Prediction and optimization of sharing bikes queuing model in grid of Geohash coding

Kui Yu¹ and Changyue Qu²

Abstract
Dockless bike-sharing systems provide parking anywhere feature and environment-friendly approach for commuter. It is booming all over the world. Different from dockless bike-sharing systems, for example, previous studies focus on rental mode and docking stations planning. Yet, due to the fact that human mobility patterns of temporal and geographic lead to bike imbalance problem, we modeled human mobility patterns, predicted bike usage, and optimized management of the bike-sharing service. First, we proposed adaptive Geohash-grid clustering to classify bike flow patterns. For simplicity and rapid modeling, we defined three queuing models: over-demand, self-balance, and over-supply. Second, we improved adaptive Geohash-grid clustering-support vector machine algorithm to recognize self-balance pattern. Third, based on the result of adaptive Geohash-grid clustering-support vector machine, we proposed Markov state prediction model and Poisson mixture model expectation-maximization algorithm. Based on data set from Mobike and OFO, we conduct experiments to evaluate models. Results show that our models offer better prediction and optimization performance.

Keywords
Bike-sharing systems, Geohash coding, grid state, queuing model, Markov chain, expectation-maximization

Introduction
With the development of technology, dockless bike-sharing systems (BSSs) have solved the last mile problem in intelligent city life.¹ BSSs are booming all over the world, especially in large cities. In the traditional self-service mode, users have to rent or return bike sharing at fixed stations. Based on mobile Internet, global positioning system (GPS), and location-based service (LBS), BSSs allow users to start or end service in community curbside, subway stations, and central business district (CBD) parking zone.

Since about 2015, the central problems for municipal administration to solve include acquiring space to park the bikes and achieve efficient use of the bikes. According to bike sharing park-anywhere feature, the core of the issue is focused on two factors:

1. Attribute to the human mobility patterns and spatiotemporal factor, the imbalance problem is difficult to model and predict in dockless BSS.
2. The phenomenon of “bike-sharing graveyard” takes place anywhere, for example, curbsides, which blocks the path of pedestrian.

This is a supply and demand planning problem that changes with temporal and geographic.² The truck-based³ and the user-based approaches are two baseline approaches to solve the bike imbalance issue.

However, the truck recycling approach depends on demand prediction and manual intervention.

Motivation and incitement
In dockless BSSs, bikes are widely used in our daily life. In recent study, many previous methods focus on demand prediction in dock BSSs. The researcher had implemented different strategies to address the occurrence of rebalance, such as sending cargo trucks to relocation bikes before rush hours. Due to lack of supervision and control strategy, the BSS is toward extreme phenomenon. That is too many illegal parking

¹School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
²College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing, China

Corresponding author:
Kui Yu, School of Computer Science, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Beijing 100876, China.
Email: yukui@bupt.edu.cn

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
at curbside, bus station, community, and so on. The imbalanced usage pattern of bikes causes over-demand and over-supply issues not only to commuter but also to cities. Motivated by the aforementioned challenge, we examine three methods: station-centric model with global features, demand prediction, and free-floating bike-sharing model.

Based on the limitations of approach, our prediction model by adaptive Geohash-grid clustering (AGC) technology to preprocess parking coordinates data. We describe the stage of imbalance changed in different Geohash grids. Then, we propose a dockless sharing bike demand modeling: over-demand model, over-supply model, and self-balance model. Expectation-maximization (EM) algorithm is derived to learn the parameters of Poisson mixture model (PMM). According to queuing theory, we modeled bike usage patterns and human mobility patterns, predicted demand, and optimized the parameters.

We also showed that the prediction and optimization algorithms improve convergence and achieve a better performance compared with existing algorithms.

**Literature review**

This section summarizes research work on modeling, demand prediction, and parameter optimization. Mathematical modeling is the first step in solving prediction and optimization problems. From the viewpoint of BSS designers, route determination and demand prediction are relocated. Parameter value is a critical decision indicator for managers and controllers to optimize BSS. In our literature review, we concentrate on spatio-temporal, demand, and rebalance problems.

There are many researchers focusing on BSS modeling. Mathematical modeling has been used in recent work such as planning model, probability model, clustering algorithms, loss function, and so on. Especially, the auto-regressive integrated moving average models and auto-regressive moving average (ARMA) models are widely used in modeling of human mobility patterns. According to BSS stations usage profile, Sayarshad et al. proposed a multi-periodic optimization formulation for planning problems. Crisostomi et al. propose a Markov chain model. In the BSS, Markov decision process was a tool to solve commuter mobility patterns. Based on the discriminative functional mixture (DFM) model, Bouveyron et al. proposed FunFEM methodology. To minimize the total cost, Hu and Liu proposed allocation model to solve bike rental stations and truck dispatching depot problems. According to travel patterns, MY Du et al. proposed a multinomial logit (MNL) model. In free-floating bike-sharing model of MNL consists of three categories, such as origin to destination pattern (ODP), travel cycle pattern (TCP), and transfer pattern (TP).

In order to redistribute the number of bikes, spatio-temporal analysis method is widely used in prediction algorithm. It is help to analyze a strategic design model for BSS. According to temporal and spatial factors, Yang and Hu proposed a spatiotemporal bicycle mobility model. The temporal and geographic mobility patterns are applied to demand prediction. Based on spatiotemporal analysis, Froehlich et al. proposed clustering technique in Barcelona’s BSSs.

From the viewpoint of bike-sharing operators, demand prediction is a critical performance indicator. We can predict the number of rental bikes according to cluster algorithm. Ciancia et al. propose station occupancy predictor. This is a data mining framework to predict the occupancy levels of the stations by Bayesian and associative classifiers. Based on station usage, long-term stability, and short-term volatility, Yao et al. proposed three-step demand estimation model. However, tensor factorization is widely used in routing prediction of BSS. YX Li et al. proposed a hierarchical prediction model by tensor factorization to extract latent user activity patterns. Based on clustering algorithm to forecast bikes’ and docks’ availability for each station, Gaussian mixture model (GMM) and PMM are common prediction schemes. LB Chen et al. propose two bikes share pool sizing techniques which guarantee bike availability with high probability. Based on resource of bike-sharing optimization, similar problems are single vehicle one commodity capacitated pickup and delivery problem (SVOCPDP), one commodity pickup and delivery traveling salesman problem (1PDTSP), Swapping Problem, and split delivery problem. These optimization algorithms above mentioned aim to find a minimum cost route for users to renting and returning bicycles. Mostly, split delivery problem and branch-and-cut algorithm are solved through a tabu search algorithm. The SVOCPDP gathers aspects from both the Swapping Problem and the 1PDTSP. The objective of optimization is to find the costless function by heuristics algorithm. In the literature, artificial immune systems support vector machine (AIS-SVM), artificial neural network (ANN)-SVM, and particle swarm optimization methods have been widely applied in optimization and classification problems.

Currently, some papers adopt Monte Carlo simulation to predict the demand of cluster. Simulation results are different from real results. In our review of the related literature, there is little research available regarding modeling and predicting user behaviors in dockless systems. In this paper, we focus on predicting
imbalance stage and optimized parameter in dockless BSSs. Based on the real-world bike-sharing data set from Beijing city, we combined Geohash coding and queuing theory approaches to improve the SVM and to optimize the parameters of the EM algorithm to adjust the weight value in PMM.

**Contribution and paper organization**

As is evident from the review, there is abundant research regarding BSSs. To tackle our problem, we carefully design solutions to overcome the above drawbacks of literature. Based on geographic-grid clustering, we proposed AGC approach preprocessing check-in/out from parking data set. Starting from parking anywhere and mobility point, we research on improved SVM (ISVM) classification and optimized EM parameter in queuing models.

