Could COVID-2019 Be Stopped Earlier?

In most cases the to predict the amount of people infected by particular disease is much simpler than to predict that certain actions taken by government (businesses, doctors, etc.) will stop the spread of it.

As the last pandemic of coronovirus shows, the last action that world can take to stop the spread of virus is to quarantine people at homes. As a result, this leads to disastrous consequences for the economy.

Since to invent the treatment against the unknown virus takes a long time, the only way to fight the disease is to detect and close the ways of its spreading on early stages.

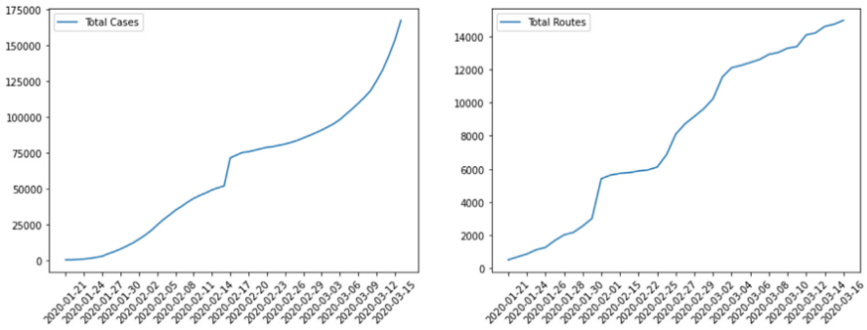
When it comes about transmission of respiratory diseases from person to person the most obvious ways of spreading are the crowded places and transportation. This becomes almost absolute true for international or even intercontinental spreading; the most (if not only single) possible way is international air routes: crowded airports and aircrafts.

To test the last assumption we can take the chronology of COVID-19 spreading across countries and the number of international air routes between countries and create such table:

Date	Country	Cum.Cases	Cum.Routes	Comment
12/1/2019	China	1	0	-
1/13/2020	Thailand	200	73	China+Thailand
1/15/2020	Japan	230	339	all previous + Japan
1/20/2020	South Korea	282	512	all previous + South Korea
1/21/2020	Taiwan, United States	...	...	all previous + Taiwan+United States
...	...	...	...	...

“Cum.Cases” is the total number of cases in all countries at current date.
“Cum.Routes” is the total number of routes between all those countries

where at least one case on corona virus was detected at this date. Now let's visualize these data:



We can admit that they are correlated and cumulative routes can be a good predictor (at least at early stages of epidemics) for cumulative cases. Indeed, using R simple linear regression and log transformation (as data has signs of exponential distribution) we are going to show that "Cum.Cases" (as independent variable) can explain >95% of variance of "Cum.Cases" (as response variable).

```
Call:
lm(formula = log(Cum.Cases) ~ log(Cum.Routes), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8704 -0.2454 -0.1616  0.1119  0.8578

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -5.31007    0.58470  -9.082  9.9e-11 ***
log(Cum.Routes)  1.81117    0.06714  26.974  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3951 on 35 degrees of freedom
Multiple R-squared:  0.9541,    Adjusted R-squared:  0.9528
F-statistic: 727.6 on 1 and 35 DF,  p-value: < 2.2e-16
```

But this relationship is not interesting in itself. What could much more promising and of high importance is the possibility to define those countries where the virus has higher probability to appear the next "time". Such early alarm system can help to stop (or at least slow down) the spreading of virus. Closing international air transportation from 3–4 already "infected" countries to those that are in list of 10 most probable next ones is much less disastrously for world economy than

current measures against COVID-19, taken after two months from the beginning of epidemic, when half of the countries have already reported about their first case.

So our goal to create such approach that can predict for “today” the next 10 countries that are more likely to be infected “tomorrow”. We start with the simplest probabilistic model: logistic regression. It will be based on the principals of forward testing, when the model is:
- retrained each specified period (daily in our case);
- using only the previous data;
- and predicts the values of the next period .

For each date train data includes features for all countries and additional label column which indicates that this country is (or not) “infected” at this date. Also it contains multiply entries for each countries defined by the number of already infected countries.

