Improving Delay Forecasts in Public Transport using Machine Learning Techniques

Henning Wobken[1], Alexander Dölling, Jon-Patric Ewelt, Niklas Howad, Florian Hustede, Hendrik Jordan, Abdalaziz Obead, Jari Radler, Sebastian Schnieder, Klaas von der Heide, Ole Wehrmeyer, Mathias Wille, Barbara Rapp, Jorge Marx Gómez,

1. Introduction

Due to the increasingly intense debate on climate change in recent years, local public transport is once again assuming an increasingly central role in the discussion on mobility concepts to substitute the focus private transportation used to have in the last decades [16]. Improving the attractivity of public bus transport is a key factor in making it competitive against car-based mobility concepts, e.g. carpooling, car sharing, and depends on the reliability and punctuality [7] of the service provided. To achieve this, the public transport service must deliver accurate information about the service of lines to its customers, enabling them to adjust their schedules or look for alternatives.

To improve the attractiveness of public transport, it is crucial to improve the information a public transport provider can relay to their customers, especially the quality of the provided delay prediction [9]. To improve the prediction, first, the factors which influence the delay of public transport services, e.g. weather and traffic, must be evaluated. Second, it must be determined how severe the influence of these factors is. To evaluate the factors influencing public transport service and to train a machine learning algorithm, this paper uses data provided by transportation providers. These providers are obligated by law to track their busses and store data about their service performance. In cooperation with AMCON, a German developer of bus data information systems [5], we are able to train a machine learning algorithm with historic service data from two transportation providers in rural and mid-level city environments. Weather and traffic data are retrieved from available interfaces.

2. Related Work

Providing accurate information on arrival and departure times in public transport is one of the key parameters for high-quality public transport. Many studies have been conducted to assess the accuracy of predictions using different data sources, methods and models. The data for the development of such models can have different sources and can either be collected historically or be available in real time: Automatic Passenger Counting (APC), Automatic Vehicle Location (AVL) and Global Positioning Systems (GPS). There are different approaches in the literature to classify the mathematical methods and models. It is possible to classify the proposed models into four categories [19]: Regression models, models for artificial

[1] Carl von Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany, henning.wobken@uni-oldenburg.de

neural networks (ANN), Kalman filter models and analytical approaches. In another comparison, the classification of analytical approaches is omitted, but the categories support vector models and nearest neighbor methods are added [4]. Another approach using a log-normal auto-regressive (AR) model approach has been successfully implemented [14] and shows that the development of a suitable model is far from complete and the classification of the models is still in discussion.

Using a regression model, correlations between several characteristics are represented by a mathematical model. A basic distinction is made between dependent and independent variables. To determine the bus travel time, different variables such as traffic flow at intersections, weather conditions or passenger demand at bus stops are decisive influencing variables [11]. Using regression models, bus travel times have been accurately predicted both under normal traffic conditions and during a temporary road closure, e.g. due to a road improvement measure [2]. In a research on the prediction of bus arrival times at bus stops with several routes, a linear regression model was successfully applied among other methods, but it turned out that in this case a Support Vector Machine (SVM) model, a statistical approach to the classification of objects, performs best for the prediction among four proposed models [30].

Another tool is the Kalman filter, which is a mathematical model for the iterative estimation of parameters based on faulty observations within a system. With the help of this filter technique, Wall and Dailey were already able to set up a first algorithm in 1999 which, in combination with GPS data and historical data, tracked the locations of public transport vehicles to predict travel times [29].

Acritical neural networks (ANN) are another approach to delay prediction, training a system inspired by the connectivity seen in the brain. Despite the theoretical background of ANN dating back to the early twentieth century [20], applying ANN especially in solving delay prediction in public transport is a novel concept [22]. To predict the delay for railway services in Germany [23], a neural network was trained and then evaluated against a rule-based system which factors in experiences and historic data. The rule-based system will use predetermined delay scenarios with the expectation that the neural network is able to abstract from known constellations causing delay, which was concluded to be the case.

Besides the actual model classification, there are many different approaches to develop dynamic arrival time prediction models, as described in a Google AI blog post [15]. In this article, the developers at Google describe the efforts the company took to improve the prediction data of the bus delays to use in their geodata service "Google Maps". The approach is splitting the bus route into multiple parts, each one gets its own delay prediction based on traffic data, and summing up each part for a total prediction on the whole route. The underlying algorithm is called long short-term memory (LSTM) and describes a special function block of recurrent neural networks, through which a kind of "long short-term memory" can be integrated [17]. The data source is provided by information the company collects from their users, used to train their machine learning algorithm. Furthermore, the discussion they conducted with members of their “Google Maps”-Team coincides with our own considerations on the issue of influence factors for bus transportation delay in general.

