World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:15, No:11, 2021

Dissecting Big Trajectory Data to Analyse Road Network Travel Efficiency

Rania Alshikhe , Vinita Jindal

Open Science Index, Computer and Systems Engineering Vol:15, No:11, 2021 publications.waset.org/10012318.pdf

Abstract—Digital innovation has played a crucial role in managing smart transportation. For this, big trajectory data collected from traveling vehicles, such as taxis through installed global positioning system (GPS)-enabled devices can be utilized. It offers an unprecedented opportunity to trace the movements of vehicles in fine spatiotemporal granularity. This paper aims to explore big trajectory data to measure the travel efficiency of road networks using the proposed statistical travel efficiency measure (STEM) across an entire city. Further, it identifies the cause of low travel efficiency by proposed least square approximation network-based causality exploration (LANCE). Finally, the resulting data analysis reveals the causes of low travel efficiency, along with the road segments that need to be optimized to improve the traffic conditions and thus minimize the average travel time from given point A to point B in the road network. Obtained results show that our proposed approach outperforms the baseline algorithms for measuring the travel efficiency of the road network.

Keywords—GPS trajectory, road network, taxi trips, digital map, big data, STEM, LANCE.

I. Introduction

TRANSPORTATION problems have a major impact on the life of people as they spend significant time commuting either for their daily needs or for entertainment purposes. These problems can be related to traveling time, distance, and/or fuel consumption. To handle these problems, the government uses many measures from time to time. They also use various digital technological innovation and provide efficiency, productivity, safety, security, and environmental outcomes for solving transport-related problems. These solutions include detecting congestion and collisions in roadworks, provides motorists with alerts for accidents and re-routing suggestions, reducing travel times, minimizing fuel consumption and energy demands, and making better use of existing infrastructure. Thus, directing our information technology capabilities towards achieving these goals is both urgent and demands time to ensure the efficiency of the road networks. Over the past decade, enormous amounts of trajectory data have been collected from traveling vehicles, such as cars, buses and, taxis, etc. due to the global positioning system (GPS)-enabled devices installed in these vehicles. This data offers a terrific opportunity to trace the movements of vehicles in fine spatiotemporal granularity. Even the researchers in their work have explored both trajectory management and trajectory data mining for finding the solution of many transport-related problems. Studies on trajectory management include trajectory compression and storage-related issues, whereas trajectory data mining includes methods and applications related to the field of data mining. Further, trajectory analysis is one of the essential parts of data mining, with a wide range of problems that have been addressed using this analysis. Examples of some problems solved using trajectory data analysis are finding the optimal path, enhancing the transportation service, anomaly detection, and causality analysis to name a few.

This study aims to address the inimitable challenges in trajectory data analysis. The resulting discoveries will provide (i) unparalleled evidence of the travel efficiency in road networks, (ii) a reliable causality analysis behind low travel efficiency in road networks across an entire city. The outputs of this paper are original and will benefit a wide range of stakeholders, from public transport service providers and urban planners, logistics businesses, and individual travelers. To achieve highly desirable solutions, the methodologies for the intended research tasks will be specifically designed to dominate any baseline solutions. Further, the solutions will target the common problems associated with trajectory analysis, such as sparse data. Properly handling these problems is essential, regardless of how advanced the analytic technologies to be applied are. In this sense, the research in this paper will contribute beyond the scope of the prescribed tasks and shed light on the field of entire trajectory data analysis.

The innovative aspects of this paper are presented in detail as follows. The first innovation is proposed statistical-based travel efficiency measurement (STEM). Taxis can be viewed as a sample of all vehicles (i.e. a sample population), and the trajectory data from taxis are, in most cases, highly sparse in fine spatiotemporal granularity, even though they can be relatively dense for some time periods and regions. Instead of a simple mean to measure traffic conditions, such as what is done in all existing studies, a null hypothesis test that considers the number of taxis and the sample variance will be applied. As a result, although calculating an exact mean of the population will still be difficult, the mean of the population can be reliably compared with a given value. To this end, the series of unique challenges to be addressed include but are not limited to the design of statistical tests, the efficiency of the process to measure numerous origin-destination (OD) pairs. Second, we have proposed the least square approximation network-based causality exploration (LANCE). To the best of our knowledge, this is the first study in this promising area. The resulting research using this novel method will be quite innovative. However, the primary objective is to determine how much each road segment contributes to the low travel efficiency in a road network by factorizing the relationships between all OD pairs and all road segments. Effective solutions must properly address the following situations: (i) when there are typically

R. Alshikhe is with the School of Computer Science and Information Technology, Royal Melbourne Institute of Technology University, 124 La Trobe St, Melbourne VIC 3000. (rania.alshikhe@student.rmit.edu.au)

V. Jindal is with the Department of Computer Science, Keshav Mahavidyalaya, University of Delhi. (vjindal@keshav.du.ac.in)

multiple paths for each OD pair, (ii) when the number of OD pairs is much greater than the number of road segments, (iii) when the trajectory data collected for some road segments are sparse, (iv) when the representation of travel efficiency in the matrix has an impact on the results and (v) when the data volumes are very large.

The remaining of this paper is organized as follows: Section II presents the related work followed by the proposed methodology discussed in Section III with the proposed statistical method to measure the travel efficiency is described. Also, it gives the details of the proposed method for network-based causality analysis method. Then Section IV presents experimentation. Next, the result and discussion of experimentation are presented in Section V that is followed with the conclusion in Section VI.

