A Novel Post Prediction Segmentation Technique for Urban Bus Travel Time Estimation

Laura Dunne
Gavin McArdle
School of Computer Science, University College Dublin
Dublin 4, Ireland
laura.dunne2@ucdconnect.ie,gavin.mcardle@ucd.ie

ABSTRACT

Buses are a vital component of an urban environment and shifting away from private cars towards public transport is important in minimising our environmental impact and creating sustainable cities. Good bus services make urban life better and safer for everyone and having reliable journey time estimates is a crucial component of a good transit service. Many techniques have been developed to predict journey times, including historical averages, statistical approaches and more recently machine learning (ML) algorithms. Several research efforts have shown that predicting the travel time of a complete bus journey is more accurate than predicting partial journeys. We propose a method of predicting travel time for a whole bus journey using ML algorithms combined with a novel post prediction segmentation technique to provide an estimate of partial journey times. This novel approach proportions the journey dynamically based on historical averages for the relevant day of the week and time of day. The ML algorithms we used to predict a whole journey time are Random Forest (RF), Support Vector Machine (SVM) and k Nearest-Neighbor (kNN). Our approach is applied to one year of data from the city-wide bus network in Dublin. Our proportioning technique gives excellent results compared to a baseline of the ratio of stop pair segments on the partial journey compared to the whole journey. The best performing ML algorithm was RF which achieved 0.16 mean absolute percentage error (MAPE) and 158 seconds mean absolute error (MAE) with our approach compared to 0.42 MAPE and 245 seconds MAE with the baseline method. The results are especially relevant on shorter journeys and on routes with large data sets. Our method achieved 0.21 MAPE on short journeys of less than 10 stops compared to 0.78 with the baseline method. This is a significant result as short journeys are challenging to predict accurately. Of the ML algorithms used, kNN required the least resources to train, whereas SVM returned the prediction quickest and required the least space to store.

CCS CONCEPTS

• **Applied computing** → **Transportation**; *Forecasting*; • **Computing methodologies** → *Supervised learning by regression.*

KEYWORDS

bus journey time prediction, bus travel time prediction, machine learning, k-nearest neighbour, random forest, support vector machine, whole journey time prediction, modelling bus networks, post prediction segmentation

ACM Reference Format:
Laura Dunne and Gavin McArdle. 2021. A Novel Post Prediction Segmentation Technique for Urban Bus Travel Time Estimation. In *The 10th International Workshop on Urban Computing Nov 01, 2021, Beijing, China.* , 9 pages.

1 INTRODUCTION

Buses have an essential role in an urban environment and are likely to become more important with increased urbanisation and an urgent need to reduce our environmental impact. The European Environment Agency reports that road transport used 12.3 million terajoules of energy in 2017, the majority of this coming from oil-derived fuels [1]. While this is a global issue there are also problems caused by the burning of fossil fuels within cities themselves, with the effect of air pollution causing significant mortality and morbidity [25]. This has been a concern for some time now and is reflected in many of the policies that are shape our world. The UN Sustainable Development Goal 11 seeks to “Make cities and human settlements inclusive, safe, resilient and sustainable”. Specifically, target 11.2 states that cities should expand public transport [28]. The Paris Agreement is a legally binding agreement to limit global warming that 196 countries signed up to in 2015. As a result, many cities now have sustainable plans in place to discourage the use of private vehicles and to encourage passengers towards sustainable transport options [6]. There are many sustainable transport options including rail, bus, cycling, walking. Cycling and walking are ideal for shorter journeys, especially in milder weather. However, buses are the most widespread form of transport because, unlike rail they don’t need an extensive infrastructure to be in place, are flexible and can be redeployed or rerouted as required. Bus services are also cost-effective compared to rail [2]. The potential benefits of a substantial proportion of urban dwellers switching to sustainable transport are dramatic. Xia et al. did an analysis in the Adelaide region of Australia on the benefit of shifting 40% of the passengers from cars to alternative methods of transport (both public transport and cycling) [34]. They found that 542 deaths/year could be prevented due to improved air quality, active transport and changes in traffic injuries. There would be further benefits in terms of improved health.

It is therefore imperative that we make bus transport an attractive option for passengers. It has been shown that punctuality

Beijing ’21, Nov 01, 2021, Beijing, China

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and timeliness of journeys have the most significant impact on passenger satisfaction [16]. Punctuality and accurately predicted journey times are synonymous and are among the most frequently requested improvements by passengers both pre-trip and during the journey [13, 14, 16, 19]. This is because passengers place greater value on low waiting time rather than decreased total journey time [15]. Kroes et al. studied the economic value of timetable change and grouped passengers into three categories [23]. Passengers either plan their journey or they do not, and planning passengers are either arrival-time constrained or departure-time constrained. An example of a planning passenger who is arrival-time constrained would be someone who must be at their office by 9 am. That same passenger might be departure-time constrained on their return journey if they want to leave the office at 6 pm. Arrival-time constrained passengers must arrive early at a bus stop to avoid missing their chosen service. They may have to take an earlier service than is necessary to guarantee they meet their arrival time constraint. Accurate journey time predictions can significantly reduce this waiting time from this type of passenger's day and encourage greater use of public transport. Cats et al. found that the potential waiting time gains associated with a prediction scheme are equivalent to the gains expected when introducing a 60% increase in service frequency [5].

Urban computing has been instrumental in improving the lives of city dwellers [39]. A huge amount of data can now be collected, stored and analysed. Examples of this include the ability to track the location of vehicles in real-time. Many bus operators equip their vehicles with GPS enabled in-vehicle devices. These Automatic Vehicle Locators (AVL) suffer some issues in urban environments but there are alternative methods such as using mobile phone network infrastructure or Wifi access points [26]. This has led to some bus operators including real-time bus tracking information at bus stops or via a mobile app [4, 32] and using real-time data to dynamically dispatch buses [38]. With our increased ability to generate and store data, so has our ability to process and analyse this data. Advances in ML methods allow us to detect subtle patterns in historical and real-time data that were not possible with statistical methods [33]. Yet, despite all the benefits of accurate journey travel time predictions and the advances in technology, predicting bus arrival time with high accuracy remains elusive. The prediction of bus arrival time is somewhat underrepresented in the literature considering how essential a mode of transport it is worldwide [30]. It is challenging to predict journey times accurately because it is a highly complex and multi-factorial problem. Bus journeys are affected by many factors, including the day of the week, time of day, weather, the volume of other traffic and passenger load. It can be challenging to disentangle the sources of travel-time variation. Yetiskul et al. analysed service characteristics, temporal and spatial dimensions and found they were all important [36]. A sensitivity analysis by Chen et al. shows that dwell time (the time the bus is stopped at a bus stop) has a greater impact on travel time than the time of the day or day of the week [10]. Dwell time is likely implicitly indicating passenger load but may also have inputs from the general congestion level and time lost rejoining the traffic flow after stopping. This likely also varies from one bus network to another and from one route to another. Chen found that 'highway' routes get higher accuracy scores and are more predictable than 'urban' routes: 94% and 78% accuracy for 'highway' and 'urban' routes, respectively [8]. Pandurangi et al. also found that 'rural' routes were more accurate than 'urban' routes [29]. Generally, congestion on a route decreases the accuracy of predictions. This is unfortunately of particular disadvantage to the urban commuter.

