ISSN 2563-7568

RESEARCH ARTICLE

Application of Big Data to Determine Traffic Congestion

Uneb Gazder[1*], Rubab Fatima[2], Muhammad Ali Ismail[3], Mir Shabbar Ali[4]

[1]*Department of Civil Engineering, University of Bahrain, Sakhir, 32038, Bahrain*
[2]*Parsons Corporation, Abu Dhabi, UAE*
[3]*Department of Computer & Information Systems Engineering & High-Performance Computing Centre & Principal Investigator at Exascale Open Data Analytics Lab, NED University of Engineering and Technology, Karachi*
[4]*Department of Civil Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan*

Abstract

This study focused on the application of big data analytics to estimate roadway congestion. The implementation of big data allows a convenient approach of data collection without any field or observer biasness. It also relieves the restriction of a limited sample. Highway Capacity Manual 2000 suggested considering travel speed as the most appropriate parameter for describing traffic congestion. Therefore, this research employs a speed performance index with the concept of big data analytics to describe traffic congestion of road. This research estimates speed performance index for different sections of road. The data was gathered for three main roads of Karachi i.e. Rashid Minhas Road, Shahra e Faisal Road and Main Korangi Road. The major achievement of this study project is to propose a novel approach to calculate the speed of a vehicle without field measurements. The data is collected through Smartphone with the aid of an application available online namely My Track. The data that was utilized for this project mainly comprises of GPS Exchange Format (GPX) routes that are converted into Extensible Markup Language (XML) to run the developed script. The script was able to determine the status of congestion on a variety of highways with the help of speed performance index. The same script can be used by administrators and other transportation service providers for estimating the congestion on a real-time basis. The methods and procedures from this study would aid in the transportation planning process, especially for route selection of individuals as well as services.

Key Words: *Big data; Vehicle speed; Road congestion; Travel time; GPS route*

Corresponding Author: *Uneb Gazder, Assistant Professor, Department of Civil Engineering, University of Bahrain Uneb Gazder Sakhir, Bahrain; E-mail: ugazder@uob.edu.bh*

Received Date: *July 10, 2024,* **Accepted Date**: *August 23, 2024,* **Published Date**: *September 18, 2024*

Citation: *Gazder U, Fatima R, Ismail MA, et al. Application of Big Data to Determine Traffic Congestion. Int J Auto AI Mach Learn. 2024;4(2):73-94.*

1. Introduction

With the advent of the latest technologies, the world is growing at a very rapid pace. These technologies produce huge magnitude of data. The data generated and shared by businesses, public administrations numerous industrial and non-profit sectors, and scientific research, has increased immeasurably [1]. These data include textual content to multimedia content (e.g. videos, images, audio) on a variety of platforms.

Report that every day the world produces around 2.5 quintillion bytes of data (i.e. 1 Exabyte equals 1 quintillion bytes, or 1 Exabyte equals 1 billion gigabytes), with 90% of this data generated in the world being unstructured [2]. Assert that by 2020, over 40 Zettabytes (or 40 trillion gigabytes) of data will have been generated, imitated, and consumed. With this overwhelming amount of complex and heterogeneous data pouring from anywhere, any-time, and any-device, there is undeniably an era of Big Data (BD) – a phenomenon also referred to as the Data Deluge [3]. Therefore, data gathering is not a challenge anymore. However, storing, processing and analyzing huge datasets become a time-consuming task. Therefore, a platform is required to effectively use these huge datasets and analyze them to give real-time information for better decision making.

Technologies such as GPS, Google Map etc. are utilized on a daily basis by many commuters. The data generated by these technologies could draw beneficial outcomes if analyzed properly. Moreover, traffic congestion has become a serious problem in many cities, especially in large cities. In order to alleviate traffic congestion and improve the levels of service and efficiency of urban transportation systems, advanced traffic control and management methods have become effective and common approaches. Evaluating traffic congestion levels of road networks is important for traffic management and control, since it could allow the corresponding agencies an accurate and clear grasping of network traffic operation status including the information of location and time for congested roads. Therefore, it is necessary to evaluate traffic congestion situations for urban road traffic networks using applicable evaluation measures.

In spite of their widespread use, the use of big data, made available through navigation technologies, for congestion mitigation and demand management is rather limited. Furthermore, the development of tools and methodologies to achieve such tasks are scarcely found. In light of the above, the main objective of this research is to determine the speed performance index to evaluate traffic congestion by applying the analysis of big data. In general, the major limitation of this study is resources. As, the availability of resources such as software and equipment highly influence the research methodology. Therefore, this research will be done according to availability of required resources. This research work is limited to urban roads only.

The following statements describe the outline of the rest of the paper. Section 2 delivers the literature review of the concept of big data and its application in traffic engineering. It also explains several different methods to evaluate a roadway with the help of different parameters and concludes with a research direction that is developed based on the literature review. Section 3 explains the selected methodology that is applied to conduct the project. It gives a detailed overview of the project methodology and also explains collected data. Section 4 is based on Data Analysis for different roadways to draw useful insight. Section 5 comprises of general discussion on major findings as well as other important aspects of the research.

Section 6 concludes the project. It identifies the major outcomes of the projects and the recommendations for further improvement.

2. Literature Review

The sources of literature review are books, research documents and the information that is available online.

2.1. Big data

Before the advent of big data, real-time traffic management was not possible. As estimation of traffic congestion requires complex calculation of volume-capacity for each segment of the network. The recorded videos can determine the performance of road, but it cannot estimate the real-time traffic condition to perform immediate response. Hence, big data plays a vital role in processing real-time information.

It is essential to understand the term big data. Big data has various definitions, however some of them are quoted in the succeeding paragraphs. According to [4] quoted that big data is still not a clearly defined term and it has been defined differently from technological, industrial, research or academic perspective. In general, it is considered as structured and unstructured datasets with massive data volumes that cannot be easily captured, stored, manipulated, analyzed, managed and presented by traditional hardware, software and database technologies.

2.1.1. Types of big data analytical (BDA) methods

For analyzing big data following analytical methods are used:

Descriptive analytics: Descriptive analytics are the simplest form of BDA method and involves the summarization and description of knowledge patterns using simple statistical methods, such as mean, median, mode, standard deviation, variance, and frequency measurement of specific events in BD streams [5]. Often, large volumes of historical data are used in descriptive analytics to identify patterns and create management reports that are concerned with modelling past behavior [6]. Most of the BDA is commonly descriptive (exploratory) in nature and the use of descriptive statistical methods (data mining tools) allows businesses to discover useful patterns or unidentified correlations that could be used for making business decisions.

Predictive analytics: This analytics is concerned with forecasting and statistical modelling to determine the future possibilities based on supervised, unsupervised, and semi-supervised learning models [5, 7, 8]. Predictive analytics are principally based on statistical methods and seeks to uncover patterns and capture relationships in data.

Prescriptive analytics: This type of analytics is performed to determine the cause-effect relationship among analytic results and business process optimization policies. Thus, for prescriptive analytics, organizations optimize their business process models based on the feedback provided by predictive analytic models [9]. In general, prescriptive solutions assist business analysts in decision making by determining actions and assessing their impact regarding business objectives, requirements, and constraints.

For the purpose of this study, different aforementioned types of analytics and their

combination can be used to achieve the research objectives.

2.1.2. Application of BD in transportation

Moreover, the rapid advancement of communication and detection technologies, low-cost and widespread sensing and a dramatic drop in data storage costs have significantly increased the amount of easily extractable information on transport and mobility. According to International Transport Forum "the volume and speed at which data are generated, processed and stored is unprecedented" [10]. In essence, Big Data is a process of gathering, management and analysis of data to generate knowledge and reveal hidden patterns [11]. The advent of Big Data has triggered disruptive changes in many fields including Intelligent Transport Systems (ITS) with a wide range of applications from smart urban planning to enhanced vehicle safety.

