Forecasting short-term taxi demand using boosting-GCRF

Xinwu Qian  
Purdue University  
550 Stadium Mall Dr  
West Lafayette, Indiana 47906  
qian39@purdue.edu

Chao Yang  
Tongji University  
4800 Cao An Rd  
Shanghai,China 201804  
tongjiyc@tongji.edu

Satish V. Ukkusuri  
Purdue University  
550 Stadium Mall Dr  
West Lafayette, Indiana 47906  
and Tongji University  
4800 Cao An Rd  
Shanghai,China 201804  
sukkusur@purdue.edu

Fenfan Yan  
Tongji University  
4800 Cao An Rd  
Shanghai,China 201804  
1334577@tongji.edu

ABSTRACT
It will be most efficient to frame operation strategies before actual taxi demand is revealed. But this is challenging due to limited knowledge of the taxi demand distribution in immediate future and is more prone to prediction errors. In this study, we develop the boosting Gaussian conditional random field (boosting-GCRF) model to accurately forecast the short-term taxi demand distribution using historical time-series demand over the study area. Comprehensive numerical experiments are conducted to evaluate the performance of boosting-GCRF as compared to five other benchmark algorithms. The results suggest that the boosting-GCRF is superior with the best modified mean absolute percentage error being 10.4%. The approach is observed to be robust based on its prediction performance on anomaly taxi demand data. In addition, the density functions generated by the boosting-GCRF model are found to well capture the actual distribution of the short-term taxi demand.

KEYWORDS
Short-term taxi demand forecasting, spatiotemporal correlations, Gaussian conditional random field, boosting

1 INTRODUCTION
Taxi is a popular mode of urban transportation due to its point-to-point service and 7-24 availability. By the end of 2014, there was approximately one yellow cab per hundred population in Manhattan, and a total of 13,587 yellow cabs served over 450,000 daily trips [1]. But such a huge industry has been notorious for its inefficient operations with long passenger waiting time and excessive vacant trips. In particular, Zhan et al. [2] suggested that we may reduce up to 60% of the market inefficiency if demand and supply are perfectly matched, and many efforts have been made to address the issue by designing recommendation systems [3–5], combining similar passengers [6–9], and framing dynamic pricing policies [10, 11]. An essential input for these efforts is the passenger demand in immediate future, which is either assumed to be known from historical observations or through trivial predictions. But we may barely improve the operation performance, if not reducing, without precise understanding of the demand level. On the other hand, it is likely to have a wide range of possible demand values over a short period of time, which emphasizes the need of modeling demand uncertainty in addition to point estimations. This motivates us to study the problem of predicting the distributions of short-term taxi demand, which will contribute to robust operation and policy making of taxi industry.

In this study, the short-term taxi demand forecasting problem focuses on inferring future demand distributions from historical observations. It is well-understood that the long-term taxi ridership is closely associated with socio-economics, demographics, as well as built environment factors [12]. But for short-term taxi demand, as urban activities and mobility may differ over time and space, it is not viable to forecast future demand based on aggregated explanatory variables. Moreover, to help real-world operations such as pre-dispatching vacant taxis and framing dynamic pricing strategies, the prediction needs to be conducted at fine spatial and temporal scale. This introduces several challenges in modeling short-term taxi demand. First, taxi demand in a small urban area is likely to be noisy and non-stationary over a short time horizon. And the prediction errors may be largely determined by whether or not the spatiotemporal correlations can be properly modeled [13]. Second, a fine level of spatial and temporal aggregations may significantly increase the dimensionality of the problem, but useful data are limited due to the rapid change of mobility patterns. While simple model may be unlikely to capture the high-dimensional nature of the problem, it is also difficult to make accurate prediction with a complex model using limited training data. Third, the taxi demand pattern may be entirely different when special events or sudden weather changes take place, and the proposed model needs to be robust against possible anomalies in addition to normal daily operations. The objective of the study is to develop a model to accurately forecast short-term taxi demand by addressing above-listed challenges.

Despite the significance of the problem, little attention has been paid to understand the nature of short-term taxi demand. To the
best of our knowledge, only Moreira et al. [14] studied the problem and proposed an ensemble model to forecast taxi demand at taxi stands for every 30 minutes, and the best MAPE reported was 23.12%. However, the demand at taxi stands may only account for a small portion of total ridership, and the best MAPE achieved may not be accurate to meet the needs of real-time market operations.

