Retrospective Clustering of COVID-19 Mortality Time Series Using Dynamic Time Warping

Murat Razi, Manuel Graña

Computational Intelligence Group, University of the Basque Country, UPV/EHU, San Sebastian, Spain

Abstract

- Retrospective analysis of the data gathered during the pandemic can be used for the preparation for future pandemics.
- It is now possible to ascertain if the pandemic response has been the same across the world.
- This paper contributes evidence that the pandemic response, as measured by the death time series of each country, was not the same everywhere.
- can be detected, such that countries in different clusters have rather different patterns of deaths.
- The cost of Dynamic Time Warping (DTW) matching provides a measure of similarity between the death time series.
- Hierarchical Clustering allows to find clusters of countries with similar death time series .
- We find a cluster of western Europe countries confirming reports following rather diverse approaches.
- Furthermore, we examine in detail the different patterns of several countries relative to Spain.

Contents

- Introduction
- Data
- Methods
- Results
- Discussion
- Conclusions

Introduction

- Half a million papers in PUBMED about COVID-19
- we think that the number of deaths is a more relevant indicator of the severity of the pandemic and the pandemic response of each country than the “cases”
- the differences in coding criteria may influence the final death toll attributed to the pandemic,
- any retrospective analysis should be grounded on actual measurements on the population

Introduction

- Recent works study the clustering of the COVID-19 death time series by different means and approaches.
- One work applies Latent Dirichlet Allocation (LDA) in order to decompose the mortality time series into time patterns (topic) of mortality evolution
  - finding that there are some latent geopolitical aggregation of countries with the same dominant topic.
- Another work carries out a comprehensive exploration of feature extraction and clustering of the mortality time series,
  - finding that some clusters of countries appear recurrently as the results of different approaches,
  - specifically the western Europe countries together with the United States are a robust cluster of mortality responses.

Introduction

- In this paper, first we carry out a selection of countries under study on the basis of the accumulated number of deaths,
  - restricting the study to the countries with a death toll higher than a defined percentile of the accumulated number of deaths.
  - This selection avoids computational artifacts introduced by small size countries.
- Then, we compute the pairwise distance induced by the Dynamic Time Warp (DTW) between selected countries, which is used for the Hierarchichal Clustering of the countries.
- We discuss the clustering findings, and some specific relations between the mortality patterns of some countries relative to that of Spain.

Introduction

- This work provides additional evidence that the pandemic has been a rather spatially heterogenous event,
  - contrary to the assumption of an homogeneous response due to a single cause associated to a specific pathogen.
- In other words, cultural and geopolitical factors appear to have played some significant role in the pandemic response.

Data

- Our world in data” (OWID) is hosted by the University of Oxford.
- The OWID site was capturing mortality time series from the John Hopkins University COVID-19 site until march 2023
- We have used “new_deaths_smoothed” data for the analysis After this date most countries stopped reporting.
- We consider the seven day moving average of the population normalized mortality time series as published in the OWID site.
- The raw mortality time series has some artifacts such as zero deaths in weekends, or even some negative reports due to corrections made by the local agencies, hence the need for the time series smoothing.

Methods

- We have computed the dynamic time warping (DTW) [11,7] among the smoothed death signals of the selected countries.
- The DTW is a dynamic programming algorithm that looks for the optimal alignment of two signals.

$$X = \{x_1, \ldots, x_N\} \text{ and } Y = \{y_1, \ldots, y_N\}, \text{ with } x_i, y_j \in \mathbb{R}^n \qquad C_{X,Y}(i,j) = \|x_i - y_j\|,$$

The DTW distance can be computed by the following recursive equation:

$$d_{X,Y}(i,j) = C_{X,Y}(i,j) + \min\{d_{X,Y}(i-1,j-1), d_{X,Y}(i-1,j), d_{X,Y}(i,j-1)\}$$

Methods

- Search has been limited to a specific time window, in order to avoid excessive distortion of the signals.
  - Hence, the effective area for the search of matching points is a diagonal area of the cost matrix.
- Signals can have different lengths, but in our case the death time series have the same length.
- The distance used to build up the cost matrix isthe Euclidean distance.
- The search is a greedy one, so there is no guarantee that the algorithm finds the global matching optima.

