TÉCNICO LISBOA

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

**João Filipe Alfaia Subtil**

Thesis to obtain the Master of Science Degree in

## Biomedical Engineering

Supervisors: Professor Manuel Luís Castro Ribeiro

Professor Leonardo Azevedo Guerra Raposo Pereira

### Examination Commitee

Chairperson: Professor João Miguel Raposo Sanches

Supervisor: Professor Manuel Luís Castro Ribeiro

Members of the Committee: Doctor André Peralta Santos

Professor Maria João Correia Colunas Pereira

**October 2021**

## Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

## Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

# Preface

The work presented in this thesis was performed at the Instituto superior Técnico (Lisbon, Portugal), during the period March-October 2021, under the co-supervision of Prof. Professor Manuel Luís Castro Ribeiro and Prof. Leonardo Azevedo Guerra Raposo Pereira.

# Acknowledgments

Firstly, I would like to thank my co-supervisors, Professor Manuel Ribeiro and Professor Leonardo Azevedo, for their support and guidance over the past few months, and for being always available to help me developing my dissertation.

I would also like to express my gratitude to the Direção-Geral da Saúde for providing the material essential for the development of this work.

Thanks to my friends who have accompanied me in the past few years.

Last, but not least, I would like to thank my family, in particular my parents and brother, for all their support and encouragement.

## 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

## Resumo

A Análise Funcional de Dados (FDA) é uma ferramenta de análise estatística que permite analisar séries temporais como funções. Apesar de ser usada em várias áreas, apenas dá os primeiros passos na epidemiologia. Nesta tese, FDA é aplicada para analisar a associação da incidência com mobilidade e variáveis sociodemográficas em Portugal Continental. O Modelo Concorrente é usado para analisar a associação entre uma variável resposta funcional (incidência COVID-19) e uma variável explicativa funcional (Mobilidade Google). O Modelo de Análise da Variância é aplicado para medir a associação entre a incidência da COVID-19 e variáveis explicativas escalares (Variáveis Sociodemográficas). Os resultados permitem identificar tendências na forma das curvas dos dados funcionais. O sinal mais forte encontrado para a associação entre COVID-19 e mobilidade foi na mobilidade Residencial. A mobilidade Retalho e Transportes Públicos apresenta resultados interessantes, sendo que a mobilidade nas Mercearias, Parques e Local de Trabalho apresentam sinais fracos. Os resultados fortaleceram a ideia de que o desfasamento para a mobilidade ter efeito na incidência é cerca de 15 dias, conforme sugerido pela literatura. Mostrou-se que certas condições sociodemográficas poderão influenciar a propagação da COVID-19, como o rendimento, estrutura etária da população, densidade de escolas ou setores de atividade. As técnicas aqui usadas sugerem que a FDA pode ser uma ferramenta adicional para a epidemiologia da COVID-19 e podem ser replicadas para a mortalidade por COVID-19 ou outras doenças ou pandemias. A FDA é uma área ampla, pelo que uma análise mais profunda pode ser feita recorrendo a outras ferramentas.



**Palavras-Chave:** Análise Funcional de Dados, Incidência COVID-19, Mobilidade Google, Variáveis Sociodemográficas, Modelo de Análise da Variância, Modelo Concorrente

**Contents**


Acknowledgments ..... vii

Abstract ..... ix

Resumo ..... xi

List of Tables ..... xv

List of Figures ..... xvi

Acronyms ..... xix

Introduction ..... 1

Material ..... 7

2.1 COVID-19 Incidence Data ..... 7

2.2 Google Mobility Data ..... 8

2.3 Sociodemographic Data ..... 10

Methodology ..... 12

3.1 Pre-Processing ..... 12

3.1.1 Stationarity ..... 12

3.1.2 Imputation (predictive mean matching) ..... 13

3.1.3 NAs Removal ..... 14

3.1.4 Data Analysis per Wave ..... 14

3.2 FDA ..... 15

3.2.1 Basis Functions ..... 15

3.2.2 Adding Coefficients to Bases to Define Functions ..... 20

3.2.3 Regression Splines: Smoothing by Regression Analysis ..... 21

3.3 Linear Models ..... 26

Results ..... 33

4.1 Functional Responses with Functional Covariates: Concurrent Model ..... 35

4.2 Functional Responses with Functional Covariates: General Concurrent Model ..... 38

4.3 Functional Responses with Scalar Covariates: Analysis of Variance Model ..... 41

Discussion ..... 46

5.1 Functional Responses with Functional Covariates: Concurrent Model ..... 46

5.2 Functional Responses with Functional Covariates: General Concurrent Model ..... 47

5.3 Functional Responses with Scalar Covariates: Analysis of Variance Model ..... 50

5.4 Limitations ..... 51

Conclusions ..... 52
6.1 Summary and Conclusions ..... 52
6.2 Future Work ..... 54
Bibliography ..... 55

# List of Tables

Table 1 - Data structure of incidence data provided by DGS ..... 8
Table 2 - Data structure of incidence data used for functional data analysis ..... 8
Table 3 - Data structure of mobile data used for functional data analysis ..... 10
Table 4 - Sociodemographic Variables Description ..... 11

# List of Figures

Figure 1 - Linear combinations of spline basis functions of orders 2, 3 and 4, which aim to fit a sine function and its first derivative (J. Ramsay et al., 2009) ..... 19
Figure 2 - Representation of 13 B-Splines basis functions (J. Ramsay et al., 2009) ..... 20
Figure 3 - Incidence Rate Curves (National and Municipalities) ..... 33
Figure 4 - Mobility Variation Curves (National and Municipalities): a) Grocery b) Station c) Parks d) Retail ..... 34
Figure 5 - Linear Coefficient Function for the Velocity of Mobility Variation with 95% pointwise confidence intervals (Residential Class, 2nd wave, 16-day lag) ..... 35
Figure 6 - Linear Coefficient Function for the Velocity of Mobility Variation in the 3rd wave with 95% pointwise confidence intervals: a) Grocery (15-day lag) b) Parks (16-day lag) c) Stations (15-day lag) d) Stations (16-day lag) ..... 36
Figure 7 - Linear Coefficient Function for the Velocity of Mobility Variation in the 3rd wave with 95% pointwise confidence intervals: a) Residential (16-day lag) b) Retail (15-day lag) ..... 37
Figure 8 - Linear Coefficient Function for the Velocity of Mobility Variation with 95% pointwise confidence intervals (Workplace Class, 3rd wave, 15-day lag) ..... 37
Figure 9 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Residential Class): a) 2nd wave b) 3rd wave c) 2nd and 3rd wave ..... 38
Figure 10 – Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Grocery Class): a) 2nd wave b) 3rd wave ..... 39
Figure 11 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Parks Class): a) 2nd wave b) 3rd wave ..... 39
Figure 12 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Retail Class): a) 2nd wave b) 3rd wave ..... 40
Figure 13 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Stations Class): a) 2nd wave b) 3rd wave ..... 40
Figure 14 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Workplace Class): a) 2nd wave b) 3rd wave ..... 41
Figure 15 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Population Density ..... 42
Figure 16 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Schools Density ..... 43
Figure 17 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Deprivation Index ..... 43
Figure 18 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Elderly Population ..... 44

Figure 19 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Youth Population ..... 44
Figure 20 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Proportion of Guaranteed Minimum Income Beneficiaries ..... 45
Figure 21 - Linear Coefficient Functions estimated for predicting COVID-19 incidence from Working Population per Service Sector ..... 45

# Acronyms

| | |
|---|---|
| **WHO** | World Health Organization |
| **COVID-19** | Coronavirus Disease-2019 |
| **ICTV** | International Committee on Taxonomy of Viruses |
| **FDA** | Functional Data Analysis |
| **FPCA** | Functional Principal Component Analysis |
| **Rt** | Effective Reproduction Number |
| **ADI** | Area Deprivation Index |
| **CSV** | Comma Separated Value |
| **INE** | Instituto Nacional de Estatística |
| **TXT** | Text File |
| **DGS** | Direção Geral da Saúde |
| **PD** | Population Density in Urban Areas |
| **DI** | Deprivation Index |
| **YP** | Youth Population |
| **EP** | Elderly Population |
| **PS** | Working Population in Primary Sector |
| **SS** | Working Population in Secondary Sector |
| **TS** | Working Population in Tertiary Sector |
| **GMI** | Guaranteed Minimum Income |
| **SD** | Schools Density |

# Chapter 1

# Introduction

At the end of 2019, several cases of a contagious pneumonia of unknown origin were identified among the population of Wuhan, a region of China, related to a seafood and live animal market. (Cavalcante et al., 2020). Despite the measures implemented in China, with the city of Wuhan subjected to quarantine, this pneumonia caused several outbreaks and quickly spread to most of the countries around the world (Aleta & Moreno, 2020).

The first cases in Portugal appeared on March 2, 2020 (Marques da Costa & Marques da Costa, 2020). As of October 26, 2021, 1.086.280 positive cases have been detected in Portugal, causing 18.141 deaths. Worldwide, the number of positive cases detected rises to 235 million, causing a number of deaths close to 5 million. 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 (Cavalcante et al., 2020). From an epidemiological point of view, this is very important, because although individuals do not present symptoms (cough, shortness of breath, etc.), they still can transmit the virus. Therefore, strategies to contain the virus are difficult to implement, since people can transmit it without knowing that they are infected, and the quarantine (only) of symptomatic cases is not effective.

Due to the rapid spread of COVID-19 worldwide, with the number of positive cases and affected countries growing exponentially, a pandemic was declared on March 11, 2020. Due to this unprecedented situation in the recent history of humanity, the governments of these countries were forced to take preventive measures, which varied from country to country. The use of masks became widespread, distancing was recommended, gatherings were prohibited. In Portugal, throughout the various COVID-19 waves that emerged, mobility restrictive measures were also used. These measures included, for example, the closure of certain activities, store's occupancy limitation, teleworking, land border closing, social-distancing and mandatory lockdown. All these measures were taken on the basis that the main route of COVID-19 transmission is via respiratory droplets. The objective was to reduce the flow of people, avoiding gatherings, reducing the probability of infection, and thus delaying the spread of the virus in the community and preventing the collapse of the national health system. (Jones et al., 2020; Nouvellet et al., 2021).

According to the European Centre for Disease Prevention and Control, coronaviruses are mainly transmitted via respiratory droplets, for example when coughing and speaking. Transmission through contact is considered possible, but it has not yet been proven for SARS-CoV-2. There is no evidence of transmission through contact with blood, and SARS-CoV-2 has been detected in respiratory and faecal specimens. Reports showed that COVID-19 transmission can be effective in crowded, confined indoor spaces, and that poor ventilation in these spaces is associated with increased transmission of COVID-19. Studies have also shown that most of the working time is spent indoors and that variations in the socioeconomic and demographic characteristics influence working patterns indoors. It was also found that outbreaks in workplaces are often characterized by the slow implementation of hygiene measures and disease control, and that the lack protective personal equipment increases the risk of infection.

These measures were taken due to the inexistence, at the time, of vaccines or effective treatments against COVID-19. Their development requires months of work, which is why it is important to protect the population by using effective containment measures, delaying the spread of the virus, and saving time for the development of vaccines and treatments. (Aleta & Moreno, 2020). In Portugal, for this to be possible across the several waves, a great joint effort of various entities was necessary, including health authorities, security forces, national health service, etc.

Despite everything, it is not always possible to know to what extent the restrictive measures applied in each country prevent the spread of COVID-19. It is necessary to understand if these measures are necessary and effective, and if there are different approaches that would allow an effective protection of the population while reducing the negative impact of these measures on society. The use of open access COVID-19 data allows researchers to use statistical methods to characterize the spread of COVID-19 (Abbas et al., 2021). Therefore, the data available for COVID-19 should be analysed, relate them to other variables that may have an influence on the virus' behaviour, draw conclusions from the results obtained, and better prepare for future pandemic waves.

Nowadays, more than ever, several types of data are essential to guide governments in health care planning. In order to obtain good quality data, statistical tools have been applied in diverse areas of science, such as medicine, biomechanics, or public health. Multivariate Data Analysis has been the most used approach to deal with time series data. Multivariate Data Analysis uses data from multiple measurements made on different experimental units, and analyses the relationships between these multivariate measurements, exploring the data structures and its possible patterns. (Goodman et al., 1979)

However, this approach ignores the functional behaviour of the generating process that underlies the data. Functional Data Analysis (FDA) arises here as an alternative methodology that has been increasingly applied in areas such as public health and biomedicine, that allows to model time series data and extract additional information.

But what is Functional Data Analysis, and what is its usefulness in the analysis and statistical treatment of data?

In Functional Data Analysis, each record that comprises the functional data is called a functional datum. Evttin and colleagues (Evttin et al., 2007) defines a functional datum not as a single observation but as a set of measurements along a continuum that, taken together, are considered as a single entity, curve, or image. Normally, this continuum is defined as being temporal, in which case the data is called longitudinal. Ramsay (J. O. Ramsay, 2016) emphasizes the fact that these observations may not be equally spaced, and their argument values can be different across records, providing flexibility.

Because this approach considers each curve as a single entity, possible correlations between repeated measurements are no longer a problem, one that persisted in multivariate data analysis (Ullah & Finch, 2013)

However, Functional Data Analysis can have numerous applications, within which the continuum is not always defined on a time scale. As Ramsay (J. O. Ramsay, 2016) states, functional data can be distributed over one or more spatial dimensions when the observation is an image, over both space and time, or may be taken from any other continuum.

Ullah and Finch (Ullah & Finch, 2013) describes functional data analysis as a way to analyse, model and predict time series data. According to them the objective of functional data analysis is to transform discrete observations from time series into a function that represents the entire measured data as a single observation (a functional datum). Then, concepts from multivariate data analysis are applied to these data, in order to develop models that allow deriving important information.

An important feature of functional data is the data resolution, that allows us to identify important characteristics of the curves, such as peaks, valleys, or some other patterns of interest.

Another very important aspect in functional data analysis is the smoothness of the curves. When collecting records, which involve discrete observations, it is expected that the latter somehow reflect the smooth process that gives rise to this data. Smoothness is defined as the characteristic of a curve that can be differentiated to a certain degree (Evttin et al., 2007). The functional observations (curves) can be transformed, through derivation, in order to observe the different forms of variation of each function (velocity and acceleration), and thus identify important comparison points in the curves.

Some applications of FDA include using curves as data, images as data, and data points as shape representations of body parts. In the systematic review of Ullah and Finch (Ullah & Finch, 2013) the application of FDA has been identified across various scientific fields including analysis of child size evolution, climatic variation, handwriting in Chinese, acidification processes, land usage prediction based on satellite images, medical research, behavioural sciences, term-structured yield curves, and spectrometry data.

Functional data analysis is a process that involves the application of several tools. These tools, whose applications are diverse and depends on the specificity of the research problem include smoothing and interpolation for data representation and pattern recognition (Hyndman & Shahid Ullah, 2007), derivation (usually up to the second derivative) to identify underlying patterns not visible before (Mas & Pumo,

2009), principal component analysis for data reduction (Croux & Ruiz-Gazen, 2005), clustering to identify patterns or individuals involved in similar processes (Song et al., 2007) or linear regression to model relations between functional data (Goldsmith & Schwartz, 2017).

Thus, FDA is a powerful tool, which has been increasingly used in many different areas. These days, it can play an important role as a statistical analysis tool to help investigate COVID-19 and its impact on society. Since the appearance of COVID-19 at the end of 2019, several studies on this coronavirus have been developed, using several mathematical methods, including the FDA. The association of COVID-19 with mobility and sociodemographic factors, the theme of this dissertation, is one of the points that have been addressed by these studies.

Boschi and colleagues (Boschi et al., 2020) used FDA to explore the COVID-19 mortality in Italy, and its association with covariates such as mobility, and socio-demographic variables. The study found that mobility and positivity are associated with COVID-19 mortality and identified schools and workplaces as having higher risk of contagion. Another study utilized FDA to study the COVID-19 positivity and mortality in the United States, and their association with Google search trends for COVID-19 symptoms. The methods used include functional principal component analysis (FPCA), dynamic correlation, canonical correlation, and clustering (Abbas et al., 2021). Moreover, other study utilized FDA methods to investigate the COVID-19 spread in the United States. Functional principal component analysis, canonical correlation, cluster analysis, dynamic FPCA and functional modelling are the techniques used, and the results demonstrate that measures such as stay-at-home orders are essential to slow the growth rate of the disease (Tang et al., 2020). Carroll and colleagues (Carroll et al., 2020) applied FDA to model the cumulative cases of COVID-19 trajectories in several countries. This study uses FPCA, Rank Dynamics, Dynamic FPCA and Functional Concurrent Regression, and shows that a decrease on workplace mobility is correlated with reduced doubling rates (with a 2-week delay).

Several other studies were developed with the aim of analysing the impact of restrictive measures applied by governments, namely lockdown measures, in containing the spread of COVID-19.

Segmented regression analysis was applied to identify shifts on the trajectory of the effective reproduction number (Rt) in Spain, and the results show a sharp decrease in Rt when lockdown is applied (Santamaría & Hortal, 2021). Sebastiani and colleagues (Sebastiani et al., 2020) analysed the incidence of COVID-19 in Italy, and the containment measures implemented. Evidence was found that the measures implemented contributed to reduce the spread of the COVID-19 in the country. Another research applied a susceptible-infected-recovered (SIR) model to COVID-19 data to calculate the weekly transmission rate (β) in several countries. Gradient boosted trees regression analysis is used to analyse the association between mobility and these β values, and according to the results distancing measures are effective in reducing the spread of the disease (Delen et al., 2020). Aleta and colleagues (Aleta & Moreno, 2020) implemented a SEIR metapopulation model that traces the spread of COVID-19 in Spain using data-driven stochastic simulations, to study the most effective methods to reduce COVID-19 transmission. The results show that early detection and isolation of symptomatic individuals, followed by public interventions, are the best approaches. Moorthy and colleagues (Moorthy et al., 2020)

developed an interrupted time series study to assess the effectiveness of lockdown in reducing the confirmed/death cases from COVID-19 in China. The results demonstrate that the social distancing measures had a positive impact in slowing the spread of the disease, reducing the incidence and mortality rates, and that the impact on incidence occurs between 7 to 17 days after the application of the measures. Other study applied a mechanistic model of COVID-19 transmission to study the spread of the disease in some countries and to investigate the impact of containment measures. The study evidences that the intervention measures had a positive impact in the disease reproduction number (Rt) (Fernández-Recio, 2020). Kumar and colleagues (Kumar et al., 2021) analysed the efficacy of diverse NPIs (non-pharmaceutical interventions) that countries can apply to slow the COVID-19 incidence and mortality. The study demonstrated that prevention-focused interventions reduces COVID-19 incidence and promotion-focused interventions enhances the response to medical emergencies. A study developed by Nouvellet and colleagues (Nouvellet et al., 2021) described the association between transmission and mobility and found evidence that mobility patterns correlate with the intensity of transmission. Also, it was noticed that the association between mobility and transmission changer over time, suggesting that smaller reductions in mobility can contribute to slow down the spread of the disease, due to other social distancing habits. Srivastava and Chowell (Srivastava & Chowell, 2020) utilized statistical tools to study shapes of COVID-19 growth rate curves and divides them into clusters. This analysis identifies the dominant incidence trajectories across Europe and USA, and can be used to characterize the transmission dynamics, improving public health decision making.

