A Prediction Model of Passenger Demand using AVL and APC Data from a Bus Fleet

Patroklos Samaras
Department of Informatics
Aristotle University of Thessaloniki
psamaras@csd.auth.gr

Anestis Fachantidis
Department of Informatics
Aristotle University of Thessaloniki
afa@csd.auth.gr

Grigorios Tsoumakas
Department of Informatics
Aristotle University of Thessaloniki
greg@csd.auth.gr

Ioannis Vlahavas
Department of Informatics
Aristotle University of Thessaloniki
vlahavas@csd.auth.gr

ABSTRACT

In this paper we present the passenger demand prediction model of BusGrid. BusGrid is a novel information system for the improvement of productivity and customer service in public transport bus services. BusGrid receives and processes real time data from the automated vehicle location (AVL) and the automated passenger counting (APC) sensors installed on a bus fleet and assists their operator on the improvement of bus schedules and the design of new bus routes and stops based on the expected demand. For the prediction of passenger demand in any bus stop, the raw sensor data were pre-processed and several different feature sets were extracted and tested as predictors of passenger demand. The pre-processed data were used for the supervised learning of a regression model that predicts people demand for any given bus stop and route. Experimental results show that the proposed approach achieved significant improvements over the baseline approaches. Knowledge representation, through the proposed feature set, played a key role on the ability of the prediction model to generalize well beyond its training set, to new bus stops and routes.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Miscellaneous

Keywords

Machine Learning, Prediction Models, Public Transportation, Passenger Demand Prediction

PCI 2015, October 01 - 03, 2015, Athens, Greece

ISBN 978-1-4503-3551-5/15/10...$15.00

DOI: http://dx.doi.org/10.1145/2801948.2801984

1. INTRODUCTION

Today, Public Transportation organisations play a significant role in the every day quality of life and provide their services to an ever increasing number of people. Along this, there is a need for improving the productivity and quality of these services. However, strictly improving a bus route productivity can have detrimental effects on the quality of service. Passengers need more buses with enough space to have a comfortable and safe travel to their destination. On the other hand, Public Transportation Companies have to discover optimized bus schedules, so that during operation no bus will be relatively empty or exceed a desired minimum number of passengers, leading to unprofitable lines and significant financial losses. As an example, uninformed or empirical bus scheduling can lead to overcrowded buses (low quality of service) followed by empty ones (low productivity). BusGrid produces solutions that tackle this trade-off by consistently optimizing schedules to consistently achieve a relatively stable average number of on board passengers.

Specifically, BusGrid is an integrated information system for productivity and customer service improvement in the Public Transportation Companies which receives real time data from installed sensors on the vehicles of Public Transportation Companies in order to:

- Calculate quality of service key performance indexes (KPIs) and improve bus line's productivity
- Analyse and produce useful information which will assist the prediction of critical factors (such as bus demand in each bus stop)
- Actively support the decision process of bus routes improvement and optimal response to passenger demand

In this paper we present the prediction models along with their corresponding feature sets that constitute the BusGrid's subsystem Prediction of Demand. Prediction of Demand has the goal of predicting the actual count of passengers, who, given a specific time and place, will board an arriving bus that reaches their desired destination. This work also concerns the generalization of the proposed model, when it comes to predict boarding passengers in a new route or bus stop.

The main contributions of this work are to:

1. Present the data flow from the the automated vehicle location (AVL) and the automated passenger counting (APC) systems, their pre-processing and database storage process.
2. Propose and describe two feature sets for bus passenger demand prediction.
3. Propose novel features to encode the possible passengers direction when boarding a specific vehicle.
4. Evaluate selected Machine Learning (ML) algorithms to learn a generalized model capable of approaching the real passenger demand, both on existing and unknown bus stops.
5. Provide experimental results on predicting demand, both for known routes and bus stops as also in newly introduced bus stops and directions, demonstrating the efficiency of the proposed methods and feature sets with respect to the bus demand prediction.

Finally, the results presented in this paper support the hypothesis that we can quite accurately estimate the passenger demand for arbitrary bus stops.

The rest of the paper is organized as follows: In section 2 a description of the integrated information system BusGrid is presented. Furthermore, BusGrid subsystems and its main work flow are described. In section 3, we present the proposed method for predicting bus passenger demand, from data acquisition and pre-processing, to feature extraction and learning. Section 4 presents the experimental setup and the results obtained from applying the proposed method to real AVL and APC data. In section 5 we discuss our work in the context of the related literature, while in section 6 we conclude to the main outcomes and contributions of our work and propose future research directions.