Methods

- we carry out Hierarchical Clustering (applying the Ward method) over the DTW distance computed for each pair of selected countries.
- Ward's method minimizes the total within-cluster variance. at each step it finds the pair of clusters that leads to minimum increase in total within-cluster variance after merging.
- At the initial step, all clusters are singletons (clusters containing a single point).
- The initial cluster distances in Ward's minimum variance method are therefore defined to be the DTW distance between points.
- In order to select the number of clusters we explore maximizing the value of the silhouette coefficient, but also maximizing the entropy of the cluster sizes

Results

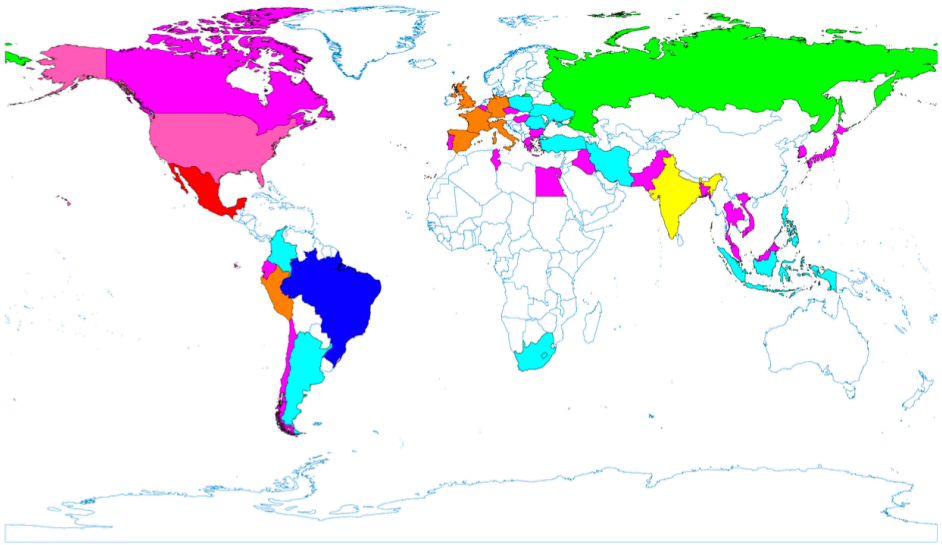
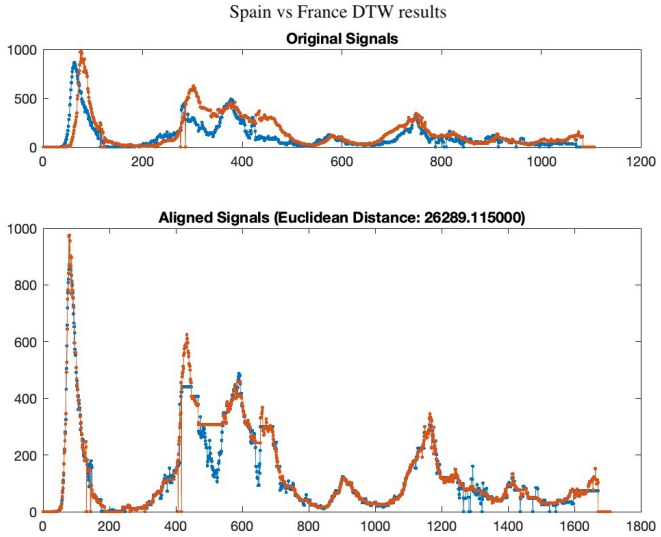


Fig. 1. World map of the detected country clusters by hierarchical clustering of DTW distances among mortality time series