Table 1: Terminology definitions

Term	Definition
Network	A group of connected bus routes. They may be connected geographically (E.g. Dublin) or by bus operator (E.g. Dublin Bus).
Route	A named (or numbered) series of stops in a particular direction.
Segment	A generic term for a sub part of a bus route.
Consecutive stop pair segment	A sub part of a bus route defined by two consecutive designated stopping places on a bus route.
Whole Journey	The journey from the first stop on a route (the origin stop) to the last stop on a route (the terminus stop).
Partial Journey	Any journey between two stops on a single route that is not the whole route.
Proportion	A ratio representing the part of a whole journey that a partial journey represents.
ML Model	A trained ML algorithm.
Model	The conceptual model of the bus network and how it is broken down into segments for ML modelling.

We aim to contribute to the solution for this problem by predicting whole and partial bus journey times with a low computational and storage burden. This paper presents a ML whole route journey time prediction method with a novel post-prediction proportioning technique to estimate a given passenger's journey. This paper contains three main contributions:

- A review of the relevant literature on bus travel time prediction using ML and bus route modelling.
- A new post ML segmentation approach to predict travel time on a bus network and evaluation against a naive approach.
- A comprehensive demonstration of the scalability of the novel approach by applying the method to a year of data from a city-wide bus network .

The remainder of this paper is organised as follows: Section 2 discusses relevant previous work in the area of bus journey time prediction, Section 3 describes our approach, and Section 4 contains the results and a discussion. Finally, Section 5 presents some conclusions and areas for future work.

2 LITERATURE REVIEW

There has been a large body of work and much research effort in the area of estimating journey times within the bus network. The research traditionally relied on a statistical analysis of historical journey times and the consensus is that such approaches are not performing as hoped because the data is non-linear [27, 31, 33]. Many ML approaches have been used to predict travel time with reasonable success. The approach taken can be impacted by the objective (e.g long-term or short-term prediction) of the research as well as the data and resources available. Studies are carried out on different data sets in different cities which hides the transferability of approaches and can lead to conflicting results. Furthermore, existing studies have used a variety of terminology which makes direct comparison of approaches and results non-trivial. We try to be explicit and consistent in our use of terminology in this paper, but for clarity, definitions of important commonly used words and phrases are given in Table 1. In this section, we examine the state-of-the-art in terms of the objective of the intervention, the conceptual model of the bus network used and the application of ML.

2.1 Objective of Applications

Many of the proposed approaches on this topic are not directly comparable because they have different primary objectives. It is important to distinguish between short-term travel time prediction and long-term travel time prediction [12]. Short-term travel time predictions are designed to predict the time a bus will arrive based on monitoring its current position after its journey has begun and constantly updating its arrival time. This type of prediction is also used for the scheduling schemes by the operations team [38]. Short-term travel time prediction is useful for the passengers waiting to board the bus or for passengers planning to connect with other forms of transport. Updated estimates reduce the uncertainty of arrival for passengers and improve passenger satisfaction but it is not designed to provide a fixed estimate of when a passenger will get to their destination before the journey begins [40].

Long-term travel time prediction is required for timetable preparation and scheduling services [18]. An objective of long-term prediction is to reduce the burden on short-term prediction as much as possible. Long-term travel time prediction tries to account for cyclical traffic patterns, general weather trends, and the dwell times at various stops based on repeating trends in passenger numbers. However, it will never predict unpredictable events like road traffic accidents, roadworks or bus breakdowns. Events like this are best monitored separately and adjusted for using short-term travel time prediction. The approach we propose is focused on long-term travel time predictions. Accurate long term travel predictions are normally what is meant by punctuality and, as mentioned earlier, is highly sought after by the urban commuter.

2.2 Conceptual Model of Bus Networks

There is very little discussion in the literature about the conceptual model of a bus route or network, although there is some evidence that it has an impact on the accuracy of predictions. The most common approach within the literature is to conceptualise a bus network as a series of routes made up of consecutive stop pair segments [12, 17, 38]. This natural segmentation of a bus route can have different implications depending on the number of stops on a bus route. The distance between stops on a particular route or network would likely have an impact as it will affect the granularity of the network and the size of the segment being predicted. Some networks, like Dublin, have consecutive stops about 200m apart, whereas other networks can have stops multiple kilometres apart [27]. A significant implication of modelling a route by consecutive stop pairs is that you need multiple models for each route, usually one less than the number of stops on the route. This adds to the training and prediction time, as numerous models must be retrieved and multiple predictions made, as well as increasing the storage required for the models. Another major implication is that making predictions based on smaller sections of a route may be less accurate. Pandurangi et al. [29] found that whole network modelling was more accurate than a consecutive stop pair approach. Chien et al. [11] used Artificial Neural Networks (ANN) with their 'link-based' approach and their 'stop-based' approach, which is segments of consecutive stop pairs and found that the longer segment of the 'stop based' approach was superior. Chen et al. [9] define 'link' travel time as the time from arrival at one stop to arrival at the next stop on the route. 'Sections' are composed of multiple such 'links'. The longer 'sections' were modelled more accurately than the individual 'links', and the error reported decreased with increased 'links' per 'section'. The common pattern in all of these studies, predicting for a longer segment resulted in a smaller relative error.

