sustainability

Article

Effects of Data Characteristics on Bus Travel Time Prediction: A Systematic Study

Hima Shaji [1], Lelitha Vanajakshi [2,*] and Arun Tangirala [3]

[1] Department of Civil Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[2] Department of Civil Engineering/Robert Bosch Centre for Data Science and Artificial Intelligence, Indian Institute of Technology Madras, Chennai 600036, India
[3] Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600036, India
* Correspondence: lelitha@iitm.ac.in; Tel.: +91-44-2257-4291

Abstract: The prediction of bus travel time with accuracy is a significant step toward improving the quality of public transportation. Drawing meaningful inferences from the data and using these to aid in prediction tasks is always an area of interest. Earlier studies predicted bus travel times by identifying significant regressors, which were identified based on chronological factors. However, travel time patterns may vary depending on time and location. A related question is whether the prediction accuracy can be improved with the choice of input variables. The present study analyzes this question systematically by presenting the input data in different ways to the prediction algorithm. The prediction accuracy increased when the dataset was grouped, and separate models were trained on them, the highest accurate case being the one where the data-derived clusters were considered. This demonstrates that understanding patterns and groups within the dataset helps in improving prediction accuracy.

Keywords: travel time data analysis; bus travel time; clustering; prediction; machine learning techniques

Citation: Shaji, H.; Vanajakshi, L.; Tangirala, A. Effects of Data Characteristics on Bus Travel Time Prediction: A Systematic Study. *Sustainability* **2023**, *15*, 4731. https://doi.org/10.3390/su15064731

Academic Editors: Mario Marinelli and Aleksandra Colovic

Received: 12 January 2023
Revised: 23 February 2023
Accepted: 3 March 2023
Published: 7 March 2023

1. Introduction and Background

The development of an efficient public transportation system is an integral part of developing a smart city. An efficient public transportation system is a possible solution to the ever-increasing urban mobility and increased air pollution in cities. It is expected that if more people shift from private vehicles to high occupancy public transportation modes, it will lead to better air quality and lesser congestion on roads, finally improving the quality of travel. Hence, a smart and sustainable public transportation system is essential for developing a smart city.

Buses are the most popular public transport in many countries, including India. The lack of reliability and the higher waiting times are some major issues concerning the buses. This causes discomfort to the passengers, thereby causing them to switch to private vehicles. This can be addressed by providing real-time bus arrival information to users, which necessitates precise bus travel time prediction. The trajectory data generated from Global Positioning System (GPS) devices fitted on transit buses serve as a rich historic database. They can be used to understand patterns in the data and make more inferences about the traffic system. Important traffic variables such as travel time, vehicle speed, delay at intersections, and signals can be extracted from the GPS data. Out of these, the most preferred variable of interest is travel time, as both users and planners understand it easily.

The studies that have been published on the prediction of travel time using GPS bus data can be broadly grouped into physics-based and data-based approaches [1–7]. Data-based approaches are widely being used as a huge amount of data is available. The most popular data-driven prediction techniques include time series [8–10] and machine learning [11–14]. These methods study the patterns in the data and assume that these patterns can be used for better forecasts. However, whether a large amount of data can be used to draw meaningful inferences from the data and aid in prediction tasks is a question that needs to be studied. Another question is whether the accuracy of these prediction methods increases with the "correct" input variables being used. The current study concentrates on these questions.

Travel time, in general, follows specific patterns such as trip-wise, daily, weekly, monthly, and yearly, and it is assumed that they are recurrent [15]. Lee et al. [16] used historical bus trajectories similar to the current trip trajectory to predict bus travel times. Kumar et al. [17] analyzed trip-wise, daily, and weekly patterns of bus travel times and developed a bus arrival prediction model. Kumar et al. [18] studied the travel time patterns for different days of the week using GPS data obtained from public transit buses. Another study by Vlahogianni et al. [19] showed pattern-based prediction to be more accurate than the classical time-series approach for short-term traffic prediction. These studies aimed at identifying proper regressors for prediction by identifying patterns within the data. However, in these studies, grouping the regressor data for the prediction was done manually using chronological factors. Travel time on any day depends on the time of the day and the characteristics of the stretch under consideration. For example, for the same section of the road, the travel times may be different on different days of the week. The travel time for a section may also be different for the same day across different weeks. Hence, assuming static patterns based on chronological factors may not be an efficient method to identify the patterns in the case of highly varying traffic variables such as travel time. This time-varying nature of traffic necessitates using automated techniques to group travel time data and capture the varying patterns in the data rather than separating them manually.

The present study analyzes the effects of data characteristics on travel time prediction systematically by presenting the input data in different ways to the prediction algorithm, such as presenting the data without grouping, dividing the dataset manually into fixed clusters based on chronological factors, and using clustering algorithms to form data-derived clusters. Clustering algorithms, based on unsupervised learning, learn from the data and help in identifying groups in the data automatically with minimum human intervention. In clustering, clusters or groups are created that are similar to the points in the same group and dissimilar from those in the other groups. Chung [20] used a clustering algorithm to group the historical travel time data for prediction and reported a significant reduction of computation time once the data was clustered. After grouping, a similar segment of the historical database only needs to be searched to identify the patterns. Van Der Voort et al. [21] predicted traffic flow on a motorway by dividing the dataset manually into weekdays and weekends and clustering using Kohonen maps. It was shown that the clustering-based grouping worked better than when the dataset was divided manually. Park et al. [22] used fuzzy *c*-means clustering and Kohonen SOM to cluster historical link travel times and calibrated individual artificial neural network (ANN) models for each class. Li et al. [23] proposed a hierarchical prediction model to predict the number of bikes on rent in a bike-sharing system in New York and Washington, D.C. The bike-sharing stations were divided into two groups based on geographical locations and transition patterns, and a gradient-boosting regression tree (GBRT) was used to predict the number of bikes on rent. Clustering helped reduce the irregular fluctuation issue at each station, and prediction errors were reduced by 0.23, especially for anomalous hours. Based on these reported results, it was decided to use clustering algorithms in the present study for grouping the historical data. Important clustering algorithms used in traffic-related applications include the *k*-means clustering algorithm [24–28], hierarchical clustering algorithm [23], and Kohonen SOM [21,22]. A prior study [29] compared the performance of the clustering algorithms, *k*-means, hierarchical, and Kohonen SOM in predicting bus travel time trends. The results showed that the *k*-means clustering performed better than the other clustering algorithms.

None of the above studies paid attention to how the group characteristics can be used for a more accurate prediction of travel time. The present study focuses on this by studying the effect of data characteristics on the prediction of bus travel time. The paper mainly examines whether using such grouping information can be used to improve the final objective of improving the travel time predictions. A comparison of performance between the case when a single model is trained on the complete dataset and multiple models are trained, each on the identified groups will be undertaken in this study. The primary tasks in the study include:

1. To perform exploratory analysis of the data and identify patterns in data across space and time;
2. To develop suitable models for the prediction of bus travel time with and without grouping the data;
3. To compare the accuracy of prediction across different groupings considered and understand the effect of incorporating data characteristics in bus travel time prediction.

