Fourteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2021)

Quantifying the Impact of Air Travel on Growth of COVID-19 Pandemic in the United States

Lu Dai, Ivan Tereshchenko, Mark Hansen
Institute of Transportation Studies
University of California, Berkeley
Berkeley, CA, USA
dailu@berkeley.edu, terivan2006@berkeley.edu, mhansen@ce.berkeley.edu

***Abstract*— This paper develops models to quantify the dynamics of the impact of air travel on the spread of the COVID-19 pandemic, using a wide range of datasets covering the period from March to December 2020. With the help of flight operation data, we first develop a novel approach to estimate the county-level daily air passenger traffic, which combines passenger load factor estimates and information about the air traffic distribution. Cross-sectional models using aggregated county-level variables are estimated. While this study focuses on air travel variables, we also control for potential spatial autocorrelation and other relevant covariates, including vehicle miles traveled (VMT), road network connectivity, demographic characteristics, and climate. The model results indicate that air travel has a strong and positive impact on the initial pandemic growth rate for both case-based and fatality-based aggregate models.**

Keywords-pandemic spread; cross-sectional model; spatial autocorrelation; air passenger traffic; network connectivity

I. INTRODUCTION

The COVID-19 pandemic in the U.S. surged at the beginning of the summer of 2020, after a slow decline in the late spring. Changes in air travel may have foreshadowed this surge back in May. As shown in Figure 1. , the blue bars represent the 2020 Transportation Security Administration (TSA) screened passengers at all the U.S. airports, compared with passengers screened on the same weekday one year ago in grey. The orange curve shows the percentage decline from 2019 to 2020 in daily throughput. On March 1st, the TSA screened over 2.2 million air passengers at all U.S. airports. It was 99% of the total number of passengers screened on a comparable weekday in 2019. Then the TSA checkpoints throughput followed a mainly downward trajectory in March-April. During the second quarter of 2020, the air traffic volume decreases by up to 96%. However, starting from May, the air traffic volume began to creep up. As of July 21st, the air traffic rebounded to about 25% of the passenger throughput in 2019.

The movement of people is an essential factor in increasing the spread of disease, such as the SARS-CoV-2, where human-to-human contact is the primary transmission mechanism. There were few researchers investigating the role of air travel in affecting the spread of COVID-19. Early in February, Lau et al. [1] investigated the relationship between international/domestic air traffic and coronavirus outbreak in China by comparing the air passenger volume and flight routes to the distribution of COVID-19 cases. The results indicate a strong linear correlation between the cases and air passenger volume. However, they used the annual passenger throughput data from 2013 to 2018, which might be inadequate since the flight schedule could have dramatically changed in 2020, especially after many flights had been canceled during the pandemic. Zhang et al. [2] estimated a gravity model to investigate the influencing factors of pandemic spread in Wuhan, and found that the frequencies of air flights and high-speed train services are significantly associated with the number of COVID-19 cases in the destination cities. By correlating an air network mobility model to air passenger traffic data, Linka et al. [3] found that the air passenger travel data can be used to predict the emerging global diffusion pattern of a pandemic at the early stages of the outbreak. Focusing on international air travel, Zhang et al. [4] proposed the imported case risk index, which accounts for the foreign country's pandemic condition and the air connectivity with China, using the air ticket reservation data provided by a Chinese aviation service company. This is one of the few studies that utilized real flight data to measure the air travel's impact on the COVID-19 pandemic.

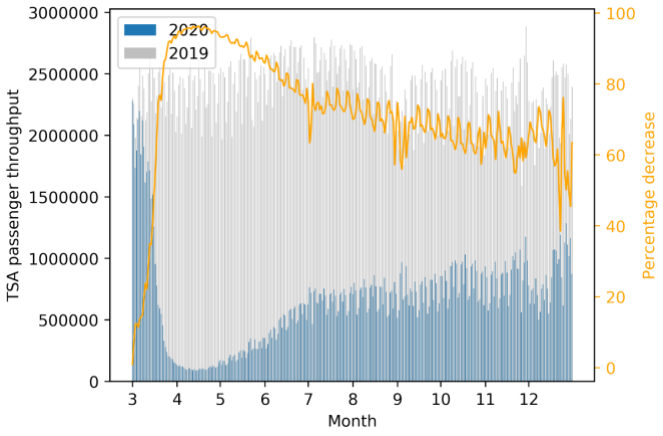

Figure 1. The daily TSA passenger throughput in 2019 and 2020

While these studies notice a dependence of pandemic spread on air traffic mobility dynamics, none has developed a reliable

estimate for measuring air passenger traffic and quantified the dynamics of the impact of such traffic on the pandemic spread. This paper aims to fill the gap by integrating data analytics and statistical modeling to provide insights about how and to what extent air travel may contribute to the COVID-19 pandemic spread in different periods. Our contribution is three-fold: first, a wide range of datasets are collected and fused together to derive factors that may potentially influence the pandemic spread in the United States. While we will focus on the air travel variables in this study, other relevant covariates related to ground transportation, demography, geography, and climate are also derived and accounted for in the model. Second, we develop a novel approach for estimating the county-level air passenger traffic, which combines the passenger load factor estimate and traffic distribution. This metric, which utilizes realized flight operation data, can serve as a proxy variable for air travel and support other coronavirus-related research endeavors. Third, we estimate cross-sectional models at the county level for different periods to investigate how air travel affects the pandemic spread over time. This is the first work to utilize realized flight operation data in quantifying the air travel's impact on the pandemic spread over time (from March to December 2020) while controlling for other potential covariates and spatial autocorrelation.

The rest of this paper is organized as follows. In Section II, we introduce the data sources. Section III provides details about how we derive the air passenger traffic variable from realized flight operation data. Section IV introduces the model form and other covariates we control for. The estimation results and discussion are presented in Section V. Finally, conclusions and future work are presented in Section VI.

II. Data Sources

We limit the scope of the study to the contiguous United States, which consists of the 48 adjoining U.S. states and the District of Columbia on the continent of North America. We combine a wide range of data sources covering the period from January 1st to December 31st in 2020.

