Joseph Piekos
Supervisor: **Georgios Karagiannis**

Grocery Shopping as a Predictor of COVID-19 Deaths

Durham University

SCAN FOR CODE

Introduction

Mortality curves have been a very accessible way for people to visualise the general trend of the COVID-19 pandemic, but we can extract even more information from them mathematically using Functional Data Analysis (FDA).

Using these techniques we can compare patterns in deaths, even at high data resolutions of town or council. Also investigation of predictors of mortality is possible, which can then inform Government health policy.

Primarily here we will compare data for 78 English counties, and look into grocery mobility as a predictor variable for COVID-19 mortality.

1. Notion of Functional Data

FDA considers each data point as a smooth function, compared to the normal notion of a data point being a number. Data is collected over a *continuum* (i.e. time) and then repeated in this case for each county:

$$\begin{array}{ccc} \text{Discrete Data} & \xrightarrow{\textit{smoothing}} & \text{Functional Data:} \\ y_{i,1}, \dots, y_{i,n} & \longrightarrow & y_i(t)\,,\ t \in [T_1, T_2] \end{array}$$

There are multiple **smoothing** methods, however here we will use a B-Spline basis with a Roughness Penalty (see "3. Smoothing"). Implemented in R using the '*fda*' package[].
With real life data, each curve will have key features, such as maxima, occur at different times, t. Therefore we must **align** our curves to account for *phase* (time) variation, so our resulting analysis only reflects *amplitude* variation.

With our new data set we can easily visualise patterns, and also explore variance in our data through a mean function and functional linear models shown below:

- **Mean**:

$$\bar{y}(t) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} y_i(t)$$

- **Functional Regression**:

$$y_i(t) = \mu(t) + \int \beta(t,s)x_i(s)\,ds + \epsilon_i(t)$$

2. Functional COVID-19 Data

Cumulative Daily COVID-19 deaths were retrieved from the UK Health Security website[1], and Mobility data was taken from Google's Community Mobility Reports[2] both for:

1. 78 English counties (including Greater London)
2. 180 days (15^{th} Feb to 13^{th} Aug 2020)

We can use this data to create Functional Data objects because we have 78 replicates each measured once a day over time.

3. Smoothing

Smoothing of data is done using a **roughness penalty** which requires three things:

- A set of basis functions, $\phi_1 \dots \phi_K$ (B-Spline basis in this case)
- An mth derivative to "penalise"
- A smoothing parameter λ

The squared second derivative, $D^2x_i(t)$, of a function corresponds to the **curvature** of its graph, and combined with λ we can decide how smooth we want the function to be vs how closely we want it to fit the data. Mathematically, the **fitting criterion**[3] can be written:

$$F(\underline{\mathbf{c}}) = RSS + \lambda \int [D^2x_i(t)]^2\,dt$$

- We aim to minimise the value of $F(\underline{\mathbf{c}})$, similarly to minimising just RSS in normal regression
- $\underline{\mathbf{c}}$ represents the coefficients of the basis functions in our estimated function
- $\lambda \to 0$, lets our function fit the data as closely as possible, so is less smooth.

