Data-Driven Traffic Flow Modeling with Machine Learning

Applicant: Xinyu Chen (PhD candidate, University of Montreal, Canada)

Background: With recent advances in sensing technologies, various kinds of sensors such as fixed sensing detectors (e.g., loop detectors, radar detectors, cameras, and drones) and mobile sensors (e.g., floating cars and mobile phones) are allocated on the transport network for monitoring traffic states and human mobility. Advanced information systems provide great opportunities for approaching big data in transportation. In recent years, another trend for emerging technologies such as autonomous vehicles and connected autonomous vehicles also highlight the importance of big traffic data and artificial intelligence algorithms. All of these demand us to utilize multi-source big traffic data for developing solutions to transport modeling. A large amount of open data such as PeMS traffic flow data (collected via detectors) (Chen, Petty, Skabardonis, Varaiya and Jia, 2001), NGSIM data (collected via camera)[1], Uber movement data (collected via ridesharing vehicles)[2], HighD data (collected via drone)[3], AD4CHE data (collected via drone)[4], and pNEUMA data (collected via drone)[5] has motivated various transport studies, including traffic state forecasting, travel time estimation, trip/route planning, and traffic signal control. However, these data show complicated data behaviors and characteristics, including sparsity, uncertainty, non-stationarity, non-linearity, multi-dimensionality, and high-dimensionality, in the meanwhile involving noises and anomalies to some extent. The more extreme case is handling imbalanced data with irregular sampling characterization and sparse information. Thus, it is meaningful for developing machine learning models for traffic flow modeling by fully characterizing these data.

Scientific Questions and Objectives: Due to the availability of big traffic data and the development of machine learning, it is an opportunity to reformulate traffic flow modeling problems from a data-driven perspective. However, data behaviors and characteristics of traffic flow complicate the modeling process, posing both methodological and practical challenges. The primary questions arise as 1) how to utilize spatiotemporal context to fuse different sources of traffic data? 2) how to learn from sparse trajectories collected by floating cars (e.g., taxis, ridesharing vehicles, and autonomous vehicles) for estimating traffic states? and 3) how to characterize the time-varying system behavior of traffic flow dynamics? To answer these questions, the goal of this research is to reformulate them appropriately from a data-driven perspective. Our scientific objectives are to:

- (**Objective A**) High-resolution speed field reconstruction of vehicular traffic flow with multi-source data. It requires us to first represent multi-source traffic data onto the same data space. Then, the trajectory data collected from individual vehicles and the traffic measurement data collected by fixed detectors are expected to be incorporated for reconstructing high-resolution speed field of vehicular traffic flow.
- (**Objective B**) City-wide traffic state estimation using floating car data. Floating car data (e.g., collected from taxis/ridesharing vehicles) are important for monitoring urban traffic states. However, this kind of data usually suffers from insufficient sampling of floating cars in total traffic and low-resolution positioning information. As the great advances in internet, positioning technologies, and autonomous vehicles, it is possible to gather high-resolution movement data of vehicles with very accurate positioning information. Thus, it is meaningful to take advantage of these data and find algorithmic solutions to urban traffic state estimation.
- (**Objective C**) Short-term traffic flow forecasting on the imbalanced and sparse data. Essentially, it requires us to handle imperfect data with relatively low data quality. For implementing traffic flow prediction, it demands us to find an efficient learning mechanism on the sparse inputs and in the meanwhile characterize the time-varying system behavior.

Assumption: Before the modeling process, we consider some basic assumptions on the big traffic data as follows. 1) Traffic flow data can be represented by matrices or tensors due to the spatiotemporal setting. 2) Traffic measurement data can be regarded as signals, in the meanwhile showing properties of both time series (e.g., temporal correlations). 3) High-dimensional traffic data can be projected onto low-dimensional spaces.

Methodology: Since traffic data are multi-dimensional and by nature sparse due to the collection process of these data, we consider to develop some unsupervised learning such as tensor factorization and supervised learning such as multi-linear and nonlinear regression for handling the aforementioned tasks.

- (**Objective A**) According to the spatiotemporal setting of traffic flow data, we plan to perform data fusion by taking the essential rules of traffic flow and interpolation methods. We consider to represent different sources of

[1]U.S. Department of Transportation Federal Highway Administration. (2016). Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Accessed 2023-01-01 from http://doi.org/10.21949/1504477.

[2]https://movement.uber.com/

[3]https://www.highd-dataset.com/

[4]https://auto.dji.com/mobile/ad4che-dataset

[5]https://open-traffic.epfl.ch/

traffic data onto the same space and establish the tensor that covers both spatial/temporal dimensions and data source dimension. As a result, tensor factorization is well-suited to high-resolution speed field reconstruction. In the modeling process, it is also meaningful to incorporate domain knowledge of traffic flow.