Moreover, the impact of sociodemographic variables on the evolution of COVID-19 was a central topic in the development of several studies.

Marques da Costa and Marques da Costa (Marques da Costa & Marques da Costa, 2020) intended to correlate the evolution of the COVID-19 in Portugal to its sociodemographic and demographic characteristics. This investigation shows that the virus spreads from large urban areas to the surroundings. Also, it evidences that elderly people in nursing homes constitute an extremely vulnerable part of the population, and immigrants have an increasing incidence. An investigation by Hatef and colleagues (Hatef et al., 2020) 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. Another study demonstrated that the prevalence and death rate of COVID-19 in USA are associated with sociodemographic conditions and mobility. Also, it shows that COVID-19 preferentially affects the elderly and is affecting the counties with a larger percentage of non-white population (Paul et al., 2021). Patel and colleagues (Patel et al., 2020) analysed the association of COVID-19 hospitalizations with racial and sociodemographic characteristics. 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. Khalatbari-Soltani and colleagues (Khalatbari-Soltani et al., 2020) offered an overview on how important sociodemographic factors are being overlooked, concluding that a recording of sociodemographic characteristics of patients with COVID-19 is essential to develop a strategy that protects the most vulnerable groups.

A study developed by Whittle and Diaz-Artiles (Whittle & Diaz-Artiles, 2020) fitted multiple Bayesian Besag-York-Mollié (BYM) mixed models using COVID-19, sociodemographic and health-care data. The study found associations between COVID-19 positivity rate and neighbourhoods with a large dependent youth population, densely populated, low-income, and predominantly black neighbourhoods. Cao and colleagues (Cao et al., 2020) developed a register-based ecological study using spatial regression analysis based on COVID-19 country-level data. The results show that the COVID-19 case-fatality rate is correlated with population size, indicating healthcare systems under pressure and/or lower treatment efficiency in countries with large populations, and secondary to higher transmission risk and poorer health. Another study developed a retrospective cohort study to analyse the strong 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. (Quan et al., 2021).

Thus, the studies listed above allow to draw several conclusions about the studies that investigate COVID-19, and the methods used. It is clear that, despite the multiple advantages presented by the FDA, this is a relatively recent technique, and conventional methods are often used instead. In Portugal, the application of the FDA in this field is really limited, so what this thesis proposes is something innovative, and intends to show the potential that this methodology has in the area of pandemic analysis and prevention. The objective is then to use several FDA techniques such as smoothing, interpolation and functional linear models, in order to analyse and quantify the association of COVID-19 incidence data with Google mobility and Sociodemographic data. The results obtained will help us understand the impact that mobility and sociodemographic conditions have on the spread of COVID-19 and whether the measures taken by the Portuguese government will or will not influence the containment of the virus.

The work was developed under the project SCOPE-Spatial Data Sciences for COVID-19 Pandemic, funded by Fundação para a Ciencia e Tecnologia[1]

[1] Project reference: DSAIPA/DS/0115/2020

# Chapter 2

# Material

The material used in this thesis, which was the basis for the development of the entire work, is divided into 3 large groups:

- COVID-19 Data

- Google Mobility Data

- Sociodemographic Data

These data are provided in CSV (Comma Separated Value) or TXT (Text File) formats. Due to the specificities of FDA methods in terms of data structure, and also in order to remove non-essential information, these data had to go through pre-processing.

As these types of data are essential in understanding the entire process of developing the thesis, detailed information about them will be provided below (source, structure, pre-processing, etc.). The programming language R[2] was used to perform data pre-processing steps.

## 2.1 COVID-19 Incidence Data

COVID-19 data are the focus of this work. These data consist of the daily 7-day cumulative incidece by COVID-19 in each of the 278 municipalities in Continental Portugal. They are provided by the Direção Geral da Saúde (DGS), and the latest version of these data refers to 761.906 newly confirmed COVID-19 cases reported between March 9, 2020 and February 6, 2021 (thus containing information on the first 3 waves of the pandemic).

The data are provided in a file in CSV format, with each row in this file corresponding to a new reported case of COVID-19 infection. For the reported case i, ti is the date of confirmed positive test, mi(s) is the municipality of residence. The cases are anonymous and characterized by date of confirmed positive test and municipality of residence (Table 1).

[2] R: A language and environment for statistical computing (R Foundation for Statistical Computing)

| Date | Municipality of residence |
|---|---|
| t1 | m1(s) |
| … | … |
| ti | mi(s) |
| … | … |
| tn | mn(s) |

*Table 1 - Data structure of incidence data provided by DGS*

Using R functions, this file was subjected to pre-processing, whose final product is a table where each column corresponds to municipality s (s= 1,…,M) , each row corresponds to day t (t = 1,…, 334), and cell values represent daily crude COVID-19 7-day cumulative incidence rates. Thus, this is time series data, with each column (municipality) being a time series (Table 2).

| 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 2 - Data structure of incidence data used for functional data analysis*

A final pre-processing step was applied to incidence data using the box-to-cox method and computing the second differences to these data. This final processing will be explained in more detail later in the Methodology section.

## 2.2 Google Mobility Data

The Google data used in the development of this work consist of daily mobility values in each of the 278 municipalities in Portugal mainland. This data is made available free by Google for public use (under the designation of Community Mobility Reports), and the version used here is related to the period of time between March 15, 2020 and February 2, 2021 (thus containing information on the first 3 waves of the pandemic). Most of the information detailed in this section is provided by Google[3].

The Community Mobility Reports show movement trends by region, across different categories of places. For each category in a region, reports show the changes in 2 ways:

- Headline number: Compares mobility for the report date to the baseline day. Calculated for the report date (unless there are gaps) and reported as a positive or negative percentage. It shows how visits and length of stay at different places change compared to a baseline.

[3] Google COVID-19 Community Mobility Reports

- Trend graph: The percent changes in the 6 weeks before the report date. Shown as a graph.

The mobility data is aggregated from mobile device location information from Android users, because they have the largest market share (Drake et al., 2020). No personally identifiable information, such as an individual's location, contacts or movement, was made available at any point, as the reports are powered by anonymisation technology to keep users' activity data private and secure. These reports are created with aggregated, anonymized sets of data from users who have turned on the Google Location History setting, which is off by default, so the data represents a sample of Google users. As with all samples, this may or may not represent the exact behaviour of a wider population. This mobility data is the same data used, for example, to show popular times for places in Google Maps.

The data that is included in the calculations depends on user settings, connectivity and whether it meets Google privacy threshold. When the data doesn't meet quality and privacy thresholds, you might see empty fields for some places and dates, and the headline number is the most-recent calculated change. Gaps should be treated as unknowns and don't mean that a specific place isn't busy.

Comparisons between places across regions should be avoided, since regions can have local differences in the data which might mislead.

Six mobility categories are presented. To make the reports useful, categories are used to group some of the places with similar characteristics for purposes of social distancing guidance. This includes categories that are useful for social distancing efforts, as well as access to essential services. For example, grocery and pharmacy are combined in a category since they tend to be considered essential.

- Supermarket and pharmacy: Mobility trends for places such as grocery shops, food warehouses, farmers markets, specialty food shops and pharmacies.
- Parks: Mobility trends for places such as local parks, national parks, public beaches, marinas, dog parks, plazas and public gardens.
- Public transport: Mobility trends for places that are public transport hubs, such as underground, bus and train stations.
- Retail and recreation: Mobility trends for places such as restaurants, cafés, shopping centres, theme parks, museums, libraries, and cinemas.
- Residential: Mobility trends for places of residence.
- Workplaces: Mobility trends for places of work

The data shows how visitors to (or time spent in) categorized places change compared to the baseline days. A baseline day represents a normal value for that day of the week. Here, the baseline day is the median value from the 5-week period Jan 3 – Feb 6, 2020, before widespread COVID-19 disruption. For some regions, the baseline falls during a time when COVID-19 was established, however, for Portugal, this does not happen.

For each region-category, the baseline isn't a single value—it's 7 individual values. The same number of visitors on 2 different days of the week, result in different percentage changes. So, it is important to keep in mind that it must not be inferred that larger changes mean more visitors or smaller changes

mean less visitors. Also, day-to-day changes comparisons must be avoided, especially weekends with weekdays.

To help track week-to-week changes, the baseline days never change. These baseline days also don't account for seasonality. For example, visitors to parks typically increase as the weather improves.

Apple also provides mobility data, however, unlike Google, it doesn't distinguish between different types of mobility.

These data are provided in a CSV file that is converted to a data.frame using R functions. In this data.frame, each line corresponds to a municipality. In each of these lines, there are the mobility values of a particular municipality, for the period of time referred above (March 15, 2020 to February 2, 2021).

The data.frame is transformed into one where each column corresponds to a municipality, each row corresponds to day t (t = 1,…,324) between March 15, 2020 to February 2, 2021 and cell values represent daily mobility values. Thus, this is time series data, with each column (municipality) being a time series. This final data.frame has a structure identical to the data.frame of COVID-19's daily cumulative incidence, which is essential to apply FDA tools and develop this work (Table 3).

| 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 3 - Data structure of mobile data used for functional data analysis*

As mentioned above, for confidentiality reasons it is sometimes not possible to obtain mobility data. For this reason, several cells with the value NA are found in the data.frames.

In order to circumvent the existence of NAs, an imputation method was used in order to fill in the missing values. Before that, a threshold was defined for the number of NA values per municipality. When this threshold was exceeded, the municipality would be eliminated due to lack of data (which would make it impossible to use imputation).

In a final step, the logarithm and the first difference were applied to mobility data. This final processing will be explained in more detail later in the Methodology section.

## 2.3 Sociodemographic Data

The sociodemographic data used in this thesis consist of a vast number of variables, and cover each of the 278 municipalities in Continental Portugal (Table 4). Unlike incidence and mobility data, sociodemographic data is not daily, but rather a single value for each municipality, referring to the most recent year data were available.

These data are essentially from 2 sources, INE and PORDATA, and are provided in CSV or TXT files. Using R functions, these files are converted into data.frames, adding other variables to identify the municipalities. However, the provided data are mostly absolute values.

They have to be transformed into standardized values (for example schools/km$^2$ or percentage of elderly population), allowing a comparative analysis between municipalities. In the final data.frames, each row is a municipality, and the columns correspond to sociodemographic variables.

The Deprivation Index used here is based on data from the 2011 Census published by Ribeiro and colleagues (Ribeiro et al., 2018). The Deprivation Index is a measure of poverty that was created within the scope of the European Deprivation Index project, and is composed of several different sociodemographic variables, with the aim of quantifying through a single variable (an indicator) the levels of deprivation of the various administrative regions of Continental Portugal, in this case its 278 municipalities. The index values vary between 1-5 where 1 is less deprived and 5 more deprived.

The sociodemographic variables that were used are as follows:

| Variable | Year | Description | Abbreviation |
|---|---|---|---|
| Population Density in Urban Areas | 2018 | Inhabitants / km$^2$ | PD |
| Deprivation Index | 2011 | 1 to 5 | DI |
| Youth Population | 2018 | % Inhabitants 0-19 years | YP |
| Elderly Population | 2018 | % Inhabitants 65+ years | EP |
| Working Population in Primary Sector | 2014 | % Working population in Primary Sector | PS |
| Working Population in Secondary Sector | 2014 | % Working population in Secondary Sector | SS |
| Working Population in Tertiary Sector | 2014 | % Working population in Tertiary Sector | TS |
| Guaranteed Minimum Income | 2018 | Proportion of Guaranteed Minimum Income beneficiaries | GMI |
| Schools Density | 2018 | Schools / Km$^2$ | SD |

*Table 4 - Sociodemographic Variables Description*

Later, in order to apply the essential methods for the work, the created data.frames had to be transformed. Unnecessary information has been removed, and essential information has been added.

# Chapter 3

# Methodology

## 3.1 Pre-Processing

### 3.1.1 Stationarity

Although COVID-19 and mobility data are in the correct format to work with, they still went an additional pre-processing step based on transformation of incidence and mobility time-series into stationary ones. It should be noted that this transformation is only carried out for modelling the association between COVID-19 and mobility. When modelling the association between COVID19 and sociodemographic variables, this transformation was not performed, and daily incidence curves are used.

When talking about stationary time-series, ones are referring to time-series whose properties do not depend on the time at which they are observed. When the points from a time-series are not stationary, they have means, variances, and covariances that change over time. This means that, generally, stationary time-series do not have predictable patterns. (Grami, 2016; Hyndman & Athanasopoulos, 2018.)

Non-stationary data, generally, are more complicated to be modelled (Horváth et al., 2014). The results obtained by using non-stationary time series may indicate a relationship between two variables where one does not exist. In order to receive consistent results, the non-stationary data needs to be transformed into stationary data.

The first step of this transformation is to submit the data to a log transformation. However, the data in question contains “zero” values, which makes it impossible to carry out any log transformation, and they have to be treated differently.

In this case, a two-parameter version of the Box-Cox transformation was used, which allows for a shift before the data is transformed

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

Parameters $\lambda_1$ and $\lambda_2$ 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 calculation of first differences. In both cases a time lag (seasonality) of 7 days was used.

$$\textbf{First difference:} \quad \Delta_m z_t = (1 - B)(1 - B^m) z_t \tag{2}$$

$$\textbf{Second difference:} \quad \Delta^2{}_m z_t = (1 - B)^2 (1 - B^m) z_t \tag{3}$$

Here, $z_t$ is an observation of the time series at time t, m is the time lag, and $\Delta z_t$, $\Delta^2 z_t$ represent the first and second differences of $z_t$ respectively and $B$ is the backward shift operator usually defined in time-series. In the context of the data that is being analysed, the transformation that takes place modifies the meaning of the data. In the case of data on the incidence of COVID-19, the second differences correspond to the weekly acceleration in the incidence rate of COVID-19. On the other hand, the first differences of Google Mobility data correspond to the velocity of mobility variation.

### 3.1.2 Imputation (predictive mean matching)

As mentioned in the Materials section, the Google mobility data, for confidentiality reasons, contains several cells with missing data (in a data.frame of R these cells contained the value NA). To apply Functional Data Analysis tools, provided by the *fda* package of R (Buuren & Groothuis-Oudshoorn, 2011), to the data, it is necessary to eliminate the existence of missing data. The use of imputation methods allows replacing the missing data cells by values estimated through specific functions. The imputation method used here is called predictive mean matching, and the package *mice* contains all the necessary functions for this transformation to be carried out.

Since this imputation method is simple and versatile, it was the first choice to cope with the missing data problem. It allows for discrete variables (as in the case of time-series), and is based on real values, providing reliability to the estimation. This method works better with large samples and provides imputations that have characteristics of the complete data. Also, the imputation is limited to range of the observed data, and so unrealistic predictions will not occur. Finally, 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. Essentially, the predictive mean matching method calculates the predicted value of some target variable. For each missing entry in the data, this method creates a small set of candidate donors from all complete cases that have predicted values closest to the predicted value for the missing entry. One donor is drawn from the candidates, and the observed value of the donor replaces the missing value. This is done based on the assumption that the distribution of the missing cell is the same as the observed data of the candidate donors. The missing values are imputed using values from the complete cases matched with respect to some metric, and several metrics are possible to determine the distance

between these cases. The predictive mean matching metric is ideal to deal with missing data because it is optimized for each target variable (van Buuren, 2018).

### 3.1.3 NAs Removal

Before the imputation method is applied to the data, it is necessary to filter the data. The imputation will be applied in all municipalities (each column of the time series), and for it to be reliable, it is necessary that there is sufficient data about them. In other words, it is necessary to apply a threshold to the amount of missing data in each municipality, from which that municipality is removed. The application of imputation methods in municipalities with very large amounts of missing data can lead to estimates that are unrealistic. Also, the municipalities in which the time series does not have values different from 0 will be eliminated, due to the danger of distorting the results of the analysis.

The threshold referred above was structured considering, in each time series:

- Existence of NA sequences longer than 15 days
- Number of NA sequences longer than 4 days (<5% of time series length)
- Total number of NA cells (<25% of the time series)
- Existence of cells with a value other than 0 or NA

### 3.1.4 Data Analysis per Wave

The evolution of the incidence of COVID-19 cases in the Portuguese population, over the period under analysis, presents as its main characteristic the existence of pandemic waves. In each wave, the incidence always behaves similarly: the virus spreads, the incidence increases until it reaches a peak, and then decreases until it stabilizes at minimum values. Considering this specificity in the behaviour of the virus, the time series that constitute the COVID-19 and Google Mobility data were divided in 3 parts, corresponding to each of the waves. This division allows FDA techniques to be applied not only to the entire pandemic period, but also to each of the waves individually (or combining consecutive waves). This makes it possible to obtain a greater combination of results and facilitates their analysis, due to the use of shorter time-series that allows us to focus on one wave at a time.

This approach can be very important, as the behaviour of the population varied rapidly over the period of the pandemic, as well as the rigidity of the containment measures applied. Thus, the individual study of each wave reduces the complexity of the results and allows us to analyse and discuss the results more efficiently, in light of the population's behaviour during that precise period of the pandemic.