The main trend in the literature to date is that while absolute error predictably increases while predicting for longer segments, the percentage (or relative error) decreases significantly [9, 17, 40, 41]. Outside of bus journey time predictions, research on traffic prediction, in general, has also established that long term prediction is less susceptible to random disturbances and shows more regular patterns than short-term predictions [7]. We build on these findings with our approach and predict the whole journey time from origin to terminus stop in a given direction. Having only a single model for each route reduces the storage and computational resources required. However, the accuracy of a given passenger's journey time prediction is of more importance than whole journey time metrics. Whole journey prediction may be the most accurate but this approach creates a problem of accurately predicting the journey time for partial journeys as most bus users do not travel from origin to terminus stops. To our knowledge, there is no literature addressing the specific problem of segmenting a whole bus journey prediction into partial journeys.

In this paper, we compare two approaches for estimating the travel time for a partial route from a ML whole journey prediction (post ML segmentation). As a baseline, we use a straightforward approach similar to the baseline used by Pandurangi et al. and use a ratio of the number of consecutive stop pair segments travelled compared with the number of consecutive stop pair segments on the route [29]. For example, if the whole bus journey consists of 50 stop pair segments and a partial journey consists of 10 stop pair segments, the proportion of the whole journey represented by the partial journey would be 10 divided by 50, or 0.2. We call this the static proportion because it stays the same regardless of the day or time, which are factors known to affect the journey time. For our proposed method, we borrow ideas from the historical averages approach [27], and apply it to calculate the proportion of the whole

Table 2: Example of the data after cleaning and preparation

Trip ID	First Stop	Second Stop	First Stop Arrival	Second Stop Arrival	Month	Day	Time	Journey Time
2071806	1	2	36027	36067	1	0	7	40
2071806	2	3	36067	36174	1	0	7	107
2071806	3	4	36174	36236	1	0	7	62
2071806	4	5	36236	36245	1	0	7	9
2071806	5	6	36245	36259	1	0	7	14

journey time that each consecutive stop pair stop represents. We do this for each day and time represented in our data set and store the results as a reference data set. This allows us to quickly compute the proportion of a whole bus journey that any partial journey is likely to take based on the day and time it occurs. We propose that this approach will leverage the accuracy of whole journey time predictions while also accounting for the cyclical variability that characterises urban transport in general and bus network in particular.

2.3 Machine Learning

As mentioned earlier, ML approaches have been shown to outperform historical averages and statistical methods for predicting journey travel time. The main ML approaches used in the literature for predicting journey times are ANN, SVM, kNN and RF [12].

It is usually not possible to make direct comparisons between studies predicting bus travel time, as the objective of the research and the conceptual model of the network varies widely as does the type and quality of data available to researchers [27]. For example. depending on the locality, supplemental traffic information may or may not be available. It is difficult to compare the performance of ML models trained on different input features (weather, headway, traffic information, passenger numbers, etc.). Throughout the literature, we see different results on different routes even with the same bus network as the characteristics (E.g. rural or urban, level of congestion, the number of intersections etc.) of the route change [3, 8, 38].

Gal et al. [17] compare many different ML models, including RF on data from Dublin. RF performed well, especially with more data. RF is very scalable compared to other methods, and in studies where it is not the best performing ML model, it still performs well. Yu et al. [37] proposed an interesting approach combining kNN and RF on a data set from Shenyang. They found this hybrid method outperformed kNN, SVM and RF, but only slightly outperformed RF, and the authors acknowledge the increased computational burden of this method. Zhang et al. [38] compared a hybrid method of SVM with a Kalman Filter (KF), RF and an ARIMA method in Shenzhen. The SVM-KF was found to outperform the other methods. Across the eight routes evaluated, the RF model's mean absolute percentage error (MAPE) was within 1% for SVM-KF. It outperformed SVM-KF during the peak morning period and was significantly worse during mid-week off-peak and weekend periods.

There are also studies that compare ANN with SVM with or without KF. Bai et al. [3] found that for three stop-pair segments in Shenzhen, ANN, SVM and KF performed similarly and were significantly outperformed by an ANN-KF and a SVM-KF. Maiti et al. [27] found SVM to outperform ANN. However, a few studies have found ANN to outperform other methods. Lin et al. [24] found ANN and hierarchical ANN to outperform KF significantly. Jeong et al. [21] found ANN outperformed historical data (HD) approaches and statistical regression approaches on data from Houston. Julio et al. [22] found ANN (Bayesian Regulation back-propagation training function) to outperform SVM on data from Santiago. The majority of the recent literature on this topic focuses on ANNs. A promising recent approach is Long Short-Term Memory (LSTM) ANN, which seems to be performing well [20, 35, 38].

However, ANNs take a long time to train and lack scalability [12]. In one example, Jeong et al. [21] looked at ANN with fourteen different training functions, and researchers chose to use Levenberg-Marquardt Backpropagation over Bayesian regularisation despite better results using Bayesian regularisation due to excessive running time on their small experimental dataset of 340 unique trips. Maiti et al. [27] demonstrated comparable levels of error with large differences in training and test time between ANN, SVM and HD approaches.

No consensus on the best ML model for bus journey time prediction has been reached. It is not even clear in which situations a particular ML model may be preferred. Even if the most accurate ML model could be identified, it may not be the deciding factor for the best algorithm, as speed and scalability are also important. Any method employed to predict journey times will need to be updated regularly and this should be considered in the design of potential solutions. The resources required for expanding the prediction approach from a limited number of routes to the entire bus network should be considered. For this reason, we chose to include RF, kNN and SVM in our experiments, but not ANN. In the next section, we detail the implementation of our proposed method.

3 METHODOLOGY

The data set used in this study was provided by the National Transport Authority (NTA) in Ireland and contains details of all of the journeys on the Dublin Bus network of 253 routes from 1st January 2018 to 31st December 2018. This represents over ten million unique bus journeys and serves a population of approximately 1.3 million people. We describe our approach in four stages: The initial stage is data preparation, followed by ML modelling with the training data for the whole bus journey with RF, kNN and SVM. A reference data set is then generated from the training data. It contains the average proportion of each consecutive stop pair segment on each route for each day/time combination. Finally, withheld test data is used to randomly generate partial journeys which are evaluated for error with our approach and baseline method.

Table 3: Example of the data at the point of ML modelling

Month	Day	Time	Total Journey Time
8	3	21	1762
11	4	12	1751
8	5	12	2151
1	2	20	1732
6	4	21	2270

Table 4: Example of the proportions reference data set

Day	Time	First Stop	Second Stop	Mean Proportion	Sample Size
0	7	1	2	0.021494	88
0	8	2	3	0.021106	96
0	9	3	4	0.022436	177
0	10	4	5	0.025301	210
0	11	5	6	0.025802	194

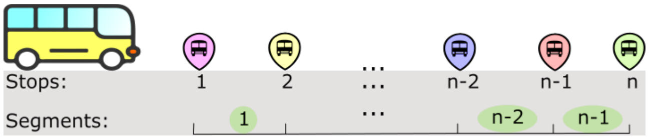

Figure 1: Bus route divided into segments by consecutive stops.