In light of the results from previous works and considering the difficulties in modeling short-term taxi demand, this study investigates predicting short-term taxi demand using the Gaussian Conditional Random Field (GCRF) model. We use historical time series of taxi demand as input, and model the distributions of taxi ridership one or multiple steps ahead for each individual location in the study area. The problem can be categorized as the multi-target probabilistic regression, and we develop a boosting approach which specifically tailors the needs of the problem. The advantages of the proposed model can be summarized from four aspects. First, the model captures both historical and future spatiotemporal correlations of short-term taxi demand. Second, the model is capable of generating forecast multiple steps ahead as a structured output, thus avoiding constructing multiple predictors for each individual output and being more accurate. Third, the model conducts probabilistic estimations rather than point estimations, which helps to quantify uncertainties of future demand and is more suitable for modeling noisy short-term demand. Finally, the model uses regularization to overcome data availability issue, and improves point and probabilistic estimation performances through the proposed boosting approach. We compare with the base GCRF, bagging-GCRF, ANN, AdaBoost regression tree, and gradient boosting tree methods using the real-world taxi data to evaluate the performance of the boosting-GCRF model. We find that the boosting-GCRF achieves the best mMAPE of 10.4% and improves the base GCRF model by 5.63% on average. The model is also found to be robust under anomalies, and the resulting probabilistic distribution may well capture the actual distribution of short-term taxi ridership.

The rest of the paper is organized as follows. Section I introduces the data and the study area, discusses about data processing, and analyzes the spatiotemporal characteristics of short-term taxi demand. Section II discusses the GCRF model in detail, presents the boosting approach, and also introduces the bagging method and benchmark algorithms. Section IV presents comprehensive numerical experiments in assessing the model performances. Finally, section V concludes the study followed by future works.

2 DATA

2.1 Data processing

In this study, we use 2015 New York City (NYC) taxi trip data to explore forecasting short-term taxi demand. The data were collected by the New York City Taxi and Limousine Commission (NYCTLC) for the year of 2015. It was reported that there were 13,000 medallion cabs by the end of 2015 [1], which produced over 20 Gigabytes of trip data. Each trip record contains geo-coordinates and timestamps for trip origin and destination. It also includes trip distance, trip duration, number of passengers, as well as payment information. Complete trip trajectories are not available due to privacy concerns.

We processed the data by first removing erroneous information, e.g. if a trip lies outside of the study area, if the trip duration is either too short (< 30s) or too long (> 2h), and if the mean travel speed is excessive (> 70mph). Then we grouped the trip records based on trip origin locations and timestamps. In this study, we choose Manhattan borough of NYC as our study area since the borough alone covers more than 90% of total taxi trips in NYC [15]. The trip records are aggregated spatially into Zip Code Tabular Areas (ZCTAs) with the time horizon of 15-minutes time interval. There are initially 106 ZCTAs in Manhattan, and some of the ZCTAs only represent a single building where no taxi trips originates inside. We merge these tracts with peripheral areas, which results in the final study area of 48 ZCTAs as shown in Figure 1.

![Study Area and NYC ZCTA Map](image_url)

**Figure 1:** Study area and NYC ZCTA Map. (The dots in study area denote selected tracts in lower, middle, and upper Manhattan for spatiotemporal characteristic analyses)

(a) 15-minutes taxi demand distribution (b) Corresponding auto correlation coefficients (Lower, middle, and upper Manhattan) (c) Semivariogram plot for spatial autocorrelation (lag at the scale of 10^2 meters)

**Figure 2:** Sample demand distribution, temporal autocorrelation plot, and spatial autocorrelation plot
2.2 Spatiotemporal correlations of short-term taxi demand

To model future demand accurately, the first priority is to analyze and understand the underlying characteristics of the data. Figure 2 presents the sample demand distributions of three selected tracts, as well as the evaluations of the spatial and temporal correlations. In this section, we discuss the data characteristics of three subareas selected from lower, middle, and upper Manhattan (see Figure 1).

It can be observed from the demand distributions at 15-minutes scale that the short-term taxi demand data are noisy and highly non-stationary. While there are signs of seasonality for lower-Manhattan and mid-Manhattan locations as revealed by the periodical hills and valleys, no obvious trend can be observed for the upper Manhattan location. Moreover, the periodical patterns are only observable at longer time-scale (daily), it is difficult to justify short-term correlations from the figure. To improve our understanding, we then plot the autocorrelation coefficients for temporal correlations and the Semivariogram for spatial correlations at these locations. As shown in Figure 2(b), all three plots suggest that the time-series demand is not random and the autocorrelation coefficients are significant even with large lags. There is clearly a strong positive association for consecutive short-term demand at all three locations, where the increase in short-term demand will likely lead to higher demand in immediate future and vice versa. But the strength of the association is heterogeneous over the space.

