VA of Aviation accidents in USA

Sigillo Luigi 1761017
Pocci Mattia 1688770
Martinelli Giuliano 1915652

05/07/2021

Contents

1 Introduction

In this report we will describe various visualization techniques realized to assist aircraft accidents visualization. We decided to focus on USA aircraft accidents because an analysis of this data can be useful to different kind of users that want to have insights about aircraft crashes.
In order to do that, we needed to find a way to represent information such that accidents could be easily plotted, analyzed and understood, supporting NTSB (National Transportation Safety Board) authorities in monitoring the situation and, if necessary, conceiving plan of actions.
Also pilots that want to buy a new aircraft can be helped in the decision by the different kind of visualizations: in particular the information on manufacturers (i.e. Fatalities, Weather Conditions, Death Rate, etc.) could be very useful when deciding which aircraft to buy.
Therefore, we realized four types of visualization:

- **Aircraft Crashes Map** is used to understand the geographic aircraft crashes distribution, identifying hotspots, and highlight states with different death rate and fatalities. Once a particular area has been detected, a more precise study can be carried on by analyzing the description in the upper right side of the page.
- **Parallel Coordinates** are specifically addressed to find patterns, analyzing every aircraft crash composing a subject (states/manufacturer/month/phase) over time. This information can be compared with the overall average values of the dataset entries.
- **Multidimensional Scaling Projection** was inserted to better visualize the relationships among the data points without having to compare each of them one by one. The possibility of having interaction between the user and the visualizations allows to better inspect some data and to give information about their relative differences.
- **Bubble Plot** has been designed and implemented in order to be highly customizable and flexible in displaying the data. In fact, the user can interact with this visualization, by modifying the axes, radius and the aggregation to be displayed, in order to highlight and extrapolate different characteristics from the data.

The code of the project is available `here`, moreover to try the system a live demo is available `here`.

2 Related work

Throughout the development of the project, to have an idea of which kinds of visualizations to use, we analyzed the literature, studying papers on both the aviation accidents domain and statistical data analysis.
For what concerns some of the employed views, we took inspiration from **Visualizing the FAA Aviation Accident Database** [1].
Here, the authors capture which aircraft manufacturers and models are involved when accidents take place. Their tool is mainly based on map view, line graphs and bar charts using Google Public Data Explorer. We took their idea to visualize an animation of the accidents over time on the map chart using a play button and a slider. Their map chart displays an animated view of the change in total fatalities for each state over time, indicated by bubbles of variable size. You can move the pointer over a bubble to get the state's name and its fatalities. Our approach instead took the map visualization and extended its capabilities: in fact, with our tool, we can decide which filter to apply on the data for displaying the desired information and we chose directly to colour the state (not involving any bubble in the view). For example, we can specify to display the information about the Death Rate associated to each state, and the map view will color them differently.
[1] employed an online visualization tool, Many Eyes, to examine relationships between the levels of damage (destroyed, substantial, minor, etc.) and the weather conditions.
We decided to focus our attention to the weather conditions (IMC, Instrumental Meteorogical Conditions or VMC, Visual Meteorogical Conditions) too. As described in one of the case studies, we decided to match up this info with the month of the flight using our bubble plot, in order to discover if there was some sort of correlation.
Fatal weather-related general aviation accidents in the United States [2] was of great inspiration for this approach in particular: in it, the authors correlated the accidents and the weather conditions using line charts.
For what regards the visualization of data over time, we took inspiration from **Analysis of trends in aviation maintenance risk: An empirical approach** [3] and decided to use a parallel coordinates, in this way we display an attribute's value (for example the number of fatal accidents) for each year.
From **Data and Information Visualization Methods, and Interactive Mechanisms: A Survey** [4] we took inspiration for the correspondent mouseon function of the parallel coordinates: in fact, in our plot the right line will be highlighted in green while the others will be left in grey.
We used **Accidents App** [5] in order to retrieve additional information about some particular flights, such as the 2011 accident of a NORTH AMERICAN vehicle, which is an outlier on our MDS as we report later on.
After reading this article on **Data Science Bowl** [6], we decided to focus our analysis also on the phase of

flight and the aircraft damage. In particular, as done in the article but with different visualization techniques, we were able to see the evolution over the years of the attributes per phase of flight (i.e. Fatalities, Death Rate, etc.).

3 Dataset

The **NTSB Aviation Accident Database** [7] contains information from 1962 on about civil aviation accidents within the United States, its territories and possessions, and in international waters.

3.1 Data Preprocessing

We decided to preprocess the dataset to shrink the dimensions and use only the useful data for our work. Initially we decided to drop the following columns: `'Accident Number','Airport Code', 'Airport Name','Registration Number','Engine Type','FAR Description','Schedule', 'Air Carrier','Report Status','Publication Date'`, because we thought they were not so meaningful for our intended analysis. Moreover, a lot of them contained sparse data (i.e. not every dataset entry had a value for those attributes).
We removed the incidents and focused only on the accidents, which are incidents associated with the operation of an aircraft which takes place between the time any person boards the aircraft with the intention of flight until such time as all such persons have disembarked. We took only the rows with the airplane damage. Furthermore we removed the crashes not related to airplanes (helicopters, balloon, etc.), and the amateur built airplanes. We removed the crashes without the weather conditions and the phase of flight. We took only the flights with a manufacturer associated. Furthermore we removed the airplanes that do not crash inside a USA state. We dropped the injury severity column because it can be derived by the Fatal value. We had to deal with the different names of same manufacturer or different declensions of the same name. We have also decided to create the column Month.
At the end of this preprocessing we ended up having only accidents occured after the year 2001. The final dataset is composed by 15 columns and 2632 rows. So respecting the AngeliniSantucci index with $AS = 39480$.

