Visualisation and Classification of Novel COVID-19 Dataset

Baki Emre Bulut
Department of Electrical and Electronics Engineering
Hacettepe University,
Ankara, Turkey
bakiemrebulut@gmail.com

***Abstract*—This document is written for describe methods of visualization and classification algorithms for novel COVID-19 case data of different countries.**

***Index Terms*—Classification, Visualization, COVID-19**

I. Introduction

The first goal of this project is visualizations with graph from daily COVID-19 data of several countries by change graph and cumulative graph of death, recovered, confirmed and active patients in a user interface. Also, whether is graph proportional to population or counts of each data could be selected in this interface.
The other goal is classifying countries with: stage of epidemic; counts of death cases and confirmed cases per population; days until first peak time from begin of pandemic.
In this case, different daily data of COVID-19 are easily interpreted and deduce for a country from countries has same characteristics.
Thus, continuum and similarities of pandemic could easily observable.

II. Related Study

At the kaggle link of dataset [1], kernels of different visualizing methods and prediction models are actual but for daily updating dataset, success rate isn't goodly defined. In the other hand, I didn't find mostly same project with mine.

III. Used Data

I used Novel COVID-19 Dataset [1] and countryinfo dataset [2] in this project.
Before begin using these two data, I pre-processed these. Firstly, I insert new columns to COVID dataset they are basically: active patient number, first observation date, elapsed day from first patient for each country.
After that, I add sum of all Province Data for each date of countries to newly generated rows and remove all provinces data. Because, some countrys haven't got a provience and it is prevent comparison of countries.
Finally, clear countryinfo dataset for only using with populations of countries and renamed as population. I compare these data with COVID data. I corrected differently named same countries in each datasets. Then, remove non intersecting countries data from two dataset.Now, data sets are ready for use.

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
42248	42249	06/14/2020	Xinjiang	Mainland China	2020-06-15 03:33:14	76.0	3.0	73.0
42256	42257	06/14/2020	Yunnan	Mainland China	2020-06-15 03:33:14	185.0	2.0	183.0
42261	42262	06/14/2020	Zhejiang	Mainland China	2020-06-15 03:33:14	1268.0	1.0	1267.0

Fig. 1. From Original Dataset: Last 3 Data Row of Mainland China.

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
40233	40234	06/12/2020	NaN	Turkey	2020-06-13 03:33:14	175218.0	4778.0	149102.0
40962	40963	06/13/2020	NaN	Turkey	2020-06-14 03:33:15	176677.0	4792.0	150087.0
41691	41692	06/14/2020	NaN	Turkey	2020-06-15 03:33:14	178239.0	4807.0	151417.0

Fig. 2. From Original Dataset: Last 3 Data Row of Turkey.

	SNo	ObservationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered	Active	First	Days
16732	40268	06/12/2020	Mainland China	2020-06-13 03:33:14	83075.0	4634.0	78367.0	74.0	01/22/2020	143
16733	40997	06/13/2020	Mainland China	2020-06-14 03:33:15	83132.0	4634.0	78369.0	129.0	01/22/2020	144
16734	41726	06/14/2020	Mainland China	2020-06-15 03:33:14	83181.0	4634.0	78370.0	177.0	01/22/2020	145

Fig. 3. From New Dataset: Last 3 Data Row of Mainland China.

	SNo	ObservationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered	Active	First	Days
14426	40234	06/12/2020	Turkey	2020-06-13 03:33:14	175218.0	4778.0	149102.0	21338.0	03/11/2020	94
14571	40963	06/13/2020	Turkey	2020-06-14 03:33:15	176677.0	4792.0	150087.0	21798.0	03/11/2020	95
14716	41692	06/14/2020	Turkey	2020-06-15 03:33:14	178239.0	4807.0	151417.0	22015.0	03/11/2020	96

Fig. 4. From New Dataset: Last 3 Data Row of Turkey.

