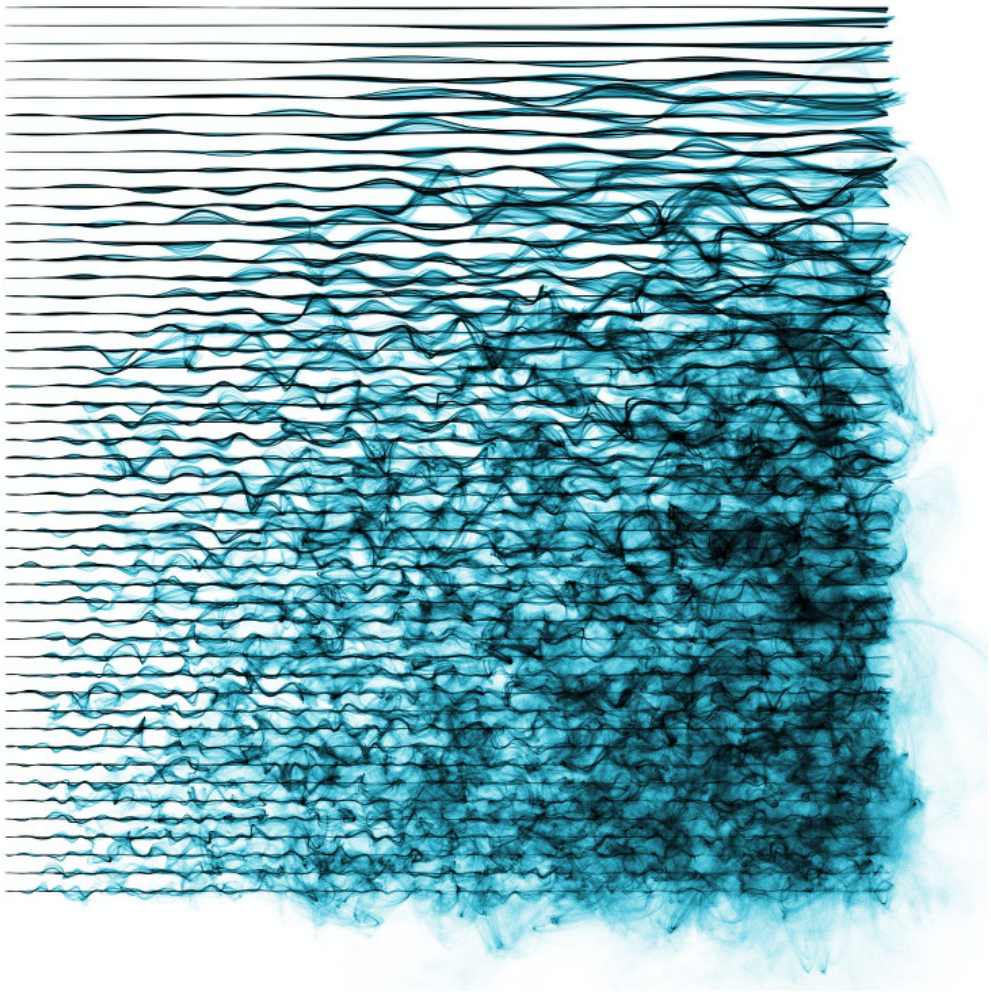
Functional Data Analysis of the Agricultural Production and Temperature Changes Across Selected European Countries

Hamed Davoodi

11267A

Functional and Topological data analysis

Alessandra Micheletti



April 2024

1- Introduction

Climate change is a pressing global issue, and its effects on agriculture are far-reaching. In this comprehensive project, we investigate how temperature variations influence crop yields across the European Union (EU) countries. Our goal is to understand the functional relationship between temperature changes and agricultural production.

Temperature change plays a pivotal role in shaping agricultural productivity, particularly through its direct impact on crop yields. As global temperatures continue to rise, understanding the intricate relationship between temperature fluctuations and crop production becomes increasingly critical. There are various approaches to address to the theoretical relationship between temperature changes and agricultural productions.

1-1- Temperature Optima for Crop Growth

Each crop species exhibits specific temperature requirements for optimal growth and development. While moderate temperature increases can enhance photosynthesis and accelerate plant metabolism, extreme heat stress can inhibit physiological processes, leading to reduced yields. Understanding the temperature thresholds at which different crops thrive is essential for predicting how shifting climate patterns may influence their productivity.

1-2- Thermal Time Requirements

Crop phenology, including stages such as germination, flowering, and maturity, is intricately linked to temperature conditions. Warmer temperatures can hasten crop development, shortening the time from planting to harvest. However, if temperatures exceed the crop's optimal range during critical growth stages, it can disrupt normal phenological progression, affecting yield quantity and quality. Conversely, cooler temperatures may prolong the growing season, providing extended opportunities for crop development in certain regions.

1-3- Heat Stress and Yield Losses

Prolonged exposure to high temperatures can induce heat stress in crops, leading to physiological damage and yield losses. Heat stress during flowering and grain filling stages is particularly detrimental for cereal crops such as maize, wheat, and rice, as it can impair pollination, reduce grain set, and decrease kernel weight. Additionally, extreme heat events can cause irreversible damage to crop tissues, exacerbating yield variability and economic losses for farmers.

1-4- Geographic Variability and Adaptation

The impact of temperature change on crop yields varies geographically, influenced by factors such as latitude, altitude, and local microclimates. While some regions may experience yield benefits from moderate temperature increases, others may face heightened risks of heat stress and yield reductions. Agricultural adaptation strategies, including breeding heat-tolerant crop varieties, adjusting planting dates, and implementing irrigation and shading techniques, are crucial for mitigating the adverse effects of temperature change and maintaining crop productivity in diverse agroecological contexts.

2- Methodology

In this study, we are utilizing functional data analysis (FDA) to explore the relationship between scalar dependent variables and functional independent variables. Our methodology comprises several essential steps, each designed to address the challenges and complexities inherent in functional data analysis. Additionally, we are employing a specific R library called "fda," authored by James Ramsay et al. and published in 2024.

2-1- Preprocessing Data

Prior to analysis, it is crucial to preprocess functional data to ensure data quality and consistency. This preprocessing involves several steps, including data cleaning to remove outliers and errors, filtering to reduce noise, and alignment to ensure temporal or spatial consistency across observations.

In order to conduct analysis using the "fda" library in R, it is necessary to convert raw data into a functional data object. One approach to achieve this is by initially preprocessing the data in a user-friendly manner using tools like Microsoft Excel. The preprocessed data can then be saved in CSV file format for further manipulation. Subsequently, the CSV file can be imported into R Studio and converted into an object recognizable by the "fda" library, utilizing the "Data2fd" function. This conversion process is foundational, as all subsequent steps of analysis depend on these functional data objects.

2-2- Basis Functions System

In functional data analysis (FDA), basis function systems are pivotal for constructing functions that represent continuous data curves. James Ramsay's 2005 work comprehensively reviewed these systems, shedding light on their significance and application. Based on his work, we can use basis function as a set of functional building blocks as ϕ_k with $k = 1, \dots, K$. Then $x(t)$ can be expressed as:

$$x(t) = \Sigma c_k \phi_k(t) = c'\phi(t)$$

This is the basis function expansion, and the parameter c is the coefficient of the expansion. Two commonly employed basis function systems in FDA are B-splines and Fourier basis functions.

B-splines: These are piecewise polynomial functions offering local control over curve shape. They are defined recursively using the Cox-de Boor recursion formula.

$$\phi_{k,1}(t) = \begin{cases} 1, & if\ t_k \le t < t_{k+1} \\ 0, & otherwise \end{cases}$$

$$\phi_{k,i}(t) = \frac{t - t_k}{t_{k+i-1} - t_k}\phi_{k,i-1}(t) + \frac{t_{k+i} - t}{(t_{k+i} - t_{k+1})}\phi_{k+1,i-1}(t)$$

With knots denoted by $t_1, \dots, t_{n+i}$. This recursion formula defines each B-spline basis function in terms of its lower-order counterparts.