Our main contributions are as follows:

1. For bike flow pattern parking-anywhere problem, we formally define the state of over-demand, over-supply, and self-balance by processing coordinate data, transition state modeling by queuing theory.
2. For demand prediction, we improved bi-classification algorithm to solve three-stage classification problem. Based on Markov state prediction (MSP), we propose AGC-SVM cope with dynamic demand.
3. For rebalance systems, we proposed PMM-EM model by PMM-EM algorithm.

The optimized parameters $\pi_L$ and $\lambda_L$ guarantee high probability of bike system running.

4. In order to validate our methods, we use real-world data under eight baseline methods, such as perceptron, decision tree, gradient boosting regression tree (GBRT), $k$-nearest neighbor ($k$-NN), $k$-means, ARMA, Kalman filter, and hidden Markov model (HMM). We adopt measure accuracy of classification and regression algorithm through root mean squared error (RMSE), root mean squared logarithmic error (RMSLE), error rate, precision, recall, and F1 to evaluate that our models outperform significantly.

This paper is organized as follows: in section “Overview,” we define the terms used in this paper as follows.

**Definition 1: Geohash coding.** Geohash is a public domain geocode system invented in 2008 by Gustavo Niemeyer. The Geohash encoding generated by latitude is stored in List 1, whereas that generated by longitude is stored in List 2. Lists 1 and 2 encodings are merged. The odd number positions denote the latitude, whereas the even number positions indicate the longitude. A total of 32 characters, namely, $0$–$9$ and $b$–$z$ (remove $a$, $i$, $l$, and $o$), are used for the base 32 encoding, that is, List 3. The purpose of privacy protection is to publish different encoding lengths. For example, the 6-bit code can represent a range of approximately 0.34 km$^2$. A length of 7 bit can represent the range of 76 m$^2$, as shown in Table 2.

**Definition 2: Geohash grid.** Geospatial index technique is a search method that efficiently deals with the roads, streets, and districts data. The grid index uses a hash data structure. Each grid corresponds to a bucket of the hash map (Figure 5).

**Definition 3: Bike flow patterns of check-in/out.** In a given time window, $[t, t + \Delta t]$ is defined as a tuple $X_t = \{X_{in}, X_{out}\}$, where $X_{in}$ and $X_{out}$ are the number of bikes’ start and end services from Geohash grid during $[t, t + \Delta t]$, respectively. We define $X_{in}$ and $X_{out}$ as the start and end bike services, shown as $X_{in}\{x_{in}^1, x_{in}^2, \ldots, x_{in}^n\}$ and $X_{out}\{x_{out}^1, x_{out}^2, \ldots, x_{out}^m\}$. Check-in/out values mean the number of bikes that activate or terminate sharing bike service in $\Delta t$, denoted as $X_{in}^{\Delta t}$ and $X_{out}^{\Delta t}$.

$$\sum_{i=1}^{n} \Delta X_i = \sum_{i=1}^{n} |x_{in}^{out} - x_{in}^{in}|$$  \hfill (1)

**Definition 4: Grid state.** Sharing bikes is used by commuters who pick it up anywhere and anytime. According to statistic consequence from data set $X_{in}^t$ and $X_{out}^t$, we find two states and one quasi-state. We define $R_s$ grid state as over-demand, $R_b$ grid state as over-supply, and $R_c$ grid state as self-balance.

**Definition 5: Queuing theory.** Queuing theory is the mathematical study of waiting lines or queues. In this paper,
we explore queuing theory by modeling and analyzing the number of sharing bikes for service, waiting times, and so on. In Geohash grid, sharing bikes are considered customers and parking is defined as entering the queuing system. Queue represents customers or sharing bikes waiting for service. We propose three queuing models, described by Kendall’s notation (the standard system used to describe and classify a queuing node).

For example: $M/M/1/\infty$ describes over-demand stage, $M/M/1/K$ describes over-supply stage, and $M/M/S/K$ describes self-balance stage.

**Definition 6: Time Window $\Delta t$.** In dynamic patterns of bike sharing, we researched the range of time 06:00 a.m–24:00 p.m–01:00 a.m. It was divided into $36\Delta t$ (30 min per section) in training data set.
Figure 3. Over-supply stage modeling.

\[ p^{(A)0} = \frac{1}{1 + \sum_{n=1}^{K} \rho^n} = \left\{ \begin{array}{cl} \frac{1 - \rho}{(1 - \rho)K}(p \neq 1) \\ \frac{1 - \rho}{1 - \rho K}(p = 1) \end{array} \right. \] (4)

When \( K \neq 1 \) and \( K \neq 0 \). The number of customers waiting in line for bikes is obtained as

\[ L_A = \sum_{n=1}^{K} np_n = p_0 \rho \sum_{n=1}^{K} np^n = \rho \left( \frac{1}{1 - \rho K} \right) = \frac{(K + 1)\rho^{K + 1}}{1 - \rho K + 1} \] (5)

- Shown as (8): It means the systems are kept at over-demand stage.
- Shown as (3): Following the changed bikes’ flow pattern, arrow (3) means \( x_{i+1}^{out} < x_{i}^{in} \) in BSS. The grid state is transformed from A stage to C stage.
- Shown as (1): Following the growth of commuter terminated service, arrow (1) means \( x_{i+1}^{out} < x_{i}^{in} \) in BSS. The grid state is transformed from A stage to B stage.
- Label B: Over-supply stage

Label \( R_B \) describes over-supply stage. This state means a lot of sharing bikes arriving at this region \( x_{i+1}^{out} < x_{i}^{in} \) in \( \Delta t \), as shown in the following equation

\[ x_{i+1}^{out} - x_{i}^{in} < -\beta \] (6)

\[ \sum_{i=1}^{N} x_{i}^{out} - \sum_{i=1}^{N} x_{i}^{in} < -\beta \] (7)

For example, in subway station region, we named over-supply stage as \( R_B \). The capacity of the \( R_B \) is \( K \). We proposed over-supply \( R_B \) stage by \( M/M/1/K \) queuing theory, as shown in Figure 3.

In this model, the idle probability of sharing bikes is

\[ p^{(B)0} = \frac{1}{1 + \sum_{n=1}^{\infty} \rho^n} = \left( \sum_{n=0}^{\infty} \rho^n \right)^{-1} = 1 - \rho \]

\[ p^{(B)n} = p^n p_0 = (1 - \rho)p^n (n = 0, 1, 2 \ldots) \] (8)

Figure 4. Self-balance stage modeling.

\[ L_B = \sum_{n=0}^{\infty} np_n = n(1 - \rho)\rho^n \]

\[ = (\rho + 2\rho_2^2 + 3\rho_3^3 + \cdots) - (\rho_2 + 2\rho_3 + 3\rho_4 + \cdots) \]

\[ = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda} \] (9)

- Shown as (9): It means that the systems are kept at over-supply stage.
- Shown as (2): Following the changed bikes’ flow pattern, arrow (2) means \( x_{i+1}^{out} > x_{i}^{in} \). The state is transformed from B stage to C stage.
- Shown as (6): Following the growth of commuter, arrow (6) means \( x_{i+1}^{out} > x_{i}^{in} \) in BSS. The state is transformed from B stage to A stage.
- Label C: Self-balance stage

\[ R_C \] stage is self-balance. It means \( x_{i+1}^{out} = x_{i}^{in} \) as shown in the following equation

\[ \left| x_{i+1}^{out} - x_{i}^{in} \right| \leq \pm \beta \] (10)

\[ \left| \sum_{i=1}^{N} x_{i}^{out} - \sum_{i=1}^{N} x_{i}^{in} \right| \leq \pm \beta \] (11)

Different from defined \( R_A \) and \( R_B \) before, this queuing system have several service centers. The capacity of the \( R_C \) stage is \( K \). We named self-balance stage as \( R_C \) stage by \( M/M/S/K \) queuing theory. In self-balance stage, the grid of \( R_C \) has high user density and high demand for sharing bikes. The working of the model is shown in Figure 4.

