WORKING PAPERS

N° 1410

February 2023

"Discrete and Smooth Scalar-on-Density Compositional Regression for Assessing the Impact of Climate Change on Rice Yield in Vietnam."

Michel Simioni, Christine Thomas-Agnan and Huong Trinh Thi

Toulouse School of Economics

Discrete and smooth scalar-on-density compositional regression for assessing the impact of climate change on rice yield in Vietnam.

Huong Trinh Thi[a], Michel Simioni[b,c], Christine Thomas-Agnan[c]

[a] *Faculty of Mathematical Economics Thuongmai University Hanoi Vietnam*
[b] *MoISA INRAE University of Montpellier France*
[c] *Toulouse School of Economics University of Toulouse Capitole Toulouse France*

Abstract

We compare several approaches to scalar-on-density regression. With a discrete point of view, the densities can be viewed as histograms whose frequency vectors belong to a simplex $\mathcal{S}^D$ and then classical compositional regression can be used. An alternative with a functional point of view is to consider density functions as infinite dimensional compositional objects, elements of the so-called Bayes space $\mathcal{B}^2$, and then compositional scalar-on-density regression can be performed. In the second approach, since the density covariate data is originally available as an histogram, these first need to be sent to $\mathcal{B}^2$ using a smoothing step performed by CB-splines smoothing. It is then interesting to investigate the potential advantage of the smooth approach with respect to the discrete one. We compare them through an application about the assessment of the impact of climate change on rice yield in Vietnam, where density covariates are the distributions of maximum daily temperatures during 30 years, from 1987 to 2016, in 63 Vietnamese provinces. Additional covariates such as precipitation, regional dummies and a time trend are added to both models. Scenarios of climate change are modelled with perturbations of the initial density by a chosen change direction producing a shift of the densities towards higher temperatures. The impact on rice yield is then obtained in both models by computing a simple inner product, in $\mathcal{S}^D$ and respectively $\mathcal{B}^2$, of the parameter of the density covariate with the change direction. The comparison shows that the smooth approach outperforms the discrete one by a better evaluation of the phenomenon scale which the discrete approach may fail to uncover.

Keywords: compositional scalar-on-density regression, Bayes space, compositional splines, climate change, rice yield, Vietnam.

1. Introduction

With the increasing complexity of recorded data, one finds nowadays models involving more elaborate data objects such as random densities. We are focusing

here on regression models where such density objects appear as explanatory variables.

It is often the case that density are recorded in an aggregated fashion as histograms (see e.g. Carter et al., 2018, for a survey of applications in climate change econometrics). With this discrete approach, the sample space can be described by the set of vectors of bin frequencies, which have positive components adding up to one. These vectors are called compositions and their space is called a simplex. The simplex can also be endowed with an Euclidian structure as is done in compositional data analysis, see Aitchison (1986) or Pawlowsky-Glahn et al. (2015) for an introduction. Scalar-on-composition regression models using the simplex representation are described for example in Hron et al. (2012). They are constructed by transforming the simplex vectors into vectors of an unconstrained linear space $\mathbb{R}^k$ (for some adapted k) and using on the right hand side of the regression equation the inner product of the resulting vectors with a parameter vector in this space. It is possible to use for example the so-called centered log-ratio transformation.

On the other hand, Petersen et al. (2022) reviews the different approaches for building regression models involving samples of probability density functions with a functional perspective. In the world of functional data analysis, densities are particular functions due to the constraints they must satisfy. For one-dimensional densities, the sample space is taken to be the space $\mathcal{D}$ of functions with positive values and integral equal to one. One of the two main approaches described by Petersen et al. (2022) uses the representation of densities in the so-called Bayes spaces $\mathcal{B}^2$. Bayes spaces were introduced in Van den Boogaart et al. (2014) by endowing the space $\mathcal{D}$, for densities with a finite support $[a, b]$, with a Hilbert space structure. Functional scalar-on-density regression models using this representation are presented for example in Talská et al. (2021). However if the data densities are initially recorded as histograms, one then needs a preliminary step allowing to represent these histograms as smooth functions. Machalová et al. (2021) propose a new class of splines, which they call compositional splines or CB-splines, to approximate probability density from histogram data taking their constraints into account.

Note that the second approach, which we call the smooth approach in the sequel, uses the same kind of strategy to construct the sample space: the Bayes space and its operations can be viewed as continuous versions of the simplex and its operations. Similarly, the scalar-on-density regression model uses a transformation of the density functions called the functional centered log-ratio which is the functional counterpart of the classical centered log-ratio transformation for vectors of a simplex.

However the functional (smooth) approach is more complex to implement and a natural question in practice is to assess the possible advantage that one gets by using the functional model. Our objective in this work is to explore this comparison with an original application to the study of the impact of climate change on rice yield in Vietnam.

Using regression models to relate agricultural yield and climate descriptors is not new, see Fisher (1924). Climate change affects the different stages along

food systems (food production, storage, processing, distribution, retail and consumption) directly and indirectly (Davis et al., 2021). Due to its direct exposition to weather conditions, crop production is all the more sensitive to climate change. In countries such as Vietnam, crop production plays a vital role in both the country's economy and the well-being of its people. For instance, rice growing activity occupies 63% of total agricultural land in Vietnam and is also essential to the livelihoods of 63% of Vietnamese farming households. Moreover, rice production reached 43.4 millions of tons in 2019, making Vietnam the fifth rice producing country in the world and the second largest rice exporter. Rice production in Vietnam is threatened by climate change. Sea level rise could adversely affect Vietnam's prime rice-growing region, namely the Mekong River Delta which accounts for 54.47% of rice-planted area in this country. Sea level rise could reach 84cm under a high greenhouse gases global emissions scenario, causing large parts of the Delta plain whose estimated average elevation is around 80cm to fall below sea-level by the end of the century (see Chapters 1 and 3 in Espagne et al., 2021). On the other hand, temperature projections (from $\sim 1.3°C$ under a scenario of low greenhouse gases global emissions to $\sim 4.2°C$ under a high emissions scenario, with faster increases on the North of the country than in the South) make it possible to anticipate what might be considered as chronic heat stress in some areas that could also affect rice production, even under lower emissions pathways.

In the econometrics literature, the assessment of the impact of climate change for a given economic sector relies on the specification and estimation of a damage function that relates an outcome specific to that sector to climate indicators. Hsiang et al. (2017) present empirical, micro-founded sector-specific damage functions for a number of sectors: agriculture, crime, health and labor. Several of these damage functions consider crop yield as the outcome of interest and relate that yield to temperature and precipitation. Those proposed by Schlenker and Roberts (2009) are particularly interesting. A recent and complete survey can be found in Ortiz-Bobea (2021). They are based on the assumption that temperature effects on yields are cumulative over time and that yield is proportional to total exposure. This implies temperature effects are additively substitutable over time. This assumption can be mathematically translated by specifying the link between crop yield and temperature as a linear functional regression of a scalar response, crop yield, on the probability density function, thus involving an integral of the temperature density against the regression parameter which is itself a function of temperature. The regression parameter captures the sensibility of crop yield at different levels of temperature. The estimation strategy adopted by Schlenker and Roberts (2009) consists of approximating that integral. The damage function is then expressed as the regression of crop yield on the numbers of days belonging to the different bins defining the histogram of the temperatures observed over the crop growing season. As in the treatment of dummy variables, one bin is removed from the list of regressors to take into account that the sum of the regressors is fixed and equal to the total number of days in the crop growing season. The impact of one more day in a given temperature bin is therefore measured with reference to the deleted

bin. Several works have applied this estimation strategy, mainly after its use by Deryugina and Hsiang (2017) in their paper about the identification of the marginal effect of climate and their application to income in the US (see, for instance, Aragón et al., 2021, in their study of subsistence Peruvian farmers' response to extreme heat).

