). The additional complexity is the same as for a non-spatial model hence for the sake of simplicity we did not consider this extra layer in this section.

### 3.2 Writing the LAG regression model in the simplex space

Starting now with a sample of compositional vectors $\mathbf{Y}$ in $\mathbf{S}^D$, and given an ilr transformation, we postulate a model like (3) for the ilr coordinates of $\mathbf{Y}$. Applying the ilr inverse transformation to each of the equations of model (3) with $M = D - 1$ and using Proposition 1, we easily get that the system of equations (3) is equivalent to the system

$$\mathbf{Y}_{i.} = \mathbf{R}^T \boxdot (\mathbf{W} \triangle \mathbf{Y})_{i.} \bigoplus_{k=1}^{K} \mathbf{X}_{ik} \odot \boldsymbol{\beta}_k \oplus \boldsymbol{\Gamma}^T \boxdot \mathbf{Y}_{i.} \oplus \boldsymbol{\epsilon}_{i.} \tag{5}$$

where $\mathbf{R}$ is a $D \times D$ matrix of parameters and where the model is now written at the individual level for all components simultaneously whereas in (3) it was at the component level for all individuals simultaneously. A more global way of writing the model for the whole matrix $\mathbf{Y}$ would be the following

$$\mathbf{Y} = (\mathbf{W} \triangle \mathbf{Y}) \boxdot \mathbf{R} \oplus \mathbf{X} \odot \boldsymbol{\beta} \oplus \mathbf{Y} \boxdot \Gamma \oplus \boldsymbol{\epsilon}, \tag{6}$$

where for a $n \times D$ matrix $\mathbf{U}$ whose rows are simplex valued and a $D \times D$ matrix $\mathbf{B}$, the product $\mathbf{U} \boxdot \mathbf{B}$ is an extension of the matrix product understood as the $n \times D$ matrix whose rows are the $\mathbf{B}^T \boxdot \mathbf{U}_{i.}$ vectors.

One can write relationships between parameters in coordinate space and parameters in the simplex. The classical relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ remains the same as in non spatial models (see for example Filzmoser et al., 2018)

$$\boldsymbol{\beta}_k = \text{ilr}^{-1}(\boldsymbol{\beta}_k^*) = \mathcal{C}(\exp(\mathbf{V}_D \boldsymbol{\beta}_k^*)).$$

Considering $\mathbf{R}^T \boxdot (\mathbf{W} \triangle \mathbf{Y})_{i.}$ and $\boldsymbol{\Gamma}^T \boxdot \mathbf{Y}_{i.}$ as compositional explanatory, the relationships between $\mathbf{R}$ and $\mathbf{R}^*$ and between $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}^*$ are as follows (see Chen et al., 2017)

$$\mathbf{R} = \mathbf{V}_D \mathbf{R}^* \mathbf{V}_D^T \qquad \text{and} \qquad \boldsymbol{\Gamma} = \mathbf{V}_D \boldsymbol{\Gamma}^* \mathbf{V}_D^T. \tag{7}$$

Equations (7) also allow to establish the link between the parameters in coordinate space for two different ilr transformations. Coming back to the simplex after fitting the model allows to get rid of the possible arbitrary choice of ilr transformation. These relationships are true for population parameters and the next question is whether they still hold for the estimated parameters. Indeed the result holds because the two steps of S2SLS are based on ordinary least squares and we know that this method preserves the relationship for estimated parameters (see e.g. Nguyen, 2019).