For example, the first run of the model (on 13 January of 2020) could be made when the second country (Thailand) reported its first case. The training data would have the following structure (several additional features were added to “routes” one):

Routes	Cases	Dist	Phys	PopulDens	Popul	Water	Label	Country	Pr_Country	Date
0	1.00	3315	1.79	148.35	0.16	0.43	0	Afghanistan	China	1/13/2020
0	1.00	7040	1.79	148.35	0.16	0.43	0	Albania	China	1/13/2020
0	1.00	8596	1.79	148.35	0.16	0.43	0	Tanzania	China	1/13/2020
72	1.00	2245	1.79	148.35	0.16	0.43	1	Thailand	China	1/13/2020
0	1.00	10629	1.79	148.35	0.16	0.43	0	Togo	China	1/13/2020
0	1.00	9668	1.79	148.35	0.16	0.43	0	Zambia	China	1/13/2020
0	1.00	9963	1.79	148.35	0.16	0.43	0	Zimbabwe	China	1/13/2020

“Pr_country”—already infected country before this date.
“Country”—the list of all countries.
“Label”—defines if current country is infected at this date.
“Routes”—the number of air routes between current country and already infected (at 1/13/2020 only China reported about COVID-19).
“Cases”—share of all cases that are registered in already infected country (at 1/13/2020 only China reported about COVID-19, thus “1.00” defines 100% of all world cases).
“Dist”—distance from current country to already “infected” (km).
“Phys”—the number of physicians in already infected country (per 1,000 people).
“PopulDens”—the density of population in already infected country (citizens per sq. km of land area)
“Popul”—share of the population of already infected country among all countries.
“Water”—annual freshwater withdrawals (1000 liters per 1 citizen).

Trained on such data model is used to produce probabilities for the rest “non-infected” countries (based on this features), which then sorted from the highest to lowest probability and Top-10 countries from this list (all countries minus China and Thailand) are considered as the next that will be infected.

As the accuracy of the model we can take the percentage of countries that were right predicted compared to the next real 10 ones. The next picture presents this idea. Based on different combination of features we get lists of 10 most likely countries (“Country” column) to be infected on 1/16/2020 (as you may note the training process was run on data before 1/16/2020 and prediction is made for this date). “Real Next 10” column contains 10 next countries that were really infected (starting from 1/16/2016 and further—see above pictures). The country that reported the first case exactly on this date is colored by green color; yellow color represents predicted countries that exists in real next 10; white color—wrong predictions.

"Routes"

Country	Probs	Date	Real Next 10
Taiwan	7.31%	1/16/2020	Japan
South Korea	5.62%	1/16/2020	South Korea
Japan	5.20%	1/16/2020	Taiwan
Hong Kong	4.42%	1/16/2020	United States
Singapore	2.86%	1/16/2020	Singapore
United States	2.80%	1/16/2020	Vietnam
Russia	2.73%	1/16/2020	Australia
Malaysia	2.73%	1/16/2020	France
Macau	2.62%	1/16/2020	Nepal
Vietnam	2.36%	1/16/2020	Malaysia

"Routes", "Phys", "Cases", "Water"

Country	Probs	Date	Real Next 10
Taiwan	12.62%	1/16/2020	Japan
South Korea	9.59%	1/16/2020	South Korea
Japan	8.87%	1/16/2020	Taiwan
Hong Kong	7.70%	1/16/2020	United States
Malaysia	6.43%	1/16/2020	Singapore
Singapore	6.05%	1/16/2020	Vietnam
India	4.95%	1/16/2020	Australia
Russia	4.86%	1/16/2020	France
Australia	4.74%	1/16/2020	Nepal
Burma	4.53%	1/16/2020	Malaysia

"Routes", "Phys", "Cases", "Water", "PopulDens", "Popul", "Dist"

Country	Probability	Date	Real Next 10
Laos	10.33%	1/16/2020	Japan
Taiwan	8.83%	1/16/2020	South Korea
Hong Kong	8.76%	1/16/2020	Taiwan
Cambodia	8.42%	1/16/2020	United States
Burma	8.06%	1/16/2020	Singapore
Vietnam	7.26%	1/16/2020	Vietnam
Macau	5.88%	1/16/2020	Australia
South Korea	5.75%	1/16/2020	France
Malaysia	5.02%	1/16/2020	Nepal
Bangladesh	4.99%	1/16/2020	Malaysia

Generally speaking, if governments of Japan, South Korea, United States, Singapore and rest 6 had closed the air traffic with Thailand and China, this might cause the decrease in spreading of virus, because 70% were correctly predicted (left table).

