A Geomagnetic Sensor Dataset for Traffic Flow Prediction

Huanchen Wang[1,*], Quanjun Chen[2,*], Zheng Dong[1], Xuan Song[1], Hao Tian[3], Donglong Yang[3], Manxia Liu[3]
[1] Southern University of Science and Technology
{11810419,11812804}@mail.sustech.edu.cn, songx@sustech.edu.cn
[2] Center for Spatial Information Science, The University of Tokyo
chen1990@iis.u-tokyo.ac.jp
[3] Transport Bureau of Shenzhen Municipality
[*] Equal Contribution.

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020763

***Abstract*—Traffic state prediction is essential in Intelligent Transportation Systems for surveillance, management, and daily commuting. For developing high-accuracy prediction models, real-world traffic state datasets are necessary for training model parameters and evaluating prediction results. However, limited by the existing traffic collection devices, most of the current open datasets for traffic state prediction cannot obtain accurate traffic flow information. In contrast, some datasets directly use detection devices in freeway systems, so they cannot reflect complex urban traffic states. Therefore, a dataset from advanced devices that can record the flow from point to point on an urban road network attracts more attention and drives the progress of research on traffic state prediction models. To deal with the above issues, we introduce a Suburban Traffic Flow dataset using Geomagnetic sensors, or STF-G dataset, constructed for traffic flow prediction. The STF-G dataset consists of 2.5 billion vehicle driving scenarios and 319 corresponding geomagnetic sensors. The data was collected over 20 months and processed with two regional road graphs. We also do the Benchmark experiments in STF-G for analyzing and evaluating the performance of graph neural network models in traffic flow prediction and compare them to the other datasets with the same baseline.**

***Index Terms*—Geomagnetic Sensor, Dataset, Benchmark, Traffic State Prediction, Graph Neural Networks**

I. Introduction

Traffic state prediction is a critical part of Intelligent Traffic Systems (ITS), which is essential in government and citizens' traffic surveillance, management, and analysis. To develop an advanced prediction model, a dataset, which can reflect the real-world traffic state, is indispensable for training model parameters and evaluating results. Especially with the wide application of Graph Convolution Networks (GCN) in traffic forecasting [1], the corresponding road connection is also needed in the traffic state datasets.

However, getting an accurate and detailed traffic state on a city scale is not easy, even in modern society. Although researchers have tried to construct several open datasets for traffic prediction tasks, they have respective problems in reflecting real urban traffic states, which can be concluded as:

(1) Data acquisition. As shown in Figure 1, there are several types of sensors to collect traffic state. Vehicle devices with the Global Positioning System (GPS) could acquire real-time location so that they can be used to calculate traffic state. However, not all vehicles have GPS devices, nor can all GPS data be aggregated because of privacy issues. Hence traffic flow collected by GPS could be a sampling of real-world situations. On the other hand, surveillance cameras can capture overall road conditions, including cars and pedestrians, then traffic flow can be counted through object detection methods in computer vision. It cannot be recognized with 100% accuracy because of challenges such as varying perspectives, lighting conditions, and sources of occlusion. Only loop detectors and geomagnetic sensors can catch the most real traffic state information as they detect each vehicle's movement directly.

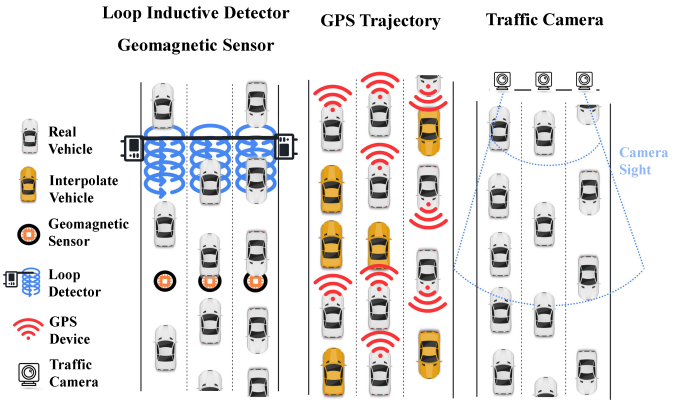


Fig. 1: Comparison of Device

(2) Road network coverage. In Table I, traffic state datasets widely used in traffic prediction research are listed. As analyzed above, data collected by loop detectors can reflect more real traffic states than GPS and cameras. Thus, the representative datasets provided by Caltrans Performance Measurement System (PEMS) [2] are the first choice in many studies. However, those loop detector datasets are based on the freeway system and cannot cover a city's regular road network, which has much more turns, intersections, and traffic signals. Therefore, models can achieve good results at loop detector datasets but can hardly be transferred to real-world applications.