## 4 Simulation

A simulation study of the performance of the S2SLS and S3SLS methods can be found in Das et al. (2003) but it is restricted to the case of a single dependent variable. For this reason, we now investigate by simulation the properties of the estimators $\boldsymbol{\beta}^*$, $\mathbf{R}^*$ and $\boldsymbol{\Sigma}^*$ of the S2SLS and S3SLS methods in the multivariate spatial autoregressive model. We consider the $n = 283$ cantons of the Occitanie region in France with a neighborhood structure based on 10 nearest neighbors. All the **R** code, graphs and tables are gathered in the supplementary material available online.[1]

For a number of replications $N = 1\ 000$, we simulate three explanatory variables $\mathbf{X_1}, \mathbf{X_2}$ and $\mathbf{X_3}$ following the Gaussian distributions $\mathcal{N}(0, 9)$, $\mathcal{N}(0, 6)$ and $\mathcal{N}(0, 9)$ respectively. When simulating the two dependent variables, we include all explanatory variables $\mathbf{X_1}, \mathbf{X_2}$ and $\mathbf{X_3}$ in each of the two equations.

[1] See `http://www.thibault.laurent.free.fr/code/spatial_coda/`

The parameter $\boldsymbol{\beta}^*$, the covariance matrix $\boldsymbol{\Sigma}^*$ and the matrix $\mathbf{R}^*$ are respectively assigned the following values

$$\boldsymbol{\beta}^* = \begin{bmatrix} \beta_{01}^* & \beta_{02}^* \\ \beta_{11}^* & \beta_{12}^* \\ \beta_{21}^* & \beta_{22}^* \\ \beta_{31}^* & \beta_{32}^* \end{bmatrix} = \begin{bmatrix} +3 & -3 \\ +2 & -3 \\ +1 & -2 \\ -1 & +3 \end{bmatrix}; \; \boldsymbol{\Sigma}_{\bar{d}}^* = \begin{bmatrix} \sigma_{11}^{2*} & \sigma_{12}^* \\ \sigma_{21}^* & \sigma_{22}^{2*} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.09 \\ 0.09 & 0.1 \end{bmatrix}; \; \mathbf{R}_{\bar{d}}^* = \begin{bmatrix} R_{11}^* & R_{12}^* \\ R_{21}^* & R_{22}^* \end{bmatrix} = \begin{bmatrix} 0.5 & 0.6 \\ 0.4 & 0.3 \end{bmatrix}$$

Alternative diagonal matrices for $\boldsymbol{\Sigma}^*$ and $\mathbf{R}^*$ are also considered

$$\boldsymbol{\Sigma}_d^* = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.1 \end{bmatrix}; \; \mathbf{R}_d^* = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.3 \end{bmatrix}.$$

It is important to note that such a model is defined primarily in the simplex and has different representations in coordinate space according to the choice of ilr transformation. Therefore, constraining the matrix $\mathbf{R}^*$ to be diagonal for a given ilr transformation does not imply that, for the same model in the simplex, the matrix $\mathbf{R}^*$ would be diagonal with a different choice of ilr.

Note that the results are not too sensitive to the simulation framework except for the estimates of the error variances when the noise to signal ratio becomes too large. We consider four data generating processes (DGP) respectively denoted by $\boldsymbol{\Sigma}_{\bar{d}}^* \mathbf{R}_{\bar{d}}^*$, $\boldsymbol{\Sigma}_{\bar{d}}^* \mathbf{R}_d^*$, $\boldsymbol{\Sigma}_d^* \mathbf{R}_{\bar{d}}^*$ and $\boldsymbol{\Sigma}_d^* \mathbf{R}_d^*$ according to the choice of matrices $\boldsymbol{\Sigma}^*$ and $\mathbf{R}^*$. For each DGP, we calculate a Monte Carlo performance measure of the estimators proposed in Section 3 and for the MLE estimators in the diagonal case. The performance is measured by the relative root mean squared error (RRMSE), which is defined for an estimator $\hat{\theta}$ of a parameter $\theta$ by:

$$\text{RRMSE}(\hat{\theta}) = \frac{\text{RMSE}(\hat{\theta})}{|\theta|} \quad \text{with} \quad \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \hat{\theta}^{(i)} - \theta \right)^2}.$$

