Clustering Spatio-Temporal Waves of Covid-19

Kevin Quinn, Evimaria Terzi, Mark Crovella

October 10, 2025

Abstract

It is commonly recognized that Covid-19 fluctuates in waves of infection activity, but much less is known about how waves differ or coincide when they are observed from different geographical locations. In this paper we aim to study the geographical patterns of infection waves throughout the Covid-19 pandemic. We propose a novel methodology for doing so, which both segments infection time-series data into waves and then clusters them together. With it we study US state and county level data, as well as data from European countries. From a clustering, we aim to bring understanding to viral spread by measuring geographical proximity and (for states and countries) similarity in terms of public health response.

1 Introduction

Understanding and communicating patterns from data which varies in both space and time is a difficult but important problem, appearing in similar forms across domains such as climate science, biology, or ecology. Prompted by the recent Covid-19 public health crisis, we study this problem in an epidemiological setting where data is collected from a set of spatial locations with temporally evolving *wave*-like infection patterns. As Covid-19 progressed the idea of the virus spreading in waves became increasingly popular, with infections rising and falling due to changes in season, patterns of contact, or viral evolution [38]. For example, figure 1 shows infection time series for two US states which have multiple peaks with high infection levels.

Throughout our study we refer to *waves*, *wave segments*, or simply *segments* as contiguous subsets of an infection time-series which are characterized by abnormal or increased viral activity. We argue that in order to effectively understand the spatio-temporal patterns of viral spread it is important to distinguish waves from their time-series as a whole. To understand why, refer again to the example in figure 1 to see that the states Massachusetts and Vermont experienced the pandemic in different but sometimes similar ways. Massachusetts had early waves around April 2020 and January 2021 which were largely non-present in Vermont. Later on in the time series, however, the two states *do* show similar patterns with waves in January and May of 2022 that are only slightly different in scale or time-alignment. We argue that while it should make sense to group the locations together during the waves in 2022, it makes much less sense to claim that these locations experienced the pandemic similarly as a whole. In this case, accounting for more temporal granularity gives deeper insight to the relationship between locations. One should allow locations to evolve over time in different but sometimes overlapping ways. However, there is also something to be said for having too much granularity – with the extreme case being an analysis that only considers single points from each time series. In that case, minor differences in timing or scale could distract from the possibility of two locations sharing similar trends on a larger time scale.

In this study we use waves to characterize the evolving spatio-temporal patterns of Covid-19. To do so we introduce a methodology which 1) splits each location's time-series into wave segments and then 2) clusters these waves together in a way which respects their position within the global time-frame. To find wave segments we modify the classic k-segmentation problem to fit wave-like unimodal or SIR representative functions to k contiguous, non-overlapping segments of each time series. We then cluster these wave segments using a hierarchical algorithm with added time constraints and a dynamic time warping distance. In our experiments we study clusters from datasets of US states, European countries, and US counties. For states and countries, we also use auxiliary geographic and public health data in an attempt to understand results and validate with comparisons against random baselines. We find that in the case of US states we are able to show that clusters

are significantly similar in both geography and governmental response. For our the European country dataset, however, the same conclusions did not hold. Locations grouped together as clusters were less likely to be similar with respect to these validation measures, suggesting a possible difference in how the two geographical regions experienced the pandemic. Despite these unclear analytical difficulties, we argue that our study introduces a novel problem, contributes a methodology to address it, and brings attention to the computational problems associated with it. We provide a full implementation that includes our experimental results. [1]

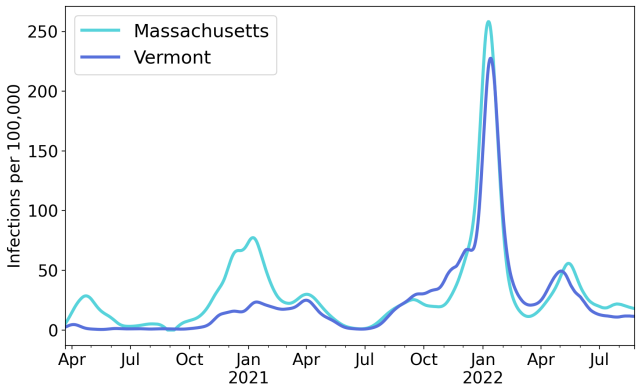

Figure 1: Time-series of daily new infections for US states Massachusetts and Vermont. Infections are reported on normalized (per 100,000 persons) scale, and have been pre-processed to remove noise. They are shown for a time period between late march of 2020 and late August of 2022.

2 Related Work

Our study draws from previous work on Covid-19, epidemiology, time-series segmentation, and time series clustering. We argue that the proposed methodology pieces these topics together in a novel way.

Covid-19: Our work is prompted by the devastating impact of Covid-19 and is inspired by the scientific work that has been done to address it. In particular, we draw from a line of work which attempts to unravel the complex geographical patterns of spread seen throughout the pandemic. Work from [10][34][35] attributes differences in infection curves to differences in factors ranging from geographical distance, population size, population density, population ages, and patterns of movement or contact. Furthermore, we reflect ideas from [3] to emphasize that local public health policy differences also play a significant role in producing epidemiological differences between locations.

We also address the fact that these geographical relationships are *non-static*. As the pandemic evolved, infection levels, contact patterns, and public health policies were consistently changing in part due to actual viral mutations of the disease. The result of such was the observed occurrence of time localized *waves* of spread, which [38] [2] [6] help to define and characterize.

Work from [22][37][13] uses clustering as a way of analyzing the complicated geographical patterns seen from their data. However, their clustering is static in the sense that they don't consider localized variations in time. Other studies from [8][28] use matrix factorization techniques to analyze spatio-temporal patterns, but are again limited by the fact that they don't allow locations to to evolve or change their cluster membership over time.

Other methods do allow for more temporal granularity, for example the work done by [23][11] leverages spatial auto-correlation tests to study both spatial and temporal patterns in Covid-19 data. This methodology originates from a long line of work [20][29][25] that searches for spatial or spatio-temporal clusters by enumerating over all possible location pairs within heuristically chosen (often evenly spaced) time-windows, and searching for those with statistical significance under some chosen model of concentration or similarity. We argue that these are limited by the need to create new clusterings for every specific time point or for every heuristic choice of time window. We distinguish our work by explicitly defining, modeling, and clustering *waves*, or short periods of pronounced infection. Doing so not only allows us to capture and visualize geographical patterns, but also gives us the freedom to let them evolve over time in a more unified and structured way.

[1] https://github.com/kevin-q2/wave_cluster/tree/main

Segmentation: Our work is in principle very similar to work for the k-segmentation problem [4] or more specifically the unimodal k-segmentation problem studied by [15]. They introduce and optimally solve the problem of fitting k piece-wise, constant segments to a time series with the constraint that the concatenation of their segments forms a unimodal curve. We employ a similar approach, adopting their dynamic programming techniques, with the important modification of requiring every *segment* to be a unimodal curve. We also contribute the idea of using an epidemiological approach that fits segments with Susceptible, Infected, Recovered (SIR) models. To do so, we follow work from [9][7] which studies the problem of using non-linear least squares regression to fit parameters to the SIR model.

Other approaches to the time series segmentation problem have greedily fit representative functions to k segments of the data by iteratively searching for segments with small error of fit [18]. We also note that work from [33] makes improvements in efficiency to the dynamic programming approach in [15], introducing faster approximation algorithms that perform well in practice.

Recent work on segmentation in the context of Covid-19 by [17] introduces a new algorithm for segmenting time-series data with re-occurring wave patterns. We include the use of their algorithm in our experiments for comparison and give a brief explanation for how it works in section 4.1.

Time Series Clustering: Clustering of time-series is a very challenging problem when data is lengthy, irregular, or noisy. There have been many different approaches to the problem which is nicely surveyed by [1] [21]. Most methods can be characterized by 1) how they choose to represent or find distances between time-series and 2) the algorithm they use to partition them. Previous approaches have typically focused on finding good notions of distance or otherwise transforming the data into some comparable feature space, and then applying standard k-means or hierarchical algorithms for clustering. Methods which are most relevant to our work are denoted by [1] as 'shape' based approaches because they employ the use of dynamic time warping distance [30] in order to allow for comparison of time-series data which is mis-aligned in time. Doing so gives more importance to the overall shape of their patterns. A recent study by [22] uses dynamic time warping distance in combination with hierarchical clustering in order to find clusters of Covid-19 time series data.

Our work makes similar use of dynamic time warping with hierarchical clustering, but is fundamentally different from many previous approaches, including [22], because we focus on clustering segments or sub-sequences taken from each location's time-series data. While most work on time-series clustering focuses on clustering whole time-series, substantially less has been done to cluster *segments* of them. Work by [14], for example, includes a similar sub-sequence clustering methodology, but only does so as a step towards their goal of clustering whole time-series. This trend seems to be in part because of influence from [19] which argues against taking such an approach. We claim, however, that the issues they propose are irrelevant for the setting of our study. This is in particular because of the fact that they focus on clustering sliding-window segments from a *single* input time-series whereas we cluster carefully modeled segments from *multiple* locations, never allowing two segments from the same location to be clustered together. We back this claim by validating our method with experiments comparing against randomized clustering baselines.

Our approach is also unique in the sense that we enforce that each cluster's segments must come from similar times. Because each time-series (corresponding to a location) lies within some common time period, we can record the subset of times occupied by a segment and choose to only cluster segments which are similar in time. By doing so we narrow our focus to finding clusters which are positioned both spatially and temporally within the data. We employ a hierarchical graph clustering approach in which each segment is a node, edges are present between segments which are similar in time, and clusters are required to be cliques. In that regard, our clustering algorithm is most similar to existing graph clustering algorithms [31] and community detection algorithms [27].

