Simple D3.Js Visualization for airplane delay dataset

Theo Jaunet * NASA Research
Mickael Bettinelli † Google Research
Miguel Solinas ‡ Microsoft Research

ABSTRACT

For decades, the Department of Transportation's (DOT) has served as an extensive resource used by engineers and managers to study flight delay patterns. Data visualization in the form of dynamics graphs and charts are actually used to summarize findings and knowledge discovery in large multidimensional datasets. This approach may be limited considering the fact that the DOT database is multidimensional. In order to discover valuable information and insight from the DOT database, there is need for an interactive and dynamic visual data exploration tools. This paper discusses the US Flight delays through a D3.js data web visualization used to explore the DOT data. Using a combination of bubble aggregation and pie charts, the dynamic knowledge generated have the potential to provide a greater level of insight into the DOT data.

Index Terms: [Multidimensional Dataset]: Crossfilter—Bubble Aggregation Dynamic pie chart Bar chart;

1 INTRODUCTION

Over the last years, the amount of flights and U.S population have increased reciprocally. These raises had a direct impact on supply-demand pricing, enhancing the airlines offers. As a matter of fact, there are more people who are willing to take a plane for a short distance than to travel in another transport like buses or personal vehicles. This fact came up with an increment in the number of flight routes, population concentration areas and overcrowding airport runways. Furthermore, these factors have direct consequences in the flight delay propagation and has an impact on strategic air traffic flow management [4].

Strictly speaking, these issues can be translated in new challenges in terms of traffic control regulation, security guarantees and delays prevention systems. Related works [14] addressed to the flow management problem have proposed different prevention measures and services to anticipate any indecent or hindrance during the passenger's trip.

This work was conceived as an alternative of the related proposals [1]. Particularly, we think that enhancing the visual understanding by new visualization methods allow the user to expand their perceptual capabilities. As an example, an inexperienced person can understand the situation in order to make efficient decisions as the data comes up.

We have assumed that the information will be treated and processed in real time over its final implementation. In order to validate our data visualization approach a dataset from the U.S. Department of Transportation's (DOT) was studied and explored. The DOT dataset tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled and diverted flights appears in DOT's monthly Air Travel Consumer Report published about 30 days after the month's end. DOT have began collecting details on the causes of flight delays since June 2003. This version of the dataset was compiled from the Statistical Computing Statistical Graphics 2009 Data Expo [16].

In order to study the datasets, different models have been proposed to analyze their composition and variable importance. However, these models are very general, they are based on the dataset implementation and require considerable technical and programming efforts. Moreover they are usually not quite intuitive. Consequently,*Tableau Public* [18] and *d3js* [3] tools are increasingly popular alternatives to visualize and predict the behavior of the datasets.

In general, tools like Crossfilter were proposed in [15] and extended in [13] as a method to explore interactively different datasets. From a user perspective, this type of tools can be conveniently implemented as a manner to have a fast and intuitive information of how the data is distributed. Furthermore in data mining areas, it can be essential to highlight knowledge in large multidimensional datasets. It is well known that visualization of information broaden human perceptual capabilities [6]. In the same way, visual representation of multidimensional data set supplies better insight into deep datasets allowing more efficient management decision making.

In general, this work results particularly attractive since it combines a good controllability and observability of the data with the ability to show in a single click a multidimensional information.

In these works, a multidimensional bubble chart aggregation is introduced as an evolution of the methodology that tackles the aforementioned weaknesses of the multi variable representation methods.

This work is based in three different representations methods implemented on the D3.js JavaScript library. It allows a straightforward design of extensive multi-variable visualization while facilitating the identification of sensible delay patterns.

This work is structured as follows. Background and previous works are reviewed in Section 2. Next, a well know technique Crossfilter is adapted to explore the DOT dataset in section 3. Tendency hypothesis methodology is detailed in section 4. The new Visualizations multidimensional bubble chart aggregation and its underlaying techniques are described in Section **??**. The discussions of these techniques are presented in Section 6. Finally, the conclusions of the paper are drawn in Section 7.

2 RELATED WORK

Many areas of research analytics for Air Traffic Management (ATM) involve the processing of big data and could benefit from the architectures promoted by the big data tool sets [12].