Fourier basis functions: Derived from trigonometric sine and cosine functions, Fourier basis functions are ideal for analyzing periodic data or data exhibiting cyclical patterns.

$$\phi_1(t) = 1$$
$$\phi_2(t) = \sin(\omega t)$$
$$\phi_3(t) = \cos(\omega t)$$
$$\phi_4(t) = \sin(2\omega t)$$
$$\phi_5(t) = \cos(2\omega t)$$
$$\dots$$

Where the constant $\omega = \frac{2\pi}{T}$ for the period T.

2-3- Smoothing

Smoothing techniques are applied to the functional data to reduce noise and variability while preserving important features. This step is crucial for enhancing the signal-to-noise ratio and improving the interpretability of the data. In other words, smoothing a dataset involves creating an approximate function aimed at capturing key patterns while filtering out noise and intricate structures. During smoothing, individual data points within a signal are adjusted to reduce discrepancies between neighboring points, typically caused by noise, resulting in a more uniform signal. Smoothing serves two crucial purposes in data analysis: first, it facilitates the extraction of pertinent information from the data when the assumption of smoothing is reasonable, and second, it enables the provision of analyses that are both adaptable and resilient. Various algorithms are employed in smoothing to achieve these objectives.

Differentiating smoothing from the concept of curve fitting reveals distinct characteristics. While curve fitting typically entails fitting an explicit function to the data, smoothing yields immediate results in the form of smoothed values without necessarily adhering to a specific functional form. Smoothing aims to offer a broad depiction of gradual changes in value,

with less emphasis on precise data point matching, contrasting with the meticulous pursuit of an optimal fit pursued in curve fitting. Moreover, smoothing methods often incorporate a tuning parameter to regulate the degree of smoothing, whereas curve fitting adjusts multiple parameters to achieve the best fit.

The general idea can be summarized as:

$$\hat{c} = argmin_c\{[\Sigma y_i - x(t_j)]^2 + \lambda \int [Lx(t)]^2 dt$$

Since we have $x(t) = c'\phi(t) = \phi'(t)c$, then we can write:

$$\hat{c} = argmin_c\{[\Sigma y_i - \phi'(t_j)c]^2 + \lambda c' \left[\int L\phi(t) L\phi'(t) dt\right] c$$

Here λ is the penalty parameter and $\lambda c'[\int L\phi(t)L\phi'(t)dt]c$ is the penalty term.

The generalized cross-validation (GCV) is a criterion commonly used for selecting the smoothing parameter λ It is calculated as follows:

$$GCV(\lambda) = (\frac{n}{n-df(\lambda)})(\frac{SSE}{n-df(\lambda)})$$

Also, we can use Hat matrix notation to get the same results, when a hat matrix is $H = \phi(\phi'\phi + \lambda R)^{-1}\phi'$ and the degree of freedom can be represented by $df(\lambda) = trace[H(\lambda)]$. Hence:

$$GCV(\lambda) = \frac{1}{n}\Sigma\left(\frac{y_i - \hat{y}_i}{1 - \frac{trace(H(\lambda))}{n}}\right)^2$$

2-4- Functional Regression:

Functional regression models are utilized to explore the relationship between functional predictors and scalar response variables. These models extend traditional regression frameworks to accommodate functional data, allowing for the estimation of functional coefficients that represent the effect of the predictors on the response variable. Functional

regression techniques account for the inherent variability and uncertainty in the functional predictors, providing robust estimates of the relationship between variables.

One of the simple but most important types of the Functional Regressions is scalar to function type meaning that the response or famous y_i is scalar and x_i is functional. We can write the regression model as:

$$y_i = \alpha_0 + \int x_i(t_j)\beta_j + \epsilon_i$$

In our case, y_i is the average annual agricultural production of country i, and $x_i(t_j)$ is the smoothed functional form of annual temperature changes in Celsius of the country i.

2-5- Functional PCA:

Principal Component Analysis (PCA) is often our initial step following descriptive statistics and plots when exploring data. The aim is to identify primary modes of variation within the dataset and assess their significance. Similar to multivariate statistics, the eigenvalues derived from the variance-covariance function represent the importance of these principal components.

In Functional PCA, each eigenvalue corresponds to an eigenfunction rather than an eigenvector, describing significant variational components. Rotating these functions can yield a more interpretable depiction of the dominant variation modes in the functional data while preserving the total common variation.

The main equation is the expression for the probe score variance associated with a probe weight ξ. It is defined as the maximum value of the sum of squared probe scores subject to the constraint that the squared integral of the weight function ξ=1 over its domain. Mathematically, it is expressed as:

$$\mu = \max\left\{\sum \rho_{\xi}^2(x_i)\right\} \text{ subject to } \int \xi^2(t)dt = 1$$

Additionally, the construction of eigenvalue/eigenfunction pairs through eigen-analysis, as well as the determination of the optimal number of harmonics to use in PCA through visual

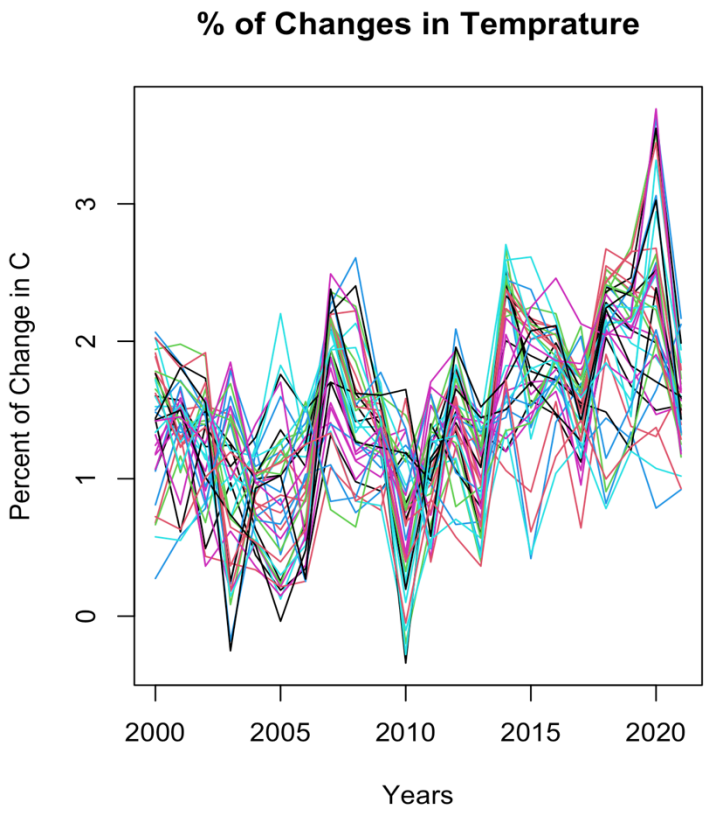
inspection of eigenvalue plots, referred to as scree plots. It also touches upon the interpretability of PCA results and the potential for alternative optimal basis systems.

2-6- Data Collection and Preprocessing

Our data covers the years from 2000 to 2021 and includes information from 38 EU countries. We carefully collected historical temperature data for each country and matched it with the average agricultural production. To make sure everything is consistent, we made the data uniform and fixed any missing information.

