On the Influence of Grid Cell Size on Taxi Demand Prediction

Merlin Korth[1(✉)], Sören Schleibaum[1], Jörg P. Müller[1], and Rüdiger Ehlers[2]

[1] Department of Informatics, Clausthal University of Technology, Julius-Albert-Straße 4, 38678 Clausthal-Zellerfeld, Germany
{merlin.korth,soeren.schleibaum,joerg.mueller}@tu-clausthal.de
[2] Institute for Software and Systems Engineering, Clausthal University of Technology, Julius-Albert-Straße 4, 38678 Clausthal-Zellerfeld, Germany
ruediger.ehlers@tu-clausthal.de

Abstract. Accurate taxi demand prediction has the potential to increase customer satisfaction and hence the usage of ride-sharing by predicting the number of taxis needed at a certain place and time. When reviewing the related work on demand prediction, we observed that in taxi demand prediction different grid topologies – e.g. rectangular subdivisions of an area – and sizes are applied. However, it is not clear how and why the grid cells are configured the way they are and a systematic comparison of different topologies and sizes as regards their influence on urban demand prediction is lacking.

In this paper, we compare the influence of different grid cell sizes – 250 m, 500 m, and 1000 m – on the prediction accuracy of different types of deep learning-based taxi demand prediction approaches, such as convolutional neural networks, recurrent neural networks, and graph neural networks. Therefore, we select five deep learning-based approaches from related work and evaluate their performance on the New York City TLC taxi trip dataset and three different evaluation metrics. Our results show that approaches with a grid cell of size 1000 m and 500 m achieve a higher prediction accuracy. Furthermore, we propose to consider the grid cell size as a tunable parameter in demand prediction models.

Keywords: Taxi demand prediction · Grid cell size · Deep learning

1 Introduction

About 99% of the world's population breathes air that does not meet the World Health Organizations' air quality guidelines [23]. A major cause is urbanization, which includes the concentration of pollution sources in a relatively small

This work was in part supported by the Deutsche Forschungsgemeinschaft under grant 227198829/GRK1931. The SocialCars Research Training Group focuses on future mobility concepts through cooperative approaches.

I. M. Pires et al. (Eds.): GOODTECHS 2022, LNICST 476, pp. 19–36, 2023.
https://doi.org/10.1007/978-3-031-28813-5_2

(urban) area. In this context, a substantial part of air pollution is caused by the transportation sector [22].

One option to contribute to a reduction of air pollution in cities is to strengthen mobility-on-demand services [29] like ride-sharing in which multiple passengers share a vehicle or taxi. To decrease customers' waiting time and thereby increase the service's popularity, the number of requests at certain locations or the demand for taxis can be predicted; this prediction can be used to proactively reposition idle taxis to locations in which the demand exceeds the supply.

Predicting the demand for taxis is challenging as the temporal and spatial imbalance between demand and supply increases e.g. with a loss of experienced drivers due to demographic change [5,7]. Additionally, taxi drivers generally have a less efficient passenger search strategy in less known neighborhoods [5,7]. As the popularity of mobility-on-demand services like taxi ride-sharing increases [2,21,33], predicting the demand for taxis becomes even more relevant.

Usually, the demand is predicted for a short term like the next 30 min based on the previous demand, for instance in the last two hours. Many approaches train neural networks based on historic datasets with millions of taxi trips. Typically, the area of a city is spatially separated into multiple non-overlapping areas – like a 1000 m square grid cell – and the number of trips in each of these areas is predicted for the next time step. The grid cell size is understood as the edge length of the cell, e.g. a 1000 m sized square grid cell corresponds to an area of 1 km^2.

In Sect. 2, we present related work about taxi demand prediction to show that while using square/rectangular grids to spatially structure the demand is common, the chosen grid cell sizes vary from 150 m to 4900 m. In Sect. 3, we present the methodology for studying the influence of the chosen grid cell size on the prediction accuracy of multiple demand prediction approaches. The experimental results are presented in Sect. 4 and discussed in Sect. 5. Finally, a conclusion is given in Sect. 6.

2 Related Work

In recent years, researchers have made numerous advances in the field of taxi demand prediction [34]. Various approaches have been developed, which use, e.g., statistical time series analysis or making forecasts about the future by analyzing traffic data and using neural networks. As shown by [24] and in Table 1, while the commonly used evaluation metrics – Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), or Mean Relative Error (MRE) – are obvious, there is a wider variation of the chosen network types. The most common network types are Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), and Graph Convolutional Neural Network (GCN). In the following, we describe related work on short-term taxi demand prediction for each of these three network types.

2.1 Convolutional Neural Networks

CNN models are mainly used in image and speech recognition. As taxi demand data organized by a square grid has a structure similar to an image [36], CNNs

Table 1. Overview of grid configurations of existing demand prediction approaches.

Reference	Grid topology	Cell size in meter	Evaluation metric	Basic method	Code
Chu, Lam, and Li [4]	Rectangualr	220 × 170	RMSE, SMAPE	LSTM & CNN	✗
Davis, Raina, and Jagannathan [5]	Rectangular	1200× 600, 4900× 4900	RMSE, SMAPE, MASE	LSTM	✗
Jin et al. [8]	Rectangular	560× 475†	RMSE, MAE	GCN	✗
Ke et al. [9]	Polygons	N/A	RMSE, MAPE, MAE	GCN	✗
Ke et al. [10]	Polygons	N/A	RMSE, MAPE, MAE	GCN	✗
Ke et al. [11]	Square	4770 × 4810	RMSE, MAE, R-squared	LSTM & CNN	✗
Lee et al. [13]	Square	700× 700	RMSE, MAPE	GCN	✗
Li and Axhausen [14]	Square	500× 500	RMSE, SMAPE	GCN	✗
Oda and Joe-Wong [16]	Square	150× 150	RMSE	CNN	✗
Pian and Wu [17]	Square	1000× 1000	RMSE, MAPE, MAE	GCN	✓
Wang et al. [26]	Rectangular	2650× 2600	RMSE, SMAPE	CNN	✓
Wu, Zhu, and Chen [27]	Square	700× 700	RMSE, MAPE	GCN	✗
Xu and Li [30]	Square	300× 300	RMSE, MAE, R-squared	GCN	✗
Xu et al. [28]	Square	153× 153	RMSE, SMAPE	LSTM	✗
Yao et al. [31]	Square	700× 700	RMSE, MAPE	LSTM & CNN	✓
Ye et al. [32]	Rectangular	720 × 420	RMSE, PCC	LSTM & CNN	✗
Zhang et al. [34]	Rectangular	890 × 910†	RMSE, MAPE, MAE	RNN	✓
Zhang et al. [35]	Rectangular	250 × 334†	RMSE, MAE	LSTM	✓
Zhang, Liu, and Zheng [37]	Square	1600 × 1600	RMSE, MAPE, MAE	CNN	✗

