A functional data analysis of COVID-19 incidence and relations with mobility and sociodemographic factors

João Filipe Alfaia Subtil

Master Science Degree in Biomedical Engineering, Instituto Superior Técnico

joao.subtil@tecnico.ulisboa.pt

Abstract

Functional Data Analysis (FDA) is a statistical analysis tool that allows to analyse time-series data as functions. However, despite being successfully applied in several scientific domains, FDA is still rarely applied in epidemiology analysis. In this thesis, FDA is applied to analyse the associations of COVID-19 incidence with mobility and sociodemographic variables in Portugal mainland. The Concurrent Model is applied to analyse the association between a functional response variable (COVID-19 incidence) and a functional explanatory variable (Google Mobility). The Analysis of Variance Model is applied to assess the association between COVID-19 incidence functional data and scalar explanatory variables (Sociodemographic Variables). The results enabled to identify some relevant trends in functional data curve shapes. The strongest association was found between COVID-19 and Residential mobility, while Mobility in Retail and Public Transports also presented significant results. Mobility in Grocery Stores, Parks and Workplaces showed weak associations. In addition, the results strength the idea that the lag for mobility to have an effect on incidence is around 15 days, as referred in literature. Results also suggest that certain sociodemographic conditions influence the spread of COVID-19, such as income level, population's age-structure, density of schools, or prevalent sectors of activity. The techniques used here suggest FDA can be considered an additional tool for epidemiological analysis of COVID-19 incidence that can be replicated for mortality data or other disease or pandemics. The FDA is a broad area, so further analysis can be done using other FDA tools.

Keywords: Functional Data Analysis, COVID-19 Incidence, Google Mobility, Sociodemographic Variables, Analysis of Variance Model, Concurrent Model

1. INTRODUCTION

At the end of 2019, several cases of a contagious pneumonia were identified in Wuhan, a region of China. [1]. Despite the quarantine implemented in China, this pneumonia caused several outbreaks and spread to most of the countries around the world [2]. Laboratory analysis concluded that this pneumonia was caused by a novel coronavirus (CoV) named 2019-nCoV. The World Health Organization (WHO) named the disease Coronavirus Disease-2019 (COVID-19), and the International Committee on Taxonomy of Viruses (ICTV) named this novel coronavirus SARS-CoV-2. A characteristic of this virus is the existence of asymptomatic cases [1], and strategies to contain the virus are difficult to implement. In Portugal, mobility restrictive measures were used, including the closure of certain activities, teleworking, and mandatory lockdowns. The objective was to reduce the flow of people, reducing the probability of infection and consequently delaying the spread of the virus. [3], [4].

However, it is not always possible to know to what extent the restrictive measures applied prevent the spread of COVID-19. It is necessary to understand if these measures are necessary and effective. Therefore, open access COVID-19 data should be analysed, relate them to other variables that may have an influence on the spread of the disease, draw conclusions from the results obtained. Nowadays, several types of data are used to guide governments in health care planning. Statistical tools have been applied in diverse areas of science, with Multivariate Data Analysis (MDA) being the most used approach. However, this approach ignores the functional behaviour of the generating process that underlies the data. Functional Data Analysis (FDA) arises as an alternative methodology that has been increasingly applied, allowing to model time series as functional data. But what is FDA? In FDA, each record that comprises the functional data is called a functional datum [5]. [6] defines a functional datum as a set of measurements along a continuum that are considered as a single curve. Normally, this continuum is defined as being temporal.

Because this approach considers each curve as a single entity, possible correlations between repeated measurements are no longer a problem, one that persisted in MDA [7]. According to [7], the goal of FDA is to transform time series into a function that represents the entire data as a single observation. Then, concepts from MDA are applied to these data. FDA processes may involve several tools [6],[8], including Smoothing, Registration, Functional Principal Components Analysis (FPCA), Functional Principal Differential Analysis or Functional Linear Models.

2. LITERATURE REVIEW

FDA was used to explore the COVID-19 mortality in Italy, and its association with covariates such as mobility, and sociodemographic variables. The study found that mobility and positivity are associated with mortality and identified schools as having higher risk of contagion [8]. A study utilized FDA methods to investigate the COVID-19 spread in the United States, and the results demonstrate the effectiveness of stay-at-home orders [9]. FDA was applied to model COVID-19 trajectories in several countries and showed that a decrease on workplace mobility is correlated with reduced doubling rates (with a 2-week delay) [10].

The incidence of COVID-19 in Italy was analysed and was found that the measures implemented contributed to reduce the spread of the COVID-19 [11]. A susceptible-infected-recovered (SIR) model was applied to COVID-19 data to calculate the weekly transmission rate (β) and the association between mobility and these β values was analysed. According to the results distancing measures are effective in reducing the spread of the disease [12]. A study developed an interrupted time series study to assess the effectiveness of lockdown in reducing confirmed/death cases from COVID-19 in China. The results demonstrated that the social distancing measures had a positive impact in slowing the spread of the disease, and that the impact on incidence occurs between 7 to 17 days after the application of the measures [13]. A study described the association between transmission and mobility and found evidence that mobility patterns correlate with the intensity of transmission [4].

Other study intended to correlate the evolution of the COVID-19 in Portugal to its sociodemographic and demographic characteristics. It showed that the virus spreads from large urban areas to the surroundings. Also, it evidenced that elderly people in nursing homes constitute an extremely vulnerable part of the population, and immigrants have an increasing incidence [14]. A study developed the Area Deprivation Index (ADI), to rank neighbourhoods by their sociodemographic characteristics and evaluate their impact on the COVID-19 prevalence in the US. The results demonstrate that some neighbourhoods with higher ADI (more disadvantaged) presented higher COVID-19 prevalence [15]. In other study, the association of COVID-19 hospitalizations with racial and sociodemographic characteristics was analysed. It was observed an association between the hospitalization risk and Townsend Deprivation Index and income, and that Black and Asian people have a higher risk of hospitalization [16]. A retrospective cohort study was developed to analyse the correlation between the patient sociodemographics and COVID-19 health outcomes, and it was shown that neighbourhood disadvantage, which is closely associated with race, is a predictor of poor health outcomes [17].

