Lecture Notes in Networks and Systems 1094

Michel Kadoch
Kejie Lu
Feng Ye
Yi Qian *Editors*

Proceedings of the International Symposium on Intelligent Computing and Networking 2024

(ISICN 2024)

Springer

A Novel ML Method for Temporal Evolution of Geographic Clusters of Disease Spread Patterns

Will Casey[1(✉)], Leigh Metcalf[2], Heeralal Janwa[3], Shirshendu Chatterjee[4], and Ernest Battifarano[5]

[1] US Naval Academy, Annapolis, MD, USA
wcasey@usna.edu
[2] Carnegie Mellon University, Pittsburgh, PA, USA
lmetcalf@andrew.cmu.edu
[3] The University of Puerto, Rico San Juan, USA
heeralal.janwa@upr.edu
[4] City University of New York, New York, NY, USA
shirshendu@ccny.cuny.edu
[5] New York, NY, USA

Abstract. We aim to carry out effective clustering of geographically distributed time series data. This work uses singular value decomposition (SVD) for dimension reduction and spectral clustering. This model approach provides insight that there may be a common structure to disease spread, which contradicts early common narratives about how SARS-CoV-2 spread. In particular, our model indicates greater normalized rates of spread in rural areas, rather than the urban centers and transportation hubs that were widely thought to have higher rates of infection during the pandemic. Our novel method provides a temporal evolution of clusters derived from sliding window SVDs via spectral clustering. When applied to SARS-CoV-2 case load data, we determine geo-temporal patterns of disease spread. The method can be a tool to help an expert discern and interpret various patterns of disease spread. We present applications to the disease datasets from the US counties and Italian regions.

Keywords: time series · geographical clustering · machine learning · singular value decomposition · sliding window · temporal evolution

1 Introduction

Longitudinal studies of SARS-CoV-2 spread have already shown a geographical correlation along various factors such as rural and urban [7]. Here, we extend the geographic correlation studies by analyzing the modalities of the SARS-CoV-2 outbreak dynamics among various population areas of the United States. Our method considers the temporal continuity of the clustering results for 3,133 US

M. Kadoch et al. (Eds.): ISICN 2024, LNNS 1094, pp. 149–164, 2024.
https://doi.org/10.1007/978-3-031-67447-1_11

counties based on their recent changes in the cumulative percentage of populations infected. The results of our method reinforce previous longitudinal studies, indicate a limited set of SARS-CoV-2 outbreak modalities, and most excitingly offer insights into dynamic events we term *mode reselection events* where counties realign their cluster identity or switch dynamic mode for SARS-CoV-2 spread. Our findings of a persistent set of three distinct spread modes are interesting from a theoretical and practical point of view. Natural descriptive factors for pathogen spread include transmissibility, population density, and social encounter odds; the last factor is partially modulated by social behavior and policy, which switched at differing times in various counties. The analysis indicates a trade-off in these factors, a finding that could be a useful tool for policy formation.

We further hypothesize that the key events to mode re-selection events are described by and could possibly be predicted by social events such as policy shifts or changing social narratives and behaviors within a population.

Our clustering method will perform a singular value decomposition (SVD) on county data. The SARS-CoV-2 virus has been a dynamically evolving virus, mutating into multiple strains, some of which were resistant to vaccinations and previous infections in the US. In addition, attitudes, behaviors, and policies throughout the pandemic also proved adaptive. As a result, the dynamics of infection for SARS-CoV-2 is non-trivial. Nonetheless, we use a method combining singular values decomposition and spectral clustering to calculate a lower-dimensional representation of county data, and then apply K-means clustering to illuminate several primary modes of spread throughout counties of the United States. Our method focuses on a sliding window over time to see the local effects of the virus, rather than a global evaluation.

2 Infection Modeling Background

The SARS-CoV-2 virus spread geographically on a scale not seen before. The virus shut down most of the world, halting manufacturing and businesses in a way that affected not just local areas or countries, but the entire world. In summary, the epidemic infection dynamics brought attention to the difficulties and limitations of standard mathematical modeling[3,6].

Our aim is to contribute to better modeling of the pandemic spread, in particular when considering time series that are spread across multiple geographic locations. The SARS-CoV-2 virus spread in ways that were unknown in the last great global epidemic, the Spanish Flu [18]. Mass transit was, if not in its infancy, in its childhood in 1917. The number of annual passenger miles today dwarfs the number of annual passenger miles at that time and some passengers transport the virus. [While mass transit and media were a new concept at that time, the scale of what is available today far exceeds the mechanism to spread the virus at that time.]

3 Method

Clustering of time series [1] encompasses problems involving the study of multiple time series data sets. The Single Value Decomposition (SVD) has been used in prior work for clustering time series problems, most notably in the context of text data [8] and databases [10,14]. Additionally, SVD has also been used in the analysis of the time evolution of genes [9] and stock prices [19]. The SVD has also been used in myriad other data science applications, such as the classification of feature-rich data sets and as a dimensional reduction technique for data [5].