As mentioned above, the data available corresponds to the period of time between March 9, 2020 and February 6, 2021, covering 3 waves of the pandemic. The division of the 3 waves was done as follows:

- 1st wave: March 9, 2020 to July 31, 2020
- 2nd wave: October 24, 2020 to December 6, 2020

- 3rd wave: December 14, 2020 to February 6, 2020

## 3.2 FDA

The first step of the FDA methodology, essential in projects that are based on Functional Data Analysis methods, is to build functions from the available data. The function construction process used in this thesis is based on the structure proposed in the (J. Ramsay et al., 2009) book. As previously mentioned in the Materials section, the COVID-19 data and the Google mobility data (after pre-processing) are time-series data. So, in this first step, the goal is to transform these time series data into functional data.

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

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{4}$$

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

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

Below, this topic will be discussed in more detail, showing how these basis functions, combined with specific coefficients, can be used to effectively create functions.

### 3.2.1 Basis Functions

COVID-19 incidence and mobility time-series data tends to be very sensitive to the behaviour of populations, and their time-series curves may become unpredictable and difficult to estimate. Thus, 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 $\phi_k$ 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 below, which has already been referred to, defines the construction of any function $x(t)$, and is called *basis function expansion*:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \boldsymbol{c}'\boldsymbol{\phi}(t) \tag{5}$$

The parameters $c_1, c_2, \dots, c_k$ are the coefficients of the basis function expansion. In the expression $\boldsymbol{c}'\boldsymbol{\phi}(t)$, $\boldsymbol{c}$ refers to the vector of *K* coefficients and $\phi(t)$ is a vector of length *K* that contains the basis functions. In this work the incidence and mobility datasets have $N$ time series, corresponding to the $N$ municipalities of Portugal mainland, which will be transformed into $N$ functions. So, instead of considering the construction of a single function, the objective is to construct $N$ functions, and the expression (5) is replaced by

$$x_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), i = 1, \dots, N \tag{6}$$

and in this case matrix notation for (6) becomes

$$\mathbf{x}(t) = \mathbf{C}\boldsymbol{\phi}(t) \tag{7}$$

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).

(J. Ramsay et al., 2009) provides us with a simple example of a basis system, that allows a better understanding of this concept

$$Basis\ system\ example: x(t) = 18t^4 - 2t^3 + \sqrt{17}t^2 + \pi/2 \tag{8}$$

This example is a linear combination of the *monomial* basis functions 1, $t$, $t^2$, $t^3$ and $t^4$ with coefficients $\pi/2$, 0, $\sqrt{17}$, -2, and 18, respectively.

But what are the basis functions used in this work? There are several types of basis systems, and some of them include Spline series, Fourier series, constant basis system and monomial basis system. Most of the problems in functional data analysis are solved using Spline and Fourier series, and in this work

Spline Series have been selected to construct the basis functions. In order to understand this choice, these two systems are explained below with more detail, along with the situations in which they should be used.

## Fourier Series

Fourier series are periodic functions, that is, functions that repeat their values at regular intervals T (period). The human blood pressure is an example of a real-life periodic process. To transform a time-series with blood pressure records into a function, a Fourier series is suited to build this function.

The Fourier series is

$$\begin{aligned} \phi_1(t) &= 1 \\ \phi_2(t) &= \sin(\omega t) \\ \phi_3(t) &= \cos(\omega t) \\ \phi_4(t) &= \sin(2\omega t) \\ \phi_5(t) &= \cos(2\omega t) \\ &\dots \end{aligned} \tag{9}$$

where the constant $\omega$ is related to the period T by the relation

$$\omega = 2\pi/T \tag{10}$$

Therefore, a Fourier Series consists of a constant basis function and successive sine/cosine function pairs (9). In each function pair, the argument is multiplied by an integer, starting at 1 and going up to a limit $m$. To create a Fourier basis system, it is required to define the period $T$ and the number of basis functions $K$. Usually, the number of basis systems is $K = 1 + 2m$.
As already mentioned, a periodic function repeats its values at regular intervals $T$. Thus, at the end of each interval $T$, the basis functions of the Fourier basis system created repeat themselves.

## Spline Series

Dealing with unpredictable, possibly non-periodic data makes the process of transforming this information into functions more complex. In this case it is required to work with more flexible basis functions, and that's where splines come from. Splines are polynomials, and their use gives greater flexibility to the construction of functions, allowing to estimate any curve feature.

### 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, which act as boundaries between the intervals. At each of these break points, knots are found, and each break point has at least one knot.

The splines 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 order/degree of the splines functions is fixed, however the behaviour of each polynomial changes depending on the subinterval in which it is found.
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 $order - 2$. However, at the start and end point of the observation interval, as many knots as the order defined for the splines are placed.
Thus, any basis system with order greater than 2 will have at least one derivative matching ($1^{st}$ derivative), and the function will have smooth continuous behaviour at all break points. It is common to resort to splines of order 4 (polynomials of degree 3), where there are two matching derivatives, $1^{st}$ and $2^{nd}$ derivative.
Figure 1 illustrates the role of spline order/degree in function estimation. Three examples of spline functions are presented, which are based on linear combinations of spline basis functions of orders two, three and four, which aim to fit a sine function and its first derivative.
It is possible to observe that the order two spline function presents some weaknesses in estimating the sine function, and that its derivative presents even worse results in estimating the derivative of the sine function. As the order increases, the fit the sine function process is enhanced, creating smoother, better-fitting curves.

$$number\ of\ basis\ functions\ =\ order\ +\ number\ of\ interior\ knots \tag{11}$$

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.

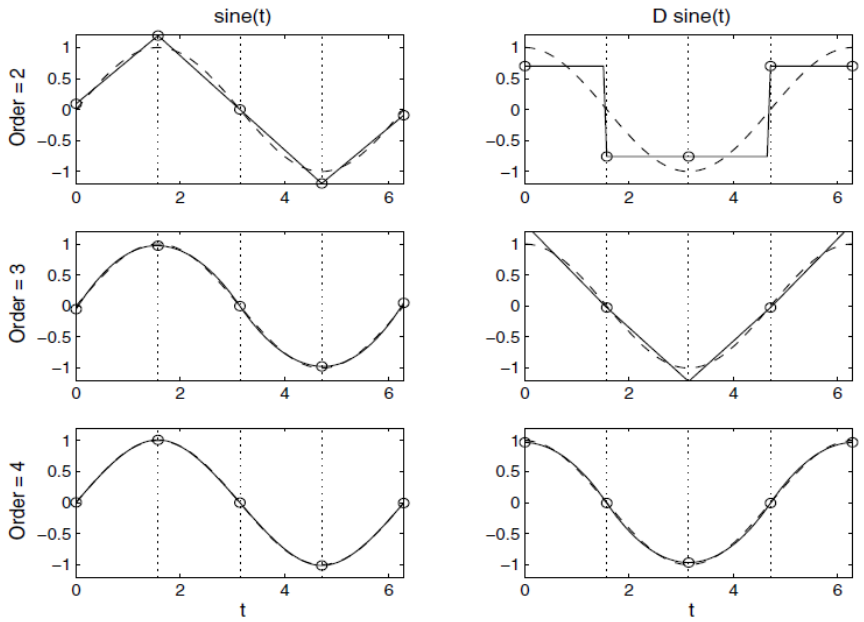



*Figure 1 - Linear combinations of spline basis functions of orders 2, 3 and 4, which aim to fit a sine function and its first derivative* (J. Ramsay et al., 2009)

There are several types of splines, but in this work only B-splines were considered ( Figure 2). Some specific characteristics of B-splines must be referred. In general, each basis function begins at zero and, and when they reach a specific knot, they increase to a peak before decreasing to zero. However, the first and the last basis functions show a different behaviour: they ascend from the first and last interior knot to a value of one on the right and left boundary, respectively, but are otherwise zero.

All of these basis functions are positive over at most four intervals, providing the system with a compact support. This feature can be crucial in computational terms, as it contributes to an increase in the efficiency of the entire function construction process. On basis systems that do not have this property, the computational effort is proportional to $K^2$. In this work, this effort is reduced, becoming proportional to $K$.

However, this type of function also has its drawbacks. In cases of unpredictable and non-periodic data, spline basis functions can sometimes create unstable fits at the beginning and end of the data observation interval. This makes, for example, the derivative estimation more complicated, because the higher the order of the derivative to estimate, the more unstable the behaviour of the functions becomes.

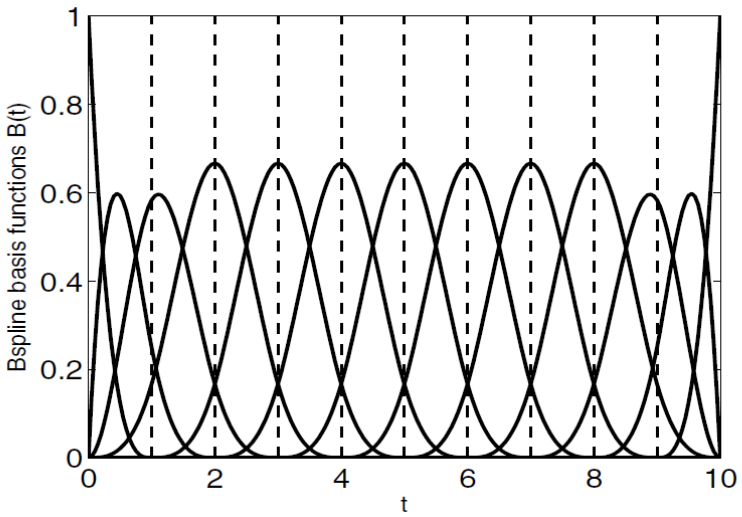



*Figure 2 – Example of representation of 13 B-Splines basis functions* (J. Ramsay et al., 2009)

In addition to the basis systems presented above (Fourier Series and Spline Series), there are other types of basis systems, such as:

- Constant Basis
- Monomial Basis
- Exponential Basis
- Polygonal Basis
- Power Basis

For a detailed description of these types of basis systems, please refer to the works of Kokoszka or Ramsay (Kokoszka & Reimherr, 2017; J. Ramsay et al., 2009).

### 3.2.2 Adding Coefficients to Bases to Define Functions

To build a function it is required to estimate coefficients after setting a basis function system. Once a basis system is set, it is required to supply coefficients in order to estimate the function as a linear combination of basis functions. If there are $K$ basis functions, a coefficient vector of length $K$ for each function is required. To estimate $N$ functions, the coefficients must be arranged in a matrix with dimensions $K$ by $N$, where

- $K$: "number of basis functions"
- $N$: "number of functions or functional observations"

In this work, incidence and mobility curves for 278 municipalities will be computed meaning $N$ = 278 functions. As for the number of basis functions $K$, it depends on the order of polynomials and internal

knots. Break points spaced 7 days apart will be used, 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.

### 3.2.3 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.

Commonly, data fitting is defined as the minimization of the sum of squared errors or residuals

$$SSE(x) = \sum_{j}^{n} \left[ y_j - x(t_j) \right]^2 \tag{12}$$

When the basis function expansion (5) is used to define function $x$, the minimization problem described above becomes

$$SSE(\mathbf{c}) = \sum_{j}^{n} \left[ y_j - \sum_{k}^{K} c_k \phi_k(t_j) \right]^2 = \sum_{j}^{n} \left[ y_j - \phi(t_j)' \mathbf{c} \right]^2 \tag{13}$$

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

$$y_j = x(t_j) + \varepsilon_j = \mathbf{c}' \phi(t) + \varepsilon_j = \phi'(t_j)\mathbf{c} + \varepsilon_j \tag{14}$$

where the true errors or residuals $\varepsilon_j$ are statistically independent and have a normal or Gaussian distribution with mean 0 and constant variance. Despite being a simple error model, the least squares estimation process can be defended on the grounds that it tends to give nearly optimal answers relative

to “best” estimation methods so long as the true error distribution is fairly short-tailed and departures from the other assumptions are reasonably mild. Model (14) is recognized as the standard regression analysis model, along with its associated least-squares solution. Using matrix notation, let the n-vector $y$ contain the $n$ values to be fit, vector $\varepsilon$ contain the corresponding true residual values, and $n$ by $k$ matrix $\Phi$ contain the basis function values $c_k\phi_k(t_j)$. Then

$$\mathrm{y} = \Phi\mathbf{c} + \varepsilon \tag{15}$$

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

$$\hat{\mathbf{c}} = (\Phi'\Phi)^{-1}\Phi'\mathbf{y} \tag{16}$$

The coefficient estimate $\hat{\mathbf{c}}$ in (16) is calculated $\mathrm{y}$ 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:

$$\mathrm{y2cMap} = (\Phi'\Phi)^{-1}\Phi \text{ so that } \hat{\mathbf{c}} = \mathrm{y2cMap}\,\mathbf{y} \tag{17}$$

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 occurs when the constructed function extracts some of the residual variation contained by the data. Overfitting generates less smooth curves, making their analysis difficult, especially if it is necessary to analyse derivatives of these same curves, which can present a very unstable behaviour. Below is described a strategy that allows to build functions with smoother curves, and thus obtain better results for analysis.

## 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. In this case, even with a large number of basis functions compared to the number of observations, the risk of overfitting is reduced, and smooth curves can be obtained. This happens because a roughness penalty is applied that allows to eliminate some of the residual variation contained by the data, smoothing the function.

### 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 $t$. Considering a function that has no curvature, such as a straight line, it is easy to spot the absence of curvature by the fact that its second derivative is zero. Thus, (J. Ramsay et al., 2009) defines the function's roughness as its integrated squared second derivative or its total curvature

$$PEN_2(x) = \int [D^2x(t)]^2 \, dt \tag{18}$$

$PEN_2(x)$ 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.
As mentioned above, it is sometimes essential to analyse, in addition to the function itself, its derivatives. So, it is essential that these same derivatives are smooth, and roughness penalties should also be applied there. In order to analyse, for example, the second derivative $D^2x$ of $x$, the roughness measure used is

$$PEN_4(x) = \int [D^4x(t)]^2 \, dt \tag{19}$$

which will then penalize the curvature of the second derivative.
Having defined the measure of the roughness of the fitted curves, the goal is now to minimize a fitting criterion that strikes a balance between curve roughness and underfitting.

Whatever roughness penalty used, a multiple of it is added to the error sum of squares to define the fitting criterion. For example, using $PEN_2(x)$ the fitting criterion will be:

$$F(\mathbf{c}) = \sum_j \left[y_j - x(t_j)\right]^2 + \lambda \int [D^2 x(t)]^2 dt \qquad (20)$$

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

The smoothing parameter $\lambda$ controls the amount of importance placed on each of the two competing goals: as $\lambda$ approaches infinity, curvature becomes increasingly penalized, and the curve approaches the straight line obtained from a linear regression, since $D^2 x$ will be essentially 0. On the other hand, as $\lambda$ approaches zero, the curve approaches to a direct interpolation of the data, since the function is free to fit the data as closely as possible with observed data, it can lead to curves with very large and sudden variations. More generally, differential operator $L$ can be used to define roughness. In this case, it will result in $\lambda \to \infty$ forcing the fit to approach more and more closely a solution to the differential equation $Lx = 0$. If $L = D^m$, this solution will be a polynomial of order $m$.

Thus, $\lambda$ allows you to effectively control the smoothing process, giving us the opportunity to define the meaning of smooth depending on the different problems that may arise.

**The Roughness Penalty Matrix R**

Bearing in mind that the regression smoothing technique discussed above did not consider the use of roughness penalties, there is a need to adapt this methodology to the roughness penalty smoothing, providing a new way to estimate the coefficient vector $\hat{\mathbf{c}}$.

The roughness penalized fitting criterion (20) is generally defined as

$$F(\mathbf{c}) = \sum_j \left[y_j - x(t_j)\right]^2 + \lambda \int [Lx(t)]^2 dt \qquad (21)$$

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

$$F(\mathbf{c}) = \sum_j \left[y_j - \phi'(t_j)\mathbf{c}\right]^2 + \lambda \mathbf{c}' \left[\int L\phi(t) L\phi'(t) dt\right] \mathbf{c} \quad (22)$$

The order K roughness penalty matrix is defined as

$$\mathbf{R} = \int \phi(t)\phi'(t)dt \quad (23)$$

From this, it is possible to define the coefficient vector $\hat{\mathbf{c}}$ as

$$\hat{\mathbf{c}} = (\Phi'\Phi + \lambda \mathbf{R})^{-1}\Phi'\mathbf{y} \quad (24)$$

As before, it is essential to define the matrix $\mathrm{y2cMap}$. This matrix is used for computing confidence regions for the estimated functions, and is obtained through the following expression

$$\mathrm{y2cMap} = (\Phi'\Phi + \lambda \mathbf{R})^{-1}\Phi' \quad (25)$$

## Choosing Smoothing Parameter $\lambda$

The generalized cross-validation measure GCV is designed to locate the best value for smoothing parameter $\lambda$. This criterion is defined as follows

$$\mathrm{GCV}(\lambda) = \left(\frac{n}{n - \mathrm{df}(\lambda)}\right)\left(\frac{\mathrm{SSE}}{n - \mathrm{df}(\lambda)}\right) \quad (26)$$

This is a twice-discounted mean square error measure. The right factor is the unbiased estimate of error variance $\sigma^2$, and thus represents some discounting by subtracting $\mathrm{df}(\lambda)$ from $n$. The left factor further discounts this estimate by multiplying by $n/(n - \mathrm{df}(\lambda))$.

## 3.3 Linear Models

Linear modelling is an approach for modelling the relationship between a response variable and one or more explanatory variables (or covariates). In this work, the response variables are the COVID-19 incidence curves, and the explanatory variables are the mobility data from Google or the ones provided by sociodemographic data. In both cases, the dependent or response variable is functional.
First, the situation where the explanatory variables are scalar (sociodemographic data) is considered. Then, the case when the explanatory variable is functional (Google mobility curves) is addressed. In the latter, two variants of the Concurrent Model are used. For the simpler concurrent model, the value of the response variable $y(t)$ is predicted only by the value of the functional covariate at the same time $t$. In the more general (and complex) concurrent model, the functional variable contribute to the prediction for all possible time values.

