WORKING PAPERS

N° 1162

October 2021

“Covariates impacts in spatial autoregressive models for compositional data”

Thibault Laurent, Christine Thomas-Agnan and Anne Ruiz-Gazen

Toulouse School of Economics

Covariates impacts in spatial autoregressive models for compositional data

Thibault Laurent[1], Christine Thomas-Agnan[2*] and Anne Ruiz-Gazen[2]

[1]Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, 31000, France.
[2]Toulouse School of Economics, University of Toulouse Capitole, Toulouse, 31000, France.

*Corresponding author(s). E-mail(s): christine.thomas@tse-fr.eu; Contributing authors: thibault.laurent@tse-fr.eu; anne.ruiz-gazen@tse-fr.eu;

Abstract

Spatial simultaneous autoregressive models have been adapted to model data with both a geographic and a compositional nature. Interpretation of parameters in such a model is intricate. Indeed, when the model involves a spatial lag of the dependent variable, this interpretation must focus on the so-called impacts rather than on parameters and when moreover the dependent variable of this model is of a compositional nature, this interpretation should be based on elasticities or semi-elasticities. Combining the two difficulties, we provide exact formulas for the evaluation of these elasticity-based impact measures which have been only approximated so far in some applications. We also discuss their decomposition into direct and indirect impacts taking into account the compositional nature of the dependent variable. Finally, we also propose more local summary measures as exploratory tools that we illustrate on a toy data set and a case study.

Keywords: Elasticities, direct impact, local impact, indirect impact, semi-elasticities, simplicial regression

JEL Classification: C10 , C39 , C46 , C65 , M31 , Q15

1 Introduction

Data with share vectors exhibiting spatial dependence can be found in many applications such as political science or land use studies, see for example Katz and King (1999) and Yoshida and Tsutsumi (2018). Modelling these vectors with covariates requires complex multivariate regression models that can accomodate their spatial and compositional dimensions. Focusing on the model introduced in Nguyen, Thomas-Agnan, Laurent, and Ruiz-Gazen (2021), we will show how to assess the impact of covariates.

Impact of covariates in a classical (non spatial and non compositional) regression model is based on the parameters of the model as follows: if a given covariate $\mathbf{X}$ increases by a given additive amount δ, all things equal, it results in an additive increase of the expected dependent variable $\mathbf{Y}$ equal to the product of δ by the corresponding parameter β_X of that covariate. This results from the fact that the expected value of the dependent variable, $\mathbb{E}(\mathbf{Y} \mid \mathbf{X})$, is the sum of the linear term $\beta_X \mathbf{X}$, and other terms independent of this particular variable $\mathbf{X}$.

If one now considers a spatial simultaneous autoregressive model, this simple interpretation gets more complex due to the fact that the link between the expected value of the dependent variable and the covariate of interest then involves a so-called filter matrix. For spatial autoregressive models with a lag component, this type of interpretation has been introduced by LeSage and Pace (2009).

In a compositional regression model, some variables (dependent and/or independent) may be vectors of parts conveying relative information on some parts of a total abundance characteristic. In such a model with compositional variables possibly on both sides of the regression equation, Morais and Thomas-Agnan (2021) show that an interpretation of variations in the simplex is possible through the use of elasticities or semi-elasticities depending on the particular type of model. However, their derivation is for the case of independence between individuals and has to be extended now in the framework of cross-correlations between statistical units (locations here).

Finally, in a model of the spatial autoregressive type for compositional data, such as the one introduced in Nguyen et al. (2021), the two difficulties are present and one needs to combine the two approaches. Another problem arises in the compositional framework when trying to define the decomposition of these impact measures because the traditional decomposition into direct and indirect impacts is based on the classical addition in $\mathbb{R}$.

In Section 2, we review the basic tools of compositional regression models. Section 3 recalls the impacts computations in the spatial autoregressive models with a single dependent variable and in the compositional regression models. In Section 4, we extend the exact evaluation of the impacts to the spatial regression framework with a compositional dependent variable, which is by nature a multivariate framework. Section 5 addresses the question of their decomposition into direct and indirect impacts and illustrates their computations and interpretations on a toy example. Finally, Section 6 takes up again the political science case study presented in Nguyen et al. (2021), adding the decomposition and interpretation of a particular covariate in terms of effects on the vote shares vector.

2 Simplex operations reminder

Let us briefly recall some tools for working with compositional data. A D-composition $\mathbf{u}$ is a vector of D parts which can be represented in the so-called simplex space

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

where ^{T} is the transposition operator. For any vector $\mathbf{w} \in \mathbb{R}^{+D}$, the closure operation is defined by $\mathcal{C}(\mathbf{w}) = \left(w_1 / \sum_{m=1}^{D} w_m, \cdots, w_D / \sum_{m=1}^{D} w_m \right)$. The perturbation and the powering operations are given, for a scalar λ and simplex vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{S}^D$, by

$$\mathbf{u} \oplus \mathbf{v} = \mathcal{C}(u_1 v_1, \ldots, u_D v_D), \quad \text{and} \quad \lambda \odot \mathbf{u} = \mathcal{C}(u_1^\lambda, \ldots, u_D^\lambda).$$

For a $D \times (D-1)$ contrast matrix $\mathbf{V}$ (see e.g. Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015), one can define an isometric log-ratio transformation traditionally called ilr. As advocated by Martín-Fernández (2019), we will rather use the name olr (orthogonal log ratio) for this transformation, defined by: $\mathbf{u}^* = \text{olr}(\mathbf{u}) = \mathbf{V}^T \ln(\mathbf{u})$, where the logarithm of $\mathbf{u} \in \mathbf{S}^D$ is understood componentwise. The inverse transformation is given by: $\mathbf{u} = \text{olr}^{-1}(\mathbf{u}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{u}^*))$. A matrix-extension of the olr transformation is given by $\text{olr}(\mathbf{A}) = \ln(\mathbf{A})\mathbf{V}$, for a $n \times D$ matrix $\mathbf{A}$.
The compositional product of a matrix by a vector, denoted by $\boxdot$, is given by

$$\mathbf{B} \boxdot \mathbf{u} = \text{olr}^{-1}(\mathbf{V}^T \mathbf{B} \ln(x)), \tag{1}$$

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 satisfying $\mathbf{B}\mathbf{1}_D = \mathbf{0}_L$ and $\mathbf{B}^T \mathbf{1}_L = \mathbf{0}_D$ where $\mathbf{1}_D$ (resp. $\mathbf{0}_D$) denotes the D-dimensional column vector of ones (resp. zeros). It is independent of the contrast matrix $\mathbf{V}$.