Results

- Non-colored countries have been excluded of the clustering process.
- unitary clusters: India, United States, or Russia.
- Geopolitical aggregations.
  - The salient case is that of Spain, France, Italy, Germany, and United Kingdom (orange).
    - This cluster has been found systematically by other approaches.
  - Another cluster (cyan) includes geographically disperse countries like Canada, Portugal, Chile, or Japan.
    - Hence, person to person disease diffussion processes are very unlikely inside this cluster of countries,
    - can be characterized by a moderate death toll.
  - Another geographically disperse cluster (light blue) includes middle east and eastern Europe countries like Ukraine, Ro mania, Poland, and Turkey, and remote countries like Argentina, South Africa, and Indonesia.

Results



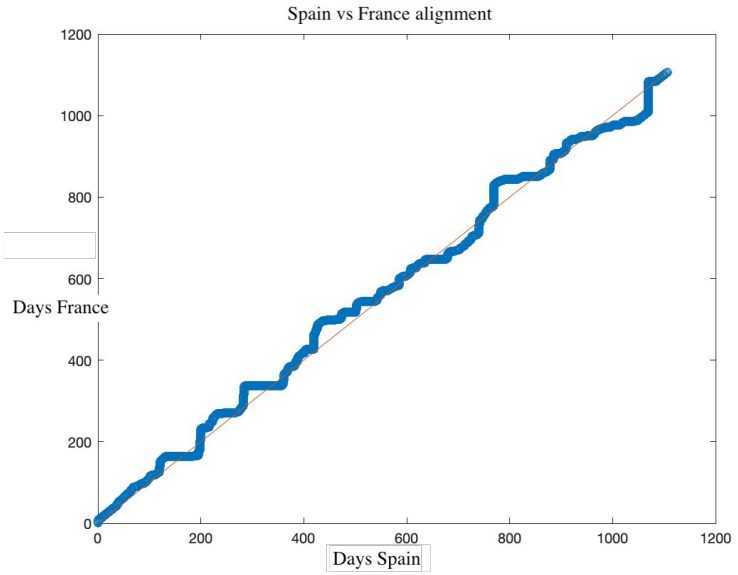


Fig. 2. Spain vs. France results of the DTW correction

Results

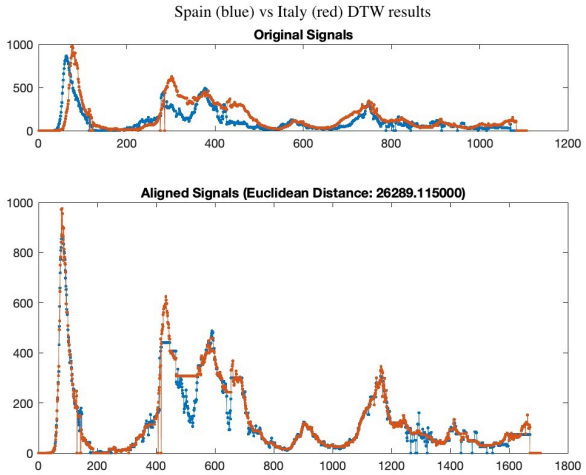
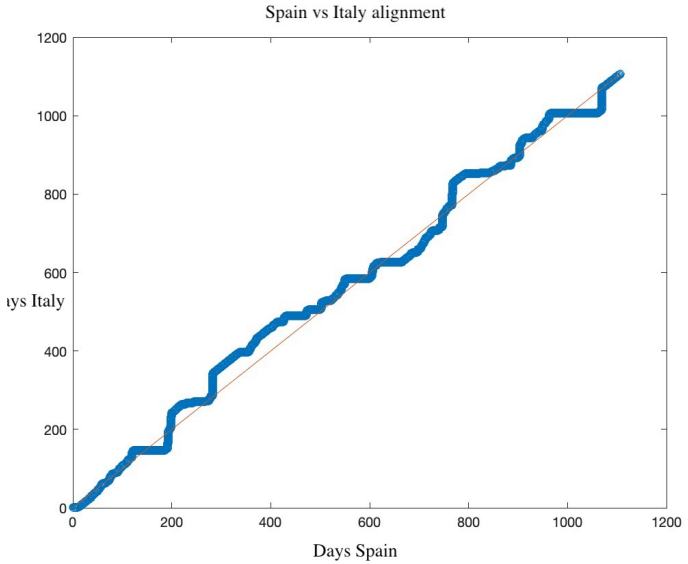
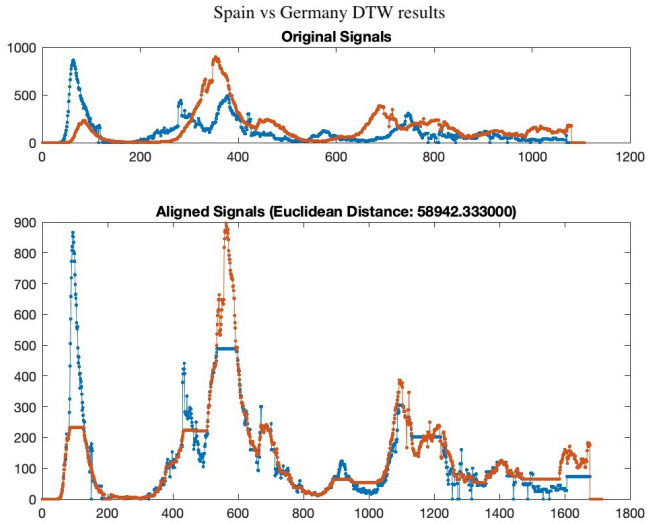


Fig. 3. Spain vs. Italy results of the DTW correction



Results



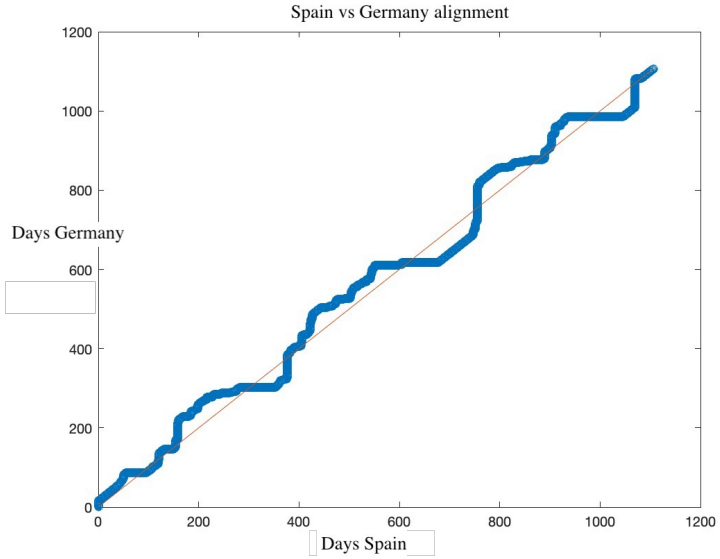
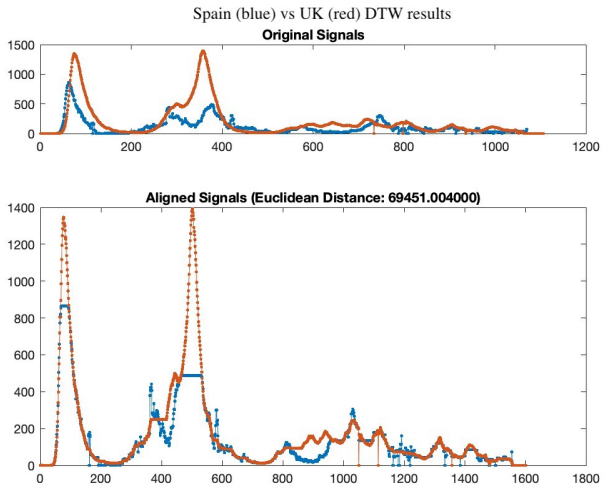


