IMPACT OF DOMESTIC AIR TRAVEL ON EARLY COVID-19 SPREAD IN THE UNITED STATES

SARA NICHOLAS, LINDSEY BLANKS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 21, 2021

ABSTRACT

CONTENTS

1 INTRODUCTION

COVID-19 has had an immense impact, infecting millions of people around the globe. The United States currently has the highest number of confirmed covid-related cases and deaths worldwide, but the spread of this virus has been very diverse. The history of the coronavirus within the United States began in January of 2020 when a Washington resident who had recently traveled back from Wuhan, China had the first confirmed case. Following the discovery of the virus within the States, some cities experienced rapid spread and a large number of confirmed cases very early on, while other cities did not see significant case tallies until months later.

As concern over the new coronavirus began to build, the federal government responded with a series of bans on international travel. These travel bans were highly controversial in the US, with some people believing they were completely unnecessary while others argued they did not happen soon enough. In March 2020, the situation in Europe, especially Italy, was rapidly deteriorating and a travel ban was officially placed on flights from Europe on March 11, 2020.

The purpose of this paper is to identify possible causes for the heterogeneity in the spread of the pandemic. Specifically, in our study we attempt to address the questions: How much of the variation in early spread of the coronavirus can be explained by airline travelers from Italy? How much can be explained by airline travelers along domestic routes? We construct a series of models that examine the relationship between air travel and number of COVID-19 cases within US states and then run log-linear regressions to examine the results.

2 PREVIOUS WORK

Due to the recency of the pandemic, the dynamics of the spread of COVID-19 are still largely unknown. Limited testing capability during the early months of 2020 means that studying the effect of travel during this time is extremely difficult and currently there is little research published about the effect of travel on the spread of the virus. Initial studies by Chinazzi et al. [1] indicate that travel restrictions placed on Wuhan, China only delayed the eventual worldwide spread by a few weeks at most. Additionally, they find that other international travel limitations had only a modest effect in reducing overall spread.

Similarly, Russell et al [2] estimated coronavirus prevalence in each country over time by comparing simple ratios of expected cases from international spread to expected cases arising from internal spread. They find that international travel restrictions had limited impact except in countries with extremely low incidence of COVID-19 and a large number of international arrivals.

Prince and Simon [5] limit their work to the impact of travel on the spread of covid-19 within the United States. For their study, they run a series of regressions to examine the effect of passengers arriving from all countries, with a focus on arrivals from China and Italy in the early months of 2020. Their results indicate that during the early stages of the pandemic, passengers from Italy were an important source of exposure, leading to significantly higher rates of infection and fatality. On the other hand, their analysis shows that places receiving more passengers from China do not experience higher case counts.

For our research, we choose to build on this body of work and examine the effect of international travelers from Italy on cases within the United States. We also expand the analysis by constructing models that include domestic air travel within the united States to examine the possible second and third order effects.

3 MODEL

3.1 Data

For COVID data, we used the New York Times COVID-19 database [4], which is a comprehensive dataset that has been used as the basis for many reputable studies on COVID. From this dataset, we specifically used state level case counts and state level testing counts.

We tested our models against COVID data for two dates: March 17, 2020 and March 31, 2020. A travel ban was implemented on March 11, restricting travel from Europe to the United States. Given the incubation period of the virus and delays in testing, we determined that case counts on March 17 were likely to still be affected by the travel that occurred immediately prior to the travel ban. Due to the shortage of tests early on, later dates have more robust data on case counts compared to the very sparse data in early March. Thus we determined that March 17 was late enough to benefit from increased testing and early enough to demonstrate effects from international travel. We also chose to test our models against case count data on March 31. At this point travel from Europe had almost entirely stopped, but infection that entered the U.S. from flights before the travel ban could still continue to spread domestically. We determined that dates much later than March 31 would not be as relevant for our models since by April, individuals began to limit their domestic travel, and mask mandates and social distancing measures became more widespread. Thus transmission of the virus on flights or in airports would not occur at the same rate as before these changes.

