BTHO COVID-19

Wangyang He, Sicong Huang, Jinhao Pan, Lu Zhang
CSCE 676 Data Mining

Introduction

Ever since 2019, the ongoing COVID-19 pandemic has affected everyone's daily lives. With the goal of helping to track the severity of viruses and seeking for deep hidden features of the existing data, we picked COVID-19 as our project topic. **Our project result could benefit the governments, people that are researching the pandemic and people are affected by or curious about the data trends of the disease.** Here are the three methods we used in the project:

- K-Means
- Outlier Detection (DeepLog)
- Generative Adversarial Network (GAN)

Dataset

The data set we chose is from the COVID-19 Data Repository by the **Center for Systems Science and Engineering (CSSE)** at Johns Hopkins University.

- Confirmed data
- Global deaths data
- Global recovered data

Method 1: K-Means for Countries

- Input: **"confirmed cases", "recovered cases", and "deaths"**
- Number of clusters = 3

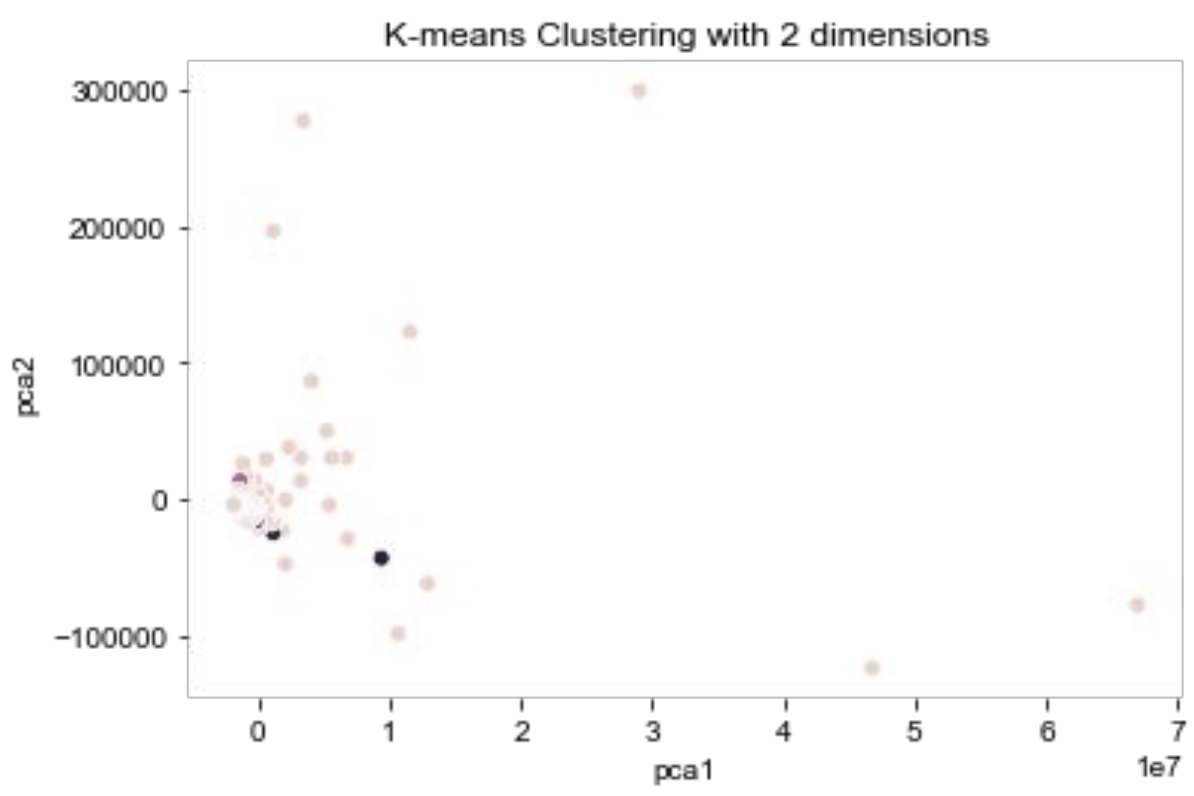


Figure 2. K-Means Clustering Visualization

In the graph above, each data point represents a country. Based on the visualization, we can see that even though the data points are separated into cluster 0-2, there are no clear groups on the graph itself. This is reasonable because all countries experienced similar situations during COVID and therefore, it is difficult to separate countries based on the given three inputs.

