Pre-print Manuscript of Article:
Bridgelall, R., Lu, P., Tolliver, D., Xu, T., "Mining Connected Vehicle Data for Beneficial Patterns in Dubai Taxi Operations," Journal of Advanced Transportation, Wiley/Hindawi, DOI: 10.1155/2018/8963234, 2018 (8963234), 8p, Sep 18, 2018.

Mining Connected Vehicle Data for Beneficial Patterns in Dubai Taxi Operations

Raj Bridgelall,[1] Pan Lu,[1] Denver D. Tolliver,[2] and Tie Xu[3]

[1] College of Business, North Dakota State University, Fargo, North Dakota, 58108, U.S.A.
[2] UGPTI, North Dakota State University, Fargo, North Dakota, 58108, U.S.A.
[3] University of Modern Sciences, Dubai, United Arab Emirates

Abstract

On-demand shared mobility services such as Uber and micro-transit are steadily penetrating the worldwide market for traditional dispatched taxi services. Hence, taxi companies are seeking ways to compete. This study mined large-scale mobility data from connected taxis to discover beneficial patterns that may inform strategies to improve dispatch taxi business. It is not practical to manually clean and filter large-scale mobility data that contains GPS information. Therefore, this research contributes and demonstrates an automated method of data cleaning and filtering that is suitable for such types of datasets. The cleaning method defines three filter variables and applies a layered statistical filtering technique to eliminate outlier records that do not contribute to distributions that match expected theoretical distributions of the variables. Chi-squared statistical tests evaluate the quality of the cleaned data by comparing the distribution of the three variables with their expected distributions. The overall cleaning method removed approximately 5% of the data, which consisted of errors that were obvious and others that were poor quality outliers. Subsequently, mining the cleaned data revealed that trip production in Dubai peaks for the case when only the same two drivers operate the same taxi. This finding would not have been possible without access to proprietary data that contains unique identifiers for both drivers and taxis. Datasets that identify individual drivers are not publicly available.

Keywords: Data quality; Data mining; Data science; Intelligent transportation systems; Mobility index; Shared economy; Taxi operations; Vehicle information systems;

Introduction

In many major cities of the world, on-demand shared mobility services are disrupting the business model of traditional street hailing and dispatched taxi services. On-demand shared mobility services involve popular transportation network companies (TNCs) such as Uber and Lyft, and micro-transit services such as Ford-owned Chariot [1]. This escalating competition for passengers has been motivating taxi companies to mine dynamic mobility data to reveal insights that could benefit operations [2], locate more customers [3], and forecast demand [4].

The *goal* of this study is to mine large-scale dynamic mobility data from connected vehicles to reveal potentially beneficial patterns that can help taxi services improve their business in the midst of growing competition from non-traditional shared mobility services. Privacy policies in many parts of the world require that, before making such data publicly available, the data owner must remove information that could identify persons. By contrast, the authors of this paper obtained a proprietary and unique dataset of Dubai taxi operations with the names of drivers replaced with a unique identification number. This enabled data mining to reveal driver-vehicle sharing patterns, which to the best of the authors' knowledge, is a gap in the available literature.

This paper *contributes* details of the Dubai case study, the proposed automated data cleaning method, and the *main finding* that a beneficial driver assignment pattern exists. This finding could inform tactics that encourage more of the beneficial assignments to help improve the efficiency and effectiveness of Dubai taxis, and perhaps shared mobility services in general.

Typical connected vehicle data from taxis include messages and variables such as geospatial position, meter-engaged, meter-vacant, door-open, idling, speed, timestamps, and dozens of other status indicators. The data size grows rapidly as tens of thousands of vehicles attempt to upload data packets every second to every few minutes. Aside from being so-called big data, dynamic mobility data can also be rather messy [5]. Data cleaning to enhance data quality is critically important in data mining, but the literature on data cleaning methods is sparse [6]. Manually cleaning large-scale mobility data is impractical. Therefore, a *primary objective* of this research is to develop and apply an automated method of data cleaning and filtering that is suitable for large-scale dynamic mobility datasets. A *secondary objective* is to develop a method for validating the quality of the cleaned dataset.

Dirty data from connected vehicles that operate in the vehicle-to-infrastructure (V2I) mode arises from many factors. They include unexpected malfunctions and various errors in the output of on-board sensors, trip meters, and the V2I communications system itself. In particular, standard global positioning system (GPS) receivers produce inaccurate or missing location coordinates because tall buildings and other occlusions distort or block the direct path of the satellite radio frequency signals [7]. Trip meters often encounter radio frequency interference as they attempt to upload data and receive acknowledgements. Hence, they tend to re-transmit and create duplicate records. Meter malfunctions or resets due to spurious electrical faults also produce inaccurate timestamps and odometer readings. To minimize the cost of data storage and communications, on-board systems seek to minimize the frequency and regularity of the geospatial position sampling [8]. Therefore, using filtering and interpolation techniques to reconstruct vehicle paths and speed profiles becomes ineffective [9].