In this model, the number of supplemented shared bikes is

\[ \mu^{(C)n} = \left\{ \begin{array}{l} n\mu(n = 1, 2, \ldots) \\ s\mu (n = s, s + 1 \ldots) \end{array} \right. \] (12)

The probability of \( n \) bikes in the platform is

\[ p^{(C)n} = \left\{ \begin{array}{l} \frac{n^n}{n!} p_0 (n = 1, 2 \ldots) \\ \frac{\rho^n}{s!} p_0(n \geq s) \end{array} \right. \] (13)

The idle probability of sharing bikes is
tagging, we transformed the data into grid layers for Geohash code. Each point in the map, as shown in Figure 5. For Geohash coding, we use the Geohash algorithm to process parking data. Latitude and longitude coordinates are transformed into grid layers with a grid shape. Geohash coding means longitude and latitude points are transformed into a grid shape. In order to find the statistic of bikes, we proposed AGC. First of all, we use the Geohash coding algorithm to process parking data. Geohash-grid clustering.

\[
p(C|0) = \left[\sum_{n=0}^{n-1} \frac{\rho^n}{n!} + \frac{\rho^s}{s!(1-\rho)}\right]^{-1}
\]

The mean number of waiting users is denoted by

\[
L_C = L_Cq + s + p_n \sum_{n=0}^{n-1} \frac{(n-s)\rho^n}{n!}
\]

- Shown as (4): It means the systems are kept at self-balance stage.
- Shown as (7): Following the changed bikes’ flow pattern, arrow (7) means \(x^\text{out}_{i_{\Delta}} > x^\text{in}_{i_{\Delta}}\) in BSS. The state is transformed from stage C to stage A.
- Shown as (5): Following the growth of commuters, arrow (5) means \(x^\text{out}_{i_{\Delta}} < x^\text{in}_{i_{\Delta}}\) in BSS. The grid state is transformed from stage C to stage B.

### Methodology

We formulate bike flow pattern as a Poisson process and modeled by queuing system. Based on predicted imbalance stage and optimized rebalance parameters, we improved AGC and AGC-SVM approaches. We proposed PMM-EM model in which key components of problems are described in the following subsection.

#### AGC

Geohash coding means longitude and latitude point data transform algorithm. It is a hierarchical spatial data structure that subdivides space into buckets with a grid shape. In order to find the statistic of bike, we proposed AGC. First of all, we use the Geohash coding algorithm to process parking data. Latitude and longitude coordinate data are transformed into grid layers on the map, as shown in Figure 5. For Geohash code tagging, we statistics \(x^\text{out}_{i_{\Delta}}\) and \(x^\text{in}_{i_{\Delta}}\) from every Geohash grid during \(\Delta t\). Bike flow pattern is denoted as \(\sum_{i=1}^{n-1} x^\text{out}_{i_{\Delta}}\) and \(\sum_{i=1}^{n-1} x^\text{in}_{i_{\Delta}}\) in the Geohash grid. We process the training data set as follows

\[
\sum_{i=1}^{n} \Delta x_i = \sum_{i=1}^{n} x^\text{out}_{i_{\Delta}} - \sum_{i=1}^{n} x^\text{in}_{i_{\Delta}} = \{\Delta x_1(x^\text{out}_{i_{\Delta}} - x^\text{in}_{i_{\Delta}}), \Delta x_2(x^\text{out}_{i_{\Delta}} - x^\text{in}_{i_{\Delta}}), \ldots, \Delta x_n(x^\text{out}_{i_{\Delta}} - x^\text{in}_{i_{\Delta}})\}
\]

A cluster consists of all density of parking points, for example, WX4EQY tagging over-supply stage and WX4EQQV tagging over-demand stage.

With the same Geohash coding length of the prefix letter, we evaluated the statistical result by threshold parameters \(\beta\) and \(\sum_{i=1}^{n} \Delta x_i\) relationship. Therefore, the statistical bikes’ flow problem and pattern recognition are done by AGC approach.

According to the number of bikes, we adjust Geohash-grid size by choosing suitable prefix lengths. Different periods have different usage stages in the same Geohash grid. The adaptive Geohash grid not only predicted imbalance stage but also protect privacy. In next subsection, we introduce stage label classification by ISVM algorithm.

#### ISVM label classification

In this subsection, we describe how to create stage label classification in training data. To predict three states of Geohash-grid bike flow pattern, we adopt three label stages in queuing models. Because of self-balance, it is a fuzzy stage. Classification of bike flow pattern is convex quadratic programming. Support vector machine is good at bi-classification problem, as the method can significantly reduce the need for labeled training instances. Intuitively, a good separation is achieved by hyperplane that has the largest distance to the nearest point in any labels classification from training data. The details of the ISVM classification algorithm are illustrated as follow:

Hyperplane of state-label classification

\[
\begin{align*}
\omega^T x - \beta &= 0; X_{label} \in R_A \lor R_B \\
\omega^T x - \beta &= 1; X_{label} \in R_A \lor R_C \\
\omega^T x - \beta &= -1; X_{label} \in R_C \lor R_B
\end{align*}
\]

Geometric margin

\[
\beta = \frac{2}{\|\omega\|}
\]

Because of fuzzy stage boundaries and label classification, we improved SVM algorithm. This solution can be distinguished as “over-demand and self-balance” and “over-supply and self-rebalance” stages. The fundamental idea behind SVM is to choose the hyperplane with the maximum margin \(\beta\), that is, the optimal canonical hyperplane. The geometric margin problem has become a convex minimization problem, as shown in...
We can obtain an equivalent formulation of minimizing $v_k^k$. Objective function is $\text{Min} \frac{1}{2} v_k^k$, from Lagrange duality transition, dual variable is $L(v, b, a) = \frac{1}{2} v_k^k - C_0 \sum_{i=1}^n a_i y_i v_i x_i$.

To do this, one needs to find the weight vector $v$ and the bias $b$ that yield the maximum margin among all possible separating hyperplanes, the state is come from the hyperplane that maximizes

$$\text{Max} \frac{1}{2} \sum_{i=1}^n a_i (y_i (a x_i - b) - 1)$$

For simplification and calculation of modeling, we named AGC-MSP. Different lengths of Geohash coding have different bike flow patterns toward AGC. AGC-MSP is a soft clustering method. It was used to predict imbalance stage and deployed bikes in advance. In application, these two scenarios for bike flow pattern prediction can be used complementarily as temporary stage scenarios and permanent state scenarios. We obtain the number of bike that come in and come out from Geohash grid. We process training data set that is recorded in the parking.