3 Covariates impacts in spatial autoregressive models and in simplicial regression models

Before combining these techniques in the next section, we first remind the reader some results about covariate impact evaluation both in spatial autoregressive models and in simplicial regression models separately. Note that since the impacts are relative to a single variable at a time, we do not index the explanatory variable of interest in order to simplify the notations. Moreover, the impacts are defined as changes in the expected value of the dependent variable which is to be understood as being conditional on the set of explanatory variables but, for the same reason, this point is not reflected in the notation.

3.1 Univariate spatial autoregressive models

Let us consider here an ordinary spatial autoregressive regression model, often referred to as the LAG model, of the following type

$$\mathcal{Y} = \rho \mathbf{W} \mathcal{Y} + \mathcal{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathcal{Y} = (Y_1, \cdots, Y_n)^T$ is the vector of observed values of the dependent variable at n locations in space, $\mathcal{X} = (X_i^k)_{i=1,\cdots n, k=1,\cdots K}$ is a $n \times K$ matrix of K covariates values observed at the same locations, $\boldsymbol{\epsilon}$ is a vector of i.i.d. disturbances with mean zero and variance σ^2, and $\mathbf{W}$ is a $n \times n$ neighborhood matrix. In spatial econometrics, the neighborhood matrix elements w_{ij} are measures of proximity between locations i and j (see for instance, Bivand, Gomez-Rubio, & Pebesma, 2008) and the lagged vector $\mathbf{W}\mathcal{Y}$ contains averages of the values of the variable $\mathcal{Y}$ in neighboring locations when $\mathbf{W}$ is row normalized. In order to assess the covariate impacts, one has to first rewrite the model equation in the so-called reduced form

$$\mathcal{Y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathcal{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}), \tag{3}$$

where $\mathbf{I}_n$ is the identity matrix of size n. Equations (2) and (3) are equivalent provided the parameter ρ is such that the filter matrix $\mathbf{A}(\mathbf{W}) = \mathbf{I}_n - \rho \mathbf{W}$ is invertible. LeSage and Pace (2009) show that due to the presence of the filter matrix, the parameters of each explanatory variable no longer measure the marginal effect of a change in this explanatory variable on the dependent variable. Since marginal effects are relative to one explanatory variable at a time, we remove the index k indicating the variable of interest. For a given variable $\mathbf{X} = (X_1, \cdots, X_n)^T$, these marginal effects are measured by the matrix of partial derivatives with terms

$$me_{ij}^X = \frac{\partial \mathbb{E}(Y_i)}{\partial X_j}$$

when i and j range in the set of n locations. For a given variable $\mathbf{X}$, these can be easily expressed in terms of the parameters and the terms of the filter matrix by $me_{ij} = a_{ij}(\mathbf{W})\beta_X$ where $a_{ij}(\mathbf{W})$ denotes the general term of the filter matrix and β_X is the parameter corresponding to variable $\mathbf{X}$. LeSage and Pace (2009) propose to summarize these n^2 impact measures separating the direct effects corresponding to $i = j$ from the indirect effects $i \neq j$. The average total impact $TE = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n me_{ij}$ then measures the average cumulative effect of increasing $\mathbf{X}$ by one unit at all locations on a typical $\mathbf{Y}$ value. The average direct impact $DE = \frac{1}{n}\sum_{i=1}^{n} me_{ii}$ (respectively the average indirect impact, also called network effect $IE = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} me_{ij}$) measures the average cumulative effect of increasing $\mathbf{X}$ by one unit at all locations on the $\mathbf{Y}$ value at that same location (respectively the average cumulative effect of increasing $\mathbf{X}$ by one unit at all locations on the $\mathbf{Y}$ value at a typical location other than its own).

3.2 Simplicial regression models

Turning now attention to simplicial regression models, we focus on the case where the dependent variable is of a compositional nature $\mathbf{Y} \in \mathcal{S}^D$ and we first assume that the explanatory variable of interest X is not compositional. For a given choice of contrast matrix $\mathbf{V}^Y$, the olr transformed vectors and parameters will be denoted by $\mathbf{Y}^*$ and $\mathbf{b}^*$. Then we have $\mathbb{E}^{\oplus}\mathbf{Y} = X \odot \mathbf{b} \oplus E$, where E contains terms involving covariates other than X. In that case, Morais and Thomas-Agnan (2021) show that the marginal effects should be measured using semi-elasticities rather than marginal effects in order to be coherent with the simplex structure of the space of the dependent variable. Let $(\mathbf{Y}_i, X_i)$, $i = 1, \dots, n$, be iid random vectors with the same distribution as $(\mathbf{Y}, X)$. The semi-elasticities are then defined by the vectors

$$se_i^X = \frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial X_i}, \tag{4}$$

where $\mathbb{E}^{\oplus}\mathbf{Y}$ is the expected value in the simplex of the simplex valued random variable $\mathbf{Y}$ (see Pawlowsky-Glahn et al. (2015) for a definition). We will denote by $se_{i:m,}^X$ for $m = 1$ to D, the D components of the vector se_i^X. Morais and Thomas-Agnan (2021) show that for each individual i, the vector of semi-elasticities se_i^X can be expressed in terms of the model parameters as follows

$$se_i^X = \mathbf{U}_{\mathbb{E}^{\oplus}\mathbf{Y}_i}\mathrm{clr}(\mathbf{b}^X),$$

where $\mathbf{U}_z = \mathbf{I}_D - \mathbf{1}_D\mathbf{z}'$ with $\mathbf{I}_D$ the $D \times D$ identity matrix, $\mathbf{1}_D$ the $D \times 1$ vector of ones and $\mathbf{b}^X$ is the vector of parameters in the simplex for variable X. Given a contrast matrix $\mathbf{V}^Y$ and the corresponding olr transformation, Morais and Thomas-Agnan (2021) also give an expression of these semi-elasticities in olr coordinate space which is independent of the particular contrast matrix: $se_i^X = \mathbf{U}^*_{\mathbb{E}^{\oplus}\mathbf{Y}_i}\mathbf{b}^{X*}$, where $\mathbf{U}^*_z = \mathbf{U}_z\mathbf{V}^Y$, and $\mathbf{b}^{X*}$ is the vector of parameters in coordinate space for variable X.