Obvious errors such as missing GPS data and incorrect timestamps are easy to detect and remove. However, errors in trip length and trip durations are not as obvious. The literature lacks studies of automated methods to clean such non-obvious errors from large-scale dynamic mobility data. Work by others affirmed that the sampling variability of vehicle position data

reduces the accuracy of link travel time and route choice estimates [5]. In general, researchers found that the inaccuracy of GPS data requires some form of data cleaning for route estimation [10]. Other studies confirmed that the non-uniform sampling of GPS data results in large gaps that reduce the accuracy of recovery methods such as linear interpolation and historical averaging [11].

The organization of this paper is as follows: the next section (Methods and Data) describes the dataset in terms of its variables, its original structure, and a distillation process to restructure the data into trip records. The methods section also describes the three key filter variables and the *layered statistical filtering* technique that automatically eliminates non-obvious errors. Section 3 (Results) validates the quality of the cleaned data by comparing the overall distribution of the key variables with their expected theoretical distributions. The results section also describes the data mining results and reveals a potentially beneficial driver assignment pattern that maximizes trip productivity while minimizing overhead. Section 4 (Discussion and Conclusions) discusses the results, concludes the study, and describes future work to leverage the uniqueness of the dataset.

Methods and Data

The Road and Transport Authority (RTA) of Dubai, United Arab Emirates provided the authors with an exclusive dataset of their taxi activity. The data records combine information from both "dispatch-only" and "street-hailed" taxis. Emirates refer to the dispatch-only taxis as Hala taxis. Dispatchers often call on non-Hala taxis when Hala taxis are unavailable. Unlike publicly available dynamic mobility datasets such as those from New York City [12], the Dubai Taxi dataset contains the unique license plate number of each vehicle. The RTA anonymized the driver information by replacing their names with unique identifiers. A literature search indicates that this is the first study to report the results of mining dispatch-taxi mobility data that contain unique identifiers for both drivers and taxis.

Data Reduction and Restructuring

The dynamic mobility data obtained from Dubai taxis covered a 185-day period from March 15, 2016 to September 15, 2016. Analysis of the data revealed that Dubai taxi companies employed nearly 21,000 drivers who operated nearly 10,000 vehicles during that period. Dubai taxis provide service any time of day, every day. Each taxi has an on-board unit that contains a trip meter, a GPS receiver, and a wireless system that enables V2I communications. On average, the on-board unit transmits the status and position of the taxi approximately every minute. Subsequently, the dynamic mobility database annually accumulates more than five billion records, each with numerous variables.

In addition to the unique taxi and driver identifiers, each data record contains the fleet identifier, the vehicle status, its speed, its position in latitude and longitude, and a timestamp. The vehicle status indicates 45 different events, one of which is when a driver accepts a dispatch request. Therefore, the first step in data distillation was to extract records of dispatched versus street-hailed trips. The status also indicates instants of trip meter engagement and vacancy. When paired, this information forms a trip record containing the times and positions of pick-up and drop-off events. Hence, the second data reduction step was to extract only those records that indicate when the meter engaged and then became vacant. Subsequently, the data distillation

process substantially reduced the data size by building approximately 3.4 million trip records from the much larger dynamic mobility dataset.

Layered Statistical Filtering

The proposed layered statistical filtering technique incorporates three layers of filtering that use *known* likelihood distribution functions for trip duration, trip length, and average trip speed. The main concept of the proposed technique is that likely errors would be outlier records that also collectively do not contribute to the formation of an overall distribution that matches the expected distribution of a key variable. The *first* layer of filtering was the *trip duration*, which is the timestamp difference between paired drop-off and pick-up events. The *second* layer of filtering was the *trip length* in terms of total path distance. However, it was not directly available because of the non-uniform sampling of the geospatial positions. That is, the high variability of position update rates resulted in distance gaps that span several kilometres. Such large gaps made it impossible to determine the actual path taken between two points that frame the street grids. However, the geodesic distance between pick-up and drop-off positions provided a suitable proxy. The algorithm used the recursive Vincenty method to derive the geodesic distance [13]. The *third* layer of filtering was the *average trip speed*. However, it was also not directly available. Therefore, the authors developed a proxy dubbed the mobility index (MI), which is the ratio of the geodesic distance to the trip duration. Intuitively, the MI represents the average rate that taxis typically cover the geodesic distance between two geospatial coordinates.