The first set of data is the county-level SARS-CoV-2 morbidity and mortality records. It is publicly available on the USAFACTS website, based on reports from state and local health agencies. This data includes the daily confirmed coronavirus infection case counts nationwide, and the number of deaths in different geographies. The data begins with the first reported coronavirus case in Washington State on Jan 21st, 2020 and is regularly updated.

The second kind of data is related to air transportation: (1) the FAA aviation system performance metrics (ASPM) database provides operated individual flight information at all departure or arrival airports with ASPM facility. Canceled flights are not in the ASPM database. Fields of interest for each flight operation including origin airport, destination airport, departure time, arrival time, air carrier, aircraft type, user class (commercial, air taxi, freight, general aviation), and tail number. We remove cargo aircraft and flights that do not arrive in the continental United States. There are over 20,000 flights without aircraft types specified. We impute the missing aircraft type by searching the database for the flight operation with the same tail number, which is physically unique to an airplane. For those that are not matched, we check the historical flight activity log on FlightAware. (2) The Freedom of Information Act (FOIA) Electronic Reading Room gives the public access to various federal records, including the Transportation Security Administration (TSA) checkpoint traveler throughput for each airport in the U.S. The data records the total number of screening passengers, not including the airport crew members, passing through TSA security checkpoints at each departing airport by the hour. (3) The airport database comprises multiple sources – the Bureau of Transportation Statistics (BTS) Master Coordinate table, the Global Airport Database (GADB), and the OpenFlights Airports Database – for the integrity of airport information. The database provides a list of domestic and foreign airport codes and their associated airport name, country, latitude, longitude, and altitude information. (4) The Aviation Encyclopedia and Pilot Booklists information on various aircraft types, manufacturer, standard seating capacity, and model names.

The third dataset is extracts of selected geographic and cartographic information from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database (MTDB). This dataset provides us with geospatial information of road networks, county cartographic boundary, and county population centers in shapefile format. The roads line shapefile provides highways, major roads, and secondary roads for national, state, and regional display. The 2019 county cartographic boundaries polygon shapefile is a simplified representation of counties with a resolution level of 1:500,000. The mean centers of population for each county are coordinates (latitude, longitude) based on the 2010 Census.

The fourth kind of data is the daily county-by-county vehicle miles traveled (VMT). It measures the sum of the number of miles traveled by each vehicle in a county over a one-day period, which we think is the best available proxy for within-county traffic. The data is provided by StreetLight Data Inc, which collects and analyzes anonymous records from location-based service (LBS) installed on mobile devices and other GPS-enabled devices. The spatial precision of LBS data varies based on which technology was used to collect user locations. It would range from 5 meters – if the GPS was on – to 50 meters for records collected using WiFi proximity and Bluetooth proximity.

The fifth kind of data is hourly weather and climate data retrieved from Google Earth Engine. We download hourly records of convective fraction, potential evaporation, shortwave radiation, specific humidity, surface temperature, total precipitation for each county.

The sixth dataset contains the cross-sectional socio-demographic characteristics of counties that may affect the rate of epidemic spread. These data include population density and urban population from the 2010 Census, and demography, educational attainment from the Bureau of Economic Analysis county-level economic data.

III. Air Passenger Traffic

Due to the coronavirus outbreak worldwide, airlines in the U.S. had a passenger load factor declined to 10% from 81% in

January 2020 [5]. The aircraft flow is an unreliable estimate and may not accurately capture the impact of air travel on the pandemic spread. However, air passenger traffic flow data availability is sparse and incomplete due to the sudden emergence of COVID-19. This work is motivated by the recognition that it is critical to developing a novel approach that can be rapidly deployed for traffic analysis with currently available data, and will be easily updated as more data become available. This section shows how we derive the air passenger traffic variable – the daily number of air travelers entering each county. This variable serves as an explanatory variable to represent the impact of air travel when we statistically model the pandemic spread using cross-sectional models presented in the next section.

A. Passenger Load Factor

Given the realized flight activity provided by the ASPM dataset, it is crucial for us to estimate the passenger load factor to convert the aircraft flows to passenger flows. One naïve approach would be to treat the daily ratio of 2020 TSA throughput to the 2019 TSA throughput as the passenger load factor, shown as the grey line in Figure 2. However, this is problematic and runs the danger of underestimation. As the traffic demand went down, the airlines also started reducing the number of operating flights to fully utilize aircraft's seating capacity in order to reduce economic loss. Therefore, the average load factor should not be simply the traffic ratio, but should also capture reductions in flight schedules during the pandemic.

Toward this end, we present an approach that can approximate the passenger load factor, taking into account both traffic volume and the flight schedule. Assuming that flights departing from the same airport on a particular day would have the same passenger load factor, we could calculate the passenger load factor of all flights that depart from airport o on day t as PLF_{ot} using TSA checkpoints throughput and aircraft seating capacity information.

$$PLF_{ot} = \frac{S_{ot}}{\sum_{m}^{M_{ot}} C_m} \quad (1)$$

S_{ot} is the total number of screening passengers, not including crew members, passing through TSA security checkpoints at the airport o on day t, which is available from the FOIA dataset.

M_{ot} is the total number of flights departing from airport r on the day t, where $m = 1, 2, \dots, M_{ot}$. The count is based on the ASPM individual flight dataset, in which canceled flights, cargo flights, and flights that do not arrive in the contiguous U.S. are removed. It varies by airport and by day.

C_m is the standard seating capacity for each flight m according to its aircraft type.

