Analysis of COVID-19 Impacts on Air-Traffic Volume

Thiam Wai Chua

Abstract—The first reported COVID-19 case in the United States was on February 6, 2020. The World Health Organisation (WHO) declared COVID-19 as a pandemic on March 11, 2020, and there have been rising cases of infection of the virus in the United States. This article presents the interactive data visualization exploring the correlation between the rate of spread of the COVID-19 pandemic and the air-traffic volume in the United States. The COVID-19 positive cases and the air traffic volume in the United States datasets used in this article were obtained from the open-source repository ”Kaggle”.

Index Terms—Data Visualization, COVID-19, Air-Traffic Volume, Correlation.

1 Introduction

Data visualization or information visualization plays a key role in scientific analysis. In 1854, John Snow has utilized the information visualization technique to investigate a neighbourhood cholera outbreak in St. James, Westminster, area of London [3]. A good visualization can represent inherent trends in data that may not be visible from raw data. The recent Coronavirus pandemic, which started in 2019 (COVID-19) pandemic poses new challenges to data scientists for its rapid and vast spread, thereby coincidentally drawing present-time parallels to the 1854 Cholera outbreak.

This paper analyzes COVID-19 and air traffic volume based on the currently available data from the Kaggle repository [9, 10]. Existing visualization techniques used to represent the ongoing spread of the pandemic are investigated, together with their correlation to the air traffic volume in the United States. In particular, we have developed and deployed a web-based interactive geographic map that indicates the spread of the COVID-19 pandemic and its effect upon air traffic volume in the year 2020.

1.1 Problem Description

The confirmed cases of COVID-19 have been drastically increasing around the globe since the outbreak. The United States, being one of the most impacted countries (based on both absolute and relative cases) [7], will be focused on in this analysis. Although the US is under a strict travel ban, based on the survey, only 16% of the American adults says they would travel abroad with a commercial airliner on the first day after officials remove all restrictions. Additionally, only 56% says they would be comfortable flying even 60 days after an ”all clear” is announced [6]. Therefore, based on this survey, a hypothetical relationship between the virus spread and air travel can be made. This article investigates the spread pattern of the virus in the US and its correlation with the airport traffic volume. This will be done through various data visualizations, depicting the progress in the year 2020. As a result, visualizations and analysis will have practical benefits on the domestic air travel policy planning. Additionally, they will contribute to controlling the spread of the virus.

2 Data Analysis

2.1 Domain Data Specification

Two datasets are utilized to study the correlation between the COVID-19 cases and airport traffic volume in the United States.

- **First dataset:** the first dataset shows daily traffic to and from certain airports as a percentage of the traffic volume from March to December 2020 based on the baseline period (from the 1^{st} of February to the 15^{th} of March 2020) [9]. It is the proportion of trips on a day as compared to the average number of trips on the same day of the week in the baseline period. Additionally, the dataset recorded daily geographical information, which is given in several attributes: geographical airport centroid (longitude and latitude), state name and airport name.
- **Second dataset:** the second dataset contains daily information about the amount of COVID-19 virus cases and deaths in 50 states in the United States from March to December 2020 [10].

Besides the above-mentioned datasets, the United States geographical map dataset [1] is merged with those datasets to create a choropleth map.

2.2 Data Abstraction: What

The COVID-19 cases and airport traffic volume datasets are recorded in tabular type in comma-separated values format.

For the COVID-19 cases table, the main attributes are date, state codes, numbers of positive cases and death cases. For the airport traffic table, the main attributes are the date, airport, state codes and the daily percentage of baseline. The date column is the temporal attribute and in sequential order, whereas the state codes and airport belong to categorical attribute type. The numbers of cases and daily percentage of the baseline are two main attributes for the correlation visualization, which belong to the quantitative attribute type. The key attribute in the table acts as an index that is used to look up the value attribute and is then encoded in a data visualization. The date and state code are used as the key attributes in the COVID-19 cases table, whereas the date and airport are used as the key attributes in the airport traffic table. The numbers of cases and percentage of the baseline are the scalar value fields since each state in the United States is treated as a unique and distinct geospatial scalar field. The availability of both datasets is static because the entire datasets are available all at once in a period between March and December 2020.

