JMS Journées de méthodologie statistique de l'Insee 2018

CODA METHODS FOR ANALYZING THE IMPACTS OF SOCIO-ECONOMIC FACTORS ON FRENCH DEPARTMENTAL ELECTIONS

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

(*) Toulouse School of Economic, UT1

(**) Department of Economic, University of Architecture, Da Nang, Vietnam

thihuongan.nguyen@tse-fr.eu

Keywords. political economy, compositional regression models

Résumé

Les proportions de votes par parti sur une subdivision du territoire forment un vecteur de données dites de composition (mathématiquement, un vecteur appartenant à un simplexe). Il est intéressant de modéliser ces proportions en étudiant l'impact des caractéristiques des unités territoriales sur l'issue des élections. Dans la littérature d'économie politique, il existe des modèles de régression qui sont restreints généralement au cas de deux partis politiques. Dans la littérature statistique, il existe des modèles de régression adaptés à des vecteurs de parts dont les modèles CODA (pour "COmpositional Data Analysis"), mais aussi les modèles de Dirichlet, les modèles de Student et d'autres. Notre objectif est d'utiliser les modèles de régression de type CODA pour généraliser les modèles d'économie politique à plus de deux partis. Les modèles sont ajustés sur des données électorales françaises des élections départementales de 2015.

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, there are regression models that 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-style 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.

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 vectors of votes shares for the French departmental election in 2015. The explanatory variables include some compositional and continuous socio-economic variables.

Some papers concentrate on the relationship between socio-economic variables and election results. Russo et Beauguitte (2012) [2] study a linear regression at three levels of aggregation (polling stations, cities and electoral districts) 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 or for each electoral division. Besides, there are some papers which use compositional data analysis in regression model (CODA model) where the dependent and independent variables may be compositional variables. Morais et al. (2017) [4] apply a CODA model to study the impact of media investments on brand's market shares. Trinh and Morais (2017) [5] use a CODA 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.

1 Data

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

Variable	Description
Vote share	Left(L), Right(R), Extreme Right(XR)
Age	Age_1840, Age_4064, Age_65.
Diplome	≤CAPBEP, BAC, Dip_SUP.
Employment	AZ, BE, FZ, GU, OQ
unemp_rate	the unemployment rate
employ_evol	Mean evolution of the yearly employment rate (2009-2014)
owner_rate	The rate of people who own assets
income_rate	The rate of people who have a salary
foreign	The rate of foreigners

Table 1: Data description

Employment has five categories: AZ (Agriculture, pêche), BE (Industrie manufacture, industrie extractive et autres), FZ (Construction), GU (Commerce, transport et service divers) and OQ (Administration publique, enseignement, santé humaine).

From the point of view of CODA, compositional data can be represented in a ternary diagram if they have three components. For instance, the vote shares of Left, Right and Extreme Right are the blue points in Figure 1. The red triangle shows that the vote shares of the Left, the Right and the Extreme Right are respectively 17.4%, 54.6%, and 28% .

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

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

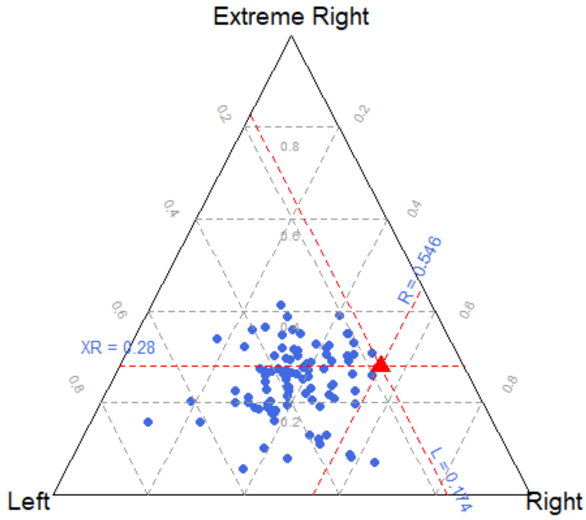

Figure 1: Explanation of vote share data

2 Compositional data analysis approach

A CODA model is proposed where the dependent variable is a compositional variable and the independent variables are compositional variables or classical continuous variables or a mixture of both. This model is based on the log-ratio transformation approach. A composition $\mathbf{x}$ is a vector of D parts of some whole which carries relative information. The D-composition $\mathbf{x}$ lies in the simplex $\mathbf{S}^D$:

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

The simplex $\mathbf{S}^D$ can be equipped with the Aitchison [8] inner product Pawlowsky [9]. Classical regression models cannot be used directly in the simplex because of the constraints that the components are positive and sum up to 1. 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). We focus here on the ilr because the coordinates in the clr transformed vector are linearly dependent, the coordinates in the alr transformed vector are not compatible with the geometry (distance between the components in the simplex space is different from the coordinates in the Euclidean space). Thus, we will use the ilr transformation in this paper.
The 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 the logarithm of its components, $\mathbf{V}_D^T$ 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 a contrast matrix $\mathbf{V}_D$ of size $D \times (D-1)$ satisfies

1. $\mathbf{V}_D \mathbf{V}_D^T = \mathbf{G}_D = \mathbf{I}_D - \frac{1}{D} \mathbf{1}_{D \times D}$ 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 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 one.

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

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

This choice of matrix defines the following ilr coordinates

$$\begin{aligned} \mathrm{ilr}_1(\mathbf{x}) &= \frac{1}{\sqrt{6}}(2\ln x_1 - \ln x_2 - \ln x_3) = \frac{2}{\sqrt{6}} \ln \frac{x_1}{\sqrt{x_2 x_3}} \\ \mathrm{ilr}_2(\mathbf{x}) &= \frac{1}{\sqrt{2}}(\ln x_2 - \ln x_3) = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3} \end{aligned}$$

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 party to the group of the Right and Extreme Right parties and the second opposes the Right to the Extreme Right. The ilr inverse transformation is given by:

$$\mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}(\exp(V_D \mathbf{x}^*))$$

where the exponential of vector $\mathbf{x}$ is the exponential of its coordinates and $\mathcal{C}(\mathbf{x}) = \left(\frac{x_1}{\sum_{j=1}^D x_j}, \cdots, \frac{x_D}{\sum_{j=1}^D x_j} \right)$ is the closure operation.

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

$$\begin{aligned} \mathbf{x} \oplus \mathbf{y} &= \mathcal{C}[x_1 y_1, ..., x_D y_D], \ \mathbf{x}, \ \mathbf{y} \in \mathbf{S}^D \\ \lambda \odot \mathbf{x} &= \mathcal{C}[x_1^\lambda, ..., x_D^\lambda], \ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D \end{aligned}$$

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 $g(\mathbf{x}) = \sqrt[D]{x_1 x_2 ... x_D}$.

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

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

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

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

It is equal to

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

2.1 Compositional regression models

The notations used in this paper are standardized in Table 2

Variable	Shares	Coordinates
Dependent	$\mathbf{Y}_i = (Y_1, ..., Y_L)$	$\text{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$
Compositional explanatory	$\mathbf{X}_{qi} = (X_{1i}, ..., X_{D_{qi}})$	$\text{ilr}(\mathbf{X}_q) = \mathbf{X}_{qi}^*$
Continuous explanatory	Z_{ki}	
General notations		
L	Number of components of the dependent variable	
$i = 1, ..., n$	Index of observations ($n = 95$)	
$q = 1, ..., Q$	Index of compositional explanatory variables ($Q = 3$)	
$k = 1, ..., K$	Index of continuous explanatory variables ($K = 5$)	

Table 2: Notations

We introduce a regression model which describes the impacts of socio-economic factors on vote shares in the French departmental election in 2015.
$\mathbf{Y} \in \mathbf{S}^L$ and $\mathbf{X}_q \in \mathbf{S}^{D_q}$ belong to the simplex spaces:

$$\mathbf{S}^L = \left\{ \mathbf{Y} = (Y_1, ..., Y_L) : Y_j > 0, j = 1, ..., L; \sum_{j=1}^{L} Y_j = 1 \right\},$$

$$\mathbf{S}^{D_q} = \left\{ \mathbf{X}_q = (X_{q1}, ..., X_{qD_q}) : X_{qp} > 0;\ p = 1, ..., D_q;\ \sum_{p=1}^{D_q} X_{qp} = 1 \right\},\ q = 1, ..., Q.$$