(S)MAPE: (Symmetric) Mean Absolute Percentage Error, MA(S)E: Mean Absolute (Scaled) Error, PCC: Pearson Correlation Coefficient, RMSE: Root Mean Square Error, N/A: Not Available

†Estimated based on coordinates

are also commonly used for taxi demand prediction due to their ability to identify spatial correlations [32].

Many authors use CNNs as demand prediction models or as building blocks for their network, such as [16,25,31,32,37]. As shown in Table 1, all CNN-based approaches use a square or rectangular grid topology, but the chosen grid cell size varies from 700 m in [31] to 1600 m in [25]. Interestingly, for most approaches CNN layers are the main building block – [16,25,37] – but some combine it with LSTM layers – [31,32]; the latter shows promising results for the prediction accuracy.

2.2 Recurrent Neural or Long Short-Term Memory Networks

While the usage of CNNs is often motivated by their ability to capture the spatial relation, Recurrent Neural Network (RNN)-based models are used to identify temporal relations in the training data [11,28,34]. While both CNNs and RNNs are able to process a sequence of input maps that represent the demand per grid cell and time step, RNNs process the input sequentially. In particular, they are able to save extracted information from previous inputs in their memory. As the classical RNN architecture suffers from the vanishing gradient problem [12], the architecture was enhanced to the Long Short-Term Memory Network architecture by Hochreiter and Schmidhuber [6]. As taxi demand prediction can be formalized as a sequence prediction problem, both RNNs and LSTM are suitable for this task [34].

Besides [31,32] that combine CNN and LSTM building blocks and were described above, [5,11,28,35] create an LSTM or an RNN – [34]. While all approaches use a square or rectangular grid, the chosen grid cell sizes vary from 153 m in [28] to 4900 m in [5]; three approaches – [31,32,34] – chose a grid cell size between 700 m and 900 m.

2.3 Graph Neural Networks

According to Li and Axhausen [14], CNN models have the limitation of using regular grids that compress the information, rather than irregularly shaped grids that may more closely approximate reality, e.g. the boroughs of a city; in contrast to CNNs, GCNs are able to handle such data. As taxi demand data are sometimes structured in irregularly shaped grids - like in the more recent records of the commonly used NYC Yellow Taxi Trip dataset [15] - and as shown in Table 1, GCNs are used to predict taxi demand. In graph neural networks the value predicted per node is computed over the values of its neighbour. GCNs represent spatial connections by non-euclidean graph structures and apply convolutional operations. The basic form of a GCN layer consists of graph convolution, a linear layer, and a nonlinear activation function [38]. While [9,10] make use of the graph networks' advantage by structuring the data with irregular polygons, [14] select a 500 m square grid. Surprisingly, all GCN-based approaches – [8,13,17,27,30] – do similar; the chosen grid cell sizes vary from 200 m to 1000 m.

2.4 Research Gap, Question, and Hypothesis

While many researchers use deep learning-based methods to predict the demand, the used grid cell sizes vary from 150 m in [16] to 4900 m in [5]; these variations also differ among the different neural network types commonly used. Surprisingly, only two approaches – [3,30] – consider the grid cell size as an optimizable parameter and optimize it: Chiang, Hoang, and Lim [3] compared grid cell sizes from 250 m to 4000 m and selected 250 m as it achieved the highest prediction accuracy measured via the metric perplexity. Xu and Li [30] compare grid cell sizes from 100 m to 400 m and select 300 m because the RMSE is the lowest.

Comparing the prediction accuracy measured on different grid cell sizes does not work per default; in Sect. 3.2 and Fig. 2 we illustrate this problem and propose an aggregation step to enable comparability. Neither [3] nor [30] describe such a step. Consequently, we assume that their methodology for choosing a grid cell size was not complete.

A change in the grid cell size affects (1) the average number of trips per cell, (2) the number of cells in which no trip occurs at a certain time step, as well as (3) the ability of a demand prediction model to capture spatial demand patterns in a city. Because of this, the high variability of grid cell sizes used by others, and the lack of a methodologically complete comparison as regards their influence, we consider the influence of grid cell sizes on the accuracy of demand prediction models as an open research gap.

This gap leads us to the following research question: **How does the grid cell size affects the accuracy of various neural network types for taxi demand prediction measured via MAE, MRE, and RMSE?**

In the context of this study, we restrict ourselves to the grid cell sizes of 250, 500 and 1000 m, as these are representative of the grid cell sizes used in the literature.

Smaller grid cell sizes might enable a demand prediction model to better capture the spatial patterns of a city; while a larger grid cell size might be sufficient to separate the demand patterns between city districts, a finer grid cell size could allow the separation of demand patterns in neighborhoods. Consequently, we select the hypothesis **H1: Smaller grid cell sizes achieve a higher prediction accuracy** to address the aforementioned research question.

