# Analyzing the impacts of socio-economic factors on French departmental elections with CoDa methods

Thi-Huong-An Nguyen, Thibault Laurent, Christine Thomas-Agnan, Anne Ruiz-Gazen

WORKING PAPERS

N° TSE - 961

October 2018

# "Analyzing the impacts of socio-economic factors on French departmental elections with CODA methods"

T.H.A. Nguyen, T. Laurent,
C. Thomas-Agnan, A. Ruiz-Gazen

Toulouse School of Economics

**Noname manuscript No.**
(will be inserted by the editor)

# Analyzing the impacts of socio-economic factors on French departmental elections with CODA methods

**T.H.A. Nguyen · T. Laurent · C. Thomas-Agnan · A. Ruiz-Gazen**

Received: date / Accepted: date

**Abstract** The proportions of votes by party on a given subdivision of a territory form a vector called composition (mathematically, a vector belonging to a simplex). It is interesting to model these proportions and study the impact of the characteristics of the territorial units on the outcome of the elections. In the political economy literature, such regression models are generally restricted to the case of two political parties. In the statistical literature, there are regression models adapted to share vectors including CODA models (for COmpositional Data Analysis), but also Dirichlet models, Student models and others. Our goal is to use CODA regression models to generalize political economy models to more than two parties. The models are fitted on French electoral data of the 2015 departmental elections.

**Keywords** political economy · compositional regression models

## 1 Introduction

Recently, models for elections focus on analyzing impacts of socio-economic factors for two-party systems using classical regression models [1]. In this paper, we propose a statistical model for studying the multiparty system using compositional data analysis (CODA) with departmental level data. The dependent variable will be the vector of votes shares for the French departmental election in 2015. The explanatory variables include some compositional and continuous socio-economic variables.

Among papers concentrating on the relationship between socio-economic variables and election results, Beauguitte et Colange (2013) [2] study a linear regression at three levels of aggregation (polling stations, cities and electoral districts)

T.H.A. Nguyen
Toulouse School of Economics
Tel.: +33695798878
E-mail: thihuongan.nguyen@tse-fr.eu

T. Laurent · C. Thomas-Agnan · A. Ruiz-Gazen
Toulouse School of Economics

and show that the socio-economic variables are significant. Kavanagh et al (2006) [3] use geographically weighted regression, which produces parameter estimates for each data point, i.e. for each electoral division. On the other hand in the statistical literature, people have developped CODA regression models where the dependent and independent variables may be compositional variables (see Mert et al. (2016) [13] for a review). Morais et al. (2017) [4] study the impact of media investments on brand's market shares with a CODA regression model. Trinh and Morais (2017) [5] use a CODA regression model to highlight the nutrition transition and to explain it according to household characteristics. Honaker et al. (2002) [6], Katz and King (1990) [7] use a statistical model for multiparty electoral data assuming that the territorial units yield independent observations.

In Section 2, we present the data set. Subsection 3.1 (resp: 3.2) recalls the principles of compositional data analysis (resp: of compositional regression models). In subsection 3.3, we implement the CODA model on the election data set and present several plots to explore the impact of explanatory variables of a classical type illustrated by the case of unemployment rate as well as variables of a compositional type illustrated by the diploma variable.

## 2 Data

Vote share data of the 2015 French departmental election for 95 departments in France are collected from the Cartelec website [1] and corresponding socio-economic data (for 2014) have been downloaded from the INSEE website [2]. Table 1 summarizes our data set.

Table 1: Data description

| Variable name | Description | Averages[3] |
|---|---|---|
| Vote share | Left(L), Right(R), Extreme Right(XR) | 0.37, 0.388, 0.242 |
| Age | Age_1840, Age_4064, Age_65. | 0.313, 0.432, 0.255 |
| Diploma | <BAC, BAC, SUP. | 0.591, 0.16, 0.239 |
| Employment | AZ, BE, FZ, GU, OQ | 0.031, 0.099, 0.049, 0.439, 0.382 |
| unemp | the unemployment rate | 0.117 |
| employ_evol | Mean annual growth rate of employment (2009-2014) | -0.145 |
| owner | The proportion of people who own assets | 0.616 |
| income_ tax | The proportion of people who pay income tax | 0.552 |
| foreign | The proportion of foreigners | 0.050 |