The Semivariogram is plotted for the whole study area at different time of the day, as shown in Figure 2(c), which is often used in spatial statistics to describe the spatial autocorrelation between a pair of sample points [16]. In particular, a smaller Gamma value indicates that the spatial correlation between two locations is stronger. In the plots, the red dots are for observed empirical gamma values with bin size equal to 20, and the curve is the fitted theoretical one. For the fitted theoretical curve, the range of x-axis under the blue curve indicates the effective range, which suggests that tracts within the lag range are spatially correlated. We can verify from the results that a strong spatial correlation does exist across different time of the day if the centroids of two tracts are within the distance of 0.05 lag, or 5 kilometers. In other words, the short-term taxi demand is found to be strongly correlated for neighborhood areas, and the closer the two tracts are the higher the correlation is. And this correlation varies based on the time of day. The range can be up to 0.05 lag for 8AM and 12PM case, while the 12AM is observed to have the smallest range of around 0.03. As a consequence, the spatial correlation is also heterogeneous with respect to the time.

In conclusion, we observed that the short-term taxi demand is non-stationary, but there exists non-trivial spatial and temporal dependencies for short-term taxi demand at various times and locations. And it is important to model the spatiotemporal correlations properly in order to obtain accurate prediction results.

3 METHODOLOGY

3.1 Gaussian Conditional Random Field

3.1.1 The basic model. We are interested in the problem of modeling the distribution of short-term taxi demand at a given location and time interval. This ensures greater flexibility for stakeholders to frame the operation strategy and helps to evaluate the level of confidence of a certain estimation. Moreover, we have shown that the short-term taxi demand presents significant spatial and temporal correlations, which need to be modeled accordingly in order to obtain accurate estimation. As a consequence, considering a study area which consists of n subareas, our focus is to model the probability:

\[ p(Y_i|X, Y_{-i}) \]

where \( Y_i = [y_{i,t+1}, y_{i,t+2}, \ldots, y_{i,t+k}]^T \) is a column vector of length k, representing the estimated taxi demand at location \( i \) up to k time steps ahead. And \( X = [X_1^T, X_2^T, \ldots, X_n^T]^T \) is a column vector of length \( tn \), which refers to the collection of observed taxi demand at \( n \) locations in the study area across t time steps into the history, where each component of \( X \) takes the form \( X_i = [x_{1i}^T, x_{2i}^T, \ldots, x_{ni}^T]^T \). Similarly, \( Y = [Y_1, Y_2, \ldots, Y_n]^T \) is the column vector of length \( kn \) for the estimated taxi demand of \( n \) locations into k future time steps, and \( Y_{-i} \) stands for all estimated demand except at location \( i \).

The conditional probability suggests two layers of dependency: (1) future demand at location \( i \) may depend on historical demand at all other places in the past \( t \) time steps, and (2) future demand of all locations may also be inter-correlated. In this study, we assume the short-term taxi demand at each location during each time interval follows the normal distribution, which we observe to be reasonable approximation for majority of the places in the data. Note this assumption is not restrictive, as one may convert any empirical marginal distribution into Gaussian distribution by conducting copula transformation and vice versa [17]. We model the structured prediction a multivariate Gaussian distribution following:

\[ p(Y|X; \Theta, \Lambda) = \frac{1}{Z(x)} \exp\left( -\frac{1}{2} Y^T \Lambda Y - X^T \Theta Y \right) \]

where \( \Theta \) is a \( tn \times kn \) matrix which maps \( X \) to \( Y \), \( \Lambda \) is a \( kn \times kn \) matrix which captures the inter-correlation among the output, and \( Z(x) \) is the normalization constant which ensures the distribution being integrated to 1 and can be calculated as:

\[ Z(x) = c|\Lambda|^{-\frac{1}{2}} \exp\left( X^T \Theta \Lambda^{-1} X \right) \]

