# A Platform for Extracting Driver Behavior from Vehicle Sensor Big Data

Sultan Ibrahim bin Ibrahim[1], Emad Felemban[2]
Ahmad Muaz Qamar[4], Abdulrahman A. Majrashi[6]
Computer Engineering Department
College of Computing and Information Systems
Umm Al Qura University, Saudi Arabia

Faizan Ur Rehman[3], Akhlaq Ahmad[5]
Institute of Consultation and Research Studies
Umm Al-Qura University, Makkah
Saudi Arabia

***Abstract*—Traffic analysis of vehicles in densely populated areas and places of public gathering can provide interesting insights into crowd behavior. Hajj is a spatio-temporally bound religious activity that is held annually and attended by more than 2 million people. More than 17,000 buses are used to transport pilgrims on fixed days to fixed locations. This poses great challenges in terms of crowd management. Using Global Positioning System (GPS) and Automatic Vehicle Location (AVL) sensors attached to buses, a large amount of spatio-temporal vehicle data can be collected for traffic analysis. In this paper, we present a study whereby driver behavior was extracted from an analysis of vehicle big data. We have explained in detail how we collected data, cleaned it, moved it to a big data repository, processed it and extracted information that helped us characterize driver behavior according to our definition of aggressiveness. We have used data from 17,000 buses that has been collected during Hajj 2018.**

***Keywords*—GPS Data; AVL sensors; hajj; big data; traffic analysis***

## I. INTRODUCTION

Open source traffic data for Saudi Arabia, provided by the Saudi Ministry of Interior, shows that rash driving is a major factor contributing to road accidents [1] in the Kingdom of Saudi Arabia[1]. The data also shows that the maximum number of incidents happen in the Makkah region, whose population is far less than the most populous Riyadh region. One of the possible reasons for this increase of accidents in the Makkah region is the fact that the annual Hajj pilgrimage happens in this area whereby more than 2 million people visit the region from all corners of the world. A large fleet of vehicles is needed to transport these people between cities and the holy places. To understand the behavior of drivers, we used the data of 17,000 buses collected in Hajj 2018.

Hajj is an annual pilgrimage of Muslims that happens every year from the 8th to the 13th of Dhul-Hijjah, the 12th month [2][3] of the lunar Islamic calendar. More than 2 million people from Saudi Arabia and around the world come to perform Hajj. International pilgrims come a few days earlier to the city of Makkah. On the 8th, they leave for Mina (bounded with a red dashed line in Fig. 1), a small dwelling of permanently installed tents near Makkah. On the morning of the 9th, they move to Arafat (bounded with a blue dashed line in Fig. 1), an open space about 17 km further south. After sunset, they come back and spend the night at Muzdalifah (bounded with a green dashed line in Fig. 1), completing their return trip to Mina by the morning of the 10th. From 10th onwards, for a period of 3 days, including an optional 4th one, the pilgrims stone the three pillars called Jamarat. They also go to slaughterhouses and visit the Grand Mosque in Makkah to perform certain obligatory rituals [2]. All this movement is restricted with respect to time and space. A fleet of more than 17,000 buses is required to move the pilgrims across the holy places, collectively referred to as Mashaer. To collect traffic data from the large number of buses utilized by the pilgrims, the General Syndicate of Cars (Naqaba[2]) has ordered the bus operators to attach Automatic Vehicle Location (AVL) sensors to their buses. Data collected from these AVL sensors has proven to be very useful in studying a number of characteristics related to traffic. We have developed a system that utilizes this data to present interactive visualization to perceive traffic activity at various hours of the day throughout the Hajj season.

This article provides an expansion of our previous research [5][6] by focusing on analysis of sensor data to extract driver behavior. In this paper we present the results of a study conducted on pilgrim bus data. The main objective of this study is to understand drivers' aggressive behavior by capturing data stamped with spatial and temporal information using Automatic Vehicle Location (AVL) devices based on Global Positioning System (GPS). The vehicles used in this study were pilgrim buses were equipped with AVL sensors that provided location-based data. From this data, we extracted speeds of various buses on different routes. We used this speed data as well as other parameters extracted from the source data to classify drivers and their driving skills according to our defined driver profiles.

This paper is divided into six sections. Section 2 discusses the state of the art in the area of traffic data collection for information extraction using AVL and GPS technology. Section 3 explains the methodology for collecting data and extracting useful information. Section 4 presents an overview of the system architecture and explains the role of different components in the system. Section 5 discusses the results of

[1] https://data.gov.sa/Data/en/organization/ministry_of_interior_-_general_directorate_of_traffic

[2] https://www.haj.gov.sa/en/InternalPages/Details/92

applying analytics on data. Section 6 concludes the paper with a summary of the whole study.

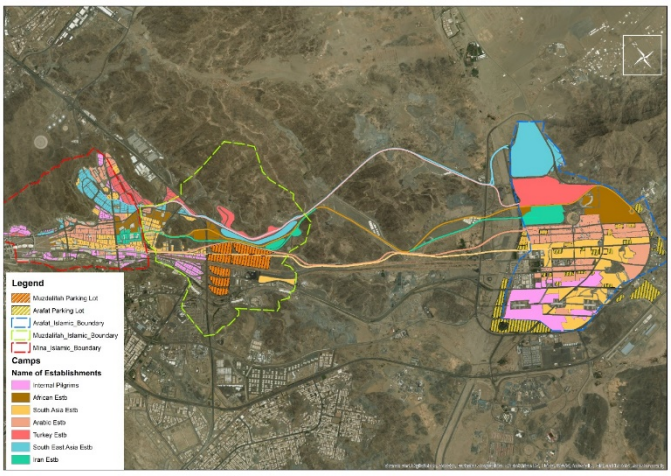



Fig. 1. Map of Mina (Red Boundary), Muzdalif (Green Boundary) and Arafat (Blue Boundary) near Makkah City (Buildings in the Lower Left).

## II. Literature Review

Our study entails using GPS data collected from AVL devices to detect driver behavior from the obtained data. We detail below the state of the art in this regard.

Grengs et al. [7] explained the procedural challenges to collect, store and design databases and to manipulate and analyze the enormous set of geocoded data captured from trips and tours for understanding driving characteristics of a single driver for a duration of a month or so. They studied 78 drivers by using an automobile and recorded their behavior on a day-to-day basis for about a month. They added the position coordinates and time stamp with each data set. Their results showed that the travel patterns were more complex when compared to traditional travels.

Necula et al. [8] utilizes a Hidden Markov Model (HMM) method and a training process and presented an interactive tool to study drivers' behaviors. The tool integrates the past real data captured from various local drivers and analyzes the routes followed by every driver utilizing time, distance height and speed information. The tool also manages the maximum likelihood to validate the next route segment for a network of roads.

Feng et al. [9] examined the merits of using accelerometer with GPS data in a transportation study. They presented three approaches by first considering accelerometer only, then GPS sensors only and lastly a combination of both accelerometer and GPS sensors. They utilized Bayesian Belief Network model to study the three different transportation modes and found that the use of accelerometer can successfully play a significant role in imputing transportation mode. The single usage of each device separately was helpful in terms of predictivity accuracy. The combined usage of both the devices reviled best performance.

