A Synthetic Dataset Analysis for Syn-Xi’an

In this section, we will perform a detailed analysis of the Syn-Xi’an dataset. By analyzing the trajectory datasets synthesized from totally different cities, we can further verify the capability of the trajectory synthesizer and the feasibility of the trajectory synthesis solution.

A.1 Dataset description

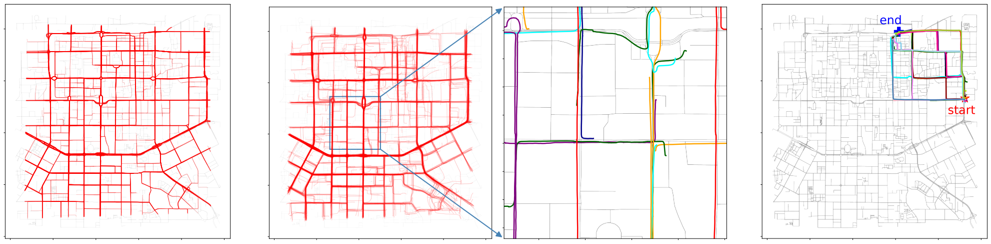
As presented in Table 5, the Syn-Xi’an dataset possessed the same data format as the Syn-Chengdu dataset. Specifically, the dataset also consists of one million trajectory records, each of which is represented in two parts: attribute information and trip trajectory. For these attribute information, such as `departure time` t_d`, trip distance (meters), trip time` t_a`(seconds), and sampling points` n, the representation and calculation are the same as for the Syn-Chengdu. This consistent representation of data can save researchers substantial time, and can also support cross-city transfer learning studies.

Table 5: Dataset description of Syn-Xi’an

Type	Description
Format	pickle / geoparquet
Size	4.66 GB
Value type	float64
Time frame	5 min
Sample interval	3 s
Spatial coverage	lat: $34.20° \sim 34.28°$ lng: $108.90° \sim 108.99°$

A.2 Trajectory geo-distribution insight

Figure 5 shows a trajectory visualization of the synthetic dataset along with a comparison of the original counterpart. Among them, the comparison between the original trajectory (Figure 5(a)) and the generated trajectory (Figure 5(b)) clearly shows that the synthetic dataset can well mirror the trajectory distribution of the original data. Further area zooming results show that there is no remarkable variation between the synthetic trajectory and the urban road network. Finally, the figure 5(c) exhibits the diversity of synthetic trajectories, where trajectories with the same start and end areas cover different paths.



(a) Original trajectories.

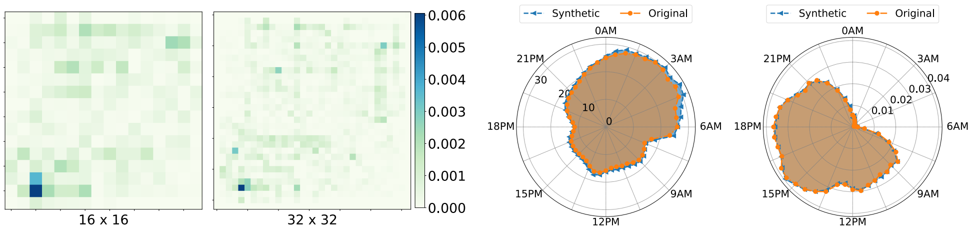
(b) Synthetic trajectories with area zoom.

(c) Same start-end areas.

Figure 5: Comparative visualizations of original and synthetic trajectories. (a) and (b) server as a reference for the quality and accuracy of the synthetic trajectories. (c) depicts the diversity of synthetic trajectories based on the same start and end areas.

A.3 Spatial-temporal distribution

We further analyze the spatial and temporal distributions incorporated in the Syn-Xi’an dataset. As shown in the heatmap of the trajectory distribution plotted by 6(a), the synthetic trajectory dataset maintains the original distribution properties well in terms of spatial distribution. For the temporal distribution of this synthetic dataset, it is clear observe that the variation in speed and trip volume is consistent with the real world (see Figure 6(b) and Figure 6(c)). These changes in urban traffic activity levels are not only reflected in Xi'an the city, but are also similar to the results presented in the Chengdu dataset, providing confidence in the cross-city urban mobility analysis.



