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

Huong Trinh Thi[1*†], Michel Simioni[2,3†] and Christine Thomas-Agnan[3†]

[1*]Faculty of Mathematical Economics, Thuongmai University, Ho Tung Mau street, Hanoi, 10000, Vietnam.
[2]MoISA, University of Montpellier, Place Pierre Viala, Montpellier, 34060, France.
[3]Toulouse School of Economics, University of Toulouse Capitole, Toulouse, 31000, France.

*Corresponding author(s). E-mail(s): trinhthihuong@tmu.edu.vn;
Contributing authors: michel.simioni@inrae.fr ; christine.thomas@tse-fr.eu;
†These authors contributed equally to this work.

Abstract

Within the econometrics literature, assessing the impact of climate change on agricultural yield has been approached with a linear functional regression model, wherein crop yield, a scalar response, is regressed against the temperature distribution, a functional parameter alongside with other covariates. However this treatment overlooks the specificity of the temperature density curve. In the realm of compositional data analysis, it is argued that such covariates should undergo appropriate log-ratio transformations before inclusion in the model. We compare a discrete version with temperature histograms treated as compositional vectors and a smooth scalar-on-density regression with temperature density treated as an object of the so-called Bayes space. In the latter approach, when density covariate data is initially available in histogram format, a preprocessing smoothing step is performed involving CB-splines smoothing. We investigate the respective advantage of the smooth and discrete approaches by modelling the impact of maximum and minimum daily temperatures on rice yield in Vietnam. Moreover we advocate for the modelling of climate change scenarios through the introduction of perturbations of the initial density, determined by a change direction curve which

induces a concentration of the densities towards higher temperature ranges. The resulting impact on rice yield is then quantified by calculating a simple inner product between the parameter of the density covariate and the change direction curve. Our findings reveal that the smooth approach and the discrete counterpart yield coherent results, but the smooth seems to outperform the discrete one by an enhanced ability to accurately gauge the phenomenon scale.

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

1 Introduction

We consider improving the linear functional regression models approach used in the econometrics literature to assess the impact of climate change on agricultural yield by properly taking into account the density nature of their functional parameter.

As the complexity of recorded data continues to grow, contemporary models increasingly involve intricate data objects including random densities. We are focusing here on regression models where such density objects serve as explanatory variables. These density objects can be treated either in a discrete fashion as histograms or in a continuous fashion as density functions, see for example [1]. True continuous observations are rarity. Density data, often available in the discrete form of histograms, are typically treated as continuous when the number of bins is exceedingly large. Consequently, a preprocessing step involving smoothing becomes necessary.

It is often the case that density data are recorded in an aggregated form as histograms, see for example [2] for a comprehensive review in the context of climate change econometrics. When adopting this discrete approach, the sample space can be described by the set of vectors of bin frequencies, with positive components that sum to one. These vectors are called compositions and their space is known as a simplex. A proper statistical treatment of this type of data can be done by compositional data analysis, see [3] or [4] for an introduction. Scalar-on-composition regression models using the simplex representation are described for example in [5]. They are obtained by transforming the simplex explanatory vectors, usually using a log-ratio transformation, to map them into an unconstrained linear space $\mathbb{R}^k$ (for some adapted value of k).

Conversely, [6] conducts a comprehensive review of various methodologies for constructing regression models involving samples of probability density functions with a functional perspective. In the realm of functional data analysis, densities stand out as unique entities due to the constraints they must satisfy. For one-dimensional densities, the sample space is defined as the space $\mathcal{D}$ of functions with positive values and a unit integral. [6] highlights one of the two primary approaches, which revolves around the representation of densities in the so-called Bayes spaces $\mathcal{B}^2$. Bayes spaces, initially introduced by [7], endow the space $\mathcal{D}$ of densities with a finite support $[a, b]$ with a Hilbert space structure. This space and structure can be viewed as a continuous version of the simplex and its associated operations. As for the log-ratio transformation,

the functional centered log-ratio serves as the functional counterpart of the classical centered log-ratio transformation for vectors of a simplex. This concept is used for example in [8] to construct functional scalar-on-density regression models. For the preprocessing step, [9] propose a new class of splines, known as compositional splines or CB-splines, specifically designed to accomodate the density constraints.

Nonetheless the functional (smooth) approach implementation is more complex prompting the natural question of assessing the potential advantage gained from using the functional model. Our objective in this work is to explore this comparison through 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 by no means a new endeavor, as evidenced by [10]. Climate change exerts both direct and indirect impacts on various facets of the food system encompassing food production, storage, processing, distribution, retail and consumption, as discussed by [11]. 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 cultivation occupies a substantial 63% of Vietnam's total agricultural land and is also essential to the livelihoods of 63% of Vietnamese farming households. Moreover, in 2019, rice production in Vietnam reached a staggering 43.4 million tons, solidifying the country's position as the world's fifth-largest rice producer and second-largest rice exporter. Unfortunately, this critical sector faces mounting threats from climate change. The rising sea levels pose a significant danger to Vietnam's primary rice-growing region, the Mekong River Delta, which accounts for 54.47% of the nation's rice-planted area. Under a high greenhouse gases global emissions scenario, sea levels could rise by up to 84 cm, potentially submerging large portions of the Delta plain whose estimated average elevation is expected to fall around 80cm below sea-level by the end of the century [see Chapters 1 and 3 in 12]. Furthermore, temperature projections (ranging from a modest increase of approximately 1.3°C under a low greenhouse gases global emissions scenario to substantial rise of around 4.2°C under a high emissions scenario, with faster increases on the North of the country than in the South) signal the possibility of chronic heat stress in some areas that could also adversely affect rice production, even under lower emissions pathways.

Within the field of econometrics, assessing the impact of climate change for a given economic sector relies on the specification and estimation of a damage function. For a specific outcome, the damage function relates a change in the climate indicators to the corresponding change in the outcome. [13] present empirical, micro-founded sector-specific damage functions tailored to various sectors, including agriculture, crime, health and labor. Several of these damage functions consider crop yield as the outcome of interest and link that yield to temperature and precipitation. Noteworthy among these contributions are the insights provided by [14], while a recent and comprehensive survey can be found in [15]. [14] build their assumptions on the premise that temperature effects on yields accumulate over time and that yield is proportional to total exposure. The consequence of this assumption is that we may use the temperature density as a functional covariate instead of using the curve of the temperature as

a function of time, in other words the order in time in which the temperatures occur has no impact on the yield. In mathematical terms, this assumption allows to specify the link between crop yield (a scalar response) and temperature as a linear functional of a probability density function. This functional incorporates an integral of the temperature density against a regression parameter, itself a function of temperature. This regression parameter encapsulates the sensibility of crop yield at different temperature levels. The estimation strategy adopted by [14] revolves around using a discrete approximation of that integral resulting from approximating the temperature density by an histogram of the number of days falling into different temperature bins over the crop growing season. Similar to the handling of dummy variables, one bin is omitted from the list of regressors to account for the fact that the sum of the regressors remains constant and equal to the total number of days in the crop growing season. The impact of an additional day within a specific temperature bin is therefore measured in reference to the omitted bin. This estimation strategy has been adopted by several researchers, gaining prominence after its use in [16]. For instance, [17] applied this approach in their study of how subsistence Peruvian farmers respond to extreme heat.

The estimation strategy proposed by [14] can be discussed in light of recent contributions to the statistical literature. The original model of [14] uses a function representation for the temperature density, making the model 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 Schlenker's treatment of their model, they approximate this integral by a finite sum resulting in a regression model on bin frequencies (excluding a reference bin). This implementation of their model is therefore comparable to a discrete scalar-on-composition model. However a significant divergence arises from this point onward. Schlenker's model uses bin frequencies (except the reference bin) as explanatory variables in a linear model. It has long been recognized in the statistical literature, see for example [18], that comparing densities is best achieved by using relative distributions, a concept with a strong scale invariance property which is a generalization of the usual scale invariance. The relative probability density function for a pair of distributions is the ratio of their two densities and it is invariant under any monotone transformation of the underlying random variable. Consequently when comparing temperature distributions, it is advisable to employ relative densities instead of absolute differences between them. Whereas using linear effects of the temperature bin frequencies as in [14] is coherent with absolute differences, in contrast, compositional data analysis use log-ratios of bin frequencies as explanatory variables, aligning with the notion of relative differences.

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 explores their main features. Section 4 presents the discrete and smooth compositional scalar-on-density regression models and their estimation results. It also provides an interpretation of the discrete and smooth parameters associated to the temperature

distribution parameters. Section 5 presents our proposal to build a climate change scenario, and 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 dataset central to our problem comprises distributions of maximum daily temperatures spanning a 30-year period, from 1987 to 2016, across 63 provinces in Vietnam. These temperature density distributions serve as key covariates within our regression model, aimed at elucidating the factors impacting rice yield in Vietnam over this timeframe. In the discrete approach, we represent these temperature covariates as compositional vectors and and we provide an overview of fundamental techniques 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. As we delve into the regression component, for the discrete approach, we employ scalar-on-composition regression techniques, as presented by [5]. In contrast, the functional approach necessitates an initial step to transform the density covariate data, originally available in histogram form, into elements of $\mathcal{B}^2$". In contrast, since the density covariate data is originally available as an histogram, the regression part of the functional approach necessitates a preliminary step to transform the 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 as compositional vectors

Let us first recall that compositional data (hereafter referred to as CoDa) vectors can be defined as vectors consisting of D positive components that sum 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 typically 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. [3].