3 Methodology

Here, we describe the dataset used in our experiment, as well as the evaluation metrics and aggregation step that enables a fair comparison of the results achieved on different grid cell sizes. Furthermore, we select five demand prediction approaches from the related work of the previous section and three baseline models. For each of the five selected models, we select the optimal grid cell size via measuring the prediction accuracy on the validation data. The final results are measured on the previously unseen test data. The parameters of each model were chosen with respect to the configurations of the authors of the respective model. Therefore, no optimization of the hyperparameters was performed. In general, the batch size is set to 256 and the demand is predicted for the next 30 min. Additionally, all external factors, such as weather information, were excluded from the prediction to allow for a consistent comparability of the grid cell configurations.

3.1 Dataset

We use the *NYC Yellow Taxi Trip Data* [15], which was created by the New York City Taxi & Limousine Commission. It is one of the most widely used datasets in the field of taxi demand prediction. We consider 18 months, starting from January 2015 and until June 2016. Almost 70% – about 12 months – of the data will be used for training the networks, about 20% – about 4 months – for validation, and the remaining 10% – about 2 months – for testing. To prepare the data and exclude outliers, we apply the same strategy as S. Schleibaum, J. P. Müller, and M. Sester [19]; that results in about 3.5% as outliers. In particular, we enhance the trip records by the indices of the grids with a cell size of 250, 500, and 1000 m. We select a square with the bottom left (40.5879, −74.0898) and the top right (40.9014, −73.6857) as the area in which the trips have to start. Figure 1 shows the absolute taxi demand in the area via a grid cell size of 250 m, which is equal to an area of 0.0625 m^2.

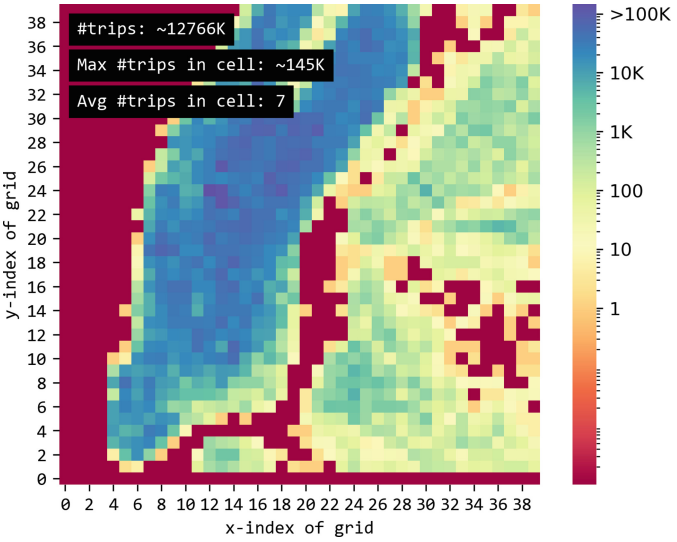

Fig. 1. Absolute number of taxi trips in New York City between January 2015 to June 2016 shown on a grid with cells size of 250 m and on a logarithmic scale

3.2 Evaluation

Metrics. To evaluate the performance, we use the three evaluation metrics commonly used: the MAE, MRE, and RMSE. In this approach y is used as the actual value, $\hat{y}$ as the predicted value, and N is the total number of values in the range of $n = 0, \ldots, N-1$. The evaluation metrics are defined as follows: MAE as $\frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$, MRE as $\frac{1}{N}\sum_{n=1}^{N}\frac{|y_n-\hat{y}_n|}{\hat{y}_n}$, and RMSE as $\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n-\hat{y}_n)^2}$.

Illustration of the Comparability Problem. As shown in Fig. 2 the grid on the left and the grid in the middle are considered as two separate predictions. For example, comparing the prediction accuracy of the two cells with the MAE – $MAE^{NoAgg.}_{1000\,m}$ of 50 and $MAE^{NoAgg.}_{500\,m}$ of 12.5 – we see that the prediction accuracy is higher for a grid cell size of 500 m. However, that is incorrect because both, the true demand of one cell and the total predicted demand of all cells, are the same in both predictions. Consequently, predictions made on different grid cell sizes are not comparable by default.

Aggregation Step. To make the results on the two grid cell sizes comparable, we aggregate the four 500 m grid cells to the right grid cell in Fig. 2 by computing the two sums $\sum_{n=0}^{3} y_n$ and $\sum_{n=0}^{3} \hat{y}_n$. Now, comparing MAE – $MAE^{NoAgg.}_{1000\,m}$ of 50 and $MAE^{Agg.}_{1000\,m}$ of 50 – we see that the prediction accuracy is the same for the 1000 m sized grid cell and for the aggregated 1000 m sized grid cell.

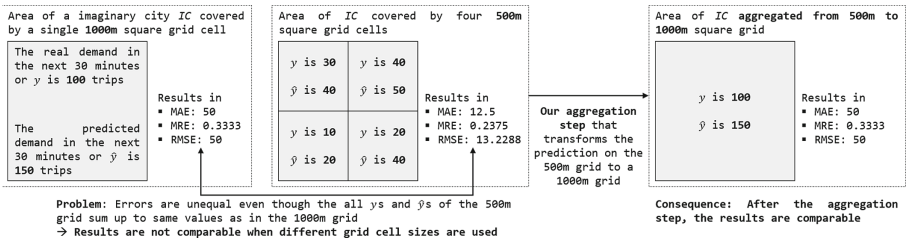

Fig. 2. Illustration of problem when comparing results from different grid cell sizes and our aggregation step to tackle this problem

Consequently, we aggregate the grid cells that occupy the same space as the largest grid cell size based on the geographical location of the grid cells. Technically, we apply two-dimensional average pooling and multiply the results by the number of cells aggregated, allowing the predicted value of a grid cell to be compared with the reference grid cell of size 1000 m.

3.3 Selection of Demand Prediction Approaches

The criteria for selecting demand prediction approaches from the related work are (C1) that we have at least one model per common network type – CNN, RNN, and GCN (C2) the availability of the source code so that we are able to reproduce the approach, and (C3) the prediction accuracy reported by the authors.