3 Datasets

We study three datasets reporting daily new infections of Covid-19 for multiple locations. Each is collected and attributed to Google's Covid-19 open data repository [36]. We'll represent each dataset D as an $m \times n$ matrix made up of n time-series vectors of length m. Each of the n time-series corresponds to a single location and is a daily record of that location's reported incidence, or number of new cases, of Covid-19 from a period starting with the beginning of March 2020 (or the beginning of February 2020 for European locations) and ending around mid-September of 2022. The different geographical regions we consider for our datasets are the

50 US states and Puerto Rico, 46 European countries, and 217 counties from the Northeastern Region of the US. We'll refer to their datasets as D^{us}, D^{eu}, and D^{co} respectively.

For each location in both of our datasets, we normalize their corresponding time-series by the specific location's population size and then report in the form of cases per 100,000 persons. That is, for each of these time series we divide every point by a static population count given by [36] and then multiply by 100,000. To remove noise due to reporting inconsistencies we use a sequential windowed average procedure that applies a 15 day windowed average of the data points 3 times. Specifically, for every location each data point is replaced by the average of itself, all the points 7 days before, and all the data points from the 7 days after. Points at coming from the first 7 or last 7 days seen in each time-series are removed for lack of information. We repeat this procedure 3 times, finding this to be sufficient for removing a few very noisy occurrences in the data. Representations of the processed data can be seen in figure 1.

For our analyses we also consider auxiliary information regarding geographic position and government response. For locations in the US we are able to use geographical coordinates for their centers of population collected by the US Census Bureau [5], but for locations in Europe we use central geographical coordinates reported in Google's repository [36]. To study trends in governmental response and public health we also incorporate information collected by the Oxford Covid-19 Government Response Tracker project [16]. For all locations from D^{us} and D^{eu} the Oxford dataset dataset provides a time-series of scores rating the local government's stringency and proactive response towards containment policies (closures for school, work, or other events – as well as restrictions on movement, travel, or transportation) and public health policies (public information campaigning, testing, contact tracing, facial coverings, and vaccinations). Scores are reported on a scale from 0 to 100 with higher scores indicating significant government intervention to stop the spread of the virus. Throughout our analyses we'll study these datasets and use the auxiliary data to validate and supplement our results.

4 Methodology

We begin by introducing the following notation that is used throughout the rest of this paper.

- Let D be a $m \times n$ dataset where each of n locations has a length m time series associated with it. We refer to D^{us}, D^{eu}, and D^{co} as the US state, European country, and US county datasets respectively. We'll also use D_ℓ to denote the ℓth column of a dataset D, which is also the time series associated with the location indexed by ℓ
- Let a wave w be a length s segment (or sub sequence) obtained from one of the time series vectors in D. We may also write w^i_ℓ to denote a wave as being the ith wave segment taken from the time series for location ℓ, or w^i for a non-specific location. We often also generally denote time series vectors as x or y, and indicate indexed subsets of them with $x_{t:t'}$, or $x_{:t}$, $x_{t:}$ for slices taken from the beginning or until the end of the vector. For a single time indexed element of x we'll write x_t
- Let $\mathcal{W}$ be the size q set of all wave segments found from each time series in D. We also write $\mathcal{W}_{\text{uni}}$, $\mathcal{W}_{\text{sir}}$, or $\mathcal{W}_{\text{wav}}$ to denote sets of waves found specifically with unimodal, SIR, or wavefinder segmentation methods. We also use $\mathcal{W}_\ell$ with size q_ℓ to denote the subset of wave segments coming specifically from location ℓ
- Let d denote any distance function. We'll use d_{dtw} or d_{comp} when referring to specific distance notions such as dynamic time warp distance or complete linkage distance.
- Let a clustering $\mathcal{C}$ be a set of r clusters $\{\mathcal{C}_1, ..., \mathcal{C}_r\}$ where each cluster $\mathcal{C}_i$ is a set of segments $\mathcal{C}_i \subseteq \mathcal{W}$ with the property that $\mathcal{C}_1 \cup \mathcal{C}_2 \cup ... \cup \mathcal{C}_r = \mathcal{W}$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for all i, j.

4.1 Segmentation

Given a length m infection time-series vector x and a number of segments k, our first task is to cut x into k contiguous and non-overlapping chunks of time (or *waves*) $w^1, ..., w^k$. To do so we search for the best $k-1$ cut-points or boundaries with which to split the data. Following the classic k-segmentation problem [4][15] our approach is to fit k wave-like representative functions to x, using dynamic programming in order to find the

optimal timing for each. In general we'll have some space of representative functions M and an approximation function $f_M : \mathbb{R}^s \to \mathbb{R}^s$ that takes as input a length $s \geq 1$ sub-sequence of x and computes a new wave vector w as the best approximation to to it from within the model space M. We use the dynamic program `k-segment()` shown in algorithm 1 to find and compute the k cuts with smallest segmentation cost c_k, measured with euclidean distance.

Algorithm 1: `k-segment()`

```
Input:  x, k, f_M
Output:  c_k
if k = 1 then
 |  c_k ← ‖x − f_M(x)‖_2
else
 |  c_k ← min_{t<|x|−(k−1)v}  ‖x_{:t} − f_M(x_{:t})‖_2 +  k-segment(x_{t:}, k − 1, f_M)
```

For this approach we'll also use a minimum wave segment size $s \geq v$, to ensure that the algorithm finds non-trivial sub-sequences. An implementation of this algorithm which finds the best segment boundaries must also keep track of costs by filling out entries of a size $m \times k$ matrix, tracing back through it in the end to find the boundaries which gave the best k-cost, c_k. With the simplest model space M of constant vectors (entries all the same value), the best approximation function f_M computes the mean of its input. In this case filling out each entry of the size $m \times k$ cost matrix takes $\mathcal{O}(m)$ time, and therefore the implementation as a whole can be computed in $\mathcal{O}(m^2 k)$. For the purposes of our study, however, we'll use this algorithm with attention restricted to the following wave models M:

Unimodal Models denoted as M_{uni} are regression functions meant to fit data which is monotonically increasing until it reaches a single pronounced local maximum, and then monotonically decreasing afterwards [12]. Both monotonically increasing and decreasing isotonic regression models are fit to the areas before and after the local maximum change point, which is found optimally as the point which produces the smallest error. Specifically, we define increasing and decreasing isotonic regression functions:

$$f_1(y) = \underset{\hat{y} \mid \hat{y}_i < \hat{y}_j \ \forall i<j}{\arg\min} \sum_i (y_i - \hat{y}_i)^2$$

$$f_2(y) = \underset{\hat{y} \mid \hat{y}_i > \hat{y}_j \ \forall i<j}{\arg\min} \sum_i (y_i - \hat{y}_i)^2$$

and concatenate their response in a single fitting function $f_{W_u}(x,t) = (f_1(x_{:t}),\ f_2(x_{t:}))$ The optimal change point t^* is then fit the data by minimizing:

$$t^* = \arg\min_t ||x - f_{W_u}(x,t)||_2$$

In general, fitting a unimodal model to a length m sequence can be computed in $\mathcal{O}(m^2)$ time [12][15], implying that `k-segment()` with the unimodal model can be computed in $\mathcal{O}(m^3 k)$.

SIR Models denoted as M_{sir} are dynamical systems meant to replicate the process of a disease by moving a fixed population of N individuals between categories Susceptible, Infected, and Recovered. The model begins with an input amount of persons in each category: S_0, I_0, R_0. It also takes as input a spread parameter $\beta \in \mathbb{R}_{\geq 0}$ which describes the average number of infection prone "contacts" that an individual has over a single unit of time (either spreading the disease to or away from the individual). Similarly a recovery rate parameter $\gamma \in [0,1]$ specifies the fraction of infected individuals which move from the infected to recovered categories within over a single unit of time. We'll denote the set of parameters necessary to define an SIR model as $\Omega = \{S_0, I_0, R_0, \beta, \gamma\}$.

At each time $t \geq 0$ the number of susceptible, infected, or recovered persons is given by the functions $S(\Omega,t), I(\Omega,t), R(\Omega,t)$ respectively, with the condition that $S(\Omega,t) + I(\Omega,t) + R(\Omega,t) = N$ for all t. These functions change over time according to the following set of ordinary differential equations:

$$\frac{\partial S}{\partial t} = -\frac{\beta SI}{N} \qquad \frac{\partial I}{\partial t} = \frac{\beta SI}{N} - \gamma I \qquad \frac{\partial R}{\partial t} = \gamma I \tag{1}$$

To construct an SIR approximation to our data we'll need to optimally choose a parameter set Ω. Since our data is a record of daily infections, we're particularly interested in finding a model with a well-fitting infection curve $I(\Omega, t)$. However, we carefully make the distinction between I which computes the *prevalence* of infections, or the total number of individuals currently infected at every t, and our data which records the *incidence* of infections, or the number of new infections at every t. To translate between these settings we'll compute differences between successive time steps:

$$f_{M_{\text{sir}}}(\Omega, t+1) = S(\Omega, t) - S(\Omega, t+1)$$

And choose parameters Ω^* to minimize the error

$$\Omega^* = \arg\min_{\Omega} \sum_{t_i} \left(x_{t_i} - f_{W_s}(\Omega, t_i)\right)^2$$

Unlike the unimodal model, choosing parameters requires expensive non-linear least squares fitting with gradient descent, which often makes this difficult to compute for long time series. Briefly we may describe the complexity of fitting a model with p parameters to a length m vector using non-linear least squares as $\mathcal{O}(\omega m p^2)$ where ω is the number of steps taken to reach convergence during the gradient descent procedure [26]. For our case of 5 parameters this gives an overall `k-segment()` complexity of $\mathcal{O}(\omega m^2 k)$, but we note that to reach convergence ω often needs to be large and can increase with the size of the input vector.