(a) Heatmap distribution of difference

(b) Average speed

(c) Trip volume

Figure 6: Spatial-temporal distribution of the synthetic dataset (SYN-XI'AN). (a) Differences in heatmap of raw-synthetic (calculated by absolute) trajectories distribution. (b) and (c) The average speed and the number of trips throughout the day.

A.4 Trajectory properties

For the analysis at the trajectory properties level, we follow the same way as introduced in Section 4.2. The results are shown in Figure 7, where the K-S statistical similarities for the four aspects of travel distance, travel time, speed and relative distance are 0.99, 0.94, 0.94 and 0.98, respectively. In addition, the visualization results can adequately demonstrate the ability of the synthetic dataset to maintain the statistical characteristics of realistic trajectories with high fidelity. This excellent result is attributed to two aspects, the superior spatial-temporal generation performance of the proposed trajectory synthesizer. Second, the high quality of the original data further enhances the robustness of the synthesized data.

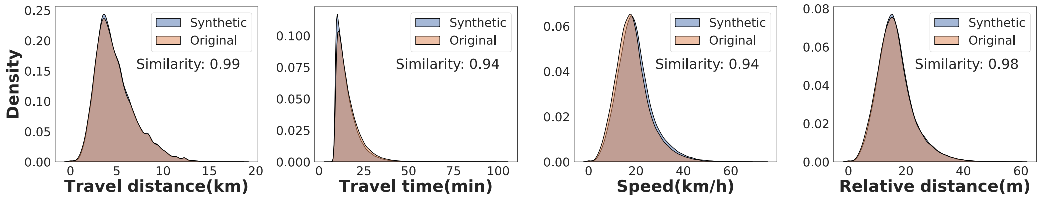


Figure 7: Comparative analysis of trajectory properties in original and SYN-XI'AN datasets.

In summary, we can conclude the following from the above analysis:

- The analysis results on the synthetic dataset of Xi'an city fully validate the feasibility of the proposed trajectory synthesis approach. It breaks a new direction for future urban mobility analytics, solving the problem of restricted use of trajectory data.
- Analytical results from two synthetic datasets demonstrate the powerful ability to employ a diffusion model as a trajectory synthesizer. This provides a novel solution for trajectory synthesis with generative models.
- Since both datasets use the same format and data style, this offers the required dataset for cross-city urban mobility analysis.

B Use Cases of Synthetic Dataset for Syn-Xi’an

In this section, we also use two case studies to test the utility of Syn-Xi’an. All the code was implemented by pytorch and run on a server with an Nvidia 2080 Ti GPU and Intel Silver CPU. All codes for these models follow the public benchmark[10] shown in the literature [16], and the results were averaged over 3 times. For performance evaluation of these tasks and models, we used the following metrics:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i}^{n}\left(y_i - \hat{y}_i\right)^2}, \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i}^{n}\left|y_i - \hat{y}_i\right|, \tag{2}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{3}$$

where y_i and $\hat{y}_i$ are the ground truth and predicted traffic value, respectively.

B.1 Traffic Demand Prediction

In the experimental setup, we used the following advanced spatial-temporal neural network models:

- **AGCRN:** This is an adaptive convolutional recurrent neural network which contains a Node Adaptive Parameter Learning (NAPL) module and a Data Adaptive Graph Generation (DAGG) module. Where NAPL is used to capture node-specific patterns and DAGG is used to infer relationships between different traffic series.
- **GWNet:** This is a traffic prediction model based on graph convolutional network (GCN) and Wavenet structure. Among them, GCN is used to capture the spatial dependency of traffic nodes and Wavenet is used to capture the temporal dependency.
- **DCRNN:** This is a advanced neural network model based on directed graphs, which models the change of traffic flow as a diffusion process.
- **MTGNN:** This is a generic graph neural network framework that combines external knowledge and relationships between variables through a graph learning module, and then captures spatial and temporal dependencies using mix-hop propagation layers and inflated inception.

Table 6: Data utility comparison by traffic demand prediction. Results are expressed as (original / synthetic / difference ratio).

