BUS ARRIVAL TIME PREDICTION USING MACHINE LEARNING

Submitted by:

ANKIT TALWAR

Supervisor:

Dr AHMED ZAHRAN

Second Reader:

PROF. KEN BROWN

MSc in Data Science and Analytics

Computer Science Department
University College, Cork

September 1, 2019

Abstract

Traffic congestion and transportation-related environmental problems are identified as severe problems all over the world. Given the negative impacts on individuals and the economic, environmental and societal costs, major capital expenditures and numerous efforts have been put in place for tackling traffic problems. Nowadays, commuters in most cities are heavily reliant on private cars. This situation motivates finding cost-effective and less polluting alternatives to efficiently serve urban mobility. The provision of faster, accurate bus arrival information would make public transportation (PT) more convenient. (European Commission 2013; Wardman 2014).

This paper focuses on developing a prediction scheme of bus travel time using the Artificial Neural Network (ANN) method. The objective is to apply machine learning techniques by identifying the travel time data suited as inputs and use them for the development of the bus travel time prediction model. However, in order to gain the maximum benefit from a neural network, there should be enough data or observations. (Zhang a, Patuwo and Hu, 1998). The model uses scheduled time table data of buses from Transport for Ireland (TFI) and obtained GPS data to predict time. The GPS data applied for developing the proposed model was collected by the REST interface to retrieve information on real-time bus information, operated by a transit agency in Ireland.

MAPE% and SMAPE error values of 2.03% and 0.0202 respectively were obtained from the study indicating that the ANN model can be used to implement an advanced and intelligent transport system. In conclusion, it is shown that bus travel time information can be reasonably and accurately provided using travel time on preceding bus stops.

Declaration

I thus proclaim that this master's thesis entitled "**Bus Arrival Time Prediction Using Machine Learning**" was completed by me for the degree of **MSc. Data Science and Analytics** under the direction and supervision of **Dr. Ahmed Zahran, University College Cork, Ireland**.

The interpretations put forth are based on my reading and understanding of the project and they are not published anywhere in the form of books, monographs or articles. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature.

Name: Ankit Talwar

Student no: 118220956

Signature:

Acknowledgement

First and foremost, I would like to extend my gratitude to my advisor **Dr. Ahmed Zahran** for his invaluable suggestions and comments in the enrichment of this thesis. The door to his office was always open whenever I had a question about my research or writing. He consistently steered me in the right direction whenever he thought I needed it. I am most grateful for his continual patience, benevolence, constructive advice, genuine interest, and unreserved guidance and direction.

My sincere gratitude also goes to those who provided me with encouragement and support throughout the development and completion of this project.

In a similar manner, I owe a great debt of respect to all my teaching staff of management who played capacitating role and librarians, and other respected members of the University who helped me grasp the necessary knowledge in my MS program study that played a key role during my work in this thesis.

Finally, I want to express my deepest appreciation to **Dr. Eric Wolsztynski,** for his encouragement and structured course on Statistical Analytics during my MS program study that enabled me to pursue this thesis with enthusiasm.

Contents

List of Figures

List of Tables

List of Acronyms

ANN	Artificial Neural Network
GTFS	General Transit Feed Specification
GPS	Global Positioning System
CSV	Comma-Separated Values
PT	Public Transport
SQL	Sequential Query Language
API	Application Program Interface
DB	Database
MAPE	Mean Absolute Percentage Error
SMAPE	Symmetric Mean Absolute Percentage Error

Chapter 1
Introduction

1.1 Background of the Study

Traffic congestion is increasing with acceleration and continually posing threat to the quality of life of people all over the world over the past decades. It increases travel time, reduces air quality and cause environmental problems, and decreases mobility of daily commuters. In order to alleviate congestion problem, different techniques have been suggested during the past years, including demand-side (congestion pricing, traffic management, etc.) and supply-side (constructing more roads, adding lanes, etc.) or their integration. (Fan and Gurmu, 2015). However, uncertainties in arrival time are quite common in public transit due to dynamic traffic conditions, particularly for highly varying heterogeneous traffic condition, contributed by various factors such as lack of lane discipline, fluctuating travel demand, incidents, signal timing and delay, bus stop dwell times, seasonal and cyclic variations.

One problem with public transport buses are that they scarcely stick to any predefined schedule while bus commuters hardly have any real-time information of the likely arrival time of the bus they are expecting. Given the lack of this crucial real-time information, commuters may end up taking private vehicles to reach their respective destinations. This may lead to reduction in usage share of public transport and increase in composition of private vehicles contributing towards the raise in congestion and other related negative impacts. Hence, prediction of travel/arrival time and informing the same to passengers is inevitable to make public transport more attractive, efficient, and competitive especially in urban areas. Such real-time information can also be used to assist commuters in making better trip-related decisions ahead of the journey, which significantly reduces anxiety levels while waiting for a bus. (Watkins et.al., 2011; Cats and Loutus, 2015).

1.2 Research Goals

1.2.1 Research Question

The research question addressed in this thesis is “To what extent can the bus travel time be reliable when historic global position system data is evaluated with the help of machine learning algorithm”.

1.2.2 Research Tasks

a) Develop a predictive statistical model for predicting bus travel time at any bus stop using the machine learning algorithm with GPS information provided. Interpret your results in technical and non-technical terms.
b) Identify different data features that can be computed from the information provided.
c) Analyse the performance of the predictive model using graphs and tables.
d) Draw a conclusion based on the research.

1.2.3 Research Objectives and Scope Overview

In this thesis, the objective is to apply machine learning techniques to develop a bus travel time prediction model. The model is developed to give real-time bus arrival information to the passenger and transit agencies for applying proactive strategies. Recent studies on bus travel time predictions reveal that the ANN model outperformed in terms of accuracy and robustness (Johar et.al., 2015). In this thesis, a dynamic model based on Artificial Neural Network (ANN) has been developed to predict bus arrival time using the Global Positioning System (GPS). The model uses scheduled time table data of buses and obtained GPS data to predict time. The GPS data applied for developing the proposed model was collected by the REST interface to retrieve information on real-time bus information, operated by a transit agency in Ireland.

For the development of the ANN model, certain features were explicitly computed from obtained GPS data to identify patterns to use in the development of a model. For extracting bus service stop positions and whole bus route trajectory from our GPS data, algorithms are presented and implemented. The developed ANN model will provide the travel time prediction between the bus's current position and the downstream bus stop along a route until the bus has reached its final stop. The model was trained, validated and tested using historic GPS data. For evaluation, mean absolute percentage error (MAPE) and Symmetric mean absolute percentage error (SMAPE) were estimated and as a result it provides a better understanding for the transit network which will lead to benefits for bus passengers and a system that can be used further used for traffic management purposes.

1.3 Limitation of the study

The generality of this thesis has been limited by different challenges. The attempts made to forecast bus travel time was only with the derived data features from the historically collected GPS datasets. Another challenge faced was the lack of well-structured and standard data, for instance, the majority of GPS generated data features values were similar. GPS data had to be filtered to eradicate redundant values to achieve accuracy in the results. Despite these challenges, I have tried to critically analyse the available data to answer the questions raised in this thesis.

1.4 Thesis Structure

The rest of the thesis is organised as follows:

- Chapter two provides an overview of existing research in similar areas and introduces the important background of datasets used in this project.
- Chapter three describes the data exploration where data is pre-processed and some technical concepts for this thesis are also presented.
- Chapter four describes the design and methodology of the research which includes machine learning algorithm implemented, the proposed development of the model and its working, algorithms presented for computing some data features and evaluation of prediction methods.
- Chapter five describes some arguments which were made during the research work of this thesis.
- Chapter six summarises the whole research, draws the conclusions and
- Chapter seven discusses future work and enhancements.

Chapter 2

Background

This chapter aims to help readers to better understand this project. Firstly, related works in the same area are reviewed and then some important prior knowledge on datasets and some important technical concepts are introduced.

2.1 Literature Review

A number of studies have been initiated in the past to address the bus arrival time prediction problem. Major techniques brought in to practice for predicting arrival time were historical based models, regression models, Kalman filter- based models, and Machine learning models.

Historic-based models were used to obtain the current and future travel time from observed historical bus travel time data of previous journeys with the assumption of stable traffic congestion. The algorithms of the historical average models were simple and required relatively small computation time. However, the performance of the models was weak and would be successful under the conditions of stable traffic congestion. The regression models required a linear mathematical function to explain a dependent variable with a set of independent variables. (Patnaik et al., 2004). Unlike the historic-based models, these are able to work satisfactorily even under unstable traffic conditions. Regression models have been used by many authors in bus travel time prediction. (Shalaby and Farhan, 2003). For example, Patnaik et al. (2004) developed a set of multiple linear regression models to estimate bus arrival times using distance, number of stops, dwell times, boarding and alighting passengers and weather descriptors as independent variables.

Kalman filtering models could be used to predict the future state of the dependent variable. They have elegant mathematical representations which can adapt to traffic changes with their time-dependent parameters (Chein et.al., 2002). These models have also been used by many authors in bus travel time prediction. (Vanajakshi et. al., 2009; Chien et.al, 2002; Shalaby and Farhan, 2003). The basic function of Kalman filtering model is to provide estimates of the current state of the model from previous time steps. They can also serve as the basis for predicting future values or improving estimates of variables at earlier times because of their capacity to filter noise. (Kalman, 1960).

A recent study focuses on the use of ANN and model-based approaches for the bus arrival prediction. (Kumar et.al., 2015). ANN models are able to deal with complex and noise data and are suitable to find nonlinear relationships between dependent variable and independent variables. (Hagan et.al. 1996). They can be used for prediction purpose, without explicitly specifying the traffic processes. (Fan and Gurmu, 2015). ANNs have recently gained popularity in predicting bus arrival time because of their ability to solve complex non-linear relationships as have been seen in many research efforts. (Jeong and Rilett, 2004; Vanajakshi, et.al., 2009; Chien et.al., 2002). ANN models have been developed by different researchers in predicting bus travel time so far used explanatory variables such as flow, speed, weather, distance etc. as inputs. (Fan and Gurmu, 2015).

With the innovation and implementation of diverse modern sensing technologies that generate large amount of data, data-driven techniques are getting more popularity. For example, these days, many urban public transportation systems deploy Automated Vehicle Location (AVL) systems like Global Positioning System (GPS) to monitor the position of buses in real time, which can provide a constantly growing database of location and timing details. (Gentili and Mirchandani, 2018).

Various studies have been reported on prediction of travel times and the methods used can be broadly classified into traffic flow-theory based and data-driven methods. Data driven approaches use larger databases to develop statistical/empirical relations to predict the future travel time without really representing the physical behaviour of the modelled system. (Zhang et.al., 2017).

Machine learning techniques, such as Artificial Neural Network (ANN) and Support Vector Machine (SVM), are some of the most commonly reported prediction techniques for travel time prediction because of their ability to solve complex relationships. (Chien et.al, 2002; Wu et.al., 2003). As a prominent approach for solving complex problems, ANNs have been recently gaining popularity in transportation. (Chang and Su, 1995; Smith and Demetsky, 1995; Wei and Wu, 1997). ANNs, motivated by emulating the intelligent data processing ability of human brains, are constructed with multiple layers of processing units, named artificial neurons. (Gurmu, et.al., 2014). Artificial neural networks (ANN) perform well in nonlinear relationships establishment. (Jeong and Rilett, 2004). An ANN model based integrated framework, which predicts the average and variance of bus travel times according to both demand and capacity factors (Mazloumi et.al., 2011). The most important demand factor, traffic flow is considered as the main input of model, while capacity factors, weather condition and schedule, are employed as auxiliary inputs. Chien et al. (2002) developed an ANN model to predict dynamic bus arrival time. They used simulated data from CORSIM including volume and passenger demand. Jeong and Rilett (2004) evaluated the performance of historical data-based model, regression model and ANN model for bus arrival time prediction and reported the ANN model performing better than the other two models.

2.2 Prior Knowledge on Datasets

2.2.1 Datasets and their sources

The arrival times of buses at stops, which can be seen at displays at bus stops, on the Bus Eireann website, and in the Realtime Ireland app is provided by real-time predictions by Bus Eireann. The Bus Eireann website makes use of a server-side real-time information REST application program interface API. Two datasets are used to predict bus arrival time prediction, the first dataset is a collection of GPS points for buses in Cork, Ireland. The second dataset from TFI is the scheduled time table data for Bus Eireann buses.

2.2.1.1 GPS Dataset

The three datasets built after accessing bus stops details from the Bus Eireann REST API, polled over 6 weeks are shown in Table 1.

Sr no.	Name	Features	Dimension
1	Bus routes	id, name, direction, number, category	258 x 5
2	Bus stops	id, name, number, latitude, longitude	5,444 x 5
3	Bus historic position	route_id, direction, vehicle_id, last_modified, trip_id, congestion_level, accuracy_level, status, is_accessible, latitude, longitude, bearing, pattern_id, has_bike_rack, category, poll_time	16 x 36,083,298

Table 1: Description of GPS datasets and their features, created by accessing REST API.

Records from the above three datasets shown for example in table 2, 3, 4 respectively.

a) Bus routes data:

Id	Name	Direction	Number	Category
7338652709907595264	208	2	208	5
7338652709907596288	215A	2	215A	5

Table 2: Example records from GPS bus routes dataset.

b) Bus stops data:

Id	Name	Number	Latitude	Longitude
7338653551721440256	Cork (Bus Station - Parn ell Place)	255021	51.89939	8.46621
7338653551721415680	Cork Airport	230061	51.84898	-8.48907

Table 3: Example records from GPS bus stops dataset.

c) Bus historic position data:

Features	Example 1	Example 2	Example 3
route_id	7338652709907595264	7338652709907595264	7338652709907595264
Direction	1	1	1
vehicle_id	7338674957838189568	7338674957838189568	7338674957838188544
last_modified	2019-03-05 06:47:51.735	2019-03-05 08:03:48.416	2019-03-05 07:54:41.204
trip_id	7338656568300882944	7338656568300882944	7338656568300887040
congestion_level	1	1	1
accuracy_level	3	3	3
Status	5	5	5
is_accessible	f	F	f
Latitude	51.87086	51.908088888888905	51.870855
Longitude	-8.54354	-8.419426111111111	-8.543541111111111
Bearing	258	164	135
pattern_id	7338650210240648192	7338650210240648192	7338650210240650240
has_bike_rack	f	F	f
Category	5	5	5
poll_time	2019-03-05 07:20:14.26448	2019-03-05 08:19:16.534591	2019-03-05 09:26:19.14732

Table 4: Example records from bus historic GPS position dataset.

2.2.1.2 Bus Eireann Time Table Schedule Dataset

Recently updated Bus Eireann data is freely downloaded in General Transit Feed Specification (GTFS) format from Transport for Ireland (TFI) official website. In this data, Bus Eireann transportation schedules and associated geographic information is useful in this thesis.

Sr no.	Name	Description	Features	Dimension
1	Agency	Transit agencies with service represented in this dataset.	agency_id, agency_name, agency_url, agency_timezone, agency_lang	1 x 6
2	Stops	Stops where vehicles pick up or drop off riders. It also defines stations and station entrances.	stop_id, stop_name, stop_lat, stop_lon	5,198 x 4
3	Routes	Transit routes. A route is a group of trips that are displayed to riders as a single service.	route_id, agency_id, route_short_name, route_long_name, route_type	309 x 5
4	Trips	Trips for each route. A trip is a sequence of two or more stops that occur during a specific period.	route_id, service_id, trip_id, shape_id, trip_headsign, direction_id	14,591 x 6
5	Stop times	Times that a vehicle arrives at and departs from stops for each trip.	trip_id, arrival_time, departure_time, stop_id, stop_sequence, stop_headsign, pickup_type, drop_off_type, shape_dist_traveled	355,747 x 9

Table 5: Description of GTFS datasets and their features, obtained from TFI website.