Data visualization methods have been applied to traffic datasets for many years. One of the earliest chart is the Marey's time-line graph [9]. Despite its rudimentary, this chart was a first step forward to understand the opportunities of what data-visualization can offer in this domain. However, this solution is difficult to apprehend as the cardinality of data increases. This chart, firstly made to study train delay, has been used in many ways. One of the most common is to alter every line thickness to fit the density also known as load in train traffic study.

Density visualizations such as heat-maps improved the understanding of spatial-temporal characteristics of traffic related data [19]. This method has become a widely applied technique for visualizing complex spatial patterns [7]. The discrete distribution of travel is

*e-mail: theo.jaunet.sio@gmail.com
†e-mail:mickael.bettinelli@etu.univ-lyon1.fr
‡e-mail:migue.solinas@gmail.com

processed into a continuous color distribution, where the travel demand intensity and trip hot spots are intuitively revealed. As such, heat maps can provide an interpretable visual representation of complex spatial distribution patterns, which brings a fresh perspective to estimate an activity.

In order to trustfully appreciate the air fights activity, some ideas were to plot the plane movements as a straight line from the take off airport to the landing one. Such method is often used to extract pattern from large dataset [10]. We have a large dataset as described in the introduction. This ability to highlights some key points out of chart is non-trivial issue. Furthermore, displaying temporal patterns is considerably harder than showing static patterns.

High observability and controllability of all the modeled variables are among the most relevant advantages of this technique [17]. The technique was implemented to represent multidimensional data using bubble chart. This one is a variation of a scatter chart in which the data points are replaced with bubbles, and an additional dimension of the data is represented in the size of the bubbles. Just like a scatter chart, a bubble chart does not use a category axis — both horizontal and vertical axes are value axes. In addition to the x values and y values that are plotted in a scatter chart, a bubble chart plots x values, y values, and z (size) values. In the same way different approaches was proposed [5] [8] which implement bubble charts incorporating another techniques to make the visualization more intuitive .

Our buuble technique is inspired by recent work on taking multi variables to explore new visual forms or aid chart reading . Because our bubble chart tool operates on D3 visualizations, it can directly modify the variables providing more control than the earlier techniques.

3 Exploring the Dataset

In order to explain and explore the dataset the Crossfilter technique was adapted taking into account some important variables like time delays, distance and dates. Before addressing the patterns identified a short description of the applied tool will be presented.

The following visualization represents an extension and an adaptation of the developed project by Square, Inc. We have used this tool called Crossfilter which was developed in order to create fast multidimensional filters for coordinated views.

Crossfilter is a JavaScript library [11] for exploring large multivariate datasets in the browser. Crossfilter supports extremely fast interaction with coordinated views, even with datasets containing a million or more records. This version of Crossfilter is a community fork of the original Crossfilter project [2].

The coordinated visualizations below Figure 2 show nearly a two million flights from early 2008: it is part of the U.S. Department of Transportation's dataset. The dataset is 260MB. It is possible to click and drag on any chart to filter by the associated dimension. The table beneath shows the eighty most recent flights that match the current filters.

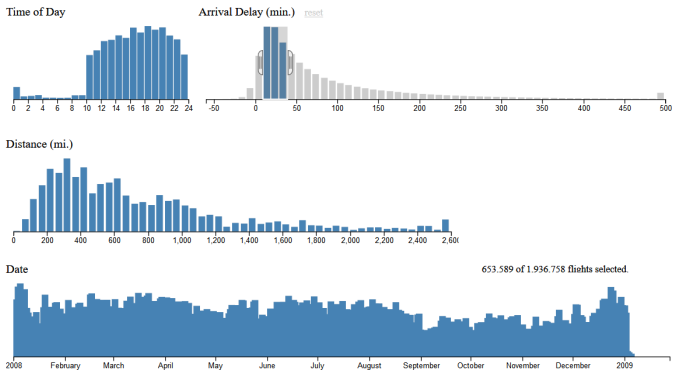

Figure 1: Average delay

Thanks to this tool it was possible to find some critics dates and periods. Specifically we can find two periods quite relevant. On one hand, selecting the delays higher than 200 min, we can discover that the dates related to this delay are during the holidays seasons. On the other hand the 25 percent of the flights are affected by delays between 10 and 30 minutes.