Choi et al. [10] considered real driving scenarios and presented a model to detect distraction due to peripheral tasks. They utilized Hidden Markov Models (HMMs) and captured drivers' characteristics using CAN-Bus (Controller Area Network) sensor. This provided them a variety of information such as steering wheel angles, brake status and brake usage with respect to time as well as breaking behavior with associated speed, etc. They defined the drivers' behavior in terms of action and distraction based on the abovementioned data.

Warwick et al. [11] presented their study on drowsiness of drivers and quoted about 100,000 crashes a year on national highways. They emphasized the development of smart a system to detect the drowsiness earlier to avoid accidents. They compared the causes of accidents because of vehicle-based issues with drivers' physiological-based approaches and found that causes were mostly due to drivers' physiology. They proposed to design a driver drowsiness detection system which utilizes wireless wearables to measure a driver's physiological data. The sensory setup provides data that can be analyzed to find key parameters related to drivers' drowsiness and generates early alerts to act in time.

Jasinski et al. [12] proposed a method that identifies real-time aggressive behavior of drivers and generate prior alerts for any dangerous behavior which may result in a severe accident. The method composed of four stages - data collection, pre-processing, semantic enrichment and calculations to compute the aggressiveness of drivers. The (TAI - Trajectory Aggressively Indicator) aggressive Indicator values varies from 0 to 100, with 0 means no aggressiveness and 100 the extremely aggressive. The proposed approach also considers the environmental conditions in calculating a better estimate of TAI.

Paefgen et al. [13] developed a method to measure the accident risk. The method utilizes GPS data collected from a large sample of traffic data from a telematics provider in northern Italy, where there were 1500 drivers with and without accidents over a period of two years. The GPS trajectories were analyzed to study the driver risk profiling problem and their findings in this regard were promising.

Khan at al. [14] presented a comprehensive survey on driving activities, the reasons for accidents, and systems to generate prior notification for drivers for their safety and comfortable drive upon early detection of an accident. Based on their findings, they suggested that a well-designed DMAS (driving monitoring and assistance system) can improve critical issues associated with drivers as well as the challenges associated with the related driving environment.

Stutts et al. [15] and Kan et al. [16] described the main reasons for majority (about 90%) of road accidents as, distraction, fatigue and aggressive driving style. Distraction refers to eating or drinking, looking at off road people, getting busy in small activities like sharing food, texting or attending phone calls, etc. and causes more than half of the accidents. Fatigue explains the physical condition of drivers like drowsiness, over acting to show up extra driving skills, etc. Aggressiveness is related to driving style, overreaction to overtaking cars, or applying breaks in front of other vehicles. Their study suggested the use of DMAS by considering the factors associated with drivers and the driving environment.

Jingqiu et al. [17], developed a hybrid model to study driving behavior and risk patterns. The model utilizes Autoencoder and Self-organized Maps (AESOM) approach to extract driving behaviors. They made 4032 observations by collecting data through GPS sensors, in Shenzhen, China, and analyzed the speed and excessive acceleration and summarize their findings as that AESOM usage may improve the quality of the driving.

Arumugam et al. [18] presented a comprehensive survey on driving behavior and addressed drivers' agressive behavior and detailed multiple incidents on short and long-term driving activities. The purpose of the survey was to explore the solution to minimize the risk on roads by considering drivers' emotional factors defining their driving behavior and provide the information to insurance companies to profile drivers' behavior and define the best possible insurance premium package for risk prevention.

Improving transportation system is one of the significant requirements for large gatherings where crowd safety is major concern [4]. For this researcher proposed several intelligent transportation systems, which in turn opened the door for research areas related to traffic data collection, data mining [19][20][21] processing and analysis [22]. GPS (global position system) sensors are valuable data sources which help in tracking the vehicles by reading their spatial information in real time [23].

## III. Methodology

We have divided our process into 5 major categories - data management, computation, behavior definition, comparison and interpretation. Each category has been further divided as shown in Fig. 2. Data management includes data collection, data cleaning, data enrichment and data visualization. Computation consists calculation of distance, speed and acceleration. Definition entails defining speed ranges with respect to roads. Comparison of the bus speed with allowed ranges with respect to road. Interpretation includes identifying other parameters, driver characteristics and classifying drivers.

### A. Data Management

Data management is a critical step that directly helps in improving the quality of extracted information for analytics. It includes cleaning, structuring, removing noise or fixing missing data and validation. In the following, we explain how we performed the above operations on our data.

*1) Data collection:* To develop any system for decision-making requires collection of a good amount of historical data. Our data source (Naqaba) is the transport authority of the Ministry of Hajj and Umrah that collected data of 17,000+ buses using automatic vehicle location (AVL) service providers in Hajj 2018. Fig. 3 shows some facts about the collected data. The data collected was for pilgrim movement on different routes during Hajj, such as, Jeddah Airport-Makkah, Makkah-Madinah, Madinah – Madinah Airport, Makkah – Mina, Makkah – Arafat, Arafat – Muzdalifah, Muzdalifah – Mina.

*2) Data cleaning:* We excluded noise from the data using the spatial boundary algorithm that removes locations outside a given boundary from the dataset. We also removed data entries where difference in distance between two points for the same bus is around zero and the location is not on main roads, assuming the bus to be parked at a pickup or drop off point or at a parking location.

*3) Data enrichment:* In our case, we collected the GPS traces of the buses during Hajj. Generally, AVL sensor providers configure GPS devices to transmit data at intervals varying from 2 to 7 minutes in length. We found that sometimes the duration between the two locations from the same bus is up to 20-30 minutes due to which some entries are lost, mostly because of connectivity issues. Fig. 4 shows the anomaly associated with a missing entry or a long distance between recorded points. The two locations from the same bus are shown in Fig. 4. The black line shows the line as per the raw data and the blue line shows an extra data point after data enrichment, i.e., after adding a missing point. The enriched data shows the actual distance and will be beneficial to extract knowledge for analytics.

*4) Data validation:* Along with data enrichment, we have also performed data quality checks to handle the GPS error issues. An error of only 10 meters can show the location of the bus on the other side of the road that will lead to a large error in calculating distance, and henceforth speed and acceleration.

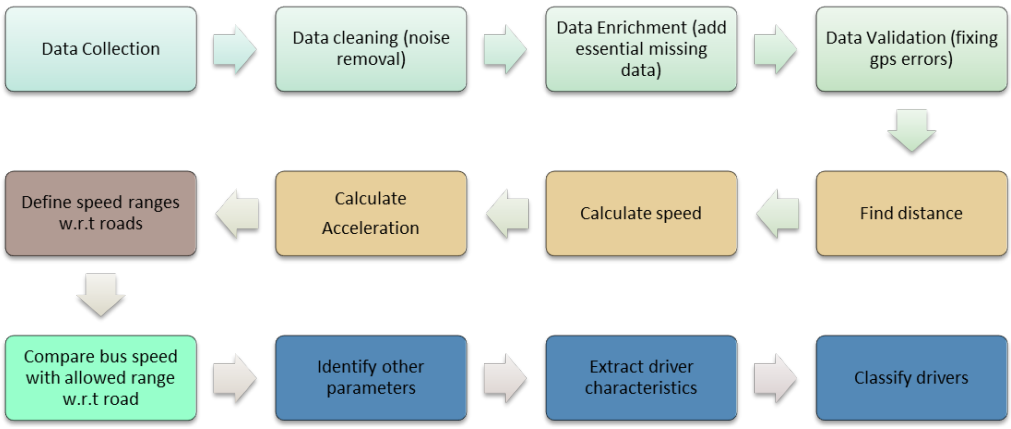



Fig. 2. Overview of the Methodology.

