LTE Connectivity and Vehicular Traffic Prediction based on Machine Learning Approaches

Christoph Ide*, Fabian Hadiji†, Lars Habel‡, Alejandro Molina†, Thomas Zaksek‡,
Michael Schreckenberg‡, Kristian Kersting† and Christian Wietfeld*

*Communication Networks Institute, TU Dortmund University, Germany
E-Mail:{Christoph.Ide, Christian.Wietfeld}@tu-dortmund.de
†Artificial Intelligence Unit, TU Dortmund University, Germany
E-Mail:{Fabian.Hadiji, Alejandro.Molina, Kristian.Kersting}@tu-dortmund.de
‡Physics of Transport and Traffic, University Duisburg-Essen, Germany
Email: {Lars.Habel, Thomas.Zaksek, Michael.Schreckenberg}@uni-due.de

***Abstract*—The prediction of both, vehicular traffic and communication connectivity are important research topics. In this paper, we propose the usage of innovative machine learning approaches for these objectives. For this purpose, Poisson Dependency Networks (PDNs) are introduced to enhance the prediction quality of vehicular traffic flows. The machine learning model is fitted based on empirical vehicular traffic data. The results show that PDNs enable a significantly better short-term prediction in comparison to a prediction based on the physics of traffic. To combine vehicular traffic with cellular communication networks, a correlation between connectivity indicators and vehicular traffic flow is shown. This relationship is leveraged by means of Poisson regression trees in both directions, and hence, enabling the prediction of both types of network utilization.**

I. Introduction

Road traffic estimation and prediction systems are typically based on the analysis of the physics of traffic (e.g., based on simulations). In order to provide a reliable and detailed prediction, these approaches leverage the data of stationary road traffic detection systems in addition to the collection of Floating Car Data (FCD) from vehicles. Typically this data contains Global Positioning System (GPS) and velocity information from the navigation system or a smartphone in the vehicle (cf. [1] for necessary penetration rates). However, the analysis of physical models for traffic is complex, particularly when the road traffic information is incomplete and data gaps in time as well as space dimension are present. To overcome this problem, we apply innovative machine learning techniques on vehicular data in order to achieve a better vehicular traffic prediction (cf. Fig. 1).

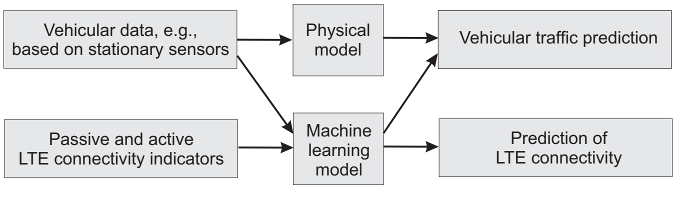
Vehicular data, e.g., based on stationary sensors
Physical model
Vehicular traffic prediction
Passive and active LTE connectivity indicators
Machine learning model
Prediction of LTE connectivity

Fig. 1: Overview of the Methodologies and Objectives.

For the machine learning methods, it is not necessarily obvious how to combine these different types of heterogeneous data and no complex physical model is required. Hence, we propose an approach that is capable of integrating different types of data and learning dependences among them. Furthermore, the learned models can be used to interpret and understand relationships between the different data sources and allow to predict missing or future values. In addition to the data from stationary road traffic detection systems, information from Long Term Evolution (LTE) cellular communication systems about the data network congestion are taken into account for the road traffic estimation. These information are gained decentralized by a single User Equipment (UE). This is the main distinction from famous existing road traffic information systems such as TomTom or Google. Those servers collect massive data and they combine information from plenty devices.

The main contributions of this paper are:

- Bringing together the physics of traffic with data analysis approaches and knowledge of cellular communication systems.
- The interpolation of data gaps by means of innovative data mining approaches based on vehicular traffic data and information of the network congestion of cellular communication systems.
- The long-term prediction based on the physics of traffic achieve a comparable performance in comparison to the data analysis approach. However, the machine learning model achieves significantly better results for a short-term prediction.
- A strong correlation between the decentralized detected connectivity of cellular communication systems and the vehicular traffic congestion on a freeway is verified.
- This correlation can be leveraged in order to predict the cellular connectivity (e.g., for vertical handover or resource-efficient data transmission decisions [2], [3]) based on vehicular traffic information and to enhance a vehicular traffic forecast based on cellular connectivity data.

