Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
`http://siba-ese.unisalento.it/index.php/ejasa/index`
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v14n1p197

Functional cluster and canonical correlation analysis of EU countries by number of daily deaths and stringency index during Covid-19 pandemic

By Keser, Kocakoç

Published: 20 May 2021

Electronic Journal of Applied Statistical Analysis
Vol. 14, Issue 01, May 2021, 197-216
DOI: 10.1285/i20705948v14n1p197

Functional cluster and canonical correlation analysis of EU countries by number of daily deaths and stringency index during Covid-19 pandemic

İstem Köymen Keser*and İpek Deveci Kocakoç

Dokuz Eylül University, Department of Econometrics
İzmir, Turkey

Published: 20 May 2021

The danger of a global pandemic, such as the new Coronavirus (Covid-19), is obvious. This study aims to investigate the behavior and relationship of the number of daily new confirmed deaths per million and the stringency index of twenty-seven European Union (EU) countries by utilizing functional cluster analysis and functional canonical correlation analysis. Functional cluster analysis was used to observe how countries cluster together according to daily deaths during the time interval between March and July 2020. Functional canonical correlation analysis was also utilized to measure the correlation between the frequency index and daily deaths, and also to determine the relative positions of countries concerning their respective variability structure. The data is obtained from OWID. Here, it is seen that Italy, Spain, Belgium, and France are particularly affected by the pandemic during the time interval within the EU countries, and the course of daily deaths is in a different position compared to other EU countries. At the same time, a very high relationship emerged between the stringency index and daily deaths as expected.

keywords: Covid 19, pandemic, functional cluster analysis, functional canonical correlation analysis, public health

[*] Corresponding author: ipek.deveci@deu.edu.tr

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

The danger of a global pandemic, such as the new Coronavirus (Covid-19), is obvious. As of July 1, 2020, the number of Covid-19 cases worldwide was 10,708,589 and 516,570 people died due to Covid-19 (Worldometer, 2020). The World Health Organization (WHO) describes Covid-19 at the High-Risk level. In order to avoid the high level of pandemic risk, countries have mainly adopted regional quarantines, restriction of going out of home, and social distance/isolation policy. As a result of this, significant restrictions and workplace closings were made in cafes, restaurants, entertainment venues, hairdressers, hotels, shopping malls, urban and intercity land/sea/air transportation services, where people often come together. The effects of these measures on the spread of the disease are still under investigation. There are many organizations that keep the track of events related to the pandemic by many variables such as daily and total cases, deaths, recoveries, and test numbers. In this study, we focused on daily confirmed new deaths per million population. According to ECDC (2020), as of 01 July 2020, 177,122 Covid-19 based deaths have been reported in the EU/EEA and the UK. Although the 7-day rolling average of daily confirmed COVID-19 deaths per million people slightly decreases (Figure 1), the course of daily confirmed COVID-19 deaths in the world based on continents (Figure 2) shows that the pandemic is not ending soon, and some policy measures are still needed.

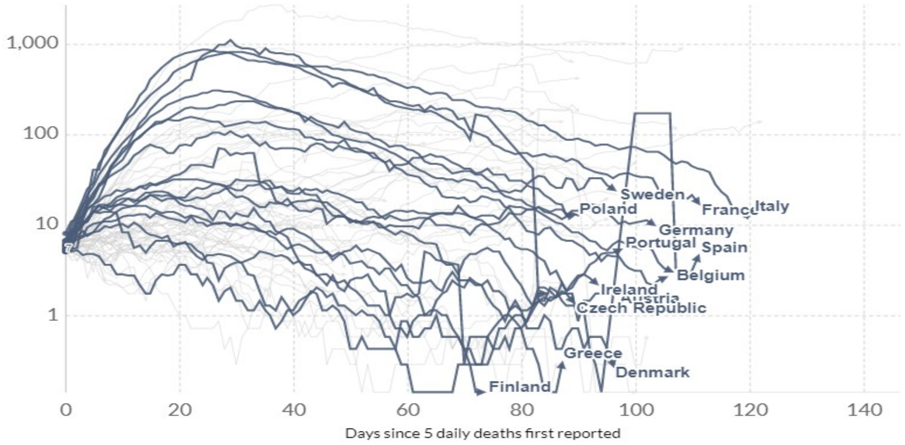

Figure 1: A 7-day rolling average of daily confirmed COVID-19 deaths per million people Source: https://ourworldindata.org/grapher/covid-daily-deaths-trajectory-per-million

To date, non-pharmacological interventions (NPI) have been the mainstay for controlling the coronavirus disease-2019 (COVID-19) pandemic Chowdhury et al. (2020). The Oxford COVID-19 Government Response Tracker (OxCGRT) systematically collects information on several different common policy responses that governments have taken to respond to the pandemic on 17 indicators such as school closures and travel restrictions. Data is collected from public sources by a team of over one hundred Oxford University

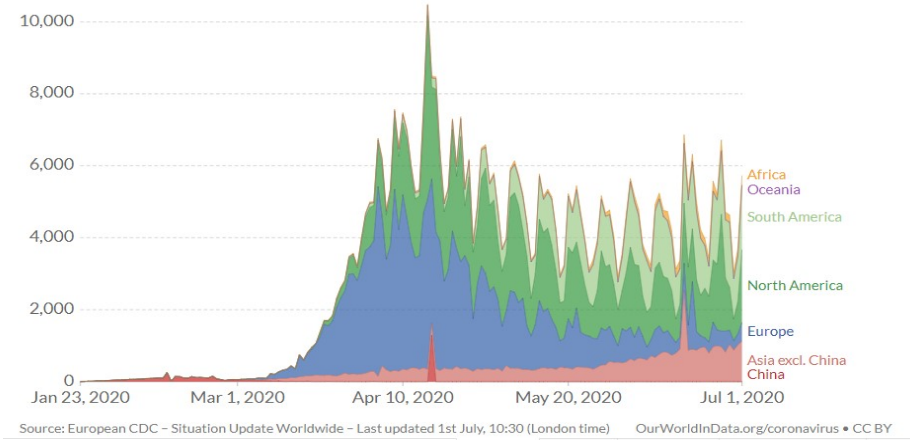

Figure 2: Daily confirmed COVID-19 deaths
Source: https://ourworldindata.org/grapher/daily-covid-deaths-region

students and staff from every part of the world. The Government Response Stringency Index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) (Oxford University Government Response Tracker, 2020). This study aims to investigate the behavior and relationship of daily new deaths per million and the stringency index of the twenty-seven European Union (EU) countries by utilizing functional cluster analysis and functional canonical correlation analysis. Functional cluster analysis was used to observe how countries are clustered together according to the number of new deaths over the period of interest. Functional canonical correlation analysis was also used to measure the correlation between the stringency index and new deaths, as well as to determine the relative positions of countries in relation to the variability structure. In Figure 3, the relationship between the number of new cases and the stringency index is examined for one day. However, since the number of new cases is closely related to the number of tests performed, it may not necessarily reflect the impact of the measures taken. Therefore, in this study, we examined the relationship between the number of deaths, a more robust variable, and the stringency index for a period of time.

