Taxi Demand Forecast using Real-Time Population generated from Cellular Networks

Shin Ishiguro
NTT DOCOMO, INC.
Sanno Park Tower 11-1, Nagata-cho 2-chome, Chiyoda-ku,
Tokyo, Japan
shin.ishiguro.tb@nttdocomo.com

Yusuke Fukazawa
NTT DOCOMO, INC.
Sanno Park Tower 11-1, Nagata-cho 2-chome, Chiyoda-ku,
Tokyo, Japan
fukazawayu@nttdocomo.com

Satoshi Kawasaki
NTT DOCOMO, INC.
Sanno Park Tower 11-1, Nagata-cho 2-chome, Chiyoda-ku,
Tokyo, Japan
satoshi.kawasaki.vx@nttdocomo.com

UbiComp/ISWC '18 Adjunct, October 8-12, 2018, SINGAPORE. Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3575-1/15/09...$15.00.
http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

Abstract

For efficient operation of taxis, it is important to provide taxi drivers with detailed information about passenger demand. In this paper, we propose a future taxi demand prediction algorithm by using real-time population data generated from cellular networks. We evaluated the effects of real-time population data on the accuracy of taxi demand prediction by using stacked denoising autoencoders. The results of an offline experiment conducted herein indicate that when real-time population data were used, the root mean squared prediction error of the proposed algorithm was 1.370 as opposed to 1.513 when population data were not used. In addition, we conducted a field test. We implement a real-time prediction system based on real-time population data, the first such online real-world test conducted worldwide. In the trial, 26 participant drivers tried our demand forecast system. The results showed that the sales of participant drivers improved by 1,409 JPY per person per day, which represents a 3.9% increase in sales on average compared to the drivers who did not use the system.

Author Keywords
Urban Mobility; Intelligent Transportation System; Machine Learning; Transportation Demand Forecast; Real-Time Population; Mobile Spatial Statistics

ACM Classification Keywords
J.4 Social and Behavior Sciences.

Introduction

In recent years, owing to the aging of taxi drivers, training of young drivers is an urgent task from the viewpoint of maintaining the number of workers in the future [1]. Experienced drivers are implicitly aware of where to find more passengers to earn high profits. By contrast, less experienced drivers do not have enough knowledge, which means they end up picking only a few passengers, and therefore, their profits tend to be low. As a result, the possibility of them resigning could increase. Even skilled drivers lack knowledge about unfamiliar areas in which they drop-off passengers. There is a high possibility of them driving their taxi empty until they return to a familiar area. To fill up these gaps in profit due to lack of knowledge, taxi companies have attempted to document the knowledge of experienced drivers and use the documented knowledge as training materials for unexperienced drivers. However, doing so increases human labor costs, and the resulting materials cannot provide drivers with information about real-time passenger demand, which changes dynamically based on the activity of people in a given area. To provide drivers with passenger demand information, future taxi demand prediction methods have been proposed [2][3]. Using predicted demand information, drivers can efficiently operate even if they lack knowledge about the area in which they are driving. In addition, passengers benefit from the timely availability of taxis.

In the present study, we propose a method to predict taxi demand based on real-time population data of each area by processing massive amounts of mobile users' location data obtained from cellular networks. We considered that the demand for taxis is greatly affected by population of the area in which a taxi is operated. First, it is assumed that the demand for taxis will increase when a large event ends. If one tracks the transition population in real-time, the increase in taxi demand can be seen. Event-related demand is evident from the population fluctuation in the area without advance knowledge of the local event. Second, it is thought that taxi demand will increase when a traffic accident occurs or when train or metro services are delayed. It is thought that increases in taxi demand can be discovered quickly by tracking population flows in real time. Therefore, in the present study, we propose a method to predict taxi demand by using real-time population data generated from cellular networks.

Figure 1 (up) and (down) show graphs comparing the transitions of the numbers of people boarding taxis and the population in a given area. Figure 1 (up) shows the number of taxi rides and the population in a station, and it suggests that the two numbers move in a synchronized way. Here, we can confirm that the taxi demand in a given area is generally correlated with the population in that area. By contrast, in Figure 1 (down), the number of taxi rides increases after 5 h since the increase in population. The area considered here is a downtown area, and it is thought that in this area, taxis are hailed mainly in the night for a ride to home. The transitions of the population and the number of taxi

rides are thought to be closely related. However, the correlation is not uniform in all areas, and it is necessary to learn these differences by using the proposed demand prediction model.

