

Research of COVID-19 impact and trend in USA using statistical approach.

Artsiom Naslednikau
Email: art.naslednikau@gmail.com

Abstract—This paper presents a research of a public dataset United States COVID-19 Cases and Deaths by State over Time from Centers for Disease Control and Prevention containing 1 721 928 instances with goals to forecast a trend and to explore an impact on population of USA.

To determine research results statistical approach was applied which consists of methods such as attribute selection, aggregation, pattern recognition, feature classification, multiple regression models, k-means clustering and visualization.

To improve and to cross-validate accuracy of research results linear and isotonic regression models were used.

To determine deviations of inflated states data have been grouped and classified by mean feature.

The visualization part of research consists of US density states maps and k-means cluster.

I. INTRODUCTION

COVID-19 caused global pandemic mode to be enabled in 2019 [cov20] [arx].

The first case was identified in Wuhan, China, in December 2019.

SARS-CoV-2 is a cause which is defined as severe acute respiratory syndrome coronavirus 2.

COVID-19 is an effect of coronavirus disease 2019.

Previously it was discovered 4 pandemics for the last century: in 1918, 1957, 1968 and 2009 [cdc].

The research is dedicated to the fifth pandemic mode of century - COVID-19.

Technical part of research consists of applying following computer science methods over the United States COVID-19 Cases and Deaths by State over Time dataset.

- attribute selection
- aggregation
- pattern recognition
- mean feature classification
- linear and isotonic regression models
- density map plotting
- k-means clustering

II. NORMALIZATION

Attribute Selection helps to clean data.

Dataset consists of 15 unique properties.

After applying method of attribute selection only 3 properties left.

Total number of confirmed instances with submission date allows to gain more accurate numbers for predictor model.

Other properties have missed or inconsistent values.

| Property | Description | Type |
|-----------------|----------------------------------|-------------|
| submission_date | Date of counts | Date & Time |
| conf_cases | Total confirmed cases | Number |
| conf_death | Total number of confirmed deaths | Number |

Table I: Selected dataset attributes [cdc20]

The dataset has 1 721 928 number of instances to work with.

In our research actual range was used - 04/15/20 - 12/27/20 with 0 as starting point at 04/01/20.

Total discovered points to works with is 55.

Selected attributes were grouped by 1, 5, 10, 15, 20, 25 days of each month with have medium values.

Numbers have cumulative feature which is shown in their increasing exponential progression III.

Recovery formula 1 has been introduced to increase accuracy of research and observe deviations.

$$(1) \quad Recovery = Case - Death$$

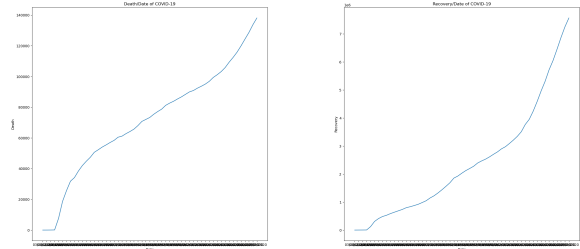


Figure 1: Raw Death/Recovery Rates

The slope of first image has sharp slope in the beginning and linear slope form in the tail.

Meanwhile beginning of another image has linear line slope and the tail has sharp curve form.

Both functions growth bounded above by $n \log n$ and below by \sqrt{n} asymptotically.

$$(2) \quad f(x) \in O(n \log n)$$

$$(3) \quad f(x) \in \Omega(\sqrt{n})$$

To get rid off increasing progression that the slope is too steep the shatter semi-curve segmentation algorithm was developed 1 1.

Algorithm 1 Shatter semi-curve segmentation algorithm

```

function SHATTERSEGMENTATION(DictionarySet  $F$ )
   $Result \leftarrow F$  ▷ Copy Data
  for ( $i = 0$ ;  $i < 1\,721\,928$ ;  $i++$ ) do
     $A \leftarrow F_i$  ▷ Current Day
     $B \leftarrow F_{i+1}$  ▷ Next Day
     $Result_i \leftarrow A - B$  ▷ Difference
  end for
  return  $Result$ 
end function

```

Applied segmented algorithm results 2.

