Performance comparison of nine site characterization methods

Yacoub Najjar* & Sam Mryyan**
** Kansas State University, Manhattan, Kansas, USA*
*** Adjutant General's Department, Topeka, Kansas, USA*

2011 Pan-Am CGS Geotechnical Conference

ABSTRACT
Various site characterization methods have been developed in the past to profile specific parameters in soil media and/or groundwater. These methods vary in their ability to make precise and accurate predictions. This paper highlights the differences between the following nine profiling methodologies: Inverse Distance to a Power, Kriging, Minimum Curvature, Modified Shepard's, Nearest Neighbor, Polynomial Regression, Radial Basis Function, Local Polynomial, and Artificial Neural Networks (ANNs). Because each method uses an individualized logic, the accuracy of the methods' predicted profiles is expected to vary. To illustrate this, a hypothetical data-rich contaminated site is used for this purpose. Accordingly, a small fraction of the available data (about 1%) is presented to each method for site profiling. A comparative study of the models' site profiling outcomes/predictions is then performed in order to assess the most accurate site profiling methodology. Overall, ANN-based characterization outperformed the performance of the other eight well-known profiling methodologies.

RÉSUMÉ
Des diverses méthodes de caractérisation de site ont été développées dans le passé pour faire une description quantitative des paramètres spécifiques dans le sol et dans les nappes phréatiques. Ces méthodes varient dans leur capacité à faire des prédictions précises et exactes. Cet article souligne les différences entre les neuf méthodes de caractérisation suivantes: La pondération inverse a la distance, krigeage, courbure minimum, la méthode modifiée de Shepard, la méthode des voisins naturels, la régression polynomiale, fonctions a bases radiales, Le polynôme local, et les réseaux de neurones artificiels (RNA). Parce que chaque méthode utilise une logique spécifique, l'exactitude des caractéristiques quantitatives prédites par les méthodes peut varier. Pour illustrer ceci, un site contaminé, riche en données et hypothétique est utilisé à cet effet. En conséquence, une petite fraction des données disponibles (approximativement 1%) est utilisée par chaque méthode d'interpolation pour prédire les caractéristiques du site. Une étude comparative des résultats/prédictions obtenus par chaque méthode pour le site d'étude fut réalisée pour évaluer la méthode de caractérisation la plus exacte. Dans l'ensemble la caractérisation avec la méthode RNA a donné de meilleurs résultats que les autres méthodes de caractérisation bien connues.

1 INTRODUCTION

Environmental contaminants in geologic media such as soil and groundwater are of great concern in today's society. Millions of dollars are spent each year in efforts to clean up areas contaminated by pollutants from industrial and public waste such as solvents, fuels, and processing waste. Before cleanup efforts can begin, the area and extent of contamination must first be determined. This can be done by using one of several profiling methodologies. Typically, these methodologies use field data and specific mathematical algorithms to predict the areas and levels of contamination. In this paper, the profiling performance of eight well-known profiling methodologies and Artificial Neural Networks (ANNs) are compared.

When trying to identify an area of contamination, the most accurate and precise means is to perform soil/groundwater sampling at regular designated intervals throughout the entire area in question. This, however, is not a practical method due to cost and time restraints. Instead, a limited number of samples are collected throughout the area in question. Many factors determine where and how samples are collected. Generally, sample locations are determined using the professional judgment of the site investigation team. Once known data is collected, one of several profiling methodologies can be used to predict areas of contamination. Eight of the most highly utilized methodologies (available in Surfer® 8.0 Software: http://www.goldensoftware.com/) for 2-D profiling are: Inverse Distance to a Power (IDP), Kriging, Minimum Curvature (MC), Modified Shepard's (MS), Nearest Neighbour (NS), Polynomial Regression (PR), Radial Basis Function (RBF), and Local Polynomial (LP). ANN-based methodology is investigated herein as an alternative efficient 2-D and 3-D profiling methodology. Although all profiling methodologies function in similar manner, some methodologies produce more accurate profiles than others.

In order to utilize any profiling methodology, a known set of data must be present. For the purpose of this paper, a hypothetical data-rich contaminated site scenario is utilized. Data was generated to compare 2-D & 3-D pollution concentration profiles generated via different profiling methodologies.

