# Analysis of Aggregate USAF Bombing Runs: THOR Dataset (1964-1975)

Jarret Flack, Jordan Peterson, Samuel Kaessner

*CS 455: Introduction to Distributed Systems*

*Colorado State University Spring 2018*

*Website: http://www.cs.colostate.edu/~cs455*

*INSTRUCTOR: Shrideep Pallickara*

## I. Introduction

As today's military environment becomes increasingly complex, we must look to data analytics in order to make educated decisions when taking action. We believe this can be done through machine learning and comprehensive data visualization. By examining the USAF (United States Air Force) kinetic and non-kinetic bombing trends from the Vietnam War, a recommendation system can be created in order to aid our military strategists in the present. Though the equipment and munitions may be out of date for modern interactions, the techniques used for this research are adaptable for current situations. For this experimentation, we are using the THOR dataset of about 5 million Vietnam bombing runs. The paper is split into seven sections: (I) Problem Characterization, (II) Dominant Approaches to the Problem, (III) Methodology, (IV) Experimental Benchmarks, (V) Insights Gleaned, (VI) Transformation of the Problem Space for the Future, and (VII) Conclusion. This ordering of sections provides a logical flow of conception of the problem to the testing and results of the experimentation. In these sections we highlight the process of using Apache Spark in order to process, analyze, train machine learning models, and graphically represent the dataset aforementioned. There are two experiments that are described in this report, the first being a resource recommendation system and the second being a data visualization system in order to aid pattern recognition. This is made possible through an online tool called FusionTables. FusionTables allows us to show the data in terms of scatter plots and heatmaps. As a result of this visualization framework we were able to represent trends of military strategies over the length of the war by mapping location of events with respect to time. We were also able to infer the types of resources best used with respect to the terrain and target of the mission. For the second experiment we created a machine learning and data mining driven recommendation system. This system took advantage of Apache Spark machine learning library Mlib. Using this technique yielded an application that will recommend the most likely used resources used for a user defined mission with respect to the Vietnam War. The data generated by this system may be useful in understanding the mission planning decisions made in the past. Therefore, we believe this technology will also have a great effect on current military planning when implemented for modern mission datasets. Though these data analytic models were used for military applications, the experiments performed in this report may reach further into the field of environmental impact research and manufacturer production decisions. By looking at graphical representations of airstrikes by type of munition, one would be able to compare long term effects of the different airstrikes on the surrounding areas. Furthermore, the environmental data gained through these analysis can be used to develop more environmentally conscious strategies and munitions. In regards to aircraft and munitions manufacturers, they may be able to use our recommendation system in order to analyse what situations their products were used in. This data would be useful in improving the quality of their products by tailoring them to how they were used in military operations.

## II. Problem Characterization

This section proposes the challenges involved when performing these two experiments. The next section will provide the solutions to the challenges explored here. When dealing with multiple gigabytes of data the first challenge is processing the information in a timely matter. This was a challenge because we had to move through the whole dataset and look at each record individually in order to scrape specific needed fields. Once the necessary data had been filtered out we passed it to our two experiments. There are two subsections presented for this section: A. "Challenges of Designing and Implementing a Recommendation System" and B. "Challenges of Producing Graphical Representations". Each section will discuss the challenges that were specific to the respective experiments:

### A. Challenges of Designing and Implementing a Recommendation System

The first challenge of creating a machine learning model was converting the fields taken from the full dataset and converting them into a machine learning friendly format. This requires taking categorical data and mapping them to quantitative values. For example, taking values “250LB M-57”, “M-3”, and “750LB BLU-1/27” then mapping them to 0,1 and 2. Next we had to create a machine learning model that will be able to differentiate between different mission types and their specific attributes. After the model was created we needed to train it with the bombing run data. This involved splitting up the data and feeding 70% of it through the model. Once the model was trained we had to validate the the model with the other 30% of the data.

### B. Challenges of Producing Graphical Representations

The first challenge of visualizing the data was filtering the desired attributes from the dataset. Not only did we want only specific attributes from our dataset, but we needed a way to quickly produce these visualizations without having to modify the source code every iteration. On top of this our dataset has some incomplete data, so it was a challenge to parse the data of this incomplete data.

