Aligning Time-series by Local Trends: Applications in Public Health

Ajitesh Srivastava

University of Southern California
ajiteshs@usc.edu

Abstract

Individual models of infectious diseases or trajectories coming from different simulations may vary considerably, making it challenging for public communication and supporting policy-making. Therefore, it is common in public health to first create a consensus across multiple models and simulations through ensembling. However, current methods are limited to mean and median ensembles that perform aggregation of scale (cases, hospitalizations, deaths) along the time axis, which often misrepresents the underlying trajectories – e.g., they underrepresent the peak. Instead, we wish to create an ensemble that represents aggregation simultaneously over both time and scale and thus better preserves the properties of the trajectories. This is particularly useful for public health where time-series have a sequence of meaningful local trends that are ordered, e.g, a surge to an increase to a peak to a decrease. We propose a novel alignment method DTW+SBA, which combines a representation of local trends along with dynamic time warping barycenter averaging. We prove key properties of this method that ensure appropriate alignment based on local trends. We demonstrate on real multi-model outputs that our approach preserves the properties of underlying trajectories. We also show that our alignment leads to a more sensible clustering of epidemic trajectories.

Code — https://github.com/scc-usc/DTW_S_apps

1 Introduction

In supporting policy-making to assess the impact of decisions and scenarios during an epidemic, a single model is seldom used (US SMH 2022, 2020; Borchering et al. 2021). Instead, multiple models are consulted, each of which produces multiple potential trajectories of epidemic burden (e.g., projected deaths and hospitalizations over time). Individual models or trajectories coming from different simulations may vary considerably, making it challenging for public health communication and supporting policy-making. Therefore, it has become common in public health to first create a consensus across multiple models and simulations through ensembling. However, current methods are limited to mean and median ensembles that suffer from several issues. They perform aggregation of severity (cases, hospitalizations, deaths) along the time axis. These ensembles are designed to capture the mean value at time t, i.e., for n trajectories $\mathbf{a}_1, \ldots, \mathbf{a}_n$, where $\mathbf{a}_i = [a(1), \ldots, a(n)]$ the ensemble is $\bar{a}(t) = \sum_{i=1}^{n} a(t)/n$. As an unintended consequence, informative aspects of individual trajectories may be lost. As an example, consider Figure 1. Two models produce almost identical projections but they are shifted in time, and they have the same peak. The ensemble produces a trajectory that has a peak that is significantly lower and wider than individual models. In public health communication, this can cause a misjudgment of the severity of the epidemic. While the ensemble correctly summarizes the expected outcome at time t, the reader tends to infer other information, such as peak timing and severity. Thus, the current approach results in an ensemble that misrepresents the underlying trajectories, e.g., underrepresenting the peak.

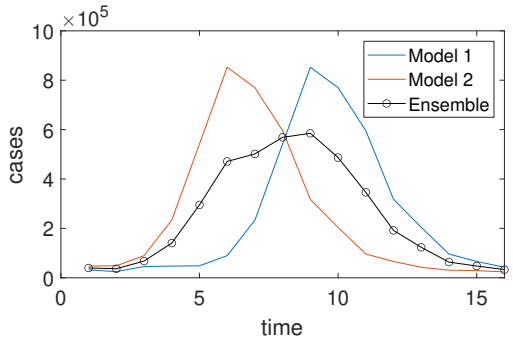


Figure 1: Failure of the mean ensemble in capturing the properties of individual time-series – much lower peak.

Our objective is to build ensemble models that are better representations of underlying trajectories. Based on our collaborations with public health experts, a necessity in any new ensemble approach is its **interpretability**. That is, it should be easily understood to someone who is not a computational scientist, how the ensemble is being created. We propose an ensemble approach that is able to summarize information in two dimensions – across scale as well as time (e.g., when a peak is expected to happen and how high). To do so, it finds an alignment of points across the trajectories that represent "similar local trends". The key to the alignment is the identification of similar trends. We use the Shapelet-space representation (Srivastava, Singh, and Lee 2022) to

transform the time-series into an interpretable matrix, where each column represents the local trend. The alignment is performed across the columns of all the trajectory representations similar to Dynamic barycenter averaging (Petitjean, Ketterlin, and Gançarski 2011). The approach is being considered for ensemble development in various Scenario Modeling Hubs (US SMH 2022; European CDC 2022; US SMH 2020) that inform public health experts and policymakers.

We ensure that our approach is grounded in theory, and yet the results are easy to interpret. We develop a theoretical notion of interpretable similarity measure and show that the similarity measured in shapelet-space representation (Srivastava, Singh, and Lee 2022) which forms the core of our alignment method, satisfies this notion. We compare our alignment method against several baselines, including currently used ensembling methods in public health on 12 real-world datasets.

Contributions. (1) We *prove necessary and sufficient conditions* for the shapelet space representation (Srivastava, Singh, and Lee 2022) to be closeness-preserving – two local trends are similar if and only if they are mapped to nearby points (Section 3.2). This forms the core of aligning local trends. (2) We develop an ensembling technique DTW+S BA that combines our trend similarity, Dynamic Time Warping (Müller 2007) and barycenter averaging (Petitjean, Ketterlin, and Gançarski 2011). It can simultaneously summarize time-series in scale and time (Section 3.3), and we demonstrate its utility on ensembling epidemic curves (Section 4). (3) We demonstrate that the distance obtained by our alignment (DTW+S) results in a more sensible clustering (Section 5.2) and improves the measurement of similarity between epidemic projections.