II. Related Work

Traffic anomalies are caused by various events, such as accidents, campaigning, protests, and sporting events. These anomaly detections along with causality analysis can be used for the trajectory data investigation. Several studies exist to solve this problem by identifying the spatiotemporal outliers that significantly deviate from normal conditions. In [1], the authors argue that recent developments in high-precision GPS technologies and infrastructure have made urban centers and cities smarter. According to the authors, one of the primary applications of urban traffic analysis is for detecting traffic anomalies based on traffic data. The detection is based on the outlier detection techniques, which involve observing inconsistencies on a set of traffic data. The authors' categories existing outlier detection approaches into two main groups, flow outlier detection, and trajectory outlier detection. Whereas in this paper, we detect the road segment with an indicator of the high contribution values that cause low travel efficiency. We review two strands of research that are relevant to our research. These are (i) statistical modeling of the traffic network and (ii) mathematical modeling of the traffic network.

The statistic is considered a tool for predicting unexpected failures. A statistical anomaly detector pursuit for things that seem unusual and then sends an alert. The research [2], [3] highlights the flaws of using GPS trajectories of taxicabs traveling in urban areas. The study is evaluated by utilizing real data generated by Beijing in 2009 and 2010. The paper identifies the problem of less effective planning in a city due to GPS trajectories of taxicabs. It utilizes "Traffic modeling", "Flaw detection" and "Real evaluation" techniques for detection. Firstly, a matrix with a set of features representing the connection between two regions is created for the taxi traffic each day. Secondly, the flawed region pairs are detected from the matrix. The pattern of these pairs is further analyzed using graphs and sub-graphs. The result from the graph represents both flawed planning and the relationship between them. Lastly, the method is evaluated using Beijing traffic data. The technique relies on calculating a simple mean for the traffic volume, speed, and detour ratio of all taxis in the city for every origin-destination pair. The means are then compared to the average conditions across the city. Another study [4], [5] also considering the mean value as the measure of travel efficiency and route with high variation from the mean value are considered anomalous routes or links depending on the problem definition. The study [4] has grouped existing techniques into different categories like classification, clustering, statistical, information-theoretic, and spectral categories. For each category, the author has identified key assumptions, which are used by the techniques to find the anomalous traffic routes. Anomaly detection by partitioning a city into a grid and then likelihood ratio test statistic is applied in the study to detect the traffic anomalies by using the count of the number of taxis in the grid cells [5], [6]. In another, study [7], PCA is used to detect anomalies based on taxi trajectories. Authors in [8] used the tensor factorization for anomaly de-tection, considering the network traffic as a tensor. Switching hidden Semi-Markov models for recognizing and detecting anomalies has been proposed in [9]. Later a regression-based anomaly detection method called the GARCH model became a popular method of anomaly detection in traffic networks [10]. In, TRAjectory Outlier Detection (TRAOD) method, each trajectory is partitioned into smaller segments called a base unit of trajectory. Then the adjusting coefficient of each partition is computed if the computed value is greater than 1; the partition of the trajectory is tagged as an outlier. Otherwise, it is considered as a normal partition [11]. In Prediction-based anomaly detection setting, prediction errors obtained from the LSTM model are fit to a Gaussian distribution. The mean and variance of the distribution are computed using the Maximum Likelihood Estimation [12]. In the study of the Trajectory Fragment Outlier (TF-Outlier) method, local difference density is used to determine the trajectory fragment outliers [13].

With the high dimensionality dataset, the complexity of the anomaly detection increases. So, to reduce the dimension and provide accurate methods to process the data, mathematical perspective is important. The research [14], [15] aims to propose a framework that helps to infer the anomalies that appear in the road traffic data. The results of the study are evaluated using Beijing taxi data observed for over three months. The framework utilizes a two-step process, such as mining and optimization. This helps to investigate the possible reasons behind the anomalous behavior of the road traffic data. Further, the data is modeled to study the traffic between regions. The road network is modeled as a directed graph (N) with a set of regions (v) bounded by major roads. A set of directed links (L) that connect two regions. A binary matrix (A) is used to define the relationship between the links and the routes. A link flow vector (b) contains traffic flow information of the links, and a vector (x) contains the flow of information of the routes. Underbalanced conditions the relationship between A, b, and x can be modeled as $Ax = b$. The mathematical modeling of the traffic network also takes into consideration the congestion on each link (b_i) using a delay function $D_i(b_i)$. For each link, the delay function is calculated, and it is generally increasing function of (b_i). Hence the time travel for each route (ji) is given by summation $(D_i(b_i)ji)$. Mining is performed by applying the PCA technique to mine for link anomalies from L, and then the L1 optimization technique is applied on $Ax = b$ to identify the routes that might have caused these anomalies. Finally, the results are evaluated by utilizing a large GPS trace consisting of nearly eight hundred million data points. Various Mathematical algorithms have been developed to support the different scenarios of anomaly detection. The road segment weight is supported by the A* algorithm and Dijkstra's algorithm, where the road segment weight corresponds to the travel distance of the segments. Recommending the fastest path relies on the modeling of road segments based on the travel times. The algorithm can find the fastest route when modeling the travel time on segments as a linear function [16]. Another algorithm discussed in [17] uses the road segment weighted value to indicate the driver preference. The authors provide techniques that find the appropriate edge weights reflecting how the driver would like to use the edges based on the selected real trajectory dataset from 52,211 taxis in Beijing as a personal reference. In [18], the Select Link Entanglements (SELES) algorithm is used to find the crucial road segment as link pairs in the anomalous region through the optimization link entanglement search algorithm. They detect the occurrence of anomalies through a high-deviation link pair. Then the link entanglement was found to check the effect of the high deviation ratio by using the SELES. Finally, the crowd's driving routes were analyzed to infer and explain the anomalies. They used a real-world GPS dataset created by tens of millions of D_i taxi drivers for seven days. All the aforementioned proposals do not explore how to detect the road segment value nor weight from past trajectories to improve the travel efficiency. Further, they do not consider the value of contribution of each road segment for the whole city. In contrast, our proposed work utilize the drivers' past trajectories to identify the road segment contribution of the low travel efficiency for all entire city not to personal drivers' reference or crowd driving paths.

