A Data-Driven Approach to Understanding and Predicting the Spatiotemporal Availability of Street Parking

Mingxiao Li
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
limx@lreis.ac.cn

Song Gao
Geospatial Data Science Lab, University of Wisconsin, Madison
song.gao@wisc,edu

Yunlei Liang
Geospatial Data Science Lab, University of Wisconsin, Madison
yunlei.liang@wisc.edu

Joseph Marks
Geospatial Data Science Lab, University of Wisconsin, Madison
jbmarks@wisc.edu

Yuhao Kang
Geospatial Data Science Lab, University of Wisconsin, Madison
yuhao.kang@wisc.edu

Moyin Li
Pace Suburban Bus, Chicago
moyin.li@pacebus.com

ABSTRACT

Searching for a parking spot in metropolitan areas is a great challenge comparable to the Hunger Games, especially in highly populated areas such as downtown districts and job centers. On-street parking is often a cost-effective choice compared to parking facilities such as garages and parking lots. However, limited space and complex parking regulation rules make the search process of on-street parking very difficult. To this end, we propose a data-driven framework for understanding and predicting the spatiotemporal availability of on-street parking using the NYC parking tickets open data, points of interest (POI) data and human mobility data. Four popular types of spatial analysis units (i.e., point, street, census tract, and grid) are used to examine the effects of spatial scale in machine learning predictive models. The results show that random forest works the best with the highest accuracy scores for the spatiotemporal availability classification across all four spatial analysis scales.

CCS CONCEPTS

• **Information systems → Geographic information systems**.

KEYWORDS

data fusion, machine learning, urban computing, spatiotemporal analytics

ACM Reference Format:
Mingxiao Li, Song Gao, Yunlei Liang, Joseph Marks, Yuhao Kang, and Moyin Li. 2019. A Data-Driven Approach to Understanding and Predicting the Spatiotemporal Availability of Street Parking. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359366

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359366

1 INTRODUCTION

Parking is an important element in the transportation system and plays an important role in people's travel decisions. Parking availability information and pricing can influence people's departure and arrival time, travel mode choices, and activity duration. Most on-street parking is free or underpriced compared to garages and parking lots and therefore it is often over-demanded. This makes searching for a parking spot in metropolitan areas a great challenge comparable to the Hunger Games, especially in highly populated areas such as downtown districts, job centers, etc. First, the supply and demand of parking spaces is unbalanced with the increasing number of vehicles but limited parking facilities in urban areas. Second, the urbanization process is accelerating in most metropolitan areas and attracting more job opportunities, human flows, business and social activities. These popular destinations together with underpriced parking generates more travel demand and parking needs. Third, parking availability is highly variable spatially and temporally.

On-street parking is often a cost-effective choice compared to parking facilities such as garages and parking lots. However, limited space and complex parking regulation rules make the search process very difficult. Moreover, there is almost no real time information about available on-street parking spots. Even if the driver is a local resident, compound parking rules can still surprise the driver and generate tickets due to various reasons such as street cleaning schedules, proximity to a fire hydrant, and no standing or no stopping rules during certain time periods. New York City (NYC) is among the most ticketed and the highest ticket cost cities in the United States[1]. There are over 10.8 million parking violation tickets generated in NYC in the fiscal year 2017 and 11.7 million in fiscal year 2018. The motivation of this research is to first understand the spatiotemporal patterns of on-street parking violation tickets and then build reliable machine learning models to predict the spatiotemporal availability of on-street parking using the NYC open data[2].

To this end, we propose a data-driven framework for understanding and predicting the spatiotemporal availability of on-street parking by training machine leaning models using the

[1] https://www.spotangels.com/blog/nyc-parking-tickets-the-most-ticketed-neighborhoods-in-nyc/
[2] https://data.cityofnewyork.us/

NYC parking tickets open data. In addition, four popular types of spatial analysis units (i.e., point, street, census tract, and grid) are used to examine the impact of spatial scale in classic machine learning predictive models.

2 DATA

2.1 Parking Violation Tickets Open Data

As mentioned above, we downloaded over 10.8 million parking violation tickets generated in NYC in the fiscal year 2017 and 11.7 million in fiscal year 2018 from the NYC Open Data platform. Each ticket contains information including a summons number, violation code, street address, ticketing time, vehicle plate, etc.