2 Background and Related Work

Multi-model Scenario Modeling Hubs Several Scenario Modeling Hubs exist around the world (US SMH 2020; European CDC 2022; US SMH 2022) that identify the policy questions of interest regarding respiratory viruses and generate scenario projections (e.g., health burden under potential transmissibility and vaccine uptake) to answer these questions. These projections are contributed by multiple independent teams, for instance, 10 teams, each contributing 100 trajectories for each scenario. An ensemble is built using the mean of outcomes at each time step to derive the consensus from these trajectories. The ensemble is then used to communicate to the policymakers. As shown in Figure 1, even if crucial information is present in the underlying models, the ensemble can ignore it. This poses a great challenge in policymaking, and there is a need for novel ensembles that overcome these issues. Our work solves this issue through a theoretically founded alignment technique.

Note that this work is not about forecasting. These projections are generated under specific scenarios that do not necessarily represent reality (due to some inherent uncertainty) but are intended to explore the impact of uncertain parameters and decisions. The goal of the ensemble is to summarize these projections. Therefore, our evaluation is not based on a comparison against some ground truth. Instead, the evaluation is on how well the ensemble captures the properties of the underlying projection trajectories. The trajectories we consider are generated by independent teams for Scenario Modeling Hubs (US SMH 2022; European CDC 2022; US SMH 2020) to inform public health policies.

Shapelet Space Representation Srivastava et. al. (Srivastava, Singh, and Lee 2022) introduce the idea of the shapelet space representation to compare short-term forecasts of epidemics. The motivation is to compare the shape of the forecasts rather than exact numerical values. Further, they wish to make the representation interpretable. Each dimension represents the similarity of the vector with one of the chosen shapes of interest, such as an increase $(1, 2, 3, 4)$ and peak $(1, 2, 2, 1)$. These shapes of interest are termed Shapelets.

Definition 1 (Shapelet)**.** *A shapelet* $\mathbf{s} = [s_1, \ldots, s_w] \in \mathbb{R}^w$ *is a vector that represents a shape of interest.*

Definition 2 (Shapelet-space Representation)**.** *Given* d *shapelets* $\{\mathbf{s_1}, \ldots \mathbf{s_d}\}$*, a Shapelet-space Representation of a vector* $\mathbf{x}$ *is a* d*-dimensional point* $P_x = (p_1, p_2, \ldots, p_d)$ *capturing the shape of* $\mathbf{x} \in \mathbb{R}^w$*, where the co-ordinate* $p_i = sim(\mathbf{x}, \mathbf{s_i})$ *for some measure of similarity. The function* $f : \mathbb{R}^w \to \mathbb{R}^d$ *is the Shapelet-space Transformation.*

The similarity function is to be chosen in such a way that two shapes are considered similar if and only if one shape can be approximated by translation and scaling of the other. However, this may cause an issue – when the shape is close to a "flat", small noise can cause it to become similar to other shapes. It is argued that there is an inherent concept of flatness in the domain of interest. For instance, in influenza when the number of hospitalizations is stable at a very low value, that shape is to be considered flat and not to be considered similar to any other shape when hospitalizations are higher. Therefore, a desirable property is the following.

Property 1 (Closeness Preservation)**.** *Two vectors have similar representation, if and only if (i) none of the vectors are "almost flat" and one can be approximately obtained by scaling and translating the other, or (ii) both vectors are "almost flat".*

They propose an approach that first identifies how similar a shape is to what we could consider "flat", and then updates the similarities of the shape with respect to other shapelets. For some constants $m_0, \beta \geq 0$, define "flatness" as $\phi = \exp(-\beta(m - m_0))$, if $m > m_0$, otherwise $\phi = 1$. Here m is the average absolute slope of the vector $\mathbf{x}$ whose shapelet-space representation is desired, i.e., if $\mathbf{x} = (x_1, x_2, x_3, x_4)$, then $m = (|x_2 - x_1| + |x_3 - x_2| + |x_4 - x_3|)/3$. The constant m_0 enforces that a vector with a very small average absolute slope is considered flat and receives a 0 similarity in all other dimensions. The constant β represents how quickly above the threshold m_0, the "flatness" should reduce. Now, the co-ordinates of shapelet-space representation are defined as

$$sim(\mathbf{x}, \mathbf{s_i}) = \begin{cases} 2\phi - 1, & \text{if } \mathbf{s_i} \text{ is "flat"}, \\ (1 - \phi) corr(\mathbf{x}, \mathbf{s_i}), & \text{otherwise.} \end{cases}$$

It is shown that this definition satisfies Closeness Preservation (Property 1) with w or more shapelets including the "flat" shapelet. *We prove that* w *shapelets are not only sufficient but necessary to satisfy this property (Theorem 2).* We

use Shapelet-space Representations of moving windows on the given time-series to capture local trends over time.