Taking in account that incubation period of COVID-19 is 14 days, the next Japan had might be infected before it closed routes with China and Thailand, thus on 1/16/2020 we would run the model once more. The training data for this date would looked like this:

Routes	Cases	Dist	Phys	PopulDens	Popul	Water	Label	Country	Pr_Country	Date
0	0.005	3885	0.81	135.90	0.01	0.83	0	Afghanistan	Thailand	1/16/2020
0	0.995	3315	1.79	148.35	0.16	0.43	0	Afghanistan	China	1/16/2020
0	0.005	8096	0.81	135.90	0.01	0.83	0	Albania	Thailand	1/16/2020
0	0.995	7040	1.79	148.35	0.16	0.43	0	Albania	China	1/16/2020
...	...	...	...	...	...	...	...	...	...	...
0	0.005	16231	0.81	135.90	0.01	0.83	0	Jamaica	Thailand	1/16/2020
0	0.995	14011	1.79	148.35	0.16	0.43	0	Jamaica	China	1/16/2020
22	0.005	4315	0.81	135.90	0.01	0.83	1	Japan	Thailand	1/16/2020
140	0.995	3047	1.79	148.35	0.16	0.43	1	Japan	China	1/16/2020
...	...	...	...	...	...	...	...	...	...	...
0	0.005	7682	0.81	135.90	0.01	0.83	0	Tanzania	Thailand	1/16/2020
0	0.995	8596	1.79	148.35	0.16	0.43	0	Tanzania	China	1/16/2020
72	1	2245	1.79	148.35	0.16	0.43	1	Thailand	China	1/13/2020
...	...	...	...	...	...	...	...	...	...	...
0	0.005	8662	0.81	135.90	0.01	0.83	0	Zambia	Thailand	1/16/2020
0	0.995	9668	1.79	148.35	0.16	0.43	0	Zambia	China	1/16/2020
0	0.005	8761	0.81	135.90	0.01	0.83	0	Zimbabwe	Thailand	1/16/2020
0	0.995	9963	1.79	148.35	0.16	0.43	0	Zimbabwe	China	1/16/2020

As you may note, for each country, that were not already infected on 1/16/2020, and for Japan, that announced first case on this date, there are two entries with values related to China and Thailand. To give the model more information about positive cases, here is also previous entry for Thailand labeled as “1”. For this time the predictions are the following:

"Routes"

Country	Probs	Date	Real Next 10
Taiwan	20.68%	1/20/2020	South Korea
South Korea	16.08%	1/20/2020	Taiwan
Hong Kong	7.27%	1/20/2020	United States
United States	4.79%	1/20/2020	Singapore
Singapore	2.64%	1/20/2020	Vietnam
Russia	2.38%	1/20/2020	Australia
Malaysia	2.30%	1/20/2020	France
Macau	2.20%	1/20/2020	Nepal
Vietnam	1.95%	1/20/2020	Malaysia
Australia	1.83%	1/20/2020	Canada

"Routes", "Phys", "Cases", "Water"

