Air Quality Forecast based on topological Potential in Mining Areas

Yi Lin*

School of Computer, Wuhan University, Wuhan, Hubei, 430079, P.R. China

Abstract
Air quality forecast is very important for production and life in mining area. This paper provides a novel method based on the theory of data field, which give the reasonable real-time air quality forecast in sense of uncertainty. The basic idea is that by using the topological potential distribution to model the interaction of air quality in adjacent region, and calculating the value of topological potential which corresponding to minimum potential entropy, the real-time air quality at any position can be predicted. The experiments from the Wuhan environment publishing system shows that this algorithm can discover the intrinsic relationship of air quality between different monitoring stations without extra parameters and background.
Keywords: Air quality forecast, Data Field, Topological Potential, Entropy, Data Mining

1. Introduction

Mining involves a wide range of human survival, from providing all kinds of necessary fuel, metals and minerals, which is of great significance. But, in fact, the increase production and environment protection did not coordinate. The promulgation of strict laws concerning environment protection makes it become the most difficult problem in face of mining industry. Taking effective and practical remedial measures to monitor and reduce the air pollution of mining industry is no delay.

Open pit mining has great effect on air quality in mines and its surroundings. Controlling the air quality is not only conductive to the long-term development of mining industry, but also for nearby residents, especially children [1][2] [3] [4-6]. Many air quality models have been based on physics studies, and generally involve the research of topological relationship. In this paper, we proposed a method of air quality index (AQI) forecast in sense of uncertainty, which is based on the topological potential of data field, which is reasonable and relatively simple.

At present, there are three categories of prediction techniques and skills, including simple empirical approaches, parametric or non-parametric statistical approaches and physically-based approaches[7]. Many of them have been discussed in literature [8-11]. The υ-support vector regression approach[12] is considered a good choice because it more offers a parameter $\upsilon \in (0,1]$ to control the number of support vectors compared to ϵ-support vector regression[13], which plays a very important role in this study.

Intuitively, given a set of AQI value of monitoring stations at some point, each of them is equivalent to a particle with certain quality in air space, which exists around an interaction field, any monitoring station in the field will receive the combined effects of other monitoring stations, thus it determined a data field on the entire space. According to [14], Minimizing potential entropy of total data field means to minimize the uncertainty of system and gives the most reasonable potential field distribution.

This paper follows an experimental method, and its organization unfolds as the core problems involved in the algorithm steps. Section 2 provides an essential introduction to data field theory. Section 3 describes the data on WHEPB[15]. Section 4 elaborates how to use the data field theory to estimate the distribution of air quality. Section 5 explores the approach how the value of potential converted to AQI. Section 6

2. Important Concepts

2.1. Data Field Theory

In the world of physical, whether universal gravitation, static electricity of Charged Body, or mag-

netic force between the magnets, even nucleus force, it must be achieved by transferring to some intermediary, namely field. Modern physics even believe field is one of the basic forms, and any physical particles cannot be stand-alone without related field[16]. A field is a physical quantity that has a value for each point in space and time[17].

Inspired by the development of field theory, data field theory[18] introduces the interaction between particles and its description method into the abstract data space. The following gives a very brief introduction to data field theory. Given a data set $D=\{x_1,x_2,...x_n\}$ with n objects in space χ, where $x_i=(x_{i1},x_{i2},...,x_{ip})',(i=1,2,...,n)$ and then each object is equivalent to a partial or nucleon with a certain quality in the p-dimensional space. According to whether the value of field at each point is a scalar, a vector or a tensor, a field can be classified as a scalar field, a vector field or a tensor field respectively. In whole space, there is a data field and any object in the field receives the interaction of the other objects. With reference to the physics field, the data field of static data, which does not depend on time, can be viewed as a stable active field. Due to the simplicity and vivacity of scalar computation, the scalar potential function is selected to describe the properties of data field.

According to the characteristics of potential function of stable sourced field in physics, basic criterion of potential function morphology of data field can be given as follows[16, 19].

[Theorem 1]Given the data object x in space Ω, $\forall y\in\Omega$, $\varphi_x(y)$ is the potential value of object x generated at point y, it obeys:

(1) $\varphi_x(y)$ is the continuous, smooth and finite function defined on space Ω;

(2) $\varphi_x(y)$ has the properties of isotropy;

(3) $\varphi_x(y)$ is the monodromic decreasing function of distance $\|x-y\|$. When $\|x-y\|=0$, $\varphi_x(y)$ reaches the maximum value, but not infinite; while $\|x-y\|\to\infty$, $\varphi_x(y)\to 0$.

