Measuring the Continuing Impact of the COVID-19 Pandemic on Domestic Air Travel

Adam Packer and Aaron Newman
SIADS 593 Winter 2023

Motivation

When the Coronavirus Disease 2019 (COVID-19) pandemic took hold in early 2020, people quickly stopped traveling. The transportation sector was drastically impacted, as far fewer people were leaving their homes for anything but basic needs, much less traveling long distances for business or pleasure. In recent months, travel has recovered, but the threat from COVID-19 has not gone away in 2023. New variants that are resistant to current vaccines continue to appear in the United States, and recent case growth in China may translate to further illness around the world as travel restrictions have been eased.

During the height of the pandemic, United States air carriers were beneficiaries of funds provided by the federal government through the 2021 CARES Act and the American Rescue Plan Act of 2021. That money allowed airlines to continue to pay staff so that they would be ready to operate when demand increased. That kind of financial support is no longer available, so it is a challenge for air carriers to plan for variations in demand that may be exacerbated by public health challenges such as COVID-19.

What if data science techniques could be used to predict changes in travel demand based on surges in COVID-19 (or other widespread illness)? Then, airlines could optimize crews and equipment and better manage impacts on on-time performance, revenue, and profit.

In this project, we explored this possibility using data on United States domestic air travel by market and COVID-19 case counts and deaths. We set forth the following questions:

1. Can we discern a correlation between changes in domestic air travel volume and changes in COVID-19 indicators?
2. If so, can we observe a lag between a change in COVID-19 indicators and a change in air travel volume?
3. How do the correlation and lag differ based on choice of geographic market?
4. With that information, can we construct a model that could predict future changes in demand?

Data Sources

United States Domestic Air Travel

The Office of Airline Information within the Bureau of Transportation Statistics (BTS, part of the United States Department of Transportation) maintains the Air Carrier Statistics database. The database includes a wealth of information collected monthly from U.S. and some foreign air carriers. One table within the database focuses on monthly passenger traffic, cargo carried, and mail carried between different domestic markets. We obtained data from this table to understand the volume of domestic passenger air traffic from January 2020 (at the start of the pandemic) through November 2022 (the most recent data available at this writing.) The BTS website allows for downloads of up to one year of data at a time as ZIP archives containing comma-separated values (CSV) files [1].

Ultimately, we used data from 3 such ZIP archives, for 2020, 2021, and 2022, totalling approximately 16.4

megabytes. Each CSV file includes 36 columns, with each row representing the amount of passengers enplaned, mail carried, and cargo carried by a single carrier between an origin and destination market. A city market is defined as a metropolitan area served by one or more airports. For example, the New York City market includes travel from and to John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA), and Newark Liberty International Airport (EWR). We processed a total of 664,694 records accounting for all travel throughout the time range. For each record, we focused on the date (month and year), carrier, origin and destination markets, and number of passengers carried. In addition, we used reference files also provided by the BTS website that mapped city markets to a five-digit code [2], and that mapped two- or three-character airline codes to the carrier's name [3].

COVID-19 Statistics

The Act Now Coalition is a nonprofit organization that was founded shortly after the COVID-19 pandemic took hold. The organization established the COVID Act Now website to consolidate data from all reporting states in the United States and provide a central repository to educate and inform the public. On the website, users can observe COVID-19 caseloads, hospitalizations, vaccinations, and deaths, broken down by time, by state, and locality. The website also offers an application programming interface (API) to access the information directly, which we used to obtain the data we needed. [4]

We obtained API keys using the directions provided on the website and downloaded two CSV files: one with COVID-19 statistics on a daily basis rolled up to a national level, and one with the same statistics by core-based statistical area (CBSA). The CBSAs are defined by the U.S. Census Bureau and include 362 metropolitan statistical areas (with 50,000 or more residents) and 560 micropolitan statistical areas (with fewer than 50,000 but with 10,000 or more residents). Each is identified with a five-digit code that we later mapped to the BTS city markets.

We downloaded the national-level data as a 232 kilobyte CSV file with 61 columns and 1074 rows (as of 15 February 2023). Each row represents a day, starting from 9 March 2020. The columns include statistics on case counts (daily, cumulative, raw, and normalized by 100,000 population, etc.), hospitalizations, deaths, community spread, and vaccination status. Our focus was on the date and daily new case and death count.

