Industrial Big Data Visualization: A Case Study Using Flight Data Recordings to Discover the Factors Affecting the Airplane Fuel Efficiency

Ming LI, Qin ZHOU SMIEEE
Accenture Labs
Beijing, China
ming.a.li@accenture.com

Abstract – **The dawn of the Industrial Internet era has shed light on the value of the big data associated with many industrial areas, such as aviation, resources, and manufacturing. Although the promising industrial big data could offer competitive advantages, it also poses profound challenges to the traditional analytics methods. The exploration and visualization capabilities are falling short for the fast-growing industrial data. Moreover, new value-adding insights need to be discovered in the fields where the domain experts have been consistently and scientifically improving. This paper introduces an industrial big data visualization case study using flight recorder data to discover the factors affecting the airplane fuel efficiency. The visualization challenge, the overall solution and the outcome are described in detail. In the meanwhile, the supporting methodology and tools for industrial data visualization are summarized into five components, including: agile data preparation, interactive exploration, large data visualization, statistical validation and report automation.**

Keywords – **Large Data Visualization; Industrial Analytics; Flight Recorder Data; Big Data Competition**

I. INTRODUCTION

Compared to the consumer domain, the big data associated with the industrial areas such as the jet engines holds more potential business value, and it has been stated that only one percent in cost reduction in jet fuel could save nearly $2 billion per year for the global commercial airline business [1]. However, since the times of commercial aviation, experts have been optimizing the aircraft design and operation to improve the fuel efficiency, raising the bar for the data scientists to mine new value-adding insights. Therefore, the industrial data analysis should attempt from different analytical perspectives, introduce the best practices in data science, and carry out solid validation for the found insights.

This paper introduces an industrial big data visualization case study based on a big data visualization competition hosted by CrowdANALYTIX to discover the factors affecting fuel efficiency using the Flight Recorder Data [2]. By exposing the problem to a wide audience, data competitions offer an effective way to reach the frontier of what is possible from a given dataset [3].

Big data visualizations have been successfully applied on large and complicated datasets in different industries, however many of them are realized using the domain specific or proprietary software. In this paper, the whole data visualization process has been implemented using the R language, leveraging the existing packages in the open source R eco-system [4]. In order to facilitate future industrial data analysis and visualization, the methodology which effectively helped in the data processing and visualization are summarized into five components: agile data preparation, interactive exploration, large data visualization, statistical validation and report automation. All the components have supporting tools in the form of R packages, so they seamlessly work together. The framework and methodology could be reused and further enhanced for quick exploration and experimentation in industrial data visualization.

This paper has been organized as follows: Section II describes the competition problem, identifies the challenges, and describes the overall approach. Section III describes the data preparation, analysis, visualization and validation in detail. Section IV is the conclusion for the case study and the suggestions of the future work.

II. THE CHALLENGES AND THE OVERALL APPROACH

In the visualization contest, the goal is to produce a visualization report explaining why some airplanes or flying instances are more fuel efficient than others. For example, to find what makes an airplane perform at higher levels of fuel efficiency during the difference phases of a flight (Taxi, Takeoff, Climb, Cruise, Approach, Rollout) and how can the flight data recordings be used to understand the drivers of fuel efficiency and derive the best practices that makes the flight fuel efficient under different conditions [2].

Multiple challenges could be encountered among the similar industrial data visualization projects. Firstly, as the industrial equipment creates large volumes of data, a heavy burden is placed on data preparation. Secondly, new insights need to be discovered beyond the existing knowledge, which requires multiple experimentations on the large data sets. Thirdly, the industrial sensor data values are usually fallen within particular ranges, which could lead to long rendering times and incomprehensible graphs with extremely dense data points. Fourthly, sometimes the patterns revealed in visualization could have happened by chance, therefore further analysis and verification are needed before applying them. Last but not the least, due to the criticality of many industries and the complexity of analysis procedure, reproducibility is highly required to make the review of the results and the whole procedure possible.