Dynamic Time Warping (DTW) DTW is a distance measure between two time-series that allows warping (local stretching and compressing) of the time component so that the two time-series are optimally aligned. Given two time-series $\mathbf{a} = [a(1), a(2), \dots]$ and $\mathbf{b} = [b(1), b(2), \dots]$, the objective of DTW is to minimize $\sum_{i \leftrightarrow j} \mathcal{D}(a(i), b(j)$, for some distance measure $\mathcal{D}$, and where $i \leftrightarrow j$ represents aligning index i of $\mathbf{a}$ with index j of $\mathbf{b}$. The alignment is done under some constraints – (1) if $a(i)$ and $b(j)$ are aligned then $a(i+1)$ cannot be aligned with $b(j')$ for some $j' < j$. (2) Every index is present in at least one alignment. (3) The first index of both time-series are aligned with each other. (4) The last index of both time-series are aligned with each other. Further, a window constraint can be added (Ratanamahatana and Keogh 2004) suggesting that indices i and j can only be aligned if $|i - j| \leq w$, for some non-negative integer w.

Shapelets In time-series literature, "shapelets" have been used to refer to informative motifs that occur in time-series (Ye and Keogh 2009). A feature vector for time-series can then be constructed by similarity of the best matching subsequence of the time-series to these motifs. The motifs are selected based on their representativeness of a class. In contrast, we use the term shapelet as *a pre-determined shape of interest coming from the domain*. We prove that a specific class of shapelet sets and similarity measures is needed to develop a representation to satisfy the closeness-preserving property. We encode all local trends of the time-series into a matrix representation, which is used in conjunction with Dynamic Barycenter Avergaing for alignment.

3 Methodology

3.1 Definitions

First, we define some terms used throughout. We start with the idea of a "trend descriptor" that formally defines the idea of assigning arbitrary label (e.g., increase, decrease, peak, etc.) to a part of a time-series. This is intended to emulate a human using categories (implicit or explicit) to interpret a pattern in the time-series. This concept will help us define interpretability of a representation and similarity measure.

Definition 3 (Trend Descriptor)**.** *A trend descriptor is a function $\mathcal{L}$ that maps any vector $\mathbf{x} \in \mathbb{R}^w$ to a label in set L denoting a shape description of $\mathbf{x}$.*

For instance, a trend descriptor may map any given 4-element vector to one of slow increase, rapid increase, exponential increase, going to peak, going past a peak, rapid decrease, flat, and unknown. Some of these labels may be more similar to each other, e.g., slow and rapid increases are more similar to each other then rapid increase and decrease.

Definition 4 (Local Trend)**.** *For a time-series $(a_1, a_2, \dots, a_T)$, a window w and some trend descriptor $\mathcal{L}$ we define the local trend at location i as $\mathcal{L}(a_i, \dots, a_{i+w-1})$.*

Definition 5 (Interpretable)**.** *We say a representation is interpretable if it is possible to identify the local trends based on the values in each dimension of the representation.*

Definition 6 (Ordered Local Trend)**.** *A class of time-series has ordered local trends if the order of the local trends appearing in the time-series conveys key information. In other words, the similarity between two time-series implies similarity between the sequences of local trends.*

In Figure 5(c) both orange and blue curves have the same sequence of local trends that can be described as a sequence of increases then peak, followed by a decrease, then a surge, and increase, another peak and decrease. While the gray line is only a sequence of increases. Such characterization is important in understanding and communicating long-term projections of epidemics as they represent specific events of interest (Howerton et al. 2023; Borchering 2023). The interpretability and ordered local trends play a central role in our novel ensemble methods. Before we use an alignment for ensembling, we develop the theory to be able to capture local trends (shapes) of the time-series for any implicit trend descriptor. To do so, we extend the notion of Shapelet-space Representation to a time-series with a sliding window.

Definition 7 (Shapelet-space Representation - Time-series)**.** *Given a time-series $\mathbf{a} \in \mathbb{R}^{T_1}$, a window w, and a Shapelet-space Transformation $f : \mathbb{R}^w \to \mathbb{R}^d$, the Shapelet-space Representation of $\mathbf{a}$ is the matrix $\mathbf{A} \in \mathbb{R}^{d \times (T_1 - w + 1)}$ whose i^{th} column is the Shapelet-space Representation of the vector $(a_i, \dots, a_{i+w-1})$.*

This matrix encodes how the time-series changes over time in an interpretable manner. Algorithm 1 summarizes this approach. For each time-series, $\mathbf{a}_i \in \mathbb{R}^{T_1}$, we first find its Shapelet-space Representation – the matrix $\mathbf{A}_i \in \mathbb{R}^{d \times (T_1 - w + 1)}$. Each column (of size d) of these matrices is obtained by sliding a w-length window on the respective time-series and obtaining its d-dimensional Shapelet-space Representation (SSR). Figure 2 shows the SSR obtained from a time-series. The SSR is built using four dimensions representing "increase", "peak", "surge", and "flat". A yellow color represents a high positive value and a blue represents a high negative value (e.g., a negative increase is a decrease). Also, note that "surge" and "increase" are similar shapes and hence seem to have a high correlation. The representation tells us that this time series has a sequence of surges/increases leading to a small peak (green in "peak" and "flat") around time step 5, followed by stability, then increase, leading to a sharp peak (bright yellow around time-step 13), followed by rapid decline (dark blue in "inc") and then stability (flatness).