For flight data, we used a large dataset which compiled the 2015 full year total passengers for each flight route which started or ended in the U.S [3]. This was the most comprehensive open source air travel dataset we were able to locate, and it provided the level of detail required to construct both the domestic airport network and the incoming traffic from Europe. We focused specifically on international travel from Italy (and its bordering countries) since Italy had a substantial early outbreak before COVID became widespread in the United States. We did not focus on China because a travel ban was implemented on January 31, 2020, which leads to negligible effects caused by travel from China [5]. Additionally, we note that since the flight data is from 2015, it does not document endogenous changes in travel patterns caused by the pandemic. This is another reason we chose dates fairly early in the pandemic, as we determined that these effects would not yet be as strong.

3.2 Baseline Model

As a baseline model, we test the correlation between the number of COVID cases in a state and the number of passengers arriving in that state via flights from Italy. Let c_i denote the number of total confirmed COVID cases in state i on a given date, and let x_i denote the normalized number of arrivals to state i from Italy. Also let t_i denote the normalized number of tests conducted in state i. We then test the regression model

$$\log(\mathbf{c}) = \alpha\mathbf{x} + \eta\mathbf{t} + k$$

for $\mathbf{c} \in \{\mathbf{c}^1, \mathbf{c}^2\}$, where $\mathbf{c}^1$ is March 17, 2020, and $\mathbf{c}^2$ is March 31, 2020. Let y_i denote the normalized number of passengers arriving in a state via flights from Italy and its bordering countries (France, Switzerland, Austria, Slovenia). We also test the regression model

$$\log(\mathbf{c}) = \gamma\mathbf{y} + \eta\mathbf{t} + k$$

3.3 Neighbors of Hubs Model

In our baseline model, we consider how arriving passengers from Italy into an airport affect the case count in that state. However, it is also feasible that some of those arriving passengers are simply connecting through that airport to their final destination in another state. Additionally, in March 2020 before social distancing measures and mask mandates were widespread, arriving infected passengers could infect other passengers who they come into contact with inside the airport, with those secondary infected passengers then continuing on to fly to other states. Further, flight crew members (pilots, flight attendants, etc) also tend to fly multiple routes in a row, and thus infected crew members could carry the virus along their next route.

Thus we want to construct a model that considers these secondary effects. The model is explained mathematically below, but we offer an intuitive high level (simplified) explanation first. If A is the set of airports that receive a high number of passengers from Italy, we expect these states to have high case counts early on (this is the baseline model). If B is the set of airports that receive a high number of passengers from airports in A, we also expect travel from Italy to have some positive effect on case counts in states represented in B. For example, New York City receives a high number of passengers from Italy, and Boston receives a high number of passengers from New York City. Thus we expect high case counts in New York, as well as elevated case counts in Massachusetts due to these secondary effects. We call this the "Neighbors of Hubs Model" (note that "neighbors" denotes neighbors on the air travel network as opposed to states that physically border one another). The exact coefficients n for this model are computed using the process below on the air travel network.

We construct the graph representing domestic air travel in the United States. Nodes represent airports, and an edge between nodes i and j denotes a direct flight connecting the two airports, with w_{ij} equal to the number of passengers that fly that route annually (by convention, $w_{ij} = 0$ if i, j are not connected by an edge).

As before, let x_i denote the normalized number of arrivals from Italy. We then compute the "neighbor rank" of a node j by multiplying the arrivals from Italy to node i to the weight of edge (i, j), and summing over all neighbors:

$$m_i = \sum_j w_{ji} x_j$$

To cluster by state, let n_i denote the neighbor rank for a given state, using $n_i = \sum_{j \in i} m_j$ where j denotes airports and i denotes the state. Normalize the n_i according to $n_i = \frac{n_i}{\max\{n\}}$. As before, also let t_i be the normalized number of tests, and c_i be the weekly case count. We then test the regression model

$$\log(\mathbf{c}) = \alpha\mathbf{x} + \eta\mathbf{t} + \sigma\mathbf{n} + k$$

for $\mathbf{c} \in \{\mathbf{c}^1, \mathbf{c}^2\}$, where $\mathbf{c}^1$ is March 17, 2020, and $\mathbf{c}^2$ is March 31, 2020.