3. OBJECTIVES

Literature review showed that FDA is a relatively recent technique in epidemiology fields, and more conventional methods are often used instead. Therefore, the objective is to use several FDA techniques such as smoothing, interpolation and functional linear models to analyse the association of COVID-19 incidence data with mobility and sociodemographic data. The results obtained in this work contribute to point out that FDA techniques can be considered an additional tool to more traditional epidemiological analysis contributing to provide new insights about the impact of potential risk factors on the spread of disease.

4. MATERIAL

Data used for analysis includes COVID-19 Data, Google Mobility Data and Sociodemographic Data. These data are provided in CSV (Comma Separated Value) or TXT (Text File) formats. Due to the specificities of FDA methods, these data had to go through pre-processing.

4.1 COVID-19 Incidence Data

COVID-19 related data are the focus of this work and consists of the daily notification COVID-19 cases by municipalities (n=278) in Continental Portugal. Data is provided by the Direção Geral da Saúde, and refers to the period between March 9, 2020 and February 6, 2021.

COVID-19 data refers to a set time-series with tabular structure where each column corresponds to municipality s, each row corresponds to day t, and each cell represent daily 7-day cumulative incidence rates (Table 1).

Date	m(s=1)	…	m(s=j)	…	m(s=M)
t = 1	rate(t=1, s=1)	…	rate(t=1, s=j)	…	rate(t=1, s=M)
…		…		…	
t = i	rate(t=i, s=1)	…	rate(t=i, s=j)	…	rate(t=i, s=M)
…		…		…	
t = T	rate(t=T, s=1)	…	rate(t=T, s=j)	…	rate(t=T, s=M)

Table 1- Data structure of incidence data used for functional data analysis

These data were processed as described in Section 5.1 before being correlated with the mobility time-series data.

4.2 GOOGLE Mobility Data

The Google data used in this work consist of daily mobility values in each of the 278 municipalities in Portugal mainland. These data is made available free by Google for public use and the version used here is related to the period of time between

March 15, 2020 and February 2, 2021. Most of the information detailed in this section is provided by [18].

The data show movement trends by region, across different categories of places. For each category in a region, reports show the changes by comparing mobility for the report date to the baseline day. Here, the baseline day is the median value from the 5-week period Jan 3 – Feb 6, 2020, before widespread COVID-19 disruption in Europe. For each region-category, the baseline is not a single value—but 7 individual values, one for each week day. A number is calculated for the report date and reported as a positive or negative percentage. It shows how visits and length of stay at different places change compared to a baseline. The mobility data are retrieved from anonymous mobile device location information from Android users [19] and aggregated by municipality. When data doesn't meet quality and/or privacy thresholds, a missing data occurs. 6 different mobility categories are analysed: *Grocery; Parks; Stations; Retail; Residential; Workplaces.*

Mobility data consists of time-series with mobility values in tabular format where each column is a municipality, each row corresponds to day t and cell values represent daily mobility values.

Date	m(s=1)	…	m(s=j)	…	m(s=M)
t = 1	mob(t=1,s=1)	…	mob(t=1,s=j)	…	mob(t=1,s=M)
…		…		…	
t = i	mob(t=i,s=1)	…	mob(t=i,s=j)	…	mob(t=i,s=M)
…		…		…	
t = Tm	mob(t=Tm,s=1)	…	mob(t=Tm,s=j)	…	mob(t=Tm,s=M)

Table 2- Data structure of mobile data used for functional data analysis

For confidentiality reasons some data are omitted. In these cases, an imputation method was used to fill-in the missing values. For analysis of relationships with COVID-19 data, a further data-processing step was conducted, which is described in Section 5.1.

4.3 Sociodemographic Data

The sociodemographic data analysed covers each of the 278 municipalities in Portugal mainland. Unlike COVID-19 or GOOGLE data, sociodemographic data are not time-series, but rather a single value for each municipality (annual statistics, from which the most recent year available was used).

These data were collected from Instituto Nacional de Estatística (National Statistics) and PORDATA (Fundação Francisco Manuel Dos Santos, Open Data provider Fundação). The Deprivation Index, which is a relative measure of poverty. computed using an European standardized approach [20]., was provided by Ribeiro [21], Sociodemographic data were transformed into population proportions or population densities to reduce the impact of different population sizes and allow a comparative analysis between municipalities.

Sociodemographic data was structured in tabular format, where rows refers to municipalities and columns to sociodemographic variables (Table 3).

Variable	Description	Abbreviation
Population Density in Urban Areas	Inhabitants / km^2	PD
Deprivation Index	1 to 5	DI
Youth Population	% 0-19 years	YP
Elderly Population	% 65+ years	EP
Working Population in Primary Sector	% Working Pop. In Primary Sector	PS
Working Population in Secondary Sector	% Working Pop. in Secondary Sector	SS
Working Population in Tertiary Sector	% Working Pop. in Tertiary Sector	TS
Guaranteed Minimum Income	Proportion of Guaranteed Minimum Income beneficiaries	GMI
Schools Density	Schools / Km^2	SD

Table 3 - Sociodemographic Variables Description

To use sociodemographic variables with FDA techniques, it was necessary to transform these values (proportions or densities) into categorical ones by grouping municipalities into tertiles (except for one variable, which already grouped municipalities into quintiles), to be used as dummy variables.

5. METHODOLOGY

5.1 Pre-Processing

5.1.1 Stationarity

For this work, time-series are transformed into stationary series. This transformation is only carried out for modelling the association between COVID-19 and mobility. Stationary time-series are time-series whose properties do not depend on the time at which they are observed, and they have means, variances, and covariances that don’t change over time [22], [23]. Non-stationary data, generally, are more complicated to be modelled [24].

The first step of this transformation is to submit the data to a log transformation. However, the data in question contains “zero” values. Instead, a two-parameter version of the Box-Cox transformation was used, which allows for a shift before the data is transformed:

$$g(y;\lambda_1,\lambda_2)=\begin{cases}\dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & when\ \lambda_1\neq 0\\ \log(y+\lambda_2) & when\ \lambda_1=0\end{cases}\qquad (1)$$

Parameters λ_1 and λ_2 are estimated using an R function called *boxcoxfit*.