The United States map dataset is recorded in tabular and geometry (spatial) types in the shapefile format. The main attributes are: state code and the geometrical lines (polygons) to represent the state border. The state code belongs to a categorical attribute type. This dataset is merged with the COVID-19 and the airport traffic volume datasets to create the choropleth map.

3 Task Analysis

3.1 Domain-Specific Tasks

As discussed in section 1.1, there exists a hypothesized negative relationship between the number of COVID-19 cases and the volume of airport traffic. It is supposed that this is a one-way relationship, such that the amount of COVID-19 cases affect the airport traffic volume, but not vice-versa (i.e. that an increase in airport traffic volume does **not** lead to a decrease in COVID-19 cases).

Following this hypothesis, the domain to which this analysis appeals most is the federal government, the state government, the Federal Aviation Administration and the public health official. With this application, governmental institutions can explore how and at what pace the virus

• *Thiam Wai Chua* {*t.w.chua1985@gmail.com*}

spreads through the country through aircraft. Subsequently, the visualizations aid in determining the effectiveness of measures taken against the spread of the virus. Since the visualizations map the airport traffic volume against the positive COVID-19 cases, the visualizations are especially useful when the measures against the spread contain cutting the volume of aeroplane movements, as an anti-spread measure. Additionally, policies could be updated as to whether air travel is continued or cut. With the visualizations presented in this report, at least the following questions can be answered:

- How quickly does limiting aeroplane movements result in a decrease of the viral spread?
- How quickly should a governmental organisation press a ban or limitation of aeroplane movements?
- What is the best policy to preserve original flight movements volume, whilst also stopping the spread of the virus?

Given that the visualizations in this report are created with COVID-19 data, whilst also still being in the pandemic, the visualizations can aid in backing future measures against the spread of COVID-19. Additionally, the visualizations can prepare the governments to face the future local or global health emergency. The authors of the report warn the reader that the report and its visualizations should be **purely indicative** in case the latter use is employed, given that another viral outbreak and its circumstances will supposedly not be similar to the COVID-19 viral outbreak.

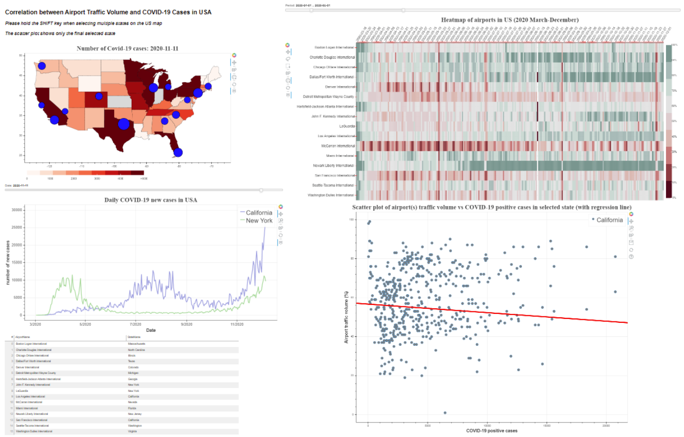
3.2 Task Abstraction: Why

The domain-specific tasks, as mentioned in 3.1 can be generalized into domain-independent tasks, such that visualization principles can be applied. To do so, an abstraction of the tasks is made. Generally, a visualization exists for the sake of consuming and producing information. This overarching goal can be broken down into two more levels of actions, resulting in a total of three step-wise levels of actions. Besides actions, there exist targets. Whereas actions are verbs, targets are nouns, and they explicitly focus on the bottom two levels of the actions [4].

According to Munzner [4], three levels of actions define the user's goals. The highest-level action describes how the visualisations are used to *analyze* by either *consuming* or *producing* data. In this report, the visualizations need to do both. On the one hand, data is consumed for computation. This can be further distinguished into the need to *discover* new ideas. This is used in, for example, policy evaluation or to observe how the virus has spread throughout the months. Additionally, one would discover how the volume of air traffic is impacted. The *discover* goal includes both finding new ideas and hypotheses, together with verification of pre-existing information. The other two distinctions (or uses) of consuming data, next to discovering, are "presenting" and "enjoyment". These do not apply to the visualisations in this report.