II. Related Work

Machine learning techniques, and specifically probabilistic graphical models, have been used for traffic prediction before and it is a research topic of great interest in these fields. For example, Piatkowski *et al.* [4] introduce spatio-temporal random fields and apply these models to similar traffic data as we do below. However, we distinguish ourselves from this approach in several ways. Most importantly, we do not discretize the data but instead predict traffic flows directly.

An improved traffic forecast can be achieved by the collection of FCD. The data transmission of vehicular information is one example of Machine-Type Communication (MTC), which can be integrated into public cellular networks [5]. The support of MTC is one of the main goals in the standardization process of LTE-Advanced [6]. In addition to the collection of information like the position and velocity (cf. [7] for an approach of vehicular speed estimation based on passive connectivity indicators) of single vehicles, communication systems can be used in order to get an overview of the vehicular traffic situation. E.g., in [8], a method to detect traffic jams based on the congestion of WiFi networks is proposed. In the context of the correlation between the congestion of cellular communication systems and vehicular traffic congestion, most approaches work centralized by analyzing network operator information. Cellular handover information are used in [9], in order to estimate the vehicular traffic congestion using machine learning methods. In contrast, we use decentralized connectivity indicators of the communication network congestion to estimate the vehicular traffic flow in a certain area.

III. Methodology and Models

The physical traffic simulation, the machine learning model and the measurements of cellular communications systems are described in this section.

A. Physical Traffic Simulation

Vehicular traffic is simulated using the OLSIMv4 framework we presented in [10]. OLSIMv4 comprises of three components: A cellular representation of the real-world highway network to be simulated, a database and interface to incorporate real-time traffic data and a microscopic traffic model describing the movement of individual vehicles.

The distribution service further provides real-time traffic data (i.e., flow, velocity, occupation time, percentage of trucks) from the aforementioned traffic detectors. This data is updated every minute and subsequently stored in a database, so that the simulation can query it at runtime. During the storing process, obviously erroneous data is filtered out. Quite often, not every detector is able to provide data at all. Apart from simple technical issues, this is often a problem around road works. There, detectors often stay offline for a long time. However, reliable data would be very valuable, because of the possible impact of road works on traffic. Therefore, dealing with missing values and their replacement is necessary. Currently, OLSIMv4 uses historical traffic data [11] for the calculation of replacement values. In this paper, we focus on the prediction algorithm from OLSIMv4. We analyze, where machine learning can improve the prediction and generates better input data for the physical simulation.

The prediction model currently used in OLSIMv4 [12] comprises of a classification by day, a moving average and exponential smoothing over multiple days. To predict traffic data for a specific timestamp a heuristic is stepwise processed. At first, assign the day of the timestamp to one of the following classes:

- Monday till Thursday except holidays or days before holidays
- Friday and days before holidays
- Saturday exept holidays
- Sunday and holidays

The days belonging to each of these classes where shown to exhibit similiar traffic behavior [12]. From these classes, the most recent accessible days t (up to 30 days) are selected. The daily timeseries for these days are smoothed each with a moving average with three minute time horizon. As a final step the smoothed timeseries are merged by exponential smoothing:

$$y_t^* = 0.8 y_t + 0.8 \sum_{i=1}^{t-1} (1-0.8)^i y_{t-i} + (1-0.8)^t y_0 \ .$$

y is a vector of vehicular traffic data that contains the flow j, the velocity v, and the occupation p for each traffic detector. y_t^* is the predicted timestamp, y_t is the most recent historical timestamp.

B. Data Mining Model

We now present an alternative approach to modeling and predicting the traffic flow based on probabilistic graphical models. Let us now assume a random variable J_a represents the traffic flow measure by detector a and j_a represents an instantiation of this random variable, i.e., a particular flow count. Hence, the traffic flow for one timestamp can be modeled by a vector of random variables $\mathbf{J} = (J_1, \ldots, J_n)$. Here, n is the number of detectors for which flow counts are available. To learn a model describing this dataset, we use *Poisson Dependency Networks (PDNs)* [13] in which each detector is modeled as a random variable over the natural numbers, i.e., $j_a \in \mathbb{N}$. PDNs are a probabilistic graphical models specifically tailored towards count data. Other approaches such as multinomial or Gaussian random fields are not well suited for the task at hand because the models fail to capture the appropriate nature of the variables. Intuitively, PDNs are Dependency Networks [14] for multivariate Poisson distributions and consist of a local model for each random variable:

$$p(j_a | \mathbf{j}_{\setminus a}) = \frac{\lambda_a^{j_a}(\mathbf{j}_{\setminus a})}{j_a!} e^{-\lambda_a(\mathbf{j}_{\setminus a})} \ .$$