Figure 1: A picture of the dataset

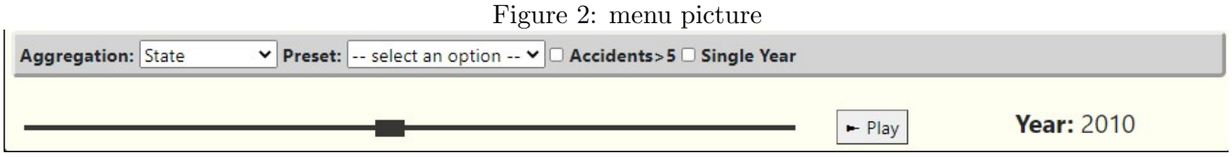
Event.Id	Event.Date	Location	Aircraft.Damage	Make	Purpose.of.Flight	Total.Fatal.In	Total.Serious.In	Total.Minor.I	Total.Uninjured	Weather.Condition	Broad.Phase.of.Flight	Crash.Location	Crash.Country	Event.Month
20200502X81	02/05/2020	PALMYRA, IL	Destroyed	Yakovlev	Personal	1.0	0.0	0.0	0.0	VMC	MANEUVER	39.408889,-89.9	Illinois	May
20200413X13	11/04/2020	Eagle River, A	Substantial	CESSNA	Personal	0.0	2.0	2.0	0.0	VMC	MANEUVER	61.35194399999	Alaska	April
20200326X82	25/03/2020	Waxahachie,	Substantial	CESSNA	Personal	0.0	1.0	0.0	0.0	VMC	STANDING	32.448055,-96.9	Texas	March
20200324X34	23/03/2020	Swansboro, I	Destroyed	Maule	Personal	2.0	0.0	0.0	0.0	IMC	UNKNOWN	34.416945,-77.0	Unknown	March
20200318X22	18/03/2020	Eagle Creek,	Destroyed	PIPER	Personal	0.0	1.0	0.0	0.0	VMC	TAKEOFF	45.352778,-122.	Oregon	March
20200317X10	16/03/2020	Pinedale, WY	Substantial	PIPER	Instructional	0.0	1.0	1.0	0.0	VMC	APPROACH	42.796944,-109.	Wyoming	March
20200313X13	13/03/2020	Sylmar, CA	Substantial	MOONEY	Personal	2.0	0.0	0.0	0.0	IMC	APPROACH	34.330556,-118.	California	March
20200311X62	11/03/2020	Sterling, MA	Substantial	CESSNA	Personal	1.0	0.0	0.0	0.0	VMC	TAKEOFF	42.430278,-71.7	Massachusetts	March

4 Visualization Techniques

4.1 Menu and Slider

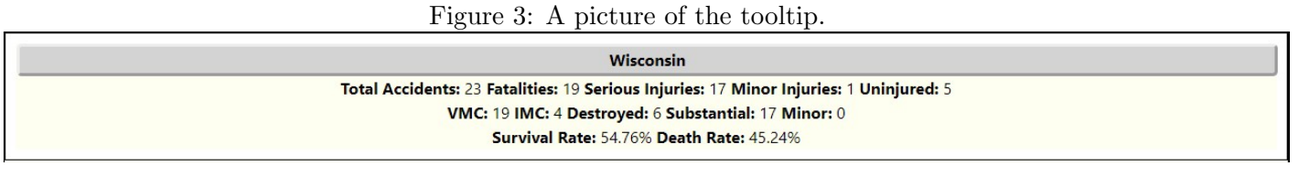
From the menu we can interact with all the visualizations and trigger runtime computations. Here we can change the type of aggregation between: State, Manufacturer, Phase of Flight or Month. This value will affect every visualization, by grouping at runtime the dataset tuples on the attribute of interest.
There are also some presets available to see immediately some insights. Then we have the checkbox options: the single year will select only the crashes of the current year, "Accidents>5" can remove the entries that do not have more than 5 accidents (a sort of outliers removal). The latter can be used to filter out data points with too few information. The slider permits to move across the years to see the change in the data, it is possible to move the slider in a manual way or we implemented a play button to see in a automatic way the data evolution. In order not to give an abrupt feedback to the user, we decided to update the visualizations when the user is dragging the slider.

Figure 2: menu picture



When we point to a state on the map or a line on the parallel coordinates or a point in the bubble or mds plot, we can see with a tooltip (showed below) the information about that particular point.