Let $\boxdot$ be the compositional matrix product, which corresponds through the ilr transformation to the matrix product in the Euclidean geometry

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

where $\mathbf{B}$ is a parameter matrix according to the compositional explanatory variables, $\mathbf{Y}_i \in \mathbf{S}^L$ denotes the compositional response value of the ith observation, $\mathbf{X}_{qi} \in \mathbf{S}^{D_q},\ q = 1, ..., Q$, denotes the qth compositional covariate of the ith observation, $Z_{ki},\ k = 1, ..., K$, denotes the kth classical continuous covariate of the ith observation.

The regression model in the simplex can be written as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}_q \boxdot \mathbf{X}_{qi} \bigoplus_{k=1}^{K} Z_k \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, ..., n \tag{1}$$

where $\mathbf{b_0}, \mathbf{B}_1, ..., \mathbf{B}_Q, \mathbf{c}_1, ..., \mathbf{c}_K$ are the parameters satisfying $\mathbf{b_0} \in \mathbf{S}^L$, $\mathbf{B}_q \in \mathbf{S}^{D_q}$, $\mathbf{c}_k \in \mathbf{S}^L$, $q = 1, \cdots, Q$, $k = 1, \cdots, K$, $\mathbf{j}_L^T \mathbf{B}_q = \mathbf{0}_{D_q}$, $\mathbf{B}_q \mathbf{j}_{D_q} = \mathbf{0}_L$, $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, $\mathbf{j}_L^T$ is the transposition of $\mathbf{j}_L$; $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows the normal distribution on the simplex.

This regression model can be rewritten in the ilr coordinate space as

$$\text{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \text{ilr}(\mathbf{X}_{qi}) \mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ki} \mathbf{c}_k^* + \text{ilr}(\boldsymbol{\epsilon}_i) \tag{2}$$

where $\text{ilr}(\mathbf{Y_i}), \text{ilr}(\mathbf{X}_{qi})$ are the ilr coordinates of $\mathbf{Y}_i, \mathbf{X}_{qi}$ $(q = 1, ..., Q)$ respectively; $\mathbf{b_0}^*, \mathbf{B}_q^*, \mathbf{c}_k^*$ are the parameters in the coordinate space, and $\text{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals. The CODA regression

model consists in assuming that ilr($\boldsymbol{\epsilon}$) follows the multivariate normal distribution with zero mean and covariance matrix $\mathbf{\Sigma}$.
It is classical to estimate model (2) using OLS and 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. We generalize this result to the case of an additional non-compositional covariate:

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

Because the interpretation of the parameters of these models is quite complex [4], we turn attention to understand the relationship between the predicted vote shares and the explanatory variables. The prediction for the above models are given by (4):

$$\hat{\mathbf{Y}}_i = \hat{\mathbf{b}}_\mathbf{0} \bigoplus_{q=1}^{Q} \hat{\mathbf{B}}_q \boxdot \mathbf{X}_{qi} \bigoplus_{k=1}^{K} Z_{ki} \odot \hat{\mathbf{c}}_k \quad i = 1, ..., n \tag{4}$$

where $\hat{\mathbf{b}}_\mathbf{0}, \hat{\mathbf{B}}_q$ and $\hat{\mathbf{c}}_k$ are the estimated parameters. We can rewrite (4) as

$$\hat{\mathbf{Y}}_i = \mathcal{C}\left[\hat{\mathbf{b}}_\mathbf{0}.(\prod_{q=1}^{Q} \mathbf{X_{qi}}^{\hat{B}_q}).(\prod_{k=1}^{K} \hat{\mathbf{c}}_k^{Z_{ki}})\right] \quad i = 1, ..., n \tag{5}$$