In Table 1, we list only one RNN-based model - corresponds to (C1) - proposed by Zhang et al. [34] and henceforth called *M1-MLRNN*. As the source code is available (C2) and the model outperformed simple LSTM models as well as ConvLSTM [20] and St-ResNet [36] (C3), we select this model. The authors first build a Spearman Correlation Coefficient matrix from historical taxi demand data, and thereby derive several clusters of taxi zones using a deterministic annealing algorithm. Afterwards, several modules composed of RNN and Fully Connected (FC) layers are built – a global one and one per cluster. Per grid cell, the global and cluster-based predictions are averaged. Thereby the *M1-MLRNN* combines global predictions with the cluster ones to predict the overall demand for pickups in a city.

As shown in Table 1, we list three purely and four partialy LSTM-based models (C1). Still, there is only one purely LSTM-based model proposed by Zhang et al. [35] – referred to as *M2-pmlLSTM* – of which the source code is available (C2). The authors were able to outperform a model that combines CNN and LSTM layers (C3). In contrast to [34], temporal classification is used instead of spatial clustering. The corresponding classifier time-feature encoder consists of FC layers that map demand data to a time class. Additionally, the demand is enhanced by denoising and passed through an LSTM layer. Both outputs are fused via another LSTM and FC layer and combined to the final prediction. Interestingly, the pick-up and drop-off demand are predicted and used as input.

The third selected model proposed by Pian and Wu [17] – named *M3-STDGAT* – represents the network type GCN (C1). We select this model as it is the only GCN-based network of which the code is available (C2) and it also outperforms a relatively simple GCN model and the DMVST model of Yao et al. [31], which consists of both LSTM and CNN layers (C3). The *M3-STDGAT* is based on a dynamic graph structure to identify dynamic time-specific spatial relations throughout the timeline. Therby, the authors use a Graph Attention Network (GAT) to identify adaptive importance allocation for neighboring regions based on pair-wise calculated correlations. The *M3-STDGAT* is composed of a spatial and a temporal module as well as a prediction layer, which combines the output of the former two. The spatial module aims to capture spatial patterns and is composed of several GAT blocks. The temporal module is based on an LSTM to determine temporal patterns in the demand.

Table 1 shows that the fourth selected model – presented by Wang, Hou, and Barth [25] and henceforth called *M4-CNNFC* – is the only CNN-based model (C1) of which the source code is available (C2). With this relatively simple CNN-based model, the authors were able to outperform an LSTM-based model (C3). After each layer of the *M4-CNNFC*, a pooling layer is applied to compress the information. After the two CNN layers, a FC is used to incorporate additional weather information and generate the demand prediction.

Although we have already introduced a CNN-based model, we select a second CNN-based model, which is presented by Oda and Joe-Wong [16]. It consists solely of three CNN layers and in contrast to [25], no pooling layers are used (C1). Therefore, we could successfully reproduce the model (C2). Despite the simple design, very good results were achieved (C3). This model does not use any additional features.

3.4 Baseline Models

The decision tree-based approach XGBoost of Chen and Guestrin [1] – referred to as *M6-XGB* – is selected as the first baseline model. Here, different machine learning concepts are applied, such as ensemble learning and gradient boosting. Further, XGBoost was also used by [9,17,34] as a baseline model. Additionally, we use relatively simple regression models as baseline: *M7-Ridge* and *M8-Lasso*. Both were previously used as baselines for demand prediction models, for instance by [9,17,31].

4 Experimental Results

The results of the experiments of the five selected models are described below in ascending order of model numbering: *M1-MLRNN*, *M2-pmlLSTM*, *M3-STDGAT*, *M4-CNNFC*, and *M5-CNN*. An overview of the results described in this section is given in Table 2.

Table 2. Comparison of the results of the experiments with validation data.

Model	Grid size	MAE [#trips]	MRE	RMSE
M1-MLRNN-C4	1000 m	1.4054	0.2911	3.172
M1-MLRNN-C3	500 m	**1.2114**	**0.2529**	**2.6177**
M1-MLRNN-C3	250 m	1.2544	0.2616	2.6248
M1-MLRNN-C1	1000 m	**1.1827**	**0.2424**	**2.55**
	500 m	1.1905	0.2425	2.5726
	250 m	1.2428	0.2581	2.7089
M2-pmlLSTM	1000 m	**1.1891**	**0.2443**	**2.5978**
	500 m	1.1971	0.2446	2.5967
	250 m	1.2570	0.2566	2.7619
M3-STDGAT	1000 m	1.1907	0.2483	2.5492
	500 m	$N/A^{\dagger}$	$N/A^{\dagger}$	$N/A^{\dagger}$
	250 m	$N/A^{\dagger}$	$N/A^{\dagger}$	$N/A^{\dagger}$
M4-CNNFC	1000 m	1.3064	0.2705	2.8142
	500 m	**1.2942**	**0.2655**	**2.7404**
	250 m	1.3393	0.2761	2.8516
M5-CNN	1000 m	**1.2082**	**0.2484**	**2.678**
	500 m	1.2224	0.2552	2.7274
	250 m	1.3146	0.2705	2.9446

† Execution not possible

4.1 M1-MLRNN: Multi-level Recurrent Neural Network

To rebuild *M1-MLRNN*, we had to adapt the model to our dataset, which differs in size and cell configuration from the experiments conducted in [34]. As argued in the previous section, we exclude weather and time information initially used in the approach. As the number of clusters depends on the dataset, we reproduce the deterministic annealing approach used. The results per grid cell size and for up to 12 clusters are shown in Fig. 3.

Suprisingly, for all grid cell sizes the use of a single cluster or no clustering achieved the highest prediction accuracy. Overall, we notice a strong fluctuation of the achieved MRE values. In Fig. 4 the pick-up demand aggregated over all cells of cluster are shown for six clusters and grid cells of size 1000 m. While some clusters are clearly separable – e.g. 2 from 6 – others – e.g. 3 from 6 – are not. The non-uniform behaviour of a cluster over time may cause the model to learn a distorted assumption about the behaviour of all cells. A prediction specific to this cluster based on the supposed uniform behavior will lead to a less accurate prediction.