Methods	AGCRN	GWNet	DCRNN	MTGNN
RMSE	6.51 / 6.08 / 6.61%	6.52 / 6.39 / 1.99%	6.52 / 6.41 / 1.69%	6.31 / 5.97 / 5.39%
MAE	4.47 / 4.16 / 6.94%	4.47 / 4.38 / 2.01%	4.49 / 4.37 / 2.67%	4.41 / 4.11 / 6.80%
MAPE	30.35 / 30.55 / 0.66%	30.82 / 32.53 / 5.55%	30.83 / 32.65 / 5.90%	29.19 / 29.94 / 2.57%

As presented in Table 6, we extend our evaluation to a different synthetic dataset by gauging the performance of AGCRN, GWNet, DCRNN, and MTGNN models on traffic demand prediction tasks. The metrics employed for performance evaluation include RMSE, MAE, and MAPE. Overall, the performances of these models on both the original and new synthetic datasets are close, endorsing the ability of synthetic datasets to replicate real-world scenarios in traffic demand prediction. To be specific, the difference ratio across performance metrics typically falls within a modest range, highlighting the robustness of our synthetic dataset. It is worth observing that the MAPE for GWNet and DCRNN show a slightly increased difference ratio. Despite these model-specific variances, the synthetic dataset continues to exhibit promising utility. In summary, these findings confirm the usability of our Syn-Xi’an dataset for traffic demand prediction tasks, and also expand its potential applicability in different urban traffic scenarios.

[10] https://github.com/deepkashiwa20/DL-Traff-Graph

B.2 Travel Time Estiamtion

For the travel time estimation task, we used the following methods to evaluate the utility of the dataset:

- **TEMP:** The method counts trips with the same or nearby origin and destination areas and then estimates the travel time by averaging all related trips.
- **XGBoost:** The method takes travel information (e.g., distance, departure time, etc.) for each trip as input, and then estimates the travel time using an ensemble learning approach.
- **WDR:** This is a popular travel time estimation method, which estimate travel time through a combination of wide network, depth network, and recurrent network.
- **DepTTE:** This is one of the representative travel time estimation methods. It first converts the original GPS trajectory into a series of high-dimensional features, and then applies RNN to capture the spatial-temporal dependence.

Table 7: Data utility comparison by travel time estimation. Results are expressed as (original / synthetic / difference ratio)

Methods	TEMP	XGBoost	WDR	DeepTTE
RMSE	345.35 / 325.44 / 5.77%	282.11 / 266.57 / 5.51%	221.90 / 244.16 / 10.03%	169.26 / 172.51 / 1.92%
MAE	230.34 / 215.86 / 6.29%	190.42 / 179.46 / 5.76%	130.51 / 137.82 / 5.60%	108.25 / 112.71 / 4.12%
MAPE	22.21 / 21.14 / 4.82%	18.95 / 18.09 / 4.54%	12.12 / 12.60 / 3.96%	10.94 / 11.10 / 1.46%

Table 7 showcases the performance of the synthetic dataset in the context of travel time estimation for Xi’an city. Again, TEMP, XGBoost, WDR, and DeepTTE are employed for comparison, and the results are evaluated based on RMSE, MAE, and MAPE metrics. Although the absolute performance figures change due to the distinct characteristics of the new city data, the consistency between the original and synthetic datasets remains robust across the metrics and models. The difference ratio remains in a similar range, indicating flexibility and reliability of the Syn-Xi’an dataset when applied to different cities. Noteworthy is the reversed performance difference in RMSE for the WDR model, showing a 10.03% discrepancy. This divergence can be attributed to the unique spatial-temporal characteristics of the different city data, indicating the need for context-specific fine-tuning when transferring models between cities. Nonetheless, the synthetic dataset still provides a reasonable approximation for most models and metrics, demonstrating its generalizability across different urban environments. In summary, these findings further reinforce the utility of our synthetic dataset, establishing its value not only for travel time estimation but also for the broader scope of urban mobility research in various cities.

C Additional Visualization for Use Cases

This section shows the visualizations of use case results (Section 5 and Appendix B).