In principle, the function morphology satisfied above conditions can be used to define the potential function of data field. Reference to the potential function formulas of gravitational field

$$\varphi_x(y)=\frac{G\times m}{\|r\|}$$

and nuclear field $\varphi(r)=V_0\cdot e^{-\left(\frac{\|r\|}{R}\right)^2}$, it can draw two alternative potential function morphology:

Quasi gravity field potential functions:

$$\varphi_x(y)=\frac{m}{1+\left(\frac{\|x-y\|}{\sigma}\right)^k} \tag{1}$$

Quasi nuclear field potential functions:

$$\varphi_x(y)=m\times e^{-\left(\frac{\|x-y\|}{\sigma}\right)^k} \tag{2}$$

Where, $m\geq 0$ represents the field source strength, which can be regarded as the quality of the data object; $\sigma\in(0,\infty)$ used to control the interaction range between objects, called impact factor; $k\in N$ is the distance index.

2.2. Potential

In classical electromagnetism, the electric potential at a point is the amount of electric potential energy that a unitary point charge would have when located at that point[20][19][20][20], which is numerically equal to move this unitary point from somewhere to reference point. The distribution of potential field is defined by the relative position between the interacting particles and independent of existence of particles. The potential function is usually regarded as a monodromic function of spatial location.

[Definition 1] If data set $D=\{x_1,x_2,...x_n\}$ contains n objects in the space $\Omega\subseteq R^P$, then the potential value of data field of any point $x\in\Omega$ in space can be expressed as

$$\varphi(x)=\varphi_D(x)=\begin{cases}\sum_{i=1}^{n}m_i\times\varphi_{gi}(x-x_i)\\ \sum_{i=1}^{n}m_i\times\varphi_{ni}(x-x_i)\end{cases} \tag{3}$$

Where $\varphi_{gi}(x-x_i)$ is the potential value for x generated by quasi gravity field potential function, $\varphi_{ni}(x-x_i)$ corresponds to the potential value of quasi nuclear force field for object x, $m_i\geq 0$ is the mass of object x_i, the hypothesis obeys the normalization condition $\sum_{i=1}^{n}m_i=1$ [18].

2.3. Zero Distance Case

It can draw two alternative potential function morphology: It can easily calculate the potential value of any location in space with the formula (1) (2) (3). However, it is needs to note the potential value of data field at the particle location. Because $\|x-x\|=0$, it needs function limit to deal with this zero distance case. The potential value of two different data field generated by object x at point x include following,

Quasi gravity field:

$$\varphi_x(y)=\lim_{\substack{\|x-y\|\to 0\\ m,k\geq 0}}\frac{m}{1+\left(\frac{\|x-y\|}{\sigma}\right)^k}=m \tag{4}$$

Quasi nuclear field:

$$\varphi_x(y) = \lim_{\substack{\|x-y\|\to 0 \\ m,k\geq 0}} m * e^{-\left(\frac{\|x-y\|}{\sigma}\right)^k} = m$$

It is a coincidence that these two different data field function give the same limit result with $\|x-y\| \to 0$.

3. The Data and its Representation

The quality of air reflected the degree of air pollution, which is determined by the level of pollutant concentration in air. The main pollutants in air quality evaluation includes ozone (O_2), particulate pollutants, PM10 (particle size less than or equal 10 μm), particulate matter PM2.5 (particle size less than equal to 2.5 μm) of carbon monoxide, carbon dioxide (NO_2), sulfur dioxide (SO_2) and etc. The air quality index (AQI) is a popular used number by national government agencies, which transforms the concentrations of above pollutants into a single conceptual numerical morphology. For above reasons, our first choice is the air quality index which is described presently.

3.1. Air Quality Index

In spite of the fact the evaluation through air quality index is a very simple approach; it has been widely recognized in the worldwide especially for city's short-term air quality status and trends, where AQI generation has to be as simple and as fast as possible. Information of specific ingredients is lost, meaning that specific concentration of ingredients cannot be used. This also ensures full utilization and practical convenience.