As show on the left side of Fig. 1, the overall approach includes data preparation, data exploration, fuel efficiency potential evaluation, fuel efficient analysis by flight phase, insights validation, and report generation. A series of reusable methods and tools are listed on the right side to help address the challenges in the visualization, including: agile data manipulation that supports quick experimentation on large

This data visualization competition was sponsored by Honeywell.

2324-9013/17 $31.00 © 2017 IEEE
DOI 10.1109/Trustcom/BigDataSE/ICESS.2017.322

data sets; lightweight interactive visualization web applications that can be built quickly for data exploration and familiarization; descriptive statistics gives a summary of fuel consumption by flight phases; large data visualization that could handle millions of data points which are challenging both perceptually and computationally; statistical validation that tells the statistically sound and actionable insights from the accidental patterns in the data; and report automation that ensures the reproducibility of the analysis.

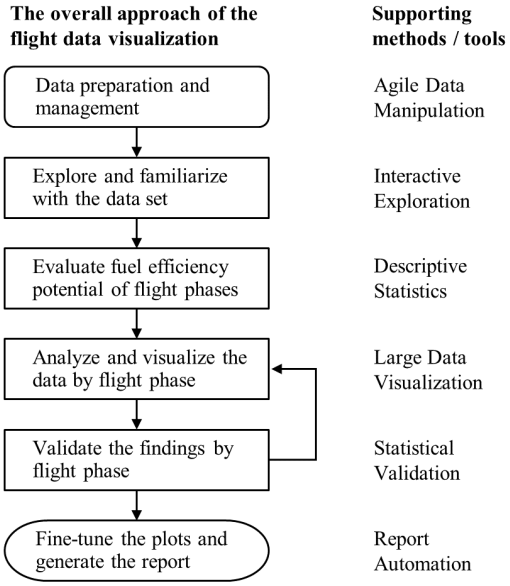


Figure 1. The overall approach and the supporting methods / tools

III. Flight Data Analysis and Visualization

A. Data preparation and management

In the competition, a data set containing 33 flight tail numbers are provided. Each flight tail is a different aircraft and has 25 flying instances.

In many data projects, data preparation is time consuming, and it could be repeated many times over the course of analysis as new problems and ideas come to light [5]. A few R packages, such as data.table [6] and dplyr [7] are employed to facilitate data preparation, efficiently reduce the programming and execution time. These tools also played an important role during later data analysis and experimentation.

The provided data come in 831 csv files and in a relatively large volume of 10.2GB, including readings of numerous sensors on aircrafts - detailed aircraft dynamics, system performance, and other engineering parameters, such as various altitude, speed, acceleration, and position measurements. The data are parsed and populated into an open source database [8]. Then, new features are derived, e.g. the departure and arrival locations and the traveled distance from the geo-coordinates [9], and the pre-flight Taxi Phases based on the flight phase sequences. The processed data set had a dimension of more than 200 columns and over 5 million rows. With the efficient R packages and an optimized workflow, the data preprocessing could be carried out on a mainstream laptop computer within several hours.

Moreover, external airport data are downloaded from an open data repository, and then the small and other types of airports were filtered out, leaving the large and medium types of airports for consideration [10]. Then the departure and arrival geo-locations of each flight instance are matched with the airports by the nearest distances. With the matched departure and arrival airports of all flight instances, the flight route and related airports can be visualized as in Fig. 2.

