Analyzing the Similarity of Covid-19 Time Series Pattern

Xuanhao Cao
Matrikelnummer 6003173
xuanhao.cao@student.uni-tuebingen.de

Dorothee Sigg
Matrikelnummer 4108173
dorothee-maria-barbara.sigg
@student.uni-tuebingen.de

Abstract

In this analysis, we are looking at the time dynamics of the Omicron wave in different European countries. Do some countries exhibit a similar time series pattern? To this end we use time series clustering algorithms to properly capture the dynamics in time. We find that clustering the countries yields interesting insights about the increase of cases, the severity and whether the wave is already broken.

1 Introduction

Covid-19 is an ongoing topic – currently the newspaper are full of articles about omicron. On 26 November 2021 the World Health Organization classified omicron as a variant of concern. Since every country has its own policy on how to deal with it, the dynamics of the case numbers are very different. Some countries struggle more than others, some seem to be already over it whereas others were never really hit that hard. This results in very different patterns in the time series of cases for each country. We want to compare these different patterns in time and find out which countries can be grouped into one category of similar omicron dynamics.

Data We use the weekly cases per million data [1] from *Our World in Data*. We restrict our analysis to Europe (geographically), of which our data covers 45 countries. As a start date we use the 26 November 2021. The data is in the form of time series per country.

2 Method: Time Series Clustering

2.1 Time Series K-Means

In order to analyze the similarity of time series patterns, it seemed obvious to use some kind of clustering algorithm. Clustering belongs to the class of unsupervised ML since there is no correct label. We chose to use a variant of the well-known k-means algorithm (modified that it can process time series), because of the simplicity and popularity of k-means [2]. For that, it needs a suitable distance measure and a way of computing averages – the cluster centers.

Distance measure Regarding the distance between two time series it is not clear how to do that, because time series are often noisy and have outliers and shifts. One naive approach would be to use euclidean distances between the two coordinates belonging to one point in time. However, there are

[1] https://github.com/owid/covid-19-data/raw/master/public/data/jhu/weekly_cases_per_million.csv

[2] Due to restricted space we do not explain k-means here.

more elaborate methods (for an overview see Aghabozorgi et al. [2015]), for instance dynamic time warping (DTW), which is more elastic and better at recognizing shifts.
DTW is calculated as follows: Given two time series $x = (x_0, ..., x_{n-1})$ of length n and $y = (y_0, ..., y_{m-1})$ of length m. The DTW distance can be described as the following optimization problem:

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j) \in \pi} d(x_i, y_j)^2}$$

where $d(x, y)$ refers in our case to the euclidean distance and $\pi = [\pi_0, ..., \pi_K]$ is a path that satisfies the following properties:

- π is a list of index pairs $\pi_k = (i_k, j_k)$ with $0 \leq i_k < n$ and $0 \leq j_k < m$
- starting point $\pi_0 = (0, 0)$ and endpoint $\pi_K = (n - 1, m - 1)$ are fix
- $\forall k > 0$: $\pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
  - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

DTW calculates the similarity between two time series data x and y in the sense that it best matches each coordinate and matching multiple coordinates is allowed (i.e. "warped", non-linear in time). We illustrated this in Figure 1.

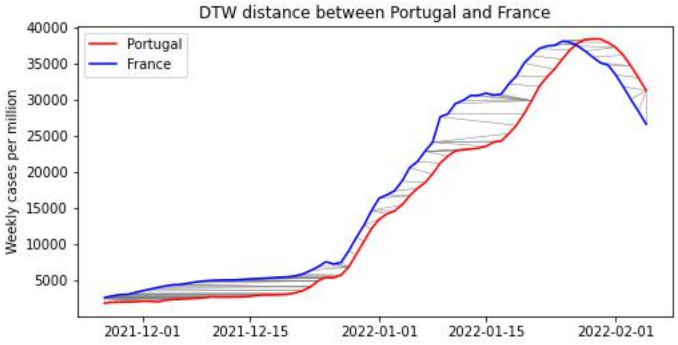

Figure 1: From the path found by DTW it gets clear, that the algorithm is able to recognize shifts – a desirable property for our analysis (which euclidean distance lacks of).