Algorithm 1: Extracting shapelet space representation of time-series using a moving window

procedure TIMESERIESSHAPE($\mathbf{a}, \mathbf{S}$)
  $w \leftarrow$ length of each shapelet in $\mathbf{S}$
  $T \leftarrow$ length of the time-series $\mathbf{a}$
  for $t = 1$ to $T - w + 1$ **do**
    $\mathbf{A}[:, t]$ = ShapeTransform($\mathbf{a}[i : i + w - 1], \mathbf{S}$) ▷ Transform a window of length w into the shapelet space
  return $\mathbf{A}$

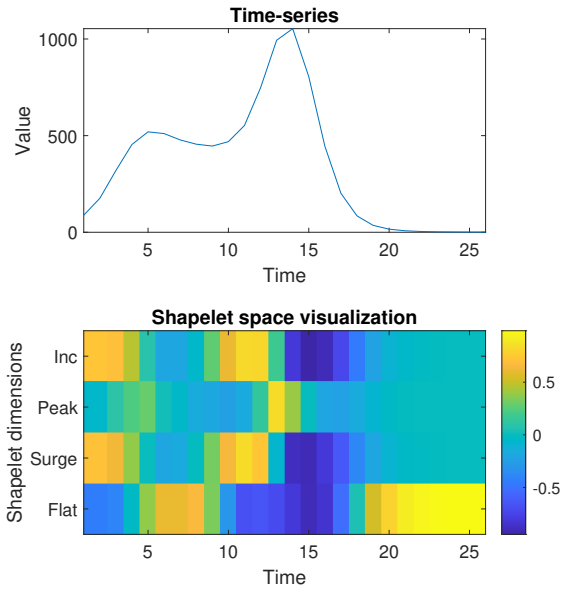


Figure 2: Shapelet-space Representation of a time-series.

3.2 Choosing the Shapelet-Space

While Srivastava et. al. provide some indication of how to choose the set of shapelets, they only prove that for vectors with w elements, w shapelets are sufficient. While one may choose more number of shapelets, e.g., one for each trend descriptor in mind, having a large number (d) of shapelets also impacts the space and time complexities as both scale linearly with d. What is the minimum number of shapelets needed? Here, we show that w shapelets are not only sufficient but also necessary to satisfy the closeness-preserving property. Let $f : \mathbb{R}^w \rightarrow \mathbb{R}^d$ be a shapelet transformation obtained by a set of linearly independent vectors $\mathbf{s_1}, \ldots, \mathbf{s}_{d-1}$ and the flat vector $\mathbf{s}_0$. Consider two vectors $\mathbf{x}$ and $\mathbf{y}$ of length w. Suppose $\mathbf{x}'$ and $\mathbf{y}'$ represent the corresponding normalized vectors obtained as: $\mathbf{x}' = \frac{\mathbf{x}-\mu_\mathbf{x}}{\|\mathbf{x}\|}$ and $\mathbf{y}' = \frac{\mathbf{y}-\mu_\mathbf{y}}{\|\mathbf{y}\|}$, where $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ are the mean of elements in $\mathbf{x}$ and $\mathbf{y}$, respectively. The following can be shown (Srivastava 2023).

Theorem 1. *Property 1 is satisfied with any set of* $w-1$ *linearly independent shapelets and the "flat" shapelet, i.e., with this choice* $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq \epsilon$ *iff (i) both* $\mathbf{x}$ *and* $\mathbf{y}$ *are "almost" flat, or (ii)* $\|\mathbf{x}' - \mathbf{y}'\| \leq \delta$*, for some small* ϵ *and* δ*.*

Theorem 2. *At least* $w-1$ *linearly independent shapelets are necessary with the "flat" shapelet to satisfy Property 1.*

Next, we show that the proper choice of shapelets results in the desired expressive power of distinguishing any two labels of an arbitrary trend descriptor.

Theorem 3 (Expressive power)**.** *A proper choice of Shapelet Space Transformation can discriminate any trend descriptor.*

Proof. Due to the closeness preserving property, it follows that with w shapelets as described by Theorems 1 and 2, we can distinguish between any two local trends taken from an arbitrary choice of scale-free trend descriptor $\mathcal{L}$. By scale-free, we mean a trend descriptor that does not distinguish based on scale, e.g., "high increase" vs "very high increase". On the other hand, it should be noted that we do not completely ignore the scale information. The flat dimension can be rewritten as $sim(\mathbf{x}, \text{flat}) = 2\phi - 1 = 2\exp(-\beta m)$, choosing $m_0 = 0$. Given, the value in the flat dimension, we can uniquely find the average absolute slope m and thus we are able to discriminate scale-based trend descriptors as well. Therefore, by choosing the w shapelets appropriately, we can discriminate any two local trends taken from an arbitrary choice of $\mathcal{L}$. □

Remarks. Any representation obtained by a function that maps the input into an element of a finite set cannot have the desired expressive power. Suppose the cardinality of the range of this function is k. Then, any trend descriptor of $k+1$ labels will have at least two trends that cannot be distinguished by this function. For instance, consider a transformation from a vector to two categories based on the positive and negative slope (increase vs decrease). Then, a "peak" cannot be distinguished from either. While the shapes of interest may be pre-determined, we wish to be able to distinguish them regardless of their precise definition. Further, note that the simple measure of absolute error also has the same expressive power, but it may also discriminate between the same local trends (both "increasing" but with slightly different slopes). Shapelet space transformation puts more emphasis on scale-free discrimination.

It follows that if we align two time-series that have the same ordered local trends but possibly at different times, Dynamic Time Warping on their SSR (DTW+S) will be able to align them, leading to a low distance (high similarity).

