Time series classification of COVID-19 dynamics in the United States

Chris von Csefalvay*

August 1, 2020

Abstract

The statistical dynamics of a pathogen within a population depend on a range of factors: population density, the effectiveness and investment into social distancing, public policy measures and non-pharmaceutical interventions (NPIs) are only some examples of factors that influence the number of cases over time. This paper outlines a time series cluster analysis of confirmed COVID-19 cases in the United States, using a soft-DTW (Dynamic Time Warping) k-means clustering and a k-shape based clustering algorithm to identify internally consistent clusters of case counts over time. The identification of disjoint time series clusters can use past patterns to discern the future of infectious dynamics in an area, and through identifying the most likely cluster-wise trajectory, inform public health decision-making.

1 Introduction

The emergence of SARS-CoV-2, and its associated viral syndrome COVID-19, has raised important questions about the ways we analyse and identify dynamic temporal processes. In particular, by identifying similarities in principal time-dependent indicators of epidemic dynamics, such as prevalence (the number of confirmed cases over time), we can gain insight into similarities that are likely to emerge across various regions. Through this, time series clustering has the potential to play a significant role in understanding the dynamic processes that drive an outbreak.

Clustering is the wider set of algorithms within unsupervised learning that identify similar patterns among data in arbitrarily high-dimensional spaces, effectively taking a set $\mathcal{P}$ of N vectors in an n-dimensional space, and assigning to each of these a label from the set $\mathcal{L}$, so that the assignment of each element of $\mathcal{P}$ to the groups defined by the labels comprising $\mathcal{L}$ minimise some objective function (typically referred to as the distance metric of the clustering). Cluster algorithms are widely used today and their practical applications are

*Starschema Inc., Arlington, VA. Correspondence: csefalvayk@starschema.net.

ubiquitous, ranging from identifying clinical phenotypes in clinical medicine[1–5] through fraud detection[6–10] to image segmentation.[11–16]

Time series clustering presents a particular complication of this problem insofar as the subject of clustering is not a vector representing a single value, but rather a time series. These time series are typically not in synchrony, but rather exhibit a range of delays, lags and leads, and may depend on extrinsic hidden variables. We may formulate the essential task of time series clustering as follows. Let X comprise n time series $x_{1\ldots n}$, and let k denote the cardinality of the set $\mathcal{L}$ – in other words, the number of partitions we wish to split the data into, with $k \leq n$. Then, the mapping $f : X \to l \mid l \in \mathcal{L}$ is a clustering if it assigns to any element $x_i \in X \mid i \leq n$ one (and only one) cluster $l_i \in \mathcal{L}$, so as to minimise an objective function (typically referred to in this context as a distance metric) J within the cluster.

This paper examines the use of two time series clustering algorithms – soft-DTW k-means clustering and k-shape clustering – to identify different patterns in COVID-19 prevalence in the continental United States, and comparing the results of the classifiers for inter-classifier consistency. By isolating the barycenters of the time-shifted clusters, we can identify consistent patterns in prevalence dynamics across multiple states, quantifying the overall effect of pre-existing characteristics, population dynamics and non-pharmaceutical interventions (NPIs) between states.

2 Methods

2.1 Source data

Source data for the 48 states of the continental United States was obtained from the Starschema COVID-19 Data Set,[17] and filtered only for confirmed case counts. Data was loaded into Python 3.7 using `pandas`,[18] and values were scaled using `tslearn.preprocessing`'s `TimeSeriesScalerMeanVariance` to $\mu = 0$ and $\sigma = 1$. The results of this transformed raw data set are laid out, by state, in Figure 1.

2.2 Soft-DTW k-means clustering

Since first described by Sakoe and Chiba (1978),[19] the dynamic time warping algorithm has been expressed in multiple formulations. The presentation below is based on Cuturi and Blondel's 2017 paper introducing Soft-DTW, with the marginal difference of using $J(\cdot, \cdot)$ instead of δ to represent the distance function.[20]

Given two time series $x_t : t_x \in \mathbb{Z}$ and $y_t : t_y \in \mathbb{Z}$, there exists a cost matrix $\Delta(\mathbf{x}, \mathbf{y})$ for the distance function J, from which we can derive the cost matrix

$$\Delta(\mathbf{x}, \mathbf{y}) = [J(x_i, y_j)]_{ij} \in \mathbb{R}^{t_x \times t_y} \tag{1}$$

Figure 1: Scaled raw data of prevalence by state.

