Assessing the Role of Temperature and Humidity in the Spread of COVID-19

Namrata Mali
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
namratam@vt.edu

Nikitha Donekal Chandrashekar
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
nikitha@vt.edu

ACM Reference Format:
Namrata Mali and Nikitha Donekal Chandrashekar. 2021. Assessing the Role of Temperature and Humidity in the Spread of COVID-19. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 Abstract

COVID-19 has become a prominent crisis of the world and the disease is rapidly spreading all over the world. Analyzing the trend of the daily count of COVID cases for the countries with similar climatic conditions would be helpful in forecasting and other applications. In this project we have implemented various data analytics techniques like clustering, Dynamic Time Warping, barycenter to evaluate the effect of climatic conditions like temperature and humidity on the spread of the COVID cases. World wide analysis will reveal the correlation between increase or decrease of COVID cases and other climatic, socioeconomic and health factors.

2 Introduction

Various studies have been conducted to understand the relation between climatic conditions and spread of respiratory tract infections. These studies help in identifying the future trend of the spread of these diseases. Researchers have investigated the influence of climatic factors on the spread of respiratory tract infections like SARS-COV2 and have found a relationship among them due to the seasonal outbreak of these diseases. [5]

On the similar front, there are several analyses being conducted on the COVID data due to its similarity in nature with other respiratory viral infections. One of the prominent research topic is to analyze and evaluate whether climatic changes affects spread of COVID or not. However the results of these investigations have been contradictory, because COVID has been transmitted across the world from cold and dry to hot and humid climatic regions. [4] [8] [3]Additionally, we observed that the previous work done in this domain are restricted to a particular country or a particular group of countries like European countries or Asian countries. No study has been done to analyze world wide trend of spread of COVID

The aim of this project is to assess and evaluate the effect of climatic conditions on the spread of COVID-19 using data analytics techniques like clustering and time series analysis. For the study, we will be considering only temperature and humidity as climatic factors or features for clustering the countries. We have considered almost 90 countries spread across the world to reduce the biased nature of the results obtained.

2.1 Keywords

dataset, COVID-19, Clustering, Time-Series, Dynamic Time Warping. Barycenter

3 Dataset

In this section, we describe the datasets constructed and utilized for the project. For our analysis , we have considered data for 90 countries extended across the world. Two datasets were considered for the project, one characterizing the COVID cases for countries and the other depicting the data of a country's climatic conditions. For scope of this project, we only consider temperature and humidity as climatic parameters.

The first dataset is an open source dataset consisting of the number of COVID cases for 90 countries around the world dated from February 25 2020 to February 28 2021. Dataset contains each day count of total number of cases, active cases, recovered cases and count of deaths. It also consists of the geographic, economic, population and demographic data of the countries which can be used for further analysis. [2] https://www.kaggle.com/sambelkacem/covid19-algeria-and-world-dataset

The second is a public domain dataset, collected from the world weather online website which contains the average

Conference'17, July 2017, Washington, DC, USA

https://doi.org/10.1145/1122445.1122456

temperature, hours of sunlight, percent humidity and the wind speed in the capital city of 156 countries. The data is dated from January 22 2020 to March 21 2020.[12] https://www.kaggle.com/winterpierre91/covid19-global-weather-data.

4 Methodology

The dataset required for analysis is a set of countries containing data of both the climatic factors namely average temperature and average humidity and the total daily COVID cases for each country. To obtain the required dataset, we merged the data for the countries common to both the datasets. To get the data for each country, we summed the COVID cases of all the countries for each month and computed the average of the climatic factors collected over the duration.

The obtained data was partitioned in the ratio of 0.8 and 0.2 for training and testing of our model respectively. In the training phase we cluster all the countries in the train data into 3 clusters using K-Means. After clustering the countries, the time series for all the counties in a cluster were aggregated to obtain a time series for a particular set of climatic conditions.