Preprocessing

In each csv file mentioned previously, the data were collected by regions under each country. During the preprocessing step, we processed data based on countries. Here are the steps of this process:

- Read three csv files into pandas data frames
- Group by country and calculate sum for each country
- Calculate additional information based on raw data:
    - **Active Cases** = Confirmed Cases - (Recovered Cases + Deaths)
    - **Daily Increase** = Today's Confirmed Cases - Yesterday's Confirmed Cases
    - **Mortality Rate** = Deaths / Confirmed Cases

The following table is an example visualization of the overall dataframe information.

Country Stats for COVID 19

Country	Confirmed	Active	Recovered	Deaths	Daily Increase	Mortality Rate
US	49085361	48296998	6298082	788363	34221	1.606
India	34633255	34159929	30974748	473326	8895	1.367
Brazil	22143091	21527455	17771228	615636	4844	2.78
United Kingdom	10523316	10377261	24693	146055	43361	1.388
Russia	9630296	9354472	5609682	275824	32013	2.864
Turkey	8903087	8825257	5478185	77830	19357	0.874
France	8021237	7900718	415111	120519	42252	1.502
Germany	6200937	6097813	3659260	103124	22945	1.663
Iran	6134465	6004265	3444798	130200	3109	2.122
Argentina	5340676	5224030	4615834	116646	1294	2.184
Spain	5202958	5114799	150376	88159	0	1.694
Italy	5109082	4974887	4144608	134195	15010	2.627
Colombia	5081064	4952284	4615354	128780	2077	2.535
Indonesia	4257685	4113818	2907920	143867	196	3.379
Mexico	3901263	3606060	2270427	295203	3811	7.567

Table 1. Example Country Stats for COVID-19

The following map is another example of visualization of daily increase cases color coded by country