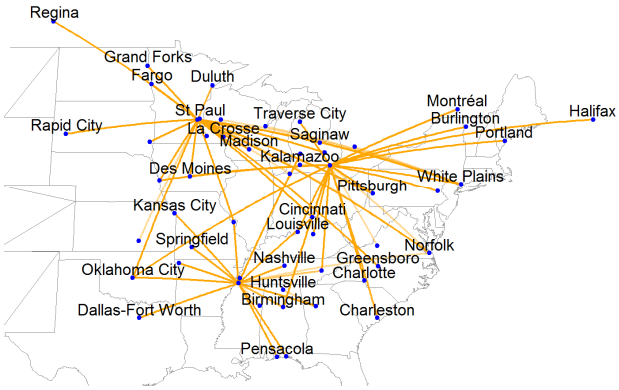


Figure 2. The flight route and related airports of the flight instances

B. Interactive flight data exploration

To effectively explore the flight data, an interactive visualization web application for the flight recorder data exploration and familiarization is implemented with the R package Shiny [11], and a time series charting R package dygraphs [12]. The interactive visualization app could read flight data from the database or separate csv files. Users could quickly browse time series data, easily zoom in and out by clicking or dragging, observe the dynamic changes during flight phases transition, compare multiple scaled variables side by side, calculate variable correlations and make the comparisons between flight instances.

C. Fuel consumption summary by flight phases

In this paper, the fuel consumption is calculated as the integration of the fuel flow. Table I is an overall summary of fuel consumption and fly duration by flight phase. Per the statistics, little fuel was consumed during the Unknown, Preflight and Rollout Phases, while the rest four flight phases consume more than 98% of the fuel, indicating greater potential for fuel conservation.

TABLE I. A Summary of Fuel Consumption by Flight Phase

Flight Phase	Fuel (lbs)	Duration (hours)	Fuel Ratio	Duration Ratio
Unknown	451	0.7	0.01%	0.05%
Preflight	88	10.9	0.00%	0.76%
Taxi	400,614	301.2	6.66%	20.89%
Takeoff	86,562	8.3	1.44%	0.57%
Climb	2,211,249	309.9	36.76%	21.49%
Cruise	2,458,985	501.4	40.88%	34.77%
Approach	850,040	304.4	14.13%	21.10%
Rollout	7,447	5.4	0.12%	0.37%

Therefore, we focused on the Taxi, Climb, Cruise and Approach Phases. As shown in the overall approach in Fig. 1, the visualization and validation steps were carried out on these phases respectively. Further analysis and various visualization techniques are applied to discover insights, moreover, regression and hypothesis testing are used to validate some of the findings.

D. The Taxi Phase

In the data set, a flight instance could have one or two Taxi Phases (a pre-flight and a post-flight). The duration and the fuel consumption during the Taxi Phase are summarized by flight instances. The result is shown in Fig. 3, where the color of each hexagon bin represents the number of flight instances with corresponding taxiing duration (x-axis) and fuel consumption (y-axis).

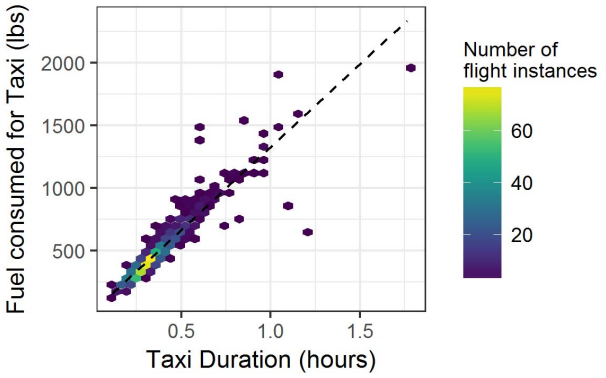


Figure 3. The time and fuel comsumed in the Taxi flight phase

Fig. 3 shows that the fuel consumed is linearly correlated to the taxi duration, which could be validated by a linear regression and a hypothesis testing on the regression coefficients. As the taxi time is usually controlled by the airport instead of the airplane itself, Fig. 4 is made to show the mean taxi time before the Takeoff Phase by airports. Due to the fair amount of fuel consumed during the Taxi phase, appreciable savings could be achieved by optimizing the taxiing operations of the airports with long taxiing times.