2 Methods

2.1 Functional data analysis

Finite discrete time series are typically viewed as multivariate data. This approach, however, ignores significant details about the stochastic method that created the data. Thus, considering a more general element as the unit of measurement under analysis is often beneficial. FDA provides the techniques to analyze, model, and predict time data series when the intrinsic structure of the data is functional, i.e. when there is an underlying function that generates the observed data (Matabuena et al., 2019). With

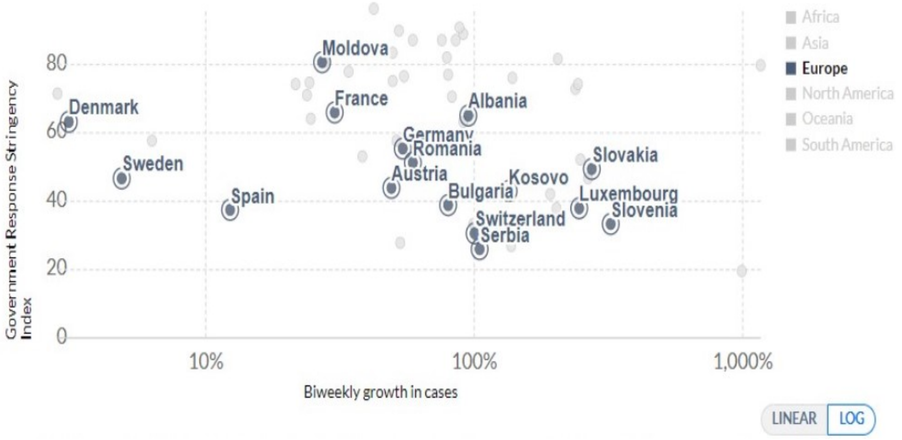

Figure 3: Biweekly growth in cases and stringency index in Europe as of 1st of July, 2020
Source:https://ourworldindata.org/grapher/government-response-stringency-index-vs-biweekly-change-in-confirmed-covid-19-cases?time=2020-07-01

the development of technology, especially in data storage capacities, the required functional equivalents of classical statistical methods have been developed. The concept of functional data first entered the literature with Ramsay (1982) and Ramsay and Dalzell (1991). Ramsay and Silverman (1997) provided and identified examples of functional data formulation of many statistical methods such as linear regression analysis, principal components analysis, canonical correlation analysis, and discriminant analysis. The basic philosophy of functional data analysis is approaching observed data functions not as a line of consecutive individual observations, but as unique inputs (Keser, 2014).

In this study, curvilinear data will be dealt with as functional data. In the functional data analysis approach, the underlying curve is assumed to be smooth. If the measured process is well controlled and measured accurately, some kind of interpolation of the observed data values may be sufficient to provide a smooth curve. However, if the data values are subject to measurement errors, a method that generally smoothes the data will be required to ensure the function's smoothness. By assuming that the underlying curve is smooth, additional information can be obtained by getting the derivatives of the functions. One of the strongest aspects of this method is that the desired degree derivatives of the curves of interest can be taken. Functional data is often observed intermittently and consists of n pairs as a functional observation (y_j, t_j) of a single function $x(t)$ that can be expressed as

$$y_j = x\left(t_j\right) + \epsilon_j \quad j = 1, 2, \ldots, \mathrm{n}. \tag{1}$$

Here y_j is a scaler and the observation of the function $x(t)$ at time t_j and ε_j denotes errors. Also, t is assumed to be limited in the range $T = [t_1, t_n]$. The function $x(t)$ provides a logical smoothing in T. One way to obtain smooth functions in the form of $x(t)$ during the time t is the linear combination of the K basis functions in form of

$(\phi_1, \phi_2, \ldots, \phi_K)$.

$$x(t) = \sum_{k=1}^{K} c_k \phi_K(t) = \mathbf{c\Phi(t)} \quad \forall \boldsymbol{t} \in \boldsymbol{T}. \tag{2}$$

$\phi_K(t)$ are known basis functions defined in the same time interval as $x(t)$, and $\mathbf{\Phi(t)}$ is the vector of these basis functions. c_k is the basis function coefficients and $\mathbf{c}$ is the coefficients vector.

There are different types of basis functions with different strengths and weaknesses. The best choice largely depends on the nature of the underlying smooth curve. For example, if a regular periodic structure is available, the Fourier basis is suitable, while less powerful, with strong local properties.

There may be undesirable high-frequency fluctuations that affect derivatives between the observation points of many bases that work well for function prediction. Typically for this, one or more derivatives of the basis function approach are selected to act logically. For computational reasons, it is preferable to select the K value, which indicates the number of basis functions, as small as possible (Strandberg, 2013). However, there are other systems.

While obtaining some function coefficients, methods such as the least-squares method or roughness penalty method are used. In this study, the roughness penalty method that provides smoothing with a certain smoothing parameter is used.

In the roughness penalty method, while obtaining the roughness penalty estimates of the coefficients, let the function $x(t)$ be a function whose derivative can be defined in the range $T = [t_1, t_n]$ and λ be the smoothing parameter. Penalized least squares $PENSSE_\lambda$ can be defined as

$$\text{PENSSE}_{\lambda} = \sum_{J} (y_j - x(t_j))^2 + \lambda \left\|D^2 x\right\|^2 = \sum_{J} (y_j - x(t_j))^2 + \lambda \int \left(D^2 x(t)\right)^2 dt. \tag{3}$$

Here, $\|D^2 x\|^2$ is a globally accepted way to measure the roughness of a curve and can be expressed in terms of vector-based as

$$\text{PENSSE}_\lambda = [\mathbf{y} - \boldsymbol{\theta}\mathbf{c}]^T [\mathbf{y} - \boldsymbol{\theta}\mathbf{c}] + \lambda \mathbf{c}^T \mathbf{R} \mathbf{c}. \tag{4}$$

Here $\mathbf{R}$ is the roughness penalty matrix. Curve estimation that minimizes the Penalty Sum of Squares is the best compromise between smoothing and goodness of fit.

2.2 Functional Cluster Analysis

Cluster analysis is an analysis used to find logical interpretable subgroups of individuals or objects. The aim is to separate the sample of objects (individuals or objects) into a small number of mutually exclusive groups based on similarities. In other words, the objects in a data set should be grouped so that the objects in the same set should be similar and the objects in different sets should be dissimilar.

In classical cluster analysis, there are two basic clustering techniques, k-means, and hierarchical clustering. Hierarchical clustering involves the establishment of a tree-like structural hierarchy using combiner or separator procedures. In the k-means method, when the number of clusters to be arranged is determined, the objects are assigned to the clusters (Giraldo et al., 2012). As stated by Di Battista and Fortuna (2016), since the functional centroids are representative of a cluster, they should belong to the same function space as the observed functions. In functional cluster analysis, clustering algorithms are used to create subgroups for a set of curves. While the curves in clusters are similar to each other in clustering, they are not as similar to other clusters as possible (Strandberg, 2013).