3.1 Data preparation

The data set was originally structured detailing the arrival of a bus at a bus stop. The data was restructured to represent consecutive stop pair segments rather than arrival events. We cleaned and analysed the data set, identified unique bus journeys and removed constant or irrelevant columns and any missing or corrupted data (E.g. negative journey times). We divided the data set by bus route and calculated the journey time between each consecutive stop pair. An example of the prepared data is shown in Table 2. For our purposes, we consider buses with the same headsign (E.g. 46A) going in different directions (E.g. northbound or southbound) as separate routes. One challenge was to identify a sequence of stops that defined a bus route. On some bus routes, there were over 1000 unique combinations of stops. Some reasons for this were due to arrival events being duplicated in the data set, or missed. There are also some route variations at different times of day on some routes. We determined the sequence of stops that would allow us to include the maximum amount of unique bus journeys. The data quality varied between routes and had on average 70% usable data. This ranged from 17% to 95%, with a median value of 75%. We then prepared our data set for modelling. The features in our data set after processing are three ordinally encoded categorical features for the day of the week, time of day and month. We decided not to include additional features, such as weather, to easily allow future comparisons, perhaps in different regions. An example of our data at the time of ML modelling is shown in Table 3. The three temporal features should allow the ML model to detect cyclical patterns in traffic volume and passenger numbers and make accurate predictions. The Month feature is intuitively encoded 1 to 12. The Day feature starts at 0 for Monday and ends at 6 for Sunday. The time group feature has a variable granularity. Suspected peak periods are 30 minutes long, and off-peak periods are 1 hour long and are encoded 0 to 29 starting at midnight. The reason for this is we suspected that patterns of travel might be lost if the granularity was too large at peak times, but sufficient data for predictions at off-peak times when bus services are less regular is also required. Our data structure was bench-marked against the data set with one-hot encoding, and one-hot encoding was found not to be beneficial for these ML models on this data set. Eighty five percent of the data for each route was used for both ML modelling and the precalculated reference proportions data set (Section 3.3) and the remaining 15% was withheld for testing.

3.2 Machine Learning

Whole route ML modelling was performed on all the routes on the bus network with three traditional ML algorithms (RF, kNN and SVM) using SKLearn. These algorithms are scalable algorithms that have been shown to perform well in previous studies. The ML models were optimised using a limited GridSearchCV with three-fold cross-validation. Prior to the final training of the ML models, the GridSearchCV hyper-parameter options were bench-marked using ten representative routes to determine the best range of hyper-parameters for routes in this data set. From these, 24 combinations of hyper-parameters were selected for each ML algorithm to allow for comparability of computational resources. However, the absolute minimisation of error was not the primary focus of this experiment. The purpose of the ML model is to provide a reasonable baseline prediction with which to compare the post modelling segmentation techniques.

3.3 Generating the Proportions Data set

After training the ML models, the same data was used to calculate the dynamic proportions for each consecutive stop pair on all of the routes. Table 2 shows the data after cleaning and preparation for a particular bus route. This data is at an earlier stage than the data used for ML modelling in Table 3. Each row in this data set refers to a consecutive stop pair segment as shown in Figure 1. Each segment is bounded by two consecutive stop pairs and there is one less segment on the route than the number of stops on the route. For each route, the training data set (Table 2) was filtered to the first segment (between stop 1 and 2). The data set now contains all incidences of this segment for the year and nothing else. The data set is further filtered to the first day and time combination present in the data set (E.g. Day 0 (Monday) and Time 7 (7-7:30 am)). The Trip IDs represent unique bus trips. All of the Trip IDs of buses that travelled on that segment on that day and time are retrieved and the segment journey time and the whole journey time are calculated for each unique trip. The segment proportion for each unique trip is calculated by dividing the segment journey time by the whole journey time. The average of these proportions is then saved to the new reference data set (Table 4). This is repeated for every day and time combination (E.g. Monday at Time 8, Monday at Time 9, etc.), and then the process is repeated for the remaining segments on the

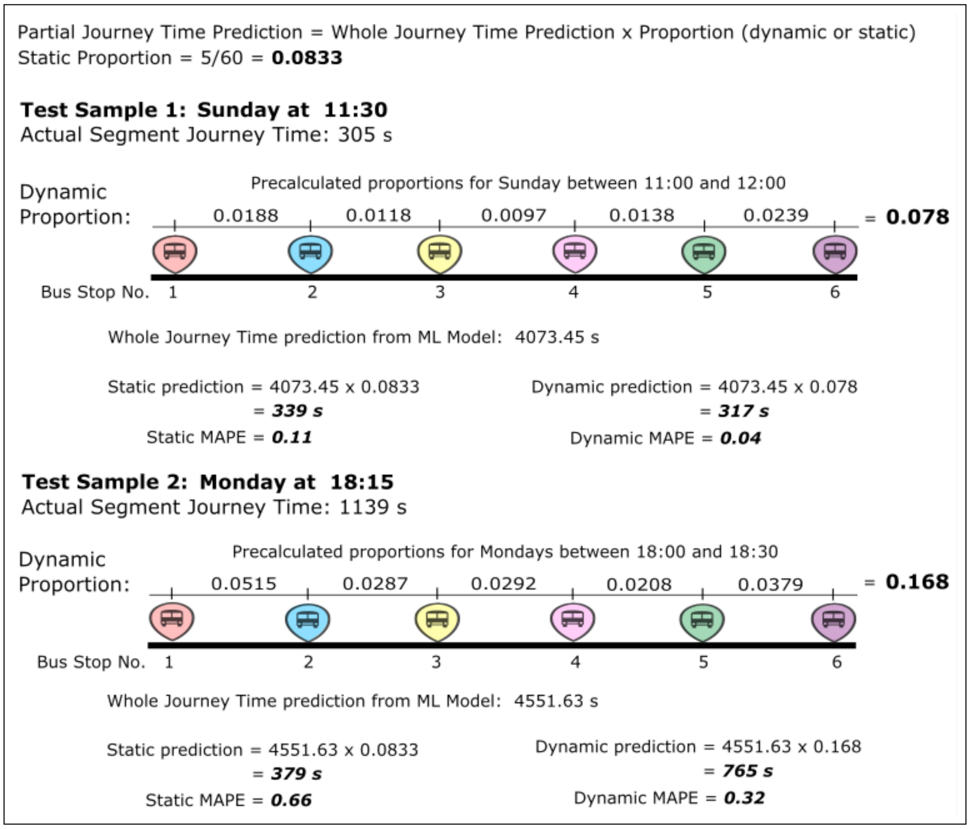
Partial Journey Time Prediction = Whole Journey Time Prediction x Proportion (dynamic or static)
Static Proportion = 5/60 = 0.0833
Test Sample 1: Sunday at 11:30
Actual Segment Journey Time: 305 s
Dynamic Proportion:
Precalculated proportions for Sunday between 11:00 and 12:00
0.0188
0.0118
0.0097
0.0138
0.0239
= 0.078
Bus Stop No. 1 2 3 4 5 6
Whole Journey Time prediction from ML Model: 4073.45 s
Static prediction = 4073.45 x 0.0833 = 339 s
Static MAPE = 0.11
Dynamic prediction = 4073.45 x 0.078 = 317 s
Dynamic MAPE = 0.04
Test Sample 2: Monday at 18:15
Actual Segment Journey Time: 1139 s
Dynamic Proportion:
Precalculated proportions for Mondays between 18:00 and 18:30
0.0515
0.0287
0.0292
0.0208
0.0379
= 0.168
Bus Stop No. 1 2 3 4 5 6
Whole Journey Time prediction from ML Model: 4551.63 s
Static prediction = 4551.63 x 0.0833 = 379 s
Static MAPE = 0.66
Dynamic prediction = 4551.63 x 0.168 = 765 s
Dynamic MAPE = 0.32

