Location Recommendation for Enterprises by Multi-Source Urban Big Data Analysis

Guoshuai Zhao, Tianlei Liu, Xueming Qian, *Member, IEEE*, Tao Hou, Huan Wang, Xingsong Hou, and Zhetao Li

Abstract—Effective location recommendation is an important problem in both research and industry. Much research has focused on personalized recommendation for users. However, there are more uses such as site selection for firms and factories. In this study, we try to solve site selection problem by recommending some locations satisfying special requirements. There are many factors affecting it, including functions of architecture, building cost, pollution discharge etc. We focus on the specific site selection of meteorological observation stations in this paper with leveraging the factors of functions of architecture and building cost from multi-source urban big data. We consider not only recommending the locations that can provide more accurate prediction and cover more areas, but also minimizing the cost of building new stations. We design an extensible two-stage framework for the station placing including prediction model and recommendation model. It is very convenient for executives to add more real-life factors into our approach. We have some empirical findings and evaluate the proposed approach using the real meteorological data of Shaanxi province, China. Experiment results show the better performance of our approach than existing commonly used methods.

Index Terms—Big Data, Location Recommendation, Recommender System, Site Selection, Urban Computing

1 Introduction

Personalized location recommender systems mostly focus on exploring user information, which includes user's profiles, locations, and trajectories. Moreover, location recommendation can be used for site selection for firms and factories which are the new target audiences. Site selection affects the rationality, dependability, and efficiency of them. Thus we must take full account of the factors of functions of architecture, building cost, pollution discharge etc.

Recently, people concern not only general weather conditions such as sunny, windy, rainy and snowy but also the more detailed and accurate weather condition such as $PM_{2.5}$, PM_{10}, and NO_2. To some extent, the existing meteorological observation stations cannot satisfy people's requirements anymore. Therefore, it is urgent for us to construct new observation stations. Nevertheless, constructing a new observation station is both costly and time-consuming, which means that we cannot set up new stations as much as the existing stations in a short time. Thus, in this paper, we mainly focus on the specific site selection of meteorological observation stations by answering a practical question: How to recommend a few candidate locations to take the lead in constructing new observation stations?

In reality, there are several challenges. First, we need to consider the cost of building new observation stations in the recommended locations. Second, we prefer that the recommended locations are homodisperse in the map. Otherwise, the result may be a set of locations concentrated together which is obviously not proper. So our main task is to recommend locations that can make prediction accurate, construct new observation stations with low cost, and cover more areas.

In this paper, we propose a two-stage framework. Figure 1 is the overview. According to different personalized requirements, the multi-source data and multi-factors are taken into consideration. By training our recommendation model and prediction model, the scores of different locations are learned which denotes the importance of locations. Then the rank of candidate locations is obtained. Compared with our previous work [1], we 1) add more related works, 2) present more details about the concepts of total distance and relative area, 3) show the details about our model training, 4) show some interesting findings on spatial-temporal information, 5) show the actual result of our model for the demonstration. The main contributions of this work are:

- We solve the problem of how to recommend the locations to construct new meteorological observation stations by leveraging the factors of functions of architecture and building cost from multi-source

- *G. Zhao and X. Qian are with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, Xi'an 710049, China, and also with the SMILES Laboratory, Xi'an Jiaotong University, Xi'an 710049, China.*
  E-mail: zgs2012@stu.xjtu.edu.cn; qianxm@mail.xjtu.edu.cn
- *Tianlei Liu is with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China.*
  E-mail: tianleiliu2015@gmail.com
- *H. Wang and X. Hou are with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.*
  E-mails: wang.huan.1006@stu.xjtu.edu.cn; houxs@mail.xjtu.edu.cn
- *T. Hou is with the Shaanxi Provincial Lightning Protection Center, Xi'an 710049, China.*
  E-mail: 263346028@qq.com
- *Z. Li is with the College of Information Engineering, Xiangtan University, Hunan 411105, China.*
  E-mail: liztchina@gmail.com

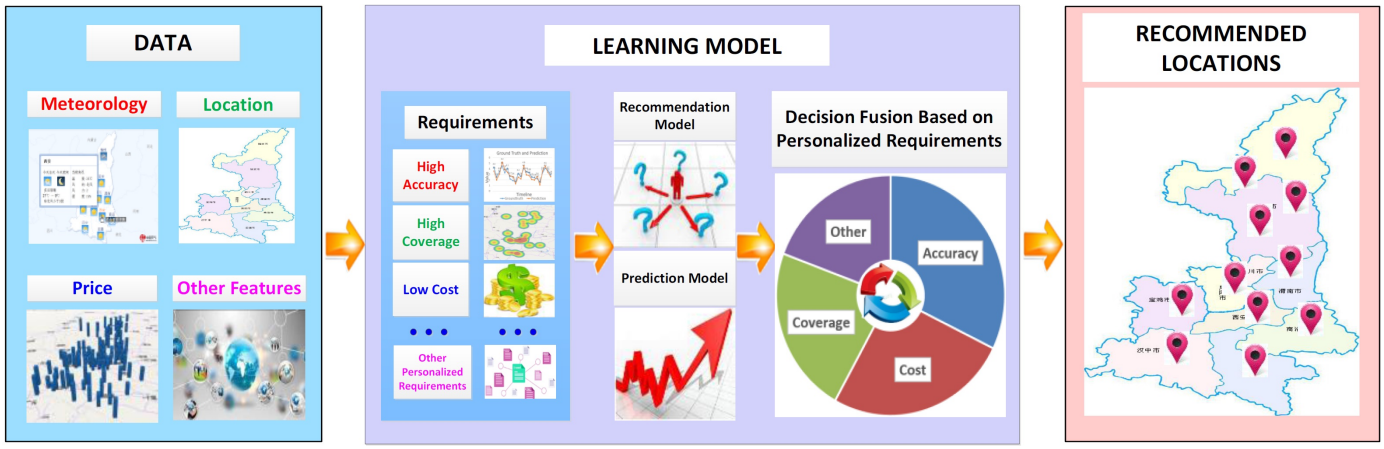


Fig. 1. The overview of our framework. According to different personalized requirements, the multi-data and multi-factors are taken into consideration. By training our recommendation model and prediction model, the scores of different locations are learned. Then the scores of locations are calculated and the solution is obtained.

urban big data. More specifically, the functions of architecture should make more accurate prediction and cover more areas. We propose a prediction model and a recommendation model which can fuse multi-source data.

- Geographical location information is explored to improve the accuracy of our prediction model because of the assumption that the more close two locations are, the more similar their meteorological data becomes. In the recommendation model, we would like to select the locations that can cover more geospatial areas. In addition, we utilize the concept of relative area to remove the border location.
- Besides the factor of geographical location, in the recommendation model, we take building cost into account. We would like to select the locations whose benchmark prices of industrial land are low. These factors are fused into our approach to learn the importance of locations to satisfy the personalized requirements of decision-makers.

The rest of this paper is organized as follows: In section 2, we present some related works on sample selection, environmental prediction, and location recommender systems. In section 3, our models proposed in this paper including recommendation model and prediction model are described in detail. Experiments and some empirical findings are introduced and the evaluation of our approach is given in Section 4. Finally, we conclude the paper in Section 5.

2 Related Works

The research is a part of urban computing [2]. Here, we review some related works on site selection.

2.1 Sample Selection

Many selective sampling problems were solved based on information entropy theory [3], [4], [5], [6] and probability [7], [8]. Hsieh et al. [3] established new stations at the locations those can minimize the uncertainty of the prediction model. They pick the location with the lowest entropy and then put it into the prediction model as known data. Then pick the second-to-last location which is the location with the lowest entropy in the new prediction model and keep running this circle. Finally, select the top k ranked locations as the location to build new stations. Du et al. [9] aimed to find a set of locations for sensor deployment to best measure the surface wind distribution over a large urban reservoir. They solve this problem by finding locations with the largest mutual information with others based on some heuristics. Erdös et al. [10] aimed to deploy sensors in an information delivery network to optimize the detection of duplicate data contents. Wang et al. [11] leveraged the spatial and temporal correlation among the data sensed in different sub-areas to significantly reduce the required number of sensing tasks allocated (corresponding to budget), yet ensuring the data quality. Karamshuk et al. [12] aimed to find a set of locations so that the placement of new retail stores can bring a maximum number of customers. They formulate the task as a learning-to-rank problem based on geographical and human mobility features. Krause et al. [13] proposed to find a set of locations such that the wireless sensors can best predict some future events, such as road speeds on a highway. Ordinary Kriging (OK) proposed in [14] is one of the most widely used interpolation models. Pourali et al. [15] utilize Bayesian brief network to find a set of functional locations such that the placement of sensors can best monitor a complex power systems.