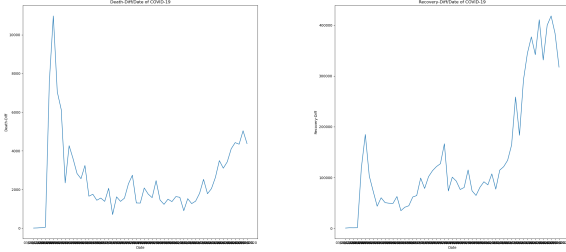


Figure 2: Segmented Death/Recovery Rates

The first image has strong noticeable wave in the beginning of the line.

The following distribution of the first image is almost normal except the rising wave in the tail.

The second image has 3 noticeable waves.

First and second waves are almost similar except the biggest wave size in the tail.

Alternatively, square-normalization approach was applied to reduce overall value of Y-axis and to preserve curve angles as a vector ratio scale 3.

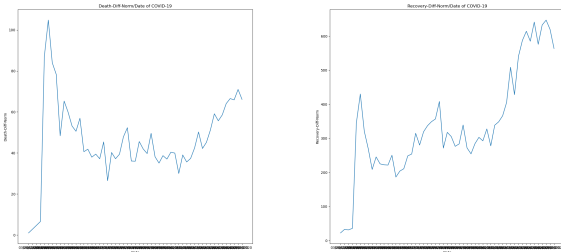


Figure 3: Square-normalization Death/Recovery Rates

Results of the applied shatter segmentation 1 and square-normalization algorithms.

Waves became less perceptible and more smoothly because of angles acute forms.

Furthermore, Segmentation algorithm 1 was applied to population dataset 4.

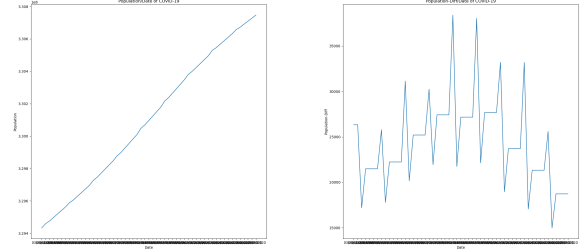


Figure 4: Raw and Segmented population of USA 2020

The first function bounded above by $n \log n$ 2 and below by \sqrt{n} 3 asymptotically.

The second image has noticeable wave in the middle of image.

III. RESULTS

A. COVID-19 impact on population of USA

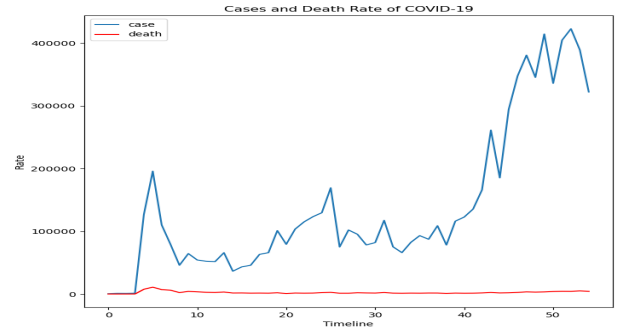


Figure 5: Case and Death Rate

Death rate curve has only one small wave in the beginning of the line.

Case rate has 3 waves with the strongest wave in the tail of the curve.

The death rate function has linear form with constant 4 asymptotic.

$$(4) \quad f(x) \in O(1)$$

Only small impact on population is noticeable in the beginning of the timeline 5.

The death rate function 6 has 1 wave in the beginning of the timeline.

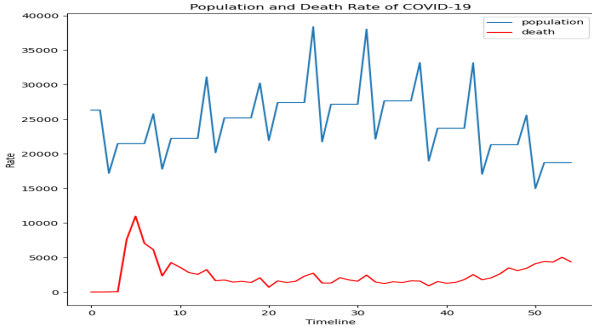


Figure 6: Population and Death Rate Correlation

Same as on the previous figure, the death rate function 6 has linear form with constant 4 asymptotic.