Data Reduction and Cleaning Approach

The first step of the data cleaning method removed obvious errors. More than 1.5% of the records contained obvious errors such as zero trip times, negative trip times, duplicate records, and blank identifiers (Table 1). The second step was the layered filtering. Repeated attempts to remove records and re-test distributions for fitness is computationally intensive. Therefore, the strategy used was to run the algorithm on substantially fewer records, namely those within the lower and upper outlier 2.5 percentile. This approach effectively adds outlier records to the 95% confidence interval only if they likely contribute to an overall distribution that matches the expected distribution of the filter variable for that layer.

Previous research demonstrated that the expected distribution of trip lengths and trip times is lognormal [14], [15], [16]. Figure 1 shows the distributions of only the upper and lower 2.5 percentile of each filter variable. The algorithm then identified a cut-off point to eliminate those outlier records that likely fell outside of the overall expected distributions. For the trip duration distribution (Figure 1a), a significant number of records clustered very close to zero minute. A minimum slope analysis (first derivative) of that distribution identified a threshold of approximately one minute and one percentile (the arrow) below which it is unlikely that those outlier records (1% of the low outliers) belong to the overall expected distribution. The minimum slope, which was zero in this case, corresponded to the lowest point of a parabola that fits that portion of the distribution. Hence, the algorithm automatically eliminated them. In a similar manner, the upper 2.5 percentile (Figure 1b) had a large number of trips with durations near 62 minutes. The outliers that followed were obvious errors because their trip durations exceeded 60,000 minutes.

