Traffic Simulations with Empirical Data: How to Replace Missing Traffic Flows?

Lars Habel, Alejandro Molina, Thomas Zaksek, Kristian Kersting and Michael Schreckenberg

Abstract For the real-time microscopic simulation of traffic on a real-world road network, a continuous input stream of empirical data from different locations is usually needed to achieve good results. Traffic flows for example are needed to properly simulate the influence of slip roads and motorway exits. However, quality and reliability of empirical traffic data is sometimes a problem for example because of damaged detectors, transmission errors or simply lane diversions at road works. In this contribution, we attempt to close those data gaps of missing traffic flows with processed historical traffic data. Therefore, we compare a temporal approach based on exponential smoothing with a data-driven approach based on Poisson Dependency Networks.

1 Introduction

Microscopic road traffic simulations based on a real-world topology usually need many preparations to deliver reliable results. At first, a promising simulation model has to be chosen and the topology has to be converted into a model-friendly representation. When the simulation shall use traffic data from real-world detectors, they

L. Habel (✉) · T. Zaksek · M. Schreckenberg
Physik von Transport und Verkehr, Universität Duisburg-Essen,
47057 Duisburg, Germany
e-mail: lars.habel@uni-due.de

T. Zaksek
e-mail: thomas.zaksek@uni-due.de

M. Schreckenberg
e-mail: michael.schreckenberg@uni-due.de

A. Molina · K. Kersting
Fakultät für Informatik, Technische Universität Dortmund,
LS VIII, 44221 Dortmund, Germany
e-mail: alejandro.molina@tu-dortmund.de

K. Kersting
e-mail: kristian.kersting@cs.tu-dortmund.de

V.L. Knoop and W. Daamen (eds.), *Traffic and Granular Flow '15*,
DOI 10.1007/978-3-319-33482-0_62

and the belonging areas to fill in (or remove) vehicles according to the data have to be implemented as well. For complex topologies, this means that the simulation results then not only depend on the quality of the model and the topology representation, but also on a possibly huge number of empirical traffic detectors.

Usually, empirical traffic detectors provide new traffic flow data every minute. This data is then used in the simulation to reproduce all the recent traffic in- and outflows of the real-world system. Therefore, the permanent availability of empirical data is necessary especially at on- and off-ramps. Unfortunately, the reliability of empirical detectors is often not good enough to ensure this requirement minute by minute. This contribution provides a comparison of two approaches to close the resulting gaps in empirical data, one working on temporal level, the other one on level of dependencies between multiple detectors.

2 Methods

In case of missing data, a decision has to be made how the simulation shall handle this issue. Principally, different strategies are possible: Some detectors are redundant, so the missing data simply could be ignored because the coverage of neighbouring detectors is sufficient enough. However, it is often difficult to decide whether a detector is important or not. The importance of such a redundant detector can also rise when neighbouring detectors go off-line. Additionally, each detector often has a complex neighbourhood, which is sometimes not fully known because the given location data is lacking precision or is outdated. The same issues can also occur on temporal level because it is usually unknown how long a detector will be off-line. This is especially problematic when the simulation is used in a real-time context, i.e. in a traffic information system [1], and thus new empirical data is queried by the simulation at run time.

The described problem is somewhat related to short-term traffic forecasting methods and to interpolation methods for incomplete time series in general. These topics have already been addressed by numerous approaches (see e.g. [8] for a summary). However, these are often complicated to understand or to apply, or they also need e.g. a complete set of historical data or a working detector neighbourhood. As described, these preconditions are often not met. Because of this, we focus on two simple and resilient methods for filling these gaps in real-time.

2.1 Exponential Method

The temporal approach [2] is based on exponential smoothing a set $\mathbf{j}$ of historical traffic flows. $\mathbf{j}$ comprises previously collected traffic flows from up to 30 timestamps t measured at the particular detector, which are chosen by a clustering algorithm that distinguishes between different weekdays, school holidays and public holidays. The predicted flow j_t^* is then obtained by

$$j_t^* = \alpha j_t + \alpha \sum_{i=1}^{t-1} (1-\alpha)^i j_{t-i} + (1-\alpha)^t j_0 \; , \tag{1}$$

where j_t is the most recent historical traffic flow. We use $\alpha = 0.8$ for long-term gap filling [3].