In the testing phase, each country is assigned to a cluster that it has similar climatic conditions with. Then the time series of the country and the computed time series for the cluster are analysed for their similarity. The time series are processed and plotted to qualitatively visualize their similarity. For quantitative analysis, we compute the Dynamic time warping (DTW) distance between the two time series and evaluate their similarity.To test and evaluate our model, we average the DTW distance for all the countries in the test data for a cluster and perform 5- fold validation.

4.1 Data Preprocessing

Data preprocessing is an important task to be performed before applying any data analytics technique on the data. Firstly, data cleaning was performed on the dataset to remove redundant columns and the missing values from the data. After cleaning the data, both datasets are pivoted to visualize them in the index/ columns format, where the index is country name for both the datasets. The next operation performed on the data, was to roll up the values of temperature and humidity for each month by its average.

At this stage the first dataset contains 90 countries and their corresponding count of COVID cases and the second dataset contains 156 countries and values of the average temperature and average humidity in the capital cities of these countries.These two datasets were merged over the common countries using the inner join. Now we have the dataset which contains the country wise values of the average temperature, humidity and the count of COVID cases for each day dated from Feb 2020 to Feb 2021. For building and evaluation of the model we divided the data in the ratio of 0.8 and 0.2 as training set and testing set for our model respectively.

4.2 Dimensionality Reduction

Dimensionality reduction is a technique which is used to transform data from higher dimensional space to lower dimension such that its representation in the lower dimensional space should preserve the meaningful information of the original data and be close to its intrinsic dimensions. Linear dimensionality reduction techniques include Principle component analysis (PCA) and Multidimensional scaling (MDS). Non- linear approaches include Isomap, locally linear embedding(LLE), stochastic Neighbourhood embedding(SNE) and t-distributed stochastic Neighbourhood embedding (t-SNE). Manifold learning is a class of non linear dimensionality reduction method, developed to learn lower dimensional representation. It also allows us to visualize and work with the lower dimesional representation of the data which is comparatively less expensive. Many real world datasets can be approximately represented as lower dimensional manifolds that are embedded in higher dimensional space.

The combination of the data for temperature and humidity combined can be represented as a point in the 6 dimensional space. Performing clustering of data points in 6D space though possible is expensive and not realizable. For the purpose of our project, dimensionality reduction was implemented on the values of temperature and humidity to make them observable. The data points of temperature and humidity of a country was reduced from a six dimesional space to a two dimesional space. This improved the interpretability of our model as people can observe and visualize the clustering of the countries based on their climatic parameters like temperature and humidity. We have used t-SNE manifold method for our analysis as t-SNE performs better than the other non linear dimensionality reduction methods for a complex dataset. 3-d visualization of the training data after dimensionality reduction is shown in Figure 1.

4.3 Clustering

Clustering is a method of grouping objects with objects in the same group having identical features and different features with objects in other groups. Clustering as a technique has applications in various fields across the domain. Mathematically, clustering is performed such that the intra cluster distances are minimized and the inter cluster distances are maximized. This is in accordance with the concept that the intra cluster density should be more and the inter cluster density should be less. K-means clustering, hierarchical clustering, density based clustering, graph based clustering,

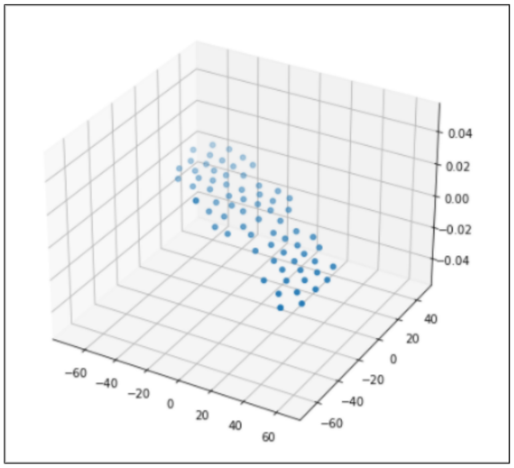

Figure 1. 3D visualization of the training data after dimensionality reduction

probabilistic clustering are a few examples of the clustering algorithms frequently used.