3.3 Ensemble Generation

The existing ensemble methods are designed to aggregate individual projections over time, thus measuring the scale (e.g., number of hospitalizations) at time t. They are not designed to aggregate when will an event (e.g., a peak) take place. However, a viewer tends to interpret both the scale and timing from the ensemble plot. We are interested in – given n time-series, $\mathbf{a}_i = [a_i(1), a_i(2), \ldots, a_i(T)], i \in \{1, \ldots, n\}$, find an "ensemble" time-series that captures the aggregate behavior in both time and scale. To address this, we assume that *each trajectory tries to estimate a sequence of latent "events"*. With this perspective, for some sequence of event $e_1, e_2, \ldots$, a time-series captures the timing of e_j and its scale. Therefore, the time-series can be interpreted as $\mathbf{a}_i = [(t_i(e_1), a_i(e_1)), (t_i(e_2), a_i(e_2)), \ldots]$. For any given event e_j, the aggregate time-series can be obtained by averaging both the timing and the severity dimensions:

$$(\bar{t}(e_j), \bar{a}(e_j)) = \left(\frac{\sum_i t_i(e_j)}{n}, \frac{\sum_i a_i(e_j)}{n} \right) \quad (1)$$

However, we do not observe these "events" explicitly. We define an event to be reflected in the time-series by a local trend. When two time-series are aligned by DTW+S, each alignment corresponds to an event. Formally, in the shapelet

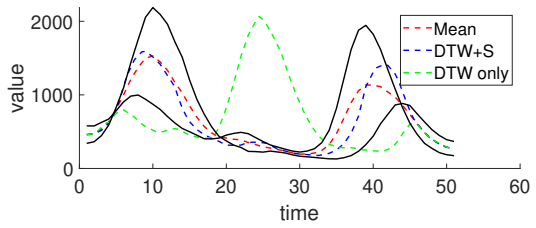


Figure 3: Applying mean, DTW, and DTW+S to develop ensemble of two time-series.

space representation $\mathbf{A}$ and $\mathbf{B}$, if columns τ_1 and τ_2 are aligned, then the local trend at time $[\tau_1, \tau_1 + w - 1]$ in time-series $\mathbf{a}$ and that at time $[\tau_2, \tau_2 + w - 1]$ in time-series $\mathbf{b}$ are defined to be corresponding to the same "event". Suppose we use DTW+S to align n projections. Then, each projection i contributes one point $(t_i(j), a_i(t_j))$ for each alignment j. This results in a set of points, one for each alignment j, using Equation 1. Finally, if desired, we can interpolate these points to estimate the value of $\bar{a}(t)$ for $t \in \{1, 2, \ldots, T\}$. Note that this approach is based on the following assumptions. First, each time-series has a similar sequence of shapes but may differ in timing and severity/scale. Second, the interpolation assumes smoothness in the desired ensemble.

Figure 3 demonstrates this approach for two time-series (in black solid line) that have two peaks each with different scales. We compare our approach against the mean ensemble and DTW (without SSR). Note that the mean ensemble results in a small second peak which is closer to the smaller peak among the input time-series. Further, its timing is biased towards the larger peak. The DTW ensemble results in one large peak by aggregating scale and timing of the first peak of one and the second peak of the other time-series. The DTW+S ensemble results in two peaks as expected, where each peak correctly averages the corresponding timing and scale of the input time-series. While similar aggregation has been discussed in the literature (Gupta et al. 1996; Petitjean, Ketterlin, and Gançarski 2011), two key differences exist. First, our application is motivated by visual interpretation of "events", such as when the epidemic peaks. Second, our approach uses an interpretable shape-based measure instead of directly using the Euclidean distance on time-series. This allows us to define an "event".

Optimally aligning multiple time-series is NP-Hard with some approximation algorithms including DTW Barycenter Averaging (Petitjean, Ketterlin, and Gançarski 2011). In this approach, the initial 'base' time-series is selected as the time-series among the set of trajectories that has the lowest distance from other trajectories. Then, in each iteration, we compute the pairwise alignment of all time-series with respect to the base time-series and the result becomes the new 'base' time-series. The alignment is traditionally computed using DTW, which we replace with DTW+S. Algorithm 2 summarizes this approach. We demonstrate that the DTW+S based Barycenter Averaging better captures the properties of the trajectories (Section 4).

Algorithm 2: Barycenter Averaging with DTW on Shapes

procedure DTW+SBA($\mathbf{a}, \mathbf{b}, \mathbf{S}, \tau$)
  for $i = 1$ to n **do**
    $\mathbf{A_i} \leftarrow$ TimeSeriesShape($\mathbf{a}_i, , \mathbf{S}$)
  $\mathbf{b} \leftarrow$ medoid($\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$)
  while not converge **do**
    $\mathbf{B} \leftarrow$ TimeSeriesShape($\mathbf{b}, \mathbf{S}$)
    for $i = 1$ to n **do**
      $\mathbf{t}_i \leftarrow$ DTW_align($\mathbf{A}_i, \mathbf{B}, \tau$) ▷ $\mathbf{t}_i$ holds the indices of $\mathbf{A}_i$ aligned to $\mathbf{B}[:,1], \mathbf{B}[:,2], \ldots, \mathbf{B}[:,n]$.
    for $j = 1$ to T **do**
      $\mathbf{b}[j] \leftarrow$ mean($\mathbf{a}[\mathbf{t}_1[j]], \mathbf{a}[\mathbf{t}_2[j]], \ldots, \mathbf{a}[\mathbf{t}_n[j]]$)
  return $\mathbf{b}$

4 Experiments

We conducted a series of experiments to demonstrate that DTW+SBA leads to a more reasonable ensemble that captures the scale and timing of events. In all experiments, unless stated otherwise, we used the following set of shapelets: (i) 'increase': $[1, 2, 3, 4]$, (ii) 'surge': $[1, 2, 4, 8]$, (iii) 'peak': $[1, 2, 2, 1]$, and (iv) 'flat': $[0, 0, 0, 0]$. According to Theorems 1 and 2, these shapelets satisfy Property 1. These were chosen because they are easily interpretable in the domain of epidemics. Some other sets of shapelets satisfying Property 1 were also tried, and their results were similar.