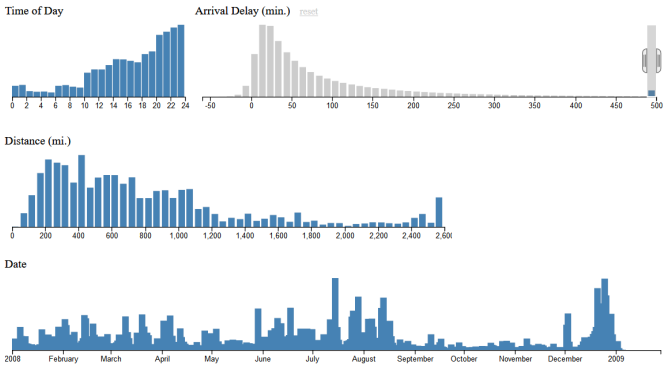

Figure 2: Busy periods

As a summary of this section, we can focus our attention to the more critical point (the holidays) to improve the Air Traffic Management. In the next section some tendency and behaviors will be treated and studied.

4 Tendency Hypothesis

4.1 Bar chart

In order, for us, to explore the dataset and determine if there are correlations between days and flights delay, we made some first steps in data visualizations with our dataset. This batch of simple charts aims also to introduce beginners and experts to tendencies and averages this dataset may offers. To determine if there is any company that was more often late in average, we made a bar chart where the horizontal axis contains US's carriers names, while the vertical axis represents the average delay in minutes. This chart differs from most common bar charts. Indeed, bar charts usually have different colors per columns to help any user to easily distinguish every column and thus extract wanted information faster. However, we chose to color those columns with a scale representing the traffic density (the more flights, the darker). This coloration allows its users to trustfully compare two carriers. As we are dealing with global average, some outliers element maybe swarm into the pool of data. As an example, a special carrier with only few flights may have an average which does not represent the reality as the cardinality of

flights is too low.
As discussed earlier, one of the limits of this chart is that with an average as main feature, it only shows information with a high level of abstraction and thus, the chart may contain some elements which can differ from what can be observed in reality. One space related issues is that this type of chart can only display so many information. As the number of columns grows up, it becomes harder to use this chart, because it may takes some time to find the wanted information, it is also difficult to compare some columns as they may have a consequent amount of bars between them. To resolve this issue, an interactive bar chart was made where you can select column and where there is a tool-tip available to help the user. Such visualizations may however miss direct the attention of a user as they may be unfamiliar with these features. Because we only have a limited number of columns, we chose not to implement any interactive features in order to remain the more neutral possible as this chart is for general purposes. Bar chart also can be difficult to apprehend if the categories names are too long. This situation can be avoided with spacing columns in order to display the names properly. Such method can only be used with few columns as the chart becomes quickly spacious with merely no information compared to empty space. Another solution is to put the names in diagonal to gain some horizontal space. However, those names becomes harder to read, the more vertical they are. We simply choose to use acronyms as many carriers sometimes refers to themselves like that and the US's carries names a distinct enough. Finally, due to the lack of any lines or marks to help the user distinguishing easily the actual value which represent the height of a bar, a user may have troubles producing precise study of this chart. Due to the purpose on those being only to have a global view of the dataset and extract tendencies, we chose to ignore this issue for two reasons. First, the lack of precision may lead the users to use this chart with a higher level of abstract than it is usually done with precise metrics. Secondly, markers may over crowd this chart and discourage users to use it as it becomes harder to understand. This situation can not be allowed in a visualization which is here to help understand the dataset.

