DNA 20

Cure 100%

PandAIRmic

An analysis of the impacts of air travel on the spread of COVID-19

Jian Cong Loh, Joshua Neronha, Stephen Sun, Tzuhwan Seet

Hypothesis

We hypothesize that the number of domestic air passenger arrivals in a US state is positively correlated with the number of new COVID-19 cases in that state.

Data

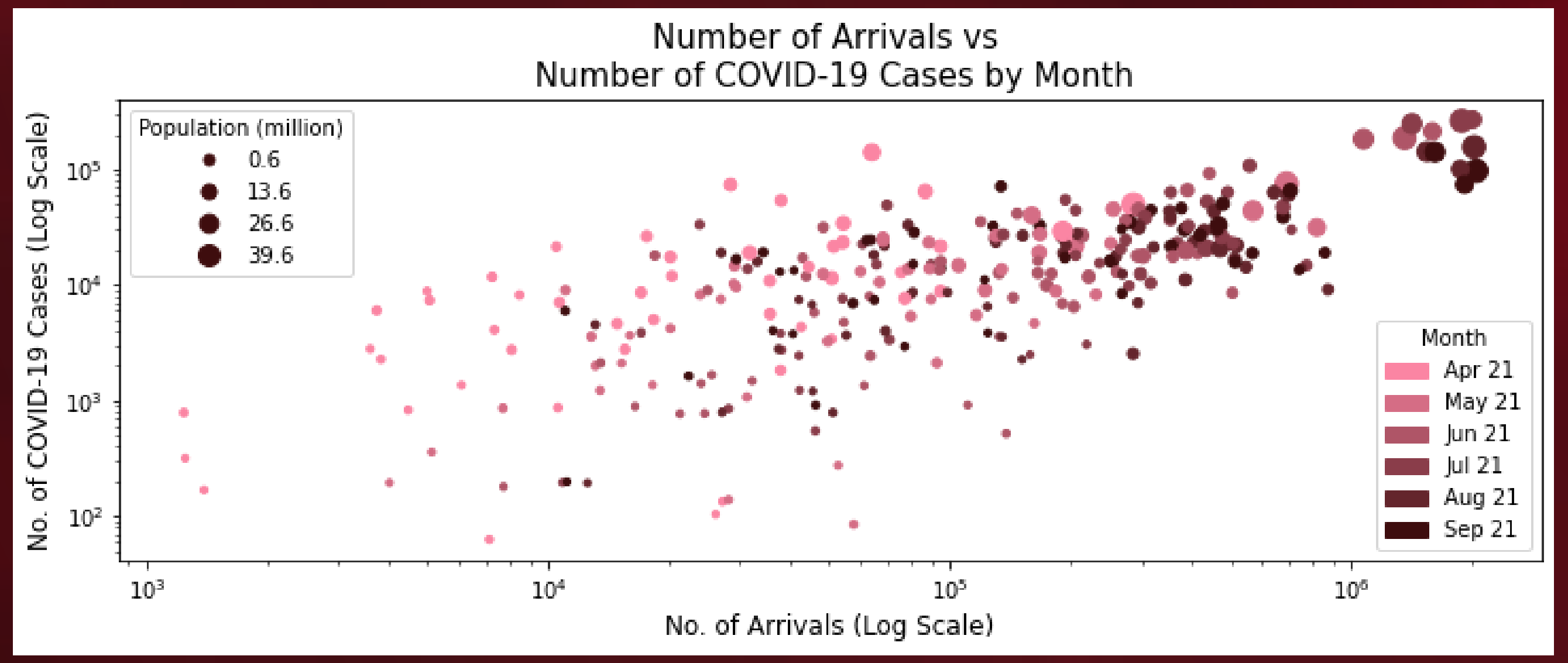
- Our **COVID-19 data** is obtained from *The Atlantic's* **The COVID Tracking Project**. It gives the daily number of positive cases for each of the 50 states, five territories and D.C., and our data spans Jan 20 and Feb 21.
- Our **air passenger arrival data** is compiled from two datasets from the USDOT. The **DB1B** gives a 10% sample of all itineraries by quarter, and covers Q1 to Q3 2020. The **T100** breaks down the number of passengers flying between two airports (including layovers) by month, and covers Jan to Nov 2020. We combined the two datasets to estimate the number of passengers travelling between every pair of states per month, excluding layovers.
- Our **policy data** comes from the **COVID-19 US State Policy Database** developed by the Boston University School of Public Health. It tracks the dates when each US state implemented COVID-19-related social safety net, economic, and social distancing policies.