The estimation strategy proposed by Schlenker and Roberts (2009) can be discussed in light of recent contributions to the statistical literature. We can see that the original formulation of the model of Schlenker and Roberts (2009) uses a function representation for the temperature density and is therefore directly comparable to the functional scalar-on-density approach: in both cases the density function appears on the right hand side of the regression equation in a linear fashion through an integral term. In the later treatment of their model however, Schlenker and Roberts (2009) approximate their integral by a finite sum arriving at a regression model on bin frequencies (after removal of a reference bin). The model they implement is therefore comparable to a scalar-on-composition model but an important divergence then appears. The Schlenker and Roberts (2009) model uses the bin frequencies (except the reference bin) as explanatory variables in a linear model. By removing one bin they take into account the relative nature of bin frequencies as is done in compositional data analysis. However in compositional data analysis, people usually use transformations of the vector of frequencies in order for the postulated linear model to be linear with respect to the vector space structure of the simplex which is based on operations different from the classical sum and multiplication by a scalar of $\mathbb{R}^D$.

The paper is organized as follows. Section 2 reviews the methodological tools involved in these discrete and smooth compositional models (simplex space and Bayes space structures, centered log-ratio transformations) as well as the construction of the compositional splines. Section 3 presents the rice yield data and the weather data and their main features through exploratory analysis. Section 4 presents the discrete and smooth compositional scalar-on-density regression models and their estimation results. Section 5 presents our proposal to build a climate change scenario, derives the corresponding formulas for computing its impact. An illustration of these impacts on the dataset allows to reveal the interest of the smooth approach. Section 6 then concludes.

2. Methodological reminders

The density data at hand in our problem is a set of distributions of maximum daily temperatures during 30 years, from 1987 to 2016, in 63 Vietnamese provinces. These densities will play the role of covariates in a regression model explaining rice yield in Vietnam during that period. In the discrete approach, we use compositional vectors to represent the covariates and we remind the basic tools for working with compositional vectors in Section 2.1. In the smooth approach, we use smooth densities and we remind in Section 2.2 the construction of the Bayes space $\mathcal{B}^2$ of densities. For the regression part, in the case of the discrete approach we can use scalar on composition regression as Hron et al. (2012). For the regression part of the functional approach, since the density

covariate data is originally available as an histogram, a first step is necessary to smooth these histograms into $\mathcal{B}^2$ elements using CB-splines smoothing. We briefly review CB-splines in Section 2.3 and CB-splines smoothing in Section 2.4.

2.1. Discrete densities representation as compositional vectors

Let us first remind that compositional data (CoDa thereafter) vectors can be defined as vectors of D positive components adding up to one, elements of a simplex denoted $\mathcal{S}^D$. A discrete density function associated to a random variable with a finite number of outcomes is represented by its probability mass function, or equivalently by the vector of probabilities of each of these outcomes which satisfies the same constraints as a CoDa vector. This space can be equipped with a vector space structure using the following operations, see e.g. Aitchison (1986).

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

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

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

$$\text{For} \quad \lambda \in \mathbb{R}, \mathbf{u} \in \mathcal{S}^D \quad \lambda \odot \mathbf{u} = \mathcal{C}(u_1^\lambda, \ldots, u_D^\lambda),$$

where $\mathcal{C}$ denotes the closure of a vector (division by the sum of its components). The above operations allow to define a proper average of a sample of n compositional vectors $\mathbf{u}_i$ (for $i = 1$ to n) by $\bar{\mathbf{u}} = \frac{1}{n} \odot (\mathbf{u}_1 \oplus \ldots \oplus \mathbf{u}_n)$ (thus the average components are just obtained by a geometric average of the sample's components).
The clr transformation of a vector $\mathbf{u} \in \mathcal{S}^D$ is defined by

$$\text{clr}(\mathbf{u}) = \mathbf{G}_D \ln \mathbf{u},$$

where $\mathbf{G}_D = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_D {\mathbf{1}_D}^T$, $\mathbf{I}_D$ is a $D \times D$ identity matrix, $\mathbf{1}_D$ is the D-vector of ones and where the logarithm of $\mathbf{u} \in \mathbf{S}^D$ is understood componentwise. For a vector $\mathbf{u}$ in the orthogonal space $\mathbf{1}_D^\perp$ (orthogonality with respect to the standard inner product of $\mathbb{R}^D$), the inverse clr transformation is defined by

$$\text{clr}^{-1}(\mathbf{u}) = \mathcal{C}(\exp(\mathbf{u})).$$

The simplex $\mathcal{S}^D$ of dimension $D-1$ can be equipped with the Aitchison inner product

$$< \mathbf{u}, \mathbf{v} >_A = < \text{clr}(\mathbf{u}), \text{clr}(\mathbf{v}) >,$$

where the right hand side inner product is the standard inner product in $\mathbb{R}^D$.

2.2. Continuous densities representation as elements of the Bayes space

As in Van den Boogaart et al. (2014), density functions can be viewed as elements of the so-called Bayes space denoted by $\mathcal{B}^2([a,b])$ composed of positive functions integrating to one on a bounded interval $[a,b]$ and whose log-transform is square integrable. This space can first be equipped with a vector space structure using the following operations. For any positive function h on $[a,b]$, let us define the closure $\mathcal{C}(h)$ of h to be the unique density proportional to h. Then for any two functions f and g in $\mathcal{B}^2([a,b])$ and any real α, one can define

- perturbation as $(f \oplus g)(t) = \mathcal{C}(f(t)g(t))$
- powering as $(\alpha \odot f)(t) = \mathcal{C}(f(t)^\alpha)$

The centered log-ratio (clr) transformation is defined for $f \in \mathcal{B}^2([a,b])$ and t in $[a,b]$ by

$$\mathrm{clr}f(t) = \log f(t) - \frac{1}{b-a}\int_a^b \log f(u)du \tag{1}$$

By construction the clr transformation maps $\mathcal{B}^2([a,b])$ into the space $L_0^2([a,b])$ of square integrable functions on $[a,b]$ with a zero integral. Its inverse exists and is expressed as follows for a function $g \in L_0^2([a,b])$,

$$\mathrm{clr}^{-1}(g)(t) = \mathcal{C}\exp\left(\mathrm{clr}g(t)\right).$$

$\mathcal{B}^2([a,b])$ can then be equipped with the following inner product which makes the clr transformation isometric when the classical inner product is used in $L_0^2([a,b])$.

$$< f, g >_{\mathcal{B}^2} = \int_a^b \mathrm{clr}f(t)\ \mathrm{clr}g(t)dt =< \mathrm{clr}f, \mathrm{clr}g >_{L_0^2([a,b])}. \tag{2}$$