	Method	Mean ensemble	DTW BA	DTW (z-norm) BA	**DTW+S BA**
Peak Size	Set 1	-0.37	-0.04	-0.01	-0.02
	Set 2	-0.29	-0.33	-0.29	-0.01
	Set 3	-0.38	-0.31	-0.38	-0.03
	Set 4	-0.28	-0.22	-0.28	-0.01
	Set 5	-0.38	-0.23	-0.38	-0.03
	Set 6	-0.28	-0.21	-0.29	-0.02
	Set 7	-0.37	-0.18	-0.37	-0.03
	Set 8	-0.11	-0.13	-0.11	-0.02
	Set 9	-0.11	-0.13	-0.11	-0.02
	Set 10	-0.12	-0.14	-0.12	-0.02
	Set 11	-0.11	-0.10	-0.11	-0.02
	Set 12	-0.12	-0.11	-0.12	-0.02
Peak Timing	Set 1	0.08	-0.05	-0.01	0.08
	Set 2	-0.08	-0.10	-0.08	-0.03
	Set 3	-0.15	-0.11	-0.15	0.07
	Set 4	-0.09	-0.08	-0.09	-0.03
	Set 5	0.10	-0.09	-0.10	0.07
	Set 6	-0.09	-0.08	-0.09	-0.02
	Set 7	-0.10	-0.08	-0.10	0.07
	Set 8	-0.11	-0.17	-0.11	0.05
	Set 9	-0.13	-0.18	-0.13	0.03
	Set 10	-0.13	-0.18	-0.13	0.02
	Set 11	-0.14	-0.09	-0.14	0.02
	Set 12	-0.15	-0.10	-0.15	0.01

Table 1: Fractional error in estimation of peak size and timing. More yellow indicates higher error.

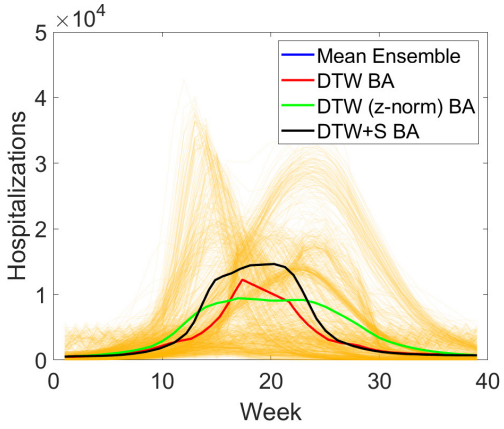
Datasets: We consider 12 sets of trajectories, for comparing our ensemble method with baselines. These sets include the following. Set 1: 75 trajectories from a model for In-

fluenza hospitalization. The length of each trajectory is 26 (weeks). Set 2-7: A total of 1000 trajectories (per set) from 10 influenza models for 6 scenarios each, extracted from Influenza Scenario Modeling Hub Round 4. The length of each trajectory is 39. Set 8-12: A total of around 1100 trajectories (per set) from 11 RSV models for six scenarios each extracted from RSV Scenario Modeling Hub Round 1. The length of each trajectory is 26 (weeks).

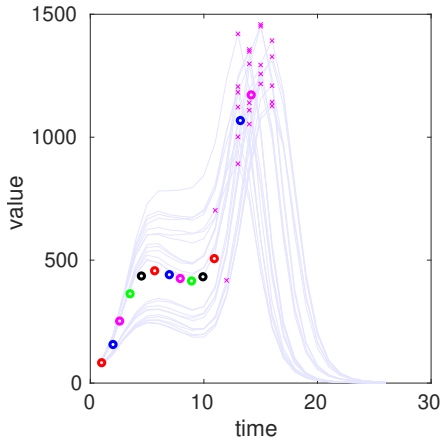
Methods: We compare the following approaches for ensembling. Mean ensemble: the most popular ensemble approach that simply averages values at each time-point (US SMH 2020; European CDC 2022); DTW BA: barycenter averaging with DTW; DTW (z-norm) BA: barycenter averaging with DTW after applying z-normalization to all trajectories; and DTW+S BA: barycenter averaging with DTW+S.

The resulting ensembles for Set 7 are shown in Figure 4(a). We observe that DTW+S BA produces the highest peak among all ensembles. Mean ensemble and DTW (z-norm) BA produce almost identical results (the green line overlaps with the blue line) and flatten the peak. To quantitatively compare the ensembles, we measure the fractional error of the ensembles in representing the peak timing and the peak size (i.e., the value and the time) at which the peak occurs. The ground truth is obtained by extracting the peak values (and timing) of all trajectories and averaging them. The results are presented in Table 1. Note that our approach is the only one that captures the peak timing and size of the underlying trajectories well for all sets.

Figure 4(b) shows an intermediate step of computing alignments for a subset of trajectories in Set 1. While DTW+DBA is not specifically designed to identify peaks, almost all peaks among different time-series are aligned (marked with a pink 'x'). The circle denotes the centroid, i.e., the ensemble point of these points. Thus, the ensemble provides a better estimate of the average size of the peak.