Here, we propose a modified view to analyze multiple time series data using SVD. In particular, we refine the use of SVD to windowed data-segments to increase sensitivity to small changes detectable at smaller time scales, enabling data-driven insights into time-localized patterns occurring at critical points in time. In particular, we are motivated by the challenge to determine subtle features in data showing how the spread of SARS-CoV-2 actually operated. As such, we are interested in analysis tools that help to confirm or refute commonly held notions of how the pandemic unfolded. These types of questions remain important for policy decisions. As such, the analysis of multiple time series such as geospatial Covid case loads seems to be an important and timely topic for data scientific methods. We construct SVD methods to cluster localities by Covid-19 infection behaviors and compare the global SVD clusters to clusters obtained in smaller timescale sliding windows of data [11]. Our methodology is developed and results are discussed. Our method discovered patterns consistent with many of the widely held views of the Covid-19 pandemic but are obtained through the novel methods presented here.

3.1 Data

The data required for this method is a SARS-Cov-2 case count by region. Each region is the source of time-series, namely the case count per day during the year 2020.

Starting from the New York Times Coronavirus Data set [17] that records case counts by day, we consider all US counties with 2019 census population estimate [4] and extract cumulative case counts for the entire SARS-CoV-2 data set. Thus, for each county, we have a time series in the form of a count function. Our data stretches over more than 900 consecutive days that record the total number of cases reported up to each day, however, most of the analysis presented here will be presented for the year of 2020.

It is important to note that we are considering cumulative cases as our data set. The cumulative cases clearly illustrate the trend in diagnosed cases, as shown in Fig. 1.

To address a few irregularities in the data, such as irregular reporting by some counties, we applied a kernel smoothing method that smooths counts over two-week periods. This is performed by using a symmetric triangular-shaped convolution kernel with two-week support. The results are simply a smoother version of the original daily count data; smoothing sub-week variability and

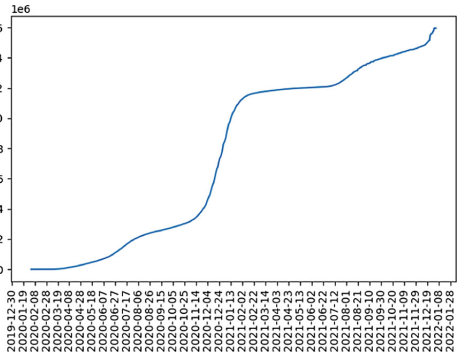


Fig. 1. Los Angeles SARS-CoV-2 Cases, Above the cumulative case numbers are plotted as a function of date starting Dec 30, 2019 through Jan 28, 2022. We observe multiple surges that correspond to different viral strains active at different times. The question arises: Will other geographic locations experience similar dynamic conditions, and what mathematical techniques can be leveraged to cluster the localities into groups that have similar dynamics?

irregularities (i.e. mostly delayed reports and sometimes corrections to reports). The smooth data does not differ qualitatively from the original data, but resolves some of the bigger issues with data regularity. We randomly spot-checked county instances for quality and results of the data smoothing process. For each county, the smoothed count data is then normalized by the 2019 population estimate. After these steps, we refer to our data as X having m rows (one for each county) and n columns (one for each sample date). The values X_{ij} represent the fraction of county i which had been infected (once or more) at time j. Note also that the 2019 population estimate data will not take into account births and deaths and migration occurring during 2020. Further the Covid case count data may include multiple counts for individuals that were re-infected, it may also exclude counts for individuals that were infected but didn't report their infection. Notwithstanding these discrepancies, due to ignoring reinfection, deaths, births, and potential erroneous counts, the normalized data will be assumed to be a monotonically increasing fraction of the population who had reported SARS-CoV-2 up to each day.

As locations with different population densities behave differently, we also consider an alternate analysis performed by multiplying the infections by the local density. This allows us to know not only the percentage of the population as a whole but also the density of SARS-CoV-2 infections.

3.2 Algorithm

Once we have our basic data set, the fractional counts are then stored in a data matrix, X, having m rows and n columns where m is the number of counties in the data and n is the number of days that the data set covers. All entries are

fractional values. Each row represents one of the m counties, and each column represents one of the n days of SARS-CoV-2 case data.

Algorithm 1. The Clustering Algorithm

Let X be the data matrix.