Below are the examples of TFI data listed in Table 5.

a) Agency:

Id	Agency_ name	Agency_url	Agency_timezone	Agency_lang
1	Bus Éireann	http://www.transportforireland.ie	Europe/Dublin	EN

Table 6: Transit agency using GTFS service.

b) Stops:

Stop_id	Stop_name	Stop_lat	Stop_lon
8370B2441301	UCC Park	51.8768338	-8.4846504
8370B2441502	CUH (Bishopstown Rd)	51.8819388	-8.5103715
8370B244161	Washington St (Costigans)	51.8973928	-8.4800911

Table 7: Example records from TFI bus stops dataset.

c) Routes:

Route_id	Agency_id	Route_short_name	Route_long_name	Route_type
10-60-e16-1	1	208		3
10-302-e16-1	1	220		3
10-54-e16-1	1	205		3

Table 8: Example records from TFI bus routes dataset.

d) Trips:

Route_id	Service_id	Trip_id	Shape_id	Trip_headsign	Direction_id
10-60-e16-1	y100m+G	11092.y100m.10-60-e16-1.2.O	10-60-e16-1.2.O	Ashmount (Turning Circle) - Curraheen Village	0
10-60-e16-1	y100m+G	11680.y100m.10-60-e16-1.22.I	10-60-e16-1.22.I	Curraheen Village - Cork City Hall	1
10-68-e16-1	y100m+G	8896.y100m.10-68-e16-1.1.O	10-68-e16-1.1.O	Patrick Street - CUH A and E	0

Table 9: Example records from TFI bus trips dataset.

e) Stop_times:

Trip_id	Arrival_time	Departure_time	Stop_id	Stop_sequence	Stop_head_sign	Pickup_type	Drop_off_type	Shape_dist_traveled
12081.y1010.10-10-e16-1.3.O	09:00	09:00	8310B5620001	1		0	0	0
12081.y1010.10-10-e16-1.3.O	09:02	09:02	8310B1378101	2		0	0	924.2335

Table 10: Example records from TFI bus stop times dataset.

2.3 Preliminary Data Analysis of Collected GPS Dataset

For the purpose of collecting data, under Transport for Ireland (TFI), Bus Eireann buses in the city of Cork, Ireland, were used. The obtained GPS data includes the trip id of the bus on a route, time stamp, and latitude & longitude of the location at which the entry was made. The frequency of GPS data obtained through API was at a frequency of every 20 seconds for 41 days in the months of February and till 21 March 2019, which has information about all bus trips over different routes and different directions either upstream or downstream. A total of 36,083,298 rows GPS data records were collected, which was found to have data for 39 unique route buses (31 bus routes data that operates in city Cork, Ireland and other 8 that has transited from side cities of Cork). The processed GPS data of all bus trips are in a Comma-Separated Values (.CSV) file and with the help of Sequential Query Language (SQL), this data is then loaded in a database table. Table 11. shows the extracted route number of 30 different buses whose GPS data points were recorded.

Route Number	Start - End Points
202	Hollyhill (Apple) - Mahon Point Omniplex
215	Cloghroe (Coolflugh Terminus) - Bros Del Rd (Opp Blackpool Shop Ctr)
248	Glenville - Cork Bus Station
243	Charleville - Cork Bus Station
233	Ballingeary (Eastbound) - Cork Bus Station
236	Castletownbere - Cork Bus Station
40	Rosslare Harbour - Tralee Bus Station
51	Galway Bus Station - Cork Bus Station
216	CUH (Bishopstown Rd) - Mount Oval (Monswood Est)
223	Haulbowline - CIT Campus
241	Trabolgan - Cork Institute of Technology
245	Outside Rail Station - Cork Institute of Technology
260	Ardmore (Opp Post Office) - Cork Institute of Technology
237	Goleen (Post Office) - Skibbereen
226	Town Car Park - Cork Institute of Technology
235	Rylane - Cork Bus Station
239	Butlerstown - Cork Bus Station
220	Fort Camden - Grange Road Terminus
201	CUH (Bishopstown Rd) - Boherboy Rd (Opp Scoil Mhuire Banrion)
206	South Mall (Danske Bank) - Grange (Dunvale)
208	Curraheen Village - Cork City Hall
214	CUH A and E - Opp Market Tavern

219	Cork Institute of Technology - Mahon Point Rd (V.H.I Clinic)
221	Knockraha (The Old Schoolhouse) - Cork Bus Station
240	Ballycotton - Cork Bus Station
261	Ballinacurra - Cork Bus Station
205	CIT Campus - Opp Market Tavern
207	Glenheights Park Terminus - Donnybrook (Scairt Cross Terminus)
213	Black Ash Park - St. Patrick Street (O2 Store)
209	Lotamore Drive - St. Patrick Street (O2 Store)

Table 11: Bus operating on different routes.

In this thesis, TFI Dataset is considered as baseline data. From the TFI dataset, details like scheduled arrival & departure time of any bus, fixed latitude & longitude positions of bus stops are extracted which in turn helps to compute different data features other than GPS data which will be discussed further. Both datasets will be used for bus travel time prediction in this statistical model.

2.4 Limitations with Collected GPS Dataset

GPS equipped bus server does not record data all the time. It is normal for a GPS server to record some errors or sometimes, it can also stop sending information, which becomes one problem. Secondly, the GPS dataset does not record any information about different factors like the number of passengers alighting or boarding at a stop, weather & traffic conditions on the route, speed of the bus at the recorded data point which majorly influences bus arrival time prediction at a bus stop. Thirdly, there is no information about the bus depots positions in the dataset, to locate the bus depot position in the data, the algorithm is presented in this thesis and discussed further. GPS data sometimes records latitude, longitude value as 0 or 180 degrees, such data points are ignored and not used in the model.

2.5 Bus Route for Testing Purpose

The testbed chosen for this paperwork is a Bus Eireann service- bus route, 208, which connects the Curraheen Village's bus depot in the western edge of city to the Ashmount's (Turning circle) bus depot in an eastern suburb of Cork city in Ireland, considering downstream direction which has a route length of 11.8 km with an average journey time of 52 minutes with 37 stops in between source and destination point (Stop0 to Stop36). The route includes varying volume, road geometric conditions. Figure 1, illustrates the selected study of bus route 208, snapshot from google maps.

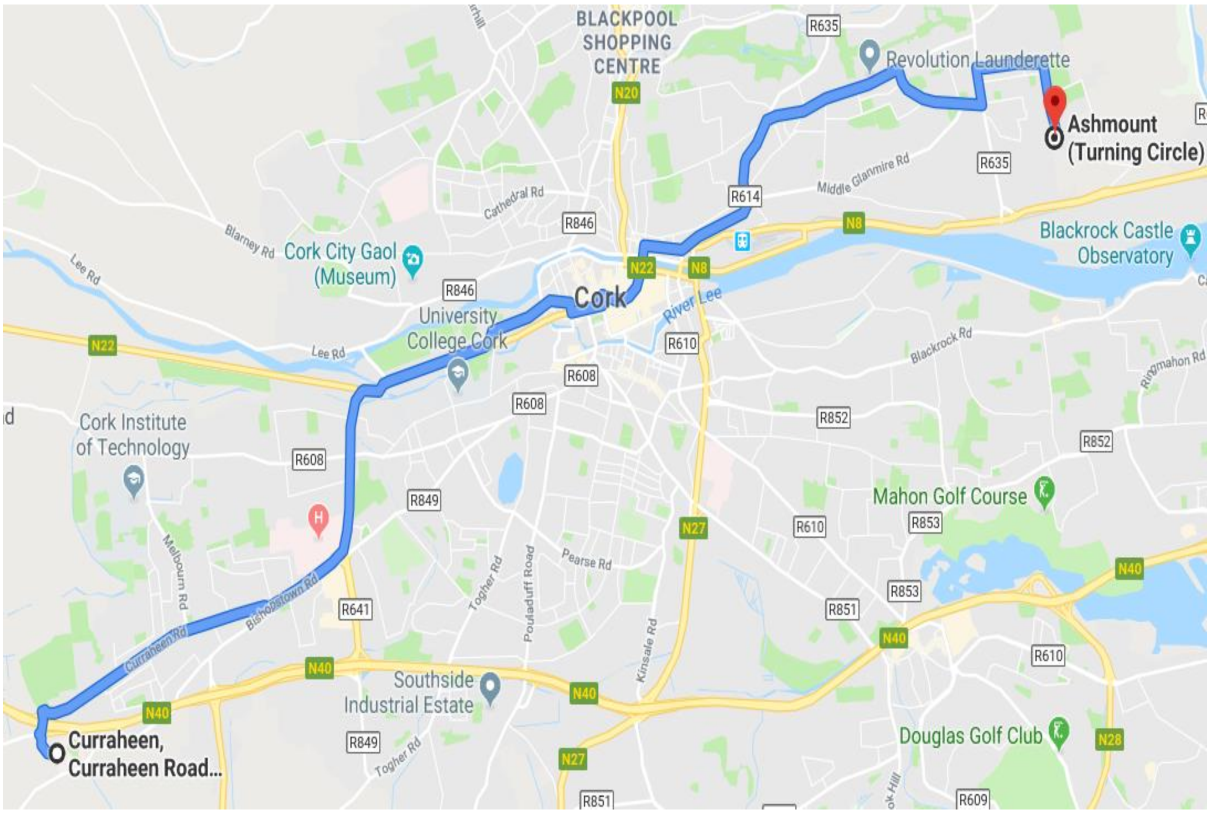

Figure 1: Bus route 208 snapshot from google maps.

Figure 2 from Moovit website, illustrates the bus stops highlighted on the chosen bus route number 208.

Figure 2: Bus service stops on route 208.

2.6 Distance Between Two Latitude and Longitude Pairs

There are a lot of calculation methods available for calculating the distance between two latitude & longitude pairs. The well-known efficient distance computation formula used in this thesis was Haversine formula (Chamberlain, 2013) and its formulation (Vivek, 2013) is given in Eq. (1).

Formulation:

$$D = 2r\ arcsin(\sqrt{Haversine(\emptyset 2 - \emptyset 1 + \cos\emptyset 1 \cos\emptyset 2\ Haversine(\lambda_1 - \lambda_2))}\ ) \qquad \text{Eq. (1)}$$

Where r is the radius of the earth (6378.1 km). D is the distance. $\emptyset 1$, $\emptyset 2$ indicates the latitude of point 1 and point 2. λ_1, λ_2 indicates the longitude of point1 and point 2.

Chapter 3
Data Exploration

3.1 Data Pre-Processing

In general, this section describes the approaches implemented for data exploration. Collected data is pre-processed so that it is ready for data analysis. Both datasets are loaded in local Db tables which makes easy extraction from both datasets. GPS & TFI dataset was made into three and four Db tables respectively as shown in figure 3.

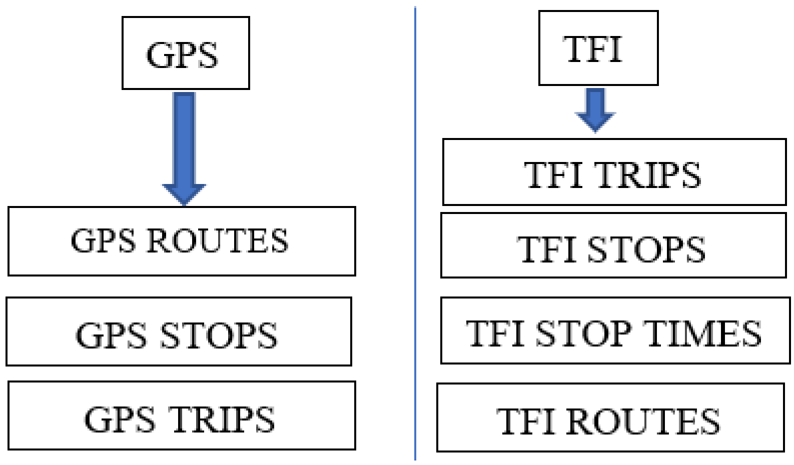

Figure 3: Datasets loaded in respected Db tables.

As mentioned earlier there are a total of 36,083,298 data points recorded overall in GPS data, with the help of SQL statement using GPS route id of bus route 208, distinct trip ids of the selected bus route are obtained from GPS dataset and further reduce the overall data points to 3,089,340. There are 807 (403 upstream and 404 downstream) unique trips recorded in the GPS data in the span of 41 days initiating from 06:30 AM to 11:30 PM. Each trip id obtained operates on the same route at different times of the day. Each GPS data records the bus's latitude, longitude and other 14 features of a bus at a time instant, for clustering the extracted data together, we group the data of bus route 208 by trip id, direction and order the data by poll-time at which the data was recorded, this will give the route trajectory for the same vehicle. Data for a single bus that operates in our selected study can be retrieved from the route by giving a unique trip id in a SQL statement. Figure 4 demonstrates the work flow of extracting single day trip.

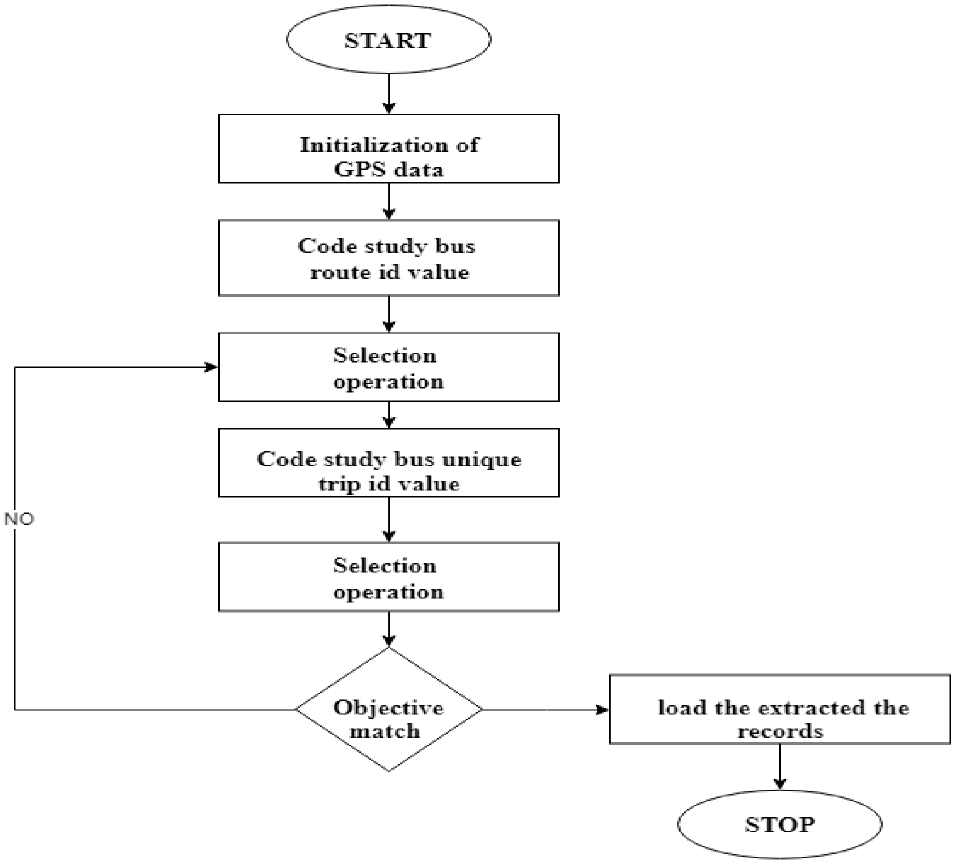

Figure 4: Flowchart for extracting data records for a single day.

Fetched data records from a single trip id contain the data for a whole day till the GPS entry was recorded. Retrieving one trip data out of a whole day trips data require some manual task, from the TFI dataset we can get the scheduled time of bus 208 running periodically. Now, for example, let's assume the obtained data is for bus starting at 8:00 AM, from GPS, recorded entries, the timestamp that is closest to selected scheduled time of bus 208, which implies that the bus would have initiated its journey at this timestamp. This data is extracted and further analysed to compute all data points of a single bus journey.