2.2 Poisson Dependency Network (PDN)

For the dependency-based gap filling, we use the recently proposed Poisson Dependency Networks [4]. Dependency networks are graphical models, meaning that each graph node represents a single detector and each edge between nodes describes dependencies between them. Note that neighbouring detectors on the road do not have to be strongly connected in the PDN.

Here, the set $\mathbf{j}$ comprises traffic flows from other detectors, but measured at the same time. The probability function to obtain a traffic flow for detector a given all the other flows $\mathbf{j}_{\setminus a} = \mathbf{j} \setminus \mathbf{j}_a$ at that time is then denoted as

$$p(j_a|\mathbf{j}_{\setminus a}) = \frac{\lambda_a^{j_a}(\mathbf{j}_{\setminus a})}{j_a!} e^{-\lambda_a(\mathbf{j}_{\setminus a})} \; , \tag{2}$$

where $\lambda_a(\mathbf{j}_{\setminus a})$ is a function which contains all knowledge about correlations between detector a and the others. In this contribution, each λ is modelled by Poisson regression trees which have been learned by the R-package `rpart`.

3 Comparison Setup

For the comparison, we use empirical traffic data from the Cologne orbital motorway network in Germany, which is formed by the motorways A1, A3 and A4 and is about 100 km long. Traffic data is provided by 187 detectors at 95 cross-sections.

Both approaches use historical traffic data in a certain sense. In the Exponential Method, there is usually a window of 1 week between each time stamp t, because Eq. (1) is not suitable for intra-day traffic forecasting, as it does not take the intraday shape of a traffic flow time series into account. For the calculation of Eq. (1), historical timestamps with missing values and timestamps of different classes (i.e. holidays, see Sect. 2) have to be removed, meaning that the effective window between two timestamps is sometimes more than 1 week.

To create a level playing field for the comparison, the PDN approach uses data from the preceding week as well to learn $\lambda_a(\mathbf{j}_{\setminus a})$, see Fig. 1 for a graphical explanation. A sufficient number of timestamps has to be in the training set to reflect the correlations, so we decided to use a 60-min window from that week. Note that in contrast to the Exponential Method, the PDN gets data from n detectors. The set of all 187 detectors is resized dynamically to n for each prediction, because detectors without passed

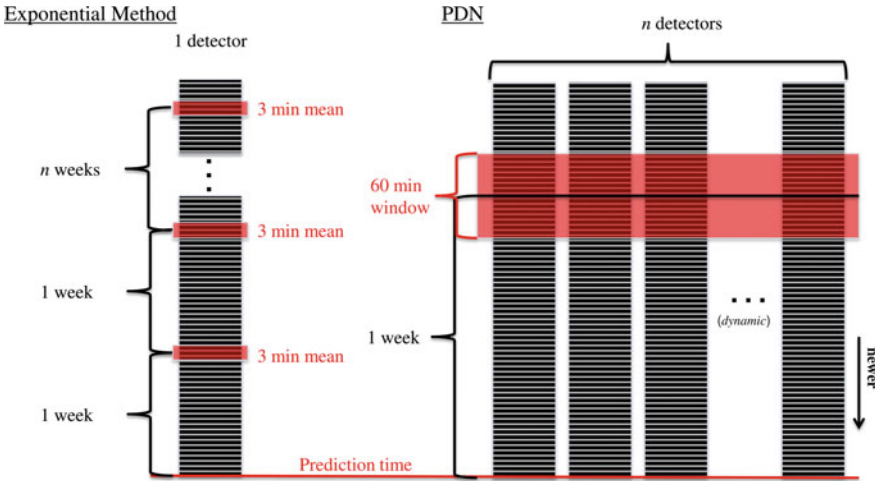

Fig. 1 Visualisation of different data sets used for the predictions. Each column represents a time series of a single detector. Highlighted rows are used for the prediction at the next new time stamp

traffic (i.e. with $\mathrm{var}(\mathbf{j}_a) = 0$) during the 60-min window have to be excluded. Also, missing values have to be removed. This can be done row-wise by removing the whole time stamp, but usually a huge number of missing values is produced by a small subset of detectors. We have excluded them column-wise first, if more than 5 % of their values were missing.

We used traffic data from 21/09/2015 to 27/09/2015 to test the predictions and data from the preceding week to train the PDN. The exponential method got data from the preceding week and up to 30 weeks before.[1]

4 Results

To compare the accuracy of both strategies, we calculated the root-mean-square error between all N predicted traffic flows $\mathbf{P}$ and observed traffic flows $\mathbf{O}$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{3}$$

and its normalised variant

$$\mathrm{NRMSE} = 100\frac{\mathrm{RMSE}}{\mathrm{sd}(\mathbf{O})}, \tag{4}$$

where $\mathrm{sd}(\mathbf{O})$ is the standard deviation of all observations.