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

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

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

[3] geometric average in the case of compositional variables

and SUP for people with a university diploma. The Age variable 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. For the vote share variable, the Cartelec website provides a very detailed information. The list of political parties which present candidates at that election is higher than 15. However, at the end of the election, it is common to present the results by grouping the political parties into three main components : Left, Right and Extreme-Right.[4]

From the CODA point of view, when compositional data have three components, they can be represented in a ternary diagram. For instance, the vote shares of the 95 departments for the Left and Right wings and the Extreme Right party are the blue points in Figure 1. The red triangle corresponding to the Aube department on Figure 1 shows that its vote shares of the Left wing, the Right wing and the Extreme Right party are respectively 17.4%, 54.6%, and 28% . Figure 2 illustrates the positions of the French departments on the ternary diagram whose components correspond to the three levels of the diploma variable, and the red triangle figures the geometric mean (adapted mean for compositional data) of all departments.

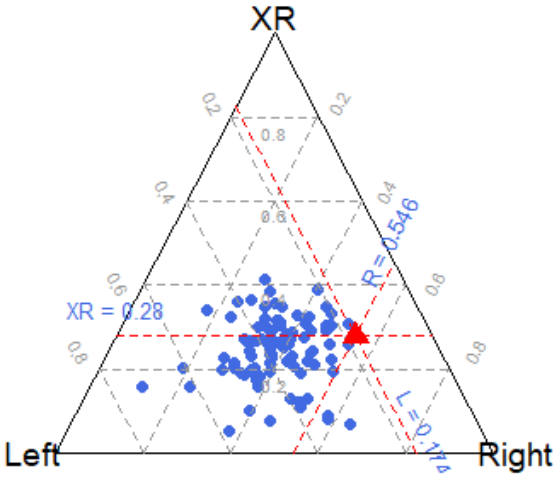


Figure 1: Vote shares in the 95 departments (blue points) with the Aube department as the red triangle

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

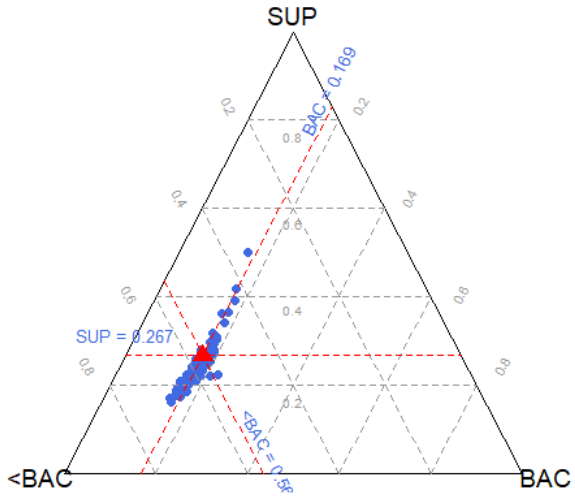


Figure 2: Components of Diploma in the 95 departments (blue points) and their geometric mean (red triangle)

## 3 Compositional data analysis approach

In order to analyze the impacts of the socio-economic factors on the election results, a CODA regression model is proposed where the dependent variable is a compositional variable (vote shares) and the independent variables are compositional or classical variables or a mixture of both. This model is based on the log-ratio transformation approach.

### 3.1 Principles of compositional data analysis

A composition $\mathbf{x}$ is a vector of $D$ parts of some whole which carries relative information. A $D$-composition $\mathbf{x}$ lies in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \{\mathbf{x} = (x_1, ..., x_D)' : x_j > 0, j = 1, ..., D; \sum_{j=1}^{D} x_j = 1\}$$

The simplex $\mathbf{S}^D$ can be equipped with the Aitchison inner product ([8] and [9]) in order to define distances. Classical regression models cannot be used directly in the simplex because the constraints that the components are positive and sum up to 1 are not compatible with their usual distributional assumptions. To overcome this difficulty, one way out is to use a log-ratio transformation from the simplex space $\mathbf{S}^D$ to the Euclidean space $\mathbb{R}^{D-1}$. The classical transformations are alr (additive log-ratio transformation), clr (centered log-ratio transformation), and

ilr (isometric log-ratio transformation). The coordinates in the clr transformed vector are linearly dependent, and the coordinates in the alr transformed vector are not compatible with the geometry (distance between the components in the simplex space is different from distance between the coordinates in the Euclidean space). For these reasons people generally use one of the ilr transformation for compositional regression models.
An isometric log-ratio transformation (ilr) is defined by:

$$\text{ilr}(\mathbf{x}) = \mathbf{V}_D^T \ln(\mathbf{x})$$

where the logarithm of $\mathbf{x}$ is understood componentwise, $\mathbf{V}_D^T$ is a transposed contrast matrix [9] associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

