2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Forecasting Dynamic Public Transport Origin-Destination Matrices with Long-Short Term Memory Recurrent Neural Networks

Florian Toqué, Etienne Côme, Mohamed Khalil El Mahrsi, Latifa Oukhellou
Université Paris-Est, IFSTTAR, COSYS-GRETTIA, 14-20 Boulevard Newton, F-77447 Marne-la-Vallée, France

Abstract— **A considerable number of studies have been undertaken on using smart card data to analyse urban mobility. Most of these studies aim to identify recurrent passenger habits, reveal mobility patterns, reconstruct and predict passenger flows, etc. Forecasting mobility demand is a central problem for public transport authorities and operators alike. It is the first step to efficient allocation and optimisation of available resources. This paper explores an innovative approach to forecasting dynamic Origin-Destination (OD) matrices in a subway network using long Short-term Memory (LSTM) recurrent neural networks. A comparison with traditional approaches, such as calendar methodology or Vector Autoregression is conducted on a real smart card dataset issued from the public transport network of Rennes Métropole, France. The obtained results show that reliable short-term prediction (over a 15 minutes time horizon) of OD pairs can be achieved with the proposed approach. We also experiment with the effect of taking into account additional data about OD matrices of nearby transport systems (buses in this case) on the prediction accuracy.**

I. Introduction

Several digital footprints are nowadays generated during daily trips. These sources of data can be used to implement new modeling approaches for urban mobility analysis. Different issues can be addressed, namely identifying trip purposes, revealing mobility patterns or achieving better understanding and predictions of passenger flows. Forecasting mobility demand is indeed a central problem for the organisation of any transport system. The ability to anticipate that is provided by a predictive algorithm means that suitable supply can be planned beforehand. Forecasting is therefore very frequently the first step in any attempt to optimise the distribution of resources. Currently, urban public transport supply is already to some extent determined by demand forecasting methods that the operators adjust on the basis of calendar and contextual data and typical ridership estimated on the basis of surveys and sensor data (e.g. automatic ticket checking). In this context, smoothing the data construction chain, the detailed analysis of ticketing data and the forecasting of the consequent loads on the network provide possible ways of improving the quality of service provided by transport systems.

Statistically, this problem can be tackled by a variety of methods and tools which formalisation requires to specify the relevant temporal and spatial horizons. For example, the problem of short-term forecasting based on OD matrices can be handled by tools that deal specifically with multivariate time series or machine learning methods. This paper investigates the opportunity to explore an innovative way of forecasting dynamic OD matrices in a public transport on the basis of long Short-term Memory (LSTM) Recurrent Neural Network. In several application areas, thanks to its ability to take into account long time lags in the prediction horizon, LSTM have shown better performances than other methods. In operational terms, a key challenge is to provide short-term prediction of load profiles which can be useful to improve the agility of public transport operation. Comparison with traditional tools such as calendar methodology or Vector Autoregressive (VAR) approach is carried out on a real dataset provided by a public transport operator of the Rennes métropole.

The remainder of the paper is organized as follows. A review of related work is provided in Section II. The smart card dataset used throughout our study is described in Section III. Our approach to forecasting dynamic subway OD matrices is presented in Section IV. Experimental results are detailed and discussed in Section V. Section VI concludes the paper and gives some perspectives of further works.

II. Related Work

A. Digital Footprints and Urban Mobility

In recent years, a plethora of data sources, such as ticketing logs (or smart card data) collected by Automated Fare Collection (AFC) systems in public transport, mobile phone data, geo-tagged status updates and check-ins published on social networks were leveraged to study human mobility.