This conditional probilistic distribution of \( Y|X \) is known as the GCRF model, which was original developed from the CRF model [18] for structured regressions, and has been applied to various problems including energy load forecasting [19], image denoising [20], and modeling patients’ behavior [21]. In particular, the distribution has the mean of \( -\Lambda^{-1} \Theta^T X \) and the covariance being \( \Lambda^{-1} \), and the two parameter matrices \( \Lambda \) and \( \Theta \) for the short-term demand forecasting problem can be understood from the graph representation as shown in Figure 3. To model the conditional distribution \( p(Y|X) \), we need to infer the most likely \( \Theta \) and \( \Lambda \) from training data, which is equivalent to solving the following maximum likelihood estimation (MLE) problem:

\[ \max_{\Theta, \Lambda} \ - \log|\Lambda| - \frac{1}{2} \text{tr}(Y^T Y \Lambda + 2Y^T X \Theta + \Lambda^{-1} \Theta^T X^T X \Theta) \]

There are two main benefits of GCRF: 1) the model is the discriminative form of Gaussian Markov Random Field model, thus being computationally more efficient and resulting in lower error compared with generative models [22], 2) while the covariance matrix
may be dense due to complex spatial and temporal correlations, the model infers the inverse of the covariance matrix which is likely to be sparse since it models the conditional independency.

3.1.2 Regularization. The basic GCRF is not readily applicable for estimation due to overfitting issue. Considering the scenario where the study area has 30 tracts, and we would like to estimate the demand 8 steps into the future using 16 steps of historical observations, where each time step corresponds to a 15-minutes time interval. The resulting input X has the number of features being 30 (tracts)×16 (time steps) = 480, and the length of Y is 30×8 = 240. In the worst case, we will have to estimate 480×240 + 240×(241)/2 number of parameters to infer Θ and Λ, which suggests that the GCRF model has high complexity. To avoid overfitting, the size of training samples is expected to be significantly larger than the number of features. As one day of trip data may only contribute to one training sample, it may require years of trip data to sufficiently train the GCRF model. To obtain the data is not an issue, nevertheless, due to rapidly changing urban landuse over time, historical trip data long time back are likely to serve as outliers for estimating demand at present. Adding regularizers to estimate model parameters is the viable solution in light of the issues. We consider estimating model parameters with l1 regularizer, which gives rise to the following maximum likelihood estimation equation:

\[
\text{maximize}_{\Theta, \Lambda} \quad -\log|\Lambda| - \frac{1}{T} \text{tr}(Y^T Y \Lambda + 2Y^T X \Theta + \Lambda^{-1} X^T X \Theta) + \lambda(||\Theta|| + ||\Lambda||) \tag{5}
\]

In Equation 5, \(\lambda\) is the l1 regularization coefficient. A large \(\lambda\) increases the number of zero elements in \(\Lambda\) and \(\Theta\), while a small \(\lambda\) will result in dense estimations of model parameters.

Equation 5 is no longer differentiable, and numerical approaches such as Newton’s method may suffer from slow convergence due to the high-dimensional nature of the problem. We adopt the second-order active set approach in [23] for parameter inference. The algorithm proceeds by constructing the second-order approximating of Equation 5 without l1 regularization term, and solving the descent directions \(\Delta_{\Theta}, \Delta_{\Lambda}\) using the coordinate descent algorithm. The solution algorithm has super-linear convergence rate and works especially well for high dimensional data. Readers may refer to [23] for analytical and theoretical details.

Finally, the mean value of future demand can be calculated as

\[
\tilde{Y} = -\Lambda^{-1} \Theta^T X \tag{6}
\]

and the covariance matrix for this multivariate Gaussian distribution is simply \(\Lambda^{-1}\). This suggests that future demand can be predicted in no time, due to efficient sparse matrix inversion and matrix multiplications with existing techniques.

3.2 Boosting-GCRF

The basic GCRF is a strong learner which tries to model the correlations between every pair of elements in input and output data. But the regularized GCRF model may be considerably weaker than the basic model, due to reduced number of parameters to avoid overfitting. And a relatively weaker model is unlikely to best fit the whole data, especially considering that each training sample may differ significantly from the others. This motivates us to boost the GCRF model for more robust and accurate demand prediction.

In particular, we are interested in the adaptive boosting (AdaBoost) approach [24] to build a strong learner Boosting-GCRF. Ada-boost was originally proposed for classification problems, where the weak model is trained using different distributions of the same training data and different outputs are combined to derive the strong model. Despite its success in classification problems, less efforts have been made in regression problems and there are few studies in combining boosting with probabilistic multi-target regression models. We follow the AdaBoost.RT algorithm [25] to develop our adaptive boosting algorithm for the GCRF model.