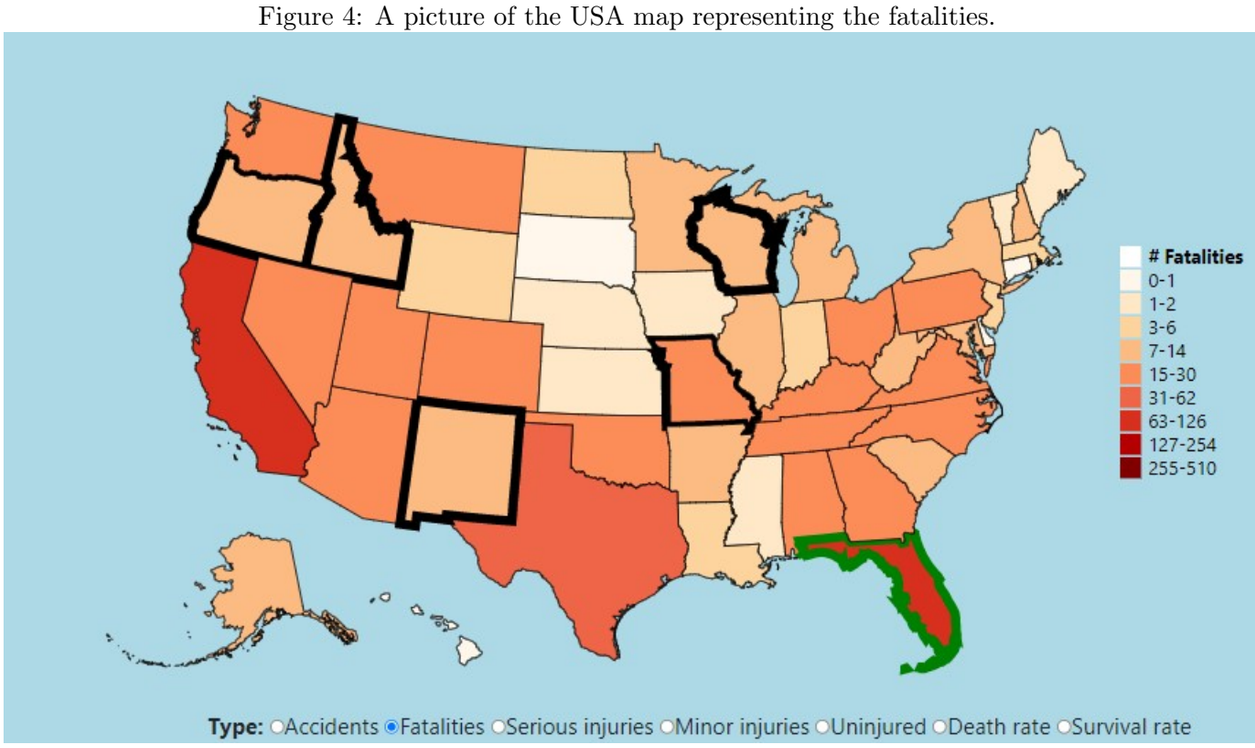
Figure 3: A picture of the tooltip.



4.2 USA Map

The map is implemented with a geojson. The USA map allows to visualize data through a color scale.
We used **colorBrewer2** [8], which permits to choose a precomputed sequential scale. Each color could represents the variation of these data on the single state of the USA: Accidents, Fatalities, Serious or Minor injuries, Uninjured and Death or Survival rate. Every state is coloured according to the values specified on the adjacent legend. The latter permits to understand the quantity of the selected attribute (such as Fatalities) for each state, given a color. The attribute to be visualized can be chosen from the menu below the map. If we hover the mouse on a particular state, we obtain its name, and trigger the tooltip that shows information about it (like the number of fatalities). The entire map is interactive with all others visualizations. In the image below we brushed the states in black and we are triggering the mouseon function on the Florida state that is in green.

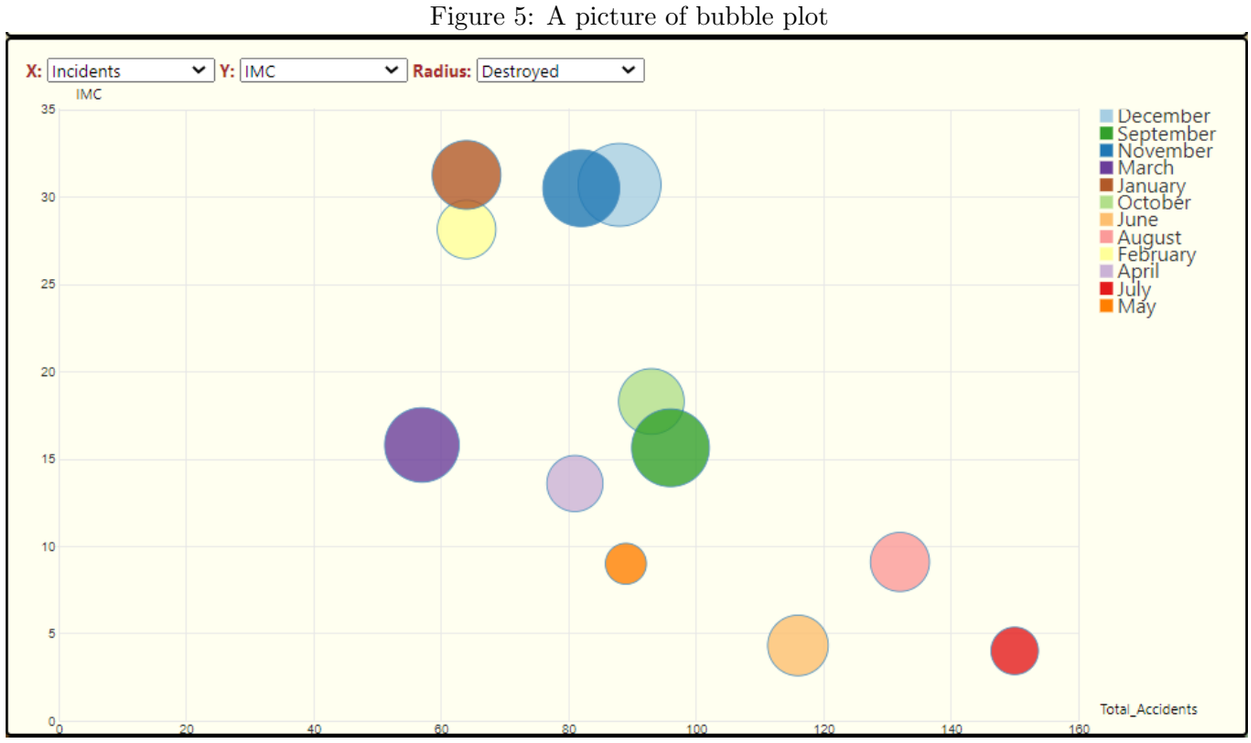
Figure 4: A picture of the USA map representing the fatalities.