3. Status Quo

To illustrate the status quo, a brief introduction to the organizational structures of the German public transport system is required. In 2018, there were 2.208 bus companies operating scheduled public transport services in Germany [9]. The line networks are managed by transport associations, of which there were 648 for public transport in 2018 [27, 24]. However, these associations do not cover the entire area of Germany. In those areas without an association, the networks are managed by the district in cooperation with bus companies. Due to the complex structure of German public transport, several different bus companies can operate in a city or region. The passenger has the requirement that the timetables can be viewed centrally. Furthermore, there are apps and displays at the stops that show the planned arrival times of the buses. The buses can now be operated by different companies and use different independent software systems. This requires central data hubs to which all systems report the current timetable situation so that end customer apps can query them.

These central data hubs require standardizations for the transmission of data. These are adopted and administered by the "Verbund Deutscher Verkehrsunternehmen" (VDV). This is an industry association of the public transport sector, comprising transport companies, transport associations, clients and manufacturers [27, 24]. The VDV consists of 450 transport companies that handle 90% of the total volume of public transport in Germany. The adopted standards are called VDV writings. A list of all standards is published on the website of the VDV [28].

The central data hubs use VDV453 [25] and VDV454 [26]. These standards indicate that the target and actual timetable data are sent by the respective software system of the bus enterprise. The forecasts for arrivals at the stops are thus calculated by different systems and can also be of varying quality. The central data hub only receives this data and makes it available to other services, e.g. an end customer app. Against this background, the optimization of the forecast takes place in the context of one bus operator.

Based on an expert interview with Olaf Clausen, Managing Director of AMCON, we collected information about how predictions are determined in the current system and which data sources are used for this prediction. The data basis is the target timetable, which contains all trips and stops with GPS coordinates. Added to this, the buses report their position every ten seconds and whether they have left a stop. The current forecast of arrivals and departures is based on a simple algorithm. When a bus starts its journey, the predicted arrival times are set equal to the target timetable. At each departure at a stop, the arrival times for the following stops are recalculated. The basis for this calculation is the deviation from the target timetable. If, as an example, a bus departs with one-minute delay, a delay of one minute is set for the following stops.

4. Data Source

This section explains our data procurement and collection. This data is prepared in a self-modelled data science process using the data warehouse reference architecture of Bauer and Günzel [6]. For this purpose, the data was divided into four dimensions (actual arrival data, target arrival data, traffic and weather) of a star schema and merged into a fact table. Based on this fact table the models could be trained. Figure 1

shows the data science process. Various data sources for weather, traffic and the bus operator's data are examined. ETL processes integrate the data into the data warehouse. The data is analyzed to determine various factors for the forecast. Training and test data are extracted and applied to machine learning models. However, only part of the possible data for machine learning is extracted to test the models for unknown data. Finally, the predictions of the models are evaluated using different metrics, results of other algorithms and the actual delays.

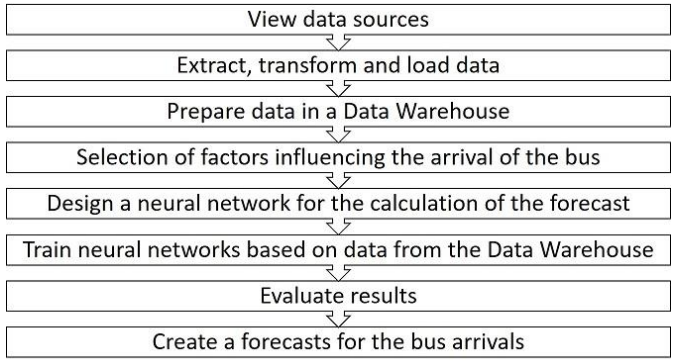

Fig. 1. Processes of data preparation and model generation

The fact table has a data set of more than five million rows with various columns for dimensions, for example jam-factor, temperature, precipitation and datetime. Weather data is provided for the entire data set, but only 6,536 rows contain information on traffic flow. Therefore, the influence of traffic on bus arrival cannot be considered in its entirety. Different columns from the fact table were discussed for using them as features for the MLP models. The analysis of [19] was also considered. Time, day of the week, month, precipitation, traffic and temperature were defined as influencing factors on a bus arrival time.