The calculating of air quality index can be divided into two parts: First, reference to diverse-level limit of each pollutant, the individual air quality indexes (IAQI) was calculated according to the measured concentration of each pollutant. It is calculated as shown in equation (6)[21]. The second step is to select the maximum value determined as AQI from various pollutants $IAQI$, namely equation (7). When the AQI is greater than 50, the pollutant with maximum $IAQI$ is the primary pollutant. If there is two or more pollutant with maximum $IAQI$, they are tied for the primary pollutants. The pollutant with $IAQI$ greater than 100 is standard exceeded pollutants.

$$IAQI_P = \frac{IAQI_{Hi}}{BP_{Hi} - BP_{Lo}}\left(C_p - BP_{Lo}\right) + IAQI_{Lo} \qquad (6)$$

Where, $IAQI_P$ - individual air quality index of pollutants p

C_p - concentration of pollutants p

BP_{Hj} - high-value of pollutant concentration limit similar to C_p

BP_{Lj} - low-value of pollutant concentration limit similar to C_p

$IAQI_{Hi}$ - individual air quality index corresponding to $IAQI_{Lo}$

$IAQI_{Lo}$ - individual air quality index corresponding to BP_{Lo}

$$AQI = \max\left\{IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n\right\} \qquad (7)$$

Where, $IAQI$ - individual air quality index

n – pollutant number

In accordance with air quality index, air quality is divided into five levels, corresponding to six categories of air quality, as shown in table 1. The larger index, the higher level means the more serious pollution and greater hazards on human.

Table 1. Six categories of air quality

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	*..air quality conditions are:*	*...as symbolized by this color:*
0-50	Good	Green
51-100	Moderate	Yellow
101-150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

3.2. Getting data from WHEPB

The data were hand-picked from Wuhan environment protection bureau system(WHEPB)[15]. It provides the real-time air quality index of ten national air quality monitoring stations in Wuhan city. The monitoring stations includes Hankou Huaqiao, Wuchang Ziyang, Zhuankou district, Wujiashan, East Lake Liyuan, Hanyang Moon Lake, East Lake Gaoxin, Qingshan Ganghua, Chenhu Qihao(control point). The data was published the real-time report of each monitoring station on the hour and lagged not more than 1 hour.

4. Selecting the Potential Distribution

4.1. Data Field based Forecasting

Inspired by the idea of data field, we introduce the interaction between the data and its field description method into realistic air space and try to use the field knowledge of data field to estimate the air quality between different monitoring stations. Each monitoring station was equivalent to a particle or a nucleon with a certain quality and the value of AQI correspond to the mass of a particle at that time. Any monitoring station will be subject to the combined effect of other

monitoring stations, thereby a stable and active data field was determined in space.

As shown in figure 1, it was the equipotential lines maps of data field with different potential functions form and parameters, which was generated according to the air quality of Wuhan's 9 monitoring stations at a moment. It can be found that, except the gravity field function when k = 1, as long as the appropriate impact factor was selected, the equipotential line distribution of potential field with different field function morphology is almost the same.

Though the idea of real-time air quality estimation using data field is very simple, which is according to the size of potential value, nevertheless, it has to be selected carefully since an inappropriate potential distribution of data field can lead to poor estimation. There are not techniques available to "learn" the optimal distribution of data field potential now.