As early as 2004, Bagchi and White [1] investigated the possible role smart card data can play in the analysis of travel practices and considered their potential to supplement or even replace more conventional data sources (such as houshold travel surveys). One problem already identified in the study is that of missing data, such as the personal details about cardholders (e.g. age, gender, income, etc. which are omitted due to privacy concerns), trip purposes, and alighting locations (which are rarely collected by the AFC system). In order to overcome the latter limitation, Barry et al. [2] proposed an approach to infering destinations based on two assumptions: (i) a given trip's destination is often located near the boarding location of the passenger's next boarding, and (ii) for the last trip of the day, passengers tend to alight near the same location where their first trip of the day took place. These hypotheses were widely used in subsequent research on trip reconstruction and OD matrix estimation [3], [4], [5], [6]. Trip reconstruction (i.e., grouping of transactions belonging to a same "logical" trip) is another important facet of smart card data enrichment. To this end, Chu and Chapleau [7] proposed a methodology that involves not only inferring missing alighting locations but also detecting transfers. This is conducted using thresholds on the elapsed time between two consecutive transactions and the spatial distance between their alighting and boarding locations respectively. The authors showed how the generated trip chains can be used to reconstruct bus load profiles and infer anchor points (i.e., frequently visited destinations, often corresponding to a passenger's home or work location). Finally, we can mention the recent and innovative work by Zhang et al. [8] who proposed a probabilistic approach to reconstructing trip chains.

Ticketing logs can be used to discover group patterns and unravel frequent travel behaviours. Agard et al. [9] used statistical techniques and partitioning tools (such as *k*-means and hierarchical clustering) in order to obtain more accurate representation of public transport user profiles. More advanced approaches, such as DBSCAN, Non-negative Matrix Factorization (NMF), and mixture models were also used to the effect of identifying key passenger behaviors in public transport networks [10], [11], [12], [13].

Alternatively, smart card data can also be harnessed to study travel behaviour at an individual level. Lathia et al. [14] analyzed travel data of London Underground users in order to estimate individual travel times and rank stations on the basis of frequentation to identify for each user a set of stations of interest that can be used to provide useful future travel updates. Ceapa et al. [15] used ticketing data to address the problem of station crowdedness. They showed that the overcrowding occurring at different underground stations is an extremely regular and easily forecastable, phenomenon. The authors have also shown that crowding takes place over short time frames. Earlier work on the same topic was conducted by Utsunomiya et al. [16] in order to show the frequency and regularity of individual use of public transport network. In addition to ticketing data, these authors had access to meta-data on users. Very recently, several authors have attempted to predict use. For example, Foell et al. [17] have attempted to predict trips on a bus network with fairly encouraging initial findings. A description of current knowledge about the use of ticketing data for public transport planning purposes is available in [18].

B. Passenger Flow and OD Matrix Forecasting

1) Passenger Flow Prediction:

The problem of flow prediction consists in predicting passenger counts at each station regardless of where they come from or where they are headed. A considerable amount of works in the literature are dedicated to flows prediction: [19] used empirical mode decomposition and gray support vector machines for short term forecasting of high-speed rail demand. Later, Chen et al. [20] improved on this method by using clustering. Methods based on neural network were also used in [21] and [22] with different optimization and mixing models in order to predict passenger flows.

2) OD Matrix Forecasting:

Contrary to the flow prediction problem, OD matrix forecasting consists in predicting trip counts between the different nodes (stations in the case of public transport) of the network. The problem was approached under different angles in the literature. Ashok and Ben-Akiva [23] used Kalman filters to predict time-dependent OD flows of drivers, based on traffic volume and average speed data collected using sensors installed in a roadway near Hayward, California and a freeway encircling Amsterdam, Netherlands. A similar approach was also proposed in [24] for short-term forecasting of OD Matrices using smart card data of bus boardings in China. An autoregressive integrated moving average (ARIMA) model was used in [25] for dynamic forcasting of time dependent OD flows in a dutch passenger rail. One limitation of the aforementioned linear models is that they cannot represent data with a large and general class of function like non-linear models do. Furthermore, in order to extract more information from data, linear models requires a feature engineering step that is not necessary for non-linear models.

