International Journal of
Environmental Research and Public Health

MDPI

Article

Associations between Google Search Trends for Symptoms and COVID-19 Confirmed and Death Cases in the United States

Mostafa Abbas [1], Thomas B. Morland [2], Eric S. Hall [1] and Yasser EL-Manzalawy [1,*]

[1] Department of Translational Data Science and Informatics, Geisinger, Danville, PA 17822, USA; mmhamza@geisinger.edu (M.A.); ehall3@geisinger.edu (E.S.H.)
[2] Department of General Internal Medicine, Geisinger, Danville, PA 17822, USA; tmorland@geisinger.edu
* Correspondence: yelmanzalawi@geisinger.edu

Abstract: We utilize functional data analysis techniques to investigate patterns of COVID-19 positivity and mortality in the US and their associations with Google search trends for COVID-19-related symptoms. Specifically, we represent state-level time series data for COVID-19 and Google search trends for symptoms as smoothed functional curves. Given these functional data, we explore the modes of variation in the data using functional principal component analysis (FPCA). We also apply functional clustering analysis to identify patterns of COVID-19 confirmed case and death trajectories across the US. Moreover, we quantify the associations between Google COVID-19 search trends for symptoms and COVID-19 confirmed case and death trajectories using dynamic correlation. Finally, we examine the dynamics of correlations for the top nine Google search trends of symptoms commonly associated with COVID-19 confirmed case and death trajectories. Our results reveal and characterize distinct patterns for COVID-19 spread and mortality across the US. The dynamics of these correlations suggest the feasibility of using Google queries to forecast COVID-19 cases and mortality for up to three weeks in advance. Our results and analysis framework set the stage for the development of predictive models for forecasting COVID-19 confirmed cases and deaths using historical data and Google search trends for nine symptoms associated with both outcomes.

Keywords: COVID-19 spread and mortality in US; functional data analysis; SARS-COV-2; Google COVID-19 search trends symptoms

Citation: Abbas, M.; Morland, T.B.; Hall, E.S.; EL-Manzalawy, Y. Associations between Google Search Trends for Symptoms and COVID-19 Confirmed and Death Cases in the United States. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4560. https://doi.org/10.3390/ijerph18094560

Academic Editor: Paul B. Tchounwou

Received: 16 March 2021
Accepted: 20 April 2021
Published: 25 April 2021

1. Introduction

In December 2019, an outbreak of the coronavirus disease (COVID-19) caused by the spread of the 2019 novel coronavirus (SARS-CoV-2) originated in the city of Wuhan in China. Due to the exponential worldwide spread of the virus, the World Health Organization (WHO) declared COVID-19 a pandemic on 11 March 2020. As of 11 January 2021, and according to John Hopkins University COVID-19 Dashboard, available at https://coronavirus.jhu.edu/map.html (accessed on 11 January 2021), the number of COVID-19 worldwide confirmed cases reached more than 90 million (including more than 22 million cases in the US) and the number of global deaths reached more than 1.9 million (~376,000 in the US). To date, the COVID-19 pandemic has caused a huge negative global impact on the economy [1], health [2,3], and education [4–6], with the US as one of the top countries affected by this pandemic.

Extensive open access COVID-19 data, including test results, present outstanding opportunities for data scientists to apply a variety of statistical and machine learning approaches to characterize underlying factors governing variation in the pattern and rate of COVID-19 spread across the US. For example, Tang et al. [7] applied functional principal component analysis (FPCA) to examine the modes of variation for COVID-19 confirmed case trajectories for 50 US states, quantified the correlations between confirmed case and death trajectories using functional canonical correlation analysis (FCCA), and grouped the 50 states into five subgroups where states shared similar COVID-19 spread

patterns. Chen et al. [8] used nonnegative matrix factorization (NMF) followed by a k-means clustering procedure applied to the NFM coefficients to cluster the COVID-19 confirmed case trajectories for 49 US states into seven groups and investigated the dynamics of the clustering results over time. Similar analyses had been applied to various regions in the world. For instance, Carroll et al. [9] used FPCA to investigate the modes of variation in COVID-19 case and death trajectories for 64 countries for an interval of 64 days and used functional regression analysis to show a significant association between reduced workplace mobility and lower spread of the virus as well as an association between baseline demographic (i.e., population density and percentage of population above 65 years) and doubling rates of mortality. Boschi et al. [10] used a functional clustering approach based on probabilistic *k*-mean with local alignment [11] to investigate the patterns of COVID-19 mortality across 20 Italian regions and applied functional regression analysis [12] to show strong associations between COVID-19 mortality and both local mobility and positivity.

Recently, Google released the Google COVID-19 search trends symptoms dataset (version 1.0) [13]. The dataset provides daily (and weekly) time series of the numbers of searches related to a specific symptom (e.g., fever) within a given geographic location (e.g., state or county). These counts were normalized using the total search activity in the target region. The main objective of this database is to provide an open-access dataset to accelerate scientific and public health insights into the spread and impact of COVID-19 while maintaining the privacy of Google users. Traditional approaches for forecasting the spread of a virus (e.g., seasonal flu) infection were found to predict results with one or two weeks delay [14]. Since its presentation in November 2008, Google Flu Trends (GFT), a web service that uses aggregated Google search queries data to provide daily estimates of the occurrence of flu two weeks in advance [15–17], had been shown to improve the predictive performance of traditional surveillance and forecasting systems [14,18,19]. Motivated by these results and several other studies demonstrating the viability of Google search trends in infectious disease surveillance and early epidemic prediction [20,21], we hypothesize that Google search trends for COVID-19-related symptoms could be used in combination with historical COVID-19 data to predict future COVID-19 spread and mortality rates.

To the best of our knowledge, no models for forecasting COVID-19 spread or mortality using the Google COVID-19 search trends symptoms database have been developed yet. The major aim of this study is to establish the foundation for developing such prediction models via identifying relevant symptoms associated with COVID-19 confirmed cases and deaths. Additionally, we apply functional data analysis techniques to investigate modes of variation and patterns of COVID-19 positivity and mortality as well as Google search trends for selected COVID-19-related symptoms in the US.

2. Methods

2.1. Data

The numbers of daily COVID-19 confirmed cases and deaths were obtained from the Centers for Disease Control and Prevention (CDC) at https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36 (accessed on 6 November 2020). The data were collected at the state level including the 50 states as well as the District of Columbia (DC) (hereafter, the 51 US states). The study period included the 245 days beginning 1 March 2020 (when the COVID-19 outbreak was declared a national emergency in the US) through to 31 October 2020, which was the time we conducted our analysis. To account for variation in population size in each state, we normalized the number of cases to count per million people in each state using 2019 census population estimates.

In addition to the two COVID-19 time-series datasets collected from CDC, we experimented with 422 time-series datasets extracted from the Google COVID-19 Search Trends Symptoms dataset (in short, symptoms dataset) [13]. The symptoms dataset was released by Google on 2 September 2020 and is publicly available at: https://github.com/google-research/open-covid-19-data/ (accessed on 6 November 2020). The dataset includes ag-

gregated state-level daily Google searches for 422 symptoms and conditions that may relate to COVID-19. The data were normalized by scaling the daily symptom search count proportional to the total search activity in each state during the day. We also included data extracted from the 2019 US Census data (available at https://www.census.gov/data.html, accessed on 6 November 2020) for four risk factors [22–24] representing the percentage of the state-level population with: black race; age greater than 65 years; unemployment status; and income below the poverty level.

2.2. Statistical Analysis Framework

In our analyses, we modeled the COVID-19 state-level data from CDC and Google search trends as functional data. Specifically, we transformed per-state time-series CDC and Google data collected during the study time interval into smoothed function curves (see Figure 1 as an example). Using these curves as data atoms, functional data analysis (FDA) provides tools analogous to statistical and multivariate analysis tools where random variables are replaced with random functions [12]. We used functional principal component analysis (FPCA) [12,25] to analyze variations in the data and detect patterns of COVID-19 spread and mortality in the US. To quantify the similarity between two functional datasets (e.g., COVID-19 case trajectories and Google search trends for one of the COVID-19-related symptoms such as hypoxemia), we used two functional correlation analysis methods. First, we used the robust and relatively fast dynamic correlation procedure [26,27] to quantify and rank the associations between COVID-19 trajectories and each of the 422 symptoms search trends. Second, we used the functional canonical correlation analysis (FCCA) method [12,28] to depict the pairwise correlations between COVID-19 trajectories and each of the top nine symptoms search trends identified from the first step. In what follows, we provide brief summaries of the functional data analysis techniques used in this study.

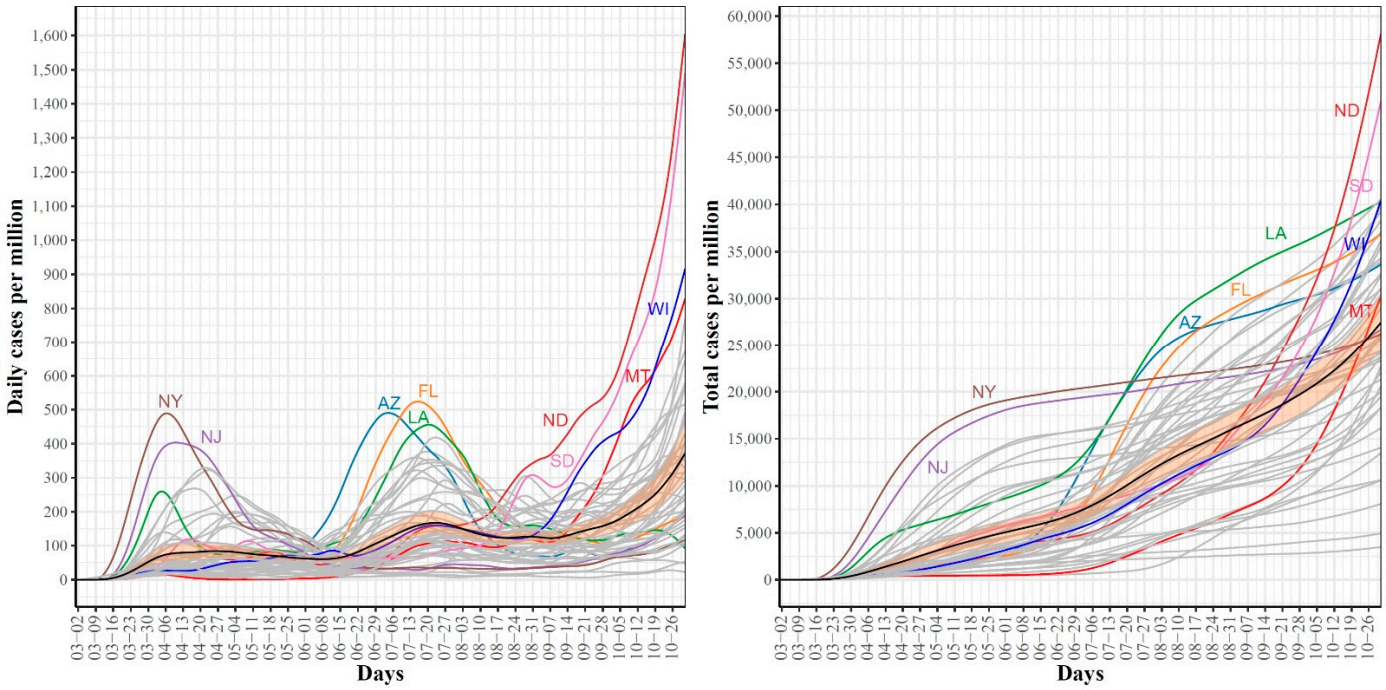

Figure 1. Trajectories for COVID-19 daily (**left**) and total (**right**) per million confirmed cases for the 51 US states. The mean curve is highlighted in black and the orange ribbon corresponds to the 95% confidence band.

2.3. From Time-Series to Functional Data

Let $\{t_{ij} : i = 1, 2, \ldots, N \text{ and } j = 1, 2, \ldots, T\}$ be an observed time-series dataset for $N = 51$ states over $T = 245$ days. We transformed each time-series into a function using a weighted linear combination of K basis functions (i.e., $x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t), \ i = 1, 2, \ldots, N$ where ϕ_k is the kth basis function). In our experiments, we used the cubic B-spline as the basis functions and set K equal to T (i.e., $K = 245$). The coefficients c_{ik} were estimated using the roughness penalty [25,29,30] on the second derivative of each curve.

Using this data representation, we: (i) examined the trajectories, and mean and variance curves to identify trends in the data; (ii) applied several FDA approaches to the three functional datasets considered in this study.

2.4. Functional Principal Component Analysis

Functional principal component analysis (FPCA) is a powerful and widely used tool for linear dimensionality reduction of functional data [12,25]. In multivariate analysis (MVA), PCA is used to: (i) reduce the number of (potentially) correlated variables to a smaller number of uncorrelated variables (principal components); (ii) explore the latent relationships between variables. PCA determines the weights for combining the input variables into principal components with maximized variations. Specifically, the principal component scores f_i of a sample x_i are determined as $f_i = \sum_{j=1}^{p} \beta_j x_{ij}$ for $i = 1, \ldots, N$, where the unknown weight vector $\beta_j = (\beta_1, \ldots, \beta_p)$ is estimated using the following stepwise procedure [12]:

1. Determine the first eigenvector ξ_1 that maximize $f_{i1} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \xi_{j1} x_{ij} \right)^2$ such that $\sum_{j=1}^{p} \xi_{j1}^2 = \| \xi_1 \|^2 = 1$.
2. Repeat this step to compute subsequent eigenvectors ξ_m for $m = 2, \ldots, p$. On the mth step, compute ξ_m such that $\frac{1}{N} \sum_{i=1}^{N} f_{im}^2$ is maximized and subject to the following constraints:

$$\| \xi_m \|^2 = 1$$

$$\sum_{j=1}^{p} \xi_{jk} \xi_{jm} = 0, \ \ k < m$$

For functional data, the preceding procedure could be used with functions replacing vectors, continuous index t replacing the discrete index j and integrals replacing summations. For more details, interested readers are referred to [12].