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

Note that such a contrast matrix $\mathbf{V}_D$ of size $D \times (D-1)$ satisfies the following property:

1. $\mathbf{V}_D \mathbf{V}_D^T = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_{D\times D}$ where $\mathbf{I}_D$ is the $D \times D$ identity matrix, $\mathbf{1}_{D\times D}$ is a $D \times D$ matrix of ones.
2. $\mathbf{V}_D^T \mathbf{V}_D = \mathbf{I}_{D-1}$ where $\mathbf{I}_{D-1}$ is the identity matrix with dimension $(D-1)$.
3. $\mathbf{V}_D^T \mathbf{j}_D = \mathbf{0}_{D-1}$ where $\mathbf{j}_D$ is a $D \times 1$ column vectors of ones.

The following $D \times (D-1)$ matrix $\mathbf{V}_D$ defined by Egozcue et al (2003) [10] is an example of contrast matrix for $D = 3$

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

This particular matrix defines the following ilr coordinates

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

The first ilr coordinate contains information about the relative importance of the first component $x_1$ with respect to the geometric mean of the second and the third components $g = \sqrt{x_2 x_3}$. The second ilr coordinate contains information about the relative importance of the second component $x_2$ with respect to the third component $x_3$. In our case, the first ilr coordinate opposes the Left wing to the group of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party. The inverse ilr transformation is given by:

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}(\exp(V_D \mathbf{x}^*)) \text{ for } x^* \in R^{D-1}$$

where the exponential of vector $\mathbf{x}$ is understood componentwise and

$$\mathcal{C}(\mathbf{x}) = \left( x_1 / \sum_{j=1}^{D} x_j, \cdots, x_D / \sum_{j=1}^{D} x_j \right) \text{ is the closure operation.}$$

The vector space structure of the simplex $\mathbf{S}^D$ is defined by the perturbation and powering operations:

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D),\ \mathbf{x},\ \mathbf{y} \in \mathbf{S}^D$$
$$\lambda \odot \mathbf{x} = \mathcal{C}(x_1^{\lambda}, \ldots, x_D^{\lambda}),\ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D.$$

The compositional inner product (C-inner product) of $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{S}^D$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle_c = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \log \frac{x_i}{x_j} \cdot \log \frac{y_i}{y_j} = \sum_{i=1}^{D} \log \frac{x_i}{g(\mathbf{x})} \cdot \log \frac{y_i}{g(\mathbf{y})}$$

where $g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \ldots x_D}$ is the geometric mean of the components.
The compositional distance (C−distance) between $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{S}^D$ is derived from the inner product

$$\begin{aligned} \mathrm{d}_c(\mathbf{x}, \mathbf{y}) &= \left( \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \left( \log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2 \right)^{1/2} \\ &= \left( \sum_{i=1}^{D} \left( \log \frac{x_i}{g(\mathbf{x})} - \log \frac{y_i}{g(\mathbf{y})} \right)^2 \right)^{1/2} \end{aligned}$$

The expected value $\mathbb{E}^{\oplus}\mathbf{Y}$ of a simplex-valued random composition $\mathbf{Y} \in \mathbf{S}^D$ (Pawlowsky [9]) is defined by

$$\operatorname{argmin}_{\mathbf{z} \in \mathbf{S}^D} \mathbb{E}(d_c^2(\mathbf{Y}, \mathbf{z}))$$

and it is equal to

$$\mathbb{E}^{\oplus}\mathbf{Y} = \mathcal{C}(\exp(\mathbb{E} \log \mathbf{Y})) = \mathrm{clr}^{-1}(\mathbb{E}\,\mathrm{clr}(\mathbf{Y})) = \mathrm{ilr}^{-1}(\mathbb{E}\,\mathrm{ilr}(\mathbf{Y})) = \mathrm{ilr}^{-1}(\mathbb{E}\mathbf{Y}^*)$$

where $\mathbf{Y}^* = \mathrm{ilr}(\mathbf{Y})$.