On the other hand, this report's visualizations need to *produce* data. The end-user needs to be able to *annotate*, *record* and *derive* information from the visualisations. The former refers to adding graphical or textual details, such as the adding states to a selection in this case. The second refers to capturing information and artefacts about the visualization session, which can be achieved by saving the visualizations in their current condition as a screenshot, for example. The latter, *derive*, refers to producing new data elements based on current data items. In these visualizations, this could be achieved by combining multiple states in the US, for example, after which differences between states can be derived.

The mid-level actions are classified as *search* actions. Search actions are subdivided into two parameters (*identity* and *location* of the items of interest) with two values (*known* and *unknown*) each. Since users using our visualizations do most likely not know where they are going to find what they need, the location is unknown. Also, the identity of what exactly they are looking for is unknown, since there is a wide range of information to be discovered from the visualizations. Thus, the mid-level search action is characterized as an *explore* alternative.


Correlation between Airport Traffic Volume and COVID-19 Cases in USA
Please hold the SHIFT key when selecting multiple states on the US map
The scatter plot shows only the final selected state
Number of Covid-19 cases: 2020-11-11
Date: 2020-11-11
Heatmap of airports in US (2020 March-December)
Boston Logan International
Charlotte Douglas International
Chicago O'Hare International
Dallas/Fort Worth International
Denver International
Detroit Metropolitan Wayne County
Hartsfield-Jackson Atlanta International
John F. Kennedy International
LaGuardia
Los Angeles International
McCarran International
Miami International
Newark Liberty International
San Francisco International
Seattle-Tacoma International
Washington Dulles International
Daily COVID-19 new cases in USA
California
New York
number of new cases
30000
25000
20000
15000
10000
5000
0
3/2020
5/2020
7/2020
9/2020
11/2020
Date
Scatter plot of airport(s) traffic volume vs COVID-19 positive cases in selected state (with regression line)
California
Airport traffic volume (%)
COVID-19 positive cases

#	AirportName	StateName
0	Boston Logan International	Massachusetts
1	Charlotte Douglas International	North Carolina
2	Chicago O'Hare International	Illinois
3	Dallas/Fort Worth International	Texas
4	Denver International	Colorado
5	Detroit Metropolitan Wayne County	Michigan
6	Hartsfield-Jackson Atlanta International	Georgia
7	John F. Kennedy International	New York
8	LaGuardia	New York
9	Los Angeles International	California
10	McCarran International	Nevada
11	Miami International	Florida
12	Newark Liberty International	New Jersey
13	San Francisco International	California
14	Seattle-Tacoma International	Washington
15	Washington Dulles International	Virginia

Fig. 1. Web-based interactive dashboard designed by utilizing Python Bokeh.

Now that the search action has been defined, we can move on the lowest-level action: the *query*. This level is subdivided into three scopes: *identification*, *comparison*, and *summarizing*. All three apply to this report's visualizations: the former because one wants to for example find where in the timeline a characterizing action has happened (no aeroplane movements, high positive tests for example). The second applies since the course of two different anti-spread measures (or viral outbreaks) can and need to be compared. The latter, *summarize*, applies since one needs to be able to have an overview of what exactly happened to be able to evaluate and learn from past actions or happenings. Also, the 'journey' (so far) through a viral outbreak can be summarized, this way. This aids in updating policies. Furthermore, the correlation between positive COVID cases and airport traffic volume per state is of great importance to test hypotheses and answer the questions in section 3.1. Additionally, (dis)similarity between states needs to be considered, for policy updates and evaluation.