2. Data Collection

The present study used data collected using GPS units fitted on Metropolitan Transport Corporation (MTC) buses in Chennai, the capital city of the state of Tamil Nadu, India. Figure 1 shows the northbound 19B bus route, which connects Kelambakkam, a suburb of the city, to Saidapet, a major commercial area of the city, which was chosen as the study stretch. With around 20 bus stops and 14 signalized intersections along the route, it is one of the busiest routes in the city. As the route connects an urban and suburban area, varying traffic features such as traffic volumes and land use characteristics can be expected across the route. During a single trip, a bus might face low traffic volume in the suburban area, leading to low travel times across those road sections and high travel times due to congestion in the city area. Table 1 details all the bus stops across the study stretch. A total of 1231 trips were collected for 45 days. The date, timestamp, latitude, and longitude of the bus's location were all included in the GPS data that was collected. Haversine formula [30] was used to calculate the distance between two consecutive GPS points. The route which spanned a total length of 29.4 kms was divided into 500 m sections for the purpose of analysis.

Figure 1. 19B bus route (source: Open Street Maps).

Table 1. Bus stops on the 19B bus route.

S. No.	Name of The Bus Stop	Distance between Bus Stops (km)	Cumulative Distance from the Origin (km)
1	Kelambakkam	0	0
2	Hindustan Engg. College	2.51	2.51
3	SIPCOT	3.4	5.91
4	Navallur	1.61	7.52
5	Navallur Church	2.5	10.02
6	Semmencheri	1.01	11.03
7	Kumar Nagar	1.28	12.31
8	Sozhinganallur P.O.	1.43	13.74
9	Karapakkam	1.81	15.55
10	TCS	0.41	15.96
11	Mootachavadi	1.46	17.42
12	Mettupakkam	0.79	18.21
13	Thorapakkam	0.6	18.81
14	Tirumailai Nagar	1.25	20.06
15	Kandanchavadi	1.66	21.72
16	Lattice Bridge	1.73	23.45
17	Women's College	1.35	24.8
18	Madhya Kailash	1.02	25.82
19	Engineering College	0.82	26.64
20	Saidapet	3.3	29.94

Data Processing

The data from the GPS units fitted in MTC buses were communicated to a remote server. The reported GPS data includes the date, timestamp, latitude, and longitude of the bus location and is reported every 10 s. From the details of latitudes and longitudes obtained from the GPS, the distance between two consecutive GPS points was calculated using the Haversine formula [30]. The Haversine formula calculates the great-circle distance between two points on a sphere, given their latitudes and longitudes. It assumes the spherical shape of the Earth and gives accurate results for most purposes. It is used extensively for navigation to calculate great-circle distances as it is not computationally expensive.

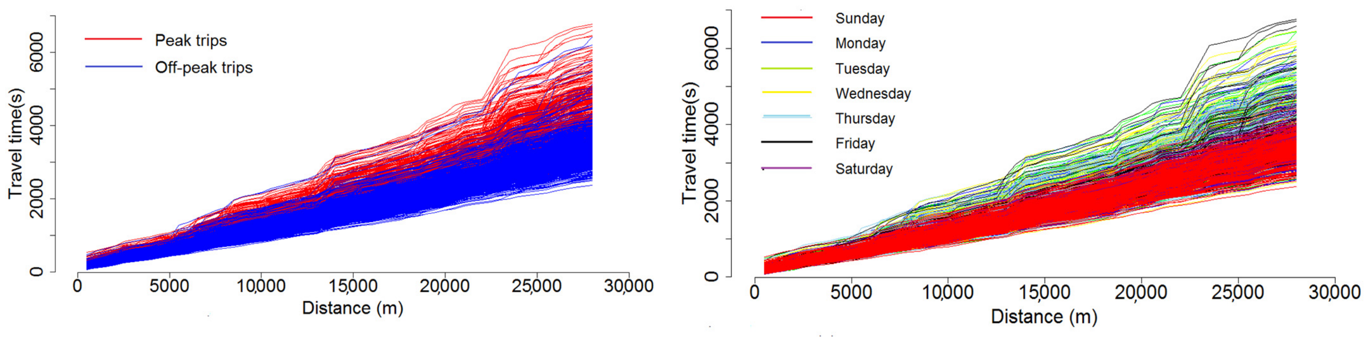
The time required to travel from point 1 to point 2 was calculated using the difference between the corresponding time stamps. Once the distance between two points and the time taken to travel were obtained, the cumulative distance and the total time taken were calculated. For analysis, the route was divided into equal sections, each 500 m, as the bus stops are spaced around 500 m apart in the study stretch. This led to 56 sections across the route. The 500 m section travel times were then calculated using interpolation. The data belonging to Section 1 is not considered as the section is part of the bus depot, and the travel times may be influenced by the layover times at the bus depot. Table 2 shows sample data with timestamps and 500 m section travel times that are used for further analysis.

Table 2. Details of the processed GPS data.

Date	Trip Number	Starting Time (hh:mm)	Section 1	Section 2	Section 3
14-Sep	1	4:58	19.6597	29.4636	15.8478
14-Sep	2	6:13	42.0619	37.3889	14.5982
14-Sep	3	6:54	31.3162	21.2192	3.5410
14-Sep	4	7:13	29.5301	86.2175	24.2950
14-Sep	5	7:15	40.5597	37.5838	26.1327
14-Sep	6	7:23	21.7128	39.0298	14.7819
14-Sep	7	7:56	28.4588	66.8901	26.2100

3. Preliminary Data Analysis

Travel times vary both spatially and temporally across the study area. In the first part of the analysis, a trajectory analysis was carried out to analyze the patterns in peak and off-peak trajectories and daily trajectories. In the previous studies made under similar traffic conditions [18], it was assumed that the peak and off-peak trip trajectories could be identified manually, and such groups of trips were significantly different from each other. Based on that, for the case with fixed clusters, trips between 8:00 am and 10:59 am and between 3:00 pm and 7:59 pm were considered peak trips and others as off-peak trips. Figure 2a shows the time-space plot for peak and off-peak trip trajectories. Though some peak and off-peak trips are distinct, most of them show similar patterns causing overlap of the time-space plot. Figure 2b shows the time-space plot for daily trip trajectories. It can be seen that weekend (Sunday) trajectories have low travel times compared to the other days of the week. However, the trajectories of other days of the week do not indicate any clear patterns, indicating that separating travel times manually into days of the week may not be efficient. These overlaps observed in the trip trajectories raise the question of whether manual grouping can account for these highly varying travel time patterns.

