mathematics

MDPI

Article

Wildfires Vegetation Recovery through Satellite Remote Sensing and Functional Data Analysis

Feliu Serra-Burriel [1,2], Pedro Delicado [2,3,*] and Fernando M. Cucchietti [1]

[1] Barcelona Supercomputing Center, 08034 Barcelona, Spain; feliu.serra@bsc.es (F.S.-B.); fernando.cucchietti@bsc.es (F.M.C.)
[2] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
[3] Institut de Matemàtiques de la UPC—BarcelonaTech (IMTech), 08028 Barcelona, Spain
* Correspondence: pedro.delicado@upc.edu

Abstract: In recent years, wildfires have caused havoc across the world, which are especially aggravated in certain regions due to climate change. Remote sensing has become a powerful tool for monitoring fires, as well as for measuring their effects on vegetation over the following years. We aim to explain the dynamics of wildfires' effects on a vegetation index (previously estimated by causal inference through synthetic controls) from pre-wildfire available information (mainly proceeding from satellites). For this purpose, we use regression models from Functional Data Analysis, where wildfire effects are considered functional responses, depending on elapsed time after each wildfire, while pre-wildfire information acts as scalar covariates. Our main findings show that vegetation recovery after wildfires is a slow process, affected by many pre-wildfire conditions, among which the richness and diversity of vegetation is one of the best predictors for the recovery.

Keywords: causal inference; functional data analysis; functional principal components analysis; function-on-scalar regression; landsat; NDVI; remote sensing; synthetic controls; time series decomposition; wildfires

Citation: Serra-Burriel, F.; Delicado, P.; Cucchietti, F. M. Wildfires Vegetation Recovery through Satellite Remote Sensing and Functional Data Analysis. *Mathematics* **2021**, *9*, 1305. https://doi.org/10.3390/math9111305

Academic Editors: Antonio Di Crescenzo and Ana M. Aguilera

Received: 2 May 2021
Accepted: 31 May 2021
Published: 7 June 2021

1. Introduction

Wildfires are becoming a major concern for societies around the globe, and research shows that changes in climate are going to alter the amount and size of wildfires in specific regions [1–4]. The effects are diverse depending on many factors, like weather conditions, vegetation affected, land cover, land management before and after the incident, the geographical region affected, or human vegetation management and risk mitigation. Wildfires occur by a combination of conditions created either by human intervention (e.g., power lines failures [5]) or by unpredictable events (such as lightnings [6,7], and thus are much harder to anticipate). As natural environments become more vulnerable to this kind of events, cities and inhabitable places need to be made more resilient, as they are likely to become more frequent due to changes in climate [3]. The result of these events increasing in size and frequency is hard to capture, as the amount of ecosystems and populations affected by these is very large.

Remote sensing can be defined as "the science of observation from a distance" [8], including many types of sensors. In this study, we are particularly interested in satellite images. These have become an invaluable and increasingly popular research field of study in the last few decades. Observation of the Earth from a distance has enormous potential. It allows monitoring and capturing changes in environments around the world, enabling their detection, quantification and possible prevention, which makes the modification of human environments more sustainable. Historically, natural disasters have played an important role in shaping societies, as these pose a significant threat in some regions on Earth. In order to create resilient and sustainable communities, remote sensing tools can

help adapt to these events [9–11], and help build environmental policies to protect Earth as we know it [12]. In this work, we are focusing on how wildfires affect vegetation and how environments recover from these catastrophic events. Remote sensing plays a critical role for assessing the impact of wildfires and learning to coexist with these events [13]. We use Functional Data Analysis (FDA, [14]) to analyze wildfire dynamics from remote sensing data. This work is part of the growing literature on FDA for remote sensing data (see, e.g., [15–17] among others).

Information from remote sensing provides a very important temporal component that allows studying and quantifying the dynamical evolution of the effects of wildfires and recoveries over time. Specifically, ref. [18] uses various sources of remote sensing data, combined with synthetic controls for assessing the vegetation impacts of wildfires over time. In this study, we analyze these recoveries processes as functional data. Each observation measures over several years how the vegetation evolves in a specific region that suffered from a large wildfire, and it represents the decrease or loss of vegetation (that will be defined in the next sections) from each wildfire, as a function of time t, starting at the time of the wildfire up until 7 years after the wildfire, showing the recovery of vegetation from these events.

Hence, the aim of this study is to explain the effects of wildfires on vegetation from remote sensing (satellite) images through FDA, as an alternative approach to classical regression methodologies used to study the effects of wildfires. Classical models usually summarize the whole recovery by comparing few periods of time, pre- and post-wildfire [19]. We take advantage of remote sensing technologies and modern statistical tools to answer questions like the following:

(i) What are the effects of wildfires on different kinds of environments?
(ii) Do the wildfire effects evolution depend on the vegetation of the burned area?
(iii) Can we explain recoveries of vegetation from wildfires using pre-wildfire observable covariates?

This study is focused on medium to large wildfires ($\geq$1000 acres, or 404 hectares) in California throughout a time-span of two decades (1996–2016). We explain the recoveries of vegetation from wildfires using pre-wildfire vegetation conditions and other characteristics of the affected lands using FDA. One of the main advantages from this methodology is that we can use the whole recovery process as a function of time.

Previous studies use differences between pre- and post-wildfire occurrence, showing relative difference between values over fixed time periods, or comparisons of few wildfires (e.g., a dozen wildfires [20], 3 or 5 years after the event [21,22]). This results in raster maps of differences between few time periods, gaining insights on the exact locations where vegetation has decreased. However, this approach lacks the temporal nature of the problem, as vegetation changes over time in a continuous manner.

In order to estimate the dynamical causal effects of wildfires, causal inference through synthetic controls was used in [18]. This methodology comes from the combination of Econometrics and Political Science, and it consists on the estimation of a hypothetical scenario (a counterfactual) with the absence of a wildfire (the intervention). Thus, in the present case, health vegetation indices were estimated in places where there were wildfires, as if the wildfires had not happened, using a Generalized Synthetic Control (GSC) methodology [23]. Then, the wildfire effect was estimated as the difference between the observed indices and the estimated counterfactuals. Usually the size of the wildfire effect decreases over time so we also refer as wildfire recovery to the wildfire effect as a function of time.

We use seasonality adjustment techniques to extract the trend of the wildfire effects estimated in the previous study. Then proceed to regress these effects, measured over time, using Functional Regression Models. More precisely, we regress functional responses on scalar covariates. This results in estimated coefficients changing over time that provide insights into different questions, as the ones stated above.

This paper is structured as follows. First, we introduce the used data and their pre-processing, as well as the algorithms used to obtain the outcomes to be predicted. Next, we explain the methodology that will be used in this study. Then, we show the attained results and summarize the key findings derived from this study. Last, we discuss the potential impact of these results and conclude with final notes.

2. Data Gathering

The study area of this paper is California over the period 1996–2016. There are three main data sources used for this study. First, perimeters from large wildfires ($\geq$404 hectares) were obtained from the Monitoring Trends in Burn Severity (MTBS) program [24] conducted by the United States Geological Services (USGS). Second, the Normalized Difference Vegetation Index (NDVI) Surface-Reflectances coming from several Landsat satellites was derived and aggregated using Google Earth Engine platform (GEE) over the areas of interest, as well as meteorological conditions over the areas of interest, that were obtained from GridMET [25] during the observed time span. Third, we use the results from a previous analysis in [18], where the effects of wildfires were estimated using the above two mentioned data sources. Details on these data sources are expanded below.

2.1. Wildfires Data

Perimeters from large wildfires ($\geq$404 hectares) that occurred over the considered time span were obtained from MTBS [24] program, as it provides a consistent source of wildfire perimeters for this period. Additionally, only perimeters of wildfires that did not overlap each other over the time period studied have been considered, because the synthetic control methodology used in [18] is not able to deal with units that experiment more than one intervention (multiple wildfires, in this case). After pruning the wildfires that either occurred too early and thus do not have enough pre-wildfire periods (at least 5 years) to estimate the counterfactual vegetation, and the wildfires that do not have enough follow-up years after the wildfire (at least 7 years), we end up with 243 wildfires. Figure 1 shows the perimeters of the burned areas. As an example, the upper right corner of Figure 1 shows the perimeter from a 2008 wildfire in the Mendocino County, officially named MEU LIGHTNING COMPLEX (MIDDLE). This fire burned 2087 acres, and the predominant land cover was evergreen forest. We have chosen this wildfire as an example because it corresponds to the modal median for 2008 (the deepest function in 2008 according to the modal depth [26]) and 2008 was the year with the largest amount of wildfires.

Moreover, several spatial covariates were obtained from MTBS: latitudinal and longitudinal centroid of the polygons, the year that the fire occurred, the month when it started, and the acres or size (in acres) of the burned areas. Lastly, another covariate indicating the average elevation of the burned areas was obtained from the National Elevation Dataset (NED) from the USGS. Table 1 shows a summary of the used covariates in this study.