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 enable the definition of a meaningful 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 components of this average are just the geometric average of the corresponding sample's components).

The clr transformation of a vector $\mathbf{u} \in \mathcal{S}^D$ is defined by

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

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

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

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

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

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

2.2 Continuous densities as elements of the Bayes space

As outlined in [9], density functions can be considered elements of the so-called Bayes space denoted by $\mathcal{B}^2([a,b])$ and comprising positive functions integrating to one on a bounded interval $[a,b]$ whose log-transform is square integrable. This concept corresponds to a particular case of that introduced in [7] for the reference measure being the Lebesgue measure. Discrete compositional data analysis is often justified by the scale invariance property of the CoDa vectors which make necessary the use of ratios of components. Let us briefly develop the meaning of scale invariance for densities. [19] proved that, when two r.v. are continuous with respect to Lebesgue measure, the p.d.f. of the relative distribution of one random variable (the comparison r.v.) with respect to another one (the reference r.v.) coincides with the ratio of the two densities evaluated at a given quantile of the reference density (see also [18]). This space can first be equipped with a vector space structure using the following operations. For any positive function on $[a,b]$, let us define its closure $\mathcal{C}(s)$ of s to be the unique density proportional to it. Subsequently, for any two functions f and g in $\mathcal{B}^2([a,b])$ and any real α, the following operations can be defined

- 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

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

Through its 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. The inverse transformation is well defined and can be expressed as follows for a function $f_0 \in L_0^2([a,b])$,

$$\mathrm{clr}^{-1}(f_0)(t) = \mathcal{C}\exp\left(\mathrm{clr} f_0(t)\right).$$

$\mathcal{B}^2([a,b])$ can then be equipped with the following inner product rendering 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

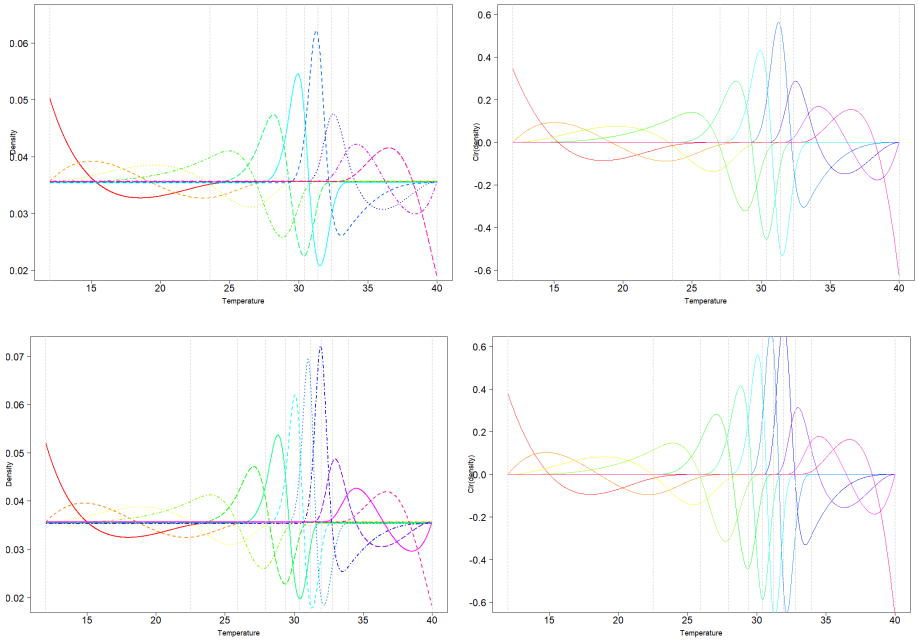
Spline functions are constructed by piecing together segments of polynomials of a specified degree connecting at specified knots points while adhering to prescribed smoothness conditions [see e.g. 20]. In our context, aimed at approximating density functions, we require a specific type of constrained splines. One approach to constructing them is described in [9] using the so-called ZB-splines in $L_0^2([a,b])$ and corresponding CB-splines in $\mathcal{B}^2([a,b])$. As is common in many CoDa techniques, the procedure is based on a log-ratio transformation, specifically the clr introduced in Section 2.2. The process starts by constructing a basis of spline functions that fulfill the integral constraint within $L_0^2([a,b])$. These basis functions are then pulled back to $\mathcal{B}^2([a,b])$ by the inverse clr transformation. The ultimate system of B-splines is entirely characterized by a sequence of knots (points where polynomial pieces connect) and an order (equal to the 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 introduced 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. Consequently the dimension of the ZB-splines basis (a 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 $g + d$. Notably,one dimension is lost for the ZB-basis due to the integral constraint. The inverse clr of the ZB-basis functions are termed the CB-basis functions. For this application, we use exclusively 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. Equation (17) in [9] establish a correspondence between the representation of any function in Z_k^Λ within both basis systems. This correspondence proves invaluable as it facilitates the manipulation of ZB-splines using conventional code originally designed for B-splines.

In our subsequent application, the temperature data will first be processed into a set of histograms, each depicting daily maximum and minimum temperatures for a specific province and year. For maximum temperatures, the data is discretized into 28 bins of length 1 within the interval $[a,b] = [12, 40]$. To approximate the underlying densities represented by these original histograms, we employ cubic splines ($k = 3$) and set $g = 7$ (respectively $g = 9$) as the number of inside knots. Consequently, the dimension of the ZB-spline basis becomes $7 + 3 = 10$ when using 7 inside knots (respectively $9 + 3 = 12$ for 9 inside knots). For minimum temperatures, the data is discretized into 22 bins of length 1 within the interval $[a,b] = [7, 29]$. To approximate the underlying densities represented by these original histograms, we employ cubic splines ($k = 3$) and set $g = 9$ as the number of inside knots.

In both cases, the positioning of the knots is determined relative to the data points position using quantiles as argued in [9].

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 vertical dotted lines on the plots indicate the knots position. We observe that the inclusion of two additional knots in the lower plots results in an increased number of basis functions that concentrate around the mode of the distribution. This enhancement enables a more precise approximation of the densities, particularly in regions where our dataset features a higher density of temperature data points.

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



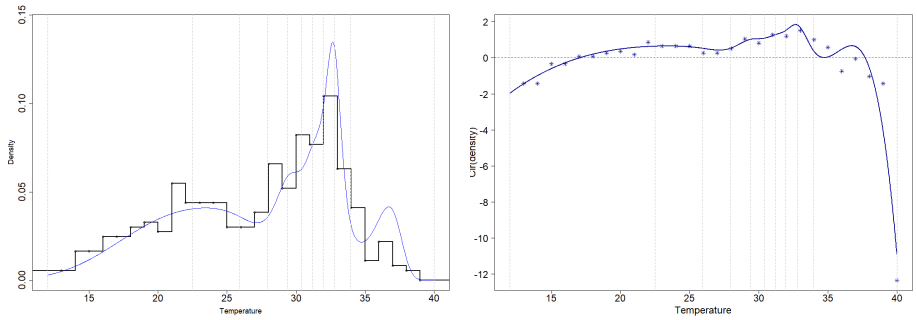
2.4 Smoothing histograms with CB-splines

Our original temperature data comprises sequences of daily maximum temperatures. In order to apply the same technique as [9], we first preprocess the data into intermediate histogram representations. Subsequently, we transform these histograms into smooth density functions using CB-splines as in [21]. The CB-spline smoothing step involves 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, we explain the clr transformed histogram frequencies by covariates derived from evaluating the ZB-spline basis functions at the midpoints of the histograms bins. To ensure the existence and uniqueness of the least squares problem (full column rank of the collocation matrix), we enforce an upper limit on the number

of knots. This upper bound is dictated by the Schoenberg-Whitney conditions (see [22]). In our application, the condition, both for maximum and minimum temperature, stipulates that the number of knots must be less than or equal to the number of bins minus 3 (degree of splines). Smoothing with ZB splines does not accommodate bins with zero counts because of the log transformation. To address this limitation, we implement a simple zero-replacement procedure: any zero count is substituted by 10^{-7} after which we apply the closure operator. For the selection of the smoothing parameter, we opt for a generalized cross-validation using a regular grid of 100 points on a log-scale.

As an illustrative example, Figure 2 displays the 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.

Fig. 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 concerning rice yield is sourced from the International Rice Research Institute[1]. The data set contains comprehensive information on annual rice production, harvested area, and rice yield at provincial level from 1987 to 2016. Rice yield is quantified in tons per hectares. Figure 3 provides an overview of the overall evolution of rice yield over the considered period. After a period of stagnation between 1987 and 1992, rice yield has exhibited consistent growth since 1992, affecting all Vietnamese provinces. This growth may be attributed to the progress of agronomic techniques over the years. While we lack a direct proxy for this progress, we will account for it through the incorporation of a linear time trend. This choice is supported by Figure 4, which reports the evolution of average rice yields for the six different agronomic regions in Vietnam. In this figure, we use the following acronyms for the regions: NMM

[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/

for Northern Midland and Montainous region, NCC for North Central Coast region, CHR for Central Highlands region, SR for Southeast region, MDR for Mekong Delta River region and RRD for Red River Delta region.

