IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 11, 2017, accepted November 13, 2017, date of publication November 16, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2774449

A Hybrid Algorithm for Estimating Origin-Destination Flows

XIANGHUA LI[1,2,3], JÜRGEN KURTHS[2,3], CHAO GAO[1], JUNWEI ZHANG[1], ZHEN WANG[4], AND ZILI ZHANG[1,5]

[1]College of Information and Computer Science, Southwest University, Chongqing 400715, China
[2]Potsdam Institute for Climate Impact Research, 11473 Potsdam, Germany
[3]Institute of Physics, Humboldt University of Berlin, 12489 Berlin, Germany
[4]Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China
[5]School of Information Technology, Deakin University, Geelong, VIC 3220, Australia

Corresponding authors: Chao Gao (cgao@swu.edu.cn) and Zili Zhang (zhangzl@swu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61402379 and Grant 61403315, in part by the Fundamental Research Funds for the Central Universities under Grant XDJK2016A008 and Grant XDJK2016B029, in part by CQ CSTC under Grant cstc2015gjhz40002, and in part by CCF-Tencent Open Fund.

ABSTRACT With the development of intelligent transportation systems, the estimation of traffic flow in urban areas has attracted a great attention of researchers. The timely and accurate travel information of urban residents could assist users in planning their travel strategies and improve the operational efficiency of intelligent transportation systems. Currently, the origin-destination (OD) flows of urban residents are formulated as an OD matrix, which is used to denote the travel patterns of urban residents. In this paper, a simple and effective model, called NMF-AR, is proposed for predicting the OD matrices through combining the nonnegative matrix factorization (NMF) algorithm and the Autoregressive (AR) model. The basic characteristics of travel flows are first revealed based on the NMF algorithm. Then, the nonlinear time series coefficient matrix, extracted from the NMF algorithm, is estimated based on the AR model. Finally, we predict OD matrices based on the estimated coefficient matrix and the basis matrix of NMF. Extensive experiments have been implemented, in collected real data about taxi GPS information in Beijing, for comparing our proposed algorithm with some known methods, such as different kinds of K-nearest neighbor algorithms, neural network algorithms and classification algorithms. The results show that our proposed NMF-AR algorithm have a more effective capability in predicting OD matrices than other models.

INDEX TERMS Origin-destination matrix, nonnegative matrix factorization, autoregressive model, GPS, prediction.

I. INTRODUCTION

In recent years, the estimation and prediction of traffic flows has become an important issue for the traffic management and traffic control in intelligent transportation systems [1]. An important input for estimating the traffic flows and patterns of urban residents is the demand for travel, which is commonly formulated as the origin-destination (OD) matrix. An OD matrix is extracted based on the counts of travels from one area to another [2]. As an input data in transport engineering, the OD matrix has attracted lots of research in the last years. Moreover, an accurate and timely prediction of OD matrix can provide reliable travel information for residents, and can help traffic management departments to optimize the traffic signals and emergency dispatch [3]. Therefore, how to predict the OD matrix accurately and timely has become a hot topic in the field of intelligent transportation design.

Many models and methods have been proposed for estimating and predicting the OD matrix. Based on different techniques for obtaining the OD information, these models and methods can be broadly classified into two categories: the static and the dynamic OD matrix prediction. Traditional OD information is collected and extracted directly by conducting surveys, which is time-consuming and high in cost [1]. Additionally, such method ignores the temporal information of a trip and the measurements may quickly become outdated. Therefore, the studies on OD matrix prediction are subject to a static model analysis [4]. Lots of statistical models and formulations, such as information minimization,

entropy maximization, maximum likelihood, Bayesian inference and generalized least squares for networks without congestion, and bi-level programming for networks with congestion, were proposed for estimating and predicting the OD matrix [3], [5], [6]. However, these models are not suitable for estimating and predicting the OD matrix in short-term and dynamic traffic networks, but for that in long-term and stable traffic networks (e.g., trip rates over a long period of time on a network are stationary). That is because these static approaches and models require that all trips should be finished in the same time period or in the same time windows, such as one week, one month, month to date, year to date and more.

On the other hand, some models for estimating the dynamic OD matrices relax the assumptions of stationary demand and incorporate stochasticity in their models, i.e., the trip rates change dynamically in a network. Therefore, the estimated and predicted OD matrix is more suitable for real traffic networks. One of the most commonly used models for estimating the OD matric is based on the prior information (e.g., the historical OD information) and the observed traffic flows. The dynamic nature of the problem is formulated as an autoregressive process [7]. Considering the temporal characteristics and sparse features of the dynamic OD matrix, some techniques, such as the Kalman filter approach [8] or the least-square modeling approach [7], are used to predict the traffic information. The stochastic stacking of historical data is used to verify the proposed model [9], [10]. For example, Ying *et al.* have combined a polynomial trend model and the Kalman filtering theory for estimating and predicting the dynamic OD flows [11]. Based on the dynamic flows of Bluetooth and Wi-Fi, Barceló et al. have applied the optimized Kalman filtering approach to estimating the OD matrix [8]. Meanwhile, a general benchmark platform for estimating the dynamic OD flows is proposed [12]. Some typical algorithms, such as LSQR, SPSA AD-PI, SPSA CG-TR and the Kalman filter approach, are included and compared in this platform. Moreover, some enhanced SPSA methods, such as weighted simultaneous perturbation stochastic approximation (W-SPSA) [13] and cluster-wise simultaneous perturbation stochastic approximation (c-SPSA) [14], are proposed for improving the stability of SPSA and reducing the noise generated by the uncorrelated measurements in the gradient approximation. Kostic *et al.* have implemented some experiments on various kinds of traffic flows in order to discuss the advantages and disadvantages of datasets, as well as the efficiency of optimization algorithms [15]. However, with the development of urban transportation, an explosive growth of the scale of traffic flows occurs. Therefore, an accurate and timely estimation and prediction for the large-scale and high-dimension OD matrix has become the focus of recent research [13]. Although some research has applied the principal component analysis (PCA) to reducing the dimension of the OD matrix, this method cannot still guarantee the nonnegative feature of the OD matrix in the analysis process [16]. The negative values in the OD matrix disobey the initial physical meaning and further disturb the accuracy of prediction.

In order to overcome the high-dimension feature and ensure that the values of the OD matrix are nonnegative, some techniques, such as nonnegative matrix factorization, have attracted the attention of researchers. On the one hand, such techniques can reveal the characteristics of large-scale data under the condition of keeping the nonnegative feature of a matrix [17], [18]. On the other hand, some methods (e.g., regression and neural networks) for predicting the short-term traffic flows can be used to estimate the dynamic OD matrix, because both problems have the same input data extracted from GPS data, and the same type of output, i.e., the prediction of traffic situation in the future. Although such methods provide effective traffic information for drivers, it is not enough for providing travel information of residents for taxi drivers [19]. In this paper, we propose an integrated model by combining nonnegative matrix factorization and autoregressive model, called NMF-AR. Such model is designed to overcome the high dimensional problem of the OD matrix, and reveal travel characteristics of residents through maintaining the nonnegative characteristics of the OD matrix.

More specifically, for the problem of reducing dimension of the OD matrix, the NMF algorithm is more effective than PCA in maintaining the original characteristics of the OD matrix and providing physical meaning of resident trip information. To solve the problem of traffic flow prediction, a nonparametric model is used to estimate the OD matrix. Combining the revealed feature of travel flow based on NMF, our proposed NMF-AR algorithm has a better prediction capability than other nonparametric models. Therefore, the results, returned by our model, can provide more valuable pick-up hotspots information for taxi drivers and further reducing the empty loading ratio of taxi. Moreover, two kinds of strategies based on temporal and spatial information are applied to dividing the collected data about taxi GPS trajectories into different datasets. Extensive experiments are implemented in these datasets in order to analyze and estimate the performance of our proposed method.

