Communications for Statistical Applications and Methods
2022, Vol. 29, No. 1, 103–125

https://doi.org/10.29220/CSAM.2022.29.1.103
Print ISSN 2287-7843 / Online ISSN 2383-4757

Exploring COVID-19 in mainland China during the lockdown of Wuhan via functional data analysis

Xing Li[a], Panpan Zhang[1,b], Qunqiang Feng[a]

[a]Department of Statistics and Finance, University of Science and Technology of China, China;
[b]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, USA

Abstract

In this paper, we analyze the time series data of the case and death counts of COVID-19 that broke out in China in December, 2019. The study period is during the lockdown of Wuhan. We exploit functional data analysis methods to analyze the collected time series data. The analysis is divided into three parts. First, the functional principal component analysis is conducted to investigate the modes of variation. Second, we carry out the functional canonical correlation analysis to explore the relationship between confirmed and death cases. Finally, we utilize a clustering method based on the Expectation-Maximization (EM) algorithm to run the cluster analysis on the counts of confirmed cases, where the number of clusters is determined via a cross-validation approach. Besides, we compare the clustering results with some migration data available to the public.

Keywords: COVID-19, functional canonical correlation, functional cluster analysis, functional principal component analysis, migration

1. Introduction

A novel strain of coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Anderson *et al.*, 2020) causes a highly infectious respiratory illness, named COVID-19 by the World Health Organization (WHO). The novel coronavirus was first detected and reported in the city of Wuhan, Hubei province in China in December 2019 (CDC, 2020). Since its outbreak, the COVID-19 has spread to more than 180 countries and territories all over the world, giving rise to global public health emergency. On Mar 11, 2020, the WHO declared the outbreak of COVID-19 a global pandemic. As of May 14, 2020, COVID-19 has caused a total of 4,170,424 confirmed cases and 287,399 death cases according to COVID-19 situation report 114 by the WHO. Besides, the COVID-19 also has brought severely negative impact to global economy. For instance, according to the National Bureau of Statistics of China, China's economy has fallen 6.8% year-on-year in the first quarter of 2020, which was the first contraction in quarterly data on record since 1992. Businesses and industries related to travel and tourism all around the world (including China) are faced with tremendous economic losses that are unlikely recoverable.

Since the outbreak of COVID-19, every country has made a great deal of efforts to control or slow down the spread of the virus and mitigate its drastic impacts. Under strict control measures, Wuhan, the outbreak center with population size greater than 11 million, was put on a lockdown for 76 days by Chinese government. The imposed lockdown resulted in travel restrictions to and from Wuhan,

[1] Corresponding author: Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 423 Guardian Drive, Blockley Hall 501, Philadelphia, PA 19104, USA. E-mail: panpan.zhang@pennmedicine.upenn.edu

Published 31 January 2022 / journal homepage: http://csam.or.kr

and even local transportation, including buses, metros and all kinds commercial vehicles were halted. The residents and nonresidents (those who did not exit Wuhan before the lockdown) in Wuhan were obligated to practice home quarantine. Outdoor activities were extremely limited. Each citizen was given a permission card and only allowed to go outdoor every the other day for a maximum of 30 minutes.

Strict control measures may help reduce the spread of the disease, but the scientists and epidemiologists urge to fully understand the virus so as to eliminate it thoroughly. There remains a lot of unknowns about the novel coronavirus. As of this writing, there is not any known medication that can directly kill SARS-CoV-2, and the vaccine doses are still at an early stage of testing. Hence, many infected patients have to rely on their immune systems to recover under the help of some standard treatments. As more and more COVID-19 data become available, it is extremely critical to carry out data-driven analysis to precisely characterize the epidemic behavior of COVID-19 and to unveil its essence.

Many COVID-19 data are stored sequentially in time order. In this article, we exploit functional data analysis methods to analyze the COVID-19 data in mainland China, as it is one of the earliest countries attacked by the virus. The infectious disease has been effectively controlled in China thanks to the implementation of a series of successful intervention protocols and measures (including mandatory quarantine policies, travel bans and color-based "health codes"). Recently, there emerged a few research articles investigating the COVID-19 data from different countries over the world via functional data analysis methods, such as the United States (Tang *et al.*, 2021) and Italy (Boschi *et al.*, 2021). Our research goal is to make reliable inference through a comprehensive analysis of available COVID-19 data, in order to provide constructive suggestions to many other countries that are still enormously affected by the virus. Specifically, we would like to answer the following questions,

1. Do the public health measures implemented in China (such as strict isolation measures and mandatory mask wearing policy) help reduce the spread of COVID-19?
2. Is there any variation shared by confirmed and death cases; if so, how to quantitatively measure it?
3. Does it exist any community structure (at city or province level) during the outbreak of COVID-19 in China?

The rest of the manuscript is organized as follows. In Section 2, we give a succinct description of the collected data and the sources of the data. In Section 3, we introduce the functional data analysis methods that will be used to analyze the COVID-19 data of China throughout the manuscript. The analysis results and their interpretations are presented in Section 4. We finally address some concluding remarks and raise some discussions in Section 5.

2. Data description and validation

The first author (XL) collected the COVID-19 data for the present analysis from the official websites of the Chinese Center for Disease Control and Prevention (China CDC) and the National Health Commission of China (China NHC). Specifically, XL recorded the number of daily confirmed and death cases from a total of 31 provinces, municipalities and autonomous regions (including Hubei province) in mainland China. Especially for Hubei province, additional data were collected for the 17 cities (including Wuhan) therein from the official website of The Health Commission of Hubei Province. The data collection period is from Jan 23, 2020 (the date that the lockdown of Wuhan was announced by Chinese government) to Apr 8, 2020 (the date that the lockdown of Wuhan officially ended), constituting a complete wave cycle of the spread of COVID-19 in a seriously affected country.

In the process of data collection, XL observed an abnormal increase of the number of confirmed cases on Feb 12, 2020. Specifically, there were 14,840 newly confirmed cases in Hubei province reported on that day, compared to an average of 2,396 from Feb 5 to Feb 11, 2020, and an average of 2,124 in the following week. According to the announcement of China NHC, the significant increase in number was owing to the change of the criterion of "confirmed case" in Hubei province. More precisely, before Feb 12, 2020, the susceptible patients who were tested positive in nucleic acid amplification tests (NAATs) were regarded as confirmed cases; since Feb 12, 2020, confirmed cases would include the susceptible patients who were either tested positive in NAATs or clinically diagnosed according to chest computed tomography (CT) imaging features. This new diagnosis criterion was released by China CDC in the Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (trial fifth version) on Feb 4, 2020. Hence, the number of confirmed cases reported on Feb 12, 2020 was actually equal to the total number of individuals tested positive in NAATs or clinically confirmed(13,332 out of 14,840) from Feb 4 to Feb 12, 2020.

COVID-19 in China broke out right before the Chinese New Year (Spring Festival) of 2020. In tradition, a large amount of travels would have taken place since two weeks before the most important festival in China. Besides, Wuhan is one of the busiest public transportation hubs in China, so it may be worthy of investigating population migration departing from Wuhan during that period. Thus, XL collected the data of population migration from Wuhan (to other domestic destinations) between Jan 10 and Jan 24, 2020 (https://qianxi.baidu.com/2020/). More details about the utilization of this collection of data will be given in the sequel.

3. Methodologies

In this section, we demonstrate the methods that will be used to analyze the time series COVID-19 data in mainland China. Naturally speaking, time series data can be viewed as functional data with respect to time, allowing statisticians to exploit flexible methods from functional data analysis (FDA) in the research, and consequently, triggering a great deal of methodological developments of FDA (Shin *et al.*, 2015; Petersen and Müller, 2016; Floriello and Vitelli, 2017) in the recent years.

Additionally, in the literature, it is evident that FDA methods are prevalent in modeling time series data from public health and biological sciences (Ullah and Finch, 2013), reflected in a large body of classical and recent publications in the literature (Baladandayuthapani *et al.*, 2008; Crawford *et al.*, 2020; Hyndman and Ullah, 2007; Leng and Müller, 2006; Shen *et al.*, 2017; Yao *et al.*, 2005). It is impossible to list them all. Particularly, a few recent papers applying FDA methods to epidemiological studies (Carroll *et al.*, 2020; Rahman and Jiang, 2021) extensively motivate us to utilize standard and advanced FDA methods in the present study.

3.1. Fundamentals of functional data analysis

FDA was firstly termed in (Ramsay, 1982), and its development mushroomed in the 1990s (Ramsay and Dalzell, 1991). A variety of standard methods for FDA are summarized in the text (Ramsay and Silverman, 2002). To begin with, we give the definition of L_2 process.