The second step of the transformation into stationary series involves the application of differencing. COVID data were subjected to the calculation of second differences, and mobility data from GOOGLE subjected to the calculation of first differences, with a 7-day time lag:

$$1^{st}\ difference\text{: } \Delta_m z_t=(1-B)(1-B^m)z_t \qquad (2)$$

$$2^{nd}\ difference\text{: } \Delta^2{}_m z_t=\ (1-B)^2(1-B^m)z_t \qquad (3)$$

z_t is an observation of the time series at time t, m is the time lag, and Δz_t, $\Delta^2 z_t$ represent the first and second differences of z_t. In the case of COVID-19 incidence, the first differences correspond to the acceleration in the incidence rate. The second differences of Google Mobility data correspond to the velocity of mobility variation.

5.1.2 Imputation (predictive mean matching)

As said before, Google mobility data contains cells with missing data. Here, an imputation method called predictive mean matching is applied, using the package *mice* [25]. It allows for discrete variables, and is based on real values, providing reliability to the estimation. One of the biggest advantages of this method is that it is less vulnerable to model misspecification, since the model is implicit in the data itself. For each missing entry, a set of candidate donors is created from all complete cases that have predicted values closest to the predicted value for the missing entry. One donor replaces the missing value. [25].

5.1.4 Data Analysis per Wave

The evolution of the incidence of COVID-19 in Portugal during the analysed period is characterized by the existence of 3 pandemic waves. Considering this, the time series that constitute the COVID-19 and Google Mobility data were divided in 3 parts, allowing FDA to be applied not only to the entire pandemic period, but also to each of the waves individually (or combining consecutive waves). The individual study of each wave reduces the complexity of the results and its analysis. The division of the 3 waves was done as follows:

- 1st : March 9, 2020 to July 31, 2020
- 2nd: October 24, 2020 to December 6, 2020
- 3rd : December 14, 2020 to February 6, 2020

5.2 FDA

The first step of the FDA methodology is to build functions from the data. The function construction process used in this thesis is based on the structure proposed in [26]. In this first step the goal is to transform time series data into functional data.

The process of constructing a function is based on the following expression

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \boldsymbol{c}'\boldsymbol{\phi}(t) \qquad (4)$$

To build a function x(t) from the data it is essential to define:

- the functions ϕ_k, called basis functions
- the coefficients c_k, to construct the function $x(t)$ as a linear combination of these coefficients with the basis functions ϕ_k

5.2.1 Basis Functions

The COVID-19 and mobility data is unpredictable as they reflect the complex response of populations under a pandemic situation, and the task of transforming these data into functions that allow an accurate analysis is not an easy one. It is necessary to use tools that allow the construction of curves from any type of data, without giving up an adequate level of efficiency from a computational point of view. The basis functions ϕ_k work as a set of functional building blocks. These functions, of which there are several types, are linearly combined with coefficients, in order to estimate the intended function.

The expression (4) defines the construction of any function $x(t)$, and is called *basis function expansion*:

The parameters $c_1, c_2, \dots, c_k$ are the coefficients of the basis function expansion. In the expression $\boldsymbol{c}'\boldsymbol{\phi}(t)$, c refers to the vector of *K* coefficients and $\boldsymbol{\phi}(t)$ is a vector of length *K* that contains the basis functions.

In this work, a dataset of N time series, corresponding to the N municipalities of Portugal mainland, are transformed into N functions. So N functions are required, and the expression (4) is replaced by

$$x_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), i = 1, \dots, N \qquad (5)$$

and in this case matrix notation for (4) becomes

$$\mathbf{x}(t) = \mathbf{C}\boldsymbol{\phi}(t) \qquad (6)$$

where $\mathbf{x}(t)$ is the vector that contains the N functions $x_i(t)$, and the coefficient matrix $\mathbf{C}$ contains all the coefficients. In this work, the coefficient matrix $\mathbf{C}$ is a matrix with N rows (one row per function $x_i(t)$) and K columns (one column per basis function).

There are several types of basis systems, and some of them include Spline series and Fourier series, that solve most of the problems in FDA. In this work Spline Series were selected to construct the basis functions.

Spline Series

Splines are polynomials, able to accommodate flexible basis functions, allowing therefore to estimate any curve feature. In this research, B-Splines are applied to model time-series data, allowing to estimate smoothed curve features reflecting the complex response of populations under the pandemic situation.

Break Points and Knots, Order and Degree

When building a spline basis system, the data is divided into subintervals throughout its observation interval. This division is carried out through the application of break points between the intervals. In the break points, knots are placed, and each break point has at least one knot.

Spline functions that define the entire basis system are polynomials with a certain degree or order. The degree corresponds to the highest power of the polynomial, and the order corresponds to $degree + 1$. The purpose of the knots is to define, at each break point, the number of matching derivatives between neighbouring polynomials. Typically, each break point contains only one knot, and the number of matching derivatives is $order - 2$. Thus, any basis system with order greater than 2 will have at least one derivative matching (1st derivative), and the function will have smooth continuous behaviour at all break points.

In conclusion, spline basis systems are defined by the break points, the sequence of knots, and the degree or order of the polynomials. Increasing the order/degree of the spline basis functions allows us to get better fits. Roughness penalties (described below) are included to avoid overfitting Finally, the number K of basis functions in this basis system is determined by the relation:

$$\begin{aligned} &number\ of\ basis\ functions \\ &= order\ +\ number\ of\ interior\ knots \end{aligned} \tag{7}$$

Interior knots are the knots placed at break points which are not either at the start or end of the interval that defines the function. Here, the break points will be spaced 7 days apart, with an internal knot for each break point and polynomials of order 6 in the case of COVID-19 data and order 5 in the case of Google Mobility data.