Fig. 3 Rice yield distributions from 1987 to 2016

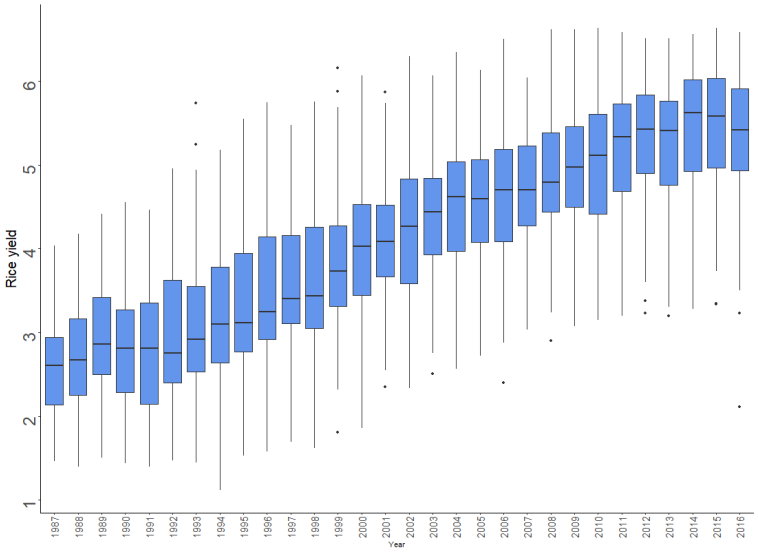


Fig. 4 Average rice yield by agronomic regions from 1987 to 2016

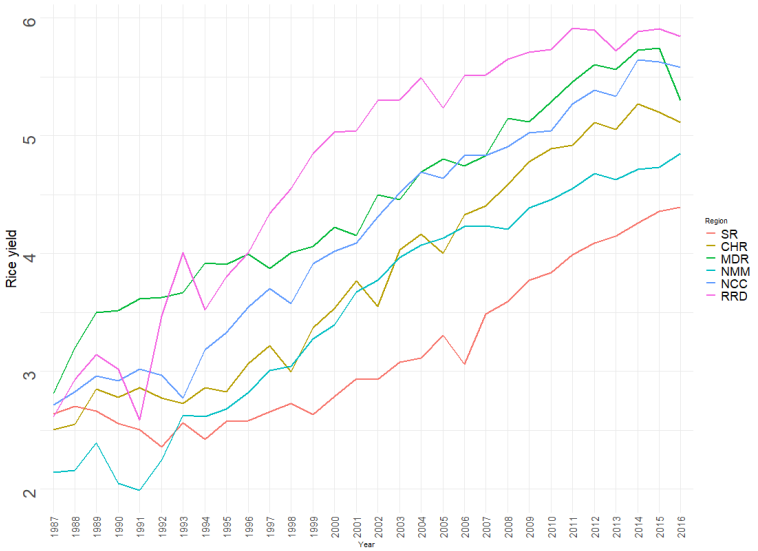
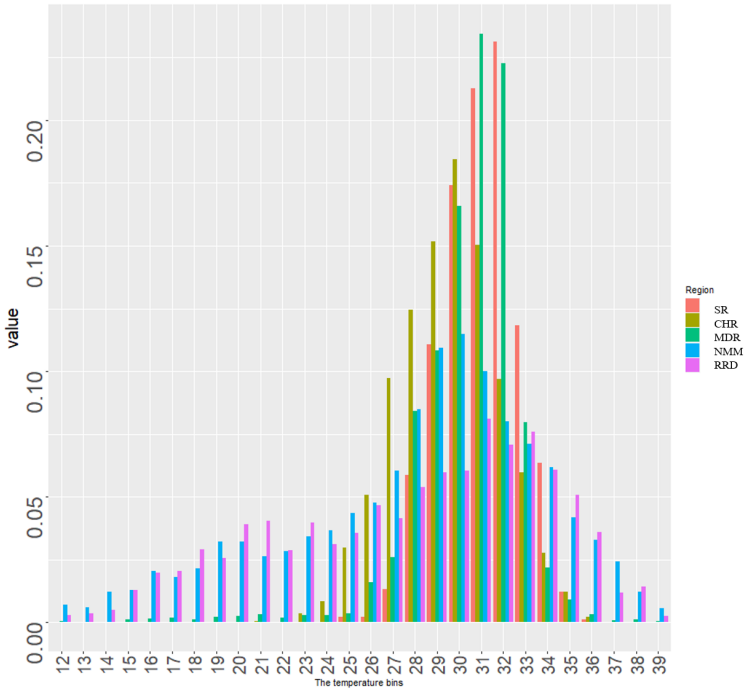


Fig. 5 Maximum temperature histograms across the Vietnamese regions in 2015



3.2 Weather data

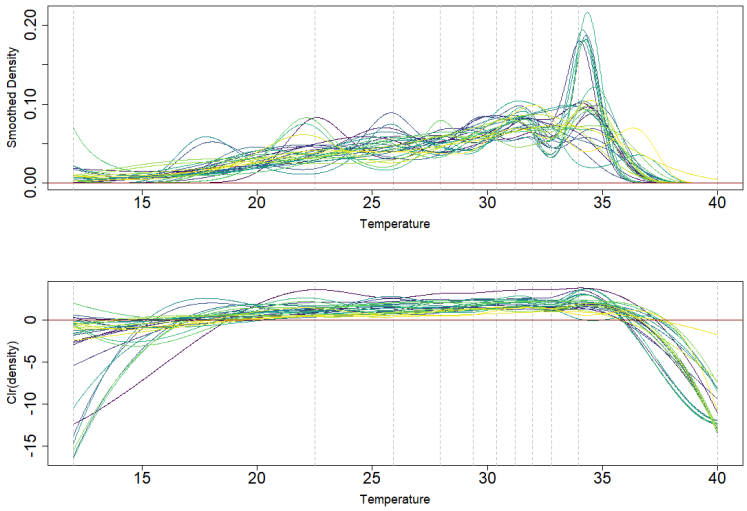
The weather data used in this study encompasses daily maximum-temperatures and precipitation records. Temperature data comes from the Climate Prediction Center (CPC) database developed and maintained by the National Oceanic and Atmospheric Administration (NOAA). We have retrieved historical information pertaining to daily maximum temperatures for a grid with a resolution of 0.50×0.50 degrees of latitude and longitude, specifically for the geographical expanse of Vietnam. Subsequently, we have transformed this data to yield the daily maximum temperature for each of 63 Vietnamese provinces and during a period of 30 years (1987-2016) (365 or 366 values for each year). The compilation yields one temperature distribution for each of 1890 statistical units.

Figure 5 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 provide a visual representation of how the range of maximum temperatures varies across different regions, emphasizing the substantial regional disparities.

Using the CB-spline smoothing tool we can also explore other aspects of the temperature densities variations across time and space. Figure 6 displays the daily maximum temperature density with 9 knots (along with its clr transform) in the province of Ninh Binh which is one of the major provinces for rice production situated in the Red River Delta region. We use the viridis color palette with 30 values, featuring 30 distinct values that transition from yellow in 1987 to dark violet in 2016 with

intermediate shades of green. The top part of Figure 6 clearly reveals the rightward shift of the temperature densities corresponding to climate change. Finally Figure 7 displays the densities and their clr transforms for all provinces in the year 2015 (9 inside knots). When examining the clr transforms, we can see groups of provinces and it would be interesting to explore their respective spatial position. It seems that they primarily differ in the range of the observed maximum temperatures.

Fig. 6 Density (top panel) and clr transform (bottom panel) of the smoothed daily maximum temperature from 1987-2016 in Ninh Binh province



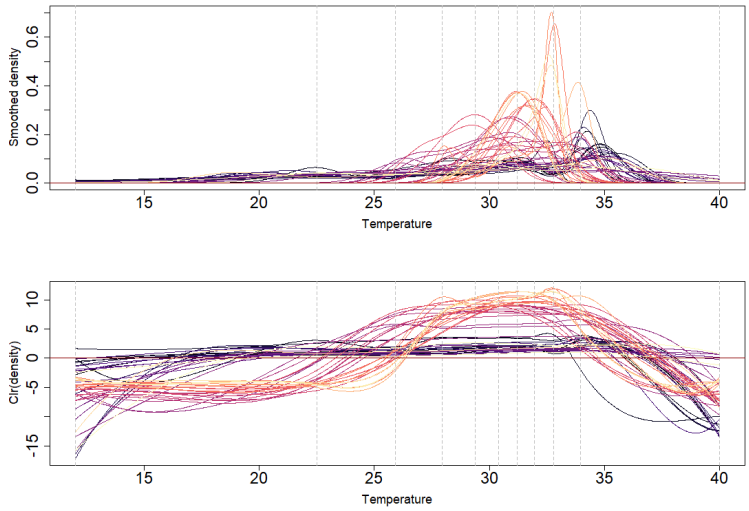
To facilitate the integration of the smoothed histograms into the subsequent regression model, it is imperative to ensure that they are expressed in a uniform basis of CB-splines. Consequently, we must employ a consistent set of knots across all $63 * 30 = 1890$ histograms. For this reason, we first pool all observations into a single distribution and place the knots at the quantiles of this global distribution.