2. THE BUSGRID INFORMATION SYSTEM

BusGrid is an integrated system with a two-fold goal of improving both quality of service and productivity of bus routes. Specifically, the system comprises five modules (see Figure 1). In the first module (E1 - left portion of Figure 1) the sensors installed on a vehicle gather a series of data from the AVL and APC systems and other installed sensors such as the engine ignition on/of, door sensors etc. Using the bus's wireless communication system (e.g. GSM), a unified data-frame containing the sensor data, is sent to the database server in real-time.

Next, a data layer is created by preprocessing the raw data of the database (E3). This stage includes data cleaning and possible regularizations and aggregations of data (e.g creating a per-day view of the data). The data-layer created, forms the data-sets used by the modules that calculate the KPIs of the bus routes and the ML modules. The KPIs (see E4 - the boxes on the left side - in Figure 1) are calculated directly from the raw data.

The Machine Learning modules are used for the prediction of critical KPIs such as passenger demand and for the adaptation of the system in various cases. The first Machine Learning module (Prediction of Demand) estimates the demand for each specific bus route at each bus stop. It uses examples of temporal and geospatial data, combined with the historical data from the APC sensors, for the supervised learning of a regression model which predicts people demand for a specific bus stop, bus and period of time. The average waiting time module produces a KPI based on the actual average waiting time weighted by the passenger demand (number of people waiting) so that waiting time is expressed in man-hours.

Finally, the RL-BUS module uses Reinforcement Learning methods to dynamically create adaptive bus schedules optimized for bus productivity and quality of service. It is a semi-supervised model learning a scheduling policy for dispatching buses on the correct time to achieve optimal bus productivity with respect to the average bus fullness and waiting time (quality of service indicators). The KPIs calculation modules and the Machine Learning modules are developed as software libraries, appropriately separating the interface tier of the system (GUI) from its business logic.

3. THE PREDICTION MODEL

This section presents the proposed method for predicting bus passenger demand, from data acquisition and feature extraction to learning the prediction model.

3.1 Data Acquisition and Preprocessing

BusGrid receives raw data from the AVL and APC sensors installed in a bus fleet. A new data record is obtained every time the buses' doors close, following that passenger counting starts each time the doors get a signal to open. We further assume that each terminal bus stop is also the initial bus stop of the returning route, and we consider that the bus will continue with its returning route or a new one. Consequently, we consider that the incoming passengers of the terminal stop belong to the returning route since their actual destination is one of bus stops belonging in the returning route. Special cases such as "off-duty" buses transferring just personnel or travelling to a service station were also considered and handled appropriately.

The raw sensor data, are stored in a database table and consists of the following fields:

- count_id,
- date and time of the counting,
- the id of the bus stop,
- the id of the current route,
- the line number that this route serves,
- the number of passengers who got on the bus,
- the number of passengers who got off the bus.

The database also contains secondary tables with information about each bus stop and route, such as the latitude and longitude coordinates of each bus stop.

Next, the pre-processing methods of BusGrid, use these raw data to:

- process and correct false entries and noisy data, which come from the AVL units
- enrich the dataset with data that are not provided by the AVL and APC systems, such as weather information
- transform and extract of new features, so that the new features will be appropriately formulated for the learning methods

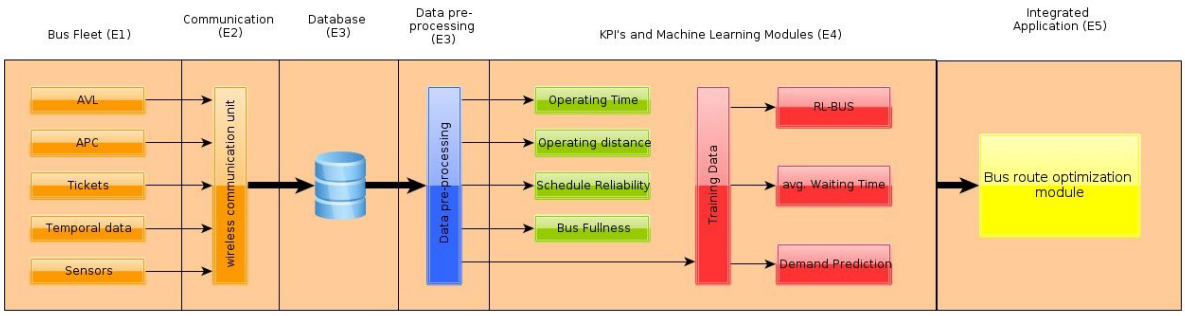


Figure 1: The BusGrid information system

- save / retrieve these preprocessed data to / from the database, so that the final dataset and its features are directly available to any other BusGrid subsystem