### 3.2 Compositional regression models

The notations used in this paper are summarized in Table 2

| **Variable** | **Notation** | **Coordinates** |
|---|---|---|
| Dependent | $\mathbf{Y}_i = (Y_1, \ldots, Y_L)$ | $\mathrm{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$ |
| Compositional explanatory | $\mathbf{X}_i^{(q)} = (X_{i1}^{(q)}, \ldots, X_{iD_q}^{(q)})$ | $\mathrm{ilr}(\mathbf{X}_{ip}^{(q)}) = \mathbf{X}_{ip}^{(q)*}$ |
| Classical explanatory | $Z_{ik}$ | |
| **General notations** | | |
| $L$ | Number of components of the dependent variable | |
| $i = 1, \ldots, n$ | Index of observations ($n = 95$ ) | |
| $q = 1, \ldots, Q$ | Index of compositional explanatory variables ($Q = 3$) | |
| $p = 1, \ldots, D_q$ | Index of the coordinates for the compositional explanatory variables | |
| $k = 1, \ldots, K$ | Index of classical explanatory variables ($K = 5$) | |

Table 2: Notations

We now describe the CODA regression model. $\mathbf{Y}_i \in \mathbf{S}^L$ denotes the compositional response value of the $i$th observation, and $\mathbf{X}_i^{(q)} \in \mathbf{S}^{D_q},\ q = 1, \ldots, Q$, denotes the value of the $q$th compositional covariate for the $i$th observation, where $\mathbf{Y} \in \mathbf{S}^L$ and $\mathbf{X}^{(q)} \in \mathbf{S}^{D_q},\ q = 1, \ldots, Q$, belong to the simplex spaces $\mathbf{S}^L$ and $\mathbf{S}^{D_q}$. $Z_{ik},\ k = 1, \ldots, K$, denotes the $k$th classical covariate of the $i$th observation. Let $\boxdot$ be the compositional matrix product, which corresponds to the matrix product in the coordinate space through the ilr transformation

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^{D} x_j^{b_{1j}}, \cdots, \prod_{j=1}^{D} x_j^{b_{Lj}}\right)^T$$

where $\mathbf{x} \in \mathbf{S}^D$ and $\mathbf{B} = ((b_{ij})),\ i = 1, \ldots, L,\ j = 1, \ldots, D$, is a parameter matrix such that the column vectors belong to $\mathbf{S}^D$, $\mathbf{j}_L^T\mathbf{B} = \mathbf{0}_D$, $\mathbf{B}\mathbf{j}_D = \mathbf{0}_L$, where $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, and $\mathbf{j}_L^T$ is the transposed of $\mathbf{j}_L$.

Let us first introduce the CODA regression model in the ilr coordinate space as follows:

$$\mathrm{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \mathrm{ilr}(\mathbf{X}_i^{(q)})\mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ik}\mathbf{c}_k^* + \mathrm{ilr}(\boldsymbol{\epsilon}_i) \tag{1}$$

where $\mathrm{ilr}(\mathbf{Y_i})$, $\mathrm{ilr}(\mathbf{X}_i^{(q)})$ are the ilr coordinates of $\mathbf{Y}_i$, $\mathbf{X}_i^{(q)}$ $(q = 1, \ldots, Q)$ respectively; $\mathbf{b_0}^*$, $\mathbf{B}_q^*$, $\mathbf{c}_k^*$ are the parameters in the coordinate space, and $\mathrm{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals. The distributional assumption is that $\mathrm{ilr}(\boldsymbol{\epsilon})$ follows the multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$.
This regression model can be written in the simplex as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}^{(q)} \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_k \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{2}$$

where $\mathbf{b_0}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(Q)}, \mathbf{c}_1, \ldots, \mathbf{c}_K$ are the parameters satisfying $\mathbf{b_0} \in \mathbf{S}^L, \mathbf{B}^{(q)} \in \mathbf{S}^{D_q},\ q = 1, \ldots, Q,\ \mathbf{c}_k \in \mathbf{S}^L,\ k = 1, \ldots, K$, $\mathbf{j}_L^T\mathbf{B}^{(q)} = \mathbf{0}_{D_q}$, $\mathbf{B}^{(q)}\mathbf{j}_{D_q} = \mathbf{0}_L$, The distributional assumption is that $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows the normal distribution on the simplex (see [8]).