IV. Methods

A. Data Visualisation Method

After the preprocessing on the dataset, I performed visualisation interface for this dataset.
By the interactively using visualisation interface, I used ipywidgets library [3]. As the shown in "Figure 5",country selected by using "dropdown box", type of graphic/ case/ graph is whether proportional or not selected with "radio buttons". And plot button generate outputs.

Fig. 5. Different Interface Outputs

B. *Classification Method*

For the classification data, I need stage of pandemic,peak day, death state and confirmed state.

Firstly, I calculate stage of pandemic. So, I filter active cases of all countries with Savgol filter [4] for smoothing data and determine the peak points of each data with ”find peak” function from scipy.signal library [5].

By using peak points, I insert a new column to dataset as ”Peak Day” and by using number of peak points, I insert a new column to dataset as ”Stage”. Examples of all this operations for three different stage are visualised in “Figure 6” as the red plots are active cases, blue plots are filtered outputs of active cases, black rectangles are peak points and related labels are top of each frame.

After that, I planned observe ”Death State” and ”Confirmed State”. They are when calculated proportional of each data with populations, which country has high, middle or low risked.

So, I observed Cummulative Confirmed Cases and Cummulative Deaths Cases per each 1 million population. Then, for classification algorithm work correctly, I discard country has lower population and hasn't confirmed case or death case yet. For classificate with these 1D datasets, I used kernel density estimation [6] method. After this classification, I obtained 3 type output and named they are as ”low”,”mid” and ”high”.

Visualisation of this clustering is shown in “Figure 7” as

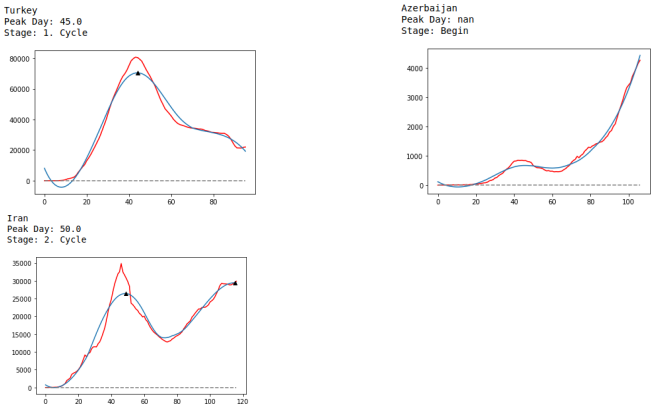


Fig. 6. Stage Detection Algorithm Output

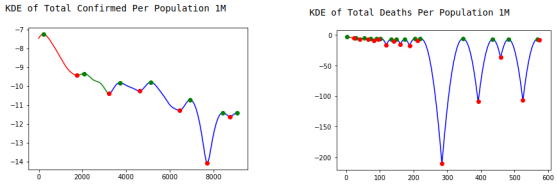


Fig. 7. Kernel Density Estimation

red markers cut classes, green markers are best estimates for the cluster centers and this plot is a linear approximation in range of input dataset and red, green blue parts are represent low,middle and high part respectively.

V. Result and Conclusion

In this project, I observe information from COVID-19 dataset. Firstly, I clean and pre processed datasets and regenerate meaningful new labels from this dataset. After that, I applied classtering algorithms to data on selected column.