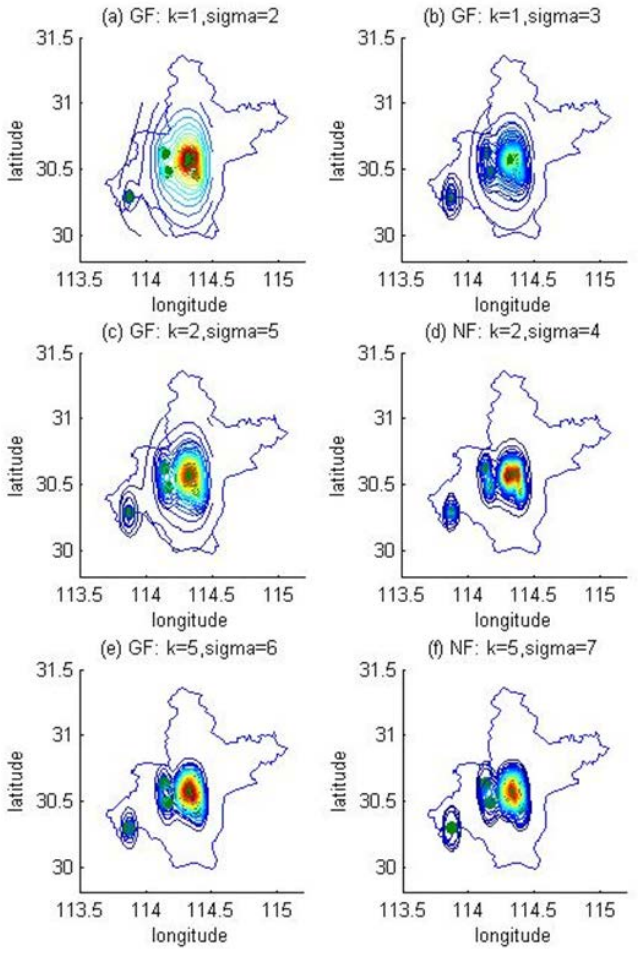

Figure 1. Potential field distribution with various function form and parameters

Where,(a) quasi gravity field(k =1, σ =2); (b) quasi nuclear field(k =1, σ =3);(c) quasi gravity field (k =2, σ =5); (d) quasi nuclear field(k =2, σ =4);(e) quasi gravity field(k =5, σ =6); (f) quasi gravity field(k =5, σ =7);

4.2. Crucial Factor

Although the best distribution of data field cannot be "learned", we can find the decisive factor of the potential distribution of data field. In the case of quasi nuclear data field generated by single object, influence radius satisfy which describe the interaction range between objects.

$$R = \sigma \times \sqrt[k]{\frac{9}{2}} \left(k \in N\right) \tag{8}$$

For $\forall k \in N$, always meets $\sigma < R \le \frac{9}{2}\sigma$. Two similar potential fields have the approximately equal radius[18].

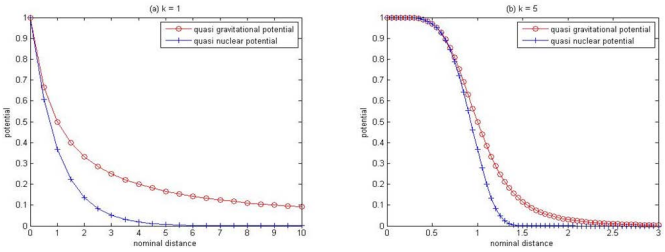

Figure 2. Comparison of potential function morphology

Figure 2 shows curve features of two kinds of potential function morphology. It can be found in figure (a), the quasi gravity field function, which can be considered as long range field, decays very slowly with the growth of distance at $k = 1$; at $k = 5$, the quasi gravity field function and quasi nuclear field function will quickly decay to zero in figure (b), which can be viewed as short range field. In the case of air quality, selection of potential function represented short range field can better describe the interaction between different regional airs. Meanwhile, taking into account the good mathematical properties and universality of Gaussian function, the quasi nuclear field function with $k = 2$ is used to describe the character of data field in air quality estimation.

Because the distribution of data field mainly depends on the interaction force range between objects, on the monitoring of formula (8) and $k = 2$,there is critical relationship between the space distribution of data field and impact factor σ . The figure3 reveals that different values of impact factor will result in three typical cases roughly: first, when sigma is small, the interaction force range is short. $\varphi(x)$ is equivalent to the superposition of n data-object-centered unimodal functions. The potential around each object is very small. In the extreme cases, there is no interaction among objects, the potential value of each object's location is $1/n$. Conversely, if the value of σ is large, there is strong interaction among objects. $\varphi(x)$ becomes the superposition of n basic functions with slow transformation and large width. Each object has

relatively large potential value around itself. In the extreme case, the potential value for each object's location is approximately equal to 1. Obviously, the potential distribution under these two extreme cases can not reflect the intrinsic distribution. Only the third case can produce meaningful estimates.

4.3. Option of impact factor

From the figure 3, it can obtain the perceptual knowledge of rational get a reasonable potential field distribution. The numerical solution can learn from the information theory. On the monitoring of the thought of information theory, Shannon entropy[14] is a measure of uncertainty. The greater entropy means the higher uncertainty, and vice versa. For the data field generated by $x_1, x_2, ..., x_3$ objects, if the potential value of each object is equal, the uncertainty of potential distribution reaches maximum, which corresponds to the maximum Shannon entropy. Conversely, when the potential distribution of each object is asymmetrical, the uncertainty of potential distribution achieves the minimum, i.e. it has the minimization entropy.