## III. Dominant Approaches

This section gives a detailed description and analysis of other works that have used the THOR Vietnam War bombing run dataset. We split this section into two groups. The majority of works performed on this data set has been data visualization or basic data analytics. For this reason, we split the review of past work into three sections: A. “Data Visualization Studies”, B. “Data Analytic Studies”, and C. “Other”. In these studies range from plotted graphs comparing two attributes of the dataset to geolocational plots of bombing runs with respect to time. None of these previously conducted studies observe the location of mission targets with respect to aircraft type, munition type or mission type. For these reasons we believe our work is substantially different from previous visualization work created through this dataset. The following is our review of previous works:

### A. Data Visualization Studies

In the graphic labeled “Number of Aerial Bombardment Missions by Military Vietnam 1965-1975 (Monthly Totals)”, Noah Rippner expresses a multiple line graph of number of missions with each line depicting the military in charge of the bombing runs. The graph shows that the United States of America dwarfs all the other militaries in number of bombing missions.[1]

In the graphic labeled “Mission Flight Hours by Time of Day Viet Nam 1965-1975”, Noah Rippner compares the flight ours of each branch of the US military and Vietnam Air Force ordered by time of day. From this we can see that the US Air Force has the most flight hours during the day and tied with the US Navy for night time.[2]

In the animated image “Visualization of 2.9 million air missions over Vietnam ‘65-’73”, the author depicts the location of all 2.9 million bombings with respect to time.[3]

In the graph “Number of Explosives Dropped by Military Vietnam War 1965- 1975 (Monthly Totals)”, Noah Rippner shows the total number of explosives dropped ordered by nation. The United States dropped far more bombs in the span of the Vietnam War.[4]

The graphic labeled “Bombing Missions of the Vietnam War” is a map of Vietnam with a plot map of all bombing runs in the dataset.[5]

### B. Data Analytic Studies

The image labeled “Counts of the descriptions of the 'KINETIC' missions” is a list of how many times the each kinetic type of bombing run appeared in the dataset. This data is ordered by most to least.[6]

The image labeled “Counts of the descriptions of the 'NONKINETIC' missions” is a list of how many times the each non-kinetic type of bombing run appeared in the dataset. This data is ordered by most to least.[7]

## IV. Methodology

This section describes the process that was taken for each experiment. For both tasks, there were common steps that were necessary for setting up the clustered framework. First we implemented a HDFS (Hadoop Distributed File System). We followed an installation and configuration guide provided by the CSU CS 455 Distributed Systems course [8].Once set up, we loaded the THOR Vietnam War bombing run data onto the HDFS. Then we used an Apache Spark setup guide provided by the CSU CS 535 Big Data course in order to set up our main processing utility[9]. The next step was to learn the main components of writing Scala programs for Apache Spark applications. A solid starting point for us was reviewing a scala programming guide provided by Apache Spark official documentation [10]. This document was out of

date, but was still useful for us to get a general understanding of programming with Scala and Spark. Following this we developed the functionality to load in a csv file and parse through each of its records. We selected specific values from the records and stored them in a DataFrame object. This object gives us an easy interface for storing and retrieving the data with sql-like statements. From this point on the methodologies for both experiments diverge into the subsections below:

### A. Methodology for Machine Learning Based Recommendation System

After we got the cluster running, the next step was to train a model that could predict the airplane used for a particular bombing run. We used scala in conjunction with Spark's MLLib libraries. We first load the data into a dataframe, and filter out all rows that don't have the features that we need to train on. These features include the target type, weapon type, weight of the weapon, period of day, the country that launched the attack, flight time, and lastly the flight hours. Once we have filtered out all the problematic rows, we then assign unique numbers to each distinct type in each column. We do this because the machine learning libraries only operate on numbers, and our data is all string type.