For the two above-mentioned series, we describe the set of matrices of all possible alignments as $\mathcal{A}_{t_x,t_y}$, which is a strict subset of $\{0,1\}^{t_x \times t_y}$. Then, DTW can be defined as the function that for any pair $(\mathbf{x}, \mathbf{y})$ identifies $A \in \mathcal{A}_{t_x,t_y}$ so as to minimise the inner product of A with the cost matrix $\Delta(\mathbf{x}, \mathbf{y})$.

$$DTW(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=\!=} \min_{A \in \mathcal{A}_{t_x,t_y}} \langle A_{t_x,t_y}, \Delta(\mathbf{x}, \mathbf{y}) \tag{2}$$

Thus, DTW can be conceived of as a search task, in which $\mathcal{A}_{t_x,t_y}$ is the search space within which we search for A so as to minimise $\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle$.

Soft-DTW universalises the notion underlying the DTW cost metric and the global alignment kernel metric

$$GAK_\gamma(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=\!=} \sum_{A \in \mathcal{A}_{t_x,t_y}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle}{\gamma}} \tag{3}$$

into a single metric.[21] Given the generalisation of the minimum metric with a smoothing factor $\gamma \geq 0$,

$$\overset{\gamma}{\min}\{a_{1\ldots n}\} \stackrel{\text{def}}{=\!=} \begin{cases} \min\limits_{i \leq n} a_i, & \gamma = 0 \\ -\gamma \log \sum\limits_{i=1}^{n} e^{-\frac{a_i}{\gamma}}, & \gamma > 0 \end{cases} \tag{4}$$

we may now define Soft-DTW as

$$sDTW_\gamma(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=\!=} \overset{\gamma}{\min_{A \in \mathcal{A}_{t_x,t_y}}} \{\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle\} \tag{5}$$

Importantly, Soft-DTW – unlike the original DTW approach by Sakoe and Chiba[19] – is explicitly differentiable. In particular, as Saigo (2006) noted,[22] the gradient of Equation (3) can be calculated quite conveniently. Let $\hat{A}$ be the average alignment matrix following the Boltzmann distribution $p_\gamma \sim e^{-\langle A_i, \frac{\Delta(\mathbf{x},\mathbf{y})\rangle}{\gamma}}$ for all $A_i \in \mathcal{A}_{t_x,t_y}$. Then,

$$\hat{A} = \frac{\sum\limits_{A_i \in \mathcal{A}_{t_x,t_y}} A_i e^{-\langle A, \frac{\Delta(\mathbf{x},\mathbf{y})}{\gamma}\rangle}}{GAK_\gamma(\mathbf{x}, \mathbf{y})} \tag{6}$$

and consequently

$$\nabla_{\mathbf{x}} DTW_\gamma(\mathbf{x}, \mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^T \hat{A} \tag{7}$$

This can be easily calculated using backward recursion, as described in Algorithm 2 of Cuturi and Blondel (2017).[20] In addition, the notion of a clustering centroid can be generalised to the metric space comprising the time series to

yield Frêchet means, also referred to in this context as barycenters. For a metric space (M, τ), $p \in M$ is a Frêchet mean of order $q \geq 1$ of the time series $x_{1\ldots n} \in M$ if it minimises the Frêchet variance, i.e.

$$p = \underset{r \in M}{argmin} \sum_{j=1}^{n} \tau(x_j, r)^q \tag{8}$$

Based on dynamic time warping distances between temporal signals, we can construct a clustering that divides the COVID-19 prevalence time series for the 48 states of the contiguous United States into a number of clusters so as to minimise distances using k-nearest neighbour clustering. Soft-DTW clustering was performed using `tslearn` 0.4.1 [23] using Python 3.7, with a γ parameter of 0.1.

2.3 k-shape clustering

k-shape clustering is a novel, robust clustering algorithm for time series that relies on iteratively refining clusters, with cross-correlation as the underlying distance metric.[24] Specifically, k-shape relies on a normalised version of cross-correlation, referred to in this context as Shape Base Distance (SBD): time series are Z-normalised (i.e. $\mu = 0$ and $\sigma = 1$), and the resulting cross-correlation sequence is divided by the geometric mean of the individual time series' auto-correlations. In this sense, k-shape can be understood as a k-means clustering that uses a cross-correlation based metric $SBD(\mathbf{x}, \mathbf{y})$. Let $mathbfx_s$ be the series $\mathbf{x}$ shifted, with zero-padding, by s, and the same be true for $mathbfy_s$ respectively, *mutatis mutandis*. For two time series of equal length $\mathbf{x}$ and $\mathbf{y}$, we recursively define shift-wise cross-correlation for shifts in the range $s \in [-m, m]$ as

$$\psi_k(\mathbf{x}, \mathbf{y}) = \begin{cases} \sum_{l=1}^{m-k} x_{l+k} y_l & k \geq 0 \\ \psi_{-k}(\mathbf{x}, \mathbf{y}) & k < 0 \end{cases} \tag{9}$$