### 3.3.1 Functional Responses with Scalar Covariates: Analysis of Variance Model

Here, variation in a functional response (COVID-19 daily cumulative 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

Using the sociodemographic data, municipalities were categorized into distinct groups. For example, in the case of the proportion of elderly population, municipalities were categorized in three groups, in which the first group contains the municipalities with the lowest percentages of elderly population and the third group contains the municipalities with the highest percentage. This way, it is possible to discuss the influence that the elderly population has on COVID-19 incidence by fitting a model of the form

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{3} x_{ij}\beta_j(t) + \varepsilon_i(t) \qquad (27)$$

where $y_i(t)$ is a *functional response*. In this case, the values of $x_{ij}$ are either 0 or 1. If the 278 by 4 matrix $\mathbf{Z}$ contains these values, then the first column has all entries equal to 1, which defines the contribution of the mean COVID-19 incidence curve; the remaining three columns contain 1 if that municipality is in the corresponding group and 0 otherwise. In addition, it is necessary to add a constraint, so that the effects of the three sociodemographic groups can be identified, which is defined as follows

$$\sum_{j=1}^{3} \beta_j(t) = 0 \text{ for all } t \quad (28)$$

In order to implement this constraint, the above equation is added to the original data as an additional 279*th* "observation" for which $y_{279}(t) = 0$.

First, a list is created, containing four indicator variables, for the intercept term and each of the groups. Applying this structure, the intercept term 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 next step, as required by (28) is to expand the COVID-19 incidence functional data object adding a 279*th* observation that takes only zero values. Then *fRegress*, an R function that performs linear regression, is called. The coefficients are extracted and plotted.

Considering the specificities of the Analysis of Variance Model, it was necessary, for each sociodemographic variable, to group municipalities into tertiles (except for Deprivation Index, which already grouped municipalities into quintiles).

### 3.3.2 Functional Responses with Functional Covariates: Concurrent Model

In order to use expression (27) for functional covariates, it can be expanded it as follows:

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{q-1} x_{ij}(t)\beta_j(t) + \varepsilon_i(t) \quad (29)$$

where $x_{ij}(t)$ is intended be a functional observation. Notwithstanding $x_{ij}$ may also be a scalar observation or a categorical indicator, simply by considering that the function $x_{ij}(t)$ is constant over time. The model (29) 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, multiplies a scalar covariate whose value is always one, and 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 $\beta_j$ are estimated, using the R function fRegress, by simplifying the problem and transforming it into the resolution of a group of linear equations. The coefficient matrix defining this linear system can then be analyzed to detect and diagnose problems related to multicollinearity among the functional covariates and the intercept. Multicollinearity (or concurvity if more than one functional covariate is involved) may generate instability and imprecision on estimating the linear coefficients, also making it more difficult to see what relative importance each covariate has in predicting the dependent variable.

Consider that the $N$ by $q$ functional matrix $\mathbf{Z}$ contain the $x_{ij}$ functions, and that the vector coefficient function $\beta$ of length $q$ contain each of the regression functions. The concurrent functional linear model in matrix notation is then

$$\mathbf{y}(t) = \mathbf{Z}(t)\beta(t) + \varepsilon(t) \tag{30}$$

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

$$\mathbf{r}(t) = \mathbf{y}(t) - \mathbf{Z}(t)\beta(t) \tag{31}$$

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

$$\mathrm{LMSSE}(\beta) = \int \mathbf{r}(t)'\mathbf{r}(t)dt + \sum_{j}^{p} \lambda_j \int \left[L_j\beta_j(t)\right]^2 dt \tag{32}$$

Consider now that the regression function $\beta_j$ has the expansion

$$\beta_j(t) = \sum_{k}^{K_j} b_{kj}\theta_{kj}(t) = \theta_j(t)'\mathbf{b}_j \tag{33}$$

in terms of $K_j$ basis functions $\theta_{kj}$. In order to express (30) and (32) in matrix notation referring explicitly to these expansions, a composite is constructed.
Defining $K_\beta = \sum_j^q K_j$ the vector $\mathbf{b}$ of length $K_\beta$ is constructed by assembling the vectors vertically, that is,

$$\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_q)' \tag{34}$$

Now, the matrix function $\Theta(t)$ is assembled as below:

$$\Theta(t) = \begin{bmatrix} \theta_1(t)' & 0 & \cdots & 0 \\ 0 & \theta_2(t)' & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \theta_q(t)' \end{bmatrix} \tag{35}$$

and $\beta(t) = \Theta(t)\mathbf{b}$. Knowing this, (31) can be adapted as follows:

$$\mathbf{r}(t) = \mathbf{y}(t) - \mathbf{Z}(t)\Theta(t)\mathbf{b} \tag{36}$$

Considering $\mathbf{R}(\lambda)$ as the block diagonal matrix with $j^{th}$ block as follows:

$$\lambda_j \int \left[L_j\theta_j(t)\right]'\left[L_j\theta_j(t)\right]dt \tag{37}$$

Then (32) can be defined as follows:

$$\mathrm{LMSSE}(\beta) = \int [\mathbf{y}(t)'\mathbf{y}(t) - 2\mathbf{b}'\Theta(t)'\mathbf{Z}(t)'\mathbf{y}(t) + \mathbf{b}'\Theta(t)'\mathbf{Z}(t)'\mathbf{Z}(t)\Theta(t)\mathbf{b}]dt + \mathbf{b}'\mathbf{R}(\lambda)\mathbf{b} \tag{38}$$

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:

$$\left[\int \Theta(t)'\mathbf{Z}(t)'\mathbf{Z}(t)\Theta(t)dt + \mathbf{R}(\lambda)\right]\hat{\mathbf{b}} = \left[\int \Theta(t)'\mathbf{Z}(t)'\mathbf{y}(t)dt\right] \tag{39}$$

This is a linear matrix equation defining the scalar coefficients in vector $\hat{\mathbf{b}}$, $\mathbf{A}\hat{\mathbf{b}} = \mathbf{d}$ , where the normal equation matrix is

$$\mathbf{A} = \int \Theta(t)'\mathbf{Z}(t)'\mathbf{Z}(t)\Theta(t)dt + \mathbf{R}(\lambda) \tag{40}$$

and the right-hand side vector of the system is

$$\mathbf{d} = \int \Theta(t)'\mathbf{Z}(t)'\mathbf{y}(t)dt \tag{41}$$

**Confidence Intervals for Regression Functions**

When using regression functions, confidence intervals are important for us to assess the quality of the estimates made. These intervals have 95% pointwise confidence and are generated using the function *fRegress.stderr*, that belongs to the fda package. The results obtained, that is, the linear coefficients with confidence intervals, are then plotted using the function *plotbeta*.

### 3.3.3 Functional Responses with Functional Covariates: General Concurrent Model

The concurrent linear model relates the value of a functional response to the current value of functional covariate(s). A general version for a functional covariate and an intercept (from now on referred to as General Concurrent Model) is

$$y_i(t) = \beta_0(t) + \int_{\Omega_t} \beta_1(t,s)x_i(s)ds + \varepsilon_i(t) \quad (42)$$

The bivariate linear coefficient function $\beta_1(t,s)$ defines the dependence of $y_i(t)$ on covariate $x_i(s)$ at each time $t$, and in this case, it is not necessary for $x_i(s)$ and $y_i(t)$ to be defined over the same range or continuum.

$\Omega_t$, the set of values to which the integration in (42) is calculated, comprises the range of values of argument $s$ over which $x_i$ is considered to influence response $y_i$ at time $t$. The subscript $t$ on $\Omega_t$ implies that this set can change from one value of $t$ to another. This can lead to backwards causation, since in this case both $s$ and $t$ are time, and $x_i(s)$ can be used to predict $y_i(t)$ when $s > t$. To prevent this from happening, it is necessary to only allow values $x_i$ to be considered before the time $t$ for which the prediction is to be made (also imposing a lower limit). Thus, this constraint takes the form of the following integral

$$\Omega_t = \{s | t - \delta \leq s \leq t\} \quad (43)$$

where $\delta > 0$ specifies how much of the previous time period is important to the prediction.

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

Let $x_i(t)$ represent daily Google mobility value in day $i$ and $y_i(t)$ represent daily COVID-19 incidence value in day $i$. The following model is proposed

$$y_i(t) = \beta_0(t) + \int \beta_1(s,t)x_i(t)ds + \varepsilon_i(t) \quad (44)$$

That is, for any day within the defined time period, the COVID-19 incidence is modelled for that day using as the functional covariate the Google mobility data for the previous days.

The regression function $\beta$ has the basis function expansion

$$\begin{aligned}\beta_1(s,t) &= \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} b_{kl}\phi_k(s)\psi_l(t) \\ &= \phi'(s)\mathbf{B}\psi(t)\end{aligned} \tag{45}$$

where the coefficients for the expansion are in the $K_1$ by $K_2$ matrix **B**. Therefore it is required to define two bases for $\beta_1$, as well as a basis for the intercept function $\beta_0$.

For a bivariate function such as $\beta_1(t,s)$ smoothness can be imposed by penalizing the $s$ and $t$ directions separately:

$$\mathrm{PEN}_{\lambda_t,\lambda_s}\big(\beta_1(t,s)\big) = \lambda_1[L_t\beta_1(t,s)]^2 ds\, dt + \lambda_2[L_s\beta_1(t,s)]^2 ds\, dt \tag{46}$$

where linear differential operator $L_s$ only involves derivatives with respect to $s$ and $L_t$ only involves derivatives with respect to $t$. A penalty to the roughness of the intercept $\beta_0$ can also be applied.

A B-spline basis is defined and used to define functional parameter objects for $\beta_0$ , $\beta_1(\cdot,t)$ and $\beta_1(s,\cdot)$ . The coefficients are penalized (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.

# Chapter 4

# Results

The evolution of the incidence and mobility curves along the 3 waves of the pandemic, both at national and municipal level, can be observed in the following figures (Figure 3 and Figure 4). There is a great heterogeneity between the curves of the municipalities (both in terms of incidence and mobility), which are represented in grey. Furthermore, it can be seen through the analysis of the national mobility curves that there is a big difference in the behaviour of the various types of mobility. The mobility variation observed in Figure 4 represent percentage of variation in relation to a baseline, represented by the horizontal axis $y$=0

This heterogeneity and difference in the behaviour of the curves contributes to obtaining distinct (and also heterogeneous) results, as it will be possible to observe in the following subsections of this chapter.

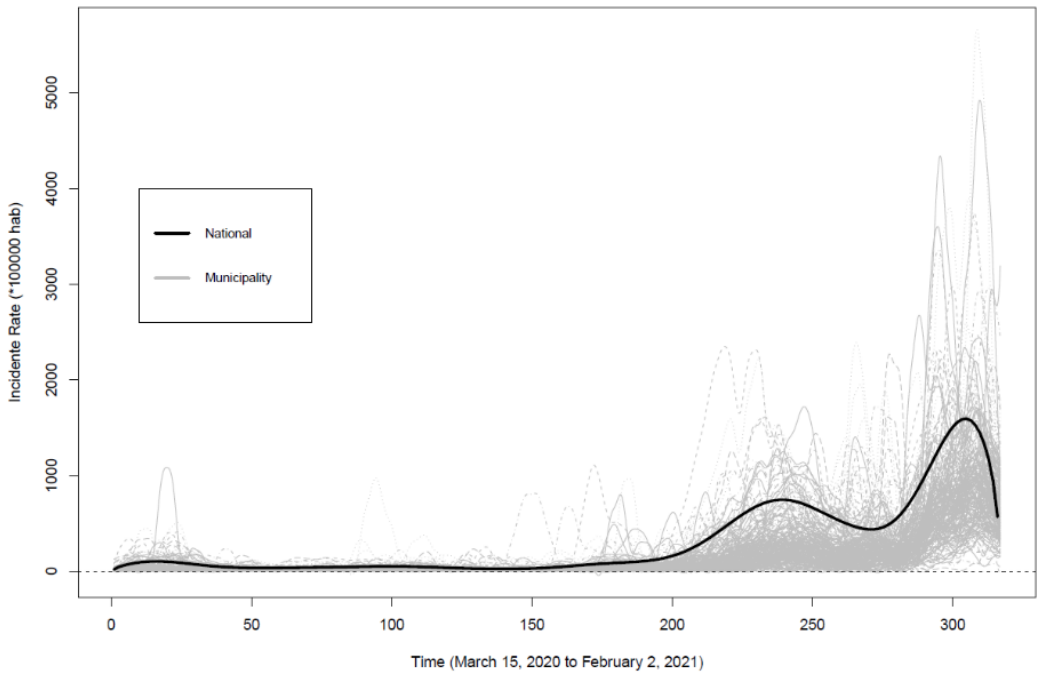



*Figure 3 - Incidence Rate Curves (National and Municipalities)*

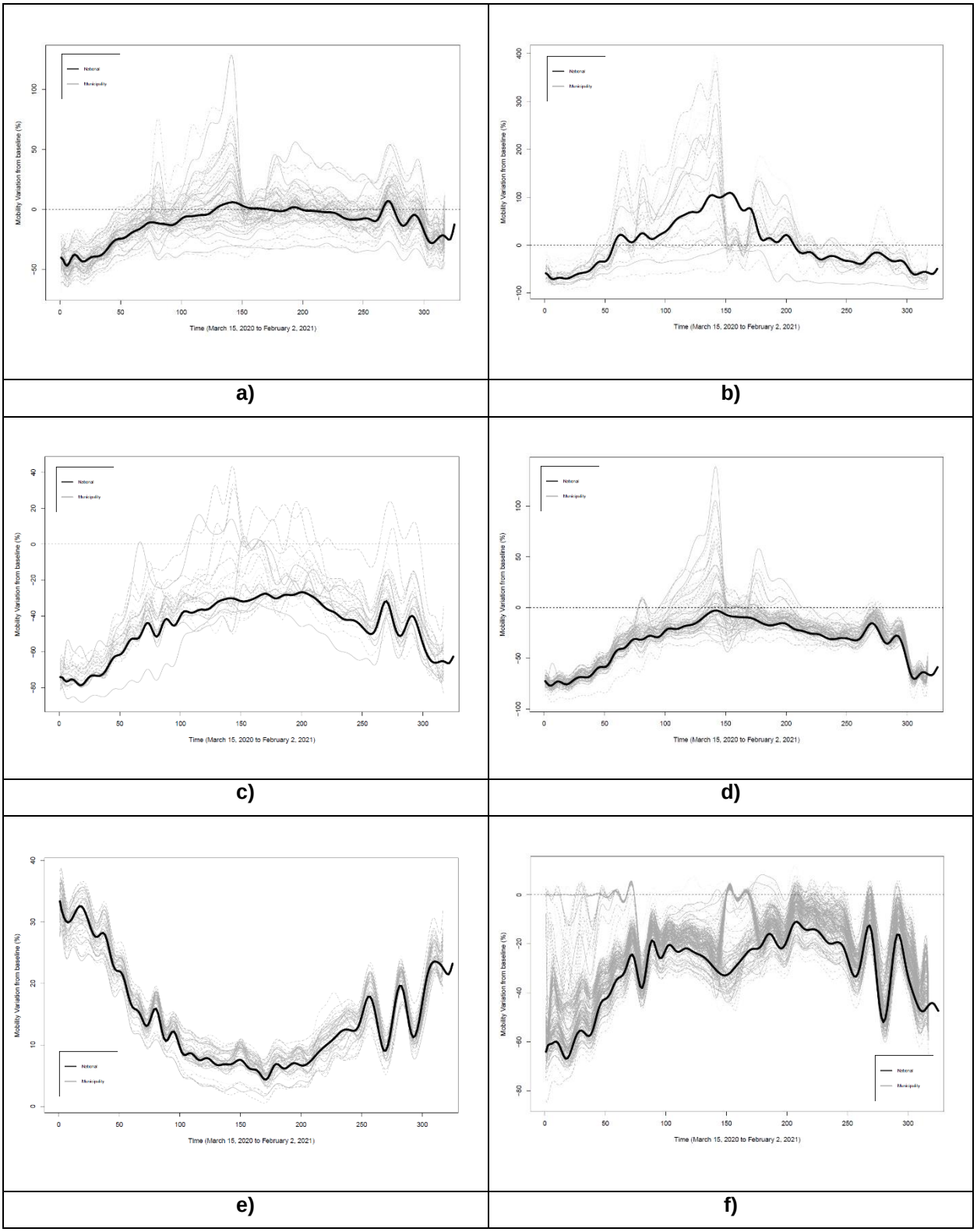



*Figure 4 - Mobility Variation Curves (National and Municipalities): a) Grocery b) Station c) Parks d) Retail e) Residential f) Workplace*

## 4.1 Functional Responses with Functional Covariates: Concurrent Model

The concurrent model was used to model the relationship between COVID-19 incidence rate acceleration (or deceleration) curves and velocity of mobility variation.

The results obtained using this method show the evolution of the linear coefficient β over a given period of time. If this coefficient is positive, it means that at that instant $t$ the two variables are related (the velocity of mobility variation influences the incidence rate acceleration).

In order to analyse the effects of mobility on incidence, it was necessary to apply a lag between the curves of both variables (because, realistically, this effect is not instantaneous, but lagged). The main objective here was to try to understand in which mobility classes showed positive linear coefficients throughout the evaluated period. In other words, to detect a similar behaviour in the velocity of mobility variation curves and in the incidence rate acceleration curves. For example, applying a lag of $x$ days allowed to assess the relationship between the velocity of mobility variation curve at time $t$, with the incidence rate acceleration curve at time $t + x$. This lag enabled to see whether or not mobility has an influence on incidence, and after how long this influence manifests itself. Here, the graphs show the linear coefficient curve, with 95% pointwise confidence intervals, where the upper curve is the upper limit of the interval, and the lower curve is the lower limit of the interval.

The results for the concurrent models fitted for the first wave and second waves showed no relevant relationships between incidence and mobility, as they revealed curves that are unstable and difficult to analyse and the application of the concurrent model to the 1st wave did not return any results. Figure 5 shows an example of a curve obtained in the 2nd wave, where the instability of the linear coefficient curve is evident.