1. Compute the SVD of $X_{m\times n} = U_{m\times m}S_{m\times n}(V^T)_{n\times n}$
2. Find the elbow in the set of singular values S, (we use the *kneed method[16]*) to capture the best low-rank approximation in the sense of best value (law of diminishing returns). This process determines the number of dominant singular values as η from S. We denote $s = [s_1, s_2, \cdots, s_\eta]$, ranked from the largest to the smallest.
3. Obtain submatrices $\tilde{U}$ and $\tilde{V}$ from U and V by choosing the first η many columns.
4. Now using s, create a low rank estimate of X, labeled $W_{m\times n} = \tilde{U}_{m\times\eta}(\text{diag}(s))_{\eta\times\eta}(\tilde{V}^T)_{\eta\times n}$. W is the best rank η approximation of X
5. Using the K-means clustering on rows of $Y_{m\times\eta} = \tilde{U}_{m\times\eta}(\text{diag}(s))_{\eta\times\eta}$. The rows of the matrix Y represent the coefficients of linear approximation over the time trends (associated with rows of V^T).
6. Test the *sum of squared errors* (SSE) for K-means. For k clusters, with k ranging from 1 to η, determine the SSE. Treating the SSE as a function of cluster size, select the best k by employing the elbow finding algorithm (the law of diminishing returns).

Matrix X is a complete, but unwieldy, representation of the infected population. With this representation of the data, it is difficult to see trends. Too many trees are clouding our view of the forest. The SVD yields a more succinct, lower-dimensional, representation of the data that lends itself to identifying trends.

Using the language of linear algebra, matrices U and V^T are unitary and therefore affect rotations of the original coordinate system. The diagonal matrix D affects dilation and contraction in the new coordinate system. Dimension reduction is achieved by considering only the largest singular values and ignoring the smallest singular values. We use *kneed* to determine which singular values to keep.

Algorithm 1 proceeds by using the matrix X as input, it then calculates its singular value decomposition (SVD). The singular value decomposition (SVD) has several properties that make it an ideal tool for data analysis and dimensional reduction. First, the partial sequence of the first k singular values can be used to form the best rank k approximation of the original data matrix X (Eckart-Young Theorem). Step two of the algorithm 1, evaluates the decay of singular values. Assuming the decay function is convex, an elbow finding method, can determine a set of singular values, or η singular values. These can used to reassemble the best rank η approximation of data in step 3. Note that the rows of V^T represent the principle traces. These are a set of data series that most effectively describe the rows of X. Next, in step four, with η established, we calculate the loading coefficients for each row of data (county) when expressed in the basis of principle

traces. The loading coefficients are Y. Our final step is to perform $K-$means on the loading coefficients as a direct method to determine cultures.

Example: To illustrate these steps of the SVD we present an example from the singular values of the SARS-Cov-2 data set for LA county 1. The singular values are shown in Fig. 2. Next, in step 2, we consider the convex decay of singular values and determine a cutoff of 11 singular values. In step 3, we need only consider 11 rows of each U, V^T to form a matrix $W_{m\times 11} = U_{m\times 11}(\text{diag}(s))_{11\times 11}(V^T)_{11\times n}$ as the best rank-11 approximation of X.

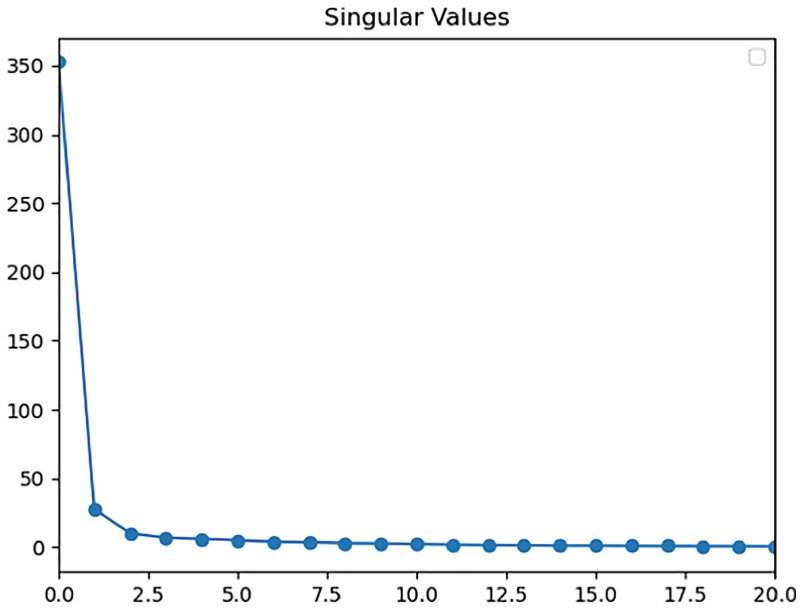


Fig. 2. Diminishing returns Calculation for singular values of SVD: The singular values decomposition is a powerful tool for data analysis. The decomposition has the property that the first k singular values can be used along with the first k columns of U and the first k rows of V^TY to determine the best rank k approximation of the original data matrix. Furthermore, the singular values are characterized by sharp convexity, indicating that the data is mostly summarized by a few inherent trends. The elbow of this graph, which we refer to as η, is a natural cutoff point for selecting the rank approximation of data, as selection of $k > \eta$ has diminishing returns.

