Finding Optimal Meteorological Observation Locations by Multi-Source Urban Big Data Analysis

Tianlei Liu*‡, Guoshuai Zhao†‡, Huan Wang†, Xingsong Hou†, Xueming Qian†§, and Tao Hou†¶
*{School of Mathematics and Statistics}, †{School of Electronic and Information Engineering},
Xi'an Jiaotong University, Xi'an, China
¶Shaanxi Provincial Lightning Protection Center, Xi'an, China
tianleiliu2015@gmail.com; {zgs2012@stu., wang.huan.1006@stu., houxs@mail., qianxm@mail.}xjtu.edu.cn;
263346028@qq.com

***Abstract*—In this paper, we try to solve site selection problem for building meteorological observation stations by recommending some locations. The functions of these stations are meteorological observation and prediction in regions without these. Thus in this paper two specific problems are solved. One is how to predict the meteorology in the regions without stations by using known meteorological data of other regions. The other is how to select the best locations to set up new observation stations. We design an extensible two-stage framework for the station placing including prediction model and selection model. It is very convenient for executives to add more real-life factors into our model. We consider not only selecting the locations that can provide the most accuracy predicted data but also how to minimize the cost of building new observation stations. We evaluate the proposed approach using the real meteorological data of Shaanxi province. Experiment results show the better performance of our model than existing commonly used methods.**

***Index Terms*—location recommendation; site selection; urban big data**

I. Introduction

In recent years, 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 any more. 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. In this paper, we mainly try to answer a practical question: Given a set of existing stations' historical observation data, how to select a small amount of 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 selected locations. Second, we prefer that the selected locations are homodisperse in map. Otherwise, the result of selection may be a set of locations concentrated together which is obviously not proper. So our main task is to select locations that can make prediction accurate, construct new observation stations with low cost, and cover more regions.

In this paper, we propose a two-stage framework. Figure 1 is the overview of our framework. According to different personalized requirements, the multi-source data and multi-factors are taken into consideration. By training our selection model and prediction model, the scores of different locations are learned which denotes the importance of locations when we select locations. Then the rank of candidate locations is obtained. We exploit a Least Squares Method based model to learn the scores of different locations. The main contributions of this paper are:

- We solve the problem of how to select the locations to construct new meteorological observation stations by multi-source urban big data analysis, including meteorological data, geographical location data and benchmark price of industrial land.
- Besides using the traditional least square methods to constrain the prediction error, geographical location data is explored to improve the prediction accuracy in our prediction model because of the assumption that the more close two locations are, the more similar their meteorological data becomes. In the selection model, we would like to select the locations that can cover more geo-spatial areas considering with the dispersity.
- Besides the factor of geographical location, in the selection 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 model to learn the importance of locations in order to meet the personalized needs of decision-makers.

The rest of this paper is organized as follows: In section 2, we present some related works on sample selection and environmental prediction. In section 3, our model proposed in this paper is described in detail. The experiments are introduced and the evaluation of our model is given in Section 4. Finally, we conclude the paper in Section 5.