2.3. Reminder on CB-splines and ZB-splines

Spline functions are made of pieces of polynomials of a given degree connecting at knots points with given smoothness conditions (see e.g. De Boor, 1978). For our purpose, because the objective is to approximate density functions, we need particular constrained splines. One way to construct them is described in Machalová et al. (2021) using the so-called ZB-splines in $L_0^2([a,b])$ and corresponding CB-splines in $\mathcal{B}^2([a,b])$. A basis of spline functions satisfying the integral constraint is first constructed in $L_0^2([a,b])$ and pulled back to $\mathcal{B}^2([a,b])$ by the inverse clr transformation. The final system of B-splines are entirely defined by a sequence of knots (points at which polynomial pieces connect) and an order (polynomial degree plus one). Let $\Lambda = \{(\lambda_1, \ldots, \lambda_g) : a < \lambda_1 < \ldots \lambda_g < b\}$ be the set of so called inside knots. For technical reasons, additional knots are necessary at the boundary: if k is the degree of the polynomial pieces ($d = k+1$ the corresponding order), k knots equal to a are added at the beginning of the interval and k knots equal to b at the end. Then the dimension of the ZB-splines basis (basis of $L_0^2([a,b])$) is equal to $g+k$ while the dimension of the B-spline

basis corresponding to the same set of knots and order is equal to $g + d$, one dimension is lost for the ZB-basis due to the integral constraint. The inverse clr of the ZB-basis functions are called the CB-basis functions. For this application, we restrict attention to cubic splines for which $k = 3$ and $d = 4$. Let S_k^Λ be the subspace of $L^2([a, b])$ generated by the B-splines basis and Z_k^Λ be the space generated by the corresponding ZB-splines basis. Machalová et al. (2021) show the correspondence between the representation of any function in Z_k^Λ in both basis systems. This correspondence is handy because it allows to manipulate ZB-splines using classical code designed for B-splines.

In our subsequent application, the data is given as a set of histograms of daily maximum temperatures, in a given province and a given year, with 28 bins of length 1 on the interval $[a, b] = [12, 40]$. We use cubic splines ($k = 3$) with $g = 7$ (respectively $g = 9$) inside knots to approximate the densities underlying these original histograms. The dimension of the ZB-spline basis is thus $7 + 3 = 10$ for 7 inside knots (respectively $9 + 3 = 12$ for 9 inside knots). The position of the knots is chosen with respect to the data points position as argued in Machalová et al. (2021) using quantiles. Figure 1 represents the two sets of basis functions thus obtained in $L_0^2([a, b])$ and in $\mathcal{B}^2([a, b])$. The knots position is indicated by vertical dotted lines on the plots. We see that the two additional knots in the bottom plots bring more basis functions with support around the mode of the distribution thus allowing a finer approximation of the densities in the region where we have more temperature data points.

2.4. Smoothing histograms with CB-splines

Our original temperature data are sequences of daily maximum temperatures. In order to use the same technique as Machalová et al. (2021), we first preprocess the data into intermediate histogram representations and then into a smooth density using CB-splines as in Machalova et al. (2016). The CB-spline smoothing step consists in choosing a ZB-spline basis in L_0^2 and viewing the estimation of the clr transformed densities expressed in the ZB-basis as a penalized least squares regression. In this regression, the clr transformed histogram frequencies are explained by the covariates obtained by evaluating the ZB-spline basis functions at the midpoints of the histograms bins. Smoothing with ZB splines does not allow bins with zero counts because of the log transformation. For this reason, we apply a simple zero-replacement procedure by first replacing any zero count by 10^{-7} and then applying the closure operator. The smoothing parameter is chosen by generalized cross-validation using a regular grid of 100 points on a log-scale.

As an illustration, Figure 2 shows an histogram of the daily maximum temperatures in 1995 in the Yen Bai province (North-East of Vietnam), as well as the corresponding smooth density obtained by the above procedure on the left plot, and the smoothed clr transform on the right plot.

Figure 1: CB-splines (left) and ZB-splines (right) with 7 inside knots (top) and 9 inside knots (bottom)

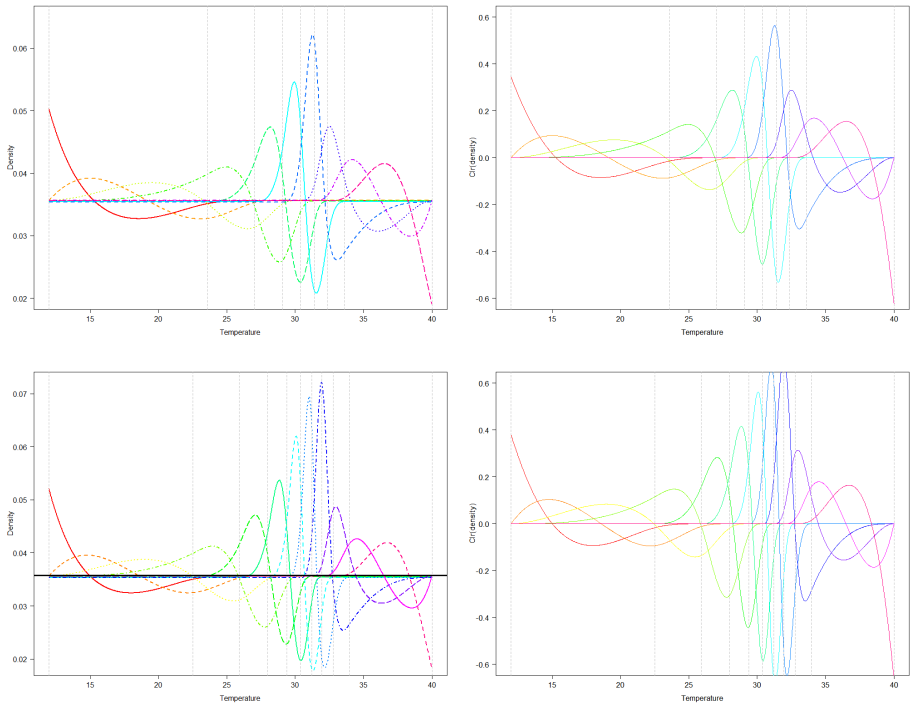
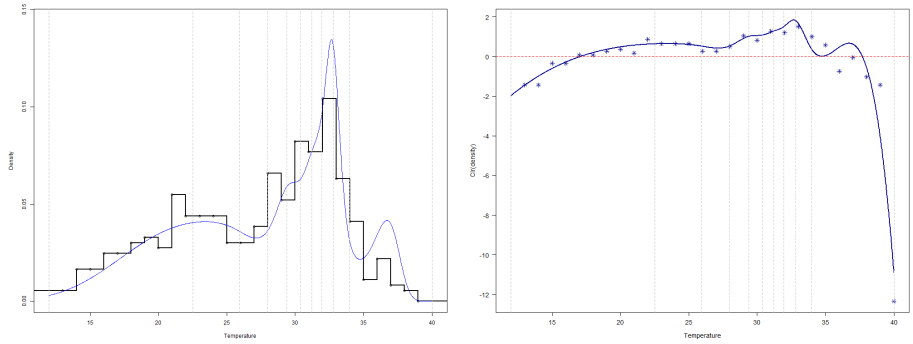


Figure 2: Density of daily maximum temperature in 1995 in Yen Bai province (left) and its clr transform (right)



3. Data and exploratory analysis

3.1. Rice yield data

The dataset about rice yield comes from the International Rice Research Institute[1]. The data set contains information on annual rice production, area

[1]IRRI is an organisation that promotes research and development of rice production in the world. Information about the institute can be found at https://www.irri.org/

harvested, rice yield at provincial level from 1987 to 2016. Rice yield is measured in tons per hectares. Figure 3 shows the overall evolution of rice yield over the considered period. After stagnation between 1987 and 1992, rice yield has grown steadily since 1992, and this growth has affected all Vietnamese provinces. This growth may be explained by the progress of agronomic techniques over the years. Since we have no proxy to measure it, we will take it into account with a linear time trend, as supported by Figure 4, which reports the evolution of average rice yields for the six different agronomic regions in Vietnam and Figure 5 for the three provinces of Yen Bai, Ninh Binh and Dong Thap.