We have reduced the fractional count data for each county to $\mathbb{R}^{11}$, a significantly lower dimension. Now we can use matrix slice notation, we let $W = U[:, : 11]S[: 11, : 11]$, a 3133×11 matrix. In W each row (county) has 11 entries that are multipliers for the characteristic fractional counts viewed in Fig. 3 allowing the best approximation of the county's actual data. We validate our work by calculating the max entry of $X - WV'[: 11, :]$ and find that no

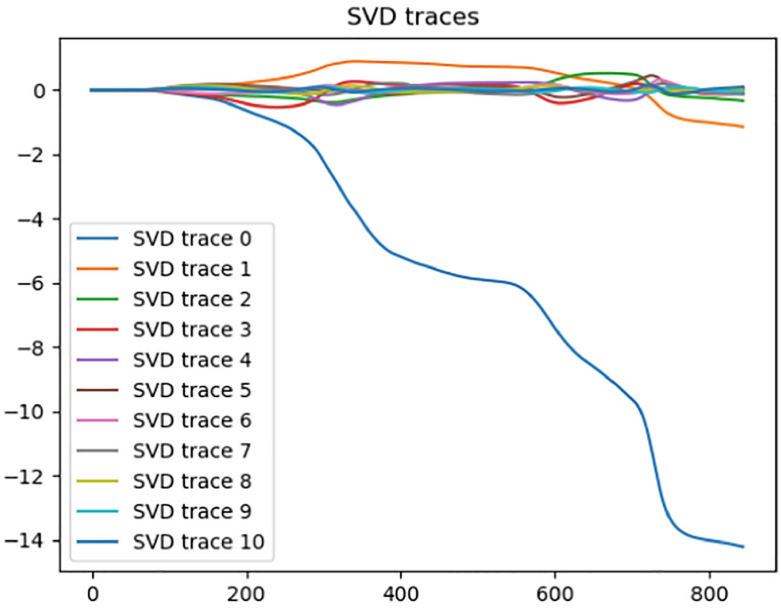


Fig. 3. Characteristic temporal traces (of fractional counts), Here we plot the first 11 rows of V'. The SVD provides a re-description for the rows of X, showing how each row can be written as a linear combination of characteristic traces. SVD also optimizes the representation in such a way as to describe the best low-rank approximation of X. For example, the best rank 11 approximation of data would re-describe each row of X as a linear combination of the traces above. Furthermore, the rank 11 approximation is best in that it minimizes the residual error (over all possible rank 11 approximations).

fractional count is off by more than 0.0485. Thus, the low-rank approximation is a good approximation of the original matrix X.

From the algorithm, we perform a K-means clustering over its row space of W. We analyze the SSE over various values of K, noting the convex shape of SSE's relation to the number of clusters, we refine an elbow detection method to determine the number of clusters η. See, for example, Fig. 4 where $K = 3$ is selected by the elbow method as an efficient choice for the number of clusters.

The clustering results in groups of localities. For two localities within a cluster, grouping the pandemic behavior is similar. For two localities in different groups, the clustering determines that the pandemic behavior has differing aspects, perhaps spread is affected by different population densities or differing policies aimed to control spread. As such, our method enables the identification of how the pandemic affected different areas in different ways, an important modeling consideration [2].

At the completion of the Windowing Algorithm, a sequence of partitions $S_1, S_2, \cdots, S_{c-W}$ which cluster the localities at each time point is calculated.

Fig. 4. Determining Number of Clusters: Similar to selecting η, the number of singular-values, we again use the law of diminishing returns to determine the number of clusters.

Here S_k refers to the partition (i.e., a collection of sets that are mutually distinct and collectively exhaustive) of the localities that were calculated for the data window k. Additionally, the centroids for each cluster can be recovered from the stored results and retrieved for analysis. Cluster centroids are a temporal data representation for all localities within the cluster. The centroid represents the best L_2 approximation over the class, as such it is similar to the centroid of an object in Newtonian physics that can be used as a single point approximation for the object itself. Similarly, when we develop K clusters from K-means, each cluster has a center position to represent all members of its grouping.

Algorithm 2. The Windowing Algorithm

W is the window size
$c \leftarrow$ number of columns in$X$.
for $t \in \{0, 1, \cdots, c - W\}$:

1. let $X_t \leftarrow X[:, t : (t + W)]$
2. Apply Algorithm 1 to X_t.
3. Store the K-means clustering results as S_t

4 Results and Discussion

Here, we apply our method in two case studies involving the outbreak of the SARS-CoV-2 virus and Covid-19 pandemic. The case studies examine data reporting the spread of case counts for SARS-CoV-2, the data is reported at local scales, in the form of:

- Case I: Covid caseload data from all counties of the United States during the year 2020,
- Case II: Covid caseload data from all regions of Italy during the year 2020.

We present results and discuss how our method can identify similar acting dynamic conditions in various localities. Additionally, we discuss how our clustering identifies a grouping of localities that are consistent with widely held expert opinions on the Pandemic outbreak dynamics. These results indicate that our method could prove a useful source of insights for a variety of complex geospatial temporal processes.