For analyzing the variation in the functional data, we applied the concept of modes of variations for functional data [31] which visualizes the range of FPC scores. The kth mode of variation is the set of functions computed as

$$\mu(t) \pm \alpha \sqrt{\lambda_k} \, \xi_k(t), \ \ \alpha \in [-A, A]$$

where $\mu(t)$ is the mean curve and λ_k is the kth eigenvalue and $A = 2$.

2.5. Functional Clustering

An interesting application of FPCA is to represent each functional curve $x_i(t)$ using its FPCA scores in order to enable the application of traditional multivariate statistical and machine learning approaches. For clustering the functional curves of the 51 states, we represented each functional curve using the normalized first four FPC scores, which

accounts for more than 90% of the variability in the data (See Section 3.7). We then used the k-means algorithm [32,33] and the Euclidean distance to cluster the data into k groups. We evaluated the clustering results for different choices of k (i.e., $k = \{2, 3, \ldots, 10\}$) using the majority rule applied to 30 different selection criteria (included in NbClust R package (version 3.0) [34]).

2.6. Dynamic Correlation

To quantify the similarity between per-state curves for confirmed cases and deaths to counterparts' curves for 422 symptoms, we used dynamic correlation [26,27]. Dynamic correlation is a non-parametric functional data analysis technique developed to measure the correlation between two random functions at individual and population levels. Given two random functions, dynamic correlation can be viewed as the cosine similarity between them after population-centering [27]. It is worth noting that the name "dynamic correlation" reflects the centering around a dynamic (time-varying) instead of a static population mean and the estimated correlation coefficient is a static value between -1 and 1. In our experiments, we used an implementation of the dynamic correlation methods proposed in [26] provided via the "dynCorr" R package (version 1.1.0) [35].

2.7. Functional Canonical Correlation Analysis

Canonical correlation analysis (CCA) [12,28] quantifies the associations between two sets of variables by transforming both of them into a common lower dimensional space with maximum correlations. Given two random vectors X and Y, the CCA finds two weight vectors u and v such that the two linear transformations $u^T X$ and $v^T Y$, also called the canonical variables, are maximally correlated. In 1993, Leurgans et al. [36] adapted the CCA for functional data. Given two random functions, $X(t)$ and $Y(t)$, the functional correlation analysis (FCCA) seeks two weigh functions $u(t)$ and $v(t)$ such that $\int_0^T uX\,dt$ and $\int_0^T vY\,dt$, called functional correlation variables, are maximally correlated. In our analysis, we used the cca.fd function from the fda R package [37] to conduct FCCA with roughness penalties on the second derivative of each curve.

3. Results

3.1. Trajectories of COVID-19 Confirmed and Death Cases in US

Figure 1 shows the trajectories for COVID-19 daily (left) and cumulative (right) per million confirmed cases for the 51 US states. In the trajectories of the confirmed daily COVID-19 cases, the mean curve (black line) shows two local maxima (i.e., waves). The first wave peaked on 6 April where the largest numbers of confirmed cases were reported in NY and NJ. The second wave peaked on 20 July where the leading states were FL, AZ, and LA. The mean curve shows that another local maximum with larger numbers of cases is expected to be formed after 31 October (end of our study time interval). In the 3rd wave, ND and SD are expected to contribute the largest numbers of confirmed COVID-19 cases. The trajectories of total COVID-19 confirmed cases complement the trajectories of daily cases and reveal several interesting patterns. First, the two leading states in the first wave, NY and NJ, had total cumulative case rates greater than the mean from the start of the pandemic until the last week of October. However, their curves seem to have flattened since mid-June. Second, AZ and FL, the two leading states in the second wave, were below the mean curve until mid-June, after which their curves were above the mean curve. The 3rd leading state in the second wave, LA, followed a pattern that was different from AZ and FL. The LA curve was always above the mean curve. Third, the leading states in the 3rd wave had total numbers below the mean and started to cross the mean curve in the last week of August (the time when most schools and universities had opened for the new academic year). Fourth, by the end of our study time duration, the cumulative numbers of confirmed COVID-19 cases in the 3rd wave leading states (ND, SD) exceeded the numbers for the 2nd wave leading states (AZ, FL), which themselves exceeded the numbers for the

leading states in the 1st wave (NY, NJ). Overall, the trajectories of the confirmed COVID-19 cases demonstrated variations in the spread of COVID-19 across the US states.

Figure 2 shows the trajectories for COVID-19 daily (left) and cumulative (right) per million death cases for the 51 states. For the daily trajectories, the mean curve had one global maximum which occurred in mid-April and starting mid-June, the mean curve was almost horizontal until mid-October where the curve seemed to start going up again. The global maxima in the mean curve of daily death cases correspond to the first wave of the COVID-19 spread in the US and therefore, it is not surprising that most deaths are reported in NY and NJ, the two leading states in the first wave. Interestingly, the NJ curve of daily death cases had a unique pattern with two peaks occurring on 20 April and 25 June. This peak reflects the sudden jump in the number of COVID-19 deaths in NJ when 1854 probable COVID-19 deaths were added. The trajectories of the total COVID-19 death cases show that from the starting date of our study period until 25 June, NY had the largest cumulative number of COVID-19 deaths. After 25 June and until the end of the study period, NJ is the leading state in terms of the total per million number of COVID-19 deaths. At the end date of our study period, the top leading states in terms of total per million number of deaths are NJ, NY, MA, CT, LA, and RI.

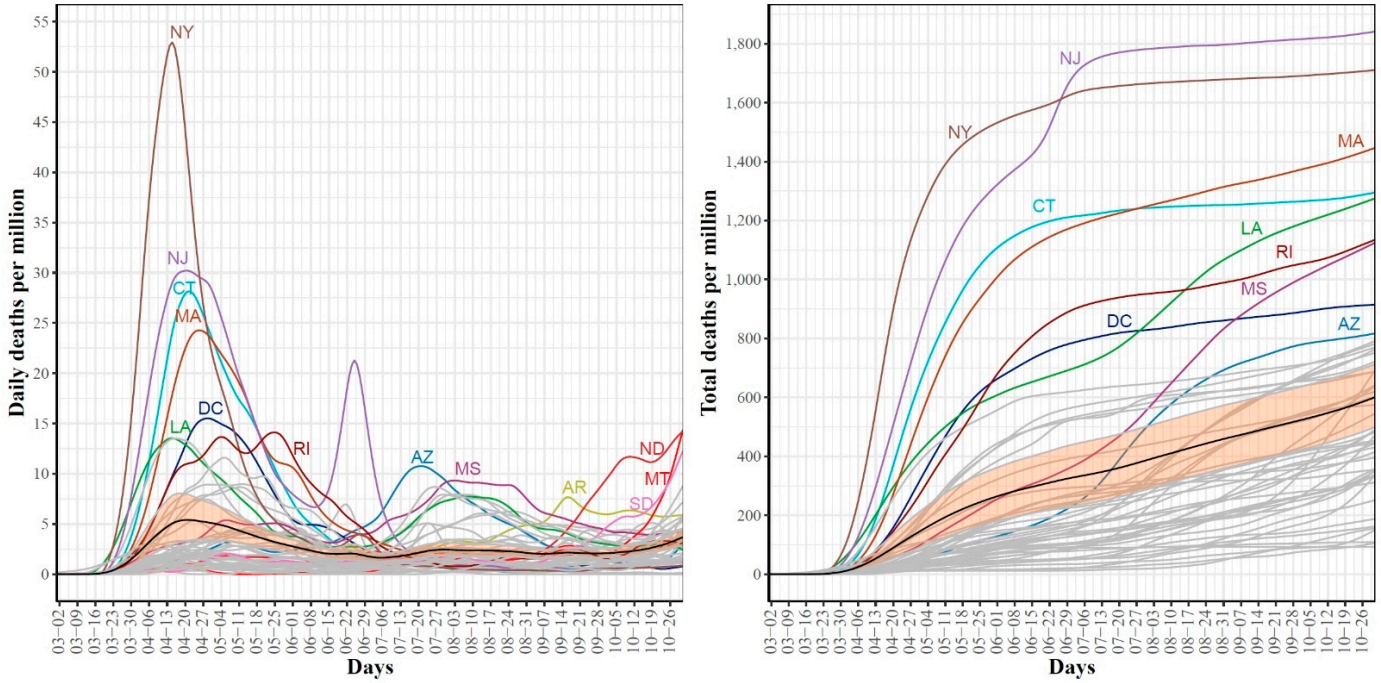

Figure 2. Trajectories for COVID-19 daily (**left**) and total (**right**) per million death cases for the 51 US states. The mean curve is highlighted in black and the orange ribbon corresponds to the 95% confidence band.

3.2. Modes of Variations in COVID-19 Confirmed and Death Cases Trajectories

To further explore the variations in the curves representing COVID-19 daily confirmed and death cases, we utilized the FPCA to project the data in a lower dimensional space. Figure 3 shows the top four eigenfunctions accounting for 95.28% of the variations in the curves shown in Figure 1. The first eigenfunction ξ_1 was less than zero until 4 June and then seemed to be exponentially increasing. This suggests the variability between states started to increase in early June and as we move forward, the variability increases more. The second eigenfunction ξ_2 had two peaks. The first peak was negative and was observed in mid-April (close to the time of the first wave peak). The second peak was positive and occurred in mid-July (close to the time of the peak in the second wave). The curve also shows that a third negative peak is about to be formed likely at a future time that is close to the time of the peak in the third wave. The third eigenfunction ξ_3 was always positive (except during the first 10 days of March) and had a global maximum in mid-April and

another local maximum in mid-July. The fourth eigenfunction ξ_4 was sinusoidal with alternating negative and positive peaks occurring in mid-April, mid-May, end of June, and end of August. Since the top two eigenfunctions accounted for 82.7% of the total variation in the data, we projected the 51 states in a two-dimensional space determined by the first two FPC scores. This projection of the 51 curves is shown in Figure 4. We found that the first wave leading states, NY and NJ, had negative FPC1 and lowest FPC2 scores. The three leading states in the second wave (FL, AZ, and LA) had negative FPC1 scores and the highest FPC2 scores. Finally, the leading states in the third wave (e.g., ND, SD, and WI) had the highest FPC1 scores but negative FPC2 scores. Figure 5 shows the spectrum of curves representing the range of FPC1 and FPC2 scores. The first mode of variation captured greater variability in the data in September and October and showed that from mid-May to the end of June, all curves were close to the mean curve. The top leading state in each of the three waves (NY, AZ, and ND) could be easily identified as outliers from the spectrum of curves for FPC1 scores. The second mode of variation captured the shape of the curves and shows that the greatest variability among curves occurred during the second wave. Two inflection points were noted at the first of June and October, respectively.