For CBSA-level data, we downloaded a 146.6 megabyte CSV file, also with 61 columns but with 1,004,964 rows (as of 15 February 2023). Again, we focused on the date and daily new case and death count, as well as the CBSA code. We also obtained a Microsoft Excel workbook from the U.S. Census Bureau that maps the CBSA codes to the names of the areas [5]. We used this information to link CBSAs to city markets in the BTS data.

Data Manipulation Methods

Our analysis started at the national level, and then moved down to the city market level. As such, we needed two different views for both the BTS and COVID-19 data. We relied heavily on the Python Pandas package for data manipulation, in particular using the *groupby* method paired with aggregation by sum.

In order to perform analysis at the national level, we created a dataframe of the United States COVID-19 daily case and daily death metrics grouped by month. For comparison to air travel, we created a similar dataframe using the BTS travel data grouping passengers across all markets and airlines by month.

To prepare for analysis at the city market level, we needed to identify those city markets and make sure we linked the CBSA code in the COVID-19 data with the city market ID in the BTS data. These were both five-digit numbers, but were not related to one another. We focused on the largest 40 domestic markets and created a CSV file with the market name, a three-character code (usually mapping to the code used by the International Air Transport Association (IATA) such as NYC for New York City, WAS for Washington, or BOS for Boston), the CBSA code, and the BTS city market ID.

With that CSV file prepared, we were able to create a dataframe from the COVID-19 CBSA data that grouped cases and deaths by month and by CBSA, adding the three-character market code. For the air travel data, we ran through each city market (again adding the three-character code for comparison) and aggregated passengers departing that market and those arriving in that market by month. Finally, we saved this data as a pickle for further use. Figure 1 below shows a few rows from our merged data set.

	date	location	loc_code	carrier	p_origin	p_dest	passengers	cases	deaths
0	2020-03-01	USA	USA	ALL	34527687.0	34527687.0	34527687.0	184104.0	4199.0
1	2020-03-01	USA	USA	04Q	274.0	274.0	274.0	184104.0	4199.0
2	2020-03-01	USA	USA	09Q	26506.0	26506.0	26506.0	184104.0	4199.0
3	2020-03-01	USA	USA	1EQ	637.0	637.0	637.0	184104.0	4199.0
4	2020-03-01	USA	USA	1QQ	417.0	417.0	417.0	184104.0	4199.0

Figure 1: Sample from Merged Data Set

At this point, we have removed all extraneous data, grouped by month and location, and have the tools to compare COVID-19 data and air travel data at the national level as well as at the market level, and can move on to analyze the data. The reader can walk through this data preparation process within our setup notebook submitted with this project.

Analysis

National-Level Comparison

We begin by looking at trends in the overall data aggregated to the national level. Figure 2 below shows deaths attributed to COVID, reported cases of COVID, and domestic air passenger counts monthly from March 2020 through November 2022. Note the differences in scale among the three plots. It is worthwhile to notice some of the peaks and troughs in the data, such as the immediate drop in air passenger volume in April 2020, the large spike in reported COVID-19 cases in January 2022, and several peaks of deaths attributed to COVID, the largest being in January of 2021.

We then compared monthly nationwide COVID-19 cases with monthly counts of nationwide air passengers. When looking at the whole series of data, which covers the time period of March 2020 through November 2022, we found a Pearson coefficient of 0.037. There is little correlation between these series, and that weak correlation is not statistically significant. Early in the pandemic, the fact that domestic air travel essentially stopped was based on actions taken by the U.S government. The pandemic had not yet spread far enough to see substantial case counts, so that would not have been a factor. However, after the height of the pandemic appeared to pass, and air travel was once again permitted, we theorized that we might see some correlation.