**Table 3.** Bikes and states data set in day 1.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|
| A | C | B | C | B | C | A | B | B | B | A | A | C | C | B | B | C |
| 57 | 9 | 16 | 4 | 8 | 20 | 7 | 38 | 21 | 11 | 9 | 38 | 82 | 6 | 3 | 8 | 14 | 1 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| A | B | A | C | A | A | C | B | B | A | B | A | A | A | A |

**Table 4.** Part of state transition in day 2.

| No. | 2nd | 17th | 30th |
|-----|-----|------|------|
| State | A  | B  | C  | A  | B  | C  | A  | B  | C  |
| Probability | 0.538 | 0.152 | 0.37 | 0.386 | 0.333 | 0.279 | 0.365 | 0.352 | 0.279 |

**Figure 6.** Fuzzy boundaries problem.

Figure 6. We can obtain an equivalent formulation of minimizing $\|a\|$. Objective function is $\text{Min} \frac{1}{2} \|a\|^2$, from Lagrange duality transition, dual variable is $L(a, \beta, \alpha) = \frac{1}{2} \|a\|^2 - \sum_{i=1}^n a_i (y_i (a x_i - \beta) - 1)$.

To do this, one needs to find the weight vector $a$ and the bias $\beta$ that yield the maximum margin among all possible separating hyperplanes, the state is come from the hyperplane that maximizes

$$\text{Max} \sum_{i=1}^n a_i (y_i (a x_i - \beta) - 1)$$

For simplification and calculation of modeling, we named AGC-MSP. Different lengths of Geohash coding have different bike flow patterns toward AGC. AGC-MSP is a soft clustering method. It was used to predict imbalance stage and deployed bikes in advance. In application, these two scenarios for bike flow pattern prediction can be used complementarily as temporary stage scenarios and permanent state scenarios. We obtain the number of bike that come in and come out from Geohash grid. We process training data set that is recorded in the parking.

Markov state transition rate matrix formula

$$P(1, n) = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}$$

**Temporary stage scenarios.** To answer the first question “which Geohash grid is imbalance state,” we first identified bikes’ flow check-out and check-in for each AGC. According to previous studies pointing out state label classification by threshold parameter $\beta$, we defined state table of transition matrix, as shown in Table 3.

The $R_A$, $R_B$, or $R_C$ state had been predicted by the Markov predicted approach. In practice, we get historical data {57, 9, –16, 4, 8, –20, 7, 38, –21, –11, –9, 38, 82, 6, 3, –8, –14, 1, 98, –15, –18, 36, 4, 51, 72, 8, –12, –9, –5, 39, –3, –7, 45, 36, 29, 22} in WX4G8C9V Geohash-gird. State labels are shown in Tables 3 and 4.

Transient evolution is
Markov state transition rate matrix is in Figure 7. Geohash coding is WX4G8C9V (19 m2), as shown in Figure 7.

Rebalance parameter optimization

The bike flow pattern can be predicted by multiple times of Markov chain iteration in Geohash grid. In grid of \( R_A \) state, the strategy is allocation bike in the morning. In grid of \( R_B \) state, the strategy was recycled bike at night. Depending on operational experience and statistical summary, the deployed and recycled problems are uncertain.

So far we have obtained the grid state obeys Poisson distribution with different parameters. Therefore, how to optimize resource of deployed and recycled is rebalance issue. From observed parking data set, we can hardly point out the distribution exactly. Throughout the day, we define that the bike usage obeys Poisson mixed distribution.

We propose a prediction method to determine the final grid of state by Markov state transition matrix. Judging from state of Geohash grid, \( \lambda_A \) and \( \beta \) are the important threshold parameters. \( \lambda \) is the bike flow pattern with regard to data sets \( X_{in} \) and \( X_{out} \) and \( \beta \) is the geometric margin in SVM algorithm.

Rebalance parameter optimization

The bike flow pattern can be predicted by multiple times of Markov chain iteration in Geohash grid. In grid of \( R_A \) state, the strategy is allocation bike in the morning. In grid of \( R_B \) state, the strategy was recycled bike at night. Depending on operational experience and statistical summary, the deployed and recycled problems are uncertain.

Because of sharing the mobility of bikes, the three states of the grid alternate every day. To answer the second questions “how to cope with rebalance issue,” we proposed PMM-EM model. In different time, the grid state obeys Poisson distribution with different parameters. Therefore, how to optimize resource of deployed and recycled is rebalance issue. From observed parking data set, we can hardly point out the distribution exactly. Throughout the day, we define that the bike usage obeys Poisson mixed distribution.

So that, the probability of sample \( X_i \) is

\[
p(x, \varphi) = \sum_{L=1}^{\infty} \pi_L P(x = k | \lambda_k) = \begin{cases}
P_A(X = k) = \frac{\lambda_A^k}{k!} e^{-\lambda_A} \\
P_B(X = k) = \frac{\lambda_B^k}{k!} e^{-\lambda_B} \\
P_C(X = k) = \frac{\lambda_C^k}{k!} e^{-\lambda_C}
\end{cases}
\]

\[
\phi = (\pi_A, \pi_B, \pi_C, \lambda_A, \lambda_B, \lambda_C)
\]

Among them, \( \pi_L \) is the coefficient, where \( \pi_L \geq 0 \), and \( \sum_{L=1}^{\infty} \pi_L = 1 \), where \( L \) is the \( i \)-th Poisson distribution.

Objectives function.

\[
\max_{\pi_L} \log L(x_i, \pi_L, \lambda) \\
\text{s.t.} \sum_{L=1}^{\infty} \pi_L = 1 \quad \varphi = \arg \max_{\lambda} Q
\]

endcondition: \( L(\varphi^{(i)}) \geq L(\varphi^{(i+1)}) \)

In Poisson mixture distribution, estimated variables \( \pi \) and \( \lambda \) is

\[
L(x_i, \pi, \lambda) = \prod_{i=1}^{n} p(x_i, \varphi)
\]
\[
M_{\text{axlog}}^{L(x), \pi, \lambda} = \sum_{i=1}^{36} \log p(x_i, \varphi) \\
= \sum_{i=1}^{36} \log \left( \prod_{j=1}^{36} \pi_j p(x_i, \varphi) \right) \\
= \sum_{i=1}^{36} \log \left( \frac{\pi_A}{\lambda_A^k} e^{-\lambda_A} + \frac{\pi_B}{\lambda_B^k} e^{-\lambda_B} + \frac{\pi_C}{\lambda_C^k} e^{-\lambda_C} \right)
\]

we define implicit variable \( \gamma \) as a \( K \)-dimensional binary random variable with only a specific value in its \( K \)-dimensional value. When \( K \) value is 1, the other elements' value is 0

\[
p(\gamma_L = 1) = \pi_L
\]

when \( p(x_1) = \sum_{L=1}^{K} \pi_L p_i(x_1, \lambda_L) \)

\[
\gamma(i, K) = \frac{\pi_f j(x_i, \lambda_k)}{\sum_{k=1}^{K} \pi_k j(x_i, \lambda_k)}
\]

\[
\text{Likelihood function}
\]

\[
L(x_1, x_2, \ldots, x_q) = L(x_n, \varphi) = L(x_n, \pi_L, \lambda_L)
\]

\[
L(x_i) = \prod_{x_i=1}^{36} \left\{ \sum_{L=1}^{K} \pi_L p(x_i, \varphi) \right\} = \prod_{x_i=1}^{36} \left\{ \pi_A \frac{\lambda_A^k}{x_i^k} e^{-\lambda_A} + \pi_B \frac{\lambda_B^k}{x_i^k} e^{-\lambda_B} + \pi_C \frac{\lambda_C^k}{x_i^k} e^{-\lambda_C} \right\}
\]

\[
\text{Log-likelihood function:}
\]

\[
\log L(\varphi) = \log \left\{ \prod_{i=1}^{36} f_i(x_i, \varphi) \right\} = \sum_{i=1}^{36} \log \left( \sum_{j=1}^{36} \pi_j p(x_i, \varphi) \right)
\]

