Spatial-Temporal Analysis of Urban Mobility using Taxi Dataset

Pratyush Kumar (pratyushkumar2901@gmail.com)
Indian Institute of Technology Bombay

Varun Singh
Motilal Nehru National Institute of Technology Allahabad

Research Article

Keywords: Uber Movement, Ride Sharing, Time Series Forecasting, New Delhi

Posted Date: November 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3630229/v1

Abstract

Ride-sharing services linked to mobile devices are innovative, on-demand transport services that aim to reduce the ill effects of private vehicles, such as pollution, congestion, etc. Ride-sharing is one of the emerging technologies which contains large data, and users with the same origin-destination and travel time are matched and share the ride. With the release of the Uber Movement Dataset for certain large cities, the spatial-temporal analysis of urban mobility using a taxi dataset is now possible. The present study used aggregated Uber trip data from 2016 to 2019 for New Delhi. This paper explores the applications of python -based techniques such as big data analytics, machine learning, and probabilistic programming for predicting travel time by utilizing the Uber Movement Dataset of New Delhi by taking several origins and destinations. Time Series Forecasting has been carried out with the help of ARIMA, Holt-Winters, Facebook Prophet, and the global model, which shows the difference between actual and predicted travel time. Spatial analysis for different wards is conducted to understand the variation in many trips. The results of this study will be helpful in urban planning and a better understanding of human mobility in New Delhi.

1. Introduction

Delhi is expanding at a rate that has never been seen before. It is the capital of the largest democracy in the world, a city with traditional values and aspirations, and a significant national hub. The growth must seamlessly incorporate historical and modern development to transform it into a worldwide metropolis and a world-class city. More outstanding planning and effective governance and administration are needed to take the necessary measures to accomplish the stated goals. Transportation connectivity is a constant concern in major cities like Delhi and Mumbai. This calls for enhanced mobility and good accessibility planning. Having accessibility alone does not guarantee mobility. The concept of mobility encompasses much more than just transportation. It means having a system of coordinated, integrated, and high-quality modes of transportation rather than simply having access to one mode. Any public transportation model implemented in the city is evaluated for efficacy and efficiency using the system's mobility.

Many transportation networks companies, such as Uber, Ola, and others, are emerging to supplement the traditional public transportation system, such as buses, trains, and metro services, by providing peer-to-peer services to the people. Ride-hailing services are growing globally because they are dependable, accessible, and affordable for every customer[1], [2]. Ride-sharing services have drastically altered the geographical layout of urban mobility landscapes over the past ten years, particularly in metropolitan areas[3]. On-demand ride-sharing services have replaced personal assets as personal transportation options[4]. This has resulted in many cases, fuel consumption savings, and per-passenger travel costs, lowering traffic congestion and emission levels[2], [5], [6].

With the release of ride-sharing data from one of the world's most popular services, Uber Movement, for a limited number of metropolitan areas, it is now possible to investigate the impact of ride-sharing services

on urban mobility at finer spatial and temporal scales[1], [3]. In India, Uber services have become more popular than conventional taxi services. It is the sixth-largest Uber market globally and operates in 58 cities in the country. In the Delhi-NCR region, UberGo is the most popular product among riders, while in cities like Bangalore and Hyderabad, UberPool has become the most prevalent choice for riders. Uber provides anonymized data through the 'Uber Movement' platform, open to everyone, from students to researchers. These services encourage private drivers to share their cars with passengers traveling to similar destinations.

Therefore, in this paper, an exploration of the daily urban mobility in the New Delhi area from Quarter 1 of 2016 to Quarter 3 of 2019, using aggregate trip count and mean travel times, has been carried out.

2. Data and Method

In this section, we see the steps of data collection and how the data is processed so that it can be analyzed.

2.1 Data Acquisition

Traffic data for this research is collected from the 'Uber Movement' platform provided by Uber [1]. This open-source website provides data for five major cities in India, including New Delhi. Data is provided zone-wise, and the zones are divided into different wards in the Uber Movement. Time of the day is segmented into a five-time zone, namely AM Peak (7 AM-10 AM), Midday (10 AM-4 PM), PM Peak (4 PM-7 PM), Evening (7 PM-12 AM), and Early Morning (12 AM-7 AM). Data includes arithmetic mean, geometric mean, and standard deviation for aggregated travel times over a selected date range and for each pair of zones. This data is open to the public and can be directly downloaded in .CSV (comma-separated value)[7] from the website for the selected origin-destination pair.

The following table shows the monthly mean travel time for a particular OD pair for each time zone of the day for the specific date range.