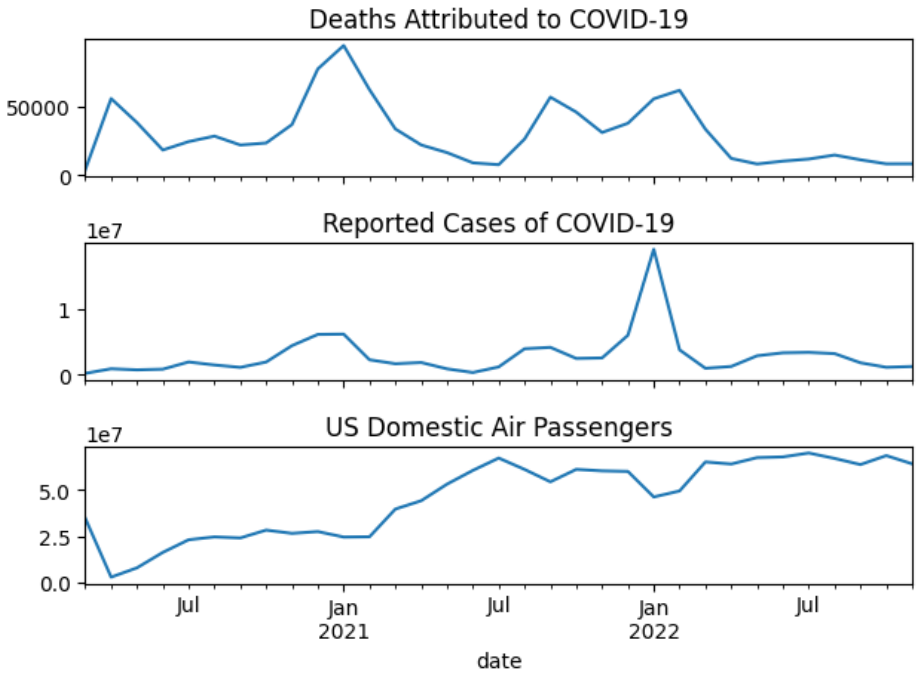

Figure 2: COVID-19 Deaths, Reported Cases, and US Domestic Air Passengers (March 2020 - November 2022)

In fact, if we limit the discussion to 2022 only (January through November), we see that the series are strongly negatively correlated, with a Pearson coefficient of -0.73 that is statistically significant to greater than 95% confidence. In general, we should be seeing a drop in passengers traveling when the case count increases, and vice versa.

We also looked at a comparison of COVID-19 deaths to nationwide air passengers. Interestingly, the entire series is somewhat negatively correlated, and that correlation (with a Pearson coefficient of -0.46) is also statistically significant to a greater than 95% confidence interval. Limiting the series to 2022 improves that correlation to -0.91.

At this scale, we also compared cases and deaths to passengers with a one-month lag. (So, we hypothesize that the case or death count in a given month is related to the passenger count in the following month.) For COVID-19 cases, we generally found improved correlation with the lag. For COVID-19 deaths, there was better correlation without the lag.

Figure 3 below shows a comparison of the correlation between case counts and passenger counts, both with and without a one-month lag. The x-axis shows the first month used in the series. As we move to the right, we remove a month from the beginning of the series to see whether the correlation changes. The Pearson coefficient is shown as dark blue bars in the figure. The orange line represents the p-value of the correlation, with the red line representing the 95% significance threshold. The orange line would need to meet or dip below the red line to indicate statistical significance. Generally, the correlation between the series does not become statistically significant until mid-2021, and the negative correlation increases up until January 2022. Here, the large spike in COVID-19 cases (evident in Figure 2 above) greatly

influences what we see in Figure 3, and beyond January, the correlation switches to positive (although without statistical significance).

In Figure 4 (also below), we show the same information, this time comparing COVID-related deaths to passenger counts. We see that the correlation (regardless of lag) remains statistically significant all the way through January 2022. The correlations remain notably larger without considering a lag and peak in absolute value in early 2022 before shifting positive and losing statistical significance.