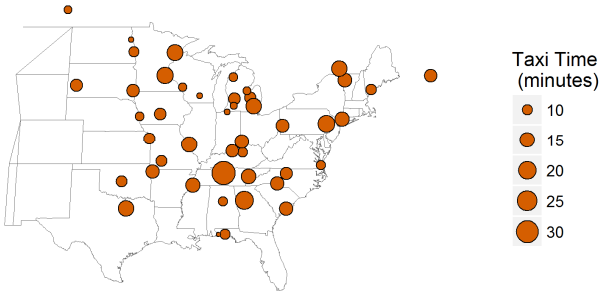


Figure 4. The mean pre-flight Taxi time by airport

E. The Climb Phase

In the Climb Phase, the airplanes behaved quite differently in terms of the flight trajectories. Fig. 5 gives a general understanding of the total fuel consumption for climbing. The Climb Phase of each flight instance is summarized into two variables: the total fuel consumption (y-axis), and the mean fuel flow rate (x-axis). A smooth line is added to show the relationship and histograms are made for the two variables respectively along the axes.

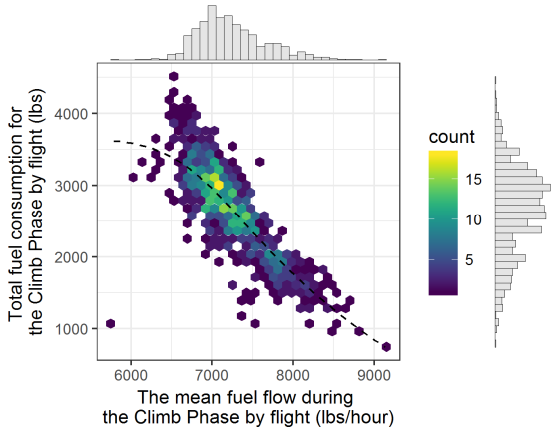


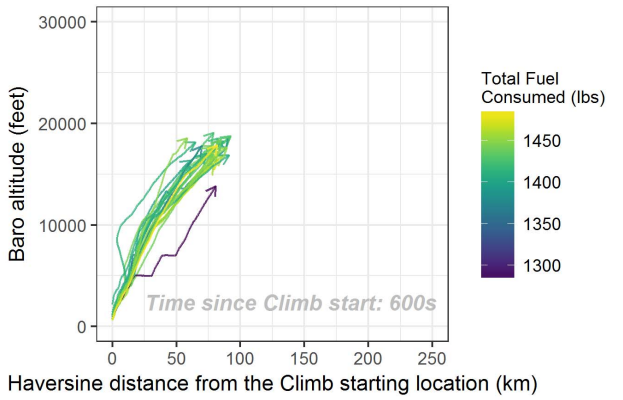
Figure 5. The total fuel consumption and mean fuel flow by flight intances during the Climb Phase

In one flight instance, climbing with higher fuel flow would result in less total fuel consumption in the Climb Phase. However, with lower fuel flow, the airplane may climb for longer time, and cover longer distance. Therefore, we may need to compare the climb trajectories, considering: time, altitude, total fuel consumption, and the distance traveled.

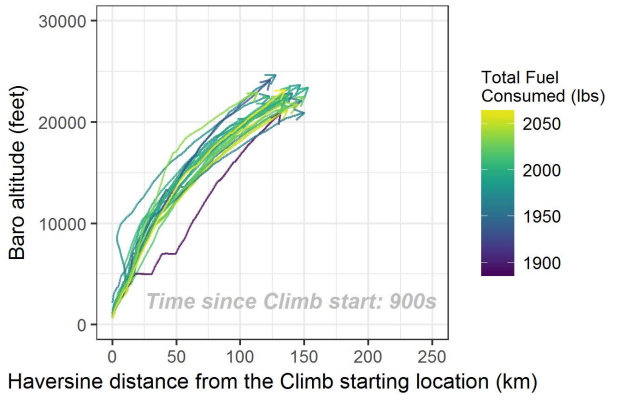
Data exploration revealed that different airplanes are different in the number of climbing phases and the final altitudes. Therefore, the climbing trajectories should be compared within one airplane or airplanes of the same model. The idea is to map the height, distance, and fuel consumption to multiple graphical elements in a visualization.