Methodology

To match the monthly arrivals data, we aggregated cases from the 15th of each month to the 14th of the next, considering the 2-week COVID-19 incubation period. This gave us **294 data points** from **49 states** (excl. DE) and **6 months** (Apr – Sep 20).

Our exploratory data analysis (see below) gave us the following insights:

- States with large pop. had more arrivals and cases – pop. is a possible confounding variable that affects both
- Linear r/s in log-log plot – possible polynomial r/s
- Cases increased over time – possible time-fixed effects



We ran simple linear regression with different versions of our variables:

- **Model 1A** – No. of arrivals vs no. of cases
- **Model 1B** – Change in arrivals vs change in cases
- **Model 1C** – Per capita (p.c.) arrivals vs per capita cases
- **Model 1C(W)** – Weighted p.c. arrivals vs p.c. cases

Using Model 1C(W) as the base, we ran several more complex regression models:

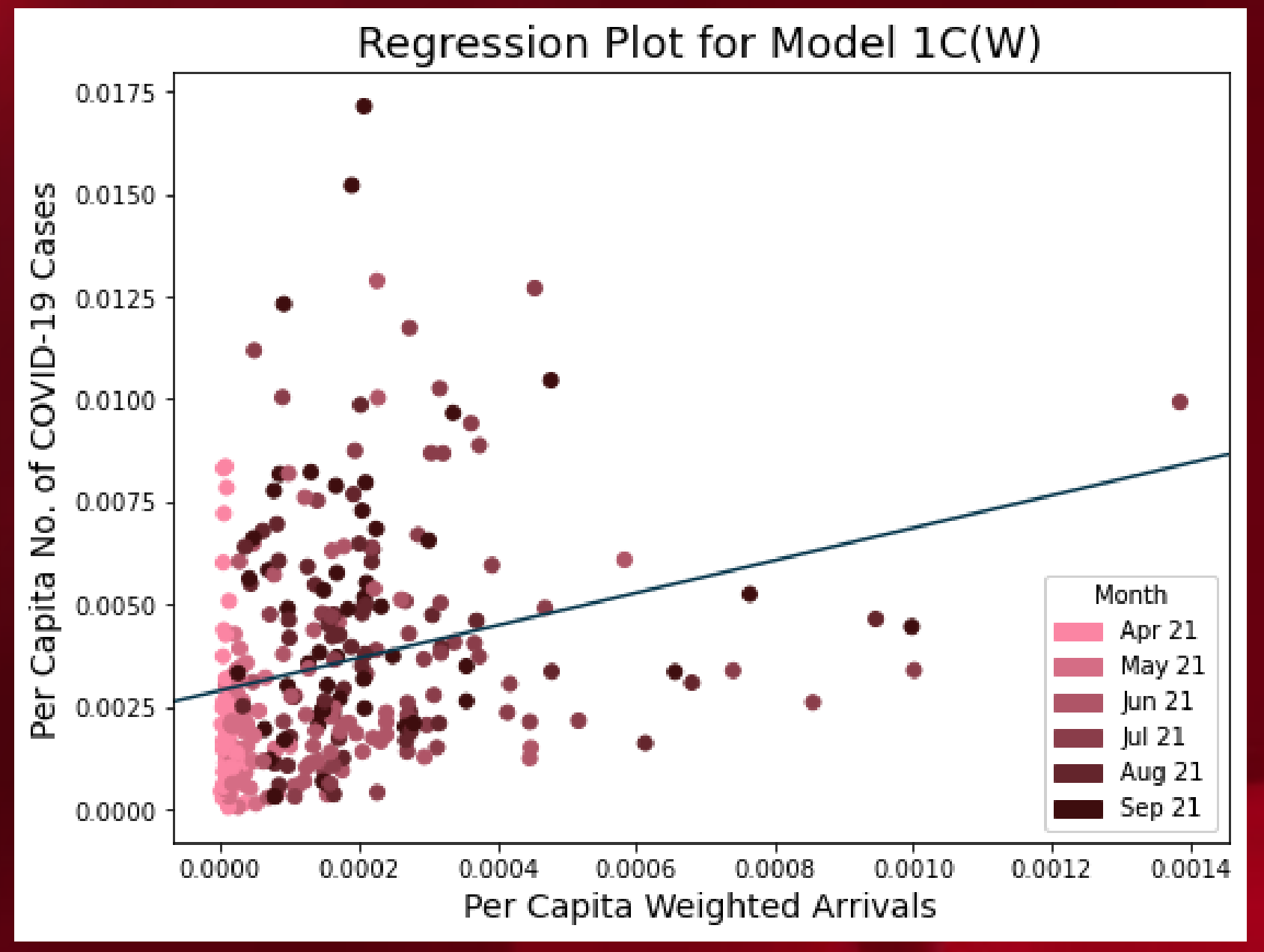
- **Model 2** – Polynomial regression
- **Model 3** – Multiple linear regression (w/ policy as controls)
- **Model 4** – Fixed effects model (entity & time)