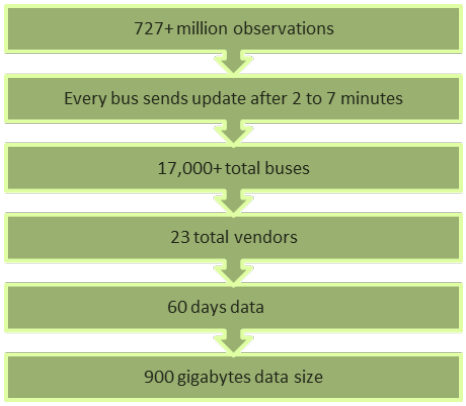



Fig. 3. Overview of Collected Data.


Fig. 4. Enriched Data in Blue Colour.

## B. Computation

After data completing manipulation, we enriched the dataset by adding acceleration and distances information as additional data columns, using mathematical equations as follows:

Calculate acceleration (a):

$$a = \frac{\Delta v}{\Delta t} = \frac{(vf - vi)}{(tf - ti)} \qquad \frac{m}{s^2}$$

Where vf = final velocity, m/s; vi = initial velocity, m⁄s; Δv = difference in velocity, m⁄s; tf = ending time in seconds, ti = starting time in seconds, and Δt = difference of time in seconds.

Calculate distance (d) between adjacent points on the globe as shown in Fig. 5 by using Haversine formula as shown in Fig. 5:

$$a = \sin^2\left(\Phi B - \Phi\frac{A}{2}\right) + \cos\Phi A * \cos\Phi B * \sin^2\left(\lambda B - \frac{\lambda A}{2}\right)$$

$$c = 2 * \operatorname{atan2}\left(\sqrt{a}, \sqrt{(1-a)}\right)$$

$$d = R \cdot c$$

Where Φ = latitude, λ = longitude, R = radius of the earth (R ≈ 6.371 km), A= ending point, B = initial point and d = the distance between two points.

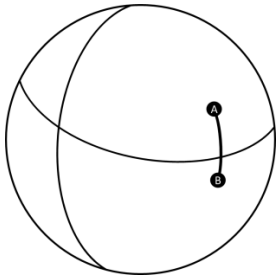



Fig. 5. Distance between Two Points on the Globe is Calculated using the Haversine Formula Due to the Curvature of the Earth's Surface.

TABLE I. ROAD TYPES

| Road | Type |
|---|---|
| Highway | Motorway |
| Highway | Trunk |
| Highway | Primary |
| Highway | Secondary |
| Highway | Residential |
| Highway | Unclassified |

## C. Definition

We use the open source Open Street Maps (OSM) and extract road related characteristics for each road such as the speed limit to calculate the speed threshold for each road and highway as shown in Table I. The speed limit varies from 60-140 kmph depending on the road type and its proximity to populated areas. We allow the driver to cross the speed limit up to 10%.

## D. Comparison

Based on the street profiles extracted above, we match the bus speed data with our speed threshold for each road on the route to classify the vehicles according to speed. The spatial queries have been used with the help of a spatial relational database. The spatial relational database stores the geometry of each road and spatial query checks whether the location of the bus belongs to that road segment or not.

## E. Interpretation

After separating the vehicles violating traffic rules from others, we apply the spatio-temporal conditions mentioned previously on data to classify the drivers into aggressive and non-aggressive behavior.

# IV. System Architecture

Fig. 6 shows the high-level view of the big data platform that we developed to analyze the bus data. We developed a data lake layer that consist a Master data service in addition to an MS SQL service. The master data service provides a visualization of all the relations in the data based on different parameters, such as Establishment, Offices or bus number (every bus is assigned to an Office which is under an Establishment related to a geographical area). The MS SQL database contains the original data we received from Naqaba.

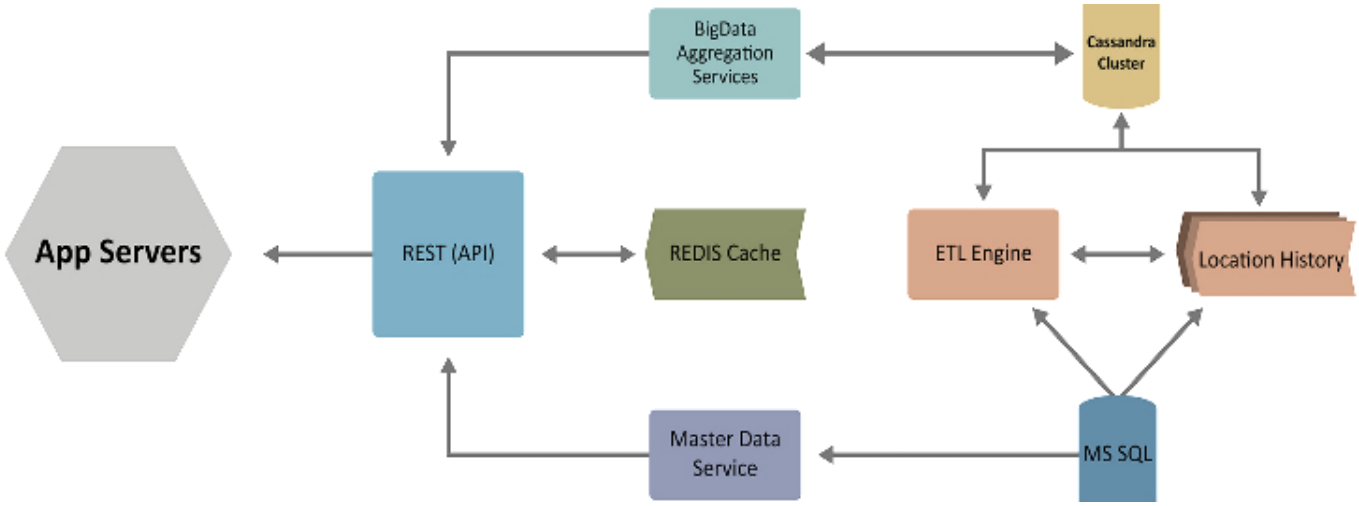



Fig. 6. Overview of the Big Data Platform.

Our big data layer is made up of a Cassandra cluster and a Big Data Aggregation service. We have migrated Location History data to the Cassandra cluster for cleaning and removing noise using the ETL engine. The benefit of using Cassandra cluster is that it increases efficiency and scalability using a distributed, wide column store, running on a NoSQL database management system.