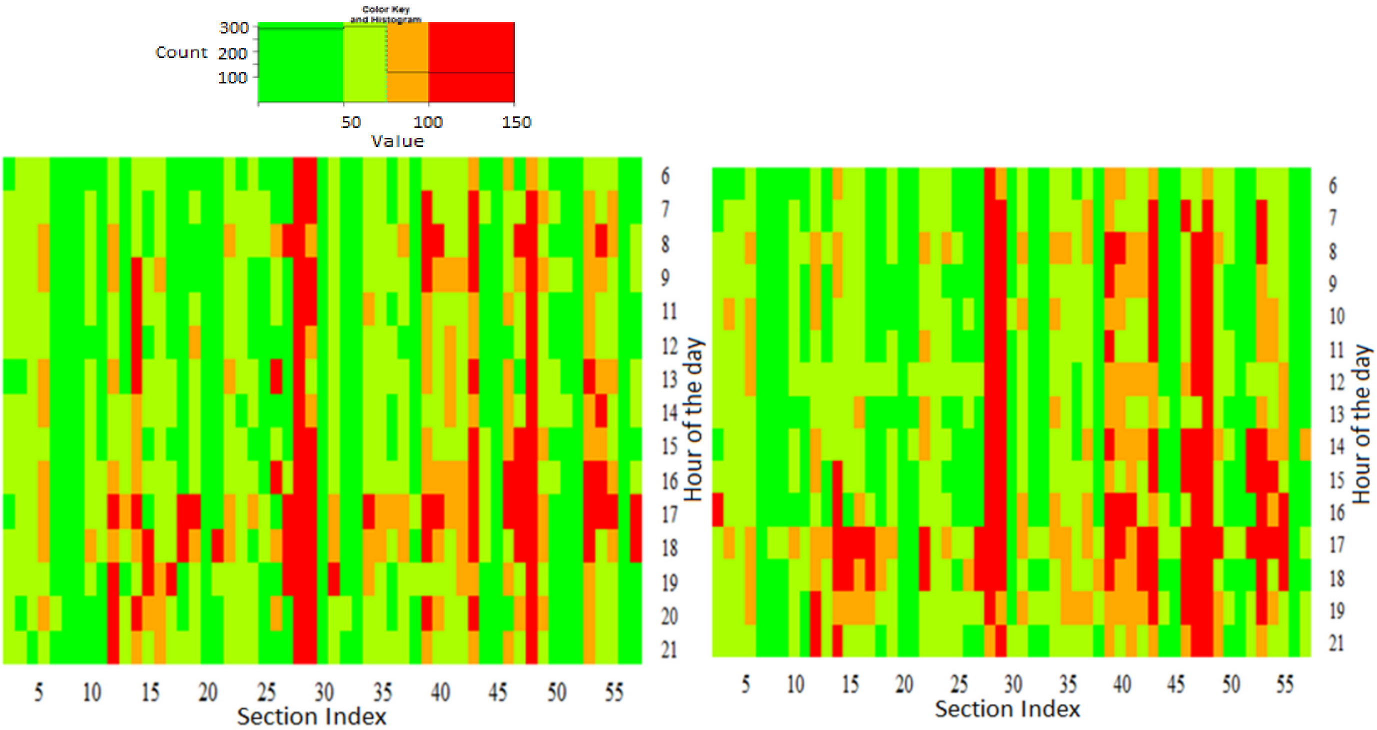


(a) (b)

Figure 2. Time-space plots for (**a**) peak and off-peak trajectories; (**b**) daily trajectories.

Previous studies from similar traffic conditions on the prediction of bus travel times [18,31] assumed that travel times followed weekly patterns. It was assumed that peak and off-peak timings remained constant over different weeks. To test the validity of this assumption, heat maps were plotted. Figure 3a,b show the heat maps of average hourly travel time for two consecutive Fridays. The heat maps are color-coded according to the travel time. It can be observed from the heat maps that the peak and off-peak timings in both cases do not remain the same. For example, consider the variation of travel times for Section 42 on both days. On week 1, travel time peaks during the time intervals 7:00 am to 10:00 am, 1:00 pm to 2:00 pm, 3:00 pm to 7:00 pm, and 8:00 pm to 9:00 pm. However, in week 2, travel time peaks during different periods (7:00 am to 12:00 noon, 2:00 pm to 4:00 pm, and 5:00 pm to

8:00 pm). Hence, to identify significant regressors for prediction, a static, manual grouping based on fixed clusters may not be efficient, as the travel time data has enough variations.



(**a**) (**b**)

Figure 3. Heat maps showing travel time variations for (**a**) October 4; (**b**) October 11.

Additionally, Figure 3 shows that the travel times on the study stretch primarily fall into the lower travel time group, as indicated by the green coded cells. In some cases, the travel times rise to a higher value, as shown in the red cells. This study looks into whether separating these would aid in improving the accuracy of travel time predictions. Data-driven clustering techniques can be applied to group this type of highly varying travel times properly. In the present study, a systematic performance comparison of the different methods in which data can be presented to a prediction algorithm and the effects on prediction accuracy is studied. The different cases considered are:

1. Predictions on base data: a single model applied to the complete set of data without considering any groups within the data;
2. Predictions using fixed clusters: grouping the dataset manually based on chronological factors;
3. Predictions using data-derived clusters: dividing the dataset into an optimum number of clusters using unsupervised learning.

In the case of predictions on base data, the whole dataset is provided to the prediction algorithm without looking into the patterns and/or groups within the data. However, it can be seen from the time-space plots in Figure 2 that the travel times on weekends are considerably lower than those on weekdays. Additionally, it can be observed that there were some distinct peak and off-peak trips on the weekdays. Hence, in the case of predictions using fixed clusters, the dataset is divided manually into three groups: trips made on weekdays during peak times, trips on weekdays during off-peak times, and trips made on weekends; three separate predictors were trained on these groups. However, the time-space plot and Figure 3 showed sufficient overlaps. This is against the assumption made in case 2, where patterns were assumed to be constant. This indicates that a manual grouping may not be sufficient. Hence, a dynamic grouping using an unsupervised learning algorithm is used to group the data based on their magnitude across space and time under case 3. Finally, predictions made from these three cases are compared to find whether the knowledge of well-defined groups in the data can improve the accuracy of the final application, which in this study are travel time predictions.

4. Travel Time Prediction Approaches

The adopted prediction methodology can be expressed as follows. Let $(Y_{t,s})_k$ denote the travel time it takes for section s to travel during trip t in the k^{th} group, where $t = 1, 2, \ldots, T$; $s = 1, 2, \ldots, S$, and $k = 1, 2, \ldots, K$, where T, S, and K denote the total number of bus trips, the total number of sections along the study stretch, and the optimum number of groups, respectively. The goal of the study is to predict the travel times for trips with $t = T + 1, \ldots .T + \Delta t$, given $(Y_{t,s})_k$. As discussed already, the present study concentrates on data-driven prediction techniques. The most commonly used methods under this category are time series techniques and machine learning techniques. The details of each of these techniques are discussed in the sections below.

4.1. Holt-Winters Forecasting

Holt-Winters forecasting [32] is one of the time series techniques that uses exponential smoothing to model and predict time-series data with a value, trend, and seasonality. Traffic variables such as travel time can be expected to show seasonality, as the values depend greatly on the time of the day and the day of the week during which the measurements were made. The travel times may be particularly high for a section of the road during particular hours of the day and may be lower for the same section of the road during a different time of the day. This may make the prediction of travel times on arterial roads an excellent candidate for applying Holt-Winters forecasting. An additive Holt-Winters forecasting method is chosen to predict travel times, and it is assumed that the seasonality in travel times remains constant through the series. The forecasted value is the sum of the baseline, trend, and seasonality components. The recursive approach to the Holt-Winters additive forecasting [32] is calculated as follows:

$$\text{Baseline: } L_t = \alpha(Y_t - S_{t-1}) + (1-\alpha)(L_{t-1} + b_{t-1}), \quad (1)$$

$$\text{Trend: } b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}, \quad (2)$$

$$\text{Seasonality: } S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-1}, \quad (3)$$

$$\text{Forecast: } \widehat{Y_{t+i}} = L_t + ib_t + S_{t+i-1}, \quad (4)$$

where Y_t is the measured travel time, $\hat{Y}_{t+i}$ is the predicted travel time, l is the length of the seasonal cycle, and α, β, and γ are the parameters of the Holt-Winters filter such that $0 \le \alpha \le 1$, $0 \le \beta \le 1$ and $0 \le \gamma \le 1$. The optimal values of α, β, and γ are found by minimizing the squared one-step prediction errors.