However, methodologies and regulations in many domains of Intelligent Transportation System (ITS) have not kept pace with the production of Big Data. Whereas, Big Data has much potential for improving the planning and management of transport activity by radically increasing the amount or near-real-time availability of mobility-related data [12].

However, there are three fields (namely operations, planning, and safety) where authorities must critically evaluate where and how new, or newly available data and data-related insights, can improve transport policy. Transportation operation services in this domain either focus on decision-making support systems for traffic operations or enhance the Advanced Traveler Information Systems (ATIS). For example, travel time prediction, traffic incident and anomaly detection, anticipatory vehicle routing, dynamic congestion charging, demand responsive parking pricing, and predicting bus bunching in network using smart card data, are among the most popular studies that have been conducted [13].

Real-time traffic information (RTTI) system is designed for the real-time control of traffic flows as well as for strategic traffic management, which essentially intends to collect, process, analyze original GPS from various independent sources, especially from floating car, mobile phone, bus, metro in and around the city, and publish real-time traffic status for public. Also, real-time traffic information can be acquired via the Internet (such as Google Map, Baidu Map), mobile phones and GPS terminals [14].

2.2. Traffic congestion index

Traffic congestion is defined as the situation when traffic is moving at speeds below the designed capacity of a roadway [15]. At present, there is no unified and fixed evaluation measure for evaluating traffic operation conditions. In fact, there are various evaluation measures in different regions. For example, Texas Transportation Institute adopted the Roadway Congestion Index (RCI) in 1994 [16]. Washington State Transportation Department published the congestion report in 2006 in which the congestion evaluation index was defined as the average peak travel time [17]. In 1985, Highway Capacity Manual (HCM) first suggested using level of service as an evaluation index of road performance [18].

In China, Ministry of Public Security chose the average travel speed of a city road as the evaluation indicator to describe congestion conditions of road traffic [19]. A considerable number of studies have explored the urban traffic state in different ways using the single valuation indicator e.g. travel speed and travel time that can be directly obtained through the loop detector, GPS, video, etc. Estimated urban traffic state using vehicle average travel speed, considering resident travel characteristics and road network capacities [20].

Studied the urban road traffic congestion evaluation index system and the motor vehicle velocity distributions using the Gaussian mixture model (GMM), for analyzing congestion characteristics [21]. In 2004, Cesar and van Beukering pointed out the benefit of travel time for measuring transportation network performance, and further discussed the methods of collecting travel time and speed data [22]. Compared travel time and travel distance, discussed the influence of various indicators on congestion quantification, and finally presented a congestion classification method based on travel time from the perspective of travelers [23].

However, considering the complexity and dynamic nature of traffic, it is difficult to comprehensively assess traffic congestion conditions of urban road networks by single evaluation indicator. As a result, several studies began to evaluate the traffic state using multiple indicators. The use of transit vehicle travel time and travel speed data in Portland, Oregon, to assess traffic conditions on arterials and freeways [24]. Mined the transit Automatic Vehicle Location (AVL) data to measure travel time and average speed on freeways and thereby quantify the corresponding traffic conditions.

Integrated the traffic volume and occupancy to form a new value for network-wide traffic states observation and analysis, and then formed a pseudo-color-image vividly representing the macroscopic traffic state [25]. Measured variability in traffic conditions by a variability index, which is computed by measuring the size (spatial volume) of the confidence regions defined by multivariate statistical quality control (MSQC) using large sets of archived traffic data (mean speed, and occupancy) [26]. Presented a comprehensive traffic state estimator derived from traffic flow variables (flows, mean speeds, and densities) [27]. Nevertheless, there are some disadvantages, such as the complex computing, difficulty from collection of data and low practical application, for evaluating network traffic congestion using comprehensive indicators. According to previous studies, both the vehicle speed, as a single evaluation index and combination with other factors to form a comprehensive indicator are important approaches to evaluate the traffic state. In this study, the speed performance index, formed by the average travel speed and the maximum permissible road speed, was selected as the traffic state classification indicator and the traffic state was defined into four categories.

As discussed desired attributes of a congestion index and suggested that a congestion index should (i) be easy to communicate, (ii) measure congestion at a range of analysis level (a route, subarea or entire urban region), (iii) measure congestion in relation to a standard, (iv) provide a continuous range of values, (v) be based on travel time data because travel time based measures can be used for multimodal analysis and for analyses that include different facility types, and (vi) adequately describe various magnitudes of congested traffic conditions [28]. There are several congestion indicators but for urban road following congestion indicators are considered.

2.2.1. Speed performance index

Vehicle speed is an important indicator for measuring the road traffic state. A large amount of vehicle speed data is detected by the loop detector from urban road traffic systems. And based on those data, Beijing Traffic Management Bureau (BTMB) has presented the speed performance index as the evaluation indicator of urban road traffic state. The index value (ranging from 0 to 100) reflects the ratio between vehicle speed and the maximum permissible speed. BTMB chooses the two values (25, 50) as the classification criterion of urban road traffic state [29]. This study uses this speed performance index to measure the road traffic state but adopts three threshold values (25, 50, and 75) as the classification criterion of urban road traffic

state. Based on this evaluation measure, we define the road segment congestion index and the road network congestion index to analyze traffic congestion of urban road networks.

2.2.2. Congestion index

Traditionally, volume-to-capacity (V/C) ratios and level of service (LOS) are implemented by transportation authorities as indicators of congestion intensity [30]. Nevertheless, traffic demand can vary substantially in both temporal and spatial dimensions. Roadway capacity can also be reduced by incidents as discussed above. In such cases, V/C ratios LOS lack the capability to capture the variability of congestion. BD from ITS facilities, on the other hand, provides professionals with much more detailed insights of congestion; they can reflect the overall performance of the whole system, zoom into specific locations or time periods, and observe the changes of congestion intensity in time and space. Above all, they can monitor congestion in real-time. By doing so, quick, precise and effective response is made possible.

Real-time congestion measurement often defines congestion based on travel time or speed. The selection of a specific congestion measure depends on the available ITS detection systems on the managed roadways. Many agencies have introduced the Automatic Vehicle Identification (AVI) systems to keep track of vehicles at different AVI locations and calculate the travel time. Travel Time Index (TTI) is widely used to measure the extra time taken during peak hours compared with that of non-peak hours [31]. Additional time-based congestion measures include hours of congestion, Planning Time Index (PTI) and delay, etc. Other prevalent ITS detection systems such as loop and radar detector systems record the spot speed information. Speed-based congestion measures are developed for these systems.

Washington State DOT (WSDOT) defines congestion based on the ratio between real-time speed detected by loop detectors and the posted speed limit [32]. Proposed congestion index (CI). Compared with choosing a fixed speed as congestion threshold, CI is a more flexible and consistent term to reflect the congestion intensity since posted speed limit can vary across roadway sections [33]. The system of interest in this study deploys numerous MVDS detectors along the expressways to continuously monitor the traffic conditions at their installed locations. Consequently, congestion index is adopted as the congestion measure for the MVDS traffic data.

Road segment congestion index: In order to measure the degree of road segment congestion, this study chooses the average road segment state and the duration of non-congestion state in the observation period to define the road segment congestion index. The non-congestion state includes two traffic states: smooth and very smooth, namely the speed performance index is larger than 50 (km/h). The value of the road segment congestion index Ri is between 0 and 1, and the smaller the value of Ri, the more congestion of road segment [34].

Road network congestion index: The road network is formed by many road segments, so this study gives the road network congestion index based on the road segment congestion index. Similarly,the value of the road network congestion index R is between 0 and 1, and thesmaller the value of R, the more congestion of road network.

2.2.3. Excess delay-ExD

Excess Delay (ExD) was introduced by TfL in the context of the congestion charging system in London [35]. Congestion is defined as the average excess or lost travel time experienced by

vehicle users on a road network. The corresponding indicator is defined as the difference between the Observed Travel Rate (TRobs) and the Reference Travel Rate (TRref).