We have used Hadoop and Presto to setup the big data aggregation service. Presto is efficient tool used for distributed SQL analytical queries on data in the Hadoop distributed file system (HDFS). Hadoop is highly beneficial for batch-based analytics while Cassandra is good for time-based. *REDIS* cache is an open-source (BSD-licensed), in-memory data structure store used as a cache.

It is good for caching a huge number of key-value pairs. We have use REDIS cache to boost the performance of the system. We have devised a RESTful API that provides a list of APIs to handle requests coming from the front-end. The front-end requests the API to fetch data from the Master data service, the big data aggregation service or from the REDIS cache and returns the results that are visualised on the screen.

The API Server provides the front-end data visualization and analysing service. It allows the user to display data based on multiple filters, including Establishment (Mo'assasa) name, Office (Maktab) number, company name, bus number and route.

Fig. 7 compares time taken to perform queries in MS SQL server and our big data platform. The orange line shows the time to fetch the records of bus id in MS SQL server while the grey line shows the time to fetch the records in our big data platform. It is clear from the Fig. 8 that time gained from moving to the big data platform is significant.

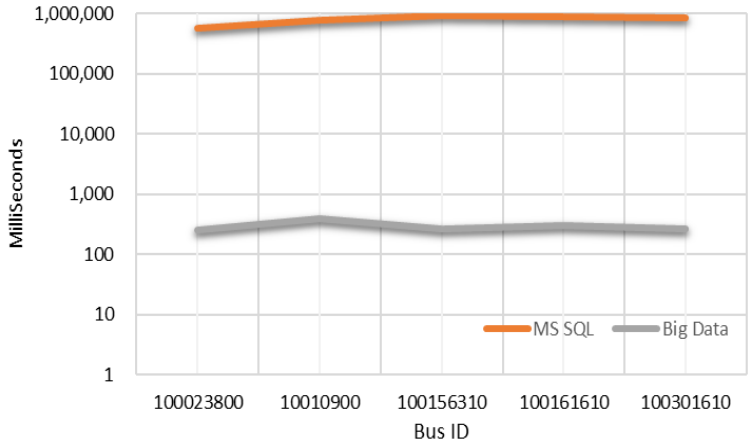



Fig. 7. Time Comparison before and after Migration to Big Data Platform.

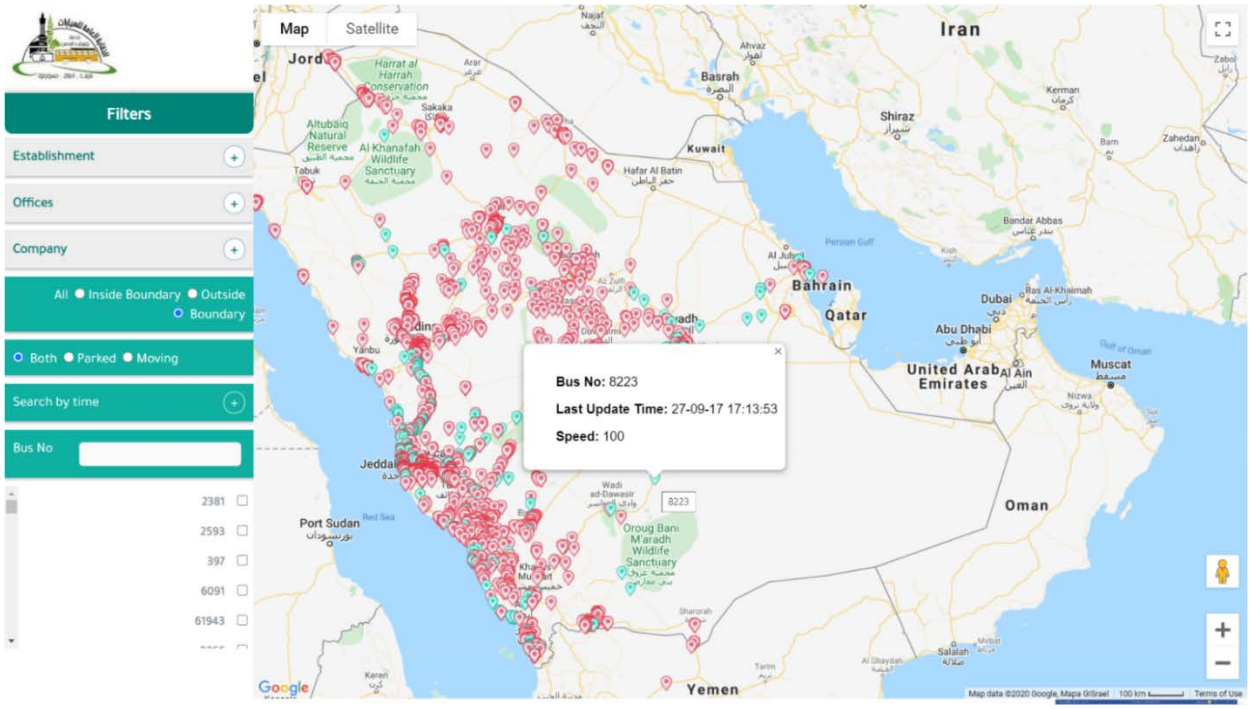



Fig. 8. Platform Visualization after Migration to Big Data Framework.

## V. Analytics and Discussion

Each bus was allotted to a single driver for the entire Hajj season. The vehicles were tracked by capturing their spatio-temporal information and the collected data was analyzed by considering each road speed limit. Table II is a snapshot of the collected data along with few violations' information.

First, we selected the continuous number of observations that exceeded speed limit threshold (80kmh) with starting & ending timestamp. Then we calculated the duration of violation from starting and ending timestamp in minutes and seconds as described in the above table. Fig. 9, is the summary of the number of violations detected by the AVL sensors. We can observe that for most of the days the driver's behavior was aggressive, crossing the threshold speed several times.

Further, we have classified the violations based on severity, and collected information regarding the frequency, duration and severity of speed limit violation by a driver. Table III details our classification of violations.

Fig. 10 summarizes the recorded violations of a driver with perspective of above classification during the entire Hajj season. We can see that normal category of violations is common, which shows that 22 times the driver violated the speed limit but just for a few seconds or so, and then reduced his speed less than the threshold. To address normal violation cases, it happened 8 times that he violated the speed limit for about 10 minutes duration. Three times he continuously violated for a duration between 10-20 minutes and twice, he committed severe violations, that is, for more than 20 minutes.