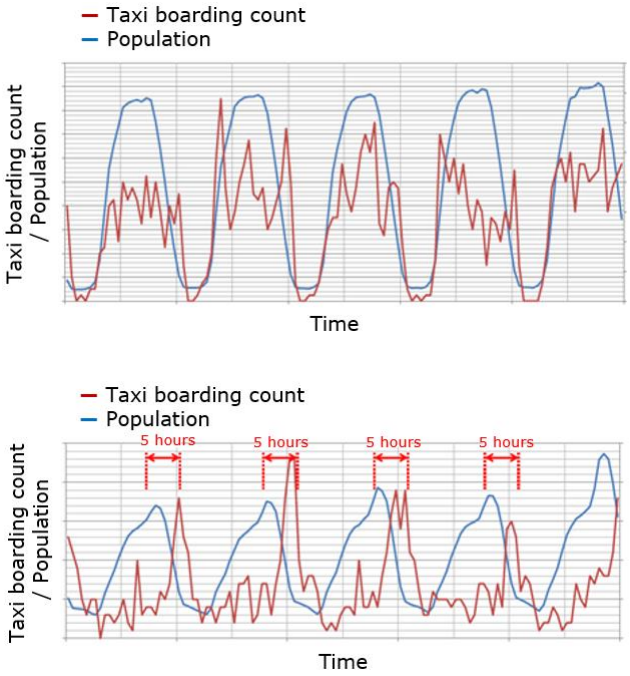

Figure 1: Taxi boarding count and population transition graph at location Shinbashi (up) and Sangenjaya (down). (Sep. 1–Sep. 5, 2016)

In addition, we conduct a field test in which real-time population data is used in an online application to solve a real problem. This is the first attempt of its type anywhere in the world. We implement a real-time taxi demand forecast system and test it with 26 participant drivers for four months across 23 wards in Tokyo**,** Musashino city, and Mitaka city. Here, we evaluate the efficiency of the proposed system from the viewpoint of sales. The contributions of the present study can be summarized as the following three points.

- To unify and abstract various data, such as taxi pick-up, real-time population, and meteorological data, a machine learning model based on stacked denoising autoencoders (SdA) is proposed.
- Taxi demand forecast accuracy is improved by using real-time population generated from cellular networks.
- A field test is conducted in which a real-time taxi demand forecast system is implemented, and this system is tested by participant drivers. The test results demonstrate the efficiency of the system in terms of improving driver sales.

Related Research

Several kinds of future taxi demand prediction methods have been proposed. Tong et al. [4] proposed methods by considering historical taxi operation data and weather data. Smith et al. [5] predicted taxi demand by using Twitter data to extract how much attention is focused in a given area at a given time. Yuan et al. [6] improved taxi prediction accuracy by using road geometries and Point-of-Interest (POI) information. Wang et al. [7] proposed the use of weather and traffic congestion information to predict taxi demand. Thus, past studies have suggested that in addition to taxi operation history, various types of data can contribute to improving the accuracy of taxi demand prediction. However, no studies have used real-time population data from a cellular network for taxi demand prediction.

Problem Setting

In the present study, we propose an algorithm to predict the taxi boarding count in each grid cell. As input data, we use real-time population, weather, and historical taxi demand data. We output the total boarding count in every 500-m grid cell over the coming 30 min. We defined prediction of the boarding count as a regression problem of continuous values. To overcome data sparsity and increase robustness to noise, we built a single model for all grids. The model shares the same neural network weights in all grid cells rather than generating multiple models for each grid cell. As the population data, we used population statistics data in which the population was estimated using data from cellular networks. These data were generated by estimating the number of mobile devices per area and at a given time with a spatial resolution of 500 m and time resolution of 10 min from the positional relationship between each base station and mobile terminals of the mobile network. In addition, we converted the actual population of Japan by adding the share rate of mobile operators. Furthermore, to consider privacy, any information or identifying individuals was deleted, and the t otal number of people in each area/at each time was calculated. With this data, it is possible to estimate the population throughout Japan with a latency of 30 minutes before the present.

Method

Data Preprocessing

UNIFYING SPATIAL AND TEMPORAL RESOLUTIONS

To create unified datasets from several types of datasets of different domains (taxi, population, and weather), we must consider the differences in spatial resolutions and temporal resolutions for each type of data. For taxi data, we calculated the total number of passengers getting on and off in the next 30 min at intervals of 10 min in each 500-m grid. We converted the population data into data of 10-min periods over 500-m grids. As the meteorological data, we used rainfall data of 1000-m grid cells. We converted these data into 500-m grid cells by dividing the meteorological figures of one 1000-m grid cell into four 500-m grid cells.