Case I. For the United States, our method routinely determined three main clusters throughout 2020. The cluster centroids (Fig. 8) tell a story of three differing rates of spread, summarized by low/medium/high rate of spread. **Data:** The original data are cumulative case counts for the year of 2020. To derive the fractional count for each county, the 2019 county population is used to normalize the cumulative count for each county. Again, the data has these limitations: reinfection is potentially multiply counted, non-reported cases are not counted, births and deaths occurring during 2020 are not accounted for when we normalize 2020 data with 2019 population totals.

The High Infection Rate Cluster: A cluster with the highest overall rate of spread is consistently observed. This class is comprised of mostly rural counties having a geographic concentration in the US south. The urban county of Miami-Dade County (in Florida) is included in this class (labeled class 2 and colored yellow in Fig. 7). Although this may seem somewhat surprising given that the virus was first witnessed in urban centers, many experts believed that the rural counties were impacted particularly badly due to older populations that had less access to health care resources such as intensive care units.

The Middle Infection Rate Cluster: Another cluster, which is consistently identified, features a medium rate of spread. This class includes some rural and some urban areas. This class includes the urban county of Las Vegas Clark County and is labeled as class 0 and colored purple in Fig. 7.

The Low Infection Rate Cluster: The lowest rate of spread is observed in a cluster concentrated in coastal areas, except the US South. Seattle King County is included in this cluster labeled class 1 and colored green in Fig. 7.

Furthermore, the classes differ dynamically at 375 d and 620 past the date of the first SARS-CoV-2 infection in the US (viewed as crossings in the load graph in Fig. 8-bottom). Finally, the geographical distribution of the three classes is given in Fig. 7.

In Fig. 4, The cluster centroids are plotted. The centroids can be thought of as the center average trend among all members of the cluster. As is illustrated, the centroids are very much alike for all three clusters, but differ by scale (or average slope). In Fig. 8 the cluster centroids coefficients are plotted over the first 10 characteristic traces (rows of V'). From this plot it appears that scale is a key differing factor, however it can also be noted that the second characteristic trace coefficients differ as well and suggest a more subtle difference among the centroid functions.

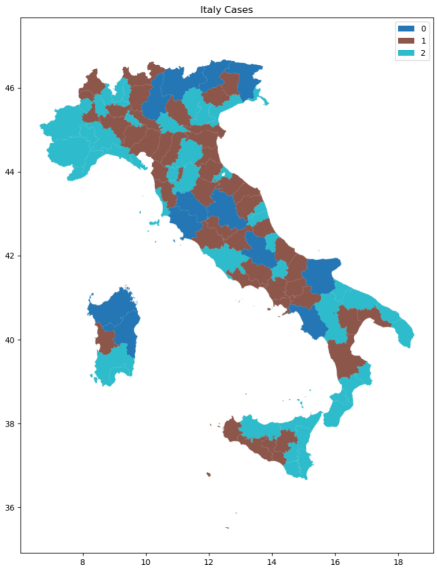


Fig. 5. SARS-CoV-2 Clusters in Italy: The analysis performed on cumulative case loads in each Italian regions, determines a clustering of regions into three descriptive trends viewed in Fig. 6, and perhaps best described as low/medium/high rates of infection.

Case II. Our second case considers local regional data in Italy during the 2020 Covid pandemic. Similarly, the same tri-banded infection rates seem to be the major feature captured by our model's evolving clusters. We determine that the Italian clusters, appearing in Fig. 5, and having cluster centroids as shown in Fig. 6 repeats the same story of three separate infection rates: low/medium/high.

Geographical Disease Progression in the Regions of Italy. Taking these images into account, it appears that the pandemic was consistent in certain provinces of Italy. However, our methods when we restrict our view to a sliding window of 30 d, reveal a slightly different result offering various retrospective insights.

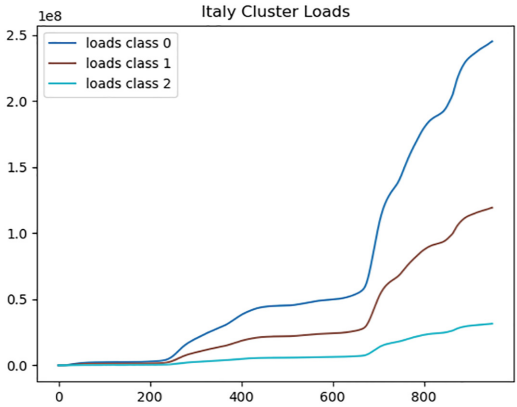


Fig. 6. SARS-CoV-2 Cluster Loads for Italy. Plotted above are the cluster centroids determined by K-means. Each centroid is a function (Cumulative number of cases by day). Further, it is a function that minimizes summed square error to the cumulative case counts of regions contained within their cluster. As can be seen, the centroids seem to capture low/medium/high rates of transmissions.