‡The first two authors contributed equally to this paper
§Corresponding author

978-1-5090-3555-7/16 $31.00 © 2016 IEEE
DOI 10.1109/CCBD.2016.52

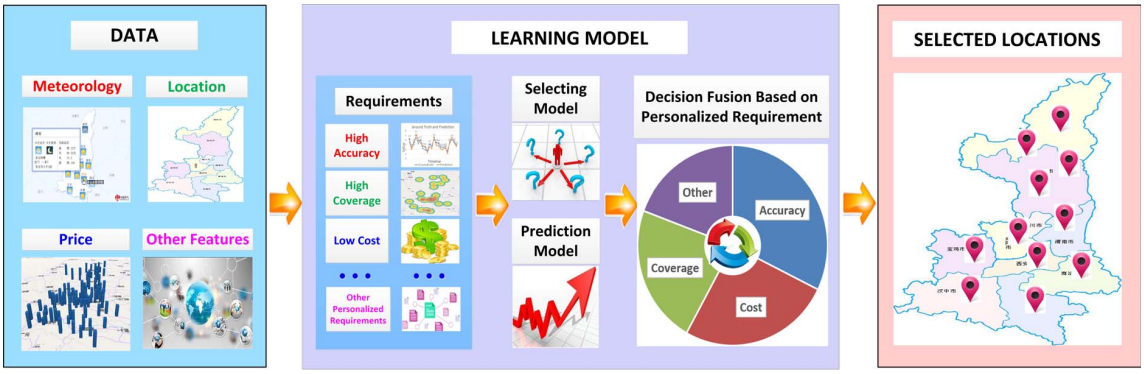

Fig. 1. The overview of our framework. According to different personalized requirements, the multi-data and multi-factors are taken into condideration. By training our selection model and prediction model, the scores of different locations are learned. Then the scores of locations are calculated and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

II. Related Work

The research is belong to urban computing [1]. Here, we review some related works on site selection.

Many selective sampling problems were solved based on information entropy theory [2], [3], [4], [5] . Hsieh *et al* [2] established new stations at the locations that can minimize the uncertainty of the prediction model. To begin with, they picked the location with lowest entropy and then put it into the prediction model as known data. Second, they picked the second-to-last location which is the location with lowest entropy in the new prediction model and keep running this circle. Finally, they selected the top k ranked locations as the location to construct new stations. Du *et al*. [6] 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*. [7] aimed to deploy sensors in an information delivery network to optimize the detection of duplicate data contents. Wang *et al*. [8] 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.

There are a lot of ways of predicted data based on different theories such as matrix factorization [9], [10], [11], [12], [13], [14], probability, cluster and similarity etc. Zheng *et al*. [15] 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*. [16] 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.

III. Our Model

As shown in Figure 1, our task is training our selection model and prediction model, and rank the locations according to the learned scores.

We divide the whole geo-spatial 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 selected 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 selection model.

TABLE I
Notations and Their Descriptions

Notations	Descriptions
A	The matrix of meteorological correlation between any two locations in prediction model
B	The matrix of geo-distance correlation between any two locations
C	The matrix of meteorological correlation between any two locations in selection model
D	The importance matrix of coverage area
E	The importance matrix of benchmark price
G	The matrix of geo-distance between any 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
T	The matrix of total distance between any location to other locations

A. Prediction Model

The observation data of meteorology in different locations are correlated with each other in spatial perspectives. Consider 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 selected locations to predict the data in un-selected locations. We propose our initial prediction model given by:

$$\min_{\mathbf{A}} \|\mathbf{R} - \mathbf{AS}\|_F + \alpha \|\mathbf{A}\|_F \quad (1)$$

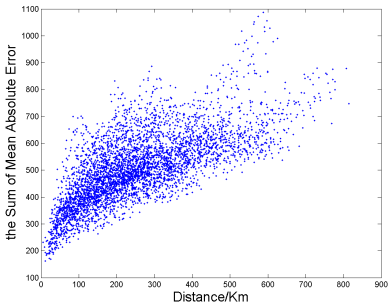

Fig. 2. The relevance between geographical distance and the difference of observation data.

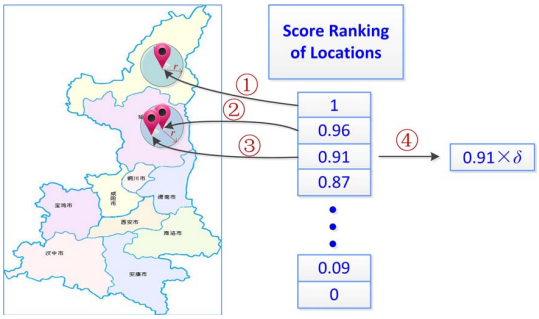

Fig. 3. Score ranking with considering the coverage of selected locations by the following steps: ① the first location which is the geographical center is mapped; ② map the second location, if the distance between it and the previous locations are larger than the predefined parameter r; ③ if the distance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