4.3 Bubble plot

The Bubble Plot is a scatter plot with a third dimension added. In fact, apart from the usual X and Y dimensions, the radius of the bubble is used to represent another attribute of the object represented by the bubble itself. In the upper menu, we can decide the type of aggregation. Hence, depending on the aggregation type, one single bubble will represent a single Month/State/Manufacturer/Phase.
A peculiarity of this chart is that we can specify which attributes to assign to the three dimensions (X, Y and R) from a list of attributes (such as number of fatalities, number of injured, death rate percentage, etc.). Moreover, we decided to use a fixed color scheme (picked from **colorBrewer2** [8]) for the Month and the Phase aggregation, given that they contain a number of elements less than twelve. For the other aggregation types, we assigned a color to each single element given its name's hash. From our perspective this was the only way not to repeat the color schema, even if some elements have similar colors (which maybe are slightly different, but the difference cannot be noticed by the human eye).
On the right side of the plot, we can find its legend: a list of colored rectangles with the name of the respective element, to identify the correct bubble in the plot.
We implemented a mouseon function, which will highlight the right bubble (the one below the user's pointer): the name of the element will appear and the description of the element will become visible in the tooltip. The mouseon function can be triggered also by passing the mouse over a particular element in the legend. This last feature, in particular, is useful to identify a bubble in the visualization when the plot is extremely filled with bubbles.
Another useful feature we decided to implement is the brush function on this plot: if we brush one or more bubbles, they will be highlighted and the other ones will be reduced in opacity. Furthermore, when brushing, the correct elements in the legend will remain at full opacity and the other ones' opacity will decrease.

Figure 5: A picture of bubble plot



4.4 Multidimensional scaling

Multidimensional scaling (MDS) is a technique used for dimensionality reduction, allowing to summarize the original n-dimensional data in a lower k-dimensional component, easier to plot and to visualize. Multidimensional scaling uses dissimilarities between pairs of objects. We want to represent these dissimilarities as distances between points in a 2 dimensional space such that the distances correspond to the dissimilarities between those points.
In our case, we decided to use MDS to plot data points representing the differences among elements of the four aggregations (Month/State/Manufacturer/Phase) or single accidents. To understand the relationship among objects we generate a proximity matrix: a symmetric matrix which encodes this dissimilarity. The bigger $c_{[i][j]}$ is, the more i and j are different.
It is possible to decide the MDS Type to display:

- **Accidents**, with this filter the coefficients of the dissimilarity matrix will be computed using total accidents, fatalities, minor injuries and serious injuries attributes of accidents. Each element is an aggregation and it is scaled form 4 dimensions to 2 dimensions.
- **Accident Percentage**, this filter is similar to the Accidents Type, but in this case the dimensions are three (Fatalities, Minor Injuries and Serious Injuries). Every measure is divided by the number of accidents.
- **Typology**, this filter is useful to compare aggregation's type of accidents, which is calculated using 16 dimensions, involving weather, destruction of aircraft and phase of flight.

At this point the user has the ability to choose the preferred typology and aggregation. Once the matrix is computed, it can be given as input to the MDS algorithm which results can be projected into a scatter plot. Another important use of MDS is to display single flights.