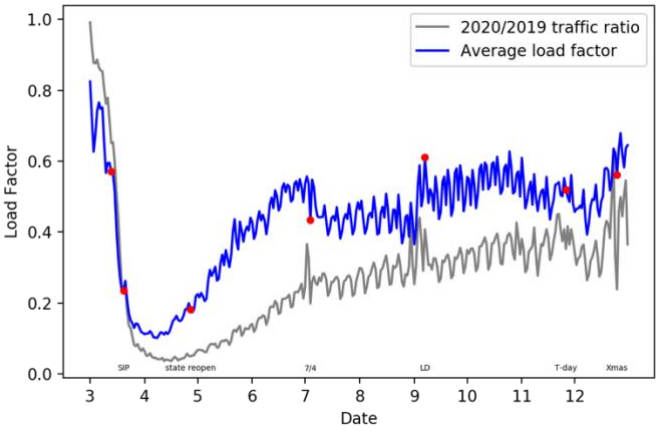

Figure 2. The passenger load factor results

The blue curve in Figure 2. represents the daily passenger load factor calculated using Equation (1), averaging over all the airports in the U.S. Our estimated passenger load factor is, in general, higher than the "load factor" approximated by the traffic ratio in grey. This reflects the fact that although traffic is reduced a lot during the pandemic, the scheduled flights are also reduced, and thus the passenger load factor would not be as low as the traffic ratio we observe. The red dots labeled on the curve are the dates with special events, such as the first Shelter In Place (SIP) restriction approved date, the first state-reopening date, Fourth of July, Labor Day, Thanksgiving Day, and Christmas. They appear to be at the peak or at the turning point, which in general matches our expectations. In early March, the traffic fell most steeply after the Center for Disease Control (CDC) issued the domestic Level 2 travel advisory. Since then, people have begun cutting down travel significantly. Our estimated passenger load factor decreases about the same magnitude, probably because airlines have not greatly grounded planes yet. Starting from March 19th, state-wide stay-at-home orders became the norm across the U.S., and that is the period when airlines canceled most of the flights. Since late April, states were gradually easing COVID-19 restrictions, and the summer travel season was kicking off. Our estimated load factor reflects this fact and has crept up steadily after the date when the first state-reopening order is announced in Colorado and Montana.

B. Airport-Level Passenger Traffic

We construct the air transportation network based on the realized flight operations collected from ASPM individual flight dataset. Our analysis is built upon the dynamic air transportation network, where the nodes are the operating airports and the links between them inferred from daily flight activity. From the previous section, we obtain the estimated passenger load factor for all flights that depart from a given airport on a given day. The estimated origin-specific daily passenger load factor can thus be applied to the realized air transportation network to approximate the volume of passengers on each link. For a given day t, we are able to calculate the total number of passengers traveling from airport o to the airport d as N_{odt}, by adding up the number of passengers on each aircraft:

$$N_{odt} = \sum_{m}^{M_{ot}} PLF_{ot} \cdot C_m \cdot \mathbb{I}(m: o \longrightarrow d) \quad (2)$$

where PLF_{ot} is the estimated passenger load factor applied to all flights departing from airport o on day t; C_m is the standard seating capacity for each flight m according to its aircraft type; $\mathbb{I}(m: o \longrightarrow d)$ is the indicator variable which equals one if the flight m flies from airport o to airport d, zero otherwise; the M_{ot} is the total number of flights departing from airport o on day t, of which the destination could be any airports in the U.S.

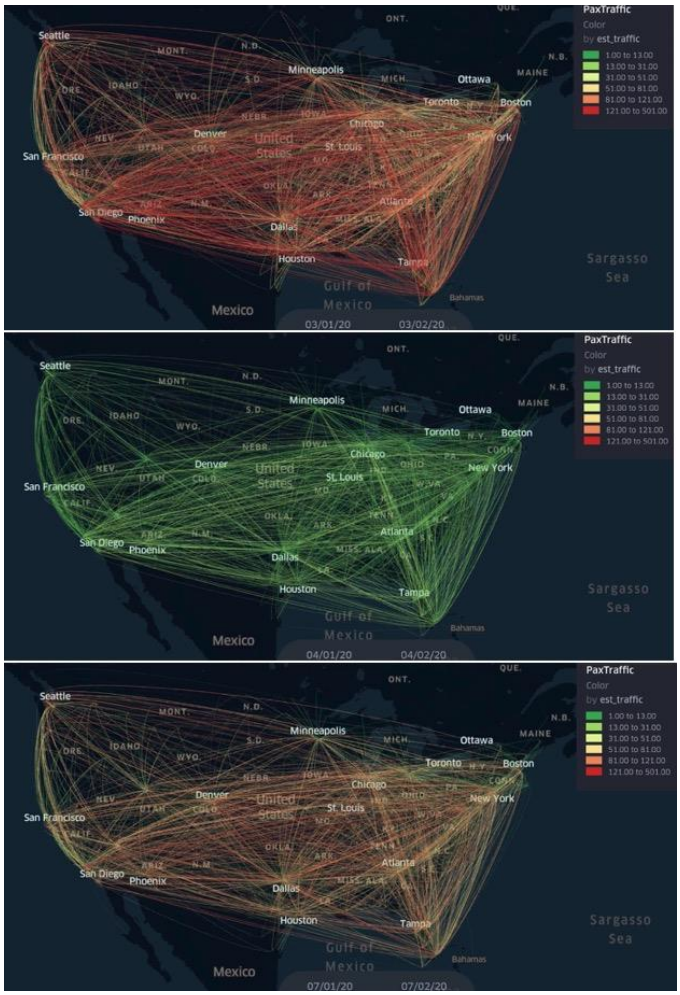

Figure 3. The air passenger flow between airports on March 1st, April 1st, and July 1st

In Figure 3. , we visualize the estimated air passenger flow between airport pairs on March 1st, April 1st, and July 1st from top to bottom, with warmer links indicating higher air passenger traffic volume. As expected, the passenger traffic was significantly dropped during late March and early April, due to the COVID-19 shocks of lockdowns and travel restrictions. There was a moderate rebound during the summer travel period, but it was not returned to the pre-pandemic levels.

After we obtain the air passenger traffic on each airport OD pair, we compute the total number of air passengers arriving at airport d on day t by summing up the traffic flows from all other airports in the U.S.:

$$N_{dt} = \sum_{o \neq d} N_{odt} \quad (3)$$

Note that we only consider direct routes due to the unavailability of data to estimate the connecting traffic. We propose the hypothesis that connecting traffic should have a smaller effect on the estimation of passenger traffic than usual, since travelers may be more inclined to fly direct routes over connecting routes during the pandemic.