Table 1
Monthly mean travel time for each time zone for selected OD pair

Time of the day	Origin Display Name	Destination Display Name	Date-range	Mean Travel Time (seconds)
Daily	IGI Airport	New Delhi Railway Station	01/03/20–31/03/20	1686
AM Peak	IGI Airport	New Delhi Railway Station	01/03/20–31/03/20	1726
Midday	IGI Airport	New Delhi Railway Station	01/03/20–31/03/20	1857
PM Peak	IGI Airport	New Delhi Railway Station	01/03/20–31/03/20	1897
Evening	IGI Airport	New Delhi Railway Station	01/03/20–31/03/20	1669
Early Morning	IGI Airport	New Delhi Railway Station	01/03/20–31/03/20	1279

3 Origins (tabulated below) and around 290 Destinations have been selected for analysis. Hence for each of the OD pairs, complete 4-year (i.e., January 2016 to March 2020) data has been downloaded from the Uber movement platform provided by Uber.

Table 2
Origins Points

S.No.	Points of Interest	Latitude	Longitude
1	Connaught Place	28.63279	77.21965
2	Indira Gandhi International Airport- T3 (IGI)	28.55616	77.09995
3	ISBT Kashmiri Gate	28.66937	77.23085

The color-coded user interface of Uber Movement is given below.

2.2 Data preparation and pre-processing

Data obtained through Uber Movement cannot be directly used. The raw data should be transformed into a format that computers can understand and analyze. Data cleaning is adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. It is the most critical pre-processing step because it will ensure data is ready for downstream needs.

Data obtained must be processed to further use. Data pre-processing using machine learning consists of various steps to achieve the required output data file, which will be used as input while processing the prediction model for travel time [8].

First, merge/concatenate all datasets into one. Here, each daily dataset is not necessarily ordered by Date because of how they are stored in the “Downloads” folder when initially downloaded with the dataset automation bot script. We also save the initial df into a CSV file, a concatenated “raw” Data frame (df). Here, the current city name is just used as a placeholder. By specifying the correct file path, the datasets of the desired city can be used. There are multiple Destination Movement IDs for one single Origin Movement ID and Date (which for now is called “Date Range”).

Table 3
Raw Data

S.I. No.	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Date Range	Mean Travel Time (Seconds)
258	132	ISBT Kashmiri Gate, Inter State Bus Terminal, …	277	700 Outer Ring Road, Bhalswa Jahangir Village, …	1/31/2020-1/31/2020, Every day, Daily Average	1279
259	132	ISBT Kashmiri Gate, Inter State Bus Terminal, …	278	Nangloi – Najafgarh Road, Brahanpuri Colon…	1/31/2020-1/31/2020, Every day, Daily Average	3778
260	132	ISBT Kashmiri Gate, Inter State Bus Terminal, …	279	Nightingale Road, Budella, Vikaspuri, Delhi	1/31/2020-1/31/2020, Every day, Daily Average	3565
261	132	ISBT Kashmiri Gate, Inter State Bus Terminal, …	280	Hastsal Road, Block A2, JJ Colony, Uttam Nagar …	1/31/2020-1/31/2020, Every day, Daily Average	3309
262	132	ISBT Kashmiri Gate, Inter State Bus Terminal, …	281	D915, New Mahavir Nagar, Tilak Nagar, Delhi	1/31/2020-1/31/2020, Every day, Daily Average	3338

This is because the travel times data is extracted from the center-most point/zone of each city to every other point/zone available at a given date. The number of available points/zones varies from Date to Date. Therefore, there is a different number of rows for each Origin Movement ID or Date. Later it will be turned into a multi-index df, indexed by Date.

Now, the necessary operations are performed to format the df, simplifying columns names, dropping unnecessary columns, creating potentially relevant columns and most importantly, indexing and sorting the df by Date.

Table 4
Processed Data

Date	NumericIndex	Origin Movement ID	Destination Movement ID	Mean Travel Time (Seconds)	Range Lower Bound Travel Time (Seconds)
2020-01-01	538	9	273	2806	2125
2020-01-01	539	9	274	2426	1603
2020-01-01	540	9	275	1489	875
2020-01-01	541	9	277	2732	1945

Lastly, formatted df is saved into a new CSV. The processed data shown above is the data that is going to be analyzed after cleaning and formatting.

2.3 Prediction model

This is the final step in analyzing the dataset. Using facts from the past and present, forecasting is the process of making predictions about the future. A type of forecasting called time-series forecasting makes predictions about future events using time-stamped data points. Time-series forecasting models come in a wide variety, each with unique advantages and disadvantages. A time series is a collection of observations made over an extended period[9]– [11]. A multivariate time series consists of the values obtained by several variables at the same periodic time. In contrast, a univariate time series comprises the values obtained by a single variable at periodic time instances. The daily variation in temperature over a day, a week, a month, or an entire year is the most basic example of a time series we encounter. These data were analyzed using ACF and PACF plots for identifying the presence of seasonality, trend, and other patterns in time series data.