In certain cases, the same kind of information is recorded in the same unit at different points in time. Each unit's data is also collected at unevenly scaled times. In order to aggregate these data , the data format and time data structure should be taken into account. These data are also obtained in a format that is unevenly sized. Because of the alignment with conventional "variable-to-variable" records, at regular times there are several blank records. These empty records are known to be lost data. Moreover, while all units are observed at the same time, typical clusters do not take into account the structure of the temporal order that is supposed to have similar values for the individuals. Functional data is available to aggregate this type of data (Tzeng et al., 2018). The infinite dimensionality of functional data is a common problem for all clustering methods and leads to some additional difficulties, such as the absence of a functional random variable's probability density function, the definition of distances, or the estimation of noisy data. Several methods have been developed to solve these problems (Jacques and Preda, 2014; Tzeng et al., 2018), which can be primarily divided into three approaches: two-stage clustering, distance-based clustering, or non-parametric clustering and model-based clustering (Léger and Mazzuco, 2020). In Jacques and Preda (2014), a thorough analysis of clustering methods can be found.

In this study, distance-based methods were used. Distance-based clustering approaches usually consist of identifying unique distances or dissimilarities for functional data and then applying a hierarchical or k-means approach as clustering algorithms (Léger and Mazzuco, 2020). In clustering functional data, it is useful to smooth out observations and cluster smoothed curves rather than observed data (Hitchcock et al., 2007). Distance-based methods prefer to use coefficients of some functions as input. A relatively modest amount of smoothing is applied to the functional data observed as the main strategy. The aim is not to change the structure of the curves before clustering, but to remove noise (Ferreira and Hitchcock, 2009).

Suppose we have an example of curves such as $x_1(t), x_2(t), \ldots, x_n(t)$, and the curves of the integrable functions defined in $T = [t_1, t_n]$ belong to the separable Hilbert space H (Giraldo et al., 2012).

In this study, the hierarchical clustering method based on the distance matrix is used. Cluster analysis methods based on the distances matrix are developed as in the classical layout, but the distances are treated as distances between the $x_i(t)$ and $x_j(t)$ curves (Giraldo et al., 2012; Henderson, 2006; Hitchcock et al., 2007; Ferreira and Hitchcock, 2009). By using the expression basis function expansion in Eq.(5),

$$d_{ij} = \sqrt{\int_{[t_1,t_n]} (x_i(t) - x_j(t))^2 \, dt}. \tag{5}$$

$$d_{ij} = \sqrt{\int_{[t_1,t_n]} (\mathbf{c_i} - \mathbf{c_j})^T \, \mathbf{\Phi}(\mathbf{t})\mathbf{\Phi}(\mathbf{t})^T \, (\mathbf{c_i} - \mathbf{c_j}) \, dt}. \tag{6}$$

$$W = \sqrt{\int_{[t_1,t_n]} \mathbf{\Phi}(\mathbf{t})\mathbf{\Phi}(\mathbf{t})^T dt}. \tag{7}$$

$$d_{ij} = \sqrt{(\mathbf{c_i} - \mathbf{c_j})^T \, \mathbf{W} \, (\mathbf{c_i} - \mathbf{c_j})}. \tag{8}$$

can be obtained.

Here, $\mathbf{c_i}$ and $\mathbf{c_j}$ are vectors of basis coefficients for i.th and j.th individuals. The Gram matrix W is the unit matrix for any orthonormal basis, such as the Fourier basis. W must be determined by numerical integration for other base functions, such as B-Splines. After calculating these distance matrices, standard combiner or separator clustering methods can be applied. However, the functions of interest may be of the desired degree derivatives, for example, first derivatives. For example, in a study on the length data of boys and girls, the main concern may be the first derivative (growth rates), or the rate of change in growth rate (i.e. acceleration) given by the second derivative (Clarkson et al., 2005).

For example, if the main point of interest is the growth rate, the distances between two growth curves functions $x_1(t)$ and $x_2(t)$ for two individuals as the square root of integrable square distances between the first derivatives of the two length curves can be defined as

$$d\left(x_1(t), x_2(t)\right) = \sqrt{\int \left(\frac{dx_1(t)}{dt} - \frac{dx_2(t)}{dt} \right)^2 dt}. \tag{9}$$

Length measurements depend on the rate of growth change rather than final lengths (Clarkson et al., 2005).

If generalized, the distances between the two curves, $x_i(t)$, and $x_j(t)$, depending on the desired order derivative, can be expressed as in Eq(10) (Huzurbazar and Humphrey, 2008; Jacques and Preda, 2014).

$$d\left(x_i(t), x_j(t)\right) = \sqrt{\int \left(x_i^l(t) - x_j^l(t) \right)^2 dt}. \tag{10}$$

2.3 Functional canonical correlation analysis

Canonical Correlation Analysis (CCA) deals with associations between two sets of random variables. CCA for data that are essential in the form of curves is different from those that are multivariate, due to the infinite dimensionality of the spaces that the data belong to (Hosseininasab et al., 2012). Functional canonical correlation analysis (FCCA) is a useful tool to quantify the relationship between paired functional data Shin and Lee (2015). Principle components analysis (PCA) may be used to examine the variation in a curve set, however, it does not give explicit information about the interaction between two curve sets such as $x_i(t)$ and $y_i(t)$ $(i = 1, 2, \ldots, \mathrm{N}, t \in T))$. $x_i(t)$ and $y_i(t)$ are assumed to be observed in a finite range (T) and all functions are integrable. Functional canonical correlation analysis (FCCA) is the functional analogous of CCA and explains the interaction between two different curve sets. For example, one may try to find out the interaction between the variation in the stringency index in the Covid-19 cycle and the variation in new deaths from the Covid-19 cycle.

The correlation surface, $r(s,t); (s,t \in T)$ in Equation (11), gives correlations between every pair of $x_i(t)$ and $y_i(t)$. The height of the surface indicates the variation of the curves at every point of time (or relevant variable) and gives a measure of covariation. However, when the data is huge, the surface becomes very complicated and hard to interpret. Then, it becomes impossible to visually reveal the dominant modes of variation between two curve sets. The same situation arises in multivariate data analysis when interpreting the variance-covariance matrix. Due to the difficulty of interpreting a multi-dimensional variance-covariance matrix, CCA is utilized. In CCA, the interaction between two sets of random vectors is explained in terms of several carefully chosen covariances.

$$r(s,t) = \operatorname{cor} r_{X,Y} = \frac{\operatorname{cov}_{x,y}, (s,t)}{\sqrt{\operatorname{var}_x(s)}\sqrt{\operatorname{var}_y(t)}}. \tag{11}$$

FCCA gives the function pair $(\xi(t), \eta(t))$ that maximizes the correlation between the canonical variates, which are linear components such as $\int \xi(t)x_i(t)dt$ and $\int \eta(t)y_i(t)dt$. For example, $\xi(t)$ and $\eta(t)$ can be assumed to be the components of variation which explains most of the interaction between stringency index in Covid-19 and new deaths from Covid-19. In other words, $\xi(t)$ and $\eta(t)$ are the canonical variable weight functions corresponding to stringency index and new deaths.