C. County-Level Passenger Traffic

To model pandemic spread at the county scale, we need to transform airport-level air passenger traffic to be county-specific. We wish to answer the following question: for all the air passengers arriving at airport d on day t, N_{dt}, how many of them or what percentage of them would go to county i? This is a traffic distribution problem that solves how air passengers at the airport will be allocated into a particular county.

Inspired by the literature on airport passenger leakage studies, we decide to consider counties within 300 miles of each airport as the targeted counties to which the air passenger flow will be allocated. Studies have found that the airport market leakage is typically in the range of 200 miles [6], up to 230 miles away [7], 250-260 miles [8], and 300 miles [9]. Given the changes in the aviation industry, we choose the most conservative and most recent estimation of 300 miles. Thus, all the air passengers arriving at a given airport would only be allocated to the neighboring counties within 300 miles.

Next, we need to decide how to allocate airport arrival passengers to all these neighboring counties. According to the data availability, we assume the traffic distribution is based on county population and distance to the airport: (a) The shorter the distance between the airport and counties, the more convenient traveling through this airport, thereby attract more arriving traffic to this county. (b) Counties with larger populations should increase the attractiveness of arriving passengers. Based on these assumptions, we define the traffic distribution probability in the form of an exponential function and calculate the number of air passengers going from airport d to county i on day t as Q_{kjt}:

$$Q_{idt} = N_{dt} \cdot \frac{P_i D_{id}^{-\alpha}}{\sum_i^I P_i D_{id}^{-\alpha}} \quad (4)$$

where P_i is the population of the county i, which is assumed to be constant over the analysis period; D_{id} is the Euclidean distance between airport d and the county i, which is also constant over the analysis period; N_{dt} is the total number of air passengers arriving at airport d on day t; α is the unknown parameter that needs to be estimated. Note that I is the total number of counties considered in the traffic distribution, which means there are I counties within 300 miles of the airport d.

We employ parametric analysis to solve α based on passenger survey reports available for the following nine airports in the U.S.: SFO, OAK, SJC, JFK, LGA, EWR, LAX, SEA, DAL. The survey was conducted by airports to obtain information about air travelers and the determinants of the travel

market. It usually included questions on journey purpose, origins/destinations, means of transport to and from airports, route flown, residence and income, etc. The survey question, or a similar question, we are particularly interested in is "from which county did you leave today?". Though the survey results are based on where people leave the county, we assume this is applicable in the opposite way. We fit Equation (4) with 1,457 samples collected from the survey responses. The optimal value of α is estimated to be 1.96 by minimizing the root mean squared error of Q_{idt}.

Lastly, our explanatory variable of the cross-sectional model – the number of air passengers traveling to county i on day t – is computed as:

$$Q_{it} = \sum_{d} Q_{idt} \tag{5}$$

This could be grasped by thinking of an example in the San Francisco Bay Area. We first distribute the arriving air passenger traffic at SFO and OAK to Alameda county according to the county population and its distance to the airport. Then we aggregate the allocated traffic from these two airports. The sum of traffic is the number of air passengers traveling to Alameda county on a given day.

IV. Model Form and Variables

In this section, we present a county-level cross-sectional model to provide insights about how and to what extent the different structural determinants may contribute to the COVID-19 pandemic spread.

A. Model Specification

In this paper, we focus on an Ordinary Least Squares (OLS) regression-based cross-sectional model:

$$y_i = \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{X_i} + \varepsilon_i \tag{6}$$

where i is the county index and $i = 1, 2, \ldots,$ N; $\widehat{\boldsymbol{\beta}}$ is the vector of estimated regression coefficients that described the sensitivity of the dependent variable to the predictors; $\boldsymbol{y}, \boldsymbol{X}$ are scaled dependent and independent variables respectively, and ε_i is the regression residual. The variables of this model are discussed in the following sections.

Cross-sectional analysis using measures aggregated over counties at the same period in time provides us with insights into the effects of structural determinants that may cause some regions to have higher or lower rates of pandemic spread. In this study, we estimate an aggregated model for the whole analysis period from March 1st to December 31st, 2020, and three cross-sectional models for three sub-periods: 03/01/2020 – 06/30/2020, 07/01/2020 – 09/30/2020, and 10/01/2020 – 12/31/2020. Note that all these cross-sectional models are estimated at the county level but aggregated over different periods. The three sub-models are analyzed to investigate whether any of the predictors have time-variant effects, and how air travel affects the pandemic spread over time.

B. Logarithm Growth Rate

The dependent variable is expected to depict the epidemic situation in each county. Thus, we compute the intrinsic growth rate of the cumulative number of confirmed cases/fatalities for each county i within a given interval of time T.

$$y_i = \ln(N_{it}) - \ln(N_{i0}) \tag{7}$$

where N_{it} is the number of confirmed cases/fatalities on the last day of the analysis period; N_{i0} is the number of confirmed cases/fatalities on the first day of the analysis period, or the last day of the previous period.

We consider both the intrinsic case growth rate and the intrinsic fatality growth rate, and run the models independently for these two response variables. The intrinsic fatality growth rate is expected to reflect the epidemic situations more accurately since the infected case data in the early phase of the coronavirus outbreak may be unreliable.

C. Ground Transportation

To control the effects of ground transportation, we derive two variables that characterize the road network connectivity and the human movements. Employing complex network theory, we explore how the location of a county with respect to the road network may influence disease spread. In Figure 4. , we visualize the logarithm case growth rate on the county choropleth map from March 1st to April 15th, with warmer colors indicating faster growth. By overlaying the national highway network on the map, we found that the logarithm case growth rate (dependent variable) is much higher in the county that is in geographical conjunctive points of freeways. This suggests that ground transportation network connectivity may be a significant causal factor in disease growth.