The Travel Rate is the inverse of the Network Speed and describes the consumption of time per kilometer travelled in the network. The Excess Delay is then the extra consumption per kilometer caused by congestion compared to the reference level.

2.2.4. Relative speed reduction - RSR

Previous studies in Sweden used the Relative Speed Reduction (RSR) as a congestion indicator for link: In empirical studies, the above indicator has shown to have the narrowest confidence interval among the presented indicators [36].

2.2.5. Queue indicator

The bottleneck approach considers, in general, design variables or apriori parameter (i.e. travel demand). Unfortunately, demand cannot be empirically observed when its value is above capacity. In this way, queues become suitable for empirical estimations using this approach. "Time in queue" is the usual indicator when dealing with links or single server. However, in road networks where vehicles circulate at speeds that vary in a continuous range, "Standing-Still-Seconds" (SSS) will be a more adequate indicator. The literature shows some studies aggregating indicators for road network areas using the critical speed for queuing as under 3 km/h with no consideration of the distance to the previous car [37]. This is a sound assumption when considering streets in the city center and it will be used in this study.

2.3. Research direction

The main objective of literature review is to formulate methodology of the research considering objectives and scope of the research. For this study, the software used is R because of its availability and user-friendliness. However, R-studio is used for a better and simplified interface. Moreover, vehicular tracking data is required in the form of GPS Exchange Formate (GPX) for finding the speed of vehicle. Additionally, these routes should be converted into Extensible Markup Language (XML) files so that it can easily be accessed in R. The vehicle track is recorded through My Track app that can be easily retrieved in smartphones. All the recorded tracks should be exported in the form of GPX and converted into XML. Afterwards, the speed performance index is calculated by utilizing aforementioned R packages. Lastly, plots and maps should be exported from R.

3. Methodology

This chapter comprises of the two main sections. Firstly, the process of data collection is described to better understand the analysis. The methodology of the research work is described in Figure 1.

However, in the next section the overall steps of data collection and analysis are discussed in a detailed manner.

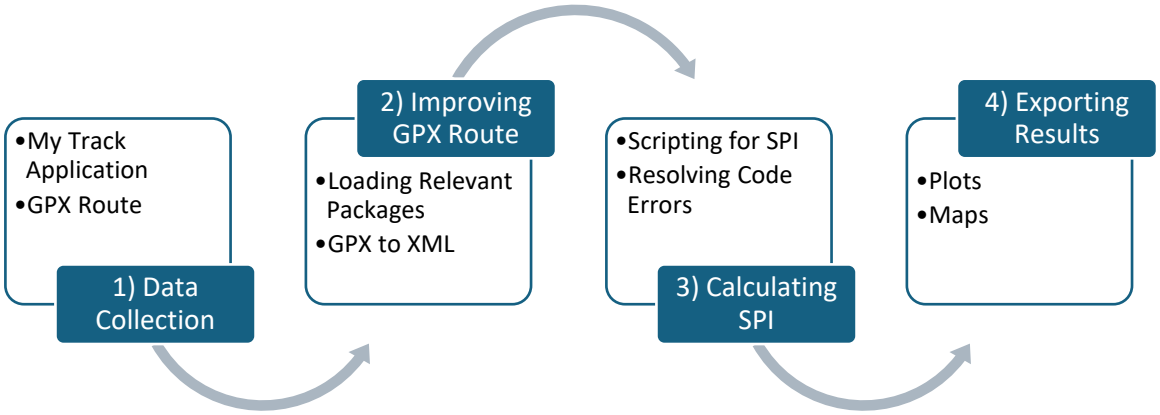

Figure 1: *Overall methodology.*

3.1. Data collection

The main task of this study is to retrieve the speed of a vehicle along a road. For this purpose, several routes are recorded within the same time frame in the form of GPX.The data comprised of 35 routes in total covering three major roads of Karachi namely, Rashid Minhas Road, Shahra e Faisal and Main Korangi Road. The time period covered for this study is morning rush hours usually between 8:30 am to 9:30am. The routes are recorded in December 2019.

The data was recorded through smartphone via online free application. The data mainly comprised of Road Application Programming Interface (API) that saves GPS co-ordinates of a vehicle along with the time and date information.

3.1.1. Road API

The road API enables us to extract information from different applications. An API is a software intermediary that allows two applications to talk to each other. In other words, an API is the messenger that delivers your request to the provider and delivers the response.

The Roads API takes up to 100 GPS points collected along a route, and returns a similar set of data, with the points snapped to the most likely roads the vehicle was traveling along. Optionally, you can request that the points be interpolated, resulting in a path that smoothly follows the geometry of the road.

3.1.2. Snap to road

When navigating via GPS, sometimes the trace or coordinates do not always match up to the road or path travelled as shown in Figure 2.

Figure 2: *Snap to road.*

With the Bing Maps Snap to Road API, use the service to specify GPS points (latitude and longitude coordinates) collected along a route to return a correspondingset of data with the points that snap to the most likely roads and corresponding road name that the vehicle or asset has travelled along. The service supports interpolating the GPS points, resulting in a route that smoothly follows the geometry of the road for map display purposes, which is a valuable tool when tracking assets and for data visualization.

3.1.3. GPX to XML

GPX is an XML schema designed as a common GPS data format for software applications. It can be used to describe waypoints, tracks, and routes. The format is open and can be used without the need to pay license fees. In other words, the GPX format is an XML data format designed for lossless storage and transfer of data for GPS devices. As such, it can contain a vast array of very detailed data (such as GPS signal strength, number of visible satellites, etc.). The format is available publicly from Topographic and is extensible by using a special extension keyword in its XML. It is used by programs that can match GPS data to events (such as geocoding photos) and is the best choice to move data between GPS devices or other than presentation.

Table 1: *Comparison between GPX and XML.*

Description	GPX	XML
Full Form	GPS exchange format	Extensible markup language
Used for	Creating maps	Defining data
Number of points	Up to 100	Unlimited
File type	Graphical	Text-based
Open in	Google earth, ArcMap, etc.	Notepad