Definition 1. (L_2 process) *Given an interval I, a stochastic process X is called an L_2 process on I if*

$$\mathbf{E}\int_I |X(t)|^2\, dt < \infty.$$

For $i = 1, 2, \ldots, n$, we denote the functional realizations of an underlying L_2 process $X(t)$ by $X_i(t)$. Let $(t_{i1}, t_{i2}, \ldots, t_{in_i})$ be the time schedule of subject i. Functional observations Y_{ij} that account for

measurement errors ε_{ij} are given by

$$Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij},$$

under the assumptions of $\mathbf{E}(\varepsilon_{ij}) = 0$ and $\mathbf{V}\text{ar}(\varepsilon_{ij}) = \sigma^2_{ij}$. Furthermore, when $\sigma^2_{ij} = \sigma^2$ is assumed across i and j, we call the measurement errors *homoscedastic*.

Although the functional data are not necessarily sampled according to the same time schedule, n_i's are not identical for all i's, the COVID-19 data for this study are collected daily from Jan 23 to Apr 8, 2020. We hence can simplify the notation for time schedule to $t_{ij} = t_j$ for $j = 1, 2, \ldots m$. For any given t_j, let $\mu(t_j) := \mathbf{E}(X(t_j))$ denote the mean function, and for any pair of t_r and t_s, let $\Sigma(t_r, t_s) := \mathbf{C}\text{ov}(X(t_r), X(t_s))$ denote the covariance function. The functional data can be viewed as an m-dimensional multivariate data structure, thus allowing us to empirically estimate the mean and covariance functions at each time point,

$$\hat{\mu}(t_j) = \frac{1}{n}\sum_{i=1}^{n} Y_{ij} \quad \text{and} \quad \hat{\Sigma}(t_r, t_s) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(Y_{ir} - \hat{\mu}(t_r)\right)\left(Y_{is} - \hat{\mu}(t_s)\right)\right].$$

The estimated mean curve and covariance surface over the entire interval I can be obtained by applying the *roughness penalty approach* (Ramsay *et al.*, 2009) to the grid points. Let $[D^2X_i(t)]^2$ measure the *curvature* in X_i at time t. Then, the the roughness penalty is defined as

$$\text{PEN}_2(X_i) := \int_I \left[D^2X_i(t)\right]^2 dt,$$

which measures the total curvature on interval I. The estimated mean and covariance functions above are obtained by minimizing the penalized residual sum of squares for some smoothing parameter λ, which can be determined by a standard cross-validation procedure. In other words, they are *penalized least square estimates*. Additionally, we get the estimate of the variance of the measurement error at time t_j as follows,

$$\hat{\sigma}^2(t_j) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_{ij} - \hat{\mu}(t_j)\right]^2 - \hat{\Sigma}(t_j, t_j).$$

For more general functional data that are not sampled from unified time schedule, one may consider nonparametric smoothers (Fan and Gijbels, 1996) for the estimates of the mean function and the covariance surface (Yao *et al.*, 2005). For instance, an alternative method is to utilize weight assignment (Li and Hsing, 2010).

3.2. Functional principal component analysis (FPCA)

Functional principal component analysis (FPCA), an extension of principal component analysis (PCA), is one of the most popular tools for dimension reduction in functional data. The essence of FPCA decomposes the functional process into a linear combination of functional principal components, which maximize the total variation of the observed curves. More specifically, FPCA is an expansion of the functional realizations in a functional basis consisting of orthogonal eigenfunctions (denoted $\phi_k(t)$, for $k = 1, 2, \ldots,$) of the covariance structure of the underlying process $X(t)$. Let $\Sigma(s, t)$ denote the bivariate covariance function at time s and t. The eigenfunctions of $\Sigma(s, t)$ are the solutions of

$$\int_I \Sigma(s, t)\phi_k(t)\, dt = \nu_k\phi_k(s),$$

where ν_k is the k^{th} eigenvalue of the covariance matrix of $X(t)$. The k^{th} *functional principal components* (FPCs) (or called scores) for subject i is given by

$$\xi_{ik} = \int_I \phi_k(t) \left[X_i(t) - \mu(t)\right] \, dt.$$

In practice, the computation of the scores is based on the following expansion,

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t),$$

where the dimension reduction is facilitated by determining a finite number K such that the sum of the first K terms provides a good approximation to the infinite sum. The selection of K is usually done through a cross-validation approach to minimize the following criterion function defined in (Yao *et al.*, 2005),

$$\text{CV}\ (K) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[X_i(t_j) - \hat{X}_i(t_j)\right]^2,$$

where

$$\hat{X}_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t).$$

There are a few other methods for choosing an appropriate K for FPCA (a number based on a desired amount of variance explained by FPCs; Burns *et al.*, 2013) and a Bayesian information criterion (Li *et al.*, 2013), but it is not our major concern in the present study.

The primary application of FPCA is to study the modes of variation of the functional curves deviated from the mean curve. In R, there are a few of well-developed packages for conducting FPCA, such as **fda** (Ramsay *et al.*, 2020) and **fdapace** (Carroll *et al.*, 2020). In the present study, we apply the functions built in **fda** to the analysis of the COVID-19 data from mainland China.

3.3. Functional canonical correlation analysis (FCCA)

In FDA, *functional canonical correlation analysis* (FCCA) is an approach to investigating how two functional data share variation, and to determining the amount of variations shared by them.

Let $X_i(t)$ and $X_i^*(t)$ be functional realizations of a couple of L_2 processes X and X^* on I_X and I_{X^*}, respectively. Consider the *Hilbert spaces* of square-integrable functions on I_X and I_{X^*} with respect to Lebesgue measure (denoted by $\mathcal{H}(I_X)$ and $\mathcal{H}(I_{X^*})$, respectively). Let $\langle \cdot, \cdot \rangle$ denote the standard operator of *inner product* such that $\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)\, dt$ for any $f, g \in \mathcal{H}(\mathcal{T})$ on some interval $\mathcal{T}$. For $k \geq 1$, the k^{th} squared *canonical correlation coefficient*, denoted by ρ_k^2, is given by

$$\rho_k^2 = \sup_{u \in \mathcal{H}(I_X),\ u^* \in \mathcal{H}(I_{X^*})} \left\{\mathbf{C}\text{ov}\left(\langle u,\ X\rangle,\ \langle u^*,\ X^*\rangle\right)\right\}^2, \tag{3.1}$$

subject to

$$\mathbf{V}\text{ar}\left(\langle u,\ X\rangle\right) = 1 \quad \text{and} \quad \mathbf{V}\text{ar}\left(\langle u^*,\ X^*\rangle\right) = 1.$$

For any given $k \geq 1$, suppose that (u_k, u_k^*) is a two-tuple solution to equation (3.1), we call (u_k, u_k^*) the pair of *probe weight functions* of ρ_k^2. Associated inner products, $\langle u, X\rangle$ and $\langle u^*, X^*\rangle$, are called *probe scores*. The calculation of (u_k, u_k^*) is usually done in an iterative manner, subject to an assumption of *orthogonality* given as follows,

$$\mathbf{C}\mathrm{ov}\left(\langle u_k,\ X\rangle, \langle u_j,\ X\rangle\right) = \mathbf{C}\mathrm{ov}\left(\langle u_k^*,\ X^*\rangle,\ \langle u_j^*, X^*\rangle\right) = \mathbf{C}\mathrm{ov}\left(\langle u_k,\ X\rangle, \langle u_j^*,\ X^*\rangle\right) = 0$$

for all $j = 1, 2, \ldots, k-1$.

Methods for conducting FCCA include, for example, local polynomial smoothing, smoothing splines, Fourier base expansions and eigenanalysis, most of which are summarized in (He *et al.*, 2004). Specifically, in Yang *et al.* (2011), the authors proposed a functional version of singular value decomposition for estimating functional canonical components, building a bridge between FPCA and FCCA. Noting that the functional data in the present study are dense and regularly recorded, we adopt the standard method that is based on Fourier expansion to determine probe weight functions and to compute probe scores. The analysis is done through a series of functions from **fda** package.

3.4. Functional cluster analysis

The goal of cluster analysis is to identify groups, where the data points in the same group are supposed to share some sort of similarity. Generally speaking, there are two classes of clustering methods: model-based clustering (Borveyron *et al.*, 2019; Handcock *et al.*, 2007) and clique-based clustering (Newman, 2006; Ouyang *et al.*, 2019). Analogously, *functional cluster analysis* (FCA) is an unsupervised learning process of functional data.