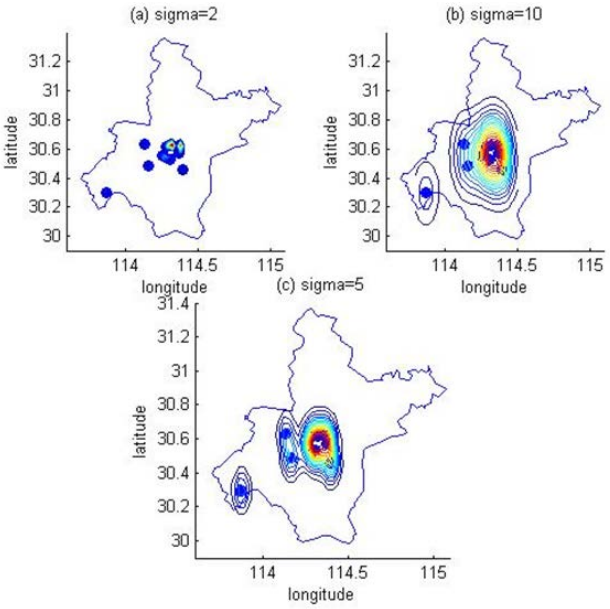

Figure 3. Three kinds of typical distribution of data field

(a) $\sigma = 1$; (b) $\sigma = 10$;(c) $\sigma = 5$.

Let $\Psi_1, \Psi_2, ..., \Psi_n$ be the potential value of objects $x_1, x_2, ..., x_n$, the potential entropy can be defined as

$$H = -\sum_{i=1}^{n} \frac{\Psi_i}{Z} \log\left(\frac{\Psi_i}{Z}\right) \qquad (9)$$

Where $Z = \sum_{i=1}^{n} \Psi_i$ is a normalization factor[18]. Understandably, the σ corresponding minimum potential entropy is the optimal impact factor.

Solving σ is essentially a single-variable nonlinear problem. There are many standard optimization algorithms can solve this problem. This paper uses the "fminbnd" function in optimization toolbox of Matlab R2011a. Figure 4 shows the distribution of data field with optimal in Wuhan City at a certain time.

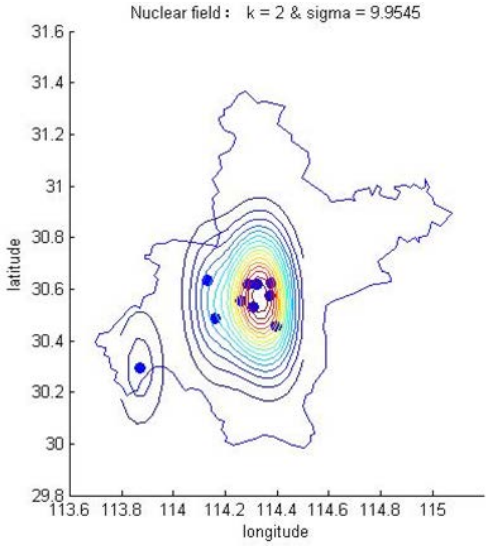

Figure 4. Optimal potential field distribution at a moment

5. Potential Value Versus Air Quality Index

The experiments performed in above section show that the spatial distribution of data potential value can lead to reasonable region division of air quality by AQI. This section explores the way to transform the potential value into air quality index. It shows that applying data field can give the remarkable real-time air quality estimation, in which case the AQI of monitoring station is needed.

5.1. Introduction

The above section has been clearly explained how to use the data field on the region division. Nevertheless, in general, the specific values of Air Quality Index (AQI) are the most concerned. The data for this purpose, displayed in Figure 5, come from the same case of figure 4 that examined the correlation between the potential value of data field and AQI at the location of monitoring station. The goal is to predict the value of AQI from potential value, which is difficult to fix by eye.

Because the outcome wanted is quantitative, this problem can be view as a regression problem in supervised learning, with one categorical predictor variable - potential value - with the response variable being the air quality index. Its purpose is to identify one best able to represent all observational data function (regression estimator), which can represent all correlation between response variable and predictor variable.