These features were validated by correlation analysis and the use of an MLP model with one feature each. The results of the analysis are shown in Table 1. We evaluate the traffic, month and day of the week as a negative factor on the arrival of the buses, because a traffic jam on a road leads to delays and there are days and months when a bus is used more often as a means of transport. Temperature and time have a positive influence as there are certain times when buses are used less, and rising temperature can lead to more use of other modes of transport. We see the precipitation as a negative influence, but we cannot prove this with our data.

Table. 1. Correlation analysis of the bus delay

column	time	weekday	month	precipitation	traffic	temperature
correlation	0.137897	-0.059135	-0.529300	0.008811	-0.016876	0.290177

The correlations can only be used as decision support for the selection of factors, since the correlation analysis did not yield clear results for one or more factors. This problem is caused by partially missing weather and traffic data. Nevertheless, positive or negative results can be interpreted for the influence of the factors on bus travel. Based on this data and selected attributes, different artificial intelligence methods were evaluated.

5. Artificial Intelligence

As shown in section 2, there have been numerous approaches, each with varying amount of success. In this chapter, we discuss different approaches and determine a suitable one to be used in the model implementation. AI is a broad field, which generally deals with enabling a computer to solve tasks which require intelligence [12]. Since we want to use the available data to automatically train a model, we need to enter the field of machine learning, which deals with self-learning artificial intelligence systems [18]. A bus arrival delay can be any real number within a reasonable time frame. This type of problem requires a numerical, continuous function to be solved. A regression under a supervised learning procedure can be used [8, 18]

There are several algorithms which can be used for regression. Because some algorithms can provide more accurate predictions than others regarding a specific problem, a suitable algorithm is determined by recommendations found in the literature and tests conducted on an early version of our data warehouse. For the tests, a model was implemented for each algorithm using Tensorflow with Keras or scikit-learn. The following four algorithms were considered:

Multilayer Perceptron (MLP) is a type of ANN. It can learn the pattern and presentation of training data, enabling it to make suitable predictions even for complex problems [3].

Support Vector Regression (SVR) is the adaption of an SVM to a regression problem [21]. Since input data of the problems discussed in this paper are not linearly separable, the SVR which would have to be used could be a two-layer neural network [18].

Decision Tree While Decision Trees are commonly used for classifications, they can also be used for regression [13]. Using the Random Forest algorithm, a more complex tree consisting of multiple decision trees can be formed for more accurate predictions [1].

Long short-term memory (LSTM) is a type of RNN that was introduced by Hochreiter and Schmidhuber in 1997 [17]. It wants to avoid the error back-flow problems of other RNNs by using gradient based learning algorithms [17]**.**

Delays in public transport can be dependent on many different factors, which might form complex relationships. An MLP model is well suited for recognizing these complex relationships [18]. In accordance with this, our tests showed that measured by the metrics R^2 and mean absolute error (MAE), the MLP model would give the best results. Therefore, MLP will be used in the model implementation. However, it must be noted that the focus of these tests was to determine a suitable algorithm quickly and not to achieve the best possible result for each algorithm. Therefore, we considered the long short-term memory approach as an additional way to make more accurate predictions for a stop. For the testing of MLP we used Tensorflow with Keras, which will also be used in the model implementation.

6. Model implementation

This section describes the process of implementing machine learning models for forecasting bus arrivals. In our work we used two different approaches one MLP model and one LSTM network.

In the determination and implementation process of the MLP model the first step is the determination of input factors, the number of layers and neurons and activation and optimizer functions. Our implementation of the genetic algorithm trains 15 models in 20 generations and uses the R^2 as the fitness metric. The first generation are 20 models whose parameters have been chosen randomly from a list of given choices. Table **2** shows the available values for our parameters.

Table. 2. Parameters for the evaluation of the model

Hidden Layers:	1, 2, 3, 4, 5
Neurons:	4, 8, 16, 32, 64, 128, 256, 512, 768, 1024
Optimizers:	adam, RMSprop, sgd, adagrad, adadelta, adamax, nadam
Activations:	relu, tanh, sigmoid, elu

After each generation the models are sorted based on their R^2 score with 1 being the best possible score. The top 40% of each generation are kept for the following generation without changing them. Of those last 60% with lower scores, 10% are also randomly kept. Those two groups make up 50% of the next generation. The other 50% are children bred from the top 40% of the previous generation. The best model of the final generation has the following parameters: hidden layers 4, neurons 768, optimizer adamax and activation elu.

The generated model was trained on a database of 2,400,000 entries of 2018 and 2019 and tested with 600,000 entries. This model has a value of $R^2 = 0.69$. The average forecasted delay was 3690.1 seconds whereas the real delay averages at around 55.76 seconds.