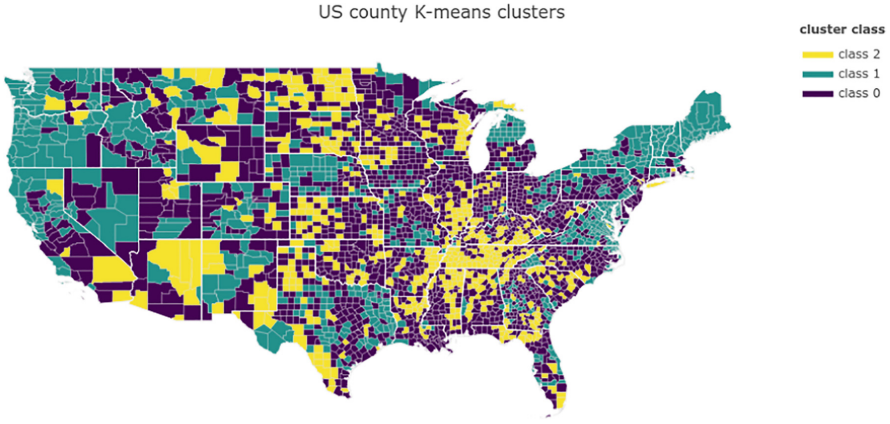


Fig. 7. US County Clustering Results tell a story of three dynamical modes: low/medium/hi rates of infections. In the graph above, all the US counties are clustered (shown by three different colors), counties having the same color are co-grouped with our method due to the similar characteristics of their dynamic case counts.

From the very beginning of the pandemic, as shown in Fig. 10a, the cases in Italy were mostly confined to northern Italy with the exception of Pesaro-Urbino. The early analysis of the pandemic discussed in [12] shows how it started in the north. Our analysis validates the results of [12], as well as brings into play the additional province in the east. The province of Pesaro-Urbino is part of the Marche region of Italy, which is known for manufacturing [13] and it is also

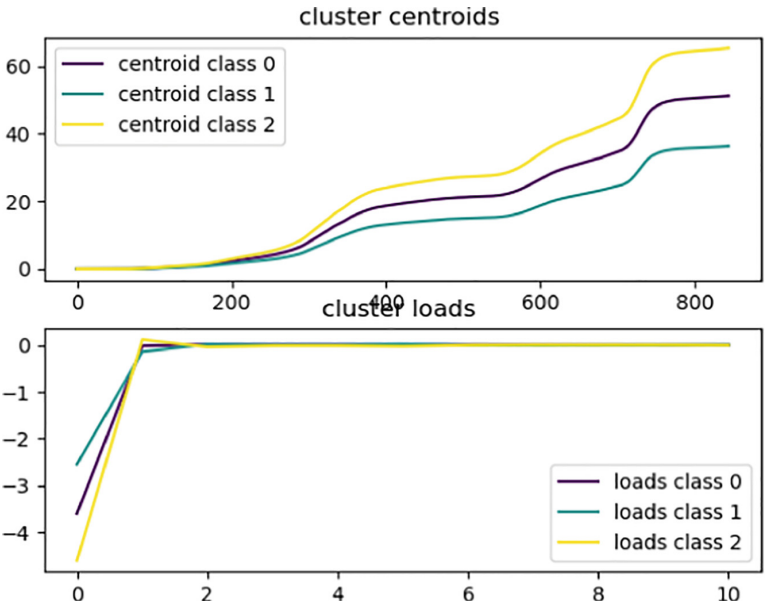


Fig. 8. US Covid infection dynamics are summarized by three centroid dynamic responses with our method. The result is a story of three separate spread rates: low/medium/high. The centroids are plotted as functions of time, and their coefficients for the first 10 characteristic traces is shown in the lower plot above.

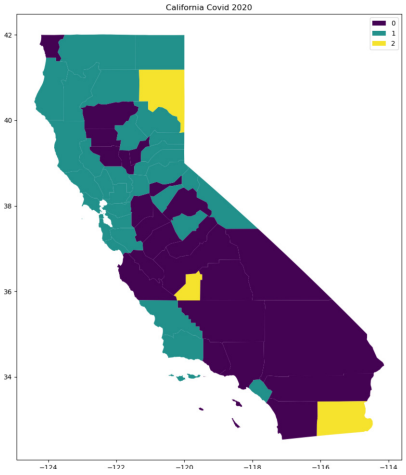


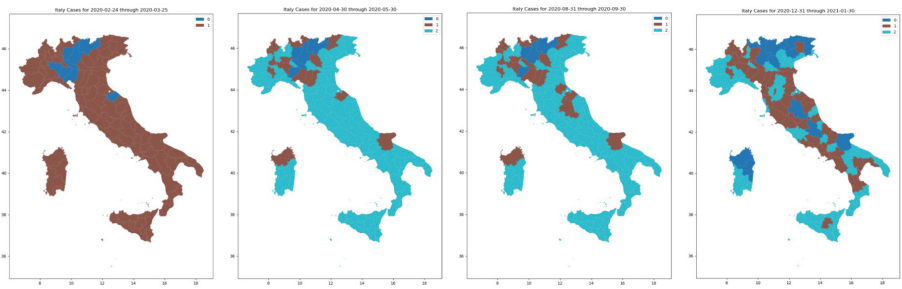
Fig. 9. California Case Clusters: Selecting the state of California from the Fig. 7 above can help to determine how the clusters can be interpreted when some geographical information is known. Note, major metropolitan or urban areas of California are clustered together, while areas that are more rural are likewise together but in a different cluster from the urban areas.