Averaging In addition to a distance measure[3], the k-means algorithm needs to compute averages. To this end, Petitjean et al. suggested the DTW barycenter averaging method (DBA), which iteratively refines an initially (potentially) random average sequence such that it minimizes the squared distances of DTW to the averaged sequence. This produces more meaningful results, see Petitjean et al. [2011] (also for the full algorithm).
For all our computations we used the tslearn package, which provides a k-means algorithm based on DTW and DBA.

Evaluating clustering In order to estimate the quality of the clustering, we use the silhouette score. It ranges from -1 to +1 and the bigger it is, the better the clustering. A short definition:

$$sc = \frac{b - a}{max(a, b)}$$

where a is the mean distance between one sample and all other points in the cluster and b is the distance between a sample an all other points in the next nearest cluster.

[3] Technically, DTW is not a metric, because it does not necessarily satisfy the triangle inequality.

2.2 Our Analysis

Regarding missing values in the dataset, we interpolated them if they were in the middle of the time series, if the were at the end we used forward filling (since we used only omicron data, missing values at the beginning did not occur). We searched for the best number of clusters k using the silhouette score to estimate the goodness of the clustering. For all our computations we used the tslearn package, which provides a k-means algorithm based on DTW and DBA.
Our code is in: https://github.com/CaoXuanHao0/Data-literacy-course-project-2022

3 Results and Discussion

Using the method described above, we classified the weekly cases per million data into 4 clusters (silhouette score = 0.44), see figure 2. The pattern for each cluster is given by averaging over the data using DBA, see figure 3. Each of cluster corresponds to one specific pattern of time series trend, i.e., the pattern of changes of new cases from the start of omicron to present.

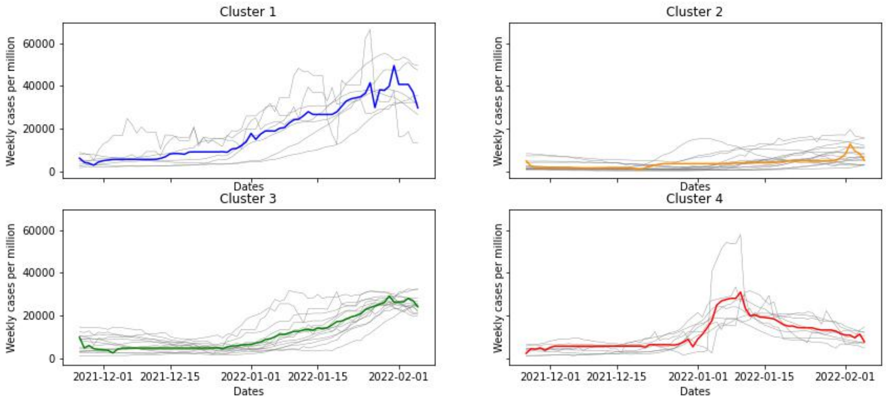

Figure 2: **Clustering Result**.
Cluster 1: Andorra, Portugal, Slovenia, Denmark, Estonia, France San Marino.
Cluster 2: Albania, Malta, North Macedonia, Ukraine, Poland, Romania, Russia, Hungary, Germany, Moldova, Finland, Croatia, Bulgaria, Bosnia and Herzegovina, Belarus, Serbia.
Cluster 3: Slovakia, Sweden, Norway, Switzerland, Netherlands, Liechtenstein, Luxembourg, Lithuania, Latvia, Iceland, Gibraltar, Czechia, Belgium, Austria.
Cluster 4: Monaco, Italy, Isle of Man, Ireland, Greece, Spain, Montenegro, United Kingdom.