Here, $\lambda_a(\mathbf{j}_{\setminus a})$ is a function which models the mean of the probability distribution of J_a and may depend on all other variables in the PDN except for J_a which is denoted as

$\mathbf{j}_{\setminus a} = \mathbf{j} \setminus \mathbf{j}_a$. Typical choices for modeling λ_a are Generalized Linear Models [15] or Poisson regression trees (PRTs). Assuming a training dataset $[\mathbf{j}^{(k)} \in \mathbb{N}^n]_{k=1,\ldots,m}$ at hand, we fit a model to the data by learning a local model for each J_a separately. More precisely, in our case we use PRTs where a regression tree is learned for each J_a assuming a Poisson distribution of the data corresponding to a leaf. Each leaf of a tree corresponds to a mean for a partition of the data. Inner nodes of the tree represent the splitting criterion for a dependent variable. Typically, the splitting criterion is based on a likelihood ratio test for a possible split. The tree is grown until no further improvement is possible and post-pruning is often applied afterwards to decrease the size of the tree and to avoid overfitting.

Having a PDN at hand, one can envision different applications based on predictions obtained from the PDN. In first place and as described above, traffic detectors might fail or are out of service due to construction sites. In this case, one or several of the J_a in $\mathbf{J}$ are missing. We can fill in the missing datapoints by querying the model for λ_a. The PDN can also be used to make predictions for a specific timestamp in the future, similar to the exponential smoothing described in the previous section. Here, one can assume that we have an observation for a point in time, i.e., $J^{(t)}$, and we want to make a prediction for $J^{(t+1)}$. This is the far more challenging problem because the prediction is now influenced not only by the current traffic but also by the dependencies between the traffic segments.

In many situations, we also want to take additional metadata into account to improve the quality of our predictions. One example for an additional data source can be the signal strength of the mobile radio network. It is likely that a high traffic appearance also leads to more congested mobile networks because of smartphones and other network-dependent accessories in the car. We will the show the correlation between road and network traffic below in more detail (see Sec. IV-B) and highlight how the data traffic can be used to make predictions about vehicular traffic and vice versa.

C. Cellular Communication Connectivity Measurements

Information about the congestion and performance of cellular communication systems are used in order to enhance vehicular traffic forecasts. The analyses of communication systems is possible with a much higher resolution in time and space dimension in contrast to stationary sensors. In this section, the methodology for estimating the congestion of cellular communication systems decentralized in the device is explained. The performance of a single user in a LTE systems is influenced by many factors. This include the coverage as well as capacity of the deployed network, the location of the user, other users in the same or neighboring cells etc. The network provide is aware of the current network congestion is able to analyze the number of active users in a certain area. In contrast to that, in this paper, we measure passive and active connectivity metrics decentralized in the UE by means of an Android app. The network congestion of the current location of an UE can be estimated by passive indicators, e.g., the signal strength that is calculated internal in the UE (we used LG G3), based on the Reference Signal Received Quality (RSRQ) and the Reference Signal Received Power (RSRP). This metric is measured without active connection to the network. However, due to the RSRQ, the signal strength takes the congestion of the LTE network into account. For the evaluation of active indicators, an active connection to the network has to be set up. Then, data can be transmitted and the data transmission time, or rather, the data rate for a fixed payload size are evaluated. The app used in this paper measures both, passive as well as active metrics by sending packets with 100 kByte File Transfer Protocol (FTP) payload and measuring the uplink transmission time. In both cases, the indicators are saved together with the GPS location and a time stamp. In the results section, the correlation of these indicators to the traffic flow of the road network is analyzed.

IV. Results

In this section, first, the comparison between physical traffic simulation and data mining methods to predict the vehicular traffic flow is presented and then, the correlation between vehicular traffic flow and LTE connectivity is illustrated.

A. Comparing Physical Traffic Simulation and Data Mining

In the following, we compare vehicular traffic flow predictions from the historical data smoothing described in Sec. III-A with the predictions made by the PDN model from Sec. III-B. The required empirical vehicular traffic data was obtained from 38 stationary detectors located at the 50 km long Cologne orbital freeway in Germany. This area is also part of the traffic information system autobahn.NRW.de, which is driven by OLSIMv4.