III. Proposed Methodology

This paper implements the methodology as described in Fig. 1, showing the method architecture that include both the proposed STEM and LANCE methods. Here, first the database is partitioned according to location IDs and regions. Then, the travel efficiency of each unique origin-destination pair is obtained, and vector l is created using the proposed statistical travel efficiency measure (STEM) method. The travel efficiency of an entire city road network is measured by evaluating the travel efficiency between all locations across the city. In this step, data preparation is already completed then travel efficiency measure is discussed and the proposed STEM method is described in detail in the next subsection. Next, matrix A (relation between road segment and OD pairs) is created from the dataset, and finally, road segments, causing the low travel efficiency are identified using the proposed least square approximation network-based causality exploration (LANCE) method. The network-based causality analysis used in LANCE method to identify which road segments are the major causes of low travel efficiency in an entire city road network with high contribution values. In the next subsection, mathematical formulation of causality analysis and proposed least square approximation network-based causality exploration (LANCE) has been described in detail.

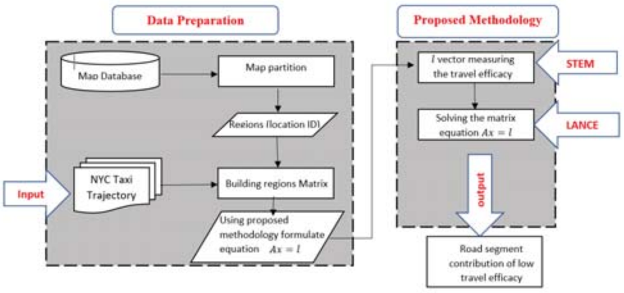

Data Preparation
Proposed Methodology
Map Database
Map partition
Regions (location ID)
Input
NYC Taxi Trajectory
Building regions Matrix
Using proposed methodology formulate equation $Ax = l$
l vector measuring the travel efficacy
STEM
Solving the matrix equation $Ax = l$
LANCE
output
Road segment contribution of low travel efficacy

Fig. 1: The Method Architecture

A. Proposed Statistical Travel Efficiency Measure (STEM) Method

A new method to reliably measure the travel efficiency of each OD pair has been developed, given that using a simple mean of taxis' travel speeds is an unreliable way to approximate the travel efficiency of all vehicles. It has been assumed that taxis may take different paths for various reasons; as a result, some will have high travel speeds, and some will travel more slowly. More importantly, taxis represent a small fraction of all vehicles, and the trajectory data are typically sparse for a particular OD pair at a given time. The number of taxis and the variance in their travel speeds play a critical role in estimating the travel efficiency of all vehicles. Hence, using a statistical test provides a far more reliable measure of travel efficiency. Determining the exact average travel speed of all vehicles is still difficult, but by using the proposed STEM, determining the travel efficiency of an OD pair is higher or lower than a given value is possible.

Let A and B be two locations. The average travel speed of taxis from A to B is $\bar{X}_{AB}$ and for all vehicles, it is $\bar{U}_{AB}$. The valid speed range is uniformly split into units, for example, $s_1 = 10\ km/h, s_2 = 20\ km/h, \ldots s_n = 100\ km/h$. Given any $s_i(1 < i < 10)$, the goal is to determine whether $\bar{U}_{AB} < s_i$ with high confidence (typically 95%) by using a statistical test based on the travel speeds of the taxis from A to B. As seen the, null hypothesis of the test is $H_0 = \bar{U}_{AB} \geq s_i$, and the alternative hypothesis is $H_A = \bar{U}_{AB} \leq s_i$. The population is not assumed to be distributed in a normal fashion, and the population variance is assumed to be unknown. A Z-test is appropriate once the number of taxis increases to 30 or more [8], [9] (the central limit theorem is in effect); otherwise, a T-test is appropriate.