Based on Munzner [4], two of the three high-level targets are used in this visualization design, which is *trends* and *features*. The patterns of the positive cases and airport traffic volume are shown in a line plot and in a heatmap, respectively. The features of interest of the datasets can be identified in the line plot and heatmap. I.e., the maximum and minimum values across the dates are clearly shown in both plots. The correlation between two attributes, positive cases and the airport traffic volume of a state are plotted in the scatter plot. Understanding the trends, distribution and correlations of these attributes, positive cases and traffic volume is extremely important for the authorities to answer the questions posed in section 3.1.

4 Visualization and Interaction Design: How

The complete view of the designed web-based interactive visualization tool by using Python Bokeh is shown in Figure 1. This visualization tool is designed according to [4]; the datasets of the COVID-19's positive cases and airport's traffic volume are stored in the spatial and chronological order as described in section 2.

- **First visual encoding:** the first visual encoding is the choropleth map. Both datasets are mapped to the US map for a specific date. The luminance in red with separated in nine magnitudes is used, for representing the daily positive cases on each state because it is a common colour often employed in practice. Furthermore, this colour is very emotive, i.e. it can easily connotate danger and death to the public. The size of the blue bubbles in the choropleth map as shown in Figure 6 represents the relative traffic volume of each airport in a specific date, i.e. the larger bubble size means higher airport traffic volume. The blue colour is selected as this colour gives an adequate contrast to the red colour on the choropleth map [4].

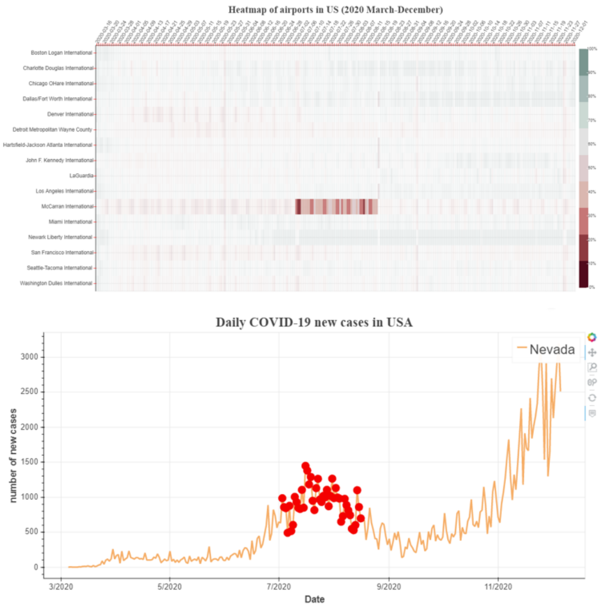


Fig. 2. The interaction of the heatmap with the line plot: the red circle marks are plotted on the line plot with the selected dates on the heatmap.

- **Interaction:** the slider acts interactively with the choropleth map for specific date selections, where the user can choose the data of the desired date to be represented on the choropleth map. The idiom is automatically updated every time the user slides the date slider. Besides that, multiple states can be selected on the choropleth map, which updates the figures of the daily positive cases on the line plot, scatter plot and in the heatmap.

- **Second visual encoding:** the line plot (Figure 7) shows the reported daily cases of selected states throughout the year 2020 from March until December. A unique colour has been assigned for each state, i.e. the line colour for every state is constant, for example, light green for California and light purple for New York, as shown in Figure 7. This colour does not change if extra states are selected, and thus directly belongs to its state.

- **Interaction:** the idiom of the line plot is updated every time the user selects multiple states in the choropleth map as mentioned in the interaction of first visual encoding. Besides that, the line plot interacts with the heatmap, as red circle marks are plotted on the line plot according to the selected dates on the heatmap by the user. Such, the user can quickly see the location in time, based on the selected data in the heatmap.

- **Third visual encoding:** the heatmap as shown in Figure 8 shows the traffic volume of every airport in the dataset. The heatmap is selected for the air-traffic volume presentation instead of the line plot because it provides an easily interpretable and meaningful comparison of the same large number of time cells (260 days) across different airports. The saturation colours ranging from red (0%) to green (100%) are used to map the volume percentage of the airport traffic volume. The range of red colours are selected between 0% and 50%, such that it grabs the attention of the authorities if there is a large decrease compared to the normal situation of 100% air-traffic volume. I.e., a darker shade of red means a lower percentage of movements.