5.2.3 Regression Splines: Smoothing by Regression Analysis

After building the spline basis system, the next step is to determine the coefficients. For this purpose, the regression analysis methodology is used, which is based on the minimization of the sum of squared errors. Data fitting is defined as the minimization of the sum of squared errors or residuals

$$SSE(x) = \sum_{j}^{n} [y_j - x(t_j)]^2 \tag{8}$$

When the basis function expansion (4) is used to define function x, the minimization problem described above becomes

$$\begin{aligned} SSE(\mathbf{c}) &= \sum_{j}^{n} \left[y_j - \sum_{k}^{K} c_k \phi_k(t_j) \right]^2 \\ &= \sum_{j}^{n} [y_j - \phi(t_j)'\mathbf{c}]^2 \end{aligned} \tag{9}$$

This approach is driven by the error model, which states that

$$y_j = x(t_j) + \varepsilon_j = \mathbf{c}'\phi(t) + \varepsilon_j = \phi'(t_j)\mathbf{c} + \varepsilon_j \tag{10}$$

where the true errors or residuals ε_j are statistically independent and have a normal or Gaussian distribution with mean 0 and constant variance. Using matrix notation, let the n-vector y contain the n values to be fit, vector ε contain the corresponding true residual values, and n by k matrix Φ contain the basis function values $c_k\phi_k(t_j)$. y is defined as follows

$$\mathrm{y} = \Phi\mathbf{c} + \varepsilon \tag{11}$$

and the least-squares estimate of the coefficient vector $\mathbf{c}$ is

$$\hat{\mathbf{c}} = (\Phi'\Phi)^{-1}\Phi'\mathbf{y} \tag{12}$$

The coefficient estimate $\hat{\mathbf{c}}$ in (12) is calculated y by multiplying the vector it by a matrix designed y2cMap. This matrix is often used to determine the variability in quantities determined by $\hat{\mathbf{c}}$, and is defined as follows:

$$\mathrm{y2cMap} = (\Phi'\Phi)^{-1}\Phi \text{ so that } \hat{\mathbf{c}} = \mathrm{y2cMap}\,\mathbf{y} \tag{13}$$

For the regression splines method to estimate functions and smooth data to work, it is necessary that the number K of basis functions be considerably smaller than the number of observations that consist of the data. One of the consequences of using a high number of basis functions is the occurrence of overfitting, which generates fewer smooth curves, making their analysis difficult, especially if it is necessary to analyse derivatives of these same curves.

Data Smoothing with Roughness Penalties

The goal of data smoothing using roughness penalties is to impose smoothness in a created function by penalizing some measure of function complexity.

Choosing a Roughness Penalty

The square of the second derivative $[D^2x(t)]^2$ is called the curvature of the function x at argument value t. [26] defines the function's roughness as its integrated squared second derivative or its total curvature

$$PEN_2(x) = \int [D^2x(t)]^2\, dt \tag{14}$$

$PEN_2(x)$ provides smoothing because if the function is highly variable, that is, it has too much curvature, the square of the second derivative $[D^2x(t)]^2$ is substantial.

Having defined the measure of the roughness of the fitted curves, the goal is to minimize a fitting criterion. Whatever roughness penalty used, a multiple of it is added to the error sum of squares. Using some differential operator L to define roughness, the fitting criterion will be:

$$F(\mathbf{c}) = \sum_{j} [y_j - x(t_j)]^2 + \lambda \int [Lx(t)]^2 dt \tag{15}$$

where $x(t) = \mathbf{c}'\phi(t)$.

The smoothing parameter λ controls the importance placed on the roughness penalty.

The Roughness Penalty Matrix R

This methodology needs to be adapted to the roughness penalty smoothing, providing a new way to estimate the coefficient vector $\hat{\mathbf{c}}$.

The roughness penalized fitting criterion (15) is generally defined as

$$F(\mathbf{c}) = \sum_j \left[y_j - x(t_j)\right]^2 + \lambda \int [Lx(t)]^2 dt \tag{16}$$

By substituting the basis expansion $x(t) = \mathbf{c}'\phi(t) = \phi'(t)\mathbf{c}$ into the equation above, results in

$$F(\mathbf{c}) = \sum_j \left[y_j - \phi'(t_j)\mathbf{c}\right]^2 + \lambda \mathbf{c}' \left[\int L\phi(t)L\phi'(t)dt\right]\mathbf{c} \tag{17}$$

The order K roughness penalty matrix is

$$\mathbf{R} = \int \phi(t)\phi'(t)dt \tag{18}$$

From this coefficient vector $\hat{\mathbf{c}}$ is defined as

$$\hat{\mathbf{c}} = (\Phi'\Phi + \lambda\mathbf{R})^{-1}\Phi'\mathbf{y} \tag{19}$$

As before, the matrix $\mathrm{y2cMap}$ is defined. It is used for computing confidence regions and is obtained through the following expression

$$\mathrm{y2cMap} = (\Phi'\Phi + \lambda\mathbf{R})^{-1}\Phi' \tag{20}$$

5.3 Linear Models

Linear modelling models the relationship between a response variable and one or more explanatory variables. In this work, the response variables are COVID-19 incidence curves, and explanatory variables are mobility data from GOOGLE or sociodemographic data.

5.3.1 Functional Responses with Scalar Covariates: Analysis of Variance Model

Here, variation in a functional response (COVID-19 daily incidence curves) is decomposed into functional effects through the use of a scalar design matrix $\mathbf{Z}$ (the covariates, sociodemographic data, are scalar).

Sociodemographic Variables Effects on COVID-19 Incidence

Sociodemographic data were divided into distinct groups. For example, in the case of the elderly population, municipalities were split into 3 groups, in which the first group contains the municipalities in the lower tertile of elderly population (youngest municipalities) and the third group contains the municipalities in the upper tertile (oldest populations). The model fitted is the following:

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{3} x_{ij}\beta_j(t) + \varepsilon_i(t) \tag{21}$$

where $y_i(t)$ is a *functional response*. Using this technique requires to add a constraint to identify the effects of the three sociodemographic groups, which is defined as follows

$$\textstyle\sum_{j=1}^{3}\beta_j(t) = 0 \text{ for all } t \tag{22}$$

In order to implement this constraint, a new row is added to the original data as an additional 279*th* "observation" for which $y_{279}(t) = 0$. The intercept term $\beta_0(t)$ is the COVID-19 mean incidence curve, and each of the other linear coefficients is the perturbation of the COVID-19 incidence mean required to fit a group's mean COVID-19 incidence curve. The R function called here is *fRegress*, that performs linear regression. The coefficients are extracted and plotted, along with the predicted curves for each group.

5.3.2 Functional Responses with Functional Covariates: Concurrent Model

For functional covariates, the model (21) can expand it as follows:

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{q-1} x_{ij}(t)\beta_j(t) + \varepsilon_i(t) \tag{23}$$

where $x_{ij}(t)$ is a functional observation. The model (23) is called the *concurrent model* and is so designated because the value of $y_i(t)$ is related to the value of $x_{ij}(t)$ only at the same time points t. $y_i(t)$ represents the functional response, the COVID-19 incidence curves. $x_{ij}(t)$ represents the functional covariate, the Google mobility curves. $\beta_0(t)$, the intercept function, captures the variation in the response that does not depend on any of the other covariate functions.