Simple Linear Regression

Model	P-Value	Coef	R^2
1A	0.000	0.083	0.611
1B	0.208	-0.095	0.005
1C	0.000	0.016	0.047
1C(W)	0.000	3.96	0.065

All simple linear regression models except Model 1B showed a **significant positive correlation** between the independent and dependent variables at the 5% significance level.

While Model 1A had the highest R^2, this likely reflects a spurious correlation as population is a confounding factor that influences both variables.



To control for population, we experimented with the percentage change (Model 1B) and per capita values (Model 1C) of both variables.

A possible reason for Model 1B's poor performance is that taking the percentage change of the variables violates the i.i.d. assumption.

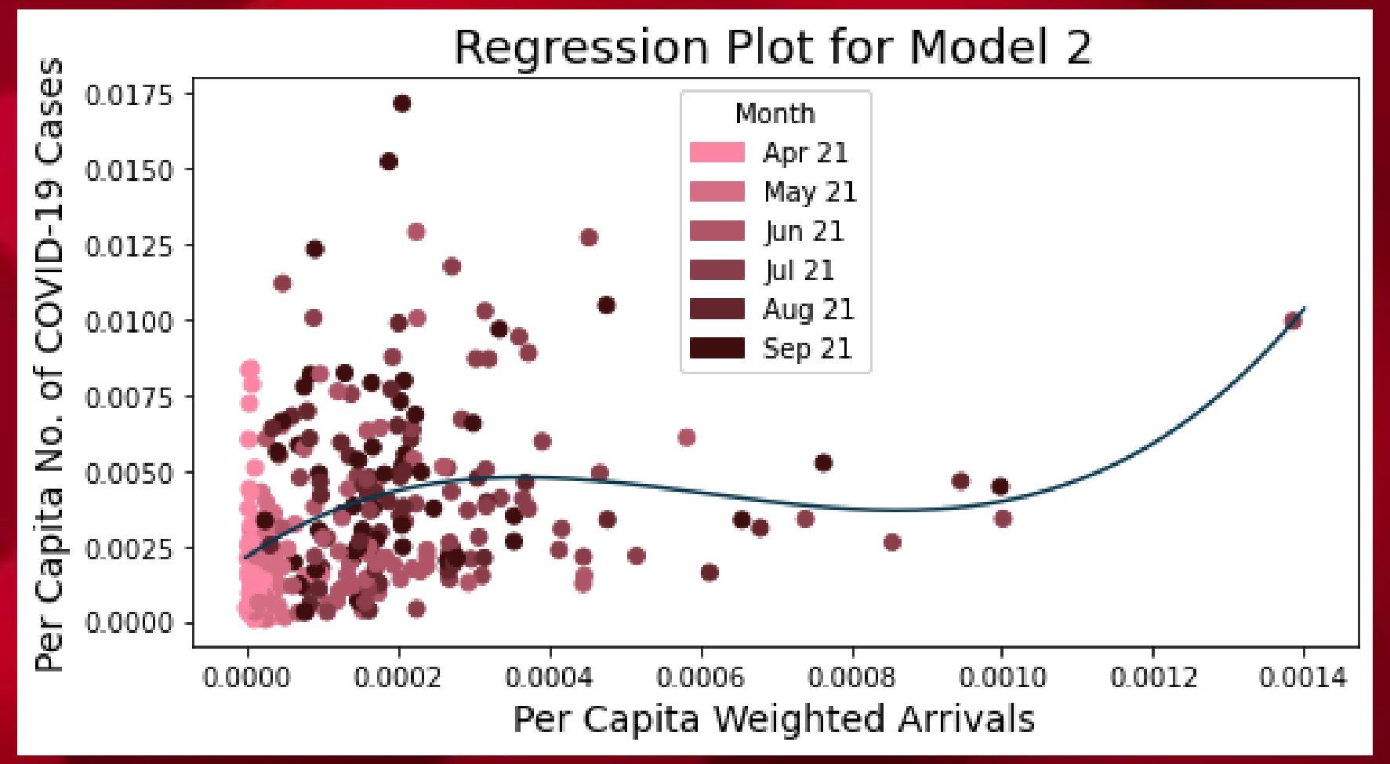
We extended Model 1C by weighing the number of arrivals by the origin states' per capita number of cases. **Model 1C(W)** gave the best R^2, and serves as the base model for our subsequent analyses.