Let us recall that the vector $se_i^X = (se_{i:1}^X, \cdots, se_{i:D}^X)$ of semi-elasticities satisfies the following property:

$$\sum_{m=1}^{D} se_{i:m}^X \mathbb{E}^{\oplus} Y_{i:m} = 0, \tag{5}$$

where $(Y_{i:1}, \cdots, Y_{i:D})$ are the D components of $\mathbf{Y}_i$ (see Morais, Thomas-Agnan, and Simioni (2018), Appendix A.4 adapted to semi-elasticities). We can associate to the vector of semi-elasticities a vector of derivatives in the simplex by

$$sd_i^X = \frac{\partial^{\oplus} \mathbb{E}^{\oplus} \mathbf{Y}_i}{\partial X_i} \tag{6}$$

with components $(sd_{i:1}^X, \cdots, sd_{i:D}^X)$. sd_i^X belongs to $\mathcal{S}^D$ and is related to the vector of semi-elasticities by

$$sd_i^X = \mathcal{C}(\exp(se_i^X)). \tag{7}$$

For a small δ, one may approximate the simplicial derivative by a finite increment

$$sd_i^X = \frac{\partial^{\oplus} \mathbb{E}^{\oplus} \mathbf{Y}_i}{\partial X_i} \sim \frac{1}{\delta} \odot (\mathbb{E}^{\oplus} \mathbf{Y}_i(x_i + \delta) \ominus \mathbb{E}^{\oplus} \mathbf{Y}_i(x_i)) \tag{8}$$

which can be rewritten

$$\mathbb{E}^{\oplus} \mathbf{Y}_i(x_i + \delta) = \mathbb{E}^{\oplus} \mathbf{Y}_i(x_i) \oplus \mathcal{C}((sd_i^X)^{\delta}) = \mathcal{C}(\mathbb{E}^{\oplus} \mathbf{Y}_i(x_j)(sd_i^X)^{\delta}),$$

where $\mathbb{E}^{\oplus} \mathbf{Y}_i(x_i + \delta)$ is the expected value of $\mathbf{Y}$ for statistical unit i when X_i increases by δ units. For a small δ, given the relationship between $sd_{i:m}$ and $se_{i:m}$, the following approximation holds:

$$\frac{\mathbb{E}^{\oplus} \mathbf{Y}_{i:m}(x_i + \delta)}{\mathbb{E}^{\oplus} \mathbf{Y}_{i:m}(x_i)} \sim 1 + \delta se_{i:m}^X \tag{9}$$

also equivalent to

$$\frac{\mathbb{E}^{\oplus} \mathbf{Y}_{i:m}(x_i + \delta) - \mathbb{E}^{\oplus} \mathbf{Y}_{i:m}(x_i)}{\mathbb{E}^{\oplus} \mathbf{Y}_{i:m}(x_i)} \sim \delta se_{i:m}^X.$$

Finally note that property (5) is important because it ensures that the vector with components $\mathbb{E}^{\oplus} \mathbf{Y}_{i:m}(x_i + \delta)$ belongs to the simplex.

4 Elasticity-based impacts for the spatial-compositional regression model

Nguyen et al. (2021) introduce a simultaneous spatial regression model of the LAG type for compositional data. Note that this is a multivariate spatial model since the dependent variable vector is in $\mathcal{S}^{\mathcal{D}}$ and they use Kelejian and Prucha (2004) for defining and estimating a multivariate version of the univariate spatial autoregressive LAG model. Thomas-Agnan et al. (2021) present an application of this model to land use data. Although it is possible to use a different set of explanatory variables for each olr coordinate, we will assume for simplicity that the same set of K explanatory variables is used in all coordinate regression equations. We also assume first that the K explanatory variables are not compositional. The model can be written in the simplex as follows

$$\mathcal{Y} = (\mathbf{W} \triangle \mathcal{Y}) \boxdot \mathbf{R} \oplus \mathcal{X} \odot \boldsymbol{\beta} \oplus \boldsymbol{\epsilon}, \tag{10}$$

where $\mathcal{Y}$ is a $n \times D$ matrix of compositional vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, $\mathbf{W} \triangle \mathcal{Y} := \mathrm{olr}^{-1}(\mathbf{W}\mathrm{olr}(\mathbf{Y}))$, $\mathbf{R}$ is a $D \times D$ matrix of autocorrelation parameters such that $\mathbf{R}\mathbf{1}_D = \mathbf{0}_D$ and $\mathbf{R}^T\mathbf{1}_D = \mathbf{0}_D$, $\mathcal{X}$ is a $n \times K$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $K \times D$ matrix of parameters, and $\boldsymbol{\epsilon}$ is a $n \times D$ matrix of compositional errors satisfying the following conditions. Denoting by $\boldsymbol{\epsilon}_{.l}$ the columns of $\boldsymbol{\epsilon}$ and by $\boldsymbol{\epsilon}_{i.}$ its rows, we assume that $\mathbb{E}(\boldsymbol{\epsilon}^*_{i.}\boldsymbol{\epsilon}^{*\,T}_{j.}) = \boldsymbol{\Sigma}^*$ if individuals i and j coincide and $\mathbf{0}$ if they are different, where $\boldsymbol{\Sigma}^*$ is a $(D-1) \times (D-1)$ covariance matrix. The matrix product $\mathbf{A} \boxdot \mathbf{R}$, for a $n \times D$ matrix $\mathbf{A}$, is an extension of the product of a matrix by a vector given in (1) and is defined by $\mathrm{olr}(\mathbf{A} \boxdot \mathbf{R}) = \ln(\mathbf{A})\mathbf{R}\mathbf{V}$ where $\mathbf{V}$ denotes the contrast matrix of the olr transform.

For this framework, Thomas-Agnan et al. (2021) give the reduced form of the model in coordinate space and in vectorized form

$$\mathrm{vec}_c\mathcal{Y}^* = \left(\mathbf{I}_{n(D-1)} - (\mathbf{R}^*)^T \otimes \mathbf{W}\right)^{-1} \left[\mathcal{X}\boldsymbol{\beta}^*_1 \ \ldots \ \mathcal{X}\boldsymbol{\beta}^*_{D-1}\right]^T + \mathrm{vec}_c\boldsymbol{\epsilon}^*,$$

where $\otimes$ denotes the Kronecker product of matrices and where $\mathrm{vec}_c\mathbf{A}$ denotes the column vectorization obtained by stacking the columns of a matrix $\mathbf{A}$. For $i, j = 1, \cdots, n$ and $l, m = 1, \cdots, D$, let us denote by $a_{il:jm}(\mathbf{W})$ the general term of the $n(D-1) \times n(D-1)$ filter matrix $\mathbf{A}(\mathbf{W}) = (\mathbf{I}_{n(D-1)} - (\mathbf{R}^*)^T \otimes \mathbf{W})^{-1}$.

The covariate impacts in Thomas-Agnan et al. (2021) are assessed using approximations of semi-elasticities based on finite differences that generalize equation (9) to the case where variable $\mathbf{X}$ is increased at location j and we look at the relative impact on $\mathbf{Y}$ at location i. We are now going to develop a formula for an exact evaluation of these quantities.

4.1 Approximate semi-elasticities

As in (4) and (6), let us introduce the semi-elasticities and semi-derivatives with the following notations

$$se^X_{ij} = \frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial X_j}$$

$$sd^X_{ij} = \frac{\partial^{\oplus}\mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial X_j}.$$

Note that these semi-elasticities vectors are now doubly-indexed due to the spatial dependence: the impact of changing an explanatory variable at a given location j is affecting all locations i. The m^{th} component of each of these vectors will be denoted respectively by $se^X_{ij:m}$ and $sd^X_{ij:m}$.