2.2 Environmental Prediction

There are a lot of ways to prediction based on different theories such as matrix factorization [16], [17], [18], probability, cluster and similarity etc. Zheng et al. [19] proposed a semi-supervised learning approach based on

a co-training framework that consists of two separated classifiers to infer the real-time and fine-grained air quality. Zheng et al. [20] reported on a real-time air quality forecasting system that uses data-driven models to predict fine-grained air quality over the following 48 hours. Satellite remote sensing [21] is a top-down approach to derive the air quality of the urban surface, which has been used for many years. Donnelly et al. [22] presented a model for producing real time air quality forecasts with both high accuracy and high computational efficiency. Shang et al. [23] used GPS trajectories of the sample of vehicles to infer the city-wide vehicular emissions. Over the past decade, some statistic models, like linear regression, regression tree, and neural networks, have been employed in atmospheric science to do a real-time prediction of air quality [22], [24], [25]. Existing air quality prediction methods in Environmental Science are usually based on classical dispersion models, such as Gaussian Plume models, Operational Street Canyon models, and Computational Fluid Dynamics [26]. These models are in most cases a function of meteorology, street geometry, receptor locations, traffic volumes, and emission factors (e.g. g/km per single vehicle), based on a number of empirical assumptions and parameters that might not be applicable to all urban environments [20], [26].

2.3 Location Recommender Systems

Recently many researchers pay more attention on recommender system [18], [17], [3], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. Zheng et al. [28] presented a comprehensive survey of recommender systems for LBSNs (Location based Social Networks), analyzing the data source used, the methodology employed to generate a recommendation, and the objective of the recommendation. The major methodologies used by recommender systems in location-based social networks can be divided into 3 categories, which are content based [47], [48], [49], link analysis [50], [51], and collaborative filtering [52], [53].

Content-based recommendation systems make recommendations by matching users' preferences [47], [54], [55], [56]. Users' preferences are discovered from users' profiles such as gender and age, and features of locations, such as tags and categories. The methods [35], [55], [56], [57], [58], [59] make recommendations by discovering users' locations and activity histories. Bao et al. [39] combined user's location and preference to provide effective location recommendations. Kang et al. [60] proposed a web service recommendation approach incorporating a user's potential preferences and diversity feature of user interests on web services.

Link analysis algorithms can find special nodes from a complicated structure which is applied for identifying important web pages for web searching. By analyzing the LBSN, link analysis algorithms extract locations meeting different needs. Zheng et al. [50] explored interests of locations based on HITS algorithm, and made recommendations by considering the interests of locations and users' travel experiences. Raymond et al. [61] made recommendations based on a random walk-based link analysis algorithm.

Recommendation systems based on collaborative filtering recommend a location to a user if this location has been visited by a similar user. It is widely utilized in products services [52], [53], [62] and travel recommendations [32], [50], [57], [63], [64], [65], [66] and service recommendation [67], [68], [69]. Jiang et al. [31] proposed a user topic based collaborative filtering approach for personalized travel recommendation. Zhang et al. [56] proposed a probabilistic framework to utilize temporal influence correlations for location recommendations by measuring the similarity between users. Sang et al. [70] [32] considered both sensor and user context to develop a contextual recommendation algorithm, and a hierarchical scheme is designed for coarse-to-fine POI recommendation. They [71] also designed a two-level solution to solve the problem of location visualization from multiple semantic themes. It first identifies the POIs and discover the focused themes, and then aggregates the low-level POI themes to generate high-level city themes for location visualization. Sang et al. [72] presented some challenges in social multimedia mining and reviews current studies on this topic. Moreover, it also presents some future directions to help to inspire audiences.

In addition, there are some other kinds of recommender systems, such as song recommendation [42], social friend recommendation [41], product recommendation [18], [37], and video recommendation [43], and service recommendation [67], [68], [69], [73], [74], [75].

2.4 Differences

In this section, we summarize the differences between location recommendation for enterprises and location recommendation for individuals. They are:

- Our target is to recommend new sites for enterprises to construct new branches, factories, or observation stations, whereas individual location recommendation aims at mining the interesting POIs that the individual user likes to visit. Our targets are quite different.
- Our data is different from the individual location recommendation. We explored meteorological data, GPS information of county, and the benchmark price of industrial land to recommend locations for building new meteorological observation stations, whereas the individual location recommendation mainly uses the user generated data, including user check-ins, comments, social circles, and other contextual information. Our data are quite different even both of us utilized GPS information.
- The basic idea of the individual location recommendation model is to find similar users and similar

POIs, and measure the preferences of users for POIs. However, in this work, we designed a prediction model and a recommendation model considering three factors, including the dispersity of the locations, building cost, and the accuracy of meteorological prediction. Our models are quite different.

Besides, compared with the related works on sample selection, the biggest difference of our work is that we solved how to recommend the locations to construct new meteorological observation stations. This is a new application scene. Moreover, we considered selecting the locations that can cover more geospatial areas and considered selecting the locations whose benchmark prices of industrial land are low. The two ideas are also different from related works.

3 Our Location Recommendation for Enterprises

We would like to recommend the locations to construct new meteorological observation stations by leveraging the factors of functions of architecture and building cost. More specifically, the functions of architecture are performing more accurate prediction and covering more areas. Therefore, we propose a prediction model and a recommendation model. As shown in Figure 1, our task is to train our recommendation model and prediction model and rank the locations according to the learned scores.

We divide the whole geospatial area into several regions by administrative divisions. Each region is the basic unit in our prediction model. In some of the regions, there is an observation station which can provide us the exact record data of meteorology in the region. The meteorological data could be represented by $\mathbf{R}_i(i = 1, 2, ..., n)$. We assume that m out of n locations will be recommended in our task to construct new stations. In addition, the real-life factors will be taken into consideration to rank candidate locations. Symbols and notations utilized in this paper are given in Table 1. **A** and **B** are the coefficient matrices to be learned in our prediction model. **C**, **D** and **E** are the coefficient matrices to be learned in our recommendation model.

TABLE 1
Notations and Their Descriptions

Notations	Descriptions
A	The matrix of meteorological correlation between every two locations in prediction model
B	The matrix of geo-distance correlation between every two locations
C	The matrix of meteorological correlation between every two locations in recommendation model
D	The importance matrix of area coverage
E	The importance matrix of benchmark price
G	The matrix of geo-distance between every two locations
P	The matrix of benchmark price
R	The matrix of meteorological data in each location
S	The matrix of meteorological data in selected locations
W	The matrix of coverage score to constrain the importance of area coverage

3.1 Prediction Model

The observation data of meteorology in different locations are correlated with each other in spatial perspectives. Considering the correlation of the meteorological data between each location in these areas, the unknown data can be predicted through little-observed data. That is to say, we use the data observed in recommended locations to predict the data in un-selected locations. We propose our initial prediction model given by:

$$\min_{\mathbf{A}} \|\mathbf{R} - \mathbf{AS}\|_F^2 + \alpha\|\mathbf{A}\|_F^2 \tag{1}$$

where **S** is the matrix of meteorological data in recommended locations. Matrix **A** consists of coefficient a_{pk} which represents the correlation of meteorological data between the recommended location p and the unrecommended location k. **AS** is the meteorology prediction of our model. The second term is used to avoid over-fitting.

The geo-distance between regions are also an important factor in our prediction model. As we can see in Figure 2, the x-axis represents the distance between regions and the y-axis indicates the corresponding difference of the meteorological data. We calculate the sum of mean absolute error of the thirty years' thunderstorm data between every two locations as the corresponding difference of the meteorological data. It shows the positive correlation, which means the distance factor is important and should be considered in prediction model, because the more close two locations are, the more similar their meteorological data becomes. We utilize matrix **B** represents the similarity between each locations' distance and the objective function of prediction model must contain the following term:

$$\min_{\mathbf{B}} \|\mathbf{G}_{min} - \mathbf{BG}\|_F^2 + \alpha\|\mathbf{B}\|_F^2 \tag{2}$$

where matrix **B** consists of coefficient b_{pk} which represents the correlation of geo-distance between p and k. **BG** is optimized to $\mathbf{G}_{min}$, which means the bigger the value of b is, the more close the two locations are, and the more similar their meteorological data becomes. Therefore, the objective function of our prediction model is given by:

$$\begin{aligned}\Phi =& \|\mathbf{R} - (\mathbf{A}+\mathbf{B})\mathbf{S}\|_F^2 + \alpha_1\|\mathbf{G}_{min} - \mathbf{BG}\|_F^2 \\ &+ \alpha_2\|\mathbf{A} + \mathbf{B}\|_F^2\end{aligned} \tag{3}$$

where the first term is used to constrain the errors. The second term is used to constrain parameter b considering with the factor of geo-distance. The third item is used to avoid over-fitting. The task is to optimize **A** and **B** by minimizing this objective function, which will be reported in Part C of Section III.

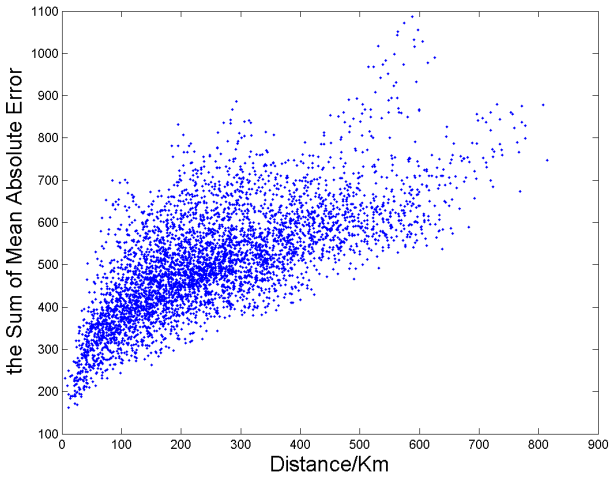


Fig. 2. The relevance between geographical distance and the difference of observation data.