Table 1 presents the RRMSE for the S2SLS method, S3SLS and ML methods for all DGPs. We did not report the relative bias in Table 1 because its value is quite similar to the RRMSE showing that the bias dominates the error. The percentage of error is generally small with a maximum of 18.16% for the variance parameters and values less than 2.15% for the other parameters. For all DGPs, the largest differences between the estimates occur for the estimation of the intercepts and for the estimation of the $\mathbf{R}^*$ matrix and these differences are small in all cases. We have done more simulations for other values of $\boldsymbol{\Sigma}^*$ and $\mathbf{R}^*$. The code is presented in the supplementary material and yields very similar results. Concerning the $\boldsymbol{\beta}$ and $\mathbf{R}$ estimates, only the S2SLS results are reported since S3SLS yields exactly the same results for the first two DGPs as proved by the theory. Concerning the variance estimates, note that there is no difference in its computation for the S2SLS and S3SLS. Finally, if we compare the three estimates in the case of DGP $\boldsymbol{\Sigma}_d^* \mathbf{R}_d^*$, we can see that the results are quite close showing that S2SLS is a practical alternative to maximum likelihood in the framework of this model.

# 5 Application to political economics

Vote share data of the 2015 French departmental election of the Occitanie region in France are collected from the Cartelec website[2]. Corresponding socio-economic data (for 2014) are downloaded from the INSEE website[3]. The number of political parties presenting candidates at that election is higher than 15. However for simplicity reasons and due to the particular interest in the extreme right party in France,

[2]https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/

[3]https://www.insee.fr/fr/statistiques

Table 1: The RRMSE (in %) for all DGPs and parameters

| **Parameters** | **RRMSE(%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\mathbf{\Sigma}_{\bar{d}}^*\mathbf{R}_{\bar{d}}^*$ | $\mathbf{\Sigma}_{d}^*\mathbf{R}_{\bar{d}}^*$ | $\mathbf{\Sigma}_{\bar{d}}^*\mathbf{R}_{d}^*$ | | $\mathbf{\Sigma}_{d}^*\mathbf{R}_{d}^*$ | | |
| | S2SLS | S2SLS | S2SLS | S3SLS | S2SLS | S3SLS | MLE |
| $\beta_{01}^*$ | 1.92 | 1.91 | 2.13 | 2.13 | 2.15 | 2.15 | 2.15 |
| $\beta_{11}^*$ | 0.28 | 0.27 | 0.29 | 0.29 | 0.29 | 0.29 | 0.29 |
| $\beta_{21}^*$ | 0.82 | 0.81 | 0.77 | 0.77 | 0.83 | 0.83 | 0.83 |
| $\beta_{31}^*$ | 0.62 | 0.63 | 0.59 | 0.59 | 0.60 | 0.60 | 0.60 |
| $\beta_{02}^*$ | 0.75 | 0.70 | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| $\beta_{12}^*$ | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |
| $\beta_{22}^*$ | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |
| $\beta_{32}^*$ | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 |
| $\mathrm{R}_{11}^*$ | 1.04 | 1.07 | 0.90 | 0.89 | 0.92 | 0.92 | 0.92 |
| $\mathrm{R}_{12}^*$ | 0.51 | 0.50 | - | - | - | - | - |
| $\mathrm{R}_{21}^*$ | 0.52 | 0.52 | - | - | - | - | - |
| $\mathrm{R}_{22}^*$ | 0.41 | 0.39 | 0.39 | 0.38 | 0.40 | 0.40 | 0.40 |
| $\sigma_{11}^{2*}$ | 8.29 | 8.45 | 8.39 | 8.39 | 8.80 | 8.80 | 8.87 |
| $\sigma_{12}^*$ | 18.01 | - | 18.15 | 18.16 | - | - | - |
| $\sigma_{21}^*$ | 18.01 | - | 18.15 | 18.16 | - | - | - |
| $\sigma_{22}^{2*}$ | 8.31 | 8.54 | 8.21 | 8.21 | 8.49 | 8.49 | 8.54 |

we have aggregated them into three main components: Left, Right and Extreme-Right[4]. The dependent variable is thus a compositional variable which contains the vote shares of Left, Right and Extreme Right party. Cantons with at least one missing value on one of the components of the dependent vector have been eliminated resulting in $n = 207$ cantons in the final dataset. The loss of information is substantial and may cause bias. However the main focus of this paper is to illustrate the methodology. We use the following contrast matrix for $D = 3$ (see e.g. Pawlowsky-Glahn et al., 2015, p. 40)

$$\mathbf{V}_3 = \begin{bmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}.$$