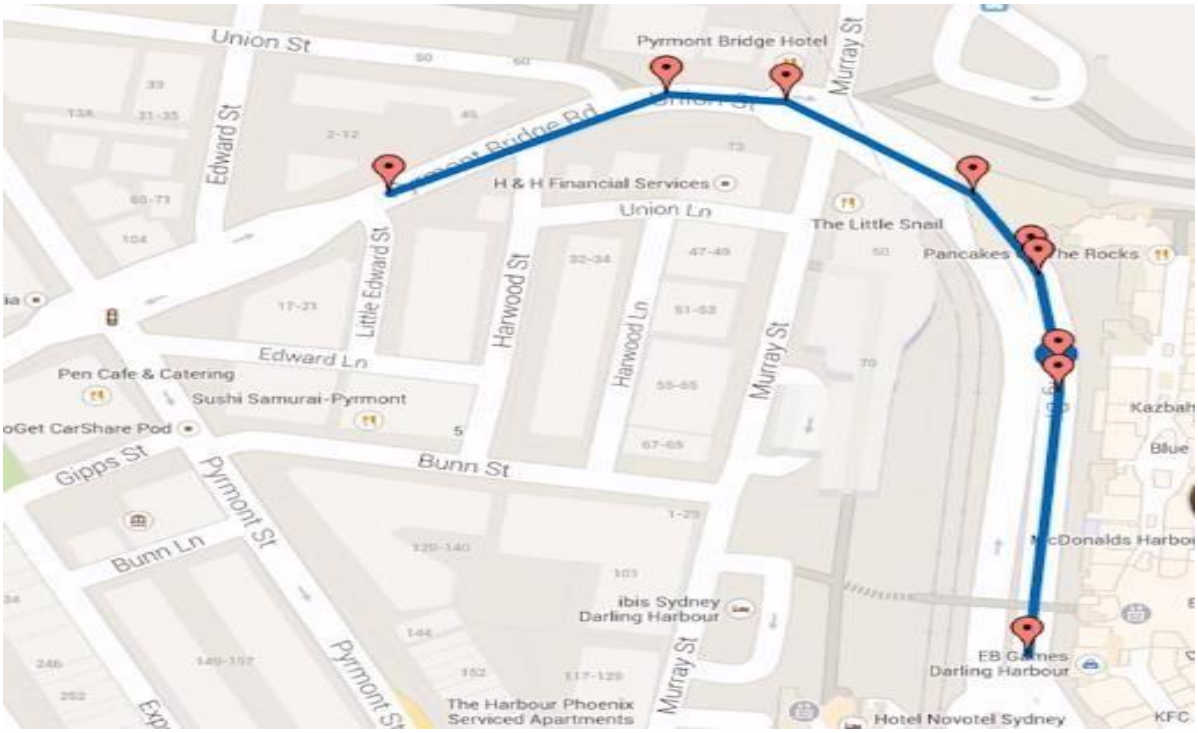

Figure 3: *GPX route before converting into XML format.*

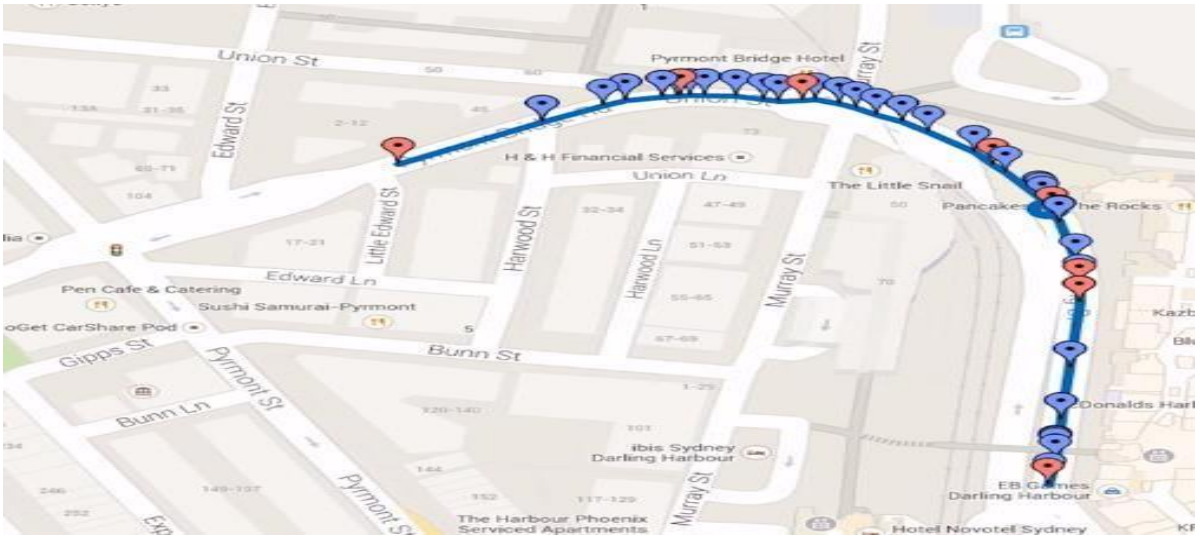

Figure 4: *GPX route after converting into XML format.*

When GPX route is converted into XML format more points are interpolated between existing data points as shown in Figures 3 and 4. Thus, increase accuracy of the result.

3.1.4. My track application

The main constraint of this research was resources; therefore, the use of freely available software is a must to proceed. Therefore, My Track Application is used to record route information in a GPX format. The application allows tracking and recording routes with several other options as shown in Figure 5.

Figure 5: *Recording GPX route through my track application.*

It can also provide speed charts and elevation charts but for paid users only. Consequently, the application is solely used for retrieving GPX routes. However, thespeed, distance travel and other parameters of analysis are required to cover in the coding process.

3.2. Creating geodatabase in R studio

In this section, the major steps regarding the R software are described. Firstly, R should be downloaded and installed for the suitable operating system. However, it would be more convenient to install R-Studio as well. R-Studio runs R tool but with a better user interface. The basic difference between working in R and R-Studio is the user interface. In R, only console window is visible whereas in R-Studio it is organizedfour main windows. The window that is available in the top left corner is to write script and run later. The bottom left window is known as console that runs all the commands of R and give the response of each command. However, the working directory, plots and packages can be managed through the window provided at bottom right side of the R-Studio. Lastly, the history and other functions can be accessed through the top right window.

For this research work, it is important to install and load all relevant packages before executing the commands. Moreover, the GPX route file should be in the working directory. All The packages which are necessary to run the script of the project are freely available on CRAN

(comprehensive R archive network).

3.3. Importing GPX routes in R

The next step is to import GPX routes in R. But the main problem in GPX format is that it only comprises of a hundred points. Therefore, it was converted into XML format. Afterwards, it is important to load all routes in the same working directory otherwise R cannot read information provided in the script.

3.4. Calculation of speed performance index

There are several different methods to evaluate road performance and congestion. However, the most suitable parameter to calculate congestion is speed as shown in Figure 6.

Congestion Measure \ Assessment Criteria		simplicity	magnitude of congestion	city comparison	continuous value	travel time	public transport
Basic measure	Total delay	Y	Y	N	Y	Y	N
	Congested travel	Y	N	N	Y	N	N
	Congested roadway	Y	N	N	Y	N	N
Ratio	Travel rate	Y	N	N	Y	Y	N
	Delay rate	Y	N	N	Y	Y	N
	Relative delay rate	Y	N	N	Y	Y	N
	Delay ratio	Y	N	N	Y	Y	N
LOS	Level of Service	Y	Y	N		N	N
Indices	Congestion index	N	Y	N	Y	Y	N
	Travel rate index	N	Y	Y	Y	Y	N
	Congestion burden index	N	Y	Y	Y	Y	N
	Roadway congestion index	N	Y	Y	Y	Y	N
	Congestion severity index	N	Y	Y	Y	Y	N
	Corridor mobility index	N	Y	N	Y	Y	N
	Lane mile duration index	N	Y	Y	Y	Y	N

Figure 6: *Congestion indices evaluation matrix.*

Therefore, a speed performance index is considered for this research project. However, it is quite troublesome to calculate the exact speed of the vehicle at different locations along the roadway. Consequently, it is suitable to record GPX route that canbe utilized to calculate speed of the vehicle.

After carrying out all the aforementioned step, the R-Studio is able to start analyzing each route. The first step was to make a geodatabase that saves the latitude, longitude, elevation, date and time information of each point. Then, the distance between two GPS points are calculated by using the distance to previous point command. This step gives the distances between all the GPS points. Now, it is essential for the geodatabase to read date and time format. This is done by specifying the date and time format in R. Afterwards, the time interval between two consecutive points was calculated by using time to previous point command.

The next step was to calculate the speed of the vehicle by simply dividing the distance with time. This step will result in the speed of each point along the route in meter per second. Then, the speed is converted from m/sec to km/hr by multiplying it with 3.6.

The next step is to calculate the speed performance index of each point. This is done by dividing vehicular speed by the maximum permissible speed of road (i.e. taken as 60km/hr)

and multiplying it by 100 as provided in equation 1.

$$Rv = {}^{v}/_{vmax} * 100 \quad (1)$$

Where R_v is speed performance index, v is speed of the vehicle and v_{max} is the maximum permissible speed of roadway. For evaluating the road condition, the speed performance index criteria are given in Table 2.