where **S** is the matrix of meteorological data in selected locations. Matrix **A** consists of coefficient a_{pk} which represents the correlation of meteorological data between the selected location p and the un-selected location k. **AS** is the 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. In Figure 2, x-axis represents the distance between regions and y-axis represents 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 put the constrain term matrix B into Equation 1. The objective function of prediction model is given by:

$$\min_{\boldsymbol{A,B}} \|\mathbf{R} - (\mathbf{A+B})\mathbf{S}\|_F + \alpha_1 \|\mathbf{G}_{min} - \mathbf{BG}\|_2 + \alpha_2 \|\mathbf{A} + \mathbf{B}\|_F \tag{2}$$

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. Matrix **B** consists of coefficient b_{pk} which represents the correlation of geo-distance between p and k. **R** is the matrix of meteorological data. In the second term **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.

B. Selection Model

Based on the above prediction model, we would like to select the locations that can help us to make more accurate prediction to build observation stations. Hence, how to select the location of these stations is an urgent issue for us now. First, considering with the accuracy of prediction, we ought to fuse the errors of prediction into our selection model. The linear combination of each location's record data $\mathbf{R}_p$ is also used to calculate the prediction. But only m of the most important locations can be selected. We use the weight of each $\mathbf{R}_p$ to represent the importance of location p. The more important the location is, The more correlation with others the location has.

Second, we select the locations those can cover most geospatial areas in map in order to make sure every location in our province will not leave the selected locations too far. It can help us to predict the more accurate meteorological data which can be proved by Figure 2. Nevertheless, coverage area is a definition that cannot be clearly measured, so we propose to employ the total distance between one location and other locations. In case that the selected locations are concentrated together, we suggest to apply the dispersity to measure the model. The process of Score ranking with considering the coverage of selected locations are given in Figure 3.

Third, we should not only consider the accuracy the selected locations can provide for the prediction model, but also regard the cost of building new stations. When the decision makers are facing this kind of selection problems, they also need to minimize the cost in the whole project. Therefore, we need to fuse the factor of cost into our model and, in brevity, we utilize the benchmark price of industrial land to represent the cost. Then our final model is given by:

$$\min_{\boldsymbol{C,D,E}} \|\mathbf{R} - \mathbf{CR}\|_F + \beta_1 \|\mathbf{T}_{max} - \mathbf{DT}\|_2 + \beta_2 \|\mathbf{P}_{min} - \mathbf{EP}\|_2 + \beta_3 (\|\mathbf{C}\|_F + \|\mathbf{D}\|_F + \|\mathbf{E}\|_F) \tag{3}$$

$$\mathbf{D} = \mathbf{W}\delta \tag{4}$$

In which **CR** represents the prediction of the meteorological data by the linear combination of the other regions' observation data. Hence, the first term in the objective function represents the square error between the real observation data and the prediction. The purpose of the first item is to constrain parameter c. The parameter $c_{\cdot i}(c_{\cdot i} = \sum_p c_{pi})$ represents the importance of the location i.

The purpose of the second term is to constrain the coverage area. Note that the value of parameter $d_{\cdot i}(d_{\cdot i} = \sum_p d_{pi})$

means the importance of the location *i* in regard of coverage. $w_{\cdot i}(w_{\cdot i} = \sum_p w_{pi})$ means the importance of the region in terms of total distance between location *i* and others. δ_i is 0.5 if the location situates at the border or belongs to the r-radius circle of previous locations. It is 1 otherwise. δ_i is the discriminant coefficient as shown in Figure 3. The total distance and the dispersity approach are leveraged together to describe the coverage. For the locations at the edge of the map, we leverage the concept of relative area [17], [18] to remove them as follows. Firstly, establish the coordinate system for each location. Secondly, record the amount of the locations in each quadrant of each location. At last, we use the four numbers to describe the relative area of each location and if one of the four numbers is 0 means it locates at the edge.