Estimation for the Concurrent Model

It is important to understand how the functional linear coefficients β_j are estimated, using the R function fRegress, by simplifying the problem and transforming it into the resolution of a group of linear equations. Consider that the N by q functional matrix $\mathbf{Z}$ contain the x_{ij} functions, and that the vector coefficient function β of length q contain each of the regression functions. The concurrent model in matrix notation is then

$$\mathbf{y}(t) = \mathbf{Z}(t)\beta(t) + \varepsilon(t) \tag{24}$$

where $\mathbf{y}$ is a functional vector of length N that contains the response functions. Let

$$\mathbf{r}(t) = \mathbf{y}(t) - \mathbf{Z}(t)\beta(t) \tag{25}$$

be the corresponding N-vector of residual functions. The weighted regularized fitting criterion is

$$\mathrm{LMSSE}(\beta) = \int \mathbf{r}(t)'\mathbf{r}(t)dt + \sum_j^p \lambda_j \int \left[L_j\beta_j(t)\right]^2 dt \tag{26}$$

Consider now that the coefficient function β_j have the expansion

$$\beta_j(t) = \sum_k^{K_j} b_{kj}\theta_{kj}(t) = \theta_j(t)'\mathbf{b}_j \tag{27}$$

in terms of K_j basis functions θ_{kj}. In order to express (24) and (26) in matrix notation referring explicitly to these expansions, a composite is constructed. After that, (26) can be defined as:

$$\mathrm{LMSSE}(\beta) = \int [\mathbf{y}(t)'\mathbf{y}(t) - 2\mathbf{b}'\Theta(t)'\mathbf{Z}(t)'\mathbf{y}(t) + \mathbf{b}'\Theta(t)'\mathbf{Z}(t)'\mathbf{Z}(t)\Theta(t)\mathbf{b}]dt + \mathbf{b}'\mathbf{R}(\lambda)\mathbf{b} \tag{28}$$

By differentiating this function with respect to the coefficient vector $\mathbf{b}$ and set it to zero, the normal equations penalized least squares solution for the composite coefficient vector $\hat{\mathbf{b}}$ is obtained:

$$\left[\int \Theta(t)'\mathbf{Z}(t)'\mathbf{Z}(t)\Theta(t)dt + \mathbf{R}(\lambda)\right]\hat{\mathbf{b}} = \left[\int \Theta(t)'\mathbf{Z}(t)'\mathbf{y}(t)dt\right] \tag{29}$$

Confidence Intervals for Regression Functions

When regression functions are fitted, confidence intervals are important to assess the quality of the estimates made. The 95% pointwise confidence intervals are generated using the function R *fRegress.stderr,* from fda package. The linear coefficients with associated confidence intervals can be plotted using the function *plotbeta* from the same package

5.3.3 Functional Responses with Functional Covariates: General Concurrent Model

A general version of the concurrent model is usually referred to as General Concurrent Model and has the following functional form

$$y_i(t) = \beta_0(t) + \int_{\Omega_t} \beta_1(t,s)x_i(s)ds + \varepsilon_i(t) \qquad (30)$$

Now, the linear coefficient function $\beta_1(t,s)$ defines the dependence of $y_i(t)$ on covariate $x_i(s)$ at each time t. In this case, it is not necessary for $x_i(s)$ and $y_i(t)$ to be defined over the same range or continuum. The set Ω_t to which the integration in (30) is calculated, comprises the range of values of argument s over which x_i is considered to influence response y_i at time t.

A Functional Linear Model for the COVID-19 Incidence and GOOGLE Mobility Data

Consider the following functional linear model where $x_i(t)$ represent daily GOOGLE mobility value in day i and $y_i(t)$ represent daily COVID-19 incidence value in day i:

$$y_i(t) = \beta_0(t) + \int \beta_1(s,t)x_i(t)ds + \varepsilon_i(t) \qquad (31)$$

For any day within the period of analysis, COVID-19 incidence is modelled as a linear combination of functional covariate GOOGLE mobility data in previous days. The regression function β has the basis function expansion

$$\beta_1(s,t) = \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} b_{kl}\phi_k(s)\psi_l(t) = \phi'(s)\mathbf{B}\psi(t) \qquad (32)$$

where the coefficients for the expansion, b_{kl} are in the K_1 by K_2 matrix $\mathbf{B}$. This requires to define two bases for β_1, as well as a basis for the intercept function β_0. For a bivariate function such as $\beta_1(t,s)$ smoothness can be imposed by penalizing the s and t directions separately:

$$\mathrm{PEN}_{\lambda_t,\lambda_s}\big(\beta_1(t,s)\big) = \lambda_1[L_t\beta_1(t,s)]^2 ds\, dt + \lambda_2[L_s\beta_1(t,s)]^2 ds\, dt \qquad (33)$$

where linear differential operator L_s only involves derivatives with respect to s and L_t only involves derivatives with respect to t. A B-spline basis is used to define functional parameter objects for β_0 , $\beta_1(\cdot,t)$ and $\beta_1(s,\cdot)$. The coefficients are smoothed, but the smoothing parameter values vary. These three functional parameter objects are placed into a list object to be supplied to function *linmod*, a function that returns the coefficients to be analysed.

6. RESULTS AND DISCUSSION

6.1 Functional Responses with Functional Covariates: Concurrent Model

The relationship between COVID-19 incidence rate acceleration curves and the velocity of mobility variation was explored using a concurrent model. Different time lags of analysis were explored allowing to estimate the time interval that elapses between a certain behaviour of the mobility curves and a (possibly) similar behaviour in the incidence curves. A lag of 15 to 16 days between the curves of both variables showed stronger associations and was set for analysis of results (the exact number of days lagged, varied according to the mobility class analysed). In 1st and 2nd wave none of mobility classes show a significant relationship with COVID-19 incidence curves. This may be due to COVID-19 data only starting at the end of March and the underreporting of positive cases. In fact linear coefficient functions in these two first waves show erratic curves with wide 95% point-wise confidence intervals that include 0 value.