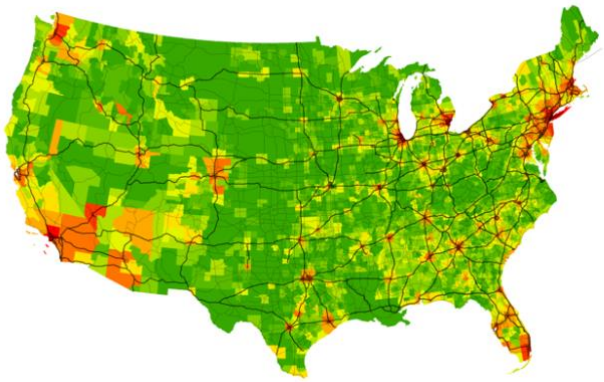

Figure 4. The county choropleth map colored by logarithm case growth rate from March 1st to April 15th, with the road network shown in black

To quantify the road network connectivity, we calculate the closeness centrality for each county. The closeness centrality measures how far a given county is from all other counties based on the shortest path through the road network. The closeness centrality metric reflects the position of the county in the road network. It indicates how "efficient" the flow of information (viruses) would be through a given county to other counties in the network. In addition, we consider the effect of population

size of the reachable counties by introducing the population multiplier. The population-weighted closeness centrality for county i is denoted as PC_i:

$$PC_i = \sum_{j \neq i}^{J_i} \frac{P_j}{D_{ij}} \quad (8)$$

where J_i represents all the counties that are reachable by county i via the road network; P_j is the population of the county j, which is assumed to be constant over the analysis period; D_{ij} is the shortest-path distance between county i and county j via the constructed road network, which is also considered to be unchanged during the analysis period.

We choose the national major roads system to construct the network in ArcGIS. As shown in Figure 5. , the nodes are the population centers for each county based on the 2010 Census. If there is a direct road connection between two counties, a link is drawn and is weighted by the distance. With this weighted undirected graph, the population-weighted closeness centrality for each county in the U.S. can be calculated using Equation (8) and colored on the nodes in Figure 5. Counties in the New York metropolitan area have a higher population-weighted closeness centrality score, which means, in general, they have the shortest distances to all other populous counties. In the model, we use the normalized population-weighted closeness centrality to improve the numerical stability.

Figure 5. Population-weighted closeness centrality (in logarithm scale)

For intracounty trips, we aggregate the daily VMT for each county and take the average over the analysis period. The county-level average daily VMT per capita is used as the explanatory variable in the cross-sectional models.

D. Climate

Studies come to conflicting results about the associations between climatic factors and the pandemic spread. Most researchers found that the transmission of COVID-19 occurs in regions exposed to cool and dry conditions, and near the lower end of the radiation gradient [10]. Most studies also reported a positive association between temperature and the COVID-19 spread due to human movements, whereas some found a negative relationship since a cold environment may provide a more favorable condition for the virus survival [11]. To investigate the climatic effects on the pandemic spread, we incorporate the average values of daily average shortwave radiation, daily cumulative precipitation, and the daily average temperature in the cross-sectional model.

E. Socio-demographic Characteristics

For each county, we include the population density, the percentage of the urban population, the number of less than high school graduates per capita, and dummy variables for different ethnic groups in the cross-sectional model. We also add dummy variables for each state to control the fixed effects at the state level, such as testing ability and lockdown policy.

TABLE I. SUMMARY STATISTICS FOR REGRESSION VARIABLES.

County-level Variable	**Description**	**Mean (standard deviation)**			
		Mar. – Dec.	***Mar. – Jun.***	***Jul. – Sep.***	***Oct. – Dec.***
Case growth rate	Logarithm case growth rate over the analysis period	7.47 (1.51)	4.48 (2.11)	1.71 (0.89)	1.38 (0.62)
Mortality growth rate	Logarithm mortality growth rate over the analysis period	3.27 (1.59)	1.30 (1.71)	0.91 (0.90)	1.15 (0.86)
Air passengers	Average number of air passengers traveling to county i, Q_i	250.26 (1643.57)	192.14 (1239.23)	273.30 (1769.04)	333.82 (2160.19)
VMT per capita	Average daily VMT per capita for county i	47.04 (15.60)	43.52 (11.19)	47.97 (11.31)	45.27 (11.89)
Closeness centrality	Normalized population-weighted closeness centrality for county i, PC_i	0.26 (0.09)			
Urban population	Percentage of urban population in county i	42.34 (30.69)			
Population density	Population density of county i	262.41 (1850.13)			
LTHSG	The number of less than high school graduate per capita	0.09 (0.04)			
Shortwave radiation	Average of daily average shortwave radiation	208.04 (17.33)	237.98 (19.23)	256.26 (14.75)	120.28 (24.23)
Precipitation	Average of daily cumulative precipitation	2.76 (1.29)	3.13 (1.44)	2.96 (1.58)	2.25 (1.41)
Temperature	Average of daily average temperature ℃	15.93 (4.31)	15.41 (4.76)	23.87 (3.21)	9.09 (5.08)

After data preprocessing and filtering, there are 3,056 observations (counties) remaining to estimate the model. The summary statistics of the variables are presented in TABLE I. The average case growth rate decreases over time, while the average mortality growth rate is relatively stable. Air travel, on average, is slowly coming back. The increasing standard deviation also suggests that some popular counties have attracted more air passenger traffic. The VMT seems to spike in the summer. All the climatic variables match our expectations.

All the explanatory variables are log-transformed except for the temperature and state dummies.

V. Results and Discussion

A. Estimation Results

In this section, we present the estimation results for the model described in the previous section. We estimate two types of models – one type, where the dependent variable is the growth of the number of cases, and another type, where the dependent variable is the growth rate of the number of deaths. For each model type, we estimate four models. First, we use the average growth rate from March to December as the dependent variable ("aggregate model"), second – growth rate between March and June, third – growth rate between July and September, fourth – growth rate between October and December. All predictors are the same in all eight estimated models.

We estimate both case-based and fatality-based models because the two available variables have their advantages and disadvantages. On the one hand, the case-specific data is available for the majority of counties, while for many smaller counties, the number of deaths is very small, which results in an apparently small growth rate of fatalities. On the other hand, the case-specific data depends on the amount of COVID testing. Early in the pandemic, the testing rates were low, which is reflected as a very low growth rate in the number of cases. The fatality growth rate more accurately reflects the spread of the epidemic for many counties. The purpose of estimating separate models for every period is to explore whether any of the predictors have varying effects without specifying a time series model.