3.4 Page Rank Model

In this model, we consider the page rank centrality measure on the network of domestic air travel. The reason for doing this is that empirically we see that cities with early outbreaks correspond to airports we expect to have high centrality (e.g. New York City). If we then view eigenvector centrality as measuring how the virus diffuses from these hotspots through the domestic airport network (due to connecting flights, transmission within airports, infection of flight crew, etc), we expect that the centrality of each node may correlate with its case count.

We again construct the graph representing domestic air travel in the United States. Nodes represent airports, and an edge between nodes i and j denotes a direct flight connecting the two airports, with w_{ij} equal to the number of passengers that fly that route annually (by convention, $w_{ij} = 0$ if i, j are not connected by an edge).

We compute the page rank of nodes on this network. We use the parameter β to inject some "free centrality" at each node, allowing that injected centrality to diffuse through the network. The details of our page rank algorithm are as follows.

Using r_i to denote the page rank of node i, we define page rank using the recursive formula

$$r_i = \sum_{j \neq i} \frac{w_{ji}}{\delta_j} r_j + \beta_i$$

where $\delta_j = \max\left\{\sum_{k \neq j} w_{jk}, \sum_{k \neq j} w_{kj}\right\}$ is used to normalize by the degree of node j. In the network of domestic airports, the difference between a node's (weighted) in-degree and out-degree represents passengers entering and exiting the network at that node. We normalize by the greater of a node's in-degree and out-degree in order to more accurately represent the number of travelers passing through an airport.

We compute two sets of ranks. In the first calculation, we set $\beta_i = 1$ for all i, such that β allows for some infected passengers to enter the network at each node but does not systematically favor some nodes over others. In the second calculation, β_i is proportional to the number of passengers arriving at airport i from Italy, such that β effectively injects infected passengers into the network via incoming flights from Italy.

For each set of ranks we calculated, we use a node's (normalized) page rank as well as its (normalized) number of arrivals from Italy as parameters for our regression model. We cluster our page ranks by state in order to test it against our case data, such that r_i represents the page rank of state i, which is simply the sum of the ranks of the airports located in state i.

To cluster by state, let p_i denote the page rank for a given state, using $p_i = \sum_{j \in i} p_j$ where j denotes airports and i denotes the state. Normalize the p_i according to $p_i = \frac{p_i}{\max\{p\}}$. Also let x_i denote the normalized number of arrivals from Italy, t_i be the normalized number of tests, and c_i be the weekly case count (as before). We then test the regression model

$$\log(\mathbf{c}) = \alpha\mathbf{x} + \eta\mathbf{t} + \mu\mathbf{p} + k$$

for $\mathbf{c} \in \{\mathbf{c}^1, \mathbf{c}^2\}$, $\mathbf{p} \in \{\mathbf{p}_S, \mathbf{p}_I\}$, where $\mathbf{p}_S$ represents the standardized ranks with $\beta = 1$ and $\mathbf{p}_I$ represents the ranks with $\beta_i = x_i$ (proportional to the number arriving passengers from Italy, where i denotes airports).

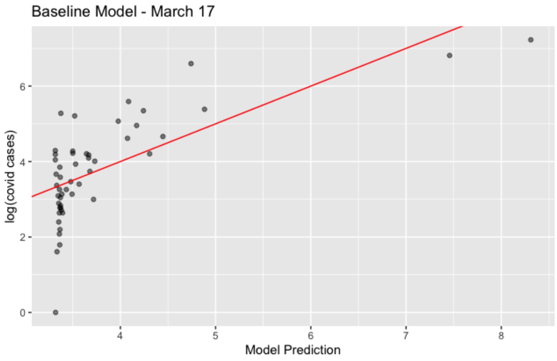
4 ANALYSIS OF RESULTS

For all our models, we test the log-linear regression model as outlined above. We use adjusted r squared as a measure of the models' fit, where higher r squared values indicate a better fit.