The overall prediction accuracies are shown in Table 1. There, every time stamp of the whole test week is included. It is obvious that both prediction methods are

[1] These values had been calculated in advance as a part of OLSIM [1].

Table 1 Overall prediction accuracy

Method	RMSE (vehs/min)	NRMSE (%)
Exponential	4.93	53.2
PDN	**4.50**	**48.6**

bold numbers show better results

not faultless, because both of them basically perform a 1-week-prediction of 1-min count data.

For a deeper analysis, it is typically better to use a subset of data from working days only. We also categorised the data by time interval[2] and created a spatial visualisation, which is shown in Fig. 2, by using the R-package ggmap [6].

The visualisation reveals that prediction problems typically are bound to topological problems: The north-eastern part of the network around the motorway junction between A1 (connects north-east and west) and A3 (connects northeast and south-east) was affected by several construction sites at time of this analysis. On the A1 a speed limit of 60 km/h had to be implemented because of repairs on a damaged bridge, also trucks were not allowed to pass that bridge. Parallel, works on the A3 started to upgrade the road cross-section from three to four lanes per direction. These required temporally closed lanes and driving on the hard shoulder. Hence, these sites and the related upstream road sections were heavily affected by congestion because of their huge bottleneck impact. They can be identified in Fig. 2 by the size of the dots, which denote the mean empirical velocity at test time.

The colours of the dots in Fig. 2 show the mean differences between predicted and observed traffic flows per lane, meaning that negative differences indicate an ongoing underestimation of flow by the prediction. As can be seen, the Exponential Method underestimated traffic on the A3 inside and upstream of the bottleneck heavily. One reason for this are the temporal lane closures. Then, the distribution of vehicles on the remaining lanes changes in contrast to the preceding weeks and thus the exponential predictions become incorrect. The predictions from the PDN clearly benefit from the learned correlations, although it was trained with historical data as well. However, the PDN sometimes overcompensated the lane closure, albeit an overestimation of traffic flow is usually less of a problem than underestimation: Traffic breakdowns happen at high traffic flows and with an underestimation, a potentially unstable traffic situation would be missed by the simulation. Also, common microscopic traffic models tend to underestimate the spatial extent of congestion [7], so that a slight overestimation usually will not harm the simulation results.

One has to note that although the mean differences are often very low outside the hotspot areas, the RMSE is usually higher, because positive and negative differences balance each other out from minute to minute. This is shown in Fig. 3, which also shows that the (N)RMSE rises heavily inside a jam. However, the PDN is always a bit more accurate.

Most of the time, time series of observations and both predictions have a quite similar shape. Figures 4 and 5 show typical examples of special situations, where

[2] Time intervals (in local time): morning 05:00–09:59, midday 10:00–13:59, afternoon 14:00–17:59, evening 18:00–21:59, night 22:00–04:59.