Table 2: *The evaluation criterion of speed performance index.*

Speed performance index	Traffic state level	Description of traffic state
[0, 25]	Heavy Congestion	The average speed is low, road traffic state is poor
[25, 50]	Mild Congestion	The average speed is lower, road traffic state is bit weak
[50, 75]	Smooth	The average speed is higher, road traffic state is better
[75, 100]	Very Smooth	The average speed is high, road traffic state is good

3.5. Plotting and exporting results

After calculating speed performance index, plots are to be made in order to visualize results. This is done by using ggplot2 and map view. The last step is to export and save all results. For this purpose, RIO (R input/output) is used.

4. Data Analysis

4.1. Setting up geodatabase

Before starting the detailed analysis of collected data, it is vital to make a database that stores all the useful information in it. For this purpose, a geodatabase should be made that reads and stores data.

Firstly, a separate folder is made and set as a working directory. This step can also be done with setwd () command by specifying the folder directory. Afterwards, commands are run to generate an empty set of columns that are subsequently used tosave results on each step of data analysis. This step is essential otherwise it would betroublesome to find the result of one step and proceed to the next. For this particular step, the following set of commands are used, as shown in Figure 7.

```
shift.vec <- function (vec, shift) {
  if(length(vec) <= abs(shift)) {
    rep(NA ,length(vec))
  }else{
    if (shift >= 0) {
      c(rep(NA, shift), vec[1:(length(vec)-shift)]) }
    else {
      c(vec[(abs(shift)+1):length(vec)], rep(NA, abs(shift))) } } }
col1 <- seq(0,100,5)
col2 <- seq(200, 100, -5)
my_df <- data.frame(c1= col1, c2= col2)
my_df
my_df$nc1 <- shift.vec(my_df$c1, -1)
my_df$nc2 <- shift.vec(my_df$c2, -1)
my_df
options(digits=10)
```

Figure 7: *Setting up geodatabase.*

After successfully setting up a data frame, the R-Studio is able to proceed further steps of analysis. However, it is vital to utilize already available R packages to simplify the process.

4.2. Loading of relevant packages

There are many packages available to work in R. However, for this study twelve main packages are to be loaded in order to get the required result. These packages have their specific command that aids in the analysis. By loading these packages, a command can be called easily.

However, it is important to note that each package should be installed in R directory before loading them in R-Studio. As discussed earlier, it is important to distinguish the function of each package. Firstly, pacman is utilized for managing all relevant packages, that means to load or unload a specific package wherever necessary. Whereas XML is used to read the GPX route information in R-Studio. However, GGMap, Mapview andOpenStreetMap enable the option to add base map.

Furthermore, Sp and Raster is mainly used to retrieve geographic information of maps. Moreover, dplyr is used for filtering data. On the other hand, lubridate is merely used to work with date and time. Whereas GGPlot2 is used as a powerful tool to plot results. Lastly, RIO is used to import results. After loading all the relevant packages, the further steps of the process can be done with reduced amount of time and complexity.

4.3. Speed performance index

4.3.1. Parsing GPX route

Firstly, all the GPX routes are converted into XML files by simply changing the file extension from .GPX to .XML. Afterwards, these routes are parsed in R-Studio. Then, commands are specified to read and export geographic coordinates from the uploaded file. Moreover, commands are also specified to read the relevant time and date format.

4.3.2. Calculating speed of vehicle

In order to calculate speed, the distance between two consecutive points should be major. This is done by using dist.to.prev command. Similarly, time interval is also calculated by measuring the time interval between two GPS points. These steps will result in measuring distance in

meter and time interval in seconds.

4.3.3. Calculation of speed performance index

The vehicular speed obtained by aforementioned steps can be used to determine the speed performance index along the road. This is done by dividing vehicular speed with maximum permissible speed along the road in km/hr and multiplying the calculated value by 100 as discussed earlier in section 3.

4.4. Data visualization through GGPlot2

After calculating vehicular speed and speed performance index graphs and maps were generated by utilizing ggplot2 package. However, to import result specific package RIO (R Input/output) is used. Lastly, tables and maps are exported through write.csv and ggsave command respectively.

5. Results and Discussion

The speed performance index along the route is obtained as a result of sequential steps. The average speed performance index along the route is found to be 27.81 that represent mild traffic congestion on three major roads of Karachi namely, Rashid Minhas Road, Shahra e Faisal Road and Main Korangi Road during morning peak hour in the month of November and December of 2019.

Firstly, the elevation along the route is plotted having all necessary legends and labels having base map at the background of the plot. The following map shows, in Figure 8, elevation along the roadway.

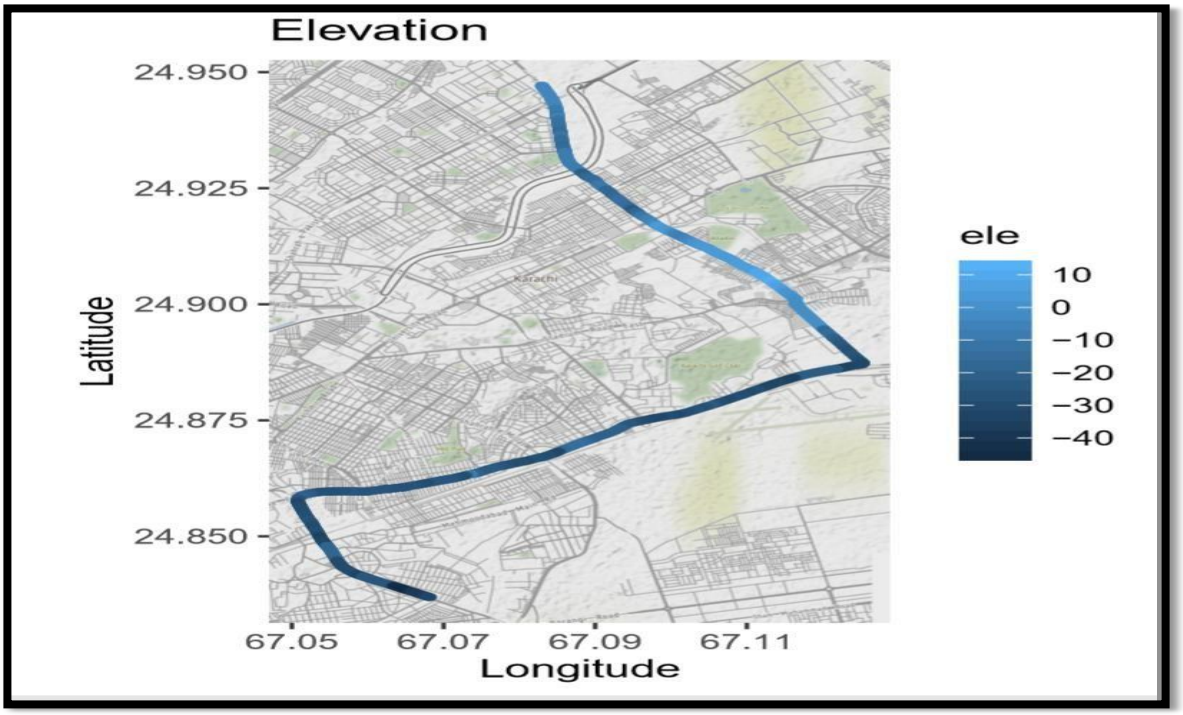

Figure 8: *Elevation along the route.*

Secondly, the calculated speed of vehicle along the road is presented in the map provided below in Figure 9.