This particular matrix defines the following ilr coordinates

$$\mathrm{ilr}_1(\mathbf{x}) = \frac{1}{\sqrt{6}}(2\log x_1 - \ln x_2 - \log x_3) = \frac{2}{\sqrt{6}}\log\frac{x_1}{\sqrt{x_2 x_3}}$$
$$\mathrm{ilr}_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\log x_2 - \log x_3) = \frac{1}{\sqrt{2}}\log\frac{x_2}{x_3}.$$

With this choice, the first ilr coordinate opposes the Left wing to the geometric mean of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party.
Our explanatory variables, presented in Table 2, include both compositional and classical variables. For the three compositional variables, Diploma, Employment and Age, the log-ratio coordinates have been calculated using contrast matrices built from sequential binary partitions (see Nguyen, 2019, for details). The categories of these variables are as follows

- Employment has five levels: AZ (agriculture, fisheries), BE (manufacturing industry, mining industry and others), FZ (construction), GU (business, transport and services) and OQ (public administration, teaching, human health),

[4]For more details, see https://fr.wikipedia.org/wiki/Elections_départementales_francaises_de_2015

Table 2: Data description.

| Variable name | Description |
|---|---|
| Vote share | Left(L), Right(R), Extreme Right(XR) |
| Diploma | <BAC, BAC, SUP |
| Employment | AZ, BE, FZ, GU, OQ |
| Age | Age_1840, Age_4064, Age_65 |
| unemp | Unemployment rate |
| nbvoter | Number of voters |
| income | Proportion of people who pay income tax |

- Diploma has three levels: <BAC for people with at most some secondary education, BAC for people with at least some secondary education and at most a high school diploma, and SUP for people with a university diploma,
- Age has three levels: Age_1840 for people from 18 to 40 years old, Age_4064 for people from 40 to 64 years old, and Age_65 for elderly.

An additional variable measuring the number of voters in each canton has been included to take into account a potential size effect.

This data set has been analyzed in Nguyen and Laurent (2019) without taking into account the spatial structure and at a different spatial scale. We use model (3) in the coordinate space with $\mathbf{\Gamma}^* = 0$. Indeed, the reason for including spatially lagged dependent variable in the equations is for taking into account the spatial dependence and this justifies terms such as $\sum_{m \in S_l^{\mathbf{WY}^*}} R^*_{ml} \mathbf{WY}^*_{.m}$ in model (3). But in our case, there is no economic reason for introducing terms such as $\sum_{m \in S_l^{\mathbf{Y}^*}} \Gamma^*_{ml} \mathbf{Y}^*_{.m}$ on the right hand side of the equations in the coordinate space. There is also no reason for assuming that the $\mathbf{R}^*$ matrix is diagonal for this particular choice of ilr transformation and therefore, we do not impose this constraint. One important consequence of this choice is that, as we mentioned before, S2SLS yield the same results as S3SLS so that we carry out the S2SLS method from Section 3 for estimating the parameters.

First of all, for comparing the independence model to the spatial dependence one, we compute Moran test statistic of the residuals as well as the LMlag test statistics (separately for each ilr) and these indicate that the LAG model is preferable. Then looking at the size effect in Table 3, the number of voters is significant in both spatial and non-spatial model in the 2nd component (indeed there is some heterogeneity in the distribution of the number of voters at the canton level). The spatial dependence parameters (elements of the matrix $\mathbf{R}^*$) are significant on the diagonal showing that a spatial dependence phenomenon is present in this data. The sign and significance of most $\boldsymbol{\beta}$ parameters are very comparable, except in few cases (the unemployment rate, proportion of people who pay income tax, diploma and age variables) for which the significance changes from one ilr coordinate to the other. Further interpretations of the model parameters in the spatial model would go through the impacts computations as in LeSage and Pace (2009). But one would have to develop this tool in a multivariate framework and this is out of the scope of the present work. When we will be able to do so, this spatial LAG model will allow to evaluate spillover effects across cantons.