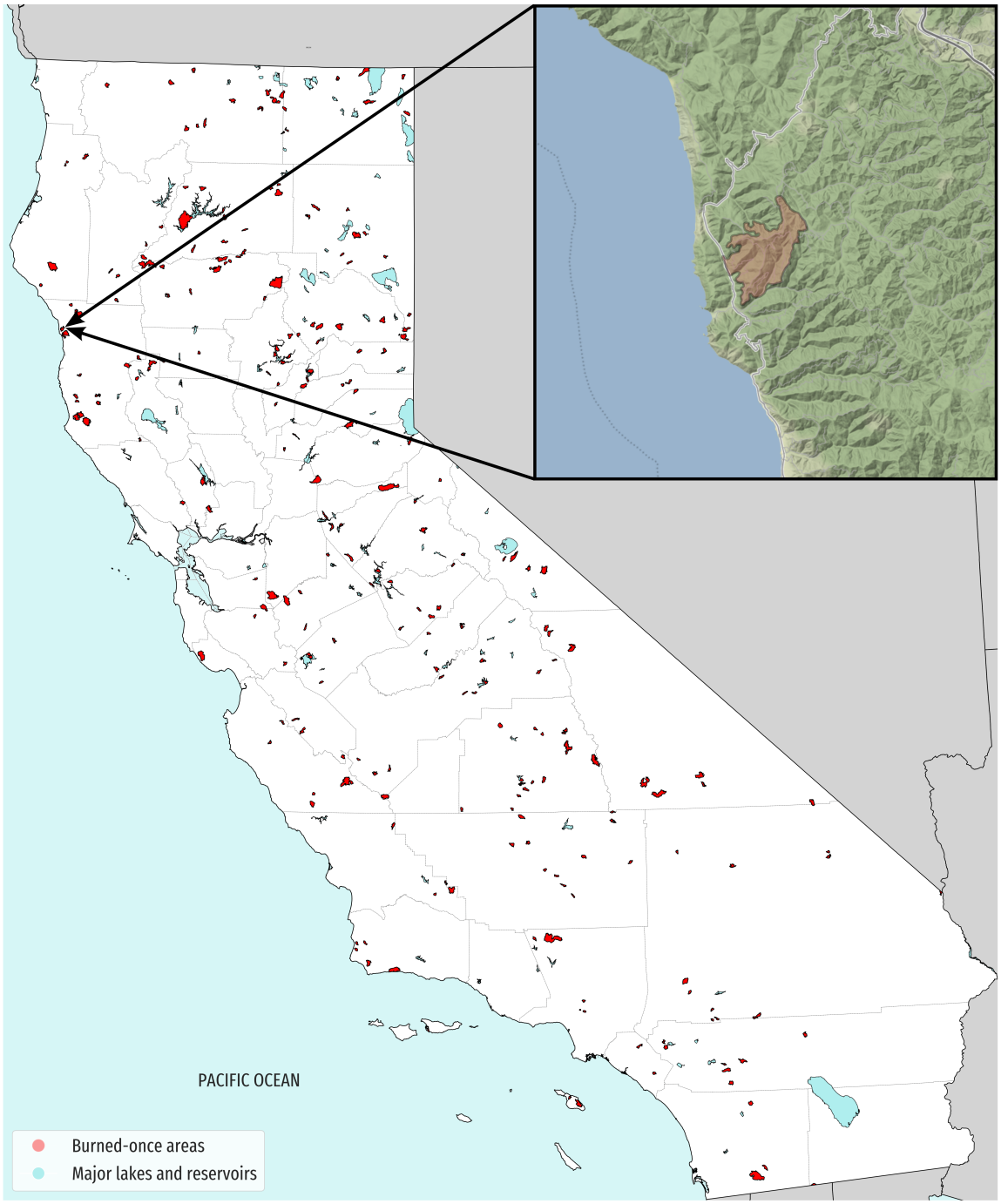

Figure 1. Map of California and the perimeters selected for this study. The upper right corner shows the perimeter of the MEU LIGHTNING COMPLEX (MIDDLE) wildfire in June 2008. The figure was generated using geopandas [27], and the data of the polygons was obtained from the Monitoring Trends in Burn Severity (MTBS) program [24], and polygons of the state and major lakes and reservoirs were obtained form the California Open Data Portal https://data.ca.gov (accessed on 6 April 2020). The top right-hand side picture was created with the contextily Python package https://contextily.readthedocs.io/en/latest/ (accessed on 20 April 2021), using the Map tiles by Stamen Deisgn, CC BY 3.0—Map data (C) OpenStreetMap contributors.

Table 1. Summary of the variables used in this study as covariates for the function-on-scalar regressions.

Variable	Description	Source
Latitude	Average of the South–North latitude coordinates for the pixels in the area of interest.	MTBS
Longitude	Average of the West–East longitude coordinates for the pixels in the area of interest.	MTBS
Avg Elevation	Average of the elevation over the sea level for the pixels in the area of interest.	NED
Year	Year the wildfire occurred.	MTBS
Start Month	Month the wildfire started.	MTBS
log(Acres)	Logarithm of the surface (in acres) of the burned area.	MTBS
Landcover	Predominant type of vegetation over the area of interest. Four categories: Shrubland/scrubland, evergreen forest, grasslands herbaceous and others.	GlobCover
Landcover Entropy	Shannon's Entropy of the distribution of Landcover among the pixels in the area of interest. Larger values indicate more variety of vegetation types.	GlobCover
Avg NDVI 5 years	Average of the Normalized Difference Vegetation Index (NDVI) for the 5 years of pre-wildfire periods (averaged over pixels).	LANDSAT
Std NDVI 5 years	Standard deviation of the NDVI for the 5 years of pre-wildfire periods (averaged over pixels).	LANDSAT
Burning Index	Burning index, a proxy for fire weather hazard, as defined in the National Fire Danger Rating System (NFDRS), averaged over pixels.	GridMET
Maximum Temperature	Maximum Temperature in Kelvin degrees (averaged over pixels).	GridMET
Rain	Daily precipitation in mm total (averaged over pixels).	GridMET
Solar Radiation	Solar Radiation in W/m^2 (averaged over pixels).	GridMET

2.2. Satellite Data

The NDVI is one of the Landsat Surface Reflectance Derived Specral Indices (LSR-DSI). For each pixel in a satellite image, it is defined as

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}},$$

where Red is the spectral reflectance measurement in the red band of the spectrum (centred near 0.66 µm), and NIR measures the reflectance in the near-infrared band (centred near 0.87 µm). Both, Red and NIR, are codified as 256 grey levels. Therefore the values of NDVI are always between -1 and 1, but in general they are non-negative. Large values of NDVI are associated with high contents of live green vegetation.

For instance, Figure 2 shows the NDVI (in red, averaged over pixels) for the MEU LIGHTNING COMPLEX (MIDDLE) wildfire example. This area was covered mainly by evergreen forest, having large NDVI values before the wildfire (they oscillate around 0.75).

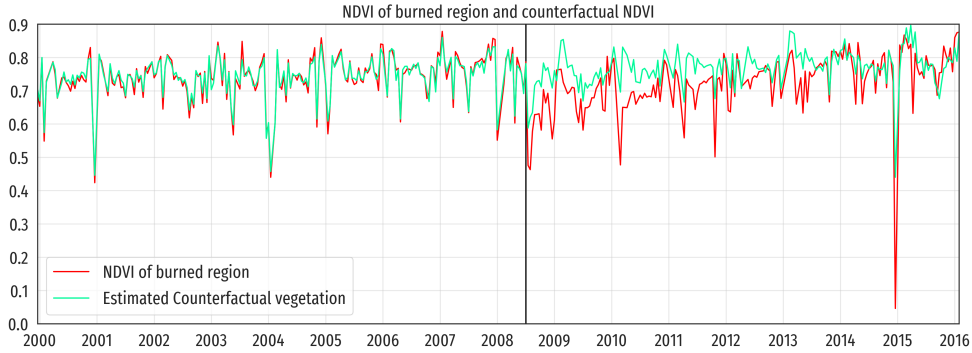

Figure 2. NDVI for MEU LIGHTNING COMPLEX (MIDDLE) wildfire. Plot of the vegetation indices NDVI of burned region and counterfactual NDVI vegetation between 2000 and 2016. This figure shows the evolution of NDVI, as well as the counterfactual estimated as explained in [18]. The data used to make this plot was obtained aggregating pixels over time of the polygons of burned areas from MTBS [24] using Google Earth Engine (GEE) [28], and the estimated counterfactual vegetation was created using the gsynth package [29].

We use the GEE platform to obtain the NDVI for images provided by three Landsat satellites (LT5, LT7 and LO8) masking clouds, shadows and snow pixels and removing pixels from water bodies such as lakes, reservoirs, rivers and creeks, as we already did in a previous study [18]. The Landsat satellites provide a consistent source of 30 m per pixel resolution, with a frequency of 16 days (approximately 26 observations per year). All the pixels within a burned region are aggregated by taking the average of each spectral index. In this way a time series of NDVI values is obtained for each region of interest. Further details can be found in [18].

Given that our main goal is predicting wildfires effects using pre-wildfire observable covariates, two additional explanatory variables were created from the spectral indices data. The average and standard deviation of the NDVI for 5 years of pre-wildfire periods were computed for all observations. These two variables work as proxies for the type of vegetation, e.g., larger NDVI values usually show forested areas, whereas lower values of the average of NDVI and larger standard deviations (associated with strong cyclical patterns) indicate grasslands or shrublands types of vegetation.

In addition, climatological covariates or weather conditions were obtained using GEE from GridMET [25]. These were also aggregated on the regions of interest, taking averages over the regions of interest on all the pre-wildfire available periods (from 1990 until the period where each wildfire occurs). This dataset has a resolution of 4 km per pixel and contains the maximum and minimum temperature (in Kelvin degrees), precipitation

accumulation (in daily milimetres), downward surface shortwave radiation (in W/m^2), and burning index from the National Fire Danger Rating System (NFDRS, [30]).

2.3. Effects of Wildfires Data

The main contribution of [18] was to estimate the effect of the studied wildfires over time.The wildfire effect was estimated as the difference between the observed spectral index and the estimated counterfactual (the values that the spectral index would have taken in a hypothetical scenario with the absence of wildfire). Counterfactuals are estimated in [18] following the proposals in [31], a way to perform GSC [23] based on matrix completion.

Figures 2 and 3 illustrate, for the MEU LIGHTNING COMPLEX (MIDDLE) wildfire example, the effect estimation process performed in [18]. Figure 2 shows the observed NDVI as well as the estimated counterfactual vegetation index. The estimated effect is the difference between these two time series and it is shown in Figure 3.

A descriptive analysis of the estimated wildfires effects is performed in [18]. Among its findings are the following. Depending on the region burned and the vegetation of these places, the effects can last from less than 2–3 years to more than a decade post-wildfire, and sometimes change the state of vegetation permanently. Serra-Burriel et al. [18] also found that the dynamical effects vary across regions, and have an impact on seasonal cycles of vegetation in later years. In order to have more conclusive results than the descriptive ones found in [18], statistical models must be proposed and estimated. A promising possibility is considering regression models with functional response (the estimated wildfire effects as functions of the time elapsed after the wildfire) and explanatory variables such as geographical location, burn severity, size of the burned area, and land cover/vegetation type. This constitutes the main contribution of the present project.

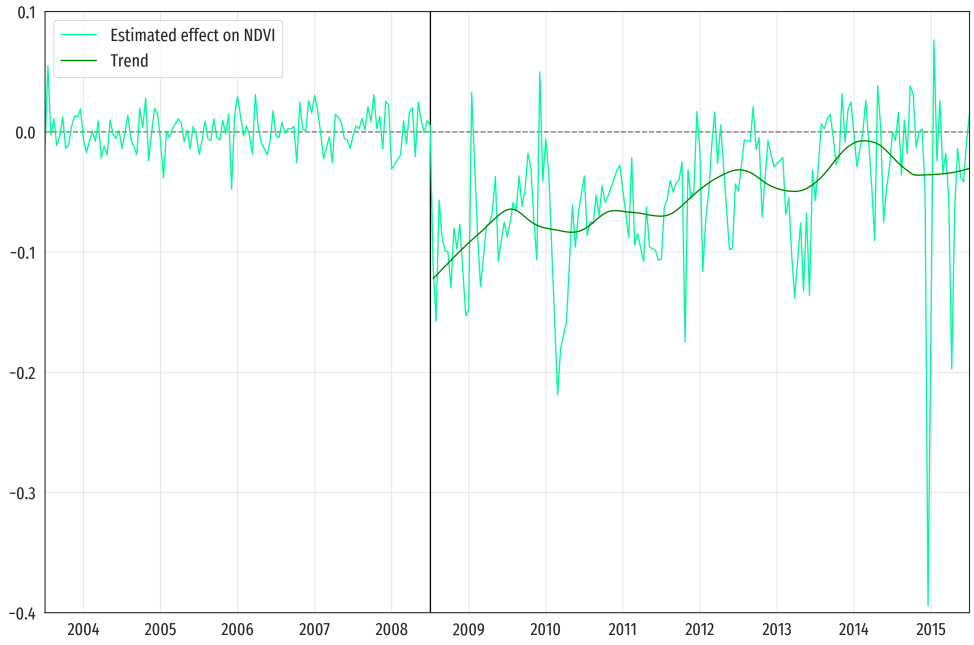

Figure 3. Plot of the effect and trend extracted for the MEU LIGHTNING COMPLEX (MIDDLE) wildfire. The estimated effect is shown in light green. The extracted trend is shown in dark green. This figure shows 5 previous years and 7 years after the wildfire, as this is the inclusion criteria in this study.