Fig. 2 shows a collection of results of both prediction approaches using empirical data measurements from Tuesday, 26 August 2014. As we compare about 54000 data points, the data is visualized using a density colorization of a set of 50x50 bins. In the upper row of Fig. 2, predictions made

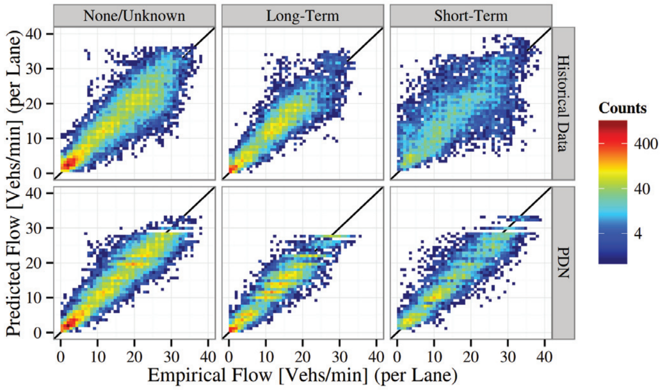

Fig. 2: Comparison between Empirical and Predicted Vehicular Traffic Flows (per Lane) on Tuesday, 26 August 2014, Using the Historical Data Approach and the PDN Model; Characterized by Event Type.

from historical data are compared with their corresponding empirical counterparts. In the lower row of Fig. 2, the prediction of the machine learning model in a leave-one-out cross validation for the task of filling in missing datapoints is shown. There, we removed one row $\mathbf{X}^j$ from the training dataset and learned a Poisson regression tree based PDN on the remaining $m-1$ instances. Then, we predicted each X_i^j assuming that all other $X_{\setminus j}^i$ are observed. We used the PRT implementation of rpart [16] for learning and predictions.

To visualize the motivation of the present work, we have split the dataset into three parts using additional meta information about traffic-related events. Therefore, we accessed our database, which stores date, location as well as reason of road works and other traffic incidents that happened at the aforementioned date based on messages from freeway maintenance authorities as well as police messages. We matched this information to each datapoint, dividing between long-term events like road works, short-term events like accidents, and a third category containing the non-assignable values. In Fig. 2, these categories are represented by the three columns.

One can see that both prediction approaches provide quite similar results when comparing them in the long-term and none/unknown event categories, although the historical data smoothing has a broader range of outliers. On the other hand, the PDN approach tends to stick to certain values instead of reproducing the natural traffic noise. This is due to the fact that each traffic segment is represented as a Poisson regression tree where each leaf corresponds to a mean value of the partitioned data. Hence, the highest leaf value represents an upper bound if no sampling technique is applied. During short-term traffic events, the situation changes and the historical data prediction is not useful anymore, while the PDNs are capable of predicting the missing datapoints even then quite well.

Comparing the physical model with the machine learning based model, one might draw the conclusion that the PDNs are superior. However, one should be careful with such an argumentation as the two models do the prediction from a different point of view. For a full comparison, one should also learn a PDN for the temporal dimension. It is interesting to see though that both models tend to over estimate the true traffic in the region of 35-40 true observations. Nevertheless, the results look promising and motivate to continue further PDN based models for traffic prediction that can then be combined with a physical simulation. For example, the prediction out the PDN can be used an input to physical simulation.

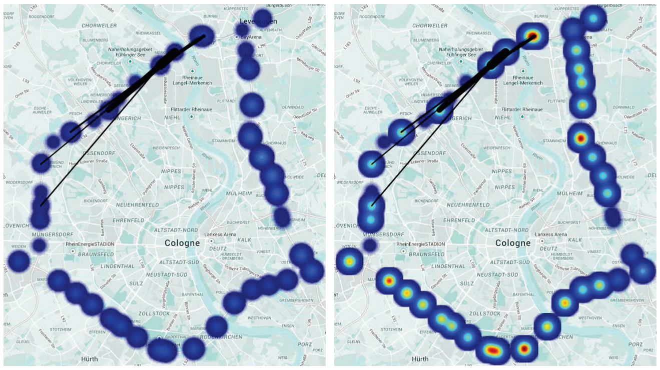

Fig. 3: Map of Average Traffic (Dark Red Segments Indicate a High Traffic) on the Highway at 1am (left) and 5pm (right) with Dependences of the Local Model Representing One Detector in the North by Black Edges. © Google.

To give an intuition how the PDN predicts the missing values, we have plotted the average traffic amount in our dataset at 1am and 5pm in Fig. 3. Additionally, we have added the dependences for the local model representing an interesting traffic detector in the north. The dependences are depicted by the black undirected edges. One can see, the traffic at this location does not only depend on neighboring traffic segments nearby, but is also influenced by segments further away. One should note that the north part of the highway is currently affected by road works which probably enforces the influence of direct neighbors. However, the influence of traffic segments in the west is also plausible. For example, cars traveling from west to east typically have to pass the bridge in the north. Hence, high road traffic flows in the west will often also increase the flows in the north.