3.2 Recommendation Model

We would like to select the locations that can help us to make the more accurate prediction, cover more regions, and have lower building cost to establish observation stations. In this study, we propose leveraging the importance of locations. It includes three aspects, i.e. the importance of locations with regard to prediction accuracy, the importance with regard to area coverage, and the importance with regard to building cost.

First, considering the importance of locations with regard to prediction accuracy, we ought to reduce the errors of prediction in our recommendation model. A linear combination of each location's record data $\mathbf{R}_p$ is utilized to calculate the prediction. But only m of the most important locations can be recommended. We use the weight of each $\mathbf{R}_p$ to represent the importance of location p with regard to prediction accuracy. The more correlation with others the location has, the more important the location is. Thus, we propose the initial recommendation model containing only the prediction accuracy as:

$$\min_{\mathbf{C}} \|\mathbf{R} - \mathbf{CR}\|_F^2 + \beta_3 \|\mathbf{C}\|_F^2 \tag{4}$$

where $\mathbf{CR}$ represents the prediction of the meteorological data by the linear combination of the other regions' data. This term is used to reduce the square error by solving the parameter c. $c_{\cdot i}$, which is equal to $\sum_p c_{pi}$, represents the importance of the location i.

Second, we recommend the locations that can cover most geospatial areas on the map in order to make sure every location in our province will not leave the recommended locations too far. It can help us to predict the more accurate meteorological data which can be proved by Figure 2. Nevertheless, the coverage area is a definition that cannot be clearly measured, so as shown in Figure 3, we propose to employ the total distance $w_{\cdot i}$ between one location and other locations. It is calculated

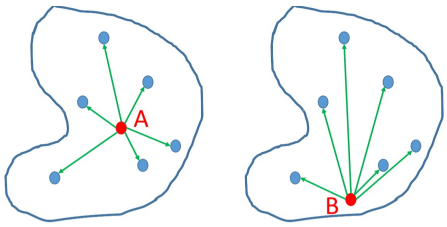


Fig. 3. The illustration of the total distance between one location and other locations. The total distance between point A and other points is smaller than that with B. It implies the smaller the total distance is, the more concentrated the point is.

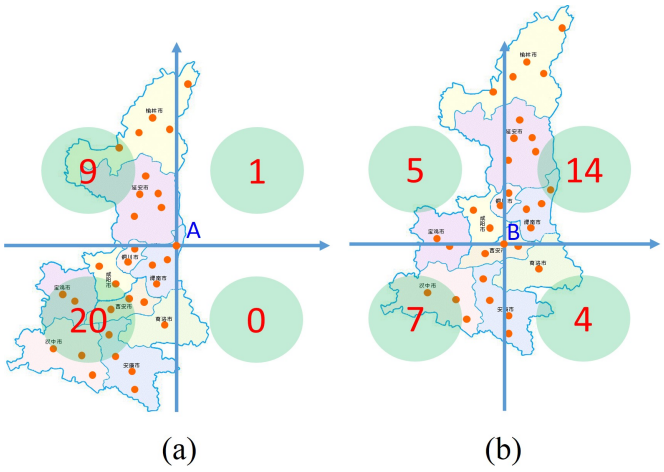


Fig. 4. The illustration of the concept of the relative area which is utilized to remove the border locations. We randomly select several locations on the map. In the case of location A in (a), first, we establish the coordinate system. Second, respectively record the number of locations in the four quadrants. Last, we use the four numbers [1, 9, 20, 0] to describe its relative area. As long as there is a parameter is 0, it indicates the location is at the border.

by $\sum_p t_{pi}$ where t is the geographical distance. The smaller the total distance is, the more concentrated the point is. In addition, for the locations at the border of the map, we leverage the concept of relative area [76], [77] to remove them as shown in Figure 4. In the case of location A in (a), we 1) establish the coordinate system; 2) respectively record the number of locations in the four quadrants; 3) use the four parameters [1, 9, 20, 0] to describe its relative area. As long as there is a parameter is 0, it indicates the location is at the border.

However, in case that the recommended locations are concentrated together, we suggest to apply the coverage score that comes from the tuned total distance according to the dispersity. The process is given in Figure 5: for the ranked total distances from largest to smallest, ① the first location that has the largest total distance (i.e. the geographical center) is mapped; ② map the second location if the distance between it and the previous

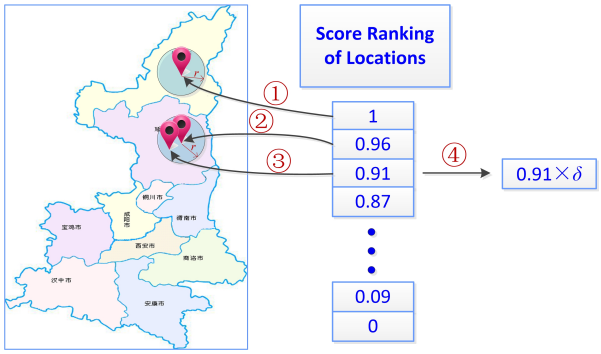


Fig. 5. Score ranking with considering the coverage of recommended locations by the following steps: for the ranked total distances from largest to smallest, ① the first location which is the geographical center is mapped; ② map the second location, if the distance between it and the previous locations are larger than the predefined parameter r; ③ if the distance between it and the previous locations are smaller than r; ④ its location score will be multiplied by a coefficient δ_i.

locations are larger than the predefined parameter r, i.e. it is not in the coverage area of the previous locations; ③ if the distance between them are smaller than r, ④ its total distance will be multiplied by a coefficient δ_i. After several rounds, we can get the final coverage scores $\mathbf{W}$:

$$\mathbf{W} = \mathbf{T}\delta \tag{5}$$

Then the importance of locations should be proportional to the coverage score:

$$\min_{\mathbf{D}} \|\mathbf{W}_{max} - \mathbf{DW}\|_F^2 + \beta_3 \|\mathbf{D}\|_F^2 \tag{6}$$

where the parameter $d_{\cdot i}$ which is equal to $\sum_p d_{pi}$ denotes the importance of the location i with regard to area coverage. δ_i is the discriminant coefficient as shown in Figure 5. It is 0.5 if the location belongs to the r-radius circle of previous locations, and is equal to 1 otherwise.

Third, we focus on the cost of building new stations. When the decision makers are facing this kind of selection problems, they try to minimize the cost in the whole project. Therefore, we need to fuse the factor of cost into our model. Brevity we utilize the benchmark price of industrial land to represent the cost. Thus, the optimization is given by:

$$\min_{\mathbf{E}} \|\mathbf{P}_{min} - \mathbf{EP}\|_F^2 + \beta_3 \|\mathbf{E}\|_F^2 \tag{7}$$

where the value of parameter $e_{\cdot i}$ which is equal to $\sum_p e_{pi}$ denotes the importance of the location i with regard to the building cost. $\mathbf{P}$ is the matrix of benchmark price. $\mathbf{EP}$ is optimized to $\mathbf{P}_{min}$, which means the bigger the value of $e_{\cdot i}$ is, the lower the cost is, and meanwhile the more important the location is.

Then the objective function of the final recommendation model is given by:

$$\begin{aligned}\Psi =& \|\mathbf{R} - \mathbf{CR}\|_F^2 + \beta_1 \|\mathbf{W}_{max} - \mathbf{DW}\|_F^2 + \beta_2 \|\mathbf{P}_{min} - \mathbf{EP}\|_F^2 \\ &+ \beta_3 (\|\mathbf{C}\|_F^2 + \|\mathbf{W}\|_F^2 + \|\mathbf{E}\|_F^2)\end{aligned} \tag{8}$$

where the first term is used to constrain solve the importance of locations with regard to prediction accuracy. The second term is utilized to solve the importance with regard to area coverage. The third term is leveraged to optimize the importance with regard to building cost. The last term avoids over-fitting.

In a word, we have three essential parts in our recommendation model. The first part selects the most important locations for the meteorological data prediction. The second part chooses the locations which possess the larger coverage and the third part opts the lower cost locations. Changing coefficients β_1 and β_2 can balance the three factors. At last, top-m biggest values of $c_{\cdot i} + d_{\cdot i} + e_{\cdot i}$ are figured out and the corresponding regions are the final locations we seek.