We need to find $\xi(t)$ and $\eta(t)$ that maximizes (12) under the constraints (13).

$$Cov\left(\int \xi X_i, \int \eta Y_i\right) = \iint \xi(s)\Gamma_{12}(s,t)\eta(t)dsdt. \tag{12}$$

$$\begin{aligned}\left(var \int \xi X_i\right) &= \iint \xi(s)\Gamma_{11}(s,t)\xi(t)dsdt = 1,\\ \left(var \int \eta Y_i\right) &= \iint \eta(s)\Gamma_{22}(s,t)\eta(t)dsdt = 1.\end{aligned} \tag{13}$$

$$\begin{aligned}\Gamma_{11}(s,t) &= \mathrm{E}(x(s), x(t)),\\ \Gamma_{22}(s,t) &= \mathrm{E}(y(s), y(t)),\\ \Gamma_{12}(s,t) &= \mathrm{E}(x(s), y(t)).\end{aligned} \tag{14}$$

Here, (14) are the covariance functions. We benefit from and combine the notations of Leurgans et al. (1993) and Ramsay and Silverman (2005) in this study.

However, applying this maximization directly does not give meaningful results. When directly applied, "there is always a pair of linear functionals of x and y that are perfect correlated on the basis of the sample". It is worth noting that the presence of unit canonical correlations is also present with discretized data when the number of readings on the "time" ordinate exceeds the number of sample curves (Kupresanin, 2008).

Optimizing at the same time with respect to two probes makes CCA an extremely greedy technique where data mining uses this concept. CCA can capitalize on the smallest difference in any set of functions in optimizing this ratio to the degree that it may be difficult to see anything of significance in the result unless we exert any control over the method. In practice , it is important to impose strong smoothness on the two weight functions of $\xi(t)$ and $\eta(t)$ to limit this greediness. This can be achieved either by choosing a low-dimensional basis for each one or by using an explicit penalty for roughness in the same way as possible for functional PCA (He et al., 2004; Kupresanin, 2008; Ramsay et al., 2009; Keser, 2014).

Unlike PCA, the need for smoothing in FCCA is not a feature specific to certain data sets but rather a general feature that must be followed for all data sets. In order to add the smoothing into the analysis, constraints in Eq(13) should be modified to include the roughness penalty term as follows:

$$\begin{aligned} \operatorname{var}\left(\int \xi x_i\right) + \lambda_1 \left\|D^2\xi\right\|^2 &= \iint \xi(s)\Gamma_{11}(s,t)\xi(s)dsdt + \lambda_1 \int \left(D^2\xi(t)\right)^2 dt = 1, \\ \operatorname{var}\left(\int \eta x_i\right) + \lambda_2 \left\|D^2\eta\right\| &= \iint \eta(s)\Gamma_{11}(s,t)\eta(s)dsdt + \lambda_1 \int \left(D^2\eta(t)\right)^2 dt = 1. \end{aligned} \tag{15}$$

Here, $\lambda_i (i = 1, 2)$ is a positive value and is referred to as the smoothing parameter chosen for regularizing the variances of canonical variates. The problem of maximizing the covariance in (12) under (15) constraints is equal to the maximization of the penalized squared correlation in (16) in terms of ε and η.

$$\operatorname{ccors} q_\lambda(\xi, \eta) = \frac{\operatorname{cov}\left(\int \xi x_i, \int \eta y_i\right)^2}{\left\{\operatorname{var}\left(\int \xi x_i\right) + \lambda_1 \left\|D^2\xi\right\|^2\right\}\left\{\operatorname{var}\left(\int \eta x_i\right) + \lambda_2 \left\|D^2\eta\right\|^2\right\}}. \tag{16}$$

This procedure is referred to as smoothed canonical correlation analysis (SCCA) (Leurgans et al., 1993). Clearly the larger value of λ_1 and λ_2, the more emphasis will be placed on the roughness penalty and the smaller will be the true correlation of the variates found by SCCA. A good choice of the smoothing parameters is essential so that we have a pair of canonical variates possessing fairly smooth weight functions and correlation that is not unreasonably low (Leurgans et al., 1993). In practice, values of λ are often taken identical.

After obtaining the first components of variation, $\xi_1(t)$ and $\eta_1(t)$, other components of variation can also be found. They must ensure the following assumptions:

- correlation between and should be high.
- weight functions should be orthogonal as in (17).

$$\int \xi_i(t)\xi_j(t)dt = 0 \quad i \neq j \quad \text{and} \int \eta_i(t)\eta_j(t)dt = 0 \quad i \neq j. \tag{17}$$

In classical canonical correlation analysis, the number of eigenvectors that maximize the correlation between two variable sets is equal to the number of variables in the smaller set. In FCCA, the number of eigenvectors is taken as the smallest one of N, which is the dimension of x_i and y_i, or K, which is the number of basis functions for ξ and η. The first canonical correlation gives us most of the information about the relation between the curve sets.

The factor having a canonical variate as its coefficient has a directional interpretation in the data space in the sense that it shows the functional change in one process along with the associated functional changes in the other process (Shin and Lee, 2015).

The studies of He et al. (2000, 2003, 2004) give alternative approaches to basis functions for canonical correlation analysis. The historical progress of FCCA is given in Hosseininasab et al. (2012). Several approaches for FCCA have been developed so far. Recent work on FCCA includes Cupidon et al. (2007, 2008); Eubank and Hsing (2008); Krzyśko and Waszak (2013); Madrigal (2017); Górecki et al. (2018, 2020).

3 Data and analysis

In this study, the number of daily deaths per million for 27 countries that are in the European Union was examined. The primary reason for the examination of the European Union countries is that the virus is more easily spread because of the free movement in the European Union members and the catastrophic start of the pandemic in Italy.

Our data source is the OWID (Our World In Data) data repository (Beltekian, 2020) for the number of daily deaths per million. OWID is a collaborative effort between researchers at the University of Oxford, who are the scientific editors of the website content; and the non-profit organization Global Change Data Lab, who publishes and maintains the website and the data tools. Stringency index data is taken from Oxford University Government Response Tracker (2020).

The number of new deaths was analyzed between 12.03.2020 -23.06.2020. Twenty-seven individual new deaths functions, one for each country, are obtained by using the roughness penalty approach via B-Spline basis functions. The lambda smoothing parameter was taken as 0.01, subjectively so as not to disturb the general appearance of the data.