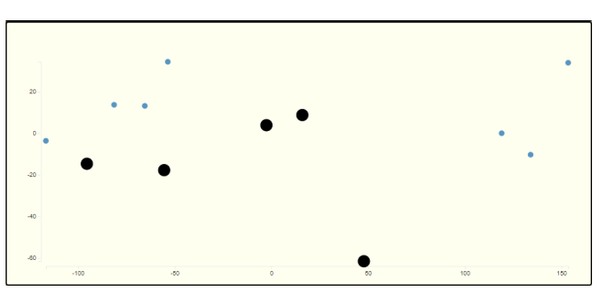


Figure 6: mds using brush

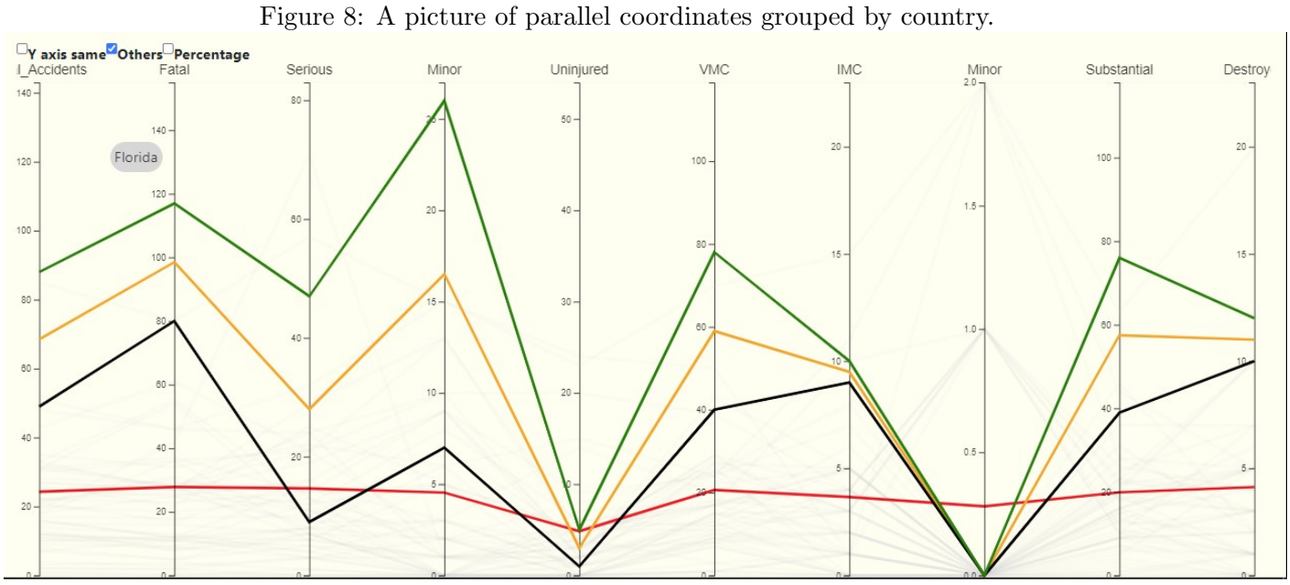
Figure 7: MDS single flights

4.5 Parallel Coordinates

Parallel coordinates is a technique for visualizing multidimensional data through parallel axes in a 2D chart. In our case, each axis could represent a year or an attribute of a crash, while the lines correspond to the aggregation type (like Manufacturer, State, etc.) intersecting each axis at its corresponding value.

This chart suffers from over-plotting, making it difficult to identify characteristics, trends or patterns. The brushing feature on the axis can filter only certain lines. It is possible to normalize the Y axis or to see the single data instead of the years on the Y axis. In addition we can choose to see the data in percentage, selecting the corresponding checkbox. The red line is the average of all the elements, while the orange line is the average of only the brushed elements, that are colored in black.
In the picture below we can see the Florida line in green because it is currently pointed by the mouse. When the mouse is on a line it provides the name of the corresponding state or aggregation selected in that moment, and a description in the tooltip.

Figure 8: A picture of parallel coordinates grouped by country.



4.6 Visualizations coordination and interaction

As previously said, the user has the ability to change various aspects of the displayed data. First of all, as said in menu section, through the specific menu it is possible to change lot of parameters. This results in a change of computation in all four visualizations. Concerning the coordination between views, they interact with one another.
The mouseon has the following effects on the views:

- On the **MDS** mouseon makes the point in the scatter plot bigger and green, making it distinguishable;
- On the **parallel coordinates** mouseon makes the line green and all the others grey and invisible;
- On the **bubble plot** mouseon makes the bubble more visible, by reducing the opacity of all the other bubbles;
- On the **map view** if the aggregation type is state, mouse on makes the state border thicker and green.

The brush has the following effects on the views:

- On the **MDS** brushing makes each selected element black;
- On the **parallel coordinates** brushing makes the correct lines black and all the others grey and invisible. At runtime, an average line for the brushed elements will be computed and displayed in orange in the plot;
- On the **bubble plot** brushing makes the brushed bubbles more visible, by reducing the opacity of the non-brushed bubbles;
- On the **map view** brushing makes the states border black and with a greater stroke width.

These two functions are not exclusive (i.e. you can trigger them at the same time), and the brush will remain if you change the year of the visualizations.

5 Case studies

5.1 Weather correlation with accidents

We can suppose that an NTSB agent could use the system to find insights about the weather correlation with the accidents. This agent could inspect the system, aggregating the data by month, he can observe that the majority of accidents takes place during summer. This will not surprise him, since the weather is more suitable to fly in these period of the year: in fact, these accidents have a low rate of IMC (Instrument Meteorogical Conditions).

As a consequence of the higher number of flights, the number of fatalities increases but the death rate is lower than the other months, and he can observed it when displaying the death rate on any of the visualizations. Furthermore, he can also observe that in this case also the number of destroyed is lower than in the other months. On the other hand, to consider the colder months, he can brush on them on the bubble, observing that the number of total accidents is lower in comparison with summer months, but on the parallel coordinates he sees that the brushed lines have a higher average death rate. He knows that during this period the overall number of flights decreases and if there is IMC weather a pilot is not allowed to fly. Another precious insight that he can extrapolate from this, is that there is no correlation between IMC weather and the number of accidents: in fact, when the weather is not appropriate, the number of accidents is significantly smaller, so it decreases accordingly.