For the clustering of the data, we used the k-Means clustering algorithm to form the clusters of the countries in the training set according to their reduced values of temperature and humidity. The data has been clustered into 3 clusters with countries in the same cluster having similar climatic conditions.

The visualization of the clusters are represented in the Figure 2 where the cluster with color blue (Cluster 0) represents the countries with low temperature - high humidity, the cluster with color green (Cluster 1) representing the countries with moderate temperature - moderate humidity and the cluster with color yellow (Cluster 2) representing countries with high temperature - low humidity. Further time series analysis techniques are applied on the time series graphs of countries in the same cluster.[4]

4.4 Model Building

Following the clustering of the countries into their respective clusters, we proceed towards analysing the daily covid cases of the countries in each of the cluster. The daily covid cases count for each country is represented as a time series as the data which consist of sequential observations or events which are changing over time can be plotted on a graph for analysis and to make the predictions. The goal for creating the model was to compute a time series that represents the trend in the daily covid cases of all the countries in a particular cluster. To achieve this goal, the time series of the countries had to be aggregated to a single time series.

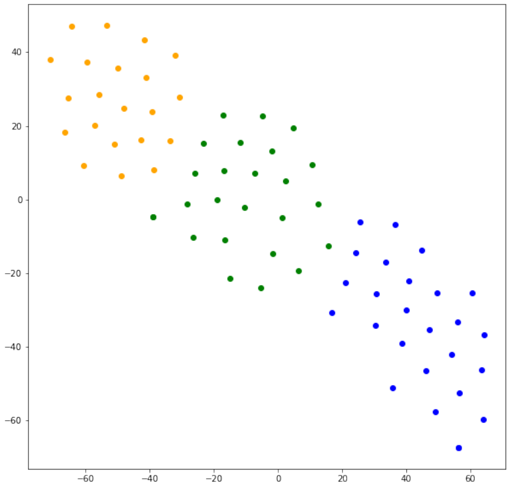

Figure 2. Graphical Visualization of the clusters in the data.

The first step in this process was to perform pre-processing techniques on the raw time series of each country. To start with, we performed amplitude scaling on the time series to remove the distortions related to the value and focus more on the trend. The linear trend for the data wasn't removed as it had an important role in computing combining the daily covid time series. This step was important as the aggregation of time series is very sensitive to the presence of linear trend. Finally, the data for the graphs were smoothed using the moving average technique which reduces the sensitivity of the time series to outlier data points.

To perform the aggregation of the processed time series, various techniques were considered. Most averaging algorithms proposed in the literature are based on pairwise averaging methods. Such strategies are really dependent of the order in which the series are taken into account with no guarantee to obtain the same accuracy with a different order. [6] Of the recent works, a global approach to have been developed is DTW Barycenter averaging (DBA) from Petitjean et al.[7] The main advantage of these methods is that series are averaged altogether and therefore there is no impact on the order of consideration of series. Barycenter is a time series which tries to minimize the sum of squares of distances of other time series in that dataset.

The tslearn module in python offers various methods to calculate the barycenters for the clustered time series which include euclidean barycenter, dtw barycenter averaging etc. [10] For our project we used the standard Euclidean barycenter computed from a set of time series.To compute the euclidean barycenter the function calculates the euclidean distance between all time series and takes the best possible distance such that it should be minimum. The graphs for the barycenter are represented in the below from Figures

[4-6]. The major drawback of using it is that all the time series should be equal lengths but, since our dataset was for a fixed duration with a definitive start date and end date no additional processing was required on the time series. Below are the images shown for the computed barycenters for all the countries in a cluster.