There are 4 components of time series, i.e., trend, seasonality, cyclic and residual. Before forecasting, these components are removed from the data as it will give errors in forecasted values.

There are 4 models used in this paper.

a. ARIMA
b. FBProphet
c. Holt-Winter’s Seasonal Method
d. Global Model

The above four models are chosen based on their established popularity and effectiveness in handling different types of time series data. They also handle the seasonality and trend effectively.

3. Results

According to Uber Movement data, Uber services have played a significant role in urban mobility, with millions of rides within and outside the metropolitan area since its launch in 2014[1]. This research includes the four-year data, which has been divided into four quarters to be analyzed separately.

3.1 Temporal Analysis

In this section, variation between actual and predicted travel time is carried out.

The above figure compares the forecasted Mean Travel Time (fcst) and the Average Mean Travel Time (AVGMTT) for all 4 models. Similarly, it has been done for 2017 and 2018.

Table 5
Comparison of error between different models

Error	Prophet	Holt-winters	Arima	Global
SMAPE	0.078759	0.056216	0.064969	0.052069

$$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

1

Where A_t is the actual value and F_t is the forecast value.

The symmetric mean absolute percentage error (SMAPE) for different models is calculated with the help of Eq. 1. SMAPE has the advantage of being symmetric, meaning it treats overestimation and underestimation errors equally and handling the zero values better. The above table shows the Symmetrical Mean Absolute Percentage Error (SMAPE) between the prophet, Holt-winters, Arima, and global model. These values show that the error in forecasting for the global model is minimum, indicating that the global model is best among these models for time series forecasting.

3.2 Spatial Analysis

It is well known that spatial information improves prediction and accuracy, especially in congested traffic and over longer time horizons. Okutani and Stephanedes pioneered the concept of capturing spatial information in time series studies of transportation-related problems[12]. In this section, spatial patterns of New Delhi are carried out depending on how the travel time has been changing. This research uses data from the uber movement in New Delhi. From Uber Movement, data for 290 wards of New Delhi was collected. The data consists of Ward Number, Ward Type, Movement ID, and Display Name, and the geometry includes the type, a polygon for all wards, and the coordinates. Data was represented quarter-

wise, i.e., for 3 months from 2016 to 2020, and the different colors show the spatial variation of travel time in the range of 500–3500 seconds.

The above figures show the spatial variation of New Delhi from 2019– 2020. Similarly, it has been done for the years 2016,2017, and 2018. Uber has not provided data from April 2020. (Fourth quarter data of 2019 has also not been provided by Uber.

3.3 Ranging of Movement IDs

The below figure shows all the movement IDs, which are in the range of 5,10,15,20 km from the selected origins of this study which include,

a. Connaught Place,
b. Indira Gandhi International Airport- T3 (IGI),
c. Kashmiri Gate ISBT

4. Discussion

In this research, uber movement data was collected from the year 2016 to 2020. The data shows a variation in all these four years, where the time taken from New Delhi to Saket in 2016 was 35 min 35 sec and 32 min 29 sec in 2020. Based on these data, four models, ARIMA, FBProphet, Holt-Winters, and Global, have been used for time series forecasting[9]– [11]. Seasonality and irregularity types have been removed from the data trend, and the forecasting is done. The global model has the best forecasting result out of the used models, as the Recurrent Neural Network (RNN) is used in this model [13]. RNN is useful for dealing with the time series and sequences more generally and does not develop separate models for each series.

The data can also be collected from Google API and there can be comparison between both the data, but that comparison is not considered in this study. Furthermore, Uber collects data only when the vehicle is carrying passengers, and Uber drivers are incentivized to complete the trip in the shortest time possible so these points can be considered while comparison.

Also, it is well established that spatial information improves prediction accuracy, particularly in congested traffic and longer time horizons. The concept of capturing spatial information in time series studies of transportation-related problems was first introduced by Okutani and Stephanedes[12]. Later, Kamriankis and Prastacos [14], [15] applied the spatial concept to predict relative velocity on significant roads in Athens, Greece. The technique is known as space-time autoregressive integrated moving average (STARIMA). The model differs from traditional ARIMA models because it uses spatial information from neighboring links to forecast traffic[16].