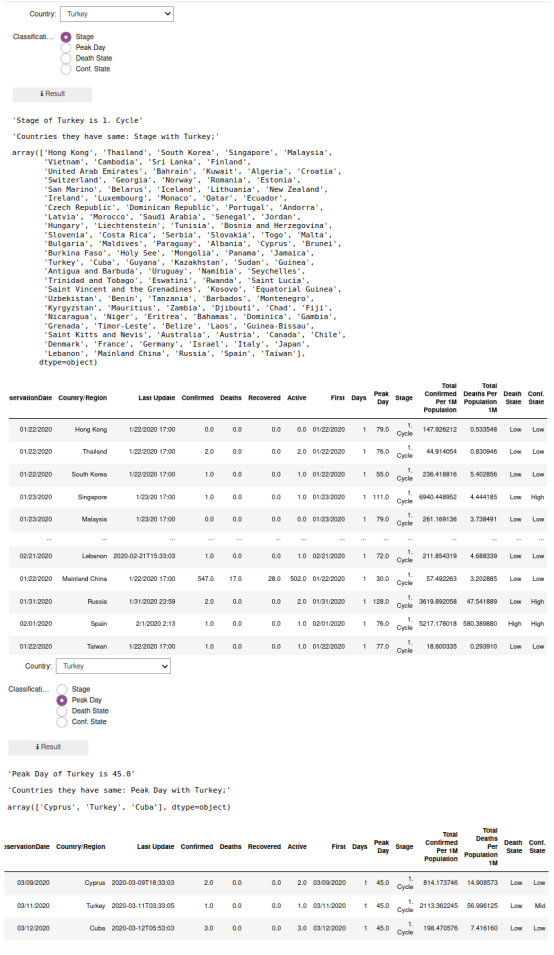
When using Kernel Density Estimation, applied bandwidth is change clustering points of subsets. I set this bandwidth for obtain closest lengths from subsets.

Also, Peak day difference has too much different value, so this situation prevent classification correctly.

In a result, I specify that 4 type classification of each countries. Due to didn't have truth table, classification outputs will help us when needed prediction next values of countries. For show the result of project, in “Figure 9”, I collect the similar countries with Turkey in each situations.

References

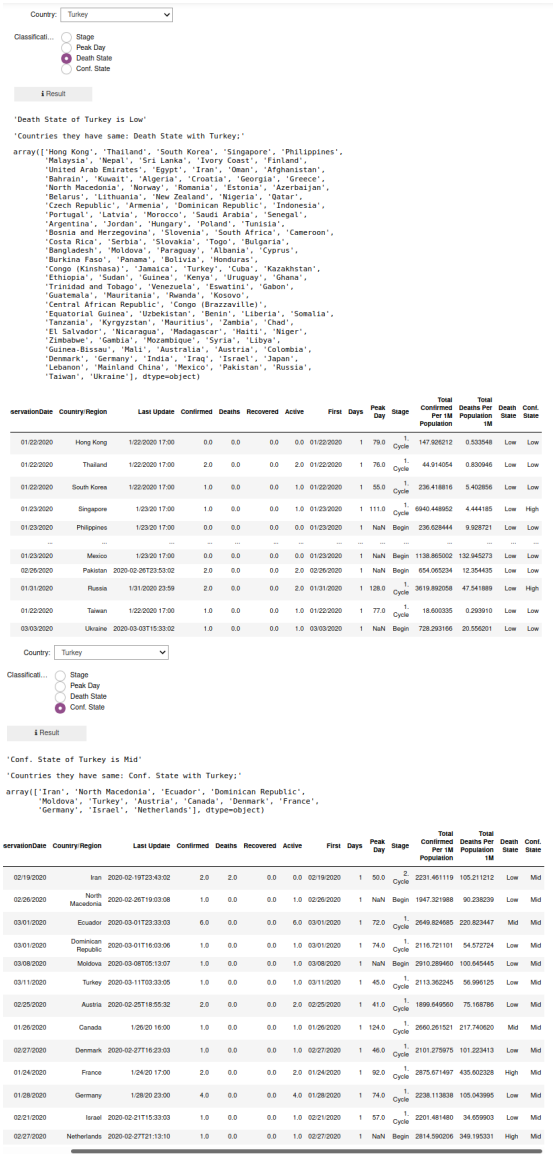
[1] Novel Corona Virus 2019 Dataset, www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
[2] countryinfo Dataset, www.kaggle.com/koryto/countryinfo
[3] IPyWidgets Library, ipywidgets.readthedocs.io
[4] Savgol Filter, docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html
[5] find_peaks Function, docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html
[6] Kernel Density Estimation, scikit-learn.org/stable/modules/density.htmlkernel-density



servationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered	Active	First	Days	Peak Day	Stage	Total Confirmed Per 1M Population	Total Deaths Per Population 1M	Death State	Conf. State
01/22/2020	Hong Kong	1/22/2020 17:00	0.0	0.0	0.0	0.0	01/22/2020	1	79.0	1. Cycle	147.926212	0.533548	Low	Low
01/22/2020	Thailand	1/22/2020 17:00	2.0	0.0	0.0	2.0	01/22/2020	1	76.0	1. Cycle	44.914054	0.830946	Low	Low
01/22/2020	South Korea	1/22/2020 17:00	1.0	0.0	0.0	1.0	01/22/2020	1	55.0	1. Cycle	236.418816	5.402856	Low	Low
01/23/2020	Singapore	1/23/20 17:00	1.0	0.0	0.0	1.0	01/23/2020	1	111.0	1. Cycle	6940.448952	4.444185	Low	High
01/23/2020	Malaysia	1/23/20 17:00	0.0	0.0	0.0	0.0	01/23/2020	1	79.0	1. Cycle	261.169136	3.738491	Low	Low
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
02/21/2020	Lebanon	2020-02-21T15:33:03	1.0	0.0	0.0	1.0	02/21/2020	1	72.0	1. Cycle	211.854319	4.688339	Low	Low
01/22/2020	Mainland China	1/22/2020 17:00	547.0	17.0	28.0	502.0	01/22/2020	1	30.0	1. Cycle	57.492263	3.202885	Low	Low
01/31/2020	Russia	1/31/2020 23:59	2.0	0.0	0.0	2.0	01/31/2020	1	128.0	1. Cycle	3619.892058	47.541889	Low	High
02/01/2020	Spain	2/1/2020 2:13	1.0	0.0	0.0	1.0	02/01/2020	1	76.0	1. Cycle	5217.178018	580.389880	High	High
01/22/2020	Taiwan	1/22/2020 17:00	1.0	0.0	0.0	1.0	01/22/2020	1	77.0	1. Cycle	18.600335	0.293910	Low	Low

servationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered	Active	First	Days	Peak Day	Stage	Total Confirmed Per 1M Population	Total Deaths Per Population 1M	Death State	Conf. State
03/09/2020	Cyprus	2020-03-09T18:33:03	2.0	0.0	0.0	2.0	03/09/2020	1	45.0	1. Cycle	814.173746	14.908573	Low	Low
03/11/2020	Turkey	2020-03-11T03:33:05	1.0	0.0	0.0	1.0	03/11/2020	1	45.0	1. Cycle	2113.362245	56.996125	Low	Mid
03/12/2020	Cuba	2020-03-12T05:53:03	3.0	0.0	0.0	3.0	03/12/2020	1	45.0	1. Cycle	198.470576	7.416160	Low	Low