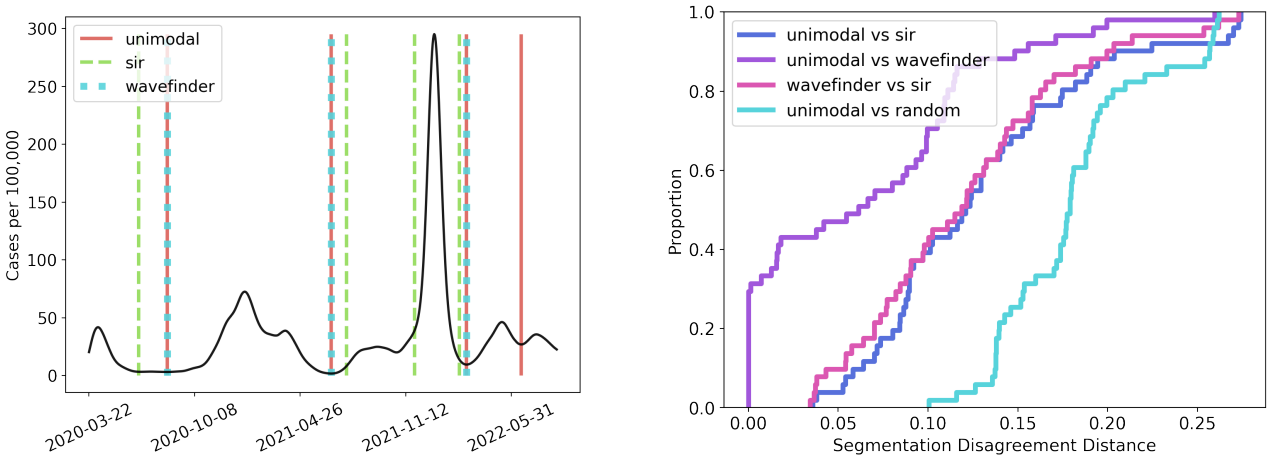

Figure 2: Comparison of Time Series segmentations with both unimodal and SIR `k-segment()` , and `wavefinder()` . (Left) A visual representation of the chosen segment boundaries for each algorithm when computed on the infection time-series for New York State. (Right) We also show a cdf of the disagreement distances for each pair of deterministic algorithms, when computed for every time series in D^{us}. With this we include a baseline comparison to random segmentation for unimodal, computed over 10,0000 samples for each location.

Finally, we'll also consider two other segmentation methods for comparison. The first is a recent method of time series segmentation proposed by [17]. Their algorithm, `wavefinder()` , finds peaks or pronounced periods of increased infection by searching local maxima and local minima and then carefully filtering out noisy occurrences. Unlike the others this algorithm does not fit representative waves to segments of the data, making it fast and generally applicable to any data with wave-like patterns. Finding local maxima and minima is done in $\mathcal{O}(m)$ and an extra parameter controlled filtering process taking only a small amount of extra time. We note that while this algorithm does have the advantage of being very fast, it lacks careful modeling for its segments and is reliant upon successful filtering. Because their segments are taken to be everything left over after the filtering process is complete, there is also very little control in choosing the number of segments.

The other baseline comparison we use is a randomized method `k-random-segment()` for splitting the data into k segments. With a minimum segment size of v this algorithm sets aside kv chunks of time, and then uniformly at random chooses $k-1$ cut points from the remaining set $\{0, 1, 2, ..., |x|-kv\}$. Let $\theta = \{\theta_1, \theta_2, ..., \theta_{k-1}\}$ be an ordered set of these randomly chosen cuts. To cut the original time series we then add back the chunks of time which were set aside. Specifically, transform each cut i as $\theta_i = \theta_i + iv$. Doing so allows us to uniformly sample from the space of segmentations that satisfy the minimum length requirements.

We compare `k-segment()` with both unimodal and SIR models to `wavefinder()` visually for a single time series vector on the left of figure 2. We also compare the segmentations more rigorously by following [24] and computing the disagreement distance between them. Briefly stated, given any time series vector x the disagreement distance between two segmentations of x is the fraction of pairs of points which belong to the same segment in on segmentation, but belong to different segments in the other. In the right hand plot of figure 2 we show the distribution (plotted as a cdf) of these distances when computed for each pair deterministic of segmentation methods over all of the time-series from D^{us}. One can notice more conclusively that the unimodal `k-segment()` algorithm and the `wavefinder()` algorithm are both very similar, often only disagreeing on $< 5\%$ of pairs for a time series. That being said, we note that *all* of the segmentation methods shown are similar enough to have distinctly smaller disagreement distances to each other than to a random segmentation with `k-random-segment()` . In the plotted figure we show the distribution of differences between the unimodal method and its random counterpart, for 10,000 samples at every location.

4.2 Dynamic Time Warping

A key feature of our methodology, and one which is often useful for time series clustering [1], is the use of dynamic time warping (dtw) to compare two waves which are not necessarily aligned in time. Let $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^{m'}$ be two waves or time series we wish to align using dtw. Let $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be some standard distance function for comparing elements of x with elements of y. For all of our experiments we use euclidean distance for d. Let $d_{\text{dtw}}(i, j)$ be this dynamic time warping distance between $x_{:i}$ and $y_{:j}$. Then we can express a penalized version of the dtw distance in the following recursive format:

$$d_{\text{dtw}}(i,j) = \min \begin{cases} \lambda_1 d(x_i, y_j) + d_{\text{dtw}}(i-1, j) \\ \lambda_1 d(x_i, y_j) + d_{\text{dtw}}(i, j-1) \\ \lambda_2 d(x_i, y_j) + d_{\text{dtw}}(i-1, j-1) \end{cases}$$

Any solution to the dtw problem creates an alignment path which we follow [32] in denoting as length q sequence of indices $(0,0), (i_1, j_1), ..., (i_{q-2}, j_{q-2}), (m-1, m'-1)$, where each i_k represents the index of x matched or aligned with the index j_k of y. The alignment is constrained so that the head $(0,0)$ and tail $(m-1, m'-1)$ of x and y must be matched together, the indices must be monotonically increasing $i_{k-1} \leq i_k$, and all indices from each vector must be represented somewhere in the path. A simple dynamic programming implementation can compute the optimal alignment in $\mathcal{O}(m \cdot m')$ time.

To this standard form we add penalties λ that force dtw to create alignment paths which are penalized for making sequential matches $(i_{k-1}, j_{k-1}), (i_k, j_k)$ where either $i_{k-1} = i_k$ or $j_{k-1} = j_k$. In other words, we want to encourage the alignment to be making forward progress in both vectors. In alignments where x or y have one index which is stagnantly matched to many indices from the other vector, the shapes of the given waves are not preserved. For our experiments we set $\lambda_1 = 5$ and $\lambda_2 = 1$, finding this to be a setting which matched our intuition for an alignment which allows for time differences, but ultimately still preserves the *shape* of each wave.

We also take careful precaution to normalize the input waves and the computed distance before performing any clustering, so that distances between waves are comparable regardless of length or scale. First we normalize both x and y by the largest value seen in either vector: $\max\{\max_i x_i, \max_j y_j\}$. This re-scaling ensures that the euclidean distance between every matched x_i and y_j is a value between 0 and 1. Then after computing the dtw distance d_{dtw} between x and y, we divide by the maximum possible distance between two 0-1 scaled vectors of the same lengths. If $\lambda_1 > \lambda_2$ this can be found simply found as $\lambda_1(m + m' - 2) + \lambda_2$.

4.3 Clique Clustering

Any wave w is associated with a ordered, consecutive subset of times from its parent time series. Let $T(w) = \{t_1, t_2, ..., t_s\}$ be this set of times for wave w. We design our algorithm to cluster waves in a way which ensures they aren't too different with respect to their position in the global time frame. For example, we wouldn't want to cluster together a wave which happened in April of 2020 with another that happened in June of 2022. We'll therefore define a time overlap fraction parameter δ and require that for any waves w_i and w_j to be clustered together, both of the following time-similarity conditions hold:

$$\frac{|T(w_i) \cap T(w_j)|}{|T(w_i)|} \geq \delta \qquad \frac{|T(w_i) \cap T(w_j)|}{|T(w_j)|} \geq \delta \tag{2}$$

For an entire set of waves $\mathcal{W}$ found by in the segmentation process we then create a time overlap graph $G_\delta(V, E)$ where V contains nodes for each wave, and edges $(i, j) \in E$ are only present if w_i, w_j satisfy the stated timing conditions in equations (2). For our purposes we will be interested in *cliques* within this graph, since they must correspond to sets of waves which all have significant overlap with each other in time. Therefore, we introduce the following hierarchical clustering algorithm which partitions $\mathcal{W}$ into r clusters corresponding to cliques in G_δ. Our algorithm keeps a list of clique clusters $\mathcal{C}$ and merges them hierarchically using a complete linkage distance d_{comp} computed with dtw:

Algorithm 2: `clique-cluster()`

```
Input: W, r, G_δ
Output: C
C ← W;
while |C| ≥ r do
    if ∃ i, j s.t. clique(C_i ∪ C_j) then
        i, j ← arg min_{i,j} d_comp(C_i, C_j) s.t. clique(C_i ∪ C_j);
        if d_comp(C_i, C_j) > η then
            return
        else
            merge(C_i, C_j);
    else
        return
```

We denote *clique*() as a function to check whether a set of nodes forms a clique in δ, terminating early if there are no clusters in $\mathcal{C}$ can be joined to form a clique. Importantly we also define a distance threshold parameter η past which we do not allow cliques to be merged. Unless one of these conditions is broken early, the process continues until we've formed r clique clusters. With d_{dtw} distances pre-computed for every pair amongst q input waves, careful tracking and updating of a pairwise distances between cliques leads to a clustering complexity of $\mathcal{O}(q^3)$. In practice we choose the number of clusters r by simply iterating until it's no longer possible to merge two cliques.