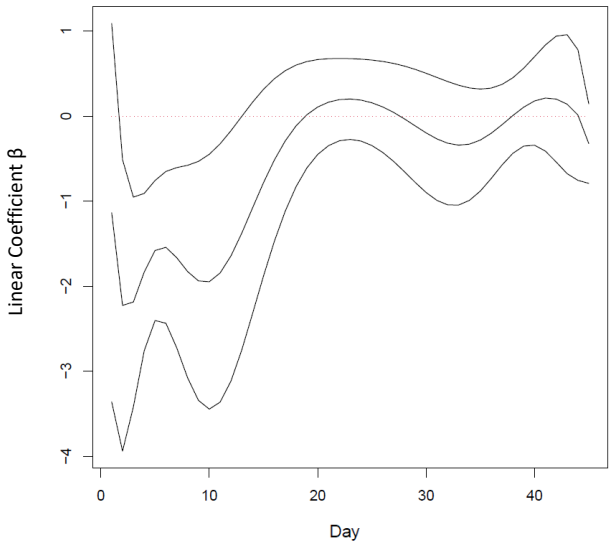



*Figure 5 - Linear Coefficient Function for the Velocity of Mobility Variation with 95% pointwise confidence intervals (Residential Class, 2nd wave, 16-day lag)*

The plots in Figure 6 show coefficients estimates obtained with the concurrent model for mobility (y-axis) along time (x-axis) within the 3rd wave. These coefficients with positive sign were estimated for Grocery (15-day delay) (Figure 6 a), Parks (16-day delay) (Figure 6 b), and Stations classes (15/16-day delay) (Figure 6 c and d). This means that an increase (decrease) of mobility rates tended to be associated with an acceleration (deceleration) in incidence rates. Despite this, the confidence intervals of these coefficient curves present an inferior limit with negative values, reducing the reliability of the results.

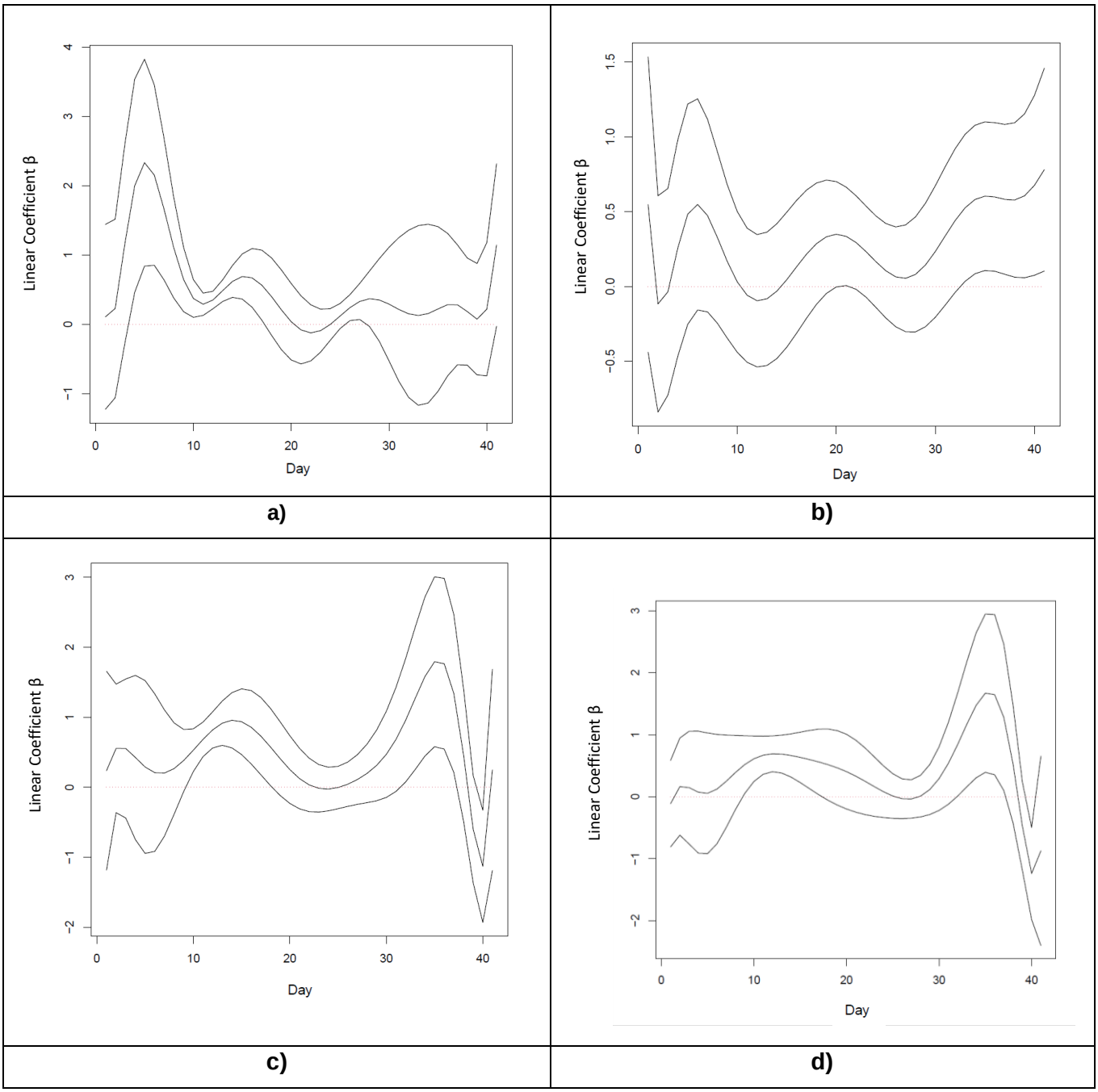



*Figure 6 - Linear Coefficient Function for the Velocity of Mobility Variation in the 3rd wave with 95% pointwise confidence intervals: a) Grocery (15-day lag) b) Parks (16-day lag) c) Stations (15-day lag) d) Stations (16-day lag)*

The strongest signal was found in the Residential class (16-day delay) (Figure 7 a). Here, throughout the 3rd wave, the estimated coefficient is always positive, with a confidence interval also always above zero. In the Retail class (15-day delay) (Figure 7 b) the behaviour is similar, although not as strong, as the confidence interval has a negative section.

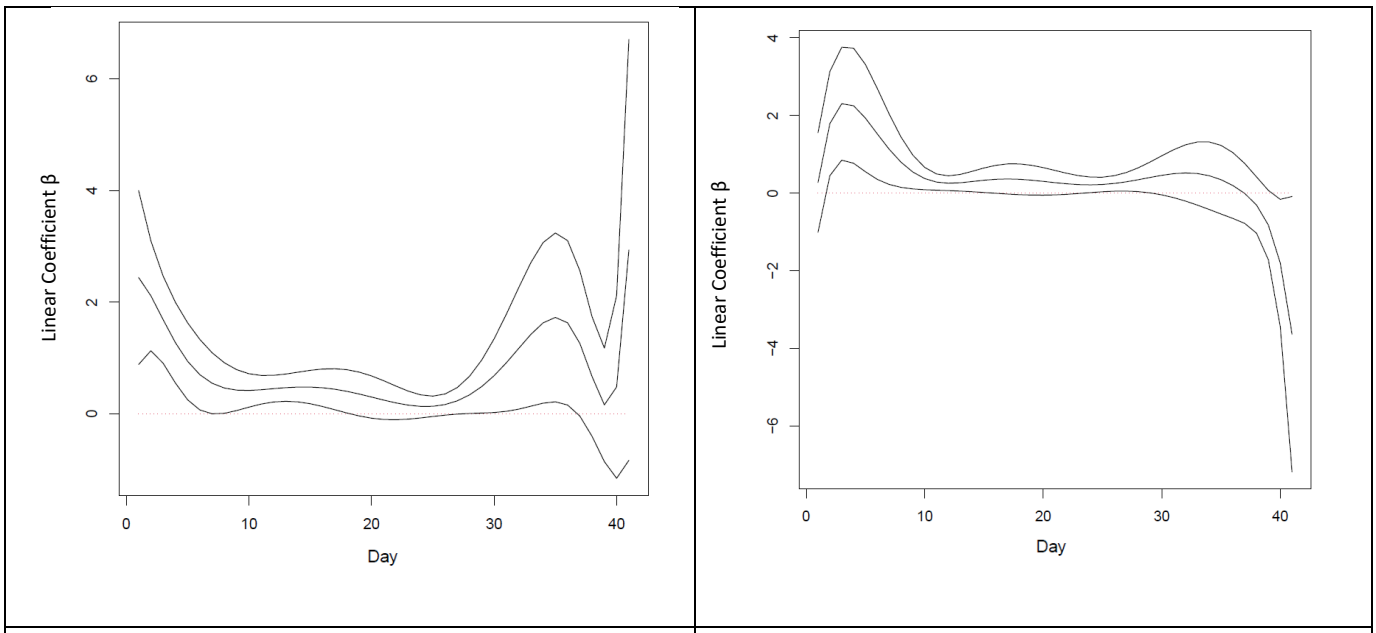



*Figure 7 - Linear Coefficient Function for the Velocity of Mobility Variation in the 3rd wave with 95% pointwise confidence intervals: a) Residential (16-day lag) b) Retail (15-day lag)*

Finally, in the Workplace class it was not possible to find a significative association during the 3rd wave (15-day delay) (Figure 8) as the curve oscillates between positive and negative values.

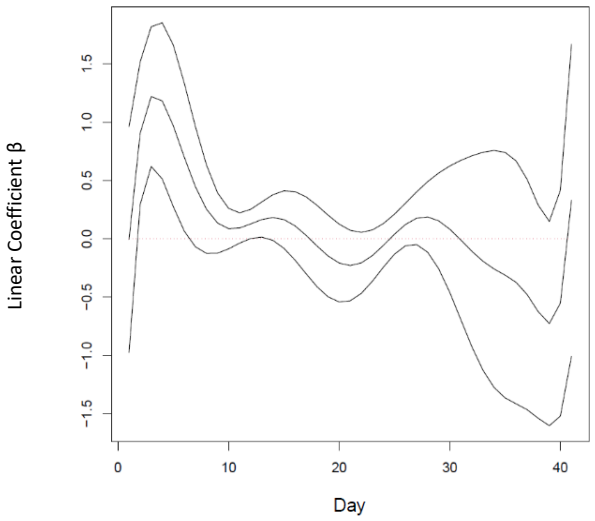



*Figure 8 - Linear Coefficient Function for the Velocity of Mobility Variation with 95% pointwise confidence intervals (Workplace Class, 3rd wave, 15-day lag)*

## 4.2 Functional Responses with Functional Covariates: General Concurrent Model

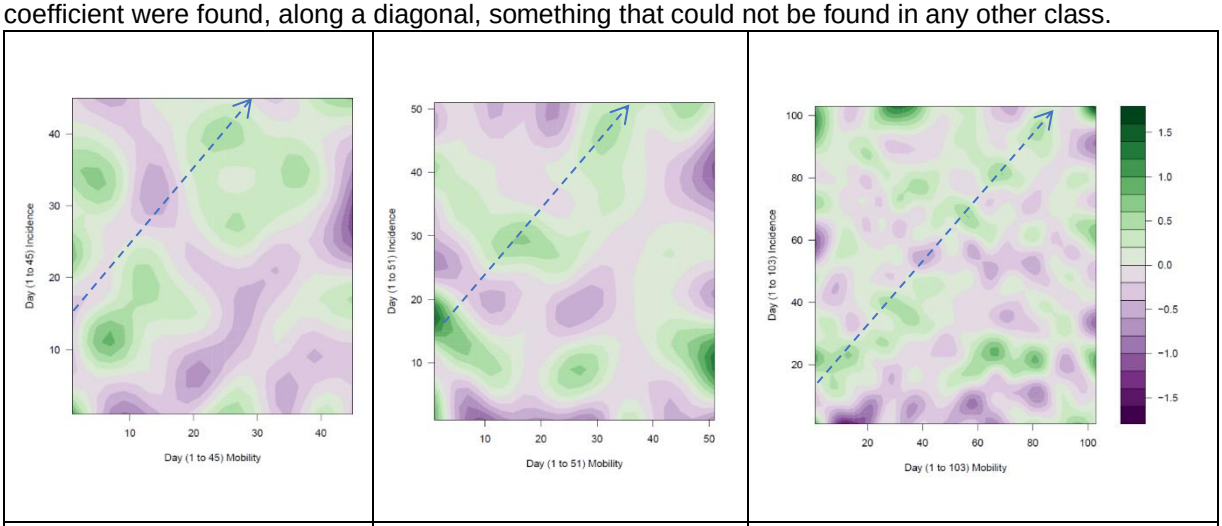
This method, similar to the previous method, explores the relationship between COVID-19 incidence rate acceleration curves (response variable) and velocity of mobility variation curves (explanatory variable). However, this method allows us to evaluate possible correlations of these variables without having to resort to lagged curves. This is because this method automatically relates the value of the response variable to all the values of the explanatory variable.

In this method, as in the previous section, the variable under analysis is the linear coefficient β. In order to facilitate the process, the analysis of the graphics is supported by a colour scale. The green colour is associated with a positive coefficient, and the purple colour with a negative coefficient. The x-axis (horizontal) is the temporal axis of mobility, and the y-axis (vertical) is the temporal axis of incidence. If at some point (x,y), the linear coefficient is positive (green colour), this means that velocity of mobility variation and the incidence rate acceleration are related.

As mentioned in the previous section, no results were found for the 1st wave. For the 2nd wave an interesting result was found, but overall, this wave provided few satisfactory results.

For Residential Mobility, in 2nd wave (Figure 9 a), few points of positive linear coefficient was found along a diagonal lagged by about 15 days (only at (25,40)), with the positive signal being weaker and more dispersed than that for the 3rd wave. The strongest signal for Residential Mobility was found in the 3rd wave. The 3rd wave graph (Figure 9 b) shows us that the linear coefficient is positive along the diagonal (but not always) referred above, that starts at (0,15) and ends at (35,50)) The dashed line used in some figures in this section illustrates the diagonal path with 15-day lag, where prevalence of positive linear coefficient functions is expected, in order to facilitate the analysis.

Also, analysing the junction of the 2nd and 3rd waves (Figure 9 b), several points of positive linear coefficient were found, along a diagonal, something that could not be found in any other class.




| a) | b) | c) |
|---|---|---|

*Figure 9 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Residential Class): a) 2nd wave b) 3rd wave c) 2nd and 3rd wave*

In the Grocery class the signal is considerably weaker, in both the 2nd and 3rd waves (Figure 10 a and Figure 10 b), although some points could be observed along the “diagonal” where the coefficient is positive.

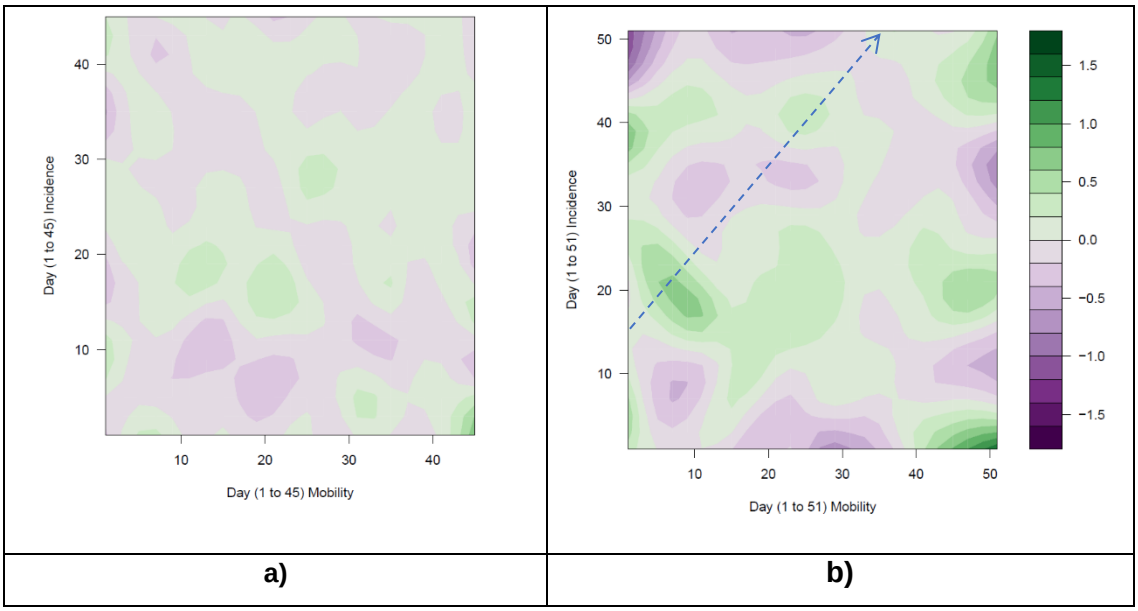



*Figure 10 – Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Grocery Class): a) 2nd wave b) 3rd wave*

In this Parks class, the signal is slightly stronger in the 2nd wave (Figure 11 a), where it is observed that the coefficient is positive at some points along the “diagonal” (for example between (20,35) and (29,30)), as opposed to the 3rd wave (Figure 11 b), where the signal is slightly weaker. Results similar to the previous section.

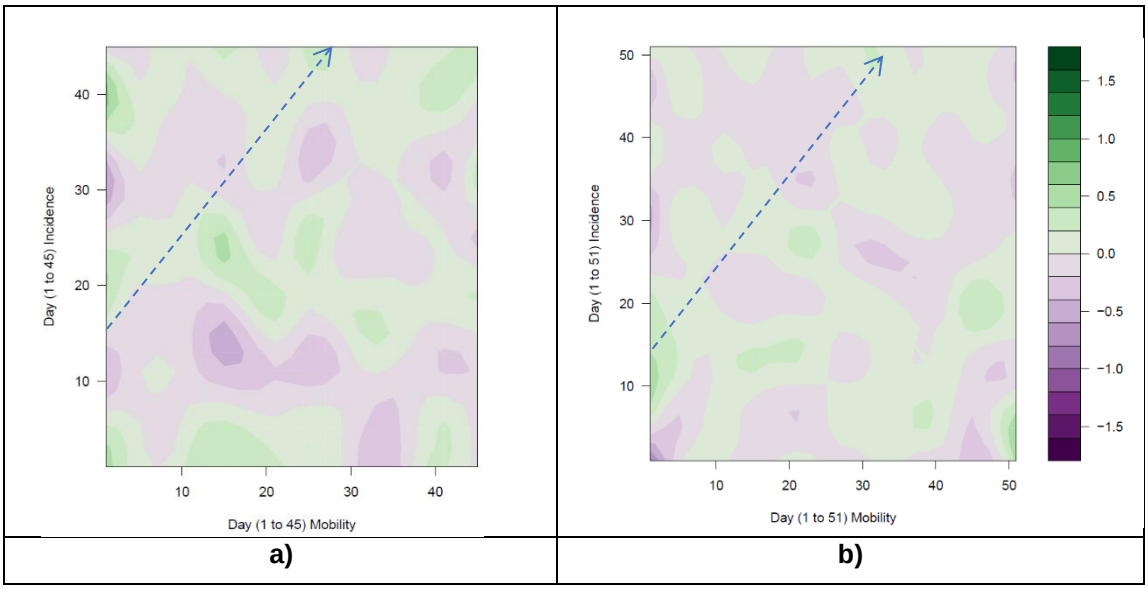



*Figure 11 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Parks Class): a) 2nd wave b) 3rd wave*

Only in the 3rd wave of the Retail class (Figure 12 a) positive coefficients could be found along the diagonal, unlike the 2nd wave (Figure 12 b), stronger at the beginning of the wave (between (0,15) and (20,35)).