Figure 2: Comparison of static and dynamic proportioning on an off-peak and peak test sample.

bus route. The five examples in Table 4 refer to Day 0 (Monday) and Time 7-11 (6 am to 10 am) and the five segments between stops 1 and 6. The mean proportion for the segments at the given time and day is shown. The sample size is the number of unique bus journeys that served this segment at this time and day during the year. It was included for future analysis.

3.4 Evaluation Method

After producing ML models that predict travel time for the whole route and generating the dynamic proportions data set, we evaluated our model's accuracy on whole journey predictions and compared our proposed dynamic proportioning method against the baseline method for partial journeys. All experiments were performed on a 2017 MacBook Pro with a 3.3 GHz Intel Core i5 processor and 16GB of memory.

We did not have access to real commuter partial journeys so we generated a random sequence of stops from each unique journey in our test set. This was achieved by choosing two random indices from a sequential list of stops for the route the unique journey is from. We check that the same index has not been chosen twice, and the lower index becomes the boarding stop and the higher index becomes the disembarking stop of the pseudo passenger.

The static proportion (baseline) was calculated as the ratio of the number of consecutive stop pair segments on the partial journey divided by consecutive stop pair segments on the whole journey. In the example shown in Figure 2, we are predicting the travel time for 5 segments on a route with 60 segments, so the static proportion is 0.0833. To find the dynamic proportion, the reference data set for this route is filtered to only the day and time when the partial journey took place. The data set is then further filtered to only the segments in the partial journey. The mean proportions that we precalculated in the proportions data set are retrieved and summed to produce the dynamic proportion. The whole journey prediction from the ML model is multiplied by the static and dynamic proportion to get their respective predictions. All partial journey predictions were compared to the actual journey time for these partial journeys, and the MAE, MAPE, root mean square error (RMSE) and the coefficient of determination (R^2) score were calculated for both the static and dynamic predictions.

In Figure 2 Test Sample 1 is an off-peak travel time test sample, and the static and dynamic predictions are similar. The true value of this partial journey was 305s, the dynamic prediction was 317s and the static prediction was 339s. This is typically what we see at off-peak times, when there is little congestion on the route and the bus makes steady progress, the static and dynamic predictions are very similar. Test Sample 2 is a peak travel time test sample on the same 5 segments as Test Sample 1 and it took 1139s, more than 3.5 times longer. The whole journey time prediction in this example is longer, but only by 10%. This is also typical, usually, only some parts of a bus route are heavily congested even during peak

Table 5: Results for whole journeys for all ML Models

Metric	RF	kNN	SVM
MAPE	**0.07**	0.08	0.12
MAE/s	**237**	261	396
RMSE/s	**320**	349	522
R^2	**0.93**	0.92	0.78

Table 6: Results for partial journeys for all ML models using both static (Stat) and dynamic (Dyn) proportioning

ML Model	RF		kNN		SVM	
Approach	Stat	Dyn	Stat	Dyn	Stat	Dyn
MAPE	0.42	**0.16**	0.43	0.17	0.45	0.18
MAE/s	245	**158**	248	161	280	194
RMSE/s	332	**281**	337	286	393	331
R^2	0.91	**0.93**	0.9	0.93	0.87	0.9

travel times. The static proportion, by its nature, remains the same, and the dynamic proportion more than doubles. The relative error for both the static and dynamic prediction increased significantly compared to the off-peak test sample, but the dynamic prediction is significantly superior.

4 RESULTS AND DISCUSSION

The error metric and performance time results obtained from our approach on data from Dublin Bus are presented here. Table 5 details the results of the three ML Models for the whole journey prediction. RF outperforms the other ML models, with an average MAPE of 0.07 and an R^2 of 0.93. These results are comparable to what is reported in the literature, although direct comparisons are not possible due to different levels of irreducible error in different data sets. It is important to note that in this study we applied our method to every route in the network and did not pre-select particularly frequent routes, or routes with high-quality data.

Table 6 shows the full results for all partial journeys. Our proposed dynamic proportioning technique outperforms the baseline across all ML models and various metrics, with 0.16 MAPE. We consider the MAPE to be the primary metric as it allows comparisons between bus routes of different lengths and different duration. RF performs best but the results from kNN are very similar. SVM performs worse, likely because SVM usually requires more hyperparameter tuning than RF or kNN. By design, all ML models had the same amount of tuning in this experiment as described in Section 3.2 as we are also considering the resources required to produce the predictions.