TABLE II. Violation Information Extracted from Source Data

| Start Timestamp | End Timestamp | No of Consecutive Observations | Violation Duration in (s) | Time Minutes |
|---|---|---|---|---|
| 2016-09-13 04:47:44 | 2016-09-13 04:55:44 | 2 | 480.0 | 8 |
| 2016-09-13 06:05:46 | 2016-09-13 06:13:45 | 2 | 479.0 | 7.98 |
| 2016-09-13 06:31:55 | 2016-09-13 06:31:55 | 1 | 0.0 | 0 |
| 2016-09-14 05:51:27 | 2016-09-14 05:51:27 | 1 | 0.0 | 0 |
| 2016-09-15 06:29:29 | 2016-09-15 06:37:29 | 2 | 480.0 | 8 |
| 2016-09-15 11:43:35 | 2016-09-15 12:19:36 | 17 | 2161.0 | 36.01 |
| 2016-09-15 12:29:43 | 2016-09-15 12:35:37 | 3 | 354.0 | 5.90 |

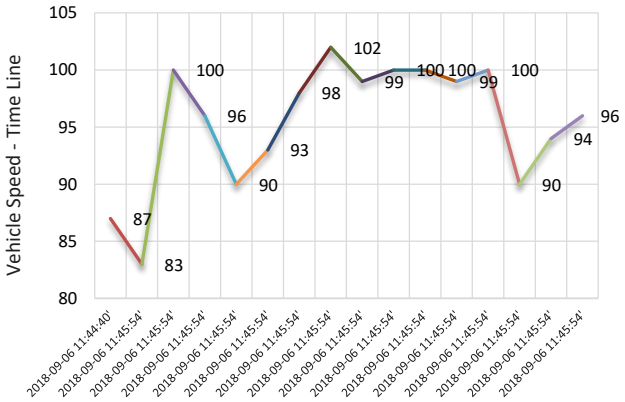



Fig. 9. Vehicle Speed - Time Line.

TABLE III. VIOLATION SEVERITY

| Severity | Explanation |
|---|---|
| Once | A violation just for a few seconds and not a continuous one for a long duration. These violations occurred only in individual observations so we cannot get starting timestamp & ending timestamp, hence we called it Once |
| Normal | If the violation's duration is less than 10 minutes, then we called it Normal violation |
| High | If the violation's duration is between 10 to 20 minutes, then we say it is High |
| Severe | If the violation's duration is more than 20 minutes, then we consider it Severe |

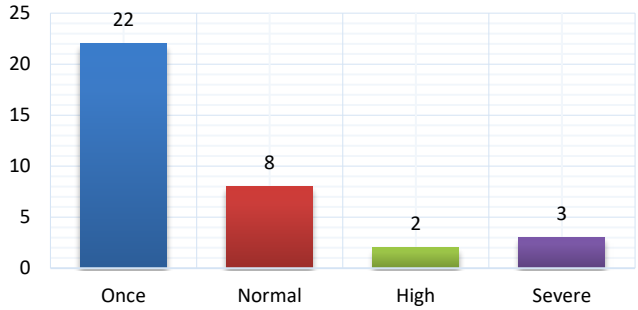



Fig. 10. Driver's behavior (Violation Classification-Bus No. 1).

On the same scale, the system analyzed mobility of several buses and found that the bus with ID 152, violated the speed limit in total 905 times, out of which mostly the violation was in once category and 197 times the bus violated for duration less than 10 minutes. No case of severe violation was recorded for bus 152. This analysis helps find out the worst cases as shown in Fig. 11.

Among all the drivers, there were some with best performance as they committed no or minor violations. Fig. 12 below is the summary of the best buses where drivers' behavior was in a satisfactory range. The best case is for bus ID who violated just for a few seconds during the entire Hajj season.

Fig. 13 shows the speed-based detection of aggressiveness. The geographical locations captured with timestamps were analyzed for one of the drivers, who was driving on C-ring road, in Makkah. Upon analysis, we discussed both the non-aggressive behavior (Fig. 13a: the green dots show that the driver's speed was within the threshold value and was not committing any violation) and aggressive behavior (Fig. 13b: the red dots show that the driver violated several times the threshold speed showing his aggressive behavior). The data points show that the driver's behavior was aggressive for 52.55% of the collected data points, which has been visualized in the figure below. The corresponding heading and acceleration information for both the cases is also detailed.

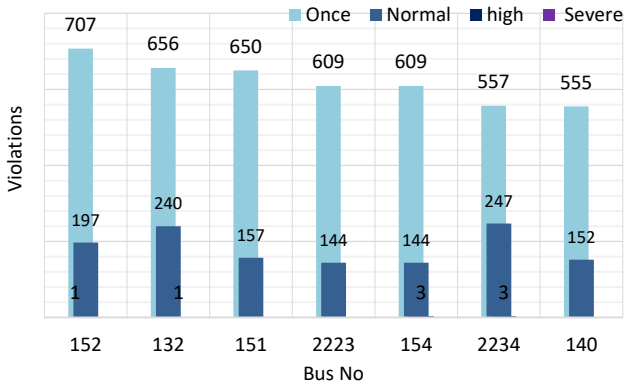



Fig. 11. Violation Severity based on Bus Number. Identification of Worst Drivers.

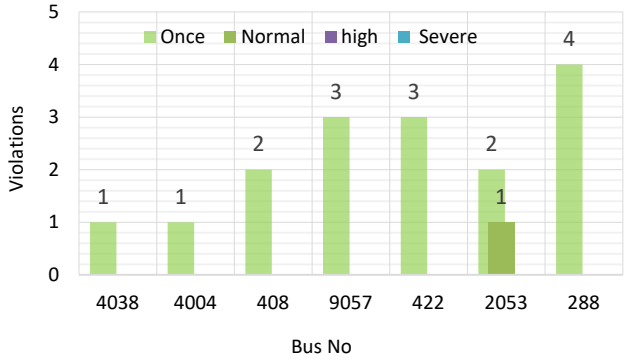
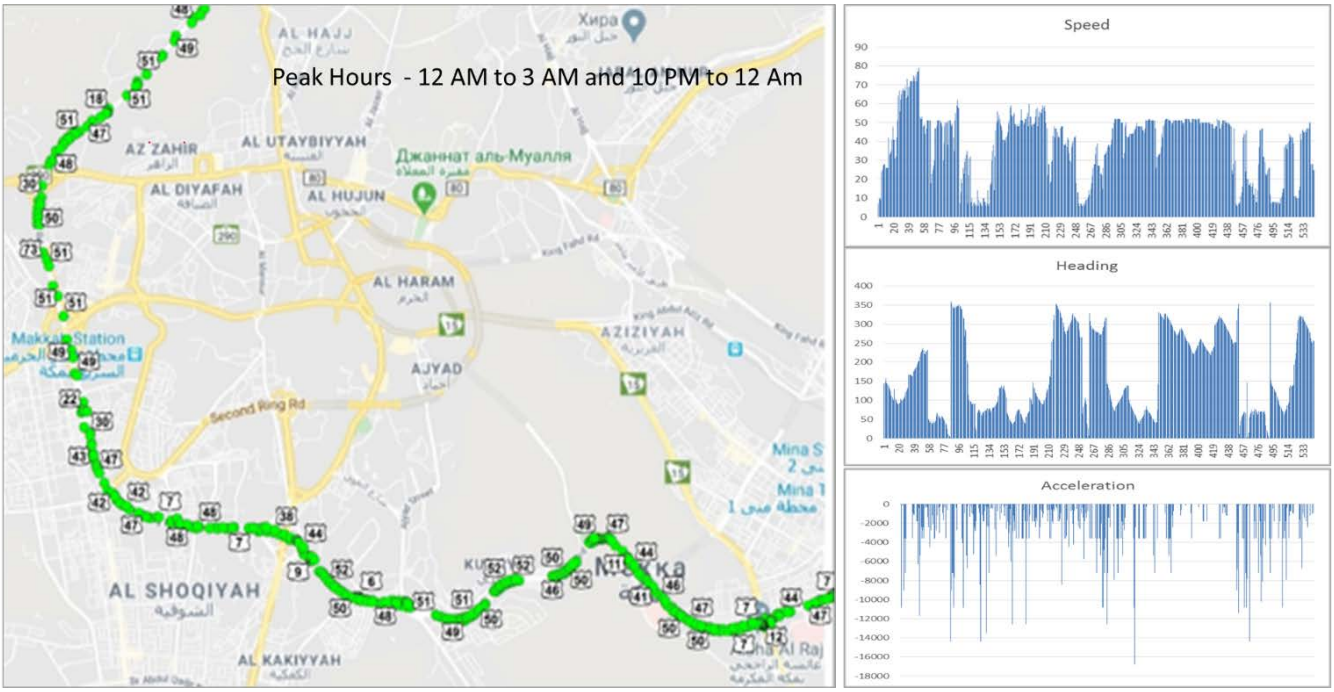