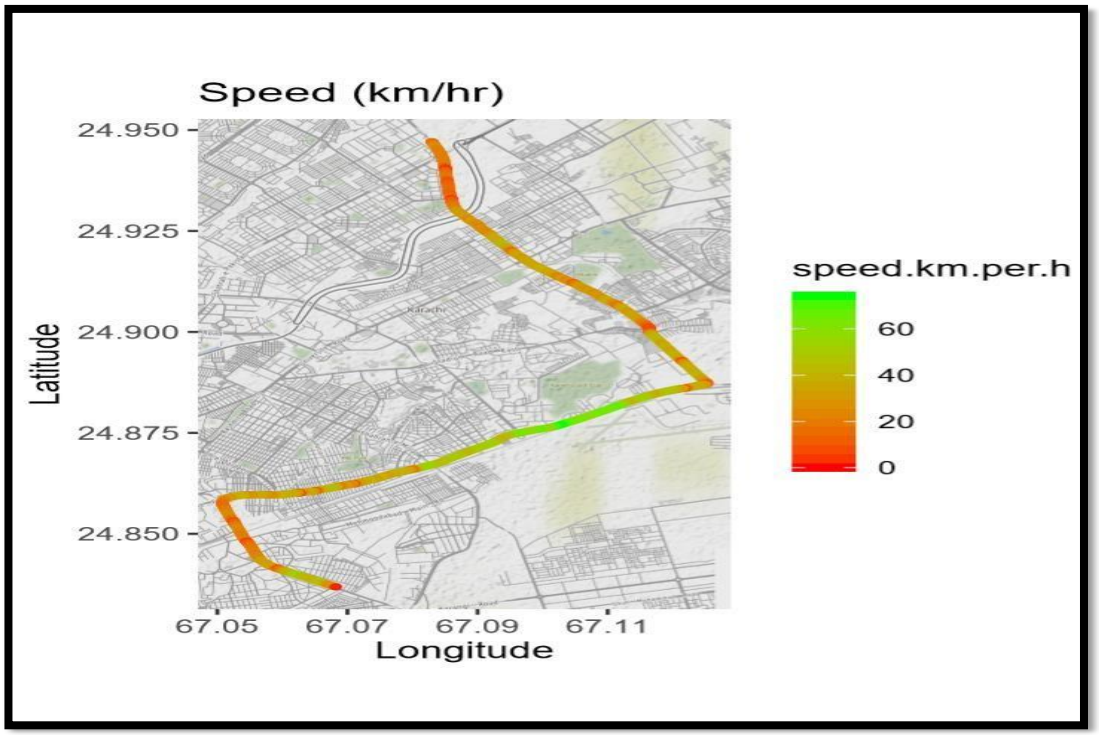

Figure 9: *Speed (km/hr) along the route.*

The speed performance index is also plotted against geographical coordinated along the route. The higher value of speed performance index indicates better roadway condition. Contrary, the lower value of speed performance index indicates poor condition of traffic.

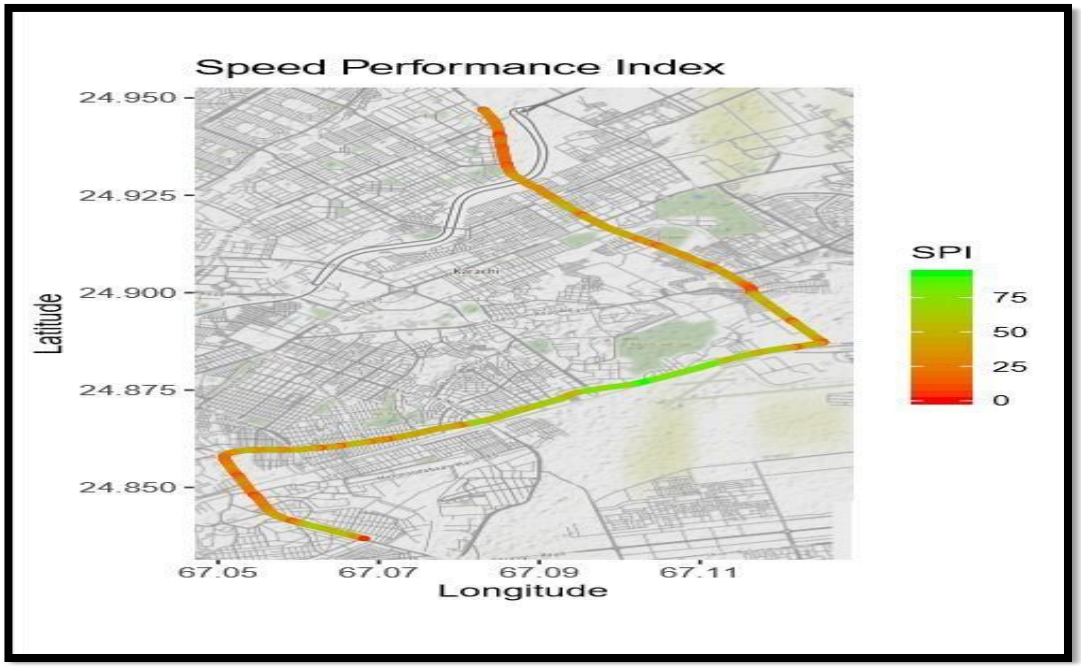

Figure 10: *Speed performance index (SPI).*

All the calculated parameters can be plotted in R-Studio by ggplot2. Therefore, elevation, speed and speed performance index of the road is extracted in the form of graphs.

The following graph, in Figure 10, represents the change in elevation along the route starting from Rashid Minhas Road to Main Korangi Road.

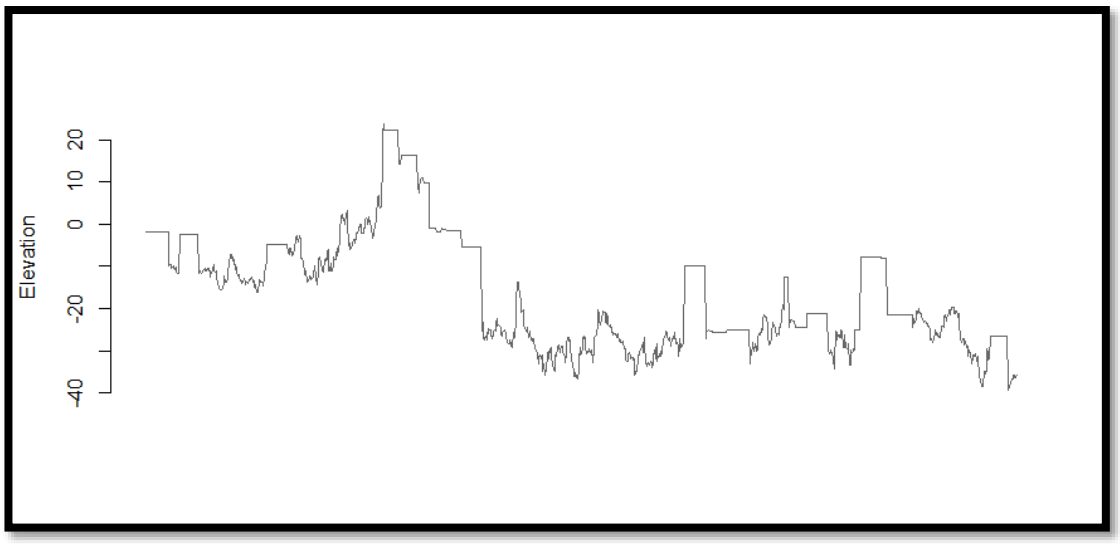

Figure 11: *Variation in elevation.*

The speed variation along the roadway can be seen in the graph provided below in Figure 12.