- **Interaction:** as explained in the interaction of the second visual encoding, the heatmap interacts with the line plot. Figure 2 shows the interaction of the heatmap with the line plot. The red circle marks on the line plot appear as the dates in the heatmap are selected by the user. Additionally, there is an interaction with the US choropleth map (Figure 6): the selection of the specific airport (blue circle) in the map will be shown in the heatmap accordingly. Finally, there is a slider to select and update the heatmap to show a specific range of dates.

- **Fourth visual encoding:** in the scatter plot as shown in Figure 9, the correlation between the airport traffic volume and the COVID-19 positive cases is plotted along with the regression line. The zero-dimensional mark type (point) is used for the data presentation and the one-dimensional mark type (line) is used for the regression line which is the derived attribute to summarize the correlation between both attributes, as discussed in [4]. With the blue colour for points and the red colour for the regression line, there is an adequate contrast for the user.

- **Interaction:** The scatter plot interacts with the choropleth map. Concretely, the idiom of the scatter plot is updated with the last-selected state in the choropleth map by the user. As merely the last-selected state is shown in the scatterplot, this is considered as a limitation of the visualization design.

5 Realization

This visualization is designed by mainly utilizing ”Bokeh” [2] and ”Pandas” [5] which is built upon Python and JavaScript. Therefore, this Python library can generate an interactive web-based visualization design. The combination of Bokeh and Pandas allows the user to interactively update *ColumnDataSource* (the fundamental data structures of Bokeh) that provides a powerful capability to connect multiple plots and hand it over to the user. Each user input (selection) filters the *ColumnDataSource* accordingly, to reflect what is shown in the figures.

The hover and selection tools available in the Bokeh enable the user to select multiple states on the choropleth and multiple cells in the heatmap and show necessary information, when the user hovers the mouse over the certain object (line plot, airport, state, etc.).

Finally, zoom, pan and reset tools available in Bokeh allows to interactively get a better and more convenient view for the user.

6 Use Cases

In this section, we explain how to use the application to answer the questions in section 3.1.

- **First question:** How quickly does limiting aeroplane movements result in a decrease in the viral spread?

  Due to the stay-at-home orders and urging people to avoid non-essential air travel to contain the spreading of COVID-19 started in March 2020, for example, at John F. Kennedy International Airport in New York, one terminal has been closed and flights were relocated, [8]. The line plot and the heatmap in Figure 3 respectively show that the positive cases decreased after the peak in April 2020 and that airport traffic volume dropped to approximately 40% in New York. These numbers are consistent with the news [8]. The number of the cases dropped after posing the restriction on the air travel in April 2020, therefore, limiting air travel effectively restrains the viral spread.

- **Second question:** How quickly should a governmental organisation press a ban or limitation of aeroplane movements?

  Figure 4 shows the daily positive cases in California, Texas, and Florida which were considered severely hit by the virus at the beginning of 2020 [11]. They share the same pattern and they took approximately 6 weeks to reach a peak of more than 10000 new cases daily. This provides an invaluable insight into the spreading speed and how quickly the governmental organization should act on the commercial aircraft movement. Based on the currently available data, the authorities should take action within 2-3 weeks when there is a sign of an outbreak. Figure 5 shows the dashboard of the daily positive cases and airport traffic volume in California, Texas and Florida in July 2020. The data shows that the airport traffic volume in those states was more than 70% during summer.