4.1 Segment Length

An analysis was performed of the results by the length of the segment. We used RF and MAPE for this analysis. Short segments consisted of less than 10 stop pairs, medium segments were 10-20 stop pairs and long segments were greater than 20 stop pairs. The average segment length was 17 stop pairs. The results are shown in Table 7 and Figure 3.

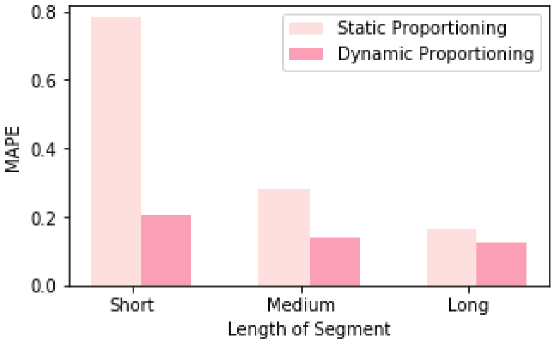

Figure 3: MAPE with RF by segment length

Table 7: MAPE by segment length using RF

Segment Length	Static MAPE	Dynamic MAPE
Short	0.78	**0.21**
Medium	0.28	**0.14**
Long	0.17	**0.12**

The results seen with dynamic proportioning are an improvement over those obtained by Gal et al. on a route from the same data set using a consecutive stop pair model and various ML models [17]. They found MAPE decreased with the length of the segment (from as high as 70% with 1-2 stops to about 15-18% for 50 stops plus). The dynamic proportioning proposed in this paper outperforms static proportioning at all segment sizes, it has the most significant performance improvement in short segments. This is the most significant result because across the literature we see far higher relative error in shorter predictions [9, 17, 41, 42]. The majority of passengers do not travel the whole bus route, so having a method to accurately predict shorter journey times would be of the most benefit to the urban population.

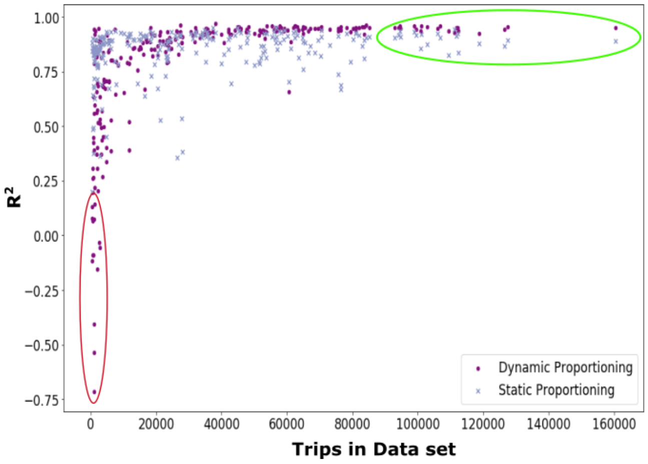

Figure 4: R^2 with RF by number of trips in route dataset

Table 8: Average resource per route required per algorithm

Algorithm	GridSearchCV Time/s	Training Time/s	Prediction Time/s	Size ML Models/MB
RF	11.66888	0.33726	0.00992	7.31
kNN	**0.32317**	**0.00179**	0.00052	0.317
SVM	22.83578	1.25737	**0.00020**	**0.165**

4.2 Size/Quality of Data

Many authors have discussed the importance of sufficient good quality data in the prediction of bus journey times and this was also an issue in this experiment [29, 41]. The overall results show the benefit of dynamic proportioning over static proportioning. The weakness of the dynamic proportioning approach can be seen in the plot of the R^2 score against the number of trips in the data set (Figure 4). All of the routes with large data sets showed very good performance, as did many of the smaller data sets. The strong performance of dynamic proportioning over static proportioning can be seen in the upper right of the plot. However, in the lower-left section of the plot, we can see some of the smaller data sets have poor R^2 scores. We looked at these routes and found that they are very infrequent services that operate an express service to universities. These routes have on average 48% usable data, whereas those routes with an R^2 greater than 0 have an average 70% usable data, but there is no absolute correlation. It is possible that some routes have data quality issues that are not detectable, perhaps less accurate sensors. With very poor quality or very small data sets using dynamic proportioning is likely resulting in error and noise in the data being compounded. There are other ways that this dynamic proportioning could be implemented. We experimented with a variant that also filtered the data frame by month. The rationale was that there are cyclical patterns in the lives of urban dwellers, like holidays or festivals, captured by season or month but not by day and time. However, this approach resulted in twelve times as many entries in the reference data set, each proportion being based on much smaller sample sizes and was found to be less accurate. Another variant that could be trialled is a coarse-grained approach where the dynamic proportions are grouped into working day or weekend and into peak or off-peak. This variant may be more appropriate on less frequent routes.

4.3 Resources Required

Any potential solution to the problem of accurately predicting journey times must be scalable and implementable for a city-wide bus network with reasonable resources. With that in mind, we analysed the average resources required to train, predict from and store the ML models used in this research. The results are shown in Table 8. We include GridSearchCV time separately to the training time with known optimal hyper-parameters. Models would likely have to be retrained daily as new data became available but the addition of a small amount of new data is unlikely to result in different optimal hyper-parameters. We expect that hyper-parameter tuning would likely be updated at a much lower frequency, perhaps monthly. The resources required by the kNN model represents the best error metrics per resources required, however the SVM has the shortest prediction time and smallest storage size. RF has an average GridSearchCV and Training time of less than SVM, however, analysis of the run times revealed that SVM is faster for the routes with smaller data sets up to around 40k-50k unique trips.

5 CONCLUSION

In this work, we have investigated the problem of long-term bus journey time prediction using ML. We evaluated the existing literature and determined the state-of-the-art methods of journey time prediction. Whole route ML modelling has been shown to have the least relative error and is a conceptual model of a bus network that is underrepresented in the literature. However, it presents a problem of accurately predicting partial journey travel times, which represent the majority of passenger bus journeys. We have proposed a method of post ML modelling segmentation based on historical averages (dynamic proportioning) and applied this method to one year of bus data on a city-wide bus network. We compared this method to a proportioning method using the number of stops on the partial journey as a ratio of the number of stops on the whole journey (static proportioning). Of the two methods compared, dynamic proportioning is more accurate than static proportioning. Dynamic proportioning is especially useful for short partial journeys which usually have poor metrics using a stop pair approach. By improving journey time predictions in a scalable way, we hope to meet the urban passenger's requirements for reliable and punctual service and encourage passengers towards public transport. The next step in our work is to compare a consecutive stop pair model of the network with the approach presented in this paper and evaluate for error, storage and computational costs. We would also like to experiment with variants of dynamic proportioning to improve results on small data sets.