(a) Ensemble comparison



(b) Intermediate step

Figure 4: Ensembling results: (a) Different ensembling approaches on Set 7. The dark yellow lines represent the individual trajectories. (b) An instance of alignment on Set 1. Pink 'x' on the individual time-series are aligned to get the ensemble point (pink circle). Previous circles represent the ensemble points obtained from previous alignments

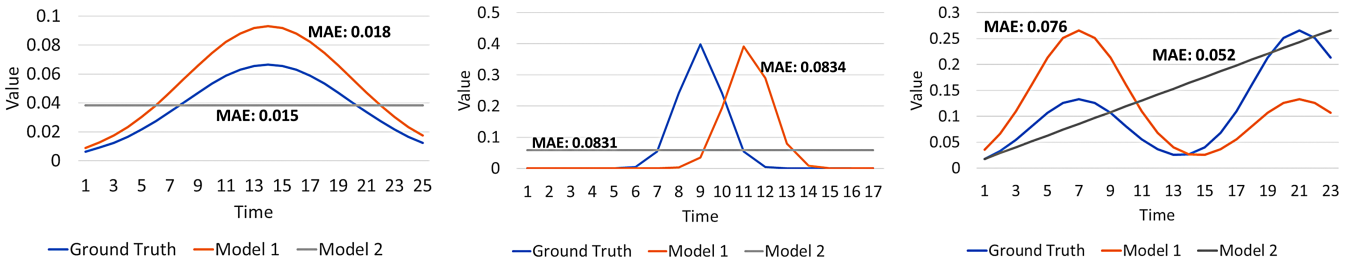
5 Discussion

5.1 Use as a Distance Measure

Consider the scenario presented in Figure 5(a). Two models perform a projection to estimate the time-series given by the ground truth. Model 1 produces a pattern that is similar to the ground truth, while Model 2 produces a flat line. If we use mean absolute error to assess which model performed better, Model 2 (flat line) will receive a better score. Although Model 1 produces identical trends and correctly predicts the peak timing, it loses to a Model 2 which conveys no information. Now, consider the scenario presented in Figure 5(b). Model 1 predicts the exact pattern but it slightly shifted in time. Again, Model 1 – a flat line, produces a lower error. Finally, in Figure 5(c), Model 1 predict the overall pattern well, it only misjudges the height of the peaks. Yet, Model 2, a straight line, is considered closer to the ground truth. Some form of a range normalization could have addressed the issue in Figure 5(a), and Dynamic Time Warping (DTW) (Müller 2007), which allows stretching the time dimension to best match two time-series, can address the issue raised in Figure 5(b). However, DTW and/or any normalization of scale cannot address the issue presented in Figure 5(c). DTW+S is an ideal measure in these cases as it measures the local trends putting more emphasis on their shapes rather than the scale. We will observe this effect next in the clustering of epidemic trajectories.

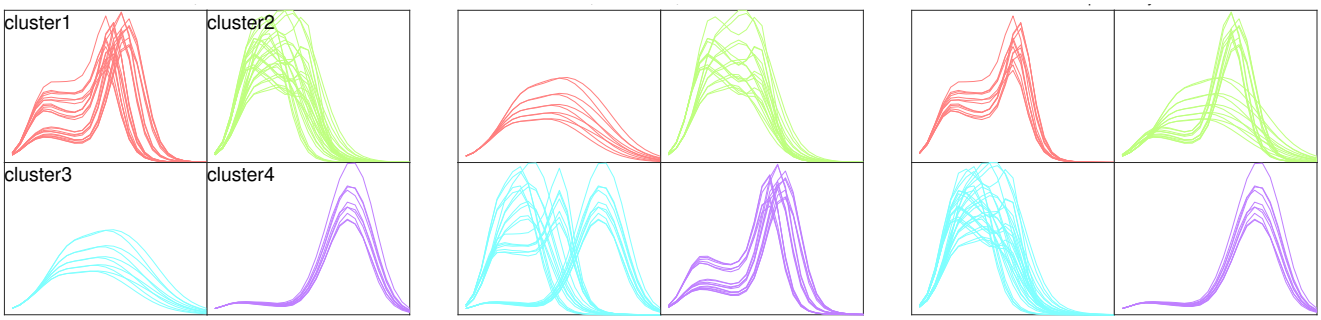
5.2 Clustering: A Qualitative Evaluation

We consider time-series projections for weekly influenza hospitalization for a US state by a model from Influenza Scenario Modeling Hub (US SMH 2022). It has 75 time-series, each corresponding to different choices of parameters and initialization. We calculated the dissimilarity matrix (all pair distances) using the following. (1) DTW+S: our method with infinite window for alignment; (2): DTW+S (cos): same as DTW+S, except that cosine distance is used instead of Euclidean for aligning SSRs; (3) DTW, normalized: Applying DTW after normalizing all time-series to zero mean and unit variance – a common normalization technique used



(a) Similar trend different scale (b) Similar scale, shifted in time (c) Similar trends with varying scales

Figure 5: Simple measures like Mean Absolute Error can be deceiving. In the three scenarios, Model 1 seems to be closer to the Ground truth, but receives a higher distance compared to a straight line.