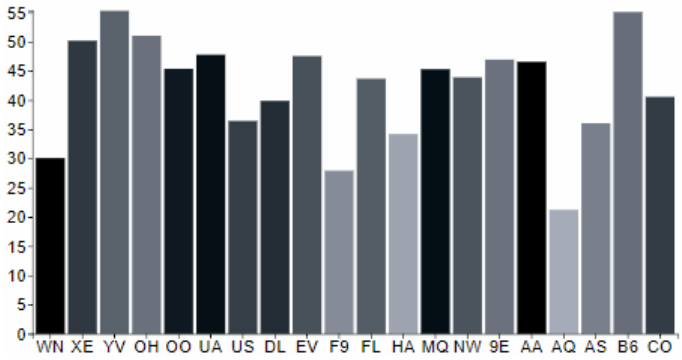

Figure 3: Average delay per carrier

4.2 Line chart

In order to understand the temporal distribution of delays and observe any kind of patterns in flights delays over the year of 2008, we have made a standard line chart with, as horizontal axis, the days of the year and vertical axis, the average delay in minutes. This visualization, even if it is far from representing the reality where any slight details may cause some delay (e.g lack of effective every Saturday at 9 pm) allowed us to put aside some theories about cumulative delay over days, and theories about correlation between days of delay. However, this visualization helped us extract some patterns. For example, we observe that the highest spike of delay occurs during summer, and that compared to the off-peak period, these months have only very few days with few delays and that these days always have more delays than days with few delays in off-peak periods. One point that can be improved is that we actually can not see the main cause of delays spikes, however, a tool-tip may help resolve this issue. A second limit of this chart is that it may be overwhelming at first glance, as there is a lot of spikes which make this chart hard to apprehend and, establish any information process such as extract tendencies. To solve this issue, we have thought about adding another line in this plot, which represents the global average per month. Such line may help a user in comparing days to the average and between themselves (using different colors for each line may help apprehend this chart).

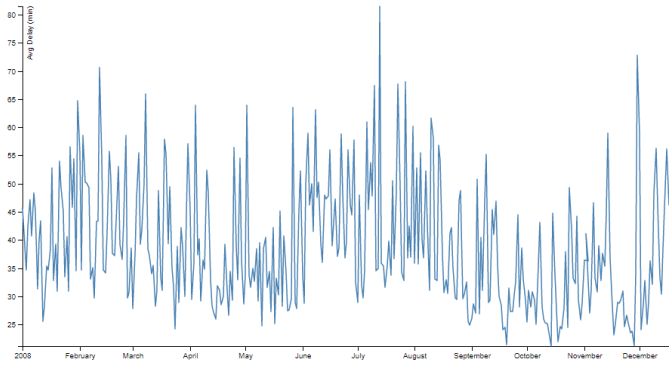

Figure 4: Average delay per day of the year

4.3 Hypothesis

Using the charts above (Fig3 and Fig4) we have established an hypothesis about a link between traffic density, vacation, and flight delays. In order to trustfully confirm this theory, we will have to compare basic result with significantly gap between two situations, as the higher the traffic density, the higher the delay. Indeed, cumulative metric can add a bias into comparative task as one minute delay repeated on severals flights in crowded days and two hundred minutes delay in one flight does not represent the same information. In order to extend our capacity to study our dataset, we have plotted the average delay and traffic density per day of the week.

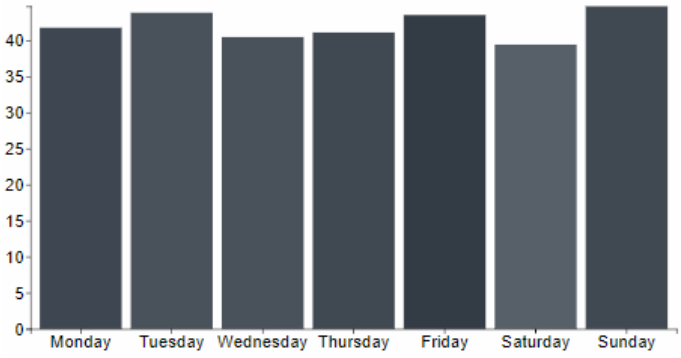

Figure 5: Average delay per day of the week

As we can see in the bar chart above, high delay days are friday, sunday and tuesday. However, the highest traffic day is friday, we may assume that this day traffic represent the population which travel for the weekend. Moreover, tuesday, a day with average traffic density and in the middle of week doesn't match our previously spoken hypothesis. The following section aims to provide tools to use this dataset and understand this outliers delays.

5 MULTIDIMENSIONAL BUBBLE CHART AGGREGATION

5.1 Presentation

As we want to supply a visualization on carriers to experts and customers, we need to show a comparison of US's carriers. From the point of view of a customer, this comparison has to point out which carrier has the less delay per flight. Indeed, they could see which carrier has the best rate of delay. They also see the main reasons about the carrier delay. Thus, customers would be able to choose the right carrier depending on their security policy. Indeed, safe carriers can choose to cancel a flight when the weather is to gloomy. On the other hand, for an expert, it has to highlight the reasons of the delay to ensure that he can understand and resolve the problem. The visualization should also take the time to plot data into account. It provides to users a way to assess the quality evolution of airlines in the time. The dataset we picked is a list of more than two millions flights across the USA. It shows for each flight the distance traveled, flight time, delay (mn) and all reasons about this delay.