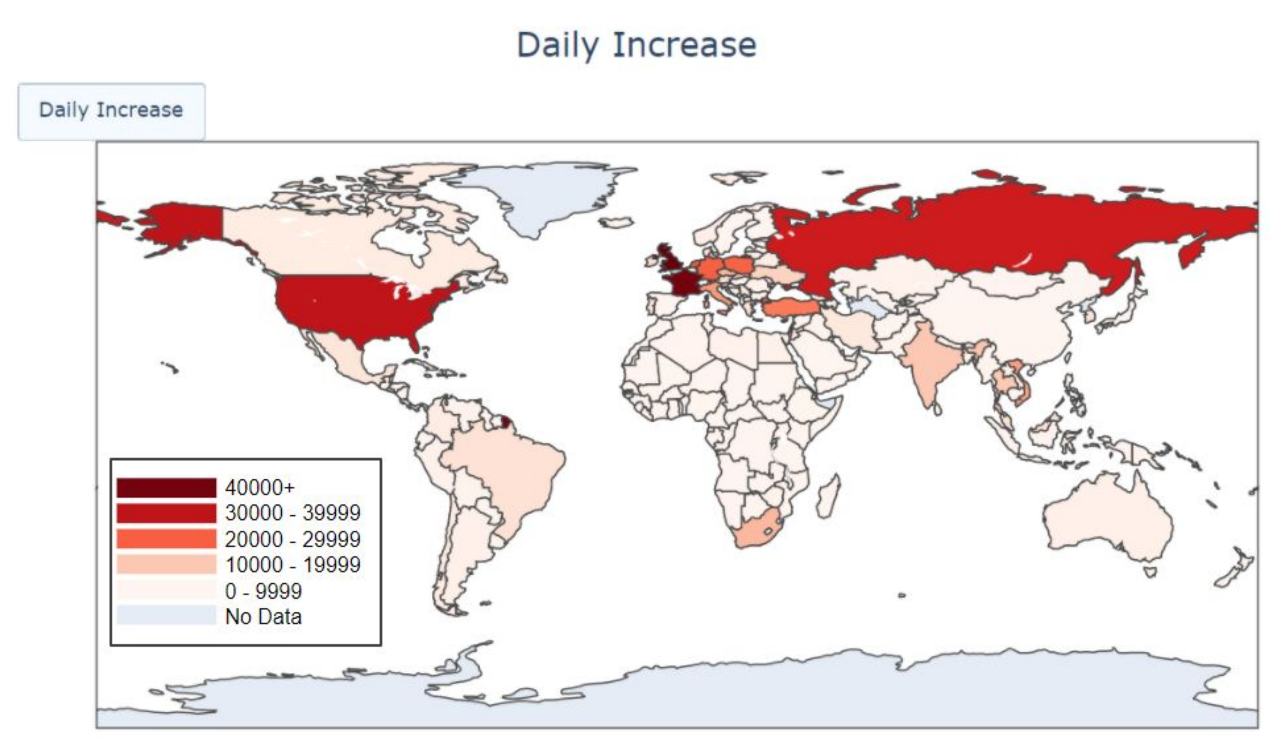


Figure 1. Daily Increase by Country, color coded

Method 2: Outlier Detection (DeepLog)

Outlier Detection using **DeepLog**. DeepLog is a deep neural network model utilizing Long Short-Term Memory **(LSTM)**, to model a system log as a natural language sequence. This allows DeepLog to automatically learn log patterns from normal execution, and detect anomalies when log patterns deviate from the model trained from log data under normal execution. The details of this implementation are:

- Input: calculated **active rate, recovered rate** and **death rate**
- Output: **outlier score for each country and outlier country index**

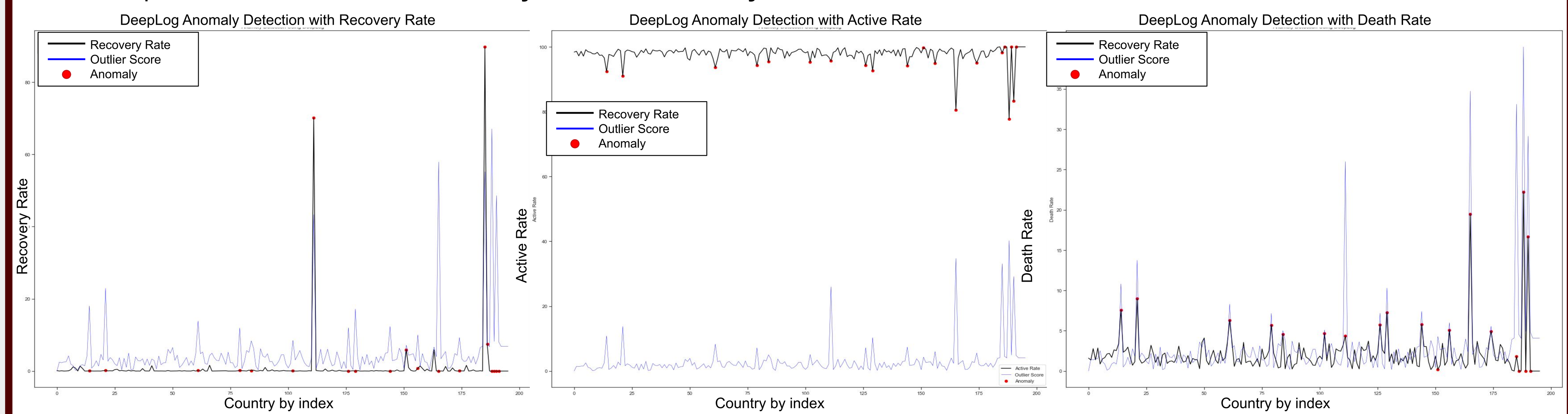
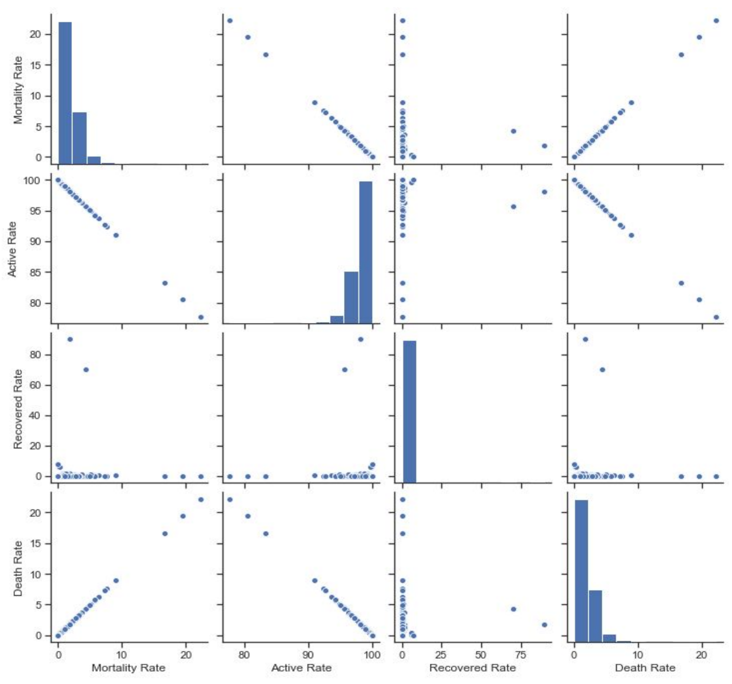