The rest of the paper is organized as follows. Sec. II presents the related work. Sec. III gives the NMF-AR algorithm in detail. Sec. IV displays some experiments conducted to illustrate the efficiency of our proposed algorithm. Sec. V concludes the main results of this paper.

II. RELATED WORK

In the era of big data, more and more data are generated from our use of public facilities, such as the social media or scientific papers. These data are stored in the datacloud and easy accessibility. Some techniques, such as the general attribute-based cryptography framework for urban data sharing [20], can ensure data security with limited computational cost and help us developing various of data-based engineering. In the filed of urban transportation, accurate and timely traffic flow prediction has been an important issue for constructing

intelligent transportation systems (ITS) [21], [22]. Therefore, based on the current and past traffic information, how to predict upcoming traffic flow is a hot topic in the field of ITS [23]. Some basic traffic information (e.g., volume, density and speed) should be considered and simulated in the prediction model in order to meet the requirement of real traffic conditions. Generally, these models are based on parametric, nonparametric and hybrid integration techniques [24].

Parametric techniques include time-series model (e.g., moving average), autoregressive moving average model and Kalman filter [25]. Although time-series models have an effective prediction capability under the condition of stable traffic flow, these models do not address the dynamic and nonlinear features of traffic flows [26]. The prediction results, returned by the Kalman filter, would be postponed compared with real situations [27]. Therefore, more and more nonparametric techniques have been developed to improve the forecast accuracy of traffic flows, such as neural network, k-nearest neighbor (KNN), and support vector regression [28]–[30]. As one of the typical nonparametric techniques, neural networks have been used for estimating traffic flows and improving the prediction capability. For example, Xia et al. have proposed an extended KNN model based on spatial-temporal features of traffic flows and estimated its performance [24], [31]. Abadi *et al.* have applied an autoregressive model to predicting the traffic flows based on real-time and estimated traffic data of a traffic network in San Francisco [32]. However, getting stuck in the local minimization will affect the performance of such models [33]. In order to overcome such shortcomings, hybrid integration techniques aim to combine nonparametric techniques with the advantages of parametric techniques, such as the statistics and neural network [34]. For example, a hybrid method combining a genetic algorithm with cross-entropy was proposed for estimating the short-term traffic congestion [35]. Autoregressive integrated moving average and genetic programming are combined for traffic flow prediction [36]. A deep-learning based approach is further applied to predicting the in-flow and out-flow of crowds in the urban area based on historical trajectory data, weather and events [38]. Although KNN-based methods are robust to noise and maintaining the randomness of traffic data, the temporal-spatial feature and inter-relationships of traffic flows are not considered in these models. Meanwhile, the traffic flow information in roads is not enough for taxi drivers to obtain the traffic flow of passengers. Therefore, the accurate travel willingness information or the travel trajectory of the urban residents should be considered in the prediction model.

The origin-destination (OD) matrix is widely used to record the travel trajectory of the urban residents, which helps taxi drivers to obtain traffic flow of passengers. However, it is difficult to deal with the high-dimension feature of an OD matrix. Currently, the principal component analysis (PCA) was used to overcome such features of an OD matrix [16]. Such technique reduces the dimension of the OD matrix through performing a linear mapping of the data to a lower dimensional space. However, the negative values, returned by PCA, cannot reflect the physical meaning of the real situation, and they disturb the explanation of the prediction results [17]. Therefore, some studies have proposed some methods based on the nonnegative matrix factorization (NMF) algorithm. In detail, the NMF algorithm can reveal the integrated feature of data by ensuring the nonnegative values in the analysis process, which has been widely used in the field of text mining, spectral data analysis and classification [17], [18]. For example, Shahnaz *et al.* have applied NMF to analyze textual data and reveal semantic features or topics through reducing the base vector of semantic features in the PCA method [39]. Additionally, NMF techniques have been successfully used to solve image classification [40], and reveal features of music and audio data [37]. Although some studies try to apply the NMF algorithm to decomposing the regional OD matrix in order to identify the function of a urban region, the travel willingness of residents are not addressed in their models [19].

To sum up, the traditional OD matrix estimation can provide traffic information for residents, but it cannot satisfy the increasing demand of traffic managers to carry more and more residents. Based on the physical meaning recorded and reflected by the OD matrix, i.e., the human mobility in urban areas and nonnegative characteristic, a hybrid algorithm is proposed in this paper. Combining the nonnegative matrix factorization with an autoregressive model, the proposed algorithm estimates the OD matrix in different temporal and spatial scales, and it reveals the dynamic changes of pick-up hotspots, which is crucial for reducing the empty loading ratio of taxi.

III. NMF-AR ALGORITHM

The framework of our proposed NMF-AR algorithm is shown in Fig. 1. It consists of three steps: (i) Data collection and cleaning are implemented in Sec. III-A. The OD matrix of each period in one day is constructed based on the GPS data of taxi obtained from the intelligent traffic management system. As the input data of NMF-AR algorithm, the OD matrix is used to reveal the travel patterns of residents in the urban area. (ii) The nonnegative matrix factorization (NMF) is applied to extracting the fundamental matrix and the temporal feature of the coefficient matrix in Sec. III-B. (iii) An autoregressive model is implemented in Sec. III-C for estimating the coefficient matrix in the $(t+1)^{th}$ period. Based on the estimated coefficient matrix and the fundamental matrix, the OD matrix in the $(t+1)^{th}$ period is generated by the decomposition-reduction method of NMF algorithm.

A. DATA PREPROCESSING

GPS data, collected from intelligent traffic systems, contain lots of track information and other noise messages. We just use some information (i.e., the *id* of a taxi, longitude and latitude, the current time of GPS, the status of a taxi) to construct an OD matrix of each period in one day. More specifically, there are three steps for us to extract an OD matrix from GPS.