TIME RELATED FEATURES

To improve the accuracy of demand forecast, we implemented features related to time. Taxi demand in most cases changes cyclically in accordance with the current time and also based on short term or long term seasonal trends. To consider cyclic change, we implemented the current time feature and the current days of the week (weekday, holiday, day before holiday) feature. Current time was expressed in two dimensions (α, β) based on formula (2). t_{hour} and t_{minute} are hour and minute of the target time. t was calculated using equation (1).

$$t = t_{hour} + \frac{1}{60} t_{minute} \tag{1}$$

$$\alpha = \sin\frac{\pi}{12} t,\ \beta = \cos\frac{\pi}{12} t,\ \{0 \leq t < 24\} \tag{2}$$

Weekday information was represented by 2-bit data, which represent the weekday/holiday of a given day and the following day. As short-term trend, we used the actual values of taxi rides, population, and weather at 30 min and 60 min before, as well as 6 h before, the current time. As a long-term trend, we calculated the averages of taxi ride volume and population data of each grid on the same day of the week in the past year.

Attribute	Descriptions
Area	Tokyo 23 wards, Musashino city and Mitaka city
Period	Apr. 1st, 2015 - Sep. 14th, 2016
Taxi data	4,400 vehicles' location point data which have pick-up and drop-off position collected every 1 minute
Population data	500m grid-wise population data generated every 10 minutes
Weather data	1000m grid-wise weather data generated every 10 minutes

Table 1: Data Description in evaluation experiment

DATA NORMALIZATION

We normalized the input data x. For each data created using the above procedure, data normalization was performed to satisfy $-1 \leq x_i \leq 1$ by using (3).

$$\widehat{x_i} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (3)$$

Deep-Learning-based Demand Prediction Model

In this section, we explain the future taxi demand forecasting method that uses stacked auto encoders [8]. An autoencoder [9] is a neural network method that attempts to restore input data. In an autoencoder, the input data is reproduced in the output layer through a hidden layer. In the stacked autoencoder model, we build the neural network structure of the multilayer by stacking autoencoders. In this model, the output result of the autoencoder is used as the input data to an another autoencoder. The first layer learns from data input into the autoencoder. After obtaining the first layer, the output of the hidden layer is used as the input to the next hidden layer. In this way, the stacking of multiple layers of autoencoders is realized. In our taxi demand forecast algorithm, the value of demand is calculated using regression. Therefore, we added a regression prediction layer to the final layer of the developed stacked autoencoder model. We conducted supervised learning by inputting the value of taxi demand as the objective variable. In addition, we fine-tuned the network.

Denoising Autoencoder

To reduce effect of noise on the input data, we adopted the stacked denoising autoencoder method [10]. A denoising autoencoder performs processing to restore original data from missing data or the data to which noise is added. In this way, it can learn the abstract representation of an input x that is robust to both noise and missing original data. In addition, it is possible to extract important information in the process of restoring the original data.

Mini-batch Training

A method called mini-batch training improves the generalization performance of the network. We use the mini-batch in our method [11]. We bundle the sample of the data generated in the previous procedure into one dataset and acquired a mini-batch randomly acquired from the data set. To reduce the influence of the differences in data distribution for each batch, data normalization was performed with batch normalization [12] before the activation function.

Evaluation

Details of Data

We applied the above deep neural network to predict future taxi demand in the Tokyo area. In the experiment, three types of datasets, taxi data, population data, and weather data, were used as input data. As taxi data, we used operation data collected from GPS devices installed in individual taxis. In this case, latitude, longitude, and passenger boarding state (0, 1) were collected at intervals of 5–10 s. As the population data, we used real-time population statistics data [13] estimated from the cellular network of NTT DOCOMO, a Japanese telecom operator. Population was estimated based on the position relationships between mobile devices and cellular base stations. The system can estimate population with a spatial resolution of 500 m and temporal resolution of 10 min. In addition, the system can estimate the total population numbers in each grid cell by adding the share rates of NTT

	Data Type	RMSE
A	Taxi	1.513
B	Taxi and population	1.370
C	Taxi and weather	1.351
D	Taxi, population and weather	1.378

Table 2: Model evaluation results for various types of data

DOCOMO and other operators. As the weather data, we used high-resolution precipitation radar data from Japan Meteorological Agency, which predicts precipitation in 1000-m grids at intervals of 10 min. The details of each type of data are given in Table 1. In the evaluation, we calculated the number of passengers who want to ride a taxi in next 30 min in each 500-m grid at intervals of 10 min. A dataset containing data for the period April 1 2015 to August 31 2016 was used for learning, and a dataset containing data for the period September 1 2016 to September 14 2016 was used for evaluation.