3.3 Model Training

Given the proposed prediction model and recommendation model, the objective functions represented in Equation (3) and (8) can be minimized by the gradient decent approach as in [16], [18], [17]. The gradients of the objective function of recommendation model with respect to the variables $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$ are given by:

$$\frac{\partial \Psi}{\partial \mathbf{C}} = -2(\mathbf{R} - \mathbf{CR}_s)\mathbf{S} + 2\beta_3 \mathbf{C} \tag{9}$$

$$\frac{\partial \Psi}{\partial \mathbf{D}} = -2\beta_1(\mathbf{W}_{max} - \mathbf{DW})\mathbf{W} + 2\beta_3 \mathbf{D} \tag{10}$$

$$\frac{\partial \Psi}{\partial \mathbf{E}} = -2\beta_2(\mathbf{P}_{min} - \mathbf{EP})\mathbf{P} + 2\beta_3 \mathbf{E} \tag{11}$$

Once we get the gradients, we update these matrices during each iteration as follows:

$$\mathbf{C}^{(t)} = \mathbf{C}^{(t-1)} - l_1^{(t)} \frac{\partial \Psi^{(t-1)}}{\partial \mathbf{C}} \tag{12}$$

$$\mathbf{D}^{(t)} = \mathbf{D}^{(t-1)} - l_1^{(t)} \frac{\partial \Psi^{(t-1)}}{\partial \mathbf{D}} \tag{13}$$

$$\mathbf{E}^{(t)} = \mathbf{E}^{(t-1)} - l_1^{(t)} \frac{\partial \Psi^{(t-1)}}{\partial \mathbf{E}} \tag{14}$$

As mentioned before, when we get the final $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$, top-m biggest values of $c_{\cdot i} + d_{\cdot i} + e_{\cdot i}$ are figured out and the corresponding regions are the final recommended locations.

The gradients of the objective function of prediction model with respect to the variables $\mathbf{A}$ and $\mathbf{B}$ are given by:

$$\frac{\partial \Phi}{\partial \mathbf{A}} = -2(\mathbf{R} - (\mathbf{A} + \mathbf{B})\mathbf{S})\mathbf{S} + 2\alpha_2(\mathbf{A} + \mathbf{B}) \tag{15}$$

Algorithm 1 The Procedure of Our Approach

Input: The matrices of our data, including matrices **R**, **S**, **G**, **T**, and **P**.
Setting the parameters, including iteration count t_1, t_2, learning rate l_1, l_2, and tradeoff parameters α_1, α_2, β_1, β_2, and β_3.
Output: The final rank of candicate locations.
The corresponding evaluation of the solution.
The building cost of this solution.
1: Initialize the variable matrices those denote the importance of locations, including matrices **C**, **D**, and **E**.
2: for $n = 1 : t_1$ do
3: Calculate the gradients of the objective function proposed in Equation (8) with respect to the variables **C**, **D**, and **E** respectively by Equation (9), (10), and (11).
4: Update matrices **C**, **D**, and **E** with the gradients by the learning rate l_1 by Equation (12), (13), and (14).
5: end for
6: Top-m biggest value of $c_{\cdot i} + d_{\cdot i} + e_{\cdot i}$ are figured out and the corresponding regions are the candidate locations.
7: Calculate the building cost and the dispersity of the locations.
8: **Output** the candidate locations, the building cost, and the dispersity.
9: Initialize the variable matrices those denote the correlation of locations in prediction model, including matrices **A** and **B**.
10: for $n = 1 : t_2$ do
11: Calculate the gradients of the objective function proposed in Equation (3) with respect to the variables **A** and **B** respectively by Equation (15) and (16).
12: Update matrices **A** and **B** with the gradients by the learning rate l_2 by Equation (17) and (18).
13: end for
14: Predict the meteorological data by the learned **A** and **B**.
15: Calculate the prediction error by RMSE and MAE.
16: **Output** the accuracy evaluation.

$$\frac{\partial \Phi}{\partial \mathbf{B}} = -2(\mathbf{R} - (\mathbf{A} + \mathbf{B})\mathbf{S})\mathbf{S} - 2\alpha_1(\mathbf{G}_{min} - \mathbf{B}\mathbf{G})\mathbf{G} + 2\alpha_2(\mathbf{A} + \mathbf{B}) \tag{16}$$

we update these matrices during each iteration as follows:

$$\mathbf{A}^{(t)} = \mathbf{A}^{(t-1)} - l_2^{(t)} \frac{\partial \Psi^{(t-1)}}{\partial \mathbf{A}} \tag{17}$$

$$\mathbf{B}^{(t)} = \mathbf{B}^{(t-1)} - l_2^{(t)} \frac{\partial \Psi^{(t-1)}}{\partial \mathbf{B}} \tag{18}$$

Algorithm 1 summarizes the whole procedure of our framework. Steps 1 to 8 show the details of our recommendation model. Steps 9 to 16 show the details of our prediction model.

4 EXPERIMENT

This section introduces the experiments in detail. The definition of the problem to be solved is that there are 22 meteorological stations to be built in Shaanxi Province, China, and then how to select the locations. We perform the proposed models to solve this problem. Here, 1) the details of our dataset are introduced, 2) some analysis of meteorological data is presented, 3) the performance measurements are reported and 4) some experimental results and discussions are given.

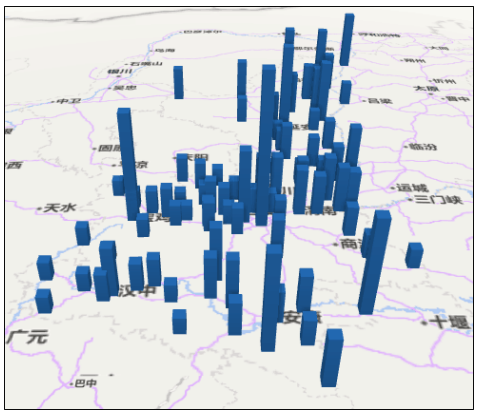


Fig. 6. The distribution of benchmark price of industrial land in Shaanxi Province. The higher the column is, the higher the price is.

4.1 Dataset Introduction

4.1.1 Meteorological Data

The meteorological data used in this paper is provided by Shaanxi Provincial Lightning Protection Center. It contains the count of thunderstorm days in each county of Shaanxi. In addition, the ten prefecture-level divisions of Shaanxi are subdivided into 107 county-level divisions. But some of them are too small so that they are merged into near divisions in the provided meteorological data. In a words, there are 96 divisions in our dataset. Moreover, the data range is from 1974 to 2011 based on one-month intervals. We utilize the meteorological data before 2000 as the training set and the other as the test set.

4.1.2 Geographical Location Data

Geographical location data is represented by Global Position System (GPS) coordinate which contains the longitude and latitude. The geographical distance between two latitude/longitude coordinates is calculated by using the Haversine geodesic distance equation proposed in [78]. We crawled the geographical location data of each county from the Internet.

4.1.3 Benchmark Price of Industrial Land

The benchmark price of industrial land released in 2010 was crawled from the Internet to approximately represent the cost of building stations. The benchmark price in Xi'an is almost 13 times higher than it in Yijun County from which we can see that it is necessary to take the benchmark price of industrial land into consideration. Figure 6 shows the distribution of benchmark price of industrial land in Shaanxi Province. The higher the column is, the higher the price is.

4.2 Some Empirical Findings

4.2.1 Temporal Findings

Figure 7 shows the trend of thunderstorm day count in Shaanxi Province. The blue line denotes the count

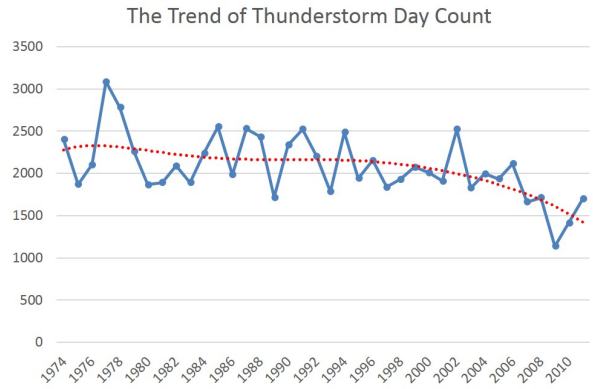


Fig. 7. The trend of thunderstorm day count in Shaanxi Province. The value of x-axis is the year. The value of y-axis denotes the count of thunderstorm days in different years.

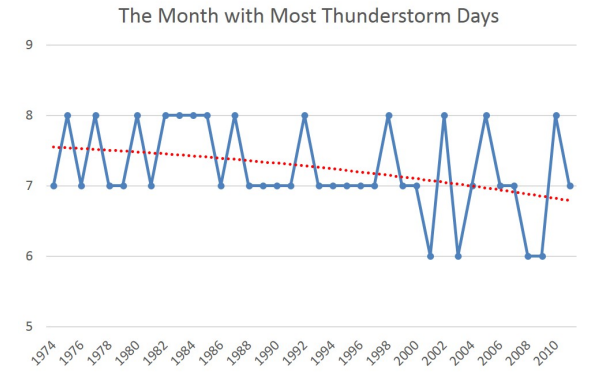


Fig. 8. The trend of the month with most thunderstorm days in Shaanxi Province. The value of x-axis is the year. The value of y-axis denotes the month of most thunderstorm days in different years.

of thunderstorm day in different years. The dotted line in red implies the trend with years. We can see that the count of thunderstorm day in Shaanxi is overall decreasing with the years. It reached the lowest level in 2009.

Figure 8 shows the Change of the month of most thunderstorm days in Shaanxi Province. The blue line denotes the month of most thunderstorm days in different years. The dotted line in red implies the trend with years based on the 2nd order polynomial. It can be seen that the month of most thunderstorm days is overall shifting to an earlier date with the years, especially after 2000.