It is seen that there are noticeable increases for all countries between 12.03.2020 and 23.06.2020. However, there are some countries that are prominent in these increases, and these can be seen on the legend in Figure 4, beginning from the highest peak points. The increase in new deaths started primarily in Italy, followed by countries such as Spain, Belgium, France, Sweden, and the Netherlands. In fact, when peak points are examined, it is seen that Spain, France, and Belgium have crossed Italy. Although the names of Italy, Spain, and France, which are among the member states of the European Union,

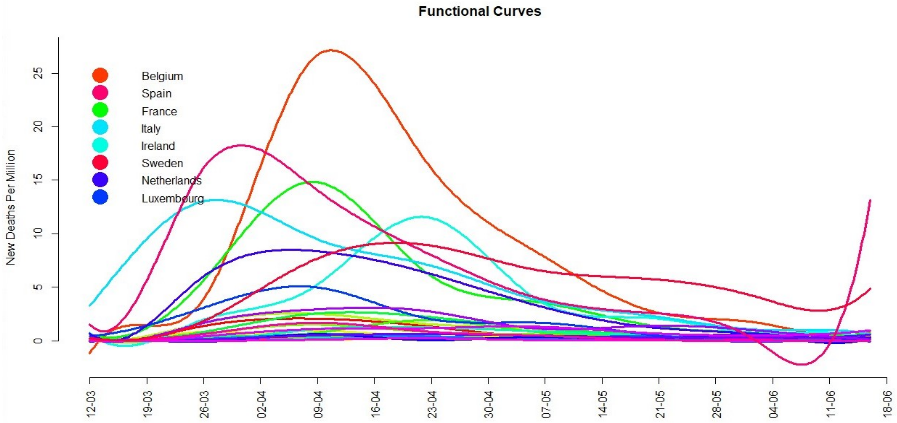

Figure 4: 27 individual new deaths curves

are in the foreground in the spread of the Covid 19 epidemic, it is seen that Belgium has a much higher mortality rate than others in a wide period.

In this study, while examining the curve corresponding to each country, one of the most important points is of course the reliability of the data. In some European Union countries, there is confusion in reporting the number of deaths. Some citizens criticize their government for not disclosing the correct numbers. In Germany, Italy, and Spain, it is stated that only those who died in hospitals are registered and those who died in nursing homes and homes are not included in these statistics, while Belgium faces a different problem. It is stated that Belgium has a different reporting method than other European Union countries and that even suspicious deaths in nursing homes are recorded as Covid-19.

In terms of new deaths, the dendrogram created according to the results of the countries' hierarchical clustering analysis for functional data is given in Figure 5 and the clustering table created according to the dendrogram is given in Figure 6.

When the dendrogram in Figure 5 is examined, it is clearly seen that Belgium is at the top as a separate cluster, as observed from the curves. It is seen that France is also a separate cluster from the others. Even though Italy and Spain seem closer than others in terms of distance, it is seen that Italy and Spain are actually separate clusters when all time intervals and curves increase and decrease rates are examined.

The reason why Italy and Spain are relatively close to each other is that they share many features such as high social citizens, beautiful weather conditions, densely populated cities, physically loving social interactions, and older elderly populations. Besides, officials in both countries have initially underestimated how quickly the virus can spread and how quickly it can push health systems to the brink of collapse.

In summary, attention should be paid to the formation of logical clusters in cluster analysis as well as statistical calculations. Here, 6 clusters have been determined by considering the dendogram, k-means cluster and the social structures of the countries together.

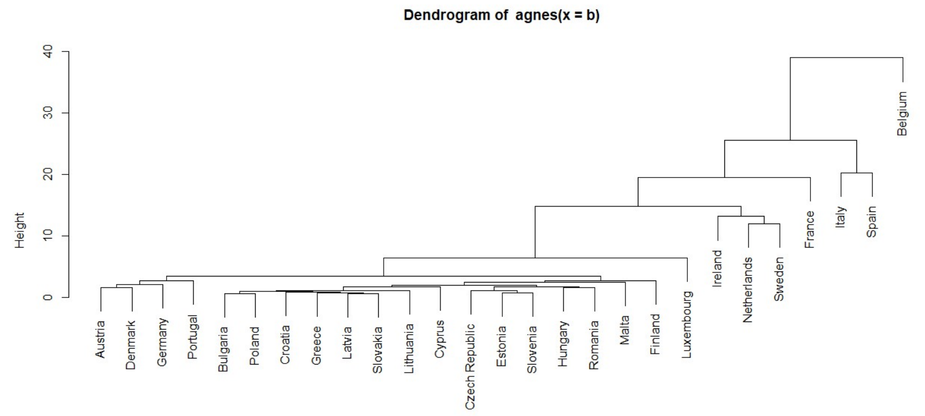

Figure 5: Dendrogram created according to hierarchical cluster analysis for functional data

When the cluster mean curves in Figure 7 are analyzed, it is seen that the mean curve of the cluster-1, which has the highest number of countries, is quite lower than the others in the entire time interval. The mean curve of the fourth cluster of Ireland, Sweden, and the Netherlands is relatively higher than the first cluster and appears to have peaked in early May. However, other clusters, in other words, Belgium, Spain, France, and Italy, in total, are far above in terms of mortality compared to these 23 countries. It was observed that the peak timings of Belgium and France and Italy and Spain were close to each other, but the number of new deaths was different, so they were in different clusters. Belgium followed the peak in Italy soon with a higher value and the peak in France with a higher value for a while. Belgium and France are two neighboring countries. These peaks and the magnitudes of the ups and downs can be observed even more comfortably in the first derivative functions (Figure 8).

When the first derivative functions of the new deaths of the countries in Figure 8 are analyzed, in other words, when the increase and decrease rates of new deaths are analyzed, the changes in the speed functions of Belgium and Spain are clearly seen. The speed curves of other countries vary in the range of ± 1 band. France and Italy, which followed Belgium and Spain in new deaths, did not go beyond the lower and upper limits of this band, although the new deaths increased and decreased. With the help of the derivative function, the size of the ups and downs of new deaths between countries can be compared more easily.

If classical cluster analysis were used, Spain and Italy would most likely be in the same cluster when a time point where the curves were close was examined, but since the two countries showed different ups and downs throughout the relevant time point, they were actually in two separate clusters. We can see them much more clearly in Figure 8, which gives the first derivative curves of the curves, in other words, the increase and decrease rates, and Figure 9, which gives the cluster according to the first derivatives. However, it is seen that Italy, Spain, France, and Belgium, which are prominent countries in the