Improving this phase of the process hinges on obtaining information about the specific starting and ending dates of the growing season within each province. However since these temporal boundaries may exhibit substantial variability across geographical regions as we have seen in section 3.2, the adoption of a standardized temperature range across all provinces would then be rendered difficult, unless we find a way of overcoming this technical constraint.

3.3 Climate change data

Let us first examine the "historical" climate change between 1987 and 2016. Using relative distributions for comparing two distributions as recommended by [18], Figure 8 showcases boxplots depicting the ratios of 2016 to 1987 densities across provinces in some regions for maximum and minimum temperatures. Notably, this analysis

Fig. 7 Density (top) and Clr transform (bottom) of the smoothed daily maximum temperature from in 2015 for all provinces



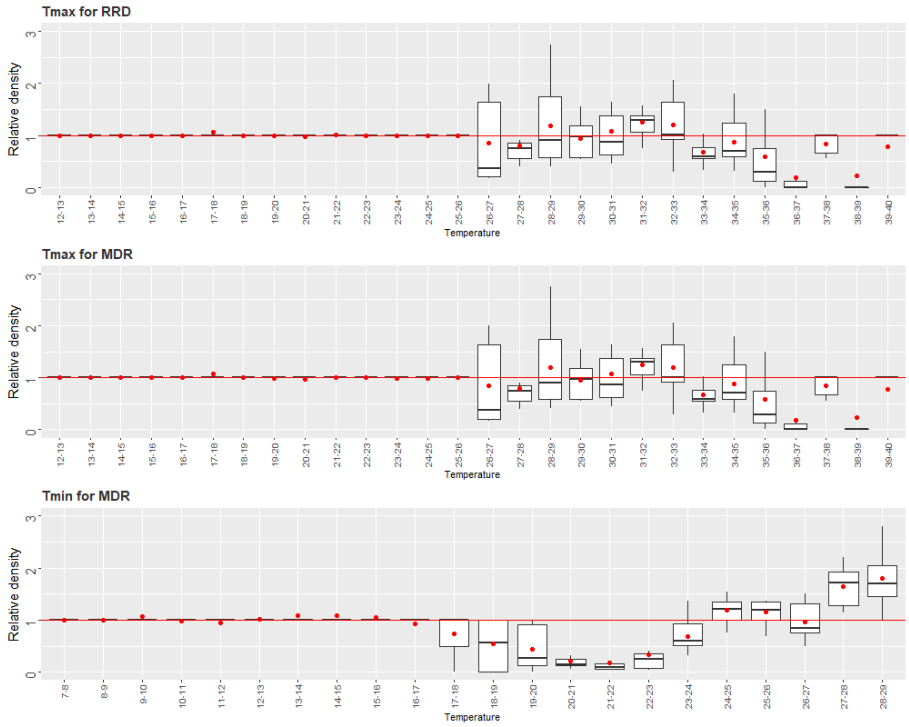
highlights temperature ranges affected by changes, specifically ranging from 26 to 40 degrees Celsius for maximum temperature and 17 to 29 degrees Celsius for minimum temperature. Moreover for maximum temperature in the Red River Delta and Mekong Delta regions, a concentration of temperature increase is observed within the 27 to 33 degrees Celsius range. In contrast, for minimum temperature, we observe an increasing trend of the ratio indicating a shift of this density to the right.

The Intergovernmental Panel on Climate Change (IPCC) provides projections of global CO_2 emissions and associated temperature distributions around the world under several scenarios associated to representative concentration pathways (RCPs) for the end of this century, see for example [23]. We will use the most optimistic RCP called RCP2.6, which projects an average increase of 1 degree Celsius relative to the period 1986-2005. In Section 5, we construct a climate change scenario for 2099 which approximates RCP2.6 and is more easily handled in our framework. The RCP2.6 data for vietnamese provinces come from [24].

4 The discrete and smooth regression models

The objective in this application is to develop a regression model to unravel the relationship between rice yield and the distribution of daily maximum and minimum temperatures for the corresponding year and province, while also controlling for additional covariates. Unlike a conventional time series model used for yield prediction in the future, our focus here is to leverage spatio-temporal variability to quantify the influence of temperature on rice yield. Therefore we decide to include a simple linear time trend in the model as a proxy for unobserved factors that may have evolved over time, such as advancements in production techniques. In view of Figures 3 and 4, the

Fig. 8 Relative distribution of daily temperature for 2016 versus 1987 in some regions



inclusion of a linear trend appears to be a reasonable choice. We further use other controlling factors namely precipitation and regional dummies. Given the distributional nature of our primary covariate, we need an adapted regression model. The choice boils down to either utilizing a histogram of daily temperatures as a compositional covariate, akin to the approach in [5], or opting for a smoothed representation of the temperature density as a continuous density covariate, following the method outlined in [8]. Before delving into the results, let us first revisit the fundamental principles behind these two models.

4.1 The discrete regression model

The scalar-on-composition regression model as presented for example in [5] constitutes a regression framework where at least one of the covariates takes the form of a compositional vector. In our discrete regression setup, the compositional vectors are temperature histograms which can also be viewed as discrete densities. 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 $< ., . >_A$ is the classical Aitchison inner product in $\mathcal{S}^D$ (see e.g. [4]). Therefore a linear model designed 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

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

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, it is essential to index all observations according to both the province i and the year k therefore the single index i of equation (3) is replaced by the two indices i and k. This adjustment allows us to define Y_{ik} as the rice yield for province i (ranging from 1 to 63) in year k (spanning from 1 to 30). Initially, the model comprises several classical scalar variables ($L = 7$) including time, precipitation and five regional dummies (reference region being CHR). In addition to these, we also incorporate two discrete densities as compositional covariates, namely the histograms of maximum and minimum daily temperature, reported with equal bins of length 1 degree Celsius. Moreover, after testing the inclusion of interactions between the two discrete densities and the six regional dummies, we decide to integrate the interactions solely for maximum temperature and three specific regions: RRD, CHR and MDR. As a result, we get $J = 5$ parameters associated with the discrete densities and denoted by $\beta^{max}_{RRD}, \beta^{max}_{NCC}, \beta^{max}_{MDR}$ and β^{max}_{other} for the maximum temperature and β^{min} for minimum temperature.

As demonstrated for example in [25], after transformation of the compositional covariates by any transformation in the log-ratio family (isometric or additive log-ratio), the estimation of such a model is done by ordinary least squares. The choice of any of these transformation will yield the same result for the discrete densities contribution when expressed as a linear combination of the logarithm of the histogram bin frequencies with a zero sum contraint on the coefficients.

4.2 The smooth regression model

Extending the model in [8] to the case of several density covariates as well as additional scalar covariates, we consider the following linear scalar on density regression model

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

where Y_i is the scalar dependent variable, β_0 is a real intercept, $\beta_j(t), j = 1, \dots J$ are curve-parameters for the effects of the densities f_{ij}, Z_l ($l = 1, \dots, L$) 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 $L^2_0([a, b])$ equipped with their respective inner products, we can rewrite the model as follows

$$Y_i = \beta_0 + \sum_{j=1}^{J} < \mathrm{clr}\beta_j(t), \mathrm{clr} f_{ij}(t) >_{L^2_0} + \sum_{l=1}^{L} \gamma_l \mathbf{Z}_{il} + \epsilon_i. \tag{5}$$

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. The expansion can be written directly in $\mathcal{B}^2([a,b])$ or equivalently for the clr transforms in $L_0^2([a,b])$. We then replace these functions by their expansions in the inner products of the model equation (4). Consequently, the inner products terms appear as linear combinations of the beta curves coordinates whose coefficients are given by the product of the Gram matrix (inner products of all pairs of basis functions) by the densities coordinates as in [8]. 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 for the province and k for the year. β_0 is a real intercept and we have the same $L = 7$ classical covariates as for the discrete model (time, precipitation and regional dummies) with their corresponding parameters γ_l. As for the discrete model, we include two smooth density covariates f_{ik}^{max} and f_{ik}^{min}, which are respectively the densities of daily maximum and minimum temperature, in province i and year k. To facilitate the comparison, we include the same interactions between densities and regional dummies. The corresponding curve-parameters will be denoted by $\beta_{RRD}^{max}(t), \beta_{NCC}^{max}(t), \beta_{MDR}^{max}(t)$ and $\beta_{other}^{max}(t)$ for the maximum temperature and $\beta^{min}(t)$ for minimum temperature. Finally ϵ_{ik} are normal errors with mean zero and standard deviation σ^2. f_{ik}^{max}, f_{ik}^{min} as well as all the curve-parameters 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. Respecting the Schoenberg-Whitney conditions of Section 2.4, we will test two different knots numbers 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.

Let us note an important difference between the discrete and the smooth model. Conventional compositional data analysis does not pay attention to the order of the components (permutation invariance). However in our case, for a temperature histogram, the components correspond to temperature bins and the order of these bins should be considered to take into account some continuity of the bin frequencies with respect to the bins positions on the temperature axis. In contrast the smooth approach does this into account.