4.1 Baseline Model

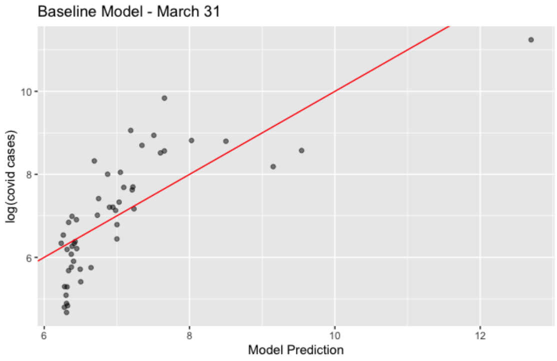
Passengers from Italy Only:

March 17



r^2 : 0.458

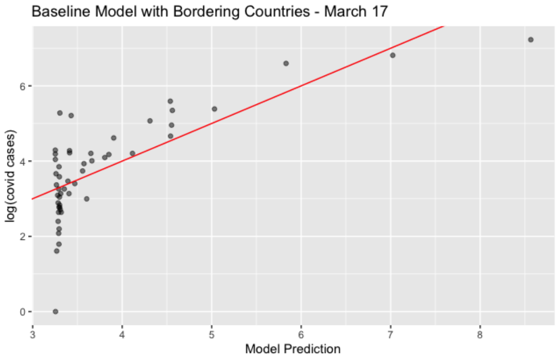
March 31



r^2 : 0.5609

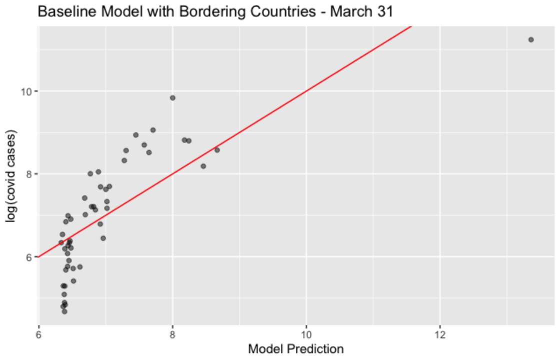
Passengers from Italy And Bordering Countries (Europe):