For our experiments we'll then compare this to a randomized version which we'll call `random-clique-cluster()` in which we simply replace the selection of clusters with minimum d_{comp} by a uniform random selection of from all cluster pairs which satisfy the clique condition.

5 Experiments

We first focus on the state-level dataset D^{us} and use the pre-computed unimodal, SIR, and wavefinder segmentations as input to our clustering algorithm. To select values for our time overlap parameter δ and distance threshold parameter η we compute silhouette scores (described later in this section) for clusterings over a range of parameter values, and select the parameter settings that produce the highest scores, and therefore have the most distinctive wave clusters. We show a clustering of D^{us} in figure 3.

Our algorithm often produces many clusters which can often be small or irrelevant, posing a problem for succinct visualization and clear understanding of of the results. We note that this issue is often inherent to complicated spatio-temporal problems, and present a simple, albeit incomplete, solution for viewing results from our methodology. In each of the rows of figure 3 we select a single time point t from the global time period of D^{us} and search for clusters which have all of their wave segments passing through or containing the time t. In the maps on the left we show the geographical locations with each cluster colored distinctly. States which are left un-colored belong to clusters which are not fully represented by the time point. In the corresponding plots on the right we display the an average wave segment for each of the clusters, with the average of a cluster being taken across the time-period shared by each of its waves. New cases per 100,000 persons are displayed on a log scale for ease of comparison between plots, and shaded regions show the standard deviation for each of the average segments. To select t values for plotting we hand-picked more or less evenly spaced time points, with some preference given to times in which the US had high infection activity.

Moving downwards in the figure 3 and following the progression of time one can notice how the pandemic evolved and how our clustering methodology is able to distinguish patterns of spread. In the beginning at time-stamp 3/22/2020 (top row) we find 5 distinct clusters with a wave starting in the north east at New York, Massachusetts, and New Jersey (dark blue), and afterwards spreading to other surrounding states (light blue) as well as parts of the midwest (light blue / green). This is later followed by a wave experienced commonly by many southern and western regions of the US (purple). In this time period of early 2020, our conclusion is that different regions of the US experienced beginning of the pandemic differently, with the south and the west lagging in time behind states like New York in New Jersey which were, at the time, epicenters of spread in the US.

The second time period we analyze is from clusters containing the time 11/27/2020 (second row). This period of the pandemic saw a large rise in cases starting in the midwest at states like North and South Dakota, and then spreading outwards across the entire US through the winter. Again we notice geographical contiguity among the midwest, south, and parts of the northeast, and attribute this to each region having significant differences in timing or scale of the wave experienced. We find fewer distinct patterns in the time period surrounding 9/23/2021 which came as a build up to the large wave experienced by nearly every US state in the winter of 2021-2022 and which we display with the time point 1/1/2022. During this time all clusters have similar waves, with distinctions only coming from very slight differences in time or scale. Admittedly we notice that our clustering algorithm performs poorly during this time, splitting up locations somewhat arbitrarily when it might have been more reasonable to just cluster all of them together. Still we'd like to hypothesize that during this time period, because restrictions began to loosen and travel opened up again, we saw much more country-wide similarity in spread compared to the geographically partitioned results of the first two plots. Finally note that not much interesting happens during the downturn period that happened around 5/31/2022, as things seemed to calm down for the remainder of the year.

While it's particularly difficult to validate our results, we can at least check that our clustering satisfies a few reasonable assumptions distinguishably better than a random clustering baseline does. To do so we will compute a few measurements, ρ for our clusters.

Assumption 1. *A cluster should have waves which are sufficiently distinct (in timing or shape) from the waves of any other cluster in the clustering*

This is a standard assumption for any clustering algorithm and to check, we compute the silhouette coefficient $\bar{\rho}_{\text{sil}}$, defined as the average of silhouette scores ρ_{sil} for each wave w_i. Let $f(w_i) = \frac{1}{|\mathcal{C}_{w_i}|-1} \sum_{w_j \in \mathcal{C}_{w_i}} d_{\text{dtw}}(w_i, w_j)$ be the average dtw distance between w_i and all other waves in its own cluster.

Also let $g(w_i) = \min_{\mathcal{C}_\bullet \neq \mathcal{C}_{w_i}} \frac{1}{|\mathcal{C}_\bullet|} \sum_{w_j \in \mathcal{C}_\bullet} d_{\text{dtw}}(w_i, w_j)$ be the average distance of w_i to some closest *other* cluster for which w_i is able to join under clique constraints. With these the silhouette score for w_i is defined as

$$\rho_{\text{sil}}(w_i) = \frac{g(w_i) - f(w_i)}{\max(g(w_i), f(w_i))} \tag{3}$$

If there is no *other* cluster which w_i is able to join under the clique constraints, we simply let $\rho_{\text{sil}}(w_i) = 0$. We can then compute the silhouette score of a clustering $\bar{\rho}_{\text{sil}}(\mathcal{C})$ as the average of $\rho_{\text{sil}}(w_i)$ scores for every wave segment in $\mathcal{W}$. Doing so gives us a numerical measure for how well our algorithm is able to distinguish the waves which it was intended to cluster. However, we will note that this should be a given. Our algorithm is

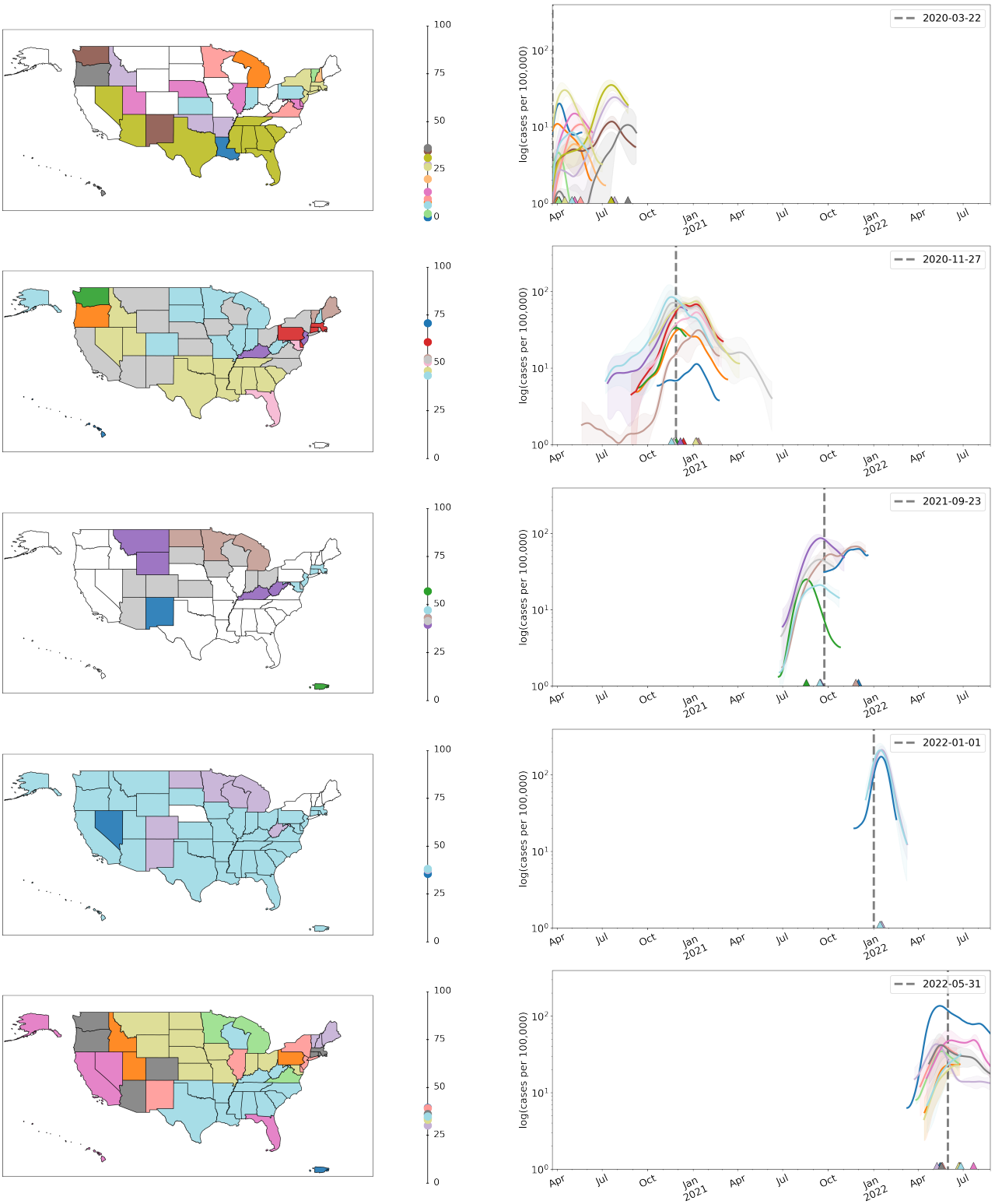

Figure 3: D^{us} clustering over time with Unimodal segmentation. In each row we show a representation of the clusters present for chosen time, evolving over time. We show a geographic map (left) where we assign a color to each of the current clusters and color the locations represented in each of them. We also report average government response scores (middle) on a scale from 0 to 100 for each of the clusters, and their average waves of infection over time on a log scale (right).

designed to cluster waves and we specifically chose parameters to boost this score. We are most interested to see how our clustering does with respect to other measurements.