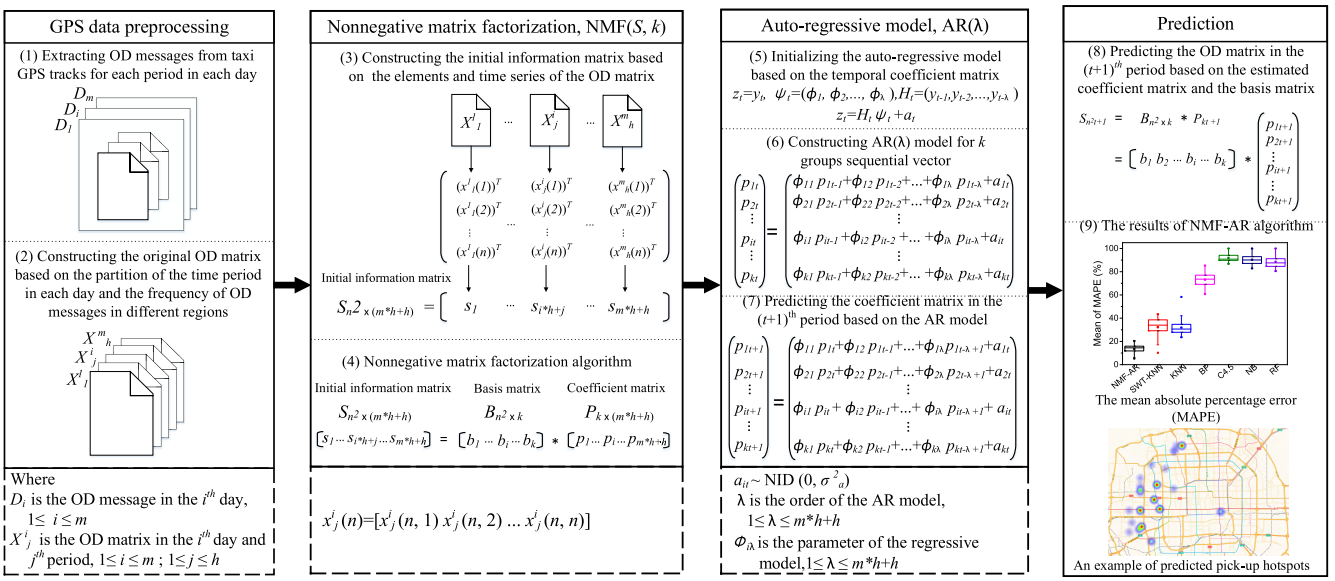


FIGURE 1. The flow chart of NMF-AR algorithm. The whole process is divided into GPS data preprocessing, nonnegative matrix factorization (i.e., NMF (S, k)), auto-regressive model (i.e., AR (λ)) and the prediction of OD matrix. As the input of NMF-AR, an OD matrix is extracted from the GPS data. The nonnegative matrix factorization (NMF) is used to reveal the pattern of OD matrix, and to extract the fundamental matrix and coefficient matrix. Based on the output of NMF, an autoregressive model (AR) is implemented to estimate the coefficient matrix in the $(t+1)^{th}$ period and predict the OD matrix in the $(t+1)^{th}$ period.

First, two kinds of division strategies are defined from the viewpoint of temporal and spatial features of residents trips. The two strategies, presented in Sec. IV-A, are used for screening the original GPS data. Then, the basic getting-on and getting-off hotspots information of travel flows are extracted and formulated. Eventually, different kinds of OD matrices are constructed with the temporal and spatial labels. To represent our method more formally, some terms are defined as follows.

Definition 1: The origin-destination matrix (OD matrix) is formulated as $X^i_j = [x_{i'j'}]_{n\times n}$, where X^i_j represents an OD matrix of the j^{th} period in the i^{th} day. x_{ab} is the total count of trips from a zone a to b ($a, b \subseteq [1, n]$).

Definition 2: The OD matrix prediction is represented formally as $X_{H+1} = F(Y^H) + W^H$, which predicts the count of residents trips in the future time based on the current or historical OD matrix. X_{H+1} denotes the predicted OD matrix. Y^H is a set of historical matrices and F denotes the selected prediction function. In addition, W^H is a corresponding set of parameters for the historical matrix.

Currently, the autoregressive model is widely used as temporal prediction model, which will also be used to predict and estimate the OD matrix in this paper. The prediction results of the OD matrix provide accurate information about the travel flows of urban residents, which can help passengers reducing the waiting time for a taxi, and help drivers reducing the empty loading ratio of taxis.

Definition 3: The measures of effectiveness (*MOEs*) for a prediction result include four metrics, i.e., mean absolute error (*MAE*), mean absolute percentage error (*MAPE*), root mean square error (*RMSE*) and maximum error (*ME*), as given in Eqs. (1) - (4).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|O_t - F_t|}{O_t} \times 100\% \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(O_t - F_t)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|O_t - F_t| \tag{3}$$

$$ME = \max_{t=1,\dots,n}|O_t - F_t| \tag{4}$$

where F_t and O_t denote the prediction and the actual OD matrices in the t^h time period, respectively. n is the number of elements in the OD matrix.

MOEs provide an in-depth and comprehensive understanding of the nature of the forecast errors [37]. More specifically, *MAE* and *RMSE* denote the statistical differences of the forecasting results. High values of *RMSE* and *MAE* indicate that there exist major changes in the prediction errors [41]. Based on existing studies, *MAPE* is an important indicator to estimate the prediction accuracy of a model [24]. The lower *MAPE* is, the higher prediction accuracy a model has. Generally speaking, the prediction capability of a model can be identified based on the range of *MAPE*. $MAPE \le 10\%$ and $11\% < MAPE \le 20\%$ suggest that a model has a high and good prediction capability, respectively. $20\% < MAPE \le 50\%$ means that a model shows a reasonable prediction capability. But $MAPE \ge 50\%$ presents an inaccurate prediction capability of a model [21].

B. NMF(S,K) ALGORITHM

The patterns of traffic flows vary with time. In particular, some specific features of traffic flows emerge during the rush hours on weekday. In order to reveal the characteristics of traffic flows, the NMF algorithm is used to analyze and decompose the OD matrix. The patterns of residents trips can be described by the fundamental matrix and coefficient matrix of the NMF algorithm. In detail, the principle of the matrix deformation is applied to constructing the initial information matrix S as shown in Eq. (5), which reflects the real-time residents trips.

$$S_{n^2\times(m*h+h)} = (s_1 \cdots s_{i*h+j} \cdots s_{m*h+h}) \tag{5}$$

The rule of matrix deformation is further formally represented in Eq. (6).

$$s_{i*h+j} = \begin{pmatrix} (x_j^i(1))^T \\ (x_j^i(2))^T \\ \vdots \\ (x_j^i(n))^T \end{pmatrix} \tag{6}$$

To represent our method formally, a basis matrix $B = [b_{ij}]_{n^2\times k}$ $(i \subseteq [1, n^2], j \subseteq [1, k])$ is defined for depicting the trip patterns of the urban residents. A coefficient matrix $P = [p_{ij}]_{k\times(m*h+h)}$ $(i \subseteq [1, k], j \subseteq [1, m*h+h])$ is used to denote the weights of the basis matrix during various periods. Therefore, the initial information matrix S can be written as Eq. (7).

$$S = BP \tag{7}$$

The two matrices B and P are unknown in Eq. (7). Although there are many matrix decomposition methods that can be used to decompose S, all elements in B and P should be nonnegative because of the restriction of the physical meaning of B and P. Hence, the NMF(S, k) algorithm is applied to decomposing S and detecting the feature of residents trips. Under the condition of known S and a positive integer $k < \min\{m, n\}$, the matrix decomposition problem can be formulated as minimization problem of the nonnegative factorization [17], as formulated in Eq. (8).

$$f(B, P) = \underset{\{B,P\}}{arg\,\min} \|S - BP\|^2 \tag{8}$$

More specifically, the detailed steps are as follows. First, $B_{n\times j}$ and $P_{j\times m}$ are initialized in order to keep the nonnegative values during the iterative process. Then, based on the cost function, as shown in Eq. (9), the two matrices are updated based on Eqs. (10) and (11), respectively. The whole iteration process will stop until Eq. (8) obtains a minimum value.

$$\|S - BP\|^2 = \sum_{i,j} (S_{ij} - (BP)_{ij})^2 \tag{9}$$

$$B_{n\times j} \sim B_{n\times j} \frac{(SP^T)_{n\times j}}{(BPP^T)_{n\times j}} \tag{10}$$

$$P_{j\times m} \sim P_{j\times m} \frac{(B^T S)_{j\times m}}{(B^T BP)_{j\times m}} \tag{11}$$

Based on the NMF algorithm, we obtain a basis matrix B and a corresponding coefficient matrix P. In the next section, a time series model is applied to analyzing and predicting the feature of P.