2. TWO-DIMENSIONAL CASE

2.1 Mathematical Equation

In order to determine the distribution of contaminates at the hypothetical site, a mathematical equation was developed to produce the pollutant concentration value at

any given (x, y) location. Note that, x and y coordinates refer to the x and y distances (in meters) for the associated observation point measured from a reference point (i.e., x = 0 m and y = 0 m).

$$V = x^{0.5} + y^{0.6} + (x^{0.3} * y^{0.2}) + 2\sqrt[4]{(x * y)} + 3\ln\left(\frac{x^{1.1} + y^{1.5} + 20{,}000}{1000}\right) \quad [1]$$

where V is the contaminant concentration value.

2.2 Databank

Two databases containing (x, y and V) values were generated for two 2-D cases at various locations across the site. The site size is 300 m in the X direction by 300 m in the Y direction. To achieve this objective, the hypothetical site was divided into two grid systems as follows:

1. 7.5 m interval case: In this scenario, 7.5 m interval (i.e., $\Delta x = \Delta y = 7.5$ m) in both x (east) and y (north) directions was used to generate a total of 1,681 sampling points. The x, y coordinates and V values of selected 17 points (about 1% of the total sampling points) were provided for the eight profiling methodologies available in Surfer software. Each methodology was then used to predict the corresponding contamination value (V) for the 1,681 designated x and y coordinates representing the site. The resulting data banks were processed to construct 8 contamination distribution contour maps and to calculate the corresponding Root Mean Squared Error (RMSE) (Equation 2) value for each methodology.
2. 3 m interval case: Utilizing a 3 m interval (i.e., $\Delta x = \Delta y = 3$ m) for both x (east) and y (north) directions it was possible to generate a total of 10,201 sampling points for the 300 m x 300 m site. Similar to the 7.5 m case, x and y coordinates and V values of selected 103 points (about 1% of the 10,201 total sampling points) were provided for the eight profiling methodologies available in Surfer software. Each methodology was then used to predict the corresponding contamination value (V) for all 10,201 designated x and y coordinates representing the site. The resulting eight data banks were processed to construct 8 contamination distribution contour maps and to calculate the corresponding RMSE value associated with each methodology.

2.3 ANN Model Development

Unlike Surfer® 8.0 profiling methodologies, ANN-based profiling model require the user to train or educate the network about the process that it is supposed to model. To train the network, a known set of input data along with the desired outcome is used [Dowla and Rogers (1995), Mryyan and Najjar (2005), Itani and Najjar (2000)]. The BackPropagation ANN methodology using the supervised training approach is used to train the desired ANN models to produce output values that are as close to the real values as possible via repeated modifications of the network's connection weights. This process continues until the error at the output layer is minimized. Once this training process is completed, the developed model can then be used for prediction tasks.

Neural Networks can reach a least-error structure by training using examples related to the problem under consideration. A least-error structure is the one responsible for producing outputs that are very close or equal to the real desired values. Reasonable training input and output vectors should cover a wide range of the sampling domain. Deriving an appropriate and representative mapping between input and output vectors reflects the effectiveness of neural networks. For proper modeling, a network should at least pass through two stages, namely training and testing stages (Najjar & Basheer,1996). Selected data with their input and output values are introduced to a network (having a certain number of hidden nodes and layers) so that the network trains itself to produce output values that are as close to the real values as possible. The training is achieved by modifying the values of the connection weights. The network stops learning when weight adjustment processes produces no improvement in the output values. The same network should be tested on data never used in training to verify its generalization capabilities. The procedures of training and testing should be repeated for networks having different numbers of hidden layers and/or hidden nodes.

When developing any ANN model, it is important to determine what input and output values will be used. For the hypothetical data-rich contaminated site case, x and y coordinates were used as the only input values to the model. The pollutant concentration value (V) was used as the output for their associated network model. X and y coordinates refer to the x and y distances (in meters), for the associated observation point, measured from a reference point (i.e., x= 0 m, y = 0 m).