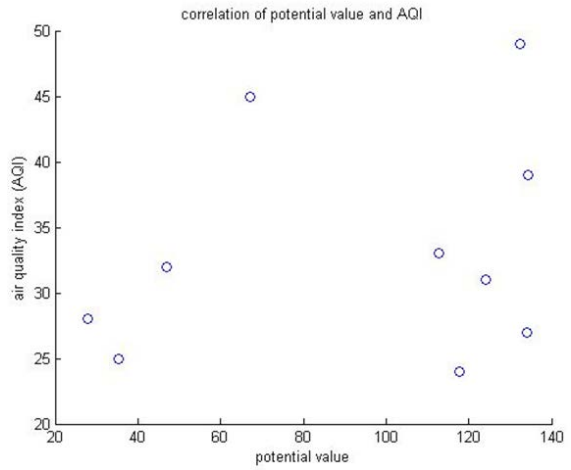

Figure 5. Scatterplot of the example in figure 4

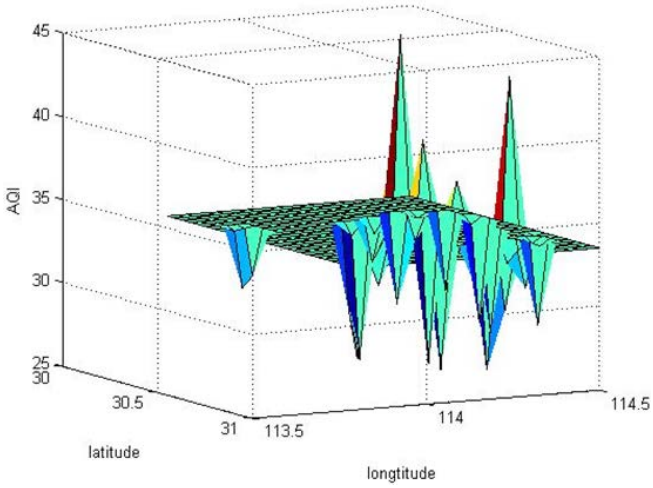

Figure 6. 3D surface of estimated AQI

5.2. Experiment

This experiment was designed to identify the reasonable AQI according to nature of actual situation. It uses the LIBSVM toolbox of Professor Chih-Jen Lin and mainly requires the choice of the type of SVR and their parameters. LIBSVM[22] is an open source machine learning libraries, which in support vector machines(SVMs), supporting classification and regression.

In order to avoid the combination explosion of the type of SVR and parameter combinations, it is important to restrict the number of experiments we try. All the experiments will be divided into two parts. The first problem is to determine the appropriate type of SVR. In LIBSVM, there are two types of support vector machines, namely, $\epsilon - \text{Support Vector Regression}(\epsilon - \text{SVR})$ and υ-Support Vector Regression$(\upsilon - \text{SVR})$. Though the $\epsilon - \text{SVR}$ model is often used, due to the less number of monitoring stations, we choose the υ–SVR model to estimate the air quality without monitoring station. It could provide a parameter $\upsilon \in (0,1]$ to control the number of support vectors. We choose the υ as 1, it means that all the points with monitoring stations are views as support vectors and each of them play a decisive role on fitting curve.

The second series of experiments attempts to find the suitable parameters. For the reasons elaborate in Figure 5, it is very easy to find a nonlinear relationship between potential value and AQI. We should first determine which type of kernel functions for nonlinear mapping. As a consequence, the radial basis function (RBF) is selected. Because the RBF kernels are one class of kernel functions without much deviation. The SVM error penalty parameter C was set to 100, which is enough for practice because the large C will hardly affect the empirical risk and have no impact on final expected risk. Figure 6 shows the estimated AQI value corresponding to figure 4.

6. Discussion

In this study, we have shown that it is feasible to use topological potential of data field to forecast the real-time air quality at that location without monitoring station. This method is simple and easy to implement without extra parameter, which can give the most reasonable AQI in the sense of uncertainty.

The results presented here demonstrate that the division by to topological potential entropy can indeed produce a very reasonable AQI predict with peaks and valleys. There are no apparent discontinuity and leaps in results except “Chenhu seven trenches” station. Taken together, this phenomenon could be due to this monitoring station located in remote areas, which fat away from other monitoring stations cluster.

Nevertheless, we show that the prediction surface is not sufficient smooth. The reason for this phenomenon may be due to the unsuitable parameters of υ–SVR or the same impact factor of data field.