It is tough to identify the bus stops in the collected GPS data, because, it is basically the latitude, longitude from GPS, entry recorded on a route at every 20 seconds and does not highlight any significant information about any bus stops position, which also implies that data does not provide dwell time of any bus at the depot. So firstly, with the help of the TFI dataset and using the haversine distance formula, the minimum distance of each fixed latitude & longitude position of bus depots to every upcoming GPS latitude & longitude data record entry on the headway route from the time of journey starts is computed. This will be help to identify bus stops in the extracted trip of GPS data. This has been done with the assumption that the time taken to reach one bus stop to another includes dwell time.

From TFI dataset, with help of SQL statements using inner joins on stops and trips DB tables, route information is extracted for the bus stops and their fixed locations for the selected study route of bus 208, shown in Table 12. Along with this SQL statements using inner joins on stops and stop times table of TFI dataset, arrival time of bus at these fixed bus stops were noted.

Bus Stop Sequence	Bus Stop Name	Bus Stop Latitude & Longitude Position
0	Curraheen, Curraheen Road (Marymount Hospice)	51.870897,-8.543022
1	Curraheen Rd (Cork Tech Park)	51.873910,-8.540734
2	Curraheen Rd (Curraheen Estate)	51.874721,-8.538362
3	Curraheen Rd (Curraheen Church)	51.876697,-8.532751
4	Curraheen Road (Spioraid Naoimh)	51.877981,-8.529048
5	Curraheen Rd (Deanshall)	51.878720,-8.524801
6	Curraheen Rd (Westgate Rd Junction)	51.879190,-8.522032
7	Curraheen Rd (Firgrove Gardens)	51.879619,-8.518609
8	CUH (Bishopstown Rd)	51.882002,-8.510096
9	Wilton Rd (Opp Credit Union)	51.884334,-8.507144
10	Wilton Road (Avoca)	51.886880,-8.506824
11	Dennehys Cross (Opp Cork Farm Ctr)	51.889505,-8.506723
12	Victoria Cross (Victoria Lodge)	51.892042,-8.506214
13	Western Rd (Opposite UCC IT Building)	51.893842,-8.499551
14	Western Rd (Opp Castlewhite Apartments)	51.894219,-8.497477
15	Sundays Well, An Oige Hostel Western Road	51.894743,-8.494621
16	Western Road (Antone Guest House)	51.895388,-8.491039
17	Mardyke Walk (St. Josephs School)	51.896640,-8.488278
18	Mardyke (Presentation College)	51.897238,-8.484884
19	Sheares Street (Mercy Hospital)	51.898298,-8.480711
20	Grand Parade (Argos)	51.898202,-8.475580
21	Cork City, Patrick Street	51.899191,-8.471086
22	Saint Patricks Quay, Mccurtain Street Cork	51.901548,-8.468350
23	Summerhill North (O Donovans Shop)	51.902069,-8.463821
24	Summerhill North (Opp Cork Chamber)	51.903009,-8.460343
25	Summerhill North (St. Lukes Cross)	51.904378,-8.457393
26	Ballyhooley Road (Windsor Terrace)	51.907165,-8.457174
27	Old Youghal Road (Dillons Cross)	51.909343,-8.454289
28	Old Youghal Rd (Opp St. Josephs)	51.910345,-8.448457
29	Old Youghal Rd (Opp Service Station)	51.911218,-8.443320
30	Iona Pk (Opp Mayfield Family Practice)	51.911776,-8.438267
31	Colmcille Avenue (Opposite Iona Green)	51.910441,-8.434911
32	Colmcille Ave (Opp Garda Station)	51.910222,-8.430709
33	Banduff, Mayfield Lotabeg	51.912674,-8.423814
34	Mayfield, Lotabeg Green	51.912675,-8.421037

35	Ashmount (Opp Junction)	51.909508,-8.419540
36	Ashmount (Turning Circle)	51.908151,-8.419207

Table 12: Details of service stops in bus route 208.

3.2 Extracting Referenced Bus Service Stops in GPS Data

Figure 5, demonstrates the flowchart of extracting referenced bus stops positions in the GPS dataset and parallelly, an array is also recording timestamp values of all the referenced bus stop positions in the dataset.

TFI & GPS data points are referred by S and G respectively, where,

S^{lat} : Array of latitude value of fixed bus stop.

S^{long} : Array of longitude value of fixed bus stop.

S^{i} : Number of Bus stops presents in the bus route where i=0,1,2,3,4, 5,…,36.

G^{lat} : Array of Latitude value from GPS dataset.

G^{long} : Array of Longitude value from GPS dataset.

G^{Time} : Timestamp when the entry was made in the GPS.

G^{k} : Length of GPS data records in a single trip, where k = 0,1,2,3,4,5,...,n.

G^{j} : Length of GPS data records in a single trip, where j = 0,1,2,3,4,5,...,n.

U^{i} : list of Computed distance from data points where i = 1,2 for getting the minimum distance for the referenced points.

P^{i} : Pair of latitude & longitude, where i = 1,2.

After processing the fetched single trip data with the flowchart presented, it depicted that sometimes, the bus stops 1 metre ~ 5 metres away (forward-backward) from the actual bus stops, for that purpose the limit is set to 5 metres. In some cases, a bus may have skipped a stop due to the lack of demand in some periods, in this situation, bus stop referenced is moved to next GPS latitude, longitude position with the minimum distance in context to fixed bus stops position.

GPS entry having the minimum distance to these fixed bus stops locations from the TFI dataset are considered to be the referenced bus stops in a single journey. After getting the referenced bus stop in the GPS data, two major tasks are performed, firstly, fetch the timestamps of each referenced bus stops whose entry was recorded and secondly, calculate the distance between each upcoming consecutive data record and referenced headway bus stops using haversine formula. Using all data records in between a pair of referred bus stop position is a better way to keep the track of bus route trajectory.

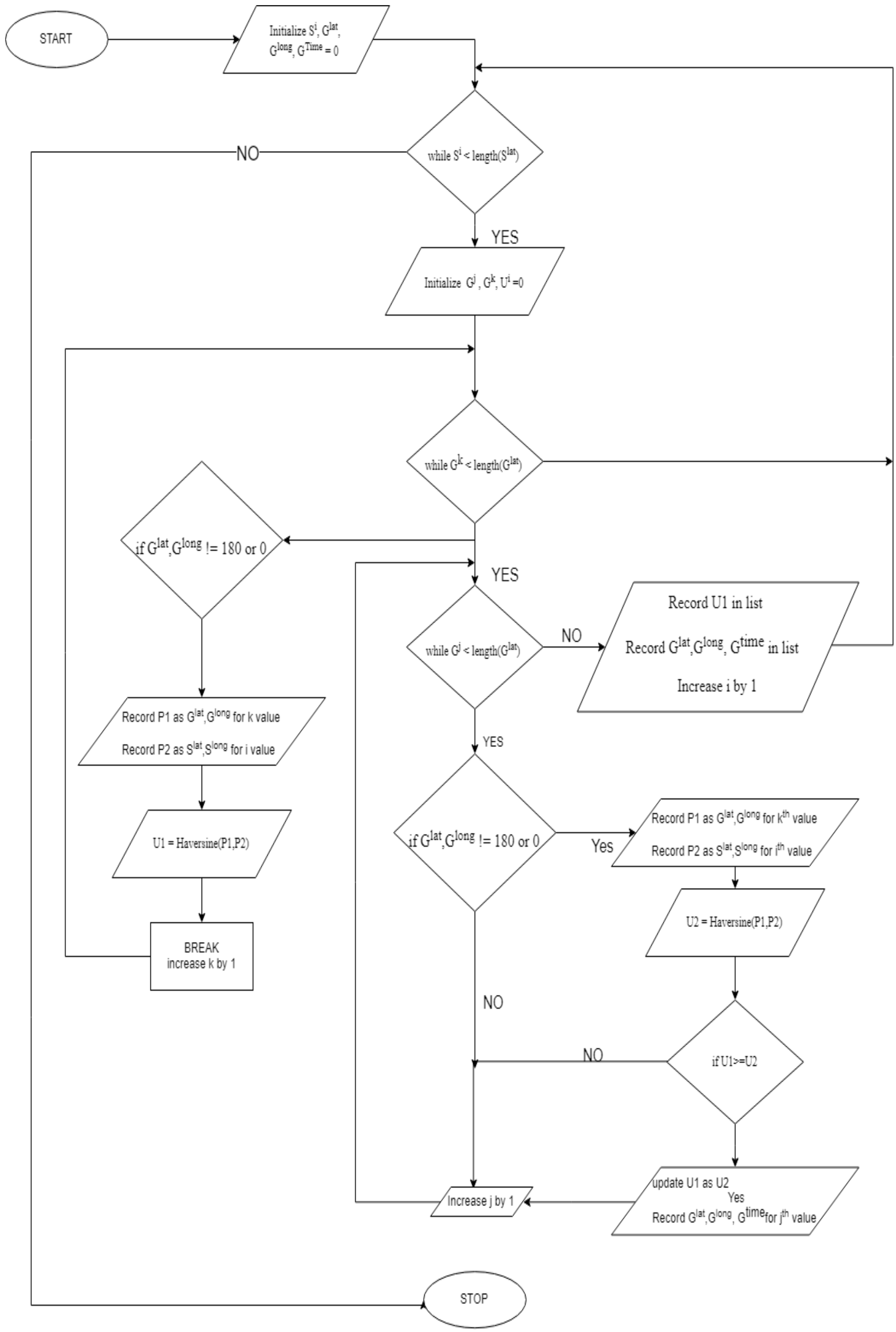

Figure 5: Flowchart for finding bus stops in fetched records.

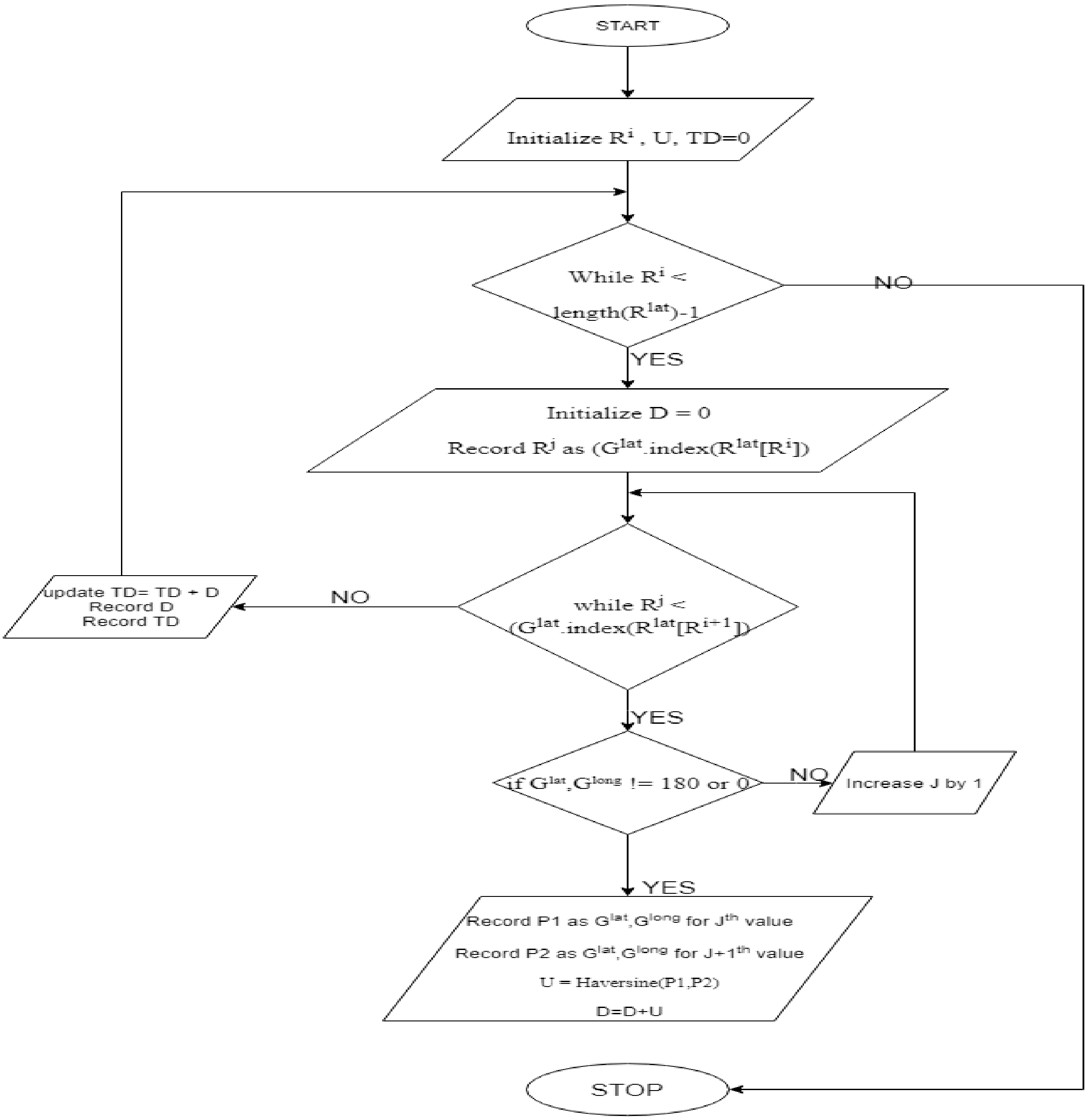

Figure 6: Flowchart for calculating distance between referenced bus stops.

3.3 Calculating Distance Between Two Referenced Bus Service Stops

Figure 6, demonstrates the workflow, how after getting reference bus stops in GPS data calculation of the cumulative distance between two referred bus stops and a total distance of the trip is obtained, using all recorded data points captured at every 20 seconds in between two bus stops, where,

R^{lat} : Array of referred bus depot's latitude from the GPS data.

R^{long} : Array of referred bus depot's longitude from the GPS data.

R^{i} : Number of bus stops in the route, where i=1,2,3,4,5,…,37.

D : Distance between consecutive bus depots.

U : Distance between two pair of latitude & longitude that falls in trajectory of GPS route

TD : Total distance of the route.

R^j : In GPS data, Indexed i^{th} position of referred latitude in G^{lat} array.

G^{lat} : Array of Latitude value from GPS dataset.

G^{long} : Array of Longitude value from GPS dataset.

P^i : Pair of latitude & longitude, where i = 1,2.

An example of calculated distance between two referenced bus stop's latitude & longitude data and cumulative distance from the source to the destination point for a single trip is shown in table 13. From the bus trajectory route obtained from GPS, on an average cumulative distance was found to be 11.67 Km, where the actual stretch of the route is of 11.79km. All calculations are proceeded with actual bus service stops sequence.

Bus Stop Sequence	Distance between two consecutive stops	Cumulative Distance (Km)
0	0	0
1	0.54	0.54
2	0.17	0.72
3	0.46	1.18
4	0.31	1.49
5	0.32	1.80
6	0.18	1.99
7	0.22	2.20
8	0.66	2.87
9	0.35	3.22
10	0.31	3.53
11	0.30	3.82
12	0.25	4.08
13	0.51	4.58
14	0.14	4.72
15	0.20	4.93
16	0.26	5.19
17	0.26	5.44
18	0.32	5.76
19	0.21	5.97
20	0.41	6.37
21	0.40	6.77
22	0.39	7.16

23	0.32	7.48
24	0.29	7.77
25	0.28	8.05
26	0.27	8.32
27	0.33	8.65
28	0.41	9.06
29	0.38	9.44
30	0.40	9.85
31	0.18	10.03
32	0.29	10.32
33	0.75	11.06
34	0.21	11.27
35	0.28	11.55
36	0.11	11.66

Table 13: Computed distance between two bus stops and their cumulative distance.

3.4 Calculating Bus Travel Time Between Two Bus Service Stops

As mentioned in section 3.1, an array of timestamp values is created of all the referred bus stops positions from the GPS data, this data feature will help to compute the time taken between stops and cumulative time taken at each bus stops.