3- Results

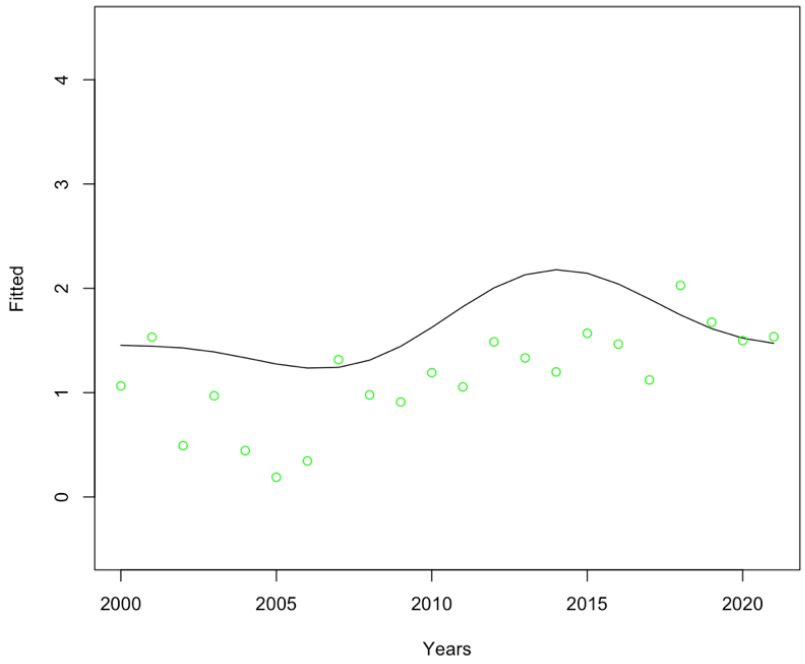
The results of our comprehensive analysis offer valuable insights into the dynamic relationship between temperature fluctuations and agricultural outcomes across the European Union (EU) countries. Our analysis identifies country-specific responses to temperature changes, highlighting the diverse agricultural landscapes within the EU. These insights lay the groundwork for targeted policy interventions and adaptation strategies aimed at bolstering the resilience of agricultural systems in the face of climate change. We can visualize the original trends of the temperature percentage change across the chosen EU countries by plotting them.



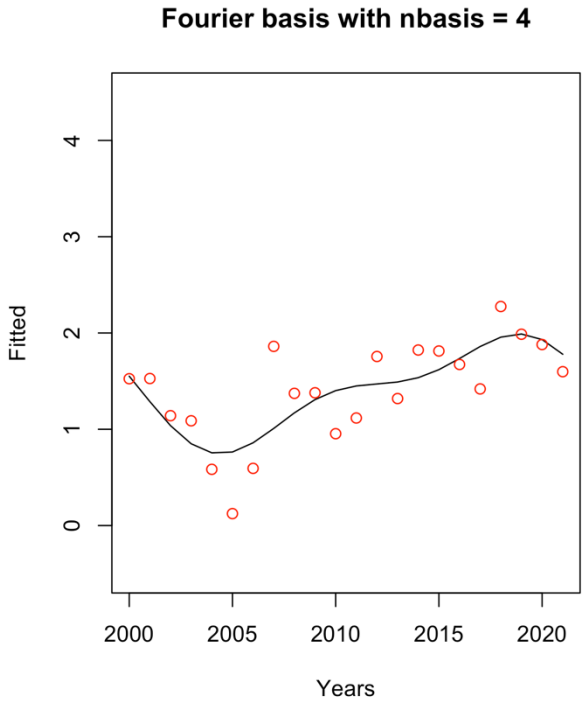
Each colored line represents the temperature percentage change trend for a specific country from 2000 to 2021. Our analysis is grounded in functional data analysis methods, yielding the following results:

3-1- Smoothing functional data

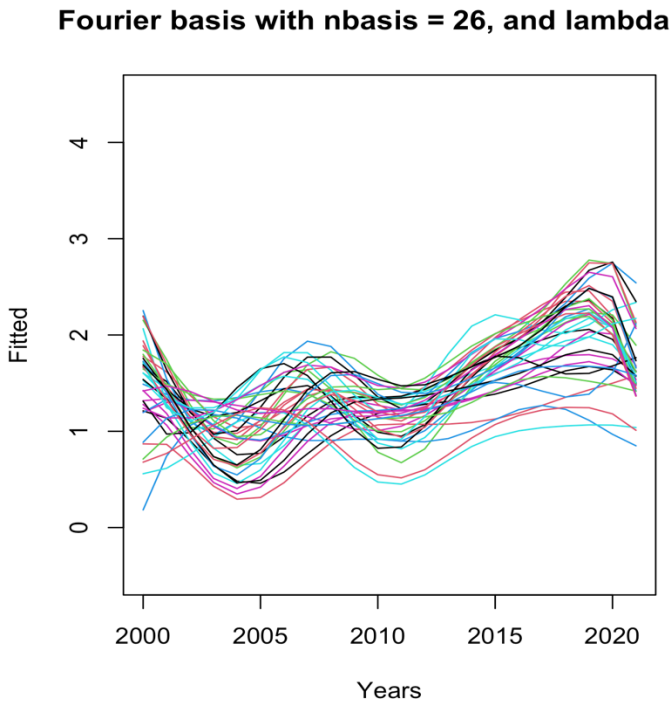
To smoothen the sharp lines and illustrate the trends more effectively, we have two main approaches: B-splines basis and Fourier basis. We can use a B-spline smoothing curve with an order of 4 for the country with an index of 5.



Also, with the Fourier basis function we can do the same with the order of 4:

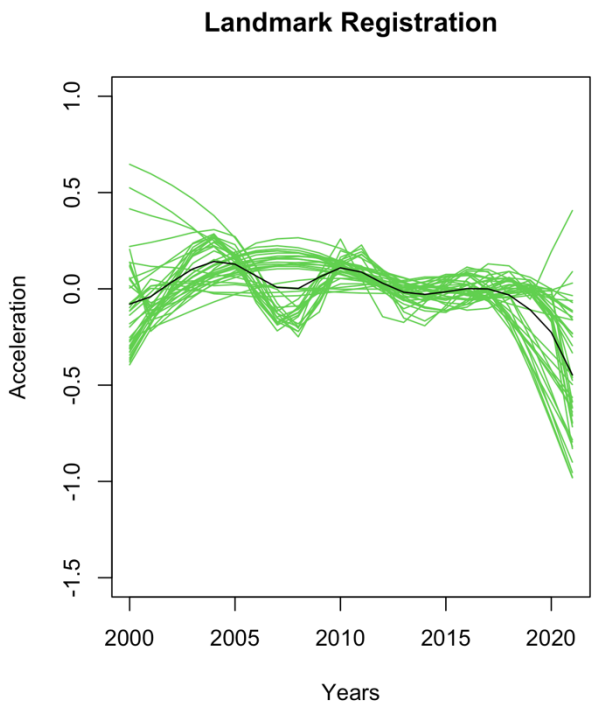


For better illustration, we can experiment with different orders and observe how the number of orders (basis) affects the curvature of the trends. Additionally, we can introduce a penalty term and apply the smoothing function to the trends of all countries.