4.2. Random Forest (RF)

Random forest [33] is a powerful machine learning tool that combines both bootstrapping and random feature selection. It builds a group of decision trees, in which the decision rules are learned from the features of the data during training and are used to build a model. Features and observations are selected randomly, and decision trees are built on the bootstrapped samples. The number of samples chosen is maintained roughly equal to the square root of the total number of predictors, $\sqrt{p}$, where p is the number of predictors, and the predictions are averaged [34]. This ensures that the built trees are not correlated and gives more reliable forecasts. The decision of how the data branches from each node are decided by the mean squared error (MSE) as given by Equation (5). The distance between the predicted value and each node is calculated, and the branch which gives a lower value of MSE is selected.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2, \quad (5)$$

where Y_i is the measured travel time, $\hat{Y}_i$ is the predicted travel time, and n is the total number of observations.

4.3. Artificial Neural Networks (ANN)

ANNs have been extensively used for the task of prediction. ANNs are trained from historical data to uncover the patterns within the data. It can handle a large amount of data and is also able to handle nonlinear relationships between the dependent and independent variables. From the preliminary data analysis, the travel time along the study stretch was found to be highly varying, and it was seen that travel time depends on many factors, such as the presence of signals and intersections. Hence, ANN was expected to work well for such highly varying systems and was chosen for the present study. The objective function here is to optimize the weights to minimize the loss function L given in Equation (6).

$$L = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2, \quad (6)$$

where L is the loss function, Y_i is the measured travel time, $\hat{Y}_i$ is the predicted travel time, and n is the total number of observations. The back-propagation algorithm [35] was used to train the neural networks. The parameters of ANN are decided based on the complexity of the problem [36]. For the present study, one hidden layer with two neurons was chosen.

5. Methodology

The study aims to predict the travel time, say t, for each section s, given the section travel times of all trips, up to $(t - 1)$. With the exploratory data analysis knowledge, three cases are considered: predictions on base data, predictions on data with fixed clusters, and predictions on data with data-derived clusters. The details of the implementation are explained in the sections below.

5.1. Predictions on Base Data

In this case, a single model is fitted on the whole dataset. As this is time series data, all the above prediction methods, Holt-Winters, RF, and ANN, are applied here. It is assumed that the trips are continuous between the starting point and the ending point of the study stretch. To train the model, 500 m section travel times from the previous n trips were used. The value of n was chosen as ten for the present study.

5.2. Predictions on Data with Fixed Clusters

The travel times are divided manually into groups in this case, based on chronological factors. The dataset is manually divided into three groups: trips made on weekdays during peak times, trips made on weekdays during off-peak times, and trips made on weekends; three separate predictors are trained on each of them. It is assumed that the patterns in travel time across all weekdays during peak and off-peak times remain constant, and those on all weekends remain constant. All three prediction techniques were trained using the 500 m section travel times from the previous n trips. The value of n was chosen as ten for the present study.

5.3. Predictions on Data with Data-Derived Clusters

Clustering-based partitioning is used in multiple-model learning when the input space partitioning is unknown [37]. Clustering helps to represent the system more accurately, provided a large amount of data are available [37]. This power of clustering in understanding the data better is used in prediction [38]. In clustering, the data are compressed into groups with similar members. Separate predictors are trained on each cluster after the data have been grouped into clusters. Hence, instead of training a single model on the complete dataset, which contains a mixture of data with varying characteristics, K different models are trained, each on a different cluster. In the present study, this method of training multiple models on the grouped dataset was expected to work better because the variable under consideration, travel time, is reported to be highly nonlinear and multidimensional. For the present study, the *k*-means clustering algorithm was chosen since it was observed from the literature review that it could be used to group large datasets efficiently [24].

5.3.1. Analysis of Cluster Memberships

The elbow method [39] was used to find the optimum number of clusters (K) for the dataset which was found to be five. The dataset was then divided into five clusters and was visualized using a heat map, color-coded according to the respective cluster memberships, as shown in Figure 4. Table 3 details the descriptive statistics of the five clusters. While most of the points belong to the very low and low travel time clusters (blue and green, respectively), the high and very high travel time clusters (orange and red, respectively) mainly correspond to trips belonging to sections with bus stops and/or intersections. The TIDEL park, Section 47, for example, has the highest proportion of high travel time clusters and is one of the busiest intersections across the study stretch.

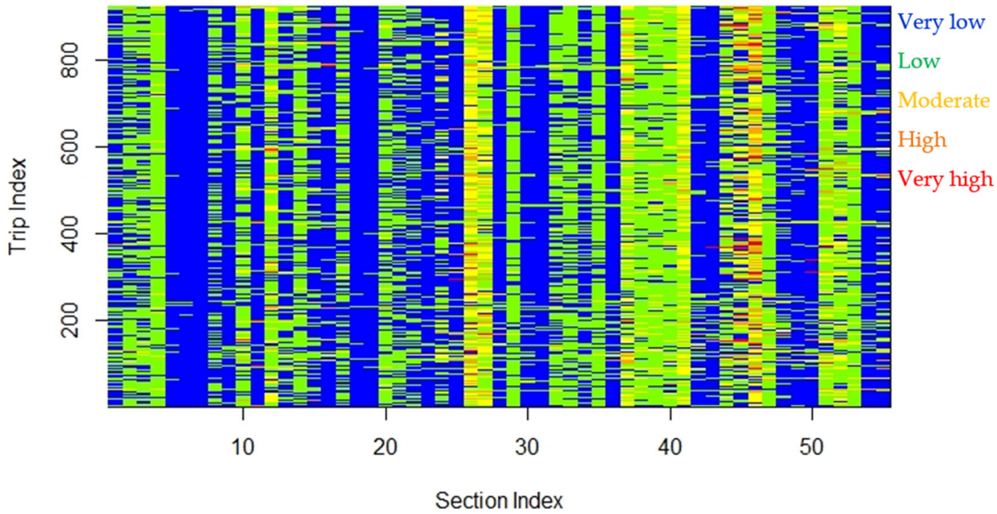


Figure 4. Cluster memberships for dynamic grouping.

Table 3. Descriptive statistics of clusters.

Cluster	Color	Number of Members	Mean of Travel Times (s)	Range of Travel Times (s)
Very low travel time	Blue	26,398	41.10	[30, 55.94]
Low travel time	Green	19,124	70.79	[55.95, 100.06]
Moderate travel time	Yellow	4019	129.37	[100.08, 188.54]
High travel time	Orange	983	247.86	[188.64, 345.41]
Very high travel time	Red	241	446.27	[347.44, 1039.06]

The next level of analysis concentrated on identifying the temporal patterns in the cluster memberships. In the first part, the hourly variation of cluster memberships was studied for a sample section with high travel time. Figure 5 shows the variation in the number of data points belonging to each cluster with the time of the day. Most of the points belong to the moderate, high, and very high travel time clusters (yellow, orange, and red-colored bars, respectively). Further, the number of such data points belonging to the high travel time clusters increases during peak hours of the day. The increase in such data points can be observed during both morning and evening peak hours.