It is classical to estimate model (2) using OLS thus assuming the independence between the ilr coordinates. Chen et al (2016) [11] give different formulas relating the parameters in the simplex to the parameters in the coordinate space. Following Chen et al.(2016) [11] (Property 2.1 and Property 2.3(3)), we extend this result to the case of an additional non-compositional covariate: we calculate the sum of the squared norm of the residuals and derive the estimators from the normal equations. We obtain that:

**Theorem 1** *In model (1)-(2), the relationship between the parameters in the simplex and their counterpart in coordinate space is given by*

$$\begin{cases} \mathbf{b_0} = \exp({\mathbf{b_0}^*}^T\mathbf{V}_L) = ilr^{-1}(\mathbf{b_0}^*) \\ \mathbf{B}_q = \mathbf{V}_{D_q}^T\mathbf{B}_q^*\mathbf{V}_L \\ \mathbf{c}_k = \exp(\mathbf{c}_k^*\mathbf{V}_L) = ilr^{-1}(\mathbf{c}_k^*) \end{cases} \tag{3}$$

*where $\mathbf{V}_L$ and $\mathbf{V}_{D_q},\ q = 1, \ldots, Q$ are contrast matrices associated to the selected ilr transformations.*

### 3.3 Impact of compositional and classical explanatory variables

Because the interpretation of the parameters of these models is not so straightforward [12], we rather concentrate on illustrating graphically the relationship between the predicted vote shares and the explanatory variables. The prediction of the dependent variable for the above models are given by (4):

$$\hat{\mathbf{Y}}_i = \hat{\mathbf{b_0}} \bigoplus_{q=1}^{Q} \hat{\mathbf{B}}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ik} \odot \hat{\mathbf{c}}_k \quad i = 1, ..., n \tag{4}$$

where $\hat{\mathbf{b_0}}, \hat{\mathbf{B}}_q$ and $\hat{\mathbf{c}}_k$ are the estimated parameters. We can rewrite (4) as

$$\hat{\mathbf{Y}}_i = \mathcal{C}\left(\hat{\mathbf{b_0}}.\left(\prod_{q=1}^{Q} \mathbf{X_i^{(q)}}^{\hat{B}_q}\right).\left(\prod_{k=1}^{K} \hat{\mathbf{c}}_k^{Z_{ik}}\right)\right) \quad i = 1, ..., n \tag{5}$$

In order to illustrate these formulas, we will focus on graphing the predicted values of the dependent variable as a function of one specific variable of interest: two cases must be considered depending on whether the specific variable is classical or compositional. In both cases, we will create a grid of potential values of the specific explanatory and fix the other explanatory variables at the values they take for one selected point of the dataset (we repeat for several selected points). For the sake of simplicity let us take $L = 3$.

For the case when the specific variable is a classical covariate $Z_{ik}$ , from (5) there exists $\hat{\mathbf{a_0}} \in \mathbf{S}^L$ (this term contains the impacts of all other explanatory but is constant when $Z_{ik}$ alone varies) such that

$$\hat{\mathbf{Y}}_i = \hat{\mathbf{a_0}} \bigoplus Z_{ik} \odot \hat{\mathbf{c}} = \mathcal{C}\left(\hat{a}_{01}\hat{c}_1^{Z_{ik}}, \cdots, \hat{a}_{0L}\hat{c}_L^{Z_{ik}}\right)$$

With $T = \hat{b}_{01}\hat{c}_1^{Z_{ik}} + \cdots + \hat{b}_{0L}\hat{c}_L^{Z_{ik}}$, we get

$$\hat{Y}_{i1} = \frac{\hat{a}_{01}\hat{c}_1^{Z_{ik}}}{T};\ \hat{Y}_{i2} = \frac{\hat{a}_{02}\hat{c}_2^{Z_{ik}}}{T}; \cdots ; \hat{Y}_{iL} = \frac{\hat{a}_{0L}\hat{c}_L^{Z_{ik}}}{T}.$$