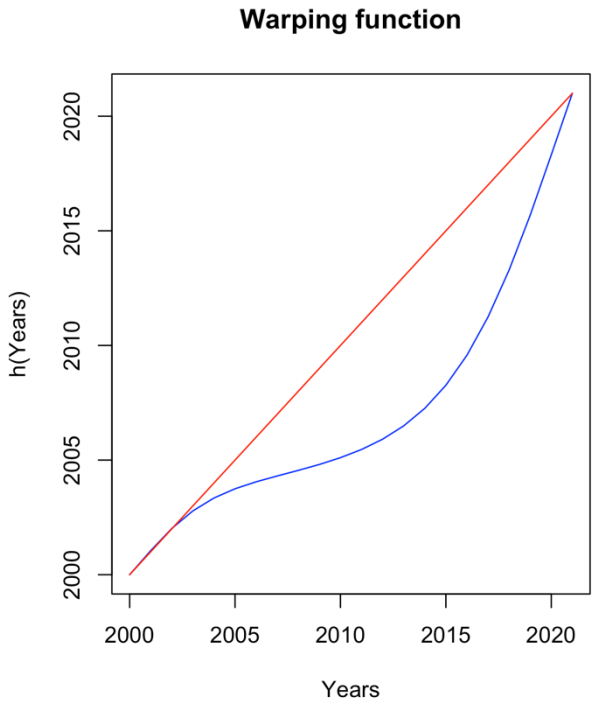


3-2- Feature (Landmark) registration and Depth analysis

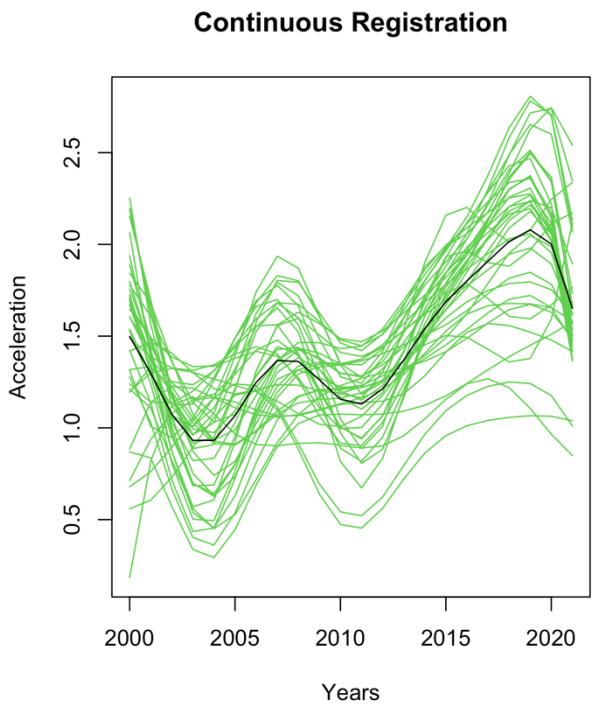
In FDA (Functional Data Analysis), landmark registration is a technique used to align functional data by identifying key features or landmarks within each curve and then registering them based on these landmarks. The curves that we have illustrated above have two specific features: Amplitude variation and Phase variation. The get general idea about these variations we should employ a specific method to capture both at the same time. However, averaging will not work. Hence, we need to align these curves based on the phase and amplitude. We have two options for this: the least squares criterion for shift alignment and landmark or feature registration. In this case, we use landmark registration. A landmark or a feature of a curve is some characteristics that one can associate with a specific argument of value T. These are typically maxima, minima, or zero crossings of curves.



When performing landmark registration in FDA and observing a change in the y-axis to acceleration, it suggests that the registration process might involve a transformation or normalization step that has converted the original data into an acceleration-like metric. After registration, we compare the aligned curves based on their acceleration profiles. This allows for a more standardized comparison of the dynamics or changes within the functional data across different curves or groups. Based on the figure above, which was generated using the minimum and maximum values, we observe fluctuations around the zero line. As time progresses, towards the end of the period, we notice a decrease in acceleration to approximately -0.5.

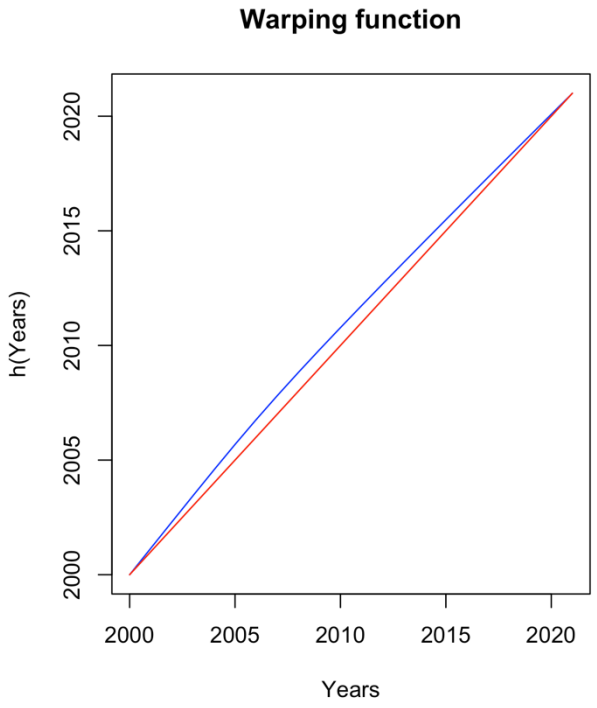


In FDA, warping functions are used to capture the transformations applied to align functional data. These functions represent how each curve is stretched or compressed in time to achieve alignment with a reference curve or a common template. Warping functions provide a mathematical framework to quantify and visualize the transformations applied during registration. In our analysis, the warping function reveals that our method adjusted each curve through compression, as indicated by values below the red line, which represents the standard benchmark of 1. Any value greater than the red line indicates stretching of the curve.



Continuous registration extends the concept of landmark registration by allowing for the registration of all points along the functional curves, rather than just specific landmarks. This technique involves estimating a smooth warping function that continuously maps points from one curve to another, optimizing alignment throughout the entire curve. Provides a more flexible and precise alignment of functional data by considering all points along the curves and enables the detection and alignment of subtle variations or features that may not be captured by landmark-based methods.

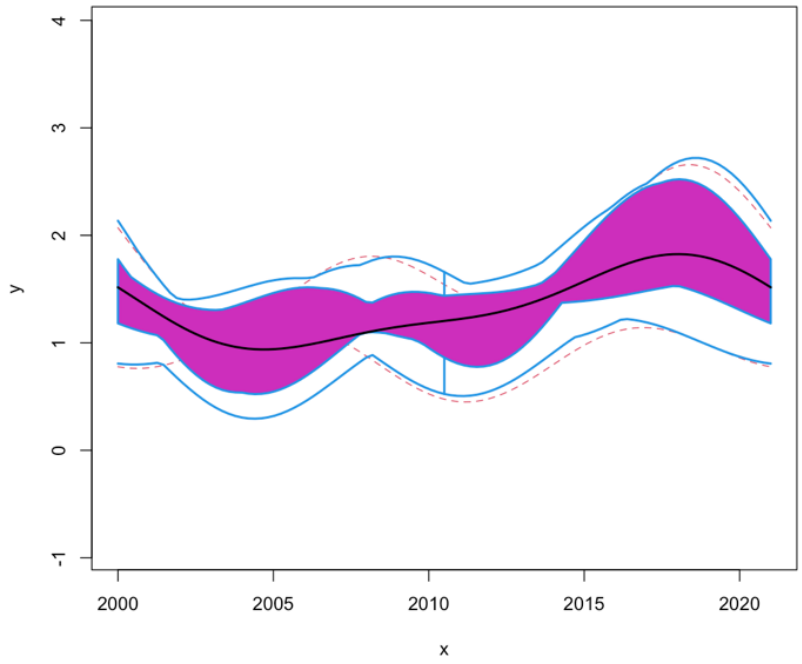
Based on our calculations, we continue to observe fluctuations in acceleration above the zero-line, indicating that in the later years, acceleration began to decline after peaking around 2017. The warping function of this continuous registration illustrates a slight stretching of the curves.



In Functional Data Analysis (FDA), depth analysis is a method used to measure centrality and outlyingness of functional data.