All coefficient estimates are for scaled predictors - centered around the mean and divided by the standard deviation of each variable. This was done in order to compare the relative importance of each variable. We do not report the coefficient estimates for the dummy variables used in the regressions as they are not the primary focus of the paper. The dummy fixed effects were included in order to reduce the potential omitted variable bias.

Let us start by examining the R^2 values. For both case-based and fatality-based aggregate models R^2 is very high – 90.2% and 80.1% of the variation is explained by the predictors. However, when estimated for individual periods, the R^2 decreases. For example, for the fatality-based October-December model, the R^2 is only 0.356. The most likely explanation for this model behavior is that there is more variation in the growth rate on shorter time spans. In such a case, we would expect the aggregate model to be better than any of the shorter-time span models, which is exactly the result we obtained. Summary statistics in TABLE I. also support this explanation.

There is also a possible explanation for why the goodness-of-fit of case-based models is higher than for the fatality-based models. In addition to capturing the pandemic dynamic, the case growth rate reflects the scale of testing efforts undertaken in individual counties. The amount of testing is influenced by the economic prosperity and cultural characteristics of locations. Our models capture this economic and cultural influence, which leads to better models.

TABLE II. Regression Estimates for 4 Cross-Sectional Models (Dependent Variable – Average Growth Rate of COVID Cases)

Period	Mar-Dec	Mar–Jun	Jul–Sep	Oct–Dec
Air Passengers	0.311*** (0.011)	0.310*** (0.021)	-0.015 (0.015)	0.006 (0.009)
VMT Per Capita	-0.433*** (0.054)	0.678*** (0.113)	-0.646*** (0.087)	-0.135*** (0.047)
Closeness	0.038*** (0.007)	0.034** (0.014)	-0.011 (0.011)	0.014* (0.006)
Urban Population %	0.141*** (0.007)	0.138*** (0.014)	0.005 (0.010)	-0.002 (0.006)
Population Density	0.390*** (0.017)	0.578*** (0.032)	-0.111*** (0.023)	-0.068*** (0.014)
LTHSG	0.278*** (0.029)	0.419*** (0.057)	-0.061 (0.043)	-0.068** (0.024)
Shortwave Radiation	-0.819* (0.415)	0.103 (0.640)	0.291 (0.553)	-0.891*** (0.159)
Precipitation	-0.114*** (0.034)	0.105 (0.063)	-0.007 (0.030)	-0.120*** (0.016)
Temperature	0.029*** (0.007)	0.003 (0.011)	0.035*** (0.010)	-0.015*** (0.005)
Adjusted R^2	0.902	0.828	0.435	0.626

In interpreting the regression results, let us first focus on the main variable of interest – the number of air passengers traveling to a given county. Note that all predictor variables, except for the dummy variables, have been log-transformed. The dependent variable is also log-transformed. This means that we can interpret the regression coefficients as elasticities of case/fatality growth rates with respect to the predictor. For example, the air passenger coefficient for the aggregate case-based model is equal to 0.331. We can interpret it the following way: for every 10% increase in the number of incoming air passengers, the ratio of COVID cases at the end of the period to the beginning of the period N_{it}/N_{i0} increases by 3.11%.

For both case-based and fatality-based aggregate models, the influence of the number of air passengers on the pandemic spread rate is strong and positive. However, if we look at the period-specific models, we see a different picture. The regression coefficient is positive for the March-June models, suggesting a positive effect of the air traffic on the pandemic spread rate. After that (June-September and October-December models), the effect is weak and not statistically significant.

There are several possible explanations for this result. First, the spread of the pandemic is determined to a significant extent at its beginning. For example, the Chinese authorities were able to curb the pandemic by introducing strict lockdowns at the very beginning, not letting the virus spread around the country.

However, air transportation played a role in the early seeding of the pandemic in countries around the world. Aviation might have played a similar role in the US. Counties with more incoming air traffic were more likely to be exposed to the virus early on. At the early phase of the pandemic (March-June), the virus spread uncontrollably throughout the country, even as the air traffic collapsed. After the lockdowns started, the pandemic spread locally, determined mainly by local factors, such as population density. However, all else equal, counties with stronger air connections had a higher early virus "load", which determined the remaining course of the pandemic.

The second explanation for the weak effect of air transportation in the later phases of the COVID pandemic is the behavioral adjustment. As the pandemic progressed, more information, such as location-specific case numbers, became available. This may have deterred travel to places with high caseloads. Moreover, travelers with even mild symptoms were urged not to travel. As a result, symptomatic spreaders became less likely to travel to other regions and spread the infection.

TABLE III. REGRESSION ESTIMATES FOR 4 CROSS-SECTIONAL MODELS (DEPENDENT VARIABLE – AVERAGE GROWTH RATE OF COVID-19 MORTALITY)

Period	Mar-Dec	Mar–Jun	Jul–Sep	Oct–Dec
Air Passengers	0.327*** (0.016)	0.409*** (0.021)	-0.017 (0.017)	-0.056*** (0.016)
VMT Per Capita	-0.368*** (0.081)	0.249* (0.117)	-0.128 (0.096)	-0.320*** (0.087)
Closeness	0.069*** (0.011)	0.030* (0.014)	0.017 (0.012)	0.027** (0.011)
Urban Population %	0.125*** (0.011)	-0.046*** (0.014)	0.106*** (0.012)	0.067*** (0.011)
Population Density	0.335*** (0.026)	0.260*** (0.033)	0.114*** (0.026)	0.017 (0.026)
LTHSG	0.477*** (0.043)	0.343*** (0.059)	0.151*** (0.048)	0.071 (0.045)
Shortwave Radiation	0.193 (0.619)	2.483*** (0.667)	0.104 (0.615)	-0.871*** (0.291)
Precipitation	-0.125** (0.050)	-0.023 (0.066)	0.002 (0.034)	-0.131*** (0.030)
Temperature	0.045*** (0.010)	0.009 (0.011)	0.037*** (0.011)	-0.038*** (0.009)
Adjusted R^2	0.801	0.717	0.325	0.356

The interpretation of the remaining variables was not the focus of this paper. We included them in the regression analysis in order to minimize the possible omitted variable bias. For example, population density has a positive effect on both the pandemic growth rate and the number of air passengers. If we do not include the population density variable, the effect of air transportation on the pandemic spread will be overestimated. To avoid such occurrences, we controlled for as many variables as possible.