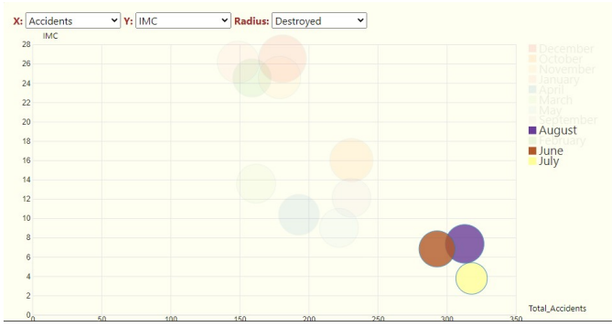


Figure 9: Bubble plot with summer monts

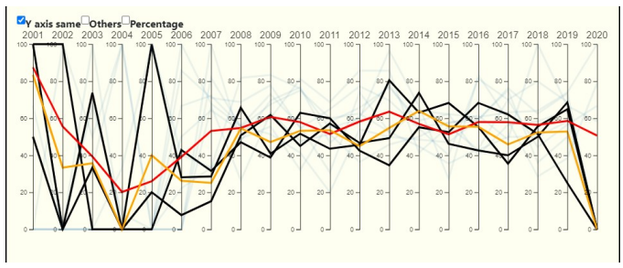


Figure 10: Parallel coordinates with summer months

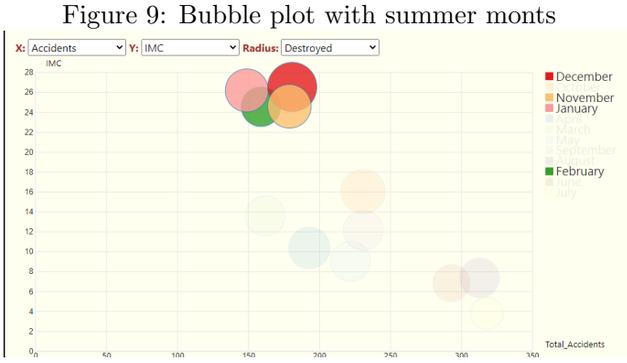


Figure 11: Bubble plot with winter months

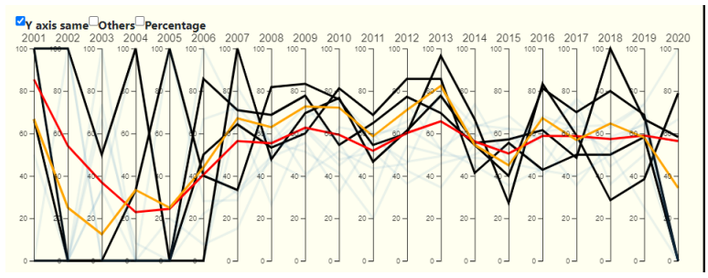
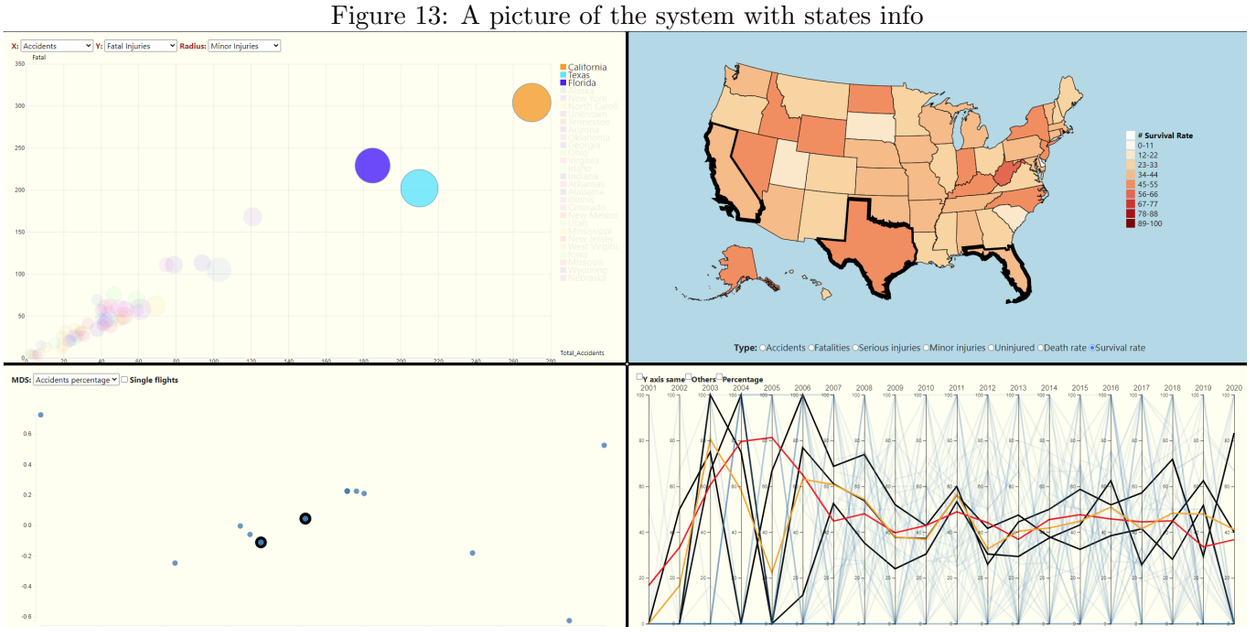


Figure 12: Parallel coordinates with winter months

5.2 States Info

Let's suppose that the same NTSB agent wants to see if there are states where the fatalities and the total accidents are higher than the average. At first glance, by observing our system and putting fatalities and accidents on the bubble plot he can see that California, Texas and Florida are outliers. He could think that the direct consequence could be that if people flies from one of these states, are more likely to be involved in a fatal accident. However, using the MDS chart in percentage, and filtering the map with survival rate, a deeper inspection reveals that the average survival rate of this places is higher than the survival rate of the whole United States. Therefore, the high fatality number is justified by the fact that in this states there are more flights than in the other ones, and naturally if the number of flights increases, the number of accidents increases.