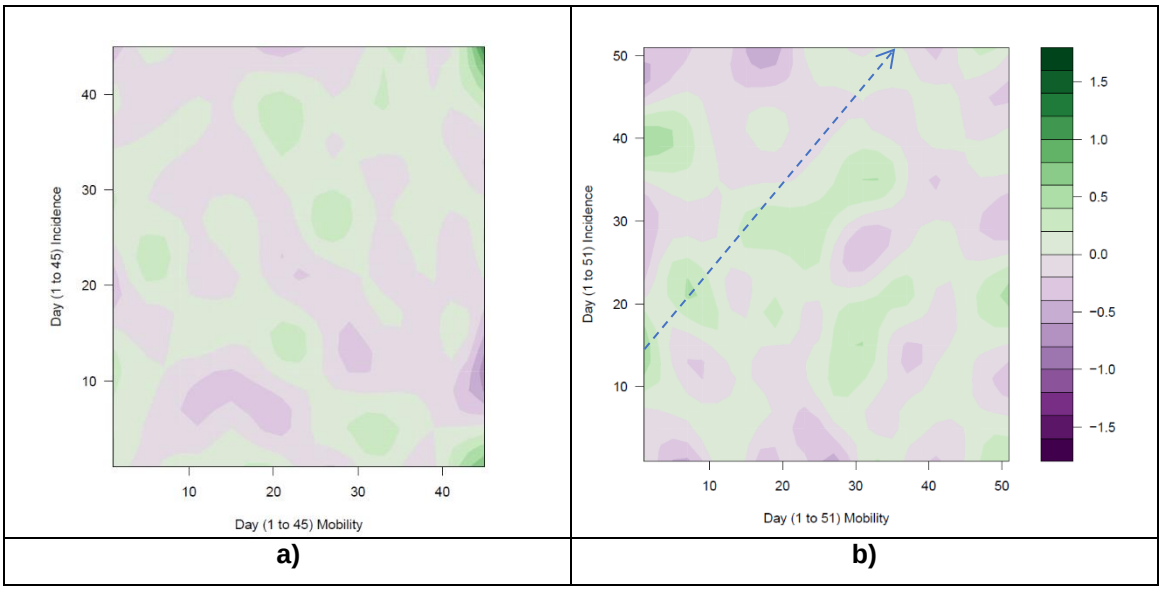



*Figure 12 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Retail Class): a) 2nd wave b) 3rd wave*

In the Stations class, graphs show us that, in the 2nd wave (Figure 13 a) and specially in the 3rd wave (Figure 13 b), there are regions of the graph that show a positive linear coefficient, along a diagonal. In the 3rd wave, the linear coefficient is positive between (0,15) and (35,50), showing that the variables are related through the entire period

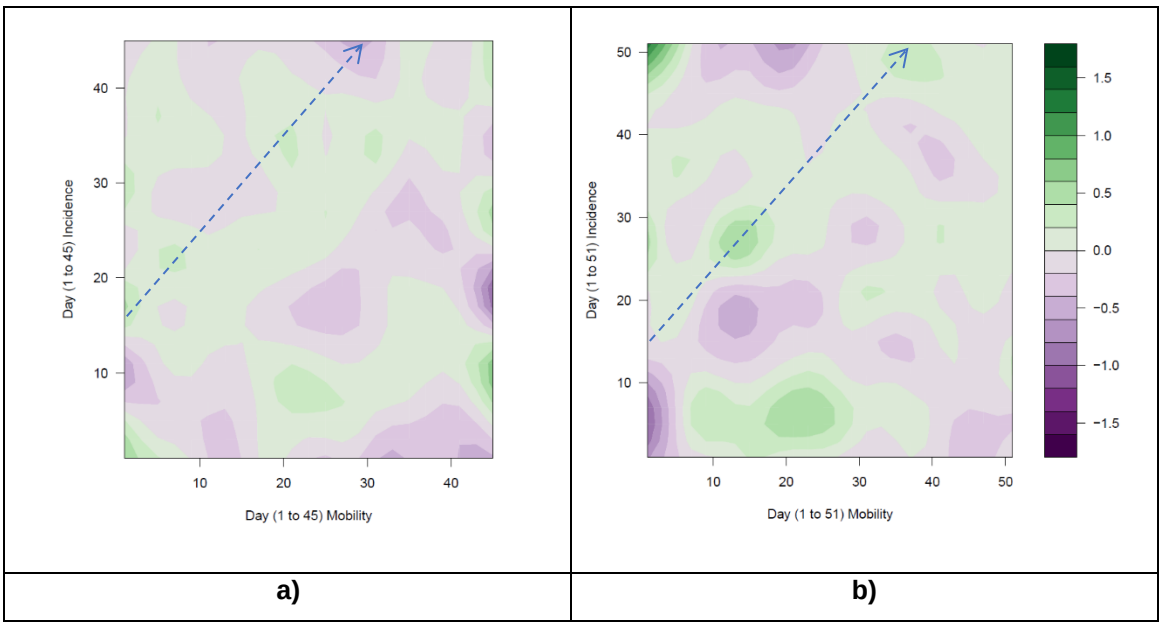



*Figure 13 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Stations Class): a) 2nd wave b) 3rd wave*

In the Workplace class, it was found some positive linear coefficient in the 3rd wave (Figure 14 b) in the diagonal represented by the dashed line (between (0,15) and (30,45)), although it is scattered and is a very weak signal.

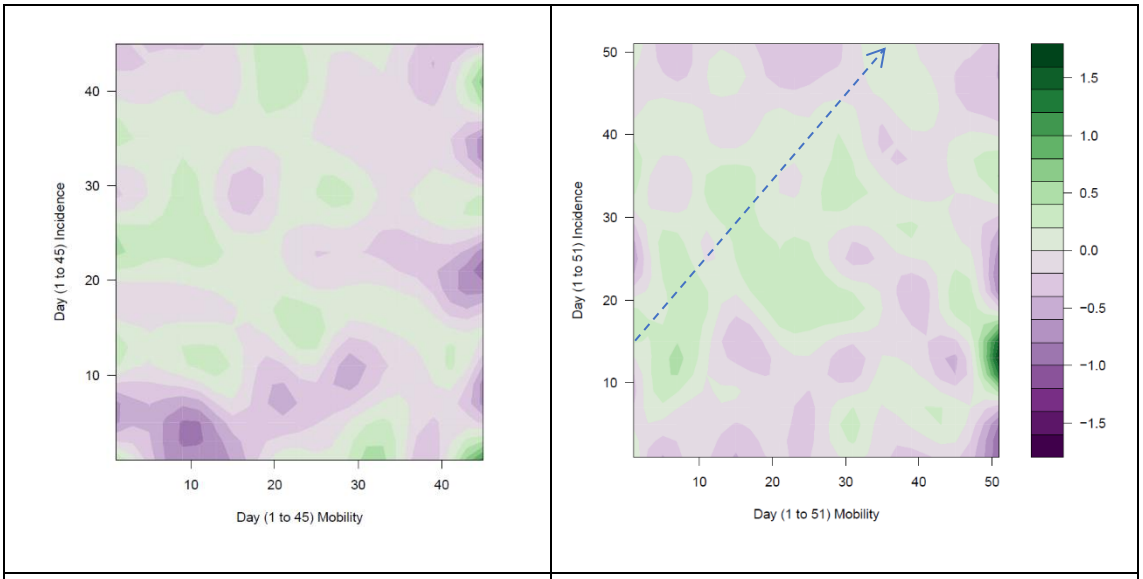



*Figure 14 - Bivariate Linear Coefficient Function for the Velocity of Mobility Variation (Workplace Class): a) 2nd wave b) 3rd wave*

## 4.3 Functional Responses with Scalar Covariates: Analysis of Variance Model

Here, the objective was to model the relationship between a scalar response variable and a functional response variable. The response variables are the COVID-19 daily incidence curves, and the explanatory variables is sociodemographic data.

For each of these sociodemographic variables, the results show us several graphs (4 to 6 graphs). The first graph corresponds to the COVID-19 mean incidence curve and is the same for all variables. As mentioned in the methodology, for each variable, municipalities are divided into classes, and each class corresponds to a graph. These graphs, whose curves represent linear coefficients, correspond to the perturbation of the COVID-19 incidence mean required to fit the class's mean COVID-19 incidence curve. What does this mean in practice? When a curve is above 0 (that is, the linear coefficient is positive) this means that, in the period in which the curve is positive, the municipalities in the same class had a COVID-19 incidence mean higher than the total average.

Likewise, if the value of the linear coefficient were negative, it means that the municipalities in the same class had a COVID-19 incidence mean lower than the total average.

The graphs obtained here are based on the joint analysis of the 2nd and 3rd waves of the pandemic, which in this case corresponds to a total of 106 days. Thus, in order to facilitate the analysis of these results, it is worth noting that:

- 2nd wave: October 24, 2020 (Day 0) to December 13, 2020 (Day 52)
- 3rd wave: December 14, 2020 (Day 52) to February 6, 2020 (Day 107)

Analysing the Population Density graphs (Figure 15), some interesting details were observed. It is possible to notice that the municipalities with lower population density (Class 1) were the most affected by COVID-19 during the 3rd wave period. On the other hand, in the municipalities with higher population density (Class 3), the opposite was verified: these municipalities were the most affected during the period of the 2nd wave, presenting curves below the average in the 3rd wave.

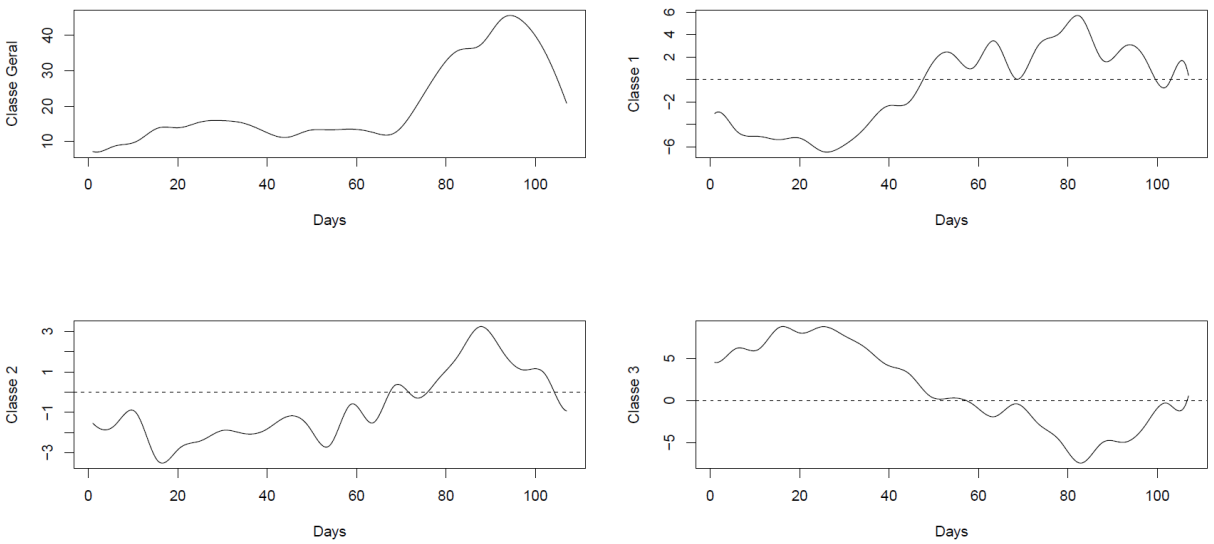



*Figure 15 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Population Density*

Turning now to Schools (Figure 16), the results show us a behaviour identical to that seen in the analysis of coefficients for population density and youth population. In this case, Class 1 corresponds to municipalities with the lowest number of schools/km2, and Class 3 corresponds to municipalities with the highest number of schools/km2. This similarity makes sense, given that municipalities with a higher percentage of youth population will tend to have more schools.

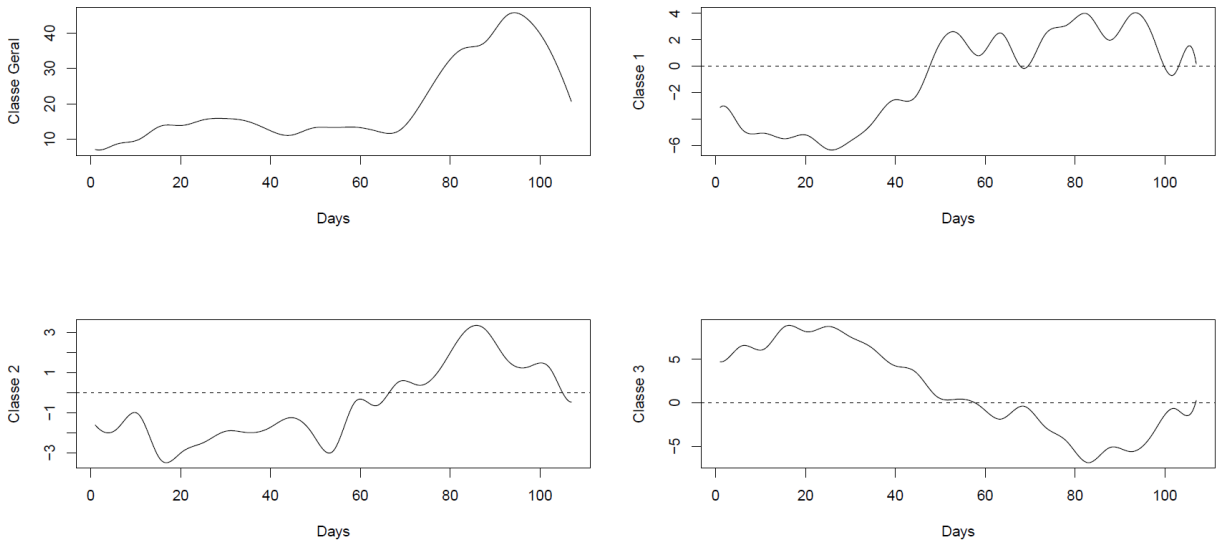



*Figure 16 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Schools Density*

In the case of the Deprivation Index (Figure 17), class 1 corresponds to least deprived municipalities, and class 5 corresponds to most deprived municipalities (with several classes in between). Contrary to expectations, class 5 municipalities had a COVID-19 incidence mean higher than the country average, throughout the 2nd and 3rd wave.

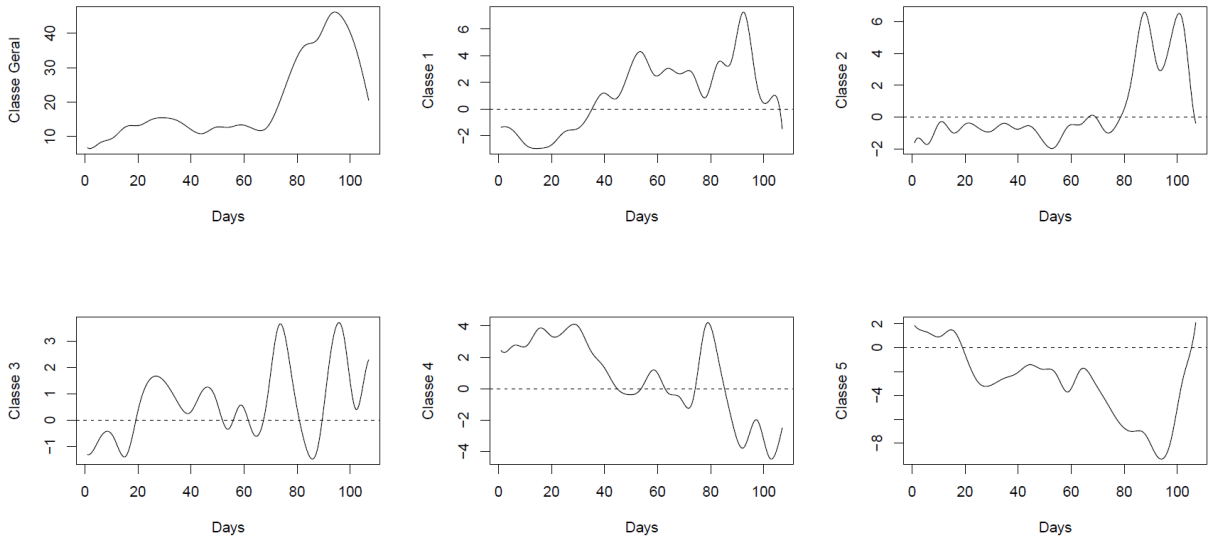



*Figure 17 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Deprivation Index*

The graphs for the Elderly Population (Figure 18), as expected, show a completely opposite behaviour to those of the Youth Population. Municipalities with a lower percentage of elderly population (Class 1) were the most affected by COVID-19 in the 2nd wave. On the other hand, municipalities with a higher percentage of elderly population (Class 3) were the most affected by COVID-19 in the 3rd wave.