Recurrent Neural Networks (RNN) were, to the best of our knowledge, first used for OD prediction in [26], in which the authors used a recurrent multilayer perceptron to predict a traffic matrix in a communications network. More recently, de Brébisson et al. [27] compared the performance of different neural network architectures, such as the multilayer perceptron, bidirectional recurrent neural networks, and models inspired by memory networks for the task of taxi destination prediction. While RNNs are powerful tools capable of representing context, they are reported to show theoretical and experimental limitations when trained for tasks involving long-term dependencies (i.e., tasks in which the temporal contingency between predictors and outputs span over extended periods) [28]. In our work, we use Long Short-Term Memory (LSTM) networks [29], which do not suffer from this limitation: LSTMs are specifically tailored to avoid the long-term dependency problem and memorize information over long periods of time.

III. Case Study Dataset

In this section, we present the dataset used for our study as well as the pre-processing step we conduct in order to make it usable for OD matrix prediction.

A. Dataset Description

We use smart card data provided by the Service des Transports en commmun de l'Agglomération Rennaise (STAR). The data is collected from 70 regular bus lines and 1 subway line serving the metropolitan area of Rennes, France over a fifteen months period (from April 2014 to June 2015). On average, 250000 ticketing logs are recorded per day. Each transaction contains an anonymized passenger id (for transactions made with smart cards), the timestamp (date and time up to the minute) and location (station) the transaction took place, the boarded line, and fare type information. Trip destinations aren't captured by the AFC system since passengers are required to validate their smart cards only when boarding. Due to privacy protection regulations, personal information about cardholders aren't available.

A map of Rennes showing the 20 most frequented stations is depicted in Figure 1. The subway line is clearly more used than the bus network, except for bus stops near “Université Rennes 2” (left of the city) and “République” (center) which are used by a more important number of bus lines compared to other bus stops (hence being frequented more than the other stops).

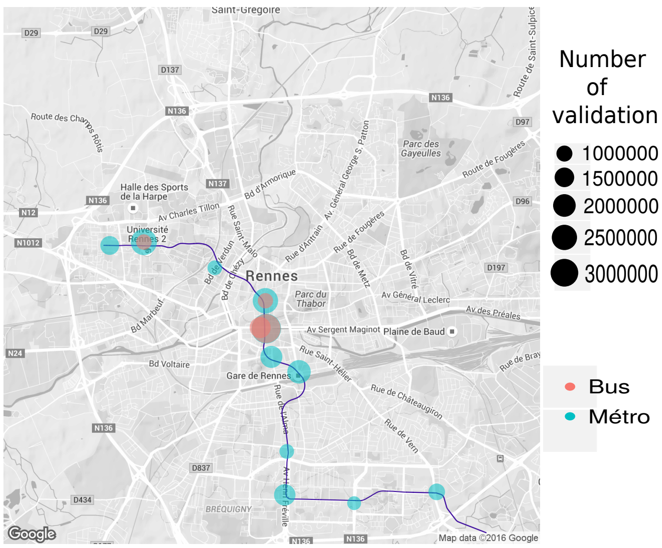

Fig. 1. The 20 most frequented stations in Rennes during the fifteen months period of study. The subway line is depicted in purple. Subway stations are indicated in cyan whereas bus stops are coloured in pink.

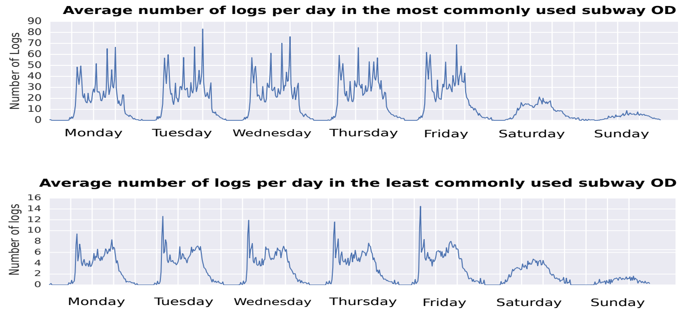

Fig. 2. Average trip counts (over 15min time bins) for the most used (top) and least used (bottom) OD pairs.

B. Data Enrichment and Aggregation Methodology

In order to be able to predict OD matrices, we need to complete the origin and timestamp information (already collected by the AFC system) by inferring the missing alighting locations. To this end, we apply the commonly used methodology proposed in [2], in which two assumptions are made: (i) the alighting location is the nearest station (serviced by the boarded bus or subway) to the next trip's origin station, and (ii) the last destination of a day is the nearest station to the one used for the very first boarding of that same day.