Clustering is an essential part of the approach of Zhang et al. [34], but we were not able to reproduce its effectiveness. Consequently, we consider both *M1-MLRNN* with and without clustering to determine the influence of the grid cell

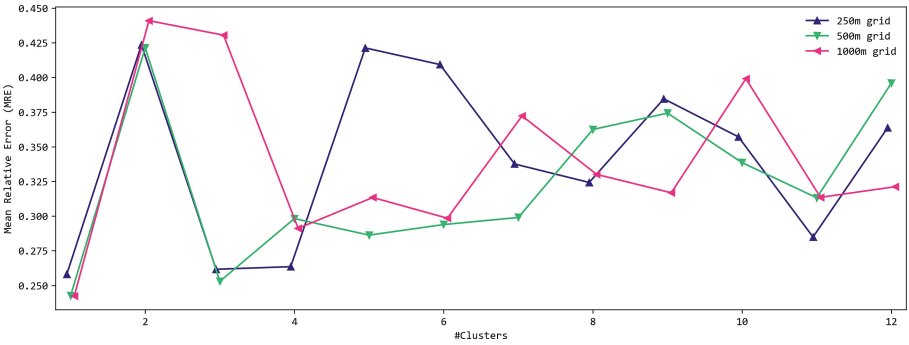

Fig. 3. *M1-MLRNN*: Selection of cluster size per grid cell size via MRE

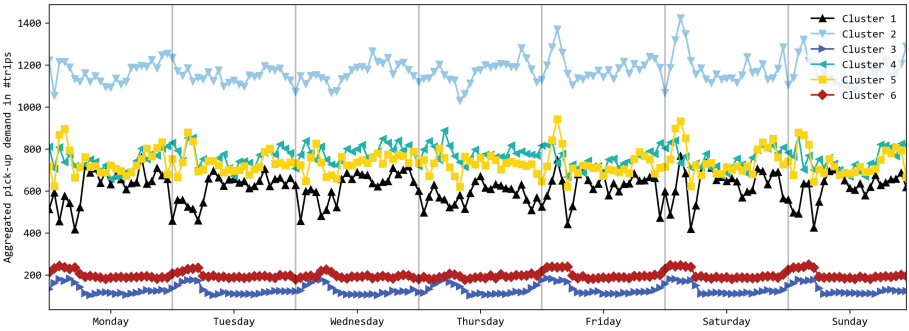

Fig. 4. Pick-up demand aggregated over a week for potential cluster sizes of the *M1-MLRNN* with a grid cell size of 1000 m

size on the *M1-MLRNN*. To distinguish between both configurations, we add the suffix -Cx, in which x represents the number of clusters used.

First, we study the results with the second-best number of clusters and secondly without clustering, which is equivalent to using exactly one cluster. Based on the results shown in Fig. 3, we choose the model configuration for the experiments with applied clustering to *M1-MLRNN-C4* in combination with 1000 m, and *M1-MLRNN-C3* with 500 m and 250 m. As shown in Table 2, for the applied clustering 500 m has the lowest MRE with a value of 0.2529 and the highest accuracy. In *M1-MLRNN-C1*, however, a different result is indicated. Here, the lowest MRE of 0.2529 is achieved with a grid cell size of 1000 m.

4.2 M2-pmlLSTM: Parallel Multi-task Learning Model

Also for the *M2-pmlLSTM* we had to do some adaptations before being able to run it on our dataset. This applies in particular to the time feature encoder. Additionally, as the model predicts both pick-up and drop-off demand, we adjusted the model to evaluate only the prediction accuracy of the pick-up demand. The

experiments conducted with *M2-pmlLSTM* provide results similar to those of *M1-MLRNN-C1*. While the difference between the 250 m configuration and the 500 m/1000 m is relatively clear in all three evaluation metrics, the difference between the 500 m and 1000 m configuration is not. Even though the differences are small, we select the 1000 m grid cell size as the difference in the MAE is relatively large. With this configuration, the *M2-pmlLSTM* achieves an MAE of 1.1891, an MRE of 0.2443, and an RMSE of 2.5978 on the validation data.

4.3 M3-STDGAT: Spatial-Temporal Dynamic Graph Attention Network

The *M3-STDGAT* receives the previous demand of the last 2.5 h as a matrix and the corresponding adjacency matrix, which represents the connection between the grid cells. While the number of trainable parameters in GAT-based networks is independent of the input size, the incorporated RNN layer that increases this number significantly – from 6.6 M to 104.9 M parameters. For this reason, among others, we were unfortunately not able to conduct the experiments for *M3-STDGAT* completely. When using a 500 m grid instead of a 1000 m one, the total number of trainable parameters increases by 1393% leading to a required memory size of more than 64 GB. Additionally, the approximated time to calculate one training epoch is about 11 hours, if we use the grid cell size configuration of 500 m in a scaled-down version of the model. Therefore, a complete experiment would take more than 92 days. Consequently, we exclude the configuration of 500 m, and 250 m from the experiments. As shown in Table 2 the remaining results are an MAE of 1.2245, an MRE of 0.2492, and an RMSE of 2.7158.

4.4 M4-CNNFC: Convolutional Neural Network with Fully Connected Layer

To conduct the experiments with *M4-CNNFC* [26] we adjusted the network and, similar to the rebuilding of the other models, removed additional features. The results show that larger grid cells do not necessarily increase the prediction accuracy. As shown in Table 2, all evaluation metrics for the grid cell size of 500 m are smaller than for 1000 m, and 250 m – $MAE_{500\,m}$ of 1.2942, $MAE_{1000\,m}$ of 1.3064, and $MAE_{250\,m}$ of 1.3393.

4.5 M5-CNN: Convolutional Neural Network

In contrast to the results of the experiments of the *M4-CNNFC*, the *M5-CNN* achieves the highest prediction accuracy with a 1000 m grid cell size. As shown in Table 2, all error metrics are smaller for the cell size of 1000 m than for 500 m and 250 m – $MRE_{1000\,m}$ of 0.2484 vs. $MRE_{500\,m}$ of 0.2552 vs. $MRE_{250\,m}$ 0.2705.