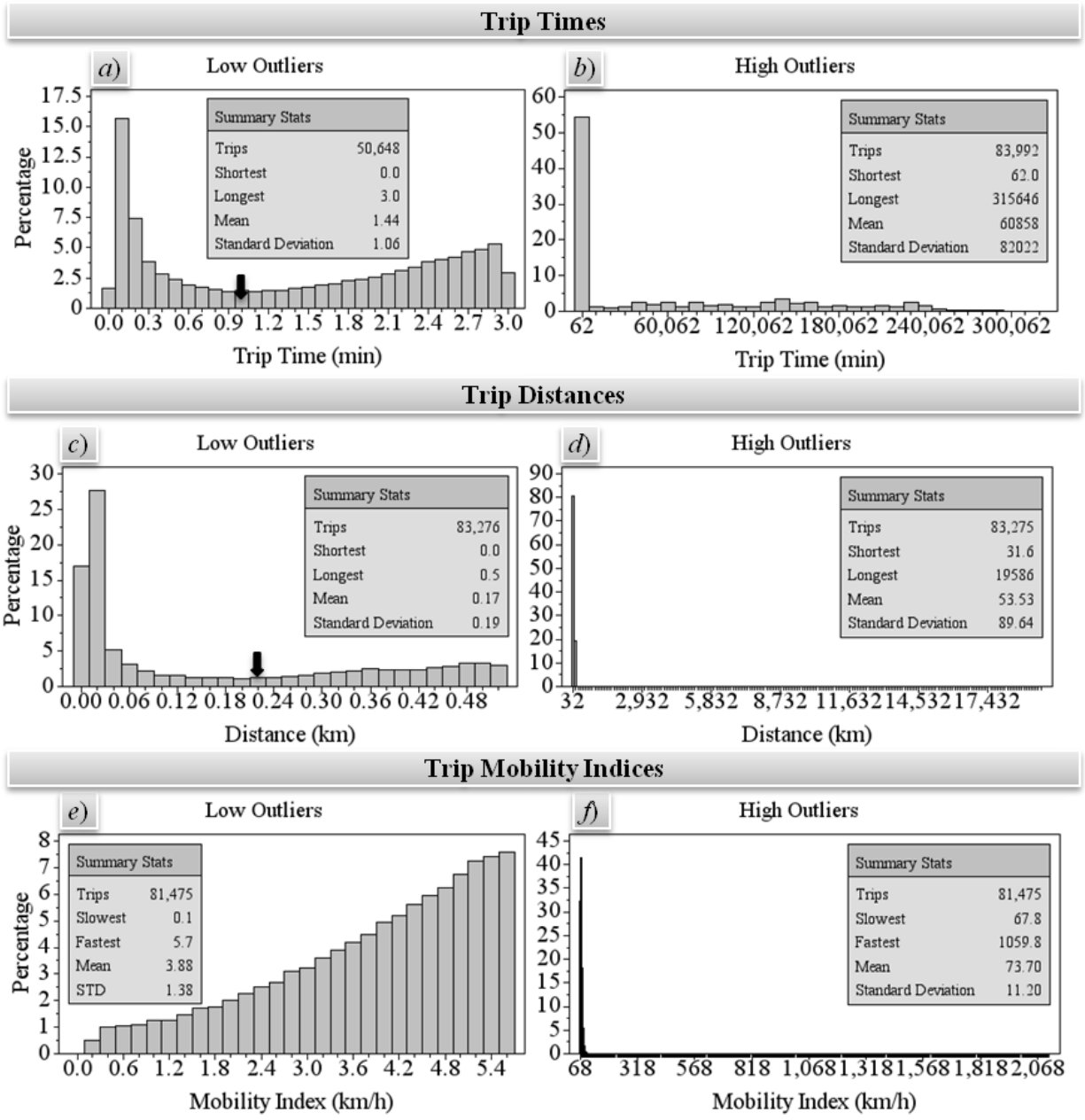

Figure 1: Distribution of outlier trip times, distances, and mobility indices.

An amplified view of the trip time distribution between one and five hours revealed a lognormal distribution followed by a very long tail of outlier records that represent less than 1% of the upper 2.5 percentile. The algorithm automatically eliminated records that did not contribute to the expected lognormal distribution. Possible sources for the extremely short trip duration errors may be passengers changing their mind about a trip after entering a vehicle, or electromagnetic noise interference in the trip meters. A possible source for the extremely long trip duration errors may be trip meter malfunctions that uploaded drop-off times from a memory buffer after restarting.

Table 1 Summary of Data Reduction and Cleansing

		Low Tail		High Tail		Reduction	
Records	Description	Count	%	Count	%	Count	%
3,444,310	185-day dispatched trip records						
3,444,304	Remove duplicate records					6	0.0002%
3,444,304	Remove invalid latitude or longitude					0	0.0%
3,391,795	Remove (pickup time) ≥ (drop-off time)					52,509	1.5%
3,391,770	Remove records missing a Driver ID					25	0.001%
3,331,032	Filter by trip time distribution	20,836	0.6%	39,902	1.2%	60,738	1.8%
3,259,001	Filter by trip distance distribution	52,400	1.6%	19,631	0.6%	72,031	2.2%
3,258,218	Filter by mobility index distribution	0	0.0%	783	0.02%	785	0.02%

Applying this statistical filtering technique to the *second* filtering variable further eliminated more than 70,000 records (Table 1). They included trip records with distances of approximately zero kilometres (Figure 1c), and obvious outliers that extended well beyond 32 kilometres (Figure 1d) that did not contribute to the expected lognormal distribution. GPS signal reflections from tall structures in Dubai may be a source for the unlikely trip distances.

In the third layer of filtering, the algorithm did not eliminate records in the lower 2.5-percentile of mobility indices (Figure 1e) because removing them did not move the overall distribution any closer to the expected theoretical distribution. However, the algorithm did eliminate outlier records in the upper 2.5-percentile tail (Figure 1f) with mobility indices that distanced the overall distribution from the expected distribution for that variable. This resulted in the elimination of records that exceeded mobility indices of 92 km h^{-1}. This is a reassuring result because the highest mobility index should be less than the highest speed limit in Dubai, which is 100 km h^{-1}. Table 1 summarizes that the overall data cleaning process, including the layered filtering, eliminated approximately 5.5% of the trip records, hence retaining those within a 94.5% confidence interval (Table 1).

Results

This section evaluates the effectiveness of the layered statistical filtering method by examining the degree of agreement between the distribution of the key variables of the remaining data and their expected theoretical distributions. This section also describes the results of the data mining on the cleaned dataset. A key lesson learned from this research is that the data collection, data preparation, and data cleaning efforts are far greater than that of the actual data mining. Consequently, the organization of this paper reflects those proportions.

Data Quality Evaluation

The criterion for evaluating the effectiveness of the proposed data cleaning method is the degree to which the distribution of the key variables of the cleaned data agrees with the expected theoretical distributions of the key variables. Figure 2 plots the distribution of the key variables of the cleaned data.