Then, for the cross-correlation sequence, we obtain the cross-correlation $\hat{\rho}_w$ for any value of $w \in 1, 2, \ldots, 2m - 1$ as

$$\hat{\rho}_w(\mathbf{x}, \mathbf{y}) = \psi_{w-m}(\mathbf{x}, \mathbf{y}) \tag{10}$$

Now, we can define the distance metric $SBD(\mathbf{x}, \mathbf{y})$ by

$$SBD(\mathbf{x}, \mathbf{y}) = 1 - \max_w \left(\frac{\hat{\rho}(\mathbf{x}, \mathbf{y})}{\sqrt{\psi_0(\mathbf{x}, \mathbf{x}) \cdot \psi_0(\mathbf{y}, \mathbf{y})}} \right) \tag{11}$$

Because of the convolution theorem, which states that under certain conditions convolution in one domain of a time series (or more generally, any signal) is equivalent to elementwise multiplication in the other domain,[25] we can efficiently compute $\psi(\mathbf{x}, \mathbf{y})$ by taking the complex conjugate of the discrete Fourier

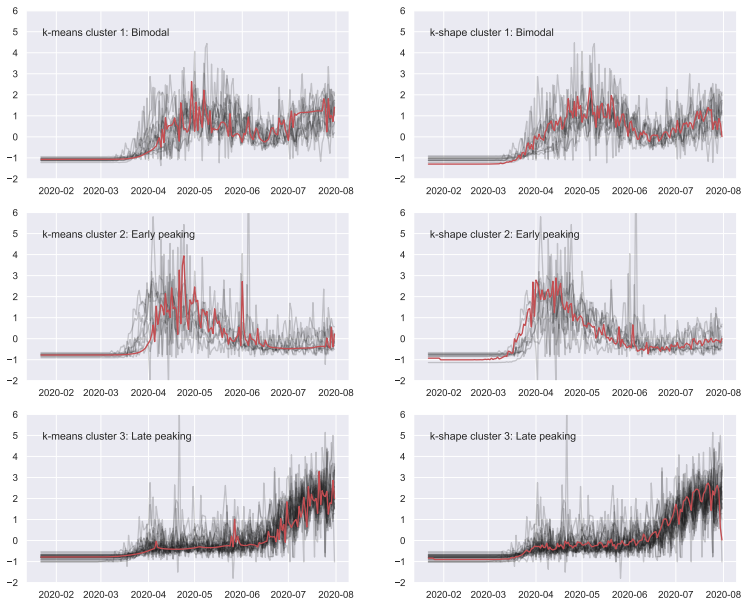

Figure 2: Mutually consistent clusters (rows) between the k-means and k-shape cluster algorithms. Data is time adjusted and barycenters are displayed in red.

transform of each series $\mathcal{F}(\mathbf{x}) \star \mathcal{F}(\mathbf{y})$, where $\star$ is the complex conjugate operator.[24] Then, given the inverse discrete Fourier transform $\mathcal{F}^{-1}$,

$$\psi(\mathbf{x}, \mathbf{y}) = \mathcal{F}^{-1}\Big(\mathcal{F}(\mathbf{x}) \star \mathcal{F}(\mathbf{y})\Big) \tag{12}$$

and as Paparrizzos and Gravano showed, Fast Fourier Transforms allow this to be calculated efficiently in $\mathcal{O}(|\mathbf{x}|log(|\mathbf{x}|))$ time rather than $\mathcal{O}(|\mathbf{x}|^2)$ time.

Similarly to the clustering effected in Subsection 2.2, k-shape clustering was performed using the `tslearn` package's `clustering.KShape` classifier, with a `n_init` setting at 16 iterations for centroid seeds, using the result with the lowest inertia, random initialization and a convergence tolerance of 10^{-6}.

3 Results

3.1 Clustering time dynamics of disease prevalence

The data set described in Section 2.1 has been separated into three clusters. As Figure 2 shows, there are three distinctly characterisable patterns based on the barycenters:

1. Late peaking (k-means cluster 1, k-shape cluster 1): states in this cluster typically have a steady, consistent pattern affected only by weekly periodicities, and begin to surge around mid-June 2020.