The null hypothesis of the test is $H_0 = \bar{U}_{AB} \geq s_i$

The alternate hypothesis of the test is $H_A = \bar{U}_{AB} \leq S_i$

With high confidence (typically 95%)

$$Z = \frac{\bar{x}-\Delta}{\frac{\sigma}{\sqrt{n}}}$$

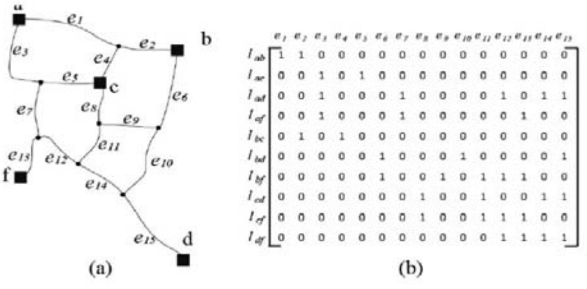
Where, $\bar{x}$ is the sample mean, Δ is the value of speed range, σ is the standard deviation, n is the number of values in the sample set. The objective of this travel efficiency measure is to find the minimum value of s_i that satisfies $\bar{U}_{AB} < s_i$. This s_i represents the minimum upper bound of the average travel speed of all vehicles from A to B. For example, s_i = 30 km/h indicates that it is generally impossible for the average travel speed of all vehicles from A to B to be over 30 km/h. Thus, the minimum upper bound is a reliable travel efficiency indicator for any OD pair. If it is less than a threshold τ, the travel efficiency is considered to be low efficiency. The next, we discuss the proposed LANCE method in detail.

B. Proposed Least Square Approximation Network-Based Causality Exploration (LANCE) Method

The proposed STEM aims to resolve the contributions of individual road segments to the low travel efficiency of OD pairs. Given that any road segment r may be found in the travel paths of different OD pairs, in general, if the travel efficiency of these OD pairs is high, the more likely it is that r has nothing to do with the low travel efficiency. However, if the travel efficiency of all the OD pairs is low, the more likely it is that r is the cause. As shown in Fig. 2a, matrix A specifies the relationships between the road segments and the OD pairs in Fig. 2b.

Given a matrix A entry a_{ij} = 1 means that the j^{th} road segment is in the travel path of the i^{th} OD pair; otherwise, a_{ij} = 0. For example, the road segment e_{12} occurs in the paths $a \rightarrow d, d \rightarrow f, b \rightarrow f$,and $c \rightarrow f$. If the travel efficiency of all these paths is low, e_{12} is likely to be the cause. The road segment e_{14} occurs in the paths $a \rightarrow d, d \rightarrow f$ and $c \rightarrow d$. If the travel efficiency of $a \rightarrow d$, $d \rightarrow f$ is low, but that of $c \rightarrow d$ is high, The proposed STEM aims to resolve the contributions of individual road segments to the low travel efficiency of OD pairs. Given that any road segment r may be found in the travel paths of different OD pairs, in general, if the travel efficiency of these OD pairs is high, the more likely it is that r has nothing to do with the low travel efficiency. However, if the travel efficiency of all the OD pairs is low, the more likely it is that r is the cause. As shown in Fig. 2a, matrix A specifies the relationships between the road segments and the OD pairs in Fig. 2b.

Given a matrix A entry a_{ij} = 1 means that the j^{th} road segment is in the travel path of the i^{th} OD pair; otherwise, a_{ij} = 0. For example, the road segment e_{12} occurs in the paths $a, d \rightarrow f, b \rightarrow f$ and $c \rightarrow f$. If the travel efficiency of all these paths is low, e_{12} is likely to be the cause. The road segment e_{14} occurs in the paths $a \rightarrow d, d \rightarrow f$ and $c \rightarrow d$. If the travel efficiency of $a \rightarrow d, d \rightarrow f$ is low, but that of $c \rightarrow d$ is high, e_{14} is less likely to be the cause of the low travel efficiency, compared with e_{12}. As such, this task involves disclosing the extent to which each road segment contributes to low travel efficiency. The matrix A represents the relationship between n OD pairs and m road segments m×n, as shown in Fig. 2(b).



$$\begin{array}{c|ccccccccccccccc}
 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} & e_{13} & e_{14} & e_{15} \\
\hline
l_{ab} & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
l_{ac} & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
l_{ad} & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
l_{af} & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
l_{bc} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
l_{bd} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
l_{bf} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
l_{cd} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\
l_{cf} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
l_{df} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{array}$$

Fig. 2: Road network and matrix A

$l = \{l_1, l_2, \ldots l_m\}^T$ is a column vector denoting a set of m, OD pairs, where entry $l_i = 1$ if the travel efficiency of the i^{th} OD pair is low; otherwise, $l_i = 0$. $x = \{x_1, x_2, \ldots x_n\}^T$ is a column vector denoting a set of n road segments, where entry $x_j \epsilon [0, 1]$ indicates the contribution of the j^{th} road segment to the low travel efficiency of the entire city road network.

For l reveals that despite the same poor traffic conditions, individual road segments do not contribute equally to the travel efficiency of the network. It models explicitly and implicitly that a segment contributes more if, compared with other road segments, it is (i) involved in more location pairs with low travel efficiency, (ii) involved in fewer location pairs with normal travel efficiency, and (iii) relatively longer. In the equation $Ax = l$, where the matrix A contains taxi trips with crossing road segments, the vector x is the indicator of the travel efficiency, and the vector l is the measuring of the travel efficiency $l_i = 1$ if the travel efficiency of the i^{th} OD pair is low; otherwise, $l_i = 0$. Here, we have A and l to find x, but the equation $Ax = l$ has no solution since the matrix A has more rows than columns. There are more equations than unknowns. However, the proposed least-squares approximation method, uses an approximate norm solution of $Ax \approx l$, was proposed to solve the problem of high accuracy and reliability. This makes it possible to add constraints to the basic norm approximation problem. Here, the vector x must satisfy positive values to be meaningful. The proposed LANCE focus to solve the matrix equation by using (1).