The third term is used to constrain parameter *e* and the value of parameter $e_{\cdot i}(e_{\cdot i} = \sum_p e_{pi})$ denotes the importance of the location *i* in terms of the cost. **EP** is optimized to $\mathbf{P}_{min}$, which means the bigger the value of $e_{\cdot i}$ is, the cheaper the cost is, and meanwhile the more important the location is.

The fourth term is used to avoid over-fitting. Finally, we have three essential parts in our selection 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 value of $c_{\cdot i} + d_{\cdot i} + e_{\cdot i}$ are figured out and the corresponding regions are the locations we seek.

C. Model Training

Given the proposed prediction model and selection model, the objective functions represented in Equation (2) and (3) can be minimized by the gradient decent approach as in [9], [11], [10]. Algorithm 1 summarizes the whole procedure of our framework. Steps 1 to 8 show the details of our selection model. Steps 9 to 16 show the details of our prediction model. The space complexity of this algorithm is $O(n \times k + 4n^2 + 4n)$, and the time complexity is $O(t_1 \times n^2 \times k + t_2 \times n \times m \times k)$, where n is the number of regions. m is the number of selected locations. k is the dimension of the meteorological data. Generally, because of $k \gg n, m$, the space complexity is $O(n \times k)$.

Algorithm 1 The Procedure of Our Framework

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 (3) with respect to the variables **C**, **D**, and **E** respectively.
4: Update matrices **C**, **D**, and **E** with the gradients by the learning rate l_1.
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 (2) with respect to the variables **A** and **B** respectively.
12: Update matrices **A** and **B** with the gradients by the learning rate l_2.
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.

IV. Experiment

In this section, we will introduce the experiments in detail. The definition of the problem is that there are 22 meteorological stations to be built in Shaanxi Province, China, and how to select the locations.

A. Dataset Introduction

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.

2) Geographical Location Data: The 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 [19]. We crawled the geographical location data of each county from the Internet.

3) Benchmark Price of Industrial Land: The benchmark price of industrial land released in 2010 is 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.

B. 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 systems [2], [16], [10] [9]. RMSE and MAE are defined as:

$$RMSE = \frac{\|\mathbf{R}_{test} - (\mathbf{A}+\mathbf{B})\mathbf{S}_{test}\|_F}{|\mathbf{R}_{test}|} \quad (5)$$

$$MAE = \frac{\|\mathbf{R}_{test} - (\mathbf{A}+\mathbf{B})\mathbf{S}_{test}\|_1}{|\mathbf{R}_{test}|} \quad (6)$$

where $\mathbf{R}_{test}$ is the real meteorological data. $\mathbf{A}$ and $\mathbf{B}$ are the matrices learned by Equation 2. $\mathbf{S}_{test}$ is the real meteorology data in selected locations. $|\mathbf{R}_{test}|$ denotes the number of data in the test set.

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

In fact, the dispersity of selected locations are also important, which has been illustrated in Figure 2. Therefore, we employ 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-selected locations, and m is the number of selected 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 model, and the lower, the better.

C. Evaluation

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). It is simply used as a classification model for meteorological data prediction.

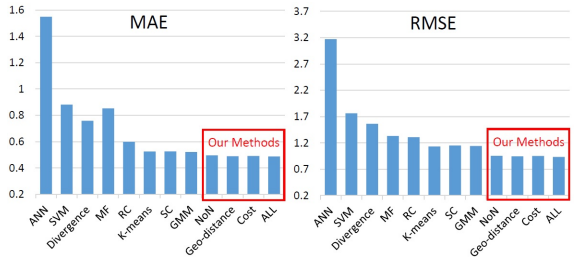

Fig. 4. Prediction performance comparison of different algorithms based on RMSE. In addition, the methods in the red box are ours.