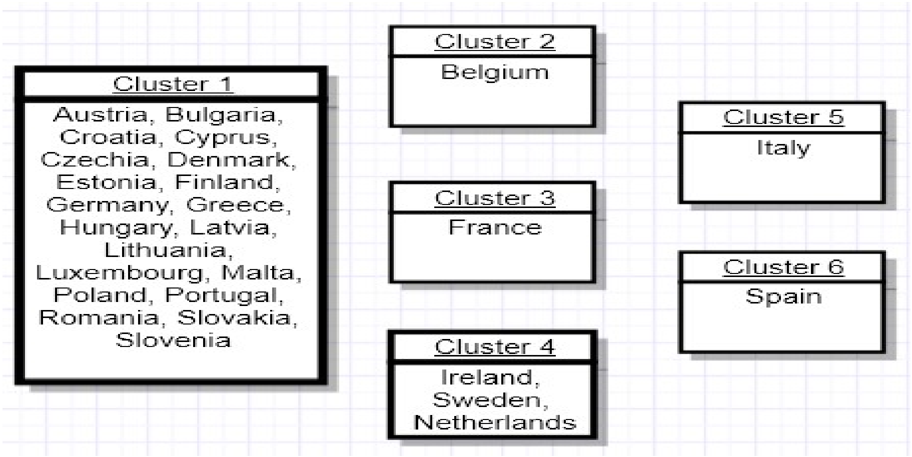
Cluster 1
Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, Germany, Greece, Hungary, Latvia, Lithuania, Luxembourg, Malta, Poland, Portugal, Romania, Slovakia, Slovenia
Cluster 2
Belgium
Cluster 3
France
Cluster 4
Ireland, Sweden, Netherlands
Cluster 5
Italy
Cluster 6
Spain

Figure 6: Clusters created according to hierarchical cluster analysis for functional data

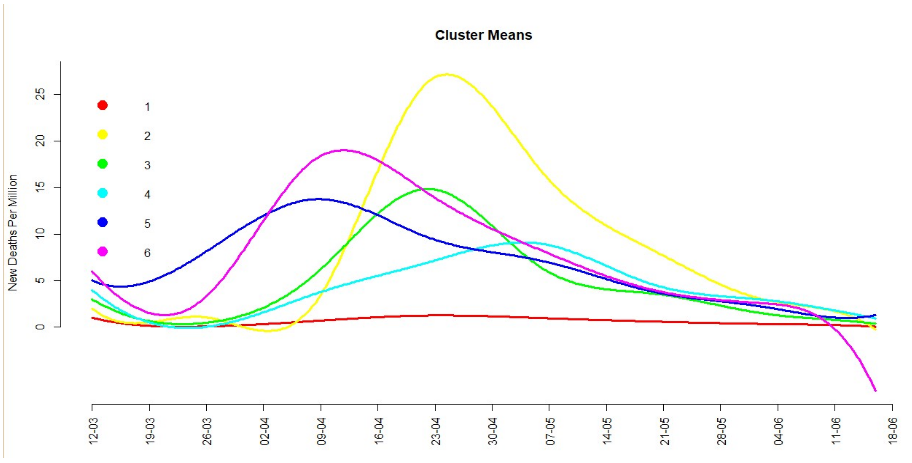

Figure 7: Cluster mean curves

outbreak of 20 countries, are clustered separately.

These results are consistent with the study of da Silva and Tsigaris (2020), which states that as countries' GDP increases, mortality rates also increase. They stated that as the GDP value of the countries increased, they travel more, and their possibility of getting sick increased. The countries that have the highest death rate increase are in the higher position in terms of GDP.

When the hierarchical cluster analysis is made according to the first derivative (Figure 9), which also takes into account the ups and downs of the new deaths, it is seen that the first cluster is not separated, in other words, the speed curves of the new deaths are similar in this period. However, Belgium and Spain are at the far end, as separate clusters, as they are the curves with the highest ups and downs. Italy and France also took place as separate clusters and even Italy and Spain diverged. Sweden and the Netherlands, which are in the same cluster here, have two closest countries again

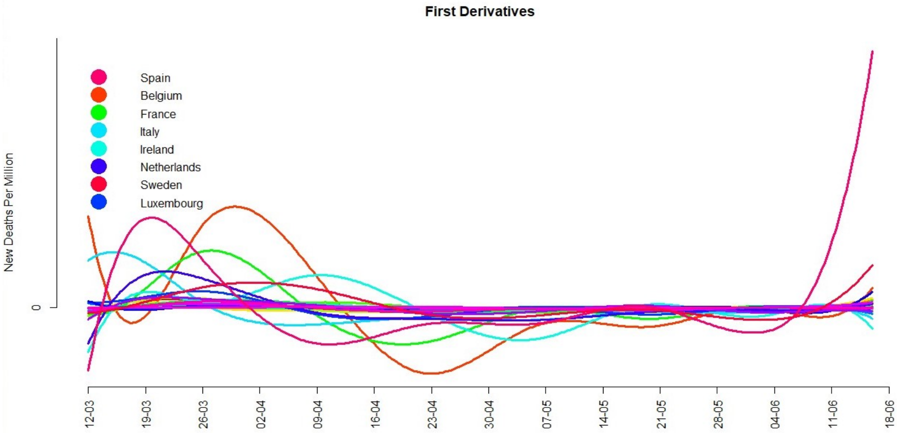

Figure 8: First derivative curves

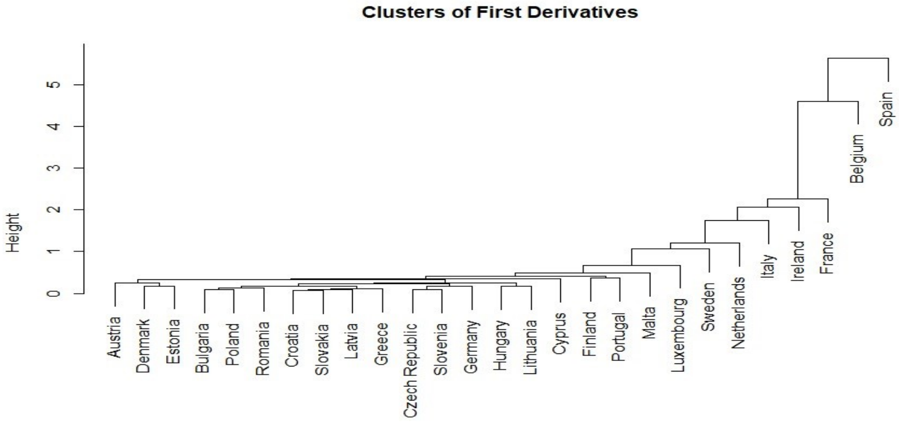

Figure 9: Functional hierarchical cluster analysis by first derivatives

and Ireland is located not far from them. In other words, in the cluster of Sweden, Netherlands, and Ireland, Sweden and Netherlands are closer to each other than Ireland in terms of the speed of new deaths.

Considering that the effects of Covid-19 may have been reflected in some countries later, curve registration, which is frequently used in functional data analysis, has been applied and it has been investigated whether this delay has an effect on the clustering of countries. When the results were examined, it was seen that there was no big change in the clusters and only the Netherlands, which is in the same cluster with Ireland and Switzerland, clustered together with France. In other words, it cannot be said that the onset of the epidemic later in other countries has a significant effect on the curves and the changes they show. The registered curves and the result of their cluster analysis are given in (Figure 10) and (Figure 11).