This array will contain DATETIME values in 'YYYY-MM-DD HH: MM: SS' format, so extraction of HH: MM: SS value is done and converted into seconds by simply adding HH multiplied by 3600 plus MM multiplied by 60 plus SS. The array is updated with these calculated times in seconds for each bus stop. Now to calculate the time taken from the current stop to reach the next upcoming bus stop, current time(s) value is subtracted from the next time(s) value.

Along with this, from TFI dataset using SQL statement with inner joins on Stops and Stops time tables, another array is created containing scheduled arrival time at every stops for bus 208, starting at 8:00 AM, This array will contain TIME values in 'HH: MM: SS' format, so this will be converted into seconds by simply adding HH multiplied by 3600 plus MM multiplied by 60 plus SS. This array is updated with these calculated times in seconds for each stop. On further exploration, the delay time of bus arrival at every bus stop was calculated by the difference between two arrays. In some cases, bus originating from source starts with some delay, this may be because of passengers boarding and buying a bus ticket. It was also seen that some buses start 1~2 minutes ahead of the scheduled time from the source.

3.5 Calculating Speed Between Two Bus Service Stops

As known, speed is the distance travelled per unit of time, after calculating the distance between each bus stop and the travel time from one bus stop to another, further calculation was done for computing average speed over the links between bus stops for all bus trips.

With this extraction of these computed features from a single trip of the bus 208 was completed. Extracted feature are shown for example in table 14.

Stop	Distance (Km) between two stops	Total distance(km)	Total time(secs) at each stop	Average Speed (Km/h)	Delay (Secs)
0	0.00	0.00	0	0.00	-143
1	0.59	0.59	80	26.66	-163
2	0.28	0.86	100	49.89	-153
3	0.30	1.17	120	54.27	-53
4	0.33	1.50	200	14.87	-73
5	0.32	1.82	260	18.91	-103
6	0.18	2.00	300	16.64	-113
7	0.13	2.14	340	12.15	-123
8	0.75	2.88	520	14.92	-123
9	0.36	3.24	580	21.41	-63
10	0.29	3.52	620	25.65	-73
11	0.29	3.81	700	12.98	-33
12	0.29	4.10	761	16.94	-64
13	0.53	4.63	861	19.13	-44
14	0.23	4.86	881	41.05	-34
15	0.12	4.98	921	10.52	-14
16	0.23	5.21	961	20.92	6
17	0.21	5.42	981	37.90	16
18	0.30	5.72	1021	27.31	36
19	0.27	5.99	1101	11.97	16
20	0.40	6.39	1281	8.07	76
21	0.39	6.79	1661	3.74	-184
22	0.39	7.18	2121	3.06	-284
23	0.33	7.51	2201	14.90	-304
24	0.24	7.75	2241	21.64	-314
25	0.28	8.03	2281	25.52	-294
26	0.26	8.29	2401	7.83	-354

27	0.35	8.65	2561	7.97	-454
28	0.41	9.06	2642	18.11	-415
29	0.38	9.43	2722	17.05	-375
30	0.14	9.58	2802	6.32	-425
31	0.49	10.06	2822	87.51	-415
32	0.27	10.33	2862	23.86	-425
33	0.64	10.97	2982	19.33	-365
34	0.19	11.16	3042	11.42	-395
35	0.40	11.56	3122	17.80	-415
36	0.14	11.70	3142	24.96	-405

Table 14: Extracted features of a single trip from the GPS data records.

This is an example of a bus scheduled to start its journey at 8:00 AM from Curraheen Village's bus depot to the Ashmount's (Turning circle) bus depot. With this approach, we have extracted data for 150 trips of bus 208 on the same route at different times of the day.

With missing factors to tell what influenced the travel time in any trip, it is assumed that if data is clubbed as per morning afternoon, evening; data might have some similarities concerning speed. So, we clubbed/categorized data into three type "A", "B", "C" as per hour of the day, according to [7am-11am, 12pm-4pm, 5pm-11pm] buses for the morning, afternoon, evening respectively.

Chapter 4
Research Design and Methodology

4.1 Research Design

This section gives an introduction of the machine learning algorithm implemented for research design in this thesis. The section firstly describes the overall design of the machine learning algorithm used and then each part of the research is discussed in detail.

4.1.1 Artificial Neural Network Model

ANN is an information processing device, which is comprised of a large number of highly interconnected processing elements that are inspired by the way biological nervous systems, such as brain process information (Hecht-Nielsen, 1987). In this information processing system, the elements are called neurons that process the information. The neuron with n inputs calculates its output as shown in Eq. (2) (Johar Amita, 2015).

$$a = f\left(\sum_{i}^{n} w_i p_i + b \right) \qquad \text{Eq. (2)}$$

Where,

p_i is the value of i^{th} input.

w_i is the value of i^{th} weight.

b is the bias and

f is an activation function of the neuron.

There are various activation functions available in R package for neural net i.e. logistic or sigmoid and hyperbolic tangent (tanh) function. The activation function is required to establish a nonlinearity in the neural network. Figure 7, demonstrates the architecture structure of the ANN model. It comprises of three layers i.e. input layer, a hidden layer, and an output layer. In the input layer, the number of processing elements is equal to the number of input variables that are required to predict the output. In the output layer, it consists of the desired variables to be predicted. The hidden layer is the connection between the given input and the desired output. On the basis of the complexity of the problem, the number of hidden neurons between input and output layers is decided by the trial and error approach.

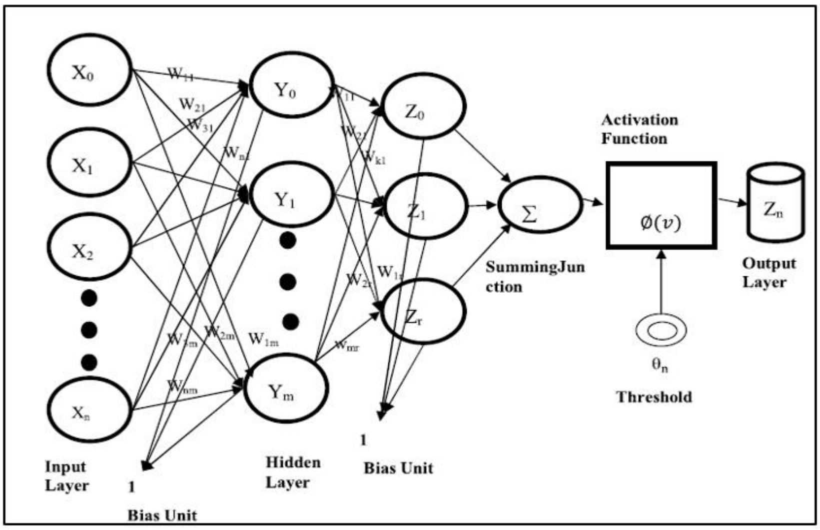

Figure 7: Artificial Neural Network structure

With trial and error approach, various type of network architectures is obtained. The processing elements are connected to each other by direct communication links, which is associated with weights. By adopting the weights of the communication links the ANN is supposed to learn a correlation between input and output. (Johar Amita, 2015).

4.2 Proposed ANN Model Development

The stop-based ANN model is developed by training stop-based data features such as the cumulative time of bus arrival over the route links between pair of stops. As the study route observation is the same for all 150 trips data, distance data feature will be the same over the route for all trips, time will be the major data feature here as it may vary during a different time of the day. An example of data records (structure of data frame) considered for analysis of our model, which will be used to predict bus arrival time prediction at upcoming bus stops are shown in table 15.

Trips	Stop 0	Stop 1	Stop 2	Stop 3	Stop 4 to 33	Stop 34	Stop 35	Stop 36	Time of day	Ave. Speed
1	0	100	140	200	-------	3462	3522	3542	A	18.24
2	0	120	160	221	-------	2902	2982	3102	A	20.39
3	0	80	100	120	-------	3042	3122	3142	B	20.89
4	0	80	120	200	-------	2362	2422	2442	A	22.84
5	0	100	160	220	-------	2801	2881	2902	A	17.53
6	0	60	100	160	-------	2902	3002	3022	C	17.20

Table 15: Structure of data frame created.

Due to the lack of explanatory factors like the number of passengers boarding or alighting at stops, traffic or weather conditions, etc. in the dataset, for testing purpose, it will feed model only with training dataset with respect to the factor for an hour of the day. For example, if it requires to get time predictions for bus 208 running at scheduled time of 1:30 pm, in order to avoid over-fit model, an ANN model will be provided with all trips from training dataset which are mapped in ‘B' hour of day, whole re-created dataset will be divided in to two parts training and testing data in the ratio 70 & 30 percentage respectively.

4.3 Proposed Working for ANN Model

In supervised learning, the desired output from output layer neurons is known, and the network adjust weight of connections between neurons to produce the desired output. During this process, the error in the output is propagated hack from one layer to the previous layer by adjusting weights of the connections. This is called the back-propagation method (Ranhee Jeong, 2004). In this thesis, the frequently used back-propagation network was implemented. The training algorithm of the back-propagation neural network involves four stages (Sivanandam et al., 2010):

- Initialization of weight,
- Feed Forward,
- Back-propagation of error signals,
- Updating weights and biases.

In the ANN structure, at the first stage, small random values are initialized to the weights. At the second stage that is feedforward, each input signal receives a unit from input variables which are transmitted to the output unit through a hidden neuron in between. If the output layer's unit does not produce the desired output then the third stage of the backpropagation method is used in which error is propagated back to all the units in the previous layer. At the last stage of the ANN structure, according to error signals, the weights and biases are updated. These steps are executed iteratively so that the error between a neural network's output and desired output can be minimized. The development procedure is completed once the network is fully trained with setting activation functions and by specifying its learning rate which is the amount that the weights are updated during training.

In order to better understand the prediction-modelling framework, Figure 8 shows a schematic diagram of a hypothetical transit route (Amer, 2004). The route presented consists of number bus service stops. When the transit bus n leaves stop *i*, the actual departure time is recorded by the GPS system. At this instant, ANN model will predict the next bus stop travel time $RT_{n(i,i+1)}$. Subsequently, the predicted travel time of the bus at the downstream bus stop *i+1* can be determined.

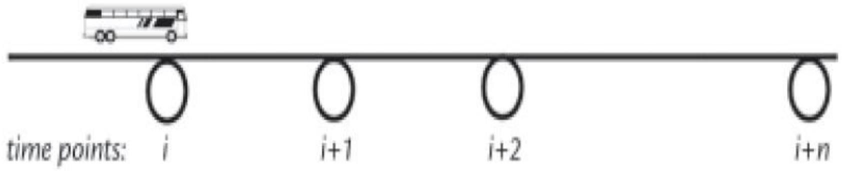

Figure 8: Schematic diagram of a hypothetical transit route.

Assuming that bus n is currently at stop i

$$AT_{n\ (i+1)} = DT_{n\ (i)} + RT_{n\ (i,i+1)}$$

Where:

$AT_{n\ (i+1)}$: is the predicted arrival time of bus n at stop *i+1*

$RT_{n\ (i,\ i+1)}$: is the predicted travel time between *i* and *i+1* from ANN model

$DT_{n\ (i)}$: is the actual departure time of bus n from stop *i*

To train the ANN model, since greater numbers of input variables can lead to longer computation times, it is inappropriate to include all the available historical data in the ANN model. Hence, to balance prediction accuracy and computation efficiency, this study only selects the historical travel time of the last five preceding bus stops that passed the target bus stop. To predict the bus travel time T_i at any bus stop *i*, we have considered last five [T_{i-1}, T_{i-2}, T_{i-3}, T_{i-4}, T_{i-5}] recent bus service stop's travel time with an average speed recorded of the trip as input variables to our ANN model. This consideration imposes a limitation to our ANN model that it cannot predict travel time for the first five bus stops. Reason of using an overall average speed of bus trips is later explained in section 5.1.

Overall, it is considered that the ANN model with three layers where a number of input nodes to the model will be six, a single hidden layer with two neurons and one output node. Reason of using two neurons in the ANN model is later explained in section 5.2. The fully connected ANN network is trained with sigmoid activation function which is usually chosen to deal with complex transportation systems (Steven, 2002). The learning rate is set lowest to 0.01 with SSE error function. Because the model is trained a sigmoid function, the linear output is set to False. Sigmoid function is given by Eq. (3):

$$f(x) = \frac{1}{1-e^{-x}} \qquad \text{Eq. (3)}$$

Proposed ANN model works best when the inputs and outputs are normalized roughly in the range [0, 1], later explained in section 5.3. For normalizing the input and output values according to the range Eq. (4) was used as shown below:

$$X_n = \frac{X - X\min}{X\max - Xmin} \qquad \text{Eq. (4)}$$

Where X_n gives the normalised value, X is the original value, X_{min} and X_{max} are the minimum and maximum value of X. In the evaluation phase, Mean Absolute Percentage Error performance metrics and Symmetric mean absolute percentage error metrics is used to estimate results from the ANN model.

4.4 Creating Random Observations for Test Records

In this thesis, it has been proposed that for testing the ANN model, some random scenarios/observations for current bus positions will be generated in the testing sample dataset, for which an algorithm is presented that will create this scenario cases where each test record will contain travel time of bus till any random bus service stops.

Algorithm for creating random scenarios/observations for current bus positions is shown in figure 9, where:

T : Testing sample data frame.

C : Current bus position data which we will initially set to zero and will be updated accordingly.

T_R : Total number of records in our test data.

C_i : Current trip from the total trips in the test data set where i = 1,2,3,4,…,T_R.

R_N : Number randomly generated in range 6 to 20, because there are 37 stops in total, maximum a random test trip can be generated where the last visited stop can be 36th stop and we can predict arrival time at 37th stop. The reason for using 20 as a maximum limit is that we want our model to predict more values rather than just one value, hence more values can be compared against actual values for a single trip for our general purpose of this research study.

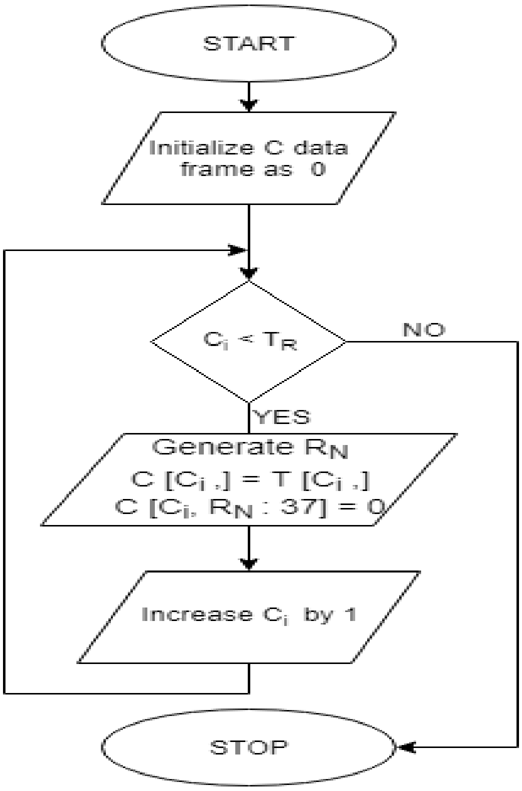

Figure 9: Flowchart for creating random last visited bus stops in our test records.

4.5 Extracting Last Visited Bus Service Stop of Test Records

An algorithm is also presented that will provide the ANN model, the last visited bus stop on any trip of the randomly created observations of last visited stops. After getting the last visited bus stop on any bus trip, the model will be able to provide predictions of bus travel time for the next bus stop which will be our target station, and it will continue predicting till the bus has reached the last bus stop.

Algorithm for finding the last visited bus stop and the target bus stop from where time travel is to be predicted till the last bus stop of the trip is shown in figure 10, where

T : Testing sample data frame.

C_T : Current(record) test trip of the test data frame.

T_C : Total number of stops in our test data set i.e. 37. (Stop0 to Stop36).

i : Stop number, $i = 0,1,2,3,\ldots,T_C$.

L_{VS} : Last visited the bus stop on the current test trip.

T_{BS} : Target bus stop from where time travel is to be predicted till the bus reaches the last stop of the trip.