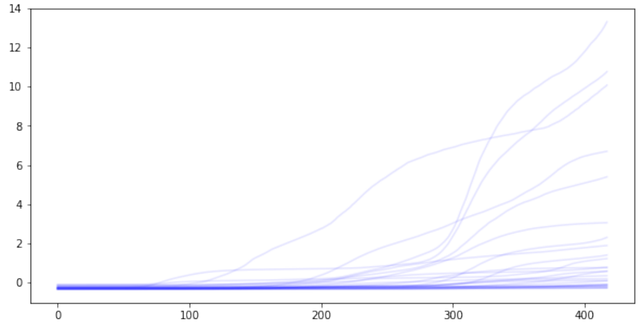

Figure 3. Graphical representation of the COVID time series for countries in cluster 0

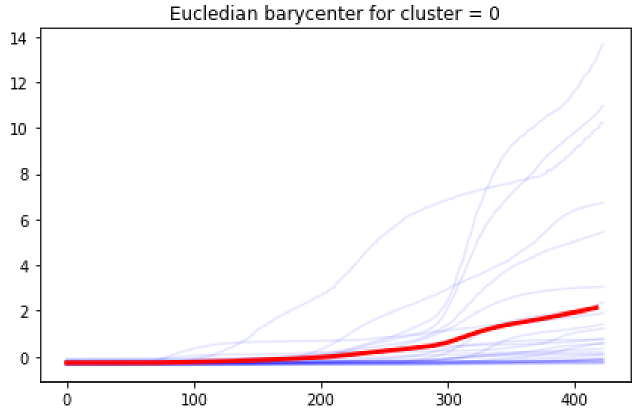

Figure 4. Graphical Visualization of the aggregated Time Series for countries in Cluster 0

4.5 Testing and Evaluation of the model

For a robust testing of a model, it is important to integrate and embrace both qualitative analysis and quantitative analysis.[11] By combining quantitative and qualitative methods, a degree of comprehensiveness may be achieved that neither approach, if used alone, can achieve. [1] Considering this concept, both qualitative and quantitative analysis is performed on the model to robustly evaluate it.

Foremost the cluster labels were predicted for the test countries using the already trained kMeans model (Clustering). For this process, the test data were first reduced to a lower dimension. Post assigning the test countries to their clusters, qualitative and quantitative analysis was performed.

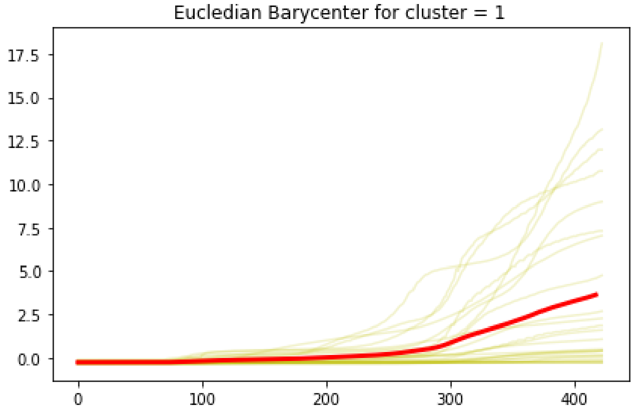

Figure 5. Graphical Visualization of the aggregated Time Series for countries in Cluster 1

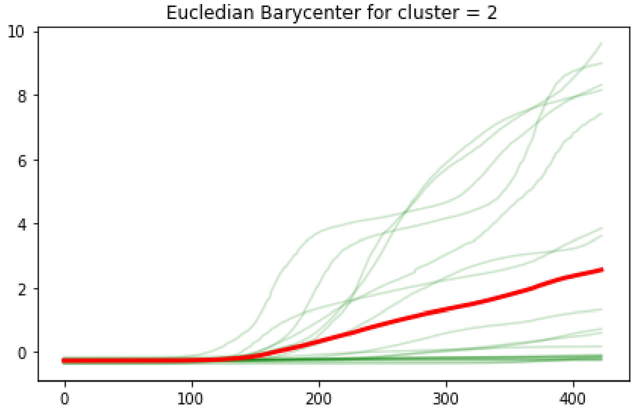

Figure 6. Graphical Visualization of the aggregated Time Series for countries in Cluster 2

As a part of the qualitative analysis, the preprocessed time series for the test countries in each cluster was plotted with the aggregated time series for the corresponding cluster. The trends and the graphs were analysed to gain information about the similarity between them. In addition to the qualitative analysis, the average DTW (Dynamic Time Warping) distance between the time series of daily covid cases for each of the country and the aggregated time series of their corresponding cluster was computed. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. To test the generalizability of the model, a 5-fold cross validation on the data.