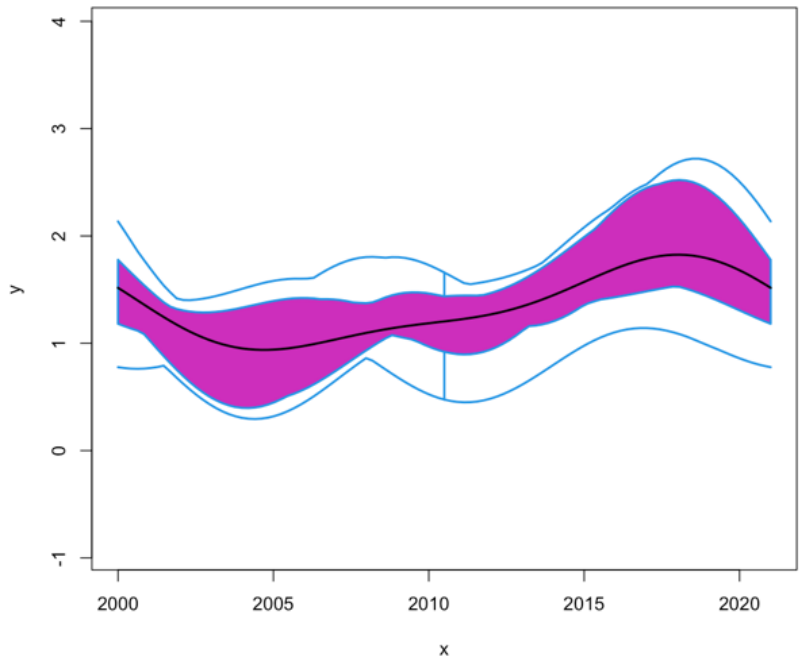
Modified Band Depth (MBD)

MBD is a depth measure that quantifies how central a functional curve is relative to a set of curves. It is a modification of the Band Depth (BD) measure, designed to handle irregularities in functional data. MBD calculates the proportion of curves that enclose a particular curve within a certain band defined by the maximum and minimum values of the set. Higher MBD values indicate greater centrality, while lower values suggest outlyingness. Below is the Mean Absolute Bending (MBD) figure depicting the percentage changes in temperature of selected EU countries, with a peak occurring in 2017 and a median fluctuation of around 1.5 percent.



Band Depth (BD)

BD is a depth measure that assesses the centrality of a functional curve within a dataset. It quantifies the proportion of curves that encompass a particular curve within their convex hull. BD provides a robust measure of centrality, particularly suitable for analyzing functional data with smooth and well-behaved curves. The shape remains consistent across both the Mean Absolute Bending (MBD) and Bending (BD) figures. Additionally, when we attempted a mixed approach combining MBD and BD, the results mirrored those of the BD figure.



The median curve is consistent across all three methods and closely resembles the trend observed in Italy.

The Wilcoxon rank-sum test, also known as the Mann-Whitney U test, is a non-parametric statistical hypothesis test used to determine if two independent samples come from populations with the same distribution. It is particularly useful when the assumptions of the t-test, such as normality and equal variances, are not met. Instead of comparing means, the Wilcoxon rank-sum test compares the medians of the two samples by ranking all the observations together and assessing whether one group tends to have consistently higher or lower values than the other. It's widely applied across various fields, especially when analyzing data that may not follow a normal distribution or when dealing with ordinal or ranked data.

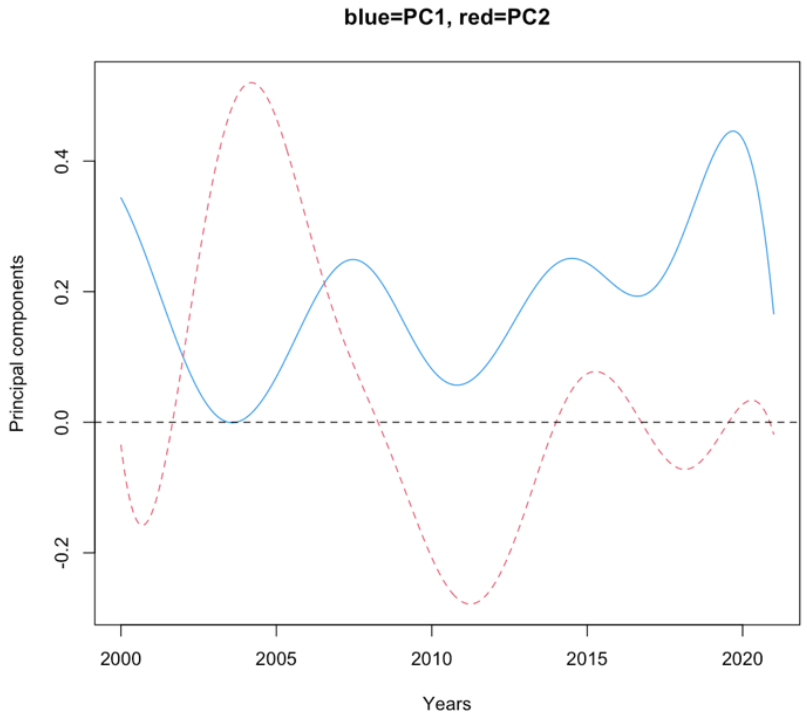
Table 1Wilcoxon rank-sum test

Region	W (Sum of the ranks)	P-Value	result
Southern EU	108	0.6533	No Difference
Western EU	92	0.8914	No Difference
Eastern EU	38	0.1667	So So
Northern EU	54	0.0168	Significant difference
Central EU	190	0.0107	Significant difference

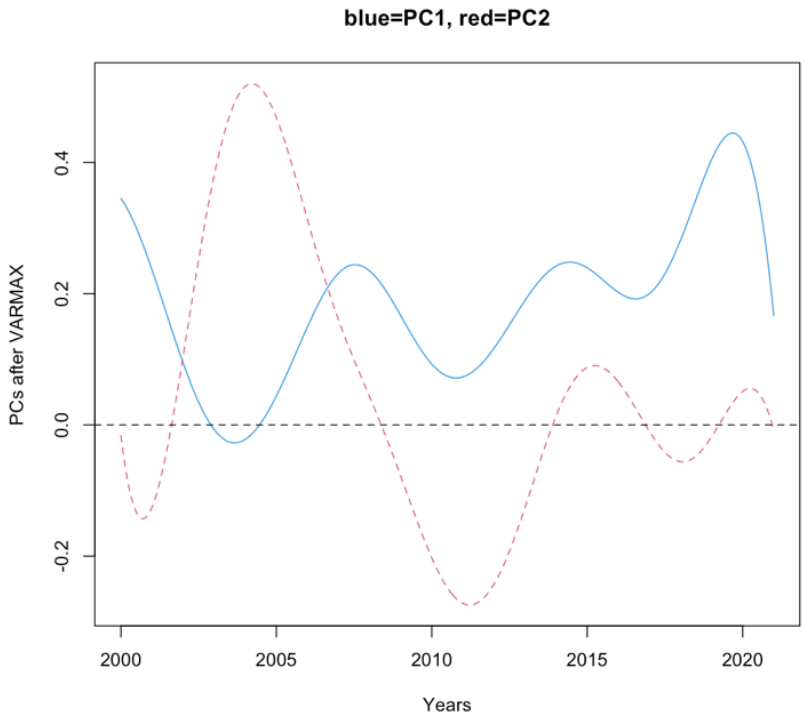
- Alternative Hypothesis: True location shift is not equal to 0

3-3- Functional PCA analysis

Functional Principal Component Analysis (FPCA) is a powerful technique used to analyze and extract meaningful information from functional data. Principal components represent the main modes of variability or patterns present in the functional data. The shape and trends of the curves along the principal component axes provide insights into how different temporal patterns contribute to the overall variability in the dataset.



The VARIMAX method is a technique commonly used for rotating the principal component functions to achieve a simpler and more interpretable representation of the data. VARIMAX rotation is a mathematical procedure applied to the PCFs to maximize their variance and make them as orthogonal (uncorrelated) and interpretable as possible. The rotation aims to simplify the interpretation of the principal components by aligning them in directions that emphasize specific patterns or features in the data.