Fig. 12. Identification of Best Driver.




(a) The Green-Dots Indicate the Places of Driver's Non-Aggressive Behaviour.

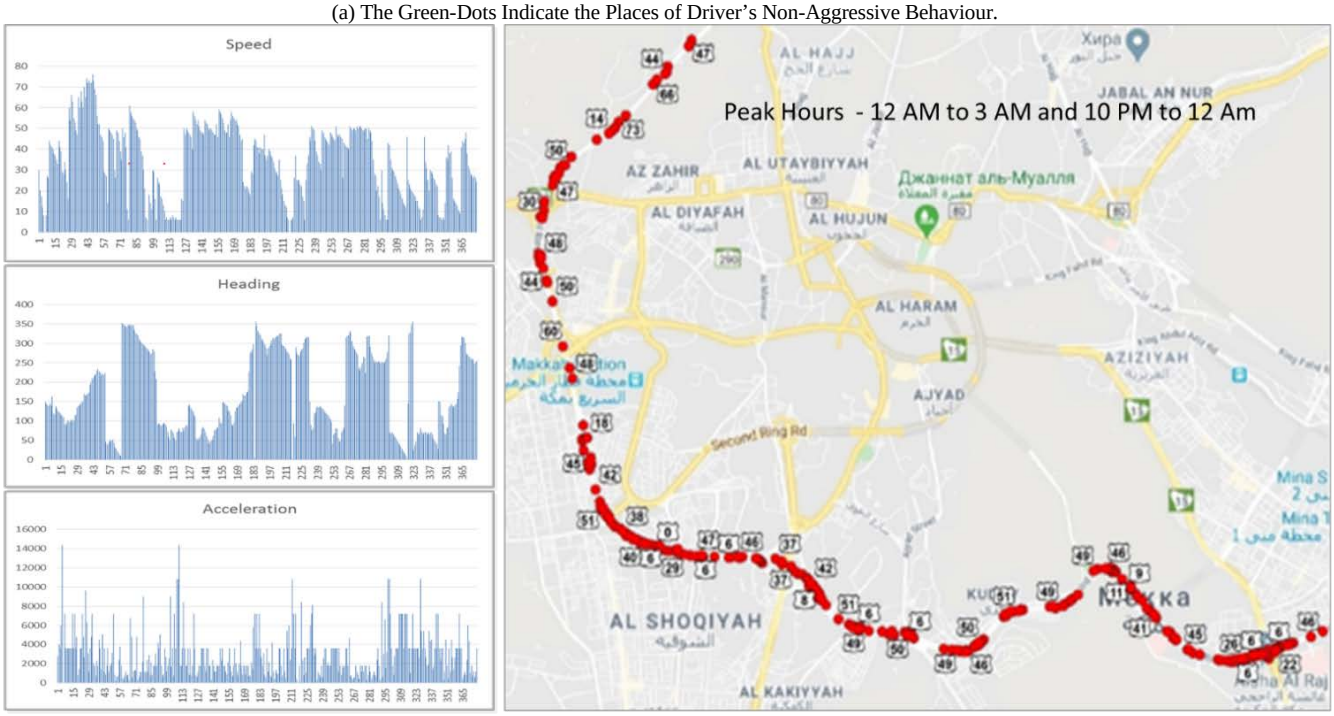



Fig. 13. (b): The Red-Dots Indicate the Places of Driver's Aggressive Behaviour.

## VI. Conclusion

In this paper we have presented a study to extract driver aggressiveness information from GPS data obtained through AVL sensors attached to a fleet of 17,000+ buses used to transport pilgrims from one place to another in the holy areas during the Hajj pilgrimage. We have explained the details of data preparation and pre-processing methodology we have adopted by moving the data to a big data platform for efficient query processing. We have also highlighted our procedure for extraction of information. One of the limitations of the experiment is that it was carried out in one particular city on drivers from outside the city. However, our technique is generic and can be used on any AVL based data in any part of the world.

## Acknowledgement

The authors would like to thank Deanship of Scientific Research and Prince Khalid Al-Faisal Chair for Developing Makkah Al-Mukarramah and the Holy Places at Umm Al-Qura University (project # DSRUQU.PKC-41-5) for the financial support. We would also like to thank Eng. Osama Fateh from Naqaba for providing us with the bus data.

## References

[1] U. B. Ghaffar and S. Ahmed, “A Review of Road traffic accident in Saudi Arabia: the neglected epidemic,” Indian J. Forensic Community Med., vol. 2, no. 4, p. 242, 2015, doi: 10.5958/2394-6776.2015.00010.7.

[2] A. Ahmad, M. A. Rahman, M. Ridza Wahiddin, F. Ur Rehman, A. Khelil, and A. Lbath, “Context-aware services based on spatio-temporal zoning and crowdsourcing,” Behav. Inf. Technol., 2018, doi: 10.1080/0144929X.2018.1476586.

[3] A. Ahmad et al., “A framework for crowd-sourced data collection and context-aware services in Hajj and Umrah,” in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), Nov. 2014, pp. 405–412, doi: 10.1109/AICCSA.2014.7073227.

[4] F. Ur Rehman et al., "A constraint-aware optimized path recommender in a crowdsourced environment," 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, 2015, pp. 1-8, doi: 10.1109/AICCSA.2015.7507185.

[5] E. Felemban, F. U. Rehman, A. A. Biabani, A. Naseer, and U. AlAbdulwahab, “Towards Building an Interactive Platform for Analyzing Movement of Buses in Hajj,” in 2019 IEEE International Conference on Big Data (Big Data), Dec. 2019, pp. 3775–3778, doi: 10.1109/BigData47090.2019.9005521.

[6] E. Felemban, F. U. Rehman, H. Wadood, and A. Naseer, “Towards Building Evacuation Planning Platform using Multimodal Transportation for a Large Crowd,” in 2019 IEEE International Conference on Big Data (Big Data), Dec. 2019, pp. 4063–4066, doi: 10.1109/BigData47090.2019.9006226.