Figure 3, 4, 5. Outlier Detection on recover rate, active rate and death rate



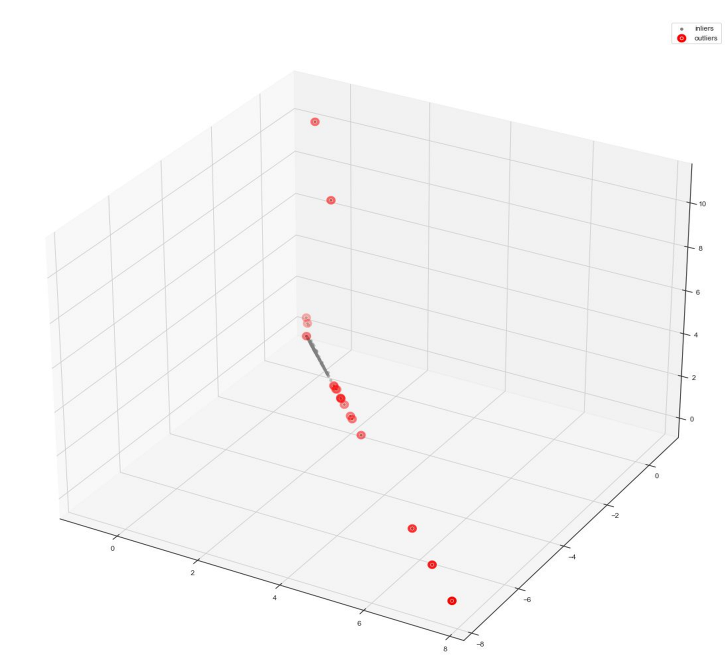


Figure 6, 7. 2d and 3d Outlier Detection Comparison

Afghanistan	Mexico
Bosnia & Herzegovina	MS Zaandam
China	Peru
Diamond Princess	Somalia
Ecuador	Sudan
Egypt	Syria
Holy See	Taiwan (China)
Iceland	Vanuatu
Liberia	Yemen
Marshall Islands	

Table 2. Outlier Countries.

Method 3: GAN

The last method we chose to implement was the **Generative Adversarial Network (GAN)**.

Benefits of GAN are:

- Ability to learn underlying layout and deep features
- Uses contest between generator and discriminator

In this project, we implemented the modified known algorithm, **Recurrent Conditional GAN (RCGAN)**. The plot below demonstrates our results compared to 1 DL baseline and 2 statistical based baselines on the same dataset.