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 = \Sigma_{i=1}^{m}(a_i^T x - b_i)^2 \\ \text{subject to} & l \le x \le u \end{array} \tag{1}$$

A bounded least-squares approximation problem is an optimization problem [10] with constraints($x > 0$), the form where with$m \geq n$ and are problem data, are the rows of A, the vector is the optimization variable and are problem parameters. CVXPY is used to implement mathematical expressions. It is a Python-embedded modeling language for convex optimization problems. It automatically transforms the problem into standard form, calls a solver, and unpacks the results [19].

IV. Experimentation

For the experimentation, this paper explores the trajectory datasets of the New York City (NYC) Taxi and Limousine Commission, containing the data of over 8 million individuals in yellow taxi trips in the city for the month of January 2018 [6]. Each record includes the date, time, and location ID of the pick-up and drop-off, the trip distance, the itemized fare, the rate type, and payment method, and the passenger count, as reported by the driver example GPS trajectory data. Travel routes are not included in this dataset; hence, each record is referred to as a trip or trajectory. Given an OD pair in New York City, the travel path would need to be recovered using a navigation system, such as TomTom or Google Maps, based on the assumption that navigation systems are widely used by drivers nowadays. We first give the details of data preparation as below.

A. *Data Preparation*

During the data preparation the task, aim is to identify valid origins and destinations from a trajectory dataset and retrieve the trajectories for each OD pair. As the total number of pick-up/drop-off locations in a trajectory dataset can be very large, valid origin/destination pairs are used to represent the pick-up/drop-off locations that are near to one another in spatial space. Therefore, all the pick-up/drop-off locations by the taxis will be identified from their trajectories by (i) location ID and (ii) all zones in New York City, including all-important location. The trajectory dataset of New York City specifies 262 zones [7], as seen in the below Fig. 3.

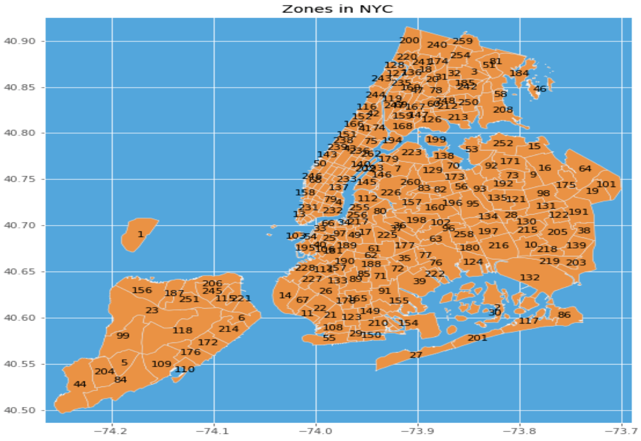


Fig. 3: The Zones in New York City

We conduct our research based on location ID and zones (small regions) for two reasons. First, zones contain rich information about the people existing there and their travel, which are the sources of the problem. Second, the flaws represented by zones contribute to both transportation planning and land use. For instance, if the connection between two zones is determined to have low efficiency, the possible solution to this after identifying the road segment could be building new roads between them (about transportation planning) or adding some businesses, such as shopping malls, in the zone outsourcing people (i.e. land-use planning) [7]. In the next section, we discuss about the results obtained.

V. Results and Discussion

All the experiments were carried out using our proposed methods for both the cases of the NYC Taxi Trips dataset. The first case covers all OD pairs in the city and the second case highlights the most popular suburb of New York City such as Manhattan and Queens, to explore the road segment, which has a high contribution to cause low travel efficiency. The result of each cases has been summarized in the following subsections.

A. *Results of the Case study of NYC Taxi Trips*

The dataset is segmented by the count number of taxis grows to 30 or larger as the final dataset contains 7211 taxi trips with unique OD paths. Our dataset increased one column display the minimum average speed of the population based on the taxi dataset as a sample to approximate measure the travel efficiency for each taxi trips to create the column vector $l = \{l_1, l_2, \ldots l_m\}^T$ is denoting a set of m OD pairs, where entry $l_i = 1$ if the travel efficiency of the i^{th} OD pair is low less than 30 km/h; otherwise, $l_i = 0$. As a result, approximately 19% of taxi trips are considered as low travel efficiency with a speed range between $20 - 30$ km/h (see Fig. 4).