Assumption 2. *The locations for a cluster's waves should be geographically similar*

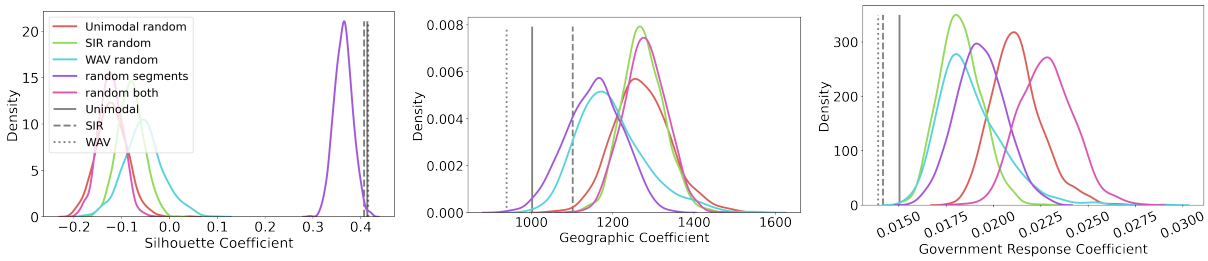

Figure 4: Comparisons against random clustering algorithms for assumptions on $\bar{\rho}_{\text{sil}}$ (left), $\bar{\rho}_{\text{geo}}$ (middle), and $\bar{\rho}_{\text{gov}}$ (right) for D^{us}. For each experiment we test with each of the unimodal, SIR, and Wavfinder segmentation methods.

Given that the causal mechanism for the spread of Covid-19 is respiratory contact, we hypothesize that clusters should be formed by locations which are close together – and for which contact is therefore more likely. We expect that at times during the pandemic long distance travel was common enough for this not to be true. However we claim that especially during periods of intense quarantine, the overall trend should be geographically similar. To check for this, we'll measure the geography with a distance in miles d_{miles} between any two locations. More specifically, for any two wave segments w_{ℓ_1}, w'_{ℓ_2} the $d_{\text{miles}}(w_{\ell_1}, w'_{\ell_2})$ distance computes the distance in miles between locations ℓ_1 and ℓ_2. To test the assumption, we'll compute a geographic score for each cluster as the average pairwise distance in miles $\rho_{\text{geo}}(\mathcal{C}_i) = \frac{1}{\mathcal{C}_i} \sum_{(w_{\ell_1}, w'_{\ell_2}) \in \mathcal{C}_i} d_{\text{miles}}(w_{\ell_1}, w'_{\ell_2})$. The geographic coefficient for a clustering, $\bar{\rho}_{\text{geo}}(\mathcal{C})$ is then taken to be the average of scores over all clusters in $\mathcal{C}$.

Assumption 3. *The locations for a cluster's waves should have similar governmental response policies during the time period of the wave*

Finally we expect that a local government's response to Covid-19 has some significant effect upon how that location experiences infection. For example, a location whose government imposed strong public health policies might have a much less intense wave of infection compared to a location which was more relaxed in its containment of the virus. To measure this, we use data from [16] which records a time series of government response scores for each location in D^{us} and D^{eu} (unavailable for D^{co}) along the same time-period. Specifically, we use datasets D^{us}_{gov} and D^{eu}_{gov} which are identical in shape to the original datasets, but which have entries replaced with government response scores. We'll then use the same segmentation boundaries as were used originally for our clustering to split these datasets into corresponding government response waves h. A government response clustering $\mathcal{C}^{\text{gov}}$ is then created by replacing every infection wave segment w with its corresponding government response segment h, while keeping the overall structure of the clustering the same. To measure if our methodology produces clusters with distinct governmental response policies, we compute average pairwise distances among response waves using the dynamic time warp distance: $\rho_{\text{gov}}(\mathcal{C}^{\text{gov}}_i) = \frac{1}{\mathcal{C}^{\text{gov}}_i} \sum_{(h,h') \in \mathcal{C}^{\text{gov}}_i} d_{\text{dtw}}(h, h')$. To measure for a clustering, we will again compute the average over all clusters as $\bar{\rho}_{\text{gov}}(\mathcal{C}^{\text{gov}})$.

These 3 assumptions are tested for each of the non-random implemented segmentation models – unimodal, SIR, and wavefinder – which are clustered using `clique-cluster()` . We then test these observed values against a few random baselines. Firstly, we experiment with keeping each of segmentation methods, but randomly clustering with `random-clique-cluster()` . Next, we replace the implemented segmentation methods with the randomized `k-random-segment()` and again cluster normally with `clique-cluster()` . Finally, we randomize both parts of the methodology by using both `k-random-segment()` and `random-clique-cluster()` . For each of the randomized methods we compute 1,000 samples to estimate distributions for each of $\bar{\rho}_{\text{sil}}$, $\bar{\rho}_{\text{geo}}$, and $\bar{\rho}_{\text{gov}}$. We limited ourselves to this relatively small number of samples given the computational difficulties associated with the task. Importantly, any time we use `k-random-segment()` we default to k wave segments used for each location in the unimodal model. In general, whenever parameters are required for segmentation we resort to whatever is used for the unimodal model. Results for the clustering of D^{us} are shown in figure 4.

By our assumptions, we expected that compared to the random methods, the non-random observed values would have larger silhouette coefficient $\bar{\rho}_{\text{sil}}$, and that they would have both a smaller geographic $\bar{\rho}_{\text{geo}}$ and government response $\bar{\rho}_{\text{gov}}$ coefficients. In this example, we do indeed see that our assumptions seem to be affirmed. The non-random implementations for all segmentation methods show statistically significant results across all three measurements when compared to their randomized versions. The difference is most striking for silhouette coefficients, where the only randomized method that comes close is random segmentation with non-random clustering – indicating that, as expected, our clustering algorithm is doing most of the work to improve this score. We note, however, that this is not the case for D^{eu}. Although the conclusions for the

silhouette coefficient are the same, the other geographic and government response validation measures no longer hold. We point the reader to the appendix for visualizations and more information on differences between segmentation methods, as well as complete results for D^{eu} and D^{co} datasets.

6 Discussion

For US state level data, D^{us}, we found strong reason to believe that waves which are close with respect to dynamic time warping distance are also similar both geographically and in terms of their public health response to the pandemic. We believe that one of the most interesting parts of our methodology is to showcase how this relationship changed over time. From results showcased in figure 3, we conclude that geographical similarity between clusters was dynamic, changing significantly as the virus evolved and the country's response fluctuated. However, results for other the other datasets D^{eu} and D^{co} are varied. For European countries we were surprised to find that clusters are relatively dissimilar in both geography and government response when compared against a random clustering baseline. This may point a difference in how the two regions of the world experienced and responded to the pandemic. However we are careful not to make any strong conclusions since it may also be an artifact of the data or of the geographical shape of the region. For US counties data D^{co} our methodology performs much worse, finding scattered and uninterpretable clusters, perhaps because the data is so similar between counties and there are very few distinguishable patterns. Despite these analytical difficulties, we argue that our methodology motivates an important problem which might even be applied to other, related spatio-temporal datasets. We argue that the work contained in this study introduces a novel, useful framework to study the progression of the Covid-19 pandemic and for spatio-temporal occurrences in general.

References

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information systems*, 53:16–38, 2015.

[2] Rozhin Amin, Mohammad-Reza Sohrabi, Ali-Reza Zali, and Khatereh Hannani. Five consecutive epidemiological waves of covid-19: a population-based cross-sectional study on characteristics, policies, and health outcome. *BMC infectious diseases*, 22(1):906, 2022.

[3] Ryan Baxter-King, Jacob R Brown, Ryan D Enos, Arash Naeim, and Lynn Vavreck. How local partisan context conditions prosocial behaviors: Mask wearing during covid-19. *Proceedings of the National Academy of Sciences*, 119(21):e2116311119, 2022.

[4] Richard Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961.

[5] US Census Bureau. Centers of population. `https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html`, 2020.

[6] Ewen Callaway. Covid's future: mini-waves rather than seasonal surges. *Nature*, 2023.

[7] Alex Capaldi, Samuel Behrend, Benjamin Berman, Jason Smith, Justin Wright, and Alun L Lloyd. Parameter estimation and uncertainty quantication for an epidemic model. *Mathematical biosciences and engineering*, page 553, 2012.

[8] Jianmin Chen and Panpan Zhang. Clustering us states by time series of covid-19 new case counts in the early months with non-negative matrix factorization. *Journal of Data Science*, 20(1):79–94, 2022.

[9] Ariel Cintrón-Arias, Carlos Castillo-Chávez, Luis Betencourt, Alun L Lloyd, and Harvey Thomas Banks. The estimation of the effective reproductive number from disease outbreak data. Technical report, North Carolina State University. Center for Research in Scientific Computation, 2008.