Fig. 6 (a), (b) and (c) compares multiple climbing trajectories at 600, 900 and 1200 seconds since taking-off. Each line in the plot represents a flight instance, the y-axis shows the heights climbed, the x-axis shows the Haversine distance traveled, the line color shows the total fuel consumption since the Climb Phase started.

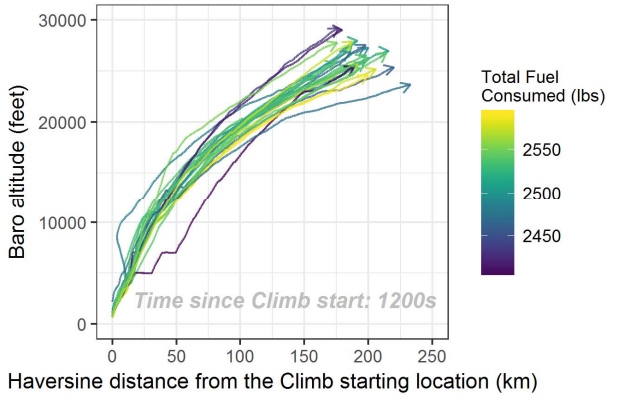
Moreover, multiple snapshots can be taken to represent the relative time since the climbing started, and a series snapshots can be made to generate an animation. Animations can be displayed in an R Shiny web application [11] or embedded in the pdf report file using the R package animation [13]. With the visualization or animation, the flight path planner could investigate the height, distance, fuel consumption and flight time, take multiple business metrics into consideration, and choose an optimum from historical data for future reference.



(a)



(b)



(c)

Figure 6. Comparing multiple climbing trajectories of one airplane

F. The Cruise Phase

The provided data set contains more than 1.8 million records of the Cruise Phase in total, and about two thirds records were in the dense area where the altitudes were between 27500 and 35500 feet and fuel flows were between 3000 and 6000 lbs/hour.

The high density is challenging both perceptually and computationally for visualization, resulting in long rendering times and incomprehensible graphs. The "bin-summarize-smooth strategy" proposed in [14] is applied to handle this situation. The overall result is shown in Fig. 7, and the densest area are magnified as in Fig. 8. Both figures differentiate the densities and can be rendered in a few seconds. The visualization tells that the altitudes were discretely set to different altitudes by thousand feet during the Cruise Phase. The results show that the fuel flow in Cruise phase decreases with the altitude.

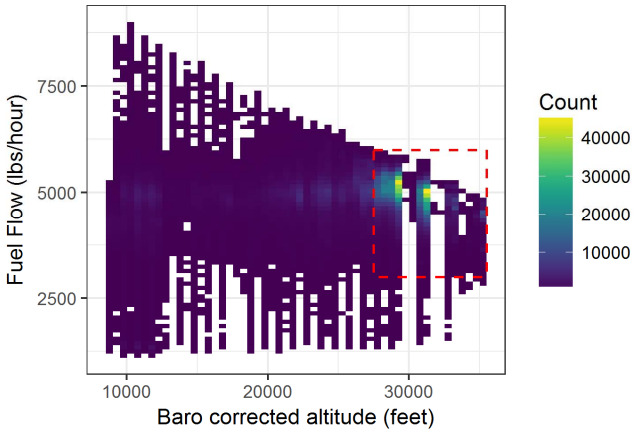


Figure 7. The altitude and the fuel flow in the Cruise Phase

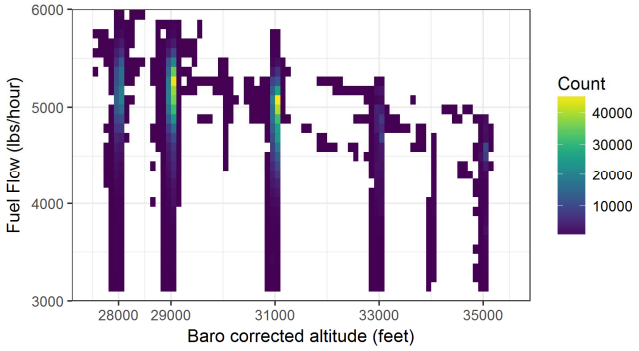


Figure 8. The altitude and the fuel flow in the Cruise Phase (magnified)