5 Observations And Results

The section consists our graphs and evaluation results of the qualitative and quantitative analysis performed in the previous step as a part of our testing and evaluation of the model.

5.1 Qualitative Analysis

In this section, the graphs will also be analysed to extract some meaningful information from the graphical visualizations.

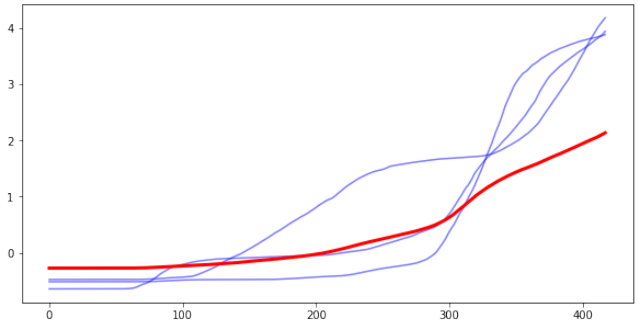

Figure 7. Graphical representation of the qualitative analysis for test data predicted to belong to Cluster 0

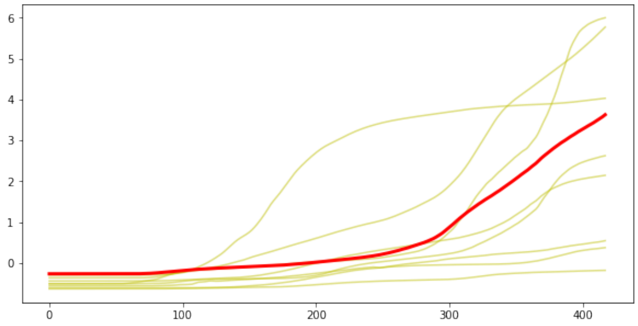

Figure 8. Graphical representation of the qualitative analysis for test data predicted to belong to Cluster 1

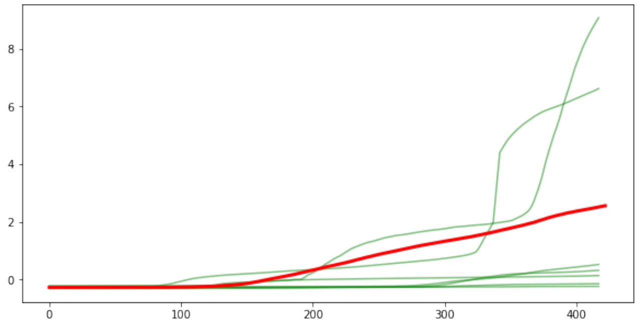

Figure 9. Graphical representation of the qualitative analysis for test data predicted to belong to Cluster 2

From the graphs, it can be observed that the aggregated time series or the barycenter for each cluster describes the global trend of the series in the cluster. Different clusters tend to have different trends. The graphs also reveal that there are countries with high variance from the barycenter. Specifically in the graph for cluster = 1, it can be observed that there are countries with almost the same trend as that of the barycenter whereas there are also countries whose, daily COVID 19 cases trend is quite different from the cluster trend.

5.2 Quantitative Analysis

In this section, we will be defining the similarity between time series depicting the daily covid case numbers of each country in the test set and their corresponding barycenter for the cluster to which the country belongs to. For numeric values we will be considering the DTW distance.Below is the table that shows the observed values for the dtw distance over the 5-fold evaluation performed.