To address these shortcomings, we introduce a Suburban Traffic Flow dataset using Geomagnetic sensors, or **STF-G**,

TABLE I: Comparison with Other Traffic State Datasets

Dataset	Location	Road Type	Device	Features	Time Period	Time Interval
SZ_TAXI [3]	Shenzhen, China	Urban Roads	Taxi Trajectory	Speed	1/1/2015 - 1/31/2015	15 min
METR_LA [4]	Los Angeles, USA	Freeways	Loop Detectors	Speed	1/3/2012 - 30/6/2012	5min
LOOP_SEATTLE [5], [6]	Seattle, USA	Freeways	Loop Detectors	Speed	1/1/2015 - 12/31/2015	5min
PEMS-BAY [7]	San Francisco, USA	Freeways	Loop Detectors	Speed	1/1/2017 - 1/30/2017	5 min
PEMSD08 [8]	San Bernardino, USA	Freeways	Loop Detectors	Flow, Speed	7/1/2016 - 8/31/2016	5 min
PEMSD04 [8]	San Francisco, USA	Freeways	Loop Detectors	Flow, Speed	1/1/2018 - 2/28/2018	5 min
STREETS [9]	Chicago, USA	County Roads	Traffic Cameras	Flow	10 weeks between 2018 - 2019	5 min
NYCTAXI_DYNA	New York, USA	Urban Roads	Taxi Trajectory	Flow	1/1/2020 - 3/30/2020	60 min
STF-G	Shenzhen, China	Urban Roads	Geomagnetic Sensors	Flow	1/1/2019 - 8/31/2020	5 min

to serve as a benchmark for traffic state prediction tasks. This dataset includes the traffic state in the temporal dimension and incorporates the sensor-based graph representing a spatial dependency, which is suitable for easily developing and evaluating graph-based models. The contributions of this paper include the following:

- We introduce a novel Geomagnetic Sensor dataset called **STF-G**, which consists of more real-world and comprehensive traffic data with corresponding Spatio-temporal information.
- We introduce a benchmark for traffic flow prediction models, especially the state-of-the-art graph-based models, that evaluates the performance and robustness of each model in suburban scenarios.

II. Related Work

Real-world traffic reporting is an essential issue in traffic information collection. Many traditional traffic collection devices acquire state indirectly, committing to image identification and data imputation. Instances consist of traffic cameras that detect and count vehicles using the computer vision method, such as STREETS [9] and GPS which interpolate the trajectory by statistical samplings after the map matching method, like the Fast Map Matching [10], such as SZ_TAXI [3]. Furthermore, the installation of loop detectors creates massive damage to roads. However, these devices are limited in terms of sight, visibility, missing data, and difficulty of the deployment factors, which prevents recording the complete real-world information of each passing vehicle.

Some recent research work [11] has begun to pay attention to geomagnetic sensor applications, especially vehicle detection, vehicle classification, and vehicle counting. The geomagnetic sensor can detect the change in the surrounding geomagnetic field caused by each passing vehicle. Therefore, it can accurately record vehicle length and passage time which can acquire the traffic state easily. In addition, the installation of the geomagnetic sensor is quick and convenient. These works apply geomagnetic sensors to collect traffic information but do not fully utilize data on state prediction, and their geomagnetic sensor datasets are very limited [12].

III. Suburban Traffic Flow using Geomagnetic Sensors

In this section, we present a dataset **STF-G**, which reflects Suburban Traffic Flow using Geomagnetic sensors, and explain what features are different from other traffic datasets for flow prediction tasks.