Concerning data enrichment we retrieve weather data in order that our model accounts also for the weather conditions that may affect the actual passenger demand. The forecast.io[1] website provides a free to use API. Using that, BusGrid retrieves weather features, such as the Apparent Temperature and Weather Summary, which encodes the general outlook of the day (eg. sunny, clear, cloudy and windy, cloudy etc.). The weather API output is a JSON string. We implemented a method which retrieves and parses this JSON string and aggregates the weather data in an hourly base. Finally, the hourly weather data are stored in a separate table in the database and merged on demand with the final feature set produced from the AVL and APC data.

3.2 Feature Selection and Extraction

A significant part of our work was to construct and evaluate suitable features for passenger demand prediction for buses able to efficiently represent bus stops and routes not present in the training dataset, that is how the regression model will be able to generalize well with respect to its input representation.

Features like route number or bus line can not be included in such a feature set. Since these features are qualitative and codified with arbitrary numbers which contain no quantitative relation in them, e.g. line numbers 13 and 14 may serve very different bus stops (geospatially distant). Such qualitative features could be used only for predictions in existing routes. BusGrid prediction model should be able to approach passenger demand on a specific location, even if the location refers to a yet, not-existing bus stop.

Consequently, the bus stop id feature was replaced with the actual geographic coordinates (latitude and longitude). The bus route code (line id) was replaced by the lambda and beta coefficients of a linear regression model describing the direction of the route following the current bus stop. The lambda coefficient represents the slope of the fitted line and beta represents the constant term. The feature set also includes the latitude and longitude coordinates of the terminal bus stop.

In a second version of the feature set we replace the lambda and beta features representing route direction with the mean latitude and longitude of the remaining bus stops. Remaining stops are the following stops of the current route, from the current bus stop to the terminal one.

The summary feature from the weather data was transformed to five ordinal features, Cloudy, Foggy, Breezy, Rainy and Windy. The new features have values from 0 to 3 according to the intensity of the weather condition they describe. For example, a Windy value of 0 means no Wind at all, 1 means barely windy, etc. A value of zero to all of these features represents a clear weather outlook.

The date and time features are included in the raw data and are very important for an accurate demand prediction model. However, a naive quantitative representation of the date and time features would wrongly return a maximum distance between dates such as 2015-01-01 and 2014-12-31. In order to correctly represent periodicity we extract only the number of day in a week (Sunday = 0, Monday = 1, ... , Saturday = 6) and the minute in day (01:12 is 60 +12 = 72nd minute of the day) and represent them in polar coordinates using their sine (1) and cosine (2). The pair of sine and cosine values of the transformed date and time features, represents them correctly as points on a circle, so that close days are also close in terms of distance.

$$weekday\_sin = \sin(sum\_of\_day * 360/7 * \pi/180) \quad (1)$$

$$weekday\_cos = \cos(sum\_of\_day * 360/7 * \pi/180) \quad (2)$$

Following the same pattern, minutes were also represented using the trigonometric transformations, formulated for the number of minutes in a day.

The features mentioned above form the final feature sets. We define the Route Regression feature set (RRf), which contains:

- lambda and beta,
- current stop 's latitude and longitude,
- last stop's latitude and longitude,
- the sine and cosine of the weekday,
- the sine and cosine of the minute of the day,
- the temperature and
- the ordinal factors Cloudy, Foggy, Breezy, Windy and Rainy.

