Flight Data to Predict COVID-19 Cases by Machine Learning

Ghadah Alshabana
George Mason University
galshaba@gmu.edu

***Abstract*—Coronavirus can be transmitted through the air in close proximity to infected persons. Commercial aircraft is a likely way to both transmit the virus among passengers and move the virus between locations. Our team utilized machine learning to determine if the number of flights into the Washington DC Metro Area had an effect on the number of cases and deaths reported in the city and surrounding area.**

***Index Terms*—coronavirus, Washington DC, virus prediction, machine learning**

I. Introduction

The ongoing pandemic has resulted in the rapid necessity of understanding how the novel coronavirus is transmitted and the various factors that allow it to rapidly disperse through a community, city, or nation. Among the risk factors to a given location, air travel may well be a factor, or have historically played a factor in allowing additional spread of the coronavirus from more infected regions to those with limited or no prior infections. The importance of learning about where and how coronavirus has entered the United States will help further our understanding of the disease. According to CDC [1] , the first coronavirus case in the US has been identified in Washington state, and that was due to air travel from Wuhan, China. The most common way Covid-19 can spread is by human interaction, through respiratory droplets such as talking, coughing, sneezing, and more. Air travelers can come from countries or areas with a high rate of infection and may very well be at risk of being exposed to the virus. Therefore, as they reach the United States, the virus could easily spread. In our analysis, we intend to use the OpenSky dataset records and combine it with CDC data to determine if the number of flights into or out of the Washington DC metro area may have impacted the number of coronavirus deaths reported in those counties and the region surrounding the respective airports in question. Other analyses have concluded that coronavirus can travel via flight and there is an inverse relationship between distance to an airport and how many coronavirus cases result from travel into the region. We suspect that the District of Columbia will show different results as a significant portion of business and political activity in the region is focused on the US federal government.

II. datasets

In this paper, we will utilize two dataset sources. The flight dataset is obtained from OpenSky, showing the air traffic during the coronavirus pandemic. Table I summarizes the OpenSky attributes [2]. The coronavirus dataset is obtained from the New York Times and shows the number of cases and death in the United States. Table II summarizes the New York Times [3].

TABLE I
OpenSky attributes

Variable Name	Description	Type
Callsign	the identifier of the flight	String
Registration	the aircraft tail number	String
Origin	the origin flight airport represented with four letters.	String
Destination	the destination flight airport, represented with four letters.	String
Firstseen	UTC timestamp of the first message received by the OpenSky Network.	String
Lastseen	UTC timestamp of the last message received by the OpenSky Network.	String

TABLE II
New York Times attributes

Variable Name	Description	Type
Date	the date of the reported Covid-19 cases and deaths.	Date
State	the name of the state.	String
County	the name of the county.	String
Fips	standard geographic identifier.	Number
Cases	the total number of cases of Covid-19.	Number
Deaths	the total number of deaths from Covid-19.	Number

For our analysis, the specific data we intend to use is as follows. From the OpenSky data: the destination (filtering for Baltimore International Airport, Dulles International Airport, and Reagan National Airport), and Last seen as that will give us an indication of the date the flight was occurring. From the New York times dataset, we intend to use the date, state, and county to filter down to the specific counties surrounding each of the airports mentioned earlier, and the area surrounding Washington DC itself. We further intend to use the cases and deaths to calculate the number of new cases and deaths occurring each day.

III. Related work

The rapid spread of coronavirus cases across the world motivates us to discover the number of flights effect on

coronavirus deaths rate. As in almost every country, the first infection cases of coronavirus were brought by travelers. While travel restrictions have been applied in many countries, they had a modest effect on limiting the spread of coronavirus cases [4], [5]. These restrictions were effective in only delaying the transmission of coronavirus [6].