In some mobility classes, like Grocery (15-day delay), Parks (16-day delay), and Stations (15-day delay), coefficients functions are more stable and positive, but still not significative, as the confidence intervals of these curves include zero value.

In 3rd wave, significant associations were found with some mobility classes. The most significant result is found in the Residential class (16-day delay) (Figure 1). where linear coefficient function is always positive and the confidence interval fails to include zero.

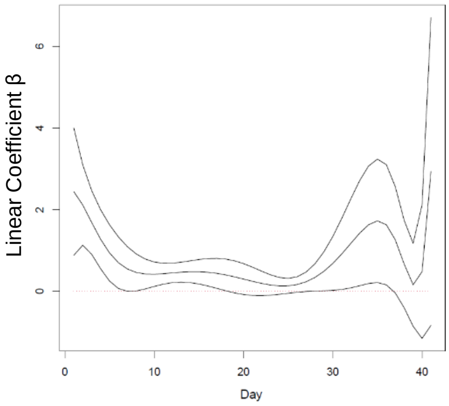

Figure 1 - Linear Coefficient Function for the Velocity of Mobility Variation with 95% pointwise confidence intervals (Residential and Retail Class, 3rd wave, 16-day lag)

Under the same concurrent model, a significant linear coefficient function was also identified in the Retail class (15-day delay) but in this case, confidence intervals include zero in one section (results not shown here).

The Workplace mobility class showed no relationship with COVID-19 incidence evolution.

It is noteworthy that outside the 15-16 days range, the results were not significant or erratic, strengthening the idea that the ideal lag (for which mobility influences incidence) may be about 15 days. A detailed analysis of the results obtained for each of the mobility classes is performed in the next section, since results were in line with this section.

6.2 Functional Responses with Functional Covariates: General Concurrent Model

This method is similar to the previous and allows to evaluate possible associations without having to resort to lagged curves and relates the value of the response variable to all the values of the explanatory variable. Suppose that the effect of mobility on incidence has a lag of about 15 days (a value that is referred in some literature, and that is also found in results of this work).

Based on this assumption it could be reasonable to expect that a graph representing this signal with 15-day lag would show a diagonal region, between the origin of the graph and the upper right corner, in which the association was positive, and shifted to the left about 15 days. However, results obtained were not that linear and this is not surprising as mobility and COVID-19 incidence variations are unpredictable as they reflect the complex situation of populations responding differently to external stimuli, under a pandemic situation.

Nevertheless, using General Concurrent Models it was possible to detect pattern or regions of positive associations. Despite not being a perfect diagonal, these regions show a lag between the velocity of mobility variation curves and the respective effect on the incidence rate acceleration curves.

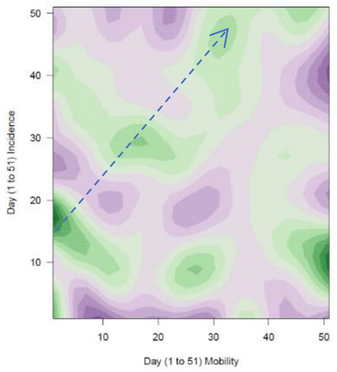

Figure 2 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Residential Class, 3rd wave). The dashed line illustrates the diagonal path with 15-day lag showing prevalence of positive linear coefficient functions.

Once again, only results in 3rd wave were significant so only results for the 3rd wave will be presented and commented below.

It was in Residential Mobility that the strongest signals were found. The 3rd wave graph (Figure 2) shows that the linear coefficient is positive along the diagonal lagged by about 15 days (the diagonal starts at (0,15) and ends at (35,50)). Retail and Stations mobility classes also showed relationship with incidence, despite the signals being weaker. Retail mobility class is related to a type of mobility aimed at obtaining non-essential goods (for example, going to shopping centres). Therefore, it is expected that the use of these spaces varied significantly according to the relief/increase of restrictions and may reflect a more relaxed/careful behaviour of the population. As the 3rd wave is related to Christmas, this strong initial signal may be related to the large agglomerations that occur at this time in commercial surfaces. This behaviour in retail mobility may therefore influence the spread of COVID-19. The signal for the Stations mobility class in the 3rd wave, may show the influence that the use of public transport has on the spread of the virus. Thus, the strongest signal in the 3rd wave could be related to the Christmas and New Year's Eve season (fewer people working) and the application of mandatory teleworking and lockdown during the 3rd wave.

Also, for the other mobility classes, such as Grocery or Parks, positive linear bivariate coefficients functions occurred but only appear in some regions along the "diagonal" (results not shown here). One hypothesis for results obtained in these classes is that mobility of the population in these classes is less sensitive to variations in incidence (compared to other classes). Grocery class is related with mobility in supermarkets and pharmacies. In these places populations obtain essential goods, and have to comply with strict protection measures, such as mask usage. On other hand, Parks mobility class are usually associated with good weather. As the 3rd wave corresponds to a winter period, parks mobility is lower so it is not expected that restriction measures would cause a significant variation in the use of these spaces. Furthermore, parks are outdoor places, where virus transmission is much lower, which may explain the weakening of the signal in the analysed waves.

Results of General Concurrent Model show that residential mobility, a sensitive variable to lockdown measures, has a strong relationship with the COVID-19 incidence rate acceleration (with a lag of around 15 days), compared to all other mobility classes. The 3rd wave coincided with the Christmas and New Year celebrations, with a large movement of people away from their homes at Christmas, and the application of lockdown measures to combat these movements in the New Year's Eve. Thus, this might have influenced the behaviour of the incidence curves. Also, the 2nd wave is related to a period of school activity and greater relaxation of measures, reducing residential mobility and increasing interaction outside households, which may also influence the spread of the virus

Finally, in the Workplace class, it was found a positive linear coefficient in the 3rd wave in the diagonal, but is a very weak signal. Fluctuations in the frequency of workplaces during the pandemic period did not have a major influence on the spread of the virus. The already widespread use of telework, and the improvements in safety conditions may have contributed to this type of mobility having no impact on the incidence rate acceleration curves.

6.3 Functional Responses with Scalar Covariates: Analysis of Variance Model

The Analysis of Variance Model was applied to model the relationship between COVID-19 daily incidence rates and sociodemographic potential risk factors, between October 24, 2020 and February 6, 2020 (only the 1st and 2nd waves were analysed).