# 6 Conclusion

Motivated by an example in political economics, we develop a simultaneous spatial autoregressive model for compositional data combining the simultaneous systems of spatially interrelated cross sectional equa-

Table 3: Multivariate independent and spatial regression models with compositional and classical explanatory variables

| | *Independence model* | | *Spatial dependence model* | |
|---|---|---|---|---|
| | $\mathbf{Y}^*_{.1}$ | $\mathbf{Y}^*_{.2}$ | $\mathbf{Y}^*_{.1}$ | $\mathbf{Y}^*_{.2}$ |
| Constant | $-4.12(1.17)^{***}$ | $-4.59(0.58)^{***}$ | $-2.98(1.16)^{**}$ | $-2.47(0.6)^{***}$ |
| diplome_ilr1 | $-1.29(0.5)^{*}$ | $-0.3(0.25)$ | $-0.72(0.46)$ | $-0.47(0.24)^{*}$ |
| diplome_ilr2 | $-0.02(0.61)$ | $-0.96(0.3)^{**}$ | $+0.07(0.53)$ | $-0.58(0.27)^{*}$ |
| employ_ilr1 | $-0.18(0.14)$ | $-0.1(0.07)$ | $-0.11(0.12)$ | $-0.12(0.06)$ |
| employ_ilr2 | $+0.5(0.16)^{**}$ | $-0.02(0.08)$ | $+0.35(0.15)^{*}$ | $+0.04(0.08)$ |
| employ_ilr3 | $-0.21(0.11)$ | $+0(0.06)$ | $-0.18(0.1)$ | $+0.08(0.05)$ |
| employ_ilr4 | $+0.21(0.06)^{***}$ | $+0.02(0.03)$ | $+0.13(0.05)^{**}$ | $-0.02(0.03)$ |
| age_ilr1 | $-1.17(0.38)^{**}$ | $+1.02(0.19)^{***}$ | $-0.9(0.4)^{*}$ | $+0.31(0.21)$ |
| age_ilr2 | $+0.5(0.31)$ | $-1.27(0.16)^{***}$ | $+0.66(0.32)^{*}$ | $-0.64(0.17)^{***}$ |
| unemp | $+0.22(2.36)$ | $+8.93(1.18)^{***}$ | $-0.88(2.86)$ | $+1.82(1.48)$ |
| income | $+4.45(0.91)^{***}$ | $+1.28(0.45)^{**}$ | $+3.39(0.82)^{***}$ | $+0.71(0.42)$ |
| nbvoter | $+0.04(0.09)$ | $+0.22(0.04)^{***}$ | $+0.07(0.08)$ | $+0.15(0.04)^{***}$ |
| $R^*_{.1}$ | - | - | $+0.65(0.16)^{***}$ | $-0(0.08)$ |
| $R^*_{.2}$ | - | - | $+0.18(0.18)$ | $+0.63(0.09)^{***}$ |
| $\Sigma^*_{.1}$ | $+0.21$ | $-0.02$ | $+0.15$ | $-0.03$ |
| $\Sigma^*_{.2}$ | $-0.02$ | $+0.05$ | $-0.03$ | $+0.04$ |
| Nb. Obs. | 207 | 207 | 207 | 207 |
| Moran's $I$ test | $7.98^{***}$ | $5.26^{***}$ | - | - |
| LMlag | $55.01^{***}$ | $38.72^{***}$ | - | - |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

tions of Kelejian and Prucha (2004) and the compositional regression models (Filzmoser et al., 2018, e.g.). We propose an implementation using spatial two-stage and three-stage methods which are easy to implement.