Classical functional clustering methods are divided into two classes: regularization methods and filtering methods (James and Sugar, 2003). In the same source, the authors proposed a model-based approach to clustering sparsely and irregularly sampled functional curves. In an independent work (Abraham *et al.*, 2003), the authors approximated functional data via B-splines and the clustering procedure was done via a k-means algorithm. In (Garcia-Escudero and Gordaliza, 2005), a trimmed k-means algorithm was introduced, and the proposed algorithm was shown to be more robust than standard k-means algorithms.

In the present study, we adopt an approach based on the *Expectation-Maximization* (EM) algorithm (Dempster *et al.*, 1977). The EMCluster algorithm (Chen and Maitra, 2019) assumes a finite mixture Gaussian distributions. The model is given by

$$f(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.2}$$

where, for each $k = 1, 2, \ldots, K$, $\pi_k \in (0, 1)$ is the mixing proportion such that $\sum_{k=1}^{K} \pi_k = 1$, and $\phi(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}_k$. The *log-likelihood* of the model in equation (3.2) is optimized by utilizing the EM algorithm. The setups of the E-step and the M-step are standard (Lee and Scott, 2012; McLachlan and Peel, 2000). The algorithm requires a predetermination of the number of clusters K prior to clustering. Conditional on K, the partitioning of functional data into clusters is done by solving the following optimization problem

$$\arg\max_k \hat{\pi}_k \phi\left(\boldsymbol{x}_i \,|\, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k\right),$$

for each $i = 1, 2, \ldots, n$.

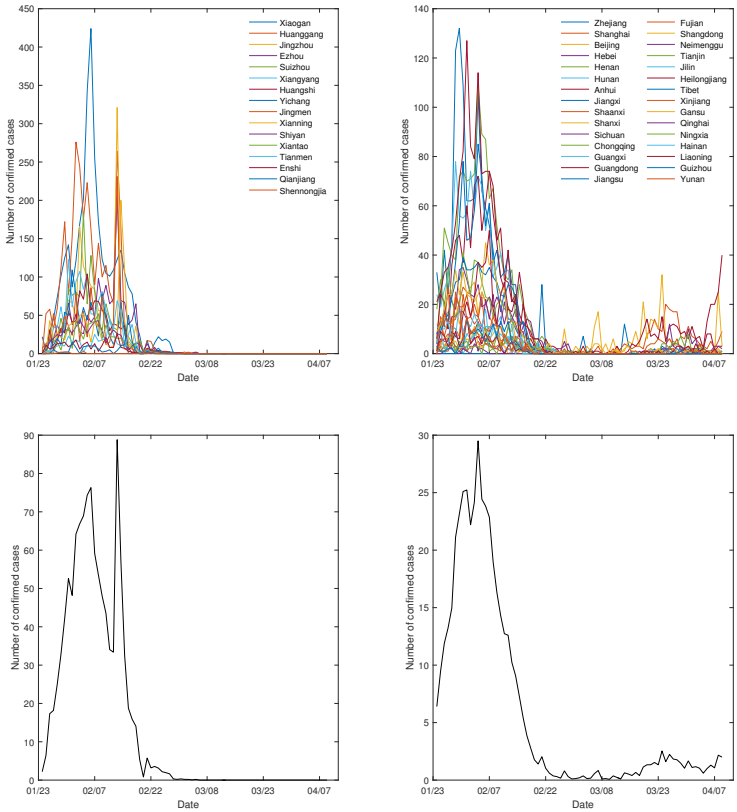

Figure 1: *Top left panel: functional data of daily confirmed cases of 16 cities in Hubei province (excluding Wuhan); top right panel: functional data of daily confirmed cases of 30 provinces (excluding Hubei province); bottom left panel: fitted mean curve of daily confirmed cases of 16 cities in Hubei province; bottom right panel: fitted mean curve of daily confirmed cases of 30 provinces.*

In the application of FCA, we utilize the EM-based algorithm to detect the community structure in the functional data of 16 cities in Hubei province (excluding Wuhan) and the counterpart of 30 provinces, municipalities and autonomous regions (excluding Hubei province) in mainland China, respectively. Specifically, the implementation of the algorithm is done through the available functions in **fdapace** package. The results are presented in Section 4.4.

4. Results

4.1. Graphic visualization

In this section, we present the analysis results by applying the FDA methods to the collected COVID-19 data in China. In general, it is necessary to account for weekend or holiday effects to the collected time series data. However, since the local public health administrations and institutions were required to update all COVID-19 information on a daily base to the central government, we do not consider weekend and holiday effects in the present study. Additionally, notice that the city of Wuhan was

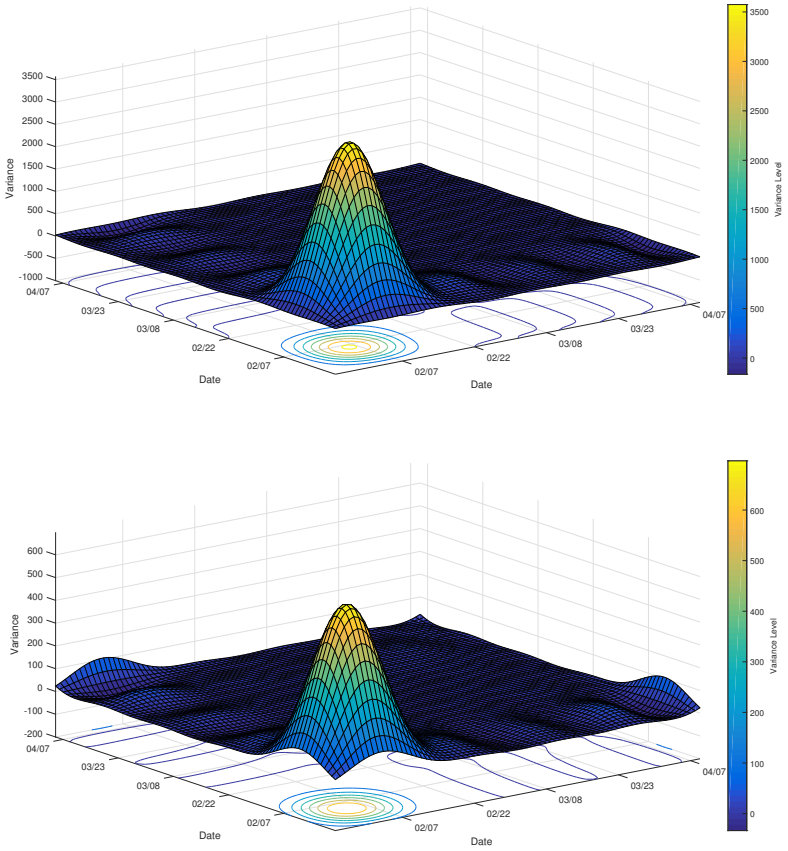

Figure 2: *Top panel: the variance-covariance surface for the number of newly confirmed cases in 16 cities of Hubei province (excluding Wuhan); bottom panel: the variance-covariance surface for the number of newly confirmed cases in 30 provinces in China (excluding Hubei province).*

the center of the outbreak of the pandemic, and that the numbers of confirmed and death cases in Hubei province appeared to be significantly higher than the counterparts of other provinces. Thus, our data analysis is divided into two groups: the analysis based on the data of 16 cities in Hubei province (excluding Wuhan), and that of 30 provinces, municipalities and autonomous regions (excluding Hubei province), respectively. For simplicity, we use "province" to represent all the provincial regions (including municipalities like Shanghai and autonomous regions like Tibet) in the remainder of the manuscript. The majority of the data analysis is done in R, whereas some of the plots are generated via MATLAB for a better visualization. All the codes associated to the present study are publicly available on a GitHub repository https://github.com/panpanzhang99299/fdacovidchina).

To begin with, we provide a graphical visualization of the functional data as well as some basic statistics. In Figure 1, we plot the number of daily confirmed cases in 16 cities of Hubei province and the counterparts in 30 provinces in China, respectively. Their fitted mean curves are also given correspondingly.