2. Early peaking (k-means cluster 2, k-shape cluster 2): states in this cluster display a rapid-onset initial peak in April to May 2020, thereafter tapering off.

3. Bimodal (k-means cluster 3, k-shape cluster 3): within this cluster, states appear to exhibit a steady number of cases and the beginnings of a bimodal distribution over time, with a peak in April-May 2020 that subsides in June, then follows on to another rise in July and August.

The geographical distribution of the permutation of k-means and k-shape classifications merits mention. As Figure

3.2 Cross-cluster agreement

In order to ascertain cross-cluster agreement, the Adjusted Rand Index (ARI) was used to quantify consensus between the k-shape and soft-DTW k-means classifiers.[26] This index, first proposed by Hubert and Arabie in 1985, is symmetric, thus it can be used to identify consensus between clusters with different metrics. At 0.864, the ARI indicates strong concurrence between the soft-DTW k-means and the k-shape classifiers.

Cross-cluster agreement is illustrated in Figure 4. As it is evident therefrom, over half of the states fall into the late-peaking (k-means cluster 3, k-shape cluster 3) category, with relatively few cases and no pronounced peaks until June 2020, after which the data evidences an oscillating but gradually increasing case count.

4 Discussion

k-shape and soft-DTW k-means classification strongly concur in identifying the three fundamental behavioural clusters of confirmed COVID-19 case count in the 48 states of the continental United States: a bimodal pattern, an early peaking pattern and a late, slower pattern that is largely stationary until approx. June 2020, then displays a rapid rise of cases.

Figure 3: Combined time traces of k-means and k-shape classifications for the major consensus groups. Bimodal behaviour accounts for 21% of states, early-peaking behaviour covers 17% and late-peaking, ascending behaviour accounts for over half (56%) of states. Three states do not fall within the major consensus groups.

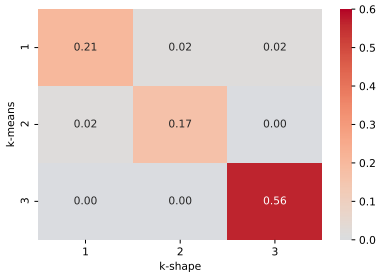

Figure 4: Inter-classifier agreement between k-shape (`k-shape`) and soft-DTW k-means (`kmeans`) classification.

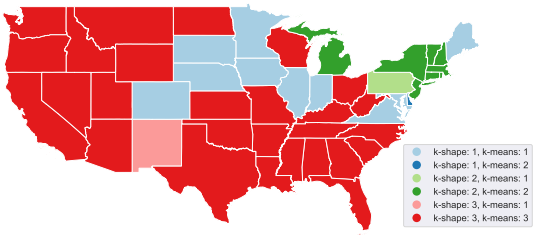

Figure 5: Choropleth map of the United States displaying the permutations of k-shape and soft-DTW k-means clustering results by state.

The geographical distribution of these is worth noting. As Figure 5 shows, most of the area of the continental United States currently follows the late peaking regime, and these states are currently poised to experience further growth in case counts. Only a few states (green shades) have followed an early outbreak by a significant reduction in cases and no further resurgence, which may be considered evidence of successful mitigation/suppression efforts on their part. Finally, a number of states (blue shades) have experienced early outbreaks and are exhibiting a bimodal pattern, whereby an initial surge in April to late May 2020 has been followed not by successful suppression but a reduction followed by yet another uptick in the number of reported cases of COVID-19.

As this paper has shown, time series clustering that allows for finding commonalities between time series that are by necessity out of synchrony can be helpful in illuminating geographical and regional patterns of disease dynamics. In particular, by using two different methods – a soft-DTW based, time-shifted k-means classifier and the correlation-based k-shape classifier –, the significant consensus between such classifications has been demonstrated where the number of confirmed COVID-19 cases in the continental United States is concerned.

Thus, by identifying the case count response, we can recognise different internally consistent clusters of case count progression over time. This may assist in understanding the governing patterns and dynamics of the SARS-CoV-2 pandemic, and assist in tailoring responses to the needs of individual areas and communities based on the temporal patterns of epidemic dynamics they exhibit.

Competing interests

The author declares no competing interests.

Supplementary data

All simulations, code and data are available on Github and under the DOI `tbc`. Shape files for the choropleth diagram in Figure 5 have been obtained from the United States Census Bureau, and are included in the data set noted above.