There are several directions we could consider to go further in this framework. We could first of all consider alternative estimation methods. For example partial least squares procedures for the Spatial LAG model have been proposed in Wang et al. (2019) but for a single dependent variable. Similarly and with the same restriction, Spatial regression trees are developed for the LAG model in Wagner and Zeileis (2019). In a different direction, the aggregation of political parties in three blocks could be reconsidered. On the one hand, this aggregation avoids the zero problem due to the absence of some parties in some cantons but on the other hand it results in an information loss: imputation methods could be used to solve this as in Palarea-Albaladejo and Martín-Fernández (2015). An extension to multivariate spatial error models (SEM) does not seem too complex but would need an additional step in the S3SLS procedure as in Kelejian and Prucha (2004). Finally, concerning the interpretation of the parameters of the spatial model, two possibilities have to be explored further. The first one is the interpretation in coordinate space. In a spatial model, interpretation of the parameters goes through the computation of impacts but to our knowledge, this has never been done in the multivariate LAG model. The second one is to obtain interpretations of the parameters in the simplex as in Morais et al. (2018) which implies being able to write the reduced form of the model in the simplex space.

# 7 Acknowledgements

We acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010

# 8 Appendix

***Proof of Proposition 1.*** Let $\mathbf{Y} \in \mathbf{S}^D$, $\mathbf{W}$ be a $n \times n$ matrix, $\alpha$ be a scalar, and let $(\mathbf{W} \triangle \mathbf{Y})_{i.}$ denotes the $i^{th}$ row of $\mathbf{W} \triangle \mathbf{Y}$, $i, j = 1, \ldots, n,\ l, m = 1, \ldots, D$.
1. $\text{ilr}(\mathbf{W} \triangle (\alpha \odot \mathbf{Y})) = \text{ilr}(\mathbf{W} \triangle \mathbf{Y}^\alpha) = \alpha \mathbf{W} \ln^T(\mathbf{Y}) \mathbf{V}_D = \alpha \mathbf{W} \text{ilr}(\mathbf{Y}) = \alpha\ \text{ilr}(\mathbf{W} \triangle \mathbf{Y})$.
2. We have

$$\text{ilr}(\mathbf{W} \triangle (\alpha \odot \mathbf{Y})) = \text{ilr}(\mathbf{W} \triangle \mathbf{Y}^\alpha) = \alpha \text{ilr}(\mathbf{W} \triangle \mathbf{Y})$$

then

$$\text{ilr}^{-1}(\text{ilr}(\mathbf{W} \triangle (\alpha \odot \mathbf{Y}))) = \text{ilr}^{-1}(\alpha \text{ilr}(\mathbf{W} \triangle \mathbf{Y})) = \alpha \odot (\mathbf{W} \triangle \mathbf{Y})$$

Thus,

$$\mathbf{W} \triangle (\alpha \odot \mathbf{Y}) = \alpha \odot (\mathbf{W} \triangle \mathbf{Y}).$$

3. We have

$$\begin{aligned}(\mathbf{W} \triangle \mathbf{Y})_{i.} &= \text{ilr}^{-1}(\text{ilr}((\mathbf{W} \triangle \mathbf{Y})_{i.})) \\ &= \text{ilr}^{-1}(\mathbf{W} \text{ilr}(\mathbf{Y}))_{i.} \\ &= \mathcal{C}(\exp(\mathbf{W} \text{ilr}(\mathbf{Y}) \mathbf{V}_\mathbf{D}^\mathbf{T})_{i.})\end{aligned}$$

where

$$(\mathbf{W} \text{ilr}(\mathbf{Y}) \mathbf{V}_\mathbf{D}^\mathbf{T})_{i.} = \left( \ln \prod_{j=1}^{n} \left( \frac{Y_{j1}}{\prod_{l=1}^{D} Y_{jl}^{1/D}} \right)^{W_{ij}} ; \ln \prod_{j=1}^{n} \left( \frac{Y_{j2}}{\prod_{l=1}^{D} Y_{jl}^{1/D}} \right)^{W_{ij}} ; \ldots ; \ln \prod_{j=1}^{n} \left( \frac{Y_{jD}}{\prod_{l=1}^{D} Y_{jl}^{1/D}} \right)^{W_{ij}} \right).$$