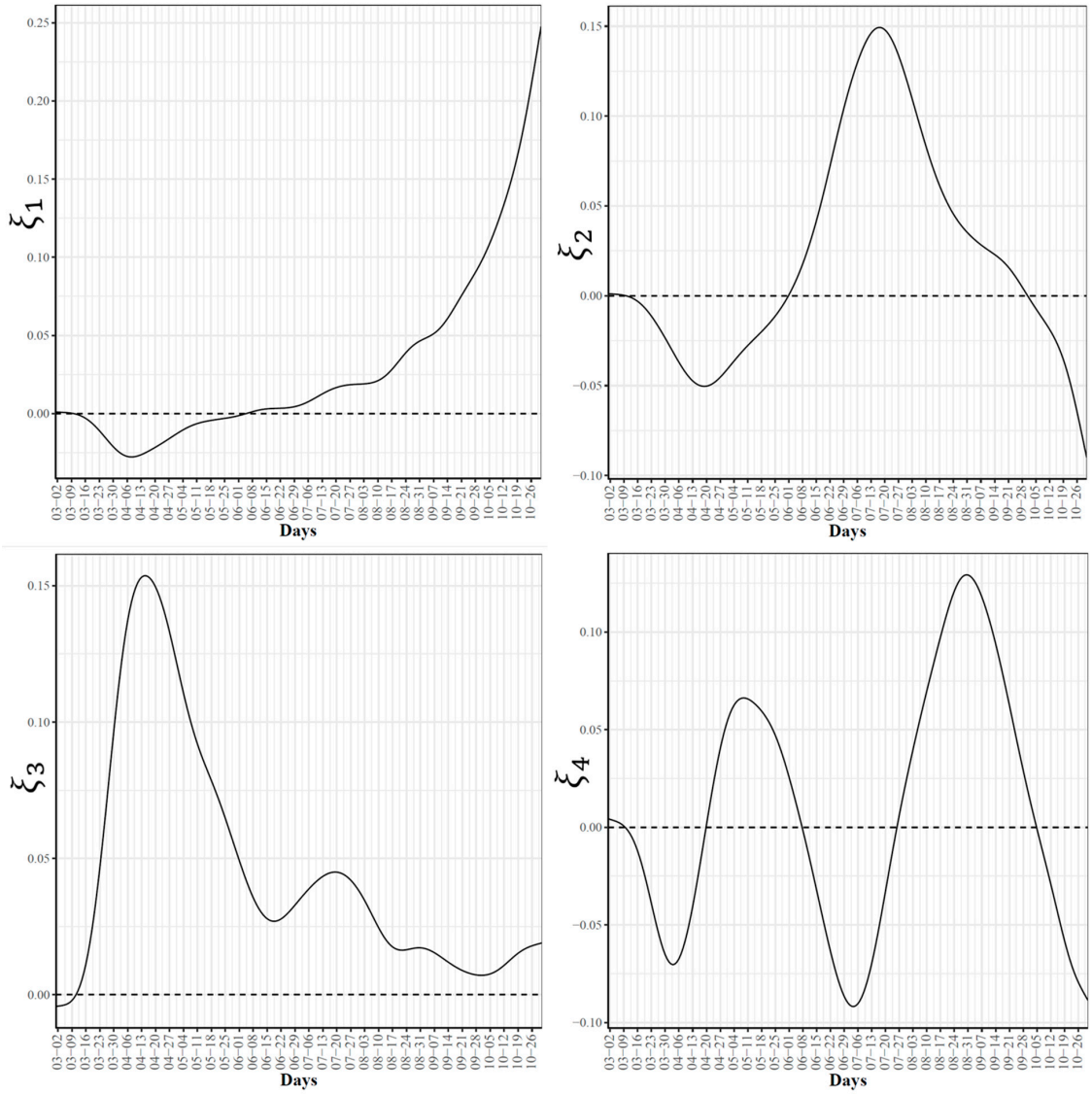

Figure 3. First four eigenfunctions of COVID-19 confirmed case trajectories.

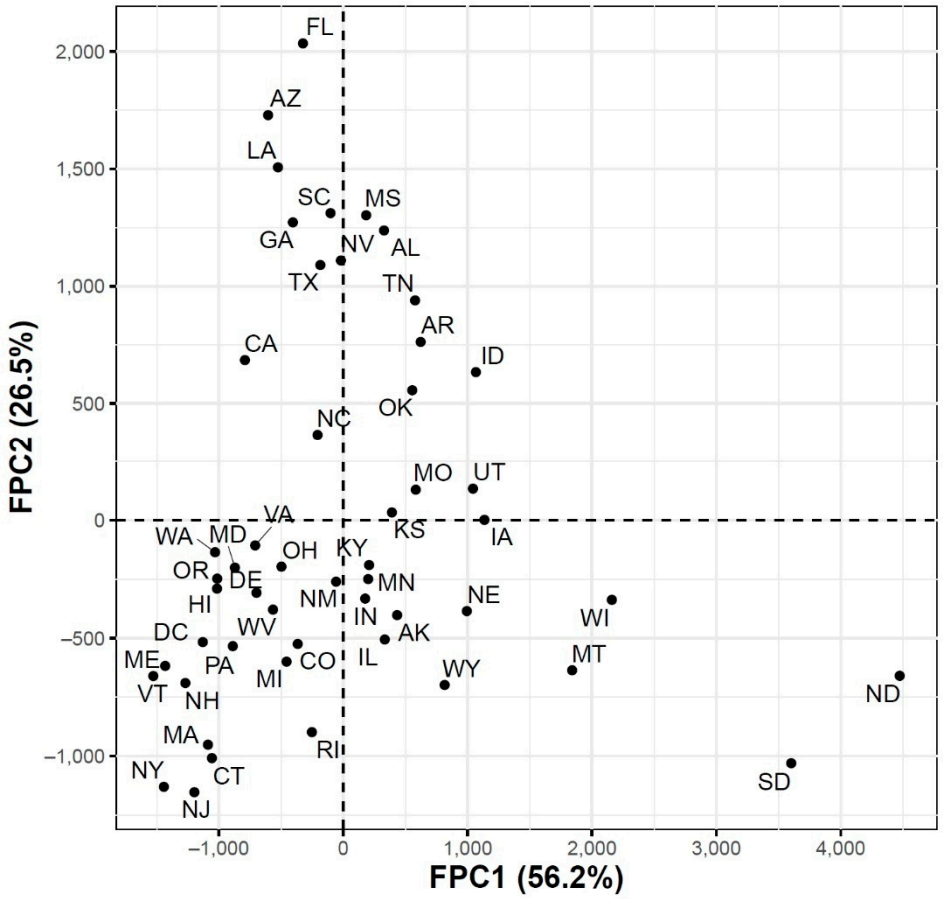

Figure 4. Scatter plot of the 51 US states in a two-dimensional space defined by the first two functional principal component scores of the COVID-19 case trajectories.

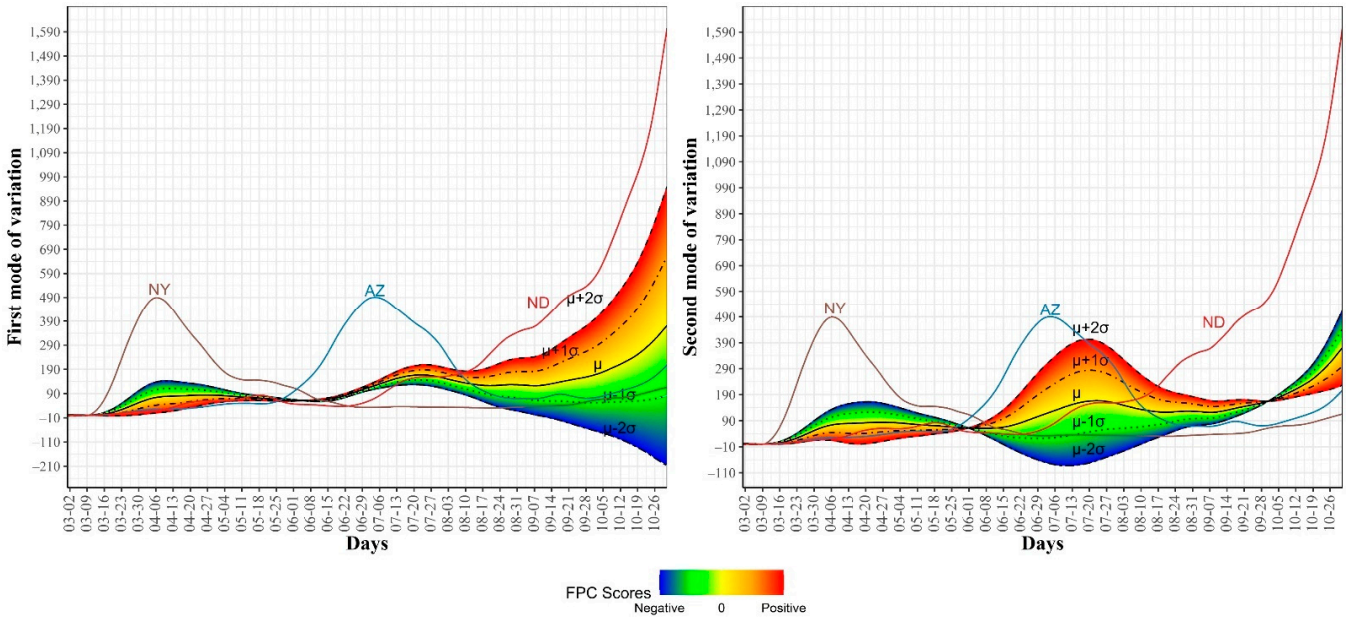

Figure 5. First and second modes of variations for the COVID-19 confirmed trajectories. The sample standard deviations of the first and second FPC scores are 1179.94 and 810.53, respectively. The curve for the leading state in each wave is also shown.

To examine the modes of variation in the COVID-19 daily death cases, we report the top four eigenfunctions (Figure 6) as well as the two-dimensional representation of the 51 US states in the space defined by the first two FPC scores (Figure 7). In Figure 6, we noted that the first eigenfunction ξ_1 had two positive peaks corresponding to the two peaks in the trajectories of NY and NJ, respectively (see Figure 2). We also noted that the second eigenfunction ξ_2 had a negative peak in mid-April and two positive peaks in mid-May and late July. These two eigenfunctions explained 83.8% of the variability in the COVID-19 daily death trajectories. Figure 7 shows the 51 US states in the two-dimensional FPC space. Most of the states seemed to form a single cluster and outlier states represented the states with high rates of COVID-19 mortality. Although NY and NJ were two clear outliers, they were not close to each other in the FPC space which reflects the different patterns in the daily death trajectories of the two states (see Figure 2). Figure 8 shows the first and second modes of variations for COVID-19 death trajectories. The first mode of variation shows that the greatest variation in the data occurred in the third week of April. Starting mid-July, the variability in the data was consistently low. The second mode of variation seems to modulate the shapes of the curves and the inflection point at the third week of April corresponds to the time when the death rates in the first wave started to drop. In both modes of variation, the top leading state in each of the three waves (NY, AZ, and ND) appeared as clear outliers.

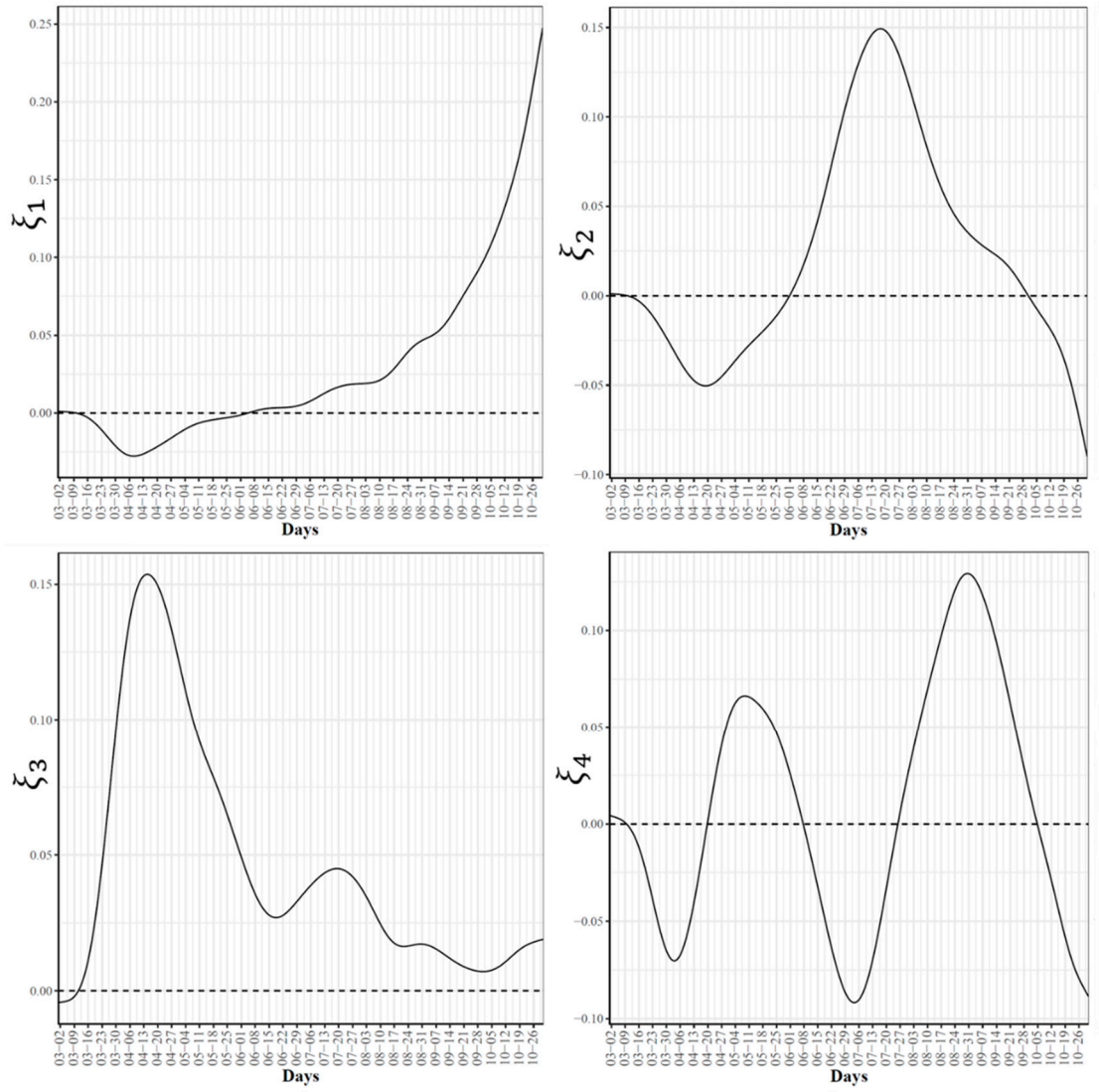

Figure 6. First four eigenfunctions of COVID-19 death trajectories.