Now for the case when the specific variable is a compositional variable $\mathbf{X_i^{(q)}}$, let us take for the sake of simplicity $D_q = 3$. As before, from (5), there exists $\hat{\mathbf{a_0}} \in \mathbf{S}^L$ (this term contains the impacts of all other explanatory but is constant when $X_i^{(q)}$ alone varies) such that

$$\begin{aligned}\hat{\mathbf{Y}}_i &= \hat{\mathbf{a_0}} \bigoplus \mathbf{X_i^{(q)}}^{\hat{B}_q} \\ &= \mathcal{C}(\hat{a}_{01} {X_{i1}^{(q)}}^{\hat{b}_{11}^{(q)}} {X_{i2}^{(q)}}^{\hat{b}_{12}^{(q)}} {X_{i3}^{(q)}}^{\hat{b}_{13}^{(q)}}, \hat{a}_{02} {X_{i1}^{(q)}}^{\hat{b}_{21}^{(q)}} {X_{i2}^{(q)}}^{\hat{b}_{22}^{(q)}} {X_{i3}^{(q)}}^{\hat{b}_{23}^{(q)}}, \\ &\quad \hat{a}_{03} {X_{i1}^{(q)}}^{\hat{b}_{31}^{(q)}} {X_{i2}^{(q)}}^{\hat{b}_{32}^{(q)}} {X_{i3}^{(q)}}^{\hat{b}_{33}^{(q)}})\end{aligned}$$

We now fit a CODA regression model describing the impacts of socio-economic factors on vote shares in the 2015 French departmental election.
After including all explanatory variables from our data set in the regression model, and eliminating one by one the variables which are not significant, we obtain the

results in Table 3. This model shows that the age of people, the proportion of people who own assets, the proportion of foreigners do not have any impact on the vote shares. However, the levels of education, the working areas, the unemployment rate and the proportion of people who pay income tax really affect the result of the French departmental election in 2015.

| | *Dependent variable:* | |
|---|---|---|
| | y_ilr[, 1] | y_ilr[, 2] |
| Diplome_ilr1 | $-2.06(0.54)^{***}$ | $-1.51(0.46)^{**}$ |
| Diplome_ilr2 | $-1.28(0.80)$ | $-2.07(0.67)^{**}$ |
| Employ_ilr1 | $-0.05(0.30)$ | $-2.12(0.34)$ |
| Employ_ilr2 | $+0.12(0.37)$ | $-2.62(0.46)^{**}$ |
| Employ_ilr3 | $+0.30(0.30)$ | $-2.12(0.34)$ |
| Employ_ilr4 | $+0.13(0.11)$ | $-2.62(0.46)^{\cdot}$ |
| unemp_rate | $-7.65(3.16)^{*}$ | $-2.12(0.34)^{***}$ |
| income_tax_rate | $+2.04(1.37)$ | $-2.62(0.46)^{***}$ |
| Constant | $-2.324(1.15)^{*}$ | $-4.80(0.97)^{***}$ |
| $R^2$ | 0.30 | 0.62 |
| Adjusted $R^2$ | 0.23 | 0.59 |
| Residual Std. Error (df = 86) | 0.30 | 0.26 |
| F Statistic (df = 8; 86) | $4.602^{***}$ | $17.85^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 3: Regression with compositional and classical variables

In order to illustrate the impact of unemployment on the predicted shares, we choose three departments Ariège, Cantal and Bas-Rhin with different characteristics: Ariège has the maximum Left wing share, Cantal the maximum Right wing share and Bas Rhin has the minimum Left wing share. We fix the values of the covariates at the values of each of the three departments and create a grid of fictive values of unemployment rates. Figure 3 shows the predictions of vote shares in these departments Ariège, Cantal and Bas-Rhin as a function of unemployment rate (its minimum and maximum in the data base are figured by the dotted vertical lines). We first of all see the non linear nature of the relationship, and the fact that they differ from one department to the other. Note that the predicted shares using this model satisfy the constraint of unit sum and it clearly shows on the graph. In all cases, when the unemployment rate increases up to a given threshold of around 15%, the Left wing and the Extreme Right party gain votes at the expense of the Right wing. However, if unemployment keeps increasing beyond 15%, the Left wing starts loosing votes while the Right wing keeps decreasing and the Extreme Right keeps increasing. Overall, we can say that as unemployment rate varies, the Left wing proportion is more stable than the other two parties and that the other two parties affect each other like interconnecting pipes. Even though the three departments curves have the same general shape, we note differences: the maximum of the Left wing share is the highest in Ariège and lowest in Bas-Rhin; it is striking that the point at which the Left wing share and the Right wing share are equal is obtained at approximately the same value of unemployment rate in the three department but corresponds to different values