The objective of this analysis was to understand the impact of different sociodemographic classes on the shape of the incidence curves. As a previous step, it was required to transform the values (quantitative data) of sociodemographic variables into classes (categorical data). The models were fitted using dummy variables for different categories of each variable and the regression coefficients for each category and variable were estimated as illustrated in Figure 3.

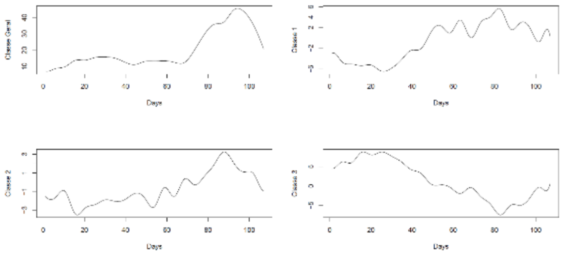

Figure 3 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Population Density

For each variable, municipalities are divided into classes or categories, and each class corresponds to a plot, whose curves represent the perturbation of the COVID-19 incidence mean required to fit the class mean COVID-19 incidence curve. This mean that in the period when a curve is positive/negative, the municipalities in the same class had a COVID-19 incidence mean higher/lower than the total average.

In Figure 3, the first plot (upper left corner) corresponds to the COVID-19 mean incidence curve in the period analysed, and the other three correspond to the impacts of municipalities with lower population density (Class 1), average density (Class 2) and higher density (Class 3). It is possible to identify from the plots that Class 1 had an higher impact on COVID-19 curves during the 3rd wave and the municipalities with higher population density (Class 3) were the most affected in the 2nd wave.

This analysis was performed with all the other variables. However, the plots are not shown here for sake of simplicity. Elderly population (EP) showed that municipalities with lower percentage of EP (Class 1, younger municipalities) were the most affected in the 2nd wave and those with higher percentage of EP (Class 3) were the most affected in the 3rd wave. As expected the opposite trends occurred in the analysis of Young Population (YP) as municipalities with lower percentage of YP (older municipalities) were the most affected in the 3rd wave and municipalities with higher percentage of YP were most affected in the 2nd wave. At this stage it is important to pinpoint that the linear coefficients curve shapes of PD, YP and EP can be explained in light of the behaviour of the population in the 3rd wave. During the month of December many moved from urban areas to non-urban areas and inland, due to Christmas. The municipalities in these areas have high percentage of EP and low PD. These movements, that preceded the 3rd wave, may explain the exponential increase in COVID-19 cases.

Turning now to schools density (SD), the results show us a behaviour identical to PD. Here, Class 1 corresponds to municipalities with the lowest SD, and Class 3 to the ones with the highest SD. This show that the behaviour of the incidence curves may also be related to school activity, as the return to in-person classes before the 2nd may have contributed to the spread of COVID-19.

In the case of the Deprivation Index (DI) results provided a counter intuitive result as the group of municipalities classified as having a lower degree of deprivation had an incidence mean higher than the country average. Because DI is a variable composed of many variables, it may be subject to higher variability, weakening the signal. In addition, the DI is based on the 2011 Census, so the data may be outdated.

Results concerning the Guaranteed Minimum Income (GMI) variable show that municipalities with higher proportion GMI in population (poorer on average) had a higher COVID-19 incidence mean than the country average in the 2nd wave and part of the 3rd wave. These results are in line with literature, as the lack of financial resources hinders the population's access to housing, health care, education, etc., making it more vulnerable to COVID-19.

When analysing the coefficients for the percentage of Working Population per activity sector, municipalities were classified into municipalities where primary activities, like agriculture or fishing are predominant (Class 1), secondary activities like industry are predominant (SS, Class 2) and services like banking or education are predominant (TS, Class 3). The results show that the Class 2 municipalities showed an above-average behaviour over the 2nd and 3rd wave, and in Class 1 municipalities, the behaviour of the incidence curves was below average. The SS is the activity sector that involves manufacturing, essential activity that involve work in closed spaces, contributing to the spread of COVID-19. The PS includes many activities that are not carried out in closed spaces. The TS, the services sector, does not require as much proximity as the SS, also through the use of telework. The latter has very low linear coefficient values and so the impact of this sector on the spread of the virus will be small.

7. CONCLUSIONS AND FUTURE WORK

The objective of this thesis was to use FDA to analyse and quantify the association of COVID-19 incidence data with Google mobility and Sociodemographic data to understand the impact that mobility and sociodemographic conditions have on the spread of COVID-19. Despite several limitations of data related with accuracy and level of aggregation, some relevant trends in functional data curve shapes could be identified. The results show that measures must be taken to protect the most vulnerable and disadvantaged populations. Also, they strengthened the idea that the ideal lag for which mobility has some effect on the incidence is about 15 days. Additionally, the results reinforce the effectiveness of restrictive measures such as lockdown. The variation of different mobility classes can be used (residential, retail and stations) to try to predict the spread of the virus. Some results suggest that sudden changes in mobility (mass movements of the population, and mandatory lockdown) have greater association with the evolution of the incidence of COVID-19. This may mean that there is a threshold in mobility behaviour, from which this association becomes stronger, and that can help to contain the spread of COVID-19.

The FDA is a very broad area with a lot of potential yet to be explored. Other approaches can be tried, for example, using FPCA and Functional Principal Differential Analysis. The methodology of this work can be applied in future waves of COVID-19 (or future pandemics). It may be interesting to study the association of other variables with the COVID-19 incidence, including other sociodemographic variables, meteorological data, other mobility data, or vaccination rates. Another suggestion would be to apply the methodology of this work using COVID-19 mortality data. This strategy may be relevant in the sociodemographic field, due to the important role they typically play in the population's health outcomes.

8. BIBLIOGRAPHY

[1] J. R. Cavalcante *et al.*, “COVID-19 no Brasil: evolução da epidemia até a semana epidemiológica 20 de 2020,” *Epidemiologia e servicos de saude : revista do Sistema Unico de Saude do Brasil*, vol. 29, no. 4, p. e2020376, 2020, doi: 10.5123/s1679-49742020000400010.