Similar to any AdaBoost algorithms, the Boosting-GCRF starts with the input training and test data \(X, Y\), the weight of the sample data \(W\), the weak learning model \(M\), and the number of machines \(T\). For each machine \(t\), we prepare the training data \(D_t\) from the X and train the machine with the data. After each machine is constructed, we evaluate the fitness of the model on the entire training data \(X\) and update the weight of each training sample based on the corresponding error rate. For the GCRF model, the error rate of a training sample \(x_i\) is measured as:

\[
\text{Err}_t(i) = \frac{|y_i^t - y_i|_2}{\sqrt{k_n}} \tag{7}
\]

where \(y_i^t\) is the predicted output from machine \(t\), and the error rate is simply the rooted mean squared deviation.

Different from point estimation regression, the output of the Boosting-GCRF may not be generated by simply taking the weighted average of the results from the \(T\) machines. The reason is that each machine \(t\) models a Gaussian distribution, and the combination of the machines is no longer a Gaussian distribution but a Gaussian mixture model. For Gaussian mixture models, the weighted mean and covariance are calculated as:

\[
E(Y|X) = \sum_{t=1}^{T} p_t E(Y|X, M_t) \tag{8}
\]

\[
\text{Var}(Y|X) = \sum_{t=1}^{T} p_t \text{Var}(Y|X, M_t) + \sum_{t=1}^{T} p_t \left(\left(E(Y|X, M_t) - E(Y|X)\right)^2\right) \tag{9}
\]
where $p_t$ is the associated weight of machine $t$, and the variance of the mixture model is therefore the mixture of the variance plus a dispersion term of the weighted means. As a consequence, this variance will be greater than any of the variance from the individual model. Even if the mean prediction gets improved by boosting, the resulting variance is likely to be over-conservative which provides little value for understanding the true distribution of taxi demand.

To overcome this issue, we introduce the combination rule where the weights are adjusted based on prediction uncertainty [26]. If a machine is uncertain about the prediction, which is equivalent to that the machine has larger variance, the importance of the machine should be discounted. And the combination of the $T$ machines is generated by the product rule rather than the weighted averaging rules following Equation 8 and 9:

$$p(Y|X) = \prod_{t=1}^{T} p(Y|X, M_t)^{p_t} \quad (10)$$

And the expectation and variance of the mixture model can be calculated as:

$$Var(Y|X) = \left( \sum_{t=1}^{T} p_t Var(Y|X, M_t) \right)^{-1} \quad (11)$$

$$E(Y|X) = Var(Y|X) \sum_{t=1}^{T} p_t E(Y|X, M_t) Var(Y|X, M_t)^{-1} \quad (12)$$

In this regard, we avoid the overdispersion on the Gaussian mixture variance and the resulting variance is no worse than the machine with the least prediction uncertainty.

Finally, we present the Boosting-GCRF algorithm as follows.

**Table 1: Boosting-GCRF algorithm**

1. **Input:** Training and test data $X, Y$, the weak learner GCRF with regularizer $\lambda$, number of machines $T$, initial weight $W$.
2. for $t=1:T$
   3. Generate training and test data $X_t$, $Y_t$ based on weight $W$;
   4. Train the machine $M_t$: $f_t(X_t) \rightarrow Y_t$;
   5. Calculate error $Err(i)$ for each sample $X_i$ in X following Equation 7;
   6. Measure the maximum error rate of machine $t$ as $\epsilon_T \leftarrow \max_i Err(i)$
   7. Calculate the exponential loss function for each sample: $L_t(i) \leftarrow 1 - \exp(-Err(i)/\epsilon_t)$
   8. Calculate the average loss function for the machine $L_t \leftarrow \sum_i W(i)L_t(i)$
   9. Set $\beta_t \leftarrow \sum_i W(i)/L_t(i)$
   10. Update the weight $W(i) \leftarrow \frac{W(i)\beta_t^{L_t(i)}}{\zeta_t}$, where $\zeta_t$ is the normalization constant.
3. Calculate model importance $p_t \leftarrow \log(Y_t/\zeta_t)$
4. Calculate $E(Y|X), Var(Y|X)$ following Equations 12 and 11.
5. **Output:** $E(Y|X), Var(Y|X)$

### 3.3 Features

The input and output data of the model are the number of taxi trips per location per time interval. And the input features consist of a vector of trip observations across all the locations during the observed time periods, and the prediction is the forecasted number of trips in the same set of locations over a pre-specified time horizon. Instead of using the actual trip counts, we first scale the counts into the range $[-1, 1]$. The scaling is applied to both $x_{ti}$ and $y_t$ across the all observations, which helps to reduce the variations among the number of trips at different time and locations.