Polynomial Regression

We used a **third degree polynomial** model, as fitting the examples to polynomial curves of increasing degrees resulted in an increase in R^2, but this improvement plateaus for degrees > 3.

All polynomial terms have **significant correlation** with p.c. number of cases at the 5% confidence level.

The **R^2 is 0.118**, which is higher than that for simple linear regression. The relationship between weighted p.c. arrivals and p.c. COVID-19 cases is therefore likely to be polynomial rather than linear.



Regressor	P-Value	Coef
X	0.002	16.9
X^2	0.000	-3.30×10^4
X^3	0.001	-1.80×10^7

Multiple Linear Regression

Regressor	P-Value	Coef
Weighted p.c. Arrivals	0.033	1.99
Stay at Home	0.032	-0.001
Business Closure	0.067	-0.001
Facemask Mandate	0.088	-0.001
Quarantine Mandate	0.000	-0.002

All policy variables are **negatively correlated** with weighted p.c. number of cases, with the relationship being significant at the 5% level for **Stay at Home** and **Quarantine Mandate**.

The **R^2 is 0.211**, which is higher than that for simple linear regression. Adding states' COVID-19-related policy to the model reduces omitted variable bias.

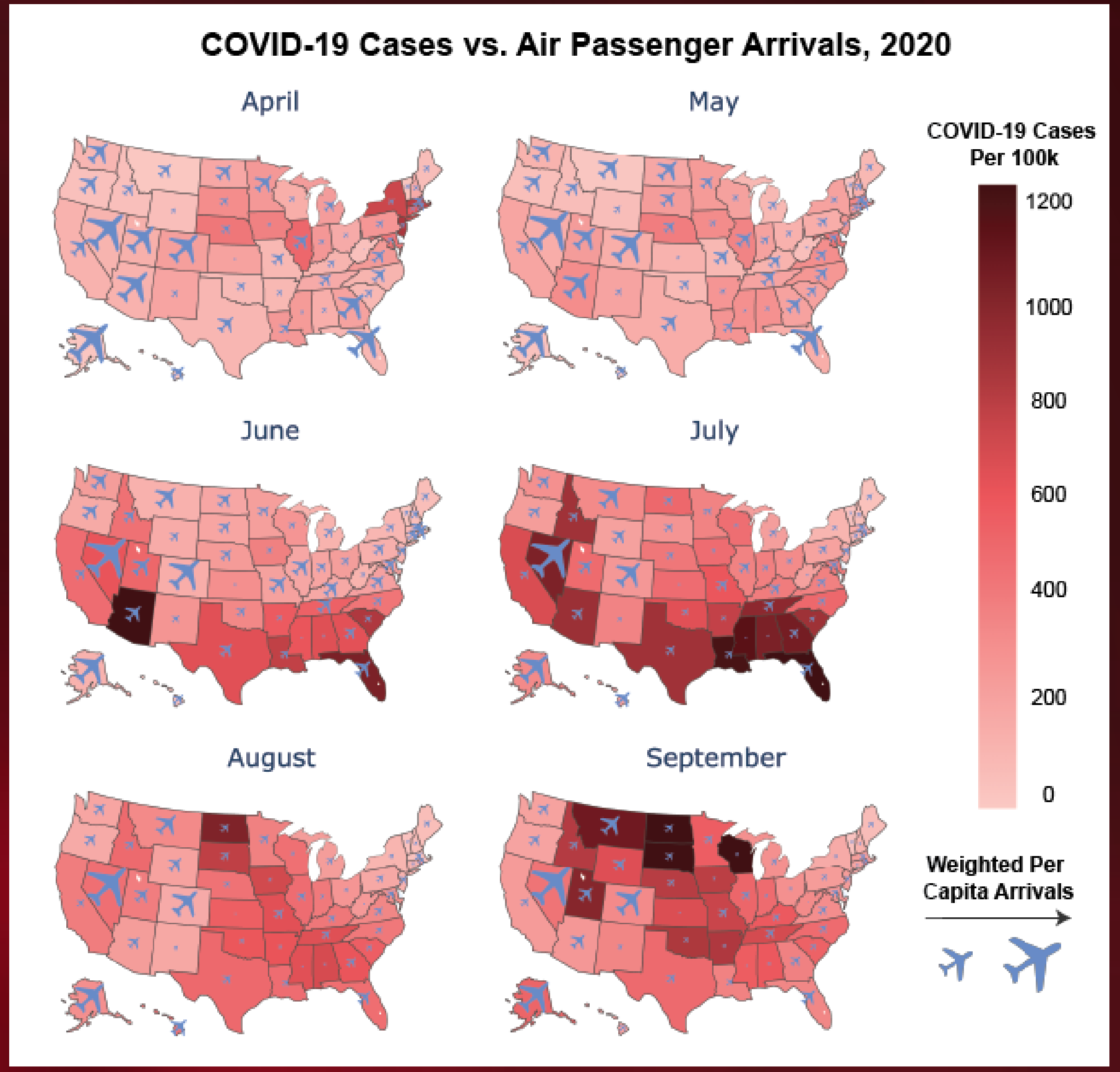
Fixed Effects Model

A fixed effect model controls omitted variables in panel data that vary across entities/states (e.g. public health systems, political attitudes) or time (e.g. federal COVID-19 policy).

Including **entity-fixed effects** to Model 3 gives an **R^2 of 0.274,** but weighted p.c. arrivals is no longer significantly correlated with p.c. cases (**p-value 0.469**). Adding **time-fixed effects** too does not change the variables' coefficients or p-values, but gives a lower **R^2 of 0.0343**.

The results indicate that there is omitted variable bias from unobserved variables that vary across states but not time, but not from those constant across states that change with time.