Figure 3: Rice yield distributions from 1987 to 2016

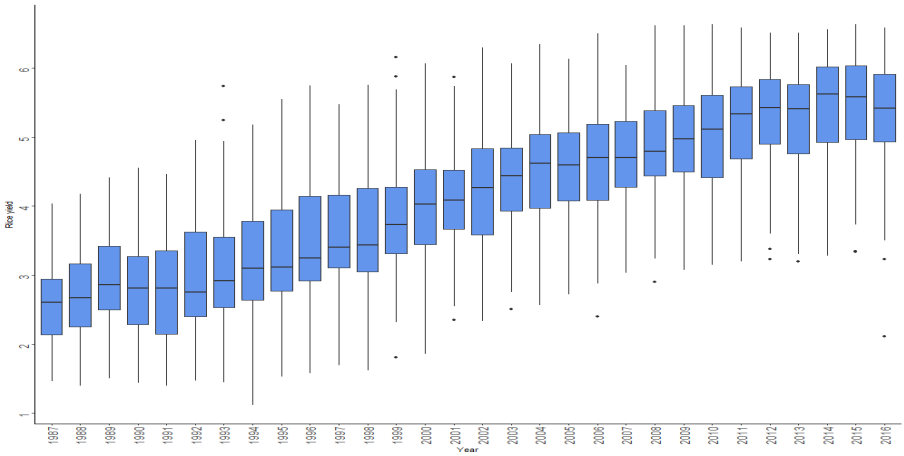


Figure 4: Average rice yield by agronomic regions from 1987 to 2016

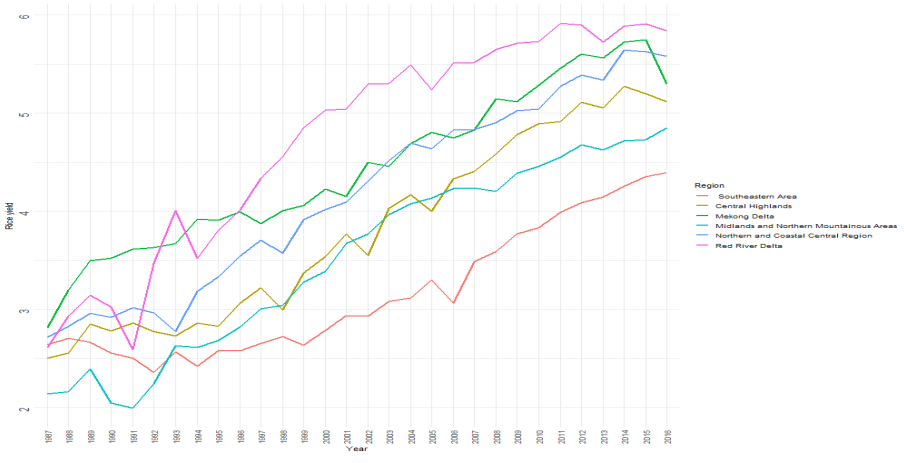
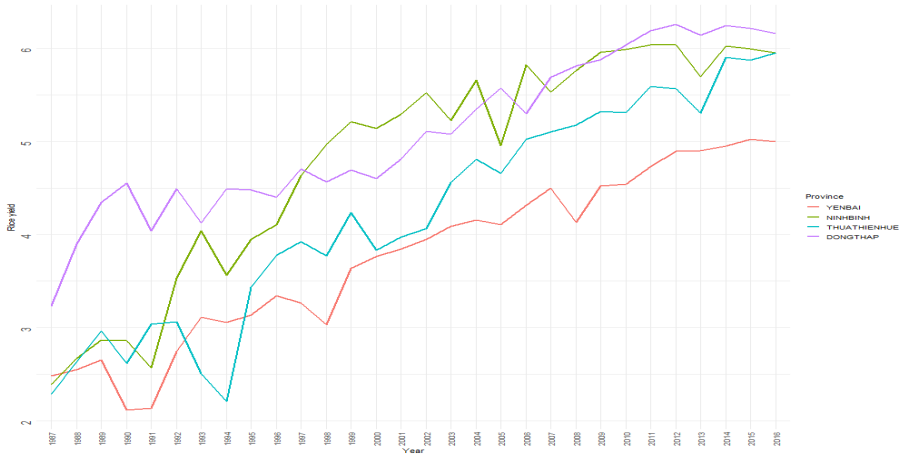


Figure 5: Average rice yield in Yen Bai, Ninh Binh and Dong Thap province from 1987 to 2016



3.2. Weather data

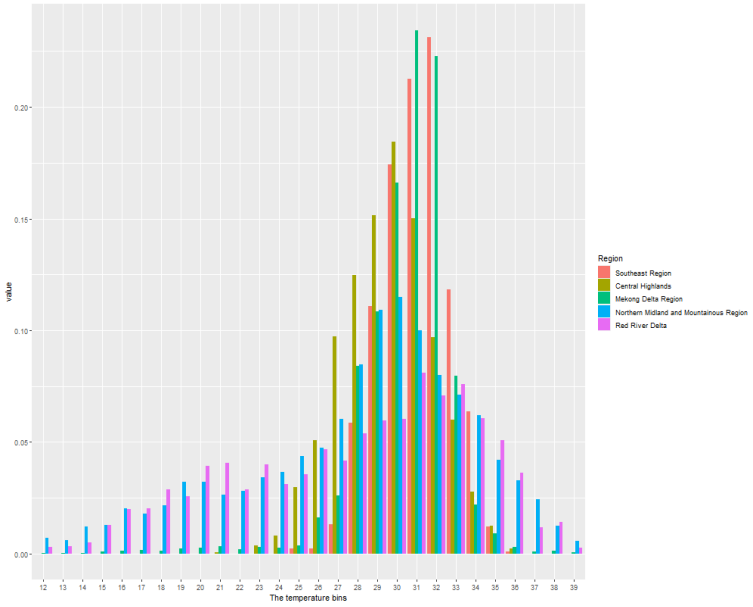
Weather data used in this study are daily maximum-temperatures and precipitation. Temperature data comes from the Climate Prediction Center (CPC) database developed by the National Oceanic and Atmospheric Administration (NOAA). We extract historical data on daily maximum temperature for a grid of 0.50×0.50 degree of latitude and longitude for Vietnam. The data is converted into the daily maximum temperature for each of 63 Vietnamese provinces and during 30 years (1987-2016) (365 or 366 values for each year) yielding one temperature distribution for each of 1890 statistical units.