Fig. 4. Spain vs. Germany results of the DTW correction

Results



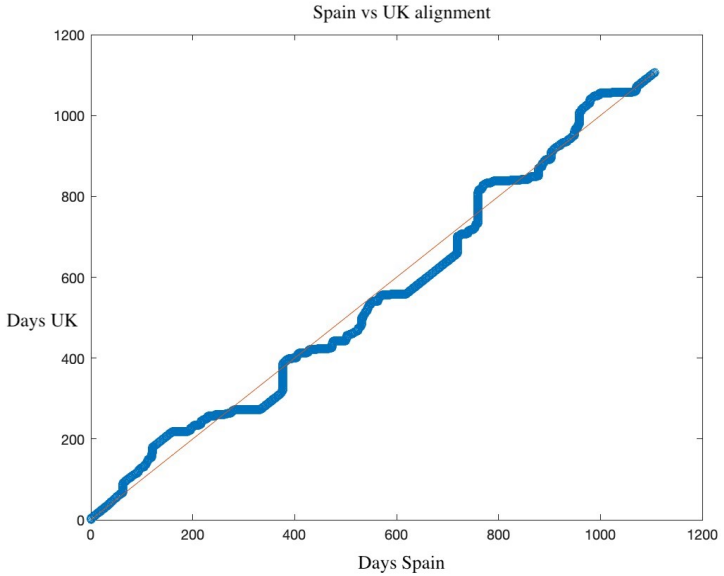


Fig. 5. Spain vs. UK results of the DTW correction

Results

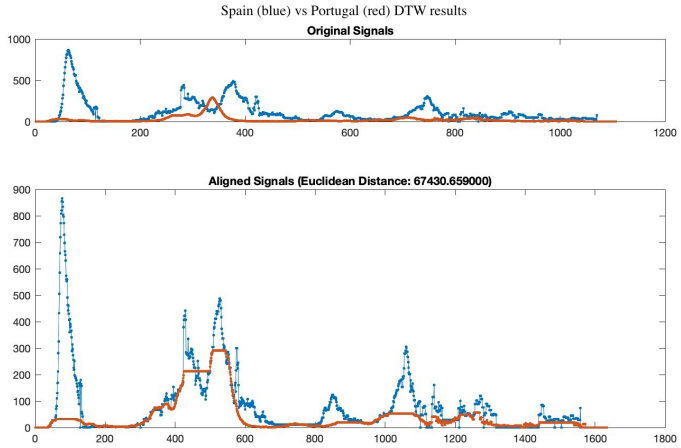
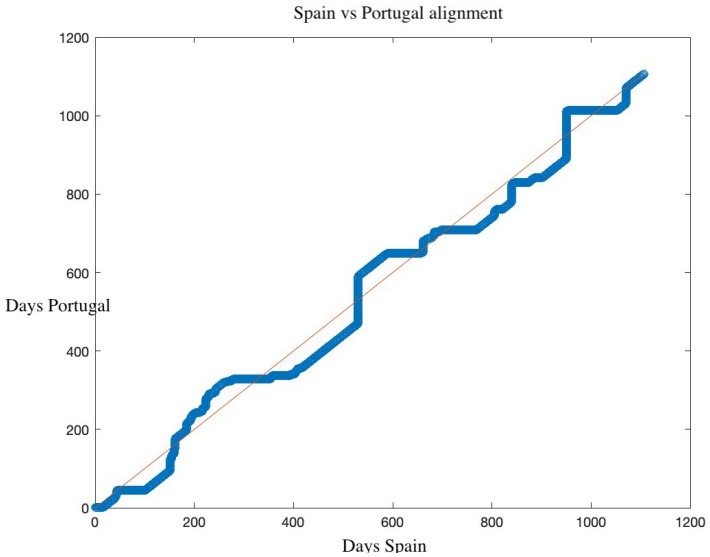
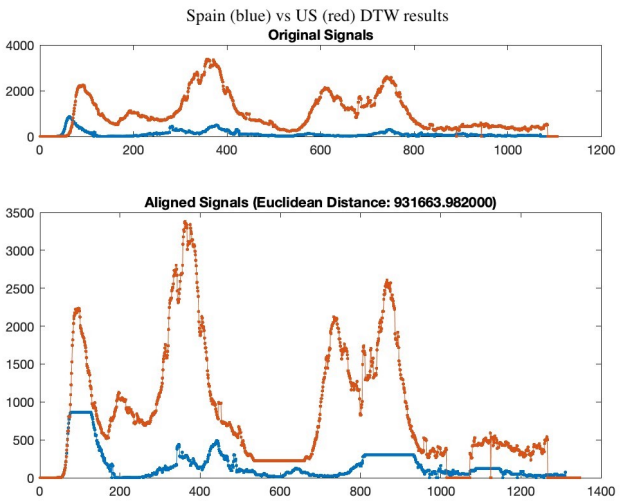