5. Conclusion

In developing countries like India, the population is growing at an unprecedented rate. In such a scenario, offering a hassle-free transportation system in urban areas is crucial to address all transportation-related issues, including those linked to the economy, ecology, and health[6], [17]. The development of transportation infrastructure and better management of it is required to accomplish these goals. Congestion, or prolonged travel times, is India's main transportation issue, especially in major cities like Delhi. Numerous factors could be to blame, including a large population, poor infrastructure, an inefficient management system, improper land use or future planning, a lack of connection, and an underdeveloped public transportation system. One must thoroughly analyze the current situation and offer appropriate suggestions to enhance the transportation system. This thesis aims to analyze the spatial-temporal urban mobility using the appropriate prediction model by predicting trip time between distinct origin-destination combinations. The present study provided census tract-level data for aggregate Uber trips through Uber Movement for New Delhi during 2016–2019. The data were available at the quarterly level. The essential conclusions drawn from the research can be summarized as follows-

1. The study results show that the global model for time series forecasting is considered the best among various models for analyzing spatial-temporal urban mobility.
2. According to the research findings, the built prediction model for the city's current traffic conditions can anticipate journey time with more than 90% accuracy.
3. Based on the training dataset, the model can automatically identify necessary parameters like learning rate, number of iterations, etc.
4. These models can assess the significance of each independent variable's contribution to the predicted target value. This is the most significant benefit in this case because the independent variables do not need separately developed empirical relations.
5. To improve the accuracy of the results, outliers can be found and eliminated using the created model.

Declarations

Acknowledgement:

The authors acknowledge the opportunity provided by the 7th Conference of the Transportation Research Group of India (CTRG-2023) to present the work that formed the basis of this manuscript.

Conflict of Interest:

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. S. sen Roy, J. Perlman, and R. C. Balling, “Analysis of urban mobility in South Florida using Uber Movement,” Case Stud Transp Policy, vol. 8, no. 4, pp. 1393–1400, (2020).

2. A. Amey, J. Attanucci, and R. Mishalani, “Real-time ridesharing: Opportunities and challenges in using mobile phone technology to improve rideshare services,” Transp Res Rec, no. 2217, pp. 103–110, (2011).
3. J. Cramer and A. B. Krueger, “Disruptive change in the taxi business: The case of uber,” American Economic Review, vol. 106, no. 5, pp. 177–182, (2016).
4. G. Dudley, D. Banister, and T. Schwanen, “The Rise of Uber and Regulating the Disruptive Innovator,” Political Quarterly, vol. 88, no. 3, pp. 492–499, (2017).
5. D. Fagnant, K. M. Kockelman Professor, and W. J. Murray Jr Fellow, “The Travel and Environmental Implications of shared Autonomous Vehicles, using agent-based model scenarios,” Transportation Research Part C, vol. 40, pp. 1–13, (2014).
6. S. H. Jacobson and D. M. King, “Fuel saving and ridesharing in the US: Motivations, limitations, and opportunities,” Transp Res D Transp Environ, vol. 14, no. 1, pp. 14–21, (2009).
7. B. Deb, S. R. Khan, K. Hasan, A. H. Khan, and Md. A. Khan, “2019 IEEE 5th International Conference for Convergence in Technology (I2CT).,” *Int J Forecast*, pp. 1–8, (2019).
8. G. Dudley, D. Banister, and T. Schwanen, “The Rise of Uber and Regulating the Disruptive Innovator,” Political Quarterly, vol. 88, no. 3, pp. 492–499, (2017).
9. S. L. Ho ’ ~ and M. Xie, “THE USE OF ARIMA MODELS FOR RELIABILITY FORECASTING AND ANALYSIS,” Computers ind. Engng, vol. 35, no. 2, pp. 213–216, (1998).
10. P. S. Kalekar, “Time series Forecasting using Holt-Winters Exponential Smoothing,” vol. 4329008, no. 13, pp. 1–13, (2004).
11. B. Kumar Jha and S. Pande, “Time Series Forecasting Model for Supermarket Sales using FB-Prophet,” *Proceedings – 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, pp. 547–554, (2021).
12. Y. J. Stephanedes, “DYNAMIC PREDICTION OF TRAFFIC VOLUME THROUGH KALMAN FILTERING THEORY,” vol. 18, no. 1, pp. 1–11, (1984).
13. S. Smyl, “A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting,” Int J Forecast, vol. 36, no. 1, pp. 75–85, (2020)
14. Y. Kamarianakis and P. Prastacos, “Forecasting Traffic Flow Conditions in an Urban Network Comparison of Multivariate and Univariate Approaches,” Transp Res Rec, (2003).
15. Y. Kamarianakis, “SPATIAL-TIME SERIES MODELING: A REVIEW OF THE PROPOSED METHODOLOGIES,” (2003).
16. S. S. Faghih, A. Safikhani, B. Moghimi, and C. Kamga, “Predicting Short-Term Uber Demand Using Spatio-Temporal Modeling: A New York City Case Study” (2018).
17. B. Yu *et al.*, “Environmental benefits from ridesharing: A case of Beijing,” Appl Energy, vol. 191, pp. 141–152, (2017).