For example, with a single classical variable Z_i, we have

$$\begin{aligned} \hat{\mathbf{Y}}_i &= \mathcal{C}(\hat{\mathbf{b}}_\mathbf{0}\hat{\mathbf{c}}^{Z_i}) \\ &= \mathcal{C}\left(\hat{b}_{01}\hat{c}_1^{Z_i}, \cdots, \hat{b}_{0L}\hat{c}_L^{Z_i}\right) \end{aligned}$$

where $\mathcal{C}(\hat{\mathbf{b}}_\mathbf{0}\hat{\mathbf{c}}^{Z_i}) = \hat{\mathbf{b}}_\mathbf{0} \oplus \hat{\mathbf{c}}^{Z_i}$; $\hat{\mathbf{b}}_\mathbf{0}$, $\hat{\mathbf{c}}^{Z_i}$, $\hat{\mathbf{Y}}_i \in \mathbf{S}^L$. With $T = \hat{b}_{01}\hat{c}_1^{Z_i} + \cdots + \hat{b}_{0L}\hat{c}_L^{Z_i}$, we get

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

2.2 Impact of one explanatory variable

2.2.1 Case of a classical explanatory variable

To model the vote shares of the French departmental election as described in Section 1, we use the above methodology with a simple regression model which contains one continuous explanatory covariate Z (unemp_rate or income_rate), that is $K = 1$ and $L = 3$. The results are in Table 3

	Dependent variable:			
	y_ilr[, 1]	y_ilr[, 2]	y_ilr[, 1]	y_ilr[, 2]
	(1)	(2)	(3)	(4)
unemp_rate	-6.422^{***}	12.739^{***}		
	(1.956)	(1.977)		
income_rate			1.350^{**}	-1.176
			(0.612)	(0.712)
Constant	0.787^{***}	-1.859^{***}	-0.712^{**}	0.285
	(0.232)	(0.234)	(0.340)	(0.396)
Observations	95	95	95	95
R^2	0.104	0.309	0.050	0.028
Adjusted R^2	0.094	0.301	0.039	0.018
Residual Std. Error (df = 93)	0.330	0.333	0.339	0.395
F Statistic (df = 1; 93)	10.782^{***}	41.540^{***}	4.862^{**}	2.726
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: Regression with a continuous explanatory variable

In the first model including unemployment rate, the constant and the covariate are significant for the two ilr coordinates of the dependent variable. In the second model including income rate, the covariate and the constant are not significant to explain the second ilr coordinate. For the first model including unemployment rate, we get the estimated parameters using the result of Table 3 combined with formulas (3) as below

	Left	Right	Extreme Right
Intercept	2.367e-01	7.208e-01	0.042
Unemployement	1.570e-05	1.786e-09	0.999

Table 4: Parameters in the simplex

From (5) we have that

$$\hat{\mathbf{Y}}_i = \mathcal{C}(\hat{\mathbf{b}}_{\mathbf{0}}\hat{\mathbf{c}}^{Z_i}) = \mathcal{C}\left(\hat{b}_{01}\hat{c}_1^{Z_i}, \hat{b}_{02}\hat{c}_2^{Z_i}, \hat{b}_{03}\hat{c}_3^{Z_i}\right)$$

If we let $T = \hat{b}_{01}\hat{b}_1^{Z_i} + \hat{b}_{02}\hat{c}_2^{Z_i} + \hat{b}_{03}\hat{c}_3^{Z_i}$, then we get

$$\hat{Y}_{i1} = \frac{\hat{b}_{01}\hat{c}_1^{Z_i}}{T};\ \hat{Y}_{i2} = \frac{\hat{b}_{02}\hat{c}_2^{Z_i}}{T};\ \hat{Y}_{i3} = \frac{\hat{b}_{03}\hat{c}_3^{Z_i}}{T}.$$

We can now use formula (5) to obtain the predictions and we plot them on Figure 2.
We do the same for the second model including the income rate. The CODA regression model is such that the predicted shows sum to one for each value of the explanatory variable. The plot emphasizes the fact that the relationships are neither linear nor monotone. For example, as the unemployment rate increases, the Left party vote share first increases and then decreases: as unemployment rate increases the Extreme Right party first picks up votes to the Right party and after a threshold of 15% to the Left party as well.