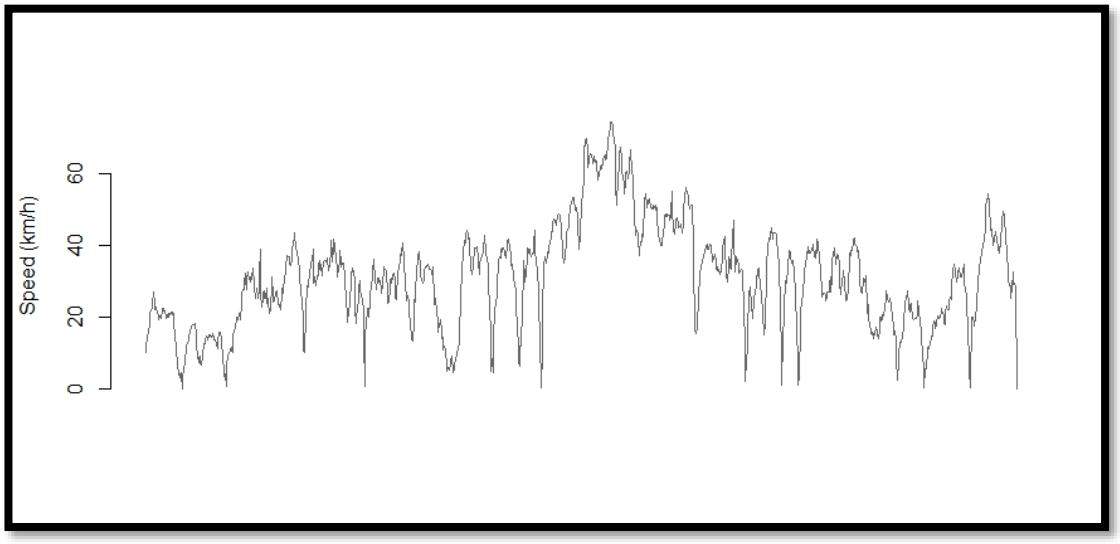

Figure 12: *Variation in speed (km/hrs.)*

The speed performance index is represented in figure 13.

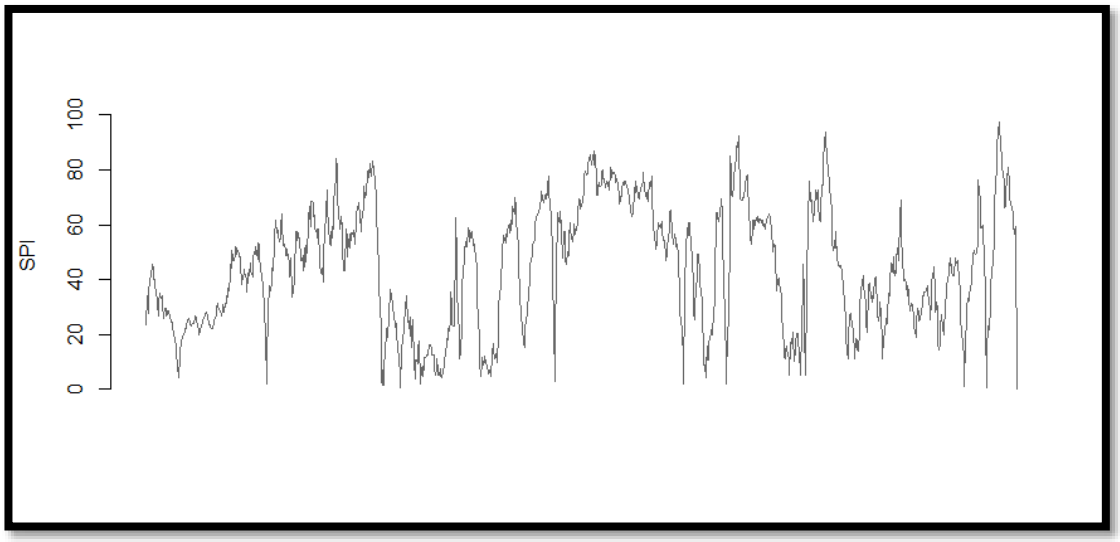

Figure 13: *Speed performance index.*

The calculated values can be saved in the form of a table. These tables can be exported from R in csv file format. These dataset thousands of entries, therefore these files can be opened partially in Excel. The glimpse of exported csv file is provided in Table 3.

Table 3: *Sample CSV file.*

	lat	lon	ele	time	dist.to.prev	time.diff.to.prev	speed.m.per.sec	speed.km.per.h	SPI
1	24.94672	67.08322	-1.7	2/21/2020 3:05	4.510603	1	4.510603	16.23817	27.06362
2	24.94666	67.08326	-1.7	2/21/2020 3:05	3.932841	1	3.932841	14.15823	23.59704
3	24.94661	67.08329	-1.7	2/21/2020 3:05	11.4148	2	5.7074	20.54664	34.2444
4	24.94644	67.08336	-1.7	2/21/2020 3:05	4.633978	1	4.633978	16.68232	27.80387
5	24.94637	67.0834	-1.7	2/21/2020 3:05	11.42733	2	5.713663	20.56919	34.28198
6	24.94621	67.08348	-1.7	2/21/2020 3:05	13.30116	2	6.65058	23.94209	39.90348
7	24.94602	67.08357	-1.7	2/21/2020 3:05	12.85524	2	6.427618	23.13943	38.56571
8	24.94583	67.08366	-1.7	2/21/2020 3:05	14.15471	2	7.077357	25.47848	42.46414
9	24.94562	67.08376	-1.7	2/21/2020 3:05	14.91731	2	7.458656	26.85116	44.75193
10	24.9454	67.08386	-1.7	2/21/2020 3:05	15.17158	2	7.585792	27.30885	45.51475
11	24.94517	67.08396	-1.7	2/21/2020 3:05	14.83667	2	7.418334	26.706	44.51001
12	24.94494	67.08406	-1.7	2/21/2020 3:05	13.38986	2	6.694929	24.10174	40.16957
13	24.94472	67.08414	-1.7	2/21/2020 3:05	6.647441	1	6.647441	23.93079	39.88465
14	24.9446	67.08418	-1.7	2/21/2020 3:05	12.7383	2	6.369152	22.92895	38.21491
15	24.9444	67.08426	-1.7	2/21/2020 3:05	5.837796	1	5.837796	21.01606	35.02677
16	24.94431	67.0843	-1.7	2/21/2020 3:05	5.466889	1	5.466889	19.6808	32.80134
17	24.94421	67.08433	-1.7	2/21/2020 3:05	5.156048	1	5.156048	18.56177	30.93629
18	24.94412	67.08437	-1.7	2/21/2020 3:05	8.912749	2	4.456374	16.04295	26.73825
19	24.94395	67.08441	-1.7	2/21/2020 3:05	10.49115	2	5.245573	18.88406	31.47344
20	24.94376	67.08446	-1.7	2/21/2020 3:05	11.64269	2	5.821346	20.95684	34.92807

21	24.94356	67.08453	-1.7	2/21/2020 3:05	5.670616	1	5.670616	20.41422	34.0237
22	24.94346	67.08457	-1.7	2/21/2020 3:05	5.499078	1	5.499078	19.79668	32.99447
23	24.94336	67.0846	-1.7	2/21/2020 3:05	5.650778	1	5.650778	20.3428	33.90467
24	24.94327	67.08463	-1.7	2/21/2020 3:05	5.199746	1	5.199746	18.71909	31.19848
25	24.94318	67.08467	-1.7	2/21/2020 3:05	8.588609	2	4.294305	15.4595	25.76583

5.1. Discussion and practical implications

There are numerous benefits to finding traffic congestion of the road. The main advantage of determining speed performance index is to evaluate the congestion of road. Speed performance index computes levels of service (LOS) of road. As delay and other congestion indicators do not have a limiting value allotted with it. However, the speed performance index enumerates the performance of road and provides better understanding.

Secondly, this procedure can be also used to determine the speed violation of particular vehicle along the route. Furthermore, roads representing continuous low speed in off peak periods usually have pavement defects such as potholes. Therefore, it provides a baseline to manage road assets in a better manner. Lastly, data obtained from this study can be enhanced and analyzed further to predict traffic situation in the projected year.

5.2. Limitations

There were several limitations for the current research which are briefly mentioned here to set up the future direction of research. The first limitation is the data format which is limited to GPX or XML in this study. However, data could be available in a great many other formats with the variety and advancement in the field of mobile technology.