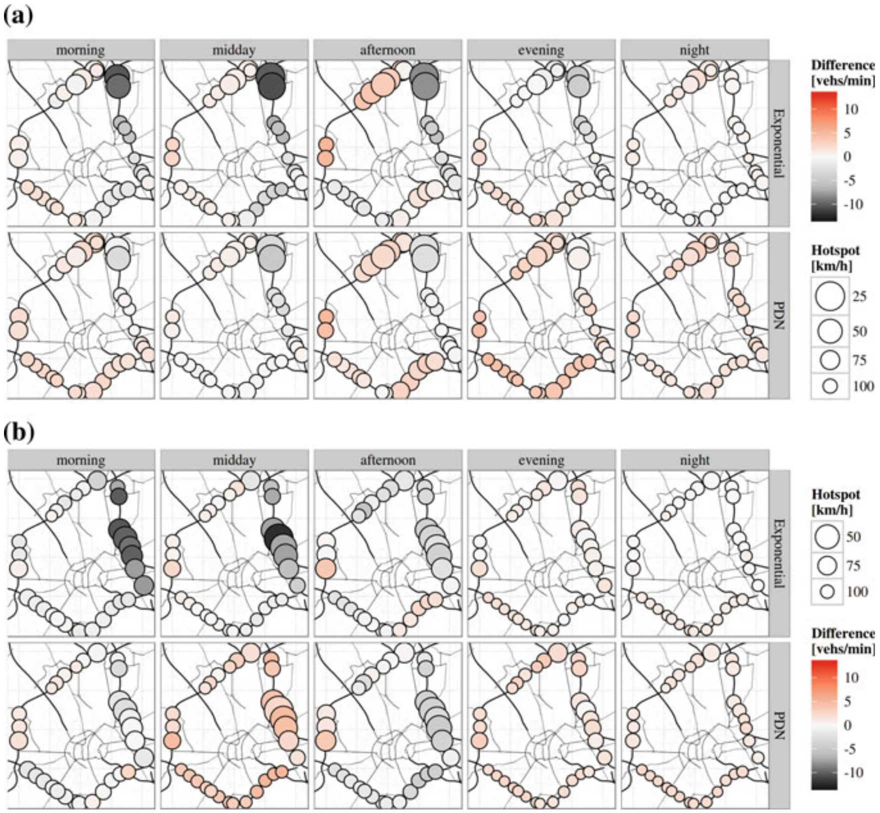

Fig. 2 Spatial visualisation of differences between predicted and observed traffic flows on working days by time interval. Each dot represents a detector cross-section in clockwise (**a**) or anti-clockwise driving direction (**b**). The size of each dot shows the mean empirical velocity measured in the corresponding time interval during the test week. *First row* Exponential Method, *Second row* PDN

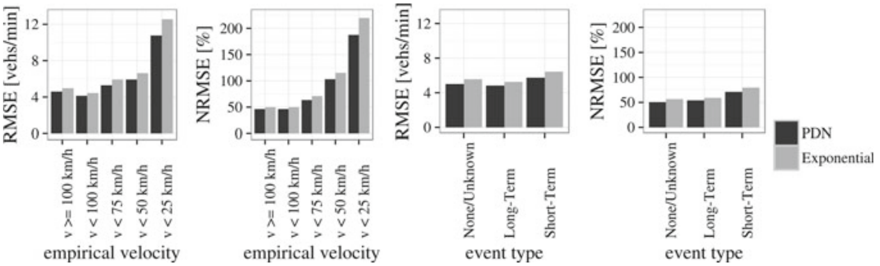

Fig. 3 Prediction accuracy in congested traffic on working days in anti-clockwise direction. Results are divided into different classes of empirical velocities and TMC event type at test time

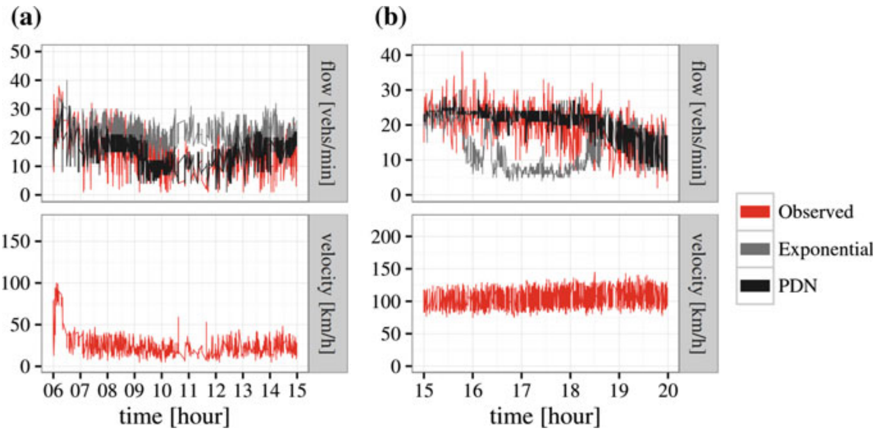

Fig. 4 Different time series with diverted traffic: less traffic than usual on the northbound A3. Some data points are missing because of missing observed data (**a**); more traffic than usual on the eastbound A4 avoiding the jammed north-eastbound A1 (**b**)