4.2.2 *Spatial Findings*

Figure 9 shows the distribution of meteorological data in Shaanxi Province. The warm color represents more thunderstorm days. The cool color implies fewer thunderstorm days. It can be seen that there are more thunderstorm days in the Plateau of Northern Shaanxi and in the mountains of southern Shaanxi than those in the plain of Central Shaanxi.

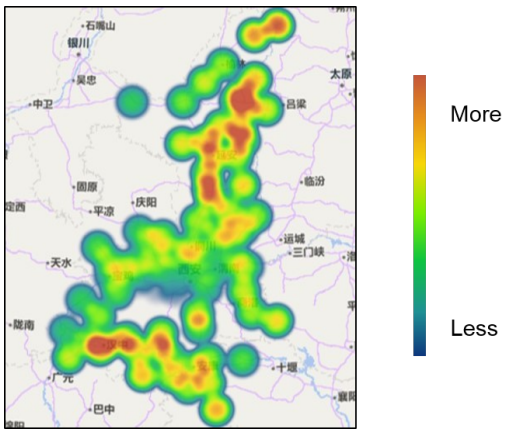


Fig. 9. The distribution of meteorological data in Shaanxi Province. The warm color represents more thunderstorm days. The cool color implies fewer thunderstorm days.

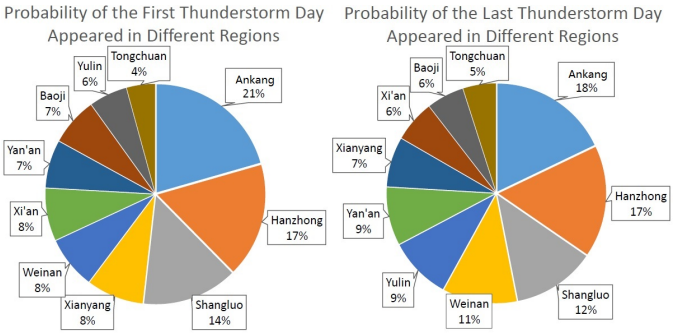


Fig. 10. The probability of the first thunderstorm day and the last thunderstorm day appeared in different regions of Shaanxi Province.

Fig. 10 shows the probability of the first thunderstorm day and the last thunderstorm day appeared in different regions of Shaanxi Province. It can be seen that early in the year the first thunderstorm day mostly appears in the mountains of southern Shaanxi, including Ankang, Hanzhong, Shangluo. However, we find that the last thunderstorm day also mostly appears in above regions. So we suppose that the monsoon should be associated for figuring this meteorological phenomenon out. Shaanxi Province belongs to monsoon region as shown in Fig. 11. When the rainy season is coming, the climate of Shaanxi is influenced by the East Asian Monsoon. Monsoon comes from the southeast and southwest. Thus, the mountains of southern Shaanxi are the first to enter rainy season. When the rainy season is fading away, the monsoon blows toward the sea. That is to say, the monsoon retreats from north to south. Finally, the last thunderstorm day mostly appears in southern Shaanxi.

4.3 Performance Measurements

The evaluation metrics of the prediction accuracy used in our experiments are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). They are the most popular accuracy measures in the literature of recommender

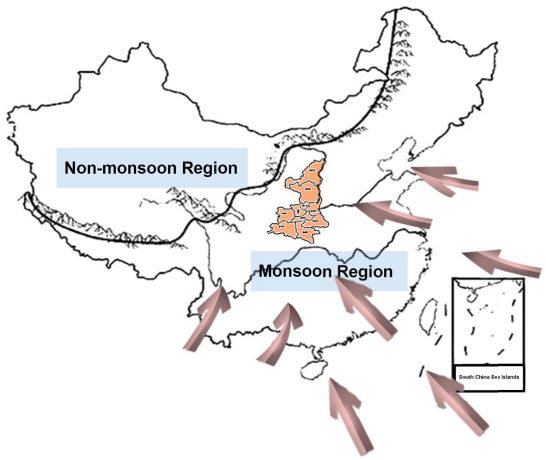


Fig. 11. Monsoon and non-monsoon regions in China. The heavy curve in black is the border of them. Shaanxi province that is in red locates in the area of the East Asian monsoon. The arrows indicate the directions of the summer monsoon. When the rainy season is coming, the monsoon comes from the southeast and southwest.

systems [3], [20], [17] [16]. RMSE and MAE are defined as:

$$RMSE = \frac{\|\mathbf{R}_{test} - (\mathbf{A} + \mathbf{B})\mathbf{S}_{test}\|_F}{|\mathbf{R}_{test}|} \tag{19}$$

$$MAE = \frac{\|\mathbf{R}_{test} - (\mathbf{A} + \mathbf{B})\mathbf{S}_{test}\|_1}{|\mathbf{R}_{test}|} \tag{20}$$

where $\mathbf{R}_{test}$ is the real meteorological data. $\mathbf{A}$ and $\mathbf{B}$ are the matrices learned by Equation 3. $\mathbf{S}_{test}$ is the real meteorology data in recommended locations. $|\mathbf{R}_{test}|$ denotes the number of data in the test set.

For cost comparison, we utilize the total benchmark price of industrial land in recommended locations to approximately evaluate the cost of building stations. It is defined as $COST = \|\mathbf{P}_{recommended}\|_1$, where $\mathbf{P}_{recommended}$ is the benchmark price of industrial land in the recommended locations.

In fact, the dispersity of selected locations are also important, which has been illustrated in Figure 2. Therefore, we propose a measurement of dispersity. The minimum of the distances between a location to others is calculated by $Dis_i = \min\{Dis_{i,1}, Dis_{i,2}, \cdots, Dis_{i,m}\}$, where i is belonged to the set of un-recommended locations, and m is the number of recommended locations. Then the variance is used to represent the dispersity as $Dispersity = var(DIS)$, where $DIS = \{Dis_1, Dis_2, \cdots, Dis_{n-m}\}$. n is the total number of regions.

In a word, four measurements including RMSE, MAE, COST, and Dispersity are utilized to evaluate our approach, and the lower, the better.

4.4 Evaluation

4.4.1 Compared Algorithms

We compare our algorithm with some other commonly used methods,including *Divergence*, *Rate of Change*, *K-means*, *Spectral Clustering*, *Gaussian Mixture Model* (GMM), *Artificial Neural Network* (ANN) with back propagation technique, *Support Vector Machine* (SVM) and *Matrix Factorization* (MF).

- *Divergence*, denoted by $\frac{\mu_1 - \mu_2}{\frac{1}{2}(\sigma_1^2 + \sigma_2^2)}$, where μ is the mean of a data set and σ is the variance of the data set. This approach selects data that have the minimum divergence value with the center data as a cluster.
- *Rate of Change* (RC), which is usually used in stock price prediction. It selects the data that have the minimum rate of change value with the center data as a cluster.
- *K-means*, which is one of the most popular methods in clustering.
- *Spectral Clustering* (SC), which is one of the most popular clustering methods based on Spectral Graph Theory.
- *Gaussian Mixture Model* (GMM), which is one of the most popular clustering methods aiming at learning probability density function for soft assignment clustering.
- *Artificial Neural Network* (ANN). We choose ANN with back propagation technique as another baseline. The ANN method is simply used as a classification model for meteorological data prediction.
- *Support Vector Machine* (SVM) is one of the most popular supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.
- *Matrix Factorization* (MF) is a factorization of a matrix into a product of matrices. It is usually used to learn the latent features in recommender systems.

Note that the last three algorithms are only used in the comparison of prediction performance. Our methods include NoN, Geo-distance, Cost, and LRE:

- NoN, which denotes the approach without any factors.
- Geo-distance, which denotes the approach with considering the factor of geo-distance.
- Cost, which denotes the approach with taking the factor of benchmark price of industrial land into account.
- LRE, which denotes the Location Recommendation for Enterprises with fusing all proposed factors.

4.4.2 Performance Comparison

Figures 12, 13, 14, and 15 show the performance comparison of different algorithms based RMSE, MAE, COST, and Dispersity respectively. Note that, the lower the four measures are, the better the performance is. It can be observed that our approaches are mostly better than

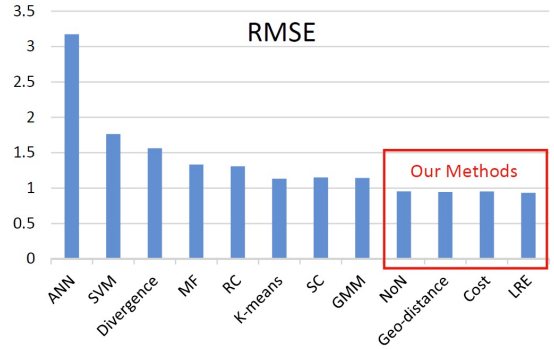


Fig. 12. Prediction performance comparison of different algorithms based on RMSE. In addition, the methods in the red box are ours.