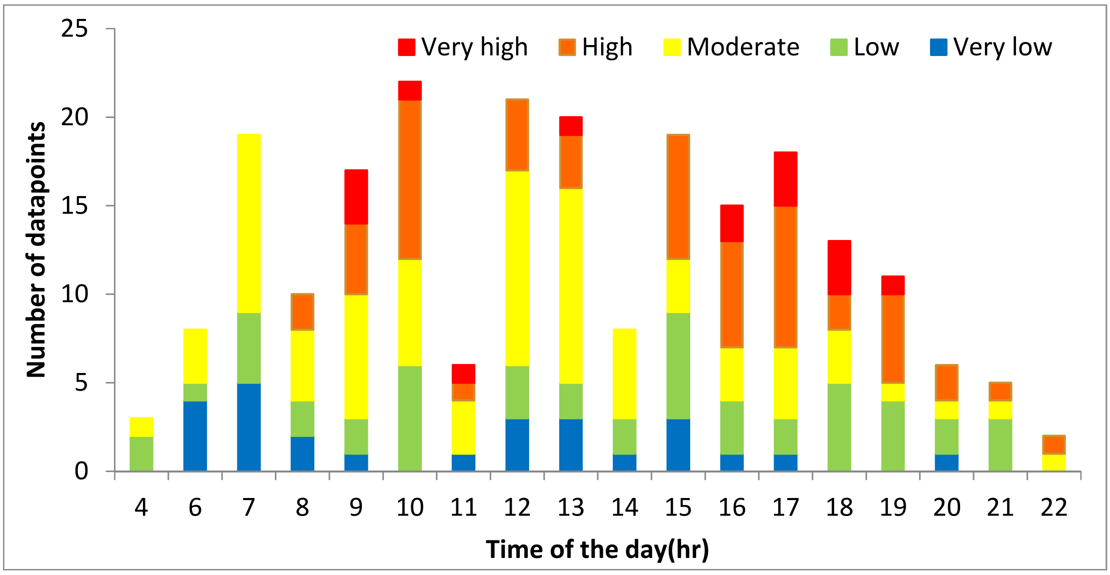


Figure 5. Cluster memberships over time of the day.

In the next part of the analysis, the cluster membership was studied across days of the week. The weekly variation of cluster memberships for a sample high travel time section is shown in Figure 6. It can be observed that Sunday does not have any points belonging to high and very high travel time clusters. This is because the traffic on Sundays is lower than other days of the week, leading to lower travel times on Sundays. The number of data points belonging to the high and very high travel time clusters is higher on other days of the week, indicating more congestion and higher travel times on these days compared to Sundays.

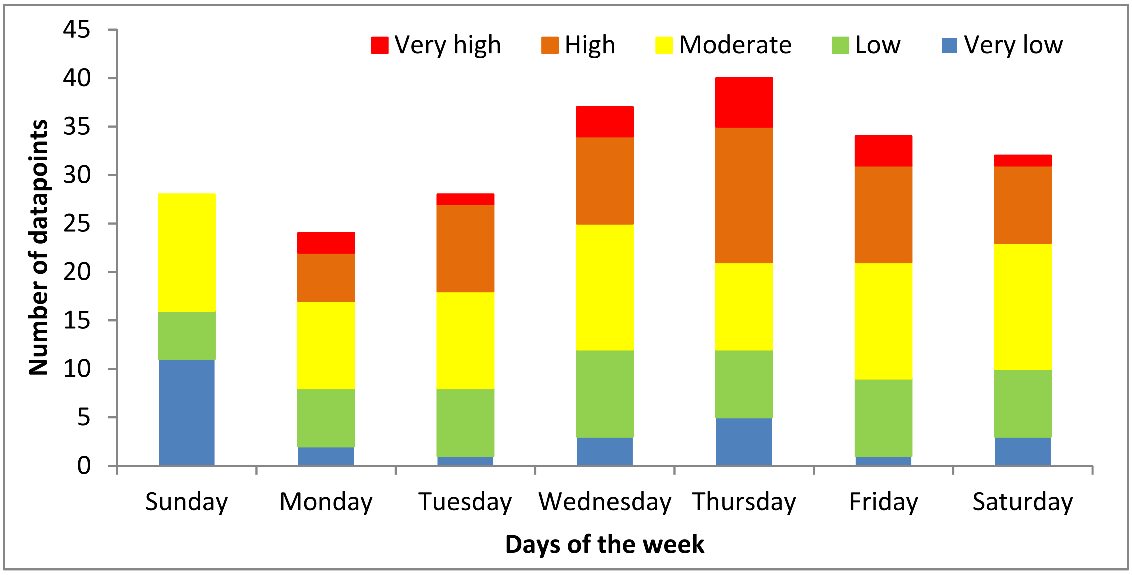


Figure 6. Cluster memberships over days of the week.

5.3.2. Travel Time Prediction

For the prediction of travel times under Case 3, both RF and ANN were chosen. Holt-Winters forecasting cannot be used for travel time prediction in this case, as once the data is clustered, it loses its continuity over time and space. Each travel time in the training dataset has three feature vectors: section index s, day index d, and 15-min interval index m. For every point in the training dataset, three neighboring points with similar day index d, 15-min index m, and section index s are searched and found within the same cluster and are used for training. If no matching training points are obtained, the database is searched to find points with any two similar feature vector values, and these travel time values will be used for further training.

These K models are used for testing after the training is over. However, the cluster to which a section travel time in the testing dataset belongs is unknown. This makes it challenging to select a suitable prediction model from the available K options. This is addressed by proposing a selection-based criterion, as discussed in [40]. In the selection-based criterion, the prediction from one of the K clusters is chosen as the best prediction for the test data based on some selection criterion. This is done by exploring the features of the clusters formed. The travel time across a study stretch is a function of the section characteristics and the time of the day at which the trip occurred. The clusters are searched to see which cluster has the maximum combination of similar section index s and 15-min index m as the test data point. Once such a cluster is found, the prediction from that cluster is used as the prediction for the section travel time under consideration. The selection-based criterion is explained in Table 4.

Table 4. Selection-based criterion.

Input:
Set of K predictions $\hat{Y} = \left[\hat{Y}_1, \hat{Y}_2, .., \hat{Y}_K\right]$ for a test point from each of the K clusters, Section for which prediction is to be made (s_{test}), Time at which prediction should be made (m_{test}). **Output:** Best prediction for the test point $\hat{Y}_{best}$.
Method: 1. In each cluster K, find the number of matching points $N = [N_1, N_2,..,N_K]$: $s_i = s_{test}$ & $m_i = m_{test}$. 2. Find $c = argmax\ (N)$. 3. Determine $\hat{Y}_{best} = \hat{Y}_c$.

6. Results and Discussions

The above-proposed methodologies were implemented in R. The mean absolute percentage error (MAPE), mean absolute error (MAE), and normalized root mean square (NRMSE) values were used to quantify the prediction accuracy. In the study, 75% of the data points were kept for training the models, and the rest of the 25% were kept for testing. The results are discussed in the next section.

6.1. Predictions on Base Data

Figure 7 shows a sample plot of the measured and predicted travel times using all three predictors. Table 5 shows the error metrics for all the cases. The prediction based on ANN yields better prediction accuracy, as seen from the values of all three error indices–MAPE, MAE, and NRMSE. Hence, it was concluded that for predictions on base data, the predictions based on ANN work best.

Table 5. Error metrics for testing (for predictions on base data).

Error Metrics	HW	RF	ANN
MAPE (%)	32.65	30.00	27.41
MAE (s)	21.78	17.87	16.67
NRMSE	0.14	0.13	0.12

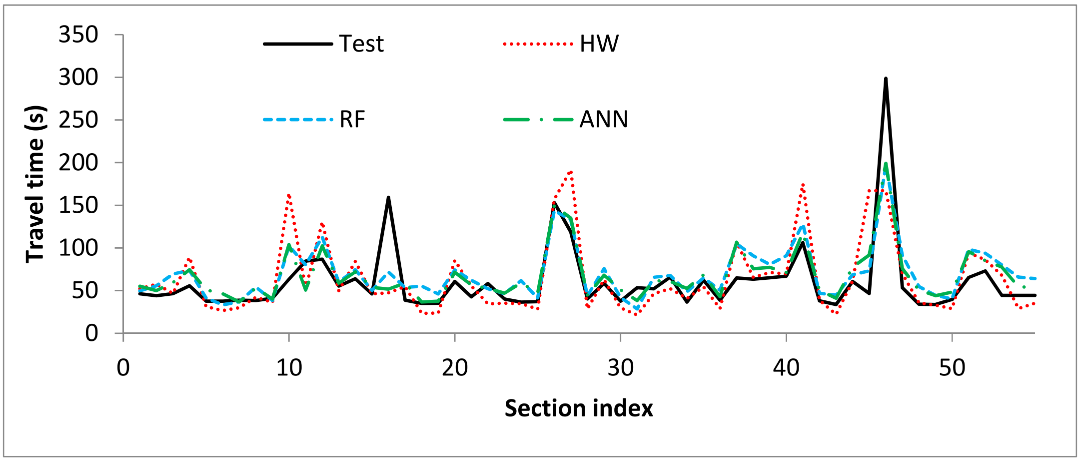


Figure 7. Measured and predicted travel times for a sample trip (for predictions on base data).