March 17



r^2 : 0.5341

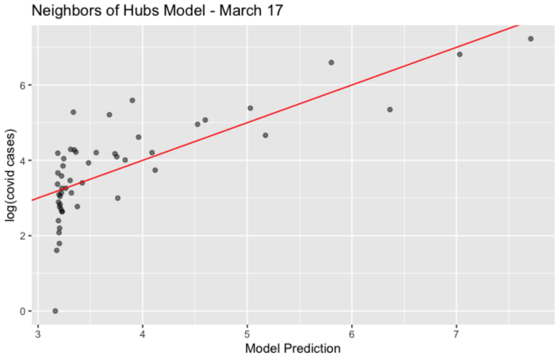
March 31



r^2 : 0.5676

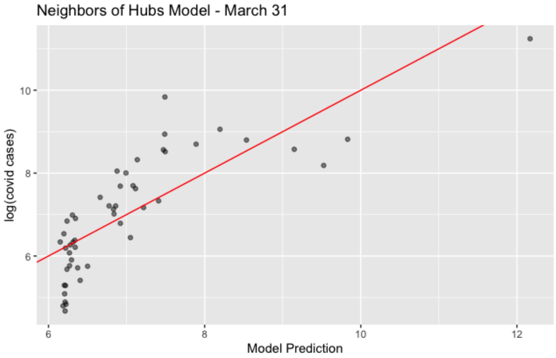
4.2 Neighbors of Hubs Model

March 17



r^2 : 0.5447

March 31

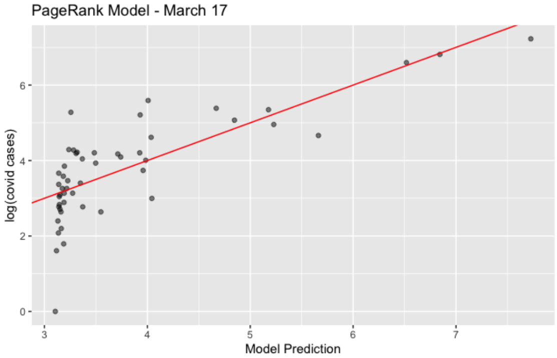


r^2 : 0.6142

4.3 Page Rank Model

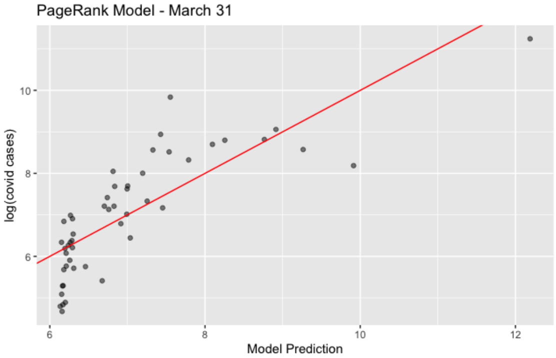
Standardized Beta:

March 17



r^2 : 0.5647

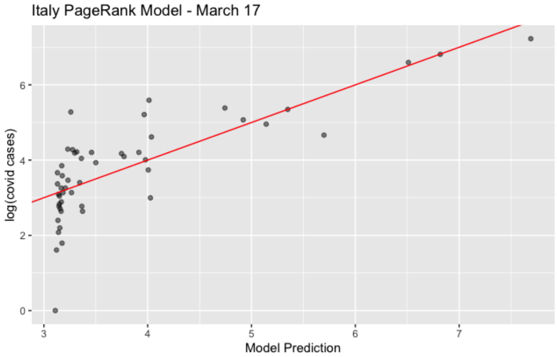
March 31



r^2 : 0.6528

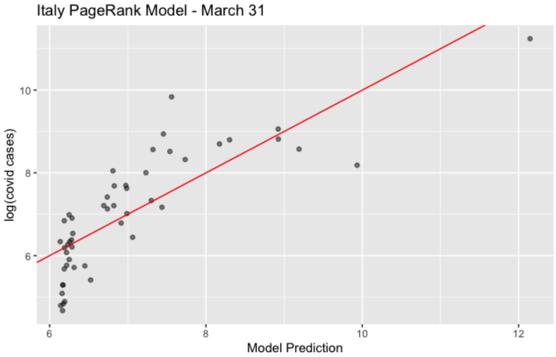
Proportional Beta:

March 17



r^2 : 0.5705

March 31



r^2 : 0.6574

4.4 Model Comparison & Analysis

Parameters:

- c_i: COVID-19 cases in state i
- x_i: incoming passengers from Italy to state i (normalized)
- y_i: incoming passengers from Italy and bordering countries (France, Switzerland, Austria, Slovenia) to state i (normalized)
- t_i: tests conducted in state i (normalized)
- n_i: neighbors rank (normalized)
- p_i: page rank (normalized)

	Baseline $\log(\mathbf{c}) = \alpha\mathbf{x} + \eta\mathbf{t} + k$ $\log(\mathbf{c}) = \gamma\mathbf{y} + \eta\mathbf{t} + k$		**Neighbors of Hubs** $\log(\mathbf{c}) = \alpha\mathbf{x} + \eta\mathbf{t} + \sigma\mathbf{n} + k$	**Page Rank** $\log(\mathbf{c}) = \alpha\mathbf{x} + \eta\mathbf{t} + \mu\mathbf{p} + k$	
	Italy	**Europe**	**Neighbors of Hubs**	**Standardized**	**Proportional**
3/17	$\alpha = 3.5713$ $\eta = 4.1384$ $r^2 : 0.458$	$\gamma = 4.0757$ $\eta = 3.6015$ $r^2 : 0.5341$	$\alpha = 2.8513$ $\eta = 3.6729$ $\sigma = 2.4398$ $r^2 : 0.5447$	$\alpha = 2.7583$ $\eta = 3.1372$ $\mu = 2.3307$ $r^2 : 0.5647$	$\alpha = 1.9353$ $\eta = 3.1730$ $\mu = 2.3671$ $r^2 : 0.5705$
3/31	$\alpha = -1.3053$ $\eta = 7.8395$ $r^2 : 0.5609$	$\gamma = 1.7172$ $\eta = 5.3582$ $r^2 : 0.5676$	$\alpha = -1.0050$ $\eta = 6.7023$ $\sigma = 2.1284$ $r^2 : 0.6142$	$\alpha = -0.8006$ $\eta = 6.2570$ $\mu = 0.6528$ $r^2 : 0.6302$	$\alpha = -1.4700$ $\eta = 6.1874$ $\mu = 2.0935$ $r^2 : 0.6351$