The wildfire effects over time estimated in [18] for 7 years post-wildfire and for each of the 243 wildfires that meet our inclusion criteria, are the base from which we construct the functional dataset that will be analyzed in this study. We perform one last step to preprocess the data, that is the trend extraction as explained in Section 3.1.

3. Methods

NDVI time series usually present seasonality, as vegetation changes throughout the seasons of the year. This is especially evident for some types of vegetation, such as grasslands or shrublands. Therefore, we expect post-wildfire NDVI time series of both, the observed and the estimated counterfactual vegetation indices, to present seasonal components. These seasonal components will have different amplitudes, since the burned region will present distinct seasonal patterns during the recovery. Therefore, the difference between the burned region NDVI and the counterfactual NDVI will presumably present a changing seasonal pattern.

Note that, when aligning all the timings of the wildfires, the seasonal pattern of each particular wildfire will present a different phase, as the timings throughout the year of wildfires are different: some wildfires occur on summer periods as opposed to the ones that occur during early spring. Hence, before aligning the recoveries for all wildfires, to conform a unique functional dataset with no mismatches in the phases of seasonality we need to extract the seasonal pattern of each wildfire separately.

In addition, several aspects of the remote sensed data can produce measurement error. Even though pixels that captured clouds were not included at the timing of aggregating multispectral data to measure vegetation, other types of noise could have potentially leaked in the data. To reduce the amount of noise and extract recoveries of vegetation from wildfires, it is suitable for this analysis to smooth the data.

Therefore, for each time series, we perform a LOcally Estimated Scatterplot Smoothing (LOESS) decomposition, that will simultaneously remove the individual seasonality from the time series, as well as remove noise from the remotely sensed data.

3.1. Trend Extraction with LOESS and Functional Representation of Data

Once the effects for each wildfire are obtained, we decompose the time series into its structural components. Trend extraction of univariate data is a wide field of study [32], where the classical decomposition model [33] is a time series decomposed in additive terms, separating trend, seasonal component and residuals. Assuming the time series can be expressed as the addition of separate terms, for a wildfire starting at calendar time t_0 (in years) we have

$$y(t) = T(t) + S(t) + R(t),$$

where $y(t)$ is the outcome observed y at time $t = t_0 + j/26$ for $j \in \{1, \ldots, N = 7 \times 26\}$, $T(t)$ is the trend component at time t, $S(t)$ is the seasonal component, which is approximately periodic with cycles of length one year (26 instants of time) in our case, and $R(t)$ is the residual component of the time series.

One method commonly used in many fields for time series decomposition is the seasonal-trend decomposition procedure using LOESS [34], that is based on local polynomial fitting. This procedure presents several advantages, such as the flexibility on the trend and seasonal components extraction or the ability to decompose series with missing values.

We use the LOESS implementation from the Python [35] library `statsmodels` [36] to extract the trend from the wildfire effects time series, removing the seasonal and the residual components at once. Figure 4 shows an example of the time series decomposition in the MEU LIGHTNING COMPLEX (MIDDLE) example. Figure 3 also shows (in dark green) the extracted trend over the estimated effect (in light green).

Finally, we align all the extracted trends at $t_0 = 0$ and represent them as functional data. Each of the 243 wildfires is now represented by a function over 7 years of recovery. Each year of data contains 26 discrete values for each observation. Figure 5 shows the functional dataset of NDVI trend recoveries, jointly with their mean function. The MEU LIGHTNING COMPLEX (MIDDLE) wildfire example is also highlighted in the figure. Raw data representation has been used for these functional data, that is, every function is represented as a column vectors with j-th element equal to the value of the function at time $t = j/26$, for $j = 1, \ldots, N = 7 \times 26$.

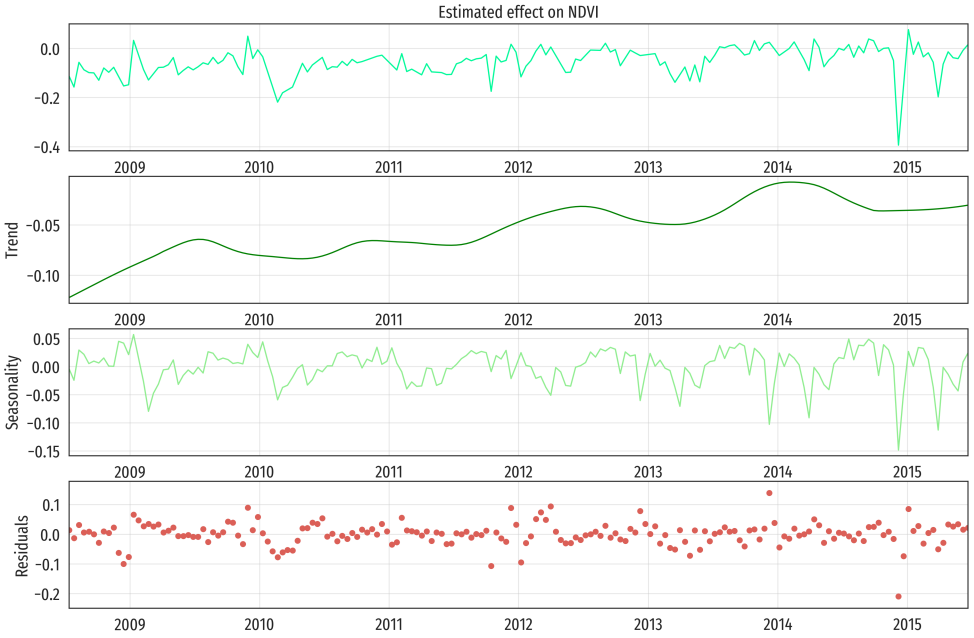

Figure 4. Plot of time series decomposition using LOcally Estimated Scatterplot Smoothing (LOESS) of the MEU LIGHTNING COMPLEX (MIDDLE) wildfire. The four graphics show, from top to bottom, the estimated wildfire effect, the extracted trend, the seasonal component, and the residuals.

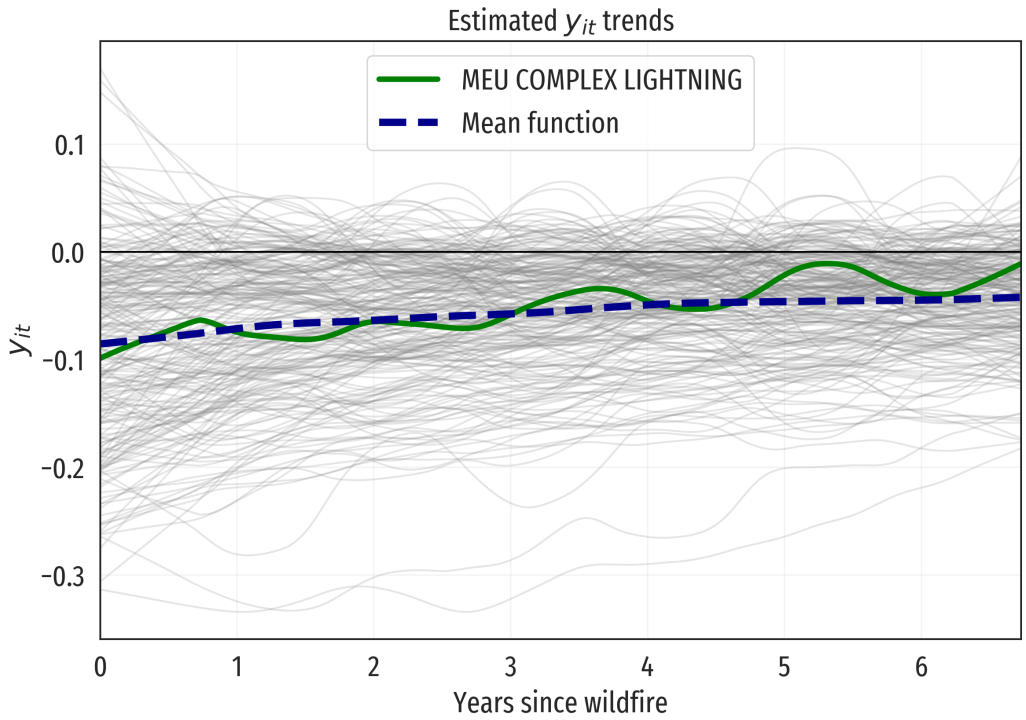

Figure 5. Plot of the functional dataset, composed by the extracted trends from the 243 estimated wildfire effects. The trend corresponding to the MEU LIGHTNING COMPLEX (MIDDLE) wildfire example has been marked in green. The functional mean is also represented (in dashed blue lines).

3.2. Functional Principal Components Analysis

Functional Principal Component Analysis (FPCA; see, for instance, [14] or [37]) is a dimensionality reduction technique for functional data that generalizes the well known Principal Component Analysis extensively used for multivariate data.

Consider a functional dataset $\{y_i(t) : i = 1, \ldots, n,\ t \in \mathcal{T} = [a, b] \subset \mathbb{R}\}$ with elements in $L^2(\mathcal{T})$, the set of square integrable functions defined on $\mathcal{T}$ equiped with the inner product $\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)dt$. It is assumed that these functional data are independent realizations of a functional random variable Y. The main objective of the FPCA is to determine the main modes of variation of the observed functions around the mean function. Formally, FPCA can be stated as follows. Let $\bar{y}(t) = (1/n)\sum_{i=1}^{n} y_i(t)$ be the mean function of the observed functional data. FPCA looks for functions $g_1, \ldots, g_q$ in $L^2(\mathcal{T})$ (principal functions) and real numbers (scores) ψ_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, q$, such that

$$\sum_{i=1}^{n} \int_{\mathcal{T}} \left((y_i(t) - \bar{y}(t)) - \sum_{j=1}^{q} \psi_{ij} g_j(t) \right)^2 dt$$

is minimum. Moreover, the functions $g_1, \ldots, g_q$ are required to be orthonormal ($\int_{\mathcal{T}} g_i(t)g_j(t)dt = \mathbb{1}_{\{i=j\}}$). In other words, we are looking for a representation of functional data in the q-dimensional space spanned by the functions $g_1(\cdot), \ldots, g_q(\cdot)$:

$$y_i(t) \approx \bar{y}(t) + \sum_{j=1}^{q} \psi_{ij} g_j(t),\ t \in \mathcal{T},\ i = 1 \ldots n.$$

For $s, t \in \mathcal{T}$, the empirical covariance function is defined as

$$\hat{c}(s, t) = \frac{1}{n} \sum_{i=1}^{n} (y_i(s) - \bar{y}(s))(y_i(t) - \bar{y}(t)).$$

It can be proven that the principal functions are the eigen-functions corresponding to the largest q eigenvalues of the sampling covariance operator, that is,

$$\hat{C}(g_j)(t) = \int_{\mathcal{T}} \hat{c}(s, t) g_j(s) ds = \lambda_j g_j(t), \text{for all } t \in \mathcal{T},\ j = 1, \ldots, q,$$

with $\lambda_1 \geq \cdots \geq \lambda_q$.