As in equation (8), for a small δ, one may approximate the simplicial derivative by a finite increment

$$sd_{ij}^X = \frac{\partial^{\oplus}\mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial X_j} \sim \frac{1}{\delta} \odot (\mathbb{E}^{\oplus}\mathbf{Y}_i(x_j+\delta) \ominus \mathbb{E}^{\oplus}\mathbf{Y}_i(x_j))$$

which can be rewritten

$$\mathbb{E}^{\oplus}\mathbf{Y}_i(x_j+\delta) = \mathbb{E}^{\oplus}\mathbf{Y}_i(x_j) \oplus \mathcal{C}((sd_{ij}^X)^{\delta}) = \mathcal{C}(\mathbb{E}^{\oplus}\mathbf{Y}_i(x_j)(sd_{ij}^X)^{\delta})$$

yielding, for a small δ, and given the relationship between $sd_{ij:m}^X$ and $se_{ij:m}^X$, the following approximation

$$\frac{\mathbb{E}^{\oplus}\mathbf{Y}_{i:m}(x_j+\delta)}{\mathbb{E}^{\oplus}\mathbf{Y}_{i:m}(x_j)} \sim 1 + \delta se_{ij:m}^X$$

also equivalent to

$$\frac{\mathbb{E}^{\oplus}\mathbf{Y}_{i:m}(x_j+\delta) - \mathbb{E}^{\oplus}\mathbf{Y}_{i:m}(x_j)}{\mathbb{E}^{\oplus}\mathbf{Y}_{i:m}(x_j)} \sim \delta se_{ij:m}^X. \tag{11}$$

Note that equation (11) was used to approximate the semi-elasticity by finite differences in Thomas-Agnan et al. (2021). But this equation is also useful to give an interpretation to the value of a semi-elasticity, computed by the exact formula that we are going to derive in the next subsection: $\delta se_{ij:m}^X$ represents the percent change of $Y_{i:m}(x_j)$ when x_j increases by δ.

4.2 Exact semi-elasticities

Combining tools of sections 3.1 and 3.2, we need to evaluate the semi-elasticities se_{ij}^X. Using the same approach as Morais and Thomas-Agnan (2021), we first have that

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial X_j} = \frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial \mathbb{E}\mathbf{Y}_i^*}\frac{\partial \mathbb{E}\mathbf{Y}_i^*}{\partial X_j}. \tag{12}$$

The first term on the right hand side is given by an application of Lemma 4.2 in Morais and Thomas-Agnan (2021). Since $\mathbb{E}^{\oplus}\mathbf{Y}_i = \operatorname{olr}^{-1}(\mathbf{Y}_i^*)$, we get

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y_i}}{\partial \mathbb{E}\mathbf{Y}_i^*} = \mathbf{U}^*_{\mathbb{E}^{\oplus}\mathbf{Y_i}}. \tag{13}$$

For computing the second term, using the reduced form of the model in coordinate space, we can write that, for each individual i and each component l,

$$\mathbb{E}\mathbf{Y}_{i:l}^* = \sum_{t=1}^{n}\sum_{m=1}^{D-1} a_{il:tm}(\mathbf{W})X_t\beta_m^* + \mathbf{E}_{i:l}, \tag{14}$$

where $\mathbf{E}_{i:l}$ is a sum of terms involving the other explanatory variables at the exception of $\mathbf{X}$. We therefore get

$$\frac{\partial \mathbb{E}\mathbf{Y}^*_{i:l}}{\partial X_j} = \sum_{m=1}^{D-1} a_{il:jm}(\mathbf{W})\beta^*_m.$$

If we denote by $A_{ij}(\mathbf{W})$ the $(D-1)\times(D-1)$ submatrix of the filter matrix $\mathbf{A}(\mathbf{W})$ corresponding to all terms involving locations i and j, we can rewrite the result for the full vector $\mathbb{E}\mathbf{Y}^*_{\mathbf{i}.}$ as

$$\frac{\partial \mathbb{E}\mathbf{Y}^*_{i}}{\partial X_j} = A_{ij}(\mathbf{W})(\beta^*_1, \cdots, \beta^*_{D-1})^T. \tag{15}$$

Combining (13) and (15), we get the following theorem.

Theorem 1 *In the spatial simultaneous autoregressive model* (10)*, the vector of semi-elasticities corresponding to classical covariates are given by*

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial X_j} = \mathbf{U}^*_{\mathbb{E}^{\oplus}\mathbf{Y_i}} A_{ij}(\mathbf{W})(\beta^*_1, \cdots, \beta^*_{D-1})^T, \tag{16}$$

where $\mathbf{U}^*_z = \mathbf{U}_z\mathbf{V}^Y$.

We can easily extend this result to the case of a compositional explanatory variable with the same technique. In that case the model equation becomes

$$\mathcal{Y} = \mathcal{X} \odot \boldsymbol{\beta} \bigoplus_{k=1}^{K_c} \mathcal{X}_k \boxdot \mathbf{B}_k \oplus (\mathbf{W} \triangle \mathbf{Y}) \boxdot \mathbf{R} \oplus \boldsymbol{\epsilon}, \tag{17}$$

where $\mathcal{X}_k$ is the k^{th} compositional explanatory variable, $\mathbf{B}_k$ the corresponding matrix of parameters ($\mathbf{B}^*_k$ will denote its olr transform), and K_c is the number of compositional explanatory variables. For such a compositional explanatory variable, omitting as before its index k, we get the following expression of the elasticities which are now replacing the semi-elasticities as in Morais and Thomas-Agnan (2021):

Theorem 2 *In the spatial simultaneous autoregressive model* (17)*, the matrix of elasticities corresponding to a compositional covariate with contrast matrix* $\mathbf{V}^X$ *is given by*

$$\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_i}{\partial \log X_j} = \mathbf{U}^*_{\mathbb{E}^{\oplus}\mathbf{Y_i}} A_{ij}(\mathbf{W})\mathbf{B}^*\mathbf{V}^X,$$

where $\mathbf{U}^*_z = \mathbf{U}_z\mathbf{V}^Y$.

5 Local impacts and impacts decomposition

In this section, coming back to the case of a classical explanatory variable, we present several ways of exploring and summarizing the semi-elasticities. Recall that we have for each component m, $(m = 1, \cdots, D)$, a $n \times n$ matrix of semi-elasticities and we will say that it is the most disaggregated level. In order to summarize these we want to aggregate them over the spatial locations. In the classical point of view for spatial autoregressive models, the aggregation is done over i and j simultaneously. We introduce intermediate steps by aggregating on one of the two indices yielding what we will call local impacts.