The average values and deviations show that the models do not recognize the actual delays. Our next step was the introduction of new input parameters. We concluded that the lacked information about how long the distance between the bus stop is, and how much of a delay there was on the previous bus stop of the line. For these new input values, the correlation values were calculated in order to obtain a comparison in the correlation between delay and input value. Those new input values have a correlation of 0.82 for the previous delay and -0.22 for the distance to the next stop.

After the calculation and introduction of those new parameters we started a second run of the genetic algorithm with 50 models in 30 generations. The model has the following parameters: hidden layers 5, neurons 128, optimizer nadam and activation elu. It was also trained on 600,000 entries of our dataset.

The resulting model has a value of $R^2 = 0.83$. The average prediction changed to -6.48 seconds and therefore has only a deviation of 62.24 seconds from the actual delays. This allowed us to determine that the previous delay is a crucial factor in achieving more efficient model. In addition, the distance has an influence on the determination of the delay, since we could only improve the model with both new input values.

As the current results are not yet satisfactory, a different approach is being evaluated. The idea is not to consider the arrivals at bus stops, but to split a bus route into a list of segments and calculate the time needed to travel in each segment. The estimate time of arrival at a bus station can then be obtained by accumulating the travel times for all segments between the current bus position and the target bus stop.
To achieve this, at first, all bus routes were split into shared segments, with the goal that each segment is both as long as possible (but never longer than the route between two stops) and shared between as many bus routes as possible. This process yielded 3,518 route segments for the analyzed area, resulting in as many models that need to be trained. An example for segments between bus stops is shown in figure 2.

Fig. 2. Example of route segmentation between two bus stops

For the preparation of the training, the time to travel the segment at a certain point in time needs to be calculated. This is being done by gathering the raw GPS data and timestamps of buses travelling there and calculating the travel time. With 170 million raw GPS datapoints, this results in around 41,000 points per segment, assuming an even distribution.

With these data, different types of models will be trained and tested. We have high hopes, that a LSTM model will perform very well in this scenario, as the memory function can for example recognize a deteriorating road state or deal better with changing weather conditions. We also hope that the increasing specialization for the model(s) can more accurately predict the conditions, and thus needed time, that the bus will have to face on that part of the road. The parameters that are going to be used for training of the models for each segment are going to be Time of day, Weekday, Month, Precipitation and Temperature.

As there are still few reliable traffic data, these are not going to be included in the model at all. In theory, the major rush-hour traffic on main roads should be accurately reflected in the historical data. The lack of traffic data also should be less of a problem for this approach due to the much higher geospatial resolution of the space the model is used for. Training one general model to factor in rush hour between certain stops of certain routes is far more difficult than training it only for the specific models of the road segments affected by rush hour.

7. Conclusion and Outlook

In this paper, bus positions and timetables data of bus provider companies in Germany were combined in a data warehouse with data from external services regarding weather and traffic. Based on these data, different machine learning models were trained using the MLP model approach. Metrics like R^2 and mean average error of the models were compared. The comparisons have shown that an MLP model can make a prediction of the delay and that there is only a small deviation from reality.

While our MLP approach involved training one generalized model, our LSTM approach involves training many specialized models. While the route segmentation needed for training the models is already finished, our LSTM models are still under development. Therefore, we could not yet present any meaningful results. But we expect the LSTM models to adapt better to the conditions of specific segments in the route network. Additionally, in the future a metamodel integrating both the LSTM model with the route segment approach and the MLP model can be tested as well. This might combine the best of the two quite different approaches.

References

[1] Abbott, D.: Applied Predictive Analytics. Wiley (2014)

[2] Abdelfattah, A.M., Khan, A.M.: Models for Predicting Bus Delays. Transportation Research Record, (vol. 1623), 8-15 (1998). https://doi.org/10.3141/1623-02

[3] Alpaydin, E.: Introduction to Machine Learning. The MIT Press Cambridge, Massachusetts, third edition edn. (2014)

[4] Altinkaya, M., Zontul, M.: Urban Bus Arrival Time Prediction: A Review of Computational Models. International Journal of Recent Technology and Engineering (vol. 2), 164-169 (2013)

[5] AMCON Software GmbH: Uber uns. https://amcongmbh.de/ueber-amcon.html (2020)

[6] Bauer, A., Gunzel, H. (eds.): Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung. dpunkt.verlag, Heidelberg, 4., uberarbeitete und erweiterte auflage edn. (2013)

[7] Boltze, M., Specht, G., Friedrich, D., Figur, A.: Grundlagen fur die Beeinflussung des individuellen Verkehrsmittelwahlverhaltens durch Direktmarketing. Tech. rep., Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL) (2002)