B. Correlation between Cellular Connectivity and Vehicular Traffic Congestion

As described in Section III-C, the number of other users in the communication network has a significant influence on the performance of the communication system. Hence, the number of other users in surrounding areas of a device can be approximated by means of connectivity indicators (e.g., the transmission time for a fixed payload size). For UEs on the road system (e.g., a freeway), the number of other vehicles dominate the communication network load due to human triggered and device triggered communication traffic cause by the smartphones of passengers. Hence, the vehicular traffic flow can be approximated, if the road network congestion is known at the road segment. Fig. 4 shows the scatterplot of LTE connectivity indicators and the traffic density of the road system for a freeway in Dortmund, Germany (GPS position: 7.3925 E, 51.4804 N). It can be seen from the figure that a strong correlation between the communication network

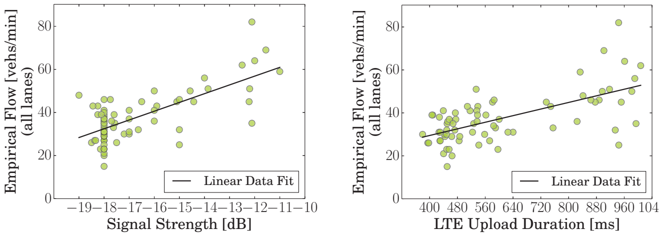

Fig. 4: Correlation between Vehicular Traffic Flow (Sum over all Lanes) and LTE Communication Connectivity. (left) Signal Strength ($R^2 = 0.50$) as Passive Indicator and (right) Transmission Duration ($R^2 = 0.41$) as Active Indicator. Each Data Point Averages Over One Minute.

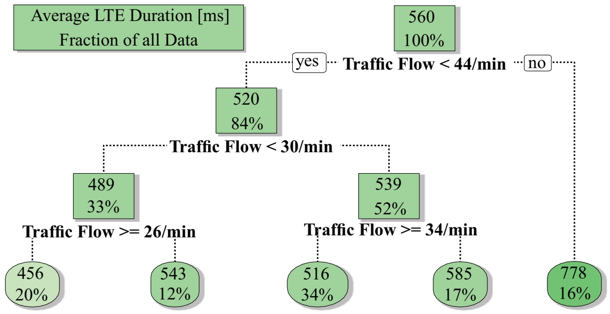

Fig. 5: **Prediction of LTE Communication Connectivity** by means of a Poisson Regression Tree, Solely learned on Current Vehicular Traffic Flow (cf. Fig. 4).

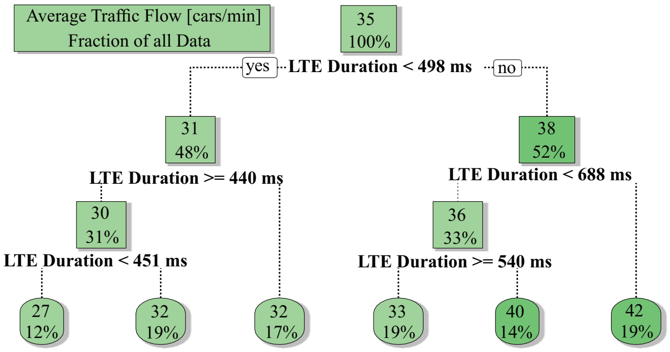

Fig. 6: **Prediction of Vehicular Traffic Flow** by means Poisson Regression Tree, Solely learned from Current LTE Communication Connectivity (cf. Fig. 4).

congestion and road network congestion can be observed for the example measurement locations. This correlation can be leveraged in both directions: A estimation of the vehicular traffic by means of LTE information and a approximation of the current LTE connectivity based on the road network congestion is suitable (cf. Fig. 1).

We built Poisson regression trees for both learning talks. Fig. 5 shows the regression tree for the prediction of the communication connectivity, solely based on vehicular traffic flow information. Such a prediction can be used, for example, in order to improve vertical handover procedures or to optimize transmission decisions for vehicular Machine-Type Communication (MTC) (cf. [2], [3]). In addition, the traffic prediction can be learned based on LTE connectivity information (cf. Fig. 6 for the regression tree). Both trees are learned based on the data presented in Fig. 4.