Figure 6 displays the average histograms of each of the 6 regions where average is understood with the simplex operations as defined in Section 2.1. These histograms show that the range of maximum temperature is quite different from one region to the other.

Using the CB-spline smoothing tool we can also explore other aspects of the variations of the temperature densities across time and space. Figure 7 displays the daily maximum temperature density (and clr transform) in the province of Ninh Binh (from Red River Delta) which is one of the major provinces for rice production. We use the viridis color palette with 30 values, changing from yellow in 1987 to dark violet in 2016 with green values in between. The top part of Figure 7 clearly reveals the right shift of the temperature densities corresponding to climate change. Finally Figure 8 displays the densities and their clr transforms in 2015 for all provinces. On the clr transforms we can see groups of provinces and it would be interesting to explore their respective spatial position. It seems that they differ mainly in the range of the observed maximum temperatures.

For using the smoothed histograms in the regression model later, we need to express them in the same basis of CB-splines, therefore we need to use the

Figure 6: Maximum temperature histograms across the Vietnamese regions in 2015

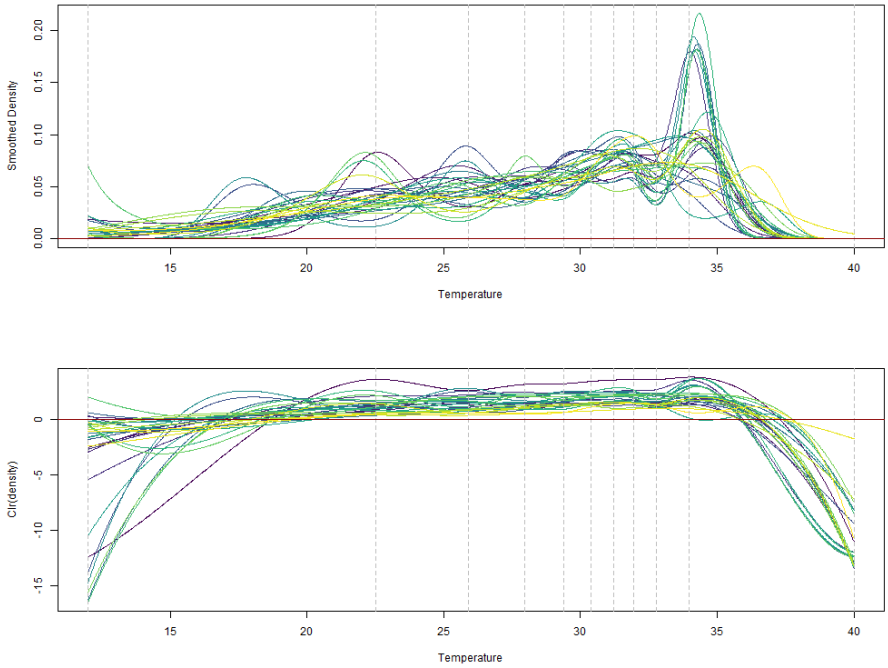


same knots position for all the $63 * 30 = 1890$ histograms. For this reason, we first pool all observations into a single distribution and place the knots at the quantiles of this global distribution.

4. The discrete and smooth regression models

The goal in this application is to construct a regression model to explain rice yield in a given province and a given year by the distribution of daily maximum temperature over that year in that province and possibly additional covariates. The purpose of this model is to understand the impact of the temperature distribution on rice yield controlling for other effects. Considering the effect of time, given our purpose, we are not interested in a time series model to predict yield in the future but rather we want to take advantage of the spatio-temporal variability in order to measure the impact of temperature on rice yield. Therefore we decide to include a simple linear time trend in the model as a proxy for unobserved factors which may have evolved with time, like the evolution of the production techniques. In view of Figures 3, 5 and 4, the linear trend seems a reasonable choice. We further use other controlling factors: precipitation and regional dummies. Because our covariate of interest has a distributional nature, we need an adapted regression model and the choice is between using an histogram of the daily temperatures as a compositional covariate, as in Hron et al. (2012), or using a smoothed version of the temperature density as a smooth density covariate, as in Talská et al. (2021). We first review the principles of

Figure 7: Density (top panel) and clr transform (bottom panel) of the smoothed daily maximum temperature from 1987-2016 in Ninh Binh province



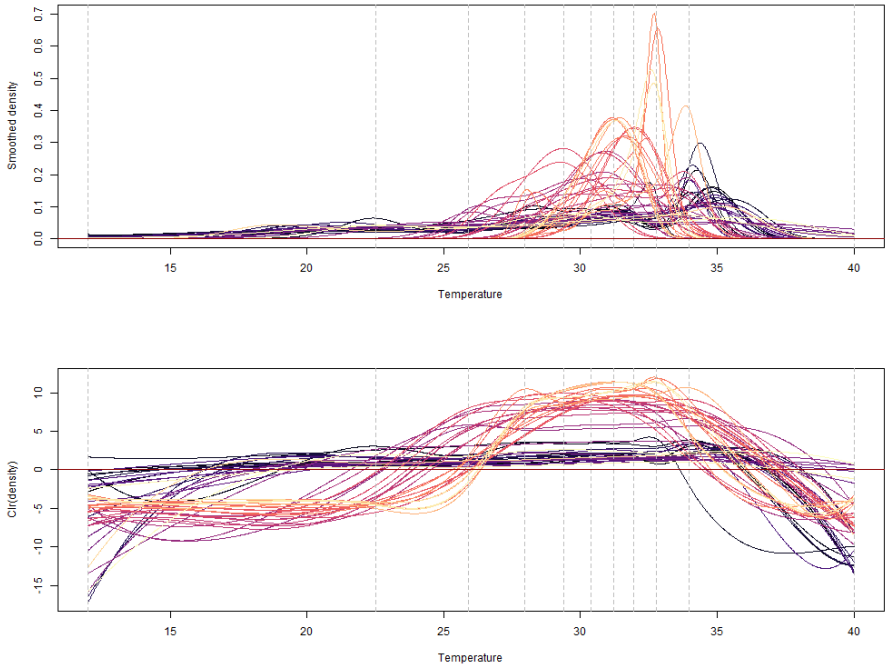
these two models before presenting their results.

4.1. The discrete regression model

The scalar-on-composition regression model as presented for example in Hron et al. (2012) is a regression model in which at least one of the covariates is a compositional vector. In our discrete regression framework we use temperature histograms as compositional vectors and these can also be viewed as discrete densities. The expression of any linear function of a compositional explanatory variable $\mathbf{X} \in \mathcal{S}^D$ must be of the form $< \beta, \mathbf{X} >_A$, where β is a parameter vector of $\mathcal{S}^D$ and we recall that $< ., . >_A$ is the classical Aitchison inner product in $\mathcal{S}^D$ (see e.g. Pawlowsky-Glahn et al. (2015). Therefore a linear model to explain a scalar variable Y with possibly several compositional variables $\mathbf{X}_j \in \mathcal{S}^{L_j}$ for $j = 1, \dots J$ and several scalar variables $\mathbf{Z}_l$ for $l = 1, \dots L$ is formulated by an equation of the form

$$Y_i = \alpha + \sum_{j=1}^{J} < \boldsymbol{\beta}_j, \mathbf{X}_{ij} >_A + \sum_{l=1}^{L} \gamma_l \mathbf{Z}_{il} + \epsilon_i, \tag{3}$$

Figure 8: Density (top) and Clr transform (bottom) of the smoothed daily maximum temperature from in 2015 for all provinces



where the parameters $\beta_j \in \mathcal{S}^{L_j}$ and the errors ϵ_i are i.i.d. gaussian variables with mean zero and variance σ^2. For our application, we need to index all observations by province i and year k therefore the index i of equation (3) is replaced by the two indices i and k so that Y_{ik} is rice yield for province i ($i = 1$ to 63) in year k ($k = 1$ to 30). In this application, we have a single discrete density as compositional covariate ($L = 1$) which is maximum temperature histogram with corresponding parameter β_{max} and several classical scalar variables ($L = 8$) with time, precipitation and six regional dummies. For the discrete covariate, the histogram of maximum daily temperature has been reported with equal bins of length 1 degree Celsius.