Fig. 6. Spain vs. Portugal results of the DTW correction



Results



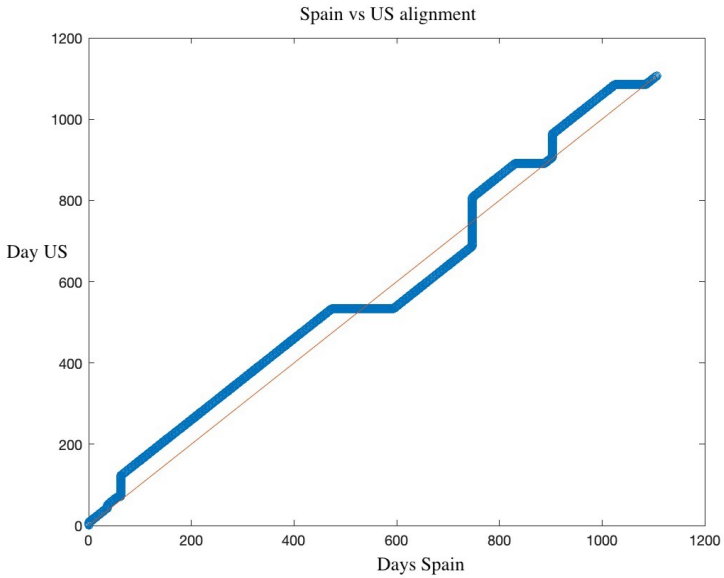


Fig. 7. Spain vs. US results of the DTW correction

Results

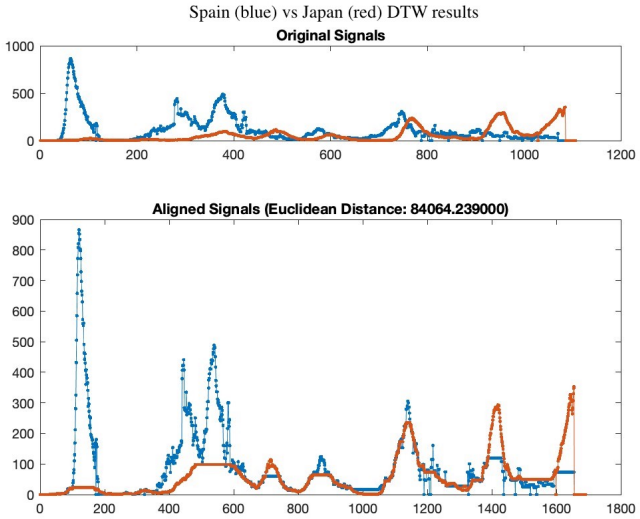


Fig. 8. Spain vs. Japan results of the DTW correction

Discussion

- One of the most salient characteristics of the COVID-19 death time series is the initial response, i.e. the mortality in the first six months of the pandemic.
  - Some countries show a strong peak of mortality in the spring of the year 2020, while others have little or no mortality response until the autumn of 2020.
  - This characteristic has a definite impact in the formation of the clusters detected by the described procedure.
    - countries in the western Europe cluster (orange) show an almost synchronized mortality peak in the spring of year 2020,
    - countries in the easter Europe cluster (light blue) have a very small and smooth response in that period.
- In light of these observation, it seems necesary to revisit the hypothesis that the COVID-19 pathogen was almost instantaneously traveling the world,
  - with cases appearing simultaneously in Australia and Spain, i.e. at two opposite points of the globe.

Discussion

- Figures 2, 3, 4, and 5 show the aligned mortality time series of Spain versus France, Italy, Germany, and United Kingdom, respectively, after computing the DTW correction.
  - Except for Germany, these countries share a large peak in the spring of 2020, with some temporal displacements.
  - We can appreciate that the peak in Spain precedes those in Italy, Germany, and France, but it is almost synchronous with that of the United Kingdom.
  - All these countries have strong mortality after the summer of 2020 until the summer of 2021, with not very different time structure.
- Mass vaccination in these countries started at the end of the year 2020.
  - However, the COVID-19 mortality did not decrease significantly until the summer of 2021,
  - and did have some rebounds, like in the so-called Omicron wave during the winter of 2021 and the beginning of the spring of 2022.

Discussion

- Figure 6 shows the aligned mortality time series of Spain versus Portugal after DTW.
  - We can appreciate why Portugal belongs to a different cluster despite being the closest spatial neighbor of Spain.
  - For some reason, there was no mortality peak in Portugal in the spring of year 2020.
- There was no synchronization with Spain,