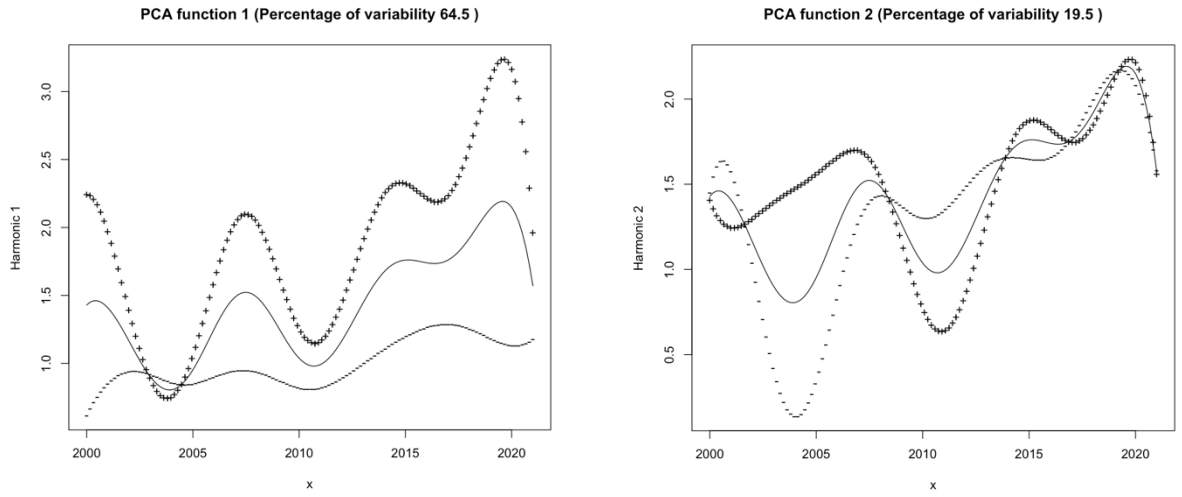


In both cases of Functional Principal Component Analysis (FPCA), PC1 and PC2 have effectively captured the majority of variation. Particularly noteworthy is the time interval from 2003 to 2007, where PC2 (the red line) in both cases accounts for the unexplained variation of PC1 (the blue line).

In addition, the detailed results of the Functional Principal Component Analysis (FPCA) reveal that the first principal component (PC1) explains the majority of the variability in the functional data, accounting for 64.5% of the total variation. This dominance suggests that PC1 captures the most prevalent and significant temporal pattern or trend present in

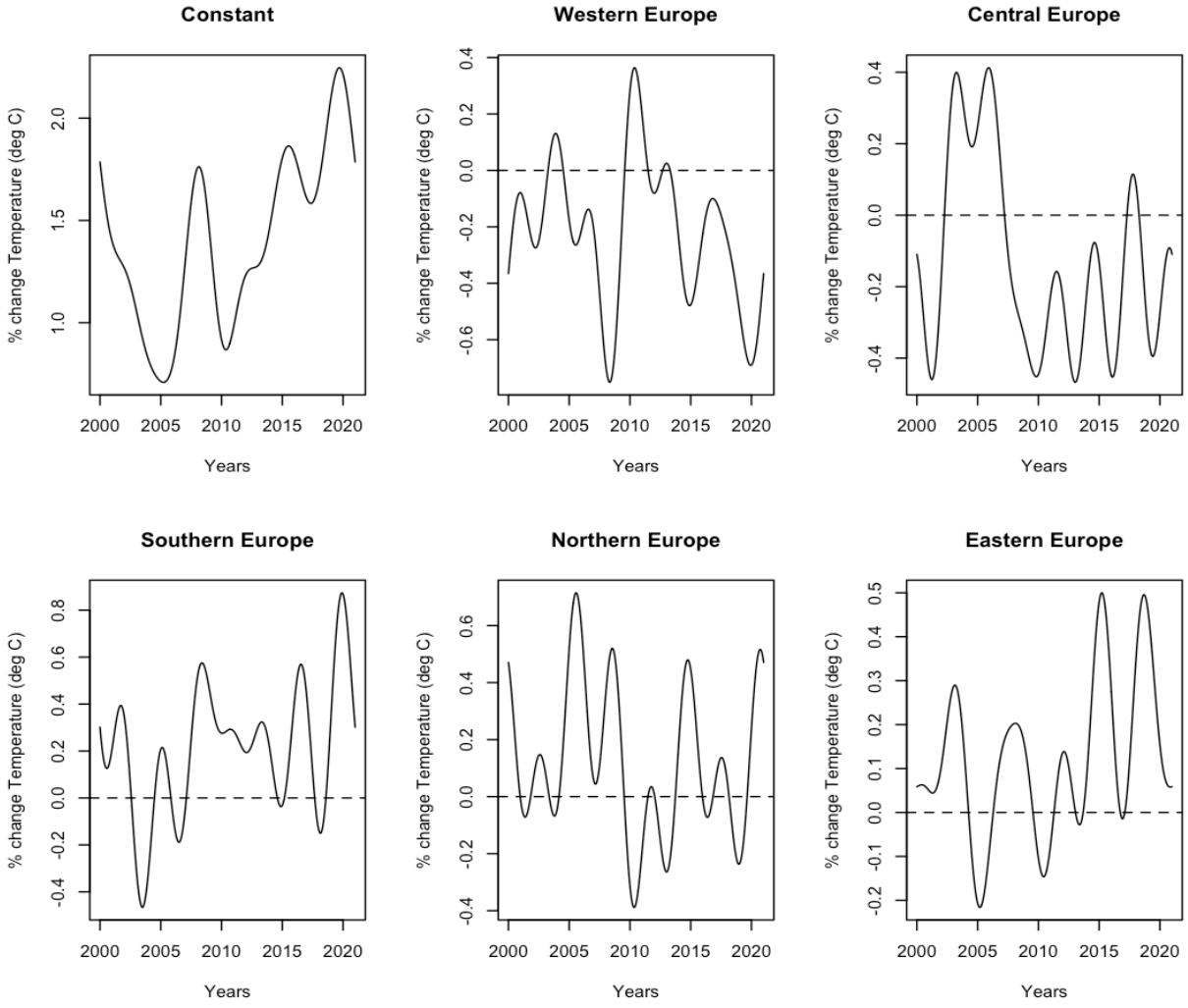
the dataset. The high proportion of variability explained by PC1 indicates its crucial role in characterizing the overall structure and dynamics of the functional data.

On the other hand, the second principal component (PC2) captures an additional 19.5% of the total variability in the data. While PC2 explains less variability compared to PC1, it still represents a significant mode of variation beyond what is captured by PC1 alone. PC2 likely captures secondary patterns or deviations from the primary trend represented by PC1.