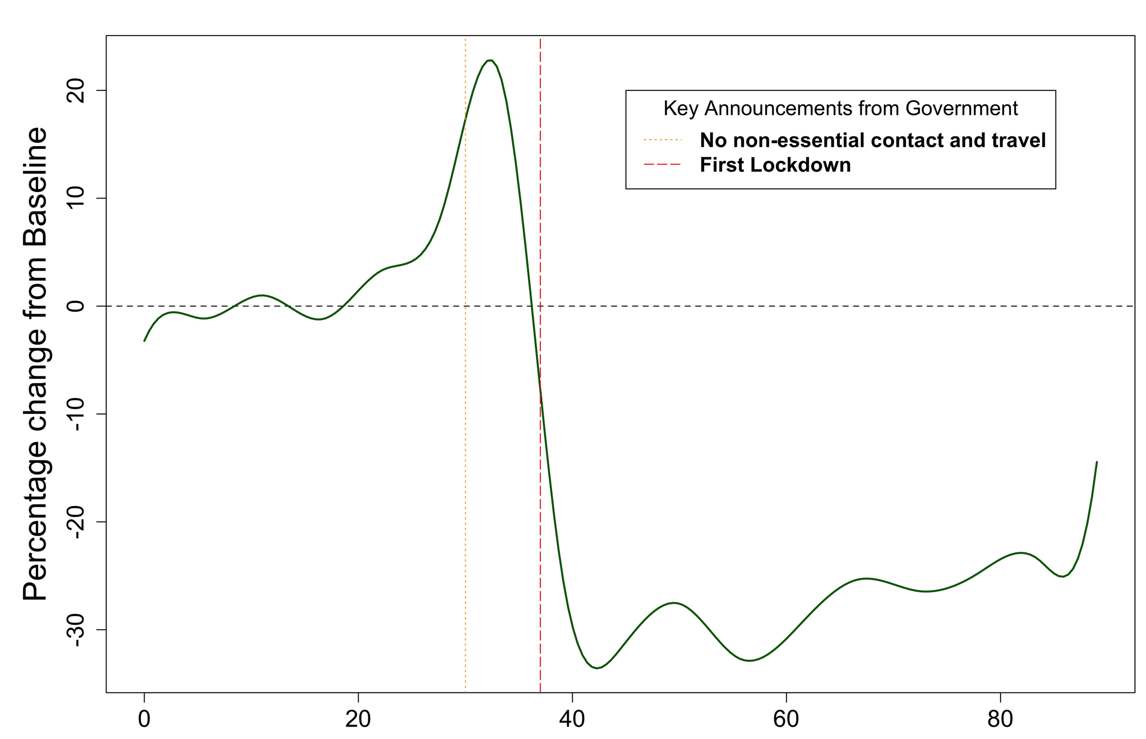

Figure 1: Mobility (to Grocery/Pharmacy) curve for County Durham

Figure 1 shows the smoothed mobility curve for County Durham, for the first 90 days. Smoothing parameter $\lambda = 10$ here.

- Steep increase around day 30, due to panic buying as fears of a lockdown become real.
- Sharp decrease to below zero, as lockdown is made official on March 23^{rd}

4. Aligning Counties' Curve

As stated earlier it is essential to **align** (or "register") all our curves, so we can analyse how much they vary in amplitude rather than in time. 50 counties' fitted and registered curves are shown below, aligned using **Continuous Registration**[3].