Since we aim to predict OD matrices over 15 minute windows, the enriched data are then aggregated into trip counts per 15 minute time bins for each OD pair. Figure 2 shows the weekly pattern of the most and least used subway OD pairs. Aside from the significant difference in amplitude between the two, we can distinguish between two distinct pattern between weekdays and the weekend. The two curves also present high frequency variations, thus reflecting important fluctuations in demand.

IV. Predicting OD Matrices with LSTM Neural Networks

We discuss the details of our approach to predicting OD matrices in this section.

The last few years, Recurrent Neural Networks (RNN) have been applied in different fields ranging from machine translation to speech recognition and language modeling. Such models are mainly used for to their ability to handle historical sequential data spanning over arbitrarily long periods of time, thus making them effective in predicting time series data. Contrary to traditional neural networks in which inputs (and outputs) are independent of each other, RNNs consider that outputs do depend on previous computations (conducted recurrently on every element of the sequence). To do so, they keep a “memory” of previous calculations under the form of a constantly updated hidden state. The hidden state s_t at time step t is calculated based on the previous state s_{t-1} and the input provided at the current step x_t using a non-linear function f:

$$s_t = f(Ux_t + Ws_{t-1}) \ , \tag{1}$$

where U and W are weight matrices assigned to x_t and s_{t-1} respectively.

The output y_t is computed from this hidden state with a function g:

$$y_t = g(Vs_t) \ , \tag{2}$$

where V is also a weight matrix.

The choice of a suitable function g strongly depends on the type of problem to be solved (classification or regression). Different functions can be used (ReLU, linear, softmax, softplus, etc.). The parameters to fit during the training phase of the RNN are U,V and W.

In practice, RNN suffer from their inability to record information of long time periods, thereby leading us to use a more complex type of RNN, the Long-Short Term Memory (LSTM) developed by Hochreiter and Schmidhuber [29] in 1997. In order to be able to retain information over longer periods of time, LSTM introduces (in addition to the hidden state) a gated memory cell (gates regulate how information propagates) in order to counter the vanishing gradient phenomenon [28] that can occur during the training phase (and which is responsible of simple RNNs not learning correctly).

Our model structure contains one LSTM layer, in which the function f (1) is a hyperbolic tangent (tanh) function and the g function (2) is a softplus function (since the problem to be solved is a regression with output in R_+).

The input attributes of the LSTM models are OD counts of the 300 timestamps preceeding the timestamp we want to forecast. What we refer to as output is the OD matrix of the timestamp that we want to predict.

V. Results and Discussion

Our experimental results are presented and discussed in this section. We evaluate ours LSTM models (i.e with and without buses network information) by comparing their prediction performance with those obtained with two traditional models: calendar model and the Vector Autoregressive (VAR) model, both presented in Section V-A. The procedure by which we train our model and perform model selection is presented in Section V-B. Our main findings are presented in Section V-C.

A. Calendar and Vector Autoregressive Model Methodologies

1) Calendar Model:

For public transport network planification, decision makers often exploit average calendar passenger counts to predict OD matrices. In its basic form, the calendar model just distinguishes between weekdays and weekends. For our experiments, we implemented a more precise calender model that takes into account a richer range of day types. The following types have been considered: Monday, Tuesday and Thursday, Wednesday, Friday, Saturday outside of school holidays (MNH, TTNH, WNH, FNH, SNH), Sunday and days off (SDO), Saturday during school holidays (SH), work days during school holidays (WDH). Depending on the category of the day and time step to predict (e.g. Thursday, June 16, 2016 at 4:00-4:15PM which belongs to category TTNH), the prediction is simply performed by averaging passenger counts for each OD for that same time step (4:00-4:15PM) over all days in the historical data belonging to the same category (TTNH). The implemented model is illustrated in Figure 3.