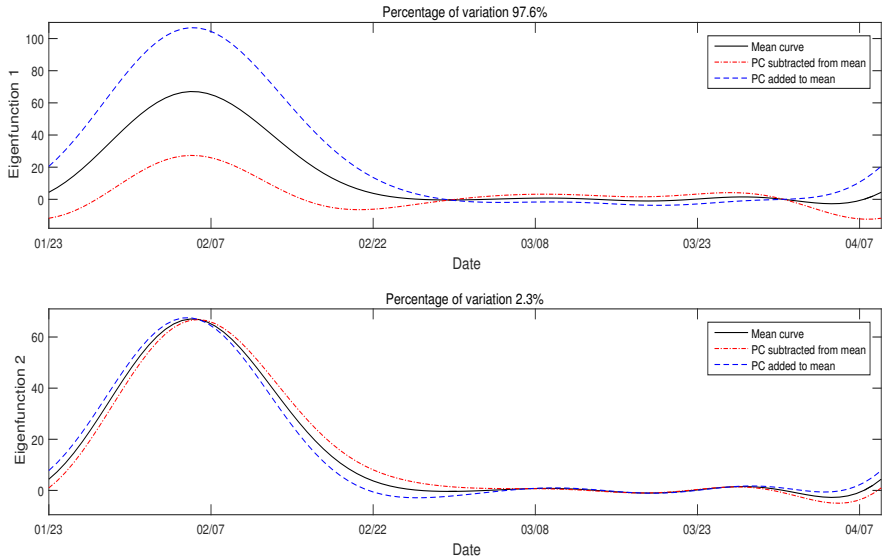

Figure 3: *Top panel: the first eigenfunction for the newly confirmed cases in the 16 cities of Hubei province with perturbations. The black curve represents the mean, whereas the blue and red curves respectively show what happens when a small amount is added to and subtracted from the mean. Bottom panel: the second eigenfunction for the newly confirmed cases in the 16 cities of Hubei province with perturbations.*

From the top left panel of Figure 1, we observe that the number of daily confirmed cases in most of the cities of Hubei province reached the peak around Feb 7, 2020, and decreased significantly in the following two weeks until arrived at a low level close to 0 around Feb 20, 2020. In some cities, a secondary peak emerged on Feb 12, 2020, owing to the change of the criteria of confirmed cases in Hubei province by Chinese officials. In the top right panel, on the other hand, we observe that the maximum number of newly confirmed cases for most provinces appeared around Feb 3, 2020. Since Feb 14, 2020, the number of newly confirmed cases (across all 30 provinces) was small. The bottom right panel shows that, in average, the spread of COVID-19 had been well controlled since early March, because of the practice of effective quarantine and isolation measures (Tang *et al.*, 2020). At the end of the current study period, we observe a small upward trend (in two right panels), which is likely to be related to the confirmed cases imported from overseas.

The smoothed variance-covariance surface of the number of newly confirmed cases of 16 cities in Hubei province (excluding Wuhan) and that of 30 provinces (excluding Hubei province) are respectively depicted in the two panels in Figure 2. The largest variance in the top panel of Figure 2 appeared around Feb 5, 2020, since a great number of newly confirmed cases were diagnosed in some cities (like Huanggang and Xiaogan), while no newly confirmed case was reported in some of the others (like Shennongjia). By scrutinizing the data, we observe that the number of newly confirmed cases arrived at the maximum value in several large-scale provinces, such as Hunan and Guangdong provinces, whereas the number of some other provinces (like Tibet) stayed as low as 0 throughout the entire study period.

For death cases, we provide an analogous graphic visualization of the data and some relevant statistics as well. For better readability, they are presented in Appendix A.1. Besides, it may be of

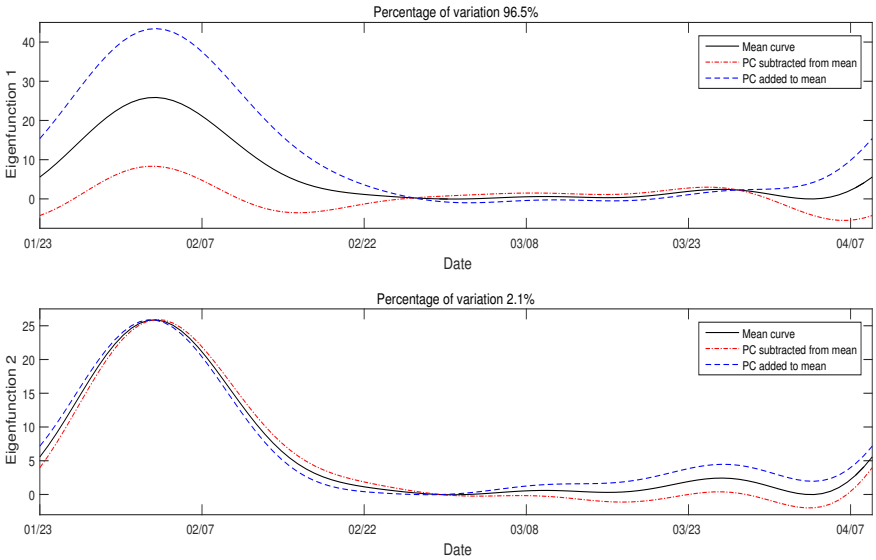

Figure 4: *Top panel: the first principal component harmonics for the newly confirmed cases of 30 provinces with perturbations; bottom panel: the second principal component harmonics for the newly confirmed cases of 30 provinces with perturbations.*

interest to look into the cross-covariance surfaces (Yang *et al.*, 2011) of confirmed and death cases in 16 cities in Hubei province as well as the counterparts of 30 provinces. These results and some related discussions are given in Appendix A.2.

4.2. Functional principal component analysis of COVID-19 data in China

In this section, we utilize FPCA to study the modes of variations in the data of daily confirmed cases (of 16 cities in Hubei province and of 30 provinces in China) over the study period. The computation results are respectively shown in Figures 3 and 4.

A total of nine orthogonal eigenfunctions are selected to constitute the functional basis, where the top two eigenfuncitons together explain 99.9% of the variations around the fitted mean curve (Figure 3), suggesting that the rest are less important. The first FPC accounts for 97.6% of the variations. As it shows, there was a large amount of shifts from the mean curve in January and February (the early times of the study period). This might be due to the overburdening of the medical system, the lack of experience in providing effective treatments and long incubation period of COVID-19 (Zhu *et al.*, 2020). A break point appeared around Feb 7, 2020, and there was a trend of decreasing since then, corresponding to the fact that a large number of national medical teams were deployed to Hubei province around that time. With the implementation of strict quarantine policies and the alleviation of the pressure to the local medical systems, the epidemic was well controlled in the 16 cities in Hubei province since March, until a (potential) second (small) wave emerged after the lockdown of Wuhan ended. The second FPC (accounting for 2.3% of the variations) shows a similar trend as the first FPC. There is a modestly negative impact (of the second FPC) on the mean function at the end of February.

The FPCA results for the confirmed cases in 30 provinces are shown in Figure 4. The top two eigenfunctions together explain 98.6% of the variations. The first eigenfunction is similar to the

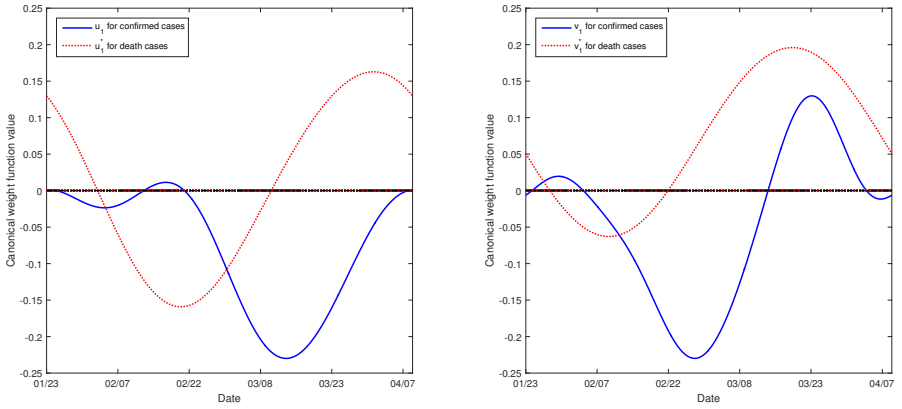

Figure 5: *Left panel: the first pair of canonical functions* (u_1, u_1^*) *correlating daily confirmed and death cases of 16 cities in Hubei province; right panel: first pair of canonical functions or probes* (v_1, v_1^*) *correlating the confirmed and death cases of 30 provinces.*

counterpart of Hubei province in trend. A relatively large variation observed at the tail (late March to early April) was probably owing to an increase of the number of confirmed cases imported from overseas. For the second eigenfunction, we observe an emergence of larger variations (compared to the bottom panel of Figure 3) since early March, indicating relatively larger positive impact from the scales of different provinces.

4.3. Functional canonical correlation analysis of COVID-19 data in China

Next, we investigate the functional canonical correlations between two functional data; namely, the amount of variations shared by the number of newly confirmed and death cases of 16 cities in Hubei province and 30 provinces in China. As stated in Section 3.3, the fundamental theory of FCCA is related to the derivations of FPCs. The FPCs for the functional data of newly confirmed cases have been computed in Section 4.2. What left is an analogous computation for newly death cases. We present the related results in Appendix A.3.