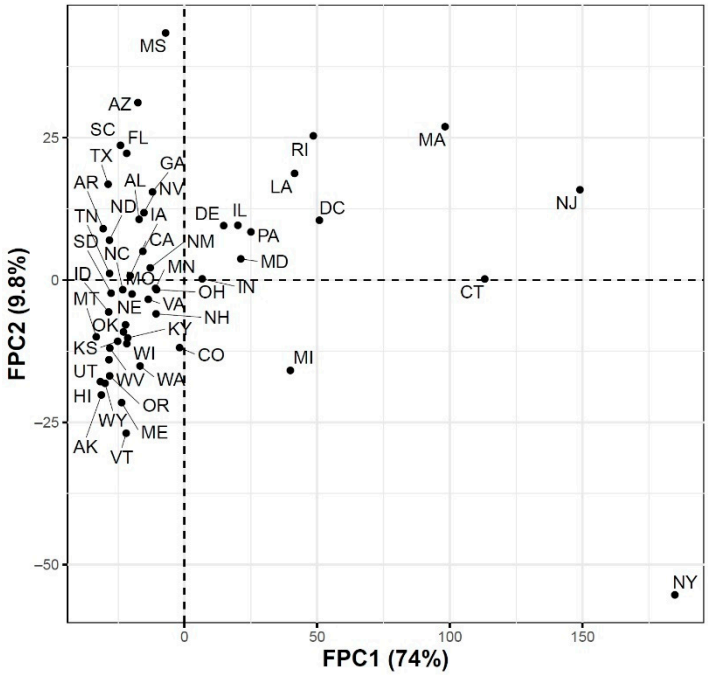

Figure 7. Scatter plot of the 51 US states in a two-dimensional space defined by the first two functional principal component scores for the COVID-19 death trajectories.

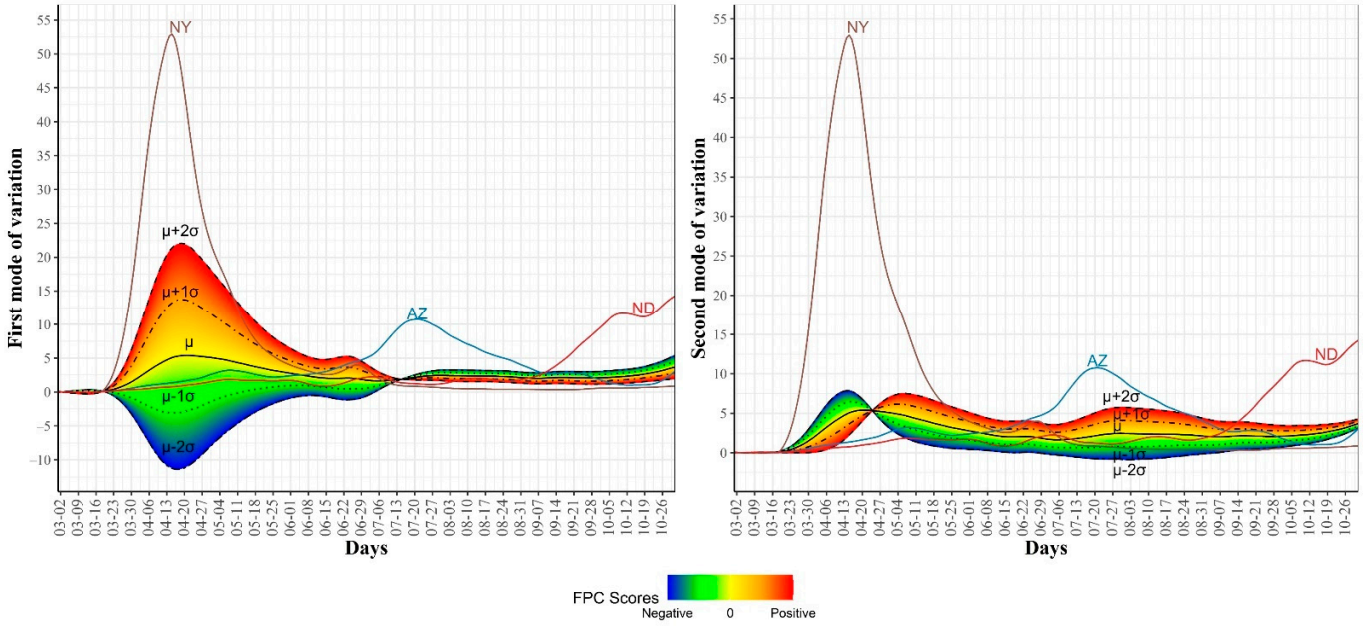

Figure 8. First and second modes of variations for the COVID-19 death trajectories. The sample standard deviations of the first and second FPC scores are 46.57 and 16.99, respectively. The curve for the leading state in each wave is also shown.

3.3. Clustering US States by Trajectories of COVID-19 Confirmed Cases

Figure 9 shows the clustering of the 51 US states based on the trajectories of COVID-19 daily confirmed cases. We found that the optimal number of clusters was seven. This large number of clusters reflects the great variations in the spread of COVID-19 across different US states due to differences in states related to demographics of its population and health policies adopted by the US states.

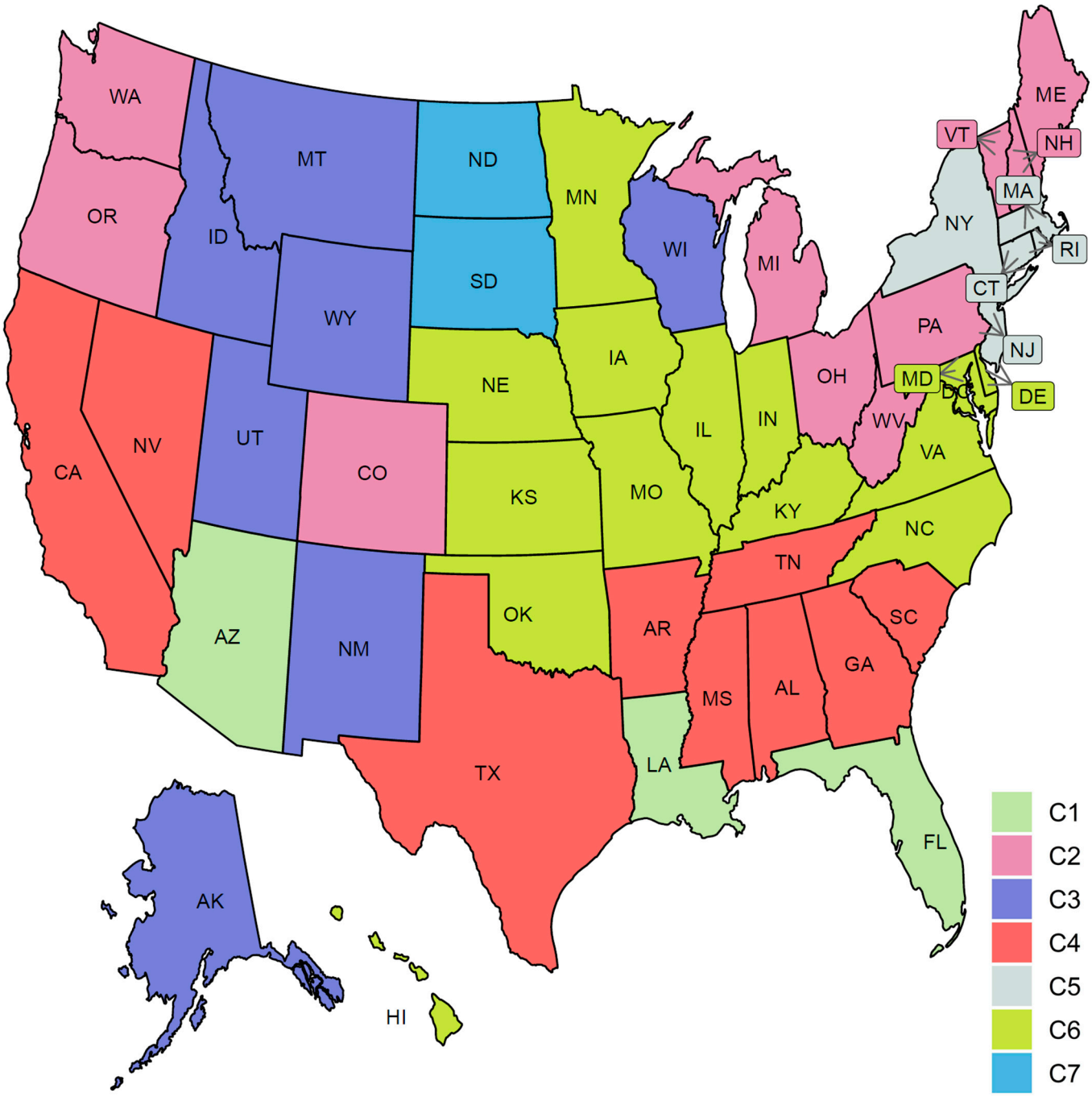

Figure 9. Clustering result of the 51 states based on the COVID-19 confirmed case trajectories. States assigned to the same cluster share the same color.

Figure 10 characterizes the member states in each of the seven identified groups from the curves of COVID-19 confirmed cases. Cluster C1 included the three leading states in the second wave AZ, FL, and LA. Interestingly, none of these three states had shared borders with the other two states. States in this group encountered two peaks in the numbers of COVID-19 cases. The first peak (with a mean of 100 cases per million) occurred in the first week of June while the second peak (with a mean of ~460 cases per million) occurred in mid-July. At the end of the study interval, another peak is about to be constructed with a mean of at least 150 cases per million. Cluster C2 included 10 states (CO, ME, MI, NH, OH, OR, PA, VT, WA, WV). This class covered states from North East, Mid-Atlantic, North West, and one isolated state, CO, from the Mid-West. Like the cluster C1 mean curve, the cluster C2 mean curve had two peaks which occurred in early June and 20 July but

the number of cases in these two peaks was around 50–60 cases per million. Cluster C3 included six western states (AK, ID, MT, NM, UT, WY) and one Mid-West state, WI. The mean curve of cluster C3 had only one peak (with 150 cases per million) that occurred on the 20 July and a new peak (with more than 600 cases per million) will be formed after the end of the study duration. Cluster C4 had seven southern states (AL, AR, GA, MS, SC, TN, TX) and two western states (CA and NV). Cluster C4 had a mean curve with a dramatic increase in the number of cases that started in the third week of May and reached a peak of 320 cases per million on the 20 July. The future expected peak is expected to have more than 250 cases. Cluster C5 had seven states (CT, DC, MA, NJ, NY, RI) which covers the two leading states in the first wave and five neighboring states from the North East and Mid-Atlantic regions. States of this cluster had an early peak in the third week of April with the average number of cases reaching 300. Then, the average number of cases dropped to less than 50 in mid-June before crossing 100 cases per day in mid-October. Cluster C6 had 14 states from the Mid-West and Mid-Atlantic regions (DE, HI, IA, IL, IN, KS, KY, MD, MN, MO, NC, NE, OK, VA). The mean curve of states in cluster C6 had three local maxima with 100, 150, and 150 cases, respectively. The average number of cases exceeded 350 daily cases at the end of our study period. Cluster C7 included two states, ND and SD, which are the top two leading states in the third wave. The mean curve of these two states crossed the 100 daily cases for the first time in mid-July and crossed the 500 daily cases in the last week of September and reached more than 1500 daily cases by the end of October. We noted that some clusters consist of two or more separated regions. For example, cluster C2 had four separated regions. This suggests that geographical location was not the key factor in shaping the pattern of COVID-19 spread across the US states due to travel restrictions and differences in the policies implemented by each state.

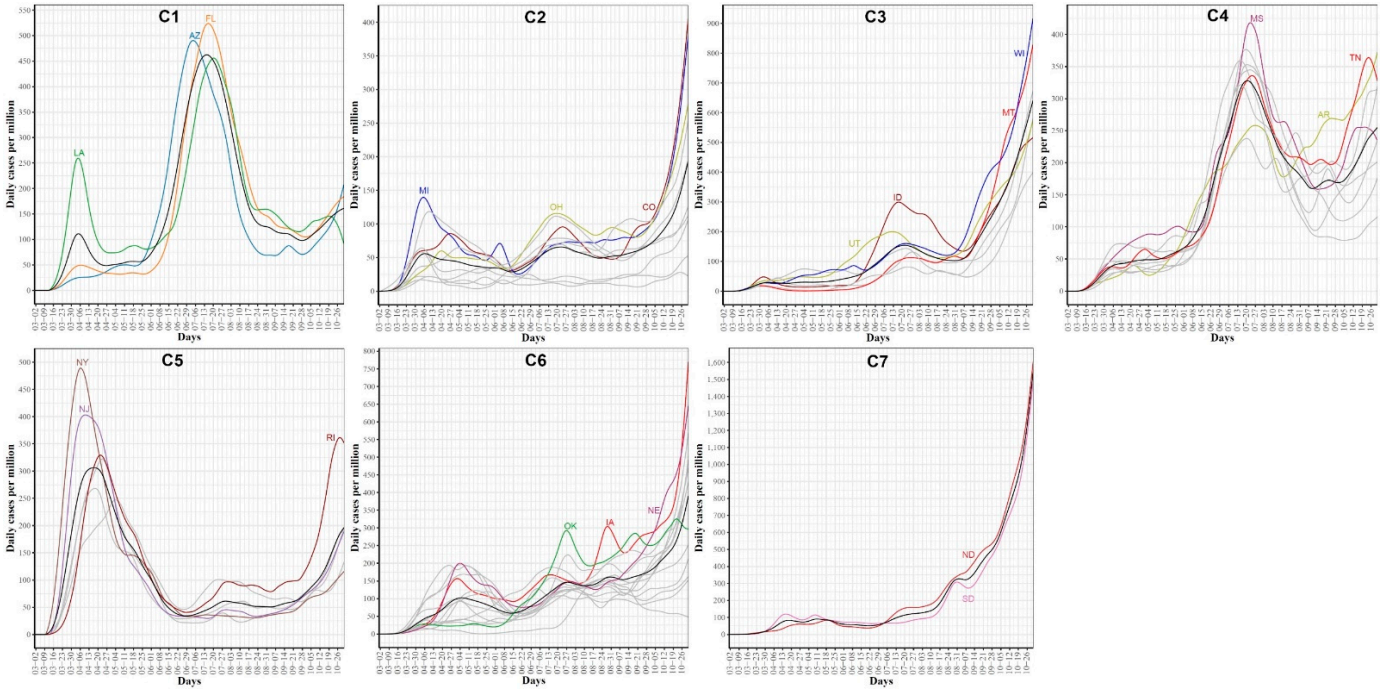

Figure 10. COVID-19 confirmed case trajectories of the states in each cluster. The mean curve for the states in each cluster is highlighted in black.