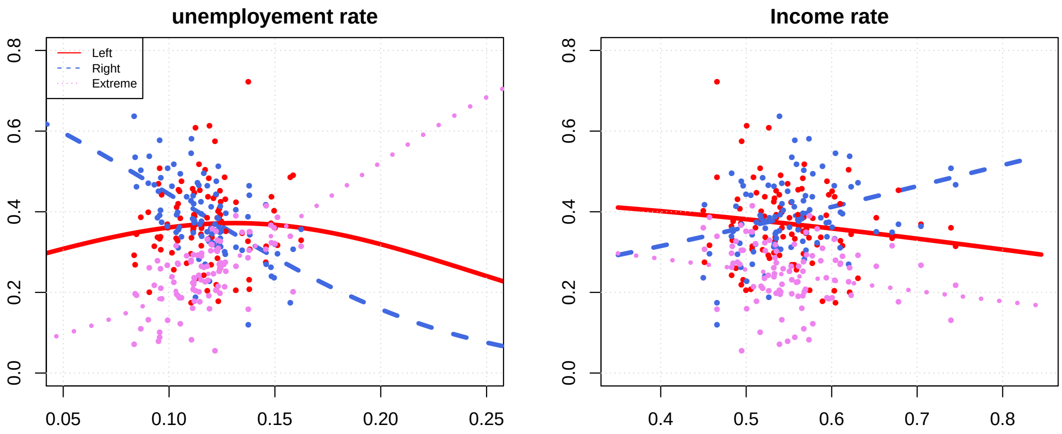

Figure 2: Curves of predictions

2.2.2 Case of a compositional explanatory variable

We use the above methodology with a simple regression model which contains one compositional explanatory covariate X (Diplome). We use the contrast matrix $\mathbf{V}_3$ as in Section 2 to transform the variable Diplome from the simplex space into the ilr coordinates space. Thus, the first coordinate Diplome_ilr1 contains information about the relative importance of Dip_SUP with respect to the geometric mean of BAC and ≤CAPBEP. The second coordinate Diplome_ilr2 contains information about the relative importance of BAC with respect to ≤CAPBEP. The regression results are in Table 5

	Dependent variable	
	y_ilr[, 1]	y_ilr[, 2]
Diplome_ilr1	$-0.857(0.34)^{*}$	$-2.117(0.341)^{***}$
Diplome_ilr2	$-1.039(0.46)^{*}$	$-2.619(0.458)^{***}$
Constant	$-0.969(0.40)^{*}$	$-2.856(0.399)^{***}$
R^2	0.064	0.299
Adjusted R^2	0.044	0.294
Residual Std. Error (df = 92)	0.338	0.337
F Statistic (df = 2;92)	3.168^{**}	19.67^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 5: Regression with a compositional explanatory variable

In this model including two ilr coordinates Diplome_ilr1 and Diplome_ilr2, the two coordinates are significant. We get back to the estimated parameters in the simplex space by using (3) and we get

	Left	Right	Extreme Right
Intercept	0.788	0.200	0.012
≤CAPBEP	1.876	-0.676	-1.201
BAC	0.342	-0.131	-0.211
Dip_SUP	-2.218	0.806	1.412

The predictions of the vote shares by departments are calculated with

$$\hat{Y}_{Li} = 0.788 * (\leq \text{CAPBEP}_i)^{1.876} * (\text{BAC}_i)^{0.342} * (\text{Dip\_SUP}_i)^{-2.218} / TD$$
$$\hat{Y}_{Ri} = 0.2 * (\leq \text{CAPBEP}_i)^{-0.676} * (\text{BAC}_i)^{-0.131} * (\text{Dip\_SUP}_i)^{0.806} / TD$$
$$\hat{Y}_{XRi} = 0.012 * (\leq \text{CAPBEP}_i)^{-1.201} * (\text{BAC}_i)^{-0.211} * (\text{Dip\_SUP}_i)^{1.412} / TD$$