Fourier basis are used as basis functions for running FCCA, as the data present periodic patterns and a fast Fourier transform finds all the coefficients efficiently (Ramsay and Silverman, 2005). For the data of 16 cities in Hubei province, we find that the estimate of the first squared canonical correlation is $\hat{\rho}_1^2 = 0.9891$. The associated pair of probe weight functions u_1 and u_1^*, respectively for daily confirmed and death cases, are displayed in the left panel of Figure 5. We observe that function u_1 is slightly negative in January until the end of the second week of February, and then becomes slightly positive in the third week of February; After that, its value remains negative in the rest of the study period. Hence, low probe scores are assigned to the cities with high number of confirmed cases in January, the first two weeks of February and March. On the other hand, function u_1^* is contrasting the period between Feb 3, 2020 and Mar 10, 2020 against the remaining days in the study period. The cities with more number of death cases in February are supposed to have low scores in this canonical variable.

For the data of 30 provinces, the estimate of the first squared canonical correlation is 0.7951. The first pair of probe weight functions (v_1, v_1^*) is shown in the right panel of Figure 5. Function v_1 is quite different from u_1 in presenting a small positive hump in January and the first week in February, but

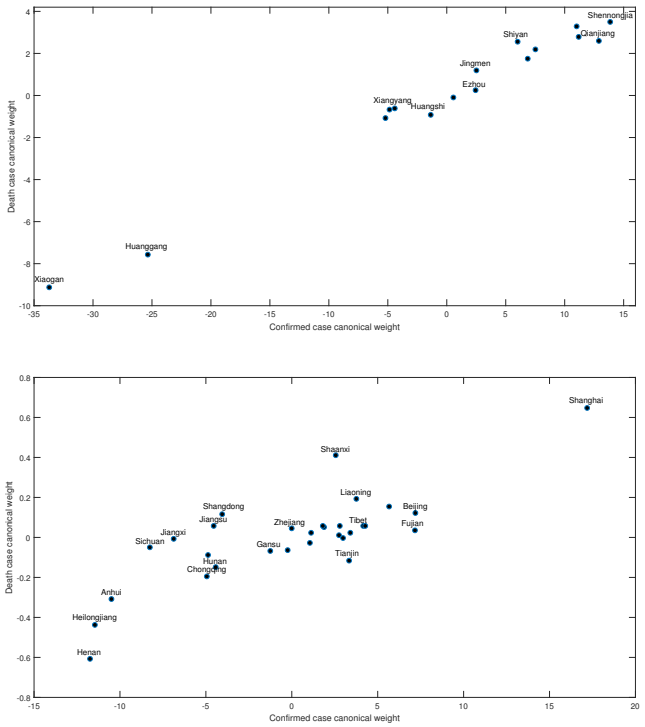

Figure 6: *Top panel: the scores of the first pair of canonical variables against each other of 16 cities in Hubei province; bottom panel: the scores of the first pair of canonical variables against each other outside Hubei province.*

staying negative in the rest of February until the end of the second week of March. After that, the function becomes positive overall. In contrast, function v_1^* is negative during the majority of February, but becomes positive since March. Based on the shape of v_1 and v_1^*, a province gets high score on both of canonical variables if it has a large number of confirmed cases at the beginning of the outbreak and late March, but a low number of death cases in the first three weeks of February.

Next, we present two scatter plots showing the probe scores of death cases against confirmed cases of 16 cities in Hubei province and 30 provinces in China, respectively in the top and bottom panels in Figure 6.

In the top panel of Figure 6, we observe that Xiaogan and Huanggang are the two cities with lowest scores in both of the canonical variables. These two cities are both adjacent to Wuhan, and respectively ranked 5 (for Xiaogan) and 2 (for Huanggang) in population according to the National Population Census of China in 2015. The two cities consistently have higher number in both confirmed and death cases than the others in Hubei province. In contrast, Shennongjia is reported to have the smallest number in both confirmed and death cases consistently. The majority of this county is a forest district with low population density, and it is also geographically far away from Wuhan. Therefore, it has positive scores in both canonical variables, and appears in the top right corner.

The bottom panel of Figure 6 shows that Henan is the province that emerges at the most bottom left corner. The province is adjacent to Hubei province, and it has the largest number of both confirmed

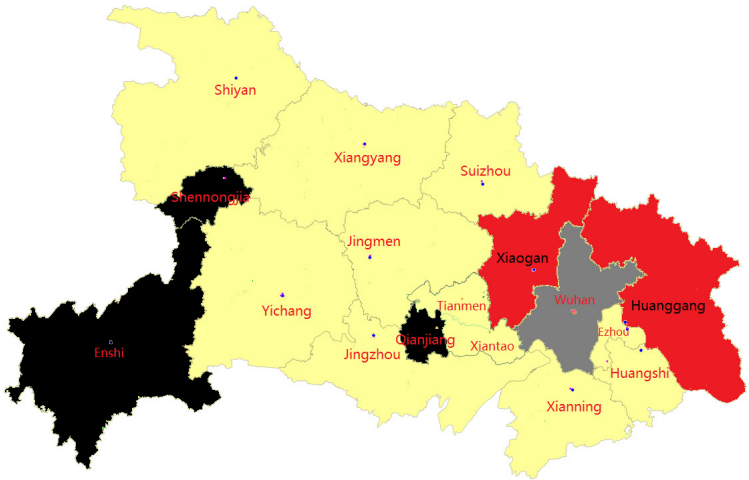

Figure 7: *Clustering result for 16 cities in Hubei province.*

and death cases (out of the 30 provinces) in the entire study period. The provinces that are adjacent to Hubei province, such as Henan and Anhui, have relatively more confirmed (and death) cases than those further away. They are close to the bottom left corner as well. One of the "abnormal" province is Heilongjiang, which does not share provincial border line with Hubei; however, the province receives negative scores in both canonical variables as well. We speculate the possible reasons as follows. There are a lot of popular winter vacation resorts in Heilongjiang, consequently resulting in massive movements in population to the province. Besides, Heilongjiang is under the risk of cases imported from overseas (especially north) after March. All these factors potentially cause the low scores in both of the canonical variables. On the contrary, Shanghai gets positive scores in both canonical variables, because there were few number of both confirmed and death cases in the first three weeks of February. In addition, in late March, a total of 145 confirmed cases (more than all the other provinces outside Hubei) and 4 death cases were reported in Shanghai, as it is one of the largest ports of entry in China, and all of these confirmed cases were from overseas.

4.4. Functional cluster analysis of COVID-data in China

In this section, we conduct cluster analysis of 16 cities in Hubei province and 30 provinces in China based on the data of daily confirmed cases. Besides, we collected the data of population migration from Wuhan to other domestic destinations between Jan 10 and Jan 24, 2020. The cluster analysis is conducted via the *EMCluster* algorithm (Chen and Maitra, 2019), which requires the knowledge of the number of clusters, denoted K. The determination of the value of K is done through an elbow method. More precisely, we compute the sum of within-cluster variations for the optimal clustering strategies corresponding to a variety of values of K, and select the most appropriate "turning point".

According to Figure B.1 in Appendix B. The numbers of clusters of the data of 16 cities in Hubei province and of 30 provinces in China are three and four, respectively.

The clustering results are respectively presented in Figures 7 and Figure 8. As stated, there are three clusters in Figure 7, where the first cluster contains Xiaogan and Huanggang. These two cities are adjacent to Wuhan and have relatively larger population densities. Besides, these two cities were indeed the most severely attacked cities (except for Wuhan) by COVID-19 in Hubei province. The

Figure 8: *Clustering result for 30 provinces in China.*

second cluster contains Enshi, Shennongjia and Qianjiang, where the former two are geographically deviated from the center of outbreak, Wuhan. Although Qianjiang is close to Wuhan, it has the second lowest population size in the province, and a small population density compared to most of the other cities. Besides, the economic development of Qianjiang is primarily based on agriculture, and the within- and across-province transportation of Qianjiang is not as convenient as many other cities in Hubei province. The remaining 11 cities in Hubei province form the last cluster. Generally speaking, these cities were modestly attacked during the outbreak of the pandemic.