The estimation of such a model is done by ordinary least squares using a transformation of the compositional covariates (any choice between ilr or alr transformation, see e.g. Coenders and Pawlowsky-Glahn (2020)).

4.2. The smooth regression model

Extending the model in Talská et al. (2021) to the case of several density covariates as well as additional scalar covariates, we consider the following linear

scalar on density regression model

$$Y_i = \quad \beta_0 + \sum_{j=1}^{J} < \beta_j(t), f_{ij}(t) >_{\mathcal{B}^2} + \sum_{l=1}^{L} \gamma_l \mathbf{Z}_{il} + \epsilon_i, \tag{4}$$

where Y_i is the scalar dependent variable, β_0 is a real intercept, $\beta_j(t), j = 1, \ldots J$ are curve-parameters for the effects of the densities f_{ij}, Z_l $(l = 1, \ldots, L)$ are real covariates with their corresponding parameters γ_l, and finally ϵ_i are normal errors with mean zero and standard deviation σ^2. The densities f_{ij} as well as the curve-parameters β_j are assumed to belong to the Bayes space $\mathcal{B}^2([a, b])$.

Using the fact that the clr transform is an isometry between $\mathcal{B}^2$ and $L_0^2([a, b])$ equipped with their respective inner products, we can rewrite the model as follows

$$\begin{aligned} Y_i = \quad & \beta_0 + \sum_{j=1}^{J} < \mathrm{clr}\beta_j(t), \mathrm{clr}f_{ij}(t) >_{L_0^2} \\ & + \sum_{l=1}^{L} \gamma_l \mathbf{Z}_{il} + \epsilon_i, \end{aligned} \tag{5}$$

In order to estimate this model, we first need to use a basis expansion of the functional parameters $\beta_j(t)$, as well as a similar expansion for the densities $f_{ij}(t)$. For the sake of simplicity, we will use the same basis system to express the functional regression parameters and the observed functional explanatory variables but it is possible to take a different basis. The expansion can be written directly in $\mathcal{B}^2([a, b])$ or equivalently for the clr transforms in $L_0^2([a, b])$. Once we replace these functions by their expansion in the inner products of the model equation (5) for example, the inner products appear as a linear function of the expansions coefficients multiplied by the Gram matrix (inner products of all pairs of basis functions) as in Talská et al. (2021). After this step, we are back to a classical linear model for ordinary covariates that we can fit with ordinary least squares.

As before in our application, all observations are indexed by province i and year k therefore the index i of equation (4) is replaced by the two indices i and k so that Y_{ik} is rice yield for province i ($i = 1$ to 63) in year k ($k = 1$ to 30). β_0 is a real intercept. We include a single smooth density covariate f_{ik}^{max} which is the density of daily maximum temperature, in province i and year k, and $\beta_{max}(t)$ is the corresponding curve-parameter. We have additionally the same $L = 8$ classical covariates (time, precipitation and regional dummies) with their corresponding parameters γ_l, and finally ϵ_{ik} are normal errors with mean zero and standard deviation σ^2. f_{ik}^{max} as well as the curve-parameter β_{max} are assumed to belong to the Bayes space $\mathcal{B}^2([a, b])$. The number of basis functions for the expansion is a function of the number of knots. In order to reduce variability, it is advisable to use a small number of knots compared to the sample size. As in Section 2.3, we will try two different knots number equal to $g = 7$ and $g = 9$ corresponding to dimensions for the corresponding ZB-basis of $7 + 3 = 10$ in the first case and $9 + 3 = 12$ in the second case.

4.3. Model results

For the smooth regression, the 7 knots model does not yield significant results for the temperature distribution while the 9 knots model all coefficients

in coordinate space are significant at the 5 percent level. The coefficients and corresponding significance of the other explanatory variables are presented in Table 1 for the discrete and the smooth with 9 knots models. Concerning the

Table 1: Estimated coefficients associated to regional dummies, total precipitation and year

Variable	Regression type	
	Discrete regression	Smooth regression (9 knots)
Region		
Northern Midland and Mountainous Region	-1.14^{***}	-1.135^{***}
North Central Coast	-0.42^{***}	-0.39^{***}
Central Highlands	-0.73^{***}	-0.70^{***}
Southeast Region	-1.59^{***}	-1.60^{***}
Mekong Delta Region	-0.23^{***}	-0.23^{***}
(Reference = Red River Delta)		
Total precipitation (Thousand ml per year)	-0.17^{***}	-0.163^{***}
Year	0.11^{***}	0.106^{***}

Note: *, **, and *** mean significant at 10%, 5%, and 1%, respectively

regional differences, Table 1 shows that these five regions have negative parameters meaning that their average rice yield is below that of the reference region Red River Delta, which is a natural result since Red River Delta is quantitatively the first most important rice-producing area in Vietnam in terms of average rice yield, see Figure 4.

Figure 9 compares the fitted and observed values of both regressions and reveal that the smooth model has a smaller residual sum of squares while having a smaller number of parameters.

Figure 10 shows on the left the histogram of the discrete β vector and on the right of the smooth β curve. Note that the bins of the histograms have been labelled using the lower bound of the bin interval. The comparison between both graphs shows the similarity between the general shape of the two graphs except that the size of the effect is different.