In order to introduce the concepts, we will use a toy data set built from a real data set that will be presented in the next section. We consider the 13 departments of the French Occitanie region and we simulate dependent and explanatory variables using a spatial-compositional regression model inspired by Nguyen et al. (2021), but at the department level instead of the canton level. We simulate for each department a single classical explanatory variable (unemployment rate, denoted "unemp") using a continuous uniform distribution on the interval $[0, 0.3]$ and we compute the corresponding vote shares of three parties: Left, Right and Extreme Right (XR) using parameter values close to those obtained in Nguyen et al. (2021). The spatial weight matrix $\mathbf{W}$ is based on the rook contiguity calculated on simplified department geometries, i.e. two polygons are neighbors if they share a common side. The simulated model in the olr space is then:

$$\mathcal{Y}^* = \mathbf{1}_n \times \boldsymbol{\beta}^{0*} + \mathbf{unemp} \times \boldsymbol{\beta}^{1*} + \mathbf{W}\mathcal{Y}^*\mathbf{R}^* + \boldsymbol{\epsilon}^*$$

where $\mathcal{Y}^*$ is a $n \times 2$ matrix, **unemp** is a $n \times 1$ vector, and

$$\boldsymbol{\beta}^{0*} = \begin{bmatrix}\beta^*_{01} & \beta^*_{02}\end{bmatrix} = \begin{bmatrix}1 & -1.1\end{bmatrix}; \; \boldsymbol{\beta}^{1*} = \begin{bmatrix}\beta^*_{11} & \beta^*_{12}\end{bmatrix} = \begin{bmatrix}-7.6 & 10.5\end{bmatrix};$$

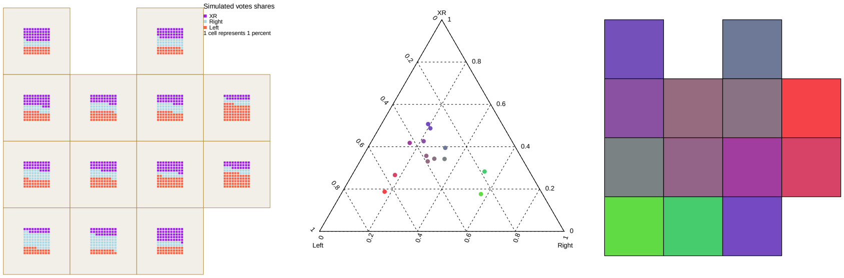
$$\boldsymbol{\Sigma}^* = \begin{bmatrix}\sigma^{2*}_{11} & \sigma^*_{12} \\ \sigma^*_{21} & \sigma^{2*}_{22}\end{bmatrix} = \begin{bmatrix}0.09 & 0.02 \\ 0.02 & 0.09\end{bmatrix}; \; \mathbf{R}^* = \begin{bmatrix}R^*_{11} & R^*_{12} \\ R^*_{21} & R^*_{22}\end{bmatrix} = \begin{bmatrix}0.65 & 0.18 \\ 0 & 0.63\end{bmatrix}.$$

The $n \times 3$ matrix of vote shares $\mathcal{Y}$ is obtained after using the olr inverse operation on $\mathcal{Y}^*$ with a constrast matrix $\mathbf{V}$ for the olr transformation given by:

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

Figure 1 shows maps of the simulated vote shares using two different representations: on the left, we plot a waffle layer using package **cartography** (see Giraud & Lambert, 2016). In the supplementary material, we present a variant of this map using a pie chart at each location. As our share data is three dimensional, we can use a ternary diagram for representing the composition in the central plot. We then associate to each point of the ternary diagram a unique color such that the ratio of its primary colors are directly connected to the values of $\mathbf{Y}$ in the simplex as explained in Laurent, Ruiz-Gazen, and Thomas-Agnan (2021): the closer the color to red, the stronger the Left share, the closer the color to green, the stronger the Right share, the closer the color to blue, the stronger the Extreme Right share. These colors are then plotted on the map, and the combination of the ternary diagram in the center with the right hand side map gives a visual point of view on the spatial distribution of the shares with their corresponding position in the simplex. We compute the

Fig. 1 Maps of the vote shares using two representations: a waffle layer and a ternary diagram with colors depending on the locations in the simplex

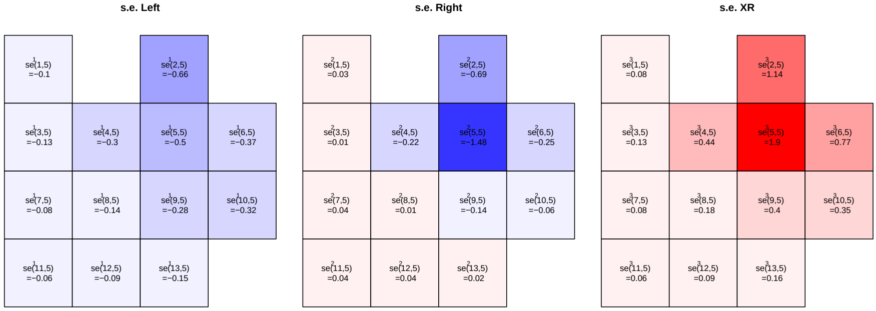


semi-elasticities se_{ij}^X with equation (16). In practice, one would first estimate the parameters of the spatial-compositional regression model; however, since we have simulated the data, we directly use the data generating process parameters. Finally, we obtain one matrix of semi-elasticities per component i.e. D matrices of size $n \times n$. At this very disaggregated level, for a given location s, it is then possible to represent the semi-elasticities $se_{is:m}^X$, $i = 1, \ldots, n$, on D maps (one map for each component m). For example, Figure 2 represents the impacts of changing the variable **unemp** in the Aveyron department, on the vote share Y at all other locations. We can see that the impacts are stronger in the neighborhood of Aveyron, the so-called spillovers in a spatial model.

Moreover, although the signs of the semi-elasticities are the same for the Extreme right (all positive) and Left (all negative), we observe both positive and negative impacts for the Right, depending on the location. The question then arises as to which of the two signs will prevail over the other in the event that the impacts were aggregated over locations. Besides, it does not seem practical to make such a plot for each of the n observations, especially if the sample size is huge. This is what motivates the following proposals, namely how can we aggregate and summarize all these semi-elasticities.