We also used four potential risk factors from 2019 census data to characterize the seven groups. Figure S1 shows the boxplots of the seven groups for each risk factor. Table 1 summarizes these results using the median and the rank of these risk factors in each group.

Table 1. Characterization (in terms of median and rank) of the seven groups of COVID-19 spread patterns across the 51 US states using four risk factors from 2019 US census data.

Census Variable	C1	C2	C3	C4	C5	C6	C7
Black race	16.9(2)	4(5)	1.5(7)	17.1(1)	13.6(3)	9.2(4)	2.9(6)
Age 65+	18.2(2)	18.5(1)	17.1(3)	16.5(5)	17(4)	16.2(7)	16.3(6)
Unemployment	5(1)	4.5(3)	3.5(6)	4.8(2)	4.5(3)	4(5)	2.7(7)
Below poverty level	13.5(2)	11.2(5)	10.4(6)	13.8(1)	10.4(6)	11.4(3)	11.3(4)

3.4. Clustering US States by Trajectories of COVID-19 Death Cases

Clustering the 51 US states based on the trajectories of COVID-19 daily death cases yielded five clusters (see Figure 11). Figure 12 characterizes the member states in each group. Cluster C1 included nine southern states (AL, AZ, FL, GA, LA, MS, NV, SC, TX). Its mean curve showed two peaks that were observed in mid-April and the last week of July with the average number of per million deaths equaling 3 and 7, respectively. Cluster C2 included nine states (CT, DC, DE, IL, MA, MD, NJ, PA, RI). IL is the only member state in C2 that does not have any shared borders with any other member states in C2. The mean curve of C2 reached its global maximum in the last week of April with 15 per million deaths. Cluster C3 had 27 states (AK, CA, CO, HI, IA, ID, IN, KS, KY, ME, MI, MN, NC, NE, NH, NM, OH, OK, OR, TN, UT, VA, VT, WA, WI, WV, WY) and the average number of per million deaths for this cluster was always less than or equal 3. Cluster C4 included five states spanning North-West (MT), Mid-West (MO, ND, SD), and South-East (AR) regions. Its mean curve demonstrated that this cluster had the largest number of per million deaths since the beginning of October 2020. Cluster C5 had only one state, NY. The trajectory curve for NY followed the same pattern as the trajectory curves in C2; however, NY reached a peak of 52 per million death cases in mid-April, while the leading state in C2, NJ, reached a peak of 30 per million death cases. Table 2 characterizes each cluster using the median of four state-level risk factors from 2019 US census data. Interestingly, cluster C1 is ranked at the top for three risk factors (race is black, unemployment, and poverty), while it had the lowest rank for the percentage of the population older than 65 years.

Table 2. Characterization (in terms of median and rank) of the seven groups of COVID-19 death patterns across the 51 US states using four risk factors from 2019 US census data.

Census Variable	C1	C2	C3	C4	C5
Black race	26.8(1)	14.6(3)	4.6(4)	3.4(5)	17.6(2)
Age 65+	16.5(5)	17.1(2)	16.8(4)	17.2(1)	16.9(3)
Unemployment	4.9(1)	4.6(2)	3.9(4)	3.8(5)	4.4(3)
Below poverty level	13.6(1)	10.8(5)	11.2(4)	12.6(3)	13(2)

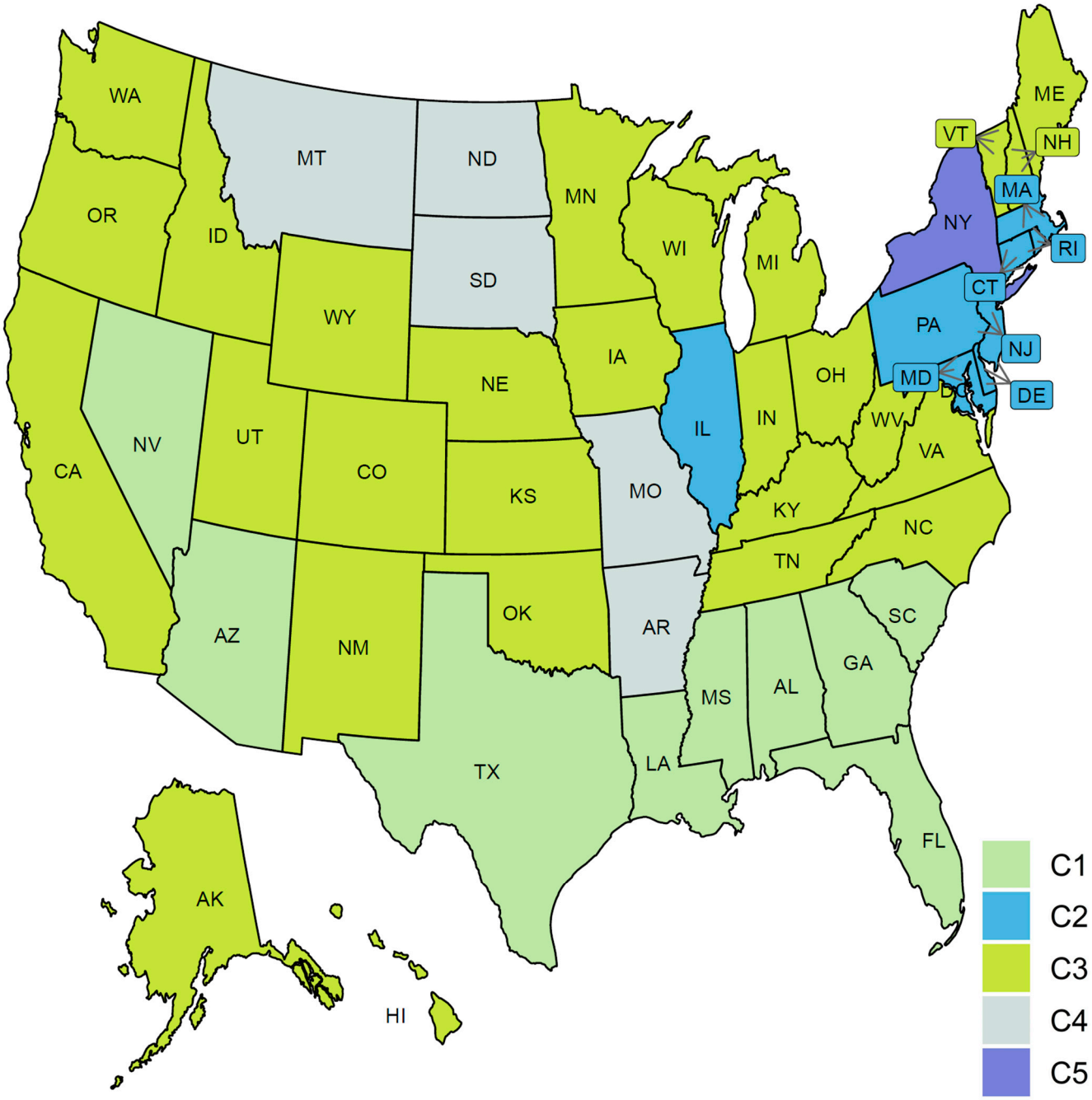

Figure 11. Clustering result of the 51 states based on the COVID-19 death trajectories. States assigned to the same cluster share the same color.

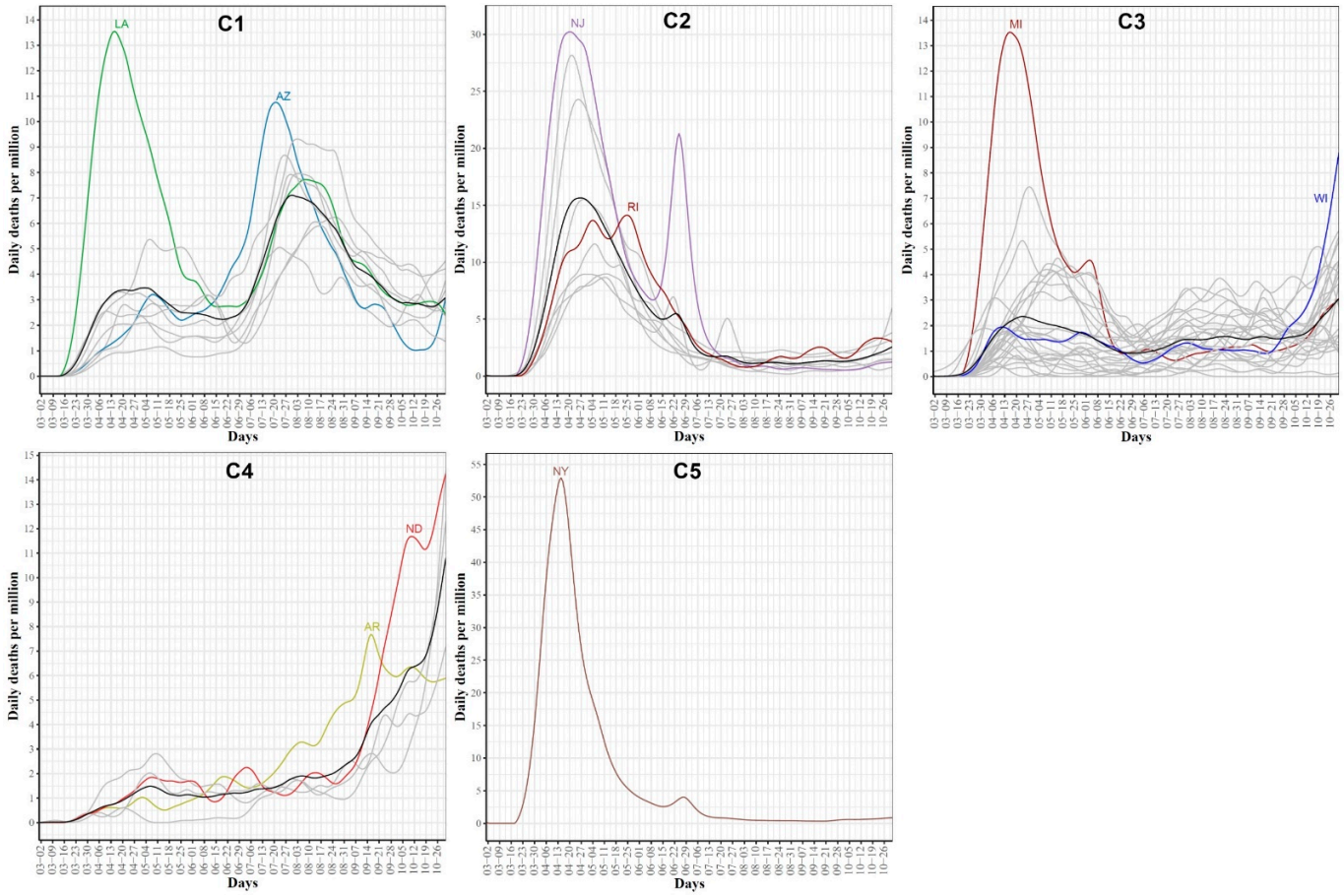

Figure 12. COVID-19 death trajectories of the states in each cluster. The mean curve for the states in each cluster is highlighted in black.

3.5. Associations between Google COVID-19 Search Trends for Symptoms and COVID-19 Confirmed and Death Cases

Anonymized and aggregated state-level Google COVID-19 daily searches for symptoms were smoothed and treated as 422 functional datasets, one dataset for each symptom. Pairwise associations between each symptom functional dataset and functional datasets for COVID-19 confirmed and death cases, respectively, were quantified using dynamic correlation [26]. The complete results are reported in Tables S1 and S2. Table 3 summarizes the top 15 symptoms, sorted by the absolute value of the dynamic correlation estimates, associated with COVID-19 confirmed and death cases, respectively. We noted that the top 15 symptoms associated with COVID-19 daily confirmed cases had relatively higher dynamic correlations compared to the dynamic correlation estimates for the top 15 symptoms associated with COVID-19 daily death cases. Interestingly, hypoxemia, ageusia, and anosmia are amongst the top 5 symptoms in both cases. The table includes four symptoms with no clear relation to COVID-19: xeroderma, bruise, nodule, and genital wart. Interestingly, the four symptoms are related to dermatologic conditions.