of the common Left wing- Right wing share; this value is lower than the maximum Left wing share in Ariège whereas it is slightly higher in Bas-Rhin. A major difference between the three departments is revealed when one looks at the highest of the three predictions: in Ariège, all realistic scenarios (between two vertical lines) result in a victory of the Left wing, in Cantal, all three parties may win depending on the value of unemployment and finally in Bas-Rhin there is no scenario leading to a victory of the Left wing . To represent this differently, we plot on Figure 4 a ternary diagram showing the curve of predicted shares as a function of unemployment rates together with a small square figuring the observed position of the given department in the triangle and a small diamond the corresponding prediction on the curve. The curve, a line in the simplex, is colored according to the value of unemployment rate. The Cantal department is better predicted than the Ariège and Bas-Rhin departments. We also note that the maximum predicted proportion for the Left wing is lower in the Bas-Rhin than in the other two departments. Finally, the triangle is divided in three parts with respect to the highest shares to highlight the winning party as in Figure 3.

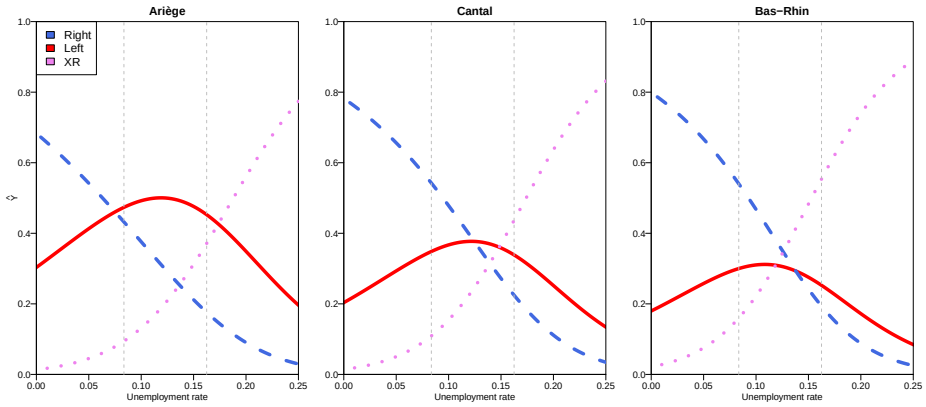


Figure 3: The vote share prediction curves in three departments: Ariège, Cantal, Bas-Rhin respectively (the grey dotted line show the minimum and the maximum observed unemployment rates)

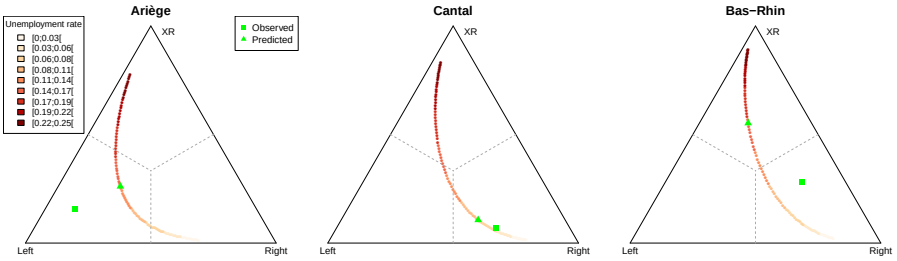


Figure 4: The vote share prediction ternary diagrams for fixed covariates given by three departments: Ariège, Cantal, Bas-Rhin respectively as a function of the unemployment rate. The green squares show the observed vote share of these departments and the green triangles on the red curve the corresponding predictions.