For the 7.5 m interval case (i.e., $\Delta x = \Delta y = 7.5$ m), a network model was developed by using the same 17 points that were used by the Surfer Software methodologies. For ANN case, 12 data sets were used for training and the remaining 5 data sets were used for testing purposes. The best performing network was determined by carrying out a number of adaptive training and online testing trials in order to arrive at the least error on the testing data sets. Overall Best Performing Network (BPN) is defined as the one having the lease error (in terms of Average Squared Error (ASE) on the testing data sets from among all evaluated trial networks. Overall BPN was achieved at ASE value of 0.010856 and a structure noted as 2-3-1 (i.e., 2 inputs representing x and y coordinates; 3 hidden nodes and one output denoting the associated value of the V variable). Once this network was established, it was then used to predict the V values at all 1,681 location points for the site. The predicted values were used to construct contamination distribution contour map and to calculate the corresponding RMSE value for this case.

For the 3 m interval case (i.e., $\Delta x = \Delta y = 3$ m), a network model was developed by using about 1% of the total 10,201 data points (i.e., the same 103 data points used by the Surfer Software methodologies). In this ANN case development, 75 data sets were used for training and the remaining 28 data sets were used for the online testing purposes. Based on various ANN work by the

Najjar and his co-workers [Ali and Najjar (1998), Hunag et al. (2006), Mandavilli et al, (2005), Mryyan and Najjar (2006), and Najjar and Felker (2003)],

It is highly imperative that training database contain all data sets that have the extreme attributes in terms of locations and values. Accordingly, the developed ANN model will always operate in an interpolation mode instead of an extrapolation mode. ANN-based prediction models are excellent when used in an interpolation model while may be unreliable when used in an extrapolation mode. Therefore, it is very important to appropriately select the distribution of the training and testing data sets. Following similar strategy as the one used for the 7.5 m interval case, the overall BPN was achieved at an ASE value of (on testing data sets) 0.000228 and a 2-2-1 structure. Once the 2-2-1 profiling network was established, it was then used to predict the V values at all 10.201 location points for the site. The predicted values were used to construct contamination distribution contour map and to calculate the corresponding RMSE value for this 3 m interval case.

Comparing ASE values for the 3 m and 7.5 m interval cases, it can be observed that the ASE value for the 3 m case has reduced by about 47 (i.e., 0.010856/0.000228) folds for the 6 fold increase in data richness (i.e., 103/17). This noted behavior is expected and logical. As more data become available, the profiling network should be able to characterize the site more accurately. Therefore, the more data is available, the more accurate is the developed profiling network. Moreover, it can be observed that the 3 m interval network only needed 2 hidden nodes to efficiently characterize the site compare to the 3 hidden nodes needed for the 7.5 m interval network.

In order to compare (rank) the prediction accuracy of the profiling methodologies used herein, the following Root Mean Squared Error (RMSE) accuracy measure was used:

$$\text{RMSE} = \frac{\sqrt{\sum_{1}^{n}(y' - y)^2}}{n} \qquad [2]$$

where:
n = number of data sets used.
y´= the output generated by the model for the V variable
y= the actual value of the V variable

Accordingly, the best performing profiling methodology is the one having the least RMSE value.

3. THREE-DIMENSIONAL CASE

One of the first steps in the remediation process of any site is to determine the characteristics of the contaminated site. This includes not only obtaining historical and geological information of the site but most importantly, determining the location and concentrations of contaminants at the site. This is done by collecting samples at selected points throughout the area of concern and analyzing the samples to determine the concentration of the contaminants. Each sample will have its own unique set of data which includes the location of the sample (latitude, longitude and depth) and a concentration. With this information, a detailed map of the contaminated site can be created (Najjar & Basheer,1996).

3.1 Mathematical Equation

Unlike the eight 2-D profiling methodologies available in Surfer Software ANN allows also for 3-D site profiling based on x, y and z coordinates. In an actual field situation, samples would be collected at various locations for lab analysis in order to obtain the associated pollutant concentration values. For the purpose of this study, the following equation was used to represent the concentration of the pollutant across the 3-D site (300 m x 300 m x 15 m):

$$V = x^{0.5} + y^{0.6} + z^{1.5} + (x^{0.3} * y^{0.2} * z^{0.1}) + 2\sqrt[4]{(x * y * z)} + 3\ln\left(\frac{x^{1.1} + y^{1.5} + +z^{2.5} + 50{,}000}{1000}\right) \qquad [3]$$

In this case, x = east, y = north, and z = depth. Accordingly, at any given location (x, y, and z) Equation 3 will produce the associated pollutant concentration value (i.e., V value). Note that, x, y and z coordinates refer to the x, y and z distances (in meters) for the associated observation point measured from a reference point (i.e., x = 0 m , y = 0 m and z = 0 m).