A. Data Collection

Shenzhen is a major sub-provincial city in China, with more than 17,633,800 resident population. Geomagnetic sensors are distributed on the roads of every district in Shenzhen. Sensors in each lane can detect the direction of vehicles to avoid roads overlapping with multiple lanes and the difficulty of 2D maps representing. In addition, the sensor can induce changes in the geomagnetic field to record the length and type of vehicles passed. The types in **STF-G** include extra-large cars, large cars, mid-cars, and micro-cars. Thus, the flow in passenger car unit (PCU) method [13] can represent more scientific conditions on the road and the standard as follows the equation:

$$Flow = 4.0 \times Extra + 2.5 \times Large + 1.5 \times Mid + 1.0 \times Micro \quad (1)$$

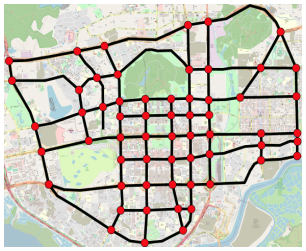
It uses the coefficient to weigh each type of vehicle, including combination, large-sized, middle-sized vehicles, and passenger cars, to result in a more reasonable traffic flow.

At the time of data collection for this work, there are 319 distinct sensors across Shenzhen for a total of 2.5 billion logs of vehicles. The **STF-G** is extracted every 5 minutes for 20 consecutive months from 1/1/2019 to 31/8/2020 among the 129 million vehicles passed. To benchmark, centers of two districts between Nanshan and Futian with the most concentrated traffic flow in Shenzhen of 50 distinct sensors in total are selected from 1/6/2019 to 30/6/2019. The missing rate is lower than 10.0% and linear interpolation is applied. In general, **STF-G** presents great reality in the direction of the roads, vehicle classification, traffic congestion, etc.

B. Sensor-Based Graph

The sensor-based graph aims to generate a spatial graph through sensor distribution and road connection. Most current mainstream state prediction models extract spatial dependency through graph structures. More and more traffic state prediction models apply graph convolution to operate the irregular

(a) Sensor Graph in Nanshan



(b) Sensor Graph in Futian

Fig. 2: Sensor Graph

spatial connection of road networks. However, several traffic datasets only provide the traffic state of different roads in the time series, lacking the graph-structure data.

TABLE II: Statistics for Each District Graph.

	Nanshan	Futian
# Vertices	19	31
# Edges	62	96
Area(sq. km)	46.05	48.42

Thus, two sensor-based graphs of **STF-G**, as shown in Figure 2, are manually established based on the connection relationship among the sensor points in real urban roads where they are located. The exact number of nodes, edges, and corresponding spatial area of each graph are listed in Table II.

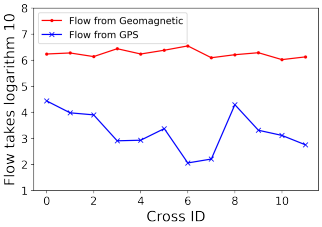
C. Data Comparative Analysis

In order to give a clear image of the advantage of **STF-G**, we do the comparative analysis between **STF-G** and other traffic state datasets of different sensor types.

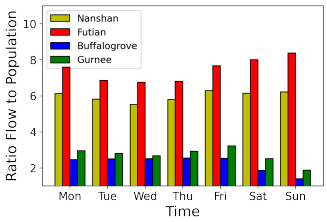
*Flow Comparison between **STF-G** and GPS.* We selected a dataset collected by taxi-GPS in Shenzhen for comparison. For these two datasets, some roads located in Futian District are selected to compare one-month traffic volume. As Figure 3a indicates, even taking logarithm 10 for flow, the volume of **STF-G** on each road is much larger than the GPS dataset.

*Flow Comparison between **STF-G** and Camera.* We selected the STREETS [9] collected by cameras in Chicago for comparison. Because the regions of STREETS and **STF-G** are different, instead of comparing flow directly, the ratio of daily flow in each dataset and the local population in the first week of June 2019 is calculated as a comparison reference. In Figure 3b, the ratios indicate that each day's flow information of **STF-G** (Nanshan and Futian) is more complete than STREETS (Buffalogrove and Gurnee).