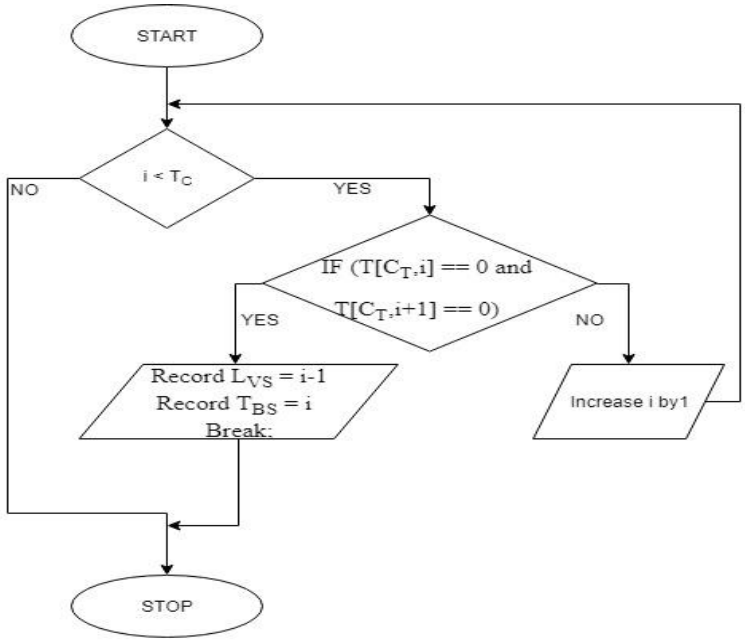

Figure 10: Flowchart for finding the last visited bus stop and the target bus stop.

4.6 Variable Importance Check

In this section, a null hypothesis for speed variable input was tested, which highlights if speed variable input is an important variable to the model or not. To check, bus data from the morning hours of a day was considered. Figure 11, demonstrates the scatter plot of time(seconds) taken by bus to reach the last stop of the route versus average speed(km/hr) of bus over the whole route. on the x-axis, y-axis we have time and average speed respectively. From the plot, it is determined that bus traveling at lower speed has taken more time to reach the final destination of the bus route and bus traveling with higher speed has taken less time. This shows there is some dependency on speed over time.

For the null hypothesis, it is assumed that the error term e in the linear regression model is independent of speed, and is normally distributed, with zero mean and constant variance. It can be decided whether there is any significant relationship between speed and time at last bus stop by testing the null hypothesis that $\beta_1 = 0$ at 0.05 significance level.

To Check,

$$H_0: \beta_1 = 0 \quad \text{Eq. (5)}$$

$$H_1: \beta_1 \neq 0 \quad \text{Eq. (6)}$$

The lm function in R to a formula that describes the variable time taken by variable speed is a pplied and save the linear regression model in a new variable Check.lm.

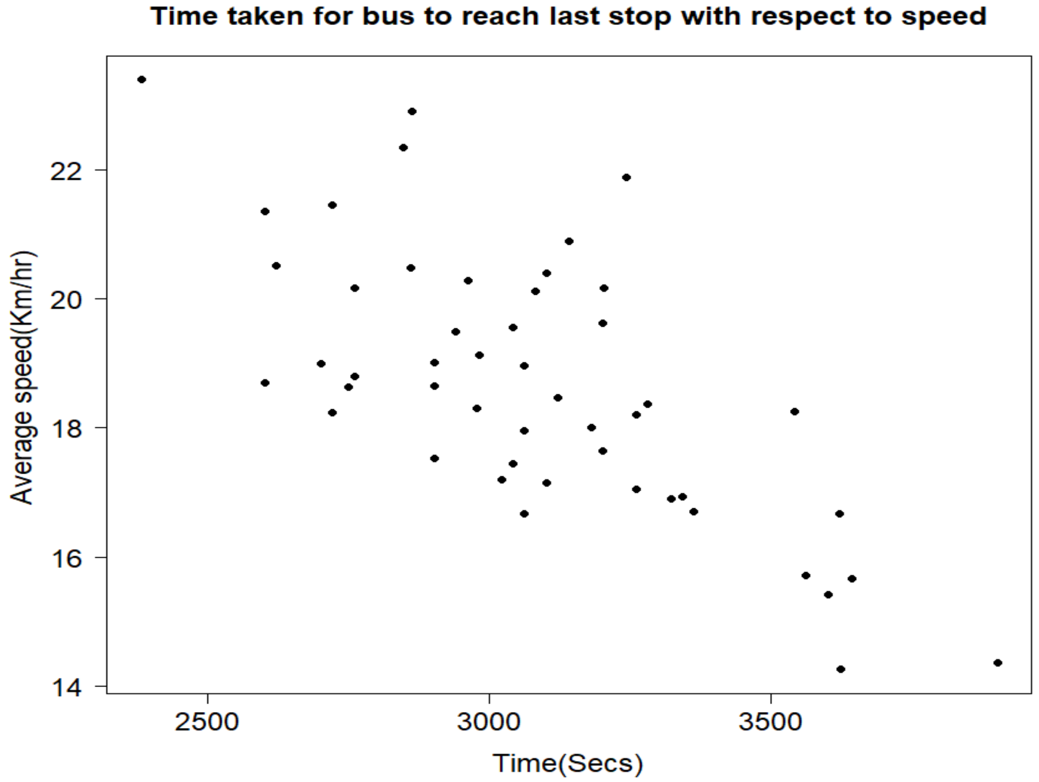

Figure 11: Plot for time taken to reach last bus stop vs average speed.

> summary(Check.lm)

Call:

lm(formula = Stop36 ~ speed, data = bus_data)

Residuals:

Min	1Q	Median	3Q	Max
-699.54	-132.10	-4.55	184.62	643.41

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5592.38	549.57	10.176	3.49e-10 ***
speed	-136.80	27.78	-4.924	5.04e-05 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 308 on 24 degrees of freedom

Multiple R-squared: 0.5025, Adjusted R-squared: 0.4818

F-statistic: 24.24 on 1 and 24 DF, p-value: 5.044e-05

From the summary results of Check.lm, it was found that the p-value is much less than 0.05. Therefore, the null hypothesis that $\beta 1 = 0$ is rejected. Hence a significant relationship can be established from the linear regression model of the data between the variables speed and time taken by bus at any stop, which suggests that speed should be included in the model.

From the summary given by R, it is prevalent that if speed of any bus traveling in morning hours of a day is increased by 1 Km/h, there will be a decrease of 136.80 seconds in time taken to reach last bus stop of the route, which implies bus will reach to the last bus stops 136.80 seconds faster.

4.7 ANN Model Processing

Below are the steps that the ANN model structure will follow:

Step 1: Generating random observation of bus.

For example, a single record from testing sample data of bus route 208 in morning hours of the day [type A] is the current trip. A random observation of this trip is generated and update zero for bus stops whose travel time is to be predicted.

Step 2: Get last visited bus stop number.

From the algorithm we presented, last visited bus stop number on this trip will be fetched, so let's assume the bus has reached Stop9 and now it will be predicting travel time for the bus at Stop10 and onwards.

Step 3: Generate a Symbolic formula and perform ANN model training.

The model is structured in a way that only the last five bus stops travel time and average speed over the route will always be provided as input nodes. So, the initial symbolic formula description of the model to be fitted will be as Eq. (7). Each time the model iterates, it will get trained with travel time from different set of last five recent bus service stop.

Stop10 ~ Stop5 + Stop6 + Stop7 + Stop8 + Stop9 + speed Eq. (7)

Next Eq.(s) will be:

Stop11 ~ Stop6 + Stop7 + Stop8 + Stop9 + Stop10 + speed

Stop12 ~ Stop7 + Stop8 + Stop9 + Stop10 + Stop11 + speed

Stop13 ~ Stop8 + Stop9 + Stop10 + Stop11 + Stop12 + speed

“ “ “ “ “ “ “ “

Similarly, the symbolic formula will be generated for each bus stops up to the final destination of bus 208 i.e. 37th bus stop which is Stop36 as per our data loaded in R.

Step 4: Computing prediction and storing the predicted value.
After the model is trained for Stop10 on training sample data, net results from the ANN model are used to compute predictions for stop10 against testing sample data from Stop5 to Stop9. We store this value is stored in a predicted data frame which will be used later for comparison of actual and predicted values.

Step 3 & 4 are repeated until the bus has reached the final destination.

Step 5: Performance metrics evaluations.
MAPE and SMAPE performance metrics are used to estimate results from the ANN model for all predicted values vs actual values from the target bus stop to the last stop of the bus trip.

4.8 Evaluation of ANN Model

Table 16 is presented to demonstrate the set of test records with random bus trips that were created by the algorithm for bus route 208, running in afternoon hours of the day which also implies that the ANN model was trained only on the class ‘B' data. With the help of the algorithm presented, the target service stop was recognized.

Test Trip	Last visited bus service stop
1	Stop11
2	Stop 8
3	Stop13
4	Stop 7
5	Stop 4

Table 16: Set of test records with randomly created last visited bus stops on a single trip.

ANN model was trained and tested with such random observations, and then MAPE and SMAPE metrics was estimated for actual GPS values and predicted bus travel time values from ANN model.

4.8.1 Performance Metrics

The MAPE formulation (Wikipedia) is shown in Eq. (8). It represents the average percentage difference between the observed value (in this case travel time at a transit stop) and the predicted value (in this case travel time at a transit stop) (Ranhee Jeong,2004).

$$\text{MAPE} = \frac{1}{n}\sum_{i}^{n}\frac{|y_i - y_0|}{y_0} \times 100\% \qquad \text{Eq. (8)}$$

Another performance metrics used was SMAPE, formulation (Wikipedia) shown in Eq. (9). It is an accuracy measure based on percentage (or relative) errors.

$$\text{SMAPE} = \frac{1}{n}\sum_{i}^{n}\frac{|y_i - y_0|}{(|y_i| + |y_0|)/2} \times 100\% \qquad \text{Eq. (9)}$$

where,

y_i = Predicted value (i.e. travel time at given transit stop).

y_0 = Observed value (i.e. travel time at given transit stop).

n = Number of fitted values.

Table represented (Lewis, 1982) containing typical MAPE% in figure 12, is used in this thesis to compare the MAPE values we estimate from the proposed ANN model

Interpretation of typical MAPE values

MAPE	Interpretation
<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

Source: Lewis (1982, p. 40)

Figure 12: MAPE value and its interpretations.

Table 17 is created to show the MAPE and SMAPE values were calculated for each test trip and the lowest MAPE & SMAPE that was estimated for the model was observed for test case trip 3 with 2.03% and 0.0202 error respectively. Figure 12 is used for interpretation of MAPE.

Test Trip	Target bus stop	MAPE %	SMAPE
1	Stop12 to 36	15.78	0.1447
2	Stop 9 to 36	6.44	0.0682
3	Stop14 to 36	2.03	0.0202
4	Stop 8 to 36	17.77	0.1551
5	Stop 5 to 36	7.25	0.0687

Table 17: Calculated MAPE and SMAPE error values.

Following is the snippet of the extracted output for test trip 3 predicted from Stop14 to Stop36 of the bus route in following manner by R.

Stop14 ~ Stop9 + Stop10 + Stop11 + Stop12 + Stop13 + speed

[1] "Actual value for stop 14 : 1080"

[1] "Predicted value for stop 14 : 1062.72"

Stop15 ~ Stop10 + Stop11 + Stop12 + Stop13 + Stop14 + speed

[1] "Actual value for stop 15 : 1160"

[1] "Predicted value for stop 15 : 1153.25"

Stop16 ~ Stop11 + Stop12 + Stop13 + Stop14 + Stop15 + speed

[1] "Actual value for stop 16 : 1200"

[1] "Predicted value for stop 16 : 1207.98"

Stop17 ~ Stop12 + Stop13 + Stop14 + Stop15 + Stop16 + speed

[1] "Actual value for stop 16 : 1220"

[1] "Predicted value for stop 16 : 1235.9"

Stop18 ~ Stop13 + Stop14 + Stop15 + Stop16 + Stop17 + speed

[1] "Actual value for stop 16 : 1240"

[1] "Predicted value for stop 16 : 1282.46"

. . . .

. . . .

. . . .

. . . .

Stop36 ~ Stop31 + Stop32 + Stop33 + Stop34 + Stop35 + speed

[1] "Actual value for stop 36 : 3622"

[1] "Predicted value for stop 36 : 3694.91"

Figure 13 is of the first neural network plot created for test trip 3 predicted for Stop14, which uses Stop9 to Stop13 and average speed as inputs, weight and error is adjusted of the bus route by R.

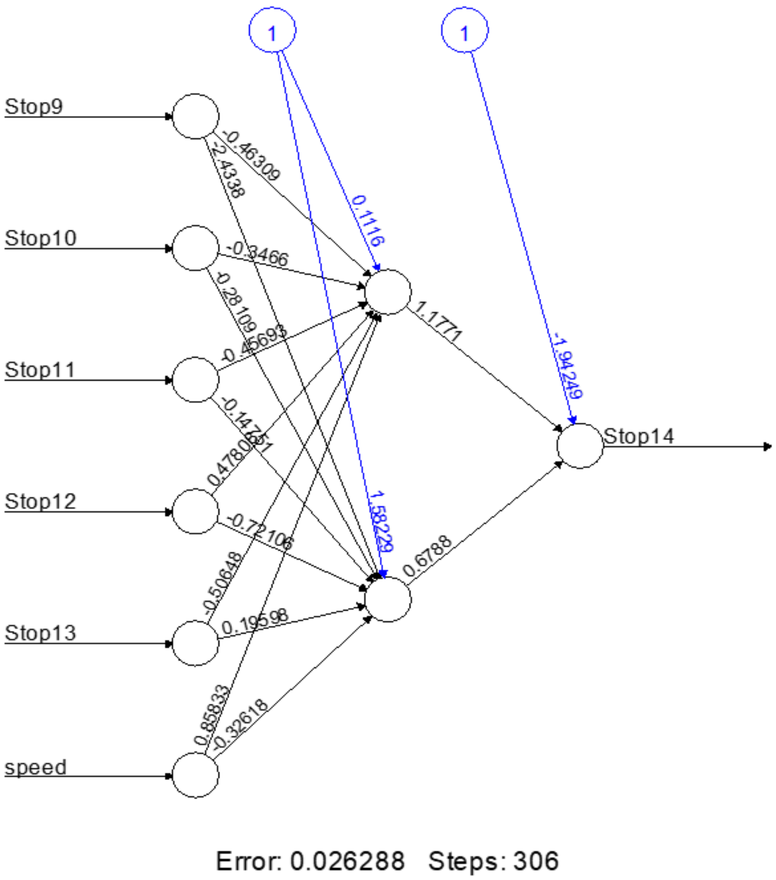

Figure 13: Neural net plot in R.

The figures (14,15,16,17,18), illustrates the GPS real-time and ANN model predicted travel time graph. On the x-axis, the bus service stop numbers are plotted, starting from the target station of each testing sample up to the last destination of the bus trip. On the y-axis, the time(seconds) taken at which a bus reaches any bus stops on its route is plotted. Each graph in the figures is labelled with test trip case, representing the cumulative time taken (seconds) for a bus from its target bus stop to each of its downstream bus stops, dashed lines depict the observed travel time from ANN model at each bus service stops marked with blue squares and continuous line represents the actual travel time from GPS data at each bus service stops marked with red triangles in the figures.