Figures

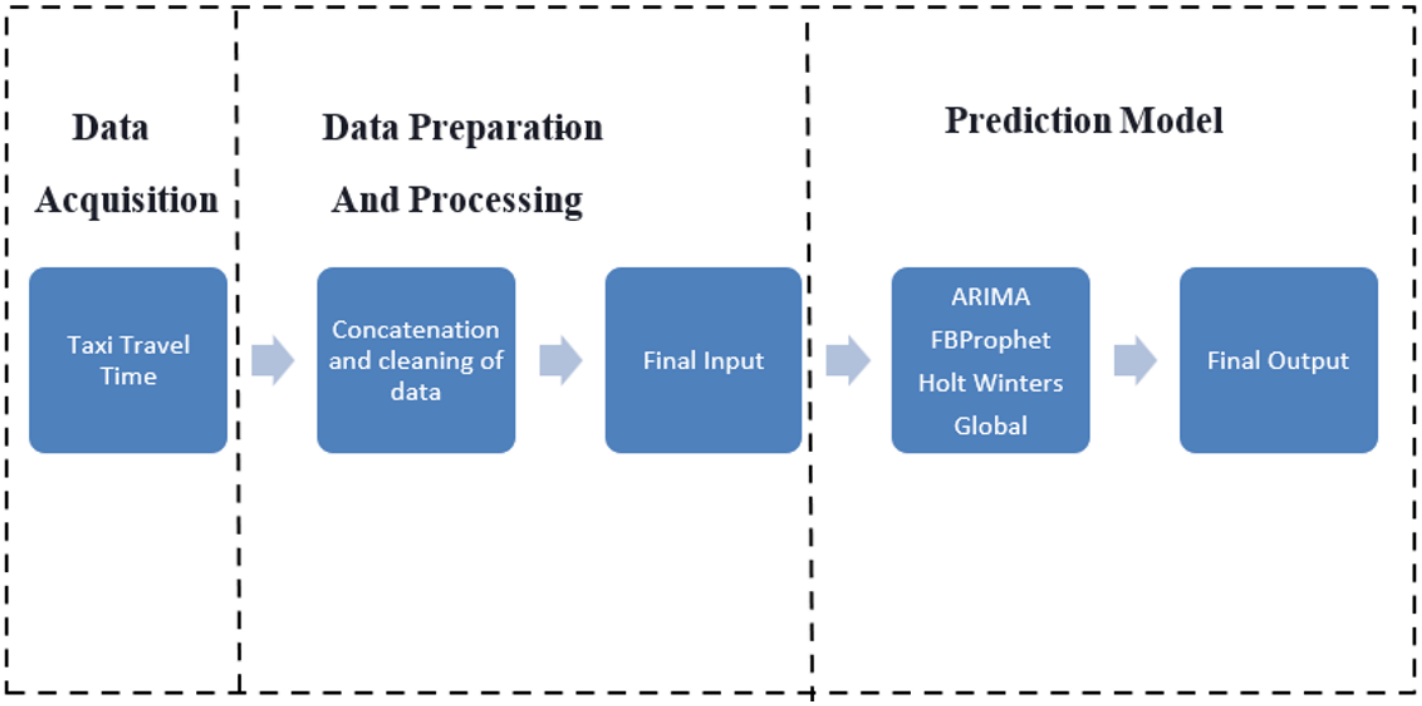


Figure 1

Schematic diagram of a methodology

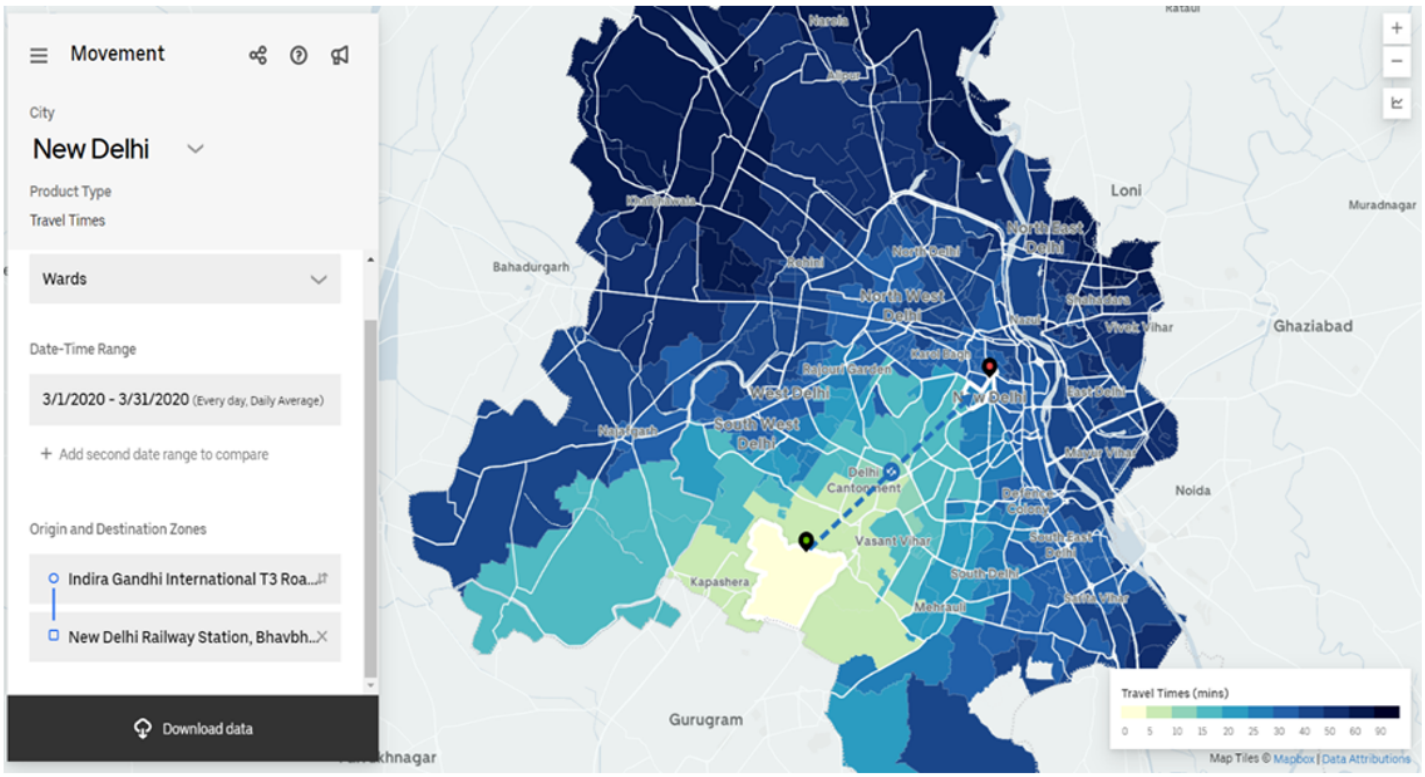