Another limitation is the sampling of data as many users or organizations may not prefer to share their data for congestion estimation ing volunteers who could drive on the selected highways in the study periods which restricted or other traffic-estimation related tasks. In this study, the data collection was done by us the data collected to a specific time period (December 2019). Moreover, the GPS data collected through the mobile devices would not be as accurate as the data collected through a static device.

The main source of error in R is to not specify the working directory correctly. The whole script cannot be run without specifying a working directory. Secondly, all the packages should be installed before loading them onto the R-Studio. Furthermore, a sound internet connection is obligatory to load base map while plotting the results. Otherwise, R-Studio showed up several lines of errors and would be unable to execute the next command. Lastly. the proposed methodology does not take into account the causes of congestion which could be accidents, special events or road conditions, in addition to the lack of capacity.

6. Conclusion and Recommendation

This study sets out to develop a tool to evaluate congestion on the road using navigation based big data. Speed performance index was used as the measure of congestion. The tool was developed using packages from R studio. Data along five different routes was collected which consisted of the major arterials of Karachi. Using the navigation data, the speed performance index was used to determine congestion on the selected routes. Furthermore, the index was also viewed in conjunction with the road profile.

It is observed that the total distance of this route is 16296.971 meters or 16.296 km. Moreover, the average speed of a vehicle along the route is calculated as 22.25 km/hr. However, the average value of Speed Performance Index is estimated as 27.81 that represents mild congestion. To conclude, mild traffic congestion is observed along these three major roads of Karachi i.e. Rashid Minhas Road, Shahra e Faisal and Main Korangi Road, during morning peak hour.

The script used in this study project can be used to find speed performance index of any route saved in GPX or XML format. However, the results of this research would provide past information about the road performance in terms of congestion. For real-time analysis, the script could be linked to the constant feed of data from the app. In such a case, it could be further extended to be connected with a route management system that helps with real-time navigation of vehicles.

Moreover, similar databases would be available from paratransit facilities in the city such as Uber, Careem, etc. Therefore, it would be easier to analyze the complete road network of Karachi by using the same script i.e. developed in this research.

As directions of future research, it is recommended that the proposed approach be used to collect a more extensive dataset which could help in determining seasonal patterns in traffic. The tool could be further incorporated with a route selection system for facilitating the operations of transportation service providers. Another future direction of research is the implementation of the proposed methodology with the support of video feed to identify the causes of congestion.

References

1. Agarwal R, Dhar V. Big data, data science, and analytics: the opportunity and challenge for IS research. Inf Syst Res. 2014;25:443-8.

2. Dobre C, Xhafa F. Intelligent services for big data science. Future Gener Comput Syst. 2014;37:267-81.

3. Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC Analyze the Future. 2012;pp:1-16.

4. Chen LC, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs. ArXiv Prepr. 2014:1-11.

5. Rehman R, Ghias K, Fatima SS, et al. Students' perception of educational environment at Aga Khan University Medical College, Karachi, Pakistan. Pak J Med Sci. 2016;32:720.

6. Assuncao MD, Calheiros RN, Bianchi S, et al. Big Data computing and clouds: trends and future directions. J Parallel Distrib Comput. 2015;79:3-15.

7. Joseph RC, Johnson NA. Big data and transformational government. It Professional. 2013;15:43-8.

8. Waller MA, Fawcett SE. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J Bus Logist. 2013;34:77-84.

9. Bihani P, Patil ST. A comparative study of data analysis techniques. Int J Emerg Trends & Tech Comp Sci. 2014;3:95-101.

10. Gupta D, Rani R. A study of big data evolution and research challenges. J Inf Sci. 2019;45:322-40.

11. Zgurovsky MZ, Zaychenko YP. Big data: conceptual analysis and applications. Springer, New York, USA. 2020.

12. Lee D, Camacho D, Jung JJ. Smart mobility with big data: approaches, applications, and challenges. Appl Sci. 2023;13:7244.

13. Vujic M, Dedic L, Furjan MT, et al. The benefits of open data in urban traffic networks. In 5th EAI International Conference on Management of Manufacturing Systems. Springer, Cham. 2022:267-82.

14. Ackaah W, Bogenberger K, Bertini RL. Empirical evaluation of real-time traffic information for in-vehicle navigation and the variable speed limit system. J Intell Transp Syst. 2019;23:499-512.

15. Dechenaux E, Mago SD, Razzolini L. Traffic congestion: an experimental study of the Downs-Thomson paradox. Exp Econ. 2014;17:461-87.

16. Zang J, Jiao P, Liu S, et al. Identifying traffic congestion patterns of urban road network based on traffic performance index. Sustainability. 2023;15:948.

17. Ewing R, Tian G, Lyons T. Does compact development increase or reduce traffic congestion? Cities. 2018;72:94-101.

18. Pandey A, Biswas S. Assessment of level of service on urban roads: a revisit to past studies. Adv Transp Stud. 2022;57:49.

19. Pucher J, Peng ZR, Mittal N, et al. Urban transport trends and policies in China and India: impacts of rapid economic growth. Transp Rev. 2007;27:379-410.

20. Jia B, Jiang R, Wu QS. The traffic bottleneck effects caused by the lane closing in the cellular automata model. Int J Mod Phys C. 2003;14:1295-303.

21. Zhu ZJ, Wu QS, Jiang R, et al. Numerical study on traffic flow with single parameter state equation. J Transp Eng. 2002;128:167-72.

22. Cesar HS, van Beukering P. Economic valuation of the coral reefs of Hawai'i. Pac Sci. 2004;58:231-42.

23. Roberts CA, Brown-Esplain J. Congestion mitigation at railroad-highway at-grade crossings. Dept of Transportation, Arizona. 2005.

24. Bertini RL, Tantiyanugulchai S. Transit buses as traffic probes: use of geolocation data for empirical evaluation. Transp Res Rec. 2004;1870:35-45.

25. Duan Y, Lv Y, Liu YL, et al. An efficient realization of deep learning for traffic data imputation. Transp Res Part C Emerg Technol. 2016;72:168-81.

26. Turochy RE, Baker SM, Timm DH. Spatial and temporal variations in axle load spectra and impacts on pavement design. J Transp Eng. 2005;131:802-8.

27. Wang Y, Papageorgiou M, Messmer A. Real-time freeway traffic state estimation based on extended Kalman filter: adaptive capabilities and real data testing. Transp Res Part A Poli Pract. 2008;42:1340-58.

28. Levinson HS, Lomax TJ. Developing a travel time congestion index. Transp Res Rec. 1996;1564:1-10.

29. Odiakose UC, Iyeke SD. Traffic congestion analysis of asaba road using volume to capacity ratio and speed performance index. J Appl Sci Environ. 2024;28:1315-26.

30. Hao N, Feng Y, Zhang K, et al. Evaluation of traffic congestion degree: an integrated approach. Int J Distrib Sens Netw. 2017;13:1-14.

31. Schrank D, Lomax T. The 2007 urban mobility report. Texas Transportation Institute, Texas A&M University System, Texas. 2007.

32. Hammond S. The effect of additional lane length on roundabout delay. Open Access Dissertations. 2014;230.

33. Dias C, Miska M, Kuwahara M, et al. Relationship between congestion and traffic accidents on expressways: an investigation with Bayesian belief networks. 40th annual meeting of infrastructure planning, Mexico. 2009.

34. Tu Y, Lin S, Qiao J, et al. Deep traffic congestion prediction model based on road segment grouping. Appl Intell. 2021;51:8519-41.

35. Santos G, Fraser G. Road pricing: lessons from London. Econ Policy. 2006;21:264-310.

36. Moran C, Koutsopoulos H. Congestion indicators from the users' perspective: alternative formulations with stochastic reference level. 12th World Conference on Transport Research, Portugal. 2010.

37. Amini B, Shahi J, Ardekani SA. An observational study of the network-level traffic variables. Transp Res Part A Poli Pract. 1998;32:271-8.