Predicting Early CoVID-19 Infections

Evan Dong(zdong6), Quinn Abrams(qabrams), Varun Senthil Nathan (vsenthil)

Introduction

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, broken down by country:

Covid-19 Case data

- daily time series from Johns Hopkins University of recorded cases, starting from 01/22/2020 and ending on 05/05/2020.

Population data

- population count and density for each country, from the U.S. Census Bureau

Flight Traffic data

- Top 100 flight routes in 2018 from Routesonline in collaboration with Sabre Market Intelligence
- 67,663 flight routes from airport to airport that were active as of 2014 from OpenFlights

Data processing

- Aggregating subdivided data
  - case data is inconsistently subdivided into smaller geographic regions (e.g. states or territories) which we aggregated to the national level
- Country name cleaning and joining
  - country names are not always well-defined, so we cleaned and standardized names for countries when we joined our tables on country names. We manually cleaned misalignments in cases of:
    - multiple acceptable names (e.g. Czechia vs Czech Republic),
    - different spellings conventions (e.g. including “the”, post-fixing “North”),
    - territories marked as countries (e.g. Puerto Rico), or
    - contested regions (e.g. Gaza Strip).

All told, we ended up with 173 of 195 countries with fully workable data.

Although there are visible outliers in for each of the variables we examine later, these outliers are features of the underlying that shouldn’t be excluded. For example, with regard to population, India and China have both have quadruple the population of the next most populous country, the United States. This similarly true with the number of incoming routes.

Methodology

Regressions:

For each regression, we held out a fifth of our data 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.

Dependent Variable:

We wanted to forecast days to first infection, so we used this as the dependent variable to run a series of linear regressions. Specifically, we defined days to first infection to be the number days until the first confirmed case starting from January 22nd, which is the first day of our data.

Methodology (contd.)

Independent Variables:

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.

For our static independent variables, we used population, population density, and 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 unfortunately static and from past years, and does not account for changes in flight volume or flight disruptions such as travel bans over the course of this pandemic. We ran a simple regression with each of these variables and then a multivariate regression with all of them.

We also tried to improve on the number of flight routes variable by only counting flight routes originating in countries with confirmed cases. This metric varies day-by-day since as the virus spreads and more flight routes start to originate in countries with confirmed cases. For example We also took only the first day’s (January 22nd’s) totals for each country from this variable and used this to see if we could improve on our regressions done with the constant variables described above.

Results and Visualizations

Overall, we found that only the number of incoming routes was at all predictive of days to first infection. All other variables showed no correlation. The baseline model is a variable that guesses that a country will not be infected until the final day for which we have data, which works out to be 104 days.

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
Number of incoming routes from infected countries	0.305	262	280

Population Density and Population Versus Days to First Infection

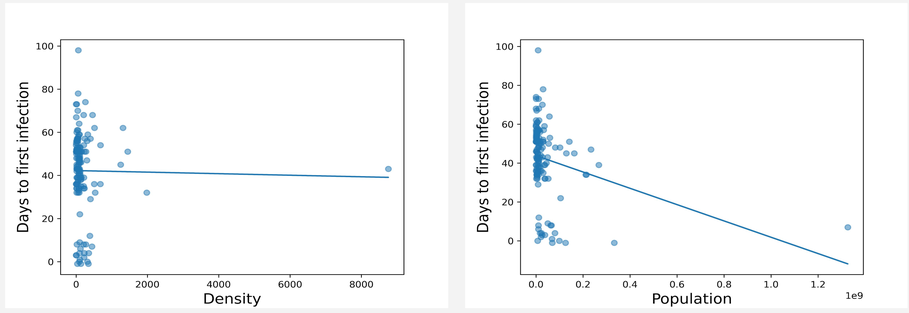

Fig 1. and **Fig. 2:** Neither population density (left) nor population (right) predicts spread of covid-19 to new countries

Number of Incoming Flights Versus Days to First Infection

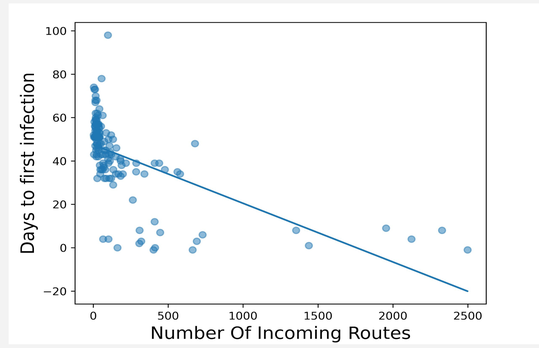

Fig 3. Number of incoming flights is the single most predictive factor, with an R-squared of 0.338

Number of Incoming Flights from Infected Countries Versus Days to First Infection

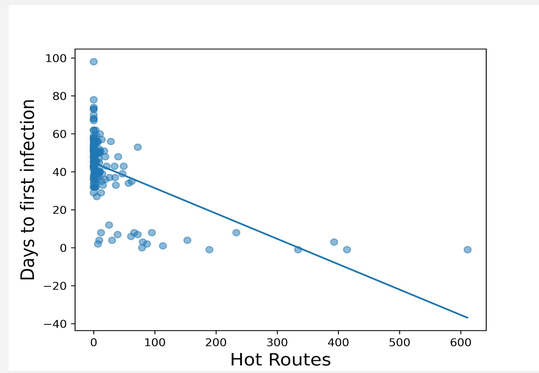

Fig 4: Number of incoming routes from infected countries with an R-squared of 0.305, which we hoped would be a refinement on the number of incoming routes, in fact, performs moderately worse.

Significance and Limitations

Ultimately, our analysis and results were limited by the data we had, and thus lack strong predictive power.

Due to cost limitations, we were forced to use historical and static flight data, as opposed to current, day-by-day information on actual flights during the Covid-19 pandemic. While the number of incoming flights proved most predictive and significant among the variables we studied, when tried to we failed to improve on this with a more nuanced metric, so it’s possible more granular data may not have helped.

Moreover, access to day-by-day information in flights may have allowed us to better utilize our time series case data, potentially weighting the significance of flight routes by the number of cases in origin countries, or other, more complex models.