ACKNOWLEDGMENTS

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] EEA. European Environment Agency. 2019. *Final energy consumption in Europe by mode of transport.* Retrieved August 21, 2021 from https://www.eea.europa.eu/data-and-maps/indicators/transport-final-energy-consumption-by-mode/assessment-10

[2] Alessandro Avenali, Giuseppe Catalano, Martina Gregori, and Giorgio Matteucci. 2020. Rail versus bus local public transport services: A social cost comparison methodology. *Transportation Research Interdisciplinary Perspectives* 7 (2020), 100200. https://doi.org/10.1016/j.trip.2020.100200

[3] Cong Bai, Zhong-Ren Peng, Qing-Chang Lu, and Jian Sun. 2015. Dynamic Bus Travel Time Prediction Models on Road with Multiple Bus Routes. *Computational*

Intelligence and Neuroscience 2015 (2015), 1–9. https://doi.org/10.1155/2015/432389

[4] Aidana Baimbetova, Kulyash Konyrova, Aigerim Zhumabayeva, and Yerkezhan Seitbekova. 2021. Bus Arrival Time Prediction: a Case Study for Almaty. *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)* 00 (2021), 1–6. https://doi.org/10.1109/sist50301.2021.9465963

[5] Oded Cats and Gerasimos Loutos. 2016. Evaluating the added-value of online bus arrival prediction schemes. In *Transportation Research Part A: Policy and Practice.* 72–80. https://doi.org/10.1016/j.tra.2016.02.004

[6] Céline Chakhtoura and Dorina Pojani. 2016. Indicator-based evaluation of sustainable transport plans: A framework for Paris and other large cities. *Transport Policy* 50 (2016), 15–28. https://doi.org/10.1016/j.tranpol.2016.05.014

[7] Chen Chen, Xiaomin Liu, Tie Qiu, and Arun Kumar Sangaiah. 2020. A short-term traffic prediction model in the vehicular cyber–physical systems. *Future Generation Computer Systems* 105 (2020), 894–903. https://doi.org/10.1016/j.future.2017.06.006

[8] Chi-Hua Chen. 2018. An Arrival Time Prediction Method for Bus System. *IEEE Internet of Things Journal* 5, 5 (2018), 4231–4232. https://doi.org/10.1109/jiot.2018.2863555

[9] Guojun Chen, Xiaoguang Yang, Jian An, and Dong Zhang. 2012. Bus-Arrival-Time Prediction Models: Link-Based and Section-Based. *Journal of Transportation Engineering* 138, 1 (2012), 60–66. https://doi.org/10.1061/(asce)te.1943-5436.0000312

[10] Mei Chen, Xiaobo Liu, Jingxin Xia, and Steven I. Chien. 2004. A Dynamic Bus-Arrival Time Prediction Model Based on APC Data. *Computer-Aided Civil and Infrastructure Engineering* 19, 5 (2004), 364–376. https://doi.org/10.1111/j.1467-8667.2004.00363.x

[11] Steven I-Jy Chien, Yuqing Ding, and Chienhung Wei. 2002. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. *Journal of Transportation Engineering* 128, 5 (2002), 429–438. https://doi.org/10.1061/(asce)0733-947x(2002)128:5(429)

[12] Teresa Cristóbal, Gabino Padrón, Alexis Quesada-Arencibia, Francisco Alayón, Gabriel de Blasio, and Carmelo R. García. 2019. Bus Travel Time Prediction Model Based on Profile Similarity. *Sensors* 19, 13 (2019), 2869. https://doi.org/10.3390/s19132869

[13] CSO. 2019. *National Transport Survey 2019.* Retrieved August 21, 2021 from https://www.cso.ie/en/releasesandpublications/ep/p-nts/nationaltravelsurvey2019/useofpublictransport/

[14] Aliasghar Mehdizadeh Dastjerdi, Sigal Kaplan, Joao de Abreu e Silva, Otto Anker Nielsen, and Francisco Camara Pereira. 2019. Use intention of mobility-management travel apps: The role of users goals, technophile attitude and community trust. *Transportation Research Part A: Policy and Practice* 126 (2019), 114–135. https://doi.org/10.1016/j.tra.2019.06.001

[15] Luigi dell'Olio, Angel Ibeas, and Patricia Cecin. 2011. The quality of service desired by public transport users. *Transport Policy* 18, 1 (2011), 217–227. https://doi.org/10.1016/j.tranpol.2010.08.005

[16] Transport Focus. 2020. *Bus Passengers Priorities for Improvement.* Retrieved August 21, 2021 from https://www.transportfocus.org.uk/publication/bus-passengers-priorities-for-improvement-2/

[17] Avigdor Gal, Avishai Mandelbaum, François Schnitzler, Arik Senderovich, and Matthias Weidlich. 2017. Traveling time prediction in scheduled transportation with journey segments. *Information Systems* 64 (2017), 266–280. https://doi.org/10.1016/j.is.2015.12.001

[18] Konstantinos Gkiotsalitis and Oded Cats. 2017. Exact optimization of Bus Frequency Settings considering Demand and Trip time variations. In *Transportation Research Board 96th Annual Meeting.*

[19] Jan-Willem Grotenhuis, Bart W. Wiegmans, and Piet Rietveld. 2007. The desired quality of integrated multimodal travel information in public transport: Customer needs for time and effort savings. *Transport Policy* 14, 1 (2007), 27–38. https://doi.org/10.1016/j.tranpol.2006.07.001

[20] Peilan He, Guiyuan Jiang, Siew-Kei Lam, and Yidan Sun. 2020. Learning heterogeneous traffic patterns for travel time prediction of bus journeys. *Information Sciences* 512 (2020), 1394–1406. https://doi.org/10.1016/j.ins.2019.10.073

[21] Ranhee Jeong and Laurence R. Rilett. 2004. Bus Arrival Time Prediction Using Artificial Neural Network Model. *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No.04TH8749)* (2004), 988–993. https://doi.org/10.1109/itsc.2004.1399041

[22] Nikolas Julio, Ricardo Giesen, and Pedro Lizana. 2016. Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms. *Research in Transportation Economics* 59 (2016), 250–257. https://doi.org/10.1016/j.retrec.2016.07.019