In order to minimize the transmissions of coronavirus during the flights, a temperature screening was implemented at several airports. However, it was an ineffective method as approximately 45% of people would be detected by airports body screening and lower result in younger people [4]. *Khanh et al.* [7] study shows that thermal imaging scans have their limitation at determining if someone certainly has coronavirus. Moreover, lack of self-disclosure of coronavirus symptoms before and after boarding leads to an increase in the spread of the COVID-19 [7]. Additionally, previous studies reported that coronavirus could be transmitted before symptoms appear [8], [9]. As people can be infected with coronavirus disease and show no symptoms, or the symptoms appear after a period of several days. Overall, *Khanh et al.* [7], *Bae et al.* [8], and *Choi et al.* [9] concluded that coronavirus could be transmitted on aircraft and consequently increase the infection risk.

A case study was applied in China to calculate the risk index of COVID-19 imported cases from inbound international flights [10]. Through this study, *Zhang et al.* [10] used global COVID-19 data and international flight data from UMETRIP, and they found that the risk index increases significantly when there are active flights associated with highly infected countries. A research on the effect of the United States local flights on COVID-19 cases indicates a high correlation, i.e., 0.8 between travelers and population and COVID-19 cases at the onset of the pandemic [1].

However, a study conducted and published in 2020 by *Desmet and Wacziarg* [11] [12] used a cross-sectional regression model to analyze the relationship between COVID-19 cases and deaths and the distance to the closest airport with direct flights from the top five affected countries and found a negative correlation. Throughout this study, *Desmet and Wacziarg* [11]] used the collected Covid-19 cases data from the New York Times and their emotion and sentiment [13]–[16] and the international flights from the Bureau of Transportation Statistics. A retrospective case series was conducted by *Yang et al.* [17].

IV. Methodology

The first step in this project was cleaning the datasets. Specifically, the Open Sky data, being crowd sourced, included errors such as duplicate entries or null origin and destination values among other things. We further cut out the irrelevant data from the OpenSky dataset (that being the time period from 2019 until the start of the pandemic) as that information is irrelevant to our analysis. At this stage in the process, the coronavirus data was also filtered to only include the relevant region. Specifically, all cities and municipalities between the airports being studied and Washington, DC itself (including the counties containing each airport).

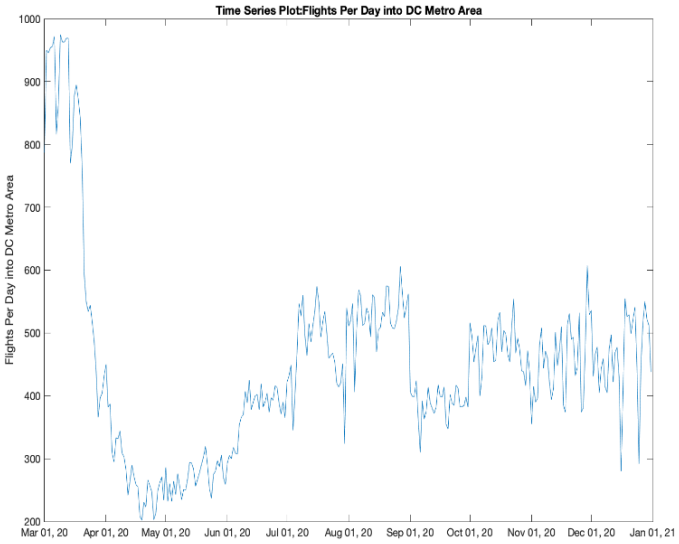

Fig. 1. Flights per day in DC Metro area

Our proposed method will involve marking all major airports in the United States and their immediately adjacent counties. We will then run a count on all flights arriving at each airport across the United States aggregated by week. Current CDC data suggests that the time from infection to onset of symptoms is 4 – 5 days [18] and research published in the Journal of Medical Virology indicates death from coronavirus has a median of 14.5 days after initial exposure [19]. As a result, based on this information, we intend to pair our flight data with coronavirus cases and deaths from one week and three weeks post the date of the flights. This should allow us to run several initial tests, including (assuming linear correlation) a Pearson Correlation Coefficient to determine if a correlation exists between number of flights and number of coronavirus deaths. We will use transfer and sentiment analysis approach in this project [20]–[24].