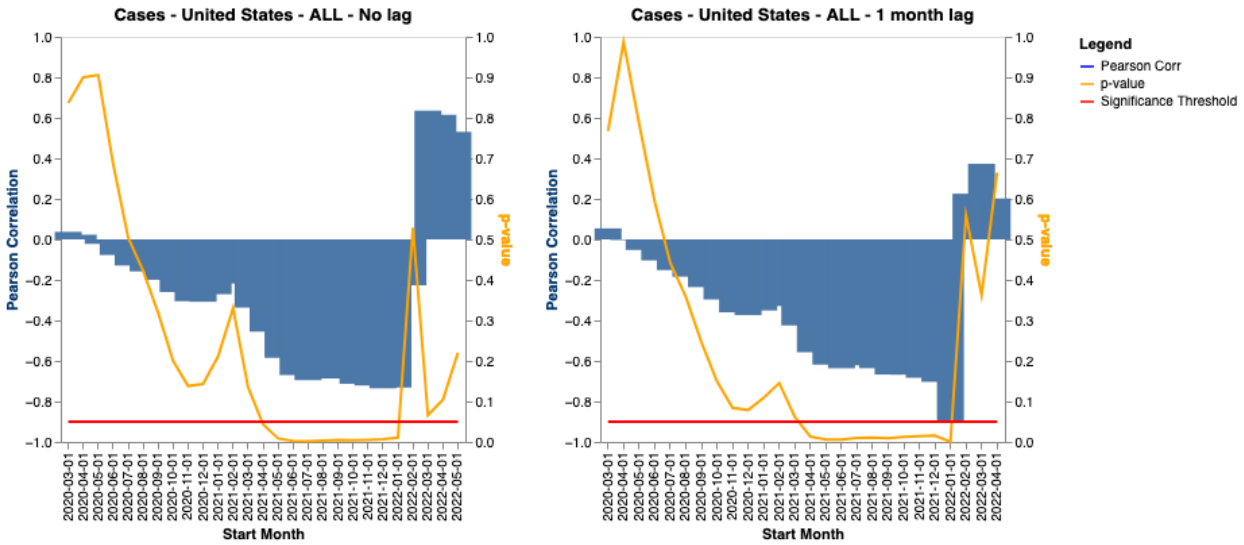

Figure 3: Correlation of COVID-19 Cases and Passenger Counts (With and Without Lag)

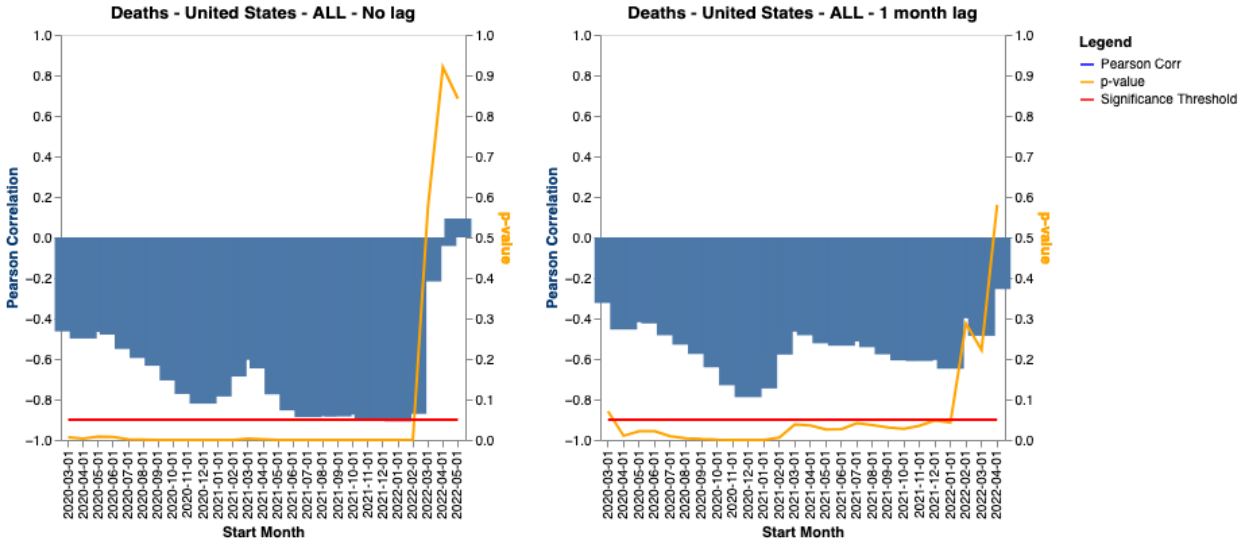

Figure 4: Correlation of COVID-19 Deaths and Passenger Counts (With and Without Lag)

National-Level Prediction

Given what we've learned about correlation between these series, could a linear regression model between cases and passengers, or between deaths and passengers, help us to predict future air

passenger travel? We used scikit-learn to fit linear regressions using the full series of data as well as limiting to 2022 data, and allowed for the possibility of a one-month lag. The results are shown in Figure 5 below:

	Type	Start Date	Lag	Series Length	R2 Score
0	cases	2020-03-01	0	33	0.001
1	cases	2020-03-01	1	32	0.031
2	cases	2022-01-01	0	11	0.535
3	cases	2022-01-01	1	10	0.000
4	deaths	2020-03-01	0	33	0.215
5	deaths	2020-03-01	1	32	0.141
6	deaths	2022-01-01	0	11	0.826
7	deaths	2022-01-01	1	10	0.814

Figure 5: Linear Regression R^2 Scores for COVID-19 Cases and Deaths vs Air Passengers

From the information in the table, it might be possible to use COVID-19 deaths as a predictor for passenger counts if we limit our scope to 2022 data. We next withheld the last two months of data as a test set, and fit the model on the first nine months of the year but discovered that it had no predictive value (with a negative R^2 score). We conclude that there is not enough data for predictive purposes and we note that this strategy neglects to account for the temporal nature of the data.