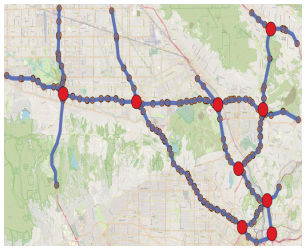
*Graph Complexity Comparison between **STF-G** and Loop Detector.* The loop detectors in most datasets are on freeways instead of urban roads, which cannot reflect the complex condition in urban. To visualize the spatial difference, We select the METR_LA [4] and PEMS-Bay [7] and represent the distributions of their detectors. According to the comparison between Figure 2 and Figure 3c 3d, it is obvious that although there are hundreds of sensors (small red dots) in METR_LA and PEMS-Bay, they are all concentrated on a few freeways



(a) Take Logarithm 10 for Flow of **STF-G** and GPS Dataset in Shenzhen



(b) The Flow to Population Ratio of **STF-G** and STREET



(c) Sensor Graph of METR_LA

(d) Sensor Graph of PEMS-Bay

Fig. 3: Comparison between **STF-G** and GPS, Camera, Loop Detector Datasets Separately

which can be simplified into graphs with only several nodes (big red dots) and edges (blue lines). While our **STF-G** distribute on major urban roads in central Shenzhen can reflect a more real and multiplex road network.

IV. Benchmark Task

In this section, we evaluate the performance of models, including statistical algorithms, time-series deep learning algorithms, and graph-based nueral networks on **STF-G**.

A. Basic Settings

The Adam [14] is used for optimization with an initial learning rate of 0.0001. The sub-datasets Nanshan and Futian are selected and divided into 70% for training, 10% for evaluation, and 20% for testing. Evaluation of the model performance through three metrics: mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). The Benchmark evaluates the classical methods including HA [15] and SVR [16], time-series deep learning as GRU [17], LSTM [18], Seq2Seq [19] and advanced graph-based models as AGCRN [20], Graph Wave Net [21], GTS [22] and DCRNN [7].

B. Results Evaluation

Table III compares the performance of our selected models for 15 minutes, 30 minutes, and 60 minutes ahead of predictions on Nanshan and Futian sub-datasets. According to the evaluation of the Tabel III, it can conclude that: (1) Generally, essential deep learning with only temporal dimension models presents a better performance in the prediction task than non-deep learning statistical methods. HA and SVR conclude preliminary results and can only work on short-term prediction. Its performance would degenerate as the step increases. Their multi-step results point to drawbacks in long-term prediction

TABLE III: Results of Each Model on Nashan and Futian Datasets

Model	Nanshan									Futian								
	15 min			30 min			60 min			15 min			30 min			60 min		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
HA	27.05	24.95%	59.95	27.05	24.95%	59.95	27.05	24.95%	59.95	20.19	23.76%	35.39	20.19	23.76%	35.39	20.19	23.76%	35.39
SVR	35.36	31.52%	67.70	37.71	33.67%	72.50	43.63	40.11%	84.00	23.95	29.45%	40.79	26.44	31.26%	48.11	32.60	37.11%	63.44
GRU	18.69	18.18%	31.92	19.61	18.69%	34.02	21.34	19.98%	38.16	17.11	22.75%	29.02	17.43	22.79%	29.87	18.12	23.07%	32.66
LSTM	19.89	19.38%	34.00	20.58	19.77%	35.70	22.04	20.94%	39.01	17.60	23.01%	30.69	17.91	23.22%	31.42	18.69	23.73%	33.79
Seq2Seq	17.95	17.66%	30.52	18.82	18.25%	32.54	20.59	19.91%	36.44	16.48	21.95%	27.89	16.88	22.23%	29.03	17.86	22.97%	31.95
GTS	17.37	17.66%	31.37	19.06	19.09%	34.21	21.75	21.41%	39.13	14.77	18.61%	24.89	15.70	19.52%	27.06	17.30	21.07%	31.09
AGCRN	17.09	17.76%	**30.17**	18.40	19.11%	33.04	**20.45**	18.60%	37.64	15.29	21.69%	26.40	16.01	20.71%	28.08	17.50	22.23%	31.52
GWNET	17.14	16.27%	30.54	**18.16**	17.31%	**32.49**	20.54	18.90%	37.50	15.13	19.84%	25.18	15.92	20.23	27.19	17.51	20.60%	30.90
DCRNN	**16.80**	**15.61%**	30.22	18.19	**16.46%**	32.56	20.51	**18.49%**	**36.93**	**14.76**	**18.30%**	**24.70**	**15.69**	**19.23%**	**26.63**	**17.29**	**20.96%**	**30.40**