Moreover the score of the i-th functional data on the j-th principal function is $\psi_{ij} = \int_{\mathcal{T}} (y_i(t) - \bar{y}(t)) g_j(t) dt$.

The numerical computation of the functional principal components can be performed in different ways. We follow the proposal of [14] (Chapters 8 and 9), based on cubic B-spline bases expansions of both, the observed functional data and the eigenfunctions of the sampling covariance operator.

Let $B_1(t), \ldots, B_K(t)$ a cubic B-spline basis on the interval $\mathcal{T} = [a, b]$. We consider the expansion of the centered functional data in this basis:

$$y_i(t) - \bar{y}(t) \approx \sum_{k=1}^{K} \alpha_{ik} B_k(t),\ t \in \mathcal{T},$$

that we write in vector notation as $y_i - \bar{y} \approx \boldsymbol{\alpha}_i^{\mathrm{T}} \mathbf{B}$, where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})^{\mathrm{T}}$ and $\mathbf{B}(t) = (B_1(t), \ldots, B_K(t))^{\mathrm{T}}$. Then

$$\hat{c}(s, t) \approx \tilde{c}(s, t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{B}(s)^{\mathrm{T}} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^{\mathrm{T}} \mathbf{B}(t) = \mathbf{B}(s)^{\mathrm{T}} \mathbf{A} \mathbf{B}(t),$$

where $\mathbf{A} = (1/n)\sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^{\mathrm{T}}$.

For a generic $f \in L^2(\mathcal{T})$, let $\sum_{k=1}^{K} \beta_k B_k(t) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{B} \approx f$ be the expansion of f in the cubic B-spline basis. Then $\|f\|^2 = \langle f, f\rangle \approx \boldsymbol{\beta}^{\mathrm{T}}\Phi\boldsymbol{\beta}$, with $\Phi_{hj} = \int_{\mathcal{T}}\int_{\mathcal{T}} B_h(s)B_j(t)dsdt$. It can be proven that the first eigenfunction of the sampling covariance operator is also the solution of $\max_{f:\|f\|^2=1} \widehat{\mathrm{Var}}(\langle Y, f\rangle)$. However,

$$\widehat{\mathrm{Var}}(\langle Y, f\rangle) = \langle \hat{C}(f), f\rangle = \int_{\mathcal{T}}\left(\int_{\mathcal{T}} \hat{c}(s,t)f(s)ds\right)f(t)dt \approx$$

$$\int_{\mathcal{T}}\left(\int_{\mathcal{T}} B(s)^{\mathrm{T}}\mathbf{A}B(t)\boldsymbol{\beta}^{\mathrm{T}}B(s)ds\right)B(t)^{\mathrm{T}}\boldsymbol{\beta}dt =$$

$$\boldsymbol{\beta}^{\mathrm{T}}\left(\int_{\mathcal{T}}\int_{\mathcal{T}} B(s)B(s)^{\mathrm{T}}\mathbf{A}B(t)B(t)^{\mathrm{T}}dsdt\right)\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}}\Phi\mathbf{A}\Phi\boldsymbol{\beta}.$$

Then $\max_{f:\|f\|^2=1} \widehat{\mathrm{Var}}(\langle Y, f\rangle)$ is (almost) equivalent to

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^K:\boldsymbol{\beta}^{\mathrm{T}}\Phi\boldsymbol{\beta}=1} \boldsymbol{\beta}^{\mathrm{T}}\Phi\mathbf{A}\Phi\boldsymbol{\beta} = \max_{\boldsymbol{\beta}\in\mathbb{R}^K:(\Phi^{1/2}\boldsymbol{\beta})^{\mathrm{T}}(\Phi^{1/2}\boldsymbol{\beta})=1} (\Phi^{1/2}\boldsymbol{\beta})^{\mathrm{T}}\left(\Phi^{1/2}\mathbf{A}\Phi^{1/2}\right)(\Phi^{1/2}\boldsymbol{\beta})$$

and the problem reduces to the diagonalization of $\Phi^{1/2}\mathbf{A}\Phi^{1/2}$. Let $\mathbf{u}_1$ be the eigenvector associated with its largest eigenvalue. Then we take $\boldsymbol{\beta}_1 = \Phi^{-1/2}\mathbf{u}_1$ and the first eigenfunction we are looking for is $g_1(t) = \sum_{k=1}^{K} \beta_{1k}B_k(t) = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{B}(t)$.

For obtaining successive eigenfunctions, it must be taken into account that two functions $g_h = \boldsymbol{\beta}_h^{\mathrm{T}}\mathbf{B}$ and $g_j = \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{B}$ are orthogonal if and only if $\boldsymbol{\beta}_h^{\mathrm{T}}\Phi\boldsymbol{\beta}_j = (\Phi^{1/2}\boldsymbol{\beta}_h)^{\mathrm{T}}(\Phi^{1/2}\boldsymbol{\beta}_j) = 0$. Therefore finding the eigenfunctions of the sampling covariance operator reduces to looking for the eigenvalues of the matrix $\Phi^{1/2}\mathbf{A}\Phi^{1/2}$.

In this approach to FPCA the smoothness of the principal functions $g_1, \ldots, g_q$ is inherited from the smoothness of the observed functional data $y_1, \ldots, y_n$ via the empirical covariance function $\hat{c}(s,t)$. Nevertheless it is possible to force smoothness in the eigenvalues explicitly performing a regularized version of FPCA (see [14] (Chapter 9)). To do so, the problem to be solved is

$$\max_f \frac{\widehat{\mathrm{Var}}(\langle Y, f\rangle)}{\|f\|^2 + \lambda\|f''\|^2}$$

for som $\lambda > 0$, where f'' is the second derivative of f and the maximization is done in the space of functions $f \in L^2(\mathcal{T})$ for which f'' is also in $L^2(\mathcal{T})$. It is easy to see that this problem with $\lambda = 0$ is equivalent to the previously considered FPCA problem, namely $\max_{f:\|f\|^2=1} \widehat{\mathrm{Var}}(\langle Y, f\rangle)$. In [14], is proved that the regularized FPCA can numerically solved by the diagonalization of the matrix

$$\Psi^{-1/2}\Phi\mathbf{A}\Phi\Psi^{-1/2},$$

where $\Psi = \Phi + \lambda\Gamma$, and Γ is the $K \times K$ matrix with generic (h,j) element $\Lambda_{hj} = \int_{\mathcal{T}}\int_{\mathcal{T}} B_h''(s)B_j''(t)dsdt$.

3.3. Functional Regression Models

Analogous to classical regression models, Functional Regression Models (FRM) regress outcomes based on covariates when using functions as either the outcomes or regressors. Hence, FRM take advantage of the nature of time changing variables, either parametrically or non-parametrically. To do so, it can use the functional representation of both regressors and/or outcomes.

3.3.1. Function-on-Scalar Regression

In this research we use the function-on-scalar regression methodology (see, e.g., [14,38,39]) as it allows us to understand the relation between the observed outcome over time, with respect to the fixed covariates observed. Let (X, Y) be a pair of random

variables, where Y is functional and $X = (X_1, \ldots, X_k)$ is a random vector of dimension k. The linear function-on-scalar regression model for Y given $X = (x_{i1}, \ldots, x_{ik})$ is stated as

$$Y_i(t) = \beta_0(t) + \beta_1(t)x_{i1} + \cdots + \beta_k(t)x_{ik} + \varepsilon_i(t), \tag{1}$$

where $Y_i(t)$ is the functional response over time $t \in \mathcal{T}$ for the observation i, x_{ij} is the value of variable X_j in the observation i, $\beta_0(t)$ is the functional intercept (it is equal to the mean function $(Y(t))$ when the k covariates are centered), $\beta_j(t)$ is the functional coefficient for the j-th covariate X_j for $j \geq 1$, and $\varepsilon_i(t)$ is the functional error for the i-th observation, a zero mean continuous stochastic process, assumed to be indpendent for different observations. The problem of variable selection in the linear function-on-scalar regression model was addressed in [40].

However, different kinds of covariates can be considered, as not all of them have a changing effect over time, or might have different effects. In order to allow the function-on-scalar regression model to admit richer covariate terms, ref. [41] introduced the functional additive mixed model (where functional covariates are also allowed). As an example, the following equation shows a function-on-scalar additive regression model with terms of different types:

$$y_i(t) = \beta_0(t) + \beta_1 x_{i1} + s_2(x_{i2}) + \beta_3(t)x_{i3} + \gamma_4(t, x_{i4}) + \varepsilon_i(t), \tag{2}$$

where $\beta_0(t)$ is the functional intercept, β_1 is constant over time, $s_2(x_{i2})$ is a smooth function of the covariate, $\beta_3(t)x_{i3}$ is the same kind of covariate-coefficient relation from Equation (1), $\gamma_4(t, x_{i4})$ is a smooth function depending on t and x_{i4}, and finally $\varepsilon_i(t)$ is the i-th error function. Variable selection is less developed for the function-on-scalar additive model than for the linear function-on-scalar model.

In the estimation of model (2), the response functions y_i have been represented as raw data, that is, as a column vector $\mathbf{y}_i \in \mathbb{R}^N$ with $y_i(t_j)$ as the j-th entry, where $t_j = j/26$. Coefficient functions $\beta_j(t)$, $j \geq 0$, are represented by their expansion in a cubic B-spline basis: $\beta_j(t) = \sum_{k=1}^{K_j} \beta_{jk} B_k(t) = \boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{B}(t)$. The smooth functions of the covariates, as $s_2(x_{i2})$, are represented also by expansions in a cubic B-spline basis over the range of the corresponding explanatory variable. For instance, $s_2(x_2) = \sum_{h=1}^{H_2} \delta_{2h} D_{2h}(t) = \boldsymbol{\delta}_2^{\mathrm{T}} \mathbf{D}_2(x_2)$. Finally, smooth functions depending on t and an explanatory variable, as x_4, are represented by their expansions in a tensor product basis. For instance, $\gamma_4(t, x_4) = \sum_{k=1}^{K_4} \sum_{h=1}^{H_4} \xi_{kh}^4 B_k(t) D_{4h}(x_4) = \mathrm{vec}(\mathbf{B}(t)\mathbf{D}_4^{\mathrm{T}}(x_4))^{\mathrm{T}} \boldsymbol{\xi}_4$, where $\boldsymbol{\xi}_4 \in \mathbb{R}^{K_4 \times H_4}$, and $\mathrm{vec}(M)$ is the vector formed by concatenation of the columns of matrix M. Using these representations, for the i-th observation, model (2) can be written as

$$\mathbf{y}_i = \mathbf{B}_N \boldsymbol{\beta}_0 + 1_N x_{i1} \beta_1 + 1_N \mathbf{D}_2^{\mathrm{T}}(x_{i2}) \boldsymbol{\delta}_2 + x_{i3} \otimes \mathbf{B}_N \boldsymbol{\beta}_3 + \left(\mathbf{D}_4^{\mathrm{T}}(x_{i4}) \otimes \mathbf{B}_N\right) \boldsymbol{\xi}_4 + \boldsymbol{\varepsilon}_i, \tag{3}$$

where 1_N denotes the N-vector of ones, $\mathbf{B}_N$ is the $N \times K$ matrix with element (j, k) equal to $B_k(t_j)$, and $M_1 \otimes M_2$ denotes the Kronecker product of matrices M_1 and M_2, and $\boldsymbol{\varepsilon}_i$ is the raw data representation of the functional noise $\varepsilon_i(t)$. Following [41], model (3) can be expressed as

$$\mathbf{y}_i = \boldsymbol{\Phi}_i \boldsymbol{\theta} + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\Phi}_i$ and $\boldsymbol{\theta}$ can be partitioned into 5 blocks, each corresponding to a term in (3). Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4)$ be the partition corresponding to the parameters.