The population rate has 1 wave in the middle of the timeline.

It can be assumed because of that the coronavirus death rate has no effect on population of the USA, otherwise the death rate suppose to have a wave in the middle.

B. Regressions Tendency

Multiple regression models were used in research to cross-validate estimated forecast results 7.

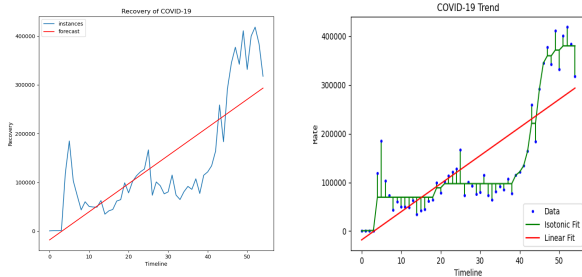


Figure 7: Recovery Rate Linear and Isotonic Regressions

Both models have similar results with small deviation.

Models show that recovery trend is growing up.

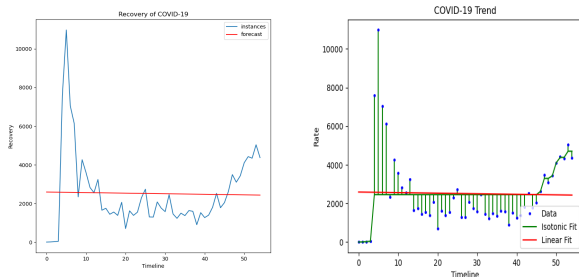


Figure 8: Death Rate Linear and Isotonic Regressions

Both functions strives for negative value and trend is death negative 8.

It can be concluded that trend is positive based on that

- death rate is going down 8
- recovery rate is going up 7
- population rate is going up 6

C. Data Classification

Data II was grouped by state with total count of 26 states V.

Each state was classified by mean feature [wik20] for death and recovery attributes as threshold to identify states with inflated rates.

Death feature classification results: 17 state are mean and 9 are overmean V.

Recovery feature classification results: 15 state are mean and 11 are overmean V.

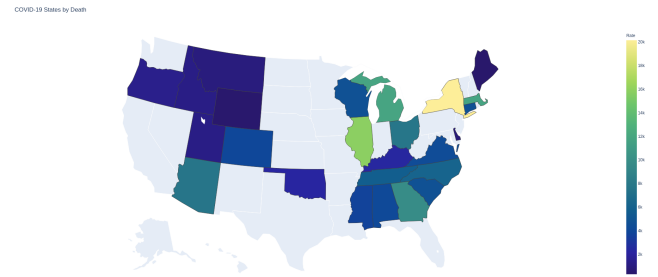


Figure 9: States Map By Death Attribute

New York state which is located in the north-east of the map has the highest death rate density between among states 9.

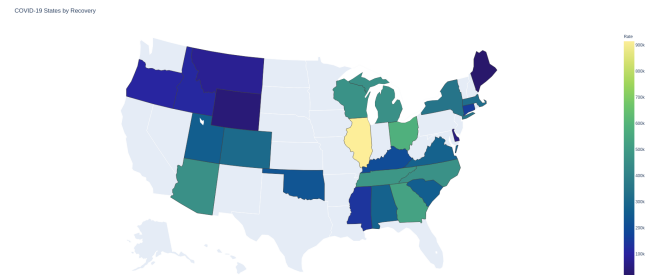


Figure 10: States Map By Recovery Attribute

Illinois state which is located in the middle of the map has the highest recovery rate 10.

D. Clustering

Total 2 sub-clusters have been generated based on k-means algorithms using unsupervised learning 11.