4.3 Model results

The histograms smoothing step and the fitting of both models are performed with the R packages *compositions* and *robCompositions*, adapting some codes from [26]. The analysis of variance table of the discrete model (Table 1) reveals that all covariates are strongly significant.

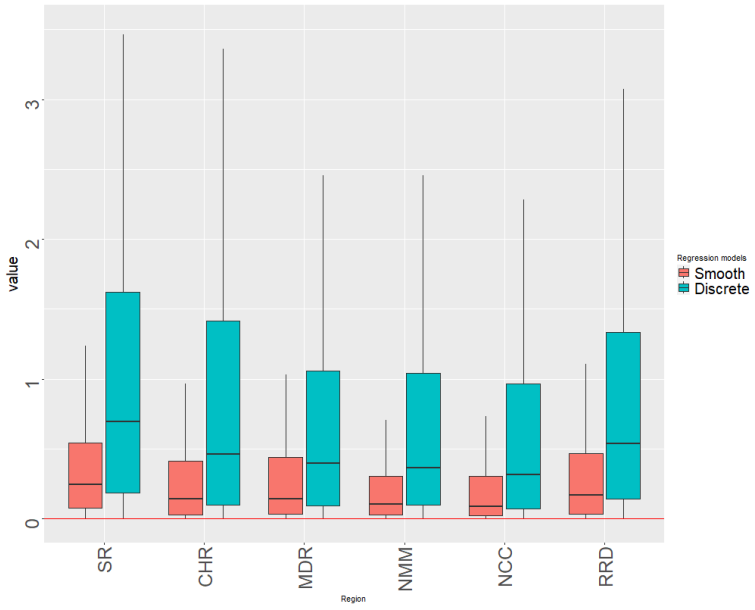
The fits of the smooth model with 7 or 9 knots reveal the superiority of the second option which we keep thereafter. The smooth model with 9 knots has a better fit than the discrete as shown in Figure 9 with the distance between fitted and observed values

for both models. The parameters estimates for classical variables displayed in Table 2 are comparable between discrete and smooth models.

Table 1 Anova for Discrete model

Variables	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	1	1522.71	1522.71	5464.70	0.0000
Total precipitation	1	12.80	12.80	45.94	0.0000
SR	1	152.21	152.21	546.27	0.0000
MKD	1	37.99	37.99	136.35	0.0000
NMM	1	155.61	155.61	558.46	0.0000
NCC	1	15.15	15.15	54.37	0.0000
RRD	1	88.14	88.14	316.33	0.0000
ilr(tmax)	27	42.53	1.58	5.65	0.0000
ilr(tmin)	21	94.26	4.49	16.11	0.0000
MKD:ilr(tmax)	27	98.64	3.65	13.11	0.0000
NCC:ilr(tmax)	27	60.53	2.24	8.05	0.0000
RRD:ilr(tmax)	27	32.60	1.21	4.33	0.0000
Residuals	1753	488.46	0.28		

Fig. 9 Aitchison distance between observed and fitted values by models and regions



The interpretation of parameters of a compositional covariate is presented for example in [25]. In the discrete case, as in [27], we interpret the difference between clr parameters as related to the influence of pairwise log-ratios. Similarly, for the smooth regression model, the ratio between the density value at two given points is related to the relative change of the density between these two points and therefore as in the

discrete case, the detection of couples of temperatures leading to the highest values of this ratio indicate the regions of most influential contrasts for the density covariate. Figure 10 shows the estimated clr parameters for maximum temperature in the discrete model. For the Red River Delta region, we see that the highest clr coefficient corresponds to the temperature bin 31-32 and the lowest clr to the temperature bin 23-24. This reveals that the contrast between these two bins, and therefore any change in the corresponding ratio, has the highest impact on rice yield in the model. A similar result appears for the North Coastal Central region. In the Mekong Delta region, the most influential contrast occurs between the bins 30-31 and 32-33. Turning now attention to the smooth model, the curves of the different functional parameters $\hat{\beta}^{max}$ on Figure 12, respectively $\hat{\beta}^{min}$ on Figure 13, are presented in the functional clr space on the right plot and in the functional Bayes space on the left plot. Comparing the right plot with Figure 10, we can see that in both models, it is the Mekong Delta and the Red River Delta regions which undergo the highest impacts of climate change. In the MDR region, the smooth model reveals a contrast between temperatures around 32 and 33 degree Celsius showing that the ratio with the highest impact is that contrasting the density at these two points. The detection of the influential ratios may be affected by the choice of parameters: in the discrete case by the bin size choice and the end-point of the first bin and in the smooth case by the number of knots of the spline approximation. However it is known that a small change in the end-point of the first bin can dramatically affect the histogram whereas the smooth approach does not suffer from that drawback. Figures 11 and 13 show these parameters for minimum temperature. For minimum temperature, the highest contrast is between 24 and 27 degrees Celsius in the smooth case whereas it is between bins 23-24 and 26-27 in the smooth case. Overall we can say that the results of both models are coherent but more precise and possibly less sensitive to parameter choices for the smooth model.

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 [25] or infinitesimal increments as in Morais et al. [28]. To simplify comparisons between the discrete and the smooth regression models, and because the changes we envision for the covariates cannot be considered as infinitesimal in the present case, we select a finite increment perspective. In order to assess the impact of a compositional covariate in a model such as (3) or (4), we imagine possible change scenarios for this covariate. In the discrete case, to be coherent with the simplex space to which the histograms belong, it would be desirable that the change scenario be linear with respect to the vector space structure of the simplex $\mathcal{S}^{28}$ whereas in the smooth case, it would be linear with respect to the vector space structure of the Bayes space $\mathcal{B}^2([a,b])$. Let us first look at what are linear changes in these two frameworks.

In the discrete case, the perturbation of a histogram f by a change scenario of direction $\varphi \in \mathcal{S}^{28}$ and intensity $h \in \mathbb{R}$ is given by

$$Tf = f \oplus (h \odot \varphi)\,, \tag{6}$$

Table 2 Estimated coefficients associated to regional dummies, total precipitation and year

	Regression type	
Variable	Discrete regression	Smooth regression (9 knots)
Constant	3.31^{***}	3.313^{***}
	(0.31)	(0.259)
Region		
NMM	-0.34	-0.43^{**}
	(0.26)	(0.21)
NCC	-1.42^{***}	-1.44^{***}
	(0.32)	(0.28)
RRD	-0.88^{**}	-0.47
	(0.36)	(0.30)
SR	-1.63^{***}	-1.67^{***}
	(0.10)	(0.09)
MDR	0.62	0.74^{**}
	(0.40)	(0.34)
(Reference = CHR)		
Total precipitation	-0.005	0.004
(Thousand ml per year)	(0.03)	(0.03)
Year	0.10^{***}	0.10^{***}
	(0.002)	(0.002)
Adjusted R^2	0.81	0.811
Residual Std. Error	0.53 (df = 1753)	0.539 (df = 1822)
F Statistic	61.04^{***}	116.921^{***}
	(df = 136; 1753)	(df = 67; 1822)
RMSE	1.10	0.53

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

where φ is a direction of change in $\mathcal{S}^{28}$. Equivalently we may write $h \odot \varphi = Tf \ominus f$, and therefore the change vector is given by

$$h \odot \varphi = \mathcal{C}(\frac{Tf_1}{f_1}, \ldots, \frac{Tf_D}{f_D}), \tag{7}$$

emphasizing the fact that the change from the initial distribution f to Tf is a relative change in the original scale of frequencies.

Similarly in the smooth case, and using on purpose the same notation for a different object, the perturbation of a density f by a change scenario $\varphi \in \mathcal{B}^2([a,b])$ and intensity $h \in \mathbb{R}$ is given by