Assuming white noise, that is $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\varepsilon}_n^{\mathrm{T}})^{\mathrm{T}} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{nN})$, the penalized likelihood criterium to be minimized for estimating the model is

$$\sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\Phi}_i \boldsymbol{\theta}\|^2 + \sum_{v \in \{0,2,3,4\}} \lambda_v \boldsymbol{\theta}_v^{\mathrm{T}} \mathbf{P}_v \boldsymbol{\theta}_v,$$

where matrices $\mathbf{P}_v$ in the the penalty terms are known positive semi-definite matrices, related with the integral of products of the second derivatives of the elements in the B-splines basis used in the smoothing terms. The smoothing parameters λ_v control the trade-off between goodness of fit to the training data, and smoothness of the non-parametrically estimated functions of t and/or x_j. The proposal in [41] is to adopt a linear mixed effects model approach to the estimation process, in which the model parameters $\boldsymbol{\theta}$ and the smoothing parameters λ_v are estimated simultaneously by restricted maximum likelihood (REML), as it is done in the R library `mgcv` ([42]) upon which [41] base the function `pffr` in their library `refund`, for estimating models as (2).

4. Results

In this section, we present the main results from this study, showing how the characteristics of the vegetation and land cover previous to the wildfire, as well as the prior weather conditions to the wildfire, affect the vegetation recovery patterns.

We start summarizing the functional dataset containing the 243 wildfire recoveries. Their mean function is represented in Figure 5, jointly with the complete dataset. The mean wildfire effect on NDVI is always negative for the 7 year period after the wildfire, and the absolute value of this negative effect is monotonically decreasing over time, going from -0.0856 at time 0 to -0.0418 seven years later, in terms of lost NDVI points, with a global average of -0.0567. In average, the burned areas are progressively recovering 0.0438 NDVI points after wildfires (approximately 10% of the range of the functional data set values, see Figure 5). It is also noticeable that, on average, it takes more than 7 years for a complete recovery of the NDVI: the value of the mean function after 7 years is still negative. The library `fda.usc` [43] in R [44] has been used for the descriptive analysis, including the choice of the MEU LIGHTNING COMPLEX (MIDDLE) as an illustrative wildfire example, as it has the modal median recovery function in 2008 (the modal year).

Next, FPCA has been applied to find the main modes of variation of the studied functional data around the average. Figure 6 shows the mean, and the mean plus/minus a constant times the first four principal functions, that have been computed using the function `pca.fd` from package `fda` [45] in R, as described in Section 3.2. In particular, the number of functions in the B-spline basis has been $K = 60$ and the penalty parameter $\lambda = 0$. These choices have been determined when using the function `fdata2fd` from library `fda.usc` with default parameters to transform the raw functional data into a `fd` class object of library `fda`.

The first principal function explains almost 90% of the variability, showing a direction of severity in the NDVI drop: wildfires with positive scores in this principal function experiment smaller drops in NDVI than those having negative scores. The second principal function (4.3% of the total variability) can be interpreted as a direction separating wildfires with faster recoveries (those with more positive scores) from those with slower regeneration capacity (wildfires with more negative scores). The following two functional components only explain less than 4% of the total variance, with no clear recovery patterns, so they should be interpreted with caution.

The main goal of this study is to quantify the influence that different pre-wildfire conditions (geographical region, climatological conditions, or vegetation types) of the burned areas have on wildfire effects over the subsequent years post-wildfire. In order to achieve this goal, function-on-scalar additive models (of the type from Equation (2)) are fitted using the function `pffr` from the library `refund` [46] in R. The list of potential covariates to be included in this model is given in Table 1. The number of functions in the B-Spline basis in all of these fitted models are the default values suggested by the `pffr` function: 20 for the functional intercept $\beta_0(t)$, 5 for smoothing terms depending on t (for instance, $\beta_3(t)$ and $\gamma_4(t, x_4)$ in model (2)), and 10 for smoothing terms depending on other covariates (for instance, $s_2(x_2)$ and $\gamma_4(t, x_4)$ in model (2)). Notice that terms as $\gamma_4(t, x_4)$ need two bases, one in the dimension of t and the other in the dimension of the explanatory variable x_4.

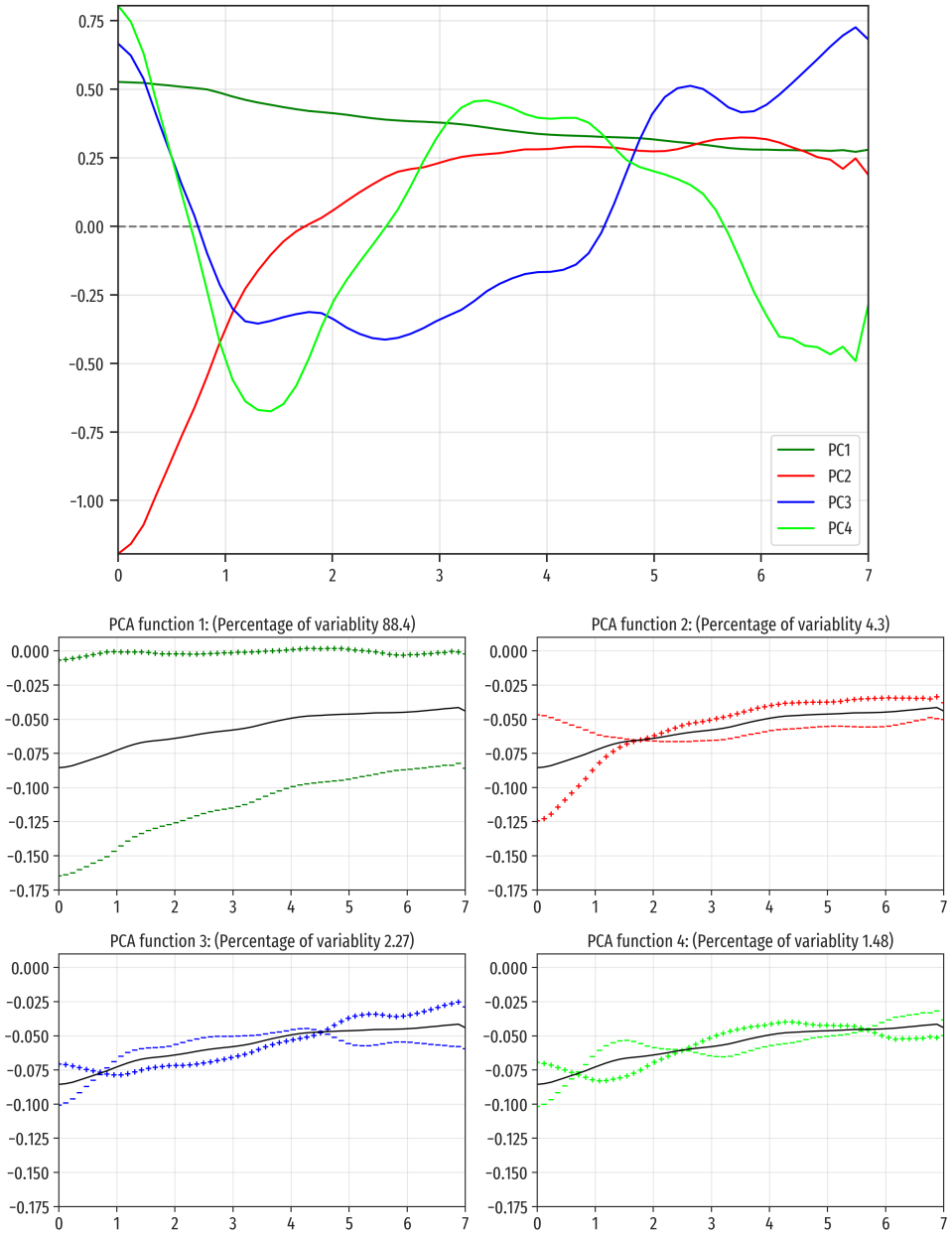

Figure 6. Functional principal component analysis results for the wildfire recoveries dataset. The upper plot shows the first four principal functions (with eigenvalues $\lambda_1 = 0.020670$, $\lambda_2 = 0.001005$, $\lambda_3 = 0.000530$, and $\lambda_4 = 0.000346$). The lower plots show the mean (black solid line), and the mean plus/minus a constant times each principal function. The percentage of variance explained by each component is indicated in the headers.

As far as we know, the variable selection problem for the function-on-scalar additive model is still an open issue, as we mentioned in Section 3.3.1. In fact, library `refund` includes a function doing variable selection for the linear function-on-scalar model (`fosr.vs`), but not for the additive extension. Additionally, each of the explanatory variables can enter in the function-on-scalar additive model in several ways, as it is illustrated in Equation (2). Therefore we have developed a heuristic model building strategy, which we describe below.

To select the way in which we introduce each covariate to the function-on-scalar additive model, five different univariate models have been fitted for each covariate separately. Exceptions were made for three pairs of covariates (longitude and latitude, average and standard deviation of NDVI for 5 years pre-wildfire, and landcover and landcover

entropy) that have been included together additively in these five single models, because both variables in each pair are jointly summarizing the same characteristic (geographic location, NDVI, and land cover). Table 2 shows the results from the $11 \times 5 = 55$ different fitted models (all of them being sub-models of Equation (2)), in terms of the percentage of observed variability explained (100 times the adjusted R^2).

Table 2. Percentage of observed variability explained from 11×5 univariate or bivariate function-on-scalar regression models. For each row, the selected model is marked in bold.