[7] J. Grengs, X. Wang, and L. Kostyniuk, “Using GPS Data to Understand Driving Behavior,” J. Urban Technol., vol. 15, no. 2, pp. 33–53, Aug. 2008, doi: 10.1080/10630730802401942.

[8] E. Necula, “Mining GPS Data to Learn Driver’s Route Patterns,” in 2014 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Sep. 2014, pp. 264–271, doi: 10.1109/SYNASC.2014.43.

[9] T. Feng and H. J. P. Timmermans, “Transportation mode recognition using GPS and accelerometer data,” Transp. Res. Part C Emerg. Technol., vol. 37, pp. 118–130, Dec. 2013, doi: 10.1016/j.trc.2013.09.014.

[10] S. Choi, J. Kim, D. Kwak, P. Angkititrakul, and J. H. L. Hansen, “Analysis and classification of driver behavior using in-vehicle can-bus information,” Bienn. Work. DSP In-Vehicle Mob. Syst., no. October 2015, pp. 17–19, 2007.

[11] B. Warwick, N. Symons, X. Chen, and K. Xiong, “Detecting Driver Drowsiness Using Wireless Wearables,” in 2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems, Oct. 2015, pp. 585–588, doi: 10.1109/MASS.2015.22.

[12] M. G. Jasinski and F. Baldo, “A Method to Identify Aggressive Driver Behaviour Based on Enriched GPS Data Analysis,” GEOProcessing 2017 Ninth Int. Conf. Adv. Geogr. Inf. Syst. Appl. Serv., no. March 2017, pp. 97–102, 2017, [Online]. Available: https://www.thinkmind.org/download.php?articleid=geoprocessing_2017_6_20_38002.

[13] J. Paefgen, F. Michahelles, and T. Staake, “GPS trajectory feature extraction for driver risk profiling,” TDMA’11 - Proc. 2011 Int. Work. Trajectory Data Min. Anal., pp. 53–56, 2011, doi: 10.1145/2030080.2030091.

[14] M. Q. Khan and S. Lee, “A Comprehensive Survey of Driving Monitoring and Assistance Systems,” Sensors, vol. 19, no. 11, p. 2574, Jun. 2019, doi: 10.3390/s19112574.

[15] J. C. Stutts, D. W. Reinfurt, and E. A. Rodgman, “The role of driver distraction in crashes: an analysis of 1995-1999 Crashworthiness Data System Data.,” Annu. Proc. Assoc. Adv. Automot. Med., vol. 45, no. May, pp. 287–301, 2001.

[16] M. Q. Khan and S. Lee, “A comprehensive survey of driving monitoring and assistance systems,” Sensors (Switzerland), vol. 19, no. 11, 2019, doi: 10.3390/s19112574.

[17] J. Guo, Y. Liu, L. Zhang, and Y. Wang, “Driving Behaviour Style Study with a Hybrid Deep Learning Framework Based on GPS Data,” Sustainability, vol. 10, no. 7, p. 2351, Jul. 2018, doi: 10.3390/su10072351.

[18] S. Arumugam and R. Bhargavi, “A survey on driving behavior analysis in usage based insurance using big data,” J. Big Data, vol. 6, no. 1, p. 86, Dec. 2019, doi: 10.1186/s40537-019-0249-5.

[19] S. Turner and L. Albert, “Its Data Quality Control and the calculation of Mobility Performance Measures,” 2000.

[20] Chaogui Zhang, Zhiyong Zheng, Fuqiang Zhang, and Jiangtao Ren, “Multidimensional traffic GPS data quality analysis using data cube model,” in Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), Dec. 2011, pp. 307–310, doi: 10.1109/TMEE.2011.6199204.

[21] G. Gecchele, R. Rossi, M. Gastaldi, and A. Caprini, “Data Mining Methods for Traffic Monitoring Data Analysis: A case study,” Procedia - Soc. Behav. Sci., vol. 20, pp. 455–464, 2011, doi: 10.1016/j.sbspro.2011.08.052.

[22] H. Wang, M. Ouyang, Q. Meng, and Q. Kong, “A traffic data collection and analysis method based on wireless sensor network,” Eurasip J. Wirel. Commun. Netw., vol. 2020, no. 1, 2020, doi: 10.1186/s13638-019-1628-5.

[23] L. Shen and P. R. Stopher, “Review of GPS Travel Survey and GPS Data-Processing Methods,” Transp. Rev., vol. 34, no. 3, pp. 316–334, May 2014, doi: 10.1080/01441647.2014.903530.