6.2. Predictions on Fixed Clusters

In this case, the dataset is divided into weekday peak trips, weekday off-peak trips, and weekend trips, and three separate predictors are trained on each of them. Figure 8 shows a sample plot of the measured and predicted travel time using all three proposed cases. It is clear that the prediction based on ANN works best, better capturing the variations in measured travel time than the other two predictors. Table 6 shows the error metrics for testing. The predictions based on ANN outperformed the predictions based on Holt-Winters and RF in this case, with a lower value of MAPE, MAE, and NRMSE.

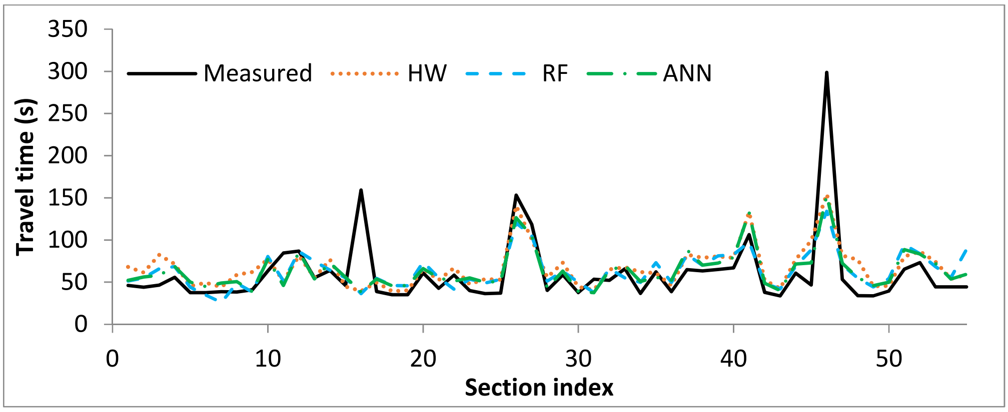


Figure 8. Measured and predicted travel times for a sample trip (for predictions on fixed clusters).

Table 6. Error metrics for testing (for predictions on fixed clusters).

Error Metrics	HW	RF	ANN
MAPE (%)	31.71	29.91	26.14
MAE (s)	19.11	17.88	16.02
NRMSE	0.12	0.12	0.11

6.3. Predictions on Data-Derived Clusters

The dataset was divided into *K* clusters, and separate prediction models were trained on each cluster. Figure 9 shows the measured and predicted travel times using both ANN and RF. Table 7 shows the error metrics for testing. Here, too, the prediction based on ANN works best, yielding lower error values.

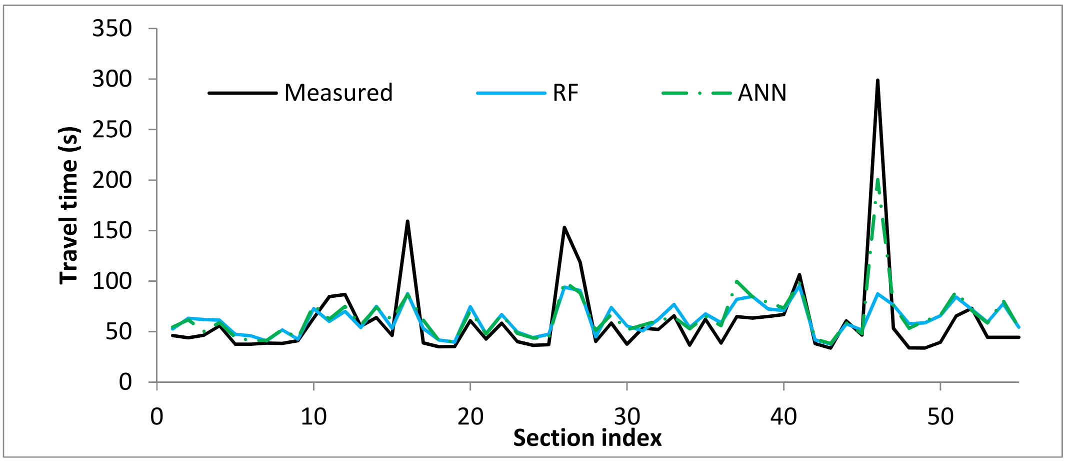


Figure 9. Measured and predicted travel times for a sample trip (predictions on data-derived clusters).

Table 7. Error metrics for testing (predictions on data-derived clusters).

Error Metrics	RF	ANN
MAPE (%)	26.119	24.884
MAE (s)	16.980	15.145
NRMSE	0.109	0.087

6.4. Comparison between Predictions on Base Data, Fixed Clusters, and Data-Derived Clusters

To check the effect of predictions on base data, fixed clusters, and data-derived clusters, the performance of all three cases was compared. In all three cases, the predictions using ANN worked best, yielding lower error values. Hence, this section compares the performance of all the predictions obtained from ANN. Figure 10 shows a sample plot of measured and predicted travel times using ANN in all three cases. Table 8 shows the corresponding error metrics. The predictions based on clustered travel time data outperformed those based on the base data and the manually grouped dataset.

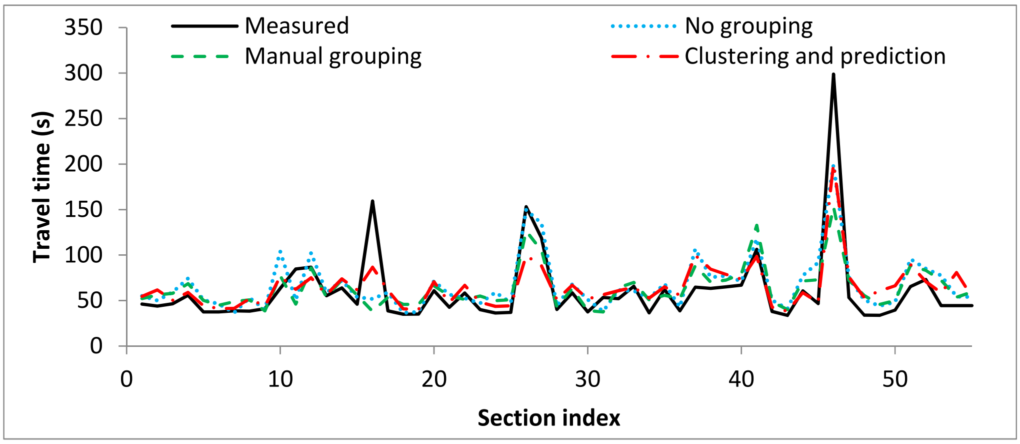


Figure 10. Measured and predicted travel times for a sample trip.

Table 8. Error metrics for testing.