$$Tf(t) = f(t) \oplus (h \odot \varphi(t)), \tag{8}$$

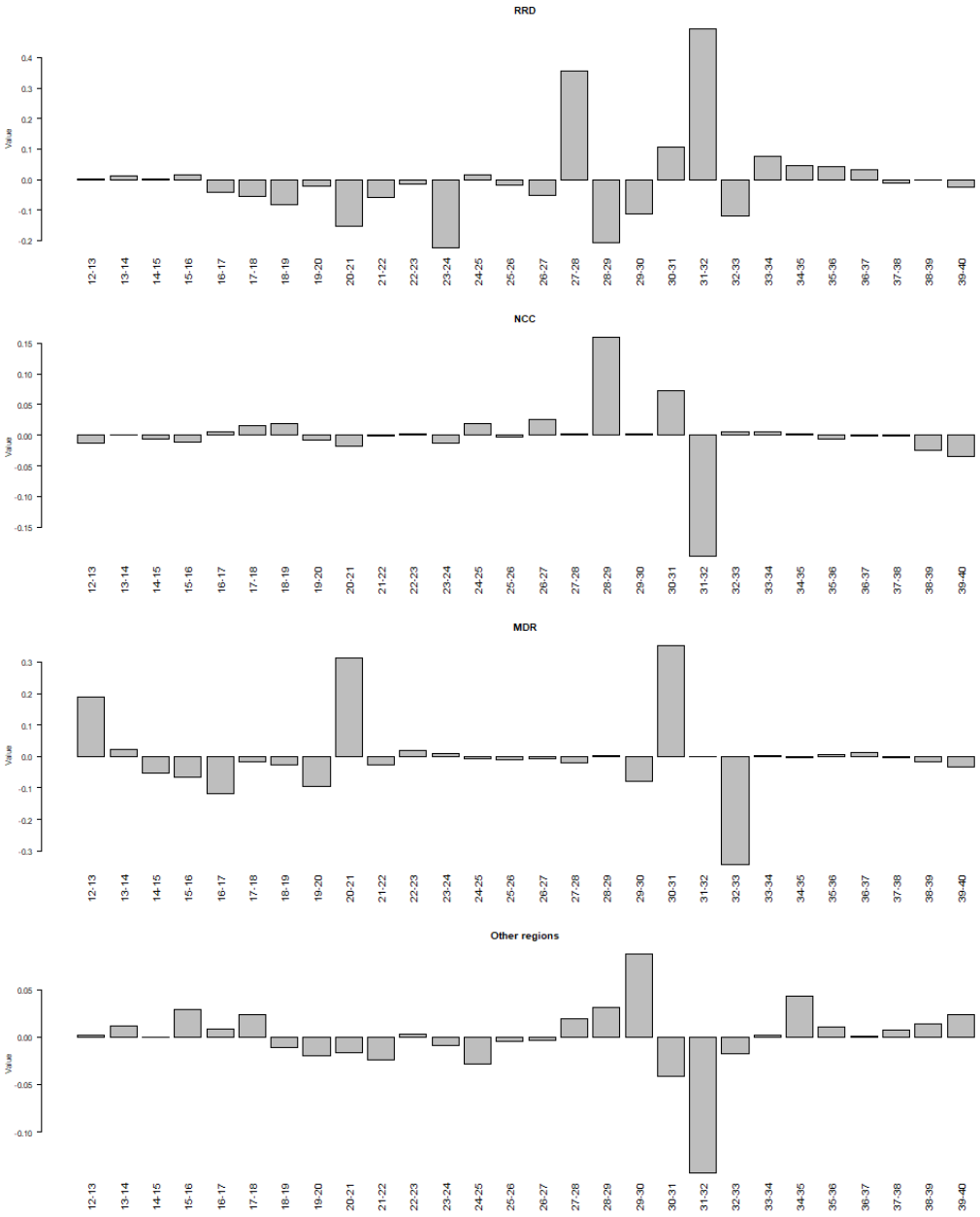
where $\varphi(t)$ is a direction of change in $\mathcal{B}^2([a,b])$. Note that, in clr space, the change writes as follows in the discrete case and

$$\mathrm{clr}Tf = \mathrm{clr}f + h\mathrm{clr}\varphi \tag{9}$$

and as follows in the smooth case

$$\mathrm{clr}Tf(t) = \mathrm{clr}f(t) + h\mathrm{clr}\varphi(t). \tag{10}$$

Fig. 10 Estimated clr coefficients of maximum temperature for some regions



Ideally, we would like to compute projections of rice yield corresponding to the IPPC projections of the temperature, see [29]. However these would give rise to a different φ vector or curve for each province, resulting in computational complexity when evaluating impact significance. Therefore we create working scenarios which are

Fig. 11 Estimated coefficients of tmax and tmin in the discrete regression - other regions

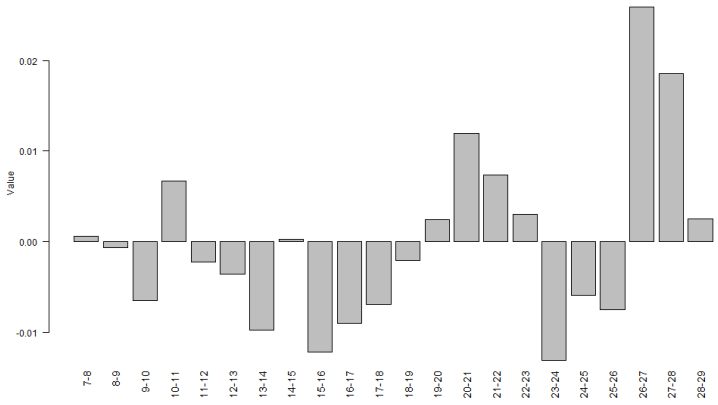


Fig. 12 Curves of $\hat{\beta}^{max}$ in the smooth regression model with interactions for all regions. Left: in B^2, right: in L^2

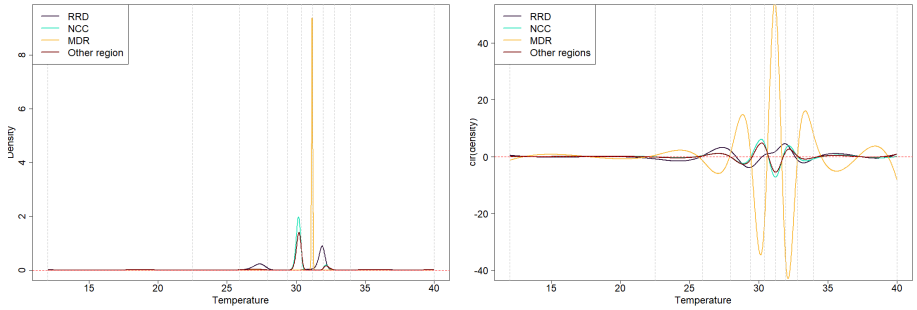
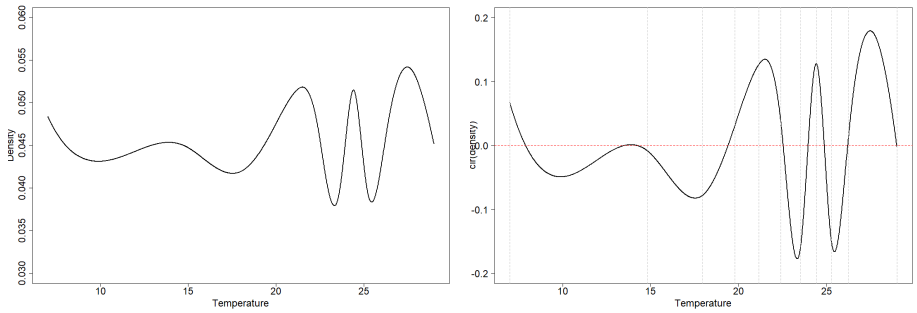


Fig. 13 Curves of $\hat{\beta}^{min}$ in the smooth regression model with interactions for all regions. Left: in B^2, right: in L^2



linear (same φ for each province) but not so far from the IPPC projections for the

end of the century. Before tuning their parameters to get approximations of the IPPC projections, let us first describe the class of working scenario.

To simplify further the comparison between discrete and smooth cases, we decide to choose a change direction curve in $\mathcal{B}^2([a,b])$ which coincides with a histogram function with bin length of one so that the change curve $\varphi(t)$ is totally determined by the vector of histogram frequencies φ used for the discrete model and therrfore the climate change is common to both cases.

In order to construct a manageable but plausible φ which would yield changes similar to those projected by RCP2.6, we propose to use the following type of φ. In the discrete case, we define φ to be the closure of the vector of ones except for bins between bin m_l and bin m_u where one is replaced by $\exp(h)$. After visual inspection of the histograms in Figures 14 and 15 of the temperature distributions as observed in 2016 (the last year of our panel) and as projected by RCP2.6 (the most optimistic IPCC scenario), we select reasonable values for m_l, m_u and h : for maximum temperature $m_l = 29, m_u = 40, h = 5$ and for minimum temperature $m_l = 23, m_u = 26, h = 5$.

For tmax Region RRD h=3 ml=19 mu= 34 Region NCC h= 4 ml= 22 mu= 34 Region CH h= 3 ml= 26 mu= 31 Region MKD h=3 ml= 30 mu=34

For tmin Region RRD h=5 ml= 14 mu= 27 Region NCC h= 5 ml= 19 mu= 25 Region CH h=5 ml=16 mu=22 Region MKD h= 5 ml= 23 mu=26

Alternative choices of working scenarios are of course possible.
From equation (7) it is easy to see that if $l' < m_l \quad \text{or} \quad l' > m_u$ and $m_u \leq l \leq m_u$, we have