believed that the initial infections were possibly caused by commuters working in the factories in the north [15].

The pandemic did not remain in those locations; we can trace the spread in terms of severity in Figs. 10b, 10c and 10d. By Fig. 10d, we see that the northern provinces that were the originally infected areas and had been in the cluster with the highest spread had moved to a slower progression of the disease. These figures illustrate the time evolution of the severity of the cluster. Returning to Fig. 5, we note that our model did not detect that variance in the disease progression.

4.1 Temporal Evolution of Geographical Clusters

Next, we consider how clustering results can evolve as we slide the window. The visualization of the temporal evolution of the geographical clusters can be observed as gifs at https://github.com/austincasey/TSCSVD.



(a) (b) (c) (d)

Fig. 10. Evolution of similar SARS-CoV-2 dynamics for case count in Italy over 2020. Our methodology provides a means to determine how clusters evolve. Here, we present clusters of regions as calculated by our method at various time points. The clusters for progressively evolving data, that is data staring in Jan 2020 but including data up to: (a) 03/23/2020, (b) 05/30/2020, (c) 09/30/2020 and (d) 01/30/2021. We argue that clusters indicate a similar dynamics at the time they are measured. For example, early on there are only two modes identified, but within two more months, a third mode is observed and sustained through several more months. The last image in (d) corresponds to changes in the dynamic clusters due to changing conditions such as new strains and ongoing policy updates.

Temporal Evolution of the Disease of Progression in the US Counties: Learning from the case of Italy, we studied the temporal evolution of disease progression in the counties in the United States. We analyzed a 30-day sliding window temporal clustering. A snapshot of three clusters is given in Figs. 11a, 11b and 11a. We observed a dramatic evolution of the disease spread, from the transportation hubs to virtually all the counties in the USA. A time frame movie of this evolution yields a mesmerizing, and scary spread of the disease over three years.

(a) (b) (c)

Fig. 11. Evolution of similar dynamic spread patterns for SARS-CoV-2 case count across counties in the United States over 2020. The temporal evolution of clusters calculated from our method is visualized with data starting from Jan 2020 up to various dates: (a) May 30, 2020, (b) September 30, 2020 and (c) January 31, 2021. With superimposed geographical information, we could interpret these clusters as indicating a different spread dynamic in urban vs rural counties across the United States.

The Case of California. We focus on California as a local example. The spread of cases illustrated in Fig. 9 for 2020 hides the local spread. The spread and temporal evolution of the disease were similar to those of Italy.

(a) (b) (c)

Fig. 12. Evolution of similar dynamic spread patterns for SARS-CoV-2 case-count across counties in California over 2020. The temporal evolution of clusters, calculated from our method, are visualized with data starting from Jan 2020 up to various dates: (a) May 30, 2020, (b) September 30, 2020 and (c) January 31, 2021.

Figure 12a illustrates the initial high spread of SARS-CoV-2 cases in the Bay Area, particularly San Francisco. In Fig. 12b, the highest density is still San Francisco, but it has spread outward from that location and also includes Los Angeles. Finally, Fig. 12c shows the temporal evolution of changing clusters of severity, that the pandemic spread to more locations with higher rates, while the majority of the state maintained a low rate of spread.

4.2 Summary

Localized analysis of case-loads data is examined with our model for two case studies: US counties and Italian Regions. Our model creates clusters of localities

with similar dynamic processes, the cluster centroids provide an average caseload function for members of the cluster. In both case studies, our method consistently determines three main clusters, differing principally by infection rate. When using the window algorithm, fine details can be observed which are localized in time, as such a window of 30 d can reveal insights into the dynamic effects. Although anecdotal, we were able to link some of the observations from a 30-day window to some aspects of the pandemic in both case studies.

5 Conclusions and Future Work

When examining the methods and results for Covid-19 pandemic spread, it seems that the model is promising and the method may prove useful for more general data-driven geospatial temporal process questions. Confined to the case studies presented here, our method produced findings consistent with several prior studies focused on Covid-19 pandemic. More refinement is possible for our methods and is planned as future work. In particular, how to select time intervals in the windowing algorithm, and how window size can relate to timescale remain open questions.