Error Metrics	Base Data	Fixed Clusters	Data-Derived Clusters
MAPE (%)	27.41	26.14	24.88
MAE (s)	16.67	16.02	15.14
NRMSE	0.12	0.11	0.09

Next, a section-level comparison of MAPE was studied. Under section-level comparison, the MAPE of all trips for every section was aggregated and compared between the three cases. The variation of MAPE values along the study stretch across various sections is shown in Figure 11. Among all three instances, predictions based on dynamic grouping using clustering and prediction work best in almost all the cases, yielding lower MAPE values. It has to be noted that the prediction errors, in general, are higher in some sections. For example, the errors are higher for Section 46 than for other sections along the study stretch. This section belongs to a major intersection, the TIDEL park intersection, along the study stretch.

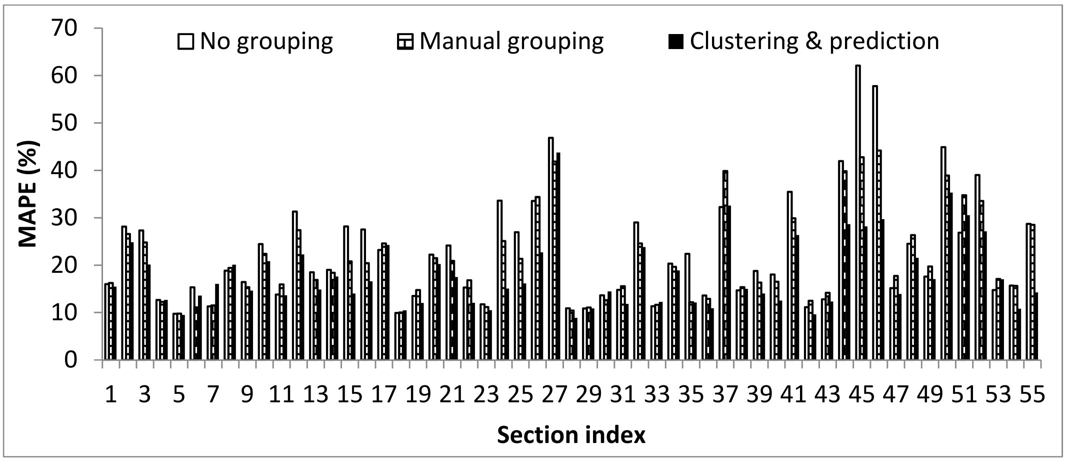


Figure 11. MAPE across sections.

The next analysis focused on the trip-level comparison. Here, the trips made on a sample day were considered, and MAPE values were plotted. Figure 12 shows the MAPE values obtained for all trips on the sample day. It can be seen here also that, among all three cases, the clustering and prediction work best in almost all the cases with lower MAPE values.

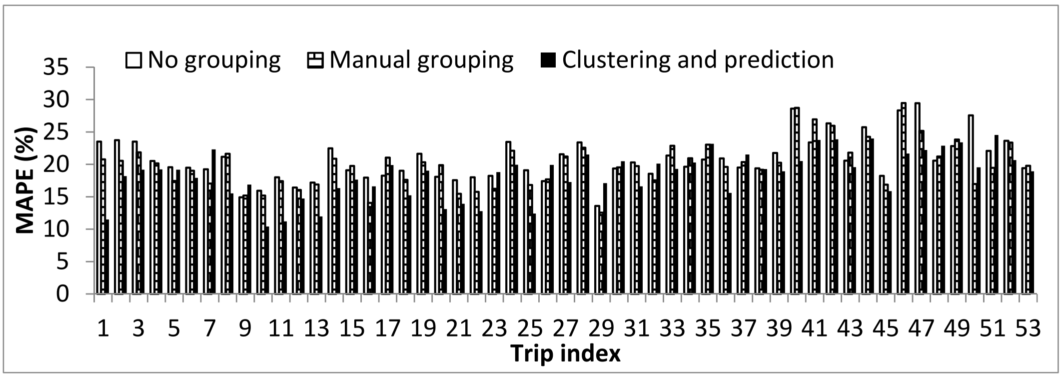


Figure 12. MAPE across all trips on a sample day.

The results from the present study were also compared to a previous study reported using similar data under mixed traffic conditions [18]. In this study, the MAPE was reported to be around 29.88%. Thus, it can be seen that the prediction error has reduced by 16.7% when predictions based on data-derived clusters were used. The results demonstrate that identifying similar points in magnitude in the dataset and training separate models on these groups helps in improving the accuracy of predictions. In addition, an automated technique such as clustering, when employed to divide the dataset, yields superior performance compared to manually dividing the dataset.

7. Conclusions

Travel time across the study stretch is highly dimensional and varies with many factors, such as the presence of bus stops, signals, and intersections. Addressing this massive variability in travel time is a challenging task. The present study compared the prediction performances of three cases, namely, predictions on base data, fixed clusters, and data-derived clusters. In the first case, no groups were considered, and a single model was trained on the complete dataset. In contrast, the dataset was divided manually based on chronological factors in the second case. In this case, three groups were considered: weekday peak trips, weekday off-peak trips, and weekend trips. In case 3, the dataset was grouped using clustering algorithms. Here, K models were trained on the K clusters, and predictions were obtained. It was seen that the clustering-based prediction approach worked best in comparison to the other methodologies. That is, instead of modeling the complete dataset, the knowledge of the presence of groups in the data can be exploited to improve the prediction accuracy. A complex traffic system can be decomposed into K groups, and then prediction models can be trained separately rather than training a model on the entire dataset for better performance.

The present study is based only on continuous tracking data from transit buses. Other data sources, such as weather data, incident data, etc., can be combined with this data to better capture unforeseen events and improve prediction accuracy. Additionally, the frequency of GPS data for the present study is 10 s. More frequent GPS data may be considered in future works. Another limitation of this study is that even for predictions based on data-derived clusters, the prediction errors were high for some sections across the study stretch. On closer observation, these sections were identified as either bus stops or intersections. A deeper analysis may be required to reduce the errors in prediction in these sections by considering more dynamic groupings within the travel time data. The study concludes that there are definite patterns in the travel time data. A systematic investigation of how these data characteristics can be used to improve the quality of travel time prediction is important. The improvement in travel time prediction leads to more accurate bus arrival time information to the users, lesser waiting times, and improvement in the quality of bus transport. The predictions from the joint clustering and prediction framework can also be used in other applications of Advanced Public Transportation Systems (APTS), such as bus priority at signals and dynamic bus scheduling, to improve the performance and efficiency of such systems.

Author Contributions: Conceptualization, H.S., L.V. and A.T.; methodology, H.S., L.V. and A.T.; formal analysis, H.S.; resources, L.V.; data curation, H.S.; writing—original draft preparation, H.S.; writing—review and editing, L.V. and A.T.; supervision, L.V. and A.T.; funding acquisition, L.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Metropolitan Transport Corporation Chennai (https://mtcbus.tn.gov.in/ (accessed on 4 January 2019)) and are available by emailing edp.mtc@tn.gov.in.