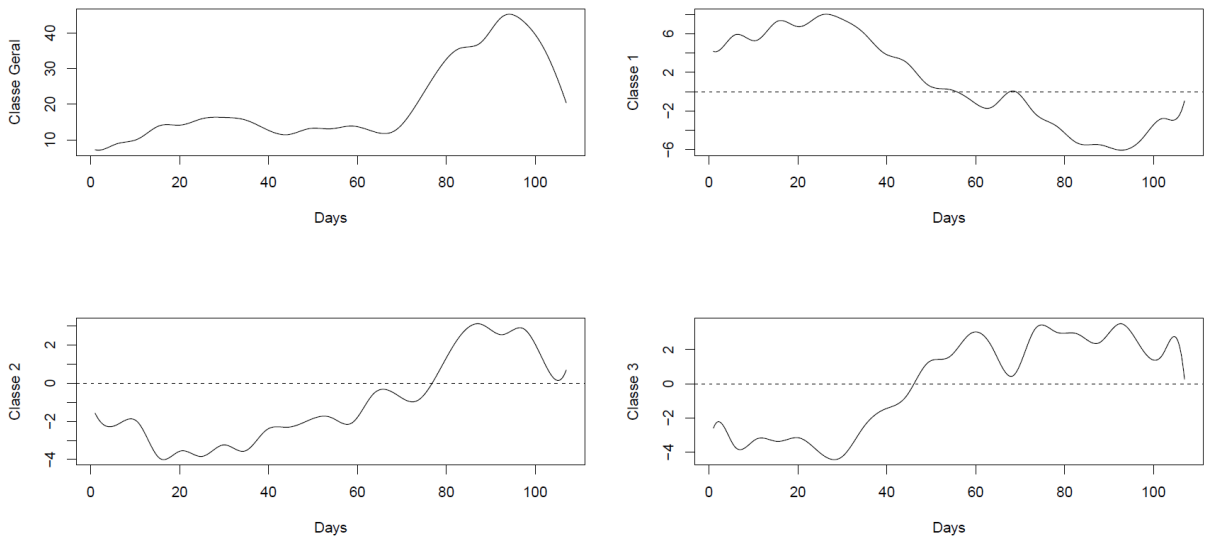



*Figure 18 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Elderly Population*

This contrast between the 2nd and 3rd wave is not exclusive to the Elderly Population. When analysing the Youth Population charts (Figure 19), the behaviour of the curves is similar: municipalities with a lower percentage of young population (Class 1) were the most affected by COVID-19 in the 3rd wave. On the other hand, municipalities with a higher percentage of Young Population (Class 3) were the most affected by COVID-19 in the 2nd wave.

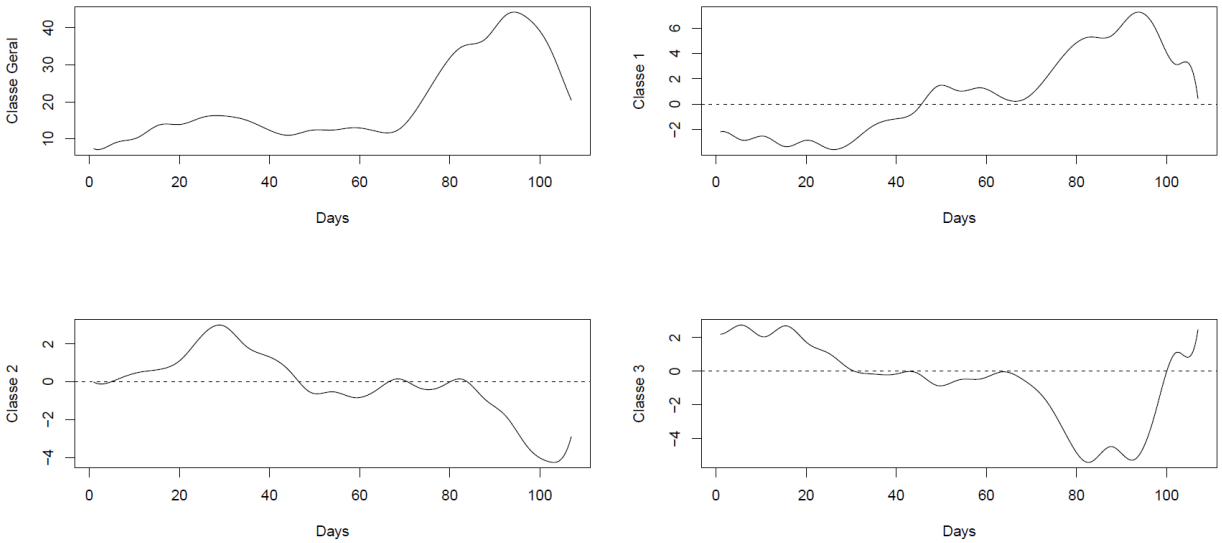



*Figure 19 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Youth Population*

For the variable Proportion of Guaranteed Minimum Income Beneficiaries (Figure 20), Class 1 corresponds to municipalities with a lower Guaranteed Minimum Income ratio, while Class 3

corresponds to those with a higher ratio. This means that in Class 3 municipalities, where the proportion of Guaranteed Minimum Income beneficiaries is higher, the level of poverty is also higher. It is possible to verify that in Class 3 municipalities, along the entire length of the 2nd wave and part of the 3rd wave, the COVID-19 incidence mean is higher than the country average.

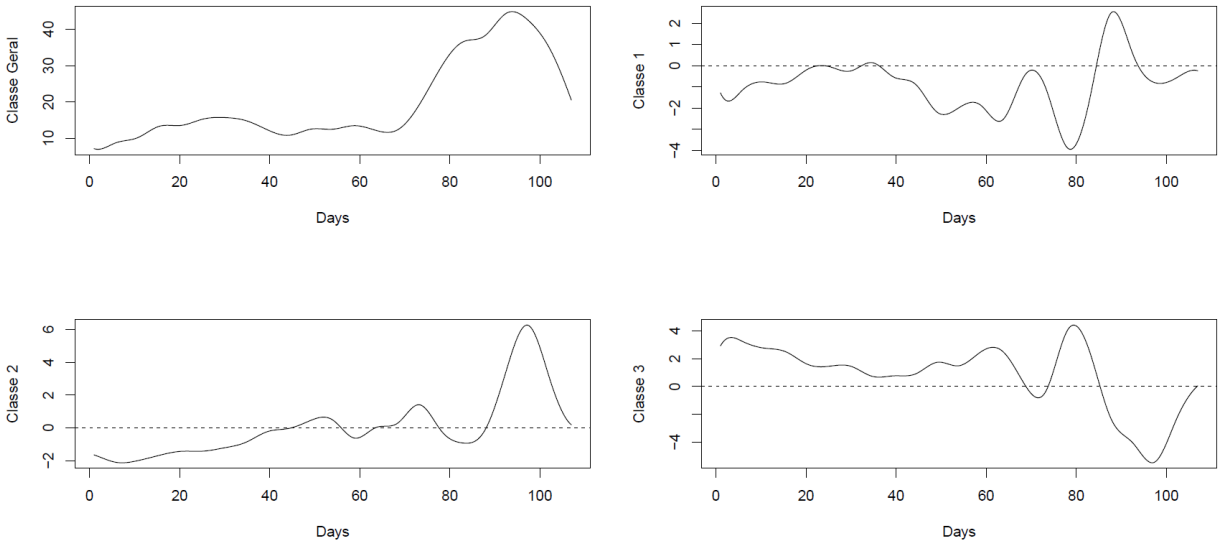



*Figure 20 - Linear Coefficient Functions estimated for predicting COVID-19 Incidence from Proportion of Guaranteed Minimum Income Beneficiaries*

When analysing the coefficients for the percentage of Working Population per Service Sector (Figure 21), class 1 includes municipalities where primary sector activities are predominant. Similarly, class 2 and class 3 refer to the secondary and tertiary sector, respectively. From the analysis of coefficients, it could be seen that the municipalities where the secondary sector is predominant (class 2) showed an above-average behaviour over the 2nd and 3rd wave. On the opposite side, in municipalities where the predominant sector is the primary, the behaviour of the incidence curves was below average.

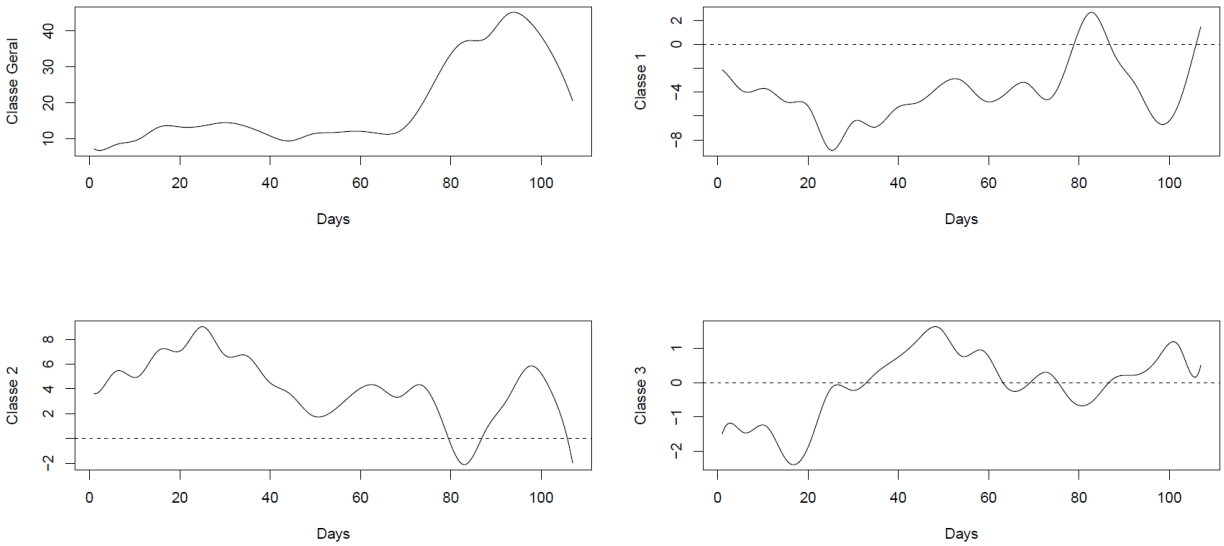



*Figure 21 - Linear Coefficient Functions estimated for predicting COVID-19 incidence from Working Population per Service Sector*

# Chapter 5

# Discussion

In this thesis FDA techniques were applied to analyse the patterns of association between COVID-19 incidence with mobility and sociodemographic variables in Portugal mainland. Despite several limitations of data related with accuracy and level of aggregation, some relevant trends in functional data curve shapes could be identified. Using the Analysis of Variance Model, the incidence was decomposed into functional effects specified by sociodemographic variables (scalars) which showed relevant impacts of population density and age-group structures on the incidence curves during 3rd wave. Moreover, Concurrent and General Concurrent Model techniques were applied to relate the incidence curves with mobility curves (both functional data) at different classes of mobility showing stronger peaks at time-lags of 15-16 days and strong patterns of mobility coefficients for varying time lags in Residential, Retail and Public Transports classes. The FDA approach have been successfully applied in several domains, but still are not very common in epidemiology analysis. In this thesis FDA was used with public open data and allowed to identify patterns of associations between COVID-19 incidence and with mobility and sociodemographic key predictors. The results obtained point out that FDA techniques can be considered an additional tool to more traditional epidemiological analysis contributing to provide new insights.

## 5.1 Functional Responses with Functional Covariates: Concurrent Model

One of the most important results observed using the Concurrent Model, whose achievement was one of the objectives of the application of this method, is related to the lag between the curves. It is noteworthy that for lags outside the 15-16 days range, the results were not good, in any of the classes. This result strengthens the idea that the ideal lag (for which mobility has some effect on the incidence) may be about 15 days, as shown in some literature ((Moorthy et al., 2020) e (Carroll et al., 2020).

The concurrent model was not applied to the 1st wave of the pandemic due to high variation of COVID-19 incidence data caused by the small number of occurrences by municipality in this period. The characteristics of the mobility and incidence functional data curves showed that they present a much higher instability in the 1st wave, compared to the other waves.

The lack of results for the 2nd wave observed here may be explained by a potential underreporting of positive cases, resulting in poorer data accuracy, and therefore not reflecting the real risk of COVID-19 infection.

A detailed analysis of the results obtained for each of the mobility classes is performed in the next section.

## 5.2 Functional Responses with Functional Covariates: General Concurrent Model

Let us suppose that the effect of mobility on incidence would have a lag of about 15 days (a value that is referred in some literature). In this section it was, therefore, expected that a diagonal region would be found in the obtained graphs, between the origin of the graph and the upper right corner, in which the linear coefficient was positive, and shifted to the left about 15 days.

However, considering that these variables concern the behaviour of populations and government measures, and that respond differently to external stimuli, the results were not that linear. Furthermore, these data have limitations in accuracy and are averaged (aggregated by municipality) have been subjected to pre-processing, and that the incidence curves correspond to the second differences of the original data, and the mobility curves correspond to the first differences. Thus, this difference in data treatment also contributes to the problem mentioned above, as the shapes of the curves do not behave in exactly the same way: if on one side there is a sharp peak, on the other side there may be a peak that is smoothed or flattened.

Nevertheless, it was possible to find, in some graphs (as in Figure 9), regions of positive associations for these variables. Despite not being able to find a perfect diagonal (for the reasons mentioned above, which cause the results to be non-linear), these regions show the existence of a lag between the velocity of mobility variation curves and the respective effect on the incidence rate acceleration curves.

It is important to point out that, in these results, it is common to find positive coefficients in places where they were not expected at first (which show, for example, correlations with a lag of 40 days, as in the top left corner of Figure 9 c) ). This is because the incidence rate acceleration and velocity of mobility variation curves have several oscillations, and apparent cause-effect relationships is found in oscillations that are 30 or 40 days apart, which is not at all realistic according to the literature (Moorthy et al., 2020).

Also, the effect of backwards causation should be disregarded. For example, if mobility on day 20 and incidence on day 10 have a negative linear coefficient, this should not be considered in the analysis. In fact, this occurred during pandemic caused by government intervention measures (e.g. lockdowns) when significant increase in incidence was followed by decrease in mobility. Analysing these relations (where mobility should be seen as the response variable and incidence as covariate) would increase dramatically the complexity of analysis.

It should also be noted that this method was found to be unsatisfactory for analysing all the pandemic as a single period, as it produced very noisy results.

**Residential Mobility**

The results obtained here were in line with the ones obtained using the Concurrent Model (from section 4.1), where it was noticed that the Residential class, in the 3rd wave, presented the best results, showing a relatively strong relationship between the variables under analysis with a 16-day delay.

This shows that residential mobility, which is very sensitive to measures of lockdown, has a strong relationship with the COVID-19 incidence rate acceleration (with a lag of around 15 days), compared to all other mobility classes. In the case of the 3rd wave, for example, which coincided with the Christmas and New Year celebrations, residential mobility plays an important role. If on the one hand there was a large movement of people away from their homes at Christmas, on the other hand there was also the application of lockdown measures in order to combat these movements in the New Year. Thus, based on the results, it can be hypothesized that Christmas and New Year might have influenced the behaviour of the incidence curves.

On the other hand, 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, especially in places such as schools.

**Grocery**

The results obtained in this class were also in line with the ones obtained using the Concurrent Model. This mobility class is related for example with mobility in supermarkets and pharmacies. As is known, these are places where the population obtains essential goods, having maintained a minimally normal functioning throughout the entire pandemic. Furthermore, the use of these spaces required people to comply with strict protection measures, such as mask usage and occupancy limitation. Thus, the behaviour of the population in this mobility class must have undergone less changes in relation to their pre-pandemic behaviour (compared to other classes), and that people were well protected. This factor may explain the weak signal found, as the evolution of Grocery mobility will possibly have little impact on the evolution of the incidence of COVID-19.

**Parks**

The results obtained here were in line with the ones obtained using the Concurrent Model. This mobility class is related for example to mobility in local parks, public beaches, and public gardens. The frequency of these spaces is usually associated with periods of high temperature and little rainfall, as is the case of Summer/Autumn, in which the 2nd wave elapsed. As the 3rd wave corresponds to a winter period, the mobility of the population in parks is typically lower (even in pre-pandemic), so it would not be expected that the relief/increase of restriction measures would cause a significant variation in the use of these spaces. Thus, the very weak 3rd wave signal can be explained in light of these factors, as it was unlikely

that parks mobility would have, in this wave, an influence on the evolution of the COVID-19 incidence rate acceleration (as opposed to the 2nd wave).

Furthermore, parks are outdoor places, where virus transmission is expected to be much lower, which may also explain the weakening of the signal in both of the analysed waves.

**Retail**

The results for this mobility class are aligned with what was shown in section 4.1, that in the 3rd wave Retail mobility presents a good relationship (positive linear coefficient) with the incidence rate variation curves, with a 15-day delay.

This mobility class is related for example with mobility in shopping centres. In this case, unlike the Grocery Class, this a type of mobility aimed at obtaining non-essential goods. Thus, it is expected that the use of these spaces varies more significantly according to the relief/increase of restriction measures, and may somehow reflect a more relaxed/careful behaviour of the population in relation to protection against the virus.

Knowing that the 3rd wave corresponds to the Christmas period, this strong initial signal may be related to the large agglomerations that occur at this time in commercial surfaces. This behaviour, which have an abnormally high influence in retail mobility curves, may therefore be related to the spread of COVID-19.

**Stations**

The signal for this mobility class, which is stronger in the 3rd wave, may show the influence that the use of public transport by the population has on the spread of the virus.

The use of public transport varies, for example, according to festive seasons, use of telework and tourism. Thus, the strongest signal in the 3rd wave could be related to the Christmas season (there are fewer people commuting to work) and the application of very restrictive measures (which includes mandatory teleworking) during the 3rd wave.

**Workplace**

These results turn out to be similar to those analysed in the section 4.1, in which it is concluded that it was not possible to obtain a strong relationship between variables.

Fluctuations in the frequency of workplaces during the pandemic period did not show evidences of 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 fact, making the oscillations in the curves of this type of mobility have no impact on the incidence rate variation curves.

## 5.3 Functional Responses with Scalar Covariates: Analysis of Variance Model

The behaviour of the linear coefficients of population density, young population and elderly population can be explained in light of the behaviour of the population in the 3rd wave. As is known, during the month of December many Portuguese moved from urban areas to non-urban areas and inland, in order to celebrate Christmas. The municipalities in these areas are characterized by two aspects: very high percentage of elderly population and low population density. The population movements observed to these locations, and that preceded the 3rd wave, can then be seen as one of the possible causes for the exponential increase in cases of COVID-19 observed in the results obtained (Figure 15, Figure 18, and Figure 19). This hypothesis is in line with the results of research from Marques da Costa (Marques da Costa & Marques da Costa, 2020) or Paul and colleagues (Paul et al., 2021) that showed that COVID-19 preferentially affects the elderly, who are much more vulnerable to the virus, hence the numbers verified in the 3rd were clearly higher.