By applying our method to case studies for Covid pandemic data, we were able to confirm common themes consistent with other Covid-19 studies, but by means of a data scientific method distinct from other methods used previously. Additionally, our method seems to reinforce the differential rates including the view that rural areas were hard hit in the US by the Covid pandemic. Our model also offers some variation or refutation to some of the common Covid Pandemic modeling, one example is found with the super-spreader in densely populated localities. This, in retrospect, appears not to have much support within our model. Our model also shows some features that differ from the popular narrative, that is, cities were not particularly critical to the narrative of the disease. Our model also suggests that the structural aspects of this disease can be modeled more effectively.

In future research, we would like to also explore more closely the relationship between clusters (as constructed by our method) and local policies. We hypothesize that for policies directly aimed at "flattening the curve," they could modify how various localities are clustered. In particular, local policy decisions by officials regarding mandates on distancing, masking, closures, and vaccinations, could potentially move a local county from one group into another.

We believe the results presented here warrant additional use and refinement for the clustering method. The method may also apply beyond pandemic modeling in areas such as cybersecurity, finance, and climate change. Most generally, the method could have utility for analyzing a variety of geospatial temporal processes, when data is accessible.

Acknowledgment. The research of SC has been supported partially by NSF grant DMS 2154564. The research of Heeralal Janwa was supported by the NASA grant 80NSSC22M0248 and also by the PR NASA Grant Consortium and its Director Dr. Gerardo Morell.

References

1. Aghabozorgi, S., Seyed Shirkhorshidi, A., Ying Wah, T.: Time-series clustering - a decade review. Inf. Syst. **53**, 16–38 (2015). https://doi.org/10.1016/j.is.2015.04.007, https://www.sciencedirect.com/science/article/pii/S0306437915000733
2. Ahammed, T., Anjum, A., Rahman, M.M., Haider, N., Kock, R., Uddin, M.J.: Estimation of novel coronavirus (covid-19) reproduction number and case fatality rate: a systematic review and meta-analysis. Health Sci. Rep. **4**(2), e274 (2021)
3. Bertozzi, A.L., Franco, E., Mohler, G., Short, M.B., Sledge, D.: The challenges of modeling and forecasting the spread of COVID-19. Proc. Natl. Acad. Sci. **117**(29), 16732–16738 (2020)
4. Bureau, USCC: County population totals: 2020-2021 (2022). https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html
5. Cassisi, C., Montalto, P., Aliotta, M., Cannata, A., Pulvirenti, A., et al.: Similarity measures and dimensionality reduction techniques for time series data mining. Adv. Data Mining Knowled. Discov. Appli., 71–96 (2012)
6. for Disease Control C, Prevention, Covid-19 forecasting and mathematical modeling (2022)
7. Economic Research Service, U.D o A: The covid-19 pandemic and rural america (2022). https://www.ers.usda.gov/covid-19/rural-america/
8. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time-series databases. ACM SIGMOD Rec. **23**(2), 419–429 (1994)
9. Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., Banavar, J.R.: Dynamic modeling of gene expression data. Proc. Nat. Acad. Sci. - PNAS **98**(4), 1693–1698 (2001)
10. Korn, F., Jagadish, H.V., Faloutsos, C.: Efficiently supporting ad hoc queries in large datasets of time sequences. ACM SIGMOD Rec. **26**(2), 289–300 (1997)
11. Li, X., Wang, S., Cai, Y.: Tutorial: Complexity analysis of singular value decomposition and its variants. arXiv preprint arXiv:1906.12085 (2019)
12. Megna, R.: First month of the epidemic caused by covid-19 in italy: current status and real-time outbreak development forecast. Global Health Research and Policy **5**(1), 43 (2020). https://doi.org/10.1186/s41256-020-00170-3
13. News, B.L.: The manufacturing is worth 98. https://www.breakinglatest.news/business/the-manufacturing-is-worth-98-of-the-exports-of-the-marche/
14. Ravi Kanth, K., Agrawal, D., Singh, A.: Dimensionality reduction for similarity searching in dynamic databases. ACM SIGMOD Rec. **27**(2), 166–176 (1998)
15. Rudan, I.: A cascade of causes that led to the COVID-19 tragedy in Italy and in other European union countries. J. Glob. Health **10**(1), 010335 (2020)
16. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a"kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 166–171. IEEE (2011)
17. Times, NY: The New York Times. (2021). coronavirus (covid-19) data in the united states. (2022). https://github.com/nytimes/covid-19-data
18. Tomes, N.: "Destroyer and teacher": Managing the masses during the 1918-1919 influenza pandemic. Public Health Rep. 125 Suppl 3(3_suppl), 48–62 (2010)
19. Xie, Y., Wulamu, A., Wang, Y., Liu, Z.: Implementation of time series data clustering based on svd for stock data analysis on hadoop platform. In: 2014 9th IEEE Conference on Industrial Electronics and Applicationsm pp. 2007–2010. IEEE (2014)