Augmenting Traffic Prediction with Simulation Based Synthetic Data

Nicolas Bent[1,3], Nicolas Saunier[2,3], Francesco Ciari[2,3], and Alejandro Quintero[1]

[1] *Polytechnique Montréal, Department of Computer Engineering, Montreal, Canada*
[2] *Polytechnique Montréal, Department of Civil, Geological and Mining Engineering, Montreal, Canada*
[3] *CIRRELT - Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation, Montreal, Canada*

Extended abstract submitted for the International Symposium on Transportation Data Modelling (ISTDM)

---

1 INTRODUCTION

Short term traffic prediction is critical for real time traffic management. While recent advances in Deep Learning (DL) methodologies have enhanced the capabilities of short term traffic prediction, these approaches necessitate extensive datasets to achieve optimal performance and reliability. However, the acquisition of comprehensive traffic flow data presents substantial challenges due to technical, logistical, and resource constraints. Synthetic data generation emerges as a promising solution to address this data scarcity, offering a systematic approach to create large-scale, realistic traffic flow datasets while mitigating the challenges associated with real-world data collection.

The application of synthetic data generation has demonstrated significant efficacy across diverse scientific domains [1]. Recent implementations emerging in transportation research predominantly utilize Deep Learning (DL) methodologies for synthetic data generation in transportation [2]. Transportation engineering has historically relied on models based on the physical simulation of vehicles and driver behaviour for traffic analysis and prediction. However, the successful deployment of physics-based models for synthetic data generation in other scientific domains suggests potential applications in transportation systems [3]. This presents a unique opportunity to enhance (DL) transportation models by leveraging synthetic simulation data.

In this work, we propose a novel methodology that integrates simulation synthetic data into DL frameworks for short term traffic prediction. Our approach leverages the rich information available in traffic simulations to enhance model training. Through comprehensive empirical evaluation, we demonstrate that this integration improves prediction accuracy and model generalization capabilities. The proposed framework offers a promising direction for advancing the field of transportation modeling by using traditional physical models to improve the performance of DL models in short term traffic prediction.

2 METHODOLOGY

We propose a comprehensive framework for enhancing transportation models through the systematic integration of synthetic and real-world data. Our methodology comprises four principal steps: calibration of synthetic data using simulation based dynamic traffic assignment, data augmentation using traffic simulation, pretraining on synthetic data and predicting on a test set of real traffic data. Predicting on real data ensures that we measure the performance of our models against on actual traffic conditions providing a true validation of our framework's effectiveness and practical applicability.

We use SUMO [4] as our simulation and software. We modify the calibration process for SUMO and employ a genetic algorithm-based (GA) optimization framework to ensure alignment

between synthetic and observed traffic patterns. This iterative procedure systematically modifies route choices and progressively converges toward solutions that minimize the deviation from observed traffic counts. This optimization continues until the root-mean-square error between simulated and observed flows falls below a predetermined threshold, ensuring that the synthetic data maintains high fidelity to real-world traffic dynamics.

Second, we implement a systematic data augmentation strategy to enhance the diversity of our synthetic dataset. For each calibrated simulation scenario, we generate multiple variants by introducing controlled perturbations to temporal and spatial parameters. Specifically, we modify departure times and route choices within physically consistent bounds, creating a comprehensive dataset that captures various traffic conditions while maintaining realistic flow dynamics.

Finally, we evaluate our approach through a training procedure across multiple state-of-the-art deep learning architectures. The models are initially pre-trained on the augmented synthetic dataset to learn fundamental traffic flow patterns. Subsequently, we fine-tune these models using real-world data. This transfer learning approach enables the models to leverage the comprehensive synthetic data while maintaining accuracy on real-world predictions. We compare our technique to the calibrated SUMO model and to models trained only on the real world data. We validate our methodology through extensive experimentation using multiple DL architectures demonstrating consistent performance improvements.

3 EXPERIMENTS

3.1 Dataset

We use the Ingolstadt dataset [5], which contains traffic flow measurements at 15-min intervals from June 2023 to present. From the available 794 loop detectors, we selected 472 detectors that maintain consistent data quality throughout the observation period. For synthetic data we use the ingolstand SUMO scenario [6] which has calibrated scenarios for 10 days in the summer of 2023. We use these calibrated scenarios as the basis for further GA calibration for all traffic days from June-August. The final Mean Absolute Error (MAE) between the calibrated synthetic and real data is between 8 and 12%. We generate a synthetic dataset by simulating the calibrated scenarios and simulating 10 scenarios per calibrated day. We modify 1% of the routes by randomly changing the start or end time by 15 min or changing the route choice. This generates 900 extra days of traffic flow. We fine tune on the real word traffic data of June-August and predict on September 2023.

3.2 GCN Models

We evaluate our method on three different graph convolution networks with temporal modules. These networks have become the state of the art (SOTA) in traffic flow prediction [7]. We compare four different models, Diffusion Graph Convolution (DGCN) [8], Spatio-Temporal Graph Convolutional Networks(STGCN) [9] Temporal Graph Convolutional Network (T-GCN) [10] and Historical Average.

3.3 Experimental Settings

We use two NVIDIA P4s for training. Each model is trained for 100 epochs, where the first 50 epochs use a 10^{-3} learning rate and the subsequent 50 epochs have a decay rate of 0.5 every 5 epochs. We use blocks of 15 min flows as input and outputs to the models. The inputs to the models are 8 15-min flows, and we predict on a 15, 30 and 45-min horizon. The evaluation metrics are MAE and Mean Square Error (MSE).