	Term Included in Each Model				
Variable	βx	$s(x)$	$\beta(t)x$	$\beta(t)x + s(x)$	$\gamma(t,x)$
Latitude, Longitude	7.05	**19.91**	7.08	19.94	17.63
Avg Elevation	19.03	**23.86**	19.32	24.15	24.91
Year	4.44	**7.93**	4.50	7.99	7.19
Start Month	6.40	**7.72**	6.57	8.39	8.17
log(Acres)	6.20	**8.91**	6.51	9.22	9.02
Landcover and Landcover Entropy	4.92	**7.25**	4.72	7.26	6.98
Avg and Std NDVI 5 years before	30.58	43.98	33.45	**46.85**	46.98
Burning Index	8.71	**14.70**	9.00	14.99	13.52
Maximum Temperature	21.74	**27.78**	22.19	28.23	28.93
Rain	22.17	29.22	23.25	**30.30**	28.34
Solar Radiation	7.60	**15.18**	7.70	15.28	16.24

The columns in Table 2 correspond to different types of models, and the rows to the variable (or to the pair of variables) used as regressors in the models. In each row, the complexity of the models increases from left to right: in the first two models, the terms depend only on the explanatory variable (linearly first, then non-parametrically), while in the other three models it depends on both, the covariate and the time index (in the third column, the term is linear in the covariate and nonparametric in time, the fourth model includes the second and third models terms additively, and finally the fifth model is nonparametric simultaneously in the covariate and the time index). In general, the models including a nonparametric term in the covariates have larger percentages of explained variability (columns 2, 4 and 5, which show an even performance) than those that are linear in the covariates (columns 1 and 3). Additionally, the inclusion of time dependent coefficients $\beta(t)$ (column 3) does not represent a large improvement with respect to the standard linear term (column 1). Therefore, for each row, a model has been selected according to a balance between explanatory power and model simplicity: a simpler model is preferred to a more complex one, if the difference in percentage of explained variability is less than 1%. At each row, the selected model is marked in bold.

Observe that the best univariate (or bivariate) fits in Table 2 correspond to the models having average and standard deviation of NDVI for the five previous years to the wildfires as covariates (almost 47% of explained variability), followed by those including rain (30%) or maximum temperature (around 28%) as explanatory variables.

Despite we do not delve any further into the results of these simple models (further comments on individual covariates effect on the response will be made below), we are going to build a multiple function-on-scalar additive model. Rather than delving further into the results of these simple models, we are going to build an additive multiple function scalar model, which in turn will provide further insights on the effect of individual covariates on the response.

We then proceed to fit a full model (using again the function `pffr` in `refund`), which includes the terms selected in Table 2. The covariates have been centered and standardized before fitting the model to force all of them to share a common scale. This way the estimated functions are comparable to each other. Tables 3 and 4, and Figures 7 and 8, summarize

the fitted model. This model explains a 72.9% of the variability observed in the response, strongly improving the best model included in Table 2 (46.98%). Tables 3 and 4 indicate that all the terms included in the model are highly significant. This fact and the large percentage of explained variability suggest that this is an adequate model.

Table 3. Full function-on-scalar additive model. Estimation of the parametric terms.

Parametric Terms	Estimate	Std. Error	t Value	Pr ($\geq \lvert t \rvert$)
(Intercept)	−0.0574	0.0003548	−155.253	$<2 \times 10^{-16}$
Landcover Grassland/Herbaceous	−0.0022	0.0003699	−4.085	4.42×10^{-05}
Landcover Shrub/Scrub	0.0031	0.0004906	6.338	2.34×10^{-10}
Landcover Other	0.0046	0.0013508	3.429	0.000606

Table 4. Full function-on-scalar additive model. Estimation of the nonparametric terms.

Nonparametric Terms	Edf	Ref.df	F	*p*-Value
Intercept(t)	13.218	19.000	364.17	$<2 \times 10^{-16}$
s_1(Latitude)	8.973	9.000	346.12	$<2 \times 10^{-16}$
s_2(Longitude)	8.984	9.000	321.41	$<2 \times 10^{-16}$
s_3(Avg Elevation)	8.632	8.960	520.30	$<2 \times 10^{-16}$
s_4(Year)	8.959	8.999	123.46	$<2 \times 10^{-16}$
s_5(Start Month)	4.977	5.000	85.87	$<2 \times 10^{-16}$
s_6(log(Acres))	8.906	8.997	9.000	$<2 \times 10^{-16}$
s_7(Entropy landcover)	8.977	9.000	346.38	$<2 \times 10^{-16}$
s_8(Avg NDVI 5 years before)	8.988	9.000	675.16	$<2 \times 10^{-16}$
$\beta_9(t)$ Avg NDVI 5 years before	3.558	3.831	377.16	$<2 \times 10^{-16}$
s_{10}(Std NDVI 5 years before)	8.825	8.989	102.17	$<2 \times 10^{-16}$
$\beta_{11}(t)$ Std NDVI 5 years before	3.962	3.999	433.37	$<2 \times 10^{-16}$
s_{12}(Burning Index)	8.940	8.998	214.318	$<2 \times 10^{-16}$
s_{13}(Maximum temperature)	8.980	9.000	487.19	$<2 \times 10^{-16}$
s_{14}(Rain)	8.966	8.999	286.88	$<2 \times 10^{-16}$
$\beta_{15}(t)$ Rain	3.585	3.844	32.83	$<2 \times 10^{-16}$
s_{16}(Radiation)	8.923	8.998	249.83	$<2 \times 10^{-16}$

We describe first the results for the parametric part of the model (Table 3), which only includes the covariate Landcover (a factor with four levels) with constant effects over time. The reference level for this factor is *Evergreen forest*. Table 3 shows that burned areas having had Grassland/Herbacious as dominant land cover experiment larger decrement in NDVI than evergreen forest areas. The opposite happens for areas at which shrubland or scrubland were dominant. Regarding the constant coefficients, the most affected areas when a wildfire happens are grassland/herbaceous (that lose 0.0596 points of NDVI on average; we noted before that the global average loss is 0.0567 NDVI points), followed by evergreen forests (losing 0.0574 points of NDVI), then shrublands and scrublands (with a reduction of 0.0543 points of NDVI), and finally areas at which other types of vegetation are dominant (where the NDVI reduction is of 0.0528 points in average). However, the landcover covariate cannot be interpreted separately from the other covariates (mainly the average and the standard deviation of NDVI, which strongly depend on types of landcover).

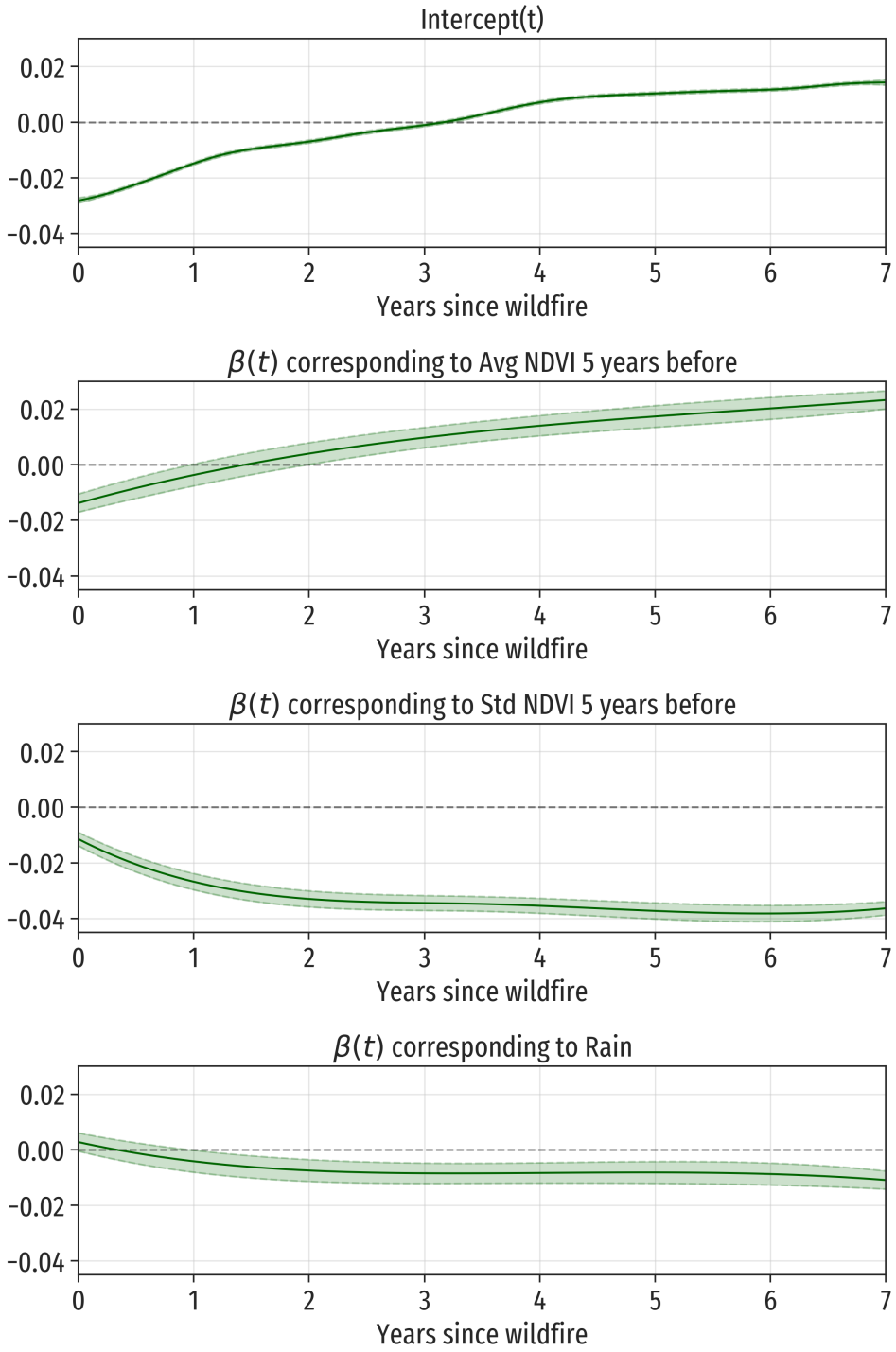

Figure 7. Full function-on-scalar additive model. Estimated functional coefficients of the form $\beta_j(t)$.