5 Discussion

We will first discuss the RNN/LSTM-based models, followed by the CNN-based ones. Then the results of the RNN/LSTM and CNN-based models are discussed among each other. Finally, limitations are pointed out and interesting options for future work are proposed. However, due to the lack of experimental results for the grid cell sizes of 250 m and 500 m of *M3-STDGAT*, no discussion can be made for this model.

5.1 RNN-Based Models

As regards the *M1-MLRNN* derivatives from Zhang et al. [34], we achieve the highest prediction accuracy when *no* clustering is applied and a 1000 m grid cell size is used. If clustering is applied a grid cell size of 500 m achieves the highest prediction accuracy. It is worth noting that we were not able to reproduce the positive effect of clustering. As mentioned in Sect. 4.1, this is probably caused by the use of deterministic annealing as a clustering method, which seems to fail to perform sufficiently. It needs to be noted that deterministic annealing cannot guarantee to be able to find a minimum if more than one local minimum exists at a given temperature [18]. For this reason, deterministic annealing may not find the optimal clustering solution in our case.

In comparison, the experiments of *M2-pmlLSTM* show that the grid cell size of 1000 m leads to the highest prediction accuracy. This confirms that for RNN-based models, rather larger grid cell sizes are preferred over smaller ones.

In the side-by-side analysis of *M1-MLRNN* and *M2-pmlLSTM*, it can be found that the difference between the cell size of 1000 m, and 500 m is relatively small. Therefore, based on the results described neither for *M1-MLRNN* nor for *M2-pmlLSTM* is it possible to unambiguously answer the question of whether a large or a medium cell size provides the best result for RNN/LSTM-based models.

The results show that for both models hypothesis H1 – smaller grid cells allow for higher prediction accuracy – has to be rejected, as either the largest or a medium grid cell size is preferred.

5.2 CNN-Based Models

The results of the experiments of *M4-CNNFC* [26] show that the grid cell size configuration of 500 m is significantly better than for other grid cell sizes. In contrast, the largest size of the grid cell for *M5-CNN* [16] is considerably better. The two models, and thus the results, differ mostly in that *M4-CNNFC* [26] uses an FC prediction layer in addition to CNN layers; whereas, *M5-CNN* [16] consists of only three CNN layers. Therefore, when using only CNN layers, larger grid cell sizes should be selected.

Similar to the RNN-based models, the results for CNN-based models show that for both models, hypothesis H1 – smaller grid cells allow for higher prediction accuracy – has to be rejected, as the largest – *M5-CNN* – and medium – *M4-CNNFC* – grid cell size is preferred.

5.3 Comparison Among Models and to Baselines

The results from the experiments with testing data shown in Table 3 reveal that the results for *M1-MLRNN-C1*, *M2-pmlLSTM*, and *M3-STDGAT* are almost the same, as they all have the highest prediction accuracy by using the largest grid cell size. Furthermore, the results show that these three more complex models perform the best followed by *M1-MLRNN-C3*, *M5-CNN*, and *M4-CNNFC*. *M1-MLRNN-C3* and *M4-CNNFC* performed best when using the medium-sized grid cell size and *M5-CNN* when using the largest grid cell size.

When comparing these five models with the three baseline models, the advantage of deep learning-based models is apparent. The best result of the baseline models was obtained with *M7-Ridge* – MAE_{M7} of 1.2997 – followed by *M8-Lasso* and *M6-XGB* as shown in Table 3. If we compare the strongest baseline model *M7-Ridge* to the four models examined, we can see that *M7-Ridge* performs better than *M4-CNNFC* in only two of three metrics – MAE and MRE –, but also undercuts the MRE of *M1-MLRNN-C1* by 0.0029. Compared to *M1-MLRNN-C1*, however, the MAE is higher by 0.0399.

Although the considered models work differently depending on the grid cell configuration, our hypothesis H1 – smaller grid cells enable a higher prediction accuracy – needs to be rejected. Instead, the grid cell size should be considered individually for each model configuration. Especially since the results are partly close to each other, we assume that the influences of the individual components can strengthen or weaken each other. This is particularly important when different models are used as base models. Assumably, in the majority of papers, the grid cell size was chosen independently of the models. The results of this paper contradict this approach and indicate that a fair comparison between models is not possible if one grid cell size is chosen for all models, which could suppress good results for some models.

In addition, the grid cell size is often predefined based on the application. Therefore, it is recommended, according to our results, to choose a model that performs best for the specific grid cell size.

Another important aspect in comparing the models is the usage of pick-up and drop-off demand by *M1-MLRNN*, *M2-pmlLSTM*, and *M3-STDGAT*. Whereas *M4-CNNFC* and *M5-CNN* are based exclusively on the pick-up demand. Thus, the pick-up demand prediction is extended by the drop-off demand feature. In this work a consideration of the drop-off data as external features was not carried out, as the dropoff data is an essential part of the model structure. Nevertheless, we expect the drop-off data to influence the accuracy.

5.4 Limitations and Future Work

In this paper, we limited the comparison to three different grid cell configurations in terms of size – 1000 m, 500 m, and 250 m. However, the results of the experiments show that medium-sized and large-sized cells have a positive influence on the prediction accuracy of different models. Therefore, intermediate grid cell sizes between 500 m and 1000 m – such as 700 m following Wu, Zhu, and Chen

Table 3. Comparison of the models on the best grid cell size among each other and to the baseline models on the test data.