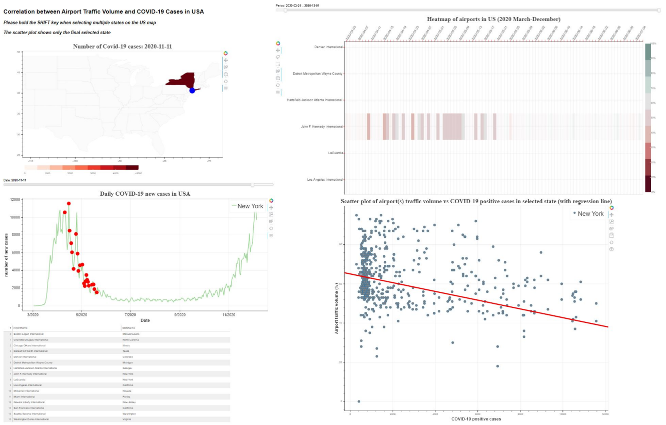

Correlation between Airport Traffic Volume and COVID-19 Cases in USA
Please hold the SHIFT key when selecting multiple states on the US map
The scatter plot shows only the final selected state
Number of Covid-19 cases: 2020-11-11
Heatmap of airports in US (2020 March-December)
Daily COVID-19 new cases in USA
New York
number of new cases
Date
Scatter plot of airport(s) traffic volume vs COVID-19 positive cases in selected state (with regression line)
New York
Airport traffic volume (%)
COVID-19 positive cases

Fig. 3. The dashboard shows the daily positive cases in New York and the airport traffic volume in April in the heatmap.

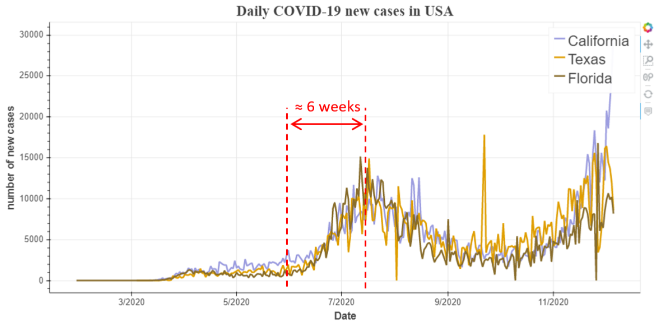


Fig. 4. A line plot of the daily positive cases in California, Texas, and Florida.

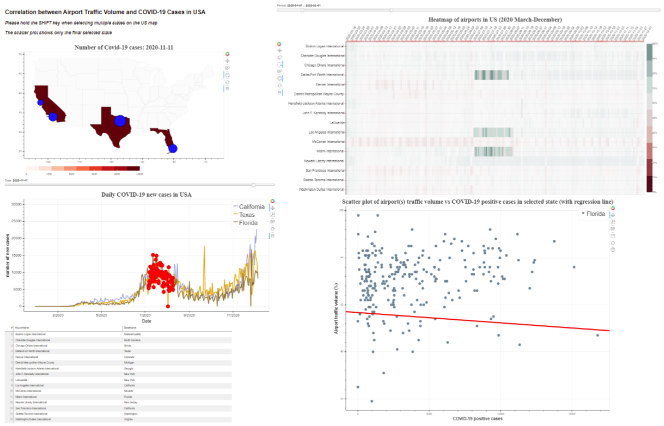

Correlation between Airport Traffic Volume and COVID-19 Cases in USA
Please hold the SHIFT key when selecting multiple states on the US map
The scatter plot shows only the final selected state
Number of Covid-19 cases: 2020-11-11
Heatmap of airports in US (2020 March-December)
Daily COVID-19 new cases in USA
California
Texas
Florida
Date
Scatter plot of airport(s) traffic volume vs COVID-19 positive cases in selected state (with regression line)
Florida
Airport traffic volume (%)
COVID-19 positive cases

Fig. 5. The dashboard of the daily positive cases and airport traffic volume in California, Texas, and Florida.

This is due to the high season for the summer holiday and the air travel in those states being relatively less strict compared to New York. Thus, this might have contributed to the rapid spread of the COVID-19 in those states during summer.

- **Third question:** What is the best policy to preserve original flight movements volume, whilst also stopping the spread of the virus?

  This is not an easy question which can be answered by using only this application. However, from the data visualization, we can see that restricting air travel is a fairly effective way of reducing the speed of the spread. In particular, as can be seen in the scatter-plot in figure 9, there is a negative correlation between the daily new cases and the airport traffic volume, i.e., the increase of new cases causes the decrease of the airport traffic volume. Thus the federal government, state government, Federal Aviation Administration (FAA) and the public health officials should post a strict air travel policy (testing before on-board), as it was shown to be fairly effective in the New York example.