Once that was done, we set up a pipeline that assembles all the results into feature vectors, which is what our multi-layer classifier operates on. This multi-layer classifier is essentially a neural network, set up to do classification. We then split the data into testing and training sets, using 70% for the train set, and 30% for the test set. This split is done by randomly shuffling the data before partitioning it, so that we don't end up testing on a subset of the data that is significantly different from the training subset. Then, we train the model on the training data, and write out the model to the disk. This allows us to quickly make predictions in the future, because we don't have to re-train the model each time we want to generate a new prediction, we just load the pre-trained model off the disk (a few megabytes at most).

To make a prediction, we load the model, as well as an input file that has rows containing the features we use to predict the type of airplane. A big advantage of using a CSV file as input instead of a textbox input or similar is the fact that we can make many predictions quickly. This input model is loaded and transformed by the same model we saved earlier, and then the predictions column is written out as a file.

### B. Methodology for Visualization of Data Trends

To visualize our data we choose Scala as our language to work in for the data visualization, the reason being that most of the data visualization libraries that we found used Scala. Using Spark SQL, we created a program to run dynamically generated queries on our dataset. This approached allowed us to visualize different types of aircraft and munitions very quickly, we were also able to visualize any quality of our data relative to that strikes' location. We then took the result from the query and parsed the data to for visualization. We ended up using a tool known as FusionTables to visualize the data in an interactive format. Even though the visualization with FusionTables had great results, we originally planned to use a third-party solution called GeoSpark. The plan was to utilize a special flavor of RDD in the GeoSpark API to organize our data spatitialy, this spatial RDD would then be plotted on to a PNG heatmap. However, we were unable to get GeoSpark to work properly, the API was not documented appropriately and it had many internal memory management issues.

## V. Experimental Benchmarks

Due to one of our experiments being a visualization of data, we were only able to perform rigorous benchmarking on the recommendation system. The benchmarking metrics for each experiment are as follows:

### A. Benchmarks for Recommendation System

Making sure that our model was correct was fairly straightforward, but the process took several steps. The first thing we did was feature analysis. After looking at the different columns available in the dataset, we made an educated guess as to which features could be possible predictors for the type of aircraft used (target type, flight time, payload, etc.). We then ran a Spark job that summed all non-null values for each feature we were interested in, to see which columns had enough data to be included in our model. Most of the columns had values for each row, but we did have to cut out two data features (target cloud cover and target weather), since only one-fourth of the dataset had these values, and training on one-fourth of our data would have significantly degraded the quality of a model we were able to build.

Next, we did a count of the number of airplanes, to understand how a random model would perform. There were approximately twenty airplanes, so a random model would have been able to predict the airplane with 5% certainty. We used 30% of our data to test our model, and used a Spark

module to calculate the model's accuracy on the test data. We found that we could get 60% accuracy on the test data - almost 12 times more effective than a random guess. This is a good percentage, considering that there were probably multiple planes capable of running any mission, and the decision on which plane to use came down to the currently available planes, current resources, etc.

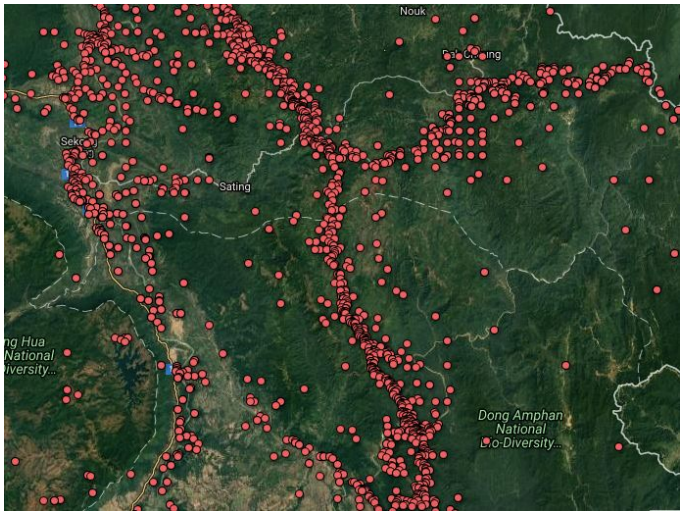
### B. Benchmarks for Visualizations

Benchmarking for our data visualization proved to be quite difficult due to there being no technically rigorous method of doing so. We had to rely on historical documents to determine if the data visualization made logical sense or not. We were also able to use our own intuition to determine the quality of our data visualization, for example the data should be seen in the Indochina region, not the middle of the atlantic ocean. With these strategies we were able to determine the quality of our visualization. We found that ur visualization was accurate, we were able to visualize historical events that occurred throughout the Vietnam War, this allowed us to extend our visual analysis to obtain other more interesting insights about our data.