To verify the insights discovered in the visualizations, Wilcoxon tests are carried out to prove that there are true differences in fuel flow rate between neighboring altitude bands, and Table II shows the test results that the fuel flowing rates when cruising at lower altitude band is greater than those at the higher altitude bands.

TABLE II. THE WILCOXON TEST RESULTS

Height 1	Height 2	p. value	Alternative
28000 feet	29000 feet	0	Greater
29000 feet	31000 feet	0	Greater
31000 feet	33000 feet	0	Greater
33000 feet	35000 feet	0	Greater

G. The Approach Phase

During the Approach Phase, the altitude and speed of a plane is gradually decreasing in order to reach the optimal

values for landing. In the due course, the potential energy and the kinetic energy is dissipated as heat due to the friction, which saves of a portion of the jet engine power output.

Made from all the Approach Phase records (~1.1 million), Fig. 9 shows that the ground speed graduate decreases when the airplane is about to land. However, some records show that the airplanes were approaching at relative high speed at low altitude, which may worth some investigation if additional data were available.

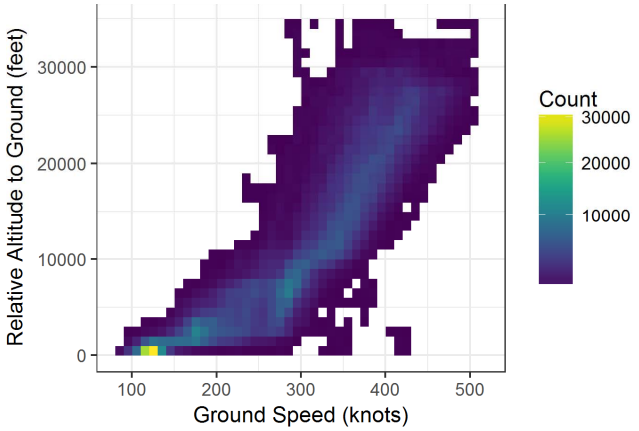
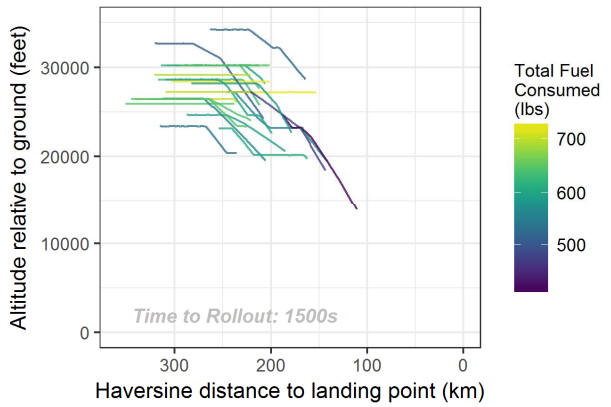


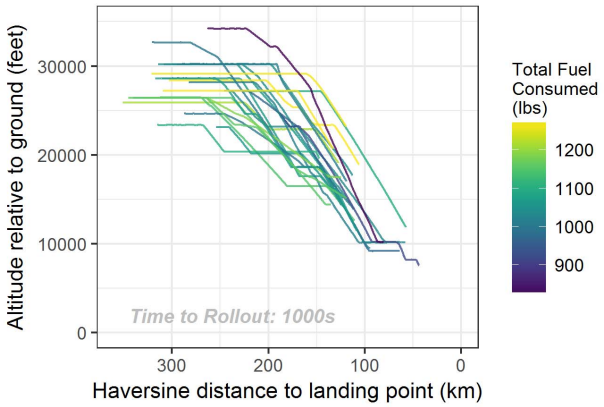
Figure 9. The ground speed and the relative altitude to ground in the Approach Phase

Fig. 10 (a), (b), and (c) are produced with the similar techniques as Fig. 6. The relative times to the Approach Phase ended were set to 1500, 1000 and 0 seconds, respectively.