Four clusters are colored differently for the data of 30 provinces, presented in Figure 8. The first cluster contains three provinces, which are Henan, Guangdong, Zhejiang, respectively. Henan is adjacent to Hubei province, and there used to be a relatively large population movements between the cities therein and Wuhan (before the lockdown started). Although Zhejiang and Guangdong are not adjacent to Hubei, there exist immense commercial tradings among these provinces, potentially causing a large amount of travels. These three provinces undergo the greatest losses in economy during the outbreak of the pandemic. There are three provinces (Anhui, Hunan and Jiangxi) in the second cluster, and all of them are adjacent to Hubei province from south and east. Besides, these three provinces have relatively high population sizes, and send a lot of labor forces to other provinces, including Hubei. Hence, these provinces were attacked not the most but quite severely by the pandemic during its outbreak. The third cluster (colored blue) contains a total of 11 provinces. They are relatively closer to Hubei province than those in the fourth cluster, except for Heilongjiang, which is located on the border of northeast of China (to Russia). As stated, one of the largest industry in Heilongjiang is tourism (especially well known for winter programs), attracting the tourists across the country prior to the official announcement of travel ban. Additionally, owing to the outbreak of COVID-19 in Russia in late March, confirmed cases imported from north have become one of the main sources in Heilongjiang, indirectly causing the lockdown of Suifenhe (a city on the inner border

of Heilongjiang) in April. The remaining 13 provinces are put into the fourth cluster. These provinces are lightly affected by COVID-19, compared to the other provinces.

In addition, we collected the data of population movements from Wuhan to other domestic destinations between Jan 10 and Jan 24, 2020 from a public online source, summarized in Table C.1 in Appendix C. The data are given in form of percentages. The table suggests that the distribution of confirmed cases is seemingly related to the population outflow from Wuhan, in support of our clustering results. The two destinations with largest percentages of population movements (from Wuhan) are Xiaogan and Huanggang, forming the first cluster in Figure 7. At the province level, we observe that the percentages of population outflow from Wuhan to Hunan, Anhui, Jiangxi, Guangdong and Henan all exceed 1%. These five provinces altogether form the first two clusters in Figure 8. Therefore, the collected migration data is in favor of our cluster analysis results, and implies that the spread of COVID-19 is likely to be related with the geographical locations of the infected regions.

5. Discussions

In this article, we conduct a series of functional data analysis on the time series data of COVID-19 in mainland China. Specifically, we apply the well-developed methods from functional principal component analysis, functional canonical correlation analysis and functional cluster analysis to the collected data.

In FPCA, we explore the functional data by looking into the modes of variations. Based on the results, we find that the practice of effective public health measures in mainland China, such as the lockdown of the center of the epidemic and strict quarantine protocols, help slow down the spread of the highly contagious disease. Besides, a lot of credits should be given to the support of national and regional medical teams from the entire country and the rapid construction of (temporary) modular hospitals, greatly relieving the pressure to the medical system, especially to hospitalization.

In FCCA, we observe high canonical correlation between confirmed and death cases inside and outside of Hubei province. Although it was reported in (Rajgor *et al.*, 2020) that the *case-to-fatality rate* (CFR) of COVID-19 is not as high as that of severe acute respiratory syndrome (SARS) that broke out in 2003. The outbreak of COVID-19 still brings a great deal of challenges to the global public health system owing to its high and rapid contagiousness as well as high canonical correlations between infectious cases and deaths.

In FCA, we find that the severity of the attack by COVID-19 is closely related to geographical locations of different regions. The cities and provinces that are closer to the center of the outbreak are more likely to be seriously attacked. Another factor that can not be underestimated is the amount of population movement. Due to the high infectiousness and the various transmission channels of COVID-19, any living stock (mainly human being) exposed to the virus would be a potential carrier.

Lastly, we address some limitations of the present study propose associated future work. Note that the data from Feb 4 to Feb 12, 2020 in Hubei province is abnormally high because of the change of the criteria of determining confirmed cases. Without the access to the accurate data, a sensible method should be considered to reasonably distribute the data over that period. The FPCA and FCCA in the present analysis are primarily based on the expansion of Fourier basis functions. For the functional data not presenting a pattern of periodicity, other basis functions, such as B-splines, can be considered. The present cluster analysis is based on an EM algorithm. EM algorithms usually undergo the limitation of slow convergence for high dimensional data, and they are extremely sensitive to initial values. The algorithm should be updated when more factors are accounted in the analysis. A possible alternative may be non-negative matrix factorization (NMF) (Chen *et al.*, 2020).

Appendix A: Parallel analysis results for death cases

In the appendix, we provide some supplementary information in support of the analysis in the main body of the manuscript.

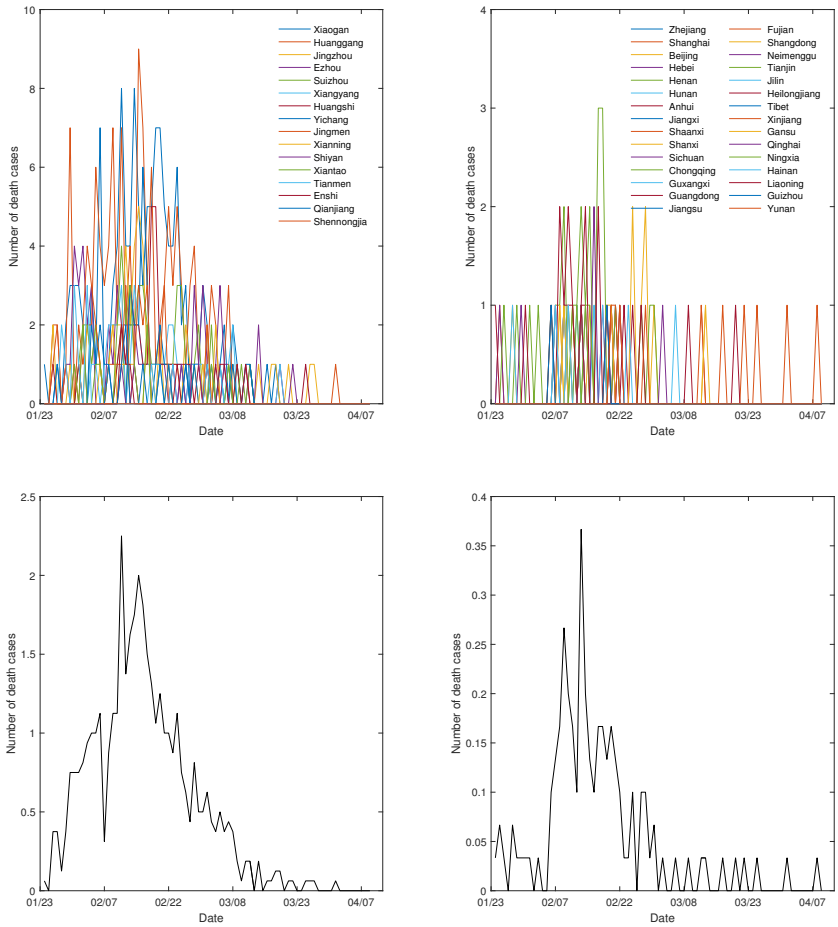

Figure A.1: *Top left panel: functional data of newly death cases in 16 cities of Hubei province (excluding Wuhan); top right panel: functional data of newly death cases in 30 provinces (excluding Hubei); bottom left panel: fitted mean curve of newly death cases in 16 cities of Hubei province (excluding Wuhan); bottom right panel: fitted mean curve of newly death cases in 30 provinces (excluding Hubei).*

Graphic visualization of death cases in China

The functional curves of death cases of 16 cities in Hubei province and of 30 provinces in mainland China are respectively displayed in the top two panels in Figure A.1; the associated fitted mean curves are plotted in the bottom panels. Compared to the counterparts of confirmed cases, we observe more fluctuations in the curves. Both of the functional data reached the peak in the second week of February,

but in a small amount (especially for the provinces outside Hubei). Since March, the number of death cases became close to 0, suggesting that the pandemic had been well controlled in China.

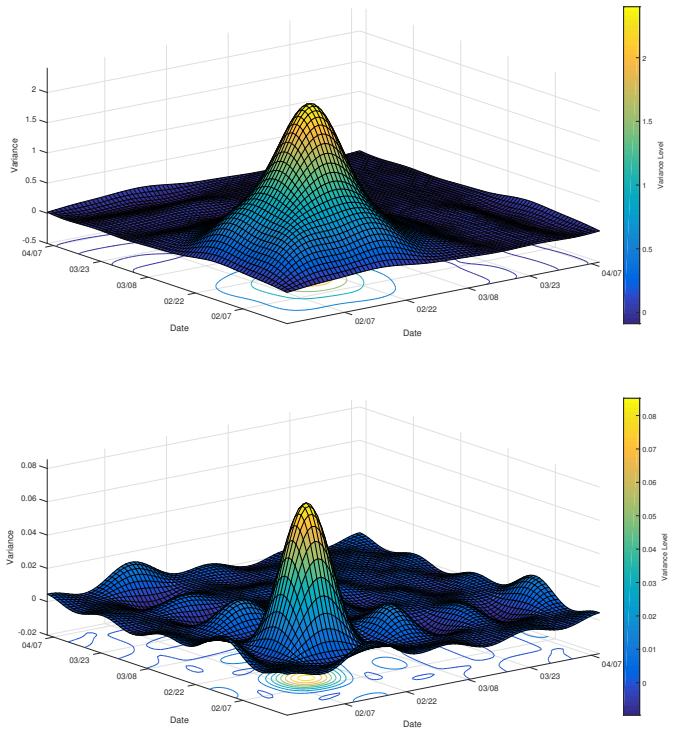

Figure A.2: *Top panel: the variance-covariance surface for the newly death cases in 16 cities of Hubei province (excluding Wuhan); bottom panel: the variance-covariance surface for the newly death cases in 30 provinces.*

From the variance surfaces in Figure A.2, we find that the variations are small in general, since the numbers of reported death cases are relatively low in both functional data (especially for 30 provinces). More "waves" are observed in the data of provinces outside Hubei because no death case was reported in quite a few provinces while some emerged in the rest.