With the functional cluster analysis, after examining the clusters of the curves accord-

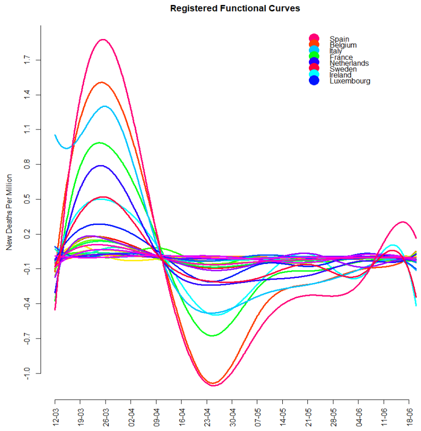

Figure 10: Registered Curves

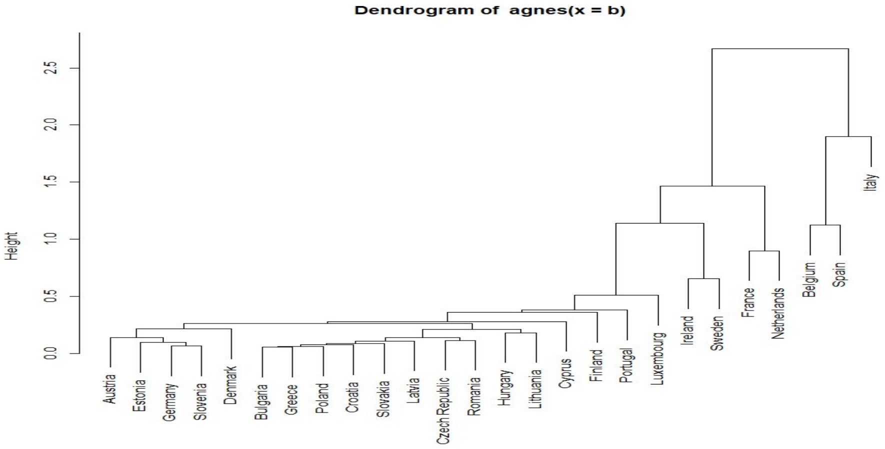

Figure 11: Functional hierarchical cluster analysis by registered curves

ing to themselves and then to their first derivatives, and determining the basic clusters, the relationship between the ”stringency index” was created according to the measures developed against Covid 19 and the number of deaths was investigated. At the same time, the positions of the countries according to their highly related variability structures were determined.

The functional curves of the stringency index for 27 countries (Figure 12) were obtained again by using the Roughness Penalty approach via B-Spline basis functions. The first canonical correlation between the index and new deaths was 0.978. Naturally, a strong relationship is expected between the new deaths and the index, which reflects strict security measures. This high relationship value meets these expectations.

In addition, when we tested the importance of the canonical correlation coefficient, the p value was found to be 0.00068, and accordingly, it can be said that the correlation coefficient is significant. For the significance test, code from Lin et al. (2017) article was

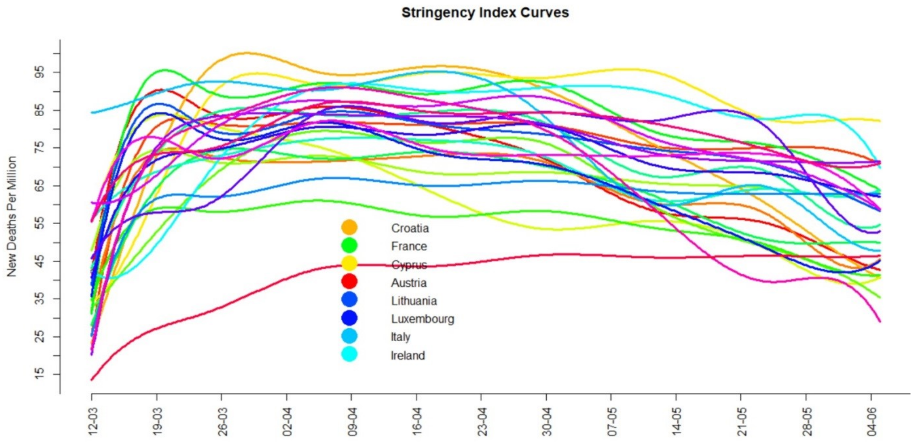

Figure 12: Stringency index for 27 countries

used.

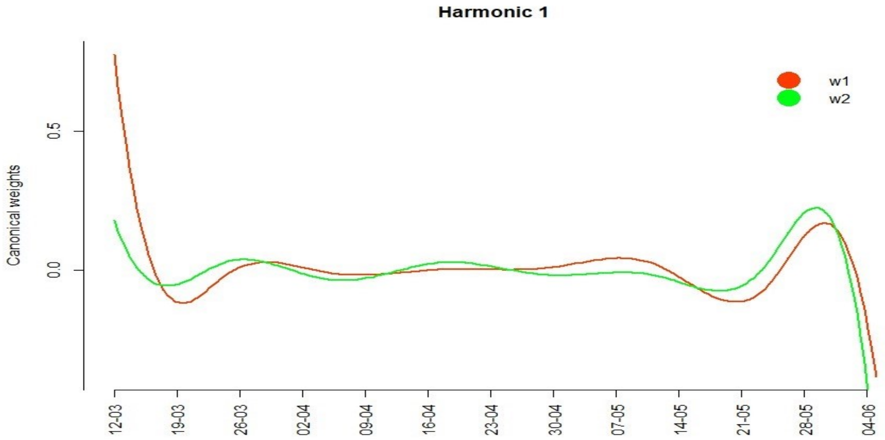

Figure 13: First canonical weight functions associated with the first canonical correlation

Figure 13 shows the two interacted canonical weight functions associated with the first canonical correlation. These two functions show related variability structures and fluctuations appear to follow each other very closely. It can be concluded that new deaths per million and stringency index are highly linearly correlated.

Figure 14 shows the relative positions of 27 EU countries due to the degree of the relationship between two variables characterizing them: the daily new deaths from Covid-19 and the stringency index, recorded in the days 12.03.2020-23.06.2020. The extreme position is occupied as expected by Italy. Especially, many countries in the first cluster were found to be in a close position here.