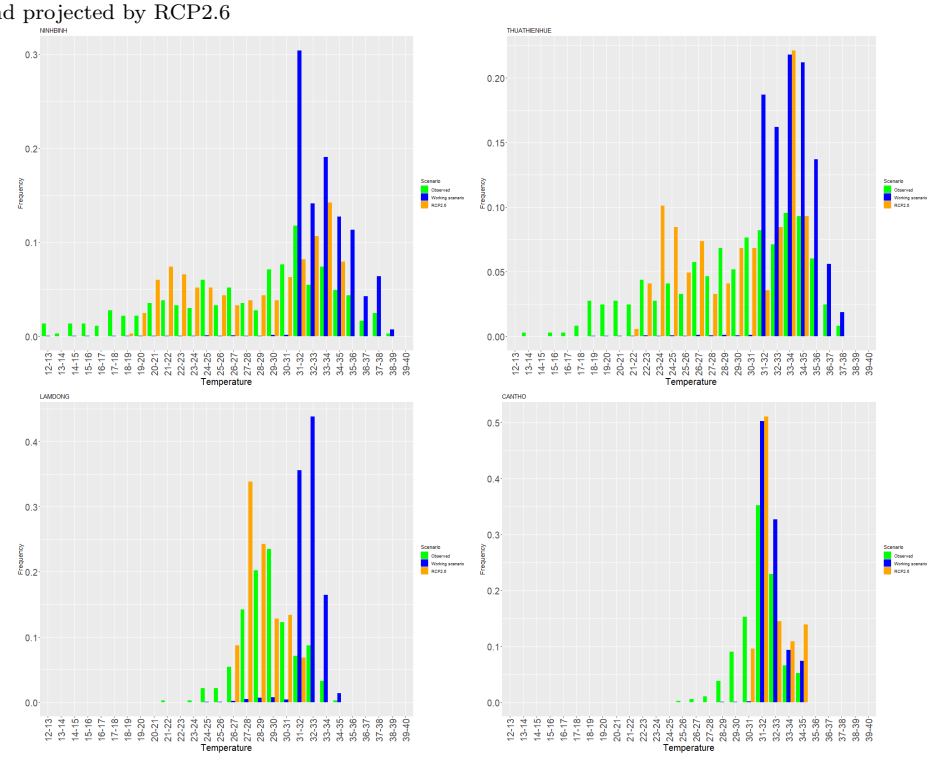
$$\frac{\frac{Tf_l}{f_l}}{\frac{Tf_{l'}}{f_{l'}}} = \exp(h) = \frac{\frac{Tf_l}{Tf_{l'}}}{\frac{f_l}{f_{l'}}} \tag{11}$$

In the smooth case, a similar formula is true replacing f_l and $f_{l'}$ by $f(t)$ and $f(t')$. Formula (11) and its smooth counterpart can be interpreted as follows: after the change, the frequency outside of the interval $[m_l, m_u]$ is equal to the frequency inside this interval before change multiplied by a factor of $\exp(-5) \sim 0.0067$. The effect is to inflate the density inside the interval $[m_l, m_u]$.

The histograms RCP2.6 of Figures 14 (for maximum temperature) and 15 (for maximum temperature) illustrate the resulting change scenario for four selected provinces (Ninh Binh for the RRD region, Thua Thien Hue for the NCC region, Lam Dong for the CHR region and Can Tho for the MDR region). We observe that the frequencies in warm temperature bins are higher after the change but we can see 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 are now in position to compute a projection of the resulting rice yield change $\hat{Y}_{ik}(h,\varphi) - Y_{ik}$, where $\hat{Y}_{ik}(h,\varphi)$ denotes the projected rice yield under the change scenario. Given that both our models are linear for the simplex structure and that $T_h f_{ik}(t) - f_{ik}(t) = h \odot \varphi(t)$, the resulting change of rice yield for a given province i and a given year k are given by

Fig. 14 Maximum temperature distributions in four provinces: observed in 2016, working scenario and projected by RCP2.6



- in the discrete regression model

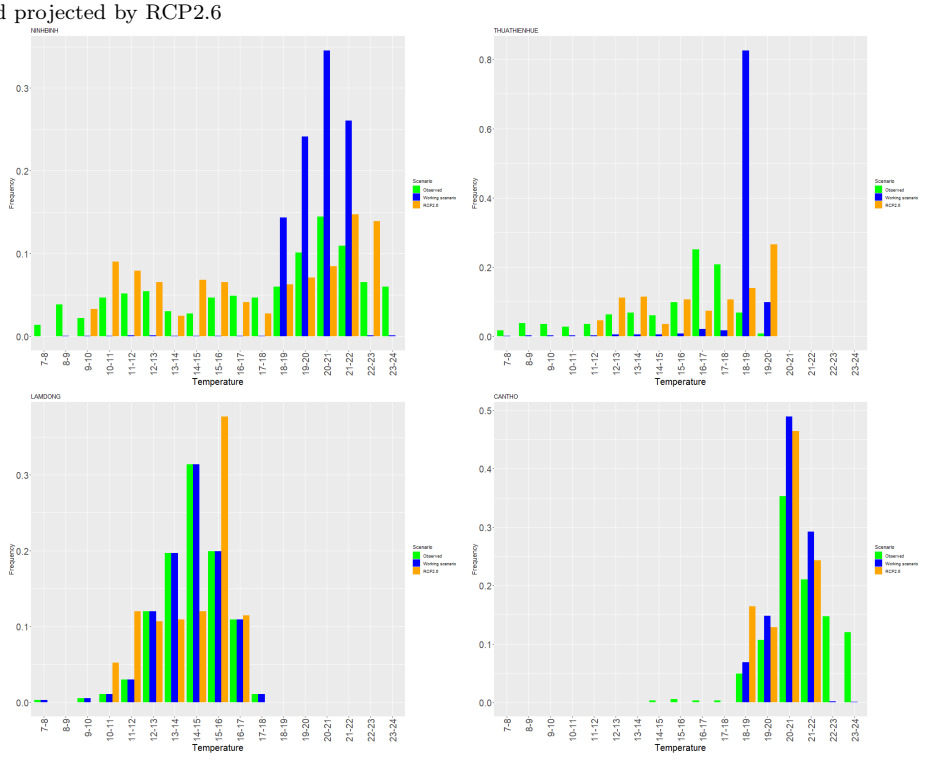
$$\begin{aligned}\hat{Y}_{ik}(h,\varphi) - \hat{Y}_{ik} &= h\sum_{r=1}^{6} \mathbf{1}_{k\in r}\{< \hat{\beta}_r^{max}, \varphi >_A + < \hat{\beta}_r^{min}, \varphi >_A\} \\ &= h\sum_{r=1}^{6} \mathbf{1}_{k\in r}\{< \mathrm{clr}\hat{\beta}_r^{max}, \mathrm{clr}\varphi >_{\mathbb{R}^{28}} + < \mathrm{clr}\hat{\beta}_r^{min}, \mathrm{clr}\varphi >_{\mathbb{R}^{28}}\},\end{aligned} \tag{12}$$

- in the smooth regression model

$$\begin{aligned}\hat{Y}_{ik}(h,\varphi) - \hat{Y}_{ik} &= h\sum_{r=1}^{6} \mathbf{1}_{k\in r}\{< \hat{\beta}_r^{max}(t), \varphi(t) >_{\mathcal{B}^2} + < \hat{\beta}_r^{min}(t), \varphi(t) >_{\mathcal{B}^2}\} \\ &= h\sum_{r=1}^{6} \mathbf{1}_{k\in r}\{< \mathrm{clr}\hat{\beta}_r^{max}(t), \mathrm{clr}\varphi(t) >_{L_0^2} + < \mathrm{clr}\hat{\beta}_r^{min}(t), \mathrm{clr}\varphi(t) >_{L_0^2}\},\end{aligned} \tag{13}$$

Note that since a given province belongs to a single region, there is indeed a single non-zero term in the right hand side sums. The inner products $< \mathrm{clr}\hat{\beta}_{max}, \mathrm{clr}\varphi >_{\mathbb{R}^{28}}$, and $< \mathrm{clr}\hat{\beta}_{min}, \mathrm{clr}\varphi >_{\mathbb{R}^{28}}$ in the discrete model, respectiveley $< \mathrm{clr}\hat{\beta}_{max}(t), \mathrm{clr}\varphi(t) >_{L_0^2}$

Fig. 15 Minimum temperature distributions in four provinces: observed in 2016, working scenario and projected by RCP2.6



and $< \text{clr}\hat{\beta}_{min}(t), \text{clr}\varphi(t) >_{L_0^2}$ in the smooth model therefore characterize the impacts of a change in temperature density in the respective models and these are constant for all provinces in both models for our working scenario. The computation of the variance of the impacts is derived in the Appendix Section.

Table 3 displays the impacts and their standard error in our application.

Table 3 Climate impact on yield using the working scenario

Type	Regions	Discrete regression		Smooth regression	
		Value	Standard error	Value	Standard error
tmax	RRD	2.6690	0.9092	20.0700	0.8626
	NCC	-1.2590	0.4896	-14.0736	0.7555
	MDR	-1.9395	0.5538	31.9298	0.5625
	Other regions	-0.2969	0.3329	-6.3677	0.6948
	Average impact*	-0.3216		6.1964	
tmin	All regions	0.0004	0.0341	−0.7754	0.0654
Overall (tmin and tmax)	All regions	-0.3212		5.4210	

*The average impact for maximum temperature is computed as a weighted average taking into account the number of provinces in each region.

6 Conclusion

We have proposed and illustrated a procedure for assessing the impact of climate change on rice yield production in Vietnam using scalar on density regression with a discrete and a smooth frameworks. 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 comes out larger than that measured by the discrete one.

Some aspects of this model correspond to preliminary choices. For example, a more realistic assessment would take into account the cropping season in each region if the cropping season data were available. However, dealing with density covariates with varying supports would have opened other issues from the methodological side since then we couldn't have used the same spline basis for all densities.