tasks. (2) The three RNN-based deep learning models, GRU, LSTM, and Seq2Seq, have the same good performance in predicting all time steps. Such results again demonstrate the importance of time series in traffic state forecasting. (3) Almost all graph-based models achieved better performance than traditional methods and time-series models on all metrics, proving that adding spatial information would bring substantial performance improvements. (4) Although the performance of the models depended more or less on the dataset, the scores of DCRNN and GWNET on most of the datasets ranked in the top two. It also proved their robustness and versatility in traffic prediction tasks. GWNET and DCRNN utilize diffusion convolutional networks, which can identify the correlation of devices on the road in different directions by themselves.

In addition, the robustness is further analyzed by referring results provided by Libcity [23] on several public datasets. As a result of the benchmark for PEMS04 in Libcity, the top two are AGCRN and GWNET, while GWNET also got top scores at our **STF-G** as analyzed above. It demonstrates the great robustness of GWNET, or self-adaptive/learnable graph models, on traffic prediction task.

V. Conclusion

In this work, we propose a novel **STF-G** dataset, which incorporates a more real-world and comprehensive flow state. It avoids the privacy issues of GPS collection, the limited sight of the camera collection, and the outdated loop detector. Furthermore, to evaluate and analyze current research on traffic state prediction, we additionally provide a benchmark for state prediction models with **STF-G**.

VI. Acknowledgments

This work was supported by Transport Bureau of Shenzhen Municipality, and the data can be obtained by applying to http://jtys.sz.gov.cn/.

References

[1] R. Jiang, D. Yin, Z. Wang, Y. Wang, J. Deng, H. Liu, Z. Cai, J. Deng, X. Song, and R. Shibasaki, "Dl-traff: Survey and benchmark of deep learning models for urban traffic prediction," in *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 2021.

[2] C. Chen, K. Petty, A. Skabardonis, P. Varaiya, and Z. Jia, "Freeway performance measurement system: mining loop detector data," *Transportation Research Record*, 2001.

[3] 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*, 2019.

[4] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Communications of the ACM*, 2014.

[5] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction," *arXiv preprint arXiv:1801.02143*, 2018.

[6] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[7] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. of ICLR*, 2018.

[8] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. of AAAI*, 2019.

[9] C. Snyder and M. Do, "STREETS: A novel camera network dataset for traffic flow," in *Proc. of NeurIPS*, 2019.

[10] C. Yang and G. Gidofalvi, "Fast map matching, an algorithm integrating hidden markov model with precomputation," *International Journal of Geographical Information Science*, 2018.

[11] C. Xu, Y. Wang, X. Bao, and F. Li, "Vehicle classification using an imbalanced dataset based on a single magnetic sensor," *Sensors*, 2018.

[12] Y. Feng, G. Mao, B. Cheng, C. Li, Y. Hui, Z. Xu, and J. Chen, "Mag-monitor: Vehicle speed estimation and vehicle classification through a magnetic sensor," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[13] J. B01-2014, "Technical standard of highway engineering," 2014.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015.

[15] J. Liu and W. Guan, "A summary of traffic flow forecasting methods," *Journal of highway and transportation research and development*, 2004.

[16] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Proc. of NeurIPS*, 1996.

[17] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[18] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[20] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Proc. of NeurIPS*, 2020.

[21] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," in *Proc. of IJCAI*, 2019.

[22] C. Shang, J. Chen, and J. Bi, "Discrete graph structure learning for forecasting multiple time series," *arXiv preprint arXiv:2101.06861*, 2021.

[23] J. Wang, J. Jiang, W. Jiang, C. Li, and W. X. Zhao, "Libcity: An open library for traffic prediction," in *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 2021.