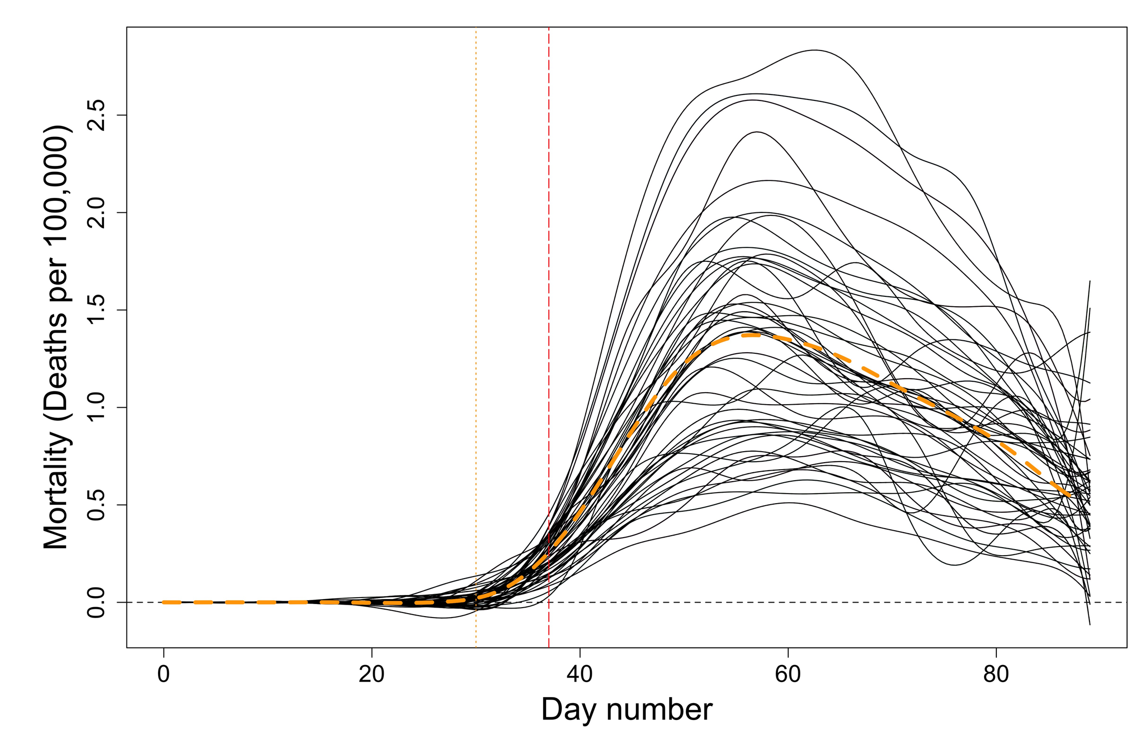

Figure 2: Aligned and smoothed mortality curves for 50 counties

The mean curve is added in orange, and we can see it seems to reflect the average of all the curves around it. From this we begin to notice some areas having more extreme mortality, prompting further investigation.

5. Predicting Deaths

Functional Regression comes in 3 forms, depending on whether the response and predictor variables are curves. Here we use Function-on-Function regression, where both mortality and mobility are curves. The model is an extension of normal linear regression:

$$y_i = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \epsilon_i$$

becomes:

$$y_i(t) = \mu(t) + \int \beta(t,s)x_i(s)\,ds + \epsilon_i(t)$$

- We can interpret the **regression function** $\beta(t,s)$ for a fixed value of t as the weight placed on the mobility each day s that is required to predict mortality on day t.

Main Result: Mobility as a Mortality Predictor

Using the "refund" package in R, we can fit the functional model, Mortality $\sim$ Mobility for the whole 180 days and 78 counties, A plot of the estimated coefficients $\hat{\beta}(t,s)$ is below

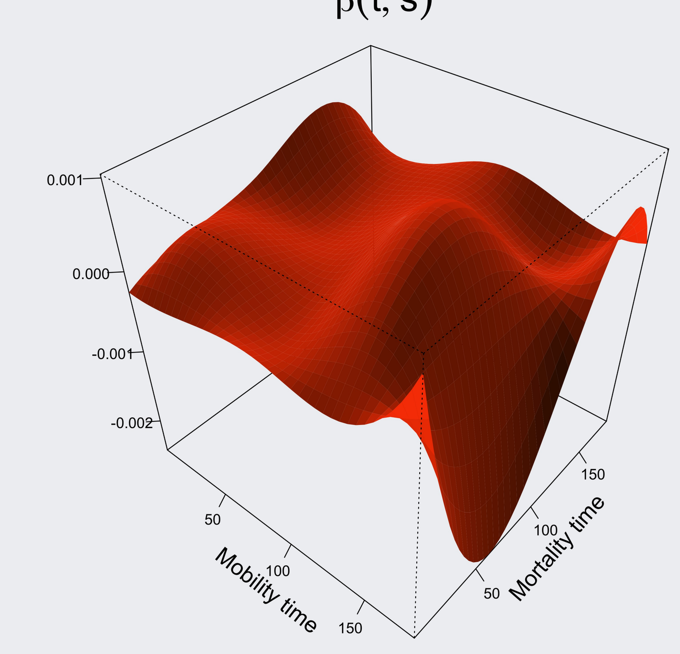

Figure 3: Coefficients for the functional linear model

- Looking at Figure 3, there are positive peaks midway and at the start of the mobility axis. This shows that local trips to Grocery shops/Pharmacies are a strong predictor of increased mortality. Later in the pandemic the surface drops below zero, as summer arrived and lockdown loosened in areas with lower deaths.
- Also, this model had an R^2 value of 0.532, meaning we captured 53.2% of the variance in mortality using only mobility data. Overall we can therefore be confident about the effectiveness of lockdowns in helping prevent COVID-19 deaths.

References

[1] UKHSA.
UK coronavirus dashboard - data.gov.uk.
https://coronavirus.data.gov.uk/details/deaths, 2022.

[2] Google LLC.
Covid-19 community mobility reports.
https://www.google.com/covid19/mobility/index.html, 2020.

[3] James Ramsay, Giles Hooker, and Spencer Graves.
Functional Data Analysis with R and MATLAB.
Springer New York, 2009.