There are five reasons of delay indexed in this dataset :

- Carrier Delay : is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, etc.
- Late Arrival Delay : at an airport is due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.
- NAS Delay : is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.
- Security Delay : is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.
- Weather Delay : is caused by extreme or hazardous weather conditions that are forecast or manifest themselves on point of departure, enroute, or on point of arrival.

5.2 Visualization

he figure 6 is a Bubble Chart where are plotted the mean delay for a high number of airlines in the USA during 2008. On the ordinate, you can see the mean flight time (mn) whereas there is the mean distance traveled (km) on abscissa during the period selected on the slider above.

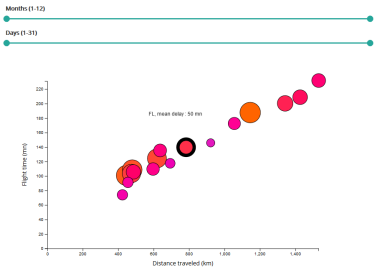

Figure 6: Bubble visualization of delay per carrier

The figure is a Bubble Chart where the mean delay for a high number of airlines in the USA during 2008 are plotted. On the ordinate, you can see the mean flight time (mn) whereas there is the mean distance traveled (km) on abscissa during the period selected on the slider above. The first goal of this visualization was to compare the flight statistic of airlines to see which one flies the most. In theory, bubbles would make a linear line on the graph and would be bigger when far from (0,0). In fact, the biggest bubbles are near on low flight times and distance traveled. It is partially explained because the delay is almost never caused during the flight time. Exploring data with this visualization allow to see that the delay is often caused by airlines before takeoff. Basically, the chart make the detection of tendency easier. Sliders above the visualization modify the display of bubbles. To illustrate, Figure 6 show data from January the 1^{st} to December the 31^{st} every day of months. If we want to see the first week of each month, we can move the range slider on values [1, 7]. It can let appear a delay tendency every first day of each month. As large bubbles are near each other, the chart can be messy. For this reason it is scalable.

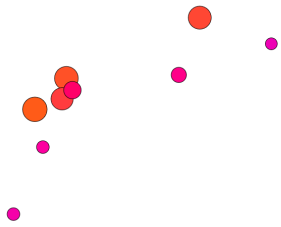

Figure 7: Zoomed bubble visualization

All radius circles are updated to fit into the screen and bring a clearer visualization of data. Their radius are bounded so they cannot be too limited neither to large and hide other circles. To ensure the draw of circles, we perform a bounded scale calculation on all of them.

Algorithm 1 scale radius

Require: *minin maxin minout maxout value*
  if *minin* = *maxin* **then**
    return *maxout*
  end if
  return (*maxout* - *minout*) * (*value* - *minin*) / (*maxin* - *minin*) + *minout*

Thus, this calculation allows to keep visibility even when the user zooms into the graph.

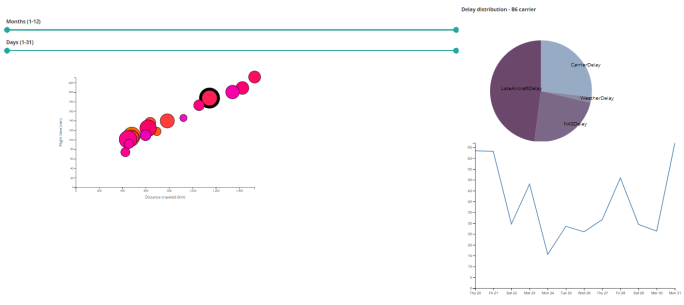

Figure 8: Reasons of delay

Moreover, we can have more information about an airline in a given time. Each circle is clickable and let appear two more information sources. The Pie chart shows reasons of delay as the few ones we explained above. This chart is useful to experts to

understand and search how to reduce delays. To illustrate, Figure 8 shows that all the delay from NW carrier is due to late aircraft delay. The Line chart helps taking some distance from aggregated data. It shows a global view and a tendency of the airline during the selected period. This tool is useful to extract patterns between carriers and find common reason to delay. For example, the Line graph could make appear some airlines with a large delay during one day. The Pie chart could explain this by a weather conditions.