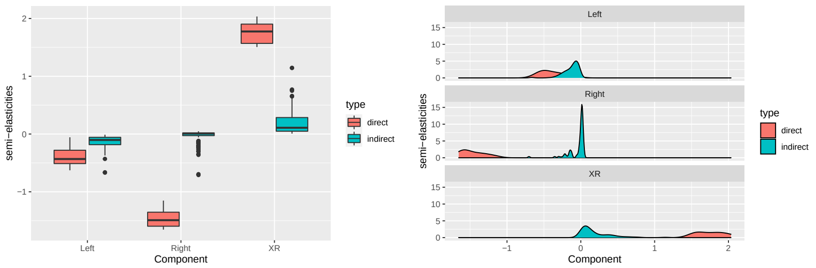
For each component, the diagonal terms of the matrix of semi-elasticities characterize direct impacts, and the extra-diagonal terms indirect impacts. Without loss of information, a first proposal is to represent parallel boxplots or density plots of the semi-elasticities, grouped by direct and indirect impacts, with respect to the component as shown in Figure 3. Direct and indirect impacts of the Extreme Right and Left components have the same sign (positive for Extreme Right and negative for Left) which leads us to the following

Fig. 2 Local impacts due to a change of **unemp** at the Aveyron department (located at 4th column, 2nd row)



interpretation: the unemployment rate has a positive impact on the Extreme Right vote share whereas it has a negative impact on the Left vote share. For the Right component, although the direct impact is positive, one cannot make a clear interpretation of an overall impact as the indirect impact can be negative (same remark has been done previously for the Aveyron department).

Fig. 3 Parallel boxplots (and density plots) of the semi-elasticities grouped by direct and indirect effects for each component

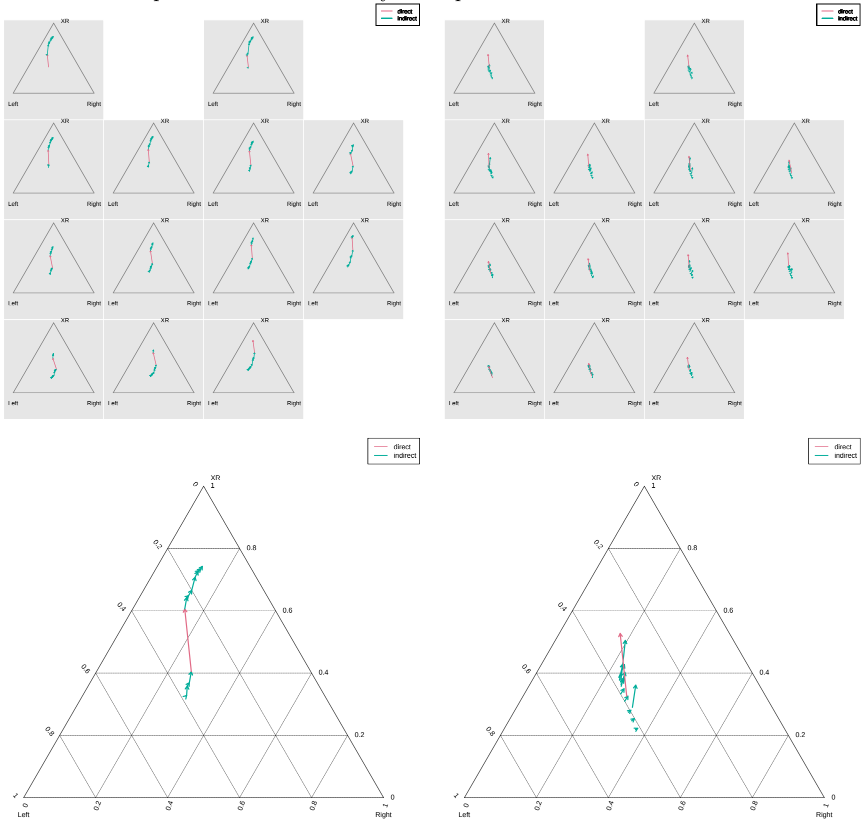


Let us now represent the changes in the simplex which can be computed using the relationship between semi-elasticities and derivatives in the simplex (7). The top part of Figure 4 presents the local impacts in the simplex for each location s using two points of view for grouping the impacts, the bottom part of Figure 4 is a zoom on the Aveyron department.

In any of the left hand side ternary diagrams, corresponding to a fixed location s, the first arrow represents the change of $\mathbf{Y}$ at s due to changing the explanatory variable at location 1, the next one is the further impact of changing the explanatory at location 2 and so on: the red arrow is when the change is at location s and the other arrows are green.

In any of the right hand side ternary diagrams, with the same convention for the colors, for a fixed location of change s, each arrow starts at a different point $\mathbb{E}(Y_i)$ and represents the change of $\mathbb{E}(Y_i)$ resulting from changing X at location s.

Fig. 4 Simplicial derivatives in a ternary diagram using two groupings: on the left by rows of the semi-derivative matrices whereas on the right by columns of the semi-derivative matrices. Bottom part: Zoom on the Aveyron departement.



Focusing for example on the Aveyron case, in the second approach (bottom right of Figure 4), with $\delta = 0.3$, we can see the expected evolution of vote shares when the unemployment rate increases by 30 points in Aveyron only. It appears that the direct impacts are stronger than the indirect ones. On the bottom left, we see the successive evolutions of the expected share of a typical department with the same characteristics as Aveyron when the unemployment rate increases by 30 points at all locations.

Going one step further in the direction of summarizing the impacts, we are first going to aggregate these doubly-indexed quantities on a single index at a time to define local effects. Then we will also bring them back to the simplex for a direct interpretation in terms of shares. For the direct impact, we define the direct effect at location s in the simplex to be

$$DE_s = sd_{ss} = \mathcal{C}(\exp(se_{ss})).$$

Note that there is a single such local effect for each s. For the indirect effects, there are two ways to associate $n-1$ indirect impacts to a given location s: adding the s-row elements or adding the s-column elements, of the matrices of semi-derivatives or semi-elasticities (excluding the diagonal elements). In both cases, adding simplex derivatives with the $\oplus$ operation in the simplex corresponds to adding the corresponding semi-elasticities with the ordinary addition in $\mathbb{R}$ thanks to $sd_{ij} = \mathcal{C}(\exp(se_{ij}))$ and to the fact that $\mathcal{C}(\exp(a)) \oplus \mathcal{C}(\exp(b)) = \mathcal{C}(\exp(a+b))$.

Adding the elements by columns corresponds to defining the local indirect impact in the simplex $IE_{s\oplus}$ by aggregating the changes on $\mathbf{Y}$ at a given location s due to changing $\mathbf{X}$ at all other locations.

$$IE_{s\oplus} = \bigoplus_{j:j\neq s} sd_{sj} = \mathcal{C}(\exp(\sum_{j:j\neq i} se_{sj})).$$

Therefore the sum of these semi-elasticities can be interpreted as a percent increase of $\mathbf{Y}$ at s resulting from all these changes.