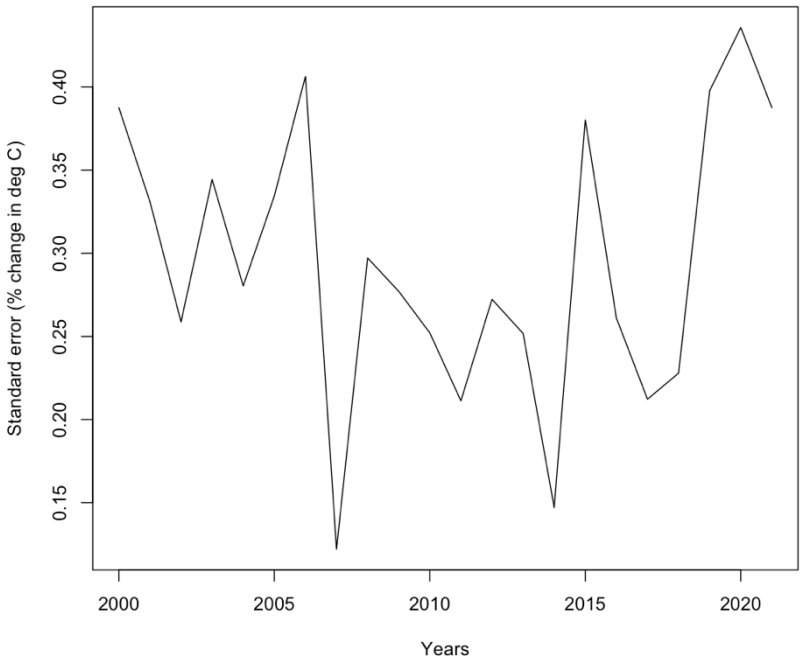


3-4- Functional ANOVA

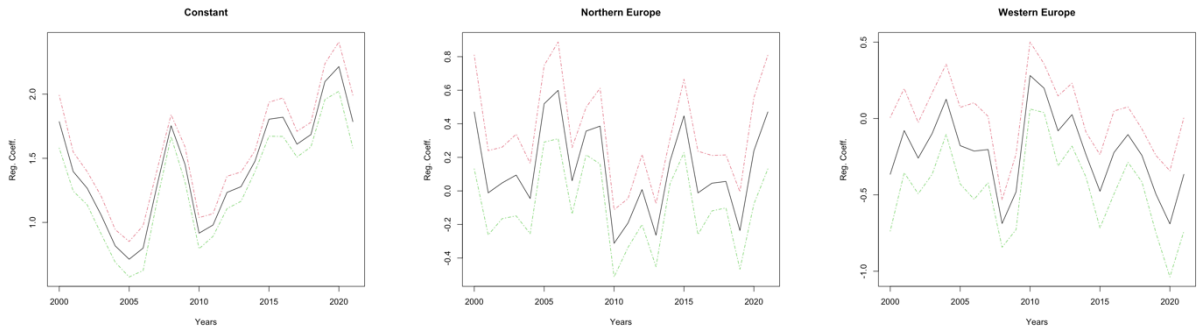
Functional ANOVA involves partitioning the variability in functional data into different sources, such as regions, and examining the contributions of each source to the overall variability. Based on the calculation, each region represents a distinct segment of the functional data where the temperature exhibits specific behaviors or patterns. This provides explanation about the magnitude of temperature variation within each region compared to the overall dataset.

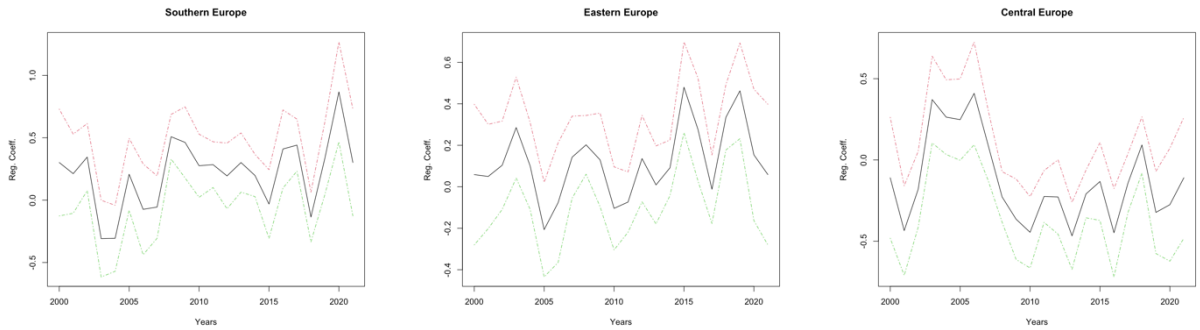


It is evident that Western and Central Europe show an overall trend below the zero-line, indicating a decrease in temperature change. In contrast, Southern, Northern, and Eastern Europe exhibit percentage change variations in temperature higher than the zero-line, suggesting an increase in temperature change.

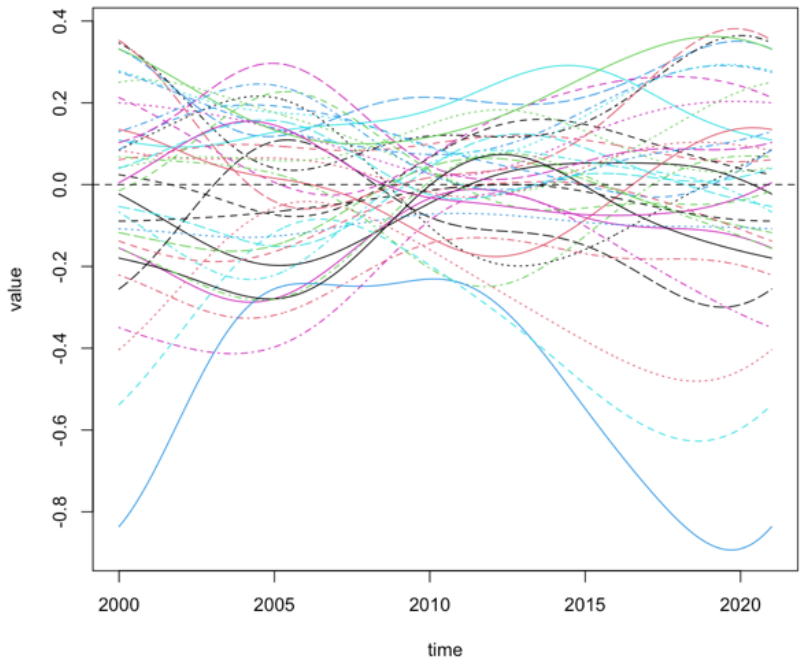


In functional ANOVA, the standard error measures the variability or uncertainty associated with the estimated effects of each region on the temperature. The figure above depicts the Standard Error of our estimation, peaking in 2020 with a minimum below 0.15. This measurement helps quantify the precision of our estimation and provides confidence intervals around the mean estimates of temperature change for all regions. Additionally, we can extract the regression coefficients according to standard errors for each region. Below is the illustration of this concept for each region. It is apparent that the regression coefficients for Western and Central Europe are mostly negative throughout the study period, whereas for other regions, the regression coefficient is mostly positive.



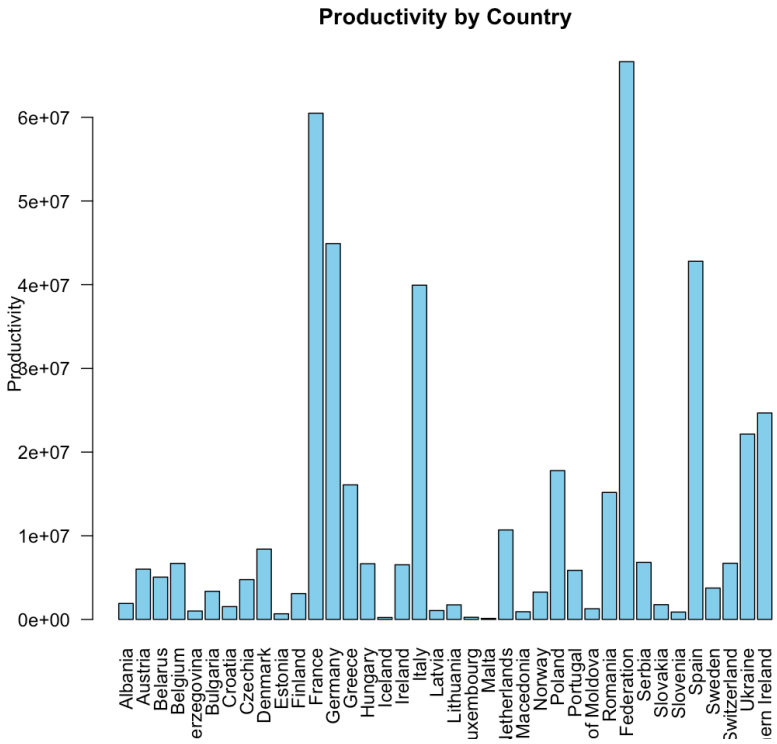


Residual values indicate the differences between the observed temperature values and those predicted by the functional ANOVA model. These residuals account for the variability in temperature that cannot be explained by the effects of the identified regions. We visualize the residual values of our estimation.