- (**Objective B**) To overcome the sparsity of partially sampled trajectory data on the transportation network, we introduce the Hankel structure in signal process to reinforce the spatiotemporal modeling of traffic flow data. The factorization on the resulting Hankel tensor (see Figure 1) is well-suited to very sparse data and can characterize complicated correlations of traffic data, hopefully producing accurate estimation of the traffic states.
- (**Objective C**) The network-wide traffic data collected through floating cars usually suffer from insufficient sampling due to the data collection mechanism. To address the imbalanced and sparse data, we plan to develop time-varying regression algorithms that quantify the data uncertainty via deep spatiotemporal priors.

Contribution: The contribution of this research would be two-fold. [**Methodological perspective**] This research could advance the development of machine learning for modeling spatiotemporal traffic data. The proposed approaches such as Hankel tensor factorization are expected to achieve state-of-the-art performance mainly due to the properly modeled domain knowledge in traffic flow. [**Practical perspective**] This research could answer some most important questions for modeling traffic flow. We reformulate the fundamental traffic flow modeling problems with machine learning and bridge the gap between data and algorithms. Therefore, this research is meaningful for supporting data-driven intelligent transportation systems and applications.

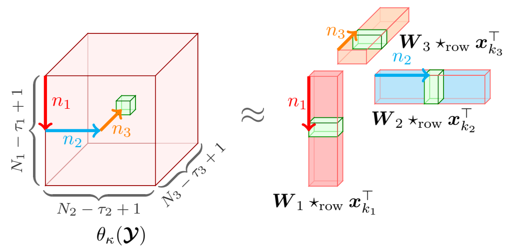

Figure 1: Illustration of Hankel tensor factorization for multi-dimensional traffic flow data $\boldsymbol{\mathcal{Y}}$. The factorization is on the samples of the tensor, and factorized components are connected via the use of circular convolution.

Significance: Understanding traffic flow dynamics is a long-standing topic in transport modeling (Treiber and Kesting, 2013), it is meaningful for drawing strong connections among data, models, and simulation. This research aims to establish efficient machine learning approaches for traffic flow modeling problems as the multi-source big traffic data are now accessible. These approaches could bridge the gap between big traffic data and real-world transport applications, helping improve the existing intelligent transportation systems. In addition, this research is expected to bring fundamental research advances to the general field of spatiotemporal data modeling and promote its application to other domains.

Prior Works: In our recent studies, we focus on spatiotemporal traffic data imputation/forecasting and spatiotemporal data pattern discovery. The proposed imputation and forecasting approaches include low-rank autoregressive tensor completion (Chen, Lei, Saunier and Sun, 2022), Bayesian temporal matrix/tensor factorization (Chen and Sun, 2022), and Laplacian convolutional representation (Chen, Cheng, Saunier and Sun, 2022). For discovering dynamic patterns of spatiotemporal data, we present a time-varying autoregression with tensor factorization (Chen, Zhang, Chen, Saunier and Sun, 2023). On the basis of these works, we developed a GitHub project—**transdim** (i.e., machine learning for transportation data imputation and forecasting)[6]—as a benchmark platform for providing publicly available data and Python implementation of state-of-the-art models (e.g., low-rank matrix/tensor methods). Overall, these would guarantee the high-quality implementation of the proposed research.

References

Chen, C., Petty, K., Skabardonis, A., Varaiya, P. and Jia, Z. (2001), 'Freeway performance measurement system: mining loop detector data', *Transportation Research Record* **1748**(1), 96–102.

Chen, X., Cheng, Z., Saunier, N. and Sun, L. (2022), 'Laplacian convolutional representation for traffic time series imputation', *arXiv preprint arXiv:2212.01529* .

Chen, X., Lei, M., Saunier, N. and Sun, L. (2022), 'Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation', *IEEE Transactions on Intelligent Transportation Systems* **23**(8), 12301–12310.

Chen, X. and Sun, L. (2022), 'Bayesian temporal factorization for multidimensional time series prediction', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(9), 4659–4673.

Chen, X., Zhang, C., Chen, X., Saunier, N. and Sun, L. (2023), 'Discovering dynamic patterns from spatiotemporal data with time-varying low-rank autoregression', *IEEE Transactions on Knowledge and Data Engineering* .

Treiber, M. and Kesting, A. (2013), 'Traffic flow dynamics', *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg* pp. 983–1000.

[6] https://github.com/xinychen/transdim (1,000+ stars & 270+ forks on GitHub)