[8] Brownlee, J.: Master Machine Learning Algorithms. v1.1 edn. (2016)

[9] Bundesverband Deutscher Omnibusunternehmer e.V.: Zahlen, Fakten, Positionen. https://www.bdo.org/zahlen-fakten-positionen (2020)

[10] Ceder, A.: Public Transit Planning and Operation, Theory, Modelling and Practice. Elsevier, Oxford (2007)

[11] Chen, M., Liu, X., Xia, J., Chien, S.I.: A Dynamic Bus-Arrival Time Prediction Model Based on APC Data. Computer Aided Civil and Infrastructure Engineering (vol. 19), 364-376 (2004)

[12] Coppin, B.: Articial Intelligence Illuminated. Jones and Bartlett Publishers, Boston, 1st ed edn. (2004)

[13] Denison, D.G.T. (ed.): Bayesian Methods for Nonlinear Classication and Regression. Wiley Series in Probability and Statistics, Wiley, Chichester, England; New York, NY (2002)

[14] Dhivya Bharathi, B., Anil Kumar, B., Achar, A., Vanajakshi, L.: Bustravel time prediction: A log-normal auto-regressive (AR) modelling approach. Transportmetrica A: TransportScience (vol. 10), 807-839 (2020). https://doi.org/10.1080/23249935.2020.1720864

[15] Fabrikant, A.: Predicting Bus Delays with Machine Learning (2019)

[16] Hayashi, Y., Matsuoka, I., Fujisaki, K., Itoh, R., Kato, H., Rothengatter, W., Takeshita, H.: Importance of intercity passenger transport for climate change issues. In: Hayashi, Y., Morichi, S., Oum, T.H., Rothengatter, W. (eds.) Intercity Transport and Climate Change: Strategies for Reducing the Carbon Footprint, 1-30. Springer International Publishing, Cham (2015)

[17] Hochreiter, S. & Schmidhuber, J.: Long Short-term Memory. Neural computation. 9. 1735-80. (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[18] Khan, S., Rahmani, H., Shah, S.A.A., Bennamoun, M.: A Guide to Convolutional Neural Networks for Computer Vision (2018)

[19] Mazloumi, E., Currie, G., Rose, G., Sarvi, M.: Using SCATS data to predict bus travel time. In: Australasian Transport Research Forum 2009. 1-14. Australasian Transport Research Forum, Auckland New Zealand (2009)

[20] McCulloch, W.S., Pitts, W.H.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics (vol. 5), 115-133 (1943). https://doi.org/10.1007/BF02478259

[21] Mechelli, A. (ed.): Machine Learning: Methods and Applications to Brain Disorders. Elsevier, San Deigo, first edn. (2019)

[22] Pekel, E., Kara, S.S.: A Comprehensive review for artificial neural network application to public transportation. Sigma Journal of Engineering and Natural Sciences (vol. 35), 157-179 (2017)

[23] Peters, J., Emig, B., Jung, M., Schmidt, S.: Prediction of Delays in Public Transportation using Neural Networks. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents. 92-97. Web Technologies and Internet Commerce (CIMCA-IAWTIC'06, Vienna (2005)

[24] Reinhardt, W.: Öffentlicher Personennahverkehr: Technik - Rechts- Und Betriebswirtschaftliche Grundlagen. Springer, Wiesbaden, second edn. (2018)

[25] Verband Deutscher Verkehrsunternehmen (VDV): VDV-Schrift 453 Ist-Daten-Schnittstelle (2018)

[26] Verband Deutscher Verkehrsunternehmen (VDV): VDV-Schrift 454 Ist-Daten-Schnittstelle - Fahrplanauskunft (2018)

[27] Verband Deutscher Verkehrsunternehmen (VDV): Daten & Fakten zum Personenund Schienengasterverkehr. https://www.vdv.de/daten-fakten.aspx (2019)

[28] Verband Deutscher Verkehrsunternehmen (VDV): Publikationsverzeichnis. https://www.vdv.de/vdv-publikationsverzeichnis-d.pdfx (2019)

[29] Wall, Z., Dailey, D.J.: An Algorithm for Predicting the Arrival Time of Mass Transit Vehicles Using Automatic Vehicle Location Data. In: Annual Meeting of the Transportation Research Board, National Research Council. 78th Annual Meeting, National Research, Washington D.C (1999)

[30] Yu, B., Lam, W.H.K., Tam, M.L.: Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C: Emerging Technologies (vol. 19), 1157-1170 (2011)