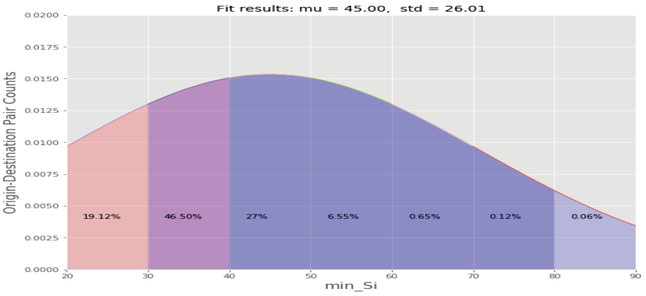


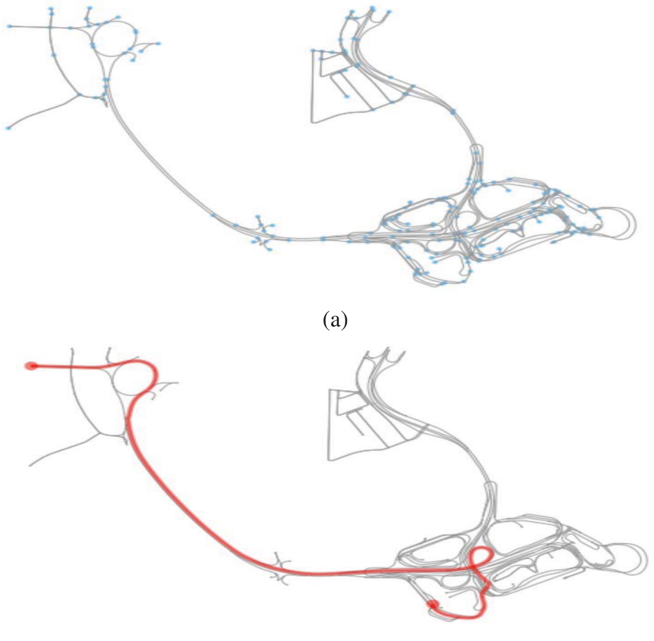
Fig. 4: Normal Distribution Data

Fig. 5a is the base map from the Open Street Map (OSM) system [11], which covers the road segments on the map, a node n is a single point described by its geographic coordinates, a unique identifier, and metadata (“tags”): , where ϕ_i is the latitude, λ_i is the longitude, i is the node identifier.

In Fig. 5b, we connected the route path along with the trajectory dataset. As a result, Fig. 6a shows an overview of all OD path coverage of all the main urban districts of the NYC network map. After creating the matrix which reflects the relationship of the taxi trips and unique road segments passed on all trajectory data. The matrix contains 7211 ODs as a row and 1628 road segments as columns. The vector$l$ represents around 53% of the data of taxi trips recorded as low travel efficiency as 3855 taxi trips recorded as 1 otherwise is 0. With matrix A and vector l are known, and the equation is solved for x vector reflect as an indicator of travel efficiency. The view of visualization of the road segment heatmap based on the result of indicator of travel efficiency display as the road network heatmap scale value from 0 to 1, So start from minimum values represent on the yellow segments then orange segments at the end the red segments with maximum values (see Fig. 6b (all road segments) and Fig. 6c (zoom of part of red road segments)).

B. *Result of the second case Taxi Trips Manhattan - Queens*

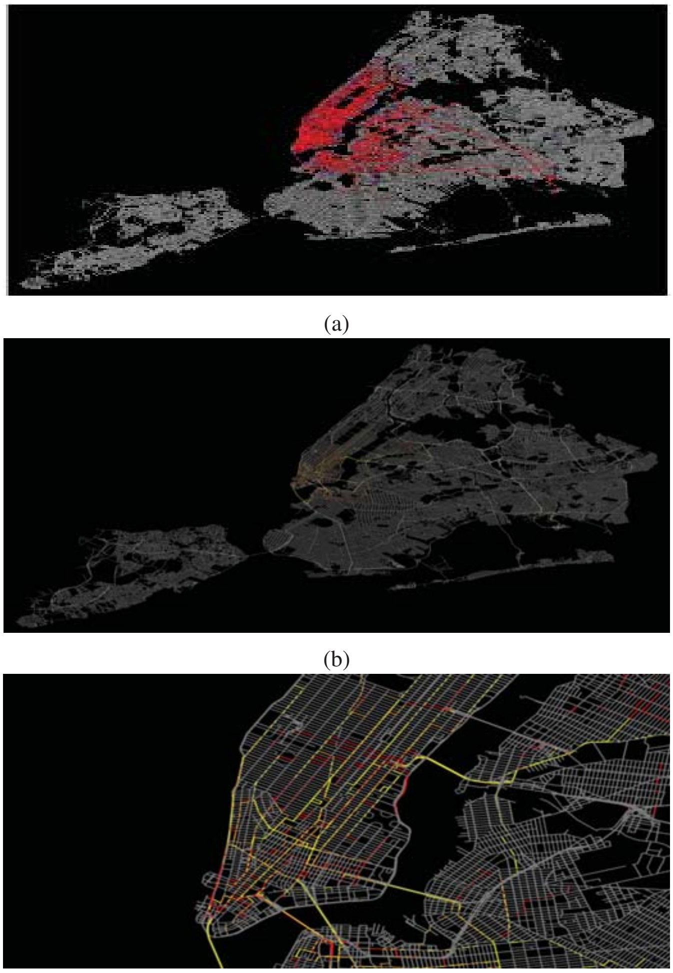
Consider the travel taxi trips from two popular suburbs in NYC city as a start point of the pick-up location from Manhattan to a drop-off location to Queens. Execution of the previous methodology present in this paper for the second case. First was used the trajectory dataset in causality analysis to pinpoint the bottlenecks that are causing the locations pairs of low travel



(a)

(b)

Fig. 5: a) Base map of OSM and b) plot of single root on base map of OSM



(a)

(b)

(c)