The estimation results underscore the importance of estimating different models for different time periods. Almost none of the predictors have the same effect across all time periods. Since the biology of the virus remained approximately the same between March and December of 2020, this result is likely the reflection of the social and behavioral response to the pandemic, population-level dynamics of the virus, such as herd immunity, and asynchronicity of the dynamics of the virus – various counties were at different stages (waves) of the pandemic at a given time. In some cases, coefficients change not only in magnitude but also in the sign. For example, in the fatality-based model for March-June, the VMT per capita coefficient was equal to 0.249, while in October-December, it was -0.320. In March-June, high VMTs contributed to the initial spread of the virus, while in subsequent time periods, when the virus was established virtually everywhere, the importance of VMT as a source of virus spread diminished. Rather, high VMT may have become more of an indicator of the degree of virus containment.

In the case of the fatality-based model that more accurately captures the pandemic spread, only a few variables had a consistent effect on the growth rate – closeness centrality, population density, and percentage of people with less than a high school degree. However, the effect of closeness centrality is very small compared to other variables. The population density effect was not statistically significant in October-December. Similarly, the effect of education was positive at first (more educated population resulted in slower growth rate) but was not significant at the end of 2020.

The effect of precipitation and temperature was small and not statistically significant for some of the models. The amount of shortwave radiation was statistically significant for the aggregate, March-June, and October-December models, but the sign of the effect was first positive then negative. The surface-level shortwave radiation reflects the amount of sunlight that reaches the surface. Sunlight increases outdoor activity, but the UV radiation might destroy Coronavirus particles. Our model shows that the average effect of sunlight over 9 months is positive, but it was large and positive in March-June (especially for the fatality-based model), and large and negative in October-December.

B. Spatial Autocorrelation

Counties with similar heterogeneity and adjacent in space may agglomerate in space. The outcome (growth rate) in a given county likely depends on both the characteristics of that county and the outcome in adjacent counties. One concern is that even after controlling for the covariates, our dependent variable may still exhibit a non-random pattern over different counties. To control for this potential spatial autocorrelation in our models, Moran's I for regression residuals ε is calculated using the following equation:

$$I = \frac{n}{\boldsymbol{W}} \frac{\sum_i \sum_j w_{ij}(\varepsilon_i - \bar{\varepsilon})(\varepsilon_j - \bar{\varepsilon})}{\sum_i (\varepsilon_i - \bar{\varepsilon})^2} \qquad (9)$$

where n is the number of counties; $\boldsymbol{W} = \sum_i \sum_j w_{ij}$ is the spatial weights matrix with zeros on the diagonal ($w_{ii} = 0$), and $I \in [-1,1]$ indicating the degree to which a spatial pattern is clumped, random, or dispersed.

Assuming that areas near each other are more similar than areas that are far apart, we first employ the Queen contiguity to define neighbors of each county. The Queen adjacency neighbors are defined as counties that share a common edge or node of the subject county. In Figure 6. , the left plot shows the histogram of the number of neighbors. Most counties in the U.S. share edges or nodes with six other counties. San Juan county, Utah, has the most neighbors. We visualize them in Figure 6. on the right.

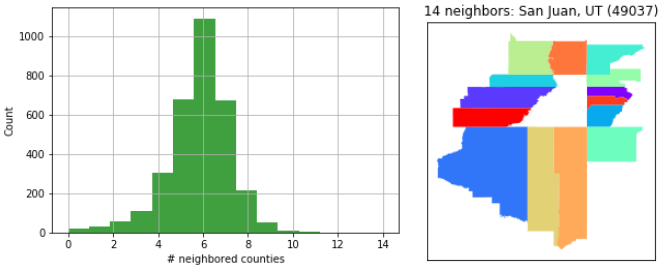

Figure 6. Histogram of Queen adjacency neighbors (left); Queen adjacency neighbors of the San Juan county (right)

Then we assign standardized weights to these neighbors by dividing the number of neighbors around the subject county. This makes sure that each neighboring county contributes the same amount regardless of how many neighbors the subject county has. Lastly, with the spatial matrix constructed, we perform the Moran's I statistical test for the fitted models after including the control covariates. The null hypothesis states that the spatial distribution of the model residuals is randomly distributed across the U.S. counties. In other words, after controlling for the covariates summarized in TABLE I. , the value of the observed growth rate is the result of a random spatial process. TABLE IV. summarizes the Moran's I statistical test results for the case growth rate models and the mortality growth rate models.

TABLE IV. SUMMARY OF SPATIAL AUTOCORRELATION OF REGRESSION RESIDUAL ACROSS U.S. COUNTIES USING MORAN'S I METRIC

		Mar. – Dec.	Mar. – Jun.	Jul. – Sep.	Oct. – Dec.
Case growth rate models	***Moran's I***	0.005	0.008	0.003	-0.008
	Z-score	0.487	0.776	0.305	-0.685
	P-value	0.313	0.219	0.380	0.247
Mortality growth rate models	***Moran's I***	-0.004	-0.005	-0.019	0.004
	Z-score	-0.314	-0.458	-1.674	0.407
	P-value	0.377	0.323	0.047	0.342

For the case growth rate models, the Moran's I values indicate fairly weak positive spatial autocorrelation across all the periods, but are not significant. We fail to reject the null hypothesis and conclude that the model residuals do not have a spatial pattern that we were unable to control for.

For the mortality growth rate models, Moran's I values indicate weak negative spatial autocorrelation for the whole analysis period, except for the period from October to December. In the second period from July to September, there is a significant negative spatial autocorrelation across the U.S. counties, suggesting a dispersed spatial pattern of the feature values. If the fatality growth rate is large in the neighboring counties, the subject county likely has a low fatality growth rate. This could be explained by the uneven distribution of nursing and medical resources among counties. Adding hospitalization features will help mitigate such spatial autocorrelation.