If bike flow patterns obey just one \( \lambda \) Poisson distribution in BSS

\[
\left\{ \begin{array}{l}
\frac{\partial \log L(\varphi)}{\partial \pi} = 0 \\
\frac{\partial \log L(\varphi)}{\partial \pi} = 0
\end{array} \right.
\]

\[
\text{Hidden variable } \gamma. \text{ In the log-likelihood function, there is a sum in the logarithm. So add hidden variable } \gamma, \text{ if } x_i \text{ from type of sample A, then } \gamma_{i,A} = 1, \gamma_{i,B} = 0, \ldots, \gamma_{i,C} = 0 \text{ means } (y_i, 1, 0, \ldots, 0)
\]

\[
L(x_i, \gamma_{i,A}, \gamma_{i,B}, \gamma_{i,C} | \pi_L, \lambda_L) = \prod_{L=1}^{36} \left( \pi_L B(x_i; \lambda_L) \right)^{\gamma_{i,L}}
\]

\[
= \left( \pi_A B(x_i; \lambda) \right)^{\gamma_{i,A}} \times \left( \pi_B B(x_i; \lambda) \right)^{\gamma_{i,B}} \times \left( \pi_C B(x_i; \lambda) \right)^{\gamma_{i,C}}
\]

\[
\gamma_{i,C} = (\pi_A B(x_i; \lambda))^1 \times (\pi_B B(x_i; \lambda))^0 \times (\pi_C B(x_i; \lambda))^0
\]
Bike flow patterns each, as we know various possibilities of optimization model suitable parameter values are obtained to construct the system and iterate maximizing Q function, and the most suitable proportion of Poisson distribution, we analyze and find out the proportion of parameter. 

Maximization. Estimate the parametric proportions of each, as we know various possibilities of $\gamma$. In order to find out the proportion of Poisson distribution, we analyze and iterate maximizing Q function, and the most suitable parameter values are obtained to construct the optimization model

$$\pi^{(i+1)}, \lambda^{(i+1)} = \arg \max Q(\pi, \lambda, \pi^{(i+1)}, \lambda^{(i+1)})$$

Derivative of Q function and equal to zero. Bike flow patterns

$$\lambda_L^{(i+1)} = \frac{E_{1L}x_1 + E_{2L}x_2 + \ldots + E_{36L}x_{36}}{\sum_{j=1}^{36} E_{ij}(y_i; y) \{ \log^{\pi_L} + \ldots + \log P_{X_{ij}} \}}$$

when

$$\begin{align*}
\lambda_A^{(i+1)} &= \frac{\sum_{i=1}^{36} (E_{AI})}{E_{AI}(y_i; k_i, \pi^{(i)}, \lambda^{(i)})} \\
\lambda_B^{(i+1)} &= \frac{\sum_{i=1}^{36} (E_{BI})}{E_{BI}(y_i, k_i, \pi^{(i)}, \lambda^{(i)})} \\
\lambda_C^{(i+1)} &= \frac{\sum_{i=1}^{36} (E_{CI})}{E_{CI}(y_i, k_i, \pi^{(i)}, \lambda^{(i)})}
\end{align*}$$

$$\pi_L^{(i+1)} = \frac{E_{1L} + E_{2L} + \ldots + E_{36L}}{36}$$

1. Mobike data: We use the data from Mobike in Beijing city, from 1 March 2017 to 30 October 2017 as the bike data. There are 4,825,118 records in train.csv and test.csv data sets. The format of data is user ID, bike ID, bike type, start time, end time, geo-coordinates start location, and geo-coordinates end location. We define them into a variable set $\{U, B, T_s, T_e, Ls, Lsi, lat, lng\}$.

2. OFO data: We use the data of OFO BSS in Hangzhou city, from 1 May 2018 to 30 June 2018 as the bike data. There are 1,462,273 records. The format of the data is bike ID, start time, end time, and geographical coordinates of starting and ending location. We define them into a variable set $\{B, T_s, T_e, Ls, Lsi, lat, lng\}$.

The experiments were executed on a computer running Windows 7, MATLAB 2014a, on a Pentium IV, Intel®, 1.84 GHz CPU, 4 GB of RAM. First, we use a statistics Spatiotemporal approach via time series analysis and get $X_{in}, X_{out}$. Second, we tagged the classification of the BSSs’ training data set. Finally, we predicted the cluster of Geohash grid.

Achievements

In this section, we experimentally evaluate the performance of AGC-MSP and PMM-EM in Geohash grid. AGC-MSP is predicted in Geohash-grid stage in next time slot.
Multi-grid scenarios. At first, we choose six districts of Beijing to divide grid of Geohash. Dongcheng district is divided into 320 Geohash grids, Xicheng district is divided into 355 Geohash grids, Fengtai district is divided into 471 Geohash grids, Haidian district is divided into 1035 Geohash grids, Chaoyang district is divided into 1344 Geohash grids, and Changping district is divided into 1833 Geohash grids. We proposed AGC-MSP in stage prediction, as shown in Figure 9.

Single-grid scenarios. The number of bikes is calculated by the function $X_{in} - X_{out}$, where positive value is demand of deployed and negative value is redundancy of recycle, as shown in Tables 5 and 6.

**Recommended solution strategies**

**$R_A$ solution.** $K$ is the capacity in $R_A$ queuing model system. When the capacity of the queues is $K = 1$, it means that one commuter searches for sharing bikes for traveling. Under over-demand stage, commuters could choose any transport tool in Geohash grid. In this system, the service rate is $\mu = 30 \text{ min/7 bikes} = 4.2 \text{ min/bike}$. The bikes arriving rate is $\lambda = 5 \text{ bikes/h}$ and the service intensity is $\rho = \lambda/\mu = 1.19$.

When the capacity of queues is $K \neq 1$, for example, users waiting for a bike, the parameters are shown in Table 7.

In the over-demand stage, queuing system service intensity is $\rho_A > 1$ in grid $R_A$, the idle probability of sharing bikes is $p_A(0) \approx 20\%$, the probability of the users loss is $p_A(k) \approx 40\% (K = 1, 2, 3...)$, the average of commuters is $L_A \approx 3$, the wait time is $W_A \approx 1 \text{ h}$, the number of bikes cannot match the demand of users based on the strong paroxysmal behavior of commuters.

**$R_B$ solution.** The over-supply state service intensity is $0 < \rho_B < 1$ in $R_B$ state. The idle probability of bikes

### Table 5. Optimization result.

| Geohash  | 1st | 5th | 11th | 16th | 19th | 23rd | 27th | 30th | 35th |
|----------|-----|-----|------|------|------|------|------|------|------|
| wx4g46j  | 9   | 2   | -5   | -10  | -13  | -17  | -8   | 20   | 13   |
| wx4g1wb  | 8   | -3  | -9   | 15   | 15   | 18   | 13   | -6   | 3    |
| wx4g1uk  | 5   | 24  | 42   | 19   | -6   | 13   | 20   | 14   | 17   |

### Table 6. PMM parameter result.

Initial value in WX4G7PT

| Stage | $R_A$ | $R_B$ | $R_C$ |
|-------|------|------|------|
| Type  | A    | B    | C    | A    | B    | C    |
| $\lambda$ | 17   | 5    | 3    | 11   | 4    | 2    |

### Table 7. Different number of commuter waiting for bike in $R_A$.

| Parameters       | $K_{commuter} = 2$ | $K_{commuter} = 3$ | $K_{commuter} = 4$ |
|------------------|--------------------|--------------------|--------------------|
| $p_A(0)$         | $1 - \rho$        | $1 - \rho$        | $1 - \rho$        |
| $p_A(0)$ _k-lost_ | $0.399$            | $0.321$            | $0.28$            |
| $L_A$            | $3.57$             | $2.55$             | $2.40$             |
| $W_A$            | $1.342$            | $0.908$            | $0.869$            |

Figure 9. Predicted imbalance stage.
is $p^{(B)}_0 = 60\%$. The commuters queuing line is $0 < L_B < 1$.