C. AR (λ) MODEL

In this section, we aim to reveal the temporal feature of the coefficient matrix P based on an autoregressive (AR) model, in order to predict the residents trips. More specifically, based on the AR model Eq. (12) and $k \geq 1$ in the coefficient matrix P, a series of AR models for each dimension in the coefficient matrix P are implemented based on Eq. (13). Moreover, the vector of the coefficient matrix in the $(t+1)^{th}$ period is estimated based on the prediction function Eq. (14).

$$z_t = \psi_t * H_t + a_t \tag{12}$$

where z_t is the observed value in the t^{th} period. H_t denotes the λ order independent variable in the t^{th} period, i.e., the observed variable from the $t-1$ to $t-\lambda$. ψ_t is the coefficient of the observed value in the t^{th} period. a_t is an independent and identically distributed (iid) constant noise following a normal distribution.

$$\begin{pmatrix} p_{1t} \\ p_{2t} \\ \vdots \\ p_{it} \\ \vdots \\ p_{kt} \end{pmatrix} = \begin{pmatrix} \varphi_{11}p_{1t-1} + \cdots + \varphi_{1\lambda}p_{1t-\lambda} + a_{1t} \\ \varphi_{21}p_{2t-1} + \cdots + \varphi_{2\lambda}p_{2t-\lambda} + a_{2t} \\ \vdots \\ \varphi_{i1}p_{it-1} + \cdots + \varphi_{i\lambda}p_{it-\lambda} + a_{it} \\ \vdots \\ \varphi_{k1}p_{kt-1} + \cdots + \varphi_{k\lambda}p_{kt-\lambda} + a_{kt} \end{pmatrix} \tag{13}$$

$$\begin{pmatrix} p_{1t+1} \\ p_{2t+1} \\ \vdots \\ p_{it+1} \\ \vdots \\ p_{kt+1} \end{pmatrix} = \begin{pmatrix} \varphi_{11}p_{1t} + \cdots + \varphi_{1\lambda}p_{1t-\lambda+1} + a_{1t+1} \\ \varphi_{21}p_{2t} + \cdots + \varphi_{2\lambda}p_{2t-\lambda+1} + a_{2t+1} \\ \vdots \\ \varphi_{i1}p_{it} + \cdots + \varphi_{i\lambda}p_{it-\lambda+1} + a_{it+1} \\ \vdots \\ \varphi_{k1}p_{kt} + \cdots + \varphi_{k\lambda}p_{kt-\lambda+1} + a_{kt+1} \end{pmatrix} \tag{14}$$

Based on the coefficient matrix P_{t+1} in the $(t+1)^{th}$ period and the basis matrix B returned by AR model and NMF algorithm, respectively, a prediction function is proposed to estimate the OD matrix $S_{n^2\times(t+1)}$ in the $(t+1)^{th}$ period as follows:

$$S_{n^2\times(t+1)} = BP_{t+1} \tag{15}$$

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL PREPARATION

In this section, some preparations for experiments are implemented for validating the accuracy of our proposed model. First, two kinds of division strategies are introduced based on the temporal and spatial features of traffic flows. Based on these strategies, six datasets are extracted from taxi GPS traces, and shown in Tab. 1 and Tab. 2. Then, some parameters and their values used in models are listed in Tab. 3.

TABLE 1. Time period division strategies.

Dataset	Division conditions	Scales of period
D_1	announcement	6 periods/day
D_2	1 hour	24 periods/day
D_3	10 minutes	144 periods/day

TABLE 2. Filter conditions of urban OD counts.

Dataset	Filter conditions	Num. of hotspots
D_4	100	40
D_5	70	155
D_6	30	490

1) DATASETS AND DIVISION STRATEGIES

In order to verify the performance of the NMF-AR algorithm, a public dataset with 1.3 billion taxi GPS traces is downloaded from DATATANG.[1] These GPS traces, belonging to 12,000 taxis in Beijing, are collected from November 1, 2012 to November 30, 2012. Most of them are sampled at a frequency of about 1 minute. Each trace is stored as ASCII text with a comma separator, such as id, trigger, status (i.e, occupied or idle) , GPS time, longitude, latitude, GPS speed, GPS direction, and GPS state. An example trace is 123456, 0, 0, 20110414160613, 116.4078674, 40.2220650, 21, 274, 1.

Based on the taxi GPS data within the fifth ring road of Beijing, a chessboard is used to denote the map of the urban area. The length of each square is 200 meters [19] and the whole map is divided into 129*129 squares. Based on getting-on and getting-off information recorded by the status of a taxi, an original OD matrix is constructed, which is used to analyze the prediction accuracy of NMF-AR in Sec. IV-B. Moreover, two kinds of strategies are implemented to divide the dataset into some subgroups in order to analyze the robustness of NMF-AR in different temporal and spatial scales in Sec. IV-C.

The first kind of division strategy aims to reflect the time-varying feature of traffic flows. Such strategy contains three divisions and obtains three datasets as shown in Tab. 1: (i) One day is divided into six periods based on the rush hours in the announcement of the Beijing Traffic Management Bureau,[2] i.e., 00:00-07:00, 07:00-09:00, 09:00-13:00, 13:00-17:00, 17:00-20:00, 20:00-24:00. (ii) One day is divided into 24 time slices. Each hour stands for a time slice. (iii) One day is uniformly divided into 144 time slices based on the period division of the Di-Tech Challenge.[3] Each time slice stands for 10 minutes.

The second kind of strategy aims to reflect the spatial feature of traffic flows. The OD matrices with different dimensions are constructed based on the statistics of hotspots within the fifth ring road of Beijing. The filter condition, i.e., the cumulative quantity of trips in one urban area, is shown in Tab. 2. More specifically, we just keep those regions (i.e, hotspots) whose number of trips (i.e, getting-on or getting-off points) is larger than the values of the filter condition.

[1] http://www.datatang.com/data/44502
[2] http://www.bjjtgl.gov.cn/zhuanti/20140328wr.html
[3] https://www.saikr.com/32943

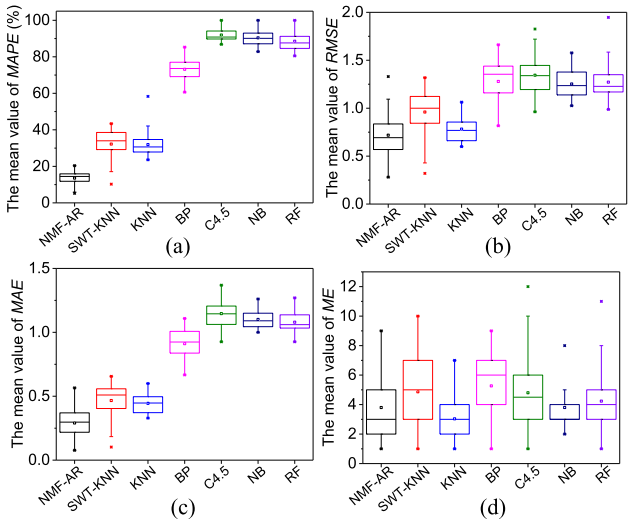


FIGURE 2. Comparisons of the prediction capability of different algorithms. Four metrics of *MOEs*, i.e., *MAPE*, *RMSE*, *MAE* and *ME* are plotted in (a), (b), (c) and (d), respectively. The box charts are the averaged results returned by the NMF-AR and other algorithms for all prediction time periods of five days in the last week of November 2012. More specifically, the bottom and top of the box denote the first and third quartiles respectively. The band and small square inside the box represent the median and the mean of the *MOEs*. From these statistics, we can conclude that the accuracy of our proposed algorithm is higher than that of other models.