The results also show that the behaviour of the incidence curves may also be related to school activity. The school term began in mid-September, with the 2nd wave of the pandemic beginning in early October. Analysing the results, it is possible to observe that the municipalities with the highest ratio of schools (Class 3) presented, in the 2nd wave, a COVID-19 incidence mean higher than the country average. Thus, it can be hypothesized that the return to in-person classes, which promoted direct interaction between students, contributed to the spread of COVID-19, and consequently to the increase in the number of cases. This is in line with the expected, and evidenced by the literature, that schools function as one of the biggest contagion hubs of COVID-19. (Boschi et al., 2020).

One of the most surprising results was the coefficients for the Deprivation Index. Contrary to expectations, class 5 municipalities (less deprived group) had a COVID-19 incidence mean higher than the country average, throughout the 2nd and 3rd wave. These results turn out to be contradictory, considering that, according to the literature, the most deprived municipalities were supposed to be among the most affected by COVID-19, which in this case is not the case. The results of (Hatef et al., 2020) for example, show that some neighbourhoods with higher ADI (more disadvantaged) presented higher COVID-19 prevalence.

The level of aggregation of available data could have play a role with this respect, since less deprived municipalities tend to be composed of more densely populated urban areas, were the (average) deprivation index tends to be lower, “masking” the existence of highly heterogeneous deprivation index values within these municipalities. It would be interesting to fit these models (Analysis of Variance Model) using higher resolution data to allow a more in-depth analysis of results.

Another possible reason that can explain this contradiction of results is the fact that Deprivation Index is a variable composed of many other variables. Thus, the signal given by this variable turns out to be much weaker, is much more subject to variability and can easily generate confusing results. In addition,

the Deprivation Index is based on the 2011 Census, so the data may be out of date and not fully correspond to reality.

In fact, results for the Guaranteed Minimum Income, on the other hand, are closer to what would be expected. In municipalities where the percentage of people with low income is higher, the spread of COVID-19 is also higher. This goes in line to what is presented in the literature, where several studies have shown that low sociodemographic conditions are a major factor contributing to the spread of the virus (Hatef et al., 2020; Whittle & Diaz-Artiles, 2020). The lack of financial resources hinders the population's access to housing, health care, education, etc., making it more vulnerable to COVID-19.

## 5.4 Limitations

One of the main limitations of this thesis was the need to use real data. Firstly, these data may be unstable, noisy, and not always reflect reality. Furthermore, the COVID-19 data applied here refer to the time period between March 15, 2020 and February 2, 2021. The fact that the 3rd wave was not analysed in its entirety, and the underreporting of cases verified in 1st wave, prevented more robust hypotheses from being postulated.

The FDA tools applied in this work, despite all their advantages, also had several limitations. Functions created by applying regression splines often can lead to overfitting of the data. This phenomenon occurs due to the extraction of residual variation contained by the data, making their analysis difficult, especially if it is necessary to analyse derivatives of these same curves, thus distorting the results obtained through the analysis of these functions. Also, spline basis functions may produce unstable fits to the data near the beginning or the end of the period over which they are defined, which may disturb the results. This instability of spline fits becomes more serious for derivative estimation, and the higher order of the derivative, the more unstable its behaviour.

Regarding the linear models applied to the data, these also had some limitations. Concurrent Model could not be applied to 1st wave, and the necessity to apply lags between curves made the process of analysis too complex and time-consuming. Additionally, the need to use regression splines to construct linear coefficient curves sometimes gives rise (as mentioned above) to unstable fits to the data near the beginning or the end of the period over which they are defined, as in Figure 8. Furthermore, both the Concurrent Model and General Concurrent Model found to be unsatisfactory for analysing all the pandemic as a single period, as it produced very noisy results. Also, the fact that the analysis was partially visual and qualitative may have induced in some subjectivity and errors in the comparisons made.

# Chapter 6

# Conclusions

## 6.1 Summary and Conclusions

The objective of this Thesis was to use several Functional Data Analysis techniques such as smoothing, interpolation and functional linear models, in order to analyse and quantify the association of COVID-19 incidence data with Google mobility (6 different classes) and Sociodemographic data (6 variables). This analysis intended, therefore, to understand the impact that mobility and sociodemographic conditions have on the spread of COVID-19 and also to determine which are the most efficient containment measures in combating the virus.

COVID-19 incidence and Google Mobility data, after pre-processing, take the form of time-series, while sociodemographic data takes the form of a single value for each sociodemographic variable. In this pre-processing, Google Mobility data is always stationary, while COVID-19 incidence is stationary only when analysing its association with mobility. These data, obtained from different sources, are municipal data, and allow this study to cover the 278 municipalities of Continental Portugal.

A literature review allowed us to realize that, despite the numerous advantages presented by the FDA methods, the functional study of time series is a recent technique. Conventional methods, such as Multivariate Data Analysis, are much more frequent. Thus, what this thesis proposed is something innovative, and intended to show the potential that the FDA has in the area of pandemic analysis. Despite several limitations of data, relevant trends in functional data curve could be identified through the application of FDA.

The first FDA tool to be applied in this work, smoothing, aimed to transform the time-series into curves. Then, linear modelling techniques were applied to analyse the association between a response variable and two explanatory variables.

The first linear modelling technique, which aimed to study the association between a functional response variable (COVID-19 incidence) and a scalar explanatory variable (Sociodemographic Variables), is called Analysis of Variance Model, that decomposes incidence into functional effects specified by the sociodemographic variables. The linear coefficient curves obtained here correspond to the perturbation of the COVID-19 incidence curve required to fit the class's mean COVID-19 incidence curve. The results showed the role that municipalities with low population density and high rate of elderly population had in the 3$^{rd}$ covid wave, with an incidence rate acceleration far above the average. In addition, the data

showed how schools may work as one of the biggest contagion hubs of COVID-19. Despite the contradictory results regarding the Deprivation Index, it was noticed that in municipalities where the percentage of people with low income is higher, the incidence rate acceleration is above average. Finally, this technique showed evidence that workers in the secondary sector (industry and construction) may be more susceptible to being infected, compared to the primary and tertiary sector. In conclusion, some sociodemographic conditions may influence the spread of COVID-19. Thus, measures must be taken to protect the most vulnerable and disadvantaged populations, and also to reduce the spread of the virus in contagion hubs, such as schools, thus helping to protect the population in general.

The second linear modelling technique, which aimed to study the association between a functional response variable (COVID-19 incidence rate acceleration) and a functional explanatory variable (velocity of Google mobility variation), is called Concurrent Model, shows the evolution of a linear coefficient over time. Two variants of this technique were applied, the Concurrent Model (which applies a lag between curves) and the General Concurrent Model. The results obtained here were mostly for the 3$^{rd}$ wave, and strengthened the idea that the ideal lag for which different classes of mobility have some effect on the incidence may be about 15 days. It was possible to notice that mobility and incidence are somehow associated, and the strongest sign found was in the relationship between residential mobility and COVID-19 incidence rate acceleration. This information may be very interesting, because it reinforces the effectiveness of the more restrictive measures applied to control the pandemic, such as lockdown. If on the one hand retail and stations mobility also showed interesting signals, on the other hand grocery, parks and workspace showed much weaker signals. Although a given mobility class is associated with COVID-19 incidence, this does not necessarily mean that it is the cause of changes in incidence. However, taking these results into account, different mobility classes (residential, retail and stations) can be used to indirectly try to predict the spread of the virus by monitoring the behaviour of the population (relaxed/cautious).

It should also be noted that some results, namely those related to Residential mobility, suggest that sudden changes in mobility (mass movements of the population, and mandatory lockdown) will have a 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, through the application of efficient measures.

Despite still not being very common in epidemiology analysis, the results obtained here point out that FDA can considered an additional tool to more traditional epidemiological analysis to provide new insights.

## 6.2 Future Work

The FDA is a very broad area, encompassing a multitude of methods, and with a lot of potential yet to be explored. In addition to the methods used in this work, other approaches can (and should) be tried, for example, using Functional Principal Components Analysis and Functional Principal Differential Analysis, referred in the Introduction. Despite that, the methodology of this work can be applied in future waves of COVID-19 (or even future pandemics) due to its easy replication.

In addition to the variables studied here, it may be interesting to study the association of other variables with the COVID-19 incidence. Possible examples include other sociodemographic variables, meteorological data, mobility data from other sources, or vaccination rates. The possibility to work with higher resolution data would also contribute to allow a more in-depth analysis of results.

Another suggestion would be to apply the methodology of this work using not the COVID-19 incidence data, but COVID-19 mortality data in Portugal. This strategy may have special relevance in the field of sociodemographic factors, due to the important role they typically play in the population's health outcomes. Proof of this is that most of the approaches found in the literature, which intended to study the association of sociodemographic conditions with the evolution of COVID-19, used mortality and hospitalization data, and found satisfactory results.

# Bibliography


Abbas, M., Morland, T. B., Hall, E. S., & EL-Manzalawy, Y. (2021). Associations between Google search trends for symptoms and COVID-19 confirmed and death cases in the United States. *International Journal of Environmental Research and Public Health*. https://doi.org/10.1101/2021.02.22.21252254

Aleta, A., & Moreno, Y. (2020). Evaluation of the potential incidence of COVID-19 and effectiveness of containment measures in Spain: A data-driven approach. *BMC Medicine*, *18*(1). https://doi.org/10.1186/s12916-020-01619-5

Boschi, T., di Iorio, J., Testa, L., Cremona, M. A., & Chiaromonte, F. (2020). *The shapes of an epidemic: using Functional Data Analysis to characterize COVID-19 in Italy*. https://doi.org/10.1038/s41598-021-95866-y

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Cao, Y., Hiyoshi, A., & Montgomery, S. (2020). COVID-19 case-fatality rate and demographic and socioeconomic influencers: Worldwide spatial regression analysis based on country-level data. *BMJ Open*, *10*(11). https://doi.org/10.1136/bmjopen-2020-043560

Carroll, C., Bhattacharjee, S., Chen, Y., Dubey, P., Fan, J., Gajardo, Á., Zhou, X., Müller, H. G., & Wang, J. L. (2020). Time dynamics of COVID-19. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-77709-4

Cavalcante, J. R., Cardoso-Dos-Santos, A. C., Bremm, J. M., Lobo, A. de P., Macário, E. M., Oliveira, W. K. de, & França, G. V. A. de. (2020). 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*, *29*(4), e2020376. https://doi.org/10.5123/s1679-49742020000400010

Croux, C., & Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, *95*(1). https://doi.org/10.1016/j.jmva.2004.08.002

Delen, D., Eryarsoy, E., & Davazdahemami, B. (2020). No place like home: Cross-national data analysis of the efficacy of social distancing during the COVID-19 pandemic. *JMIR Public Health and Surveillance*, *6*(2). https://doi.org/10.2196/19862

Drake, T. M., Docherty, A. B., Weiser, T. G., Yule, S., Sheikh, A., & Harrison, E. M. (2020). The effects of physical distancing on population mobility during the COVID-19 pandemic in the UK. In *The Lancet Digital Health* (Vol. 2, Issue 8, pp. e385–e387). Elsevier Ltd. https://doi.org/10.1016/S2589-7500(20)30134-5

Evttin, D. L. [ I., Gina, R. F., Nuzzo, L., & Ramsay, J. O. (2007). Introduction to Functional Data Analysis. *Canadian Psychology*, *48*(3), 135–155. https://doi.org/10.1037/cp20070I4

Fernández-Recio, J. (2020). Modelling the evolution of COVID-19 in high-incidence European countries and regions: Estimated number of infections and impact of past and future intervention measures. *Journal of Clinical Medicine*, *9*(6), 1–17. https://doi.org/10.3390/jcm9061825

Goldsmith, J., & Schwartz, J. E. (2017). Variable selection in the functional linear concurrent model. *Statistics in Medicine*, *36*(14). https://doi.org/10.1002/sim.7254

Goodman, L. A., Kruskal, W., H ; Rabe-Hesketh, S., & Everitt, B. (1979). Regression Model for Categorical and Limited Dependent Variables. Sage, Thousand Oaks, CA McCullagh P 1980 Regression models for ordinal data. In *Journal of the Royal Statistical Society, Series B* (Vol. 54). Springer.

Grami, A. (2016). Probability, Random Variables, and Random Processes. In *Introduction to Digital Communications* (pp. 151–216). Elsevier. https://doi.org/10.1016/b978-0-12-407682-2.00004-1

Hatef, E., Chang, H. Y., Kitchen, C., Weiner, J. P., & Kharrazi, H. (2020). Assessing the Impact of Neighborhood Socioeconomic Characteristics on COVID-19 Prevalence Across Seven States in the United States. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.571808

Horváth, L., Kokoszka, P., & Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, *179*(1), 66–82. https://doi.org/10.1016/j.jeconom.2013.11.002

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*.

Hyndman, R. J., & Shahid Ullah, Md. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, *51*(10). https://doi.org/10.1016/j.csda.2006.07.028

Jones, N. R., Qureshi, Z. U., Temple, R. J., Larwood, J. P. J., Greenhalgh, T., & Bourouiba, L. (2020). Two metres or one: what is the evidence for physical distancing in covid-19? *BMJ (Clinical Research Ed.)*, *370*, m3223. https://doi.org/10.1136/bmj.m3223

Khalatbari-Soltani, S., Cumming, R. C., Delpierre, C., & Kelly-Irving, M. (2020). Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. In *Journal of Epidemiology and Community Health* (Vol. 74, Issue 8, pp. 620–623). BMJ Publishing Group. https://doi.org/10.1136/jech-2020-214297

Kokoszka, P., & Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315117416

Kumar, V., Sood, A., Gupta, S., & Sood, N. (2021). Prevention- Versus Promotion-Focus Regulatory Efforts on the Disease Incidence and Mortality of COVID-19: A Multinational Diffusion Study Using Functional Data Analysis. *Journal of International Marketing*, *29*(1), 1–22. https://doi.org/10.1177/1069031X20966563

Marques da Costa, E., & Marques da Costa, N. (2020). 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*, 72–79. https://doi.org/10.14393/hygeia0054396

Mas, A., & Pumo, B. (2009). Functional linear regression with derivatives. *Journal of Nonparametric Statistics*, *21*(1). https://doi.org/10.1080/10485250802401046

Moorthy, V., Restrepo, A. M. H., Preziosi, M. P., & Swaminathan, S. (2020). Data sharing for novel coronavirus (COVID-19). In *Bulletin of the World Health Organization* (Vol. 98, Issue 3, p. 150). World Health Organization. https://doi.org/10.2471/BLT.20.251561

Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E. C., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cooper, L. v., Coupland, H., Cucunuba, Z. M., Cuomo-Dannenburg, G., Dighe, A., Djaafara, B. A., Dorigatti, I., Eales, O. D., van Elsland, S. L., Nascimento, F. F., … Donnelly, C. A. (2021). Reduction in mobility and COVID-19 transmission. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21358-2

Patel, A. P., Paranjpe, M. D., Kathiresan, N. P., Rivas, M. A., & Khera, A. v. (2020). Race, socioeconomic deprivation, and hospitalization for COVID-19 in English participants of a national biobank. *International Journal for Equity in Health*, *19*(1). https://doi.org/10.1186/s12939-020-01227-y

Paul, A., Englert, P., & Varga, M. (2021). Socio-economic disparities and COVID-19 in the USA. *Journal of Physics: Complexity*, *2*(3). https://doi.org/10.1088/2632-072X/ac0fc7

Quan, D., Luna Wong, L., Shallal, A., Madan, R., Hamdan, A., Ahdi, H., Daneshvar, A., Mahajan, M., Nasereldin, M., van Harn, M., Opara, I. N., & Zervos, M. (2021). Impact of Race and Socioeconomic Status on Outcomes in Patients Hospitalized with COVID-19. *Journal of General Internal Medicine*, *36*(5), 1302–1309. https://doi.org/10.1007/s11606-020-06527-1

Ramsay, J., Hooker, G., & Graves, S. (2009). Functional Data Analysis with R and MATLAB. In *Functional Data Analysis with R and MATLAB*. Springer New York. https://doi.org/10.1007/978-0-387-98185-7

Ramsay, J. O. (2016). Functional Data Analysis – Theory. In *Wiley StatsRef: Statistics Reference Online* (pp. 1–13). Wiley. https://doi.org/10.1002/9781118445112.stat00516.pub2

Ribeiro, A. I., Launay, L., Guillaume, E., Launoy, G., & Barros, H. (2018). The Portuguese version of the European deprivation index: Development and association with all-cause mortality. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0208320

Santamaría, L., & Hortal, J. (2021). COVID-19 effective reproduction number dropped during Spain's nationwide dropdown, then spiked at lower-incidence regions. *Science of the Total Environment*, *751*. https://doi.org/10.1016/j.scitotenv.2020.142257

Sebastiani, G., Massa, M., & Riboli, E. (2020). Covid-19 epidemic in Italy: evolution, projections and impact of government measures. *European Journal of Epidemiology*, *35*(4), 341–345. https://doi.org/10.1007/s10654-020-00631-6

Song, J. J., Lee, H.-J., Morris, J. S., & Kang, S. (2007). Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry*, *31*(4). https://doi.org/10.1016/j.compbiolchem.2007.05.006

Srivastava, A., & Chowell, G. (2020). Understanding Spatial Heterogeneity of COVID-19 Pandemic Using Shape Analysis of Growth Rate Curves. *MedRxiv : The Preprint Server for Health Sciences*. https://doi.org/10.1101/2020.05.25.20112433

Tang, C., Wang, T., & Zhang, P. (2020). *Functional data analysis: An application to COVID-19 data in the United States*. http://arxiv.org/abs/2009.08363

Ullah, S., & Finch, C. F. (2013). *Applications of functional data analysis: A systematic review*. http://www.psych.mcgill.ca/misc/fda/

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

Whittle, R. S., & Diaz-Artiles, A. (2020). An ecological study of socioeconomic predictors in detection of COVID-19 cases across neighborhoods in New York City. *BMC Medicine*, *18*(1). https://doi.org/10.1186/s12916-020-01731-6