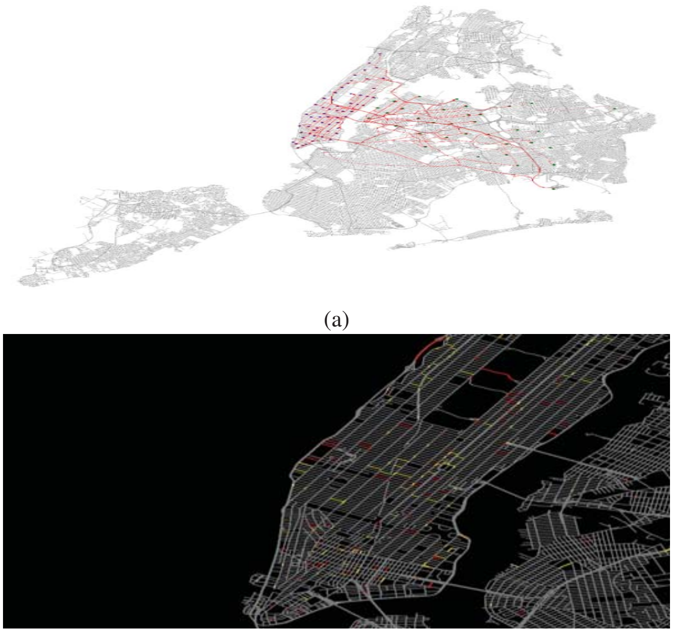
Fig. 6: a) Visualization of all routes of taxi trips on NYC network city. b) Visualization of the all road segment on NYC network map c) The zoom of the part contains the most of red segments

efficiency. Subsequently, display the visualization of ODs on the OSM map (see Fig. 7a).

The matrix A, in this case, contains 799 taxi trips start from Manhattan to Queens with 1206 road segments as columns. Given an OD, considering the probability that a road segment will be used by the driver is calculated by (count number of taxi trip passed this road segment/ total number of taxi trips), and then apply this information in the matrix equation $Ax = l$. As a result, the road segment is involved in the shortest path will be increased. If the increase of the road segment in the shortest path is higher, the probability is lower. The new matrix represents the probability of each road segment with solved the equation $Ax = l$ where x vector indicates the contribution of the road segment to the low travel efficiency. The contribution becomes smaller as x = 0 and greater as $x = 1$ as the visualization of the heatmap of the road segments (see Fig. 7b).

The results in this section aim to show the outperformance of our proposed approach using big trajectory datasets such as NYC taxi trip data. We have compared our proposed STEM approach with the baseline methods (simple Mean and PCA). The simple mean relies on calculating the mean for the traffic volume, the speed, and the detour ratio (i.e., travel distance/geodesic distance) of all the taxis in the city for unique origin-destination pair. However, this method is not accurate because the number of taxis only represents a small fraction of all the vehicles plying in the city. Further, the trajectory data with fine spatiotemporal granularity is very sparse. Moreover, the taxi drivers may also take different paths to increase or decrease the distance trips as a result the trip duration will be changed. Thus, after considering the above factors, using simple means from the trajectory data to represent travel efficiency in road networks is not a very effective method.

Next, we tried to implement the PCA method using the NYC taxi trip dataset. The original features as the speed of taxi trips, trip distance and duration time, etc are converted into Principal Components. Principal Components are the linear combination of all original features. Principal Components are not as readable and interpretable as the original ones. Moreover, the PCA used in previous studies to separate the normal links from the abnormal is often arbitrary, and the results are sensitive to the choice made. Taxis can be viewed as a sample of all vehicles (i.e., a sample population) and the trajectory data from taxis are, in most cases, highly sparse in fine spatiotemporal granularity, even though it can be relatively dense for some periods and regions. So, instead of using a simple mean to measure traffic conditions like all existing studies, the proposed STEM will be applied by considering the number of taxis and the sample variance. As a result, even though it would still be difficult to calculate an exact



(a)

(b)

Fig. 7: a) The paths of taxi trip from Manhattan-Queens. b) The zoom of the road segments heatmap Manhattan-Queens taxi trip

mean of the population, the mean of the population can be reliably compared against a given value. So, our proposed approach STEM is a much more robust and reliable technique to measure the travel efficiency of all origin-destination pairs across an entire city's road network.

The results show that our proposed approach, STEM outperforms both baseline algorithms for high-urban traffic to calculate the minimum speed of population based on the sample of our population. The efficiency of the STEM process to measure numerous origin-destination pairs, verifying the measure used, and identify the minimum speed of population to use as input in the formulate the matrix equation used and solved on the next proposed approach LANCE.

Next, we evaluated our proposed LANCE solution's effectiveness using both real case studies NYC taxi trips and Man-Queens trips as a more specific result for more popular places in New York City. To the best of our knowledge, there are no existing studies relevant to our proposed approach LANCE that plotted each road segment physically on the NYC map and computed the accurate values of the contribution of all road segments that causes low travel efficiency. The LANCE is a form of mathematical regression analysis technique developed in our proposed approach to determine the line of the best fit for our dataset, providing the approximation of the relationship between the matrix of taxi trips crossing road segments with trips speed. There exist many baseline algorithms that solved the matrix equation, such as the Non-Negative lease square (NNLS), Quadric Programming (QP), and Liner programming (linprog) in literature. To the best of our knowledge, no method is working under similar lines as that of ours. So, the proposed LANCE is the novel approach in the area. Further, we compare the two baseline algorithms with our proposed approach, LANCE, and found that LANCE output only 486 road segments with high contribution values towards low travel efficiency. On the other hand, both the baseline techniques (QP) and (linprog) provide 7336 sparse values for each taxi trip, which is a significantly higher value than the value given by the proposed LANCE method. Further, it was observed that the root causes of traffic anomalies in road networks and their propagation and influence used different techniques based on a classical neural network model, which provides other objectives different from our proposed approach, LANCE. To the best of our knowledge, no latest studies based on optimization algorithms detect the road segment with high indicators values of the contribution that cause low travel efficiency. Thus, the proposed LANCE is the comprehensive solution to improve efficiency rather than increasing the complexity of the data models as used in neural network techniques by past researchers.