In addition to feature scaling, we further construct additional features based on the time series observations using radial function. In particular, we adopt the Gaussian radial function:

$$h(x_i^t) = e^{-(x_i^t - \mu_i^t)^2} \quad (13)$$

where $\mu_i^t$ denotes the average number of trips for location $i$ at time $t$ across all observations. Equation 13 is mainly used to measure the deviation of a particular feature from the normal scenario. A large deviation usually implies the impact from unusual external events such as bad weather or traffic condition, which is likely to affect subsequent time intervals and neighborhood locations. As a result, the radial function doubles the size of input features, but contributes to extracting additional information from the time series observations.

### 3.4 Benchmark Algorithms

We introduce three benchmark algorithms which are considered as the state-of-the-art approaches in literature for short-term demand forecasting.

#### 3.4.1 Bagging-GCRF.

The bootstrap aggregating or bagging, is another popular approach for constructing strong models from an ensemble of weak learners [27]. We replace the weighted average aggregation of weak learners in conventional bagging with the same product weighting rules in Boosting-GCRF, and develop the Bagging-GCRF approach as the first benchmark algorithm.

#### 3.4.2 Artificial Neural Network.

The artificial neural network (ANN) is known to be capable of modeling the nonlinearity of the data [28], and we use the feed-forward network as the second benchmark algorithm. Theoretical work suggested that single hidden layer is sufficient to model any complex nonlinear relationship [29], and we choose the structure of single layer with 10 neurons based on our experiments. We also add the regularization term in the form of the sum of squares of the network weights to avoid overfitting issue.

#### 3.4.3 Boosting regression tree.

We follow the Adaboost.R2 algorithm [30] to construct an ensemble of regression trees for the short-term demand prediction. For each decision tree regressor, the maximum depth is set to 3. And 200 weak trees are trained using the square loss function.

#### 3.4.4 Gradient boosting tree.

The gradient boosting method can be viewed as a combination of boosting regression tree and gradient descent approach. The main idea of the gradient boosting tree algorithm is that it construct new learners to minimize the loss function and additively adding new learners to the existing
4 NUMERICAL RESULTS

4.1 Experiment Setting

All experiments are conducted using MATLAB. The built-in neural network library is used to perform the training and testing of the ANN models. We use 2015 NYC taxi trip data which contain 365 observations. We partition the data into two sets, with 75% of them being the training set and the rest 25% of the data serving as the stand-alone test set. The model performances are compared by first conducting 5-fold cross-validation over the training set, and then train the model using the complete training data and evaluate the performance on the test data. Each sample contains the taxi data of 96 time steps across the study area, where each prediction time step corresponds to the time interval of 15-minutes. We create 14 experimental scenarios by various segments of the each sample with different start and end time steps.

4.2 Performance Measures

Two metrics are used for evaluating the algorithm performances in this study. The first metric is the root means square error (RMSE), which measures the deviations of predictions from the observations and is calculated as

\[
RMSE = \frac{1}{D} \sum_{d=1}^{D} \sqrt{\frac{1}{kn} \sum_{i=1}^{n} \sum_{t=1}^{k} (y^d_{d,i} - y^{d,*}_{d,i})^2}
\]

where \(D\) is the number of observations in the validation data, \(y^d_{d,i}\) is the predicted demand at location \(i\) at time interval \(t\) for observation \(d\), and \(y^{d,*}_{d,i}\) is the observed demand correspondingly.

The second metric we used is the modified mean absolute percentage error (mMAPE). It differs from the conventional MAPE by addressing the value of observations being zero. Instead of averaging over the error of individual observations, the mMAPE measures the sum of errors for all observations over the summation of the observations as:

\[
mMAPE = \frac{1}{D} \sum_{d=1}^{D} \frac{\sum_{i=1}^{n} \sum_{t=1}^{k} |y^d_{d,i} - y^{d,*}_{d,i}|}{\sum_{i=1}^{n} \sum_{t=1}^{k} y^d_{d,i}}
\]

4.3 Discussion

We first discuss the choice of regularization coefficient for GCRF, as well as the number of learners for boosting and bagging methods. We evaluate the corresponding RMSE for different \(\lambda\) values by averaging the RMSE from the 5-fold cross-validation, and the result is presented in Figure 4(a). The result corresponds to the scenario by using data from 8AM to 1PM to predict taxi demand from 1PM to 2PM (4 steps ahead). When \(\lambda > 1\), the base GCRF model is observed to be almost completely sparse with high errors. On the other hand, as we keep decreasing \(\lambda\), the model becomes too complex and it overfits the data. Based on the results of all scenarios, we choose \(\lambda = 0.05\) and this value is also used for boosting and bagging methods. While we do not have a stopping criterion for boosting and bagging methods, the number of learners for each method is determined by choosing the value that corresponds to the smallest test RMSE based on the cross-validation results. As can be seen from Figure 4(b), the performance gets stabilized as we increase the number of learners to over 60, for both boosting and bagging methods. As a consequence, the number of optimal learners for bagging and boosting are set to 60.