## VI. Insights Gleaned

Through this study there were a variety of topics that were new to us. The immediate learning points came from developing with the Apache Spark framework. More specifically, we had to learn how to develop in Scala as there was easy integration between the two components. We found that there was a moderately steep learning curve for the inclusion of scala and Apache Spark libraries. The MLib library in particular took a long time to get accustomed to. While the documentation and examples provided by Apache were helpful, there wasn't much community support for the MLib library or resources outside of the official documentation. The GeoSpark third party library was equally frustrating, with there being little documentation or community support outside of the official webpage. The situation with GeoSpark taught us that third party libraries have varied interoperability and can be highly dependant on the configuration of the system using the library.

Regarding creating visualizations of the dataset, we came across interesting information that challenged our intuition about the Vietnam War. We learned that most the of aerial bombardments took place in Laos and Cambodia. We also noticed that the bombing placements of B-52 strikes and other aircraft strikes occur in a grid pattern. This could be from poor data collection or a product of formal strategy such as carpet bombing. When looking at the difference of targets bombed in respect to aircraft, we saw that AC-130s targeted isolated supply routes, while the B-52s targeted infrastructure and industrial sectors.




Visualizing the Ho Chi Minh Trail (Vietcong supply routes)

Regarding our recommendation system, we found that it was possible to predict features of a dataset based off the other features. This by itself is not a new conclusion, but we found that a potential useful application of this technique would be filling in gaps in old, historical datasets; this technique could even be applied to this dataset. Perhaps we wanted to visualize all the F-4 strikes, but some rows in our dataset were missing information on which plane was used. If those points still had latitude/longitude coordinates. Using our prediction system, we could fill in those missing data points with our model's accuracy. Assuming the model was accurate enough, it would be reasonable to still use the generated F-4 points in our visualization, even though we weren't explicitly told which plane those strikes were from.

## VII. Transformation of the Problem Space in the Future

Imagine a world where almost everything is connected to a network. We believe this is an inevitable reality in the near future. The trend of producing increasingly connected devices, coupled with the trend of increased autonomous data production, will lead to a vast data analytics space for various industries. More data granularity and volume of data opens many doors to the big data space, but not at a

cost. At the moment our technology will not be able to handle the future applications of predictive analytics. Though ITProPortal claims that this need for increased technology is necessary for analytics in the business space, but we believe the fundamental concepts translate to our research space [11]. The main crux of predictive analysis is using historical data to make an educated guess about the near future. This will be a sizable benefit to any government when looking to predict the effectiveness of military mission planning. In order for this future to come about, there needs to be advancements not only in processing power and efficiency but also data collection mechanisms. More so data collection mechanisms as Moore's Law will continue and quantum computing enters production. The limiting factor for the future is our network capabilities. There is a natural boundary to how fast data can be communicated that is dictated by the laws of physics [12]. If this boundary is overcome in the future, then we believe the volume of data that can be collected will increase substantially. In turn, the general public will be able to use abstracted interfaces in order to perform data analytics on what will be their past and our present. Due to the increase in devices and instruments that produce data, the future generations of the world will be able to experience much more complex analysis of war time strategies. They will also be able to create intricate models that represent asset trends for present day military operations. Another interesting experiment that will be available in the future is comparing actions taken between multiple conflicts. This comparison could be used for future strategy influence, or just be purely for historical purposes. But in reality the possibilities would be endless for how these technologies will be used. Whatever the future holds, we believe that the data will become much more complex and abundant. All creating increasingly more accurate predictions and a better understanding of the historical events that take place.