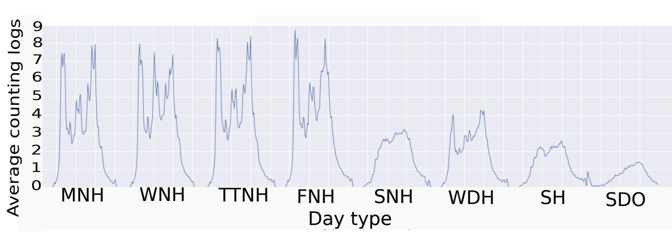

Fig. 3. Illustration of the calendar model. For each of the eight defined day categories, the prediction for a given time step is made by averaging trip counts that occurred for that step during all days belonging to the same category.

2) Vector Autoregressive (VAR) Model:

The vector autoregressive (VAR) model [30] is a generalization of the univariate regressive (AR) model, a time-series regression model allowing to predict a variable by linearly taking account of the previous observation and a stochastic term. The VAR model has the advantage of treating each variable of the structure to predict symmetrically. Therefore, it is capable of capturing the linear inter-dependencies among multiple time-series. Each variable can be predicted by resolving an equation representing its evolution based on its own lag and the lags of the other model variables.

B. Experimental Setting

1) Experimental Protocol:

Our main goal is to predict a future subway OD matrix based on previously observed subway OD matrices. We are also interested in studying the effect of including additional information about alternative public transport modes (buses in this case). Therefore, we also want to make the prediction of the matrix based on observed OD matrices combining both bus and subway. For this second objective, due to the sheer size of the network (15 subway stations and over 600 bus stops) training the model on the entirety of the data would be inefficient (the complete OD matrix contains more than 31000 entries). Consequently, in order to accelerate the training of the models, we decided to only keep the trip counts for the 2000 most frequent bus ODs in addition to the 210 subway ODs for which we aim to predict trip counts.

After OD filtering, we divide the fifteen months' worth of enriched ticketing logs into three datasets as follows. The first year is randomly divided into two datasets: a training set to train the model (80% of the data, which is roughly equivalent to 9.5 months) and a validation set (the equivalent of 2.5 months constituting the remaining 20% of the data) to perform model selection (i.e., set the model's hyperparameters to suitable values). The last three months of data constitute the testing dataset with which we evaluate the models. This scenario mimics a realistic situation in which data up to a given point of time are available to train and tune the predictive model that will be later used to predict future, unobserved periods of time.

We use the Mean Square Error (MSE) to measure and compare the performances of the four models (calendar model, VAR, subway only LSTM, and bus and subway LSTM):

$$\text{MSE} = \frac{1}{T * |OD|} \sum_{t=1}^{T} \sum_{i=1}^{|OD|} (OD_t^i - y_t^i)^2 \ , \tag{3}$$

where y_t^i is the prediction for the i^{th} Origin-Destination at time step t and OD_t^i its expected value, calculated from the data after application of the enrichement process described in Section III-B.

2) LSTM Model Selection:

For the training step, we used a validation-based early stopping [31] method. This method allows to avoid overfitting by stopping the training of the model when the loss of the validation set stops decreasing. We conduct gradient-based optimization using the ADAM [32] optimizer. Since the hyperparameters of this optimizer do not require particular fine-tuning, we focus the optimization of the LSTM on the hidden state's size: we performed a grid search over this parameter in order to find the size that leads to the best results. We do this model selection separately for both the subway-only LSTM and bus and subway LSTM. For the bus and subway LSTM model, the loss decreases when the hidden state's size grows, as can be seen from Figure 4. This is explained by the fact that the LSTM can encrypt more information in a bigger hidden state. The best hidden state size is 3000 according to the figure (minimum loss for the validation set). For the subway-only LSTM, optimality occurs at an earlier stage, with a hidden state size of 900.

Fig. 4. Evolution of the Mean Square Error (MSE) as a function of the size of the hidden state for the LSTM model using bus and subway ODs for both the training (blue) and validation (green) datasets. The validation MSE is minimized when the state size is equal to 3000.