In summary, we gave a novel and effective air quality forecast method based on topological potential. Although this method is not sufficient smooth, it offers a completely novel idea and give a reasonable real-time value without other extra parameter. It is very suitable for predicting the air quality in mining area and other domains in need of real–time monitoring of air environment.

7. Conclusion

In this paper, we proposed a novel and efficient air quality forecast method that is based on the theory of data field. By minimizing potential entropy, it can give the reasonable topological potential distribution in the sense of uncertainty. If can use the more information about specific mining area, such as the average yield of mining area, number of motor vehicles, the equipment and machinery used in mining, and so on, the model should be get more accurate results. This method maybe very useful for

monitoring air quality in mining area and other aspects of the automation of production processes, metallurgy and mining industry.

Conflict of Interest

The author confirms that this article content has no conflict of interest.

References

1. M. Franklin, H. Vora, E. Avol, R. McConnell, F. Lurmann, F. Liu, et al., "Predictors of intra-community variation in air quality," Journal of Exposure Science and Environmental Epidemiology, vol. 22, pp. 135-147, 2012.
2. W. J. Gauderman, H. Vora, R. McConnell, K. Berhane, F. Gilliland, D. Thomas, et al., "Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study," The Lancet, vol. 369, pp. 571-577, 2007.
3. R. McConnell, K. Berhane, L. Yao, M. Jerrett, F. Lurmann, F. Gilliland, et al., "Traffic, susceptibility, and childhood asthma," Environmental Health Perspectives, vol. 114, p. 766, 2006.
4. D. W. Dockery, J. Cunningham, A. I. Damokosh, L. M. Neas, J. D. Spengler, P. Koutrakis, et al., "Health effects of acid aerosols on North American children: respiratory symptoms," Environmental health perspectives, vol. 104, p. 500, 1996.
5. R. McConnell, K. Berhane, F. Gilliland, S. J. London, H. Vora, E. Avol, et al., "Air pollution and bronchitic symptoms in Southern California children with asthma," Environmental health perspectives, vol. 107, p. 757, 1999.
6. R. McConnell, K. Berhane, F. Gilliland, J. Molitor, D. Thomas, F. Lurmann, et al., "Prospective study of air pollution and bronchitic symptoms in children with asthma," American journal of respiratory and critical care medicine, vol. 168, pp. 790-797, 2003.
7. 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, pp. 632-655, 2012.
8. N. Almeida, L. Marques, and A. T. de Almeida, Multi-point systems for remote monitoring of air quality, 2003.
9. E. Agirre, A. Anta, L. J. R. Barron, and M. Albizu, "A neural network based model to forecast hourly ozone levels in rural areas in the Basque Country," in Air Pollution Xv. vol. 101, ed, 2007, pp. 109-118.
10. S. Al Maskari, D. Kumar, and T. Chiffings, DATA MINING FOR ENVIRONMENT MONITORING, 2012.
11. F. Amato, S. Nava, F. Lucarelli, X. Querol, A. Alastuey, J. M. Baldasano, et al., "A comprehensive assessment of PM emissions from paved roads: Real-world Emission Factors and intense street cleaning trials," Science of the Total Environment, vol. 408, pp. 4309-4318, Sep 2010.
12. B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," Neural computation, vol. 12, pp. 1207-1245, 2000.
13. V. N. Vapnik, "Statistical learning theory," 1998.
14. T. M. Cover and J. A. Thomas, Elements of information theory: John Wiley & Sons, 2012.
15. (2013). Wuhan environment protection bureau system. Available: http://ft.whepb.gov.cn:8090/
16. W. Gan, " Clustering:The fundamental problem in data mining reasearch," PhD, The PLA university of science and technology, 2003.
17. J. Gribbin, Q is for Quantum: Particle Physics from AZ: Universities Press, 1999.
18. D. Li and Y. Du, Artifical Intelligence with Uncetainty: National Defence Industry Press, 2005.
19. H. Lv, "Human face recognition reasearg based on data field," Master, Nanjing University of Science and Technology 2002.
20. Electric potential. Available: http://en.wikipedia.org/wiki/Electric_potential
21. Wikipedia. (2015). National Ambient Air Quality Standards. Available: https://en.wikipedia.org/wiki/National_Ambient_Air_Quality_Standards
22. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, p. 27, 2011.