Table 3. Top 15 symptoms associated with COVID-19 confirmed cases and deaths, respectively. The nine symptoms shared in the two sets are highlighted in bold.

Symptom	Correlation	Symptom	Correlation
Hypoxemia	0.78	**Hypoxemia**	0.65
Ageusia	0.69	**Hypoxia**	0.54
Anosmia	0.69	**Pneumonia**	0.48
Dysgeusia	0.65	**Anosmia**	0.46
Hypoxia	0.64	**Ageusia**	0.41
Sinusitis	0.62	Panic attack	0.40
Fever	0.61	**Fever**	0.40
Low grade fever	0.61	**Shortness of breath**	0.40
Xeroderma	0.56	Chest pain	0.40
Pneumonia	0.55	**Dysgeusia**	0.39
Chills	0.54	Nodule	−0.38
Shortness of breath	0.54	Genital wart	−0.37
Cough	0.54	Bradycardia	0.36
Nasal congestion	0.52	Bronchitis	0.35
Bruise	−0.52	**Chills**	0.35

Symptoms shared in the two sets are highlighted in bold.

Figure 13 shows the mean curves for COVID-19 daily confirmed and death cases as well as the mean curve with the top 3 symptoms associated with COVID-19 daily confirmed and death functional data, respectively. It should be noted that the correlations reported in Table 3 were not computed using the mean curves but using the state-level curves in each group. However, the mean curves summarized the trend in each group and showed interesting trends in the search curves that are correlated with the times of the three waves. For example, the mean curve for hypoxemia showed a drop in searches for this symptom in two time-intervals, 05/20–06/20 and 07/29–09/14, where a drop in the average number of daily confirmed cases is noticed in these two time-intervals. We also noted that the mean curves for ageusia and anosmia seemed to have the same trend. The dynamic correlation coefficients between the curves for COVID-19 daily confirmed cases and the curves for these symptoms were 0.69 in both cases.

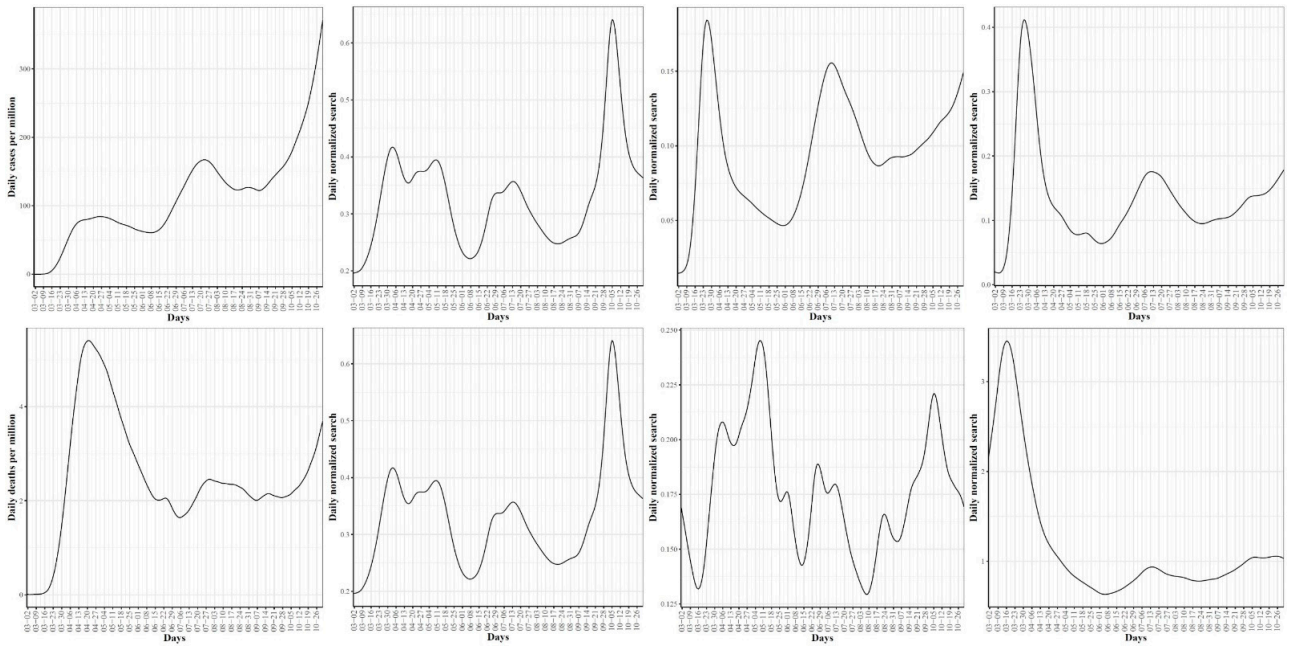

Figure 13. At the top are the mean curves for the COVID-19 confirmed cases, hypoxemia, ageusia, and anosmia searches, respectively. At the bottom are the mean curves for the COVID-19 death cases, hypoxemia, hypoxia, and pneumonia searches, respectively.

3.6. Dynamics of Associations between Top Google Search Trends and Future COVID-19 Confirmed and Death Cases

Let $= \{x_1, \ldots, x_{245}\}_1^{51}; Y = \{y_1, \ldots, y_{245}\}_1^{51}$ represents the time-series data for COVID-19 confirmed cases and Google searches for hypoxemia, respectively. As shown in Table 3, the dynamic correlation coefficient between these two functional datasets is 0.78. This high correlation score indicates a strong association between the number of searches for hypoxemia y_t and the number of COVID-19 cases x_t at any timepoint t. Additionally, we are interested in the correlation between y_t and x_{t+k}, $k \geq 1$, which represents the association between the number of searches for hypoxemia in day t and the number of COVID-19 cases observed in the future (i.e., k days after t). Figure 14 reports such dynamic correlations for the nine symptoms that are shared in the top associated 15 symptoms with COVID-19 confirmed (left) and death (right) cases for $k = 0, 1, \ldots, 30$ days. For correlations with COVID-19 confirmed cases, the highest correlations were observed at $k = 0$ for hypoxemia, hypoxia, and pneumonia, while for the remaining six symptoms, the highest correlations (which are slightly greater than the correlations at $k = 0$) were noted for $k \in [9–15]$. The only feature that had a non-slight increase in its correlation coefficient compared to its initial correlation was "shortness of breath". Correlation curves for hypoxemia, hypoxia, and pneumonia started to drop earlier than the curves for the remaining symptoms. At $k = 30$, hypoxemia and pneumonia had the lowest dynamic correlation coefficients compared with the correlations for the remaining symptoms.

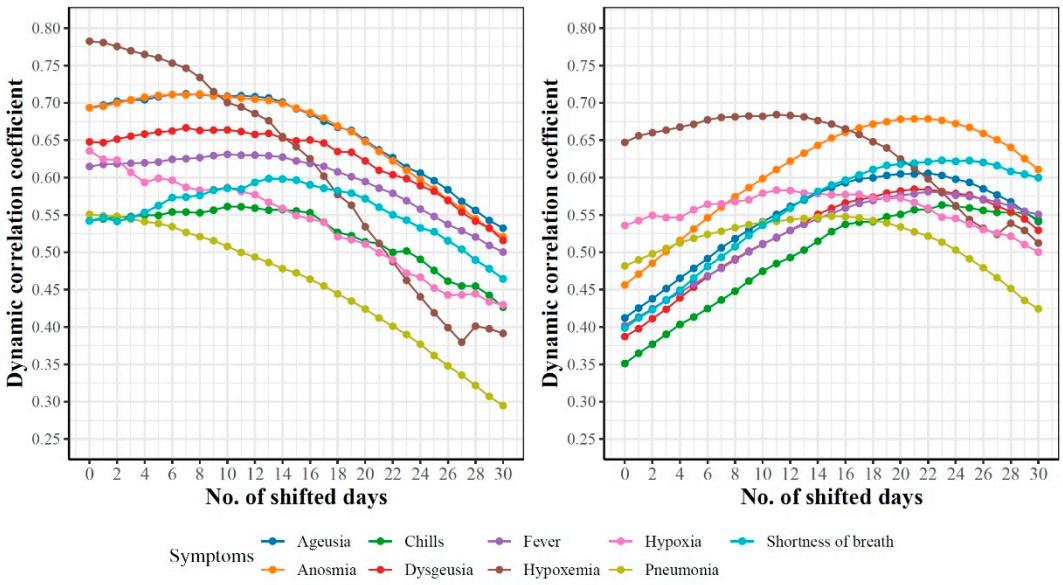

Figure 14. Dynamicity of the pairwise associations between the trajectories of nine symptoms and k-days ahead trajectories of COVID-19 cases (**left**) and deaths (**right**).

For correlations with COVID-19 death cases, we noted a consistent increase in the correlation coefficients for all nine symptoms until each symptom reaches its maximum value at $k \in [13–23]$ before the correlations start to drop. Interestingly, we noted that the correlations for hypoxemia, hypoxia, and pneumonia started to drop around $k = 15$ days, while the correlations for other symptoms started to drop around $k = 23$.

3.7. Exploratory Functional Data Analysis of Google Search Trends for Hypoxemia

Since hypoxemia is the most correlated symptom with both COVID-19 confirmed and death functional data, we report an exploratory FDA of the Google daily search trends for hypoxemia. In addition, we included the results when this exploratory analysis was applied to anosmia and shortness of breath in Supplementary Figures S3–S12.

First, Figure 15 (left) shows the smoothed curves for the 51 states. The mean curve had three waves in the searches for hypoxemia. The first wave peaked on 6 April where NY, NJ DC, and CT were the leading states with the largest observed daily normalized searches for hypoxemia. The second wave peaked in mid-July with three leading states, AZ, TX, and FL. The third wave peaked on 5 October and the top three leading states were HI, MT, and AZ. Overall, these trajectories identified the three waves of COVID-19 spread and some of the leading states in each wave. Figure 15 (right) shows the scatter plot of the 51 states in the two-dimensional space defined by the top FPC scores accounting for 76% of the variability in the curves. We noted that the four leading states in the first wave had the largest second FPC scores as well as positive first FPC scores. The leading states in the second wave were located in the fourth quadrant (positive FPC1 scores and negative FPC2 scores). The figure also shows four outlier states with negative FPC2 scores: VT, WY, AK, and ND. The trajectories of these four states as well as the mean curve for the 51 states are provided in Supplementary Figure S13. We observed that VT was a clear outlier as its curve was far below the mean curve during the time interval spanning mid-May to mid-September. The remaining three curves were the farthest curves below the mean curve from the study starting date until mid-September.

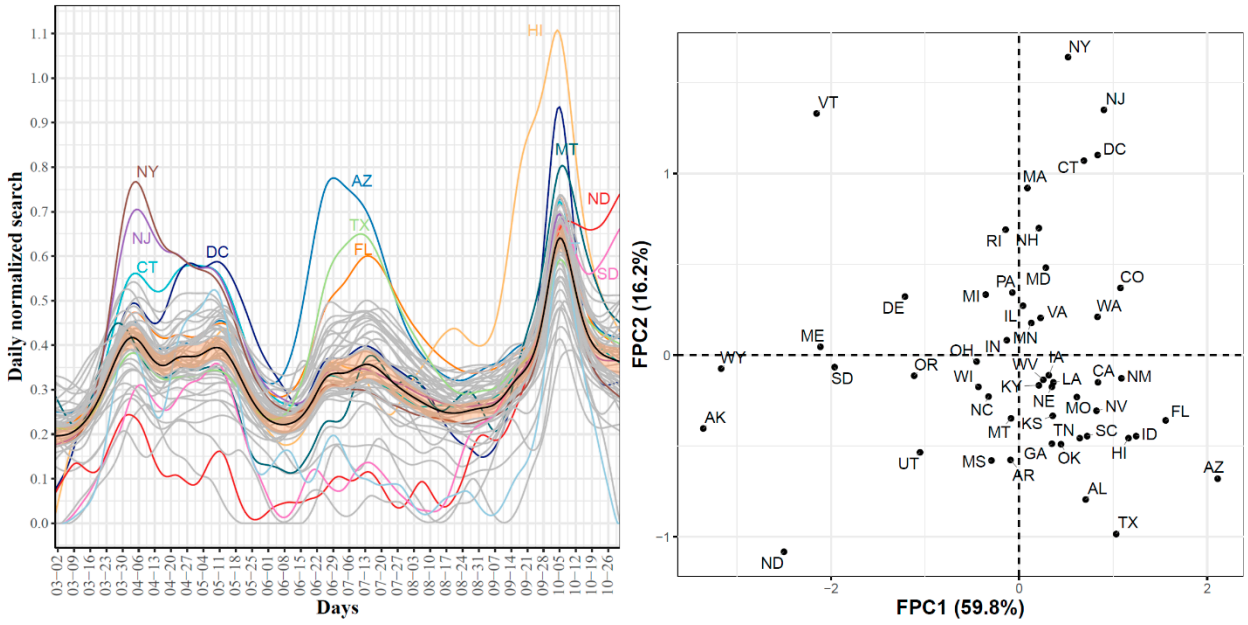

Figure 15. State-level trajectories of the Google daily searches for hypoxemia (**left**) and their projections into a two-dimensional space determined using the first two FPC scores (**right**). The mean curve is highlighted in black and the orange ribbon corresponds to the 95% confidence band.