2.2 Points of Interest

In order to understand what kind of surrounding environments are associated with more parking violation tickets, such as the presence of an employment center, retail stores, health care services, shopping centers, and so on, we collected data for over 137,000 points of interest (POIs) in NYC from the Safegraph business venue database[3]. The POIs are first classified based on the North American Industry Classification System (NAICS) 2-digit sector codes. To begin, the POIs are classified into 23 categories based on the NAICS 2-digit sector codes, including *Agriculture Forestry Fishing 11, Mining Oil and Gas Extraction 21, Utilities 22, Construction 23, Manufacturing (31,32,33), Wholesale Trade 42, Retail Trade 44, Retail Trade 45, Transportation Warehousing (48,49), Information 51, Finance Insurance 52, Real Estate & Rental Leasing 53, Professional Scientific Tech 54, Administrative Support and Waste 56, Educational Services 61, Health Care and Social Assistance 62, Arts & Entertainment & Recreation 71, Accommodation & Food Services 72, Other Services 81, Public Administration 92.* In addition, the category of *Parking Lots and Garages* (NAICS Code: 812930) is treated as a separate POI category since it is directly related to parking activity. This gives a total of 24 types of POIs at the root level of categorization as part of model features.

2.3 Human Mobility Patterns

In addition to the static spatial distribution of POIs information, we also retrieved the fine-resolution visit patterns of all POIs from the aforementioned SafeGraph database which covers dynamic human mobility patterns of millions of anonymous smart phone users. For each POI, the records of aggregated visitor patterns illustrate the number of unique visitors and the number of total visits to each venue during the specified time window, which could reflect the attractiveness of each venue. The mean hourly visits over a week were recorded as a 168-dimensional vector to show the dynamic stream of visit patterns. If a visitor stays for multiple hours, a visit will be shown in each hour during which the visitor stayed.

2.4 Data preprocessing

In this study, as shown in Figure 1, we discretized the study area into four spatial scales (point level, street level, census tract level, and 1km grid level) and the time into 168 hourly slots (7 days of a week * 24 hours of a day) to capture a snapshot of the availability of street

[3]https://www.safegraph.com

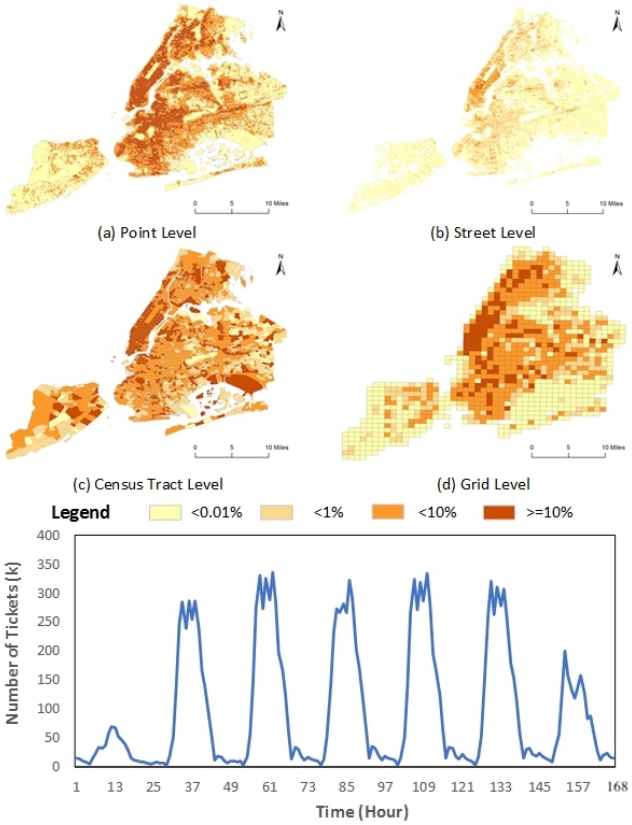

Figure 1: The spatial distribution of parking violation tickets at four spatial scales and the temporal variation curves of tickets.

parking. To perform an analysis at the selected spatiotemporal scale, we first found the coordinates of each tickets by using an online geocoding service. Then, since we focus on the street parking availability, the ticket points that were more than 50 meters from the road were deleted. Finally, we distributed the number of tickets and corresponding attributes as follows for each spatial unit and each time slot.

At the point level, the following features were chosen: the location of $p_j(x_j, y_j)$, the time of day t_h, and the day of week t_d. The corresponding number of tickets $Num_{(j,h,d)}$ was used as the label data for model training. At the street level, besides the spatiotemporal characteristics p_j, t_h, and t_d, the street width st_{wid}, street length st_{len}, street type st_{type}, and whether it is a two-way street st_{dir} were selected to characterize each street. In sum, the training features can be represented as follows: $[p_j,t_h,t_d,st_{wid},st_{len},st_{type},st_{dir}]$. As for the census tract level and 1km grid level, the features were more complex. At these levels, we not only considered their spatiotemporal characteristics p_j, t_h, and t_d, their street attributes with the summation of street length *sum_st_len* and street area *sum_st_area*, but also considered the dynamic human mobility patterns and the POI distribution in the corresponding spatial unit. The POI data was aggregated to each spatial unit and represented as 24 features. In addition, the