Thus, commuter should not wait. In $R_B$ state, there are several bikes offered to commuters to pick up in grid parking zone. The resources for bike sharing are in imbalance state. Thus, it is phenomenon of “bikes graveyard” by social news report.

In $R_B$ state, the utilization rate of sharing bikes is inefficient. After several iterations, Markov transition matrix becomes steady-state. Based on the arrival rate $\lambda$ according to a Poisson process, the number of check-in of bikes is $\mu_B = 20$ bikes/h by the historical data. The remainder bikes are in idle state in $R_B$ Geohash grid. According to PMM-EM, the optimized parameter is $\lambda_B = 8$ bikes/h. The remainder bikes will be recycled by truck dispatching strategy. The intensity of the queuing model system is $\rho_B = 0.4 < 1$. The $p^{(B)}_0$ value is larger than the $p^{(B)}_1$ value. The result shows that the users are not required to wait, as shown in Table 8.

**RC solution.** Table 9 shows that the bikes arrive at the parking platform following Poisson distribution.

In self-balance Geohash grid, the parameters are described as follows: the idle probability is $p^{(C)}_i = 17\%$, the waiting time for service is 0.09 h, the arrival rate $\lambda_C = 7$ bikes/h, the completed services rate $\mu_C = 5.67$, the intensity is $\rho_C = 1.23$, the number of users waiting in line is $L_C = 1.24$. When the number of counter is $S_{\text{counter}} = 3/5/7$, the parameter indicators are presented in Table 10.

The occasional idle situation of sharing bikes in $R_C$ causes the behavior of customer selection to become stable. No case of over-demand or over-supply was observed. The number of bikes can maintain stable self-balance.

**Systems solution**

$$MinP_{R \rightarrow \text{Adaptive}}(p) = \begin{cases} R_A = \text{Min}(L(p) \prod_{i=1}^{M} p^A_{k-\text{lost}}) \\ R_B = \text{Min}(\prod_{i=1}^{M} p^B_{i-\text{idle}}) \\ R_C = \frac{1}{2} \sum \text{Grid}_{\text{Geohash}}(x_m - x_{\text{out}})^2 \end{cases}$$ (51)

Table 8. Model parameters in region $R_0$.

| $p_{i0}^{(B)}$ | $p_{i1}^{(B)}$ | $p_{i2}^{(B)}$ | $L_B$ | $W_B$ |
|---------------|---------------|---------------|-------|-------|
| 0.6           | 0.038         | 0.028         | 0.67  | 5 min |

Table 9. Optimized value in WX4G7PT.

| $\pi$ | $\lambda$ |
|-------|-----------|
| 0.66  | 29        |
| 0.19  | 4         |
| 0.25  | 1         |
| 0.24  | 6         |
| 0.61  | 3         |
| 0.15  | 11        |
| 0.32  | 7         |
| 0.19  | 6         |
| 0.49  |           |

Table 10. $R_C$ parameters indicators.

| Parking platform | $S = n = 3$ | $S = n = 5$ | $S = n = 7$ |
|------------------|-------------|-------------|-------------|
| Idle $p^{(3)}_0$ | 0.1705      | 0.1701      | 0.1701      |
| Probability of $n$ $p^{(3)}_n$ | 0.1929      | 0.1620      | 0.1020      |
| Average number $L_3$ | 1.244       | 1.231       | 1.231       |

$$p^{A}_{k-\text{lost}} = \left(\frac{\lambda}{\pi}\right)^{K} p_0^{A}, p^{A}_0 = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}$$ (52)

$$p^{0}_{i-\text{idle}} = 1 - \frac{\lambda}{\mu} = 1 - \left(\frac{\text{number of user/h}}{1/\text{service hours}}\right)$$ (53)

$$\text{MinC}_{\text{num}}(x) = \theta_A N_A - \theta_B N_B + \theta_C N_C$$ (54)

Bike rebalance strategy in grid $R_A$

$$N_A = \pi^t + T(t \leq 36) \pi A x_{A} - T(t \leq 36) \pi B x_{B}$$ (55)

When

$$\{ \begin{cases} N_A > 0; \text{Put in} \\ N_A = 0; \text{Keep on} \\ N_A < 0; \text{Remove or stop} \end{cases}$$ (56)

where $N_A$ is the optimized number of bike put in $R_A$ and $N'_A$ is the initial quantity in $R_A$.

Bikes’ rebalance strategy in grid $R_B$

$$N_B = \pi^t + \pi x_{B} x_{t} - \pi A x_{A} x_{t}$$ and $t \in (t \leq 36)$ (57)

When

$$\{ \begin{cases} N_A > 0; \text{Remove} \\ N_A = 0; \text{Keep on} \\ N_A < 0; \text{Put in or stop} \end{cases}$$ (58)

where $N_B$ is the optimized recycle bikes from $R_B$. $N'_B$ is the initial quantity in $R_B$ and $\theta$ is the penalty factor.

BSS rebalance optimized object function is
\[
F(X(x_{\text{in}}, x_{\text{out}}), T(t_i), \pi_L, \lambda_L, \theta_L, w) = \begin{cases}
\min P_{R-\text{Adaptive}}(p) \sum_{L} R_L \\
\min C_{\text{mambas}}(x) = \sum_{L} \theta_L N_L \\
\min |x_{\text{in}}|^2 \quad (59)
\end{cases}
\]

Result

In the experiments, two models are designed as the classification and prediction frameworks. For the result on bike check-in/out, we compared famous baseline clustering and classify method perceptron, decision tree, k-NN, and k-means in bike-sharing service region. For the result on imbalance problem, we compared AGC-MSP with baseline method: ARMA and GBRT in predicted demand. For the result on rebalance problem, we compared PMM-EM with Kalman filter and HMM in parameter optimization.

Baselines and evaluation method. The models are proposed in our work to solve imbalance and rebalance problems by bike flow patterns prediction and parameter optimization. The methods are ISVM algorithm to solve multi-classification problem and AGC algorithm, based on AGC-MSP and PMM-EM. In order to confirm our models, there are eight approaches that can be compared with the proposed method as follows:

Perceptron: the perceptron is an algorithm for supervised learning of binary classifiers.

Decision tree: rules based on variable values are selected to get the best split to differentiate observations based on the dependent variable. Tree models where the target variable can take a discrete set of values are called classification trees.

GBRT: gradient boosting regression tree.

k-NN: k-nearest neighbor algorithm output depends on whether k-NN is used for classification or regression. In k-NN regression, the output is the property value for the object. This value is the average of the values of k-NNs.

Geographical grid: It means that the city was divided into several grids.

k-means: k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

ARMA: This method is used to predict the stage of bikes in future time series. It helps to find imbalance stage based on the result.

Kalman filter: It is an algorithm that uses a series of measurements observed over time.

HMM: In HMMs, the state is directly visible to the observer, and therefore, the state transition probabilities are the only parameters. The state is not directly visible, but the output dependent on the state is visible.

In order to evaluate the realistic scenario of BSS, the test data set is divided two parts: sequential hour slots and anomalous time slots.

In this paper, we proposed AGC, ISVM, AGC-MSP, and PMM-EM approaches compared with six baselines method. AGC is based on density-based clustering, ISVM is based on classification algorithm, AGC-MSP is based on classification and regression analysis, and PMM-EM is based on machine learn.