6 Discussion

The radius of bubbles is a very delicate work. It is very hard to find the right way to display them. Because the mean delay is converted into a radius, circles have not a linear proportional area. Since the formula to calculate circle area is :

$$A = \pi.R.R \tag{1}$$

A carrier which has a twice higher mean delay than an other one will have a circle area twice larger. The area calculation is a problem in a visualization where user tries to assess delay from that. A better way to scale circle would be to take in account this problem and bound them as we already did. Indeed, the visualization will be preserved from bad visibility (too tiny or too large circle) and will not be victim of bias anymore.

7 Conclusion

To conclude, our visualizations allow to find knew knowledges about the dataset. Because of sliders of the Multidimensional Bubble Chart, we can filter over more than a million of data in order to point out tendencies and patterns across them. Our chart is as useful to customers as it is to airlines experts. It brings new elements to help them to make a choice on the less delay airline or to improve their own service. Innovations are not only presented on the visualizations, a large number of optimizations were made to deal with the dataset. Indeed, we had to work hard to optimize the display of charts. The cost for reading more than a million lines is excessive and be able to do it brings to our chart more relevancy to the users.

References

[1] B. Arnesen, J. Hwang, K. Karpack, and M. Ribera. Exploring flights: Visualizing big data.

[2] M. Bostock. crossfilter, 2017. [Online; accessed 10-01-2018].

[3] M. Bostock. D3.js, 2017. [Online; accessed 10-01-2018].

[4] A. Churchill, D. Lovell, and M. Ball. Flight delay propagation impact on strategic air traffic flow management. *Transportation Research Record: Journal of the Transportation Research Board*, (2177):105–113, 2010.

[5] J. Demuynck, D. Bosteels, M. De Paepe, C. Favre, J. May, and S. Verhelst. Recommendations for the new wltp cycle based on an analysis of vehicle emission measurements on nedc and cadc. *Energy Policy*, 49:234–242, 2012.

[6] U. M. Fayyad, A. Wierse, and G. G. Grinstein. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.

[7] B. C. B. Haarman, R. F. Riemersma-Van der Lek, W. A. Nolen, R. Mendes, H. A. Drexhage, and H. Burger. Feature-expression heat maps–a new visual method to explore complex associations between two variable sets. *Journal of biomedical informatics*, 53:156–161, 2015.

[8] J. Harper and M. Agrawala. Deconstructing and restyling d3 visualizations. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pp. 253–262. ACM, 2014.

[9] R. Hranac, J. Kwon, M. Bachmann, and K. Petty. Using marey graphs to visualize transit loading and schedule adherence. In *Transportation Research Board 90th Annual Meeting*, number 11-0350, 2011.

[10] C. Hurter, S. Conversy, D. Gianazza, and A. Telea. Interactive image-based information visualization for aircraft trajectory analysis. *Transportation Research Part C: Emerging Technologies*, 47:207–227, 2014.

[11] S. Inc. crossfilter, 2017. [Online; accessed 10-01-2018].

[12] C. Kelly, K. Craig, and M. Matthews. Real-time predictive analytics to estimate air traffic flow rates. In *Integrated Communication, Navigation, and Surveillance Conference (ICNS), 2015*, pp. N1–1. IEEE, 2015.

[13] L. Meskanen-Kundu. Making data accessible: An overview of interactive data visualization using d3. js as applied to a scientific dataset: Making a static visualization interactive. 2015.

[14] A. R. Odoni. The flow management problem in air traffic control. In *Flow control of congested networks*, pp. 269–288. Springer, 1987.

[15] I. Sadeh, I. Oya, J. Schwarz, and E. Pietriga. Prototyping the graphical user interface for the operator of the cherenkov telescope array. *arXiv preprint arXiv:1608.03595*, 2016.

[16] A. Sections. Statistical computing statistical graphics, 2009 url=.

[17] S. Sirisack and A. Grimvall. Visual detection of change points and trends using animated bubble charts. In *Environmental Monitoring*. InTech, 2011.

[18] Tableau. Tableau public, 2017. [Online; accessed 10-01-2018].

[19] C. Yu and Z.-C. He. Analysing the spatial-temporal characteristics of bus travel demand using the heat map. *Journal of Transport Geography*, 58:247–255, 2017.