Moreover, we define a second feature set, Round Mean feature set (RMf), which contains all the above features,

[1]http://www.forecast.io

Table 1: Prediction Model Root Mean Square Error (RMSE)

(a) Unknown/New Routes and Stops test set

	Model	RMSE
1	Random Forest with Route Regression	6.82
2	Random Forest with Route Mean	6.83
3	Mean Model	7.31
4	Bagging with Route Regression	7.37
5	Bagging with Route Mean	7.39
6	Median Model	7.81
7	1-NN with Route Mean	11.03
8	1-NN with Route Regression	11.41

(b) Known Routes and Stops test set

	Model	RMSE
1	Random Forest with Route Mean	7.04
2	Random Forest with Route Regression	7.07
3	Bagging with Route Mean	7.53
4	Bagging with Route Regression	7.53
5	Mean Model	8.45
6	Median Model	8.77
7	1-NN with Route Mean	9.19
8	1-NN with Route Regression	9.19

but instead of lambda and beta, it uses the mean latitude and longitude of the bus stops following the current one to represent the direction of the bus.

algorithm	number of trees	feature set
Random Forest	200	RRf
Random Forest	500	RRf
Random Forest	1000	RRf
Random Forest	200	RMf
Random Forest	500	RMf
Random Forest	1000	RMf
Bagging	200	RRf
Bagging	500	RRf
Bagging	1000	RRf
Bagging	200	RMf
Bagging	500	RMf
Bagging	1000	RMf

Table 2: Models and parameters tested in the experiments.

3.3 Learning and Evaluation

The purpose of the proposed prediction model is to estimate the actual passenger demand, given a specific time and place. The model will provide useful guidance in real world bus route scheduling. These predictions will also be used internally by the BusGrid RL-bus system, to design and evaluate new bus stops or new routes that were not present in the original dataset. The prediction model will also be useful for the evaluation and improvement of existing route and bus stops.

Extensive experimentation on regression methods and parameter tuning lead to the use of the Random Forest ensemble method [3], as well as the Bagging of regression trees [2]. Some of the results obtained are presented and discussed in Section 4. The feature sets formed and presented above, are being used as predictors in the selected models. Both learning algorithms showed promising results (see section 4) and are used in the last version of BusGrid for the passenger demand prediction models. The produced models using each of the two final feature and their parameters are presented in Table 2.

The target variable of the regression models is always the number of boarding passengers. Furthermore, all features used in the training and testing process of the prediction model are appropriately scaled.

4. EXPERIMENTAL RESULTS

In this section we present the experimental procedure, and discuss the results demonstrating the models performance using the two proposed feature sets RRf and RMf, for various parameters and settings (see Table 2).

The experimental procedure involves the training, testing and evaluating of the regression methods mentioned in the previous section. Our experimental data consist of 53,450 records, which were obtained from 4 vehicles that had the AVL and APC systems installed on. The AVL and APC systems of the bus fleet were created, developed and installed by Link Technologies[2], a private company that specializes in telemetry applications and O.A.S.Th. the Public Transportation Company of the city of Thessaloniki and its extended urban area. O.A.S.Th operates more than 600 buses in 76 routes that carry over 150 million passengers annually.

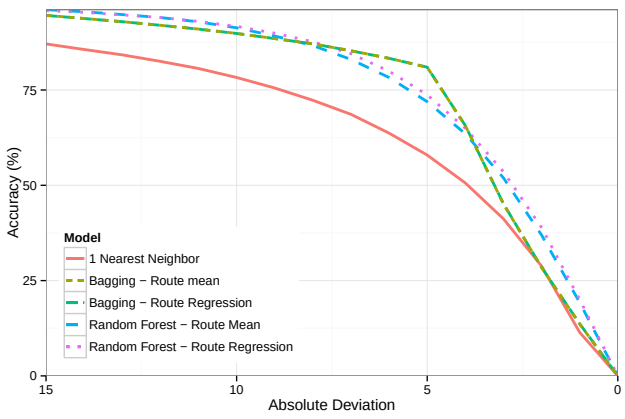
The models were trained and tested using the two final feature sets, as well as also the training and testing sets for the prediction of previously unseen bus stops or routes. Each dataset is split in 2 smaller datasets, the training set, which is a random 70% of the initial dataset, and the testing set, which is the rest 30% of the initial dataset. We used the training sets to train the selected models, random forest and bagging with regression trees. The test set was used to test the predictions of our models in other data, by comparing the actual values with the predicted ones. The performance of each model is evaluated using the Root Mean Squared Error metric (RMSE) .

A second round of experiments included the training of the same models presented in Table 2, but testing them on new routes and bus stops, not present in the training set. To produce these training and testing sets, special care was given on the selection of the test data so that it contained not only new routes but also new bus stops, not present in the training set.