VI. CONCLUSIONS AND FUTURE WORK

In this study, we estimate how air travel contributed to Covid spread in the United States, using county-level data. Toward that end, we develop a novel approach for estimating the county-level daily air passenger traffic, which incorporates estimation of passenger load factor and traffic distribution. This metric, which utilizes realized flight operation data, can serve as a proxy variable for air travel and support other coronavirus-related research projects. Cross-sectional models using county-level aggregate variables are employed to quantify the impact of air travel on the pandemic growth rate in different time periods. An aggregated model from March to December, and three sub-period models (March – June, July – September, October – December) are estimated in order to investigate the time-varying effects of the predictors. While this study focuses on the air travel variables, other relevant covariates, including VMT, road network closeness centrality, urban population, population density, education level, shortwave radiation, precipitation, and temperature, are also derived from multiple datasets. We include them in the regression analysis to minimize the possible omitted variables bias.

The aggregate model has a higher goodness-of-fit than sub-period models, likely because there is more variation in the case/fatality growth rate on shorter time horizons. In addition, the goodness-of-fit of case-based models is higher than for the fatality-based models. One possible explanation is that the case growth rate reflects the scale of testing efforts undertaken in individual counties, which is influenced by the economic prosperity and cultural characteristics captured by the model. As to the primary variable of interest – the number of air passengers traveling to a given county, we find that it has a strong, positive influence on pandemic spread in both the case-based and fatality-based aggregate models. However, the effect seems to become weak and not statistically significant for the June – September and October – December models. It is probably because air travel plays a role in the early seeding of the pandemic in the U.S. Counties with more incoming air traffic were more likely to be exposed to the virus early on. After the lockdowns started, the pandemic spread is mainly determined by local factors such as population density. Moreover, as the pandemic progressed, potential air passengers adjusted their behavior in ways that reduced risks of COVID spread. Finally, we perform Moran's I statistical test for our model residuals and find no significant spatial autocorrelation of residuals that we fail to control for.

We will continue working on deriving robust signatures of the air travel influence on COVID-19 transmission. First, the current cross-sectional model should be improved. We will develop a deep learning model – Graphical Neural Networks – to capture the dynamics of the transportation influences on the pandemic spread at more granular temporal and spatial scales. With the graph structure of transportation networks added to the hidden layers, the deep learning model is capable of learning the latent temporal and spatial dependencies and predict the pandemic spread dynamics. Second, once the model meets the requirements of prediction accuracy and response speed, attention should be given to predict how various transportation interventions might have affected the growth of coronavirus transmission, and assess their effectiveness of potential targeted transportation interventions, depending on when, where, and for how long they are deployed. The counterfactual analysis could be employed where specific transportation interventions have been taken to control the spread, or might do so in the future. For example, certain air services or even entire airports might have been closed at different times. Using such a model, we will be able to assess the potential of these changes under counterfactual interventions, which can be readily represented by changing model covariates, eliminating selected nodes and links in the graph network, and or manipulating the network topology, to assess how potential transportation interventions could be adapted to cost-effectively slow down the pandemic spread and minimize its impact on society.

REFERENCES

[1] Lau, H., V. Khosrawipour, P. Kocbach, A. Mikolajczyk, H. Ichii, M. Zacharski, J. Bania, and T. Khosrawipour. *The association between international and domestic air traffic and the coronavirus (COVID-19) outbreak.* Journal of Microbiology, Immunology and Infection, 2020. **53**(3): p. 467-472.

[2] Zhang, Y., A. Zhang, and J. Wang. *Exploring the roles of high-speed train, air and coach services in the spread of COVID-19 in China.* Transport Policy, 2020. **94**: p. 34-42.

[3] Linka, K., M. Peirlinck, F. Sahli Costabal, and E. Kuhl. *Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions.* Computer Methods in Biomechanics and Biomedical Engineering, 2020. **23**(11): p. 710-717.

[4] Zhang, L., H. Yang, K. Wang, Y. Zhan, and L. Bian. *Measuring imported case risk of COVID-19 from inbound international flights---A case study on China.* Journal of Air Transport Management, 2020. **89**: p. 101918.

[5] Mazareanu, E. *Monthly passenger load factor (PLF) on international flights by region 2020*, in *Transportation & Logistics*, Statista, Editor. 2021.

[6] KIMLEY-HORNANDASSOCIATES, I. *Airport Air Service Profile: Orlando-Sanford International Airport.* 2012.

[7] Suzuki, Y., M.R. Crum, and M.J. Audino. *Airport choice, leakage, and experience in single-airport regions.* Journal of transportation engineering, 2003. **129**(2): p. 212-218.

[8] Phillips, O.R., L.R. Weatherford, C.F. Mason, and M. Kunce. *Passenger leaks and the fate of small community air service.* Economic Inquiry, 2005. **43**(4): p. 785-794.

[9] Ryerson, M.S. and A.M. Kim. *A drive for better air service: How air service imbalances across neighboring regions integrate air and highway demands.* Transportation Research Part A: Policy and Practice, 2018. **114**: p. 237-255.

[10] Araujo, M.B. and B. Naimi. *Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate.* MedRxiv, 2020.

[11] Xie, J. and Y. Zhu. *Association between ambient temperature and COVID-19 infection in 122 cities from China.* Science of the Total Environment, 2020. **724**: p. 138201.

AUTHOR BIOGRAPHIES

Lu Dai is a Ph.D. candidate of Civil and Environmental Engineering at the University of California, Berkeley. Her research broadly focuses on developing machine learning methods and statistical analysis to solve real-world problems, with applications to air traffic management, delay prediction, and aviation safety.

Ivan Tereshchenko is a Ph.D. candidate of Civil and Environmental Engineering at the University of California, Berkeley. His research interest are in airspace demand and capacity modelling, macroscopic traffic flow models, and machine learning applications.

Mark Hansen is a Professor of Civil and Environmental Engineering at University of California, Berkeley, and co-director of the National Center of Excellence for Aviation Operations Research (NEXTOR-III). Dr. Hansen received his Ph.D. in Engineering Science in 1988 from University of California, Berkeley.