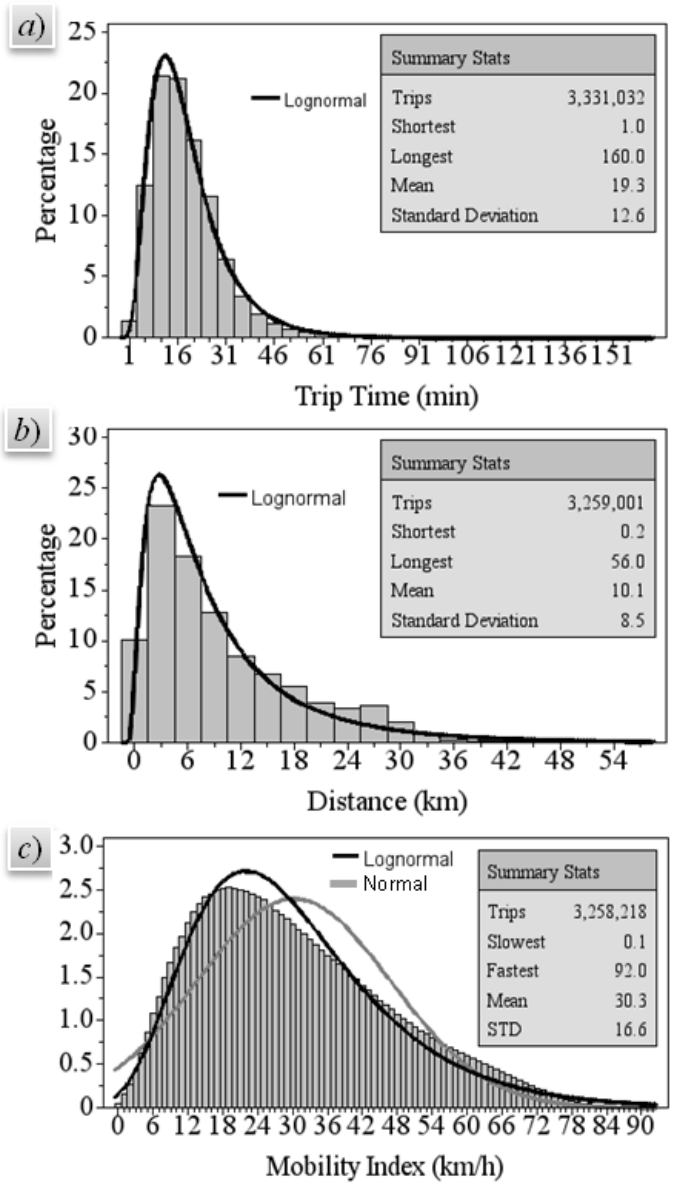

Figure 2: Distribution of trip times, distances, and mobility indices after cleaning.

The line plot is the continuous distribution that best fits the histogram of cleaned trip times (Figure 2a), geodesic distances (Figure 2b), and mobility indices (Figure 2c). The iterative Levenberg-Marquardt nonlinear least squares method of curve fitting identified the parameters of the best-fit distributions [17]. The model for estimating the best-fit lognormal distribution $D_{LN}(\xi)$ is

$$D_{\ln}(\xi) = \frac{\gamma_{\ln}}{\xi\sqrt{2\pi\sigma_{\ln}^2}} \exp\left[-\frac{1}{2}\left(\frac{\ln(\xi)-\mu_{\ln}}{\sigma_{\ln}}\right)^2\right]_{\xi>0} \quad (1)$$

The constants γ_{ln}, μ_{ln}, and σ_{ln} are estimates of the amplitude, mean, and standard deviation parameters, respectively. Trip distances are highly correlated to trip times, hence they distribute

similarly. The mobility index is a random variable derived from the ratio of the travel distance and the travel time, therefore, it also follows a lognormal distribution. Prior knowledge establishes that the mobility index cannot be zero or infinite because neither the travel distance nor the travel time will be zero or infinite. Therefore, the mobility index is limited to a finite interval. Table 2 lists the statistics of the key variables and parameters of the distributions that best fit their histograms. The variable ΔT is in minutes and ΔL is in km.

Table 2: Parameters of the cleaned distributions.

Histogram	ΔT	ΔL	MI
Bins	160	29	47
Mean	19.3	10.1	30.3
STD	12.6	8.5	16.6
Min	1.0	0.2	0.1
Max	160.0	56.0	92.2
Chi-Squared	Lognormal	Lognormal	Lognormal
$\chi 2$ *DF*	157	26	44
γ	101.6	234.6	214.4
μ	2.9	2.2	3.4
σ	0.6	1.3	0.7
$\chi 2$ Critical	187.2	38.9	60.5
$\chi 2$ Statistic	21.1	8.4	10.9
p-value (%)	100	100	100

The chi-squared goodness-of-fit test [19] indicates when there is a significant difference between the expected frequencies and the observed frequencies of the variables. The null hypothesis H_0 is that the observed distribution of the variables is the same as the candidate distribution. Failure to reject the null hypothesis will result in accepting the alternative hypothesis that there was no significant departure of the observed distribution from the candidate distribution.