The k-mean algorithm experienced error when more than 5 clusters were calculated.

| Date | Case | Death | Recovery |
|------------|---------|--------|----------|
| 03/25/2020 | 504 | 1 | 503 |
| 04/01/2020 | 1584 | 9 | 1575 |
| 04/05/2020 | 2585 | 31 | 2554 |
| 04/10/2020 | 3908 | 74 | 3834 |
| 04/15/2020 | 130619 | 7672 | 122947 |
| 04/20/2020 | 326443 | 18652 | 307791 |
| 04/25/2020 | 436615 | 25705 | 410910 |
| 05/01/2020 | 515576 | 31842 | 483734 |
| 05/05/2020 | 561470 | 34182 | 527288 |
| 05/10/2020 | 625946 | 38446 | 587500 |
| 05/15/2020 | 680173 | 42034 | 638139 |
| 05/20/2020 | 732398 | 44858 | 687540 |
| 05/25/2020 | 784045 | 47415 | 736630 |
| 06/01/2020 | 849940 | 50655 | 799285 |
| 06/05/2020 | 886338 | 52305 | 834033 |
| 06/10/2020 | 929645 | 54055 | 875590 |
| 06/15/2020 | 975503 | 55495 | 920008 |
| 06/20/2020 | 1038759 | 57050 | 981709 |
| 06/25/2020 | 1104655 | 58432 | 1046223 |
| 07/01/2020 | 1205765 | 60491 | 1145274 |
| 07/05/2020 | 1284999 | 61192 | 1223807 |
| 07/10/2020 | 1388488 | 62810 | 1325678 |
| 07/15/2020 | 1503448 | 64192 | 1439256 |
| 07/20/2020 | 1626670 | 65743 | 1560927 |
| 07/25/2020 | 1756264 | 68034 | 1688230 |
| 08/01/2020 | 1925549 | 70771 | 1854778 |
| 08/05/2020 | 2000485 | 72072 | 1928413 |
| 08/10/2020 | 2102576 | 73367 | 2029209 |
| 08/15/2020 | 2197800 | 75444 | 2122356 |
| 08/20/2020 | 2276011 | 77204 | 2198807 |
| 08/25/2020 | 2358032 | 78780 | 2279252 |
| 09/01/2020 | 2475466 | 81236 | 2394230 |
| 09/05/2020 | 2550763 | 82689 | 2468074 |
| 09/10/2020 | 2616739 | 83918 | 2532821 |
| 09/15/2020 | 2699039 | 85415 | 2613624 |
| 09/20/2020 | 2792140 | 86785 | 2705355 |
| 09/25/2020 | 2879397 | 88414 | 2790983 |
| 10/01/2020 | 2988233 | 90009 | 2898224 |
| 10/05/2020 | 3066371 | 90909 | 2975462 |
| 10/10/2020 | 3182475 | 92432 | 3090043 |
| 10/15/2020 | 3305166 | 93699 | 3211467 |
| 10/20/2020 | 3440605 | 95094 | 3345511 |
| 10/25/2020 | 3606521 | 96895 | 3509626 |
| 11/01/2020 | 3867726 | 99417 | 3768309 |
| 11/05/2020 | 4052847 | 101196 | 3951651 |
| 11/10/2020 | 4346956 | 103238 | 4243718 |
| 11/15/2020 | 4694410 | 105859 | 4588551 |
| 11/20/2020 | 5075047 | 109352 | 4965695 |
| 11/25/2020 | 5420314 | 112452 | 5307862 |
| 12/01/2020 | 5834653 | 115881 | 5718772 |
| 12/05/2020 | 6170432 | 119987 | 6050445 |
| 12/10/2020 | 6574954 | 124411 | 6450543 |
| 12/15/2020 | 6997693 | 128753 | 6868940 |
| 12/20/2020 | 7386553 | 133790 | 7252763 |
| 12/25/2020 | 7708482 | 138161 | 7570321 |