VI. Conclusion

In this paper, we have proposed a framework for dissecting big trajectory data to analyze a large GPS dataset obtained from over 8 million taxis in NYC city during one month. Our proposed methodology compromised of two proposed methods: STEM and LANCE. The first method, STEM measures the travel efficiency of road networks across an entire city and the second method, LANCE is for network-based causality analysis. After applying the STEM, the resulting data analysis will reveal the road segment causes of low travel efficiency. For measuring the travel efficiency, we have used the statistical solution based on the null hypothesis test as the taxis can be viewed as a sample of all vehicles (i.e. a sample population), and the trajectory data from taxis are, in most cases, highly sparse in fine spatiotemporal granularity, even though they can be relatively dense for some time periods and regions. In network-based causality analysis using LANCE, one finds the contribution of road segments which are the cause the low travel efficiency using the least-squares approximation method, which in turn uses an approximate norm solution of $Ax \approx l$, to solve the problem with high accuracy and reliability in transportation area.

Acknowledgment

The first author would like to acknowledge Taibah University Grants Commission for funding this work via a Ph.D. scholarship.

References

[1] Y. Djenouri, A. Belhadi, J. C.-W. Lin, D. Djenouri, and A. Cano, "A Survey on Urban Traffic Anomalies Detection Algorithms," IEEE Access, vol. 7, pp. 12192–12205, 2019.

[2] J. D. Mazimpaka and S. Timpf, "Trajectory data mining: A review of methods and applications," J. Spat. Inf. Sci., vol. 13, no. 13, pp. 61–99, 2016.

[3] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, Urban Computing with Taxicabs. 2011.

[4] A. Siffer, P.A. Fouque, A. Termier and C. Largouet, "Anomaly Detection in Streams with Extreme Value Theory," 2017.

[5] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On Mining Anomalous Patterns in Road Traffic Streams,"2011.

[6] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On Detection of Emerging Anomalous Traffic Patterns Using GPS Data," 2013.

[7] W. Kuang,S. An and H. Jiang, "Detecting traffic anomalies in urban areas using taxi GPS data." Mathematical Problems in Engineering 2015.

[8] M. Xu, J. Wu, H. Wang, and M. Cao, "Anomaly Detection in Road Networks Using Sliding - Window Tensor Factorization," 2019.

[9] T. V Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model," 2005.

[10] R. Engle, "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics," 2001.

[11] J.-G. Lee, J. Han, and X. Li, "Trajectory Outlier Detection: A Partition-and-Detect Framework," 2008.

[12] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection," 2016.

[13] J. Mao, T. Wang, C. Jin, and A. Zhou, "Feature grouping-based outlier detection upon streaming trajectories," IEEE Trans. Knowl. Data Eng., vol. 29, no. 12, pp. 2696–2709, Dec. 2017.

[14] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," Proc. IEEE Int. Conf. Data Mining, ICDM, pp. 141–150, 2012.

[15] F. Kelly, "The Mathematics of Traffic in Networks," 2006.

[16] V. Lianty, "Finding the Shortest Path of Taxi pick-up location to Customers Using A * pathfinding Algorithm," May, 2014.

[17] J. Dai, B. Yang, C. Guo, and Z. Ding, "Personalized route recommendation using big trajectory data," Proc. Int. Conf. Data Eng., 2015-May, pp. 543—554, 2015.

[18] H. Wang, Y. Li, G. Liu, X. Wen, and X. Qie, "Accurate detection of road network anomaly by understanding crowd's driving strategies from human mobility," ACM Trans. Spat. Algorithms Syst., vol. 5, no. 2, pp. 1–17, Aug. 2019.

[19] S. Boyd and L. Vandenberghe, "Convex Optimization." [Online]. Available: http://www.cambridge.org. [Accessed: 21-Aug-2020].

Rania Alshike received the bachelor's degree in computer science and technology from the Flinders University, Australia. She did her master degree in Information Technology at RMIT University, Australia. She is currently pursuing a PhD degree with RMIT University, Australia. Her current research interests include spatiotemporal data management and analysis.

Dr. Vinita Jindal Dr. Vinita Jindal is an Assistant Professor in the Department of Computer Science, Keshav Mahavidyalaya, University of Delhi since August 2001. She was Head of Department of Computer Science, Keshav Mahavidyalaya, University of Delhi from June 2017 till May 2019. She did her Doctorate in Computer Science from University of Delhi in 2018. She did her M.Phil. in Computer Science from Madurai Kamaraj University in 2007, MCA from IGNOU in 2000 and Bachelor in Mathematics from University of Delhi in 1997. She is mainly working in the area of Artificial Intelligence and Networks. Her areas of interest include Cybersecurity, Intrusion Detection Systems, Dark Web, Deep Learning, Recommender Systems and Vehicular Adhoc Networks to name a few.