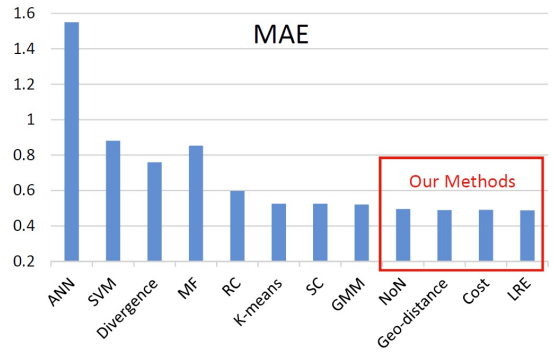


Fig. 13. Prediction performance comparison of different algorithms based on MAE. In addition, the methods in the red box are ours.

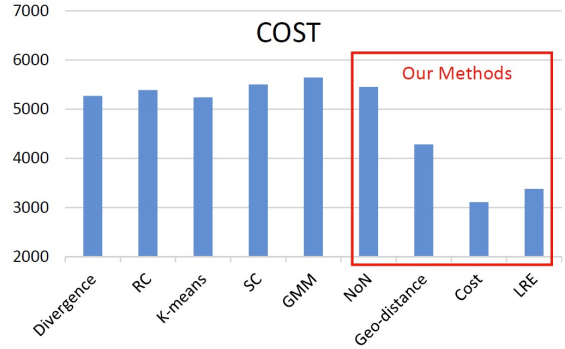


Fig. 14. Recommendation performance comparison of different algorithms based on COST. In addition, the methods in the red box are ours.

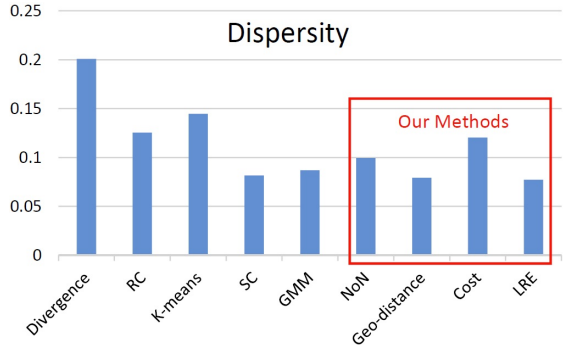


Fig. 15. Recommendation performance comparison of different algorithms based on Dispersity. In addition, the methods in the red box are ours.

the compared algorithms, especially in the comparison of COST and Dispersity. Moreover, from performance comparison, it can be seen that the factors fused in our approach are all effective. When we only consider the factor of cost, the performance of our approach on COST is much better than other algorithms. When we only take the factor of geo-distance, our approach also reaches the best performance on Dispersity. If we combine the two factors, our approach (LRE) achieves the optimal solution with balancing the two factors. Then decision makers can adjust the model according to their personalized requirements.

Figure 16 shows the actual results of K-means and our method LRE. In this figure, the rectangle landmarks in green are all the counties involved in our dataset. The counties with the pie landmarks are the recommended locations for building new meteorological observation stations. We also show their performances on RMSE, MAE, Dispersity, and COST. We can see that our method LRE not only could improve the prediction accuracy, but also could avoid selecting the border locations, such as A and B selected by K-means. Besides, compared with K-means, LRE also avoid the over-concentration of locations, such as C in K-means where three locations gather in a small scale.

With regard to the statistical comparison, as shown in Figures 12, 13, 14, and 15, we decrease 6.4% on MAE, 17.6% on RMSE, 35.6% on COST, and 5.2% on Dispersity. Note that, the lower the four measures are, the better the performance is.

4.4.3 *Discussions*

There are some parameters to balance the fused factors. In Equation (8), the parameter β_1 is the weight of the importance of geographical Dispersity. In Figure 5. The parameter r is used to avoid the concentration of recommended locations. In other words, r is designed to control the degree of dispersity directly. β_1 is served to regulate the extent of importance of dispersity. Both of them are related to the final performance on dispersity. Figures 17(a) and (b) show the impact of β_1 and r on performance. It can be seen that our approach could provide different solutions according to different requirements of dispersity. The cost of building new stations is one of the most concerned criterions. Adequate capital is the foundation of a booming company. Thus a cost-saving solution is expected. In our approach, the parameter of β_2 is set to manage the degree of importance of cost. Figure 17(c) demonstrates the effect of β_2 on performance of our approach in light of the cost of establishing new stations. Apparently, our approach offers various solutions according to different requirements of cost.

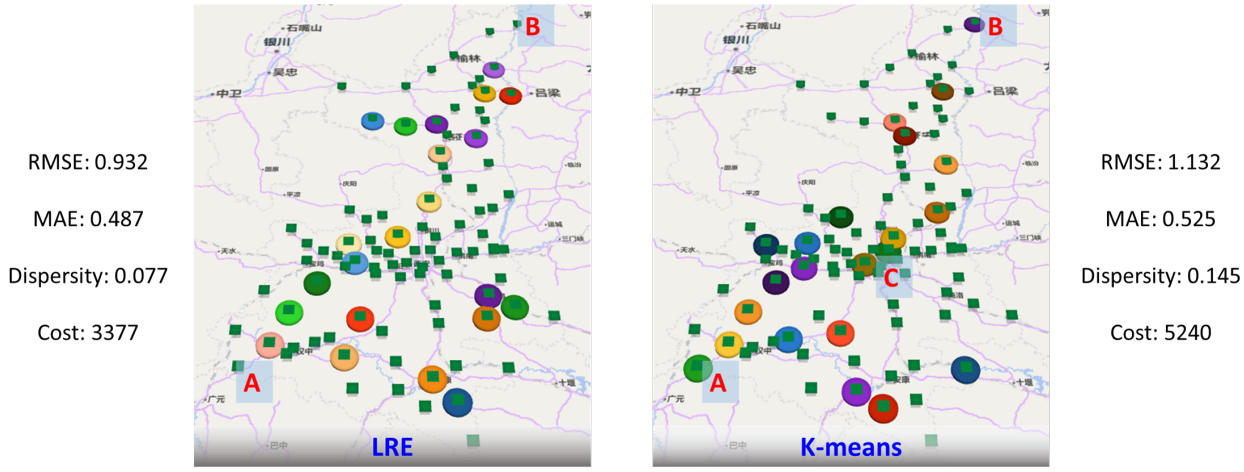


Fig. 16. The actual results of K-means and our method LRE. In this figure, the rectangle landmarks in green are all the counties involved in our dataset. The counties with the pie landmarks are the recommended locations for building new meteorological observation stations. We also show their performances on RMSE, MAE, Dispersity, and COST. Note that, the lower the four measures are, the better the performance is. We can see that our method LRE not only could improve the prediction accuracy, but also could avoid selecting the border locations, such as A and B selected by K-means. Besides, compared with K-means, LRE also avoid the over-concentration of locations, such as C in K-means where three locations gather in a small scale.

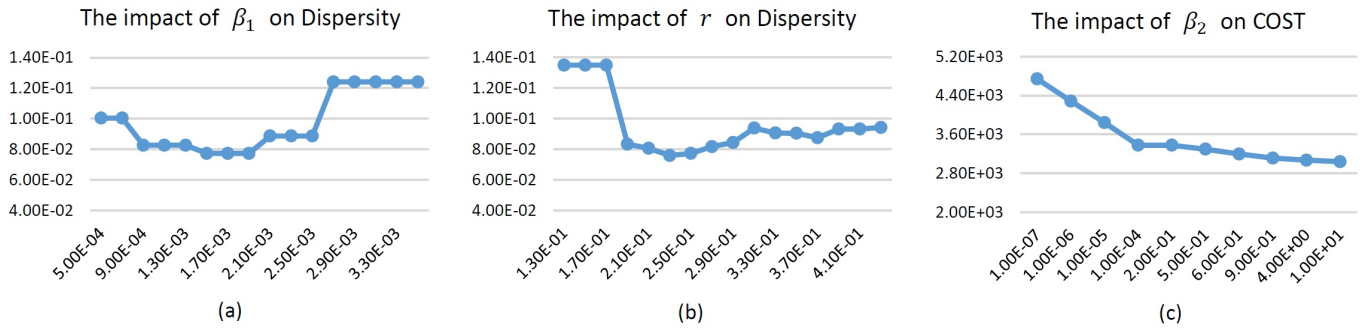


Fig. 17. Discussions on the impact of parameters on performance of our approach.

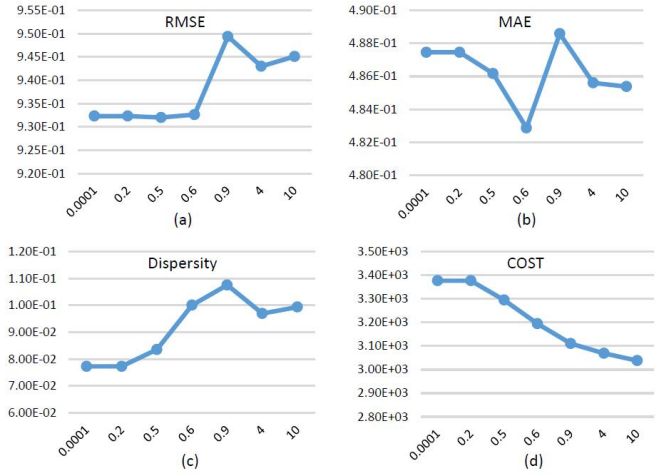


Fig. 18. The impact of factor parameter β_2 on different measures.