where L (respectively R and XR) denotes the Left (respectively the Right and Extreme Right) party and

$$TD = \sum_{j=L,R,XR} \hat{B}_{0j} (\leq \text{CAPBEP}_i)^{\hat{B}_{1j}} (\text{BAC}_i)^{\hat{B}_{2j}} (\text{Dip\_SUP}_i)^{\hat{B}_{3j}}$$

and are plotted on Figure 3. The people who have no diploma or CAPBEP tend to vote for the Left party, the people who have BAC tend to vote for the Extreme Right, and the people who have a higher diploma tend to vote for the Right party.

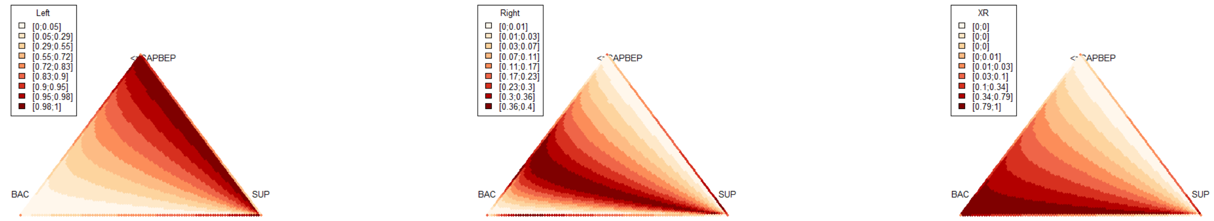

Figure 3: Predictions of vote shares according to Diploma

2.3 Impact of compositional and classical explanatory variables

In this section, we will include all of explanatory variables from our data set in the regression model, and we eliminate one by one the variables which are not significant. The result is in Table 6. This model shows that the age of people, the rate of people who own assets, the rate of foreigners do not have any impact on the vote shares. However, the levels of education, the working areas, the unemployment rate and the rate of people who have a salary 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_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 6: Regression with compositional and classical variables

3 Conclusion

The above analysis demonstrates that the CODA regression models can be useful in the context of political economy. We develop a prediction formula for these models. One of the perspective is to introduce the geographical dimension. Another perspective is to use the logistic Student distribution (Katz and King, 1999 [7]) instead of the logistic normal distribution. Moreover, we plan to compute the elasticities to characterize the impacts of the covariates.

References

[1] J. B. Lewis and D. A. Linzer, “Estimating Regression Models in which the Dependent Variable is based on Estimates,” *Political Analysis*, vol. 13, no. 4, pp. 345–364, 2005. [Online]. Available: `http://www.jstor.org/stable/25791822`.

[2] L. Beauguitte and C. Colange, “Analyser les comportements éléctoraux à l’échelle du bureau de vote,” Mémoire scientifique halshs-00839899, 2013. [Online]. Available: `https://halshs.archives-ouvertes.fr/halshs-00839899`.

[3] A. Kavanagh, S. Fotheringham, M. Charlton, and R. Sinnott, “A Geographically Weighted Regression Analysis of Election Specific Turnout in the Republic of Ireland,” University College Cork, Oct. 2006.

[4] J. Morais, C. Thomas, and M. Simioni, “Using compositional and Dirichlet models for market share regression,” *Journal of Applied Statistics*, 2017. DOI: `https://doi.org/10.1080/02664763.2017.1389864`.

[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*, vol. 10:1, 2002.

[7] J. Katz and G. King, “A Statistical Model for Multiparty Electoral Data,” *The American Political Science Review,*, vol. 93, 1999.

[8] J. Aitchison, “A General Class of Distributions on the Simplex,” *Royal Statistical Society*, 1985.

[9] G. Pawlowsky-Glahn, *Modeling and Analysis of Compositional Data*. Wiley, 2015.

[10] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, “Isometric Logratio Transformations for Compositional Data Analysis,” *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[11] J. Chen, X. Zhang, and S. Li, “Multiple linear regression with compositional response and covariates,” *Journal of Applied Statistics*, 2016. DOI: `10.1080/02664763.2016.1157145`.