Let us now turn attention to the case of a compositional explanatory variable impact. Figure 5 presents the vote share predictions according to the Diploma variable in the same three departments (Ariège, Cantal and Bas-Rhin). The principle is the same: all explanatory variables are fixed to the value of the given department except Diploma. We create of grid in the Diploma triangle and compute the predicted shares at each point of this grid. However since it is impossible to plot a function from the simplex to the simplex, we choose to summarize the predicted shares by the winning party (corresponding to the highest share) and color the triangle in the Diploma space according to the winning party color. The observed shares are also figured by black points in this ternary diagram thus showing the realistic values. This figure shows that there is a large proportion of fictive situations (in terms of diploma proportions) where the Left party would win.

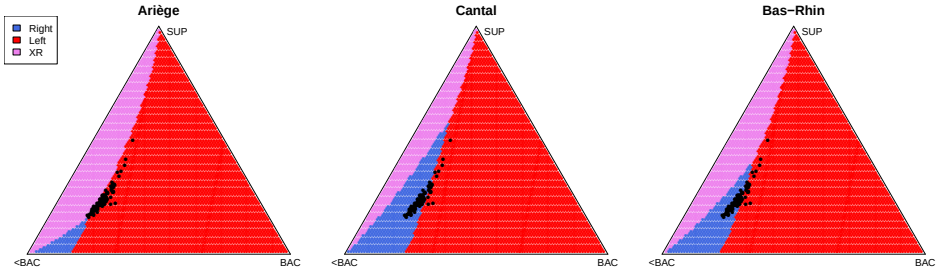


Figure 5: Predictions of vote shares according to Diploma for fixed covariates given by the departments Arige (left plot), Cantal (middle plot) and Bas-Rhin (right plot)

## 4 Conclusion

The above analysis demonstrates that the CODA regression models can be useful in the context of political economy. We analyze how the predicted values in these models vary with the predictors and propose new graphical tools to explore the impact of some socio-economic variables on election results. Our future perspectives are to introduce the geographical dimension in the model and to use the Student distribution (Katz and King, 1999 [7]) instead of the normal distribution. At last, we plan to compute the elasticities as in [12] to characterize the impacts of the covariates in a more quantitative way.

Acknowledgments. We thank professor Le Breton for introducing us to this topic of political science and for nice discussions.

## References

1. J. B. Lewis and D. A. Linzer, Estimating Regression Models in which the Dependent Variable is based on Estimates, Political Analysis, 13, 345-364 (2005).
2. L. Beauguitte and C. Colange, Analyser les comportements électéraux à l'échelle du bureau de vote, Mémoire scientifique halshs, 85 (2013).
3. A. Kavanagh, S. Fotheringham, M. Charlton and R. Sinnott, A Geographically Weighted Regression Analysis of Election Specific Turnout in the Republic of Ireland, 14, University College Cork (2006).
4. J. Morais, C. Thomas-Agnan and M. Simioni, Using compositional and Dirichlet models for market share regression, Journal of Applied Statistics, al-01558527 (2017).
5. T. H. Trinh and J. Morais, Impact of socioeconomic factors on nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis, To appear in Statistical Methods in Medical Research.
6. J. Honaker, J. Katz and G. King, A fast, Easy and Efficient Estimator for Multiparty Electoral Data, Political Analysis, 10, 1 (2002).
7. J. Katz and G. King, A statistical Model for Multiparty Electoral Data, The American Political Science Review, 93 (1999).
8. J. Aitchison, A General Class of Distributions on the Simplex, Royal Statistical Society (1985).
9. G. Pawlowsky-Glahn, J. J. Egozcue and R. Tolosana-Delgado, Modeling and Analysis of Compositional Data, Wiley (2015).
10. J. J. Egozcue, G. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barcelo-Vidal, Isometric Logratio Transformations for Compositional Data Analysis, Mathematical Geology, 35, 3, 279-300 (2003).
11. J. Chen, X. Zhang and S. Li, Multiple linear regression with compositional response and covariates, Journal of Applied Statistics, 44, 12, 2270-2285 (2017).
12. J. Morais, C. Thomas-Agnan and M. Simioni, Interpretation of explanatory variables impacts in compositional regression models, HAL, submitted.
13. M. C. Mert, P. Filzmoser, G. Endel and I. Wilbacher, Compositional data analysis in epidemiology, Sage journal, 27, 6, 1878-1891 (2016).