[10] Tad A Dallas, Grant Foster, Robert L Richards, and Bret D Elderd. Epidemic time series similarity is related to geographic distance and age structure. *Infectious Disease Modelling*, 7(4):690–697, 2022.

[11] Regis Darques, Julie Trottier, Raphael Gaudin, and Nassim Ait-Mouheb. Clustering and mapping the first covid-19 outbreak in france. *BMC Public Health*, 22(1):1279, 2022.

[12] M. Frisen. Unimodal regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 35(4):479–485, 1986.

[13] Kimiya Gohari, Anoshirvan Kazemnejad, Ali Sheidaei, and Sarah Hajari. Clustering of countries according to the covid-19 incidence and mortality rates. *BMC Public Health*, 22(1):632, 2022.

[14] David Guijo-Rubio, Antonio Manuel Durán-Rosal, Pedro Antonio Gutiérrez, Alicia Troncoso, and César Hervás-Martínez. Time-series clustering based on the characterization of segment typologies. *IEEE transactions on cybernetics*, 51(11):5409–5422, 2020.

[15] N. Haiminen and A. Gionis. Unimodal segmentation of sequences. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 106–113, 2004.

[16] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, et al. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature human behaviour*, 5(4):529–538, 2021.

[17] John Harvey, Bryan Chan, Tarun Srivastava, Alexander E Zarebski, Paweł Dłotko, Piotr Błaszczyk, Rachel H Parkinson, Lisa J White, Ricardo Aguas, and Adam Mahdi. Epidemiological waves-types, drivers and modulators in the covid-19 pandemic. *Heliyon*, 9(5), 2023.

[18] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. In *Data mining in time series databases*, pages 1–21. World Scientific, 2004.

[19] Eamonn Keogh and Jessica Lin. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and information systems*, 8:154–177, 2005.

[20] Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8):799–810, 1995.

[21] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[22] Zhixue Luo, Lin Zhang, Na Liu, and Ye Wu. Time series clustering of covid-19 pandemic-related data. *Data Science and Management*, 6(2):79–87, 2023.

[23] Troy McMahon, Adrian Chan, Shlomo Havlin, and Lazaros K Gallos. Spatial correlations in geographical spreading of covid-19 in the united states. *Scientific reports*, 12(1):699, 2022.

[24] Taneli Mielikäinen, Evimaria Terzi, and Panayiotis Tsaparas. Aggregating time partitions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 347–356, 2006.

[25] Teshager Zerihun Nigussie, Temesgen T Zewotir, and Essey Kebede Muluneh. Detection of temporal, spatial and spatiotemporal clustering of malaria incidence in northwest ethiopia, 2012–2020. *Scientific reports*, 12(1):3635, 2022.

[26] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[27] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814–818, 2005.

[28] Kevin Quinn, Evimaria Terzi, and Mark Crovella. Characterizing covid waves via spatio-temporal decomposition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3783–3791, 2022.

[29] Richard F Raubertas. Spatial and temporal analysis of disease occurrence for detection of clustering. *Biometrics*, pages 1121–1129, 1988.

[30] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[31] Satu Elisa Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

[32] Romain Tavenard. An introduction to dynamic time warping. `https://rtavenar.github.io/blog/dtw.html`, 2021.

[33] Evimaria Terzi and Panayiotis Tsaparas. Efficient algorithms for sequence segmentation. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 316–327. SIAM, 2006.

[34] Loring J Thomas, Peng Huang, Fan Yin, Xiaoshuang Iris Luo, Zack W Almquist, John R Hipp, and Carter T Butts. Spatial heterogeneity can lead to substantial local variations in covid-19 timing and severity. *Proceedings of the National Academy of Sciences*, 117(39):24180–24187, 2020.

[35] Stefan Thurner, Peter Klimek, and Rudolf Hanel. A network-based explanation of why most covid-19 infection curves are linear. *Proceedings of the National Academy of Sciences*, 117(37):22684–22689, 2020.

[36] O. Wahltinez et al. Covid-19 open-data: curating a fine-grained, global-scale data repository for sars-cov-2. 2020. `https://goo.gle/covid-19-open-data`.

[37] Vasilios Zarikas, Stavros G Poulopoulos, Zoe Gareiou, and Efthimios Zervas. Clustering analysis of countries using the covid-19 cases dataset. *Data in brief*, 31:105787, 2020.

[38] Stephen X Zhang, Francisco Arroyo Marioli, Renfei Gao, and Senhu Wang. A second wave? what do people mean by covid waves?–a working definition of epidemic waves. *Risk Management and Healthcare Policy*, pages 3775–3782, 2021.

7 Appendix

7.1 Comparison of Segmentation Methods

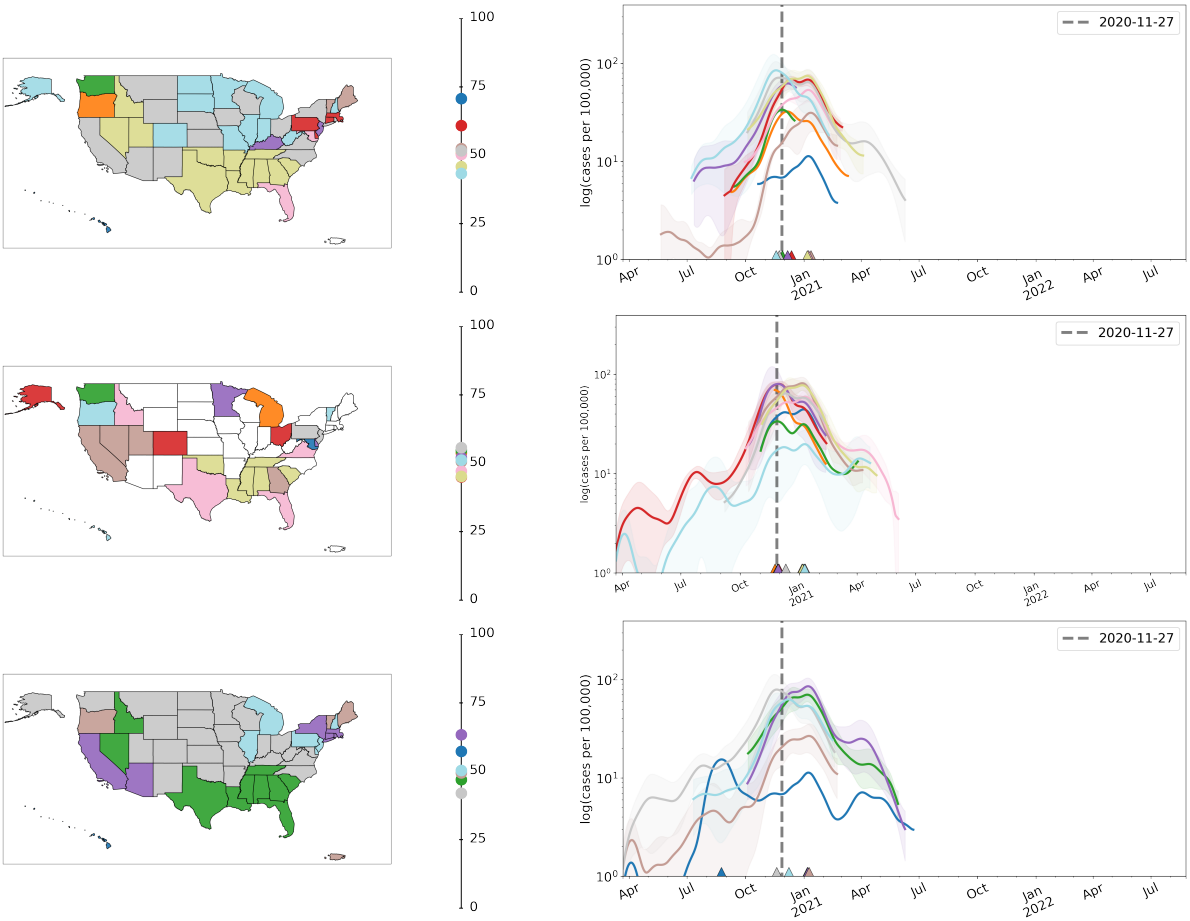

Figure 5: Clusters present at 11/27/2020 for clusterings computed with unimodal (top), SIR (middle), and wavefinder (bottom) segmentations.

Importantly we'd like to understand the differences in clustering produced by each of our segmentation methods. For each segmentation method we select values for the time overlap parameter δ and the distance threshold

parameter η by computing silhouette scores $\bar{\rho}_{\text{sil}}$ over a grid of options and selecting the pair which maximizes the score. We then run `clique-cluster()` with these selected values until we reach a stopping point where no two clusters may be joined to form a clique. Importantly, the clusterings from each of the segmentation methods are comprised of different items (wave segments) and can have very different sizes both on an individual cluster scale and as a whole. This makes it tricky to compare between clusterings and select an appropriate segmentation method.

In figure 5 we visually compare the clusterings produced by each of the segmentation methods for D^{us}. To do so we select a single time stamp, 11/27/2020, and display clusters which are fully present with maps on the left as well as their average wave segments on the right. In some form, all three clusterings capture a significant wave event happening around this time stamp. However, the geographical patterns are different for each. unimodal and wavefinder segmentation methods show the most agreement, with strongly cohesive clusters in the south and parts of the midwest. SIR, on the other hand, seems to fall out of line with the others. It produces a clustering with much less geographical cohesiveness.