However, actually, we cannot find out the locations where all the factors could perform best. We use three factors: the accuracy of the prediction, the price of the selected land and the dispersity of all the selected locations to evaluate our model. When we try to seek the best performance on COST, we find our model will perform worse on other measures as shown in Figure 18. When we change the parameter β_2 of the benchmark price, the performance on COST becomes better with the increasing parameter. However, the performances on RMSE, MAE, and Dispersity are irregular, and most of them are inverse to the performance on COST. Thus, it is difficult to seek the best performance on all the measures.

With regard to the speed and efficiency, The space complexity of our algorithm is $\mathcal{O}(n \times k + 4n^2 + 4n)$, and the time complexity is $\mathcal{O}(t_1 \times n^2 \times k + t_2 \times n \times m \times k)$, where n is the number of regions. m is the number of recommended locations. k is the dimension of the meteorological data. Generally, because of $k \gg n, m$, the space complexity is $\mathcal{O}(n \times k)$. The time complexity of K-

means is $\mathcal{O}(n \times m \times t_1 \times k)$, and the time complexity of the recommendation part of our algorithm is equal to that of GMM which is also $\mathcal{O}(n^2 \times k \times t)$ [79]. However, generally, $n \gg m$, the time complexity of our algorithm is higher than K-means. For Spectral Clustering, its time complexity is high, because it needs to figure out the eigenvectors [79]. For the prediction part of our algorithm, the time complexity is equal to MF, which is $\mathcal{O}(t \times n \times m \times k)$. Other cluster based methods all directly regard the values of centers as the predictions. Thus, they do not cost any computing, but the weakness of their performance is apparent as shown in Figures 12 and 13.

5 Conclusions

In this paper, we introduced a framework to recommend locations for solving the problem of site selection. The factors of prediction accuracy, area coverage, and building cost are taken into account. We proposed a recommendation model and a prediction model to find out the optimized candidate locations. The weights of different factors can be fine tuned according to the personalized requirements. We solved the practical optimization problem and provided the solution with more intelligence. Additionally, we had some empirical findings which show the characteristics and trends of thunderstorm days, and some latent reasons were analyzed.

In our future work, the nonlinear prediction model will be performed, and more types of meteorological data, more urban data and more real-life factors will be considered. Besides, enterprises always have different requirements, so it is difficult to balance the fused factor parameters. For now, we just use the empirical parameters for the performance comparison. We would like to do more research on this topic and try to figure out how to balance the parameters with an auto-learning method in the condition of some basic requirements to improve performance most significantly.

References

[1] T. Liu, G. Zhao, H. Wang, X. Hou, X. Qian, and T. Hou, "Finding optimal meteorological observation locations by multi-source urban big data analysis," in *Proc. CCBD*, 2016, pp. 175–180.

[2] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM TIST*, vol. 5, no. 3, pp. 38:1–38:55, 2014.

[3] H. Hsieh, S. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proceedings of the 21th ACM SIGKDD*, 2015, pp. 437–446.

[4] B. Settles, "Active learning literature survey," *University of Wisconsinmadison*, vol. 39, no. 2, pp. 127–131, 2010.

[5] P. Sollich and D. Saad, "Learning from queries for maximum information gain in imperfectly learnable problems," in *Proc. NIPS*, 1994, pp. 287–294.

[6] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies." *Journal of Machine Learning Research*, vol. 9, no. 3, pp. 235–284, 2008.

[7] S. Hwang, C. Hsu, and C. Lee, "Service selection for web services with probabilistic qos," *IEEE Trans. Services Computing*, vol. 8, no. 3, pp. 467–480, 2015.

[8] R. Ramacher and L. Mönch, "Service selection with runtime aspects: A hierarchical approach," *IEEE Trans. Services Computing*, vol. 8, no. 3, pp. 481–493, 2015.

[9] W. Du, Z. Xing, M. Li, B. He, L. H. C. Chua, and H. Miao, "Sensor placement and measurement of wind for water quality studies in urban reservoirs," *ACM Trans. Sen. Netw.*, vol. 11, no. 3, pp. 41:1–41:27, Feb. 2015.

[10] D. Erdös, V. Ishakian, A. Lapets, E. Terzi, and A. Bestavros, "The filter-placement problem and its application to minimizing information multiplicity," *PVLDB*, vol. 5, no. 5, pp. 418–429, 2012.

[11] L. Wang, D. Zhang, A. Pathak, C. Chen, H. Xiong, D. Yang, and Y. Wang, "CCS-TA: quality-guaranteed online task allocation in compressive crowdsensing," in *Proc. ACM UbiComp*, 2015, pp. 683–694.

[12] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, "Geo-spotting: mining online location-based services for optimal retail store placement," in *Proc. KDD*, 2013, pp. 793–801.

[13] A. Krause, R. Rajagopal, A. Gupta, and C. Guestrin, "Simultaneous optimization of sensor placements and balanced schedules," *IEEE Trans. Automat. Contr.*, vol. 56, no. 10, pp. 2390–2405, 2011.

[14] M. A. Oliver and R. Webster, "Kriging: a method of interpolation for geographical information systems," *International Journal of Geographical Information Science*, vol. 4, no. 3, pp. 313–332, 1990.

[15] M. Pourali and A. Mosleh, "A functional sensor placement optimization method for power systems health monitoring," in *Proc. IEEE Industry Applications Society*, 2012, pp. 1–10.

[16] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. NIPS*, 2007, pp. 1257–1264.

[17] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1763–1777, 2014.

[18] G. Zhao, X. Qian, and X. Xie, "User-service rating prediction by exploring social users' rating behaviors," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 496–506, 2016.

[19] Y. Zheng, F. Liu, and H. Hsieh, "U-air: when urban air quality inference meets big data," in *Proc. ACM SIGKDD*, 2013, pp. 1436–1444.

[20] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proc. the 21th ACM SIGKDD*, 2015, pp. 2267–2276.

[21] Y. Jiang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang, "MAQS: a personalized mobile sensing system for indoor air quality monitoring," in *Proc. UbiComp*, 2011, pp. 271–280.

[22] A. Donnelly, B. Misstear, and B. Broderick, "Real time air quality forecasting using integrated parametric and non-parametric regression techniques," *Atmospheric Environment*, vol. 103, no. 103, pp. 53–65, 2015.

[23] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proc. ACM SIGKDD*, 2014, pp. 1027–1036.

[24] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real-time air quality forecasting, part I: History, techniques, and current status," *Atmospheric Environment*, vol. 60, no. 32, pp. 632–655, 2012.

[25] ——, "Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects," *Atmospheric Environment*, vol. 60, no. 6, pp. 656–676, 2012.

[26] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: a review," *Atmospheric Environment*, vol. 37, no. 2, pp. 155–182, 2003.

[27] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan, "Active learning in recommender systems," in *Recommender Systems Handbook*, 2015, pp. 809–846.

[28] J. Bao, Y. Zheng, D. Wilkie, and M. F. Mokbel, "Recommendations in location-based social networks: a survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.

[29] C. Cheng, H. Yang, I. King, and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in *Proc. AAAI*, 2012.

[30] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in *Proc. AAAI*, 2015, pp. 1721–1727.

[31] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized POI recommendations," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, 2015.

[32] J. Sang, T. Mei, J. Sun, C. Xu, and S. Li, "Probabilistic sequential pois recommendation via check-in data," in *Proc. ACM SIGSPATIAL*, 2012, pp. 402–405.

[33] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proc. ACM CIKM*, 2014, pp. 659–668.

[34] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proc. ACM RecSys*, 2013, pp. 25–32.

[35] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated GPS traces," *ACM TIST*, vol. 2, no. 1, p. 2, 2011.

[36] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura, "Geo topic model: joint modeling of user's activity area and interests for location recommendation," in *Proc. ACM WSDM*, 2013, pp. 375–384.

[37] G. Zhao, X. Qian, and C. Kang, "Service rating prediction by exploring social mobile users' geographical locations," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 67–78, 2017.

[38] P. Lou, G. Zhao, and X. Qian, "Schedule a rich sentimental travel via sentimental POI mining and recommendation," in *IEEE International Conference on Multimedia Big Data*, 2016.

[39] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proc. ACM SIGSPATIAL*, 2012, pp. 199–208.

[40] X. Wang, Y. Zhao, L. Nie, Y. Gao, W. Nie, Z. Zha, and T. Chua, "Semantic-based location recommendation with multimodal venue semantics," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 409–419, 2015.

[41] S. Huang, J. Zhang, L. Wang, and X. Hua, "Social friend recommendation based on multiple network correlation," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 287–299, 2016.

[42] K. Mao, L. Shou, J. Fan, G. Chen, and M. S. Kankanhalli, "Competence-based song recommendation: Matching songs to one's singing skill," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 396–408, 2015.

[43] P. Zhou, Y. Zhou, D. Wu, and H. Jin, "Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1217–1229, 2016.

[44] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1910–1921, 2016.

[45] G. Zhao, X. Qian, X. Lei, and T. Mei, "Service quality evaluation by exploring social users' contextual information," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3382–3394, 2016.