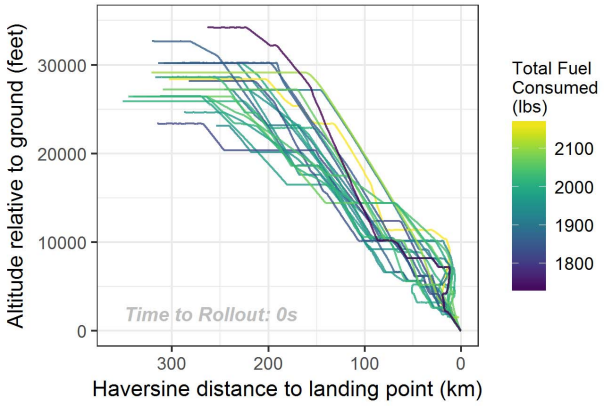
From the visualization, several patterns could be observed: 1) Descending from higher altitudes will consume less fuel due to the higher potential energy. High altitude also means more fuel were consumed in the Climb Phase, and less fuel consumed in the main Cruise Phase previously. 2) In a short period, descending fast consumes less fuel. 3) Cruising generally consumes more fuel than descending. 4) Making turns (yawing) during approach consumes more fuel. 5) In a holistic view, a smoother descending trajectory may lead to less fuel consumption.



(a)



(b)



(c)

Figure 10. Comparing multiple approaching trajectories of one airplane

H. Chart and Report Generation

The competition required a visualization report in pdf format. For high quality graphics, most charts could be produced with the pdf graphics device in R. Moreover, animations were embedded in the final report to show the dynamic changes of the Climb Phase and the Approach Phase over time.

The reproducible research methodology is applied throughout the analysis. The report is written in an R markdown file [15] and the report is generated in RStudio [16] automatically, making the convenient review of the whole process possible, and facilitating the report update.

IV. Conclusion

Visualization are made upon large volumes of flight recorder data to discover the insights on the factors affecting the airplane fuel efficiency. The whole process for the visualization are described, and several insights about the fuel consumption have been discovered, including:

- The Taxi, Climb, Cruise and Approach Phases have higher fuel efficiency potentials.
- In the Taxi Phase, the taxiing time and fuel consumption exhibit linear correlation, and there are differences among airports on pre-flight taxi time, which may need more data for the root cause analysis.

- In the Climb Phase and the Approach Phase, multi-dimensional visualization with animation revealed different trajectory patterns overtime, considering: altitude, distance, and fuel consumption, which could be useful reference for the path planning.
- In the Cruise Phase, the cruising altitudes were set to different altitudes discretely. Visualizations from the large and dense data set show that the fuel flow rate has negative relationship with the altitude, and the findings are validated by the hypothesis testing. The "bin-summarize-smooth strategy" proposed in [14] has been proved to be efficient for such situation.

With additional data, such as the flight load being available, further analysis and visualization could be made for both the report and the web application to meet the more specific needs of the possible users, e.g. the operators or the planners.

As the industrial data grow ever larger, it's important that our capability to visualize grows correspondingly. This paper summarizes the challenges in industrial data visualization, and proposed the supporting methods and tools. The current toolbox includes many helpful R packages that work together seamlessly to boost the efficiency of the data manipulation, analysis and visualization. These packages could be grouped by their roles played in the visualization process, including: agile data preparation, interactive exploration, large data visualization, statistical validation and report automation. In future, an industrial data visualization framework could be built after more case studies being conducted and additional components, such as high dimensional data visualization, being integrated.