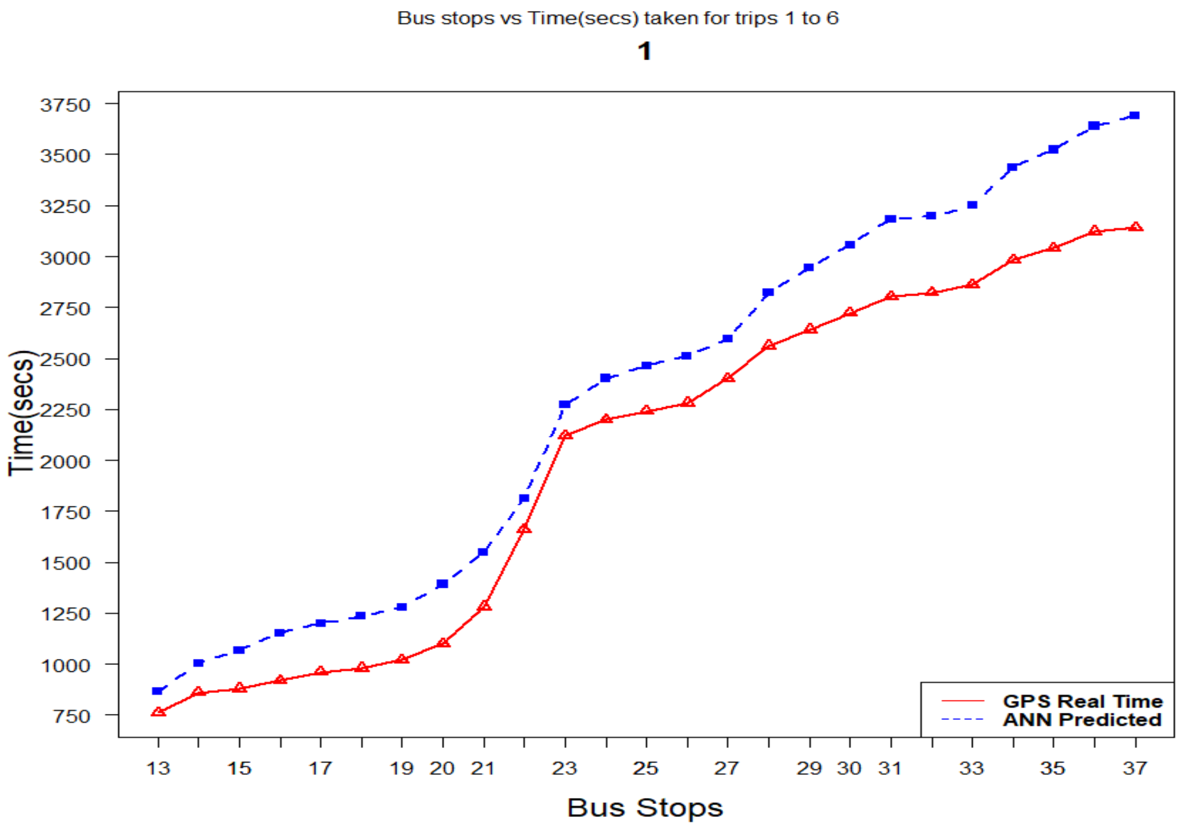

Figure 14: Plot actual vs predicted cumulative travel time for bus test trip 1.

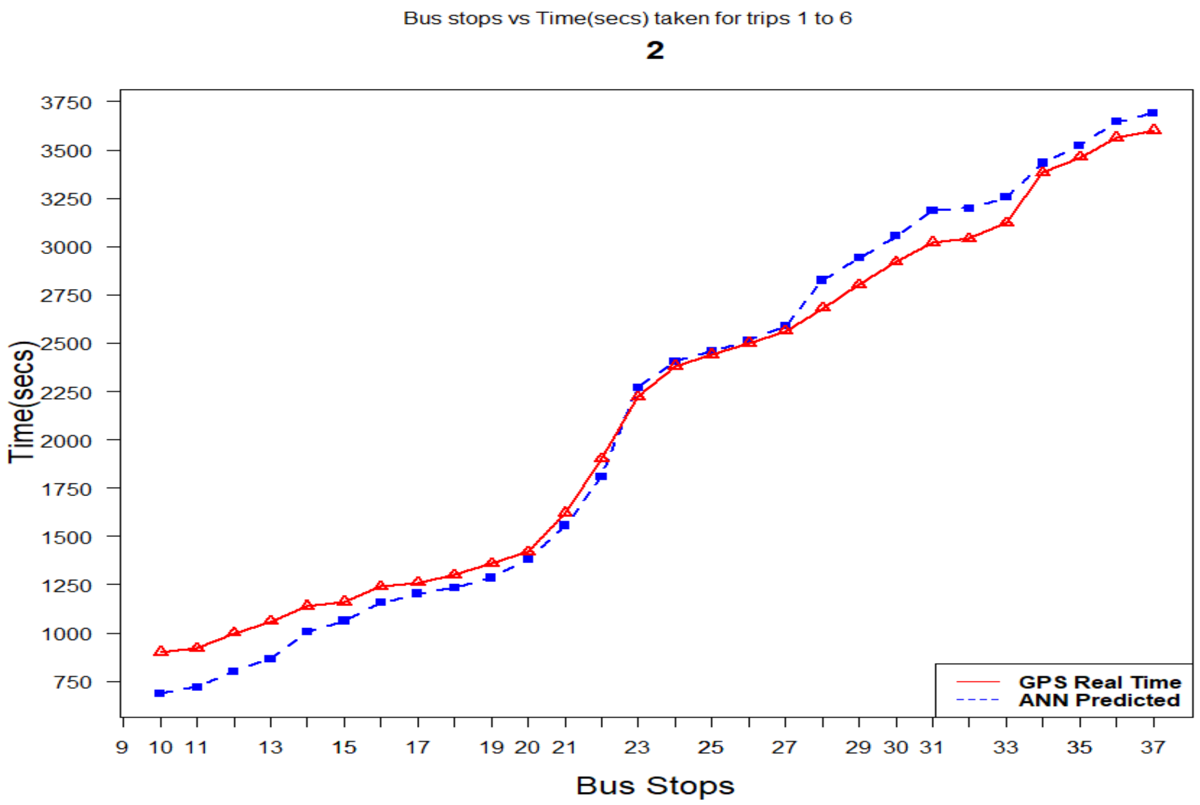

Figure 15: Plot actual vs predicted cumulative travel time for bus test trip 2.

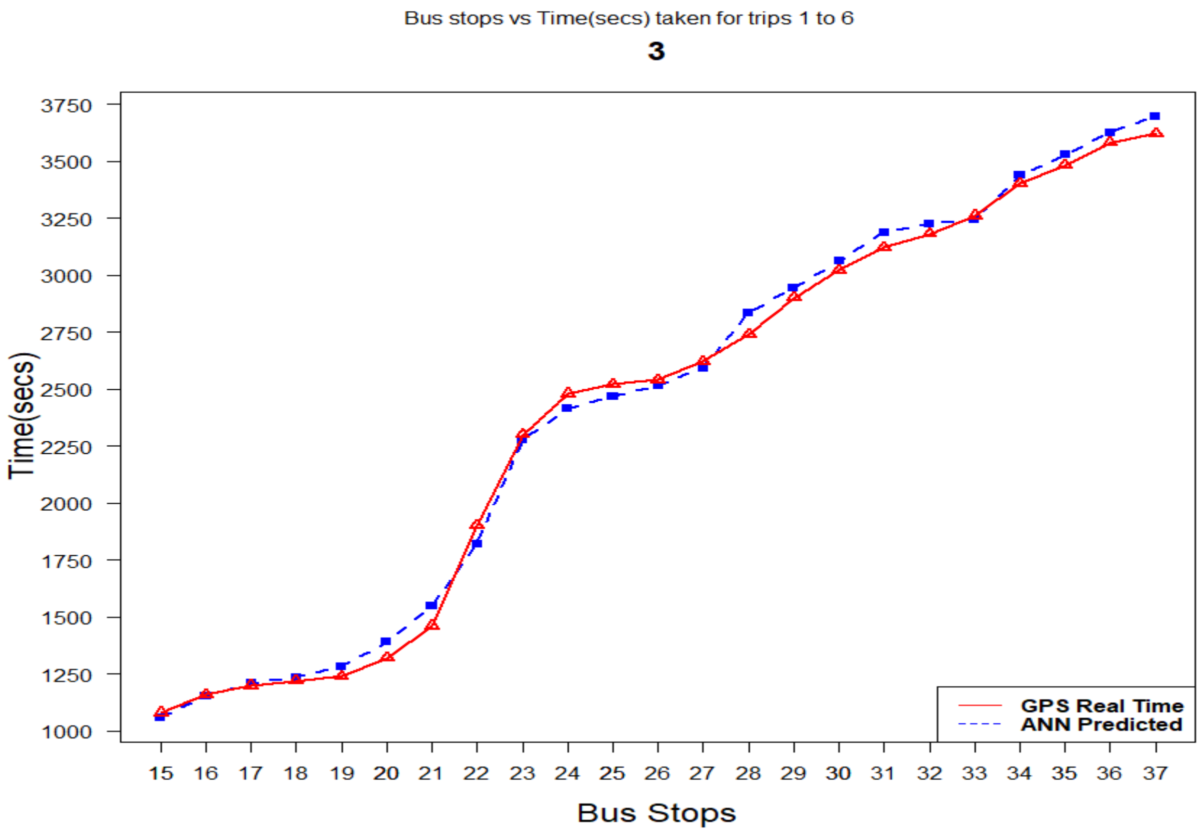

Figure 16: Plot actual vs predicted travel cumulative time for bus test trip 3.

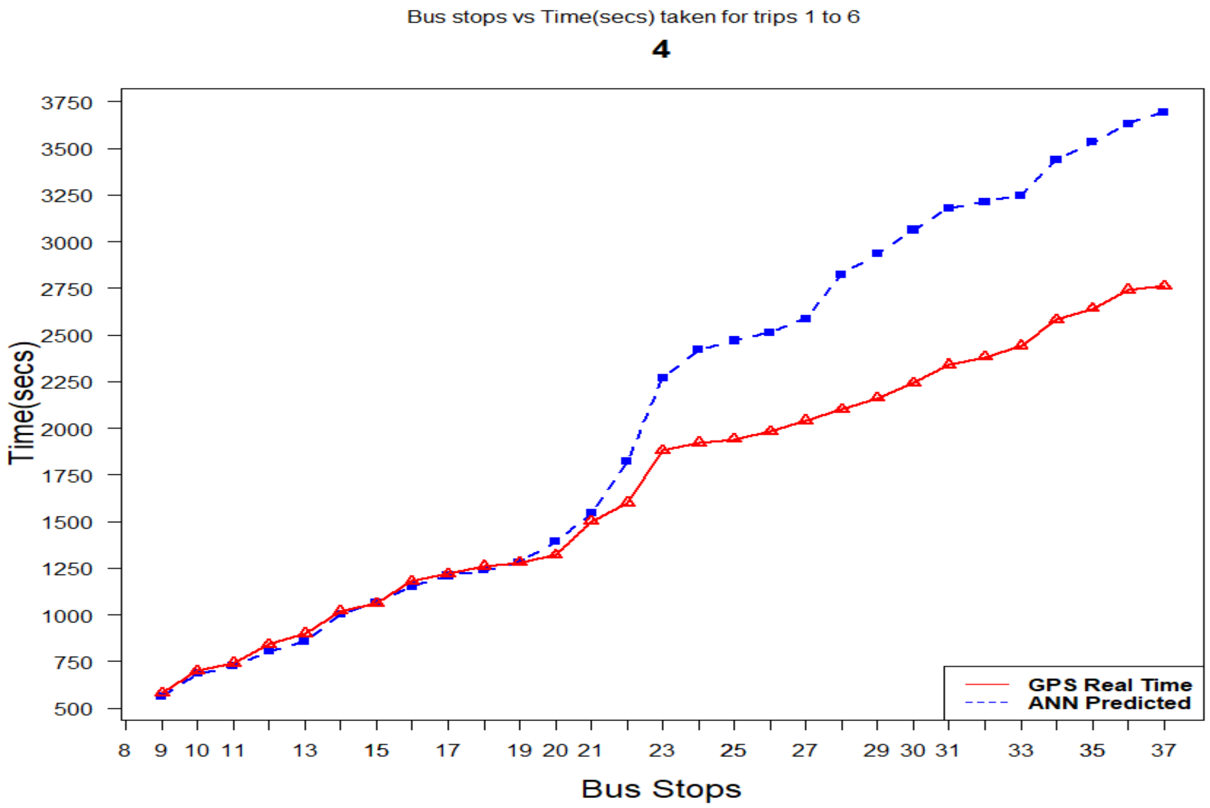

Figure 17: Plot actual vs predicted cumulative travel time for bus test trip 4.

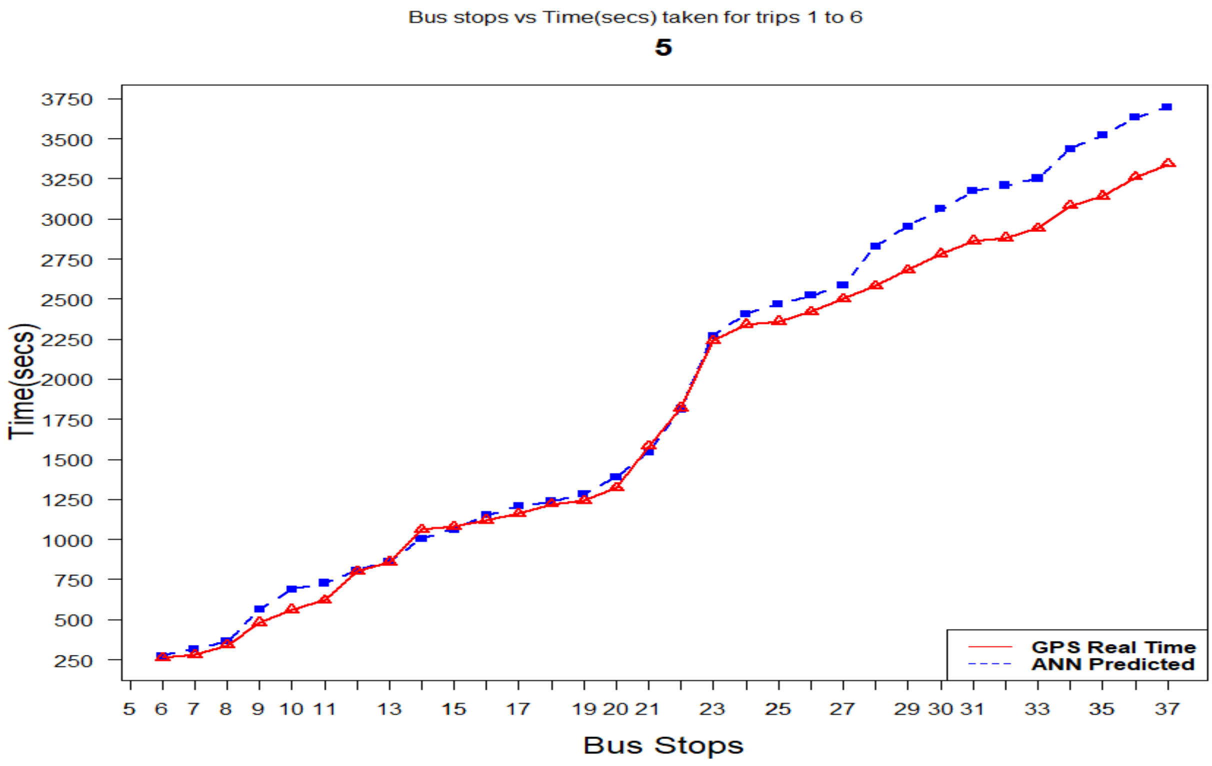

Figure 18: Plot actual vs predicted cumulative travel time for bus test trip 5.

Figure 16 is of test trip case 3, bus trip whose last visited bus service stop was Stop13, ANN model started predicting its travel time at bus service Stop14 and model was iterated until bus travel time at last service Stop36 was predicted. In table 17, MAPE% and SMAPE error values were calculated for this test trip 3. Actual GPS and ANN predicted bus travel time is presented in table 18 and table 19 respectively, it shows that bus had reached last stop at 3622 secs (60 minutes 22 seconds) where the ANN model predicted that bus would reach the last stop at 3627.22 secs (60 minutes 27.22 seconds), prediction meaning that this bus would reach last bus stop with difference of ~ 5 seconds, this would not be unexpected because factors like weather or traffic conditions are not trained in the model and hence be able to predict more accurately.