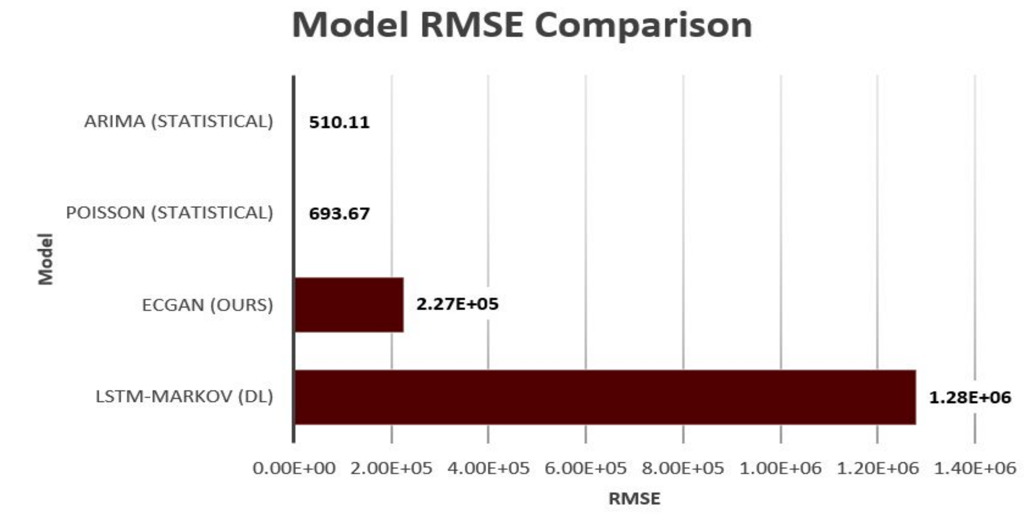


Figure 8. Model RMSE Comparison

As the result suggests, the architecture of GAN proves its advantage over conventional recurrent neural network, but the improvement is marginal and inconsequential compared to statical models.

Conclusion

Overall, we used data preprocessing and data visualization techniques as well as three methods, k-means, outlier detection and GAN to analyze the chosen COVID-19 dataset. The results from the three methods present different aspects of the pattern and trends in the given dataset.

As the COVID-19 pandemic continues to interface social activities and global trades, every part of the world still lives under the fear of outbreak. With the results of our project, we hope to help the world to understand the behavior of the virus better and to help guide authorities and citizens around the global to plan accordingly.

Related Work:

Du, Min, et al. *DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning*. https://www.cs.utah.edu/~lifeifei/papers/deeplog.pdf.

Hundman, Kyle, et al. "Detecting Spacecraft Anomalies Using Lstms and Nonparametric Dynamic Thresholding." *ArXiv.org*, 6 June 2018, https://arxiv.org/abs/1802.04431.

Zubair, Md., et al. "An Efficient K-Means Clustering Algorithm for Analysing COVID-19." *ArXiv.org*, 21 Dec. 2020, https://arxiv.org/abs/2101.03140.

Abdullah D;Susilo S;Ahmar AS;Rusli R;Hidayat R; "The Application of K-Means Clustering for Province Clustering in Indonesia of the Risk of the COVID-19 Pandemic Based on COVID-19 Data." *Quality & Quantity*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/34103768/.

Nikparvar, Behnam, et al. "Spatio-Temporal Prediction of the COVID-19 Pandemic in US Counties: Modeling with a Deep LSTM Neural Network." *Nature News*, Nature Publishing Group, 5 Nov. 2021, https://www.nature.com/articles/s41598-021-01119-3.

Ma, Ruifang, et al. "The Prediction and Analysis of Covid-19 Epidemic Trend by Combining LSTM and Markov Method." *Nature News*, Nature Publishing Group, 31 Aug. 2021, https://www.nature.com/articles/s41598-021-97037-5.

Kumar, R. Lakshmana, et al. "Recurrent Neural Network and Reinforcement Learning Model for Covid-19 Prediction." *Frontiers*, Frontiers, 1 Jan. 1AD, https://www.frontiersin.org/articles/10.3389/fpubh.2021.744100/full#h6.

Barría-Sandoval, Claudia, et al. "Prediction of Confirmed Cases of and Deaths Caused by Covid-19 in Chile through Time Series Techniques: A Comparative Study." *PLOS ONE*, Public Library of Science, https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0245414#sec005.

Esteban, Cristóbal, et al. "Real-Valued (Medical) Time Series Generation with Recurrent Conditional Gans." *ArXiv.org*, 4 Dec. 2017, https://arxiv.org/abs/1706.02633.