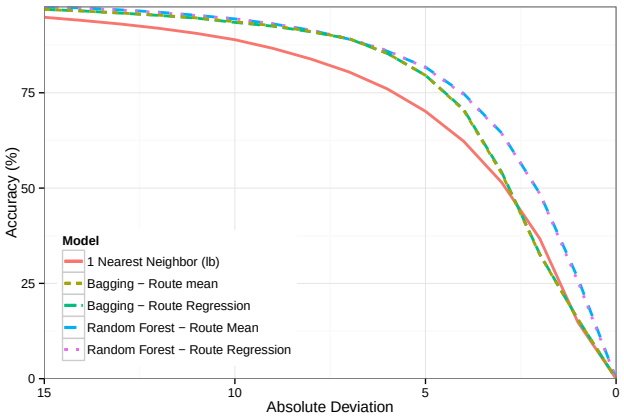
Our proposed models are tested against three baseline models. The Mean Model and Median Models always return the mean and median values respectively of the trainset's boarding passengers count. A third more complex baseline model uses a 1NN regression model finding for each instance of the testing set, the *closest* example in the training set using the Euclidean Distance based on all the quantitative features. The more complex 1NN baseline model was used with both feature sets RRf and RMf.

For the implementation of the methods and the exper-

[2]photos of the installed APC sensors, by Link Technologies: http://j.mp/BusGridSensors



(a) Uknown Stops



(b) Known Stops

Figure 2: Percentage of test set examples correctly estimated for varying levels of absolute prediction deviation thresholds.

iments we used the R statistical software[3]. The model's parameters are detailed in the rest of this section.

The evaluation of the models is depicted in the tables and plots above (see Table 1 and Figure 2). Due to space limitations we present only the models that demonstrated the best results.

Table 1(b) shows the RMSE of the examined models in known routes and bus stops. The random forest models using 200 trees performed best among all models trained and also achieved a marginally smaller error than the bagging models using the same number of trees. In both cases no significant improvement was found with more trees, and in some cases, models with more trees could overfit the training set. Trained models tested in known bus stops and routes, exhibit different performance pattern when tested in new bus stops and routes. For the models tested in known stops and routes, random forest achieves the lowest RMSE, with 200 trees and using the mean stop's latitude and longitude to represent bus direction (RMf feature set).

On the other hand, models tested in new -unseen- routes and stops (see Table 1(a)) give better results, when they use the *lambda* and *beta* features instead of the mean stop's latitude and longitude. Still, Random Forests and Bagging perform best when trained using 200 trees as in the previous experiment, with Random Forests achieving again the lowest root mean squared error. It is important to note that the 1-NN baseline models demonstrate the worse performance in this case.

Moreover, for the model with the best results, Random Forests, we examined the importance of the features as these are calculated from the method itself.

The feature importance estimations are shown in Table 3. We can observe two distinct groups of features based on their importance. The current bus stop's latitude and longitude, the terminal's latitude and longitude, the proposed lambda and beta features representing bus direction and finally the sine and cosine of the current minute of the day can be considered as the most important features, while the rest of them have smaller importance values, but their contribution as predictors is significant.

[3]http://cran.r-project.org

Feature	%Inc MSE	IncNodePurity
currentStopLatitude	51.62	279053.47
currentStopLongitude	50.36	213814.18
lambda	26.51	118222.56
beta	25.41	116149.09
endingStopLatitude	21.28	67875.20
endingStopLongitude	18.16	47138.76
apparentTemperature	9.42	285690.77
Cloudy	0.78	48501.02
Foggy	0.15	9541.096
Breezy	0.10	5964.09
Windy	2.03	13834.59
Rain	0.52	12682.73
weekdayAsSine	2.47	79054.63
weekdayAsCosine	2.94	91472.24
minuteindayAsSine	28.48	316633.48
minuteindayAsCosine	25.39	323634.34

Table 3: Feature importance as computed from the random forest model using feature set RRf. A larger value of Increased MSE indicates an important feature.

To examine the accuracy of the models for different levels of allowed error (deviation) we use two REC curves [1] depicting the number of test set examples correctly estimated (accuracy) for varying levels of allowed Mean Absolute Error (MAE) (see Figure 2). The random forest models perform best and significantly better from the baseline models in the full range of deviations up to a MAE value of ~ 2 people.