We can further take this data and produce scatterplots to visualize any correlation and determine other factors like spread and variance. Depending on the results of this analysis and visualization, we will then use machine learning in R and MATLAB to experiment with appropriate regression methods and attempt to produce a model that can determine (with a degree of confidence) the likely number of additional coronavirus deaths that may result from the ongoing flights into a given airport. Control counties with 0 flights will be added to the model as well, chosen at random, which do not contain airports and are not adjacent to counties with airports.

To allow for public review and availability of our data and ongoing progress, we will be posting the progress of our project online [25].

Additionally, at this stage we began looking at the increases in cases and deaths per day for each region. Below we can see visually the cases and deaths reported in Washington DC itself for the time period of March 2020 through March of 2021 (see figure **??**). Noting that our flight data ends on January 1st of

TABLE III
PROJECT TIMELINE

#	Task	Date	Status
1	Project ideas	2/7/2021	✓
2	Data collection	2/7/2021	✓
3	Project proposal	2/14/2021	✓
4	Data cleaning	2/25/2021	✓
5	Milestone#1	3/14/2021	✓
6	Data analysis	3/20/2021	✓
7	Milestone#2	3/22/2021	✓
8	Data Visualization	4/14/2021	
9	Project Presentation	4/19/2021	

2021, there are still several clear peaks that can be looked at (at least in this data set) to determine if the number of flights into the DC region have an effect on the resulting cases/deaths.

To accomplish this initial look at how closely correlated flights are with deaths and cases per day, we ran some preliminary analysis and attempted to create scatterplots around our initial concept for the Washington DC area. That is to say that we compared flights with the number of cases reported a week later when we anticipated those would be detected or otherwise reported to the appropriate health agencies. When looking at this data we found unexpectedly that there was very limited negative correlation. Running Pearson's correlation coefficient against the two values yielded a result of -0.29 where the more flights reported the fewer cases were reported 7 days later. This can be roughly seen in the scatterplot of these two values with a slight negative trend (see figure **??**). Note however that the large accumulation of points near the right end of the graph at 1000 flights per day reflects early flight data towards the beginning of the pandemic and may need to be considered differently as the coronavirus had not yet spread as thoroughly throughout the nation at the time.

A similar attempt was made at analyzing the relationship between flights and deaths reported 21 days later. This showed a slightly more pronounced negative correlation of -0.31 and a similarly more pronounced trend in the visible scatterplot of these values (see figure 4). Note that the same limitation of the cases applies, where the cluster of dots at the right side of the graph reflects data from early on in the pandemic before the coronavirus was as widespread.

V. Project Timeline

Although this project has given a better understanding for our initial question, there are many other tools and techniques that can be done to improve this analysis.We began with one idea and took it forward to collecting the data, cleaning and analyzing it so far. In the future we plan to add more data visualization that support our conclusions. As this is a semester long project, our time was very limited but in future research can be done on this topic to validate and improve the results. Also, as this project is done very close to the time COVID-19 has begun, our data could have been limited as well. Moving forward, there might be more data availability which provides for more detailed analysis.

VI. Conclusion

At this stage in the project, we've noted a few points of consideration that must be addressed in the final analysis. First and foremost, as our flight data includes pre-pandemic information, we may need to consider a better starting point (perhaps from the first case or first death reported) depending on further review to make sure that our data is actually reflective of how the virus has spread into the region after the pandemic started. Additionally, it initially seems that there is very little correlation between flights and either cases or deaths visible (or the correlation is very limited). This may be due to a number of factors, though we currently speculate that we might need a better way of aggregating the total cases and deaths reported to ensure that we properly capture any cases resulting from the flight (for example, perhaps an average of the 7 days following the flight). In either case, having completed early analysis we are now aware of several possible flaws in our initial approach and will be working to mitigate them to ensure an accurate and useful analysis for the final report.