ACKNOWLEDGMENT

We thank Honeywell for the approval of publishing this paper. We also thank Mohan Singh from CrowdANALYTIX for his support.

REFERENCES

[1] Evans, Peter C., and Marco Annunziata. 2012. "Industrial Internet: Pushing the Boundaries of Minds and Machines." [online] Available at: https://www.ge.com/docs/chapters/Industrial_Internet.pdf [Accessed 17 May 2017].

[2] Crowdanalytix.com. 2016. What makes airplanes fuel efficient?. [online] Available at: https://www.crowdanalytix.com/contests/what-makes-airplanes-fuel-efficient- [Accessed 17 May 2017].

[3] A. Goldbloom, "Data Prediction Competitions -- Far More than Just a Bit of Fun," *2010 IEEE International Conference on Data Mining Workshops*, Sydney, NSW, 2010, pp. 1385-1386.

[4] R Core Team. 2016a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. [online] Available at: https://www.R-project.org/ [Accessed 17 May 2017].

[5] H. Wickham, "Tidy Data," Journal of Statistical Software, vol. 59, no. 10, 2014.

[6] Dowle, M, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan. 2015. *Data.table: Extension of Data.frame*. [online] Available at: https://CRAN.R-project.org/package=data.table [Accessed 17 May 2017].

[7] Wickham, Hadley, and Romain Francois. 2016. *Dplyr: A Grammar of Data Manipulation*. [online] Available at: https://CRAN.R-project.org/package=dplyr [Accessed 17 May 2017].

[8] Wickham, Hadley, David A. James, and Seth Falcon. 2014. *RSQLite: SQLite Interface for R*. [online] Available at: https://CRAN.R-project.org/package=RSQLite [Accessed 17 May 2017].

[9] Hijmans, Robert J. 2016a. *Geosphere: Spherical Trigonometry*. [online] Available at: https://CRAN.R-project.org/package=geosphere [Accessed 17 May 2017].

[10] data.okfn.org. 2016. "List of Airport Codes, Locations and Other Information Around the World." *GitHub Repository*. [online] Available at: https://github.com/datasets/airport-codes [Accessed 17 May 2017].

[11] Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2016. *Shiny: Web Application Framework for R*. [online] Available at: http://shiny.rstudio.com [Accessed 17 May 2017].

[12] Vanderkam, Dan, JJ Allaire, Jonathan Owen, Daniel Gromer, Petr Shevtsov, and Benoit Thieurmel. 2016. *Dygraphs: Interface to 'Dygraphs' Interactive Time Series Charting Library*. [online] Available at: https://CRAN.R-project.org/package=dygraphs [Accessed 17 May 2017].

[13] Xie, Yihui, Christian Mueller, Lijia Yu, and Weicheng Zhu. 2015. *Animation: A Gallery of Animations in Statistics and Utilities to Create Animations*. [online] Available at: http://yihui.name/animation [Accessed 17 May 2017].

[14] Wickham, Hadley. 2013. "Bin-Summarise-Smooth: A Framework for Visualising Large Data." [online] Available at: http://vita.had.co.nz/papers/bigvis.pdf [Accessed 17 May 2017].

[15] JJ Allaire, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman and Ruben Arslan. 2016. *Rmarkdown: Dynamic Documents for R*. [online] Available at: http://rmarkdown.rstudio.com [Accessed 17 May 2017].

[16] RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. [online] Available at: http://www.rstudio.com/ [Accessed 17 May 2017].

Ming Li received his M.S. degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China in 2007. He is an associate manager of Industrial Analytics R&D in Accenture Labs. His research interests include industrial data analysis, machine learning and smart grid technologies.

Qin Zhou (M'92, SM'04) received his B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China in 1983 and 1986, respectively, and his Ph.D. degree in electrical engineering from Iowa State University, Ames, IA, USA in 1992. He is currently the Fellow and director of Smart Grid R&D, Accenture Labs. His research interests include energy big data anlytics, smart grid, and power system operation and optimization.