Models	MAE (15/30/45) min	MSE (15/30/45) min
HA	5.34	7.65
SUMO	4.96	7.22
DGCN	3.92/4.61/5.18	5.43/ 6.45/7.38
DGCN + Synth	3.52/4.15 /4.71	4.91/5.83/6.69
STGCN	3.85/4.52/5.11	5.31/6.32/7.29
STGCN + Synth	3.46/4.08/4.63	4.82/5.75/6.64
TGCN	3.78/4.45/5.03	5.20/6.20 /7.21
TGCN + Synth	3.41/4.04/4.58	4.73/5.65/6.58

Table 1 – *Performance of models with and without synthetic data*

Models	MAE (15/30/45)	MSE (15/30/45)
DGCN	11.1%/ 11.1% / 9.9%	10.5% /10.6% / 10.3%
STGCN	11.2%/ 10.7% / 10.3%	10.1% /9.8% / 9.7%
TGCN	10.6%/ 10.1% / 9.8%	10.5% /10.6% / 10.3%
Average	11.0%/ 10.6% / 10.0%	10.4% /10.3% / 10.1%

Table 2 – *Percent improvement of models when using synthetic data*

4 RESULTS

In Table 1 we show the results of the model trained with and without synthetic data. Incorporating synthetic data consistently improves the performance of all tested GCN variants. Our models also outperform SUMO in these time horizons. Table 2 shows the percent improvement when using synthetic and non-synthetic data, the average improvement is between 10 and 11%. The consistent improvement across different architectures suggests that synthetic data augmentation is a robust approach for enhancing traffic prediction models.

5 DISCUSSION

This work demonstrates the effectiveness of synthetic data augmentation in traffic prediction, addressing a fundamental challenge in deep learning applications: the need for large-scale, high-quality training data. While domains like computer vision benefit from vast labeled datasets, traffic flow prediction faces inherent data scarcity due to limited sensor deployment and collection costs. Our results show that calibrated synthetic data provides a viable solution, yielding consistent performance improvements of 10-11% across various GCN architectures. The success of this approach aligns with established scaling laws in deep learning, which indicate that model performance logarithmically improves with dataset size, which have recently been shown for graphs by [11]. The key advantage of our synthetic data approach lies in its calibration procedure, which ensures high fidelity to real-world traffic patterns while expanding the available training data. The consistent improvement across different prediction horizons (15, 30, and 45 min) suggests that synthetic data is helpful for different time horizons.

6 CONCLUSION AND OUTLOOK

Our study reveals that synthetic data significantly enhances traffic prediction accuracy when used with graph convolutional networks with tests across multiple GCN architectures showed consistent 10-11% improvements in both MAE and MSE metrics.

Ongoing research, which is expected to be completed by September, will help determine the best practices for synthetic data usage in short term traffic prediction. First, investigating the

relationship between calibration error and model performance could establish quality thresholds for synthetic data generation. Second, understanding the optimal ratio of synthetic to real data could help optimize training strategies. Finally, analyzing the effectiveness of synthetic data augmentation under different real-world data availability scenarios could provide insights into the most beneficial settings for his approach. Looking ahead, combining simulation based synthetic data with DL models shows promise for tackling other transportation problems. For instance, further improving short term traffic prediction by adding synthetic data for what-if scenarios or training smart dynamic traffic assignment models.

References

[1] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, "Machine learning for synthetic data generation: A review," 2024.

[2] Y. Zhu, Y. Ye, Y. Wu, X. Zhao, and J. Yu, "Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22961–22977, 2023.

[3] Z. Wang, X. Yu, C. Wang, W. Chen, J. Wang, Y.-H. Chu, H. Sun, R. Li, P. Li, F. Yang, H. Han, T. Kang, J. Lin, C. Yang, S. Chang, Z. Shi, S. Hua, Y. Li, J. Hu, L. Zhu, J. Zhou, M. Lin, J. Guo, C. Cai, Z. Chen, D. Guo, G. Yang, and X. Qu, "One for multiple: Physics-informed synthetic data boosts generalizable deep learning for fast mri reconstruction," 2024.

[4] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*, pp. 2575–2582, IEEE, 2018.

[5] M. Knezevic, A. Donaubauer, M. Moshrefzadeh, and T. H. Kolbe, "Managing urban digiital twins with an extended catalog servce," in *Proceedings of the 7th International Smart Data and Smart Cities (SDSC) Conference 2022*, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, UNSW Sydney, 2022.

[6] M. Harth, M. Langer, and K. Bogenberger, "Automated calibration of traffic demand and traffic lights in sumo using real-world observations," in *SUMO Conference Proceedings*, vol. 2, pp. 133–148, 2021.

[7] H. Li, Y. Zhao, Z. Mao, Y. Qin, Z. Xiao, J. Feng, Y. Gu, W. Ju, X. Luo, and M. Zhang, "Graph neural networks in intelligent transportation systems: Advances, applications and trends," 2024.

[8] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2018.

[9] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, IJCAI-2018, p. 3634–3640, International Joint Conferences on Artificial Intelligence Organization, July 2018.

[10] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, p. 3848–3858, Sept. 2020.

[11] J. Liu, H. Mao, Z. Chen, T. Zhao, N. Shah, and J. Tang, "Neural scaling laws on graphs," 2024.