Table III: Grouped selected attributes by [cdc20]

| Date | Population |
|--------------------|------------|
| midrule 03/25/2020 | 329433166 |
| 04/01/2020 | 329459499 |
| 04/05/2020 | 329476690 |
| 04/10/2020 | 329498178 |
| 04/15/2020 | 329519667 |
| 04/20/2020 | 329541155 |
| 04/25/2020 | 329562644 |
| 05/01/2020 | 329588430 |
| 05/05/2020 | 329606219 |
| 05/10/2020 | 329628455 |
| 05/15/2020 | 329650692 |
| 05/20/2020 | 329672928 |
| 05/25/2020 | 329695164 |
| 06/01/2020 | 329726295 |
| 06/05/2020 | 329746456 |
| 06/10/2020 | 329771658 |
| 06/15/2020 | 329796860 |
| 06/20/2020 | 329822061 |
| 06/25/2020 | 329847263 |
| 07/01/2020 | 329877505 |
| 07/05/2020 | 329899443 |
| 07/10/2020 | 329926866 |
| 07/15/2020 | 329954289 |
| 07/20/2020 | 329981711 |
| 07/25/2020 | 330009134 |
| 08/01/2020 | 330047526 |
| 08/05/2020 | 330069263 |
| 08/10/2020 | 330096434 |
| 08/15/2020 | 330123605 |
| 08/20/2020 | 330150776 |
| 08/25/2020 | 330177947 |
| 09/01/2020 | 330215986 |
| 09/05/2020 | 330238125 |
| 09/10/2020 | 330265798 |
| 09/15/2020 | 330293471 |
| 09/20/2020 | 330321145 |
| 09/25/2020 | 330348818 |
| 10/01/2020 | 330382026 |
| 10/05/2020 | 330400989 |
| 10/10/2020 | 330424693 |
| 10/15/2020 | 330448397 |
| 10/20/2020 | 330472101 |
| 10/25/2020 | 330495805 |
| 11/01/2020 | 330528990 |
| 11/05/2020 | 330546051 |
| 11/10/2020 | 330567378 |
| 11/15/2020 | 330588705 |
| 11/20/2020 | 330610031 |
| 11/25/2020 | 330631358 |
| 12/01/2020 | 330656950 |
| 12/05/2020 | 330671932 |
| 12/10/2020 | 330690661 |
| 12/15/2020 | 330709389 |
| 12/20/2020 | 330728117 |
| 12/25/2020 | 330746845 |

Table IV: Population of USA [pop20]

| State | Death | DeathClass | Recovery | RecoveryClass |
|-------|-------|------------|----------|---------------|
| NC | 6152 | OVERMEAN | 461328 | OVERMEAN |
| ID | 1191 | MEAN | 111206 | MEAN |
| WY | 373 | MEAN | 36327 | MEAN |
| NY | 20151 | OVERMEAN | 345322 | OVERMEAN |
| OH | 7687 | OVERMEAN | 582943 | OVERMEAN |
| CT | 4686 | MEAN | 157763 | MEAN |
| MS | 3718 | MEAN | 140018 | MEAN |
| AZ | 7635 | OVERMEAN | 457378 | OVERMEAN |
| IL | 15799 | OVERMEAN | 915050 | OVERMEAN |
| VA | 4272 | MEAN | 274881 | MEAN |
| TN | 5646 | OVERMEAN | 480082 | OVERMEAN |
| ME | 314 | MEAN | 18203 | MEAN |
| MT | 916 | MEAN | 78013 | MEAN |
| CO | 3948 | MEAN | 304959 | OVERMEAN |
| MA | 11706 | OVERMEAN | 316601 | OVERMEAN |
| WI | 4679 | MEAN | 463220 | OVERMEAN |
| AL | 4096 | MEAN | 273658 | MEAN |
| MI | 11775 | OVERMEAN | 458153 | OVERMEAN |
| OR | 1407 | MEAN | 104591 | MEAN |
| OK | 2218 | MEAN | 227244 | MEAN |
| KY | 2312 | MEAN | 199732 | MEAN |
| DE | 791 | MEAN | 50695 | MEAN |
| GA | 9656 | OVERMEAN | 527423 | OVERMEAN |
| SC | 4662 | MEAN | 258730 | MEAN |
| PR | 1187 | MEAN | 67274 | MEAN |
| UT | 1182 | MEAN | 259407 | MEAN |

Table V: Classified states by mean feature