Figure 2

Uber Movement color coded user interface

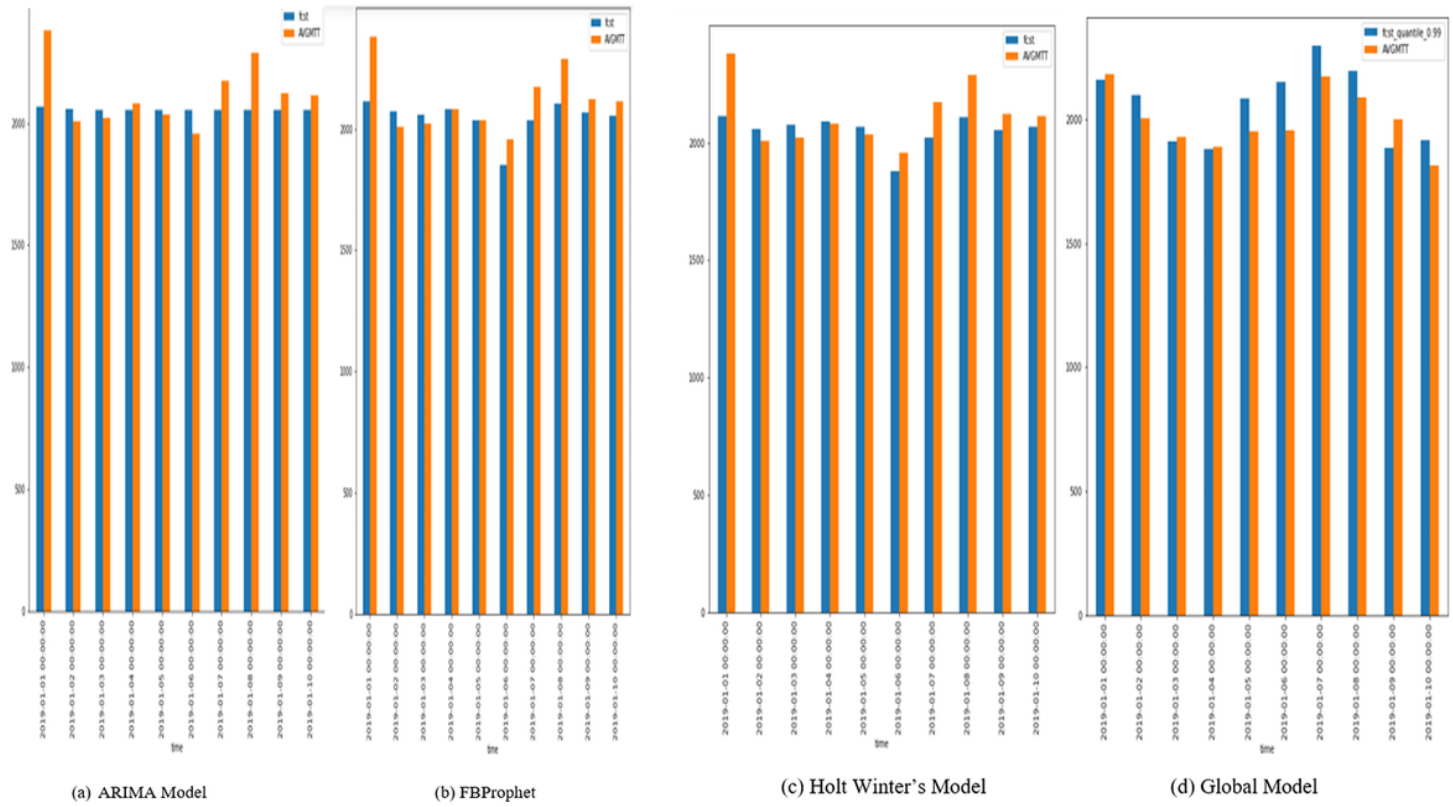


Figure 3

Forecasting for the year 2019