> *Actual*

Stop 14	Stop 15	Stop 16	Stop 17	Stop 18	Stop 19	Stop 20	Stop 21	Stop 22	Stop 23
1080	1160	1200	1220	1240	1320	1461	1901	2301	2481

Stop 24	Stop 25	Stop 26	Stop 27	Stop 28	Stop 29	Stop 30	Stop 31	Stop 32
2521	2541	2621	2741	2901	3021	3122	3182	3262

Stop 34	Stop 35	Stop 36
3482	3582	3622

Table 18: Actual bus travel time for test trip 3 from GPS data.

> Prediction

Stop 14	Stop 15	Stop 16	Stop 17	Stop 18	Stop 19	Stop 20	Stop 21
1058.72	1152.3	1208.99	1235.9	1282.46	1391.24	1550.01	1821.37

Stop 22	Stop 23	Stop 24	Stop 25	Stop 26	Stop 27	Stop 28	Stop 29	Stop 30
2277.79	2414.45	2468.34	2515.49	2593.42	2835.4	2946.4	3063.08	3188.72

Stop 31	Stop 32	Stop 33	Stop 34	Stop 35
3226.42	3244.75	3439.1	3528.88	3627.22

Table 19: ANN Model predicted bus travel time for test trip 3.

From the plot in figure 16 for trip test case 3, it can be observed that for Stop14, ANN predicted travel time is 1058.72 seconds whereas, our actual time of the historic trip is 1080 seconds. It means that the ANN model underestimated the value by ~21.28 seconds. For Stop15, the ANN model predicted less by ~8 seconds. Likewise, for Stop16 ANN model estimation was over predicted by ~9 seconds. It was observed that the ANN model has predicted either very precisely or close enough to the actual trip travel time after the first target bus service stop.

To understand more about MAPE & SMAPE and to get more insights about the travel time prediction from the ANN model, as shown in the figures (19, 20, 21, 22, 23), the observed and actual the bus travel time between each subsequent bus service stop was further analysed. Observations can be made on, how close travel time is predicted by the proposed ANN model with respect to actual bus travel time from GPS data, figure 21 is of bus test trip case 3, with MAPE and SMAPE value of 2.03% and 0.0202 error respectively, some conclusions can be made with considering the limitation of GPS dataset used that at some bus service stops, ANN model predicts either more than real travel time or less than actual real GPS travel time. It was also observed that the clustering of the whole data leads to a larger MAPE and SMAPE. This would not be unexpected because the clustering explicitly accounts for different congestion and demand levels associated with different parts of the day.

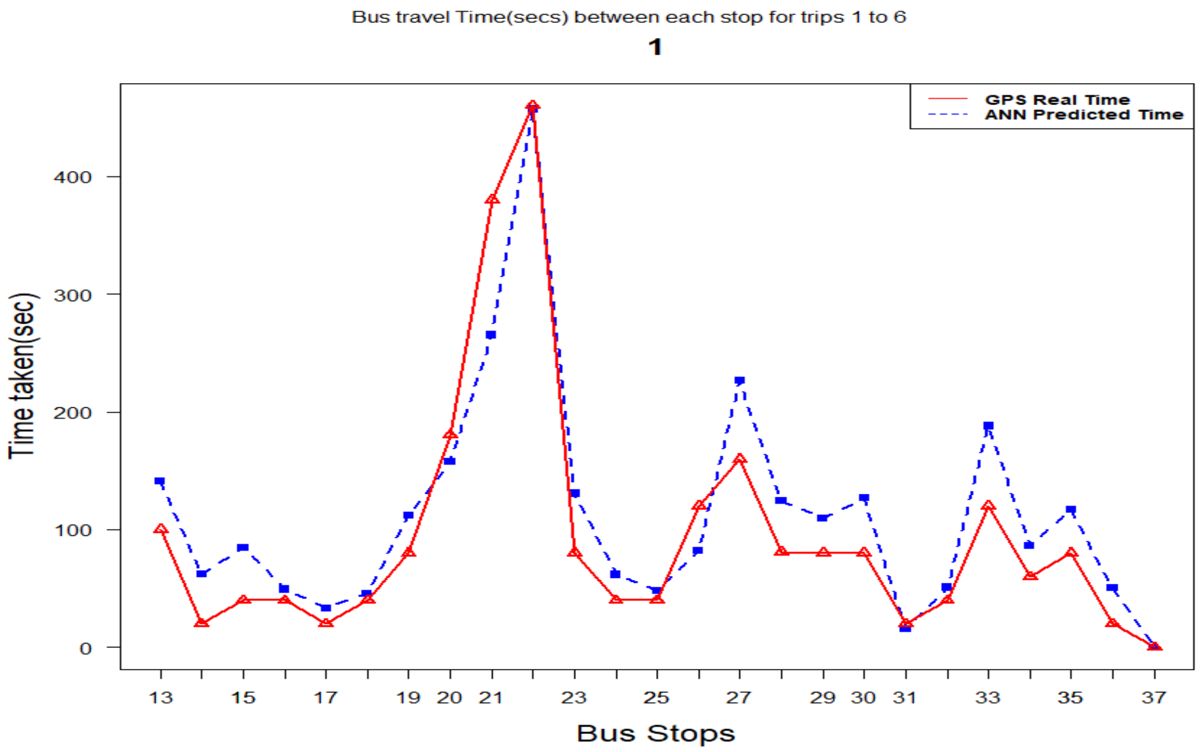

Figure 19: Plot actual vs predicted bus travel time for bus test trip 1.

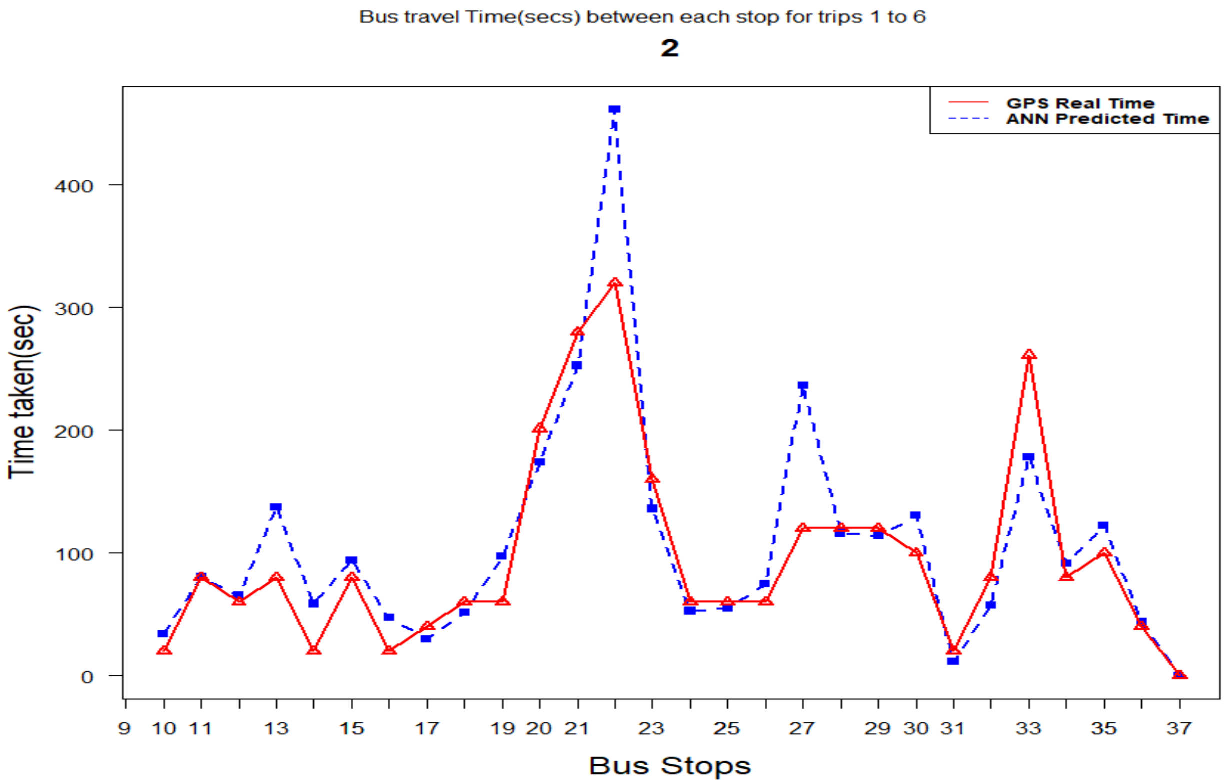

Figure 20: Plot actual vs predicted bus travel time for bus test trip 2.

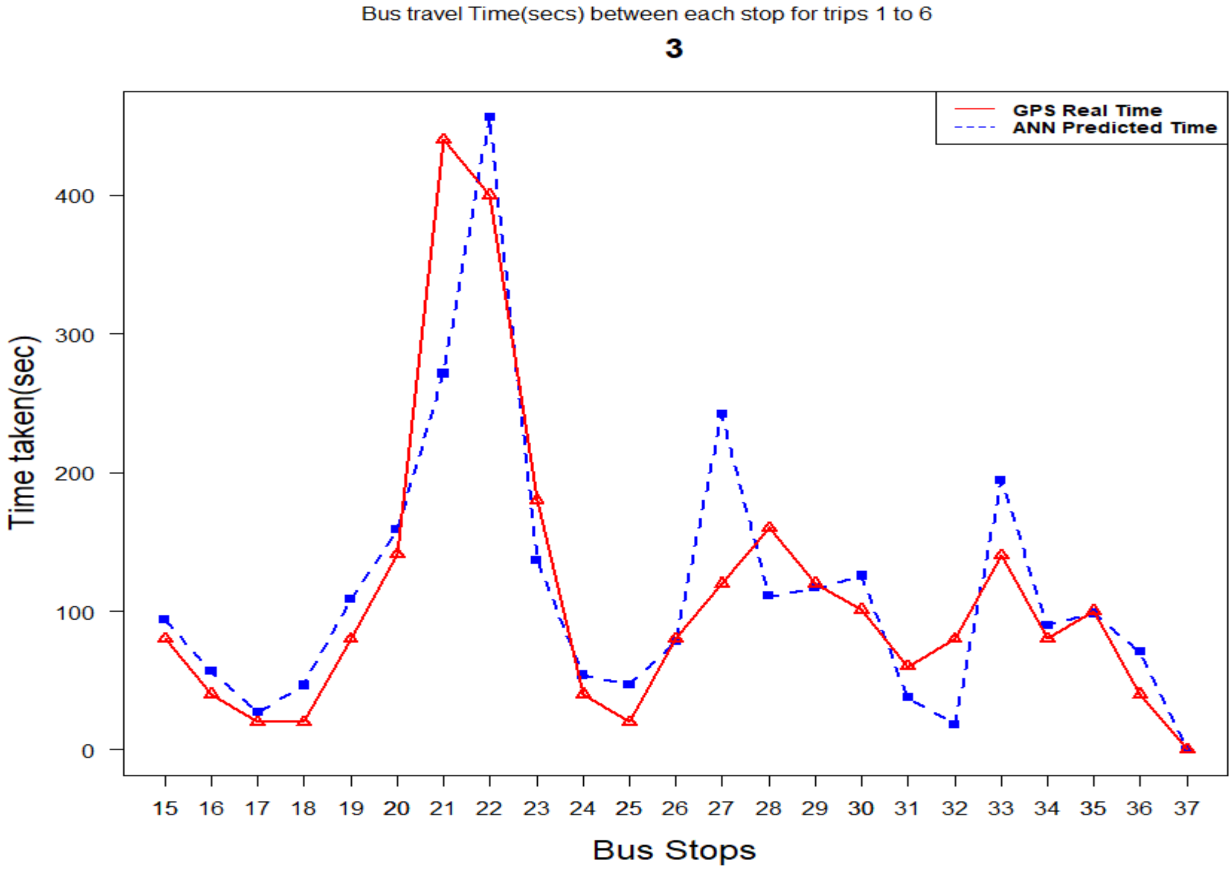

Figure 21: Plot actual vs predicted bus travel time for bus test trip 3.

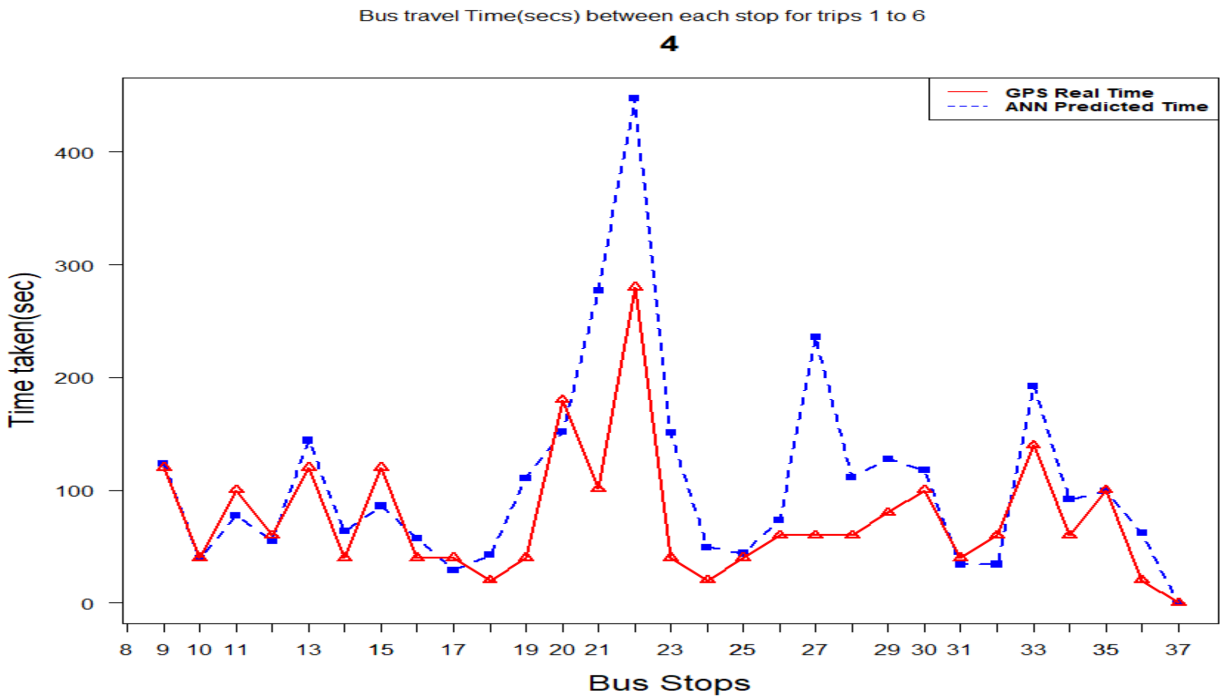

Figure 22: Plot actual vs predicted bus travel time for bus test trip 4.