V. Conclusion and Future Work

In this paper, we have applied machine learning methods for research questions of both, LTE communication connectivity and vehicular traffic prediction. We have shown that a more reliable vehicular traffic prediction can be enabled based on machine learning. In addition, measurement data about the cellular connectivity (captured in the UE) and data from vehicular traffic congestion have been correlated. It has been shown that Poisson regression trees are suitable solutions for the prediction of communication connectivity based on road network information and vice versa. However, we currently learn separate models for each objective. In the future, we will investigate a joint model for predicting both types of data and do not restrict our model to Poisson distributions only.

Acknowledgment

Part of the work on this paper has been supported by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project B4 "Analysis and Communication for the Dynamic Traffic Prognosis", as well as KE 1686/2-1 as part of the SPP1527 "Automous Learning". The NRW ministry MBWSV and Deutsche Forschungsgemeinschaft (DFG, SCHR 527/5-1) have supported OLSIMv4.

References

[1] C. Ide, B. Niehoefer, T. Knaup, D. Weber, C. Wietfeld, L. Habel, and M. Schreckenberg, "Efficient Floating Car Data Transmission via LTE for Travel Time Estimation of Vehicles," in *IEEE Vehicular Technology Conference (VTC-Fall)*, Quebec City, Canada, Sep. 2012.

[2] C. Ide, L. Habel, T. Knaup, M. Schreckenberg, and C. Wietfeld, "Interaction between Machine-Type Communication and H2H LTE Traffic in Vehicular Environments," in *Proc. of the IEEE Vehicular Technology Conference (VTC-Spring)*. Seoul, Korea: IEEE, May 2014.

[3] C. Ide, B. Dusza, and C. Wietfeld, "Client-based Control of the Interdependence between LTE MTC and Human Data Traffic in Vehicular Environments," *IEEE Transactions on Vehicular Technology, in press*, vol. PP, no. 99, pp. 1–1, 2014.

[4] N. Piatkowski, S. Lee, and K. Morik, "Spatio-temporal random fields: compressible representation and distributed estimation," *Machine learning*, vol. 93, no. 1, pp. 115–139, 2013.

[5] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: a survey," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, 2013.

[6] "3GPP TR 23.888 - System Improvements for Machine-Type Communications," 3rd Generation Partnership Project, version 1.2. [Online]. Available: http://www.3gpp.org

[7] G. Chandrasekaran, T. Vu, A. Varshavsky, M. Gruteser, R. P. Martin, J. Yang, and Y. Chen, "Vehicular Speed Estimation Using Received Signal Strength from Mobile Phones," in *In Proc. of the 12th ACM International Conference on Ubiquitous Computing*, ser. UbiComp '10. New York, NY, USA: ACM, 2010, pp. 237–240.

[8] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "VTrack: Accurate, Energy-aware Road Traffic Delay Estimation Using Mobile Phones," in *In Proc. of the 7th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '09. New York, NY, USA: ACM, 2009, pp. 85–98.

[9] W. Hongsakham, W. Pattara-Atikom, and R. Peachavanish, "Estimating road traffic congestion from cellular handoff information using cell-based neural networks and K-means clustering," in *In Proc. of the ECTI-CON*, vol. 1, Krabi, Thailand, May 2008, pp. 13–16.

[10] J. Brügmann, M. Schreckenberg, and W. Luther, "Real-Time Traffic Information System Using Microscopic Traffic Simulation," in *8th EUROSIM Congress on Modelling and Simulation*, Cardiff, Wales, Sep. 2013, pp. 448–453.

[11] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg, "Three categories of traffic data: Historical, current, and predictive," in *Proc. of 9th IFAC Symposium Control in Transportation Systems*. Pergamon, Mar. 2001, pp. 250–255.

[12] ——, "Different methods of traffic forecast based on real data," *Eur. J. Oper. Res.*, vol. 155, no. 3, pp. 558–568, jun 2004.

[13] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting, "Poisson dependency networks: Gradient boosted models for multivariate count data," *Machine learning Journal (minor revisions)*, 2015.

[14] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie, "Dependency Networks for Density Estimation, Collaborative Filtering, and Data Visualization," *Journal of Machine Learning Research*, vol. 1, pp. 49–76, 2000.

[15] P. McCullagh and J. Nelder, *Generalized Linear Models*. Chapman and Hall, 1989.

[16] T. M. Therneau, B. Atkinson, and B. Ripley, *rpart: Recursive Partitioning*, 2011.