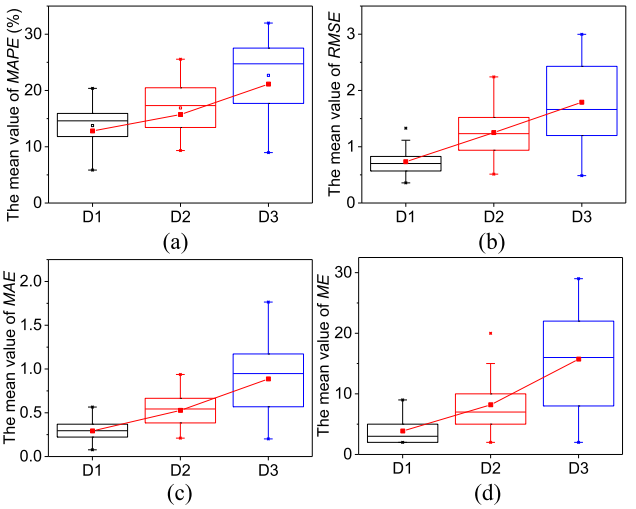


FIGURE 3. Comparisons of prediction capability of different algorithms in three temporal-based datases. Four metrics of *MOEs*, i.e., *MAPE*, *RMSE*, *MAE* and *ME* are shown in (a), (b), (c) and (d), respectively. From these results, we find that the computational efficiency of NMF-AR algorithm is sensitive to the the granularity of time division. The smaller the granularity of time division is, the lower prediction accuracy the NMF-AR algorithm has. More specifically, the NMF-AR algorithm can achieve a better prediction capability during the rush hours on weekday through comparing the *MAPE* values in three datasets.

2) PARAMETERS SETTINGS

For verifying the prediction capability of our proposed NMF-AR, some parameters are listed in Tab. 3. The values

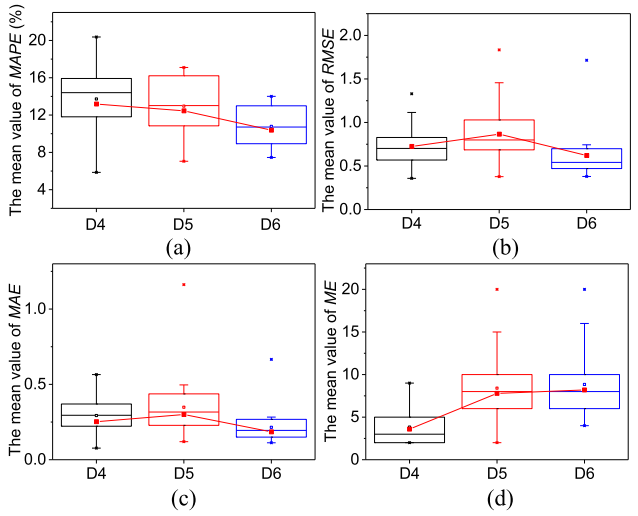


FIGURE 4. Comparisons of prediction capability of different algorithms in three spatial-based datases. Four metrics of *MOEs*, i.e., *MAPE*, *RMSE*, *MAE* and *ME* are shown in (a), (b), (c) and (d), respectively. From these results, we can conclude that the computational efficiency of NMF-AR algorithm is not sensitive to the scales of the urban areas. More importantly, the NMF-AR algorithm achieves a better prediction capability with the increase of urban areas.

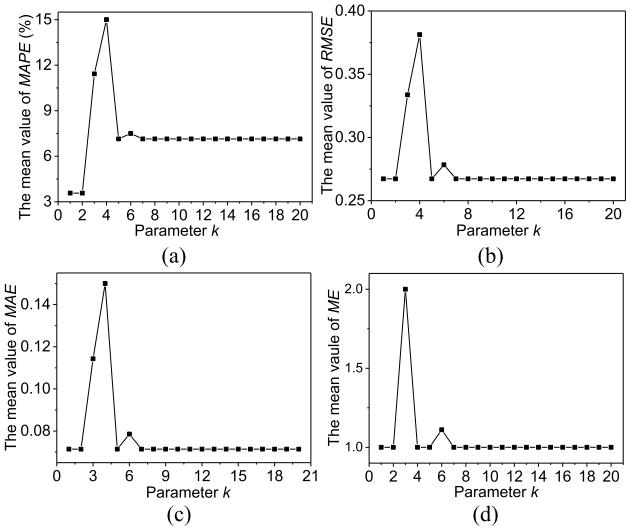


FIGURE 5. The sensitivity analysis of parameter k for NMF-AR in D_1. The dynamic changes of four metrics of *MOEs* with the k eigenvectors extracted from the initial travel information matrix are plotted in (a) *MAPE*, (b) *RMSE*, (c) *MAE* and (d) *ME*, respectively. From these analyses, we find that the prediction errors are gradually stable as k is larger than 6. The same phenomenon can be observed in other datasets. Such results show that NMF-AR algorithm has a high exploration capability.

of parameters used in our model is based on the parameter analysis in Sec IV-D. The values of the parameters, used in other models, are based on the previous work in [35].

B. PREDICTION ACCURACY

In order to evaluate the prediction capability of our proposed NMF-AR algorithm, some experiments are implemented to compare NMF-AR with six other prediction models, i.e., SWT-KNN, KNN, BP, NB, RF and C4.5 [24].

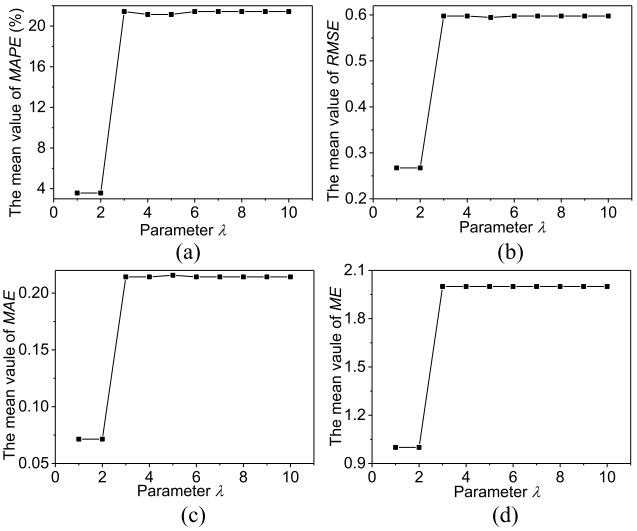


FIGURE 6. The sensitivity analysis of parameter λ for NMF-AR in D_1. The dynamic changes of four metrics of *MOEs* with the λ are plotted in (a) *MAPE*, (b) *RMSE*, (c) *MAE* and (d) *ME*, respectively. The average *MAPE* is less than 4% if $\lambda < 3$, which means that the NMF-AR algorithm achieves a high prediction capability. With the increase of λ, the *MAPE* tend to be stable around 21.23%. Results show that NMF-AR algorithm has an effective prediction capability and stability in most cases.

TABLE 3. Parameter setting and defining of all models.

Models	Parm.	Val.	Meaning
NMF-AR	k	6	The amount of basic patterns
	λ	2	The order of AR
SWT-KNN	k	21	The amount of neighbors
	a	0.4	The space weight of traffic flow
	b	0.4	The time weight of traffic flow
	r	0.9	The correlation coefficient of target and neighbors
KNN	k	15	The amount of neighbors
BP	l	2	The amount of hidden layer
	a	10	The amount of neurons in the first layer
	b	10	The amount of neurons in the second layer
RF	k	50	The amount of random tree
C4.5	a	5	The threshold of recursion

The comparison results are plotted in Fig. 2. Specifically, the box charts of the prediction results show the first and third quartiles, median and mean value of *MOEs*, which provide a comprehensive comparison of NMF-AR with other prediction models.