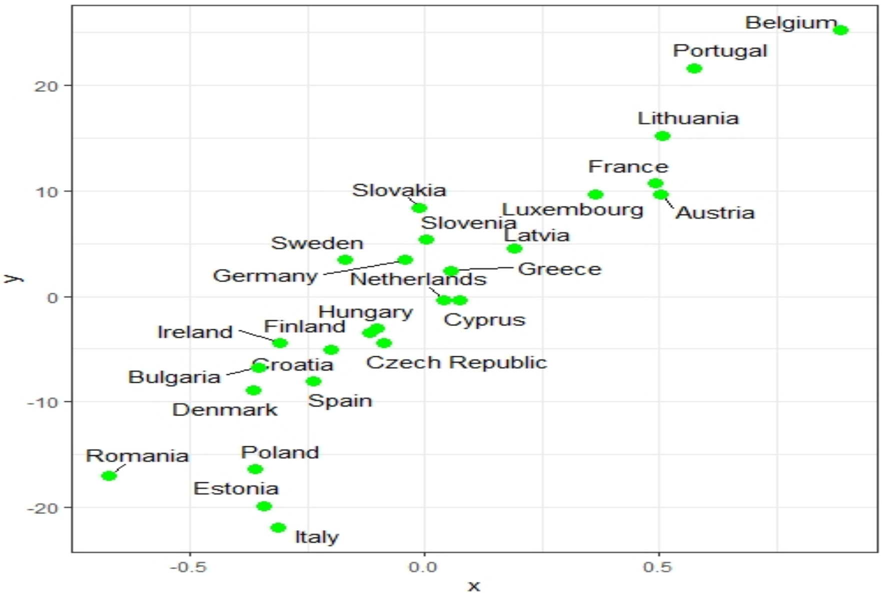

Figure 14: Canonical scores for the first pair of canonical variables

4 Discussion

When the number of new deaths per million population of the European Union member countries in the period of interest is analyzed by functional analysis, the countries that differ from and resemble each other can be clearly seen. While twenty countries in cluster 1, which is a member of the European Union, are located very close to each other in terms of both the ups and downs of the curves and their distance from each other, it is seen that the main fluctuations and extreme situations are in Belgium, Italy, Spain, and France. These countries need to be followed carefully and the measures to be increased. However, Belgium and Spain are at the far end, as separate clusters, as they are the curves with the highest ups and downs.

At the same time, a very high relationship of 0.978 between the stringency index, which reflects the measures taken against the epidemic, and the number of new deaths, as expected, has been revealed. When the positions of the countries with respect to each other are examined here, it can be seen that Italy can be evaluated as an extreme value and the positions of the countries in cluster 1 are very close.

Although these results are not very surprising, using functional data analysis instead of usual line graphs or time series analysis and making the statistical connections between stringency index and the number of new deaths might give future researchers a new perspective to use. Functional data analysis allowed the examination of the new deaths and stringency curves of the countries individually and the derivative curves taking into account the ups and downs of the curves in the relevant period, and also the clusters of the countries with the original curves could be compared.

References

Beltekian, D., G. D. G. C. (2020). Data on covid-19 (coronavirus) by our world in data.

Chowdhury, R., Heng, K., Shawon, M. S. R., Goh, G., Okonofua, D., Ochoa-Rosales, C., Gonzalez-Jaramillo, V., Bhuiya, A., Reidpath, D., Prathapan, S., et al. (2020). Dynamic interventions to control covid-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries. *European journal of epidemiology*, 35(5):389–399.

Clarkson, D. B., Fraley, C., Gu, C., and Ramsay, J. (2005). *S+ Functional Data Analysis: User's Manual for Windows®*. Springer Science & Business Media.

Cupidon, J., Eubank, R., Gilliam, D., and Ruymgaart, F. (2008). Some properties of canonical correlations and variates in infinite dimensions. *Journal of Multivariate Analysis*, 99(6):1083–1104.

Cupidon, J., Gilliam, D., Eubank, R., Ruymgaart, F., et al. (2007). The delta method for analytic functions of random operators with application to functional data. *Bernoulli*, 13(4):1179–1194.

da Silva, J. T. and Tsigaris, P. (2020). Policy determinants of covid-19 pandemic–induced fatality rates across nations. *Public health*, 187:140–142.

Di Battista, T. and Fortuna, F. (2016). Clustering dichotomously scored items through functional k-means algorithm. *Electronic Journal of Applied Statistical Analysis*, 9(2):433–450.

ECDC (2020). European centre for disease prevention and control website.

Eubank, R. and Hsing, T. (2008). Canonical correlation for stochastic processes. *Stochastic Processes and their Applications*, 118(9):1634–1661.

Ferreira, L. and Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, 38(9):1925–1949.

Giraldo, R., Delicado, P., and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, 66(4):403–421.

Górecki, T., Krzyśko, M., Waszak, Ł., and Wołyński, W. (2018). Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, 59(1):153–182.

Górecki, T., Krzyśko, M., and Wołyński, W. (2020). Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data. *Artificial Intelligence Review*, 53(1):475–499.

He, G., Müller, H.-G., and Wang, J.-L. (2000). Extending correlation and regression from multivariate to functional data. *Asymptotics in statistics and probability*, pages 301–315.

He, G., Müller, H.-G., and Wang, J.-L. (2003). Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis*, 85(1):54–77.

He, G., Müller, H.-G., and Wang, J.-L. (2004). Methods of canonical analysis for functional data. *Journal of Statistical Planning and Inference*, 122(1-2):141–159.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics: The official journal of the International Environmetrics Society*, 17(1):65–80.

Hitchcock, D. B., Booth, J. G., and Casella, G. (2007). The effect of pre-smoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation*, 77(12):1043–1055.

Hosseininasab, S. M. E., Faridrohani, M., and Golshan, A. M. (2012). Functional analysis of current and noncurrent balance facilities of iranian export development bank. *Journal of Statistical Theory and Applications*, 11(2):121–142.

Huzurbazar, S. and Humphrey, N. F. (2008). Functional clustering of time series: An insight into length scales in subglacial water flow. *Water resources research*, 44(11).

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255.

Keser, I. K. (2014). Comparing two mean humidity curves using functiona t-tests: Turkey case. *Electronic Journal of Applied Statistical Analysis*, 7(2):254–278.

Krzyśko, M. and Waszak, Ł. (2013). Canonical correlation analysis for functional data. *Biometrical Letters*, 50(2):95–105.

Kupresanin, A. M. (2008). *Topics in functional canonical correlation and regression.* PhD thesis, Arizona State University.

Léger, A.-E. and Mazzuco, S. (2020). What can we learn from functional clustering of mortality data? an application to hmd data. *arXiv preprint arXiv:2003.05780.*

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(3):725–740.

Lin, N., Zhu, Y., Fan, R., and Xiong, M. (2017). A quadratically regularized functional canonical correlation analysis for identifying the global structure of pleiotropy with ngs data. *PLoS computational biology*, 13(10):e1005788.

Madrigal, P. (2017). fccac: functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets. *Bioinformatics*, 33(5):746–748.

Matabuena, M., Vidal, J. C., Hayes, P. R., Saavedra-García, M., and Trillo, F. H. (2019). Application of functional data analysis for the prediction of maximum heart rate. *IEEE Access*, 7:121841–121852.

Oxford University Government Response Tracker (2020). Coronavirus government response tracker.

Ramsay, J. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.

Ramsay, J., Hooker, G., and Graves, S. (2009). Introduction to functional data analysis. In *Functional data analysis with R and MATLAB*, pages 1–19. Springer.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis.* Springer-Verlag.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis.* Springer.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):539–561.

Shin, H. and Lee, S. (2015). Canonical correlation analysis for irregularly and sparsely observed functional data. *Journal of Multivariate Analysis*, 134:1–18.

Strandberg, J. (2013). Cluster analysis for functional data. Master's thesis, Umeå University, Sweden.

Tzeng, S., Hennig, C., Li, Y.-F., and Lin, C.-J. (2018). Dissimilarity for functional data clustering based on smoothing parameter commutation. *Statistical methods in medical research*, 27(11):3492–3504.

Worldometer (2020). Covid-19 coronavirus pandemic.