The chi-squared statistic (χ^2) is

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \tag{2}$$

The random variables O_k are the histogram values observed in bin k and E_k are the expected values of the hypothesized distribution. The chi-squared test rejects the null hypothesis if the χ^2-statistic exceeds the critical χ^2 value of a chi-squared distribution evaluated at degrees-of-freedom *DF* and a specified significance percentage. Statisticians typically set the significance value to 0.05, which represents a low probability of 5% that the test will reject the null hypothesis when in fact it is true. The alternative approach calculates the chi-squared probability values (p-values) that correspond to the observed χ^2-statistic, evaluated at the *DF*. The tests reject the null hypothesis when the p-values are less than the selected significance percentage.

As shown in Table 2, the chi-squared tests could not reject the null hypothesis for the distributions tested. Therefore, the tests conclude that the trip times and the trip distances of the cleaned and filtered data do not depart significantly from the expected lognormal distribution. By

extension, the mobility indices do not significantly depart from the lognormal distribution because it is a dependent variable of the trip times and the trip distances. Subsequently, these tests validated the effectiveness of the proposed layered statistical filtering method of removing records that are likely erroneous.

Data Mining of Trip Production

Several different drivers can operate a taxi, and a driver can operate several different taxis. The data mining quantified the number of drivers that operated a unique taxi as the level of "taxi-split" and the number of taxis operated by a unique driver as the level of "driver-split." Given the uniqueness of the dataset, the focus of the data mining was to examine the distribution of taxi- and driver-split. Figure 3 captures the data mining results, which shows the distribution of taxi- and driver-splits by fleet type. For brevity, the figures show the frequency of *cases* for up to 20 taxis, but the maximum was actually 137. The pattern revealed was that the *productivity factor*, in terms of the number of trips-per-taxi-per-driver, peaked when only two to three drivers operated a given taxi. This two-driver taxi *case* dominated for non-Hala taxis whereas two- and three-driver taxi cases dominated equally for Hala taxis. These cases accounted for 33.5% and 15.2% of the non-Hala (Figure 3a) and Hala (Figure 3b) taxi-split cases, respectively. Scenarios of single-driver taxis accounted for only 2.5% and 0.4% of the non-Hala and Hala taxis, respectively. They were also among the least productive of cases in terms of the trips-per-taxi productivity factor.

The data mining results also indicate that a given taxi in Dubai sustains high trip production by assigning as many drivers to them as needed to minimize their parking times (Figures 3a-3b). However, driver changes incur significant overhead or off-duty time that reduces a taxi's trip production efficiency. The off-duty time includes time spent in depositing collected cash fares at specific bank locations, and then driving to designated locations to accommodate driver shift changes.

Using the productivity factor of trips-per-driver-per-taxi revealed that the factor peaked for cases of drivers operating a single taxi. These cases accounted for 35.3% and 20.5% of the non-Hala (Figure 3c) and Hala (Figure 3d) driver-split cases, respectively. Further inspection of the driver-split for the two-driver taxi pools (Figures 3a-3b) revealed that their operators came from the pool of drivers of a single taxi (Figures 3c-3d). By induction, trip productivity tends to peak when the same two drivers operate a given taxi. RTA was previously unaware that this pattern dominated. However, the pattern was not surprising because they provided two explanations for its commonness. Firstly, drivers can minimize their insurance costs by minimizing the number of different vehicles they operate. Secondly, the logistical complexity and time overhead involved with shift-changes increases with the number of drivers of a given taxi. The observed pattern along with the RTA explanation suggests that new tactics to encourage or facilitate more of the beneficial taxi assignment would likely lead to reduced overhead and enhanced trip production.