Due to the randomness of maximum and minimum values, the quartiles and means in the box charts are the core indicators for comparing different algorithms. Fig. 2 shows that all metrics of the NMF-AR are better than those of the other models. Moreover, the average of *MAPE* ranges from 5% to 20%, and the lengths of boxes (i.e., the maximum and minimum values) of the NMF-AR are shorter than those of the other models, which means that the NMF-AR has a stronger prediction capability and verifies that the robustness of the NMF-AR is better than that of the other models.

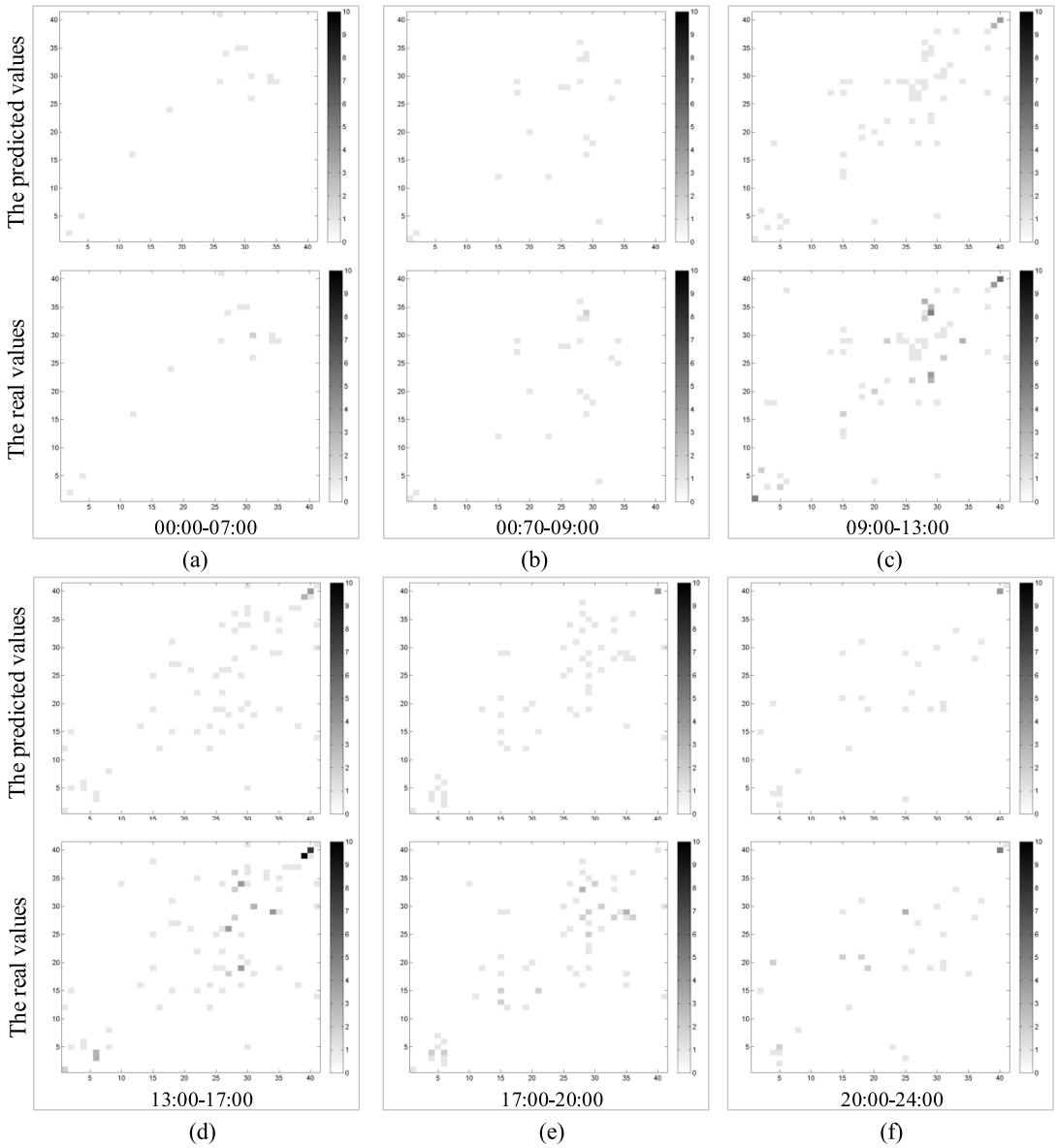


FIGURE 7. Illustration of the hotspots distribution recorded by OD matrix. The first and second rows show the prediction values and the real values during 6 time periods on 26 November, 2012, respectively. Results show that NMF-AR algorithm has a higher accuracy for predicting hotspots distribution.

C. PREDICTION SCALABILITY

In order to estimate the scalability of different algorithms, the whole dataset is divided into some subgroups through adjusting the temporal and spatial scales of the urban area. Based on two kinds of division strategies in Sec. IV-A, 6 subsets are obtained. The comparison results of *MOEs* are plotted in Fig. 3 and Fig. 4, respectively. The box charts of results returned by NMF-AR for all prediction time periods of five days in the last week of November 2012, in which the bottom and top of box are the first and third quartiles, respectively. The band inside the box denotes the median value of *MOEs*, and the small square inside the box stands for the mean value of *MOEs*. The ends of whiskers represent the minimum and maximum values of *MOEs*, respectively.

The comparison results in Fig. 4 show that *MAPE* goes down with the increase of hotspots, i.e., the improvement of prediction accuracy. Therefore, we can conclude that the prediction capability of NMF-AR is not affected by the increasing spatial scales. On the other hand, we further analyze the relationship between the prediction capability of NMF-AR and the granularity of time. As plotted in Fig. 3, the error indices returned by the NMF-AR algorithm increases with the decrease of the granularity of time, i.e., the prediction capability of NMF-AR is sensitive to the division of temporal scales. Especially, the NMF-AR algorithm can achieve

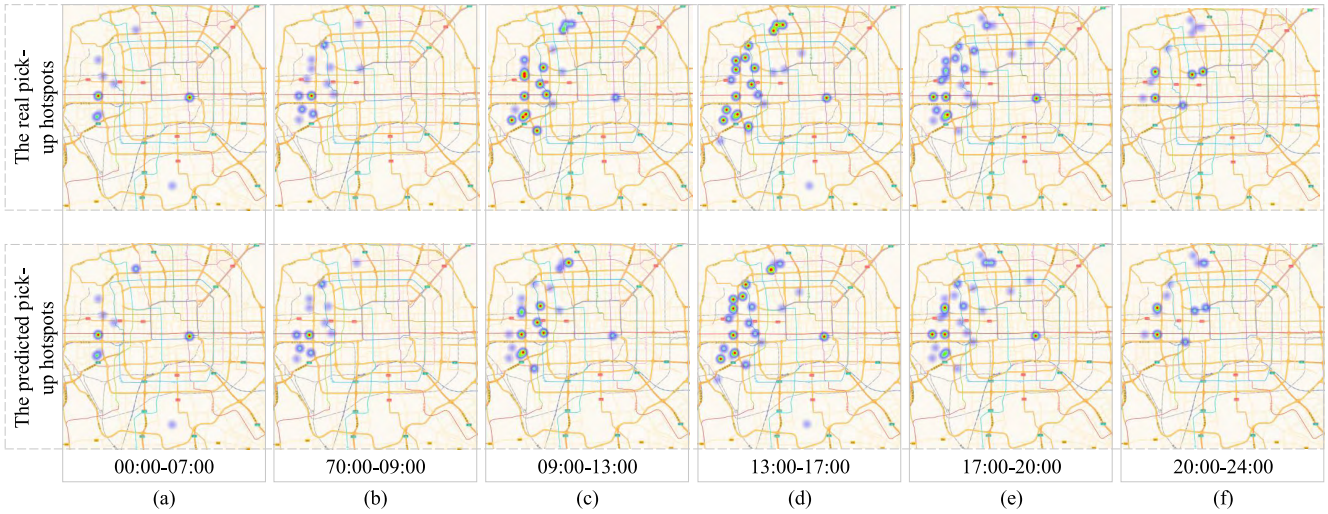


FIGURE 8. Illustration of the pick-up hotspots in the real map. The first and second rows show the real values and the prediction values during 6 time periods on 26 November, 2012, respectively. The colors varying from lavender to red denote the density level of pick-up locations ranging from low to high. Results show that the prediction errors of our proposed model are lower, which can provide an effective and timely travel information of residents for taxi drivers.