Fig. 8. Result of Classification of Turkey Cases



servationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered	Active	First	Days	Peak Day	Stage	Total Confirmed Per 1M Population	Total Deaths Per Population 1M	Death State	Conf. State
01/22/2020	Hong Kong	1/22/2020 17:00	0.0	0.0	0.0	0.0	01/22/2020	1	79.0	1. Cycle	147.926212	0.533548	Low	Low
01/22/2020	Thailand	1/22/2020 17:00	2.0	0.0	0.0	2.0	01/22/2020	1	76.0	1. Cycle	44.914054	0.830946	Low	Low
01/22/2020	South Korea	1/22/2020 17:00	1.0	0.0	0.0	1.0	01/22/2020	1	55.0	1. Cycle	236.418816	5.402856	Low	Low
01/23/2020	Singapore	1/23/20 17:00	1.0	0.0	0.0	1.0	01/23/2020	1	111.0	1. Cycle	6940.448952	4.444185	Low	High
01/23/2020	Philippines	1/23/20 17:00	0.0	0.0	0.0	0.0	01/23/2020	1	NaN	Begin	236.628444	9.928721	Low	Low
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
01/23/2020	Mexico	1/23/20 17:00	0.0	0.0	0.0	0.0	01/23/2020	1	NaN	Begin	1138.865002	132.945273	Low	Low
02/26/2020	Pakistan	2020-02-26T23:53:02	2.0	0.0	0.0	2.0	02/26/2020	1	NaN	Begin	654.065234	12.354435	Low	Low
01/31/2020	Russia	1/31/2020 23:59	2.0	0.0	0.0	2.0	01/31/2020	1	128.0	1. Cycle	3619.892058	47.541889	Low	High
01/22/2020	Taiwan	1/22/2020 17:00	1.0	0.0	0.0	1.0	01/22/2020	1	77.0	1. Cycle	18.600335	0.293910	Low	Low
03/03/2020	Ukraine	2020-03-03T15:33:02	1.0	0.0	0.0	1.0	03/03/2020	1	NaN	Begin	728.293166	20.556201	Low	Low

servationDate	Country/Region	Last Update	Confirmed	Deaths	Recovered	Active	First	Days	Peak Day	Stage	Total Confirmed Per 1M Population	Total Deaths Per Population 1M	Death State	Conf. State
02/19/2020	Iran	2020-02-19T23:43:02	2.0	2.0	0.0	0.0	02/19/2020	1	50.0	2. Cycle	2231.461119	105.211212	Low	Mid
02/26/2020	North Macedonia	2020-02-26T19:03:08	1.0	0.0	0.0	1.0	02/26/2020	1	NaN	Begin	1947.321988	90.238239	Low	Mid
03/01/2020	Ecuador	2020-03-01T23:33:03	6.0	0.0	0.0	6.0	03/01/2020	1	72.0	1. Cycle	2649.824685	220.823447	Mid	Mid
03/01/2020	Dominican Republic	2020-03-01T16:03:06	1.0	0.0	0.0	1.0	03/01/2020	1	74.0	1. Cycle	2116.721101	54.572724	Low	Mid
03/08/2020	Moldova	2020-03-08T05:13:07	1.0	0.0	0.0	1.0	03/08/2020	1	NaN	Begin	2910.289460	100.645445	Low	Mid
03/11/2020	Turkey	2020-03-11T03:33:05	1.0	0.0	0.0	1.0	03/11/2020	1	45.0	1. Cycle	2113.362245	56.996125	Low	Mid
02/25/2020	Austria	2020-02-25T18:55:32	2.0	0.0	0.0	2.0	02/25/2020	1	41.0	1. Cycle	1899.649560	75.168786	Low	Mid
01/26/2020	Canada	1/26/20 16:00	1.0	0.0	0.0	1.0	01/26/2020	1	124.0	1. Cycle	2660.261521	217.740620	Mid	Mid
02/27/2020	Denmark	2020-02-27T16:23:03	1.0	0.0	0.0	1.0	02/27/2020	1	46.0	1. Cycle	2101.275975	101.223413	Low	Mid
01/24/2020	France	1/24/20 17:00	2.0	0.0	0.0	2.0	01/24/2020	1	92.0	1. Cycle	2875.671497	435.602328	High	Mid
01/28/2020	Germany	1/28/20 23:00	4.0	0.0	0.0	4.0	01/28/2020	1	74.0	1. Cycle	2238.113838	105.043995	Low	Mid
02/21/2020	Israel	2020-02-21T15:33:03	1.0	0.0	0.0	1.0	02/21/2020	1	57.0	1. Cycle	2201.481480	34.059903	Low	Mid
02/27/2020	Netherlands	2020-02-27T21:13:10	1.0	0.0	0.0	1.0	02/27/2020	1	NaN	Begin	2814.590206	349.195331	High	Mid

Fig. 9. Result of Classification of Turkey Cases