Acknowledgments: The authors would like to acknowledge the support from the Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), Indian Institute of Technology Madras, Chennai, India.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bin, Y.; Zhongzhen, Y.; Baozhen, Y. Bus Arrival Time Prediction Using Support Vector Machines. *J. Intell. Transp. Syst.* **2006**, *10*, 151–158. [CrossRef]
2. Kumar, S.V.; Vanajakshi, L.; Subramanian, S.C. A model based approach to predict stream travel time using public transit as probes. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 5–9 June 2011; pp. 101–106. [CrossRef]
3. Zhang, M.; Xiao, F.; Chen, D. Bus Arrival Time Prediction Based on GPS Data. *ICTE 2013* **2013**. [CrossRef]

4. Chu, L.; Oh, S.; Recker, W. Adaptive Kalman filter based freeway travel time estimation. In Proceedings of the 96th Annual Meeting of the Transportation Research Board, Washington, DC, USA, January 2005.
5. Tong, D.; Merry, C.J.; Coifman, B. Traffic information deriving using GPS probe vehicle data integrated with GIS. In Proceedings of the Center for Urban and Regional Analysis and Department of Geography, Columbus, OH, USA, 17 November 2005.
6. Chien, S.I.-J.; Ding, Y.; Wei, C. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. *J. Transp. Eng.* **2002**, *128*, 429–438. [CrossRef]
7. Kumar, B.A.; Jairam, R.; Arkatkar, S.S.; Vanajakshi, L. Real time bus travel time prediction using k-NN classifier. *Transp. Lett.* **2019**, *11*, 362–372. [CrossRef]
8. D'Angelo, M.; Al-Deek, H.; Wang, M. Travel-time prediction for freeway corridors. *Transp. Res. Rec. J. Transp. Res. Board* **1999**, *1676*, 184–191. [CrossRef]
9. Williams, B.M.; Hoel, L.A. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. *J. Transp. Eng.* **2003**, *129*, 664–672. [CrossRef]
10. Cryer, J.; Chan, K. *Time Series Analysis with Applications in R*; Springer: New York, NY, USA, 2008.
11. Kalaputapu, R.; Demetsky, M.J. Application of artificial neural networks and automatic vehicle location data for bus transit schedule behavior modeling. Moving Toward Deployment. In Proceedings of the IVHS America Annual Meeting, Washington, DC, USA, 17–20 April 1994.
12. Park, T.; Lee, S.; Moon, Y.J. Real time estimation of bus arrival time under mobile environment. In Proceedings of the International Conference on Computational Science and Its Applications, Assisi, Italy, 14–17 May 2004; pp. 1088–1096.
13. Pan, J.; Dai, X.; Xu, X.; Li, Y. A Self-learning algorithm for predicting bus arrival time based on historical data model. In Proceedings of the 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, Hangzhou, China, 30 October–1 November 2012; Volume 3, pp. 1112–1116. [CrossRef]
14. Nithishwer, M.; Kumar, B.A.; Vanajakshi, L. Deep learning– just data or domain related knowledge adds value?: Bus travel time prediction as a case study. *Transp. Lett.* **2021**, 1–11. [CrossRef]
15. Kwon, J.; Coifman, B.; Bickel, P. Day-to-Day Travel-Time Trends and Travel-Time Prediction from Loop-Detector Data. *Transp. Res. Rec. J. Transp. Res. Board* **2000**, *1717*, 120–129. [CrossRef]
16. Lee, W.C.; Si, W.; Chen, L.J.; Chen, M.C. HTTP: A new framework for bus travel time prediction based on historical trajectories. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 6–9 November 2012; pp. 279–288. [CrossRef]
17. Kumar, S.V.; Vanajakshi, L. Application of multiplicative decomposition and exponential smoothing techniques for bus arrival time prediction. In Proceedings of the Transportation Research Board 91st Annual Meeting, Washington, DC, USA, 22–26 January 2012.
18. Kumar, B.A.; Vanajakshi, L.; Subramanian, C. Pattern-Based Bus Travel Time Prediction under Heterogeneous Traffic Conditions. In Proceedings of the Transportation Research Record, Washington, DC, USA, 12–14 January 2013.
19. Vlahogianni, E.; Karlaftis, M.; Golias, J.; Kourbelis, N. Pattern-Based Short-Term Urban Traffic Predictor. In Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference, Toronto, ON, Canada, 17–20 September 2006; pp. 389–393. [CrossRef]
20. Chung, E. Classification of traffic pattern. In Proceedings of the 11th World Congress on ITS, Nagoya, Japan, 18–24 October 2003.
21. Van Der Voort, M.; Dougherty, M.; Watson, S. Combing Kohonen Maps with ARIMA Time Series Models to Forecast Traffic Flow. *Transp. Res.* **1996**, *4*, 307–318. [CrossRef]
22. Park, D.; Rilett, L.R. Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Networks. *Transp. Res. Rec. J. Transp. Res. Board* **1998**, *1617*, 163–170. [CrossRef]
23. Li, Y.; Zheng, Y.; Zhang, H.; Chen, L. Traffic prediction in a bike-sharing system. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015.
24. Ladino, A.; Kibangou, A.; Fourati, H.; de Wit, C.C. Travel time forecasting from clustered time series via optimal fusion strategy. In Proceedings of the 2016 European Control Conference (ECC), Aalborg, Denmark, 29 June–1 July 2016; pp. 2234–2239. [CrossRef]
25. Julio, N.; Giesen, R.; Lizana, P. Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms. *Res. Transp. Econ.* **2016**, *59*, 250–257. [CrossRef]
26. Tang, J.; Liu, F.; Zou, Y.; Zhang, W.; Wang, Y. An Improved Fuzzy Neural Network for Traffic Speed Prediction Considering Periodic Characteristic. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2340–2350. [CrossRef]
27. Alkheder, S.; Taamneh, M.; Taamneh, S. Severity Prediction of Traffic Accident Using an Artificial Neural Network. *J. Forecast.* **2016**, *36*, 100–108. [CrossRef]
28. Su, S.; Chaniotakis, E.; Narayanan, S.; Jiang, H.; Antoniou, C. Clustered tabu search optimization for reservation-based shared autonomous vehicles. *Transp. Lett.* **2020**, *14*, 124–128. [CrossRef]
29. Shaji, H.E.; Tangirala, A.K.; Vanajakshi, L. Evaluation of Clustering Algorithms for the Prediction of Trends in Bus Travel Time. *Transp. Res. Rec. J. Transp. Res. Board* **2018**, *2672*, 242–252. [CrossRef]
30. Statistics Sweden. Design Your Questions Right: How to Develop, Test, Evaluate and Improve Questionnaires 2004. Available online: http://www.scb.se/statistik/_publikationer/OV9999_2004A01_BR_X97OP0402.pdf (accessed on 11 January 2023).
31. Kumar, B.A.; Kumar, V.; Vanajakshi, L.; Subramanian, S.C. Performance comparison of data driven and less data demanding techniques for bus travel time prediction. *Eur. Transp.—Trasp. Eur.* **2017**, *65*, 1–17.
32. Winters, P.R. Forecasting Sales by Exponentially Weighted Moving Averages. *Manag. Sci.* **1960**, *6*, 324–342. [CrossRef]
33. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

34. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013.
35. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
36. Zurada, J.M. *Introduction to Artificial Neural Systems*; St. Paul: West Publishing Company: Minnesota, MN, USA, 1992; Volume 8.
37. Adeniran, A.A.; El Ferik, S. Modeling and identification of nonlinear systems: A review of the multimodel approach—Part 1. *IEEE Trans. Syst. Man Cybern. Syst.* **2017**, *47*, 1149–1159. [CrossRef]
38. Trivedi, S.; Pardos, Z.A.; Heffernan, N.T. The utility of clustering in prediction tasks. *arXiv* **2015**. [CrossRef]
39. Thorndike, R.L. Who Belongs in the Family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
40. Shaji, H.E.; Tangirala, A.K.; Vanajakshi, L. Joint clustering and prediction approach for travel time prediction. *PLoS ONE* **2022**, *17*, e0275030. [CrossRef]