Second, Figure 16 shows the first four eigenfunctions accounting for 91.1% of the total variability in the hypoxemia trajectories. Despite the overall high dynamic correlation coefficient between the trajectories of COVID-19 confirmed cases and Google searches for hypoxemia, we noted that the four eigenfunctions summarizing the modes of variations in the two datasets were different from those summarizing trajectories of the COVID-19 confirmed cases (see Figures 3 and 16). The same observation is also valid when comparing the scatter plots of the two datasets in the two-dimensional FPC scores (see Figures 4 and 15 (right)). To investigate the modes of variability in these two datasets, we report the results of applying the functional correlation analysis to the two datasets in the following paragraph.

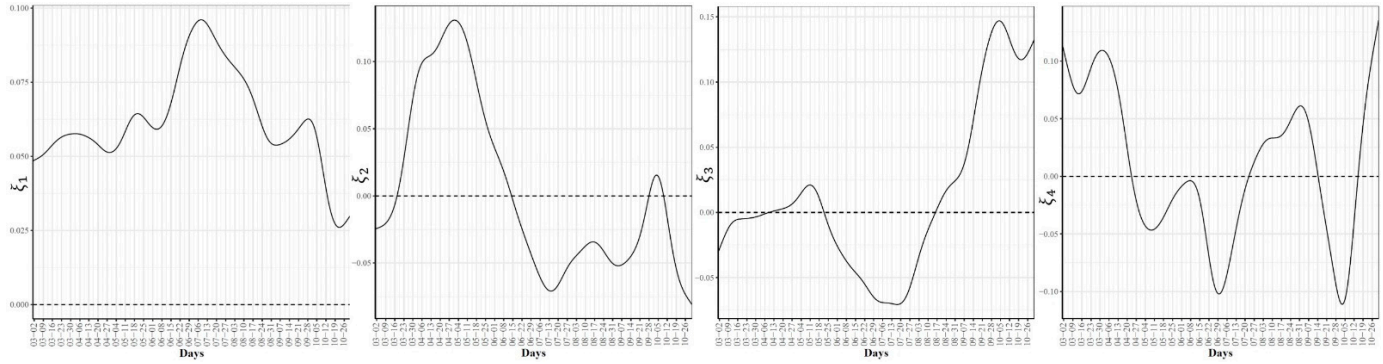

Figure 16. First four eigenfunctions of the state-level trajectories for the Google daily searches for hypoxemia.

Third, Figure 17 shows the two-dimensional scatter plots of the first and second canonical scores of the COVID-19 confirmed cases and Google searches for hypoxemia sets of curves, respectively. The correlations for the first and second canonicals were 0.99 and 0.98, respectively. The scatter plot of the 51 states in the two-dimensional space defined using the first canonical scores shows that the leading states in the second wave in the hypoxemia trajectories had the highest first canonical scores, while the leading states of in the third wave (except HI since it is a leading state in the hypoxemia trajectories but not in the trajectories of the COVID-19 confirmed cases) had the lowest first canonical scores. The scatter plot of the 51 states using their second canonical scores shows that ND and SD had the highest second canonical scores followed by the states in the second wave, while the leading states in the first wave had the lowest second canonical scores.

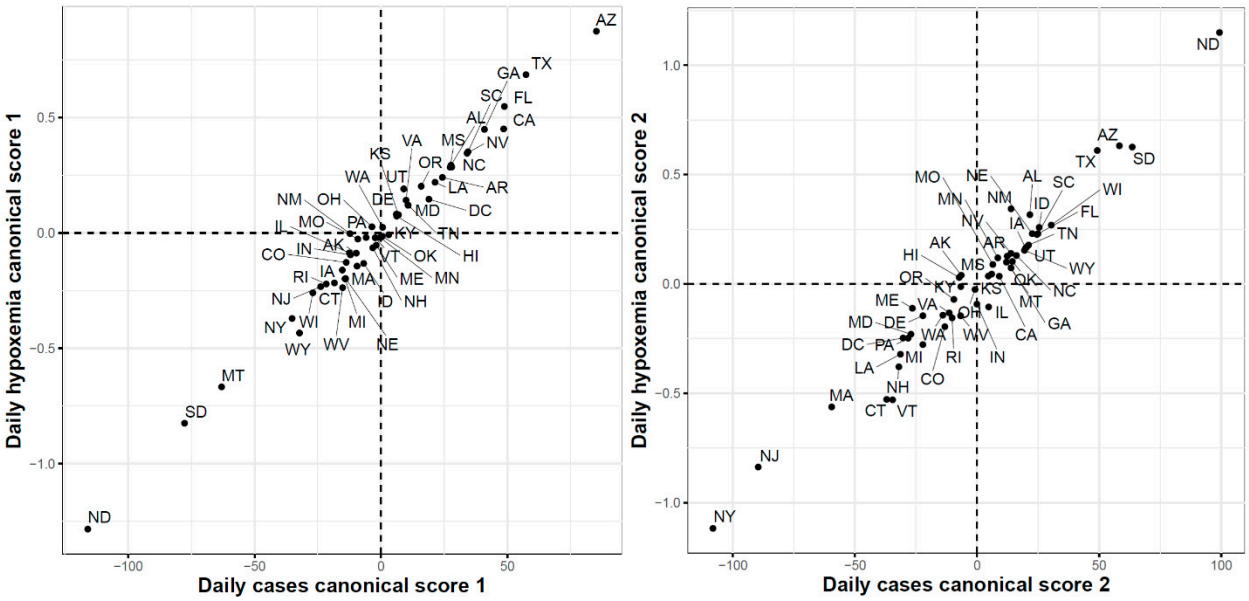

Figure 17. First (**left**) and second (**right**) functional canonical variables scores of the COVID-19 confirmed cases versus the Google daily searches for hypoxemia.

Finally, Figure 18 shows the results of the FCCA when applied to COVID-19 death cases and Google searches for hypoxemia functional datasets. The correlations for the first and second canonicals were 0.96 and 0.95, respectively. The scatter plot of the 51 states using their first canonical variables shows that the states with the highest canonical scores were the states with the largest numbers of reported death cases between March and June. Interestingly the three states with the largest numbers of death cases at the end of the study period (ND, MD, and SD) had the lowest second canonical scores.

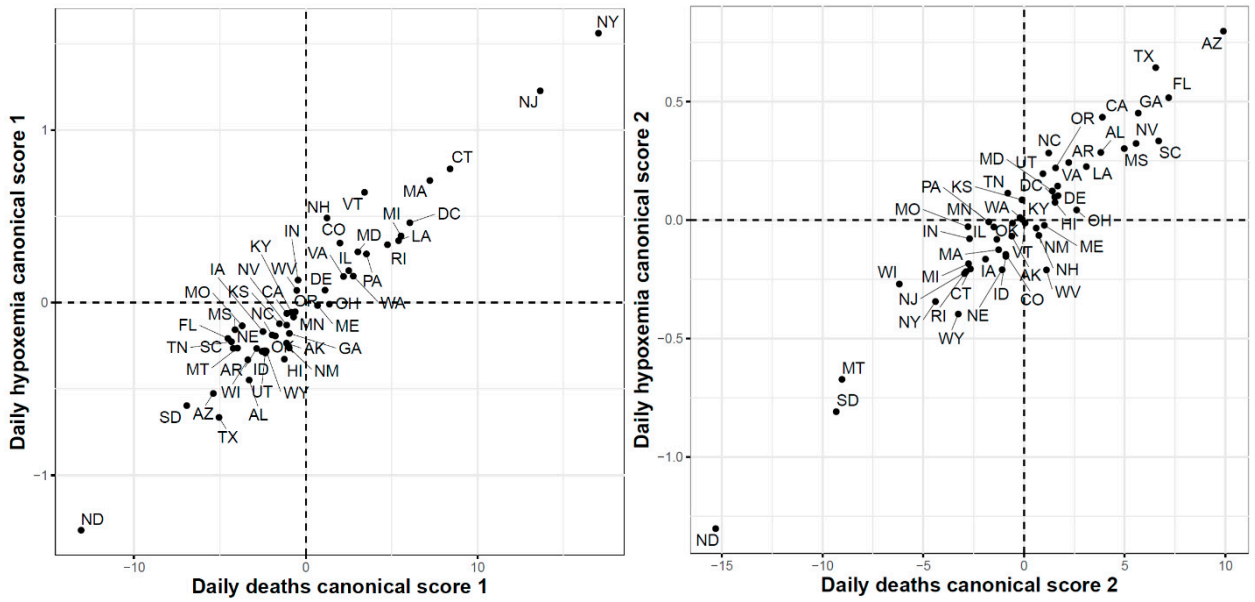

Figure 18. First (**left**) and second (**right**) functional canonical variables scores of the COVID-19 deaths versus. the Google daily searches for hypoxemia.

4. Discussion

In this study, we conducted functional data analysis of three time-series datasets: CDC COVID-19 daily confirmed cases; CDC COVID-19 daily death cases; and Google COVID-19 symptoms search trends. Particularly, we utilized different functional data analysis techniques (including functional principal component analysis, dynamic correlations, and functional canonical correlation) to analyze and categorize different patterns of the time-dynamics of COVID-19 confirmed and death cases and to identify associations between trajectories of Google COVID-19 symptoms search trends and COVID-19 trajectories of confirmed and death cases.

Analysis of the state-level trajectories of the data demonstrated variations in the time-dynamics of COVID-19 confirmed and death cases as well as Google search trends for COVID-19-related symptoms such as hypoxemia, ageusia, and anosmia. Visualization of the trajectories revealed three waves of the spread of COVID-19 in the United States. These three waves were noted not only in the COVID-19 confirmed case trajectories but also death trajectories and the trajectories of Google searches for symptoms related to COVID-19 infection.

Although the virus spread faster during the second and third waves, we noted that the death rates were the highest during the first wave of the pandemic. The reasons are not entirely known [38] but this observation has been justified (in part) by multiple factors including our improved medical knowledge on how to better treat COVID-19 patients, more preparedness in our health care system, some effects of the summer months, and the discovery of many mild or asymptomatic COVID-19 cases in later waves compared to the first wave when testing was often restricted to the sickest individuals. Other potential factors that have not been yet verified include: the virus became less deadly, vulnerable people became more protected (e.g., using social distancing and masks).

FPCA is a widely used and powerful tool in FDA. FPCA is critical for exploratory functional data analysis, dimensionality reduction, functional clustering, functional classification, and functional linear regression [39]. In this work, we found that the first two principal components can reliably identify the outlier states (i.e., the leading states in the three COVID-19 waves observed in the trajectories data). Moreover, we used functional eigenvalue analysis to explore modes of variations in the data. Finally, we used the leading four FPC scores to model each curve in a reduced dimensional space and applied the *k*-means algorithm to identify groups of states with common functional behavior. Our

cluster analysis results suggested that there were seven different patterns of the COVID-19 spread across the 51 states (see Figure 10). On the other hand, we found fewer variations across the 51 states when their death trajectories are clustered. Only five disjoint groups were identified and one of them includes a single state, NY.

All functional data analyses, discussed in the preceding paragraphs, are applicable to a single functional dataset. To explore the associations between two functional datasets (i.e., Google COVID-19 search trends for symptoms and COVID-19 confirmed/death cases), we utilized dynamic correlation and functional canonical correlation analysis. First, the dynamic correlation was used to quantify significant associations between each of the 422 symptoms functional datasets and COVID-19 confirmed/death cases functional datasets. Our results revealed two discrepancies between COVID-19 confirmed and death cases data related to their associations with Google COVID-19 search trends for symptoms: (i) Google COVID-19 search trends for symptoms had stronger correlations with COVID-19 confirmed data than COVID-19 death data; (ii) Dynamics of the associations between Google COVID-19 search trends for symptoms and COVID-19 confirmed cases were different from those of the associations between Google COVID-19 search trends for symptoms and COVID-19 death cases. Second, functional canonical correlation analysis (FCCA) was utilized to explore the relationships between Google search trends for hypoxemia and COVID-19 confirmed/death cases functional datasets. Our FCCA results acknowledged our results obtained using dynamic correlation analysis that suggests a stronger association of Google search trends for hypoxemia and COVID-19 confirmed cases than COVID-19 death cases obtained using dynamic correlation, and demonstrated the existence of common shared two-dimensional spaces between these pairs of functional datasets. Overall, our results demonstrated strong associations between Google search trends for several COVID-19-related symptoms and COVID-19 confirmed and death cases in the US.

The present study has some limitations. First, our clustering analyses have identified different patterns from the trajectories of COVID-19 confirmed and death cases in the US from the beginning of the pandemic until the end of October 2020. However, these patterns are dynamic and might change after extending the study time interval. Second, the numbers of daily positive COVID-19 cases reported during the first wave are underestimated due to insufficient testing in most of the states during the first months of the pandemic. Third, our analyses have focused on exploring the associations between Google search trends and COVID-19 spread and mortality trajectories while ignoring other potential factors such as mobility [40,41] and environmental factors [42–44] (e.g., temperature and humidity). Finally, variations in the spread of COVID-19 exist at the US county level [45]. Moreover, some government restrictions were initially implemented at the city or county level before extending to the state level [46]. Therefore, county-level analysis is a natural extension of this work.