[2] A. Aleta and Y. Moreno, “Evaluation of the potential incidence of COVID-19 and effectiveness of containment measures in Spain: A data-driven approach,” *BMC Medicine*, vol. 18, no. 1, May 2020, doi: 10.1186/s12916-020-01619-5.

[3] N. R. Jones, Z. U. Qureshi, R. J. Temple, J. P. J. Larwood, T. Greenhalgh, and L. Bourouiba, “Two metres or one: what is the evidence for physical distancing in covid-19?,” *BMJ (Clinical research ed.)*, vol. 370, p. m3223, Aug. 2020, doi: 10.1136/bmj.m3223.

[4] P. Nouvellet *et al.*, “Reduction in mobility and COVID-19 transmission,” *Nature Communications*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-21358-2.

[5] J. O. Ramsay, “Functional Data Analysis – Theory,” in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2016, pp. 1–13. doi: 10.1002/9781118445112.stat00516.pub2.

[6] D. L. [I. Evttin, R. F. Gina, L. Nuzzo, and J. O. Ramsay, “Introduction to Functional Data Analysis,” *Canadian Psychology*, vol. 48, no. 3, pp. 135–155, 2007, doi: 10.1037/cp20070l4.

[7] S. Ullah and C. F. Finch, “Applications of functional data analysis: A systematic review,” 2013. [Online]. Available: http://www.psych.mcgill.ca/misc/fda/

[8] T. Boschi, J. di Iorio, L. Testa, M. A. Cremona, and F. Chiaromonte, “The shapes of an epidemic: using Functional Data Analysis to characterize COVID-19 in Italy,” Aug. 2020, doi: 10.1038/s41598-021-95866-y.

[9] C. Tang, T. Wang, and P. Zhang, “Functional data analysis: An application to COVID-19 data in the United States,” Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.08363

[10] C. Carroll *et al.*, “Time dynamics of COVID-19,” *Scientific Reports*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-77709-4.

[11] G. Sebastiani, M. Massa, and E. Riboli, “Covid-19 epidemic in Italy: evolution, projections and impact of government measures,” *European Journal of Epidemiology*, vol. 35, no. 4, pp. 341–345, Apr. 2020, doi: 10.1007/s10654-020-00631-6.

[12] D. Delen, E. Eryarsoy, and B. Davazdahemami, “No place like home: Cross-national data analysis of the efficacy of social distancing during the COVID-19 pandemic,” *JMIR Public Health and Surveillance*, vol. 6, no. 2, Apr. 2020, doi: 10.2196/19862.

[13] V. Moorthy, A. M. H. Restrepo, M. P. Preziosi, and S. Swaminathan, “Data sharing for novel coronavirus (COVID-19),” *Bulletin of the World Health Organization*, vol. 98, no. 3. World Health Organization, p. 150, Mar. 01, 2020. doi: 10.2471/BLT.20.251561.

[14] E. Marques da Costa and N. Marques da Costa, “A PANDEMIA COVID-19 EM PORTUGAL CONTINENTAL – UMA ANÁLISE GEOGRÁFICA DA EVOLUÇÃO VERIFICADA NOS MESES DE MARÇO E ABRIL,” *Hygeia - Revista Brasileira de Geografia Médica e da Saúde*, pp. 72–79, Jun. 2020, doi: 10.14393/hygeia0054396.

[15] E. Hatef, H. Y. Chang, C. Kitchen, J. P. Weiner, and H. Kharrazi, “Assessing the Impact of Neighborhood Socioeconomic Characteristics on COVID-19 Prevalence Across Seven States in the United States,” *Frontiers in Public Health*, vol. 8, Sep. 2020, doi: 10.3389/fpubh.2020.571808.

[16] A. P. Patel, M. D. Paranjpe, N. P. Kathiresan, M. A. Rivas, and A. v. Khera, “Race, socioeconomic deprivation, and hospitalization for COVID-19 in English participants of a national biobank,” *International Journal for Equity in Health*, vol. 19, no. 1, Jul. 2020, doi: 10.1186/s12939-020-01227-y.

[17] D. Quan *et al.*, “Impact of Race and Socioeconomic Status on Outcomes in Patients Hospitalized with COVID-19,” *Journal of General Internal Medicine*, vol. 36, no. 5, pp. 1302–1309, May 2021, doi: 10.1007/s11606-020-06527-1.

[18] Google, “Covid-19 Community Mobility Reports,” 2021.

[19] T. M. Drake, A. B. Docherty, T. G. Weiser, S. Yule, A. Sheikh, and E. M. Harrison, “The effects of physical distancing on population mobility during the COVID-19 pandemic in the UK,” *The Lancet Digital Health*, vol. 2, no. 8. Elsevier Ltd, pp. e385–e387, Aug. 01, 2020. doi: 10.1016/S2589-7500(20)30134-5.

[20] G. Launoy, L. Launay, O. Dejardin, J. Bryère, and E. Guillaume, “European Deprivation Index: designed to tackle socioeconomic inequalities in cancer in Europe,” *European Journal of Public Health*, vol. 28, no. suppl_4, Nov. 2018, doi: 10.1093/eurpub/cky213.625.

[21] A. I. Ribeiro, L. Launay, E. Guillaume, G. Launoy, and H. Barros, “The Portuguese version of the European deprivation index: Development and association with all-cause mortality,” *PLoS ONE*, vol. 13, no. 12, Dec. 2018, doi: 10.1371/journal.pone.0208320.

[22] R. J. Hyndman and G. Athanasopoulos, “Forecasting: Principles and Practice.”

[23] A. Grami, “Probability, Random Variables, and Random Processes,” in *Introduction to Digital Communications*, Elsevier, 2016, pp. 151–216. doi: 10.1016/b978-0-12-407682-2.00004-1.

[24] L. Horváth, P. Kokoszka, and G. Rice, “Testing stationarity of functional time series,” *Journal of Econometrics*, vol. 179, no. 1, pp. 66–82, 2014, doi: 10.1016/j.jeconom.2013.11.002.

[25] S. van Buuren and K. Groothuis-Oudshoorn, “mice : Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, vol. 45, no. 3, 2011, doi: 10.18637/jss.v045.i03.

[26] S. van Buuren, *Flexible Imputation of Missing Data, Second Edition*. Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC, 2018. doi: 10.1201/9780429492259.

[27] J. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*. Springer New York, 2009. doi: 10.1007/978-0-387-98185-7.