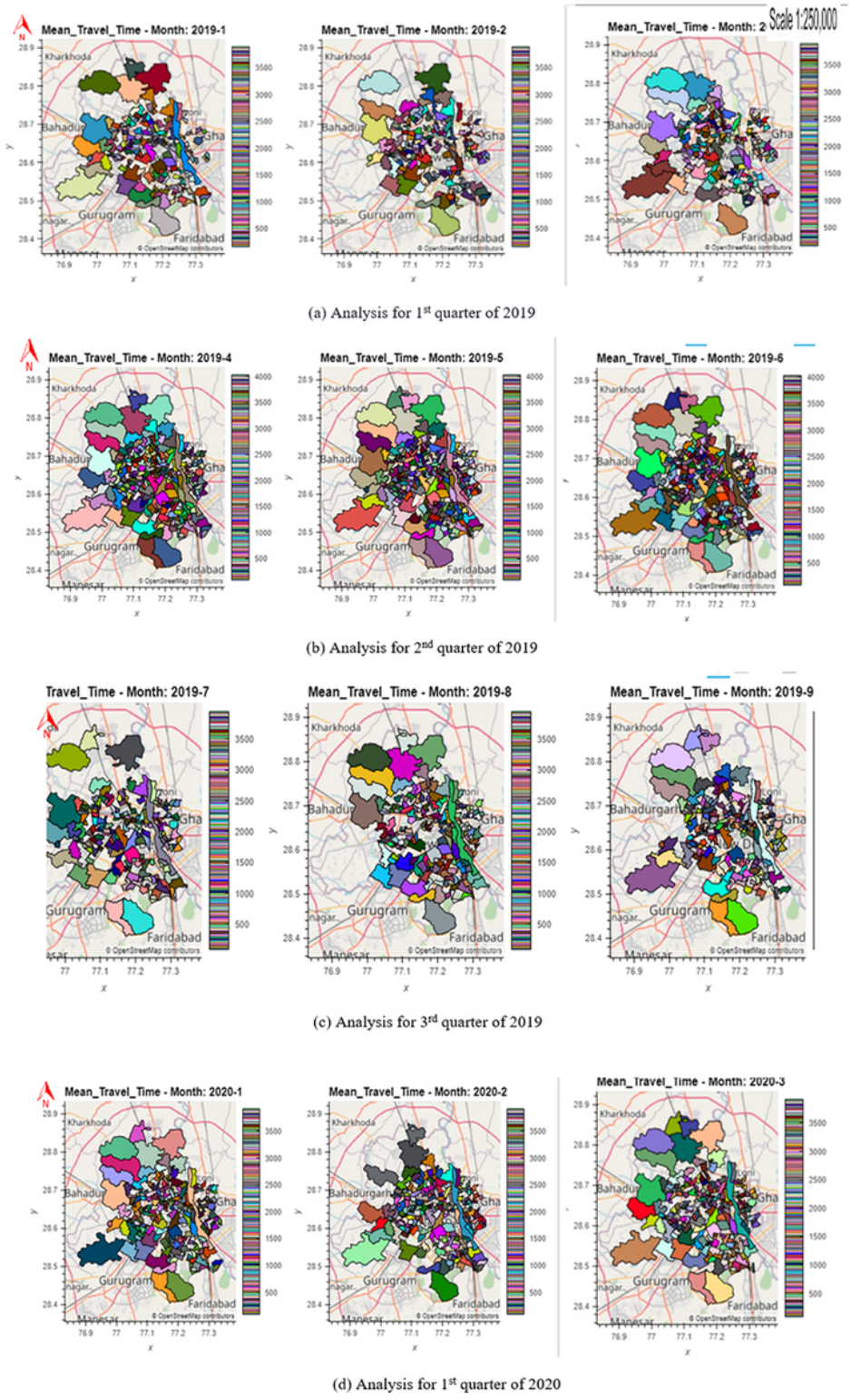
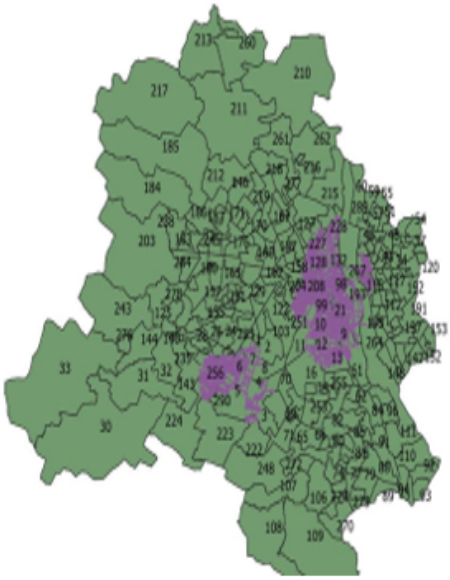
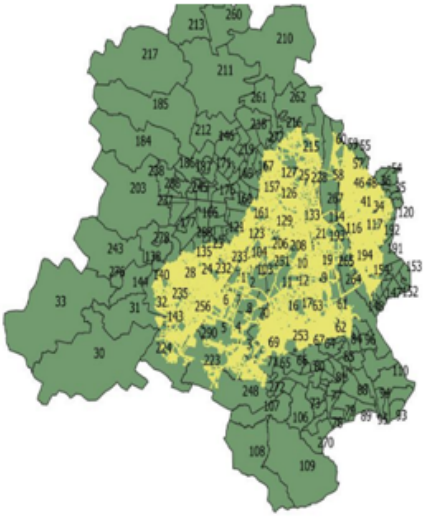