Table 2 presents the cross-validation performance of the 6 algorithms under 14 different scenarios. The 14 scenarios correspond to prediction tasks at 7 different time the day with two different prediction lengths: 4 and 8 steps into the future. The start and the end in the table define the segments of time used for training, and the predict value refers the number of time steps to forecast the taxi demand. Cases 1 and 8 can be viewed as prediction during morning peak hours (MP), cases 2-5 and 9-12 correspond to off-peak time (OP), and the rest of the cases refer to evening peak time (EP). The
The previous results are obtained from the cross-validation of the training set, and we next present the results on the stand-alone test set in Table 3. Since the total number of trips varies from time to time, the RMSE only provides a measure among different methods for the same scenario, and the mMAPE helps to compare performances between scenarios. It can be seen from the table that the average RMSE from the test set is in general consistent with that of the training set. Boosting-GCRF is still the best model and the average performance is found to be even better than that of the training data, with the improvements of 4.96% in RMSE and 5.63% in mMAPE as compared to the base GCRF model. Moreover, the best mMAPE achieved by the boosting-GCRF is 10.4%, and the biggest difference over base GCRF is 11.8% (case 4). MP cases 1 and 8 are again the two cases that are comparatively harder to predict for all models, which can be explained by the same reasons discussed above. It can be concluded from the results that the boosting-GCRF model is a superior method for modeling short-term taxi demand, and the level of accuracy (12.4% mMAPE in average) may be well suited for real-world applications.

As the average mMAPE and RMSE suggest that the boosting-GCRF model is better than the base GCRF model, we next assess how the differences look like if we zoom into locations with different demand level. Figure 5 presents the snapshots of the prediction performances corresponding to case 11, but we extend the prediction intervals from 8 to 12 time steps ahead. The results are averaged over all testing samples. We visualize the location with highest average demand, and two locations with relatively low number of ridership. We also plot the prediction performance by averaging over all the locations in the study area to understand the mean performance at each time step. Figure 5(a) indicates that both GCRF and boosting-GCRF perform well for high demand location, with worst mMAPE being lower than 8% and the predicted values resemble the shape of the observed number of trips. And this issue is alleviated if we predict 8 steps ahead instead of 4 (case 8 versus case 1), since GCRF also explores correlation among future predictions through λ, and 8 steps prediction provides more information to better calibrate the matrix compared with only predicting 4 steps. Finally, the bagging-GCRF is found to have similar performance as compared to base GCRF model in all scenarios, which suggest the ineffectiveness of bagging method in improving the well-regularized GCRF model. Note that the performance is always associated with the bias-variance trade-off, and the regularization is introduced to reduce the variance of a complex model with low bias. While bagging approach is known to improve model performance by reducing variance, the ineffectiveness of bagging for the base GCRF model implies that the base model is well regularized with the appropriate choice of regularization term. Finally, while boosting regression tree and gradient boosting tree methods are found to perform way worse than the other methods, we only focus on the comparison of other four approaches in the rest of the study.

The previous results are obtained from the cross-validation of the training set, and we next present the results on the stand-alone test set in Table 3. Since the total number of trips varies from time to time, the RMSE only provides a measure among different methods for the same scenario, and the mMAPE helps to compare performances between scenarios. It can be seen from the table that the average RMSE from the test set is in general consistent with that of the training set. Boosting-GCRF is still the best model and the average performance is found to be even better than that of the training data, with the improvements of 4.96% in RMSE and 5.63% in mMAPE as compared to the base GCRF model. Moreover, the best mMAPE achieved by the boosting-GCRF is 10.4%, and the biggest difference over base GCRF is 11.8% (case 4). MP cases 1 and 8 are again the two cases that are comparatively harder to predict for all models, which can be explained by the same reasons discussed above. It can be concluded from the results that the boosting-GCRF model is a superior method for modeling short-term taxi demand, and the level of accuracy (12.4% mMAPE in average) may be well suited for real-world applications.
And both approaches are found to perform poorly for certain low demand locations, with the mMAPE being as high as 45%. Though the absolute difference is not that significant considering the small magnitude of trip counts, cautions should still be exercised for implementing the prediction for these places. If we average over all places, as shown in Figure 5(d), the boosting-GCRF is found to outperform the base GCRF in almost all time steps.