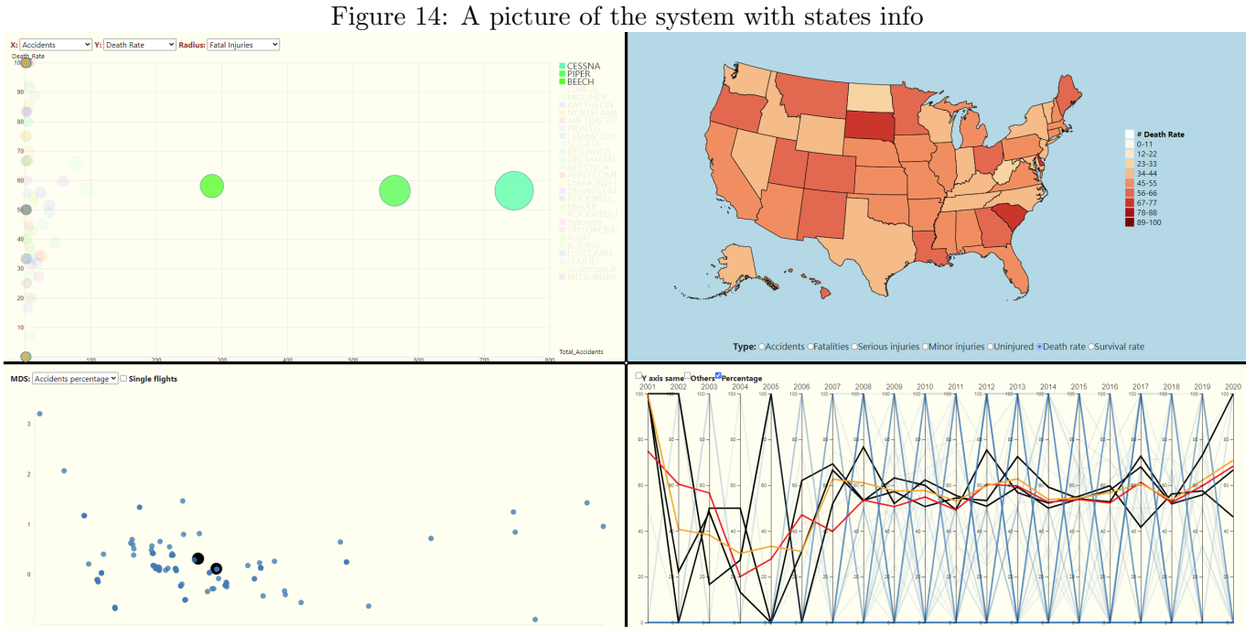
Figure 13: A picture of the system with states info



5.3 Big Manufacturers

Let's suppose that an aviator wants to buy a new airplane. Using the tool if he takes a quick inspection to the various visualizations, aggregating by manufacturer, he can see that BEECH, PIPER and CESSNA have the highest number of accidents. By brushing and focusing on these three manufacturers, he will discover an interesting fact: filtering by death rate the parallel chart and by analysing the MDS in percentage mode, he can see that they are not outliers. These three big manufacturers have more accidents simply because they are the most popular, in fact he can see that their average death rate is similar to the average death rate of all the manufacturers. There are other manufacturers (like PILATUS or SOCATA) with less accidents and higher death rate.

Figure 14: A picture of the system with states info



5.4 Phase Info

The same aviator is interested in seeing the phase that is the most dengerous. So he aggregates by Phase of flight, and can immediately observe that the majority of accidents takes place at takeoff. If he inspect further, he can see that the death rate at takeoff is less than 50%. An interesting conclusion can be expressed when he confronts the takeoff phase with the maneuver one: the latter, has less accidents (621 versus 762 of takeoff), but an higher death rate (66.91% versus 47.72%). The takeoff phase has less fatalities than the maneuver one, but more minor injuries: he can conclude that in an accident during takeoff, he is more likely to survive, while if the accident takes place at maneuver time, he is more likely to die.