Iteration	Cluster 0	Cluster 1	Cluster 2	Average DTW distance
1	18.779	20.584	4.874	14.746
2	13.664	23.900	20.640	19.401
3	20.595	15.281	16.763	17.546
4	19.012	15.539	18.780	17.777
5	15.554	3.920	18.296	12.563
Average DTW distance over all iterations				16.406

Table 1. Table to depict the results of quantitative evaluation

Though studies state that DTW distance is not a metric, it has been identified a definitive measure of similarity in almost all of its application. [9] From the computed values for the DTW distance we can observe that the values are significantly high. This observation is made as amplitude scaling has been performed on the time series for all the countries.

6 Conclusion And Future Works

In this section , we will deduce our observations to reach a conclusion and revisit our objective of the project. Since our dataset consists of countries ranging all around the globe, the results obtained in the outcome are more robust and have less bias which was one of the drawbacks in the existing works.

Firstly, the graphs for the qualitative analysis direct us towards the similarity in the trends of the count of the daily COVID cases for countries with similar climatic conditions. One of the major purposes of clustering and aggregating time series is for prediction and the accuracy of the prediction depends on the effectiveness of clustering and aggregation.

Though the initial trends are supportive of the fact that temperature and humidity have an effect and role to play in the spread of COVID 19, the quantitative values of testing progresses the project in the direction that other climatic, socioeconomic and health conditions need to be consider for analysis to get a more accurate results for the predictions of COVID cases.

7 Acknowledgments And Individual Works

We would like to thank Dr. Reddy for giving us the opportunity to work on an interesting project related to COVID. We would also like to thank Ms. Ming Zhu for the guidance she gave us throughout the semester. With respect to the work done by us Namrata worked on the data pre-processing and evaluation, and Nikitha worked on the clustering and time series aggregation parts of the project.

References

[1] Earl R Babbie. *The practice of social research*. International Thomson Publishing Services, 1998.

[2] Sami Belkacem. "COVID-19 data analysis and forecasting: Algeria and the world". In: *arXiv preprint arXiv:2007.09755* (2020).

[3] Sajad Jamshidi, Maryam Baniasad, and Dev Niyogi. "Global to USA County Scale Analysis of Weather, Urban Density, Mobility, Homestay, and Mask Use on COVID-19". In: *International journal of environmental research and public health* 17.21 (2020), p. 7847.

[4] Paulo Mecenas et al. "Effects of temperature and humidity on the spread of COVID-19: A systematic review". In: *PLoS one* 15.9 (2020), e0238339.

[5] EG Mourtzoukou and Matthew E Falagas. "Exposure to cold and respiratory tract infections". In: *The International Journal of Tuberculosis and Lung Disease* 11.9 (2007), pp. 938–943.

[6] Vit Niennattrakul and Chotirat Ann Ratanamahatana. "Inaccuracies of shape averaging method using dynamic time warping for time series data". In: *International conference on computational science*. Springer. 2007, pp. 513–520.

[7] François Petitjean, Alain Ketterlin, and Pierre Gançarski. "A global averaging method for dynamic time warping, with applications to clustering". In: *Pattern recognition* 44.3 (2011), pp. 678–693.

[8] Indrani Roy. "The role of temperature on the global spread of COVID-19 and urgent solutions". In: *International Journal of Environmental Science and Technology* (2020), pp. 1–20.

[9] Enrique Vidal Ruiz, Francisco Casacuberta Nolla, and Hector Rulot Segovia. "Is the DTW "distance" really a metric? An algorithm reducing the number of DTW comparisons in isolated word recognition". In: *Speech Communication* 4.4 (1985), pp. 333–344.

[10] Romain Tavenard et al. "Tslearn, A Machine Learning Toolkit for Time Series Data". In: *Journal of Machine Learning Research* 21.118 (2020), pp. 1–6. URL: http://jmlr.org/papers/v21/20-091.html.

[11] Marja J Verhoef and Ann L Casebeer. *Broadening horizons: Integrating quantitative and qualitative research*. 1997.

[12] Pierre Winter. *covid19-global-weather-data*. https://www.kaggle.com/winterpierre91/covid19-global-weather-data.