Spatiotemporal structure of urban traffic flows: from road network to data-driven feature selection

Dmitry Pavlyuk[1]

[1] Transport and Telecommunication Institute, Riga, Latvia
Dmitry.Pavlyuk@tsi.lv

Accurate incorporation of a spatiotemporal structure of urban traffic flows into predictive schemes is an emerging problem in the field of intelligent transportation systems (ITS) (Vlahogianni et al., 2014). Utilization of the temporal dimension of dependencies has an extensive theoretical background and accepted in the majority of modern traffic forecasting models, but the interest to the spatial dimension is rapidly growing. Existence of spatial dependencies between traffic flows at connected road network links is perfectly supported by the macroscopic traffic flow theory, but capturing of these dependencies for modelling and forecasting is a challenging task. Taking the similarity of traffic flows and fluid streams, many studies utilize a straightforward definition of spatial dependencies: a vehicle that observed at an upstream point will be later observed at a downstream point. Under this definition a structure of the road network (connectivity and distances of between road segments) is considered as a complete source of information about spatial dependencies. This structure is usually included into forecasting models (such as the space-time autoregressive integrated moving average model, STARIMA) via a static matrix of spatial weights.

Although the road network-based approach demonstrates a good forecasting performance for linear traffic flows at arterial roads, it doesn’t explain spatial phenomena in a highly interconnected structure of urban roads. In such complex environment spatial dependencies of traffic flows appear not only for adjacent, but also for remote road segments. These remote dependencies (teleconnections) occur due to many reasons. One of the reasons is common traffic patterns that affect remote road segments: for example, simultaneous traffic flows from different directions to the city centre every morning or to a football stadium at match days. Another reason is based on high availability of road information for drivers via navigation software: congestion at a road segment forces informed drivers to choose alternative ways and leads to causal spatial dependencies between remote road links. The road-network based approach is unable to capture these spatial teleconnections and thus recently they have been criticized in literature in a favour of data-driven approaches (Ermagun and Levinson, 2016; Li et al., 2017). The problem of identification of spatiotemporal dependencies between road links can be considered as a special case of the feature selection problem.

Feature selection is a machine learning technique of selecting of relevant features (predictors) for model specification. Methods of feature selection are commonly classified to filter methods (selection of features is executed before model specification), wrapper methods (selection of features utilises the model performance as an objective function), and embedded methods (selection of features is incorporated into the model estimation process) (Chandrashekar and Sahin, 2014).

In this research we concentrate on analysis of spatial specification of vector autoregressive models (VAR) in respect of different feature selection methods. VAR models are widely used to capture the linear interdependencies among multiple time series. Our choice of the VAR model is explained by its good forecasting performance, many empirical applications to traffic flows in recent publications, flexible identification of spatial and temporal dependencies, and availability of direct spatial structure extraction from estimation results. Let Y_t is a $k \times 1$ vector $(y_{1,t}, y_{2,t}, \dots, y_{k,t})'$, where $y_{i,t}$ is a value at a time point *t* and a spatial location *i*, then VAR(p) model is presented as:

$$Y_t = \sum_{\mathrm{h}=1}^{\mathrm{p}} \Phi_h Y_{t-\mathrm{h}} + \varepsilon_t$$

where Φ_h is a set of $k \times k$ matrices of unknown coefficients for every lag $h = 1, \dots, p$ that represent spatial and temporal dependencies; ε_t is a $k \times 1$ vector of i.i.d. disturbances. Feature selection for VAR models corresponds to increasing the sparsity of the Φ_h matrixes.

We utilize the following set of models:

1. Unrestricted VAR model as a baseline for benchmarking feature selection approaches
2. Spatially regularized VAR model (Schimbinschi et al., 2017), where spatial dependencies are allowed only between road links that are connected in respect of travel time between them and allowed traffic speed. This model represents the filter feature selection approach.
3. Genetically optimized sparse VAR model, where a selected set of features is defined using a genetic search algorithm (the wrapper feature selection approach).
4. Adaptive LASSO regularization of VAR model (Kamarianakis et al., 2012). This model represents the embedded feature selection approach.

The selected model specifications are estimated for real world sensor-based traffic flow data and tested for their short-term forecasting accuracy and sparsity (complexity of spatiotemporal dependencies).

Special attention is paid to an effect the size of analysed urban network. Modern ITS provide traffic flow information for thousand sensors with high temporal resolution, which is challenging for statistical modelling. Gradually increasing the spatial dimension we discovered its effect on forecasting accuracy and feature selection ability of the analysed models.

The main scientific value of the research lies in direct empirical comparison of different feature selection approaches to the statistical traffic flow forecasting model and discovered bounds of their applicability for city-wide urban networks.

Acknowledgements

This work was financially supported by the post-doctoral research aid programme of the Republic of Latvia (project no. 1.1.1.2/VIAA/1/16/112, “Spatiotemporal urban traffic modelling using big data”), funded by the European Regional Development Fund.

References

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Ermagun, A., Levinson, D.M., 2016. Spatiotemporal Traffic Forecasting: Review and Proposed Directions, in: TRB 96th Annual Meeting Compendium of Papers. Presented at the 96th Annual Transportation Research Board Meeting, USA, p. 29.

Kamarianakis, Y., Shen, W., Wynter, L., 2012. Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. Applied Stochastic Models in Business and Industry 28, 297–315. https://doi.org/10.1002/asmb.1937

Li, Z., Jiang, S., Li, L., Li, Y., 2017. Building sparse models for traffic flow prediction: an empirical comparison between statistical heuristics and geometric heuristics for Bayesian network approaches. Transportmetrica B: Transport Dynamics 1–17. https://doi.org/10.1080/21680566.2017.1354737

Schimbinschi, F., Moreira-Matias, L., Nguyen, V.X., Bailey, J., 2017. Topology-regularized universal vector autoregression for traffic forecasting in large urban areas. Expert Systems with Applications 82, 301–316. https://doi.org/10.1016/j.eswa.2017.04.015

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: Where we are and where we’re going. Transportation Research Part C: Emerging Technologies 43, 3–19. https://doi.org/10.1016/j.trc.2014.01.005