Cross-covariance surfaces

It is of interest to see how confirmed and death cases related in 16 cities in Hubei province as well as in 30 provinces outside Hubei. As suggested in (Yang *et al.*, 2011), we explore the relation of confirmed and death cases through the cross-covariance structure that underlies the functional correlation measure. The cross-covariance surfaces of the two functional data are plotted respectively in the top and bottom panels in Figure A.3. In the top panel, the largest covariance is observed in the pair of Feb 5, 2020 (for confirmed cases) and Feb 16, 2020 (for death cases), respectively coincide with their largest variance dates. The covariance surface stays positive throughout January and February, but remains close to 0 in the rest of the study period.

In the bottom panel, on the other hand, we find that the peak of the surface emerges at the pair of

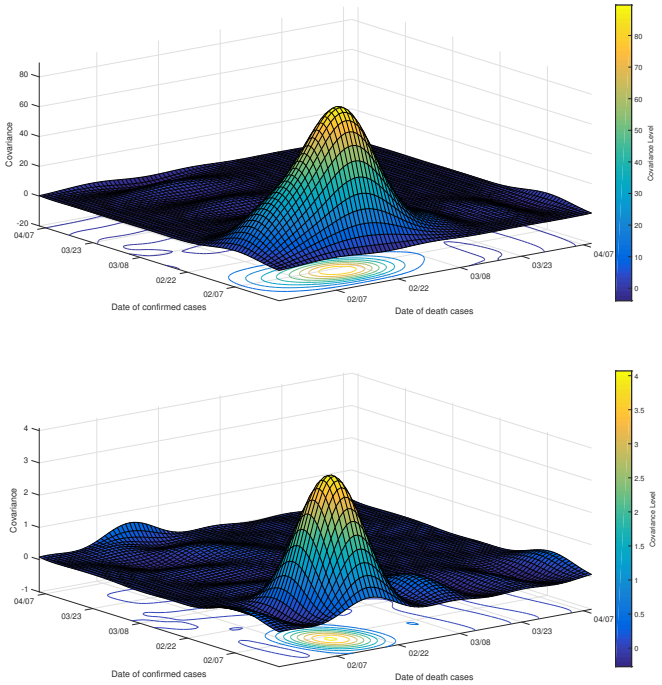

Figure A.3: *Top panel: cross-covariance surface for the newly confirmed cases and the newly death cases of 16 cities in Hubei province (excluding Wuhan); bottom panel: the cross-covariance surface for the newly confirmed cases and the newly death cases in 30 provinces in China (excluding Hubei province).*

Feb 5, 2020 for confirmed cases and Feb 13, 2020 for death cases. The overall pattern of the surface is similar to that in the above panel. However, we would like to point out that the covariance of the vertical axis is indeed small. Except for relatively large numbers of confirmed cases in several provinces in February, few confirmed cases are reported in other time periods. In addition, the numbers of death cases stay low across the provinces throughout the study period.

FPCA based on death cases

In this subsection, we conduct an analogous FPCA based on the collected data of death cases in 16 cities in Hubei province and 30 provinces in mainland China. The results are respectively shown in Figures A.4 and A.5.

For the death cases in the 16 cities in Hubei province, the top two eigenfunctions explain 98.4% of the total variation. In the top panel of Figure A.4, relatively large amount of variations were observed since the start of the study period until mid March, where the largest appeared around Feb 16, 2020. The second eigenfunction that explains 3.6% of the total variation presents an analogous pattern as the counterpart of confirmed cases in Hubei province.

For the death case data of 30 provinces, the top two eigenfunctions explain 93.6% of the total variation. The first eigenfunction is similar to the counterpart of death cases in Hubei province (the top panel in Figure A.4), but the magnitude of the variation is small in general. The second eigenfunction accounts for 7.6% of the total variation. We observe a positive impact on the mean function in January

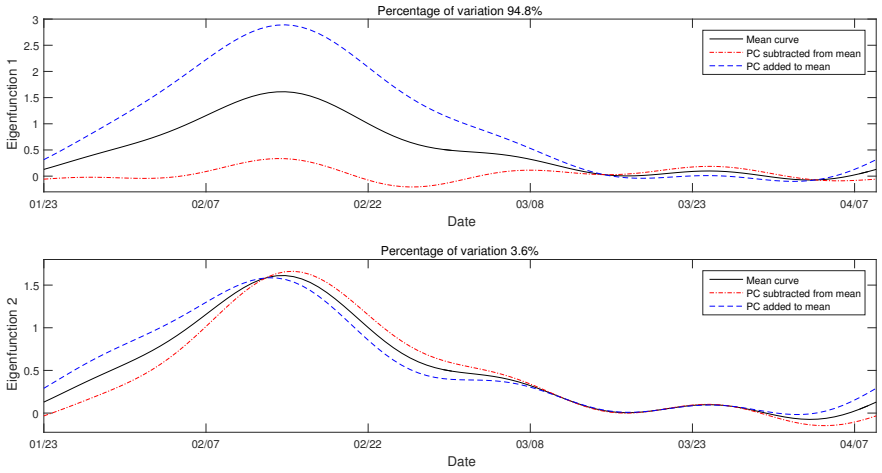

Figure A.4: *Top panel: The first eigenfunction for the newly death cases in 16 cities of Hubei province with perturbations; bottom panel: The second eigenfunction for the newly death cases in 16 cities of Hubei province with perturbations.*

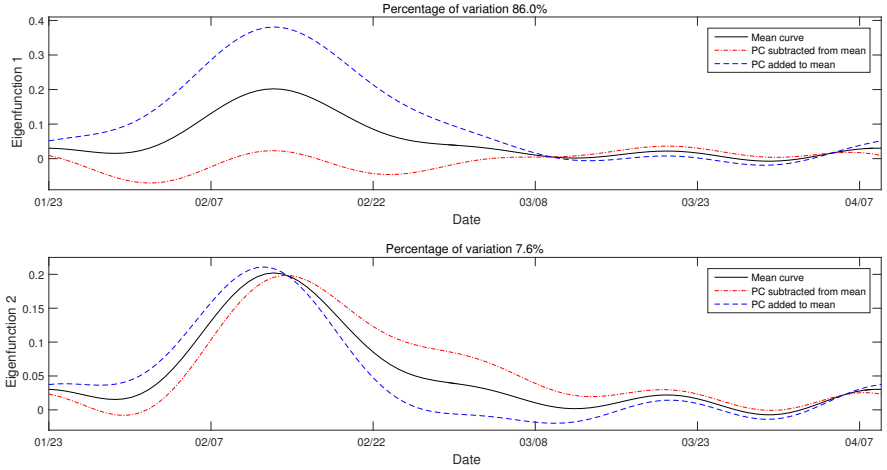

Figure A.5: *The first eigenfunction for the newly death cases in 30 provinces in China with perturbations; the black line represents the mean, where the blue line and red line respectively show what happens when a small amount of a principal component is added to and subtracted from the mean. Bottom panel: The second eigenfunction for the newly death cases in 30 provinces in China with perturbations.*

and the first half of February, but a negative impact since then until the end of March.