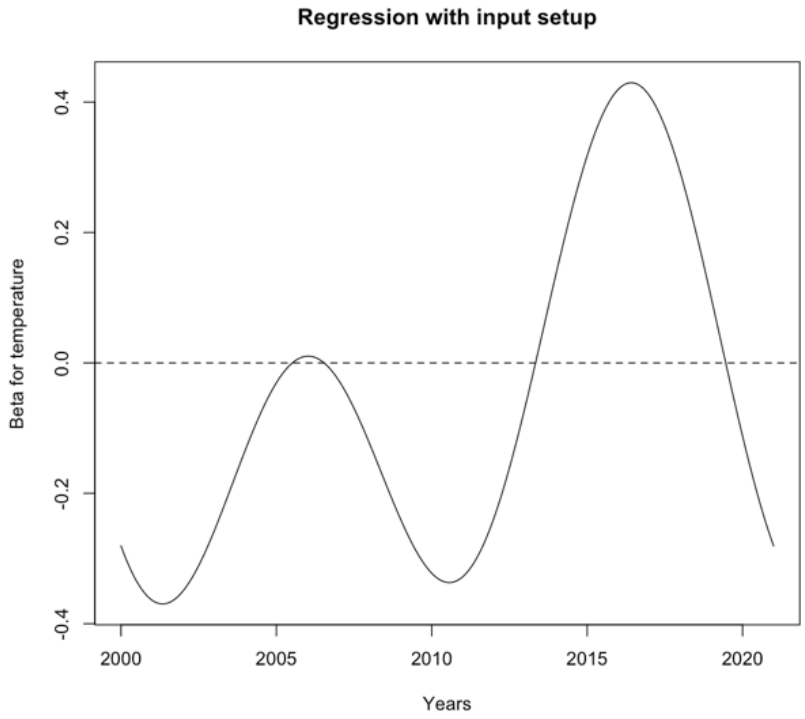


3-5- Functional Regression analysis

Functional regression involves modeling the relationship between a response variable (y) and one or more predictor variables (x) that are functional data. In our study, the y variable is the average agricultural production for each country in EU.



In functional regression, beta values represent the regression coefficients that quantify the relationship between the functional predictor(s) and the response variable. Each beta value corresponds to a specific functional predictor and indicates the magnitude and direction of its effect on the response variable. The beta values are estimated through the regression analysis and provide insights into how changes in the functional predictor(s) influence the response variable.

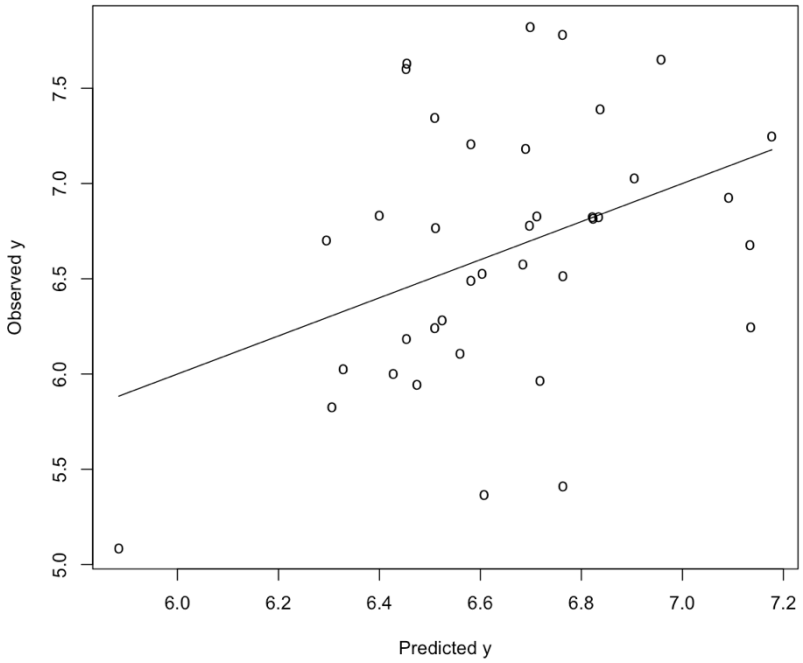


Based on the results of functional regression, we observe fluctuations in the beta values over time. From 2001 to 2014, there is a negative relationship between the percentage change in temperature and average agricultural production. However, the intensity of this relationship varies during this period, reaching peaks in 2001 and 2011. When the beta line crosses the zero-line, it indicates no relationship between the variables in our regression. Starting from 2014, the beta value becomes positive, indicating a positive relationship, reaching its maximum in 2017 at around 0.4 before declining and crossing the zero-line in 2019.

The plot below displays observed y-values as points or a line plot, overlaid with predicted y-values separately. Visual inspection of the plot allows for assessing how well the regression model captures the relationship between the functional predictors and the response variable. In our case, the line does not fit accurately based on the observed and predicted y-values, possibly due to the study period or the number of cases evaluated.

Generally, having more data improves the R^2 value. In our calculation, the R^2 is approximately 15.11% with an F-ratio of 2.13. The P-value for the test with H_0: all

coefficients are identically equal to zero, is 0.090, indicating that we can reject the null hypothesis at a significance level of 10%.



4- Conclusion

In conclusion, our analysis using functional regression revealed intriguing insights into the relationship between the percentage change in temperature and average agricultural production across the studied period. We observed a dynamic pattern in the beta values, signifying fluctuations in the relationship between these variables over time. Specifically, from 2001 to 2014, a negative relationship was predominant, with varying intensities highlighted by peaks in 2001 and 2011. However, post-2014, the relationship shifted to positive, reaching its peak in 2017 before declining.

The plotted data, comparing observed and predicted y-values, indicated a less precise fit of the regression model, potentially due to limitations in the study period or the quantity of cases evaluated. Nevertheless, our analysis yielded an R^2 value of approximately 15.11%, with an F-ratio of 2.13. The obtained P-value of 0.090 suggests statistical significance, allowing us to reject the null hypothesis, the coefficients of the regression model are identically equal to zero, at a significance level of 10%.

Upon reviewing the results from the Wilcoxon rank-sum test, we observe varying degrees of significance in regional differences regarding the relationship between temperature change and agricultural production. Notably, both the Central and Northern EU regions demonstrate statistically significant differences, suggesting distinct patterns in these areas compared to others. Conversely, the Southern and Western EU regions exhibit no significant differences, implying relatively consistent trends in temperature change and agricultural production across these regions. The Eastern EU region falls into an intermediate category, with a "So So" result indicating a less definitive distinction compared to the significant differences observed in the Central and Northern EU regions. These findings highlight the importance of considering regional nuances when analyzing the relationship between temperature change and agricultural production.

Overall, our findings underscore the complex dynamics between temperature change and agricultural production, emphasizing the importance of continued research to understand and address these relationships effectively.

5- References:

1. EU Science Hub. "Climate change and agriculture."
2. European Environment Agency. "Looking into the future of arable agriculture in Europe."
3. European Parliament. "The impact of extreme climate events on agricultural production in the EU."
4. FAO. (Food and Agriculture Organization of the United Nations). FAOSTAT database. Retrieved from http://www.fao.org/faostat/en/
5. Joint Research Centre. "Impacts of climate change in agriculture in Europe."
6. Ramsay, J. O., & Silverman, B. W. (2002). Applied Functional Data Analysis: Methods of Applied Functional Data Analysis. Springer.
7. Ramsay, J. O., & Silverman, B. W. (2005). Functional Data Analysis. Springer Series in Statistics.
8. Ramsay, J. O., & Silverman, B. W. (2006). Applied Functional Data Analysis: Methods and Case Studies. Springer Series in Statistics.
9. Ramsay, J. O., Hooker, G., & Graves, S. (2009). Functional Data Analysis with R and MATLAB. Springer.
10. 10. Springer. "Climate Change and Variability Impacts on Agricultural Production and Sustainability."