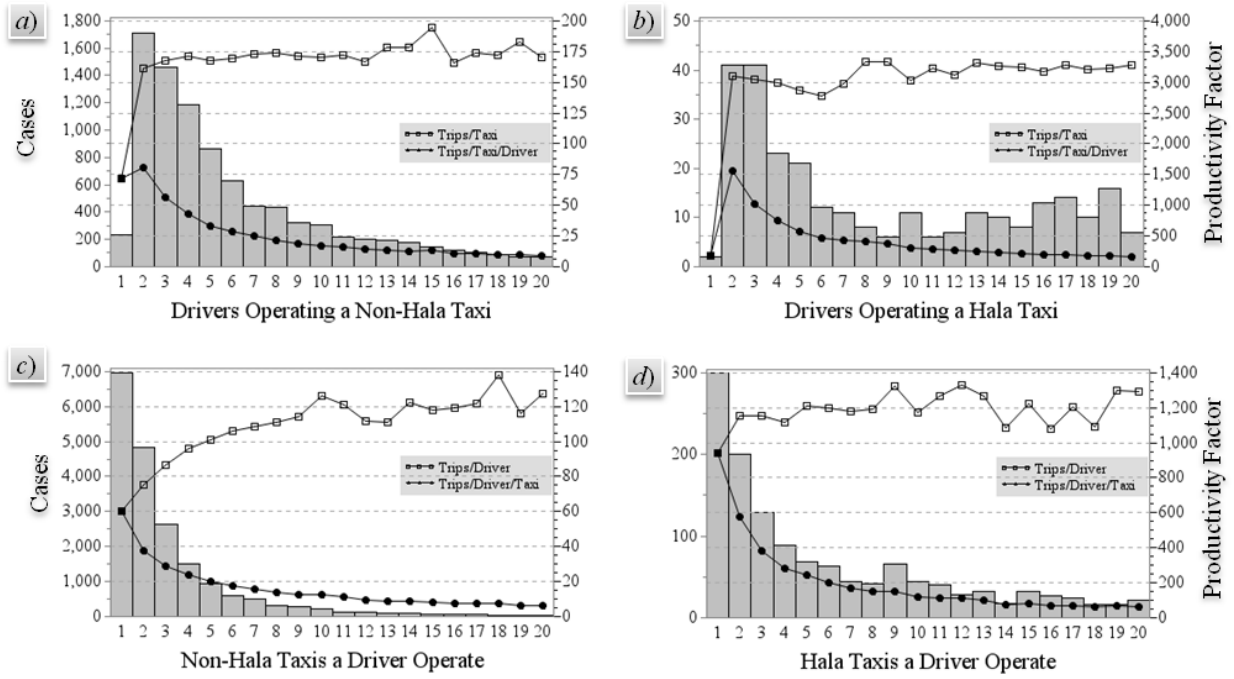

Figure 3: Distribution of taxi and driver splits by fleet type.

Discussion and Conclusions

The proliferation of shared mobility services worldwide and their growing variety has led to intense competition with traditional dispatch taxi services. Hence, the goal of this study was to mine large-scale dynamic mobility data from connected taxis to discover beneficial patterns that could inform tactics to improve the competitiveness of dispatch taxi services. However, the huge size, non-uniform composition, variable update rates, and GPS errors complicate the task of data mining. Therefore, the main *objective* and *contribution* of this research was to improve the quality of the dataset by developing an automated data cleaning and filtering method, tailored for such datasets. During the course of this research, the authors learned that the data collection, data preparation, and data cleaning efforts are far greater than that of the actual data mining effort. Therefore, the organization and relative focus of this paper reflects their relative magnitude.

The proposed *layered statistical filtering* algorithm automatically eliminated *outlier* records that contained both obvious errors and likely errors. The *main idea* of the technique was to remove records that moved the distributions further away from the known theoretical distributions of the key filter variables. Validation of the quality of the cleaned data used chi-squared statistical tests to compare the distribution of the three variables with their expected distributions. The tests determined that the overall cleaning procedure, including the filtering algorithm, removed obvious outliers and other poor quality records that represented approximately 5% of the dataset.

Subsequently, the data mining focused on examining taxi trip production as a function of taxi-driver pairing patterns. Such an analysis would not be possible without the uniqueness of the dynamic mobility dataset, which includes identifiers to distinguish individual drivers. The revealed pattern was that taxi trip production peaks for the case where only the same two to three drivers operate the same taxi. The RTA explanation was that drivers could minimize their

insurance costs by minimizing the number of different vehicles that they operate. Fewer drivers per vehicle also reduce the logistical complexity and the time overhead of shift-change and cash deposit procedures. Hence, taxi companies in Dubai can use this finding to develop tactics that would encourage more of the beneficial assignment pattern. At this point, it is unknown whether similar patterns exist for dispatch taxi services in other cities of the world.

A limitation of the proposed layered data filtering method is that it relies on known statistical distributions of the selected filter variables. This necessitates the transformation of dynamic mobility data into trip records containing the timestamps and geospatial positions of trip origins and destinations.

Future research will mine the Dubai taxi data to characterize the spatial-temporal dynamics in supply and demand to guide decisions in zonal taxi allocations. The authors will also investigate various methods of predictive analysis to guide driver recruitment, fleet acquisition, network management, scheduling, and revenue management decisions.

Data Availability

As previously described, the unique dataset used in this research is proprietary. Special arrangements with the Dubai Road and Transport Authority are necessary to gain access.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

A grant from the University of Modern Sciences, Dubai, United Arab Emirates, supported this research.