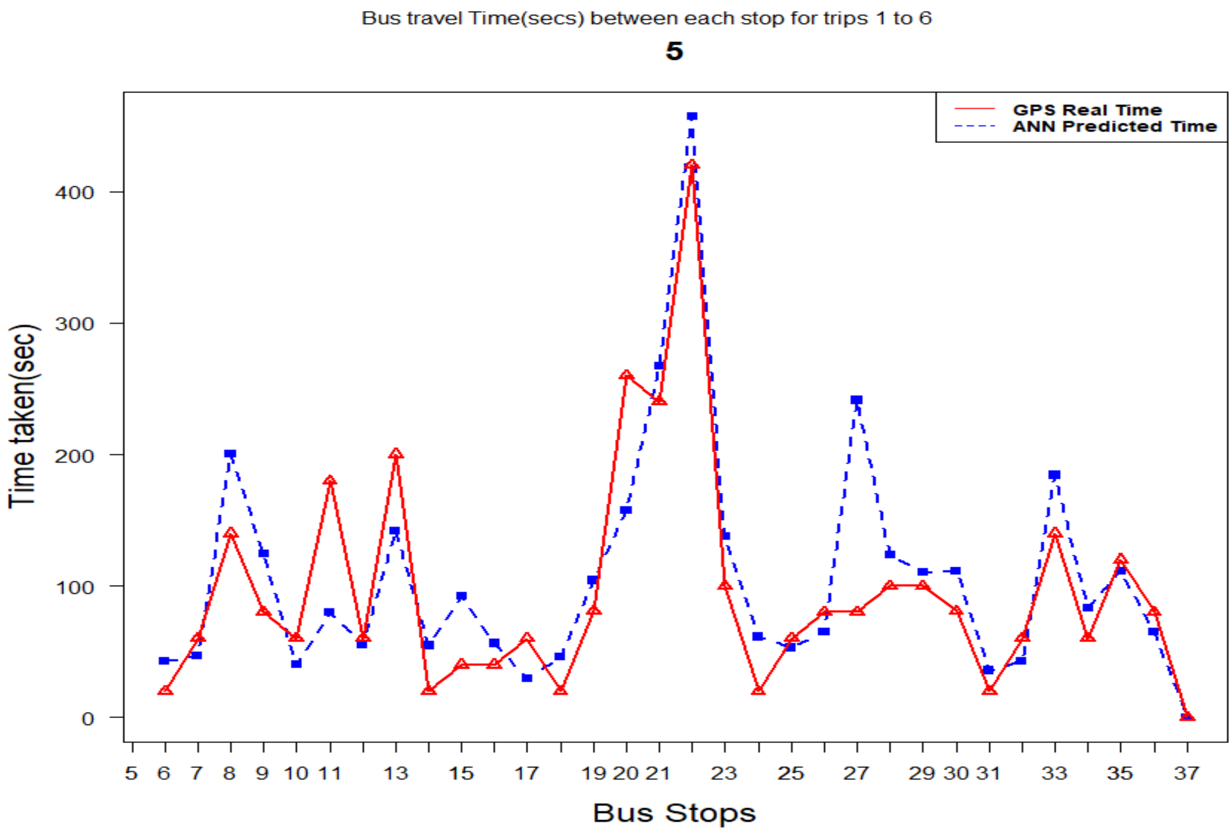

Figure 23: Plot actual vs predicted bus travel time for bus test trip 5.

Chapter 5

Arguments

Some arguments were made during the research work of this project development, which are listed out in below sections.

5.1 Average Speed Variable as an Input to the ANN Model

We have already discussed the importance of variable speed in our proposed ANN model and also checked the null hypothesis which proves that speed is significant to the model. In this thesis, we have proposed that an overall average speed of any bus trip over a route with travel times at last five recent bus service stops must be used as inputs to the ANN model during training and while evaluating the model, last visited bus service stop is identified to predict the next target bus service stop.

Ideally, for any bus traveling on any route has different speeds between different service stations at different hours of the day with which can be again due to unusual traffic or weather conditions, there can be unusual road blockage due to some reasons. For the prediction of travel time at any service stop, the ANN model must be provided with travel times and an average of different speeds between the last five recent bus stations. But as we are using historic GPS data points to train and test our ANN model, and general-purpose for this paper was to check how well the ANN model predicts travel time at bus service stop using certain parameters and check its model performance. Table 15, shows the structure of data frame used to train and test the model, using calculated average speed between any five bus service stops with different set of travel times at each iterative model training for training and testing inputs, would have an unstructured data frame and it will be difficult to predict travel time for this paper study, which limits to use an average speed over bus trips for training and testing the ANN model.

5.2 Identifying Number of Neurons in Hidden Layers

In this section, we identify the best number of neurons for the proposed ANN model. A test was performed to check the number of neurons required in the proposed ANN model. A single trip generated with a random bus stop as its last visited bus stop was taken for testing. Table 20 is created to identify the number of neurons giving the lowest MAPE% in testing.

This test was conducted by initially training our ANN model with a single neuron and kept increasing the neurons and hidden layers. Here an explicit test was carried out for the same single trip in each test to identify the number of neurons. From table 20, on using a single layer and a total of two neurons initialized in the proposed model, it was found to give the lowest MAPE% value. Also, a comparative study claims that a single hidden layer to be sufficient for ANNs to approximate any nonlinear functions (G. Zhang, 1998). Thus, a single hidden layer with two neurons can be used for the research work in training the ANN model.

Hidden Layer	Total hidden Layer neurons	MAPE%						Mean MAPE %
		Trip 1	Trip 2	Trip 3	Trip 4	Trip 5	Trip 6	
1	1	11.951	8.991	14.548	21.187	19.215	20.312	16.13
	2	11.946	8.904	14.553	21.489	19.218	20.229	16.09
	3	11.942	8.869	14.607	21.186	19.228	20.332	16.14
	4	12.155	8.856	14.564	21.237	19.143	20.193	16.17
	5	12.034	8.784	14.496	21.330	19.203	20.348	16.19
	6	11.990	8.855	14.480	21.029	19.235	20.451	16.22
2	(2+1) = 3	11.989	8.941	14.552	21.423	19.237	20.294	16.14
	(2+2) = 4	11.961	8.913	14.576	21.396	19.250	20.266	16.11
	(2+3) = 5	12.026	8.946	14.602	21.424	19.211	20.273	16.15
	(2+4) = 6	11.990	8.895	14.618	21.402	19.259	20.327	16.16
	(2+5) = 7	11.973	8.931	14.551	21.406	19.275	20.296	16.14
	(2+6) = 8	11.973	8.918	14.654	21.399	19.245	20.317	16.15

Table 20: MAPE calculation in different number of neurons.

5.3 Normalization Range of Input and Output Variables

In this section, the best normalization range of input and output variables with different activation functions was determined. A comparative study claims that the “Tanh” function produces both positive and negative values, shown in Eq. (10), which tends to yield faster training than “logistic” or “Sigmoid” functions that produce only positive values (Fan and Gurmu, 2015) Normalization range used was [-1,1]. A similar study claims the algorithm works best when the network inputs and outputs are normalized roughly in the range [-1, 1] (Johar Amita, 2015).

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \qquad \text{Eq. (10)}$$

In order to use the better normalization range with correct activation function in the ANN model proposed for this project, a test was performed where the ANN model is trained with below initial setup:

- Setup 1: Normalization range [0,1] with “Sigmoid” as activation function.
- Setup 2: Normalization range [-1,1] with “Tanh” as activation function.
- Setup 3: Normalization range [-1,1] with “Sigmoid” as activation function
- Setup 4: Normalization range [0,1] with “Tanh” as activation function.

From a practical point of view to test, morning bus trips type “A” having some random last visited bus service stops were used. Table 21, shows three bus trips as test records, which were the same in all the test setups. We have checked for the setup that gives us the lowest MAPE% and SMAPE error values, which will be considered the best setup for our proposed ANN model.

Test Trip	Last visited bus stop
1	Stop12
2	Stop6
3	Stop8

Table 21: Set of test records with randomly created last visited bus stops on a single trip.

Test Trip	Setup 1		Setup 2		Setup 3		Setup 4	
	MAPE %	SMAPE	MAPE %	SMAPE	MAPE %	SMAPE	MAPE %	SMAPE
1	9.11	0.0971	23.12	0.2917	9.91	0.1033	19.65	0.226
2	18.35	0.2193	33.49	0.4965	19.49	0.2366	26.98	0.35
3	17.98	0.1687	28.62	0.3809	18.48	0.1722	19.75	0.214

Table 22: Calculated MAPE and SMAPE values for different ANN setups.

Table 22 is created to show the MAPE and SMAPE values were calculated for each test trip, it was observed that test trip1 gets the overall lowest MAPE% and SMAPE error values in all four ANN setup, indicating that for our proposed ANN model, setup1 using “Sigmoid” activation function with input and output normalized in range [0,1] is best than other setups. Hence this setup is implemented in the project for ANN model development in this thesis.

Chapter 6

Results and Conclusion

Artificial neural networks predictive model has been implemented for a wide variety of transportation problems as a neural network automatically discover the relationship between the variables and naturally updates its weights and bias. In ANN modelling, depending upon the problem statement, one can choose the required number of variables as there is no upper limit on the number of variables. ANN provides flexibility as it can learn from a non-linear and complex model, it has generalization ability on unseen data, ANN has forecasting abilities, accuracy and some amount of fault tolerance in the prediction of travel time. For the design of neural network architecture in RStudio tool, mostly trial and error approach are used.

This paper has presented an algorithm for real-time prediction of bus arrival time using machine learning methods provided with GPS data, in an intelligent prediction system. As comparative studies prove to claim that ANN models are better in predicting bus travel time. Thus, an artificial neural network model is developed to predict bus travel information on the basis of travel time from the last five recent bus service stops and average speed over the trip to predict travel time on headway bus stops. Computed features like distance travelled, demand characteristics, and time of day, average speed, travel time between bus stops were obtained from the TFI and GPS dataset. Although the available data were limited and historic, some assumptions had to be made for which this paper has presented an algorithm to find the closest latitude & longitude pairs of bus service stops and some data points had to be ignored because of errors in it. The initial results presented here appear to be reasonable and promising.

For each of the test performed in this thesis, Using ANN model, it is possible to develop one stop-based model to estimate bus travel time prediction for all downstream bus service stops a given starting bus stop. A sample of the stop-based model with testing for all random origins of type “B” buses is shown in Table 16. Since the methodology used to develop all prediction is the same for all test trips, their MAPE% and SMAPE errors are similar and can be presented in a range. Therefore, it is redundant to discuss each trip individually in detail. This paper has presented that the performance of the bus arrival time prediction algorithm is also expected to change as the speed to the downstream station increases. To quantify this relationship, a null hypothesis was performed.

Based on performance metrics i.e. MAPE% and SMAPE errors and test we have performed; it proves our proposed ANN model can predict bus travel time of any target bus service stop using last five recent bus travel times. Table 17, with no traffic & weather condition factors involved, lowest MAPE% and SMAPE error from ANN model obtained was 2.03% and 0.0202 error value respectively, from the figure 12 of MAPE values and its interpretations (Lewis, 1982), it was found that prediction is highly accurate.

Chapter 7

Future Scope and Recommendations

In this thesis, bus data records that we have used for training the ANN model are for the buses that have stopped on every service stop with or without passengers boarding and alighting. Because there can be cases where due to different hours of the day or due to service stops that have no demand for passengers boarding and alighting, a bus may have not stopped and thus increasing a probability to reach the next bus stop earlier than the previous bus travelling over the same route. Artificial neural network embedded with such capabilities to use data records of such trips which creates a dummy value from the historic data for such bus service stops can be implemented to predict bus travel times.

Using real-time information such as current speed from the GPS server will have an advantage in predicting bus travel time at any bus service stop. Because having real-time information about bus arrival time, passengers can plan their travel, utilize their time in other works and might not have to wait for a longer time at a bus service stop. Along with using current speed as an input to the model, traffic or weather condition factor inputs will be an enhancement to the presented ANN model in this thesis.

The system can be made capable of tracking a large number of buses simultaneously, detecting their service routes and directions automatically, and predicting their arrival time to downstream stations with acceptable accuracy.

References

Amita, J., Singh, J. S. and Kumar, G. P. (2015) 'Prediction of Bus Travel Time Using Artificial Neural Network', International Journal for Traffic & Transport Engineering, 5(4), pp. 410–424. doi: 10.7708/ijtte.2015.5(4).06.

Cats, O. and Loutus, G. 2015. Real-time bus arrival information system – an empirical evaluation. Journal of Intelligent Transportation Systems.

Chamberlain, R.G. Great Circle Distance between Two Points, website link accessible: https://www.movable-type.co.uk/scripts/gis-faq-5.1.html

Chien, S.I J., Ding, Y., and Wei, C. 2002. Dynamic bus arrival time prediction with artificial neural networks. ASCE Journal of Transportation Engineering, 128(5):429–438.

Dailey, D.J. An Algorithm for Predicting the Arrival Time of Mass Transit Vehicles 27 using AVL data. In Proceedings of the 78th Annual Meeting, (CD-ROM). Transportation 28 Research Board of the National Academics, Washington, D.C.

Fan, W. and Gurmu, Z. 2015. Dynamic travel time prediction models for buses using GPS data. International Journal of Transportation Science and Technology. 4(4), 353-366.

General Transit Feed Specification (GTFS) website, link accessible from: https://developers.google.com/transit/gtfs/reference/

Gentili, M. and Mirchandani, P. B. 2018. Review of optimal sensor location models for travel time estimation. Transportation Research Part C: Emerging Technologies, 90:74–96.

Hagan, M., Demuth, H., and Beale, M. 1996. Neural network design, PWS, Boston, 1996.

Hecht-Nielsen, R. 1987. Kolmogorov's Mapping Neural Network Existence Theorem. In Proceedings of the First IEEE International Conference on Neural Networks, San Diego, 4-11.

Jeong, R. and Rilett, R. 2004. Bus arrival time prediction using artificial neural network model. in Intelligent Transportation Systems. The 7th International IEEE Conference. pp. 988-993

Jianmei, L. Dongmei, C., FengXi, L., Qingwen, H., Siru, C., Lingqiu, Z., and Min, C. 2017. "A Bus Arrival Time Prediction Method Based on GPS Position and Real-Time Traffic

Flow," *IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress,* Orlando, FL, pp. 178-184. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.42.

Kalman, R. 1960 A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering, 82 (Series D), 35–45.

Kumar V., B. A., Vanajakshi, L., Subramanian, S.C. 2013. Comparison of Model Based and Machine Learning Approaches for Bus Arrival Time Prediction.

Lewis, C.D. (1982). Industrial and business forecasting methods. London: Butterworths

Mazloumi, E., Rose, G., Currie, G., and Sarvi, M. 2011. An Integrated Framework to Predict Bus Travel Time and Its Variability Using Traffic Flow Data. Journal of Intelligent Transportation Systems, vol. 15, pp. 75-90.

Mean absolute percentage error (MAPE) Formulation accessed from Wikipedia website: https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

Moovit website, link accessible from: https://moovitapp.com/index/en/public_transit-line-208-Ireland-502-851897-228784-0

Patnaik, J., Chien, S., and Bladihas, A. 2004. Estimation of bus arrival times using APC data. Journal of Public Transportation, 7(1), 1–20.

Shalaby, A. and Farhan, A. 2003. Bus travel time prediction model for dynamic operations control and passenger information systems. CD-ROM, the 82nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Sivanandam, S.N.; Sumathi, S.; Deepa, S.N. 2010. Introduction to Neural Networks using Matlab 6.0, Tata McGraw Hill, New Delhi.

Symmetric mean absolute percentage error (SMAPE) formulation accessed from Wikipedia website: https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error

Transport for Ireland website, link accessible from: https://transportforireland.ie

Vanajakshi, L., Subramanian, S.C., Sivanandan. R. 2009. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. IET Intelligent Transport System, 3, 1-9.

Wang, J., X. Chen, and S. Guo. 2009. Bus travel time prediction model with v - support vector regression. In IEEE Intelligent Transportation Systems 3(1), Washington, D.C. pp. 655-660.

Watkins, K. E., Ferris, B., Borning, A, Rutherford, G. S. and Layton, D. 2011. Where is my bus? impact of mobile real-time information on the perceived and actual wait time of transit riders. Transportation Research Part A: Policy and Practice, 45(8):839– 848.

Wu, C.H., Su, D. C., Chang, J., Wei, C. C., Ho, J. M., Lin, K. J., and Lee, D. 2003. An advanced traveller information system with emerging network technologies. In Proceedings of 6thAsia-Pacific Conference Intelligent Transportation Systems Forum, pages 230–231, Taipei, Chinese-Taipei.

Zhang, G. P., 2000. "Neural Networks for Classification: A Survey" IEEE Transactions and reviews. 30(4): 451-462.

Zhang, G. and Patuwo, B., and M. Hu, 1998. Forecasting with artificial neural networks: the state of the art. International Journal of Forecasting, 14 (1), 35–62.

Zhang, Z, Wang, Y., Chen, P., He, Z., and Yu, G. 2017. Probe data-driven travel time forecasting for urban expressways by matching similar spatiotemporal traffic patterns. Transportation Research Part C: Emerging Technologies, 85:476–493.