We next considered the use of a Vector Autoregression (VAR) model. VAR models can be used for multivariate time series predictions, which is relevant here. The idea is that given multiple variables, the values of those variables for a given time *t* are the weighted sums of those values at time *t-1* (assuming a lag of one). It also assumes that each variable has an impact on the others. (For example, not only would COVID-19 cases impact passenger counts, but passenger counts would impact COVID-19 cases. This is honestly a bit of a stretch, but it is plausible that a higher passenger count would bring more people in contact with the virus and then lead to a higher case count. However, while we expect that COVID-19 deaths may have an influence on passenger count, it seems unlikely that the reverse is true.)

We used the Python *statsmodels* package to generate a VAR model using all three series (cases, deaths, and passengers), working with a tutorial from the Machine Learning Plus website [6]. VAR works best when the variables used are stationary. Typically, passenger air travel statistics are not stationary, because passenger volume tends to have both a trend (air travel tends to increase year-over-year) and seasonality (there is usually more travel in the summer than the winter). Nevertheless, the historical trend and seasonality appear to have been disrupted by the pandemic, so we decided to fit the model as is, without adjustments. We used all 33 months of the full time series, training the model with the first 27 and withholding the last six for test data. We were pleasantly surprised to find that the model predicted the last six months with relatively small errors. Let's look at Figure 2 again with our predictions overlaid (in Figure 6 below).

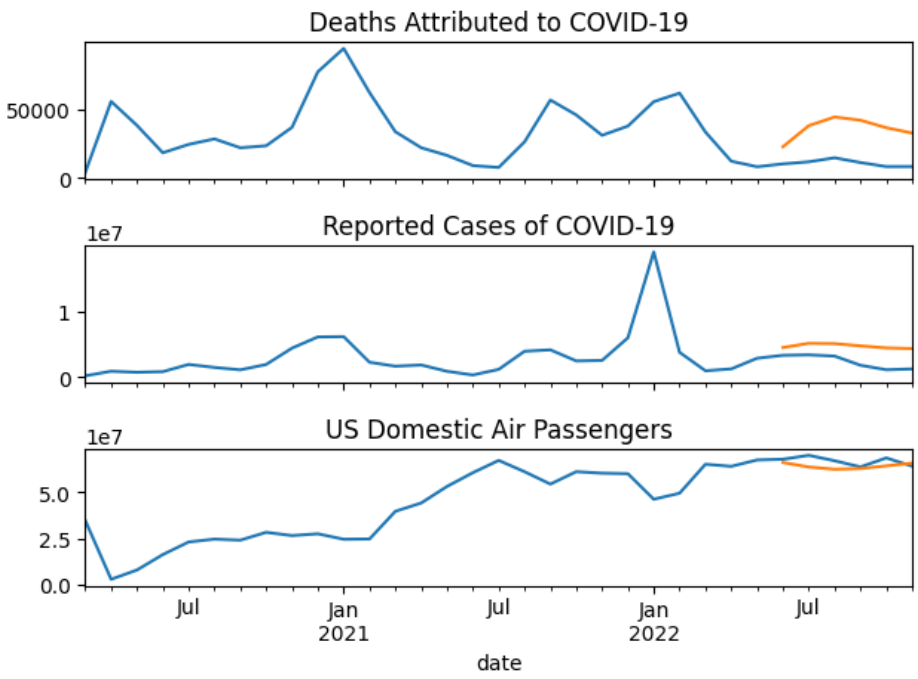

Figure 6: COVID-19 Deaths, Reported Cases, and US Domestic Air Passengers (March 2020 - November 2022, with modeled predictions for the last six months in orange)

On visual inspection, the passenger counts appear to have the least error, which is actually what we want. We calculated specific forecasting metrics, which are below in Figure 7.

variable	mape	me	mae	mpe	rmse
passengers	0.05	-2694845.93	3242777.16	-0.04	3770369.85
cases	1.44	2365055.40	2365055.40	1.44	2495145.98
deaths	2.40	25379.85	25379.85	2.40	26112.89

Figure 7: Metrics from VAR Predictions for June - November 2022

As the figure shows, the mean absolute percentage error (MAPE) for passengers is very small. In fact, it is 2.4% or less for all three variables, but predicting passenger counts is our focus. This is a promising result.