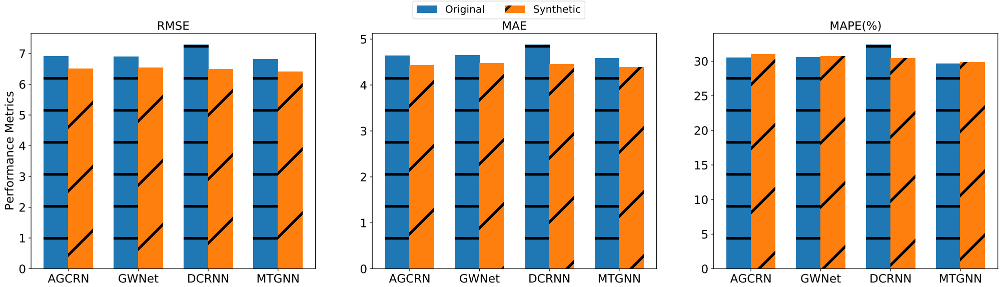


Figure 8: Performance metrics visualization for traffic demand prediction (Chengdu).

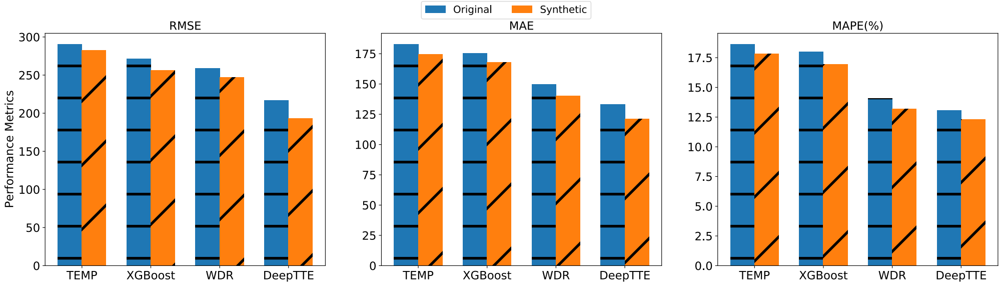


Figure 9: Performance metrics visualization for travel time estimation (Chengdu).

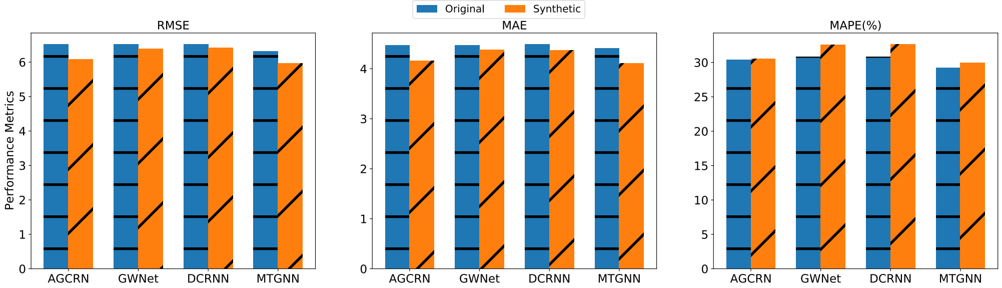


Figure 10: Performance metrics visualization for traffic demand prediction (Xi’an).

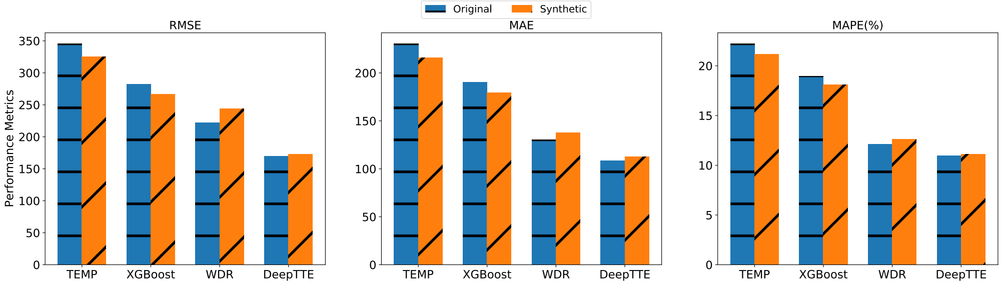


Figure 11: Performance metrics visualization for travel time estimation (Xi’an).