Adding the elements by rows corresponds to defining the local indirect impact in the simplex $IE_{\oplus s}$ by aggregating the impacts due to changing $\mathbf{X}$ at a given location s on $\mathbf{Y}$ at all other locations

$$IE_{\oplus s} = \bigoplus_{i:i\neq s} sd_{is} = \mathcal{C}(\exp(\sum_{i:i\neq j} se_{is})).$$

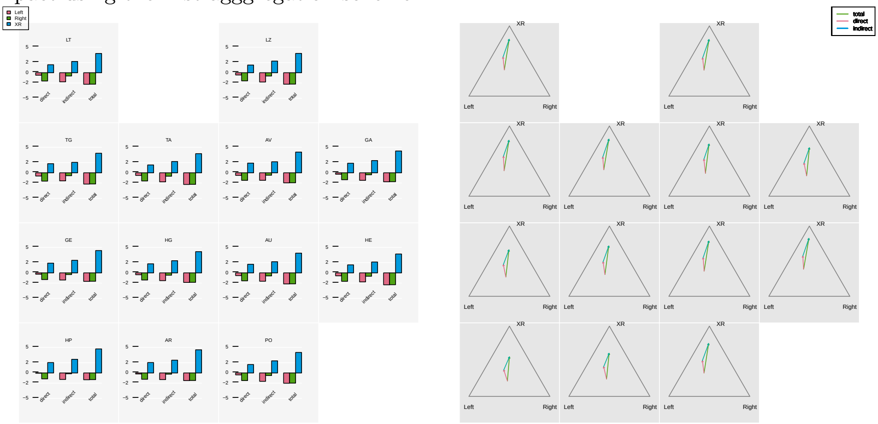
However since we consider the effects on $\mathbf{Y}$ at different locations which might have a different initial value of $\mathbf{Y}$, the interpretation is more tricky. Adding these impacts in the simplex represents the change of the total in the simplex (for the $\oplus$ operator) of the share vectors. Indeed $\oplus_i \mathcal{C} \exp \delta se_{is} = \mathcal{C}(\exp \delta \sum_i se_{is})$, therefore $\delta \sum_i se_{is:m}$ can be interpreted as the percent increase of component m of the total share vector, for the simplex $\oplus$ operator. It is also the simplex-average share vector when multiplying (with $\odot$) by $\frac{1}{n}$. Therefore the vector $(\sum_i se_{is})_{m=1,\cdots m}$ yields the vector of percent-changes of all components of the simplex average share vector.

In both cases, we can define a local total effect $DE_s + IE_{s\oplus}$ and $DE_s + IE_{\oplus s}$. Instead of having D matrices of size $n \times n$, we now have D vectors of size $n \times 1$ for direct impacts, D vectors of size $n \times 1$ for indirect impacts (for each point of view) and D vectors of size $n \times 1$ for total impacts (for each point of view).

Using the first aggregation scheme, the right hand side of Figure 5 represents the local direct, indirect ($IE_{s\oplus}$) and total impacts ($DE_s + IE_{s\oplus}$). The left hand side shows barplots of the corresponding semi-elasticities.

On the barplots, we see that the direct impacts on the Left share seem lower than on the Right share whereas it is the inverse for the indirect impacts. On the ternary diagrams, the direct impact is almost parallel to the basis of the ternary diagram which is opposed to the Left vertex, indicating that when unemployment rate increases at location s, the Extreme Right vote share increases at location s mainly at the expense of the Right. The indirect impact's

Fig. 5 Barplots of semi-elasticities and ternary diagrams of the local direct and indirect impact using the first agggregation scheme



arrow is almost perpendicular to the basis opposed to the Extreme Right vertex indicating that when the unemployment rate increases at all other locations, the Extreme Right vote increases at location s mainly at the expense of the Left.

Table 1 gives the final summary: the aggregation over both indices yielding the classical direct, indirect and total impacts. When the unemployment increases by δ points for a small δ, the Extreme right vote shares increase by $4.1\delta\%$ of which $1.8\delta\%$ is due to the direct effect and $2.3\delta\%$ is due to the indirect effect.

Table 1 Impacts summary for the unemployment variable.

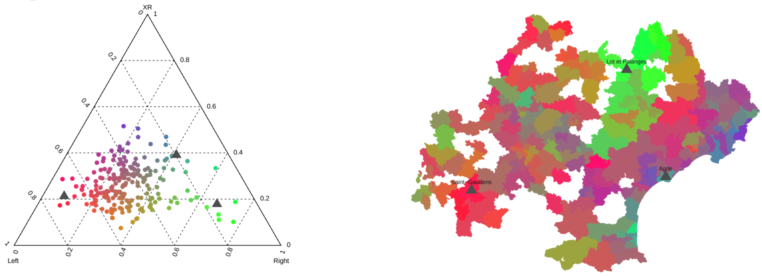
	Left	**Right**	**Extreme Right**
Direct	-0.399	-1.465	1.756
Indirect	-1.571	-0.483	2.309
Total	-1.970	-1.948	4.065

6 Application on a case

We use a case presented in Nguyen et al. (2021) illustrating the simultaneous spatial regression model of the LAG type for compositional data with a political dataset. They explain the vote shares of the Left, Right and Extreme Right parties on $n = 207$ cantons, by socio-economic explanatory variables as the diploma level, the age distribution or the proportion of people who pay income tax. The spatial weight matrix they use is based on the first order queen contiguity: i.e. two polygons are neighbors if they share a common side or a common vertex. In their work, the semi-elasticities are approximate and the interpretation is not done in the simplex. In this section, we are going to present and interpret the exact semi-elasticities.

Figure 6 shows the vote shares data in the simplex and on a map. As in Nguyen et al. (2021), the cantons with at least one missing value have been eliminated. We can identify some clusters that behave similarly: for instance, the south west contains cantons with a important Left vote share, the north tends to have cantons with a important Right vote share, and finally, the east has cantons with strong Extreme Right vote share. To illustrate the local impacts, we choose the following three cantons, namely "Saint-Gaudens" from the south-east, "Lot et Palanges" from the north, and "Agde" from the east.