number of visits observed in the specified unit sum_visit_j, the number of unique visitors $sum\_visitor_j$, the number of visits of corresponding to the time of day $visit_{(j,h)}$ and the number of visits of corresponding to the day of week $visit_{(j,d)}$ were added to model the human mobility patterns. Thus, the training features can be represented as follows: $[p_j, t_h, t_d, sum\_st\_len, sum\_st\_area, POI_1, \cdots, POI_{24}, sum\_visit_j, sum\_visitor_j, visit_{(j,h)}, visit_{(j,d)}]$. Note that the records with at least one ticket are marked as '1' for binary classification. Table 1 shows the number of samples at each spatial scale and the ratio of positive and negative cases for parking availability classification.

Table 1: The number of samples at different spatial scales and the ratio of positive and negative cases for parking availability classification.

Spatial Unit	# of samples	% positives	% negatives
Point	35,629,944	12.3	87.7
Street	6,716,976	29.7	70.3
Census Tract	356,496	70.5	29.5
1km Grid	139,440	69.2	30.8

3 METHODS

To investigate the impact of the spatial resolution on parking availability prediction using machine learning, we selected the following widely used machine learning models for this case study. It can be interpreted as a binary classification problem of whether the corresponding time and place can be parked by analyzing the historical parking violation ticket information. Therefore, a specified location and time with at least one ticket is marked as '1' to represent 'Risky Parking' and the others will be marked as '0' to represent 'Permitted Parking'.

KNN: The k-nearest neighbors algorithm (KNN) classifies an object by a vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors in feature space [4].

Logistic regression: It uses a logistic function to model the relationship between one dependent binary variable and one or more nominal, ordinal, interval and ratio independent variables [1].

Naive Bayes: It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting [6].

SVM: Support vector machines (SVM) construct a set of hyperplanes in a high-dimensional space. New samples are mapped into the same space and predicted based on the gaps which they fall into [3].

SGD: Stochastic gradient descent (SGD) classifier implements linear support vector machines with SGD learning: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule [8].

Random Forest: The random forest (RF) constructs a multitude of decision trees and outputs the results by computing the mean of the predictions of each individual tree [2]. RF is trained on different parts of the same training set, with the goal of reducing the variance.

DNN: Deep neural network (DNN) is a multi-hidden-layer artificial neural network whose artificial neurons can respond to a surrounding unit within a portion of the coverage. We constructed a DNN architecture consisting of four fully connected layers and two dropout layers with a 0.5 rate to regularize the DNN and improve the generalization error. The output layer uses a sigmoid activation function to produce a probability between 0 and 1 for binary classification using a threshold of 0.5.

Evaluation: There are four outcomes from the binary classification result [7]: true positives (TP), false positives (FP), True negatives (TN), and False negatives (FN). The following metrics including the overall accuracy, precision and recall, F1-score are used to evaluate the parking availability classification models.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

4 RESULTS

Regarding the classification results, as shown in Table 2, random forest outperforms all other models and achieved both high accuracy scores (0.82, 0.85, 0.86, and 0.88) and high F1-scores (0.82, 0.72, 0.90, and 0.88) across all four spatial scales. The KNN and the DNN also perform well and fall behind the random forest by a small margin. Note that we chose k=3 as the number of nearest neighbors in feature space with regard to the temporal autocorrelation patterns of parking availability over time. The autocorrelation coefficient for parking availability with temporal lag of 3 hours is 0.28, 0.40, 0.38, and 0.55 at the point, street, census tract, and grid levels respectively. Although the Naive Bayes model get a good accuracy and F1 scores at the point level (0.73 and 0.75) and at the street level (0.66 and 0.72), it didn't perform well at the aggregation levels with lower F1-scores at the census tract level (0.33) and at the 1km grid level (0.54).

Alternatively, one may want to check the detailed precision-recall curves with different recall rates as shown in Figure 2 to compare the model performance especially for skewed datasets [5]. It shows that the random forest outperforms all other models with the highest precision value across different recall rates in parking availability prediction at four spatial scales.

5 CONCLUSION

In this study, we propose a data-driven framework for understanding and predicting the spatiotemporal availability of on-street parking by training a set of machine leaning models using the NYC parking tickets open data. The models are tested at four types of spatial analysis units (i.e., point, street, census tract, and grid) and the results confirmed the impact of spatial scale in machine learning predictive models. The experiment results show that random forest works the best with the highest F1 scores for the spatiotemporal availability classification across all four spatial scales. Given a search location and time for on-street parking, the F1-score is 0.82 for parking availability prediction, which shows a good potential for street parking applications. Moreover, using the surrounding

Table 2: Prediction accuracy and F1-score of parking availability using all features with different machine learning models.