Thus,

$$\begin{aligned}(\mathbf{W} \triangle \mathbf{Y})_{i.} &= \mathcal{C} \left( \prod_{j=1}^{n} \left( \frac{Y_{j1}}{\prod_{l=1}^{L} Y_{jl}^{1/D}} \right)^{W_{ij}} ; \prod_{j=1}^{n} \left( \frac{Y_{j2}}{\prod_{l=1}^{L} Y_{jl}^{1/D}} \right)^{W_{ij}} ; \ldots ; \prod_{j=1}^{n} \left( \frac{Y_{jD}}{\prod_{l=1}^{L} Y_{jl}^{1/D}} \right)^{W_{ij}} \right) \\ &= \mathcal{C} \left( \prod_{j=1}^{n} Y_{j1}^{W_{ij}} ; \prod_{j=1}^{n} Y_{j2}^{W_{ij}} ; \ldots ; \prod_{j=1}^{n} Y_{jD}^{W_{ij}} \right)\end{aligned}$$

4. Let $\mathbf{Y}_1, \mathbf{Y}_2 \in (\mathbf{S}^D)^n$, and let $\mathbf{Y}_1^* = \text{ilr}(\mathbf{Y}_1)$, $\mathbf{Y}_2^* = \text{ilr}(\mathbf{Y}_2)$, then

$$\begin{aligned}\text{ilr}^{-1}(\text{ilr}(\mathbf{W} \triangle (\mathbf{Y}_1 \oplus \mathbf{Y}_2))) &= \text{ilr}^{-1}(\mathbf{W}(\mathbf{Y}_1^* + \mathbf{Y}_2^*)) \\ &= \text{ilr}^{-1}(\mathbf{W}\mathbf{Y}_1^* + \mathbf{W}\mathbf{Y}_2^*) \\ &= \text{ilr}^{-1}(\mathbf{W} \text{ilr}(\mathbf{Y}_1) + \mathbf{W} \text{ilr}(\mathbf{Y}_2)) \\ &= \text{ilr}^{-1}(\text{ilr}(\mathbf{W} \triangle \mathbf{Y}_1) + \text{ilr}(\mathbf{W} \triangle \mathbf{Y}_2))) \\ &= \text{ilr}^{-1}(\text{ilr}(\mathbf{W} \triangle \mathbf{Y}_1))) + \text{ilr}^{-1}(\text{ilr}(\mathbf{W} \triangle \mathbf{Y}_2))) \\ &= (\mathbf{W} \triangle \mathbf{Y}_1) \oplus (\mathbf{W} \triangle \mathbf{Y}_2)\end{aligned}$$

Thus,

$$\mathbf{W} \triangle (\mathbf{Y}_1 \oplus \mathbf{Y}_2) = (\mathbf{W} \triangle \mathbf{Y}_1) \oplus (\mathbf{W} \triangle \mathbf{Y}_2)$$

□

# References


Aitchison, J. (1985). A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):136–146.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London. Reprinted in 2003 with additional material by the Blackburn Press.

Billheimer, D., Cardoso, T., Freeman, E., Guttorp, P., Ko, H.-W., and Silkey, M. (1997). Natural variability of benthic species composition in the delaware bay. *Environmental and Ecological Statistics*, 4(2):95–115.

Billheimer, D., Guttorp, P., and Fagan, W. F. (1998). Statistical analysis and interpretation of discrete compositional data. *National Center for Statistics and the Environment (NRCSE) Technical Report NRCSE-TRS*, 11:39.