5. Conclusions

We have conducted exploratory functional data analysis on daily state-level CDC COVID-19 confirmed cases and deaths time-series as well as selected Google search trends for symptoms related to COVID-19 infection. Our functional clustering results of the CDC COVID-19 spread trajectories assigned the 51 US states to seven disjoint groups, each with a distinct spread pattern. Clustering of the COVID-19 death trajectories identified five distinct patterns. We have also quantified the associations between Google search trends for 422 symptoms and CDC COVID-19 trajectories for confirmed as well as death cases using dynamic correlation. Moreover, we have explored the dynamics of associations between nine top Google search trends and future COVID-19 confirmed and death cases, respectively. Finally, we have applied exploratory functional data analysis to the Google search trends for three COVID-19-related symptoms (hypoxemia, anosmia, and shortness of breath) and quantified the associations between their trajectories and the trajectories of COVID-19 confirmed and death cases using functional canonical correlation analysis. Our results and analysis framework set the stage for the development of predictive models for

predicting future COVID-19 confirmed cases and deaths using historical data and Google search trends for identified symptoms, which is the focus of our ongoing work.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijerph18094560/s1, Additional file 1: Supplementary Tables S1: Dynamic correlations between Google trends for symptoms and COVID-19 confirmed cases, Table S2: Dynamic correlations between Google trends for symptoms and COVID-19 death cases, Additional file 2: Supplementary Figure S1: Characterizing the seven groups of COVID-19 spread patterns across the US states using their boxplots of four COVID-19 risk factors corresponding to the percentage of population that are, (a) African American, (b) aged 65+ years, (c) Unemployment, (d) below the poverty level, Figure S2: Characterizing the five groups of COVID-19 death patterns across the US states using their boxplots of four COVID-19 risk factors corre-sponding to the percentage of population that are, (a) African American, (b) aged 65+ years, (c) Unemployment, (d) below the poverty level, Figure S3: State-level trajectories of the Google daily searches for anosmia. The mean curve is highlighted in black. The mean curve is highlighted in black and the orange ribbon corresponds to the 95% confidence band, Figure S4: Projections of the anosmia trajectories into a two-dimensional space determined using the first two FPC scores, Figure S5: First four eigenfunctions of the state-level trajectories for the Google daily searches for anosmia, Figure S6: First (left) and second (right) functional canonical variables scores of the COVID-19 confirmed cases versus the Google daily searches for anosmia, Figure S7: First (left) and second (right) functional canonical variables scores of the COVID-19 deaths versus. the Google daily searches for anosmia, Figure S8: State-level trajectories of the Google daily searches for "shortness of breath". The mean curve is highlighted in black and the orange ribbon corresponds to the 95% confidence band, Figure S9: Projections of the "shortness of breath" trajectories into a two-dimensional space determined using the first two FPC scores, Figure S10: First four eigenfunctions of the state-level trajectories for the Google daily searches for "shortness of breath", Figure S11: First (left) and second (right) functional canonical variables scores of the COVID-19 confirmed cases versus the Google daily searches for "shortness of breath", Figure S12: First (left) and second (right) functional canonical variables scores of the COVID-19 deaths versus. the Google daily searches for "shortness of breath", Figure S13: Trajectories of the four outlier states with negative FPC2 scores. The black curve represents the mean curve for the 51 states.

Author Contributions: Y.E. designed and conceived the research. M.A. and Y.E. ran the experiments and analyzed the data. T.B.M. and E.S.H. contributed to the interpretation of the results and manuscript writing. Y.E. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Y.E. is supported by startup funding from the Geisinger Health System. The funder had no role in the design.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The numbers of daily COVID-19 confirmed cases and deaths were obtained from the Centers for Disease Control and Prevention (CDC) at https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36 (accessed on 6 November 2020). The Google COVID-19 Search Trends Symptoms dataset is publicly available at: https://github.com/google-research/open-covid-19-data/ (accessed on 6 November 2020). The 2019 US Census data are available at https://www.census.gov/data.html (accessed on 6 November 2020).

Conflicts of Interest: The authors declare that they have no competing interest.

References

1. McKibbin, W.; Fernando, R. The economic impact of COVID-19. In *Economics in the Time of COVID-19*; Centre for Economic Policy Research: London, UK, 2020.
2. Xiong, J.; Lipsitz, O.; Nasri, F.; Lui, L.M.W.; Gill, H.; Phan, L.; Chen-Li, D.; Iacobucci, M.; Ho, R.; Majeed, A.; et al. Impact of COVID-19 pandemic on mental health in the general population: A systematic review. *J. Affect. Disord.* **2020**, *277*, 55–64. [CrossRef] [PubMed]
3. Mann, D.M.; Chen, J.; Chunara, R.; Testa, P.A.; Nov, O. COVID-19 transforms health care through telemedicine: Evidence from the field. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1132–1135. [CrossRef] [PubMed]

4. Marinoni, G.; Van't Land, H.; Jensen, T. The Impact of Covid-19 on Higher Education around the World. *IAU Global Survey Report*. 2020. Available online: https://www.iau-aiu.net/IMG/pdf/iau_covid19_and_he_survey_report_final_may_2020.pdf (accessed on 23 November 2020).
5. Ferrel, M.N.; Ryan, J.J. The Impact of COVID-19 on Medical Education. *Cureus* **2020**, *12*, e7492. [CrossRef]
6. Aucejo, E.M.; French, J.; Araya, M.P.U.; Zafar, B. The impact of COVID-19 on student experiences and expectations: Evidence from a survey. *J. Public Econ.* **2020**, *191*, 104271. [CrossRef]
7. Tang, C.; Wang, T.; Zhang, P. Functional data analysis: An application to COVID-19 data in the United States. *arXiv* **2020**, arXiv:2009.08363.
8. Chen, J.; Yan, J.; Zhang, P. Clustering US States by Time Series of COVID-19 New Case Counts with Non-negative Matrix Factorization. *arXiv* **2020**, arXiv:2011.14412.
9. Carroll, C.; Bhattacharjee, S.; Chen, Y.; Dubey, P.; Fan, J.; Gajardo, Á.; Zhou, X.; Müller, H.-G.; Wang, J.-L. Time dynamics of COVID-19. *Sci. Rep.* **2020**, *10*, 1–14. [CrossRef]
10. Boschi, T.; Di Iorio, J.; Testa, L.; Cremona, M.A.; Chiaromonte, F. The shapes of an epidemic: Using Functional Data Analysis to characterize COVID-19 in Italy. *arXiv* **2020**, arXiv:2008.04700.
11. Cremona, M.A.; Chiaromonte, F. Probabilistic K-mean with local alignment for clustering and motif discovery in functional data. *arXiv* **2018**, arXiv:1808.04773.
12. Marron, J.S.; Ramsey, J.O.; Silverman, B.W. Functional Data Analysis. *J. Am. Stat. Assoc.* **1998**, *93*, 1232. [CrossRef]
13. Bavadekar, S.; Dai, A.; Davis, J.; Desfontaines, D.; Eckstein, I.; Everett, K.; Fabrikant, A.; Flores, G.; Gabrilovich, E.; Gadepalli, K. Google COVID-19 Search Trends Symptoms Dataset: Anonymization Process Description (version 1.0). *arXiv* **2020**, arXiv:2009.01265.
14. Preis, T.; Moat, H.S. Adaptive nowcasting of influenza outbreaks using Google searches. *R. Soc. Open Sci.* **2014**, *1*, 140095. [CrossRef]
15. Cook, S.; Conrad, C.; Fowlkes, A.L.; Mohebbi, M.H. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE* **2011**, *6*, e23610. [CrossRef]
16. Santillana, M.; Zhang, D.W.; Althouse, B.M.; Ayers, J.W. What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends? *Am. J. Prev. Med.* **2014**, *47*, 341–347. [CrossRef] [PubMed]
17. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 1012–1014. [CrossRef] [PubMed]
18. Kandula, S.; Shaman, J. Reappraising the utility of Google Flu Trends. *PLoS Comput. Biol.* **2019**, *15*, e1007258. [CrossRef] [PubMed]
19. Dukic, V.; Lopes, H.F.; Polson, N.G. Tracking Epidemics with Google Flu Trends Data and a State-Space SEIR Model. *J. Am. Stat. Assoc.* **2012**, *107*, 1410–1426. [CrossRef]
20. Pervaiz, F.; Pervaiz, M.; Rehman, N.A.; Saif, U. FluBreaks: Early Epidemic Detection from Google Flu Trends. *J. Med. Internet Res.* **2012**, *14*, e125. [CrossRef] [PubMed]
21. Zhang, Y.; Milinovich, G.; Xu, Z.; Bambrick, H.; Mengersen, K.; Tong, S.; Hu, W. Monitoring Pertussis Infections Using Internet Search Queries. *Sci. Rep.* **2017**, *7*, 10437. [CrossRef] [PubMed]
22. Carethers, J.M. Insights into disparities observed with COVID-19. *J. Intern. Med.* **2021**, *289*, 463–473. [CrossRef]
23. Snowden, L.R.; Graaf, G. COVID-19, Social Determinants Past, Present, and Future, and African Americans' Health. *J. Racial Ethn. Health Disparities* **2021**, *8*, 12–20. [CrossRef]
24. Abedi, V.; Olulana, O.; Avula, V.; Chaudhary, D.; Khan, A.; Shahjouei, S.; Li, J.; Zand, R. Racial, Economic, and Health Inequality and COVID-19 Infection in the United States. *J. Racial Ethn. Health Disparities* **2020**, 1–11. [CrossRef]
25. Silverman, B.W. Smoothed functional principal components analysis by choice of norm. *Ann. Stat.* **1996**, *24*, 1–24. [CrossRef]
26. Dubin, J.A.; Müller, H.-G. Dynamical Correlation for Multivariate Longitudinal Data. *J. Am. Stat. Assoc.* **2005**, *100*, 872–881. [CrossRef]
27. Liu, S.; Zhou, Y.; Palumbo, R.; Wang, J.-L. Dynamical correlation: A new method for quantifying synchrony with multivariate intensive longitudinal data. *Psychol. Methods* **2016**, *21*, 291–308. [CrossRef] [PubMed]
28. He, G.; Müller, H.-G.; Wang, J.-L. Methods of canonical analysis for functional data. *J. Stat. Plan. Inference* **2004**, *122*, 141–159. [CrossRef]
29. Rice, J.A.; Silverman, B.W. Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves. *J. R. Stat. Soc. Ser. B* **1991**, *53*, 233–243. [CrossRef]
30. Pezzulli, S.; Silverman, B. Some Properties of Smoothed Principal Components Analysis. *Comput. Stat.* **1993**, *8*, 1–16.
31. Jones, M.; Rice, J.A. Displaying the important features of large collections of similar curves. *Am. Stat.* **1992**, *46*, 140–145.
32. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Stat. Probab.* **1967**, *1*, 281–297.
33. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1988.
34. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A.; Charrad, M.M. Package 'nbclust'. *J. Stat. Softw.* **2014**, *61*, 1–36.
35. Dubin, J.; Li, M.; Qiao, D.; Müller, H.-G. DynCorr: Dynamic Correlation Package (Version 1.1.0). 2017. Available online: https://cran.r-project.org/web/packages/dynCorr/index.html (accessed on 21 December 2020).
36. Leurgans, S.E.; Moyeed, R.A.; Silverman, B.W. Canonical Correlation Analysis When the Data are Curves. *J. R. Stat. Soc. Ser. B* **1993**, *55*, 725–740. [CrossRef]

37. Ramsay, J.O.; Silverman, B.W. *Applied Functional Data Analysis: Methods and Case Studies*; Springer: Berlin/Heidelberg, Germany, 2007.
38. Ledford, H. Why do COVID death rates seem to be falling? *Nat. Cell Biol.* **2020**, *587*, 190–192. [CrossRef] [PubMed]
39. Shang, H.L. A survey of functional principal component analysis. *AStA Adv. Stat. Anal.* **2014**, *98*, 121–142. [CrossRef]
40. Lee, M.; Zhao, J.; Sun, Q.; Pan, Y.; Zhou, W.; Xiong, C.; Zhang, L. Human mobility trends during the early stage of the COVID-19 pandemic in the United States. *PLoS ONE* **2020**, *15*, e0241468. [CrossRef]
41. Kraemer, M.U.G.; Yang, C.-H.; Gutierrez, B.; Wu, C.-H.; Klein, B.; Pigott, D.M.; Du Plessis, L.; Faria, N.R.; Li, R.; Hanage, W.P.; et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **2020**, *368*, 493–497. [CrossRef] [PubMed]
42. Qu, G.; Li, X.; Hu, L.; Jiang, G. *An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19)*; American Chemical Society (ACS): Washington, DC, USA, 2020; Volume 54, pp. 3730–3732.
43. Shakil, M.H.; Munim, Z.H.; Tasnia, M.; Sarowar, S. COVID-19 and the environment: A critical review and research agenda. *Sci. Total Environ.* **2020**, *745*, 141022. [CrossRef]
44. Scafetta, N. Distribution of the SARS-CoV-2 pandemic and its monthly forecast based on seasonal climate patterns. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3493. [CrossRef]
45. Ives, A.R.; Bozzuto, C. Estimating and explaining the spread of COVID-19 at the county level in the USA. *Commun. Biol.* **2021**, *4*, 1–9. [CrossRef]
46. White, E.R.; Hébert-Dufresne, L. State-level variation of initial COVID-19 dynamics in the United States. *PLoS ONE* **2020**, *15*, e0240648. [CrossRef]