Predicting early CoVID-19 infections

Data Detectives: zdong6,qabrams,vsenthil

Goal

During the early stages of a pandemic, when infections remain localized to just a few countries, policymakers in as yet uninfected countries could benefit from knowing how soon their country may see its first cases. Such a prediction could help policymakers decide when to make international travel restrictions such as flight bans or mandatory quarantine policies. With that goal in mind, we decided to look at whether international flight route and traffic data, basic information about counties, and CoVID-19 case data could predict how long it would take before countries saw their first infection.

Data

We gathered three types of data: Covid-19 case data, population data, and flight traffic data. The Covid-19 data is a daily time series of confirmed cases, by country (sometimes state/province), from John Hopkins University. For information about each country, we pulled population and density data from the US Census Bureau. Flight data was challenging to find. Only paid APIs provide data on traffic volume along each flight route worldwide, so the best we could come up with for free was the top 100 flight routes for 2018 from Routesonline in collaboration with Sabre Market Intelligence. We also pulled a list of 67, 663 flight routes from airport to airport that were active as of 2014 from OpenFlights. We joined the flight route and traffic data with data about airports to determine the originating and destination country of each flight route.

To do our analysis, we needed to join many of our tables on country names. We manually cleaned up misalignments due to countries with multiple acceptable names (e.g. Czechia vs Czech Republic) and due to different spellings conventions (e.g. including “the”, post-fixing “North”). In other cases, there simply wasn’t data, territories had been marked as countries (e.g. Puerto Rico), or contested regions were attributed differently (e.g. Gaza Strip). We ended up with 173 of 195 countries for which had all data.

Model+Evaluation Setup

We wanted to forecast days to first infection, so we looked at two types of independent variables: ones that hold constant across the epidemic and ones that vary as a result of the epidemic. We defined days to first infection to be a variable that counts down until reaching the day with first infection where it reached 0, and then continues to count down into the negatives.

We examined several different relationships through the use of ordinary least squares regression, looking at correlations in both simple and multivariate regression.

For our independent variables, we used population, population density, the number of incoming flight routes. These independent variables hold constant across the epidemic since they describe

attributes of countries. In the case of flight routes, our data is static and from past years, and does not account for changes in flight volume over the course of this pandemic. Our dependent variable with the number of days until the first case was found from January 22nd, the beginning of our time series data. For example, Spain's first case was recorded on February 1st, so it has a value of 10.

By combining the case data and the flight route data, we also ran a simple regression on the number of incoming flight routes originating in countries with confirmed cases each day. On January 22nd, the first day of our data, the United States, for example, had 189 incoming flight routes originating in countries with confirmed cases. Two weeks later it had 1174 and as of May 5th the last day of our data, it had 2304.

We held out a fifth of our data for each regression as a train-test split to prevent overfitting. We also used repeated random sub-sampling over 50 iterations to reduce random variation in R-squared and MSE.

Results and Analysis

Claim #1: The number of incoming routes to a country is mildly correlated with days to first infection.

Support for Claim #1:

The number of incoming routes had an R-squared of 0.338, which is not good, but shows some correlation. The training and testing MSEs for incoming routes were similar and its testing MSE was lower than the baseline, showing that this variable wasn't overfitting. All other variables showed no correlation. Our baseline was a variable that guessed that a country would not be infected until the final day of the data timeframe, which was 104 days.

(All numbers are averaged between ~50 simulations)

Variables	R-squared	Train MSE	Test MSE
Baseline	0.00	375	370
Population	0.084	340	391
Population density	0.035	371	391
Number of incoming routes	0.338	246	265

Claim #2: Running a multiple regression with all our variables fails to improve on the performance of just the number of incoming routes.

Support for Claim #2:

Running a multivariate regression on our data, we find an R-squared value of roughly 0.36, with testing MSEs of around 240 (depending on the split). This is slightly better than the incoming flights alone, but not by much.

Features	R-squared	Train MSE	Test MSE
Population, population density, and number of incoming flights	0.353	230	246
Just number of incoming flights	0.338	246	265

Claim #3: Culling routes not originating in countries with confirmed cases, failed to improve on the number of incoming routes as a metric.

Support for Claim #3:

By only counting routes from countries with confirmed cases, we hoped we could squeeze a bit more predictive power out of the incoming routes metric. We found, however, that it performed moderately worse with a lower R-squared, which was also reflected in a higher MSE.

Variable	R-squared	Train MSE	Test MSE
Number of incoming routes from countries with confirmed cases	0.305	262	280
Number of incoming routes	0.334	247	259