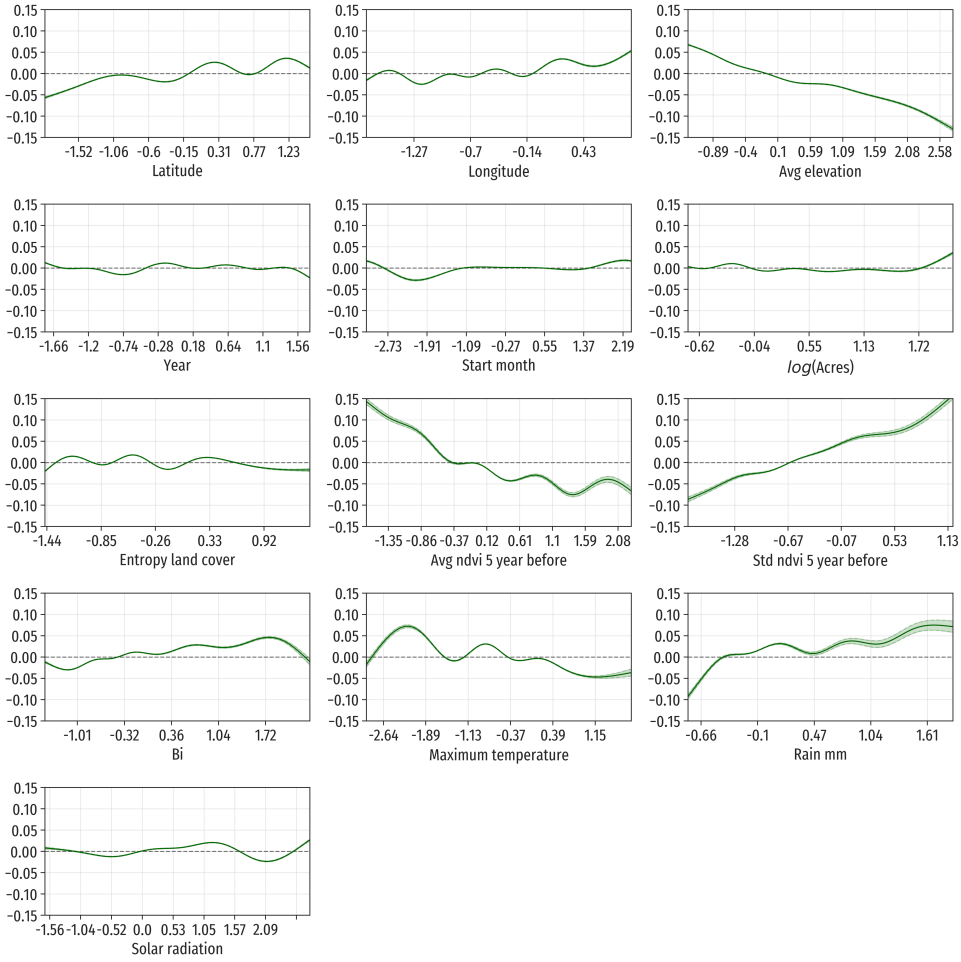

Figure 8. Full function-on-scalar additive model. Estimated smooth terms of the form $s_j(x_j)$.

We move our attention now to non-parametrically estimated terms, using the information contained in Table 4 and in Figure 7 (showing the estimation of the functional coefficients $\beta_j(t)$) and Figure 8 (which includes the estimations of the functions $s_j(x_j)$).

The estimation of the function $\beta_0(t)$ in model (2) is labeled `Intercept(t)` in Figure 7 (upper panel). Except for a vertical shift, it is approximately equal to the mean function (see Figure 5). The vertical shift should be equal to the estimated `Intercept` in Table 3 if there were no factor covariates in the model. In our case, however, this `Intercept` is referred to the level *Evergreen forest* of the factor Landcover.

There are three covariates (Avg NDVI 5 years before, Std NDVI 5 years before, and Rain) that contribute with two terms ($\beta_j(t)x_j$ and $s_j(x_j)$) to the full additive function-on-scalar model. To understand the contribution of these variables to the response recovery functions, we have to consider simultaneously the two corresponding estimated functions, where one is represented in Figure 7 and the other one in Figure 8. Regarding Avg NDVI 5 years before (average of NDVI over the 5 years before the wildfire), the estimation of its functional coefficient $\beta_j(t)$ (Figure 7, second panel) presents a monotonically increasing pattern with a total increment of 0.04 NDVI points over the 7 years. At the same time, the estimation of its term $s_j(x_j)$ (Figure 8, third row, second column) is a roughly decreasing function with a range of values of more than 0.20 NDVI points. So it follows that the contribution of the term $s_j(x_j)$ is much larger than that of the term $\beta_j(t)x_j$ for this explanatory variable. The nonparametric term $s_j(x_j)$ indicates that larger values of NDVI vegetation tend to suffer more from wildfires. For instance, in average, an area with pre-wildfire NDVI value equal to the mean plus one standard deviation loses 0.1 NDVI points more than another area with pre-wildfire NDVI value one standard deviation below the average. For

these two fictitious areas, the effect of the term $\beta_j(t)x_j$ is to add or subtract, respectively, the estimated coefficient $\beta_j(t)$. Then the area with NDVI values over the mean will have a larger decrease in NDVI the first one and a half yeas, but its recovery will be faster than in the area with previous lower NDVI values.

For Std NDVI 5 years before (standard deviation of NDVI over the 5 years before the wildfire), the relative relevance of the term $\beta_j(t)x_j$ is also much smaller than that of the term $s_j(x_j)$: their ranges are 0.03 and 0.25, respectively. The functional coefficient $\beta_j(t)$, negative for all t, is decreasing the first two years and almost constant from then on (with an approximate value of -0.04 NDVI points). The term $s_j(x_j)$ in this case is an increasing function on the standard deviation of pre-wildfire NDVI values, indicating that vegetation diversity (large values of Std NDVI 5 years before) is a protecting factor against wildfire effects. Combining both terms, the difference in loss of NDVI points between two areas with values of Std NDVI 5 years before one standard deviation over and below the mean, respectively, for t larger than two years is

$$\big(\beta_j(t) + s(1)\big) - \big(\beta_j(t) + s(1)\big) \approx (-0.04 + 0.10) - (0.04 - 0.03) = 0.05.$$

For t smaller than 2 years, the differences between these two areas are smaller than 0.05 and increasing in t.

For the explanatory variable Rain, the term $\beta_j(t)x_j$ is even less important than in the two previous cases (the range of $\beta_j(t)$ is smaller than 0.02 NDVI points, and it is almost constant from two years after the wildfire). On the other hand, the term $s_j(x_j)$, that has an approximate range of 0.17, grows rapidly at low values of the variable Rain (smaller than 0.3 times the standard deviation below the mean, approximately) and then it is almost constant or slightly increasing. We conclude that moderate or large precipitations seem to help recover or protect against the wildfire effects.

The remaining 10 explanatory variables contribute to the full additive function-on-scalar model only with a nonparametric term $s_j(x_j)$ that remains constant over time after the wildfire. The estimations of these terms are represented in Figure 8. The most relevant contribution to the model is that of the covariate Avg Elevation, which estimated term $s_j(x_j)$ has a range of 0.20 NDVI points. This function is decreasing in elevation, indicating that the wildfire effects are larger in more elevated areas, probably because elevated areas present in average richer vegetation (larger pre-wildfire NDVI values) than those with lower elevation.

Less important, although also worth mentioning, are the explanatory variables Bi (burning index) and maximum temperature. For the burning index, the estimated term $s_j(x_j)$ is an slightly increasing function in the middle part of the range of burning index values. It follows that areas with lower fire hazard will have slightly larger wildfire effects. The estimated term $s_j(x_j)$ for maximum temperature is roughly decreasing in its argument, indicating that low maximum temperatures protect moderately against the wildfire effects.

Regarding geographical coordinates contribution to the model, the wildfire effects in the South (respectively, West) are larger than in the North (respectively, East), but the differences are small (less than 0.1 NDVI points).

Lastly, we do not find clear and strong interpretable patterns of dependence between the response, the wildfire effects functions, and the rest of covariates (year, start month, log(acres), entropy land cover, and solar radiation).

The results we have shown indeed provide evidence that function on scalar regression models are a useful methodology to answer the research questions we stated in the introduction. Our findings show that wildfire effects are different depending on the kind of environment (e.g., average elevation of burned areas, average daily accumulated rain or average daily maximum temperatures), and that wildfire effects do vary according to the vegetation of the burned area (i.e., predominant landcover, type of vegetation or greenness of vegetation), as suggested by previous literature [21,47–51]. Our results also suggest that the vegetation response is influenced by the combination of previous conditions of vegetation and climatological characteristics, e.g., we see that environments

with vegetation having larger values of NDVI and less seasonal fluctuations are more severely affected. Moreover, the variance in different types of recoveries can be largely explained using pre-wildfire observable covariates, such as weather conditions, greenness, seasonality or location and elevation.

5. Conclusions and Discussion

The functional regression methodology has shown to be an effective way to study and explain vegetation recovery from wildfires, using pre-wildfire explanatory variables. The additive function-on-scalar fitted model explains 72.9% of the total variability of the responses. A large part of the explanatory power of the model goes directly to explain the recovery dynamic through the presence of regression coefficients that change over time. Nevertheless, the main part of the relationship between the explanatory variables and the wildfire effects functions is constant over time after the wildfire and, it is worth mentioning, non-linear.

The most important lessons we draw from this model are the following. On average, the recovery process after a wildfire is slow and takes more than 7 years (the time span used in this study). Each particular wildfire is a combination of a unique set of conditions that alter vegetation and ecosystems in a different manner, and it seems that all of them have an effect on the wildfire recovery process. The main risk conditions for a given area from suffering larger wildfire effects are, in this order, to have a rich and homogeneous vegetation (large and uniform NDVI, dominance of grassland, herbaceous vegetation or evergreen forest as land cover), to present a low precipitation regime, to have a large elevation over the sea level, to have low burning index, to have large maximum temperatures, and to be located in the South or West of California.

The convenience of studying outcomes changing over time, together with the estimation of the effect of several kinds of conditions pre- and post-wildfire, makes functional regression models to be a perfect methodology for this kind of studies. Previous studies use standard multiple regression models to compare absolute values of spectral indices, or comparisons of geolocated rasters such that these can include the spatial component of wildfires. However, giving estimates of the effect of these characteristics on the recovery pattern of vegetation from wildfires will allow environmental scientists and land management entities to study the characteristics that need more preservation.

It is important to notice that this methodology has only been implemented over the recoveries estimated from [18]. Nevertheless, this could be applied in many other research areas and fields, benefiting from the temporal component that this methodology includes, as everything is observed and measured over time. Expanding the study area to other fire-prone regions around the world, and increasing the time-span observed after wildfires (e.g., 15 years after each fire) would probably allow to observe full recoveries from wildfires. However, this remains outside the scope of this work.

This study tries to close the gap between satellite remote sensing and evaluation of wildfires' effects over time. It must be noted that gathering and pre-processing data, usually coming from different sources, is a crucial and highly sophisticated task when dealing with remote sensing data. Functional data analysis, and functional regression in particular, is an advanced statistical methodology well suited to analyze such rich data sets.

Author Contributions: Conceptualization, F.S.-B., P.D. and F.M.C.; methodology, F.S.-B. and P.D.; software, F.S.-B.; validation, F.S.-B., P.D. and F.M.C.; formal analysis, F.S.-B. and P.D.; investigation, F.S.-B. and P.D.; resources, F.S.-B., P.D. and F.M.C.; data curation, F.S.-B.; writing—original draft preparation, F.S.-B. and P.D.; writing—review and editing, F.S.-B. and P.D.; visualization, F.S.-B.; supervision, P.D. and F.M.C.; project administration, P.D. and F.M.C.; funding acquisition, P.D. and F.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: Serra-Burriel would like to thank the Barcelona Supercomputing Center for the Severo Ochoa Mobility Grant, and Delicado would like to thank the Spanish Ministerio de Ciencia e Innovación for the grant MTM2017-88142-P.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available at the GitHub repository: https://github.com/feliuserra/wildfires_vegetation_recovery_fda accessed on 6 April 2020.