References

[1] J. A. Ruiz-Gayosso, M. del Castillo-Escribano, E. Hernández-Ramírez, M. del Castillo-Mussot, A. Pérez-Riascos, and J. Hernández-Casildo, "Correlating USA COVID-19 cases at epidemic onset days to domestic flights passenger inflows by state," *International Journal of Modern Physics C*, vol. 32, p. 2150014, Nov. 2020.

[2] X. Olive, M. Strohmeier, and J. Lübbe, "Crowdsourced air traffic data from the opensky network 2020," 2021.

[3] The New York Times, "Nytimes/covid-19-data." https://github.com/nytimes/covid-19-data/blob/master/us-states.csv, 2021, March 13.

[4] M. Bielecki, D. Patel, J. Hinkelbein, M. Komorowski, J. Kester, S. Ebrahim, A. J. Rodriguez-Morales, Z. A. Memish, and P. Schlagenhauf, "Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: A narrative review," *Travel Medicine and Infectious Disease*, vol. 39, p. 101915, Jan. 2021.

[5] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," *Science*, vol. 368, pp. 395–400, Mar. 2020.

[6] A. Adekunle, M. Meehan, D. Rojas-Alvarez, J. Trauer, and E. McBryde, "Delaying the COVID-19 epidemic in australia: evaluating the effectiveness of international travel bans," *Australian and New Zealand Journal of Public Health*, vol. 44, pp. 257–259, July 2020.

[7] N. C. Khanh, P. Q. Thai, H.-L. Quach, N.-A. H. Thi, P. C. Dinh, T. N. Duong, L. T. Q. Mai, N. D. Nghia, T. A. Tu, L. N. Quang, T. D. Quang, T.-T. Nguyen, F. Vogt, and D. D. Anh, "Transmission of SARS-CoV 2 during long-haul flight," *Emerging Infectious Diseases*, vol. 26, pp. 2617–2624, Nov. 2020.

[8] S. H. Bae, H. Shin, H.-Y. Koo, S. W. Lee, J. M. Yang, and D. K. Yon, "Asymptomatic transmission of SARS-CoV-2 on evacuation flight," *Emerging Infectious Diseases*, vol. 26, pp. 2705–2708, Nov. 2020.

[9] E. M. Choi, D. K. Chu, P. K. Cheng, D. N. Tsang, M. Peiris, D. G. Bausch, L. L. Poon, and D. Watson-Jones, "In-flight transmission of SARS-CoV-2," *Emerging Infectious Diseases*, vol. 26, pp. 2713–2716, Nov. 2020.

[10] L. Zhang, 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*, vol. 89, p. 101918, Oct. 2020.

[11] K. Desmet and R. Wacziarg, "Understanding spatial variation in COVID-19 across the united states," tech. rep., National Bureau of Economic Research, June 2020.

[12] M. Heidari, J. H. J. Jones, and O. Uzuner, "Misinformation detection model to prevent spread of the covid-19 virus during the pandemic," 2021.

[13] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *IEEE 2021 World AI IoT Congress,AIIoT2021*, 2021.

[14] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in *IEEE 2021 World AI IoT Congress,AIIoT2021*, 2021.

[15] M. Heidari, J. H. J. Jones, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[16] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a specific business domain," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[17] N. Yang, Y. Shen, C. Shi, A. H. Y. Ma, X. Zhang, X. Jian, L. Wang, J. Shi, C. Wu, G. Li, Y. Fu, K. Wang, M. Lu, and G. Qian, "In-flight transmission cluster of COVID-19: A retrospective case series," Mar. 2020.

[18] CDC, "Interim clinical guidance for management of patients with confirmed coronavirus disease (covid-19)." https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html, 16 Feb, 2021.

[19] W. Wang, J. Tang, and F. Wei, "Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in wuhan, china," *Journal of Medical Virology*, vol. 92, pp. 441–447, Feb. 2020.

[20] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[21] M. Heidari, J. H. J. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[22] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[23] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0542–0547, 2020.

[24] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[25] M. Thompson, G. Alshabana, T. Tran, and A. Chitimalla, "Predict covid-19 cases using opensky data." http://mason.gmu.edu/ ttran81/, 2021.