Model	Accuracy and F1 (Point)	Accuracy and F1 (Street)	Accuracy and F1 (Census Tract)	Accuracy and F1 (Grid)
KNN	0.79 and 0.79	0.73 and 0.73	0.78 and 0.85	0.83 and 0.83
Logistic Regression	0.55 and 0.56	0.58 and 0.57	0.66 and 0.73	0.77 and 0.74
Naive Bayes	0.73 and 0.75	0.66 and 0.72	0.42 and 0.33	0.67 and 0.54
SVM	0.53 and 0.53	0.55 and 0.51	0.71 and 0.83	0.56 and 0.69
SGD	0.57 and 0.60	0.59 and 0.60	0.68 and 0.84	0.76 and 0.75
Random Forest	0.82 and 0.82	0.85 and 0.72	0.86 and 0.90	0.88 and 0.88
DNN	0.74 and 0.76	0.75 and 0.74	0.71 and 0.83	0.76 and 0.74

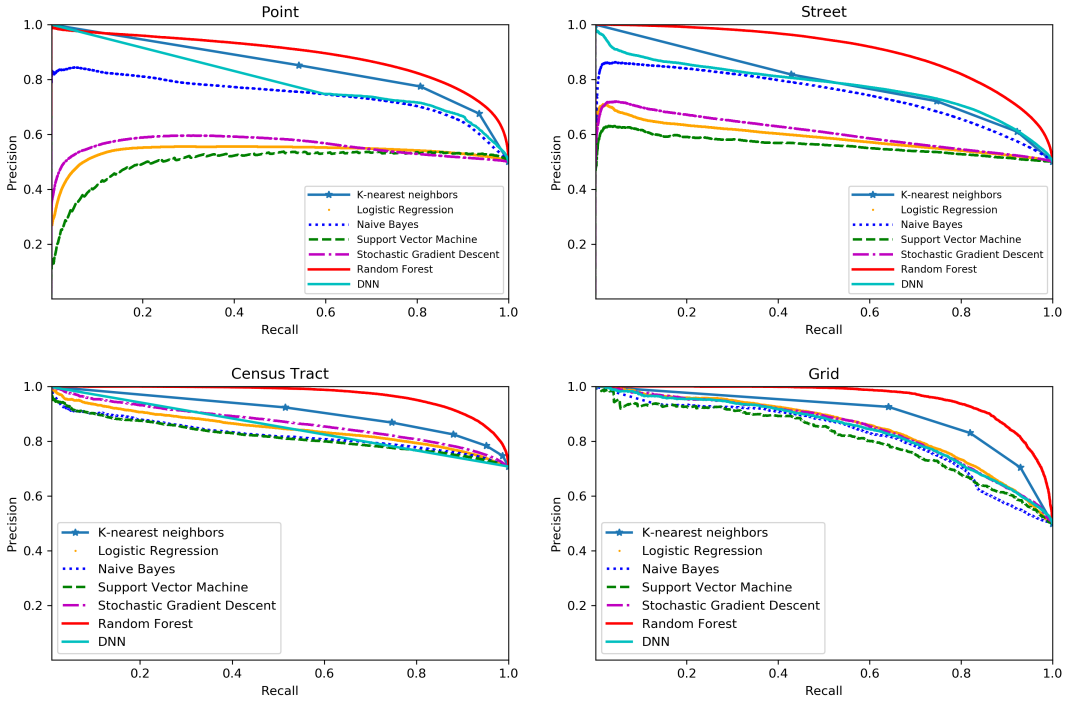

Figure 2: The precision and recall curves of the parking availability prediction results at different spatial scales.

POI context and dynamic human mobility patterns can help improve the accuracy of parking availability prediction. With better on-street parking information provided in advance, drivers can enhance their parking decision-making. Our research may offer insights into parking management policy such as parking regulation rules, pricing, and time limitation to balance the parking demand and supply at different spatial scales using open data and machine learning approaches.

ACKNOWLEDGMENTS

The authors would like to thank Safegraph for providing the anonymous location data and POI visit patterns. Support for this research was provided by the University of Wisconsin - Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

REFERENCES

[1] Joseph Berkson. 1944. Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* 39, 227 (1944), 357–365.
[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
[4] Thomas M Cover, Peter E Hart, et al. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13, 1 (1967), 21–27.
[5] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 233–240.
[6] Melvin Earl Maron. 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)* 8, 3 (1961), 404–417.
[7] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning*. MIT press.
[8] Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, 116.