APPENDIX 1: SAMPLE OF ENRICHMENT DATA

| Door_NO | Bus_No | GPS_Date | Latitude | Longitude | GPS Speed | Heading | Acceleration | Status | Distance | Calculated Speed (Km/h) | Compare speeds | Speed Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2345 | 1006 | 02-SEP-16 05.34.48.671000000 PM | 21.439383 | 39.542163 | 10 | 67 | 126.0190919 | Aggressive | 0.53333 | 8 | Calculated_Speed < GPS Speed | 80.00% |
| 2345 | 1006 | 02-SEP-16 05.34.49.461000000 PM | 21.439386 | 39.542193 | 14 | 90 | 9856.262834 | Aggressive | 0.00568 | 14 | Calculated_Speed = GPS Speed | 100.00% |
| 2345 | 1006 | 02-SEP-16 05.34.50.486000000 PM | 21.439373 | 39.54223 | 16 | 115 | 4845.222073 | Aggressive | 0.00619 | 15 | Calculated_Speed < GPS Speed | 93.75% |
| 2345 | 1006 | 02-SEP-16 05.34.51.498000000 PM | 21.439343 | 39.542263 | 17 | 137 | 2403.204272 | Aggressive | 0.00791 | 19 | Calculated_Speed > GPS Speed | 89.47% |
| 2345 | 1006 | 02-SEP-16 05.34.52.204000000 PM | 21.439301 | 39.54229 | 18 | 148 | 2990.033223 | Aggressive | 0.00602 | 18 | Calculated_Speed = GPS Speed | 100.00% |
| 2345 | 1006 | 02-SEP-16 05.34.53.428000000 PM | 21.439253 | 39.542315 | 21 | 154 | 7563.02521 | Aggressive | 0.00873 | 22 | Calculated_Speed > GPS Speed | 95.45% |
| 2345 | 1006 | 02-SEP-16 05.35.00.599000000 PM | 21.438873 | 39.542495 | 22 | 151 | 473.7465456 | Aggressive | 0.35 | 21 | Calculated_Speed < GPS Speed | 95.45% |
| 2345 | 1006 | 02-SEP-16 05.35.01.501000000 PM | 21.438823 | 39.542528 | 23 | 144 | 2398.401066 | Aggressive | 0.01001 | 24 | Calculated_Speed > GPS Speed | 95.83% |
| 2345 | 1006 | 02-SEP-16 05.35.02.475000000 PM | 21.438775 | 39.54257 | 25 | 135 | 4881.355932 | Aggressive | 0.01024 | 25 | Calculated_Speed = GPS Speed | 100.00% |
| 2345 | 1006 | 02-SEP-16 05.35.03.364000000 PM | 21.438731 | 39.54262 | 24 | 128 | -2639.296188 | Non-Aggressive | 0.00834 | 22 | Calculated_Speed < GPS Speed | 91.67% |
| 2345 | 1006 | 02-SEP-16 05.35.04.221000000 PM | 21.438691 | 39.542675 | 27 | 123 | 8845.208845 | Aggressive | 0.00848 | 25 | Calculated_Speed < GPS Speed | 92.59% |
| 2345 | 1006 | 02-SEP-16 05.35.06.698000000 PM | 21.438603 | 39.542796 | 29 | 129 | 2668.64344 | Aggressive | 0.02098 | 28 | Calculated_Speed < GPS Speed | 96.55% |
| 2345 | 1006 | 02-SEP-16 05.35.08.348000000 PM | 21.438503 | 39.542911 | 29 | 133 | 0 | Non-Aggressive | 0.01957 | 30 | Calculated_Speed > GPS Speed | 96.67% |
| 2345 | 1006 | 02-SEP-16 05.35.09.323000000 PM | 21.438445 | 39.542966 | 31 | 137 | 5442.176871 | Aggressive | 0.01103 | 30 | Calculated_Speed < GPS Speed | 96.77% |
| 2345 | 1006 | 02-SEP-16 05.35.10.404000000 PM | 21.438386 | 39.543021 | 30 | 140 | -2564.102564 | Non-Aggressive | 0.01131 | 29 | Calculated_Speed < GPS Speed | 96.67% |
| 2345 | 1006 | 02-SEP-16 05.35.12.651000000 PM | 21.438268 | 39.543116 | 27 | 146 | -4073.934364 | Non-Aggressive | 0.02136 | 29 | Calculated_Speed > GPS Speed | 93.10% |
| 2345 | 1006 | 02-SEP-16 05.35.14.719000000 PM | 21.438158 | 39.543195 | 24 | 149 | -3972.048547 | Non-Aggressive | 0.01813 | 24 | Calculated_Speed = GPS Speed | 100.00% |
| 2345 | 1006 | 02-SEP-16 05.35.15.449000000 PM | 21.438105 | 39.543226 | 24 | 152 | 0 | Non-Aggressive | 0.00886 | 22 | Calculated_Speed < GPS Speed | 91.67% |
| 2345 | 1006 | 02-SEP-16 05.35.16.349000000 PM | 21.438051 | 39.543255 | 22 | 156 | -5337.286879 | Non-Aggressive | 0.00899 | 24 | Calculated_Speed > GPS Speed | 91.67% |

APPENDIX 2: SAMPLE OF FINAL RESULTS

| Bus_No | Data Collection Time | Data Calculation Time | Total time | Number of Records | Aggressive | Non-Aggressive |
|---|---|---|---|---|---|---|
| **1006** | 0.0:0.0:0.030506 | 0.0:0.0:0.000005 | 0.0:0.0:0.030511 | 5658 | 42.08% | 57.92% |
| **1010** | 0.0:0.0:0.039205 | 0.0:0.0:0.000004 | 0.0:0.0:0.039210 | 7346 | 38.32% | 61.68% |
| **1029** | 0.0:0.0:0.023313 | 0.0:0.0:0.000004 | 0.0:0.0:0.023317 | 4368 | 40.38% | 59.62% |
| **1102** | 0.0:0.0:0.008098 | 0.0:0.0:0.000003 | 0.0:0.0:0.008101 | 1652 | 14.41% | 85.59% |
| **1109** | 0.0:0.0:0.029801 | 0.0:0.0:0.000004 | 0.0:0.0:0.029805 | 5575 | 43.34% | 56.66% |
| **1202** | 0.0:0.0:0.000013 | 0.0:0.0:0.000002 | 0.0:0.0:0.000015 | 3 | 0.00% | 100.00% |
| **1205** | 0.0:0.0:0.012375 | 0.0:0.0:0.000003 | 0.0:0.0:0.012378 | 2311 | 40.07% | 59.93% |
| **1207** | 0.0:0.0:0.058354 | 0.0:0.0:0.000005 | 0.0:0.0:0.058359 | 10858 | 43.67% | 56.33% |
| **1209** | 0.0:0.0:0.015614 | 0.0:0.0:0.000005 | 0.0:0.0:0.015619 | 2916 | 40.98% | 59.02% |
| **1210** | 0.0:0.0:0.058767 | 0.0:0.0:0.000005 | 0.0:0.0:0.058772 | 11019 | 34.42% | 65.58% |
| **1211** | 0.0:0.0:0.047687 | 0.0:0.0:0.000005 | 0.0:0.0:0.047691 | 8966 | 36.17% | 63.83% |
| **1213** | 0.0:0.0:0.026305 | 0.0:0.0:0.000004 | 0.0:0.0:0.026308 | 5010 | 36.71% | 63.29% |
| **1214** | 0.0:0.0:0.026867 | 0.0:0.0:0.000005 | 0.0:0.0:0.026872 | 5155 | 33.40% | 66.60% |
| **1215** | 0.0:0.0:0.000034 | 0.0:0.0:0.000002 | 0.0:0.0:0.000036 | 8 | 25.00% | 75.00% |
| **1216** | 0.0:0.0:0.018076 | 0.0:0.0:0.000003 | 0.0:0.0:0.018079 | 3386 | 40.84% | 59.16% |
| **1217** | 0.0:0.0:0.054429 | 0.0:0.0:0.000005 | 0.0:0.0:0.054434 | 10162 | 38.98% | 61.02% |
| **1219** | 0.0:0.0:0.025904 | 0.0:0.0:0.000005 | 0.0:0.0:0.025909 | 4848 | 36.14% | 63.86% |
| **1220** | 0.0:0.0:0.040444 | 0.0:0.0:0.000004 | 0.0:0.0:0.040448 | 7569 | 38.79% | 61.21% |
| **1222** | 0.0:0.0:0.013775 | 0.0:0.0:0.000003 | 0.0:0.0:0.013777 | 2610 | 39.46% | 60.54% |
| **1223** | 0.0:0.0:0.030309 | 0.0:0.0:0.000005 | 0.0:0.0:0.030313 | 5654 | 38.80% | 61.20% |
| **1225** | 0.0:0.0:0.011557 | 0.0:0.0:0.000004 | 0.0:0.0:0.011561 | 2162 | 38.34% | 61.66% |
| **1226** | 0.0:0.0:0.004240 | 0.0:0.0:0.000003 | 0.0:0.0:0.004243 | 817 | 33.90% | 66.10% |
| **1228** | 0.0:0.0:0.039108 | 0.0:0.0:0.000003 | 0.0:0.0:0.039111 | 7338 | 41.69% | 58.31% |
| **1229** | 0.0:0.0:0.058615 | 0.0:0.0:0.000005 | 0.0:0.0:0.058620 | 10974 | 36.66% | 63.34% |