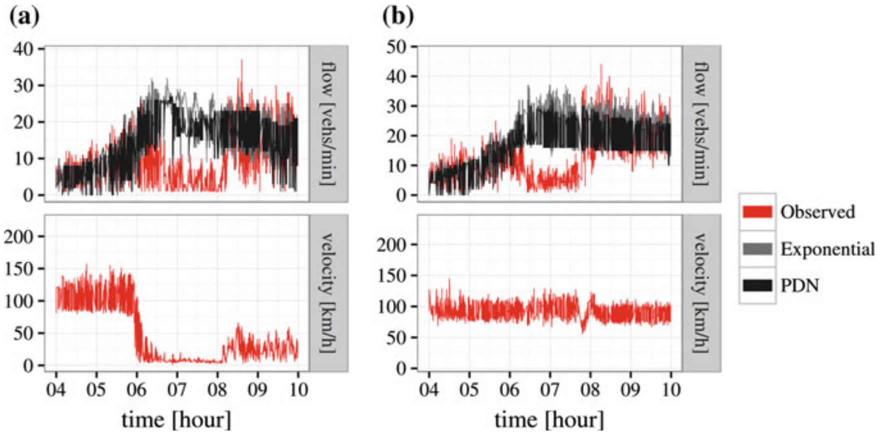

Fig. 5 Time series with an accident on the A3 in northbound direction: upstream of the bottleneck (**a**); downstream of the bottleneck (**b**)

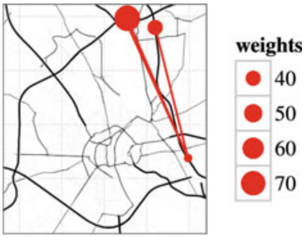

Fig. 6 Established links and the corresponding weights in the PDN for a northbound detector on the motorway A3, located in the south-eastern corner of the map. The hotspot area determines the upstream traffic conditions

the predictions differ. When traffic uses different ways than usual (see Fig. 4), the PDN approach has a clear advantage, because it implicitly detects the necessary dependencies to deal with the situation. However, it is not able to avoid any misprediction. Figure 5 shows time series with a big accident during the rush hours, where prediction accuracy of traffic flow was not only affected upstream of the incident, but also downstream, because the flow was drastically reduced by the accident. This event was not foreseeable for both methods.

5 Conclusion

In this contribution, we have analysed the prediction accuracy of Poisson Dependency Networks in the context of traffic simulations in comparison to an older approach. It is interesting to see how good the PDN performed, given the fact that we did not implement the detector network topology explicitly. Further improvements are planned, namely using a more flexible window of training data as well as trying out different strategies for learning $\lambda_a(\mathbf{j}_{\setminus a})$.

Also, a detailed analysis of the graphical structure of the PDN seems to be promising. An example for the correlations the PDN is revealing is shown in Fig. 6. With the described training window, the PDN only established up to 4 edges per node. When used with more learning data it uses more edges, even to the other side of the ring, as we showed in [5] in a different context.

Acknowledgements This contribution has been supported by 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’.

References

1. Brügmann, J., Schreckenberg, M., Luther, W.: Real-time traffic information system using microscopic traffic simulation. In: Al Begain, K., Al Dabass, D., Orsoni, A., Cant, R., Zobel, R. (eds.) EUROSIM 2013—8th EUROSIM Congress on Modelling and Simulation, pp. 448–453. EUROSIM, IEEE, Cardiff, Wales (2013)
2. Chrobok, R., Kaumann, O., Wahle, J., Schreckenberg, M.: Three categories of traffic data: historical, current, and predictive. In: Schnieder, E., Becker, U. (eds.) Proceedings of 9th IFAC Symposium Control in Transportation Systems, pp. 250–255. Pergamon (2001)
3. Chrobok, R., Kaumann, O., Wahle, J., Schreckenberg, M.: Different methods of traffic forecast based on real data. Eur. J. Oper. Res. **155**(3), 558–568 (2004)
4. Hadiji, F., Molina, A., Natarajan, S., Kersting, K.: Poisson dependency networks: gradient boosted models for multivariate count data. Mach. Learn. **100**(2–3), 477–507 (2015)
5. Ide, C., Hadiji, F., Habel, L., Molina, A., Zaksek, T., Schreckenberg, M., Kersting, K., Wietfeld, C.: LTE Connectivity and vehicular traffic prediction based on machine learning approaches. In: 2015 IEEE Vehicular Technology Conference (VTC Fall) (2015)
6. Kahle, D., Wickham, H.: ggmap: Spatial visualization with ggplot2. R J. **5**(1), 144–161 (2013)
7. Knorr, F., Schreckenberg, M.: On the reproducibility of spatiotemporal traffic dynamics with microscopic traffic models. J. Stat. Mech. Theory Exp. **2012**(10), P10018 (2012)
8. Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C.: Short-term traffic forecasting: where we are and where we’re going. Transp. Res. Part C: Emerg. Technol. **43**, 3–19 (2014)