Arguably, the best method to use for predictive modeling may be Seasonal Autoregressive Integrated Moving Average (SARIMA) with Exogenous Variables (SARIMAX). SARIMAX would allow us to account for the seasonality and trending that are usually observed in air travel statistics, and use COVID-19 case counts and/or deaths as exogenous variables that influence the value of air passenger counts, but are not part of the model themselves. While a promising approach, we found that we did not have sufficient data to produce a useful model.

Market-Level Comparison

Now that we have established correlations at the national level and applied VAR modeling to predict future air travel, we turn to looking at some of the individual city markets. Looking at correlation, we found that most of the larger city markets tended to match the national trend. There was often strong negative correlation between cases or deaths and passenger volume, particularly when focusing on 2022 data. We continued to find that correlation between COVID-19 deaths and passenger counts was often stronger than between COVID-19 cases and passenger counts. The reader can explore these results in the analysis notebook that was submitted with the project [7]. We were interested to observe that there were practically no cases of statistically significant correlation for markets in the state of Florida, to include Miami (MIA), Fort Myers (RSW), Orlando (MCO), and Tampa (TPA). We are not certain of the reason for this, but it may have been caused by several changes in how COVID-19 cases and deaths were reported by the Florida Department of Health and localities within the state over the course of the pandemic. Examples for COVID-19 cases with a one-month lag are shown below in Figure 8 and details can be found in the project's analysis notebook [9].

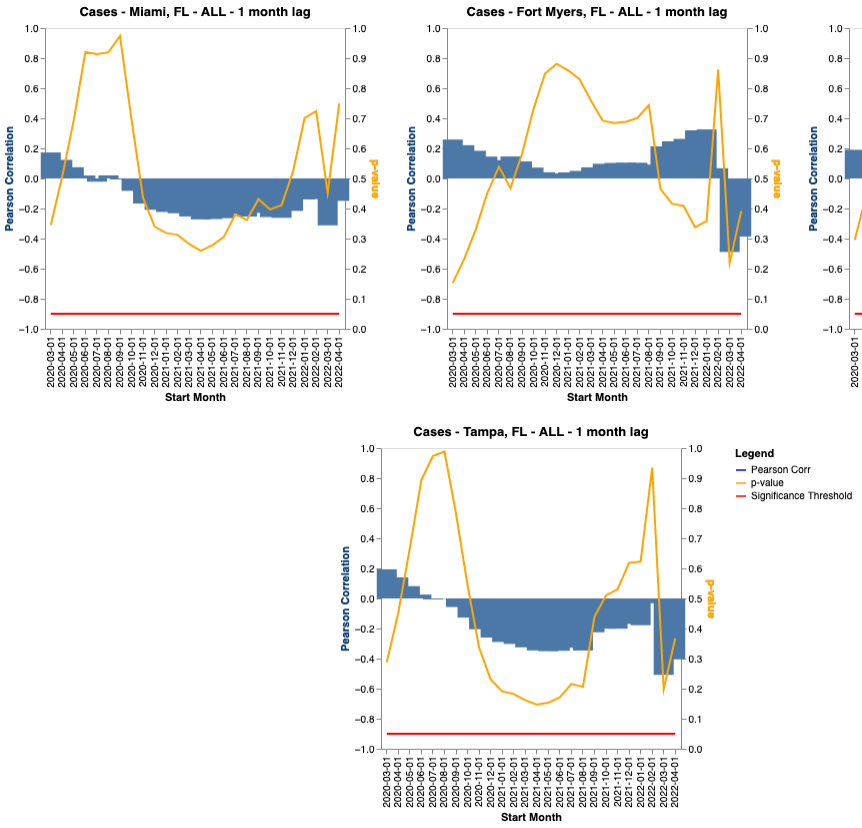

Figure 8: Correlation Between COVID-19 Case Counts and Passenger Counts in Florida Markets

In general, we also found that the major United States domestic carriers also followed the national trend. The reader can also refer to the analysis notebook for further study [8].