We observe that for both dates, each sophistication on the model does indeed improve the accuracy. Thus higher order network effects do correlate with early case counts. Because we normalized all our parameters into the range $(0, 1]$, we can compare the coefficients to determine importance of the various parameters.

On March 17, the coefficients corresponding to the domestic air travel network effects are of similar magnitude (ranging from slightly smaller to slightly larger) compared to the coefficients corresponding directly to travel from Europe. This is true across all the models that incorporate these network effects. This suggests that both international travel and domestic travel correlate strongly with early COVID-19 case counts in the U.S. Controlling for (very heterogeneous) availability of testing is also crucial to modeling early case counts, as demonstrated by the high values of η in all the models.

By March 31, we found travel from Italy to in fact have a negative correlation with state level case counts. This suggests some incompleteness in the data, as the data set we used does not account for changes to travel patterns in 2020 and thus does not accurately represent travel from Italy in late March 2020. Notably, however, the coefficients corresponding to domestic air travel are still fairly high, suggesting that while international travel played a much smaller role later on, domestic travel continued to play a role. This seems feasible as shifts in domestic travel patterns occurred much later than shifts in international travel patterns. Thus infection that had originally arrived in the U.S. via international flights could spread freely across the country via domestic flights, even after international travel had effectively ceased.

5 FURTHER STUDY

Refinements on this study would ideally seek datasets with more granular and recent data. International travel tends to have seasonal trends which are not represented in data aggregating annual passenger totals. Additionally, flight data from 2020 would eliminate uncertainty regarding the magnitude of endogenous shifts in travel behavior due to COVID.

Controlling for other parameters could also improve the accuracy of the models. For instance, population density and demographics could significantly impact infection rates and thus case counts. Similarly, heterogeneity within states creates some noise in the model (New York City likely had a much different infection rate than Albany or Rochester). If data were gathered detailing what cities or counties arriving travelers into an airport tend to live or stay in, then this data could be used to construct a model at a more granular level to address this heterogeneity.

Other functional forms for the regression model could also be tested. Similarly, we used a very simple formula $\beta_i = x_i$ to determine these inputs to the proportional Page Rank model. One could instead use gradient descent to determine the optimal formula for the β_i as a function of x_i and perhaps other parameters. This could be informative as to what factors lead to a state being an important source of centrality in the network.

6 CONCLUSION

While there are many possible sophistications on our models and opportunities for further study, we can conclude from our work that domestic air travel did in fact play a large role in COVID-19 spread in mid to late March 2020. Earlier implementation of social distancing measures and mask mandates on flights, as well as domestic travel restrictions, may have helped curb this early spread through the United States.

This result is informative as to how rapid spread of future infections can be prevented. Once an infection has entered a country from some origin, it can and will continue to spread throughout the country even if travel from the origin is halted. Thus if transmission from the origin occurs, there must be a twofold response that (1) halts travel from the origin and (2) takes steps to reduce domestic transmission.

7 APPENDIX

Data Processing & Implementation of Models:
https://github.com/saranicholas/covid-domestic-air-travel

REFERENCES

[1] M. Chinazzi et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 2020. doi: 10.1126.

[2] T. Russell et al. The effect of international travel restrictions on internal spread of covid-19. *medRxiv*, 2020.

[3] Gary Hoover. Us airline route segments 2015. https://data.world/garyhoov/us-airline-route-segments-2015, 2016.

[4] New York Times. Coronavirus (covid-19) data in the united states. https://github.com/nytimes/covid-19-data, 2020.

[5] Jeffrey Prince and Daniel H. Simon. The effect of international travel on the spread of covid-19 in the us. *SSRN*, 2020.