5. Climate change scenario and its marginal effect

Covariates impact in scalar-on-composition regression can be evaluated either using finite increments as in Coenders and Pawlowsky-Glahn (2020) or infinitesimal increments as in Morais et al. (2018). To simplify comparisons between the discrete and the smooth regression models, we choose a finite increment perspective here. In order to assess the impact of a compositional covariate in a model such as (3) or (4), that is in our case the impact of climate change, we imagine scenarios for a change in this covariate. In the discrete case, to be coherent with the simplex space to which the densities belong, the change scenario should be linear with respect to the vector space structure of

Figure 9: Fitted value versus dependent values of discrete regression (left) and smooth regression (right) using 9 inside knots

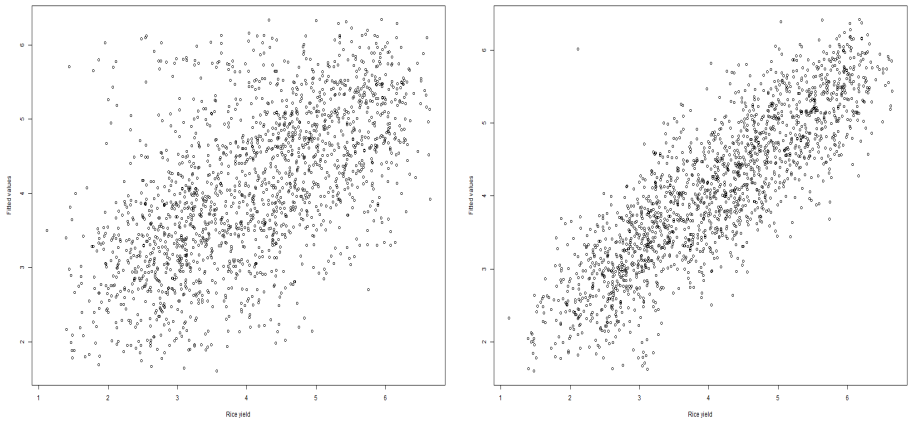
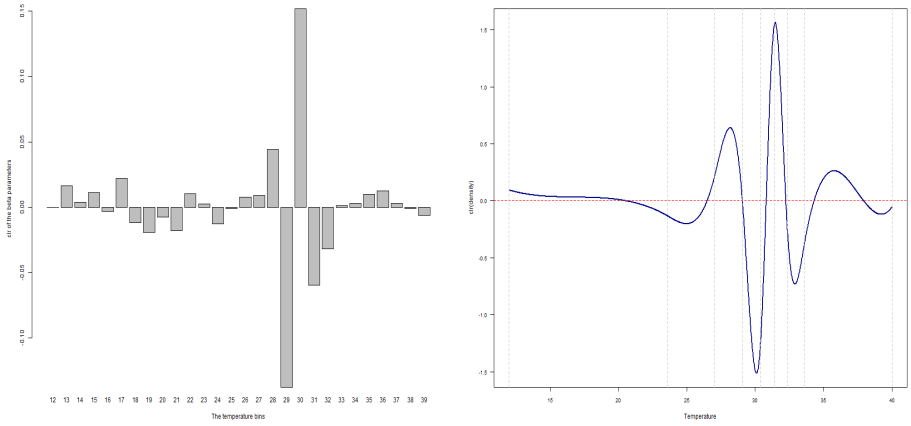


Figure 10: Beta parameters in the discrete (left) and smooth (right) regression models



the simplex $\mathcal{S}^{28}$. In the smooth case, the change scenario should be linear with respect to the vector space structure of the Bayes space $\mathcal{B}^2([a,b])$.

For this reason, we use finite linear increments to create the change scenarios as follows. In the discrete case, the change is given by

$$T_h f_{ik} = f_{ik} \oplus (h \odot \varphi), \tag{6}$$

where φ is a direction of change in $\mathcal{S}^{28}$. In the smooth case, the change is given by

$$T_h f_{ik}(t) = f_{ik}(t) \oplus (h \odot \varphi(t)), \tag{7}$$

where $\varphi(t)$ is a direction of change in $\mathcal{B}^2([a,b])$. To simplify comparison, we decide to choose a change direction curve in $\mathcal{B}^2([a,b])$ which coincides with a histogram with bin length of one so that the change curve $\varphi(t)$ is described by its vector of histogram frequencies φ which is the same as the φ used for the discrete model.

We propose the following choice for the change direction: φ is the bin frequency vector $\mathcal{C}(1,\ldots,1,e,\ldots,e)$ with the first component equal to e being component number m. Alternative choices are of course possible. The real h reflects by its sign the orientation on the direction $\varphi(t)$ and by its absolute value the intensity of the change. Note that in clr space, the change writes as follows

$$\mathrm{clr} T_h f_{ik}(t) = \mathrm{clr} f_{ik}(t) + h\mathrm{clr}\varphi(t). \tag{8}$$

For a choice of $h = 1$, the l^{th} component of $T_h f(t)$ is then given by

$$T_h f_{ik}(t)_l = \frac{f_{ik}(t)_l \varphi(t)_l}{S_m + e(1 - S_m)}, \tag{9}$$

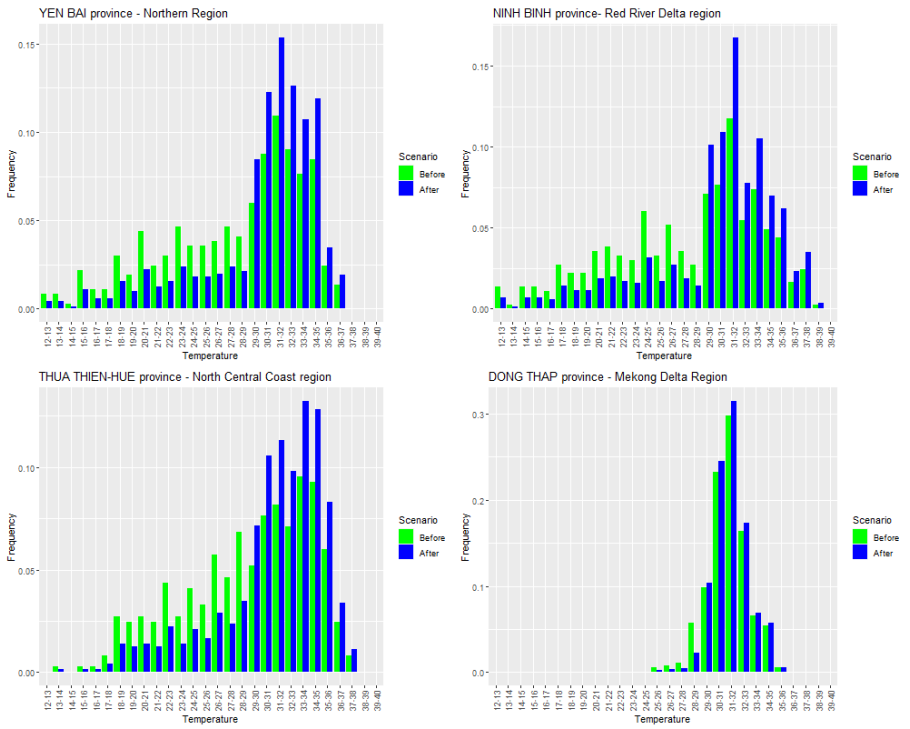
where $S_m = \sum_{j=1}^{m-1} f_{ikj}(t)$. Then it is easy to see that if $l \geq m$ and $l' < m$

$$\frac{\frac{T_h f_{ik}(t)_l}{f_{ik}(t)_l}}{\frac{T_h f_{ik}(t)_{l'}}{f_{ik}(t)_{l'}}} = \exp(1) = \frac{\frac{T_h f_{ik}(t)_l}{T_h f_{ik}(t)_{l'}}}{\frac{f_{ik}(t)_l}{f_{ik}(t)_{l'}}} \tag{10}$$

Formula (10) can be interpreted as follows: the ratio between two frequencies, one above the threshold m relative to the other below the threshold, is multiplied by a factor of $\exp(1) \sim 2.7$ after the change. The factor $\exp(1)$ can be adapted using an intensity h different from 1. Figure 11 illustrates this scenario for four selected provinces, for the threshold $m = 18$ corresponding to temperatures between 29 and 30 Celsius degrees and for $h = 1$: we can see that the frequencies in warm temperature bins are higher after the change but that the change is relative and not obtained by adding a constant to all bins after the threshold.