Evaluation metrics: To evaluate the performance of method, we adopt RMSE, RMSLE, error rate, precision, recall, and F1 which are widely used to measure accuracy of classification and regression algorithm

\[
\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (60)
\]

\[
\text{recall} = \frac{|TP|}{|TP| + |FN|}
\]

\[
F1 - \text{score} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (61)
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (x'_{\text{in}} - x_{\text{in}})^2 + (x'_{\text{out}} - x_{\text{out}})^2}{n}} \quad (62)
\]

\[
\text{RMSLE} = \sqrt{\frac{\sum_{i=1}^{n} (\log(x'_{\text{in}} + 1) - \log(x_{\text{in}} + 1))^2}{n}} \quad (63)
\]

\[
\text{Error rate} = \frac{\sum_{i=1}^{n} |X_{G_i} - X_{G_i}|}{\sum_{i=1}^{n} X_{G_i}} \quad (64)
\]

where \(X_{G_i}\) are the bikes of the check-in or check-out from Geohash-grid cluster \(G_i\) during \(\Delta t\).

In machine learning, clustering is unsupervised learning approach. It is difficult to evaluate the result of clustering. What is clusters number \(n\)? We usually chose the number of clusters by knowledge and experience. In our experiments, we cluster all parking points separately in Mobike and OFO BSSs.

Result on clustering

For simplicity and rapid modeling, we have been inspired by the most popular method of density-based spatial clustering of applications with noise (DBSCAN) and OPTICS. In the bike flow pattern clustering, we apply the idea of AGC in state label renewal. According to the number of bikes’ check-in/out, we
compare $k$-means ($k = 3$) and geographically constrained label propagation (GCLP) with the geographical grid.\textsuperscript{36}

When the demand of imbalance problem in BSS is predicted, we should be able to cluster bikes’ checkout/in regions. $k$-means is the baseline clustering algorithm by distance of object. But in dockless pattern of bike sharing, the point of parking is changing with $\Delta t$. GCLP considers geographic constraint and label propagation based on popular community detection algorithm. Table 11 reveals that the time complexity is better than GCLP. Therefore, we adopted AGC for the next step.

Result on label classification

In order to figure out the stage of bike flow pattern for each region (grid or dock station), the initial parking data will be transformed to $\{x_{iD}^{out}, x_{iD}^{in}\}$ by $\Delta t$. After preprocess, we adopt simply bi-classification method. To make it fair, we change multi-classifiers to binary classifiers. To evaluate the rebalance by parameter optimization, RMSE is evaluated as the prediction performance. In this paper, $k = 2$ is $(R_A$ and $R_C)$ or $(R_B$ and $R_C)$ binary classifiers problem and $k = 3$ is multi-classifiers problem. We used ISVM to compare with regression tree, $k$-NN, perceptron, and $k$-means. Figure 10 shows that when $k = 2$, perceptron is famous with high accuracy and when $k = 3$, our ISVM is very suitable for dealing with multi-classification problems.

Result on predicted patterns

We proposed AGC-MSP compared with ARMA, ANN, and Kalman filter predicted algorithm.

Table 11. Complex of clustering.

| Method               | Time complexity |
|----------------------|-----------------|
| AGC                  | $O(n)$          |
| Geographical grid    | $O(c)$          |
| $k$-means            | $O(n\log n)$    |
| GCLP                 | $O(n^2)$        |

AGC: adaptive Geohash-grid clustering; GCLP: geographically constrained label propagation.

Table 12. True and false positive.

| Prediction | Truth                |
|------------|----------------------|
|            | Over-demand          | Over-supply         |
| Over-demand| True positive        | False positive      |
| Over-supply| False negative       | True negative       |

In other words, the proposed AGC-MSP based on density-based clustering and Markov models can improve the prediction performance. The predicted over-demand state results are shown in Table 12.

Based on the AGC-MSP results, we predict the check-out/in stage by ARMA and GBRT. We choose two type of stages for AGC in our experiment: one is over-supply stage and the other is over-demand stage. The performances of ARMA are much better than GBRT. In addition, in all the hours, GBRT is less affected by time factor. We proposed that AGC-MSP is more accurate than ARMA and GBRT obviously, as shown in Figures 11–13. Because of sample of time window, the accuracy of AGC-MSP depends on stable $\Delta t$ prior knowledge for historical data.

Result on parameter optimization

We proposed PMM-EM’s parameter optimization compared with HMM and Kalman filter. In anomalous time series, PMM-EM performance was calculated using RMSLE and error rate.

In summary, PMM-EM based on the result of AGC, ISVM, and Markov state transition matrix is much better than baseline method. In time complexity, the PMM-EM model need to achieve stable transition by initialized values iterated as $\pi_L$ and $\lambda_L$. Like pre-classified items in clustering, these sets are often created by expert human. Fortunately, we proposed queuing model, in which the bikes’ flow obey Poisson distribution. We can choose $\lambda_L$ by historical data. Based on prior knowledge, we can find the optimal parameter in the convergence state as quickly as possible. Base on PMM-EM algorithms, the performance of RMSLE is 0.349 in OFO dataset. The result of Mobike dataset 0.371 > 0.349, Therefore, in OFO dataset, PMM-EM performs is better than Mobike dataset result, show as in Tables 13 and 14.

Figure 10. Label of static classified.
Table 13. Demand of bikes’ check-out Geohash grid.

| WX4G46j | All hours sequential data | Anomalous hours interval data |
|---------|----------------------------|-------------------------------|
|         | RMSLE                      | Error                         | RMSLE                      | Error                         |
| Company | Mobike                     | OFO                           | Mobike                     | OFO                           | Mobike                     | OFO                           |
| HMM     | 0.387                      | 0.372                         | 0.439                      | 0.451                         | 0.353                      | 0.355                         | 0.453                      | 0.489                         |
| PMM-EM  | 0.371                      | 0.349                         | 0.421                      | 0.407                         | 0.288                      | 0.282                         | 0.351                      | 0.347                         |
| Kalman filter | 0.386                      | 0.369                         | 0.423                      | 0.425                         | 0.311                      | 0.314                         | 0.371                      | 0.375                         |

RMSLE: root mean squared logarithmic error; HMM: hidden Markov model; PMM-EM: Poisson mixture model expectation-maximization.

Table 14. Demand of bikes and check-in into Geohash grid.

| WX4G46j | All hours sequential data | Anomalous hours interval data |
|---------|----------------------------|-------------------------------|
|         | RMSLE                      | Error                         | RMSLE                      | Error                         |
| Company | Mobike                     | OFO                           | Mobike                     | OFO                           | Mobike                     | OFO                           |
| HMM     | 0.624                      | 0.653                         | 0.689                      | 0.701                         | 0.681                      | 0.671                         | 0.834                      | 0.835                         |
| PMM-EM  | 0.365                      | 0.350                         | 0.408                      | 0.402                         | 0.297                      | 0.290                         | 0.353                      | 0.340                         |
| Kalman filter | 0.384                      | 0.373                         | 0.425                      | 0.419                         | 0.335                      | 0.302                         | 0.365                      | 0.359                         |

RMSLE: root mean squared logarithmic error; HMM: hidden Markov model; PMM-EM: Poisson mixture model expectation-maximization.
Conclusion and future works

Bike sharing is a means of transportation that provides services to residents through mobile Internet, LBS, e-commerce, and other technologies. With the development of the market and the increasing number of customer groups, a huge amount of data has been generated.