Appendix B: Choosing the number of clusters

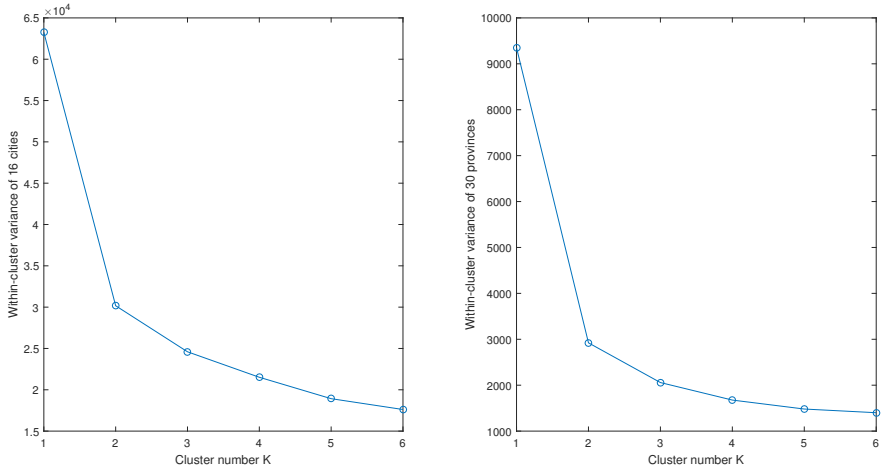

Figure B.1: *Left panel: the within-cluster variation based on the newly confirmed cases of 16 cities in Hubei province (excluding Wuhan) for given values of K. Right panel: the within-cluster variation based on the newly confirmed cases in 30 provinces in China (excluding Hubei province) for given values of K.*

In Figure B.1, we plot the within-cluster variations for the optimal clustering strategies given a set of values of $K = 1, 2, \ldots, 6$. The cluster analysis is done based on the number of daily confirmed cases of the two functional data.

Appendix C: Population migration from Wuhan to other domestic destinations

In Table C.1, we show the migration data departed from Wuhan from Jan 10 to Jan 24, 2020; top 10 cities in Hubei province and top 10 provinces outside Hubei are presented.

Table C.1: The percentages of top 10 population movements from Wuhan to other domestic destinations between Jan 10 and Jan 24, 2020.

	Cities in Hubei		Provinces in China	
Rank	City	Percentage	Province	Percentage
1	Xiaogan	13.80%	Henan	5.68%
2	Huanggang	13.04%	Hunan	3.48%
3	Jingzhou	6.54%	Anhui	2.27%
4	Xiangyang	6.33%	Jiangxi	2.12%
5	Huangshi	6.10%	Guangdong	1.94%
6	Xianning	5.01%	Jiangsu	1.46%
7	Xiantao	4.33%	Chongqing	1.27%
8	Ezhou	3.97%	Sichuan	1.24%
9	Jingmen	3.30%	Shangdong	1.09%
10	Suizhou	3.21%	Zhejiang	1.06%

Acknowledgement

We thank two anonymous reviewers for their thoughtful comments and suggestions that have improved the overall quality of the manuscript.

References

Abraham C, Cornillon PA, Matzner-Løber E, and Molinari N (2003). Unsupervised curve clustering using B-splines, *Scandinavian Journal of Statistics*, **30**, 581–595.

Anderson KG, Ranbaut A, Lipkin WI, Holmes EC and Garry RF (2020). The proximal origin of SARS-CoV-2, *Nature Medicine*, **26**, 450–452.

Boschi T, Di Iorio J, Testa L, Cremona MA, and Chiaromonte F (2021). Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy, *Scientific Reports*, **11**, 17054.

Borveyron C, Celeus G, Murphy TB, and Raftery AE (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*, Cambridge University Press, Cambridge, UK.

Baladandayuthapani V, Mallick BK, Hong MY, Lupton JR, Turner ND, and Carroll RJ (2008). Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogenesis, *Biometrics*, **64**, 64–73.

Burns DM, Houpt JW, Townsend JT, and Endres MJ (2013). Functional principal component analysis of workload capacity functions, *Behavior Research Methods*, **45**, 1048–1057.

Carroll C, Gajardo A, Chen Y, *et al.* (2020). fdapace: Functional Data Analysis and Empirical Dynamics, R package version 0.5.4, https://CRAN.R-project.org/package=fdapace

Carroll C, Bhattacharjee S, Chen Y, *et al.* (2020). Time dynamics of COVID-19, *Scientific Reports*, **10**, 21040.

Chen WC and Maitra R (2019). EMCluster: EM algorithm for model-based clustering of finite mixture Gaussian distribution, R Package, http://cran.r-project.org/package=EMCluster

Chen J, Yan J, and Zhang P (2020). Clustering US states by time series of COVID-19 new case counts with non-negative matrix factorization, arXiv:2011.14412.

Crawford L, Monod A, Chen AX, Mukherjee S, and Rabadán R (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis, *Journal of the American Statistical Association*, **115**, 1139–1150.

Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–22.

Fan J and Gijbels I (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London, UK.

Floriello D and Vitelli V (2017). Sparse clustering of functional data, *Journal of Multivariate Analysis*, **154**, 1–18.

Garcia-Escudero LA and Gordaliza A (2005). A proposal for robust curve clustering, *Journal of Classification*, **22**, 185–201.

Handcock MS, Raftery AE, and Tantrum J (2007). Model-based clustering for social networks (with discussion), *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **170**, 301–354.

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

Hyndman RJ and Ullah S (2007). Robust forecasting of mortality and fertility rates: A functional data

approach, *Computational Statistics & Data Analysis*, **51**, 4942–4956.

James GM and Sugar CA (2003). Clustering for sparsely sampled functional data, *Journal of the American Statistical Association*, **8**, 397–408.

Lee G and Scott C (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data, *Computational Statistics & Data Analysis*, **56**, 2816–2829.

Leng X and Müller H-G (2006). Classification using functional data analysis for temporal gene expression data, *Biostatistics*, **22**, 68–76.

Li Y and Hsing T (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data, *The Annals of Statistics*, **38**, 3321–3351.

Li Y, Wang N, and Carroll RJ (2013). Selecting the number of principal components in functional data, *Journal of the American Statistical Association*, **108**, 1284–1294.

McLachlan GJ and Peel D (2000). *Finite Mixture Models*,Wiley-Interscience, New York.

Newman MEJ (2006). Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8577–8582.

Ouyang G, Dey DK, and Zhang P (2019). Clique-based method for social network clustering, *Journal of Classification*, **37**, 254–274.

Petersen A and Müller H-G (2016). Functional data analysis for density functions by transformation to a Hilbert space, *The Annals of Statistics*, **44**, 183–218.

Rajgor DD, Lee MH, Archuleta S, Bagdasarian N, and Quek SC (2020). The many estimates of the COVID-19 case fatality rate, *The Lancet Infectious Diseases*, **20**, 776–777.

Rahman A and Jiang D (2021). Regional and temporal patterns of influenza: Application of functional data analysis, *Infectious Disease Modelling*, **6**, 1061–1072.

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

Ramsay JO and Dalzell CJ (1991). Some tools for functional data analysis, *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**, 539–572.

Ramsay JO and Silverman BW, *Applied Functional Data Analysis: Methods and Case Studies*, Springer-Verlag, New York.

Ramsay JO and Silverman BW (2005). *Functional Data Analysis*, Springer-Verlag, New York.

Ramsay JO, Hooker G, and Graves S (2009). *Functional Data Analysis with R and MATLAB*, Springer-Verlag New York, NY.

Ramsay JO, Graves S, and Hooker SG (2020). fda: Functional Data Analysis, R package version 5.1.4, https://CRAN.R-project.org/package=fda

Shen M, Tan H, Zhou S, Smith GN, Walker MC, and Wen SW (2017). Trajectory of blood pressure change during pregnancy and the role of pre-gravid blood pressure: A functional data analysis approach, *Scientific Reports*, **7**, 6227.

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

Tang B, Xia F, Tang S *et al.* (2020). The effectiveness of quarantine and isolation determine the trend of the COVID-19 epidemic in the final phase of the current outbreak in China, *International Journal of Infectious Diseases*, **95**, 288–293.

Tang C, Wang T, and Zhang P (2021). Functional data analysis: An application to COVID-19 data in the United States, *Quantitative Biology*.

The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team (2020). Vital Surveillances: The Epidemiological Characteristics of an Outbreak of 2019 novel coronavirus diseases (COVID-19)—China, China CDC Weekly 2, 113–122.

Ullah S and Finch CF (2013). Applications of functional data analysis: A systematic review, *BMC*

Medical Research Methodology, **13**.

World Health Organization (2020). *Coronavirus disease 2019 (COVID-19): situation report, 114*, https://apps.who.int/iris/handle/10665/332089

Yang W, Müller H-G, and Stadtmüller U (2011). Functional singular component analysis, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **73**, 303–324.

Yao F, Müller H-G, and Wang J-L (2005). Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association*, **100**, 577–590.

Zhu H, Wei L, and Niu P (2020). The novel coronavirus outbreak in Wuhan, China, *Global Health Research and Policy*, **5**,

Received July 16, 2021; Revised December 10, 2021; Accepted December 14, 2021