Were this hypothetical climate change to happen in a given province and year, we can now predict for each h the resulting rice yield change $\hat{Y}_{ik}(h) - Y_{ik}$, where $\hat{Y}_{ik}(h)$ denotes the projected rice yield under the change scenario. Given that both our models are linear for the simplex structure and that $T_h f_{ik}(t) -$

Figure 11: Climate change scenarios in four provinces



$f_{ik}(t) = h \odot \varphi(t)$, the resulting change of rice yield for a given province i and a given year k are given by

$$\hat{Y}_{ik}(h) - \hat{Y}_{ik} = h < \hat{\beta}_{max}, \varphi >_A = h < \mathrm{clr}\hat{\beta}_{max}, \mathrm{clr}\varphi >_{\mathbb{R}^{28}}, \tag{11}$$

in the discrete regression model and by

$$\hat{Y}_{ik}(h) - \hat{Y}_{ik} = h < \hat{\beta}_{max}(t), \varphi(t) >_{\mathcal{B}^2} = h < \mathrm{clr}\hat{\beta}_{max}(t), \mathrm{clr}\varphi(t) >_{L_0^2}, \tag{12}$$

in the smooth regression model. The inner products $< \mathrm{clr}\hat{\beta}_{max}, \mathrm{clr} f\varphi >_{\mathbb{R}^{28}}$ and $< \mathrm{clr}\hat{\beta}_{max}(t), \mathrm{clr} f\varphi(t) >_{L_0^2}$ therefore characterize the impacts of a change in temperature density in the respective models and these are constant for all provinces in both models. In our application, we find a decrease in rice yield of 0.055 tons per hectare in the discrete model and 2.417 tons in the smooth model. The discrete model turns out to underestimate the impact of the change. Another exploration with a smaller bin width, corresponding to an intermediate situation between the bin 1 histograms and the smooth curves, revealed an intermediate result in terms of rice yield showing that the approximation error should be the source of the difference in the model's appreciation of the impact of a change.

Alternative climate change scenarios could also be considered. For example, one could change the threshold for the direction change vector we considered or imagine a completely different vector which would weight the change differently for the modified bins i.e. replace h by a vector $(h_m, \ldots, h_{28})$. Another possibility would be to choose a change vector adapted to each region: we have seen in Section 3 that the 6 regions presented histogram shapes with different ranges of temperatures (i.e. support of the density). All of these scenarios are easy to implement with the above procedure.

6. Conclusion

We have compared two compositional approaches for the regression of a scalar on density objects by applying them to a concrete case study of the impact of climate change on rice yield production in Vietnam. The results show that the smooth or functional approach allows to keep more information from the density objects. The impact thus measured by the smooth model appears larger than that measured by the discrete one. Our model however is very simple and it would be interesting, for a more realistic application, to include the additional density of daily minimum temperatures that was available to us. Another improvement would be to take into account the cropping season in each region but from the practical side we did not have this data and this would open other issues from the methodological side since we have taken the same spline basis for all densities so far. It could be justified to include among the explanatory variables the number of days with maximum temperature above a chosen threshold which would complement the bins relative values reflected by the histograms. Alternative methods of estimation could be considered: instead of choosing a small number of knots, one can use penalized regression. However in that case it seems more difficult from the implementation point of view to include several explanatory densities.

In order to measure the impact of climate change, we have chosen to consider simple change scenarios. Further work could involve, while remaining in the realm of linear changes, to explore more complex scenarios by choosing other change directions, or consider changes non constant in space. Finally, we intend to explore further the comparison between the discrete and the smooth models by simulations which would allow to compare their respective result to a ground truth and could confirm our claim that the smooth model is yielding better results due to a smaller approximation error.

Acknowledgments The authors acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010 and the GEMMES project (Agence Française de Développement). The authors acknowledge support from the Vietnam Institute for Advanced Study in Mathematics (VIASM).

References

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

Aragón, F.M., Oteiza, F., Rud, J.P., 2021. Climate change and agriculture: Subsistence farmers' response to extreme heat. American Economic Journal: Economic Policy 13, 1–35.

Van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V., 2014. Bayes Hilbert spaces. Australian & New Zealand Journal of Statistics 56, 171–194.

Carter, C., Cui, X., Ghanem, D., Mérel, P., 2018. Identifying the economic impacts of climate change on agriculture. Annual Review of Resource Economics 10, 361–380.

Coenders, G., Pawlowsky-Glahn, V., 2020. On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT–Statistics and Operations Research Transactions 44, 201–220.

Davis, K.F., Downs, S., Gephart, J.A., 2021. Towards food supply chain resilience to environmental shocks. Nature Food 2, 54–65.

De Boor, C., 1978. A practical guide to splines. Springer-Verlag, New York.

Deryugina, T., Hsiang, S., 2017. The Marginal Product of Climate. Working Paper 24072. National Bureau of Economic Research.

Espagne, E., Ngo-Duc, T., Nguyen, M.H., Pannier, E., Woillez, M.N., Drogoul, A.and Huynh, T.P.L., Le, T.T., Nguyen, T.T.H., Nguyen, T.T., Nguyen, T.A., Thomas, F., Truong, C.Q., Vo, Q.T., Vu, C.T., 2021. Climate change in Vietnam: Impacts and adapatation. A COP26 assessment report of the GEMMES Vietnam project. Technical Report. Agence Française de Développement, Paris, France. https://www.ird.fr/gemmes-vietnam-report-climate-change-vietnam-impacts-and-adaptation.

Fisher, R., 1924. The influence of rainfall on the yield of wheat in Rothamsted. Philosophical Transactions of the Royal Society of London 213, 89–142.

Hron, K., Filzmoser, P., Thompson, K., 2012. Linear regression with compositional explanatory variables. Journal of Applied Statistics 39, 1115–1128.

Hsiang, S., Kopp, R., Jina, A., Rising, J., Delgado, M., Mohan, S., Rasmussen, D.J., Muir-Wood, R., Wilson, P., Oppenheimer, M., Larsen, K., Houser, T., 2017. Estimating economic damage from climate change in the United States. Science 356, 1362–1369.

Machalova, J., Hron, K., Monti, G.S., 2016. Preprocessing of centred logratio transformed density functions using smoothing splines. Journal of Applied Statistics 43, 1419–1435.

Machalová, J., Talská, R., Hron, K., Gába, A., 2021. Compositional splines for representation of density functions. Computational Statistics 36, 1031–1064.

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

Ortiz-Bobea, A., 2021. Chapter 76 - the empirical analysis of climate change impacts and adaptation in agriculture, in: Barrett, C.B., Just, D.R. (Eds.), Handbook of Agricultural Economics. Elsevier. volume 5 of *Handbook of Agricultural Economics*, pp. 3981–4073.

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

Petersen, A., Zhang, C., Kokoszka, P., 2022. Modeling probability density functions as data objects. Econometrics and Statistics 21, 159–178.

Schlenker, W., Roberts, M.J., 2009. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences 106, 15594–15598.

Talská, R., Hron, K., Grygar, T.M., 2021. Compositional scalar-on-function regression with application to sediment particle size distributions. Mathematical Geosciences 53, 1667–1695.