[23] Eric Kroes and Andrew Daly. 2018. The economic value of timetable changes. *Transportation Research Procedia* 31 (2018), 3–17. https://doi.org/10.1016/j.trpro.2018.09.042

[24] Yongjie Lin, Xianfeng Yang, Nan Zou, and Lei Jia. 2013. Real-Time Bus Arrival Time Prediction: Case Study for Jinan, China. (2013). https://ascelibrary.org/doi/pdf/10.1061/40632%28245%2978

[25] Cong Liu, Renjie Chen, Francesco Sera, Ana M. Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline S.Z.S. Coelho, Paulo H.N. Saldiva, Eric Lavigne, Patricia Matus, Nicolas Valdes Ortega, Samuel Osorio Garcia, Mathilde Pascal, Massimo Stafoggia, Matteo Scortichini, Masahiro Hashizume, Yasushi Honda, Magali Hurtado-Díaz, Julio Cruz, Baltazar Nunes, João P. Teixeira, Ho Kim, Aurelio Tobias, Carmen Íñiguez, Bertil Forsberg, Christofer Åström, Martina S. Ragettli, Yue-Leon Guo, Bing-Yu Chen, Michelle L. Bell, Caradee Y. Wright, Noah Scovronick, Rebecca M. Garland, Ai Milojevic, Jan Kyselý, Aleš Urban, Hans Orru, Ene Indermitte, Jouni J.K. Jaakkola, Niilo R.I. Ryti, Klea Katsouyanni, Antonis Analitis, Antonella Zanobetti, Joel Schwartz, Jianmin Chen, Tangchun Wu, Aaron Cohen, Antonio Gasparrini, and Haidong Kan. 2019. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *New England Journal of Medicine* 381, 8 (2019), 705–715. https://doi.org/10.1056/nejmoa1817364

[26] Wenping Liu, Jiangchuan Liu, Hongbo Jiang, Bicheng Xu, Hongzhi Lin, Guoyin Jiang, and Jing Xing. 2016. WiLocator: WiFi-Sensing Based Real-Time Bus Tracking and Arrival Time Prediction in Urban Environments. *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)* (2016), 529–538. https://doi.org/10.1109/icdcs.2016.31

[27] Santa Maiti, Arpan Pal, Arindam Pal, T Chattopadhyay, and Arijit Mukherjee. 2014. Historical data based real time prediction of vehicle arrival time. *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (2014), 1837–1842. https://doi.org/10.1109/itsc.2014.6957960

[28] UN. United Nations. 2015. *Make cities and human settlements inclusive, safe, resilient and sustainable.* Retrieved August 21, 2021 from https://sdgs.un.org/goals/goal11

[29] Ankhit Pandurangi, Clare Byrne, Candis Anderson, Enxi Cui, and Gavin McArdle. 2020. Design and Development of an Application for Predicting Bus Travel Times using a Segmentation Approach. In *Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management 2020.* 72–80. https://doi.org/10.5220/0009393800720080

[30] Engin Pekel and Selin Kara. 2017. A Comprehensive Review for Artificial Neural Network Application to Public Transportation. *Sigma Journal of Engineering and Natural Sciences* 35 (03 2017), 157–179.

[31] Amer Shalaby and Ali Farhan. 2004. Prediction Model of Bus Arrival and Departure Times Using AVL and APC Data. *Journal of Public Transportation* 7, 1 (2004), 41–61. https://doi.org/10.5038/2375-0901.7.1.3

[32] Mathieu Sinn, Ji Won Yoon, Francesco Calabrese, and Eric Bouillet. 2012. Predicting arrival times of buses using real-time GPS measurements. *2012 15th International IEEE Conference on Intelligent Transportation Systems* (2012), 1227–1232. https://doi.org/10.1109/itsc.2012.6338767

[33] Wichai Treethidtaphat, Wasan Pattara-Atikom, and Sippakorn Khaimook. 2017. Bus Arrival Time Prediction at Any Distance of Bus Route Using Deep Neural Network Model. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (2017), 988–992. https://doi.org/10.1109/itsc.2017.8317891

[34] Ting Xia, Monika Nitschke, Ying Zhang, Pushan Shah, Shona Crabb, and Alana Hansen. 2015. Traffic-related air pollution and health co-benefits of alternative transport in Adelaide, South Australia. *Environment International* 74 (2015), 281–290. https://doi.org/10.1016/j.envint.2014.10.004

[35] Zhi-Ying Xie, Yuan-Rong He, Chih-Cheng Chen, Qing-Quan Li, and Chia-Chun Wu. 2021. Multistep Prediction of Bus Arrival Time with the Recurrent Neural Network. *Mathematical Problems in Engineering* 2021 (2021), 1–14. https://doi.org/10.1155/2021/6636367

[36] Emine Yetiskul and Metin Senbil. 2012. Public bus transit travel-time variability in Ankara (Turkey). *Transport Policy* 23 (2012), 50–59. https://doi.org/10.1016/j.tranpol.2012.05.008

[37] Bin Yu, Huaizhu Wang, Wenxuan Shan, and Baozhen Yao. 2018. Prediction of Bus Travel Time Using Random Forests Based on Near Neighbors. *Computer-Aided Civil and Infrastructure Engineering* 33, 4 (2018), 333–350. https://doi.org/10.1111/mice.12315

[38] Xinming Zhang, Min Yan, Binglei Xie, Haiqiang Yang, and Hang Ma. 2021. An automatic real-time bus schedule redesign method based on bus arrival time prediction. *Advanced Engineering Informatics* 48 (2021), 101295. https://doi.org/10.1016/j.aei.2021.101295

[39] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Evaluating the added-value of online bus arrival prediction schemes. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 38. https://doi.org/10.1145/2629592

[40] Marko Čelan, Mitja Klemenčič, Anamarija L. Mrgole, and Marjan Lep. 2017. Bus-stop Based Real Time Passenger Information System – Case Study Maribor. *IOP Conference Series: Materials Science and Engineering* 245, 4 (2017), 042008. https://doi.org/10.1088/1757-899x/245/4/042008

[41] Marko Čelan and Marjan Lep. 2017. Bus arrival time prediction based on network model. *Procedia Computer Science* 113 (2017), 138–145. https://doi.org/10.1016/j.procs.2017.08.331

[42] Marko Čelan and Marjan Lep. 2020. Bus-arrival time prediction using bus network data model and time periods. *Future Generation Computer Systems* 110 (2020), 364–371. https://doi.org/10.1016/j.future.2018.04.077