## VIII. Conclusion

In conclusion the knowledge gained and results produced show more than a proof of concept. Because of this techniques developed can be utilized in various ways. This work is not limited to historical military mission data, it also has the potential to work in many planning or data reconstruction applications. For example, if the model was trained on historical data that had missing information from a moderate amount of records, we predict that our model would be able to make an educated guess about what the missing information would have been with respect to the trained dataset. This recommendation technique would be useful for historians and researchers in other various fields working with incomplete or corrupted datasets.

We also had success in manipulating the data such that we were able to visualize geographical patterns based on user input. The user would input the information they wanted to see represented with respect to location. An interesting result was that we were able to depict the Ho Chi Minh trail, where supplies were smuggled into Vietnam from Laos and Thailand, using only the location of AC-130 airstrikes. With the help of FusionTables, we were also able to create heat maps based on the total weight of napalm dropped on targeted areas. These results are predicted to have a beneficial impact on understanding the strategies and actions taken during the Vietnam War. The techniques used to produce these maps are also not limited to military bombing data. An application of this technique may be used with any dataset that contains geographical location data such as longitude and latitude. The user only needs to input the category (represented by columns in the csv file) and type within that category in order to produce useful maps.

Though the process of these experiments we were able to more fully understand the challenges and importance of the developed techniques stated above. We also realize the various use cases for these applications such as environmental recovery research, military strategy planning, historical event research, and humanitarian aid support. These areas are not the only fields of study that could be impacted due to the generalized nature of our solutions. As a result of this we believe the work we have done is contributing to the way we look at data. Through the implementation of our machine learning model, missing or corrupted data may not be useless after all.

## IX. Bibliography

[1] "Number of Arial Bombardment Missions by Military Vietnam 1965-1975 (Monthly Totals)", Noah Rippner, "https://data.world/datamil/vietnam-war-thor-data/discuss/vietnam-war-thor-data/5804"

[2] "Mission Flight Hours by Time of Day Viet Nam 1965-1975", Noah Rippner, "https://data.world/datamil/vietnam-war-thor-data/discuss/vietnam-war-thor-data/5804"

[3] "Visualization of 2.9 million air missions over Vietnam '65-'73",u/datadotworld,

”https://www.reddit.com/r/dataisbeautiful/comments/5jfys3/29_million_air_missions_over_vietnam_65_73_oc/”

[4] “Number of Explosives Dropped by Military Vietnam War 1965- 1975 (Monthly Totals)”, Noah Rippner, ”https://data.world/datamil/vietnam-war-thor-data/discuss/vietnam-war-thor-data/5804”

[5] “Bombing Missions of the Vietnam War”, Unknown, “https://geographicalimaginations.files.wordpress.com/2017/01/vietnamwarbombing-01.jpg”

[6] “Counts of the descriptions of the 'KINETIC' missions”, Noah Rippner, ”https://data.world/datamil/vietnam-war-thor-data/discuss/vietnam-war-thor-data/5804”

[7] “Counts of the descriptions of the 'NONKINETIC' missions”, Noah Rippner, ”https://data.world/datamil/vietnam-war-thor-data/discuss/vietnam-war-thor-data/5804”

[8] “Hadoop 2.7.3 Setup Walkthrough”, Unknown, Computer Science Department, Colorado State University, “http://www.cs.colostate.edu/~cs455/CS455-Hadoop-Setup-Guide.pdf”

[9] “Apache Spark in CSB120 Lab Installation Guide”, Naman Shah, Computer Science Department, Colorado State University, “https://www.cs.colostate.edu/~cs535/PA1-Info/Apache-Spark.pdf”

[10] “Spark Programming Guide”, Unknown, Apache Spark, “https://spark.apache.org/docs/0.9.1/scala-programming-guide.html”

[11] “Why Big Data Demands New Technology”, Philip Woods, Director, KRCS Group Ltd “https://www.itproportal.com/features/why-big-data-demands-new-technology/”

[12] “Introduction to Distributed Systems [Networking]”, Shrideep Pallickara, Department of Computer Science, Colorado State University, “http://www.cs.colostate.edu/~cs455/lectures/CS455-L2-Networking.pdf”