3) VAR Model Selection:

The result of the VAR model prediction depends on the choice of the lag parameter. This parameter represents the number of previous observations that the model is taking into account to make its prediction. We select the best lag parameter based on the best result for the validation dataset, which is achieved when the lag parameter is equal to 10 (cf. Figure 5).

C. Results

An example of results obtained using the LSTM approach is shown in Figure 6. In general, the model is well capable of predicting the trip counts correctly. However, when sudden changes are observed, the prediction is less accurate.

The MSE results obtained by the four models are depicted in Table I. The first noticeable point is that, contrary to the VAR and both LSTM models that as expected achieve better results on the validation set than on the test set, the calendar model achieves better results on the latter. This can be explained by the fact that (i) the validation set exhibits a higher variance (because of the larger period it covers) and

Fig. 5. Evolution of the training MSE (blue) and validation MSE (green) as a function of the lag parameter for the VAR model. The minimal validation loss is observed for lag values around 10.

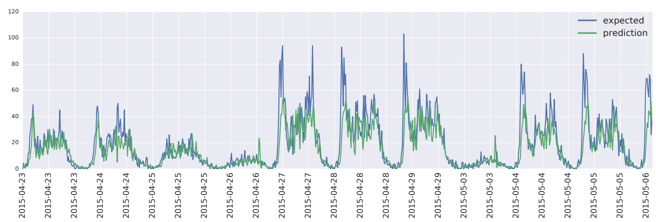

Fig. 6. Example of results obtained using the bus and subway LSTM model for the most frequent OD. Predictions are indicated by the green curve whereas real values are depicted with the blue curve.

that (ii) the OD values in the test set were coincidentally closer to the means that were calculated from the train set.

The best results are achieved by the LSTM models, with the subway and bus LSTM model outperforming the subway only LSTM model. This confirms that providing additional data about complementary transportation modes (bus in this case) helps improve the quality of the prediction for the transportation network of interest (subway). This observation is intuitive since these modes are connected and mutually influence one another.

TABLE I
MSE OF THE DIFFERENT MODELS ON EACH DATASET

	Train	Validation	Test
Calendar	11.6	12.35	8.86
VAR	4.34	5.56	5.88
LSTM (data: only subway)	4.01	4.94	5.00
LSTM (data: subway and bus)	**2.73**	**4.52**	**4.71**

VI. CONCLUSION

In the present work, a machine learning based methodology was proposed to forecast dynamic public transport matrices over a 15 minutes time horizon. Our approach is based on a LSTM recurrent neural network that is able to capture long time lags in the prediction horizon. The effectiveness of the proposed methodology was evaluated using a real smart card dataset issued from the public Transport network of Rennes Métropole in France. The evaluation of the LSTM performances was conducted by comparing it to two traditional approaches that are based on a calendar model or a Autoregregressive (VAR) model. We also investigated the effect of including additional information about the nearby bus transportation system on the forecasting accuracy. The results obtained so far are encouraging: the LSTM model that uses both subway and bus data does indeed improve on all the data set prediction performances compared to those of the Calendar, VAR, and LSTM using only subway information models.

Further investigations have to be carried out to enhance and consolidate these results. Particular attention will be given to reduce the computational time for the training phase of the LSTM, thus allowing for a better optimisation of its setup. It would be interesting to further investigate the best way to incoporate in the forecast model historical data about other contextual variables and external factors (e.g. weather, traffic incidents, etc.). In the future, we plan to extend this methodology to the whole public transport network. Scaling up the prediction methods is indeed challenging and provides a framework for forecasting mobility demand. We would also like to explore the potential of using ensemble learning (i.e., using multiple models) to enhance the prediction of OD matrices.

References

[1] M. Bagchi and P. R. White, "What role for smart card data from bus systems," in *Proceedings of the Institution of Civil Engineers. Municipal Engineer 157, 39–46. March Issue ME1*, 2004.

[2] J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda, "Origin and destination estimation in new york city with automated fare system data," *Transportation Research Record*, vol. 1817, pp. 183–187, 2002.