Parameter Search of Stacked denoising Autoencoder

We used hyperopt module in Python to perform hyper-parameter search by means of random search [14][15]. We optimized hyper-parameters; the number of nodes in each layer, noise coefficient of denoising autoencoder, regularization coefficient of sparse autoencoder, ratio of dropout, and batch size.

About Evaluation Index

To evaluate the usefulness of the proposed method, we evaluated the root mean square error (RMSE) of the method by using equation 4. Here, t_i is the actual number of people boarding taxis, and $\hat{t}_i$ is the predicted number of taxi rides.

$$\mathrm{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (t_i - \hat{t}_i)^2 \right)^{\frac{1}{2}} \tag{4}$$

Experiment Results

We performed offline evaluation to evaluate if weather data and population data are effective for improving prediction accuracy by changing the data type using the four-layer SdA. Table 2 and Figure 2 summarize and show, respectively, the results with only taxi data, taxi and population data, taxi and meteorological data, and all three types of data. According to the results, the accuracy is higher when we use population and weather data than that when we use taxi data alone. The weather data greatly contributes to the improvement of accuracy over the entire area. In the case of low-demand grid cells, the accuracy is inferior when all three types of data are used compared to that when taxi and weather data are used. By contrast, in the case of high-demand grid cells, the accuracy is superior when all three types of data are used. We consider that the population data contributes to enhanced performance, especially in locations where demand is high. A comparison of the predicted and the actual taxi boarding counts in a certain grid cell is shown in Figure 6.

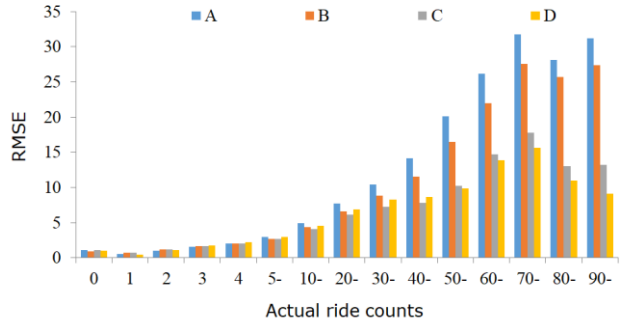

Figure 2: Predictive accuracy (RMSE) for each actual ride value of SdA with different input data. Left vertical axis: RMSE, right vertical axis: total occurrences of actual target values (logarithmic scale), horizontal axis: actual boarding value, A: using only taxi data, B: using taxi and population data, C: using taxi and rainfall data, D: using all three types of data

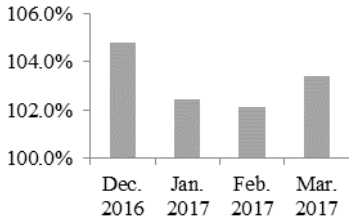

Figure 3: Sales comparison. The graph indicates the sales of the participants compared to those of non-participants. The figures suggest that the system improved the sales of participants in all four consecutive months of the field test.

Demonstration Experiment

We implemented the proposed taxi demand forecasting system in taxis and conducted field test in a real-world environment. The target area was 23 wards in Tokyo, Musashino city, and Mitaka city area. The implemented system forecasts and displays future taxi boarding numbers in each 500-m grid from present time to 30 min in the future. The predicted number is updated once every 10 min. An image of the output of the proposed system is shown in Figure 4, and a screen grab of the actual system on a tablet device is shown in Figure 5.

Figure 4: Taxi demand prediction visualization system. The system calculates future passenger boarding demands and visualizes them in real time on a tablet device screen. The tablet device is installed in the dashboard of taxis.

We conducted the field test from December 1 2016 to March 31 2017. In the experiment, 26 taxi drivers participated and operated based on the demand forecast result. We compared sales numbers between the 26 taxi drivers who used our system and 10,640 taxi drivers who did not. Figure 3 shows the results of sales comparison. Compared to the non-participants, the sales of the participants improved in all four consecutive months by 3.9% on average, which equals to 1,409 JPY per person per day. As a result, we confirmed the efficiency of our proposed method in the field test in addition to offline evaluation.