Similarity in actual *segmentations* between unimodal and wavefinder, as discussed in section 4.1, may play a role in these observed outcomes. For our experiments we strongly relied on the unimodal segmentation method, since it seemed to be both validated by similarity to wavefinder while also satisfying our desire for carefully modeled wave segments. While the SIR method had the benefit of being a epdemiologically motivated model, we did not find much evidence to support its use in our experiments.

7.2 European Countries

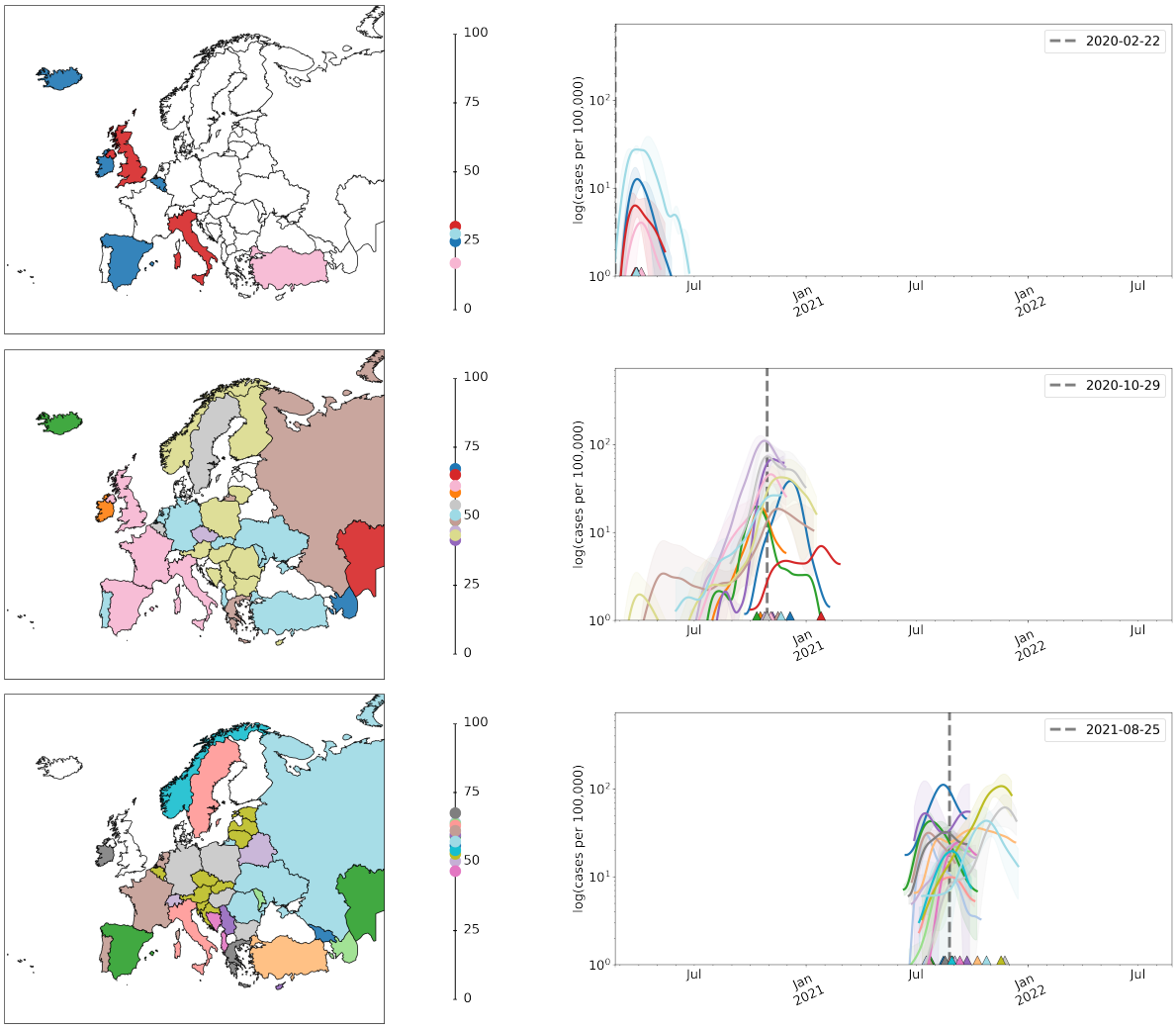

Figure 6: D^{eu} clustering over time with Unimodal segmentation. In each row we show a representation of the clusters present for chosen time, evolving over time. We show a geographic map (left) where we assign a color to each of the current clusters and color the locations represented in each of them. We also report average government response scores (middle) on a scale from 0 to 100 for each of the clusters, and their average waves of infection over time on a log scale (right).

Clustering with the unimodal segmentation method is also shown for D^{eu} in figures 6 7 and comparisons against random baselines in figure 8. Looking at results we notice some patterns which are different than for D^{us}. Some of this may be caused by the fact that this dataset has infection data which reaches back further in time than the state dataset did. Here we start our analysis in February of 2020, finding only four clusters fully present this early on, and seeing very little geographic cohesiveness (Note: the light blue cluster which is too small to see on the map is represented by locations Andorra and San Marino). As time progresses the pandemic became much more widespread across the continent. At the time stamps of 10/29/2020 we see a large wave experienced by almost all of Europe, and which does show on the map to be somewhat geographically contained. For example, locations in western Europe such as the UK, France, Spain, and Italy all cluster together. Continuing in time, however, we notice that this pattern does not hold. Throughout the rest of the pandemic we see much less geographical cohesiveness. While we occasionally we see some similarity in eastern Europe, the pattern does not seem to be broad enough to support the same kind of claims we make for US states.

In fact, looking to our validations against random baselines in figure 8, we notice that while our methods do seem to be significant with respect to the silhouette coefficient, the same cannot be said for both the geographic and government response coefficients. In other words, while we can claim that the clustering is producing groups of waves which are distinct, we cannot also say that those groups similar geographically or in terms of government response. A random clustering, may in fact produce clusters which are similar on both of these

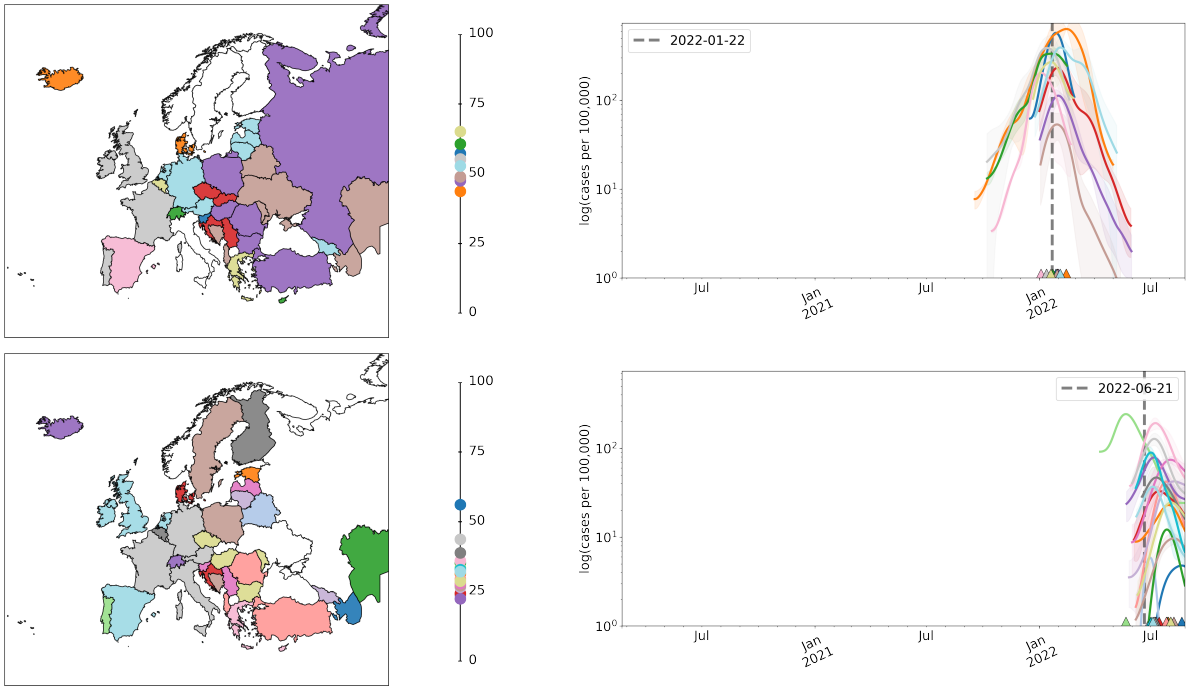

Figure 7: Extension of figure 6.

measures.

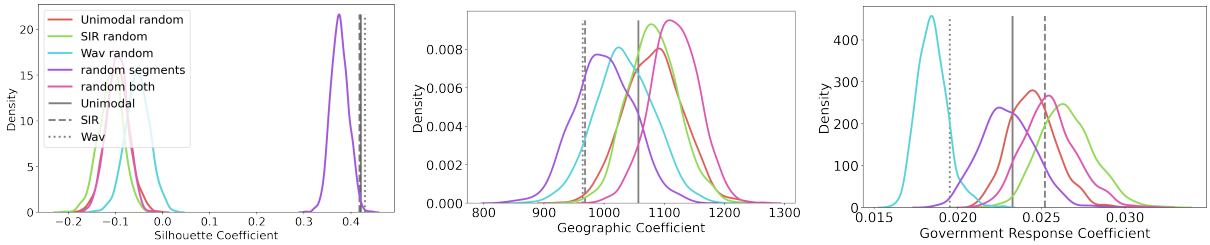

Figure 8: Comparisons against random clustering algorithms for assumptions on $\bar{\rho}_{\text{sil}}$ (left), $\bar{\rho}_{\text{geo}}$ (middle), and $\bar{\rho}_{\text{gov}}$ (right) for D^{eu}. For each experiment we test with each of the unimodal, SIR, and Wavfinder segmentation methods.