7 Discussion and Conclusion

7.1 Discussion

This application gives a valuable insight into the correlation between the COVID-19 daily cases and the airport traffic volume. This data can be used as a reference for future local or global health emergencies. This application is used to answer the questions in section 3.1. During the development, some problems were encountered. The interaction between the heatmap and the line plot cannot be achieved by using the *ColumnDataSource*, a data structure of Bokeh. This is because joining the dataset of COVID-19 cases and airport traffic volume is cumbersome. Therefore, the information of the selected dates on the heatmap is transferred to the line plot. Then, red circle marks are plotted as new glyph on the line plot.

The design of the selection of the states was initially achieved by using checkboxes. Subsequently, the application was improved by states selection on the choropleth map, This gives a better interaction between the application and the user, given that it simplifies the layout of the application.

The limitation of this application is that only a single state's data of the new cases together with a single airport's traffic volume of the last selected state on the choropleth map can be plotted on the scatter plot. I.e., the application cannot generate multiple scatter plots for each selected state. The complexity of the algorithm increases, since the authors had no prior knowledge as to how many states would be selected by the user beforehand. Plotting multiple scatter plots for each selected states could be designed in the future work. Also, a Scatterplot Matrix (SPLOM) could be a valuable asset. However, since a single scatterplot contains all the necessary information, it is therefore adequately intuitive to the user.

The data shows that the pattern of spread of the virus is identical for several states. This might be due to the identical politic, geographic and demographic factors. As future work, the datasets with these factors can be added to study the correlation of these factors with the COVID-19 and air-traffic volume data.

7.2 Conclusion

In this project, we developed an application which can be used to study the spread of the COVID-19 pandemic and its impact on the air traffic volume in the US. The datasets were obtained from the open-source repository Kaggle. The choropleth map is used to present the new cases and airport traffic volume of each state. The line plot and heatmap are used to present the daily new cases and traffic volume from March until December 2020, respectively. This application can be used by the authorities to explore, compare and summarize the information.

Using the application and the datasets, it can be concluded that there is a negative correlation between COVID-19 new cases and air-traffic volume, i.e. the air-traffic volume decreases as the new cases increases. However, the relatively high air-traffic volume during summer in California, Texas and Florida might lead to an increase of more than 10000 new cases daily. We found out that the pattern of spread of the virus was comparable for those states. Therefore, identical measures can be set for those states and variant measures for other states. This accelerates the response to within 2 to 3 weeks given that the speed of spread of the virus took approximately 6 weeks.

Overall, the application covers a set of questions, which can be used by the federal and state authorities to explore the patterns and draw conclusions of a viral outbreak, primarily focusing on COVID-19. This enables the authorities to be prepared for future local and global health emergency.

References

[1] United States Census Bureau. Cartographic boundary files - shapefile. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.

[2] Bokeh Contributors. Bokeh 2.2.3 documentation. https://docs.bokeh.org/en/latest/index.html.

[3] Tom Koch and Kenneth Denike. Crediting his critics' concerns: Remarking john snow's map of broad street cholera, 1854. *Social Science and Medicine*, 69:1245–1251, 2009.

[4] Tamara Munzner. *Visualization Analysis Design*. CRC Press, 2014.

[5] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[6] Dan Reed. The data say fear of covid-19 will hamper the travel industry's recovery for years. https://www.forbes.com/sites/danielreed/2020/05/14/the-data-say-the-fear-of-covid-19–will-hamper-the-travel-industrys-recovery-for-years/?sh=429b4a2c103f.

[7] Joost Schellevis. Nederlandse besmettingscijfers uitzonderlijk hoog, ook wereldwijd. *NOS*, 12 2020.

[8] David Shepardson. U.s. faa juggles air traffic staffing as flights plummet amid coronavirus. https://www.reuters.com/article/us-health-coronavirus-usa-aviation-idUSKBN21H3HP.