Figure 4

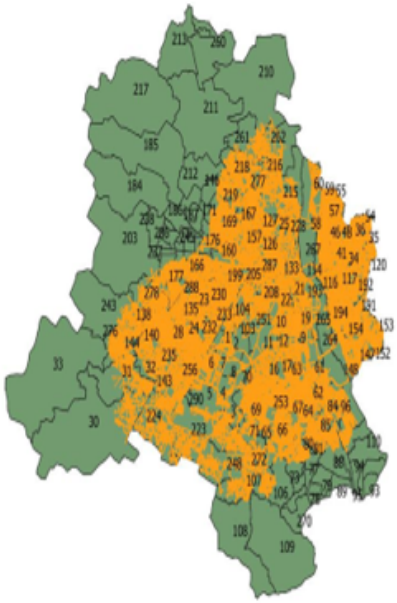
Spatial analysis of the year 2019-20



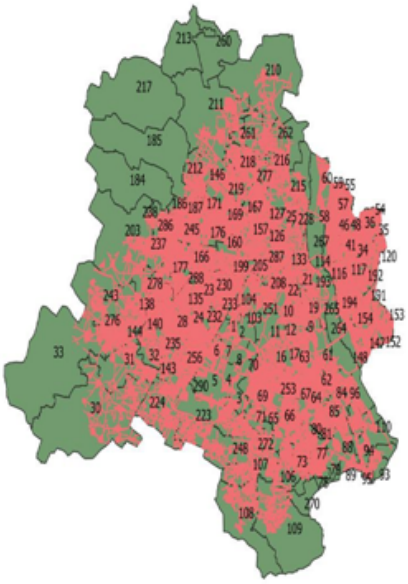
(a) Movement IDs in the range of 5 km



(b) Movement IDs in the range of 10 km



(c) Movement IDs in the range of 15 km



(d) Movement IDs in the range of 20 km

Figure 5

Various Range of Movement IDs