In additional to average performances, we are also interested in the robustness of the prediction results. There exist many anomaly events such as bad weather condition and large sports events, which may significantly affect the distribution of taxi ridership. And we present the prediction performances under anomalies in Figure 6. The results in Figure 6 are obtained from scenario 11, where we train the model using the whole training set and validate the model for forecasting short-term taxi demand. Numerical results suggest that the boosting-GCRF model is robust to anomalies, which has the mMAPE being 31.9% better than base and bagging GCRFs. And the ANN is also found to perform better than base GCRF under anomaly scenario. The most noticeable difference is for demand between 100-200, where GCRF and bagging-GCRF are found to overestimate the number of trips by a considerable margin, and the Boosting-GCRF tends to perform well for all trip values. Meanwhile, the ANN model well predicts the value for majority of the points, but it also suffers outliers with excessive errors.

Finally, as the previous results are associated with the point estimation results, we discuss how well the models represent the actual distribution of short-term taxi trips. The way to understand how well the model captures the distribution of the data is by constructing the confidence interval, and a well fitted distribution should cover a% of the data with a% confidence level. While it is easy to do so for univariate distributions, it is non-trivial to build the confidence regions of multi-variate distributions especially for high-dimensional cases in our study. One approach is to take the Cartesian product of a% confidence interval for each dimension, but the chance that the resulting confidence region does not contain the mean for at least one of the dimension will be way higher than $1-a$. To address this issue, we apply the Bonferroni correction. That is, to reach the overall confidence level of a%, we take the joint confidence interval of each dimension with the individual confidence level being $1-(1-a)/T$, where $T$ is the total number of steps to predict. We then evaluate the percentage of samples that are contained within the confidence region constructed at each location across prediction time steps. The results are presented in Table 4. We average the performance for MP, OP, and EP cases based on our discussion of Table 2. It can be seen from the results that both base GCRF and bagging-GCRF are less confident with the prediction by giving excessively wide confidence regions. This probabilistic estimation is therefore less informative and has little value for deriving stochastic operation strategies in real world. On the other hand, the Boosting-GCRF performs well in all three cases. And the coverage level are very close to all of the confidence levels in OP and EP cases. The only exception is for the 68% confidence level during morning peak, where the boosting-GCRF is also found to have worse performance according to previous discussions. A possible amendment to this issue is to train a specific model with different choice of regularization term and different number of weak learners. But in general, we may arrive at the conclusion that the boosting-GCRF is very effective in terms of both point-estimation results as well the probabilistic prediction performances.

### Table 4: Coverage level

<table>
<thead>
<tr>
<th></th>
<th>Confidence</th>
<th>GCRF</th>
<th>Boosting-GCRF</th>
<th>Bagging-GCRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>68%</td>
<td>0.970</td>
<td>0.783</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>0.996</td>
<td>0.827</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>0.999</td>
<td>0.879</td>
<td>0.999</td>
</tr>
<tr>
<td>OP</td>
<td>68%</td>
<td>0.970</td>
<td>0.783</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>0.996</td>
<td>0.827</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>0.999</td>
<td>0.879</td>
<td>0.999</td>
</tr>
<tr>
<td>EP</td>
<td>68%</td>
<td>0.970</td>
<td>0.783</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>0.996</td>
<td>0.827</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>0.999</td>
<td>0.879</td>
<td>0.999</td>
</tr>
</tbody>
</table>

5 CONCLUSION

In this study, we analyze the characteristics of short-term taxi demand based on real-world trip data, and develop the boosting-GCRF model for forecasting short-term taxi demand. Numerical results suggest that the boosting-GCRF may reach the best mMAPE of 10.4%, be robust against demand anomalies, and well model the actual distributions of the observations across space and time. For future works, we plan to fine tune the model to improve the performance for low-demand areas, introduce data fusion to incorporate other time-varying factors into our current modeling framework, and apply the model in other prediction tasks, such as short-term traffic flow and travel speed predictions.

ACKNOWLEDGMENT

The algorithmic development in this work is partially funded by the NSF grant 1520338. The authors take full responsibility for the findings and do not represent the views of NSF.
REFERENCES