[9] Terence Shin. COVID-19's impact on airport traffic. https://www.kaggle.com/terenceshin/covid19s-impact-on-airport-traffic.

[10] SRK. COVID-19 in USA. https://www.kaggle.com/sudalairajkumar/covid19-in-usa.

[11] The New York Times. Coronavirus in the u.s. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.

A Appendix

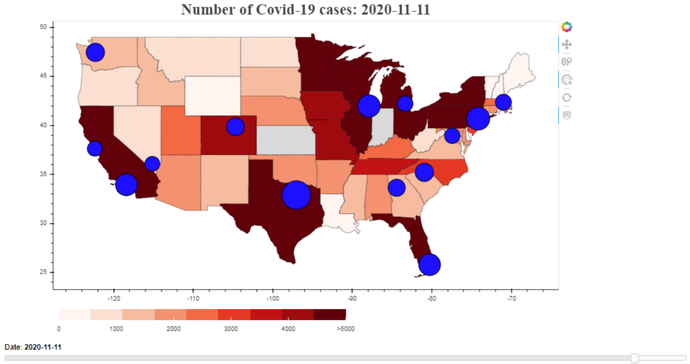


Fig. 6. A choropleth map showing the COVID-19 positive cases and percentage of the baseline of the airports' traffic volume on a specific date.

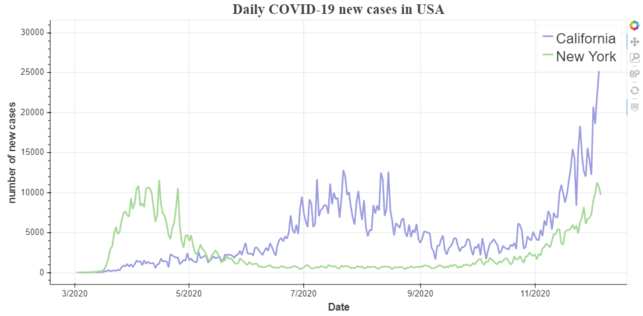


Fig. 7. The line plot of the COVID-19 daily reports positive cases of the selected states throughout 2020 March until December.

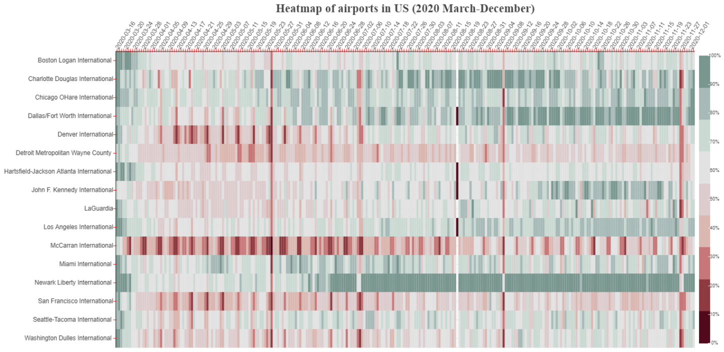


Fig. 8. The heatmap of the traffic volume of each airport throughout 2020; March until December.

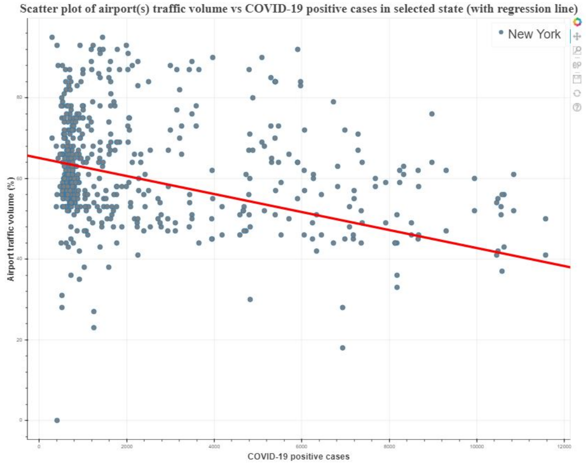


Fig. 9. Scatter plot of COVID-19 cases and airports traffic volume in New York (including the regression line).