Map Visualization


COVID-19 Cases vs. Air Passenger Arrivals, 2020
April
May
June
July
August
September
COVID-19 Cases Per 100k
1200
1000
800
600
400
200
0
Weighted Per Capita Arrivals

Given the geographical nature of the data, a map provides a useful visualization that allows us to compare across different states at one glance and more intuitively understand the relationship between arrivals and COVID-19 cases. The map supports the results of our statistical tests by showing that there is some positive correlation between the independent and dependent variables, but it is limited.

Conclusions

- There is a statistically-significant positive correlation between air passenger arrivals and new COVID-19 cases, although the independent variable explains only a small proportion of the variance in the independent variable.
- The relationship between domestic air passenger arrivals and new COVID-19 cases is likely polynomial and not linear.
- Social distancing policies are negatively correlated with COVID-19 cases, but are not the only omitted variables that vary across states.
- More variables that are neither entity- nor time-fixed are needed as controls to understand the true relationship between air passenger arrivals and COVID-19 cases.

Limitations / Future Directions

- Due to its proprietary nature, data from the USDOT is available only by month or quarter, limiting the resolution of our analysis. More granular data will allow for a more precise analysis of the relationship.
- There is likely simultaneous causality where air passenger arrivals affect COVID-19 cases, and vice versa. Instrumental variable regression can allow us to eliminate the bias.
- As an extension, the relationship between international flight volumes and COVID-19 cases can be explored. Stricter international travel restrictions may mean that flights have a greater impact on the cross border spread of COVID-19.