We can conclude that the proposed feature sets combined with the ensemble learning methods used, demonstrated significantly better results compared to our baseline models.

5. RELATED WORK

In [7] a passenger demand prediction model developed specifically for bus networks is presented. This work takes a time-series forecasting approach presenting a weighted ensemble prediction model from two Poisson models and an ARIMA model which successfully predict bus demand for a short-time period of P-minutes. The prediction model as

it formulated as a time-series one is applied only to specific bus-routes present in the historical data, whereas the prediction model proposed here allows for a highly accurate generalized prediction to unseen bus stops and routes. Moreover our model does not make any assumption regarding the distribution of the data (non-parametric) which can be especially important when dealing with concept-drift.

In [6] an interactive multiple model (IMM) is proposed, which comprises three time series, namely the weekly, daily, and 15 min time series. The filter algorithm-based model proposed predicts short-term passenger demand in contrast to our approach which can be used for an indefinite prediction horizon. This work formulates the prediction problem as a time-series forecasting one. In our work the three distinct time series representing periodicity, are incorporated as features controlling this way for the effect of time on the model's output.

In [4] a prediction system is presented as a part of a greater project that can be used in real-time transit information extraction. Similary to our work, it uses data retrieved via the APC and AVL systems that are installed on buses. The modelling system consists of two separate algorithms, which are based on the Kalman filter method. The first one uses the last three-day historical data of the current bus, as well as the running time of the previous bus at instant k to estimate the bus running time on a particular link at instant $k+1$. The second algorithm uses the last three-day historical passenger arrival rates on a specific bus stop to predict the current passenger arrival rate, in contrast to our work that predicts the actual number of incoming passengers based on a variety of features including weather data. Furthermore, the cited work can be applied only on previously known bus stops and routes.

In [5] a framework for real-time demand-responsive bus dispatching control is presented, which along with other methods, includes one for Short-term passenger demand forecasting. The study is oriented toward passenger trip generation. The cited method, as in [6], formulates the forecasting problem as a time-series forecasting one, uses ITS technology (such as APC systems) to retrieve the passenger volume in each bus stop and employs a simple exponential smoothing technique for the short-term passenger demand prediction. Our proposed model can be applied in not yet existing bus stops and directions. Moreover, the use of weather data and the variety of the features that constitute our proposed feature sets, enable a long-term passenger demand forecasting with significant accuracy.

6. CONCLUSIONS AND FUTURE WORK

This paper examined a way of improving the productivity and customer service quality in the public transport bus services. We presented BusGrid, an information system which uses Machine Learning methods to estimate productivity and QoS KPI indexes. We also proposed two feature sets of important predictors for an accurate passenger demand predictions.

Results showed significant benefit from using the proposed models over the baseline ones, producing accurate predictions for the passenger demand, both on existing bus stops and not yet existing ones.

Future work includes further experimentation on Machine Learning regression methods that will further improve the accuracy of passenger demand prediction. A study on the concept-drift characteristics of the domain will further help to produce more trend-invariant predictions. Finally, the models will be evaluated further when used by the RL-bus subsystem to create simulated passenger demand scenarios and learn the appropriate bus schedules that achieve both a better route productivity and quality of service.

7. ACKNOWLEDGEMENTS

We would like to thank Link S.A. for deploying the necessary infrastructure (AVL and APC sensors) and providing us with the data. This work has been supported by the Greek General Secretariat for Research and Technology.

8. REFERENCES

[1] J. B. Bij, R. Edu, and K. P. B. Bennek. Regression error characteristic curves. In *Twentieth International Conference on Machine Learning (ICML-2003). Washington, DC*, 2003.

[2] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] A. Shalaby and A. Farhan. Prediction model of bus arrival and departure times using avl and apc data. *Journal of Public Transportation*, 7(1):41–62, 2004.

[5] J.-B. Sheu. A fuzzy clustering approach to real-time demand-responsive bus dispatching control. *Fuzzy sets and systems*, 150(3):437–455, 2005.

[6] R. Xue, D. J. Sun, and S. Chen. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. *Discrete Dynamics in Nature and Society*, 2015, 2015.

[7] C. Zhou, P. Dai, and R. Li. The Passenger Demand Prediction Model on Bus Networks. pages 1069–1076. IEEE, Dec. 2013.