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 each having a different smoothing parameter.

In order to measure the impact of climate change, we have chosen to consider simplified change scenarios as close as possible to RCP scenarios of IPCC. Refinements of these approximations are also possible.

We evaluate seprately the impact of maximum and that of minimum temperatures. A direction of improvement could be to use the bivariate density of the maximum and minimum temperatures and evaluate their joint impact which would take into account their possible correlation.

7 Appendix

This appendix provides details about the computation of the impacts and their variance. We derive the impact variance in a general scenario where the change direction curve may depend on province i and will use it for the change scenario given by the simplex-difference between the RCP2.6 histogram and the observed histogram in 2016 for that province. The computation is very similar in spirit for both the discrete and the smooth framework, however the evaluation of the inner products involved is more intricate in the smooth framework. We evaluate separately the impact of maximum temperature and that of minimum temperature, and then add them up to get the impact of climate change. We develop the computation for maximum temperature and the result for minimum temperature is obtained in the same fashion.

In the smooth framework, the impact estimate for maximum temperature between an initial time, say 0, and time s, say $s = 2099$, is given, for a province i in region r, by

$$\hat{Y}_{is} - \hat{Y}_{i0} =< clr(\varphi_i), clr(\hat{\beta}_r^{max}) > . \tag{14}$$

Because the RCP scenarios are available as histograms, we will assume that φ_i is a step function (constant on each bin with values $(\varphi_i)_j$ for bin j.). Since $clr(\hat{\beta}_r^{max})(t) = \sum_{l=1}^{g+3} z_l(\hat{\beta}_r^{max}) Z_l^4(t)$ the impact of maximum temperature for province i in region r is

then given by

$$\hat{Y}_{is} - \hat{Y}_{i0} = \sum_{l=1}^{g+3} z_l(\hat{\beta}_r^{max}) \int clr(\varphi_i)(t) Z_l^4(t) dt, \tag{15}$$

where $z(\hat{\beta}_r^{max})$ is the $g+3$ vector of components of the $\hat{\beta}_r^{max}$ curve in the ZB-spline basis and $Z_l^4(t)$ is the l^{th} ZB-spline curve. For $i = 1$ to 63 and $l = 1$ to $g+3$, let us denote by p_{il} the integral term

$$p_{il} = \int clr(\varphi_i)(t) Z_l^4(t) dt \tag{16}$$

and therefore $\hat{Y}_{is} - \hat{Y}_{i0} = \sum_{l=1}^{g+3} z_l(\hat{\beta}_r) p_{il}$.

To compute the p_{il}, we take advantage of the fact that φ_i are constant on the bins (b_j, b_{j+1}) and then of the fact that the integral of a ZB-spline can be obtained with differences of B-splines of a higher order using equation (7) in [9] as follows

$$p_{il} = \int_{12}^{40} clr(\varphi_i)(t) Z_l^4(t) dt = \sum_{j=1}^{28} \int_{b_j}^{b_{j+1}} clr(\varphi_i)_j(t) Z_l^4(t) dt \tag{17}$$

$$= \sum_{j=1}^{28} clr(\varphi_i)_j \int_{b_j}^{b_{j+1}} Z_l^4(t) dt = \sum_{j=1}^{28} clr(\varphi_i)_j \left(B_l^5(b_{j+1}) - B_l^5(b_j) \right) \tag{18}$$

Turning now attention to the variance of the estimated impact of maximum temperature, we have
The unbiasedness of the OLS estimates in clr space implies that $\mathbb{E}(z(\hat{\beta}_r^{max})) = z(\beta_r^{max})$. Therefore we have

$$\mathrm{Var}(\hat{Y}_{is} - \hat{Y}_{i0}) = \mathbb{E}\left[\left(\sum_{l=1}^{g+3} (z_l(\hat{\beta}_r^{max}) - z_l(\beta_r^{max})) p_{il} \right)^2 \right] \tag{19}$$

Let P be the $63 \times (g+3)$ matrix of elements p_{il}. Then the variance of the impact in province i is given by

$$\mathrm{Var}(\hat{Y}_{is} - \hat{Y}_{i0}) = \mathrm{Var} P_{i.} z(\hat{\beta}_r^{max}) = P_{i.} \mathrm{Var}(z(\hat{\beta}_r^{max})) P_{i.}^T, \tag{20}$$

where $P_{i.}$ is the i^{th} row of P. We can estimate $\mathrm{Var}(z(\hat{\beta}_r^{max}))$ by the empirical variance-covariance matrix of the parameters estimates.

We will use a single scenario by region. For region r, this scenario φ_r is computed as the average (in the simplex sense) of the RCP scenarios of all the provinces in that region.

Acknowledgement(s)

The authors gratefully acknowledge VIASM. Part of this work was completed while they were visiting the Vietnam Institute for Advanced Study in Mathematics (VIASM).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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).

References

[1] Filzmoser, P., Hron, K., Menafoglio, A.: Logratio approach to distributional modeling. In: Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan, pp. 451–470. Springer, ??? (2021)

[2] Carter, C., Cui, X., Ghanem, D., Mérel, P.: Identifying the economic impacts of climate change on agriculture. Annual Review of Resource Economics **10**(1), 361–380 (2018)

[3] Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)

[4] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: Modeling and Analysis of Compositional Data. John Wiley & Sons, Canada (2015)

[5] Hron, K., Filzmoser, P., Thompson, K.: Linear regression with compositional explanatory variables. Journal of Applied Statistics **39**(5), 1115–1128 (2012)

[6] Petersen, A., Zhang, C., Kokoszka, P.: Modeling probability density functions as data objects. Econometrics and Statistics **21**, 159–178 (2022)

[7] Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V.: Bayes Hilbert spaces. Australian & New Zealand Journal of Statistics **56**(2), 171–194 (2014)

[8] Talská, R., Hron, K., Grygar, T.M.: Compositional scalar-on-function regression with application to sediment particle size distributions. Mathematical Geosciences **53**(7), 1667–1695 (2021)

[9] Machalová, J., Talská, R., Hron, K., Gába, A.: Compositional splines for representation of density functions. Computational Statistics **36**(2), 1031–1064 (2021)

[10] Fisher, R.: The influence of rainfall on the yield of wheat in Rothamsted. Philosophical Transactions of the Royal Society of London **213**, 89–142 (1924)

[11] Davis, K.F., Downs, S., Gephart, J.A.: Towards food supply chain resilience to environmental shocks. Nature Food **2**, 54–65 (2021)

[12] Espagne, E., Ngo-Duc, T., Nguyen, M.-H., Pannier, E., Woillez, M.-N., Drogoul, T.P.L.. A.and Huynh, Le, T.T., Nguyen, T.T.H., Nguyen, T.T., Nguyen, T.A., Thomas, F., Truong, C.Q., Vo, Q.T., Vu, C.T.: 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 (2021)

[13] 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.: Estimating economic damage from climate change in the United States. Science **356**(6345), 1362–1369 (2017)

[14] Schlenker, W., Roberts, M.J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences **106**(37), 15594–15598 (2009)

[15] Ortiz-Bobea, A.: Chapter 76 - the empirical analysis of climate change impacts and adaptation in agriculture **5**, 3981–4073 (2021)

[16] Deryugina, T., Hsiang, S.: The marginal product of climate. Working Paper 24072, National Bureau of Economic Research (2017)

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

[18] Handcock, M.S., Morris, M.: Relative distribution methods. Sociological Methodology **28**(1), 53–97 (1998)

[19] Holmgren, E.B.: The pp plot as a method for comparing treatment effects. Journal of the American Statistical Association **90**(429), 360–365 (1995)

[20] De Boor, C.: A Practical Guide to Splines. Springer, New York (1978)

[21] Machalova, J., Hron, K., Monti, G.S.: Preprocessing of centred logratio transformed density functions using smoothing splines. Journal of Applied Statistics **43**(8), 1419–1435 (2016)

[22] Machalová, J.: Optimal interpolatory splines using b-spline representation. Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica **41**(1), 105–118 (2002)

[23] Burgess, M.G., Ritchie, J., Shapland, J., Pielke, R.: Ipcc baseline scenarios

have over-projected co2 emissions and economic growth. Environmental Research Letters **16**(1), 014016 (2020)

[24] Tran-Anh, Q., Ngo-Duc, T., Espagne, E., Trinh-Tuan, L.: A high-resolution projected climate dataset for Vietnam: Construction and preliminary application in assessing future change. Journal of Water and Climate Change **13**(9), 3379–3399 (2022) https://doi.org/10.2166/wcc.2022.144 https://iwaponline.com/jwcc/article-pdf/13/9/3379/1114662/jwc0133379.pdf

[25] Coenders, G., Pawlowsky-Glahn, V.: On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT–Statistics and Operations Research Transactions **44**(1), 201–220 (2020)

[26] Menafoglio, A.: BayesSpaces-codes. GitHub (2021)

[27] Boogaart, K., Filzmoser, P., Hron, K., Templ, M., Tolosana-Delgado, R.: Classical and robust regression analysis with compositional data. Mathematical Geosciences **53**(5), 823–858 (2021)

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

[29] Scenarios, E.: Ipcc special report. Cambridge Univ, Cambridge (2000)