Acknowledgements

The authors are grateful to the Dubai Road and Transport Authority for their review, appreciation, and feedback on the research outcome.

References

[1] Y. M. Nie, "How can the taxi industry survive the tide of ridesourcing? Evidence from Shenzhen, China," *Transportation Research Part C: Emerging Technologies,* vol. 79, no. Online, pp. 242-256, June 2017.

[2] M. W. Ulmer, L. Heilig and S. Voß, "On the Value and Challenge of Real-Time Information in Dynamic Dispatching of Service Vehicles," *Business & Information Systems Engineering,* vol. 59, no. 3, pp. 161-171, 2017.

[3] X. Hu, S. An and J. Wang, "Taxi Driver's Operation Behavior and Passengers' Demand Analysis Based on GPS Data," *Journal of Advanced Transportation,* vol. 2018, 2018.

[4] A. W. Smith, A. L. Kun and J. Krumm, "Predicting taxi pickups in cities: which data sources should we use?," in *Proceedings of the 2017 ACM International Joint*

Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, 2017.

[5] K. Liu, T. Yamamoto and T. Morikawa, "An analysis of the cost efficiency of probe vehicle data at different transmission frequencies," *International Journal of Intelligent Transportation Systems Research,* vol. 4, no. 1, pp. 21-28, 2006.

[6] T. Dasu and T. Johnson, Exploratory data mining and data cleaning, vol. 479, New York: John Wiley & Sons, 2003.

[7] P. D. Groves, L. Wang and M. Ziebart, "Shadow Matching: Improved GNSS Accuracy in Urban Canyons," *GPS World,* vol. 23, no. 2, pp. 14-18, 2012.

[8] T. Miwa, D. Kiuchi, T. Yamamoto and T. Morikawa, "Development of map matching algorithm for low frequency probe data," *Transportation Research Part C: Emerging Technologies,* vol. 22, pp. 132-145, 2012.

[9] J. Liu, X. Yu, Z. Xu, K. R. Choo, L. Hong and X. Cui, "A cloud-based taxi trace mining framework for smart city," *Software: Practice and Experience,* vol. 47, no. 8, pp. 1081-1094, 24 August 2016.

[10] H. J. v. Zuylen, F. Zheng and Y. Chen, "Using probe vehicle data for traffic state estimation in signalized urban networks," in *Traffic Data Collection and its Standardization*, J. Barceló and M. Kuwahara, Eds., New York, Springer, 2010, pp. 109-127.

[11] Z. D. Y. T. Z. Q. H. a. X. L. Zhang, "A study on the method for cleaning and repairing the probe vehicle data," *IEEE Transactions on Intelligent Transportation Systems,* vol. 14, no. 1, pp. 419-427, 2013.

[12] C. Yang and E. J. Gonzales, "Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data," in *Seeing Cities Through Big Data*, P. ". Thakuriah, N. Tilahun and M. Zellner, Eds., New York, New York: Springer International Publishing, 2017, pp. 405-425.

[13] T. Vincenty, "Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations," *Survey Review 23,* vol. 23, no. 176, pp. 88-93, 1975.

[14] H.-C. Chu, "An empirical study to determine freight travel time at a major port," *Transportation Planning and Technology,* vol. 34, no. 3, pp. 277-295, April 2011.

[15] N. Wu and J. Geistefeldt, "Standard Deviation of Travel Time in a Freeway Network--A Mathematical Quantifying Tool for Reliability Analysis," in *CICTP 2014: Safe, Smart, and Sustainable Multimodal Transportation Systems*, Changsha, China, 2014.

[16] E. Durán-Hormazábal and A. Tirachini, "Estimation of travel time variability for cars, buses, metro and door-to-door public transport trips in Santiago, Chile," *Research in Transportation Economics,* vol. 59, pp. 26-39, November 2016.

[17] P. E. Gill, W. Murray and M. H. Wright, Practical optimization, London: Academic Press, 1981.

[18] C. Forbes, M. Evans, N. Hastings and B. Peacock, Statistical distributions, Hoboken, New Jersey: John Wiley & Sons, 2011.

[19] A. Papoulis, Probalility, Random Variables, and Stochastic Processes, New York: McGraw-Hill, 1991.

[20] A. Stocker, C. Kaiser and M. Fellmann, "Quantified Vehicles," *Business & Information Systems Engineering,* vol. 59, no. 2, pp. 125-130, 2017.

[21] T. Teubner and C. M. Flath, "The economics of multi-hop ride sharing," *Business & Information Systems Engineering*, vol. 57, no. 5, pp. 311-324, 2015.