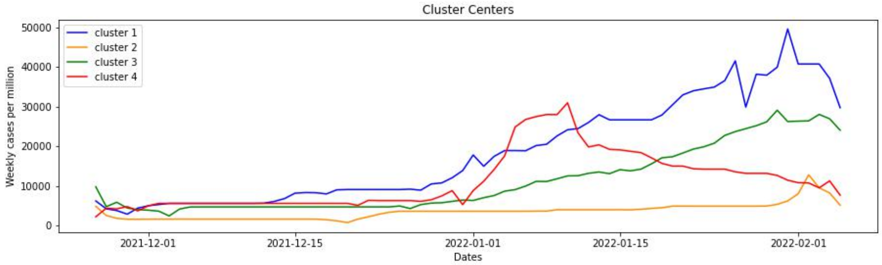

Figure 3: Cluster Centers

The first cluster (blue line) exhibits a pattern which shows a monotonic increase resulting in very large case numbers. Countries in this cluster were hit hard by omicron, but at the same time the wave seems to be broken by now.

The second cluster (yellow line) includes mostly countries in eastern Europe and Germany and Finland. Their pattern indicates that their pandemic situation kept well, but recently tarts to become a little worse. This could mean that either they were not affected that much by omicron or on the other hand they had effective policies. Regarding Germany, it is interesting to see that the situation is – in the light of an European comparison – not that bad (as it sometimes may be transported by German newspapers).

The third cluster (green line) includes countries from central and northern Europe. Their pattern indicates a steady increase and it is not that clear if the case numbers are already going down. Probably, these countries are still in the middle of the omicron wave.

The fourth cluster (red line) shows an early bump of cases by beginning of January. Subsequently, cases go down and the omicron wave seems to be under control, which means government's policy to omicron outbreak is possibly a success. This is consistent with recent reports in this countries, for example BBC news reported omicron is becoming less and less severe in UK Roberts [2022].

For a geographic visualization of the clusters see figure 4. Some geographic relationships become visible: Most of the eastern European countries seem to belong to the same cluster. The same holds for some northern European countries. Apart from that, geographic relationships are not really present.

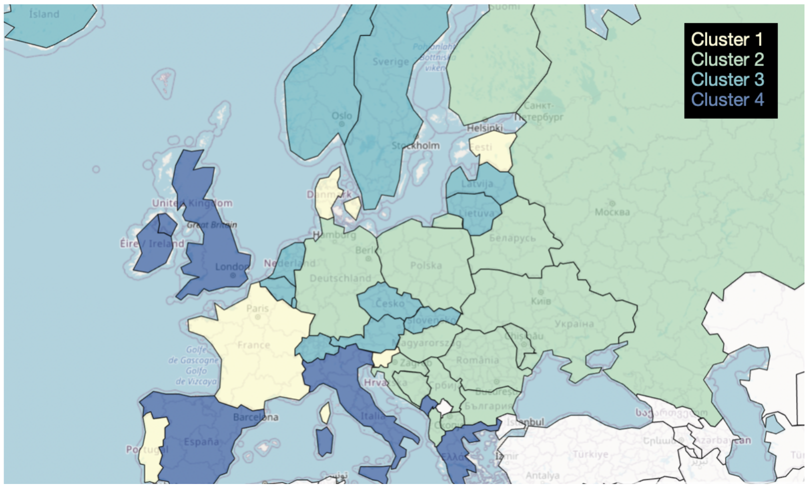

Figure 4: Map of Europe. Countries are colored according to their cluster.

Limitations One limitation may be the algorithm and the task itself. The silhouette score is at 0.44, which means that it could still be better. On the other hand, this clustering task is not easy and maybe no clear-cut clusters exist.

4 Conclusion

Our analysis shows which countries have a similar dynamics during the Omicron wave. It would be interesting to correlate that with the underlying governmental policies. Do countries which are in the same cluster also have similar regulations? However, one can also question the data: In many countries the test capacities are at its limits and the number of unreported cases may be higher than in other phases of the pandemic.

References

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

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

Michelle Roberts. High confidence – omicron is less severe in uk. *BBC News*, 2022. URL `https://www.bbc.com/news/health-59999698`.