Model	Grid size	MAE [#trips]	MRE	RMSE
M1-MLRNN-C3	500 m	1.2598	0.2588	2.7588
M1-MLRNN-C1	1000 m	1.2341	0.2437	2.6990
M2-pmlLSTM	1000 m	1.2339	0.2433	2.6911
M3-STDGAT	1000 m	**1.2338**	**0.2410**	**2.6796**
M4-CNNFC	500 m	1.3416	0.2638	2.8367
M5-CNN	1000 m	1.2630	0.2486	2.8081
M6-XGB	1000 m†	1.5887	0.3128	3.1796
M7-Ridge	1000 m†	1.2997	0.2559	2.8400
M8-Lasso	1000 m†	1.4733	0.2901	3.1122

† Other configurations were not tested

[27], Lee et al. [13], and Yao et al. [31] – and grid cell sizes larger than 1000 m – for example 1500 m or 2500 m – similar to Zhang, Liu, and Zheng [37] and Wang et al. [26] – could be chosen.

Linked to the grid cell configuration, a limitation of our approach is that we do not consider the influence of the angle at which the grid cells are placed on the city map. Possibly, there is an influence on the prediction accuracy depending on the orientation of the grid cells relative to the orientation of the main roads of a city.

Furthermore, models with different combinations of layer types could be investigated. Thus, a pure FC network could be considered, as well as, combinations of CNN layers with LSTM layers, or different kinds of GCNs with LSTM layers.

Furthermore, as already described in Sect. 5.3, *M1-MLRNN*, *M2-pmlLSTM* and *M3-STDGAT* include both pick-up and drop-off demand in the calculation, whereas *M4-CNNFC* and *M5-CNN* – as well as *M6-XGB*, *M7-Ridge*, and *M8-Lasso* – only use the pick-up demand. The usage of drop-off demand can be considered as using an additional feature, like weather, population density, and points of interest. Therefore, the influence of these external factors on the accuracy of the prediction in combination with the size of the grid cell could also be investigated.

Another aspect is the usage of different datasets. Here, only the TLC dataset from New York City is used in the experiments, which limits the general explanatory value. Using a dataset from, for example, China – like [17,31,37] – could potentially affect the results, as there is a different traffic behavior and a different topology of the city. Here, it was difficult for us to get access a large taxi trip dataset from China.

6 Conclusion

When combining different network types in a model – e.g. CNN or LSTM layers in combination with FC layers – we could obtain different conclusions. According to our results, the grid cell size should be considered individually for each model configuration. This is especially important to consider when different models are used to evaluate the performance of new models. Therefore, we propose to consider the grid cell size as a tunable parameter in demand prediction models. Still, following our results a large grid cell size should be considered for a CNN model and a large or medium grid cell size should be selected for RNN or LSTM models.

However, depending on the use case of demand prediction, a smaller or medium-sized grid cell size might even be better – despite a slight decrease in accuracy – as, for example, the repositioning of taxis in ride-sharing can be done more precisely. In future approaches, a finer subdivision of the grid cell size, the influence of drop-off demand, as well as the influence of the angle of the dataset, could be investigated more closely to further improve prediction accuracy.

References

1. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, New York, NY, USA, pp. 785–794. Association for Computing Machinery (2016). https://doi.org/10.1145/2939672.2939785. ISBN 9781450342322
2. Chen, W., Chen, J., Yin, G.: Exploring side effects of ridesharing services in urban China: role of pollution - averting behavior. Electron. Commer. Res. **12**(4), 317 (2020). https://doi.org/10.1007/s10660-020-09443-y. ISSN 1389-5753
3. Chiang, M.-F., Hoang, T.-A., Lim, E.-P.: Where are the passengers? A grid-based gaussian mixture model for taxi bookings. In: Ali, M., Huang, Y., Gertz, M., Renz, M., Sankaranarayanan, J. (eds.) Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York, NY, USA, pp. 1–10. ACM (2015). https://doi.org/10.1145/2820783.2820807. ISBN 9781450339674
4. Chu, K.F., Lam, A.Y.S., Li, V.O.K.: Travel demand prediction using deep multi-scale convolutional LSTM network. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 1402–1407. IEEE (2018). https://doi.org/10.1109/ITSC.2018.8569427. ISBN 978-1-7281-0321-1
5. Davis, N., Raina, G., Jagannathan, K.: Grids versus graphs: partitioning space for improved taxi demand-supply forecasts. IEEE Trans. Intell. Transp. Syst. **22**(10), 6526–6535 (2021). https://doi.org/10.1109/TITS.2020.2993798
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735. ISSN 0899-7667
7. Ishiguro, S., Kawasaki, S., Fukazawa, Y.: Taxi demand forecast using real-time population generated from cellular networks. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, New York, NY, USA, pp. 1024–1032. ACM (2018). https://doi.org/10.1145/3267305.3274157. ISBN 9781450359665