[3] A. Cui, "Bus passenger origin-destination matrix estimation using automatic data collection systems," Ph.D. dissertation, MIT, USA, 2006.

[4] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Intelligent Transportation Systems*, vol. 11, pp. 1–14, 2007.

[5] W. Wang, J. P. Attanucci, and N. H. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, 2011.

[6] J. Zhao, A. Rahbee, and N. Wilson, "Estimating a rail passenger trip origindestination matrix using automatic data collection systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, pp. 376–387, 2007.

[7] K. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2063, pp. 63–72, 2008.

[8] F. Zhang, N. J. Yuan, Y. Wang, and X. Xie, "Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach," *Knowledge and Information Systems*, pp. 1–25, 2014.

[9] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, 2006, pp. 17–19.

[10] X.-l. Ma, Y.-J. Wu, Y.-h. Wang, F. Chen, and J.-f. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, no. 0, pp. 1 – 12, 2013.

[11] M. Poussevin, N. Baskiotis, V. Guigue, and P. Gallinari, "Mining ticketing logs for usage characterization with nonnegative matrix factorization," in *SenseML 2014–ECML Workshop*, 2014.

[12] M. K. El Mahrsi, E. Côme, J. Baro, and L. Oukhellou, "Understanding passenger patterns in public transit through smart card and socio-economic data," in *3rd International Workshop on Urban Computing (UrbComp), ACM SIGKDD Conference*, New York, USA, August 2014.

[13] A.-S. Briand, E. Come, M. El Mahrsi, and L. Oukhellou, "A mixture model clustering approach for temporal passenger pattern characterization in public transport," in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, Oct 2015, pp. 1–10.

[14] N. Lathia, J. Froehlich, and L. Capra, "Mining public transport usage for personalised intelligent transport systems," in *IEEE International Conference on Data Mining. Sydney, Australia*, 2010.

[15] I. Ceapa, C. Smith, and L. Capra, "Avoiding the crowds: Understanding tube station congestion patterns from trip data," in *Proceeding of the 1st ACM SIGKDD International Workshop on Urban Computing. ACM press*, 2012, pp. 134–141.

[16] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transportation Research Record*, vol. 1971, pp. 119–126, 2006.

[17] S. Foell, S. Phithakkitnukoon, G. Kortuem, M. Veloso, and C. Bento, "Catch me if you can: Predicting mobility patterns of public transport users," in *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, Qingdao, China, October 8 –11 2014.

[18] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[19] X. Jiang, L. Zhang, and X. M. Chen, "Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in china," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 110–127, 2014.

[20] Y. Chen, H. S. Mahmassani, and Z. Hong, "Data mining and pattern matching for dynamic origin–destination demand estimation: Improving online network traffic prediction," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2497, pp. 23–34, 2015.

[21] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 211–234, 2005.

[22] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.

[23] K. Ashok and M. E. Ben-Akiva, "Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows," *Transportation Science*, vol. 36, no. 2, pp. 184–198, 2002.

[24] X. Chen, S. Guo, L. Yu, and B. Hellinga, "Short-term forecasting of transit route od matrix with smart card data," in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Oct 2011, pp. 1513–1518.

[25] E. Van der Hurk, L. G. Kroon, G. Maróti, and P. Vervest, "Dynamic forecast model of time dependent passenger flows for disruption management," in *12th conference on advanced systems for public transport in Santiago, Santiago, Chile*, 2012, pp. 23–27.

[26] F. Qian, G. Hu, and J. Xie, "A recurrent neural network approach to traffic matrix tracking using partial measurements," in *2008 3rd IEEE Conference on Industrial Electronics and Applications*, June 2008, pp. 1640–1643.

[27] A. de Brébisson, É. Simon, A. Auvolat, P. Vincent, and Y. Bengio, "Artificial neural networks applied to taxi destination prediction," *CoRR*, vol. abs/1508.00021, 2015.

[28] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 157–166, 1994.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[30] H. Lütkepohl, *Vector autoregressive models*. Springer, 2011.

[31] L. Prechelt, "Early stoppingbut when?" in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 53–67.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.