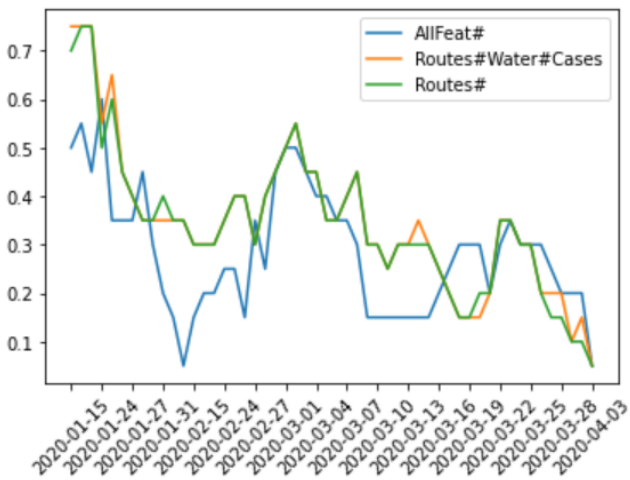
Country	Probs	Date	Real Next 10
Taiwan	21.92%	1/20/2020	South Korea
South Korea	12.74%	1/20/2020	Taiwan
Hong Kong	5.33%	1/20/2020	United States
United States	1.59%	1/20/2020	Singapore
Malaysia	1.43%	1/20/2020	Vietnam
Singapore	1.43%	1/20/2020	Australia
Russia	1.11%	1/20/2020	France
Macau	0.96%	1/20/2020	Nepal
India	0.85%	1/20/2020	Malaysia
Australia	0.84%	1/20/2020	Canada

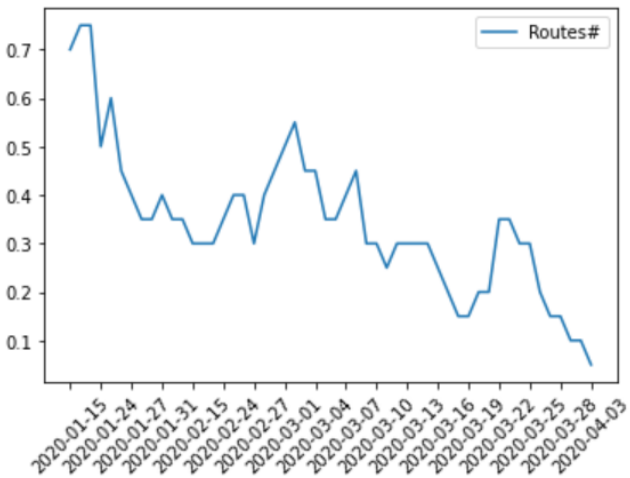
"Routes", "Phys", "Cases", "Water", "PopulDens", "Popul", "Dist"

Country	Probability	Date	Real Next 10
Taiwan	19.80%	1/20/2020	South Korea
South Korea	12.13%	1/20/2020	Taiwan
Hong Kong	7.92%	1/20/2020	United States
Macau	2.21%	1/20/2020	Singapore
Malaysia	2.17%	1/20/2020	Vietnam
Singapore	1.93%	1/20/2020	Australia
Laos	1.82%	1/20/2020	France
Vietnam	1.79%	1/20/2020	Nepal
Burma	1.76%	1/20/2020	Malaysia
Cambodia	1.64%	1/20/2020	Canada

Accuracy of model appears stable: 70% in two runs(using only routes as feature). Also, these predictions correctly detected South Korea as the country with first case at current date.

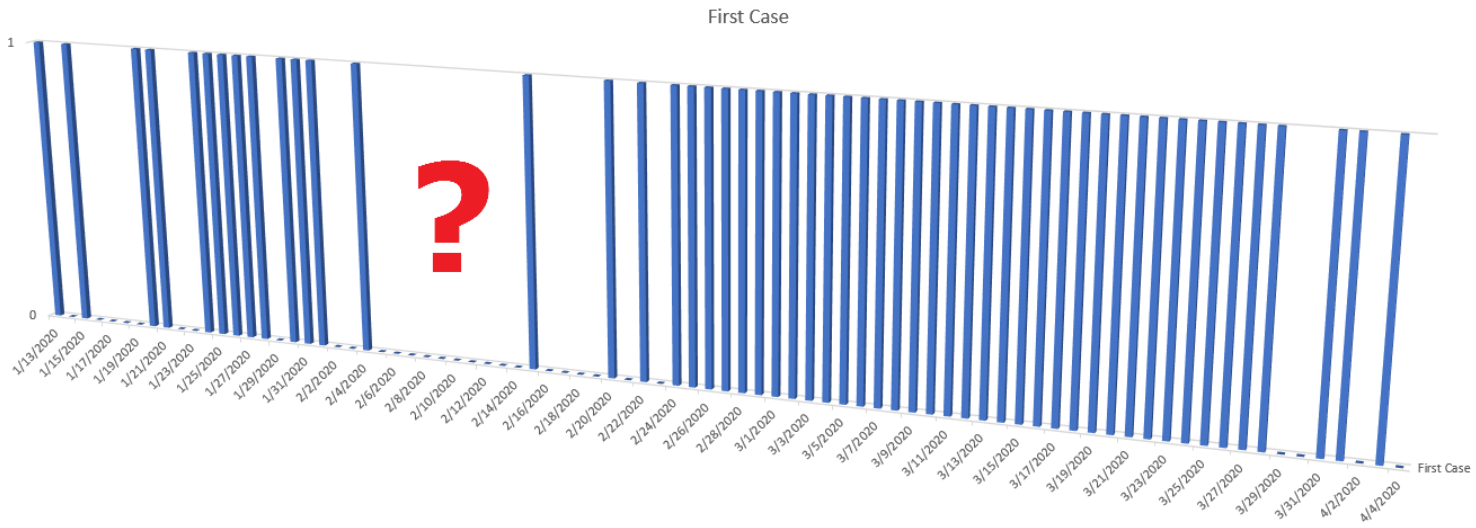
Next graph represents the accuracies of 3 models based on different combination of features for the period of two months 1/16/2020–3/16/2020:





Almost any of 6 features included in the model (with routes as a bases one) makes the model accuracy worse. Only “Water” with “Cases” somewhat helps on few days, but this could be random chance effect, as the improvement is too small.

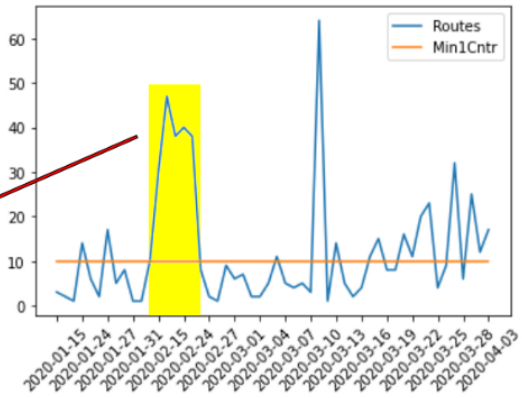
If you noticed the gap in dates in middle of February—this is not the error: “something” happened after the 4th February, so that almost two weeks no new country reported their first case:



Also, the decline in accuracy of predictions (single “Routes” graph above) looks very natural: with the virus spreading across the countries not only air traffic became the “means of delivery” of COVID-19 to new countries.

Looking more closely to this period gives insight that virus found another way to come to these countries:

Date	Min	Country
1/15/2020	3	Japan
1/20/2020	2	South Korea
1/21/2020	1	Taiwan, United States
1/24/2020	14	Singapore, Vietnam
1/25/2020	6	Australia, France, Malaysia, Nepal
1/26/2020	2	Canada
1/27/2020	17	United Arab Emirates
1/28/2020	5	Germany, Sri Lanka, Cambodia
1/30/2020	8	Finland, India, Philippines
1/31/2020	1	Italy, United Kingdom
2/1/2020	1	Russia, Spain, Sweden
2/4/2020	10	Belgium
2/15/2020	30	Egypt
2/20/2020	47	Iran
2/22/2020	38	Israel, Lebanon
2/24/2020	40	Bahrain, Kuwait
2/25/2020	38	Afghanistan, Iraq, Oman
2/26/2020	8	Algeria, Austria, Brazil, Croatia, Switzerland
2/27/2020	2	Denmark, Georgia, Greece, Macedonia, Norway, Pakistan, Romania



There is a big spike in the minimum position in the sorted list of probabilities for countries with first case at current date: for example, Egypt has value 30, meaning that model considered Egypt to be unlikely to report its first case on that date (giving higher probabilities for another 29 countries). And this is normal as air traffic covers about 57% of international passenger transportation. So, we still have undetected features that can improve the model. For example:
1. The number of tourists between countries without specifying the type of transportation they used.
2. The share of urban and rural population in the country.
3. Leave your suggestion in the comments.

Also, checking this approach on previous significant epidemics data could confirm the model viability, robustness and applicability for stopping of the spread of the virus in future.