8. Jin, G., Xi, Z., Sha, H., Feng, Y., Huang, J.: Deep Multi-view spatiotemporal virtual graph neural network for significant citywide ride-hailing demand prediction. CoRR, abs/2007.15189 (2020)
9. Ke, J., Feng, S., Zhu, Z., Yang, H., Ye, J.: Joint predictions of multi-modal ride-hailing demands: a deep multi-task multi-graph learning-based approach. Transp. Rese. Part C Emerg. Technol. **127** (2021). https://doi.org/10.1016/j.trc.2021.103063. ISSN 0968-090X
10. Ke, J., Qin, X., Yang, H., Zheng, Z., Zhu, Z., Ye, J.: Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network (2019)
11. Ke, J., Zheng, H., Yang, H., Chen, X.M.: Short-term forecasting of passenger demand under on-demand ride services: a spatio-temporal deep learning approach. Transp. Res. Part C Emerg. Technol. **85**, 591–608 (2017). https://doi.org/10.1016/j.trc.2017.10.016. ISSN 0968-090X
12. Kolen, J.F., Kremer, S.C. (eds.) A Field Guide to Dynamical Recurrent Networks. IEEE (2009). https://doi.org/10.1109/9780470544037. ISBN 9780470544037
13. Lee, D., Jung, S., Cheon, Y., Kim, D., You, S.: Demand forecasting from spatiotemporal data with graph networks and temporal-guided embedding (2019)
14. Li, A., Axhausen, K.W.: Short-term traffic demand prediction using graph convolutional neural networks. AGILE GISci. Ser. **1**, 1–14 (2020). https://doi.org/10.5194/agile-giss-1-12-2020
15. NYC Taxi and Limousine Commission. TLC Trip Record Data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
16. Oda, T., Joe-Wong, C.: MOVI: a model-free approach to dynamic fleet management. In: IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, pp. 2708–2716 (2018). https://doi.org/10.1109/INFOCOM.2018.8485988
17. Pian, W., Wu, Y.: Spatial-temporal dynamic graph attention networks for ride-hailing demand prediction (2020)
18. Rose, K.: Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proc. IEEE **86**(11), 2210–2239 (1998). https://doi.org/10.1109/5.726788
19. Schleibaum, S., Müller, J.P., Sester, M.: Enhancing expressiveness of models for static route-free estimation of time of arrival in urban environments. Transp. Res. Proc. **62**, 432–441 (2022). https://doi.org/10.1016/j.trpro.2022.02.054, https://www.sciencedirect.com/science/article/pii/S2352146522001818. ISSN 2352-1465, 24th Euro Working Group on Transportation Meeting
20. Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS 2015, Cambridge, MA, USA, vol. 1, pp. 802–810. MIT Press (2015). https://dl.acm.org/doi/10.5555/2969239.2969329
21. Uber Technologies Inc., Financials (2021). https://investor.uber.com/financials/default.aspx
22. United Nations. Sustainable Transport, Sustainable Development. Interagency Report for Second Global Sustainable Transport Conference (2021). https://sdgs.un.org/publications/interagency-report-second-global-sustainable-transport-conference
23. United Nations. Billions of people still breathe unhealthy air: new WHO data (2022). https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data/

24. Varghese, V., Chikaraishi, M., Urata, J.: Deep learning in transport studies: a meta-analysis on the prediction accuracy. J. Big Data Anal. Transp. **2**(3), 199–220 (2020). https://doi.org/10.1007/s42421-020-00030-z. ISSN 2523-3556
25. Wang, C., Hou, Y., Barth, M.: Data-driven multi-step demand prediction for ride-hailing services using convolutional neural network. Adv. Comput. Vision 11–22 (2019). . https://doi.org/10.1007/978-3-030-17798-0_2, https://dx.doi.org/10.1007/978-3-030-17798-0_2. ISSN 2194-5365
26. Wang, Y., Yin, H., Chen, H., Wo, T., Xu, J., Zheng, K.: Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In: Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., Karypis, G. (eds.) Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, pp. 1227–1235. ACM (2019). https://doi.org/10.1145/3292500.3330877. ISBN 9781450362016
27. Wu, M., Zhu, C., Chen, L.: Multi-task spatial-temporal graph attention network for taxi demand prediction. In: Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, New York, NY, USA, pp. 224–228. ACM (2010). https://doi.org/10.1145/3395260.3395266. ISBN 9781450377072
28. Xu, J., Rahmatizadeh, R., Boloni, L., Turgut, D.: Real-time prediction of taxi demand using recurrent neural networks. IEEE Trans. Intell. Transp. Syst. **19**(8), 2572–2581 (2018). https://doi.org/10.1109/TITS.2017.2755684. ISSN 1524-9050
29. Xu, Y., Li, D.: Incorporating graph attention and recurrent architectures for city-wide taxi demand prediction. ISPRS Int. J. Geo-Inf. **8**(9) (2019). https://doi.org/10.3390/ijgi8090414, https://www.mdpi.com/2220-9964/8/9/414. ISSN 2220-9964
30. Ying, X., Li, D.: Incorporating graph attention and recurrent architectures for city-wide taxi demand prediction. ISPRS Int. J. Geo Inf. **8**(9), 414 (2019). https://doi.org/10.3390/ijgi8090414
31. Yao, H., et al.: Deep multi-view spatial-temporal network for taxi demand prediction. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 2588–2595. AAAI Press (2018). https://doi.org/10.1609/aaai.v32i1.11836
32. Ye, J., Sun, L., Du, B., Fu, Y., Tong, X., Xiong, H.: Co-prediction of multiple transportation demands based on deep spatio-temporal neural network. In: Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., Karypis, G. (eds.) Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, pp. 305–313. ACM (2019). https://doi.org/10.1145/3292500.3330887. ISBN 9781450362016
33. Zardini, G., Lanzetti, N., Pavone, M., Frazzoli, E.: Analysis and control of autonomous mobility-on-demand systems. Annu. Rev. Control Robot. Auton. Syst. **5**(1) (2022). https://doi.org/10.1146/annurev-control-042920-012811
34. Zhang, C., Zhu, F., Lv, Y., Ye, P., Wang, F.-Y.: MLRNN: taxi demand prediction based on multi-level deep learning and regional heterogeneity analysis. IEEE Trans. Intell. Transp. Syst. 1–11 (2021). https://doi.org/10.1109/TITS.2021.3080511. ISSN 1524-9050
35. Zhang, C., Zhu, F., Wang, X., Sun, L., Tang, H., Lv, Y.: Taxi demand prediction using parallel multi-task learning model. IEEE Trans. Intell. Transp. Syst. 1–10 (2020). https://doi.org/10.1109/TITS.2020.3015542. ISSN 1524-9050
36. Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017, pp. 1655–1661. AAAI Press (2017). https://doi.org/10.5555/3298239.3298479

37. Zhang, K., Liu, Z., Zheng, L.: Short-term prediction of passenger demand in multi-zone level: temporal convolutional neural network with multi-task learning. IEEE Trans. Intell. Transp. Syst. **21**(4), 1480–1490 (2020). https://doi.org/10.1109/TITS.2019.2909571. ISSN 1524-9050
38. Zhou, J., et al.: Graph neural networks: a review of methods and applications. AI Open **1**, 57–81 (2020). https://doi.org/10.1016/j.aiopen.2021.01.001, https://www.sciencedirect.com/science/article/pii/S2666651021000012. ISSN 2666-6510