  - while Spain was perfectly synchronized with the United Kingdom, which is much more distant spatially.
- The first and only significant mortality peak in Portugal appears in the beginning of year 2021,
  - paradoxically, it does match a valley in the Spanish mortality time series.
  - Synchronized with the mass vaccination campaign

Discussion

- Figure 7 shows the aligned mortality time series of Spain versus United States after DTW.
  - The United States was the universal reference for guidance in the pandemic response, so it is interesting to examine how the mortality responses were aligned to this reference.
- The huge difference in the magnitude of the time series should be the main reason for the Spain and United States to be in different clusters.
  - There was some peak synchronization in the spring of year 2020.
  - However, COVID-19 mortality in the United States never came down close to zero until the end of 2021, just before the Omicron wave.
- It seems that the most influential country for health guidance did suffer the greatest death toll.

Discussion

- Finally, Figure 8 shows the aligned mortality time series of Spain versus Japan after DTW.
- Again, it is easy to appreciate that Japan did not have a mortality peak in the year 2020, and some small incidence started in 2021.
- Paradoxically, the stronger mortality peaks in Japan are appearing in the year 2022, when mortality in Spain is disappearing.
  - These mortality peaks follow the mass vaccination waves.
- Therefore, Japan shows a completely different pattern of mortality from Spain.

Conclusions

- There is growing evidence from these works that socio-cultural factors did influence the death toll of the COVID-19 pandemic, maybe more than the designated pathogen.
  - The paradigm of “one pathogen, one disease” seems to be broken in the COVID-19 pandemic.
  - The present paper does add more evident to this departure of the classical medical paradigm.
- Using DTW induced distance between time series, we found clusters of countries with rather different evolutions of the COVID-19 mortality.
  - Some clusters may support some kind of synchronization inthe propagation of the disease, while others provide counterfactual evidence.
  - For instance the two spatial neighbors of Spain, Portugal and France, belong to different clusters and show radically different time patterns of COVID-19 mortality.
- The results in this paper invite to look more deeply into the socio-cultural roots of the pandemic response, in order to learn useful lessons for the future.