References

[1] Tariq Ahmad, Michael J Pencina, Phillip J Schulte, Emily O'Brien, David J Whellan, Ileana L Piña, Dalane W Kitzman, Kerry L Lee, Christopher M O'Connor, and G Michael Felker. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *Journal of the American College of Cardiology*, 64(17):1765–1774, 2014.

[2] Pranab Haldar, Ian D Pavord, Dominic E Shaw, Michael A Berry, Michael Thomas, Christopher E Brightling, Andrew J Wardlaw, and Ruth H Green. Cluster analysis and clinical asthma phenotypes. *American journal of respiratory and critical care medicine*, 178(3):218–224, 2008.

[3] Christine Lochner, Sian MJ Hemmings, Craig J Kinnear, Dana JH Niehaus, Daniel G Nel, Valerie A Corfield, Johanna C Moolman-Smook, Soraya Seedat, and Dan J Stein. Cluster analysis of obsessive-compulsive spectrum disorders in patients with obsessive-compulsive disorder: clinical and genetic correlates. *Comprehensive psychiatry*, 46(1):14–19, 2005.

[4] M Weatherall, J Travers, PM Shirtcliffe, SE Marsh, MV Williams, MR Nowitz, S Aldington, and R Beasley. Distinct clinical phenotypes of airways disease defined by cluster analysis. *European Respiratory Journal*, 34(4):812–818, 2009.

[5] Lichuan Ye, Grace W Pien, Sarah J Ratcliffe, Erla Björnsdottir, Erna Sif Arnardottir, Allan I Pack, Bryndis Benediktsdottir, and Thorarinn Gislason. The different clinical faces of obstructive sleep apnoea: a cluster analysis. *European Respiratory Journal*, 44(6):1600–1607, 2014.

[6] Tanmay Kumar Behera and Suvasini Panigrahi. Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In *2015 Second International Conference on Advances in Computing and Communication Engineering*, pages 494–499. IEEE, 2015.

[7] Qi Liu and Miklos Vasarhelyi. Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In *29th world continuous auditing and reporting symposium (29WCARS), Brisbane, Australia*, 2013.

[8] Yi Peng, Gang Kou, Alan Sabatka, Zhengxin Chen, Deepak Khazanchi, and Yong Shi. Application of clustering methods to health insurance fraud

detection. In *2006 International Conference on Service Systems and Service Management*, volume 1, pages 116–120. IEEE, 2006.

[9] Andrei Sorin Sabau. Survey of clustering based financial fraud detection research. *Informatica Economica*, 16(1):110, 2012.

[10] Sharmila Subudhi and Suvasini Panigrahi. Use of optimized fuzzy c-means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University-Computer and Information Sciences*, 2017.

[11] Keh-Shih Chuang, Hong-Long Tzeng, Sharon Chen, Jay Wu, and Tzong-Jer Chen. Fuzzy c-means clustering with spatial information for image segmentation. *computerized medical imaging and graphics*, 30(1):9–15, 2006.

[12] Guy Barrett Coleman and Harry C Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.

[13] Xiao-Bo Jin, Guo-Sen Xie, Kaizhu Huang, and Amir Hussain. Accelerating infinite ensemble of clustering by pivot features. *Cognitive Computation*, 10(6):1042–1050, 2018.

[14] Kyle Lafata, Zhennan Zhou, Jian-Guo Liu, and Fang-Fang Yin. Data clustering based on langevin annealing with a self-consistent potential. *arXiv preprint arXiv:1806.10597*, 2018.

[15] Thrasyvoulos N Pappas and Nikil S Jayant. An adaptive clustering algorithm for image segmentation. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 1667–1670. IEEE, 1989.

[16] Zhenyu Wu and Richard Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1101–1113, 1993.

[17] Földi Tamás and Chris von Csefalvay. Starschema covid-19 data set, August 2020. URL `https://doi.org/10.5281/zenodo.3969287`.

[18] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9), 2011.

[19] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[20] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*, 2017.

[21] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In *International Conference on Artificial Intelligence and Statistics*, pages 1695–1704, 2020.

[22] Hiroto Saigo, Jean-Philippe Vert, and Tatsuya Akutsu. Optimizing amino acid substitution matrices with a local alignment kernel. *BMC bioinformatics*, 7(1):246, 2006.

[23] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21 (118):1–6, 2020. URL `http://jmlr.org/papers/v21/20-091.html`.

[24] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870, 2015.

[25] Alan V Oppenheim, John R Buck, and Ronald W Schafer. *Discrete-time signal processing. Vol. 2*. Upper Saddle River, NJ: Prentice Hall, 2001.

[26] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.