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., and Pebesma, E. J. (2008). *Applied spatial data analysis with R*. Springer-Verlag.

Borghesi, C. and Bouchaud, J.-P. (2010). Spatial correlations in vote statistics: a diffusive field model for decision-making. *The European Physical Journal B-Condensed Matter and Complex Systems*, 75(3):395–404.

Chen, J., Zhang, X., and Li, S. (2017). Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics*, 44(12):2270–2285.

Das, D., Kelejian, H. H., and Prucha, I. R. (2003). Finite sample properties of estimators of spatial autoregressive models with autoregressive disturbances. *Papers in Regional Science*, 82(1):1–26.

Egozcue, J. J., Daunis-I-Estadella, J., Pawlowsky-Glahn, V., Hron, K., and Filzmoser, P. (2012). Simplicial regression. the normal model. *Journal of Applied Probability and Statistics*, 6(1-2):87–108.

Filzmoser, P. (2020). *StatDA: Statistical Analysis for Environmental Data*. R package version 1.7.4.

Filzmoser, P., Hron, K., and Reimann, C. (2010). The bivariate statistical analysis of environmental (compositional) data. *Science of The Total Environment*, 408 19:4230–4238.

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied Compositional Data Analysis*. Springer-Verlag.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education, fifth edition.

Katz, J. N. and King, G. (1999). A statistical model for multiparty electoral data. *American Political Science Review*, 93(1):15–32.

Kelejian, H. H. and Prucha, I. R. (2004). Estimation of simultaneous systems of spatially interrelated cross sectional equations. *Journal of Econometrics*, 118(1-2):27–50.

Leininger, T. J., Gelfand, A. E., Allen, J. M., and Silander, J. A. (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):314–334.

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Mardia, K. V. (1988). Multi-dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284.

Martins, A. B. T., Bonat, W. H., and Ribeiro, P. J. (2016). Likelihood analysis for a class of spatial geostatistical compositional models. *Spatial Statistics*, 17:121–130.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). The principle of working on coordinates. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis*, pages 31–42. John Wiley & Sons.

Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47(5):1–25.

Nguyen, T. H. A. (2019). *Contribution to the statistical analysis of compositional data with an application to political economy*. PhD thesis, University of Toulouse Capitole.

Nguyen, T. H. A. and Laurent, T. (2019). Coda methods and the multivariate student distribution with an application to political economy, preprint. *Preprint*.

Overmars, K. P., De Koning, G. H. J., and Veldkamp, A. (2003). Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, 164(2-3):257–270.

Palarea-Albaladejo, J. and Martín-Fernández, J. A. (2015). zCompositions - R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96.

Pawlowsky, V. and Burger, H. (1992). Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24(6):675–691.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2016). Spatial analysis of compositional data: a historical review. *Journal of Geochemical Exploration*, 164:28–32.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Pirzamanbein, B., Lindström, J., Poska, A., and Gaillard, M.-J. (2018). Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *Spatial Statistics*, 24:14–31.

Rubio, R. H., Costa, J. F. C. L., and Bassani, M. A. A. (2016). A geostatistical framework for estimating compositional data avoiding bias in back-transformation. *Rem: Revista Escola de Minas*, 69(2):219–226.

Salazar, E., Giraldo, R., and Porcu, E. (2015). Spatial prediction for infinite-dimensional compositional data. *Stochastic Environmental Research and Risk Assessment*, 29(7):1737–1749.

Wagner, M. and Zeileis, A. (2019). Heterogeneity and spatial dependence of regional growth in the eu: A recursive partitioning approach. *German Economic Review*, 20(1):67–82.

Wang, H., Gu, J., Wang, S., and Saporta, G. (2019). Spatial partial least squares autoregression: Algorithm and applications. *Chemometrics and Intelligent Laboratory Systems*, 184:123–131.

Yoshida, T. and Tsutsumi, M. (2018). On the effects of spatial relationships in spatial compositional multivariate models. *Letters in Spatial and Resource Sciences*, 11(1):57–70.