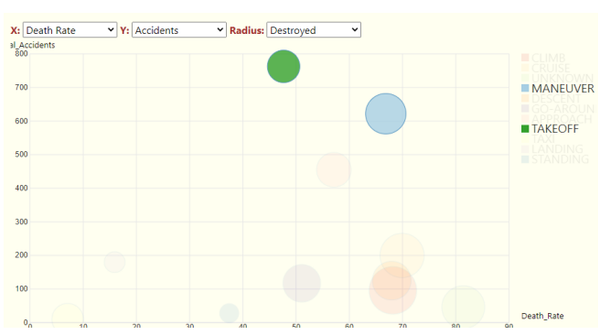


Figure 15: Bubble

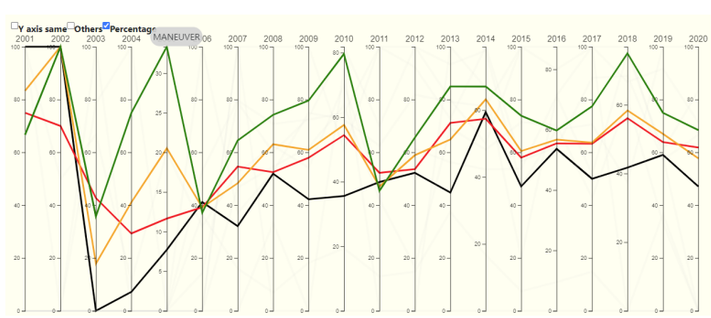
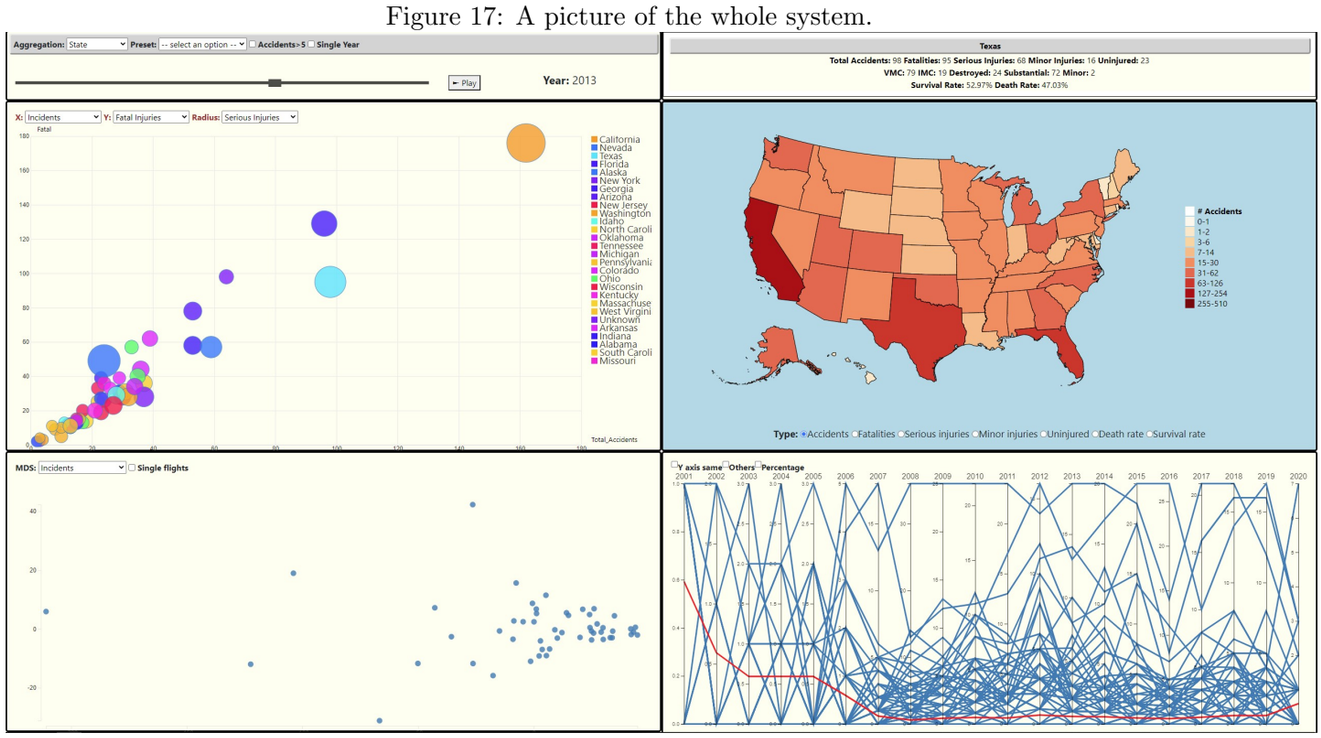


Figure 16: Parallel coordinates

6 Conclusion

We introduced this visual analytics tool to support the analysis of airplane crashes in the USA. This tool offers to a user a global view of all crashes in the USA from 2001 to 2020. He can investigate what is happening in a particular state, and see the dissimilarities with respect to others and the trend of that place for the type of information he is interested in, for example the fatalities. The work could be expanded including also USA cities or in general country outside the USA to have a more complete view on the airplane crashes. Furthermore in a future work we could include the aircraft model.

Figure 17: A picture of the whole system.



References

[1] T. Fox, Mary Ann Howell, Michael Senatore, and S. Varghese. Visualizing the faa aviation accident database. 2011.

[2] Andrew J. Fultz and Walker S. Ashley. Fatal weather-related general aviation accidents in the united states. *Physical Geography*, 37(5):291–312, 2016.

[3] Karen B. Marais and Matthew R. Robichaud. Analysis of trends in aviation maintenance risk: An empirical approach. *Reliability Engineering & System Safety*, 106:104–118, 2012.

[4] Muzammil Khan and Sarwar Shah Khan. Data and information visualization methods, and interactive mechanisms: A survey. *International Journal of Computer Applications*, 34(1):1–14, 2011.

[5] Miklos. Accidents app, 2021.

[6] Katherine Larson. Data analysis on aviation accidents. *Data Science Bowl*, 2017.

[7] NTSB. Aviation accident database & synopses, 2020.

[8] Mark Harrower and Cynthia Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal The*, 40:27–37, 06 2003.