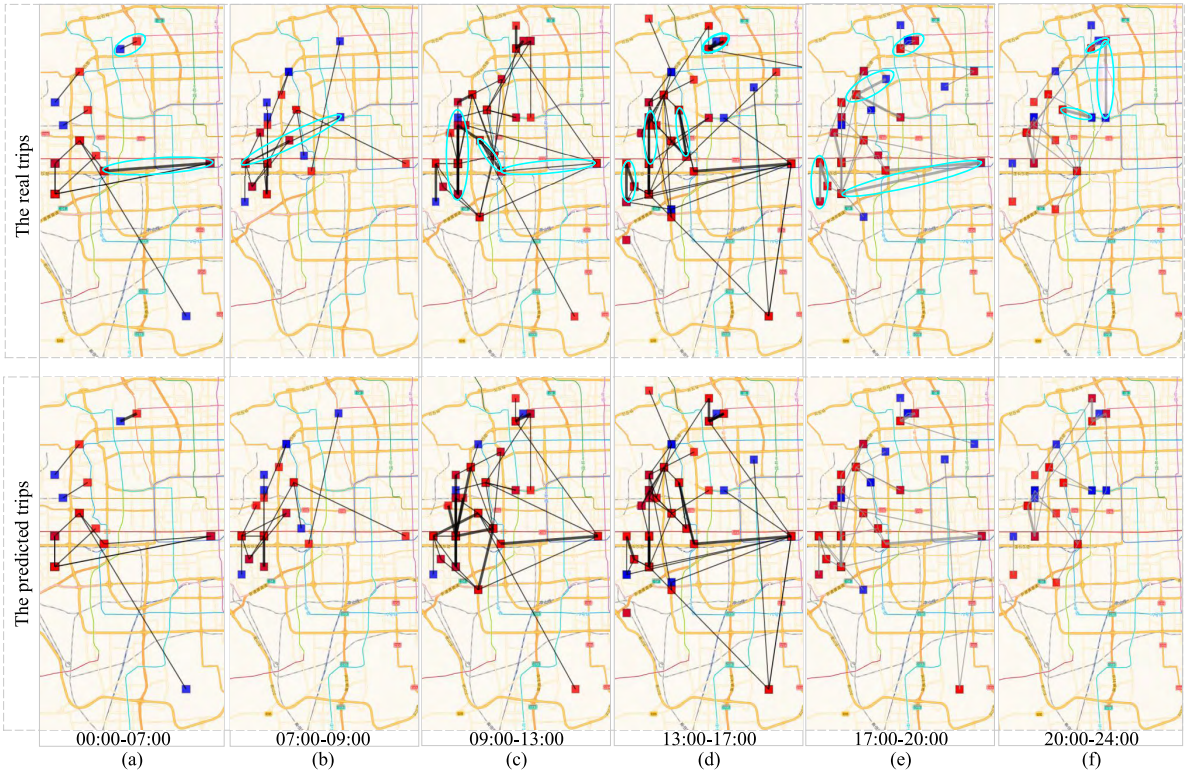


FIGURE 9. Comparison of the real trips and the prediction trips based on NMF-AR algorithm in the 6 time periods on 26, November 2012. The red and blue squares denote the destination locations and original locations, respectively. The thickness of the black lines represents the number of trips. More specifically, some differences between the real and predicted values are highlighted by circles with the nattier blue. Results show that NMF-AR algorithm has an effective prediction capability, which is very useful for providing timely recommendation information for taxi drivers.

a better prediction capability in D_1, in which the division strategy is based on the definition of rush hours released by the government.

D. PARAMETERS ANALYSIS

Some important parameters, i.e., the k in the NMF algorithm and λ in the AR model, are analyzed in this section. Fig. 5

shows that the prediction errors are gradually stable with the increasing k, i.e., our proposed model is not sensitive to k. The fluctuation of NMF-AR at the initial stage is caused by the randomness of k basic patterns revealed from the initial information matrix S. With the increase of k, the prediction result tends to be stable.

The same phenomenon occurres in Fig. 6. Because of the low-order of the AR model at the initial stage, the useful information is missing, which further causes a certain fluctuation on the condition that $\lambda < 5$. With the increase of λ, the metrics of *MOEs* are gradually stable. Therefore, we can conclude that the NMF-AR algorithm has a high exploration capability and stability.

E. A CASE STUDY

In this section, we use the NMF-AR algorithm to estimate the OD matrix of the 6 periods on November 26, 2012. The hotspot distribution between the predicted value and actual value is shown in Fig. 7. The grayscale of color ranges from shallow to deep, which denotes that the number of trips ranges from small to large. The white color represents that a region does not include any flow. From this result, we find that our proposed model has a high prediction capability if there are enough traffic flows as shown in Figs. 7(c)-(f).

Additionally, illustrations of the pick-up hotspots and trips are shown in Figs. 8 and 9, respectively. The lavender in Fig. 8 means that the density level of the pick-up locations is low. On the contrary, the red point denotes a high density level of pick-up locations. Fig. 9 further visualizes the difference between the real resident trips and the prediction results. We clearly find show that our proposed model can achieve a good prediction result and recommend suitable pick-up locations to drivers for reducing the empty loading ratio of taxis.

V. CONCLUSION

Based on the traffic information in roads, it is an important problem on how to provide a timely and accurate prediction about traffic conditions to facilitate the traffic control and overcome traffic jams during rush hours. Based on the analysis of OD matrix estimation problem, a hybrid algorithm, called NMF-AR, is proposed by combining the nonnegative matrix factorization (NMF) algorithm and the Autoregressive (AR) model. Based on real taxi GPS data in Beijing, some experiments are implemented for estimating the performance of NMF-AR. Comparing our proposed algorithm with other prediction models, such as SWT-KNN, KNN, BP, NB, RF and C4.5, we find that our algorithm has a better capability and scalability. Additionally, our proposed algorithm has a high exploration capability and stability based on the parameter analysis. Moreover, some visualization results show that our algorithm can provide timely and efficient information about pick-up locations, which are very useful for both residents and taxis drivers. In the future, more external factors, such as weather conditions, will be taken into consideration for improving the prediction accuracy of OD flow.

ACKNOWLEDGMENTS

The authors would like to thank all editors and the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] T. Tomer and T. Kolechkina, "Estimation of dynamic origin–destination matrices using linear assignment matrix approximations," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 618–626, Jun. 2013.

[2] Y. Lou and Y. Yin, "A decomposition scheme for estimating dynamic origin–destination flows on actuation-controlled signalized arterials," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 5, pp. 643–655, 2010.

[3] C. Yan, X. Ye, and Z. Wang, "A forecasting model of the proportion of peak-period boardings for urban mass transit system: A case study of Osaka prefecture," in *Proc. Transp. Res. Board 95th Annu. Meeting*, 2016, pp. 1–14.