Market-Level Prediction

Finally, we wanted to see if VAR modeling would work at the market level as well as it seemed to at the national level. Figure 9 shows a promising example for Los Angeles, California.

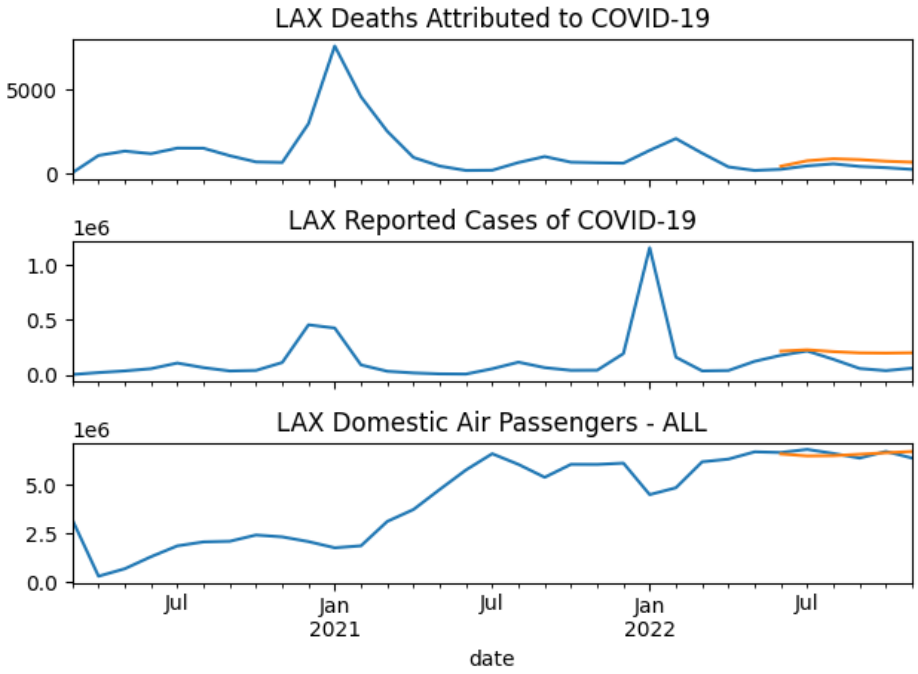

Figure 9: COVID-19 Deaths, Reported Cases, and Los Angeles Domestic Air Passengers (March 2020 - November 2022, with modeled predictions for the last six months in orange)

Figure 10 shows the metrics associated with this prediction, which remain very strong. Readers can view various combinations of locations and carriers in the predictive analysis notebook provided with this report [10].

variable	mape	me	mae	mpe	rmse
passengers	0.03	-9493.85	187752.88	-0.00	219390.25
cases	1.61	93227.86	93227.86	1.61	109199.98
deaths	0.90	334.98	334.98	0.90	344.89

Figure 10: Metrics from VAR Predictions for Los Angeles, CA (June - November 2022)

Overall Findings

We did indeed find strong correlations between COVID-19 case counts and air passenger volume, as well as correlations between COVID-19 deaths and air passenger volume for certain periods of time. We were

also able to demonstrate the ability to predict future passenger counts using Vector Autoregression. It is unclear how well this predictive modeling will work into the future as travel trends and influences will certainly change, as they have throughout the course of the pandemic. We were fortunate to be able to ignore trending and seasonality in the data for this particular time period, but that may not be the case going forward.

There are many factors that should cause of concern about these data sets:

- During the first several months of the pandemic, travel volume was impacted by COVID, but not by case or death count as the virus had not yet spread widely. The drastic drop in U.S. domestic travel volume was due to government policies greatly limiting travel in an attempt to reduce exposure. Only when those policies were relaxed did we see the potential for case/death counts to influence how many passengers actually traveled.
- There are many other reasons why travel may be impacted that are not directly related to COVID, such as weather or airline industry labor shortages causing delays and cancellations. Although not in the time frame studied, we only need to look at Southwest Airlines' software failures in December 2022 leading to many flight cancellations or the January 2023 ground stop due a failure of the Notice to Air Mission (NOTAM) system to see some of these impacts.
- There may be unobserved socioeconomic variables at play, to include personal income of potential travelers, inflation generally, and specifically the price of airline tickets.
- The data we obtained from COVID Act Now is only as good as the state and local governments who provided the raw information. Several states changed the way they reported cases and deaths during the studied time period, and that may have had an effect on the overall trend. Furthermore, as home testing has become prevalent, fewer people are reporting positive cases to public health authorities, and that is likely reflected in the more recent months in our data.
- Predictive techniques require sufficient observations to produce a viable model. Although we had daily observations of COVID-19 cases and deaths, we needed to summarize that data at the monthly level to match the BTS air travel data. Therefore, at most, we had 33 monthly observations to work with.