[46] X. Ma, X. Lei, G. Zhao, and X. Qian, "Rating prediction by exploring user's preference and sentiment," *Multimedia Tools and Applications*, Apr 2017.

[47] M. Park, J. Hong, and S. Cho, "Location-based recommendation system using bayesian user's preference model in mobile devices," in *Proc. UIC*, 2007, pp. 1130–1139.

[48] M. Ye, D. Shou, W. Lee, P. Yin, and K. Janowicz, "On the semantic annotation of places in location-based social networks," in *Proc. ACM SIGKDD*, 2011, pp. 520–528.

[49] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," *J. Ambient Intelligence and Humanized Computing*, vol. 5, no. 1, pp. 3–19, 2014.

[50] Y. Zheng, L. Zhang, X. Xie, and W. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. WWW*, 2009, pp. 791–800.

[51] H. Yoon, Y. Zheng, X. Xie, and W. Woo, "Smart itinerary recommendation based on user-generated GPS trajectories," in *Proc. UIC*, 2010, pp. 19–34.

[52] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: An open architecture for collaborative filtering of netnews," in *Proc. CSCW*, 1994, pp. 175–186.

[53] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. WWW*, 2001, pp. 285–295.

[54] O. Khalid, M. U. S. Khan, S. U. Khan, and A. Y. Zomaya, "Omnisuggest: A ubiquitous cloud-based context-aware recommendation system for mobile social networks," *IEEE Trans. Services Computing*, vol. 7, no. 3, pp. 401–414, 2014.

[55] L. Song, C. Tekin, and M. van der Schaar, "Online learning in large-scale contextual recommender systems," *IEEE Trans. Services Computing*, vol. 9, no. 3, pp. 433–445, 2016.

[56] J. Zhang and C. Chow, "Ticrec: A probabilistic framework to utilize temporal influence correlations for time-aware location recommendations," *IEEE Trans. Services Computing*, vol. 9, no. 4, pp. 633–646, 2016.

[57] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with GPS history data," in *Proc. WWW*, 2010, pp. 1029–1038.

[58] Y. Zheng, X. Xie, and W. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

[59] J. Zhang and C. Chow, "igslr: personalized geo-social location recommendation: a kernel density estimation approach," in *Proc. ACM SIGSPATIAL*, 2013, pp. 324–333.

[60] G. Kang, M. Tang, J. Liu, X. F. Liu, and B. Cao, "Diversifying web service recommendation results via exploring service usage history," *IEEE Trans. Services Computing*, vol. 9, no. 4, pp. 566–579, 2016.

[61] R. Raymond, T. Sugiura, and K. Tsubouchi, "Location recommendation based on location history and spatio-temporal correlations for an on-demand bus system," in *Proc. ACM SIGSPATIAL*, 2011, pp. 377–380.

[62] M. Deshpande and G. Karypis, "Item-based top-N recommendation algorithms," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143–177, 2004.

[63] H. Huang and G. Gartner, "Using trajectories for collaborative filtering-based POI recommendation," *IJDMMM*, vol. 6, no. 4, pp. 333–346, 2014.

[64] Y. Shi, P. Serdyukov, A. Hanjalic, and M. Larson, "Personalized landmark recommendation based on geotags from photo sharing sites," in *Proc. ICWSM*, 2011.

[65] L. Wei, Y. Zheng, and W. Peng, "Constructing popular routes from uncertain trajectories," in *Proc. ACM SIGKDD*, 2012, pp. 195–203.

[66] C. Zhang and K. Wang, "POI recommendation through cross-region collaborative filtering," *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 369–387, 2016.

[67] L. Yao, Q. Z. Sheng, A. H. H. Ngu, J. Yu, and A. Segev, "Unified collaborative and content-based web service recommendation," *IEEE Trans. Services Computing*, vol. 8, no. 3, pp. 453–466, 2015.

[68] Y. Hu, Q. Peng, X. Hu, and R. Yang, "Time aware and data sparsity tolerant web service recommendation based on improved collaborative filtering," *IEEE Trans. Services Computing*, vol. 8, no. 5, pp. 782–794, 2015.

[69] H. Sun, Z. Zheng, J. Chen, and M. R. Lyu, "Personalized web service recommendation via normal recovery collaborative filtering," *IEEE Trans. Services Computing*, vol. 6, no. 4, pp. 573–579, 2013.

[70] J. Sang, T. Mei, and C. Xu, "Activity sensor: Check-in usage mining for local recommendation," *ACM TIST*, vol. 6, no. 3, pp. 41:1–41:24, 2015.

[71] J. Sang, Q. Fang, and C. Xu, "Exploiting social-mobile information for location visualization," *ACM TIST*, vol. 8, no. 3, pp. 39:1–39:19, 2017.

[72] J. Sang, C. Xu, and R. Jain, "Social multimedia ming: From special to general," in *Proc. IEEE ISM*, 2016, pp. 481–485.

[73] X. Chen, Z. Zheng, X. Liu, Z. Huang, and H. Sun, "Personalized qos-aware web service recommendation and visualization," *IEEE Trans. Services Computing*, vol. 6, no. 1, pp. 35–47, 2013.

[74] S. Wang, Z. Zheng, Z. Wu, M. R. Lyu, and F. Yang, "Reputation measurement and malicious feedback rating prevention in web service recommendation systems," *IEEE Trans. Services Computing*, vol. 8, no. 5, pp. 755–767, 2015.

[75] Y. Zhong, Y. Fan, K. Huang, W. Tan, and J. Zhang, "Time-aware service recommendation for mashup creation," *IEEE Trans. Services Computing*, vol. 8, no. 3, pp. 356–368, 2015.

[76] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da Silva Torres, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognition*, vol. 47, no. 2, pp. 705–720, 2014.

[77] X. Qian, Y. Zhao, and J. Han, "Image location estimation by salient region matching," *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 4348–4358, 2015.

[78] R. W. Sinnott, "Virtues of the haversine," *Sky & Telescope*, vol. 68, no. 2, article 159, p. 158, 1984.

[79] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, Jun 2015.

Guoshuai Zhao received the B.E. degree from Heilongjiang University, Harbin, China, in 2012, the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2015, and is currently working toward the Ph.D. degree at Xi'an Jiaotong University, Xi'an, China.

He was a full-time intern at Social Computing Group, Microsoft Research Asia under the supervision of Dr. Ruihua Song from Jan. 2017 to Jul. 2017. He has published several research papers on journals such as IEEE TKDE, TMM, TBD, and conferences such as IEEE BigMM, MMM. His current research interests include recommender systems and social media big data analysis.

Tianlei Liu Tianlei Liu received B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2017, and is expected to get his M.S. degree from Columbia University in 2018.

His current research interests include social media big data analysis and Quantitative Finance.

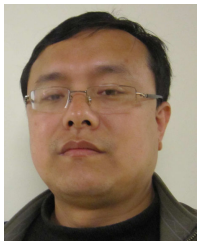
Xueming Qian (M'09) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2008.

He was a Visiting Scholar with Microsoft Research Asia, Beijing, China, from 2010 to 2011. He was previously an Assistant Professor at Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory, Xi'an Jiaotong University. His research interests include social media big data mining and search.

Prof. Qian was the recipient of the Microsoft Fellowship in 2006. He was the recipient of the Outstanding Doctoral Dissertations Awards from Xi'an Jiaotong University and Shaanxi Province in 2010 and 2011, respectively. His research is supported by the National Natural Science Foundation of China, Microsoft Research, and the Ministry of Science and Technology.

Tao Hou received the B.S. degree from Xi'an University of Posts & Telecommunications in 2009. Now he is an engineer at Shaanxi Provincial Lightning Protection Center.

Huan Wang received the B.S. in Xi'an University of Technology, in 2004, and M.S. degrees in Xi'an Jiaotong University in 2010 respectively, and is currently working toward the Ph.D. degree at Xi'an Jiaotong University, Xi'an, China. Her research interests include video/image coding, communication and transmission.



Xingsong Hou received the B.S. degree in electronic engineering from North China Institute of Technology, Taiyuan, China, in 1995, and the M.S. degree and Ph.D. degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China in 2000 and 2005, respectively. From 1995 to 1997, he was an Engineer with the Xi'an Electronic Engineering Institute in the field of radar signal processing. Now he is a professor of the School of Electronics and Information Engineering, Xi'an Jiaotong University. His research interests include video/image coding, wavelet analysis, sparse representation, sparse representation and compressive sensing, and radar signal processing.

Zhetao Li was born in Hunan province, P.R.China. He received the B.Eng. degree in Electrical Information Engineering from Xiangtan University in 2002, the M.Eng. degree in Pattern Recognition and Intelligent System from Beihang University in 2005, and the Ph.D. degree in Computer Application Technology from Hunan University in 2010. Dr. Li is a professor in College of Information Engineering, Xiangtan University. From Dec 2013 to Dec 2014, he was a post-doc in wireless network at Stony Brook University. From Dec 2014 to Dec 2015, he was an invited professor at Ajou University. For his successes in teaching and research he received the Second Prize of Fok Ying Tung Education Foundation Fourteenth Young Teachers Award in 2014. His research interests includ wireless communication and multimedia signal processing.