(a) DTW+S, win = ∞ (b) Normalized DTW, win = ∞ (c) shapelet-only

Figure 6: Comparison of clustering results obtained from DTW+S against other distance measures. DTW+S with Euclidean (default) and cosine distance produces reasonable clustering. All other measures mix different patterns into one cluster.

with DTW (Chen et al. 2015); (4) DTW: DTW without any transformation or normalization; (5) Euclidean, normalized: Euclidean distance without any time warping on standard normal time-series; and (6) Shapelet-only: Euclidean distance on the SSR without any time-warping. For DTW+S, we generate hierarchical clusters with the number of clusters selected using the Silhouette Coefficient (Zhou and Gao 2014). We use the same number of clusters for clustering using each of the above dissimilarity measures.

Figure 6 shows the results of clustering. We observe that DTW+S produces four clusters with similarly shaped time-series. DTW with standard normalization mixes clusters 1, 2, and 4. The Shapelet-only measure mixes clusters 1 and 3. Omitted for brevity, DTW+S (cos) produces the same clustering (different ordering); DTW without normalization does not produce any discernable pattern in the trends and instead seems to group together those time-series that have similar peak heights; Euclidean distance with standard normalization mixes clusters 1, 2 and 4. Note that Shapelet transformation is not sufficient to capture similar time-series due to not being flexible across time. On the other hand, DTW with simple standard normalization cannot capture similar trends that occur at different scales. However, when they are combined in DTW+S, they produce reasonable clustering.

Limitations Extending our approach to other fields may require some domain knowledge to understand the appropriate choice of shapelets. However, Theorems 1 and 2 act as guidelines to ensure that the chosen shapelets satisfy the desired property of closeness preservation. Another limitation is the implementation – currently, we use $\mathcal{O}(dwT)$ time and space for DTW on the distance matrix obtained from SSR. Here, the T is the length of individual time-series, w is the window, and d is the number of shapelets. In future work, we will explore existing optimizations of DTW (Keogh and Pazzani 2000) and attempt to transfer them to DTW+S.

6 Conclusion

We have proposed a novel interpretable representation with a theoretical foundation to improve alignment of time-series.It can capture local trends and it is closeness-preserving. Based on this representation, we have developed an ensemble method that captures both the aggregate scale and timing of the individual time-series significantly better than the currently used mean ensemble and DTW-based barycenter averaging. This is central to public health communication and decision-making. We have also shown that our approach results in better clustering compared to other measures.

Acknowledgements

This work was supported by the Centers for Disease Control and Prevention and the National Science Foundation under the awards no. 2135784, 2223933, and 2333494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation and the Center for Disease Control and Prevention. We would like to thank the US and European Scenario Modeling Hub for useful discussions.

References

Borchering, R. 2023. FluSight 2023-2024. https://github.com/cdcepi/FluSight-forecast-hub.

Borchering, R. K.; Viboud, C.; Howerton, E.; Smith, C. P.; Truelove, S.; Runge, M. C.; Reich, N. G.; Contamin, L.; Levander, J.; Salerno, J.; et al. 2021. Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios-United States, April-September 2021. *MMWR. Morbidity and mortality weekly report*, 70(19): 719–724.

Chen, Y.; Keogh, E.; Hu, B.; Begum, N.; Bagnall, A.; Mueen, A.; and Batista, G. 2015. The UCR Time Series Classification Archive. www.cs.ucr.edu/~eamonn/time_series_data/.

European CDC. 2022. European COVID-19 Scenario Hub. https://github.com/covid19-forecast-hub-europe/covid19-scenario-hub-europe.

Gupta, L.; Molfese, D. L.; Tammana, R.; and Simos, P. G. 1996. Nonlinear alignment and averaging for estimating the evoked potential. *IEEE transactions on biomedical engineering*, 43(4): 348–356.

Howerton, E.; Contamin, L.; Mullany, L. C.; Qin, M.; Reich, N. G.; Bents, S.; Borchering, R. K.; Jung, S.-m.; Loo, S. L.; Smith, C. P.; et al. 2023. Informing pandemic response in the face of uncertainty. An evaluation of the US COVID-19 Scenario Modeling Hub. *medRxiv*.

Keogh, E. J.; and Pazzani, M. J. 2000. Scaling up dynamic time warping for datamining applications. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 285–289.

Müller, M. 2007. Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Petitjean, F.; Ketterlin, A.; and Gançarski, P. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3): 678–693.

Ratanamahatana, C. A.; and Keogh, E. 2004. Making time-series classification more accurate using learned constraints. In *Proceedings of the 2004 SIAM international conference on data mining*, 11–22. SIAM.

Srivastava, A. 2023. DTW+ S: Shape-based Comparison of Time-series with Ordered Local Trend. *arXiv preprint arXiv:2309.03579*.

Srivastava, A.; Singh, S.; and Lee, F. 2022. Shape-based Evaluation of Epidemic Forecasts. In *2022 IEEE International Conference on Big Data (Big Data)*, 1701–1710. IEEE.

US SMH. 2020. COVID-19 Scenario Modeling Hub. https://github.com/midas-network/covid19-scenario-modeling-hub.

US SMH. 2022. Flu Scenario Modeling Hub. https://fluscenariomodelinghub.org/.

Ye, L.; and Keogh, E. 2009. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 947–956.

Zhou, H. B.; and Gao, J. T. 2014. Automatic method for determining cluster number based on silhouette coefficient. In *Advanced materials research*, volume 951, 227–230. Trans Tech Publ.