Acknowledgments: The code used in this work has been performed using Python 3.8.1 [35] and R 3.6.2 [44] programming languages and the Google Earth Engine (GEE) platform [28].

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Spracklen, D.V.; Mickley, L.J.; Logan, J.A.; Hudman, R.C.; Yevich, R.; Flannigan, M.D.; Westerling, A.L. Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. *J. Geophys. Res.* **2009**, *114*, 1–17. [CrossRef]
2. Bryant, B.P.; Westerling, A.L. Scenarios for future wildfire risk in California: Links between changing demography, land use, climate, and wildfire. *Environmetrics* **2014**, *25*, 454–471. [CrossRef]
3. Westerling, A.L.; Bryant, B.P.; Preisler, H.K.; Holmes, T.P.; Hidalgo, H.G.; Das, T.; Shrestha, S.R. Climate change and growth scenarios for California wildfire. *Clim. Chang.* **2011**, *109*, 445–463. [CrossRef]
4. Westerling, A.L.R. Increasing western US forest wildfire activity: Sensitivity to changes in the timing of spring. *Philos. Trans. R. Soc. B Biol. Sci.* **2016**, *371*. [CrossRef]
5. Mitchell, J.W. Power line failures and catastrophic wildfires under extreme weather conditions. *Eng. Fail. Anal.* **2013**, *35*, 726–735. [CrossRef]
6. Keeley, J.E. Distribution of lightning and man-caused wildfires in California. In *Proceedings of the Symposium on Dynamics and Management of Mediterranean-Type Ecosystems*; USDA Forest Service: Berkeley, CA, USA, 1982; pp. 431–437.
7. Amatulli, G.; Peréz-Cabello, F.; de la Riva, J. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. *Ecol. Model.* **2007**, *200*, 321–333. [CrossRef]
8. Barrett, E.C.; Curtis, L.F. *Introduction to Environmental Remote Sensing*; Psychology Press: London, UK, 1999.
9. Scheffer, M.; Bascompte, J.; Brock, W.A.; Brovkin, V.; Carpenter, S.R.; Dakos, V.; Held, H.; Van Nes, E.H.; Rietkerk, M.; Sugihara, G. Early-warning signals for critical transitions. *Nature* **2009**, *461*, 53–59. [CrossRef]
10. Verbesselt, J.; Umlauf, N.; Hirota, M.; Holmgren, M.; Van Nes, E.H.; Herold, M.; Zeileis, A.; Scheffer, M. Remotely sensed resilience of tropical forests. *Nat. Clim. Chang.* **2016**, *6*, 1028–1031. [CrossRef]
11. Liu, Y.; Kumar, M.; Katul, G.G.; Porporato, A. Reduced resilience as an early warning signal of forest mortality. *Nat. Clim. Chang.* **2019**, *9*, 880–885. [CrossRef]
12. de Leeuw, J.; Georgiadou, Y.; Kerle, N.; de Gier, A.; Inoue, Y.; Ferwerda, J.; Smies, M.; Narantuya, D. The function of remote sensing in support of environmental policy. *Remote Sens.* **2010**, *2*, 1731–1750. [CrossRef]
13. Moritz, M.A.; Batllori, E.; Bradstock, R.A.; Gill, A.M.; Handmer, J.; Hessburg, P.F.; Leonard, J.; McCaffrey, S.; Odion, D.C.; Schoennagel, T.; et al. Learning to coexist with wildfire. *Nature* **2014**, *515*, 58–66. [CrossRef]
14. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2005.
15. Acar-Denizli, N.; Delicado, P.; Başarır, G.; Caballero, I. Functional regression on remote sensing data in oceanography. *Environ. Ecol. Stat.* **2018**, *25*, 277–304. [CrossRef]
16. Militino, A.F.; Ugarte, M.; Montesino, M. Filling missing data and smoothing altered data in satellite imagery with a spatial functional procedure. *Stoch. Environ. Res. Risk Assess.* **2019**, *33*, 1737–1750. [CrossRef]
17. Sugianto, S.; Rusdi, M. Functional Data Analysis: An Initiative Approach for Hyperspectral Data. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2019; Volume 1363, p. 012087.
18. Serra-Burriel, F.; Delicado, P.; Prata, A.T.; Cucchietti, F. Estimating heterogeneous wildfire effects using synthetic controls andsatellite remote sensing. *arXiv* **2020**, arXiv:2012.05140.
19. Engel, E.C.; Abella, S.R. Vegetation recovery in a desert landscape after wildfires: Influences of community type, time since fire and contingency effects. *J. Appl. Ecol.* **2011**, *48*, 1401–1410. [CrossRef]
20. Bright, B.C.; Hudak, A.T.; Kennedy, R.E.; Braaten, J.D.; Henareh Khalyani, A. Examining post-fire vegetation recovery with Landsat time series analysis in three western North American forest types. *Fire Ecol.* **2019**, *15*. [CrossRef]
21. Casady, G.M.; van Leeuwen, W.J.; Marsh, S.E. Evaluating Post-wildfire Vegetation Regeneration as a Response to Multiple Environmental Determinants. *Environ. Model. Assess.* **2010**, *15*, 295–307. [CrossRef]
22. Steiner, J.L.; Robertson, S.; Teet, S.; Wang, J.; Wu, X.; Zhou, Y.; Brown, D.; Xiao, X. Grassland Wildfires in the Southern Great Plains: Monitoring Ecological Impacts and Recovery. *Remote Sens.* **2020**, *12*, 619. [CrossRef]
23. Xu, Y. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Anal.* **2017**, *25*, 57–76. [CrossRef]
24. Eidenshink, J.; Schwind, B.; Brewer, K.; Zhu, Z.L.; Quayle, B.; Howard, S. A Project for Monitoring Trends in Burn Severity. *Fire Ecol.* **2007**, *3*, 3–21. [CrossRef]

25. Abatzoglou, J.T. Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.* **2013**, *33*, 121–131. [CrossRef]
26. Cuevas, A.; Febrero, M.; Fraiman, R. Robust estimation and classification for functional data via projection-based depth notions. *Comput. Stat.* **2007**, *22*, 481–496. [CrossRef]
27. Jordahl, K.; den Bossche, J.V.; Fleischmann, M.; Wasserman, J.; McBride, J.; Gerard, J.; Tratner, J.; Perry, M.; Badaracco, A.G.; Farmer, C.; et al. Geopandas/Geopandas: V0.8.1. 2020. Available online: https://zenodo.org/record/3946761 (accessed on 6 April 2020). [CrossRef]
28. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
29. Xu, Y.; Liu, L. gsynth: Generalized Synthetic Control Method. R Package Version 1.1.7. 2020. Available online: https://github.com/xuyiqing/gsynth (accessed on 6 April 2020).
30. National Wildfire Coordination Group. *Gaining an Understanding of the National Fire Danger Rating System*; NWCG Fire Danger Working Team; PMS 932 2002; National Wildfire Coordination Group: Boise, ID, USA, 2002.
31. Athey, S.; Bayati, M.; Doudchenko, N.; Imbens, G.; Khosravi, K. Matrix Completion Methods for Causal Panel Data Models. *arXiv* **2021**, arXiv:math.ST/1710.10251v4.
32. Alexandrov, T.; Bianconcini, S.; Dagum, E.B.; Maass, P.; McElroy, T.S. A Review of Some Modern Approaches to the Problem of Trend Extraction. *Econom. Rev.* **2012**, *31*, 593–624. [CrossRef]
33. Brockwell, P.J.; Brockwell, P.J.; Davis, R.A.; Davis, R.A. *Introduction to Time Series and Forecasting*, 3rd ed.; Springer: New York, NY, USA, 2016.
34. Cleveland, R.B.; Cleveland, W.S.; McRae, J.E.; Terpenning, I. STL: A seasonal-trend decomposition. *J. Off. Stat.* **1990**, *6*, 3–73.
35. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
36. Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010.
37. Horváth, L.; Kokoszka, P. *Inference for Functional Data with Applications*; Springer Science & Business Media: New York, NY, USA, 2012.
38. Kokoszka, P.; Reimherr, M. *Introduction to Functional Data Analysis*; CRC Press: Boca Raton, FL, USA, 2017.
39. Goldsmith, J.; Zipunnikov, V.; Schrack, J. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics* **2015**, *71*, 344–353. [CrossRef] [PubMed]
40. Chen, Y.; Goldsmith, J.; Ogden, R.T. Variable selection in function-on-scalar regression. *Stat* **2016**, *5*, 88–101. [CrossRef]
41. Scheipl, F.; Staicu, A.M.; Greven, S. Functional additive mixed models. *J. Comput. Graph. Stat.* **2015**, *24*, 477–501. [CrossRef]
42. Wood, S. *Generalized Additive Models: An Introduction with R*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2017.
43. Febrero-Bande, M.; de la Fuente, M.O. Statistical computing in functional data analysis: The R package fda.usc. *J. Stat. Softw.* **2012**, *51*. [CrossRef]
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
45. Ramsay, J.O.; Graves, S.; Hooker, G. FDA: Functional Data Analysis. R Package Version 5.1.4. 2020. Available online: https://cran.r-project.org/web/packages/fda/index.html (accessed on 8 January 2021).
46. Goldsmith, J.; Scheipl, F.; Huang, L.; Wrobel, J.; Di, C.; Gellar, J.; Harezlak, J.; McLean, M.W.; Swihart, B.; Xiao, L.; et al. Refund: Regression with Functional Data. R Package Version 0.1-23. 2020. Available online: https://cran.r-project.org/web/packages/refund/index.html (accessed on 8 January 2021).
47. Goetz, S.J.; Fiske, G.J.; Bunn, A.G. Using satellite time-series data sets to analyze fire disturbance and forest recovery across Canada. *Remote Sens. Environ.* **2006**, *101*, 352–365. [CrossRef]
48. Wittenberg, L.; Malkinson, D.; Beeri, O.; Halutzy, A.; Tesler, N. Spatial and temporal patterns of vegetation recovery following sequences of forest fires in a Mediterranean landscape, Mt. Carmel Israel. *Catena* **2007**, *71*, 76–83. [CrossRef]
49. Cuevas-González, M.; Gerard, F.; Balzter, H.; Riaño, D. Analysing forest recovery after wildfire disturbance in boreal Siberia using remotely sensed vegetation indices. *Glob. Chang. Biol.* **2009**, *15*, 561–577. [CrossRef]
50. Bolton, D.K.; Coops, N.C.; Wulder, M.A. Characterizing residual structure and forest recovery following high-severity fire in the western boreal of Canada using Landsat time-series and airborne lidar data. *Remote Sens. Environ.* **2015**, *163*, 48–60. [CrossRef]
51. Bartels, S.F.; Chen, H.Y.; Wulder, M.A.; White, J.C. Trends in post-disturbance recovery rates of Canada's forests following wildfire and harvest. *For. Ecol. Manag.* **2016**, *361*, 194–207. [CrossRef]