Next Steps

If we could obtain data at a weekly or daily frequency, we may then have sufficient observations to support more robust predictive modeling techniques, or further reduce error in our studied techniques. The Bureau of Transportation Statistics makes flight-by-flight information available as part of its analysis of flight delays. With some effort, this data could be aggregated at the daily or weekly level to support additional analysis.

There are additional COVID-19 statistics we could consider, including hospitalizations, vaccinations, and community risk levels (these were introduced by the Centers for Disease Control in March 2022). Some combination of these statistics with the ones we have discussed in this project may yield interesting results. We might also consider smoothing out some of the outliers in the data, such as the very large spike in COVID-19 cases seen nationally in January 2022.

It would also be interesting to study the impact of other variables in addition to COVID-19 statistics on air travel, such as inflation (or consumer price index), the cost of oil, unemployment, or even changes in gross domestic product growth.

Finally, we would like to include international travel (to include by international carriers) in the study to see if that offers additional insights.

Statement of Work

- Acquisition and primary EDA of air travel data - Aaron
- Acquisition and primary EDA of COVID-19 indicators - Adam
- Review of EDA results and analysis planning - Aaron and Adam
- Initial correlation analysis - Aaron (with Adam's support)
- Visualization of correlation results - Adam (with Aaron's support)
- Experiments with time-series prediction methods - Aaron (with Adam's support)
- Initial draft of final report - Aaron
- Initial draft of polished code notebooks - Adam
- Final review and editing of all products - Aaron and Adam

References

1. T-100 Domestic Market (U.S. Carriers), *Bureau of Transportation Statistics website*, https://transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FIL&QO_fu146_anzr=Nv4%20Pn44vr45. Last accessed 15 February 2023. (Note: The referenced web page has a link to download data.)
2. City Market Definitions (L_CITY_MARKET_ID.csv), *Bureau of Transportation Statistics website*, https://transtats.bts.gov/Download_Lookup.asp?Y11x72=Y_PVgl_ZNeXRg_VQ. Last accessed 18 February 2023.
3. Unique Carrier Codes (L_UNIQUE_CARRIERS.csv), *Bureau of Transportation Statistics website*, https://transtats.bts.gov/Download_Lookup.asp?Y11x72=Y_haVdhR_PNeeVRef. Last accessed 18 February 2023.
4. The Act Now Coalition, Data API and Data Definitions, *COVID Act Now website*, https://covidactnow.org/data-api and https://apidocs.covidactnow.org/data-definitions/. Last accessed 15 February 2023. (Note: The Data API link provides instructions to obtain an API key.)
5. Core based statistical areas (CBSAs), metropolitan divisions, and combined statistical areas (CSAs), *United States Census Bureau website*, https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2020/delineation-files/list1_2020.xls. Last accessed 18 February 2023.
6. Selva Prabakharan, "Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python," *Machine Learning Plus website*, 7 July 2019, https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/. Last accessed 18 February 2023.
7. Correlations at the market level can be seen in the analysis notebook submitted with the project. Case counts compared to passenger counts with one month lag can be found in Section 4.4. (Analysis_Notebook.ipynb, also available at https://github.com/newmanar/SIADS593.) Death counts compared to passenger counts with no lag can be found in Section 4.6.
8. Correlations at the air carrier level can be seen in the analysis notebook submitted with the project in Section 4.5. (Analysis_Notebook.ipynb, also available at https://github.com/newmanar/SIADS593.) Carriers included were Delta (DL), American (AA), United (UA), Southwest (WN), Alaska (AS), and JetBlue (B6).
9. Correlations for Florida markets can be seen in the analysis notebook submitted with the project in Section 4.8. (Analysis_Notebook.ipynb, also available at https://github.com/newmanar/SIADS593.)
10. Functions to perform predictive analysis with Vector Autoregression for arbitrary combinations of market and carrier can be found in the predictive analysis notebook submitted with the project in Section 2.2. (Predictive_Analysis_Notebook.ipynb, also available at https://github.com/newmanar/SIADS593.)