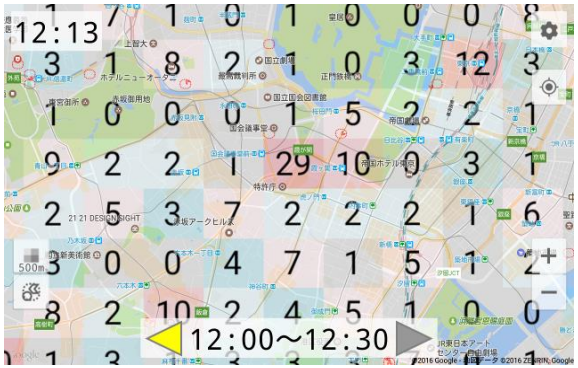

Figure 5: Output image of the taxi demand forecast system. The figures in each grid cell indicate expected the number of passengers boarding in a target 500-m mesh area in the next 30 min. Areas with high demand are colored in red and those with low demand are colored in blue or are transparent.

Conclusion

In this paper, we proposed a taxi demand forecasting method by using stacked denoising autoencoder and cellular network based real-time population data. In an offline experiment, we showed that the proposed algorithm can predict taxi demand with an error of 1.370 RMSE when using real-time population data compared to an error 1.513 when not using population data. In addition, we conducted a field test by installing the proposed system in 26 participating. The field test result showed that the sales of the participant drivers

increased by 1,409 JPY per person per day, which represents an average increase in sales of 3.9% compared to the drivers who did not use the proposed system. In the future, we aim to optimize total profit by considering total distance and time of taxi boarding, as well as efficient passenger allocation control to individual drivers.

References

1. Taxi in Tokyo 2017–Tokyo Hire-Taxi Association. http://www.taxi-tokyo.or.jp/datalibrary/pdf/hakusyo2017all.pdf, (Referred 2018-07-03)
2. Li, Bin and Zhang, Daqing and Sun, Lin and Chen, Chao and Li, Shijian and Qi, Guande and Yang, Qiang.. 2011. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. *PERCOM Workshops*.
3. Powell, Jason W and Huang, Yan and Bastani, Favyen and Ji, Minhe. 2011. Towards reducing taxicab cruising time using spatio-temporal profitability maps. *SSTD*. 242-260.
4. Tong, Yongxin and Chen, Yuqiang and Zhou, Zimu and Chen, Lei and Wang, Jie and Yang, Qiang and Ye, Jieping and Lv, Weifeng. 2017. The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms. *KDD*, 1653-1662.
5. Smith, Austin W and Kun, Andrew L and Krumm, John. 2017. Predicting taxi pickups in cities: which data sources should we use? *Ubicomp*.
6. Yuan, Jing and Zheng, Yu and Zhang, Liuhang and Xie, XIng and Sun, Guangzhong. 2011. Where to find my next passenger. *Ubicomp*.
7. Wang, Dong and Cao, Wei and Li, Jian and Ye, Jieping. 2017. DeepSD: supply-demand prediction for online car-hailing services using deep neural networks. *ICDE*.
8. Lv, Yisheng and Duan, Yanjie and Kang, Wenwen and Li, Zhengxi and Wang, Fei-Yue. 2015. Traffic flow prediction with big data: a deep learning approach. *Transactions on Intelligent Transportation Systems, IEEE* 16, 2: 865-873.
9. Hinton, Geoffrey E and Salakhutdinov, Ruslan R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313: 504-507.
10. Vincent, Pascal and Larochelle, Hugo and Lajoie, Isabelle and Bengio, Yoshua and Manzagol, Pierre-Antoine. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.Journal of Machine Learning Research 11: 3371-3408.
11. Li, Mu and Zhang, Tong and Chen, Yuqiang and Smola, Alexander J. 2014. Efficient mini-batch training for stochastic optimization. *KDD*. 661-670.
12. Ioffe, Sergey, and Szegedy, Christian. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*. 448-456.
13. Masayuki, Terada, Tomohiro Nagata and Motonari Kobayashi. 2013. Population estimation technology for mobile spatial statistics. *NTT DOCOMO Technical Journal* 14, 3.
14. Bergstra, James, and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 2012, 13.Feb: 281-305.
15. Bergstra, James; Yamins, Dan and Cox, David D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. Python in Science Conference. 2013: 13-20.