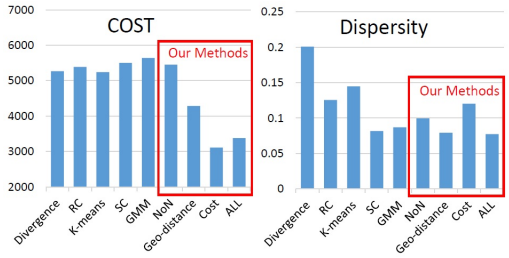

Fig. 5. Selection performance comparison of different algorithms based on MAE. In addition, the methods in the red box are ours.

- *Support Vector Machine* (SVM) is one of the most popular supervised learning models that 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 system.

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

- NoN, which denotes the appraoch 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.
- ALL, which denotes the approach with fusing all proposed factors.

2) Performance Comparison: Figures 4 and 5 show the performance comparison of different algorithms based RMSE, MAE, COST, and Dispersity. It can be seen that our approaches are mostly better than 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 model are all effective. When we only consider the factor of cost, the performance of our model on COST is much better than other algorithms. When we only take the factor of geo-distance, our model also reaches the best performance on Dispersity. If we combine the two factors, our model (ALL) achieves the optimal solution with balancing the two factors. Then decision makers can adjust the model according to their personalized requirements.

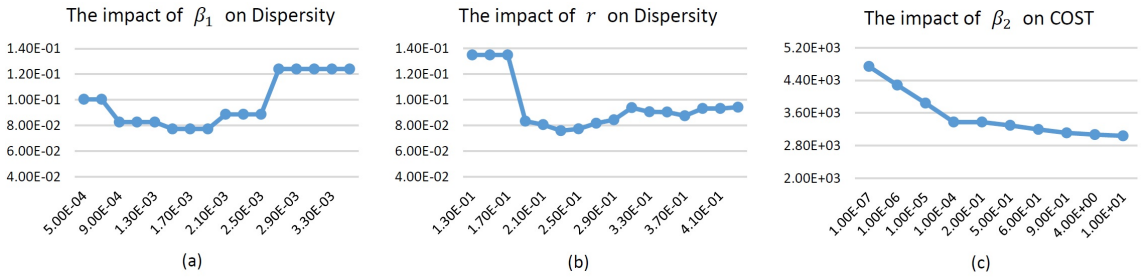

Fig. 6. Discussion on the impact of parameters on performance of our model.

3) Discussion: There are some parameters to balance the fused factors. In Equation (3), the parameter β_1 is the weight of the importance of geographical Dispersity. In Figure 3. The parameter r is used to avoid the concentration of selected 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 6(a) and (b) show the impact of β_1 and r on performance. It can be seen that our model 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 model, the parameter of β_2 is set to manage the degree of importance of cost. Figure 6(c) demonstrates the effect of β_2 on performance of our model in light of the cost of establishing new stations. Apparently, our model offers various solutions according to different requirements of cost.

V. Conclusions

In this paper, we introduced a framework to recommend locations for solving the problem of site selection, in which the factors of geographical location and benchmark price of industrial land are taken into account. It is employed to solve the practical optimization problem and provide the solution with more intellegence. The weights of different factors can be fine tuned according to the personalized requirements.

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.

Acknowledgment

This work was supported in part by Program 973 under Grant 2012CB316400, in part by Program of Guangdong Science and Technology under Grant 2016A010101005, in part by the National Science Foundation of China under Grant 60903121, Grant 61173109, and Grant 61332018, and in part by Microsoft Research Asia.

References

[1] 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.

[2] 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.

[3] B. Settles, "Active learning literature survey," *University of Wisconsin-madison*, vol. 39, no. 2, pp. 127–131, 2010.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] G. Zhao, X. Qian, and C. Kang, "Service rating prediction by exploring social mobile users' geographic locations," *IEEE Transactions on Big Data*, 2016.

[13] P. Lou, G. Zhao, X. Qian, H. Wang, and X. Hou, "Schedule a rich sentimental travel via sentimental POI mining and recommendation," in *IEEE Second International Conference on Multimedia Big Data, BigMM 2016, Taipei, Taiwan, April 20-22, 2016*, 2016, pp. 33–40.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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