7.3 Northeast US Counties

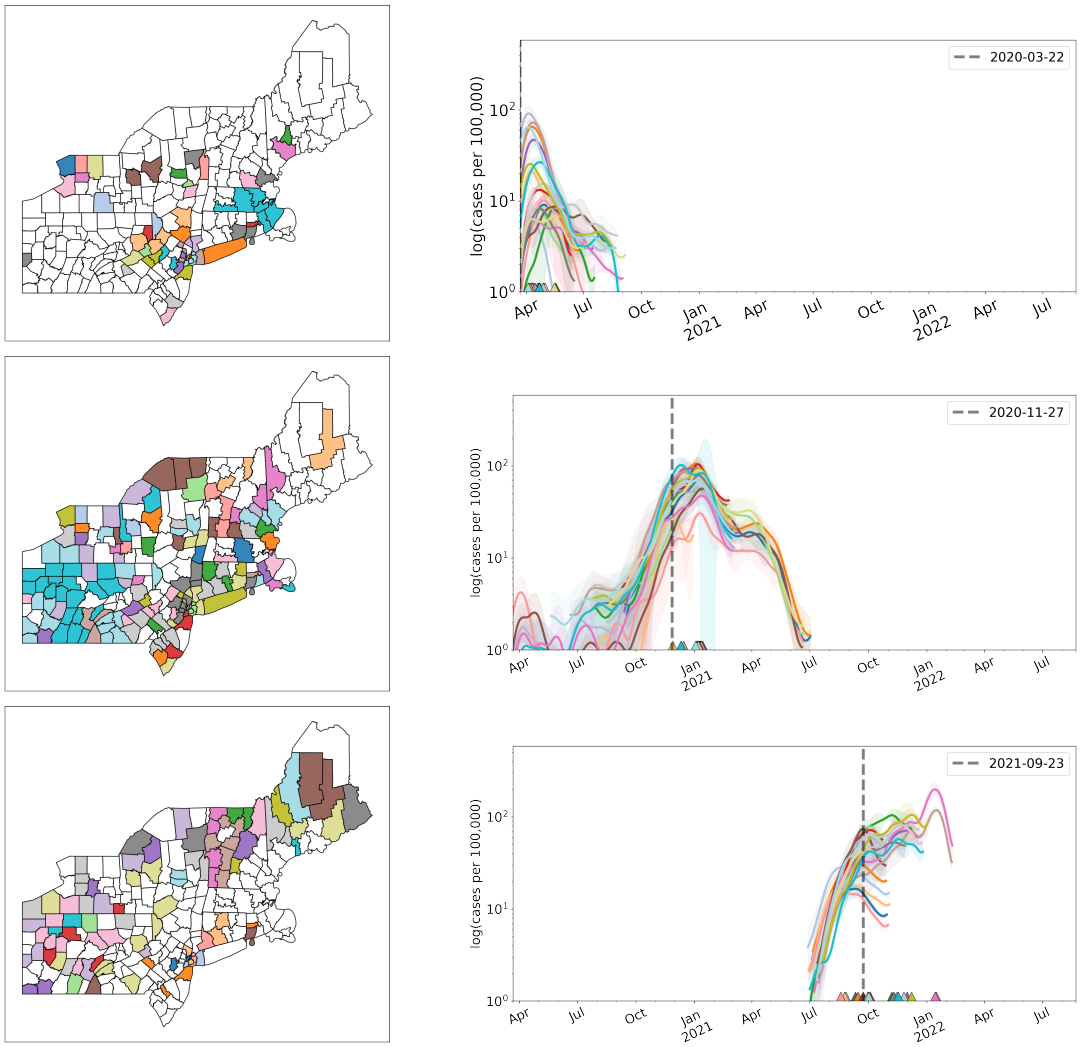

Figure 9: Northeast D^{co} clustering over time with Unimodal segmentation. In each row we show a representation of the clusters present for chosen time, evolving over time. We show a geographic map (left) where we assign a color to each of the current clusters and color the locations represented in each of them. We also show their average waves of infection over time on a log scale (right).

Taking another point of view, we were interested to see how results may change if we move from a state level to a smaller county level perspective of the pandemic. In this experiment we selected a subset of US counties coming from the northeast region of the country, including all from the states of Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont.

We find that what was most striking in this experiment was how similar the waves appear to be in each of the clusters. In the beginning of the pandemic there was good separation between each of the cluster's waves, but as time progresses things look nearly the same. This is also represented in the cluster maps, in which clusters are often scattered across the entire region.

This being said, we computed a limited random baseline test for this experiment, which only includes results from the unimodal segmentation method (the reason being that this was a much more computationally costly experiment to run) and for the silhouette and geographic validation measures (since government response data was not collected on the US county level). We observe that clusters do in fact seem to be significantly similar in geography. However, for the aforementioned reasons, its much more difficult to decide whether this is a strong conclusion or is otherwise just an artifact of the data and the geographical patterns of the region.

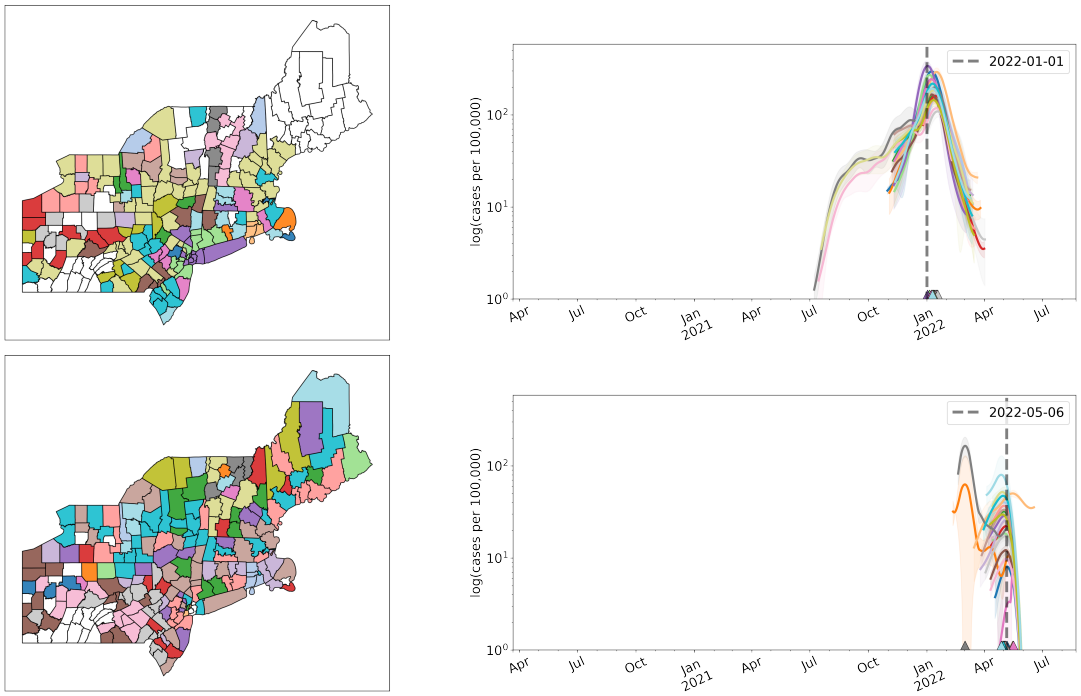

Figure 10: Extension of figure 6.

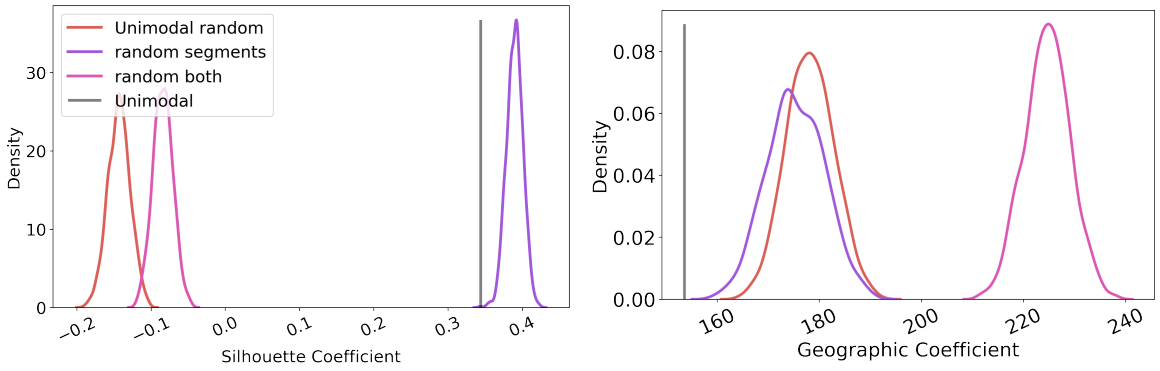

Figure 11: Comparisons against random clustering algorithms for assumptions on $\bar{\rho}_{\text{sil}}$ (left), $\bar{\rho}_{\text{geo}}$ (right). In this experiment we only tested with the unimodal segmentation method.