Fig. 6 Maps of the vote shares in the $n = 207$ cantons

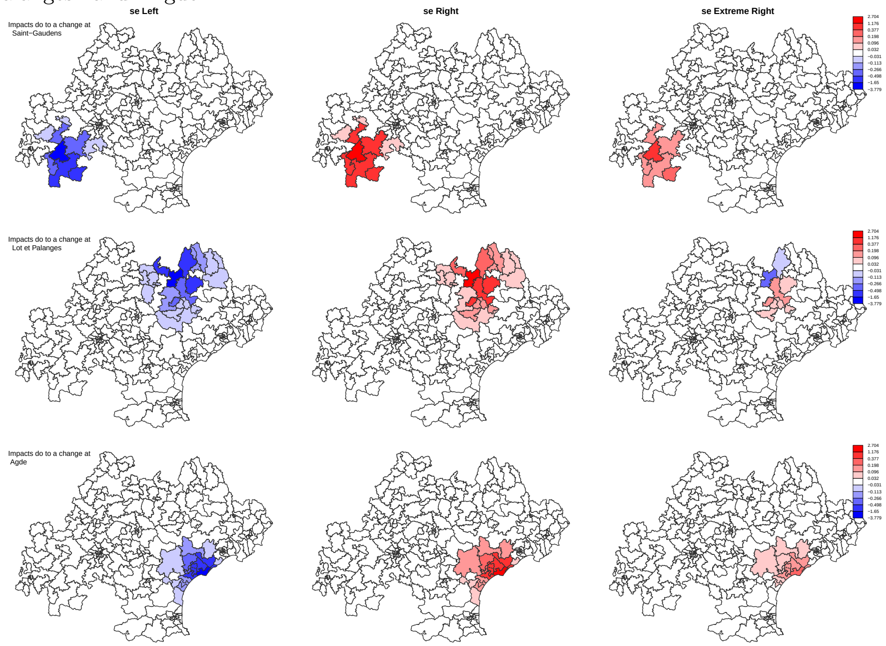


For the simultaneous spatial regression model of the LAG type for compositional data, we use the parameters estimates provided in Table 2 of Nguyen et al. (2021). Note that the estimates of the spatial autocorrelation matrix $\mathbf{R}^*$ is the same as in the previous section simulation framework. We will now focus on the interpretation of a classical variable: the proportion of people who pay income tax. As in the previous section we move from the less to the more aggregated points of view.

We first present in Figure 7 the local semi-elasticities for the three selected cantons. Whichever site we considered, the proportion of people who pay income tax has a negative direct and indirect impact on the Left vote share whereas it has a positive direct and indirect impact on the Right vote share. The direct and indirect impacts observed on the Extreme Right are positive for "Saint-Gaudens" and "Agde", but we can see that for "Lot et Palanges", impacts can be positive or negative. The closest neighbors seem to be affected as strongly as the modified sites, which suggest that the indirect impact is important.

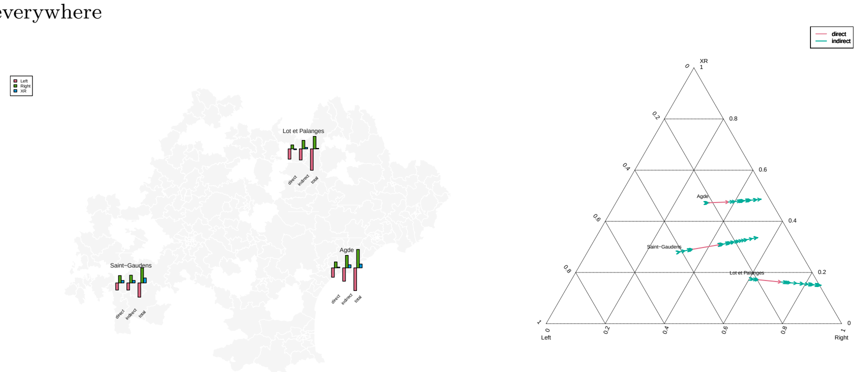
Figure 8 shows the local impacts at each of the locations "Saint-Gaudens", "Lot et Palanges" and "Agde" due to a change of X everywhere. On the left are the barplots of the semi-elasticities and on the right the cumulated impacts in the simplex with the first aggregation scheme. We observe that for the three cantons, the Right party is clearly the one that benefits from an increase of the proportion of people who pay income tax, to the detriment of the Left. However, if for "Lot et Palanges", the Extreme Right vote is weakly negatively impacted, it is the inverse for "Saint-Gaudens" and "Agde". Besides, the indirect effect is at least twice as large as the direct effect. Direct

Fig. 7 Semi-elasticities observed due to a change of X at "Saint-Gaudens", "Lot et Palanges" and "Agde"



and indirect effects influence the vote shares in the same way (barplots have the same signs and arrows are oriented identically).

Fig. 8 Impacts at "Saint-Gaudens", "Lot et Palanges" and "Agde" due do a change of X everywhere



To conclude, we can aggregate the local impacts to obtain the global summaries in Table 2. If the proportion of people who pay income tax variable increases by δ points for a small δ, the Right vote shares at all locations are predicted to increase by a total $8.2\delta\%$ of which $2.9\delta\%$ is due to the direct effect and $5.3\delta\%$ is due to the indirect effect. The Left vote share is predicted to decrease by a total of 6.9δ% of which 2.5δ% is due to the direct effect and 4.4δ% due to the indirect effect. The Extreme Right is predicted to increase by a total of 2.8δ% of which 1.1δ% is due to the direct effect and 1.76δ% is due to the indirect effect. This result underlines the importance of looking at the local impacts as we can see that, at least for one location, the impact on the Extreme Right can be locally negative.

Table 2 Scalar measures summary for proportion of people who pay income tax variable.

	Left	**Right**	**Extreme Right**
Direct	-2.516	2.927	1.081
Indirect	-4.429	5.300	1.759
Total	-6.945	8.227	2.840

7 Conclusion

In the framework of simplicial regression models involving spatial dependence in the compositional dependent variable (share vector), we have derived exact formulas for computing semi-elasticities (or elasticities, depending on the nature of the explanatory variable of interest). These are used to measure the impact of changing the value of a given explanatory variable at a given location on the expected value of the dependent variable at a potentially different location. We present exploratory tools to summarize this large amount of information with different levels of aggregation and we explain their interpretation directly in terms of the share vector. We propose several graphical representations illustrating the direct, indirect and total effects at a local level. Our future investigations will be focused on assessing the significance of these indicators which are nonlinear functions of the parameters.

Supplementary information. Supplementary material including code and additional figures can be found at http://www.anonymized.

Acknowledgments. 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.

References

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

Giraud, T., & Lambert, N. (2016). **cartography**: Create and integrate maps in your R workflow. *JOSS*, *1*(4).

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

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

Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C. (2021). Selecting colors for mapping compositional data. *Preprint.*

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

Martín-Fernández, J. (2019). Comments on: Compositional data: the sample space and its structure. *TEST*, *28*(3), 653–657.

Morais, J., & Thomas-Agnan, C. (2021). Impact of covariates in compositional models and simplicial derivatives. *Austrian Journal of Statistics*, *50*(2), 1–15.

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

Nguyen, T.H.A., Thomas-Agnan, C., Laurent, T., Ruiz-Gazen, A. (2021). A simultaneous spatial autoregressive model for compositional data. *Spatial Economic Analysis*, *16*(2), 161-175.

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

Thomas-Agnan, C., Laurent, T., Ruiz-Gazen, A., Nguyen, T.H.A., Chakir, R., Lungarska, A. (2021). Spatial simultaneous autoregressive models for compositional data: Application to land use. In P. Filzmoser, K. Hron, J.A. Martín-Fernández, & J. Palarea-Albaladejo (Eds.), *Advances in compositional data analysis: Festschrift in honour of vera pawlowsky-glahn* (pp. 225–249). Cham: Springer International Publishing.

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