[4] M. Tanaka, T. Kimata, and T. Arai, "Estimation of passenger origin-destination matrices and efficiency evaluation of public transportation," in *Proc. 5th IIAI Int. Congr. Adv. Appl. Informat.*, 2016, pp. 1146–1150.

[5] F. Viti, M. Rinaldi, F. Corman, and C. M. J. Tampère, "Assessing partial observability in network sensor location problems," *Transp. Res. B, Methodol.*, vol. 70, pp. 65–89, Dec. 2014.

[6] F. Simonelli, V. Marzano, A. Papola, and I. Vitiello, "A network sensor location procedure accounting for O–d matrix estimate variability," *Transp. Res. B, Methodol.*, vol. 46, no. 1, pp. 1624–1638, 2012.

[7] M. Bierlaire and F. Crittin, "An efficient algorithm for real-time estimation and prediction of dynamic OD tables," *Oper. Res.*, vol. 52, no. 1, pp. 116–127, 2004.

[8] J. Barceló, L. Montero, L. Marqués, and C. Carmona "A Kalman-filter approach for dynamic OD estimation in corridors based on bluetooth and Wi-Fi data collection," in *Proc. 12th World Conf. Transp. Res. (WCTR)*, 2010, pp. 1–29.

[9] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. B, Methodol.*, vol. 18, no. 1, pp. 1–11, 1984.

[10] K. Ashok and M. E. Ben-Akiva, "Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows," *Transp. Sci.*, vol. 36, no. 2, pp. 184–198, 2002.

[11] L. Ying, J. Zhu, W. Huiyan, and L. Zhenyu, "A novel method for estimation of dynamic OD flow," *Proc. Eng.*, vol. 137, pp. 94–102, Dec. 2016.

[12] A. Tympakianaki, H. N. Koutsopoulos, and E. Jenelius, "c-SPSA: Cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin–destination matrix estimation," *Transp. Res. C, Emerg. Technol.*, vol. 55, pp. 231–245, Jan. 2015.

[13] L. Lu, X. Yan, A. Constantinos, and M. Ben-Akiva, "An enhanced SPSA algorithm for the calibration of dynamic traffic assignment models," *Transp. Res. C, Emerg. Technol.*, vol. 51, pp. 149–166, Feb. 2015.

[14] C. Antoniou *et al.*, "Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 79–98, May 2016.

[15] B. Kostic and G. Gentile, "Using traffic data of various types in the estimation of dynamic O-D matrices," in *Proc. Int. Conf. Models Technol. Intell. Transp. Syst.*, Jun. 2015, pp. 66–73.

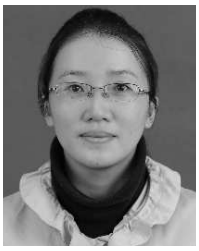
[16] T. Djukic, G. Flötteröd, H. V. Lint, and S. Hoogendoorn, "Efficient real time OD matrix estimation based on principal component analysis," in *Proc. 15th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 115–121.

[17] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Comput. Statist. Data Anal.*, vol. 52, no. 1, pp. 155–173, 2007.

[18] K. Devarajan, "Nonnegative matrix factorization: An analytical and interpretive tool in computational biology," *PLoS Comput. Biol.*, vol. 4, no. 7, p. e1000029, 2008.

[19] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò, "Collective human mobility pattern from taxi trips in urban area," *PLoS ONE*, vol. 7, no. 4, p. e34487, 2012.

[20] J. Shen, D. Liu, J. Shen, Q. Liu, and X. Sun, "A secure cloud-assisted urban data sharing framework for ubiquitous-cities," *Pervasive Mobile Comput.*, vol. 41, pp. 219–230, Oct. 2017.
[21] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.
[22] Q. Shi and M. Abdel-Aty, "Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 380–394, Sep. 2015.
[23] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 3–19, Jan. 2014.
[24] D. W. Xia, B. F. Wang, H. Q. Li, Y. T. Li, and Z. L. Zhang, "A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting," *Neurocomputing*, vol. 179, pp. 246–263, Feb. 2016.
[25] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, 2011.
[26] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 2, pp. 246–254, Jan. 2009.
[27] D. Ming-Jun and Q. Shi-Ru, "Fuzzy state transition and Kalman filter applied in short-term traffic flow forecasting," *Comput. Intell. Neurosci.*, vol. 2015, 2015, Art. no. 875243, doi: 10.1155/2015/875243.
[28] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transp. Res. C, Emerg. Technol.*, vol. 22, pp. 103–118, Jun. 2012.
[29] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, "Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 644–654, Jan. 2012.
[30] Z. Zheng and D. Su, "Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 143–157, Jun. 2014.
[31] D. Xia, H. Li, B. Wang, Y. Li, and Z. Zhang, "A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction," *IEEE Access*, vol. 4, pp. 2920–2934, 2016.
[32] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 653–662, Apr. 2015.
[33] M.-L. Huang, "Intersection traffic flow forecasting based on v-GSVR with a new hybrid evolutionary algorithm," *Neurocomputing*, vol. 147, pp. 343–349, Jan. 2015.
[34] F. Moretti, S. Pizzuti, S. Panzieri, and M. Annunziato, "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling," *Neurocomputing*, vol. 167, pp. 3–7, Nov. 2015.
[35] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 557–569, Feb. 2016.
[36] C. Xu, Z. Li, and W. Wang, "Short-term traffic flow prediction using a methodology based on autoregressive integrated moving average and genetic programming," *Transport*, vol. 31, no. 3, pp. 343–358, 2016.
[37] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.
[38] J. B. Zhang, Y. Zheng, and D. K. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2016, pp. 1655–1661.
[39] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Inf. Process. Manage.*, vol. 42, no. 2, pp. 373–386, 2006.
[40] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[41] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 61–78, May 2016.



XIANGHUA LI is currently a Lecture with the College of Computer and Information Science, Southwest University, Chongqing, China, and a Visiting Scholar with the Humboldt University of Berlin, Germany. Her current research interests include big data statistics and analysis, and intelligence algorithms.

JÜRGEN KURTHS is currently a Professor of nonlinear dynamics with the Humboldt University of Berlin, Germany. He is currently the Chair of the Research Domain Transdisciplinary Concepts with the Potsdam Institute for Climate Impact Research, Potsdam, Germany. He has authored over 650 papers with H-factor 71. His current research interests include complex networks, synchronization, and time series analysis. He is a fellow of the American Physical Society and a member of the Academia Europaea. He is an Editor-in-Chief of Chaos, and a member on the editorial board of other journals.

CHAO GAO was a Post-Doctoral Research Fellow with the Computer Science Department, Hong Kong Baptist University. He is currently an Associate Professor with the College of Computer and Information Science, Southwest University, Chongqing, China. His current research interests include complex social networks analysis, nature-inspired computing, and data-driven complex systems modeling.

JUNWEI ZHANG is currently pursuing the M.S. degree with the College of Computer and Information Science, Southwest University, Chongqing, China.

ZHEN WANG received the Ph.D. degree from Hong Kong Baptist University, Hong Kong, in 2014. He is currently a Professor with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University. He has authored or co-authored over 100 research papers and four review papers with over 4000 citations. His current research interests include complex networks, evolutionary game, and data science.

ZILI ZHANG received the B.Sc. degree from Sichuan University, the M.Eng. degree from the Harbin Institute of Technology, and the Ph.D. degree from Deakin University, all in computing. He is currently the Dean of the College of Computer and Information Science and the College of Software, Southwest University, Chongqing, China, and a Senior Lecturer with Deakin University, Australia. He has authored and co-authored over 100 refereed papers in international journals or conference proceedings, six monographs or textbooks published by Springer. His research interests include multi-agent system, bio-inspired AI, and big data statistics and analysis.