Different researchers have different attitudes toward BSSs. With the development of deep learning and ANN, the traffic dispatching that has become hot spot in international conference. However, those algorithms are not suitable for our framework. The main reason is perspective of feature and factor. For instance, genetic algorithm, simulated annealing, and the heuristic algorithm are all with good performance in nondeterministic polynomial time (NP)-hardness problem, but our approach focus on global features. In future research, we should address the problem of heuristic algorithm, such as particle swarm optimization optimized BSS. Besides the optimization, deep learning differential privacy is a newest research area. Therefore, commuter should be prevented from being tracked and intercepted. Large-scale data sets are shared, and the law of user behavior is examined. The overall arrangement of sharing bikes is continuously optimized. It provides a reference for traffic planning of urban public service and alleviates urban congestion. It provides a business sitting scheme which pushes media advertising and recommends commodities.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Key R&D Program of China under Grant No. 2017YFB1400800 and National Natural Science Foundation of China under Grant No. 61702046.

ORCID iD

Kui Yu https://orcid.org/0000-0003-0862-198X

References

1. Kaufman SM and Gordon-Koven L. Citi bike: the first two years. UTC J 2015; 24: 17–25.
2. Ciancia V, Latella D, Massink M, et al. Exploring spatiotemporal properties of bike sharing systems. In: IEEE international conference on self-adaptive & self-organizing systems workshops, Cambridge, MA, 21–25 September 2015. New York: IEEE.
3. Hu SR and Liu CT. An optimal location model for a bicycle sharing program with truck dispatching consideration. In: IEEE international conference on intelligent transportation systems, Qingdao, China, 8–11 October 2014.
4. Zeng M, Yu T, Wang X, et al. Improving demand prediction in bike sharing system by learning global features. In: Machine learning for large scale transportation systems (LSTS) @ KDD-16, San Francisco, CA, 13 August 2016.
5. Chen LB, Zhang D, Wang L, et al. Dynamic cluster-based over-demand prediction in bike sharing systems. In: UbiComp ’16 proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing, vol. 14, Heidelberg, 12–16 September 2016, pp. 841–852. New York: ACM.
6. Chang PC, Wu JL, Xu Y, et al. Bike sharing demand prediction using artificial immune system and artificial neural network. Soft Comput 2017; 1: 1–14.
7. Pal A and Zhang Y. Free-floating bike sharing: solving real-life large-scale static rebalancing problems. Transport Res C: Emer 2017; 80: 92–116.
8. Lin J-R and Yang T-H. Strategic design of public sharing bikes systems with service level constraints. Transport Res E: Log 2011; 47(2): 284–294.
9. Sayarshad H, Tavassoli S and Zhao F. A multi-periodic optimization formulation for bike planning and bike utilization. Appl Math Model 2012; 36(10): 4944–4951.
10. Crisostomi E, Faizrahmemonoo M, Schloite A, et al. A Markov-chain based model for a bike-sharing system. In: 2015 international conference on connected vehicles and expo (ICCVE), Shenzhen, China, 19–23 October 2015. New York: IEEE.
11. Bouveyron C, Côme E and Jacques J. The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann Appl Stat 2016; 9(4): 1726–1760.
12. Groschwitz N and Polyzos G. A time series model of long-term NSFNET backbone traffic. In: Proceedings of the IEEE international conference on communications, ICC’94, New Orleans, LA, 1–5 May 1994, pp. 1400–1404. New York: IEEE.
13. Box G, Pelham E and Jenkins G. Time series analysis, forecasting and control. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 1990.
14. Mingyang D, Lin C. Better Understanding the Characteristics and Influential Factors of Different Travel Patterns in Free-Floating Bike Sharing: Evidence from Nanjing, China[J]. Sustainability, 2018; 10(4): 1244–1258.
15. Yang Z and Hu J. Mobility modeling and prediction in bike-sharing systems. In: MobiSys ’16 proceedings of the 14th annual international conference on mobile systems, applications, and services, vol. 18, Singapore, June 2016, pp. 165–178. New York: ACM.
16. Kaltenbrunner A, Meza R, Grivolla J, et al. Urban cycles and mobility patterns: exploring and predicting trends in a bicycle-based public transport system. Pervasive Mob Comput 2010; 6(4): 455–466.
17. Froehlich J, Neumann J, Oliver N, et al. Sensing and predicting the pulse of the city through shared bicycling. In: International joint conference on artificial intelligence, Pasadena, CA, 11–17 July 2009, pp. 1420–1426. New York: ACM.
18. Yao X, Shen X, He T, et al. Demand estimation of public bike-sharing system based on temporal and spatial correlation. In: 2018 4th international conference on big data
19. Ye Xin Li, Yu Zheng. Traffic Prediction in a Bike-Sharing System. SIGSPATIAL '15 Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems 2015; 4:156–163.

20. Zhong W, Chen F, Fu K, et al. SAFEBIKE: a bike-sharing route recommender with availability prediction and safe routing. In: Conference '17, Washington, DC, July 2017.

21. Cai Q, Xue Z, Mao D, et al. Bike-sharing prediction system. In: A El Rhalibi, F Tian, Z Pan, et al. (eds) E-learning and games: Lecture notes in computer science, vol. 9654. Cham: Springer, 2016, pp. 301–317.

22. Froehlich J, Neumann J and Oliver N. Measuring the pulse of the city through shared bicycle programs. In International workshop on urban, community, and social applications of networked sensing systems (UrbanSense 08), Raleigh, NC, 4 November 2008, pp. 16–20, https://www.cs.umd.edu/~jonf/publications/Froehlich_MeasuringThePulseOfTheCityThroughSharedBicyclePrograms_UrbanSense2008.pdf.

23. Come E, Latifa O. Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Velib’ System of Paris[J]. Acm Transactions on Intelligent Systems, 2014, 5(3):39.1–39.21.

24. Long Biao Chen, Da Q Zh. Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems. UbiComp ’16 Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg 2016; 14: 841–852. Germany: ACM.

25. Ling P, Qing C. Rebalancing Dockless Bike Sharing Systems. Neural Computing & Applications, 2018, 23: 59–63.

26. Lin JR, Yang TH and Chang YC. A hub location inventory model for sharing bikes system design: formulation and solution. Comput Ind Eng 2013; 65(1): 77–86.

27. Ji S, Cherry CR, Han LD, et al. Electric bike sharing: simulation of user demand and system availability. J Clean Prod 2014; 85: 250–257.

28. Alvarez-Valdes R, Belenguer JM, Benavent E, et al. Optimizing the level of service quality of a bike-sharing system. Omega 2016; 62: 163–175.

29. Guoming T and Lukasz G. Bikeshare pool sizing for bike-and-ride multimodal transit. IEEE T Intell Transp 2018; 14: 5–21.

30. Chemla D, Meunier F and Wollfer CR. Bike sharing systems: solving the static rebalancing problem. Discrete Optim 2013; 10(2): 120–146.

31. Hernandez Pérez H and Salazar González JJ. A branch-and-cut algorithm for a traveling salesman problem with pickup and delivery. Discrete Appl Math 2004; 145: 126–139.

32. Anily S and Hassin R. The swapping problem. Networks 1992; 22: 419–433.

33. Archetti C, Speranza MG and Hertz A. A tabu search algorithm for the split delivery routing problem. Transport Sci 2006; 40: 64–73.

34. Szymczyk P and Szymczyk M. Classification of geologic structure using ground penetrating radar and Laplace transform artificial neural networks. Neurocomputing 2015; 148(19): 354–362.

35. Louppe G and Geurts P. Ensembles on random patches. In: Machine learning and knowledge discovery in databases, Bristol, 23–27 September 2012, pp. 346–361. Berlin: Springer.

36. El-Assi W and Mahmoud MS. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation 2018; 44(3): 589–613.