Spatiotemporal cross-validation of urban traffic forecasting models

Dmitry Pavlyuk\textsuperscript{a,}* \\
\textsuperscript{*} Transport and Telecommunication Institute, Lomonosova street 1, Riga, LV-1019, Latvia

Abstract

Spatiotemporal traffic forecasting models become a popular tool of urban transport engineering. Validity of spatiotemporal model specifications and their generalisation abilities are the key aspects that intensively addressed in practice and methodological studies. This paper proposes a spatiotemporal cross-validation approach to estimating model performance, which extends classical temporal cross-validation techniques to a complex spatiotemporal structure of traffic flow relationships. The proposed approach allows estimating model generalisation abilities in the spatiotemporal dimension – ability to forecast traffic flows at unobserved nearby road segments. Additionally, the spatiotemporal cross-validation provides clues for stability of model performance in respect to minor modifications of the spatial structure. Advantages of the proposed spatiotemporal cross-validation approach are demonstrated on a large citywide traffic data set.

1. Introduction

Urban traffic flows are characterised by a complex network of spatial and temporal relationships. These relationships appear between traffic flows at remote segments of a citywide road network and are widely used for spatiotemporal specification of predictive models like spatially regularised vector autoregressive models (VAR), spatial k-nearest neighbours (KNN) algorithm, and graph-based deep learning models, among others (Pavlyuk, 2019). Modern specifications of predictive models are usually highly parametric and suffer from the problem of overfitting, so advanced cross-validation techniques are applied for model performance estimating, hyperparameter tuning and model selection. One of the primary purposes of traffic flow models is accurate prediction of future traffic states, so majority of existing studies apply cross-validation in the temporal dimension only, while spatial and spatiotemporal generalisation power of models is rarely addressed. In this study, we focus on spatiotemporal cross-validation of traffic flow models and provide empirical evidences of its importance for the model selection process.

2. Research contribution

Complexity of the spatiotemporal structure of urban traffic flows (relationships between traffic flows at distant road segments that appear with time delays) and a large volume of available data lead to the critical issue of overfitting, which is and widely acknowledged in the traffic forecasting domain. Cross-validation (CV) is a classical approach to overcome overfitting, which is based on the multiple splitting of the research sample into training and testing subsets. By their nature, citywide traffic flows are multivariate time series, which consist of one or more traffic flow characteristics for many road segments. Thus, the majority of existing studies implement CV in the temporal dimension: training and testing sets are split on the base of a time moment. The most popular temporal CV technique is the rolling window approach (Zivot and Wang, 2006), which slides a time-based window over the research sample and splits it into the training part in the beginning of the window and the testing part at the end. This approach allows assessing of the model forecasting performance and model stability over the time, so it captures an almost uncontested

* Corresponding author. \textit{E-mail address:} Dmitry.Pavlyuk@tsi.lv

Peer-review under responsibility of the scientific committee of the 23rd EURO Working Group on Transportation Meeting.
hegemony in exiting methodological studies on traffic flow forecasting. At the same time, application of spatiotemporal traffic flow models is wider that just forecasting in time. Presence of spatial relationships allows utilising the space dimension: forecasting of traffic flows for neighbour road segments and validating model stability with respect to the background spatial structure. Although the methodology of the spatial CV is well developed, it is rarely directly applied in traffic engineering practice, presumably due to the complexity and non-regularity of the spatiotemporal structure of traffic flows. The regular block-based approach to spatial splitting of data sets does not work well for graph structures like traffic flows: it breaks the structure of relationships and does not guarantee independence of training and testing sets. In addition, performance estimates, obtained for a limited subset of spatial data, are inadequate for those spatiotemporal models that heavily rely on the completeness of the spatial structure.

This study proposes the spatiotemporal approach to cross-validation of traffic flow models. This approach is based on emerging network CV techniques and allows solving the raised issues:

- Assessing model forecasting performance for road segments without historical data,
- Evaluating the stability of model specifications and forecasting performance for different spatial structures.

The proposed approach is approbated for two popular spatiotemporal model specifications: spatially regularised VAR, $SpVAR$, (Schimbinschi et al., 2017) and spatiotemporal KNN, $SpKNN$, (Cai et al., 2016), using a citywide real-world data set.

3. Methodology and preliminary results

The research data set includes traffic flow information from a sample of 100 sensors, deployed on arterial roads of Minneapolis, US. The time period covers 40 weeks with temporal resolution of 5 minutes. Thus, data are represented by a multivariate time series of 100 spatial dimensions, pre-processed for removing periodical components, outliers, and missed observations. In this paper, we propose the spatiotemporal cross-validation approach for estimation of traffic flow models’ forecasting accuracy.

Unlike classical temporal cross-validation techniques, the proposed approach estimates a power of models for forecasting of traffic flows at neighbour road segments without historical data. We discussed existing spatial cross-validation approaches (Holland et al., 1983; Li et al., 2019) as well as distinctive features of spatiotemporal traffic flow relationships that lead to the proposed approach.

The proposed CV technique is applied to forecasting performance estimation of $SpVAR$ and $SpKNN$ models. Using the empirical results, obtained for the large real-world data set, we compare $SpVAR$ and $SpKNN$ forecasting performance in spatial and temporal dimensions. We compare the proposed spatiotemporal CV approach with classical (temporal) CV and discuss its features and advantages. In conclusion, we provide general recommendations for application of the proposed approach for the comparison of spatiotemporal urban traffic forecasting models.

Acknowledgement

The author was financially supported by the specific support objective activity 1.1.1.2. “Post-doctoral Research Aid” (Project id. N. 1.1.1.2/16/I/001) of the Republic of Latvia, funded by the European Regional Development Fund. Research project No. 1.1.1.2/VIAA/1/16/112 “Spatiotemporal urban traffic modelling using big data”.

References

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., Sun, J., 2016. A spatiotemporal cumulative k-nearest neighbor model for short-term traffic multistep forecasting. Transportation Research Part C: Emerging Technologies 62, 21–34. https://doi.org/10.1016/j.trc.2015.11.002

Holland, P.W., Laskey, K.B., Leinhardt, S., 1983. Stochastic blockmodels: First steps. Social Networks 5, 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Li, T., Levina, E., Zhu, J., 2019. Network cross-validation by edge sampling. arXiv:1612.04717 [stat].

Pavlyuk, D., 2019. Feature selection and extraction in spatiotemporal traffic forecasting: a systematic literature review. European Transport Research Review 11. https://doi.org/10.1186/s12544-019-0345-9

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

Zivot, E., Wang, J., 2006. Rolling Analysis of Time Series, in: Modeling Financial Time Series with S-PLUS. Springer, New York, pp. 313–360. https://doi.org/10.1007/978-0-387-32348-0_9