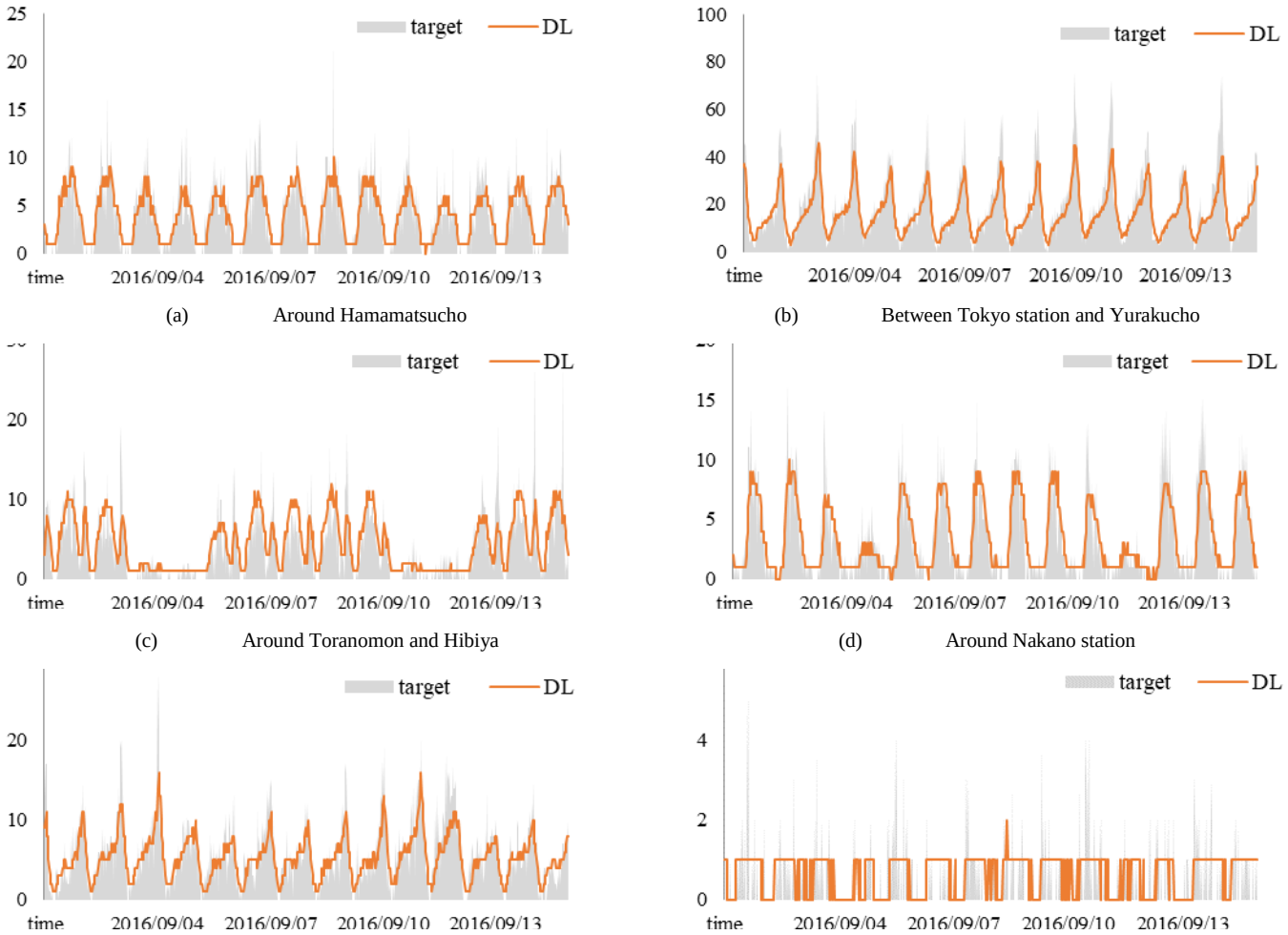

Figure 6: Comparison of predicted and actual taxi boarding counts in a given grid cell (horizontal axis: time, vertical axis: number of boarding) As for panels (a), (b), and (e), the predicted and actual values change by almost the same amount over time. In panels (c) and (d), the prediction is correct even when the demand varies depending on the day of the week. In panel (f), the demand forecast is correct even in areas with low demand.