3.2 ANN Model Development

A large database containing (x, y, z and associated V) values were generated via Equation (3). Accordingly, the (300 m x 300 m x 15 m) hypothetical site was divided into a grid system. Grid lines were set at 7.5 m intervals for both x (east) and y (north) directions (i.e., 2-D plane) and at depths z = 1.5, 4.5, 7.5, 10.5 and 13.5 m. A total of 1,681 sampling points were generated, in this case, for each depth. This produced a total of 8,405 points. In this case, x, y and z coordinates were used as input nodes while V variable is used to represent the output nodes. Eight five (85) data points (representing about 1% of the total 8,405 available data points) were selected to train and test the desired ANN model. Accordingly, 60 data points were used for training while the remaining 25 points were used for online testing in order to assess the generalization capability of the trained networks. Similar procedure to the one used for the 2-D case was utilized herein to arrive at the optimal 3-D ANN profiling model. Therefore, the structure of the 3-3-1 BPN contained 3 inputs, 3 hidden and 1 output nodes. The ASE on the testing data set associated with this 3-3-1 BPN model is 0.000300. The 3-3-1 ANN model was then used to predict the corresponding contamination values (V) for the 8,405 designated x, y and z coordinates representing the site. The resulting data bank was processed to construct various contamination distribution contour maps (at z = 1.5, 4.5, 7.5,10.5 and 13.5 m) of the hypothetical contaminated site. Moreover, the resulting data bank was used to compare the ANN predicted values with the actual values at all 8,405 location points. The resulting RMSE value calculated for the 3-D BPN is about 6.4%

3.3 Regression Model Development

Since none of the eight Surfer-based methodologies can perform 3-D profiling using x, y and z, the following regression equation was developed using the same 85 data points utilized in developing the BPN model:

$$V = - 4.11 + 0.0499*x + 0.0766*y + 7.60*z \quad [4]$$

Where V represents the desired contaminant concentration value for given x, y and z coordinates within the site.

The regression model (Equation 4) was then used to predict the corresponding contamination values (V) for the 8,405 designated x, y and z coordinates representing the site. The resulting data bank was processed to construct various contamination distribution contour maps (at z = 1.5, 4.5, 7.5, 10.5 and 13.5 m) of the hypothetical contaminated site. Similarly, the resulting data bank was used to compare the regression-based model predicted values with the actual values at all 8,405 location points. The RMSE value obtained for this case is about 17.4%.

4. RESULTS AND DISCISSION

In order to compare the performance of all methodologies utilized herein, two comparison strategies were utilized, namely: i) RMSE values, and ii) contour maps.

4.1 Comparison of RMSE Values

2-D Cas**e:** RMSE values obtained for the 2-D profiling cases and via the nine profiling methodologies (including ANN method) are listed in Table 1. By examining RMSE values in this table, it can be observed that for the 7.5 m. interval case, all nine methodologies attain high RMSE values. The model achieving the least RMSE value is the ANN-based model. It attains about 19.17% error rate. The second most accurate methodology is the Local Polynomial with a 19.3% error rate. When compared to ANN performance, this represents less than 1% difference in prediction accuracy rate. The profiling methodology that produced the least accurate profile is the Inverse Distance to a Power. It has an RMSE value of about 42.4%. This represents more than double the RMSE value for the ANN model.

For the 3 m. interval case, all of the nine methodologies attain lower RMSE values. This is logical and consistent with our intuition. As more data become available, all models will become more accurate. Again, the model with the least RMSE value is the ANN-based model. Its error rate is about 3.7%. The second most accurate methodology is the Radial Basis Function with about 4.8% error rate. When compared to ANN performance, this represents about 30% difference in the prediction accuracy rate. The profiling methodology that produced the least accurate profile is again the Inverse Distance to a Power, with an RMSE value of about 10.4%. This represents about 2.8 times the RMSE value attained via the ANN model. The only consistent thing in the RMSE comparison listed in Table 1 is that: ANN-based profiling methodology is rank best and Inverse Distance to a Power methodology is ranked worst. All other seven methods seem to vary in terms of their ranking. Therefore, in order to assure that we are using the best profiling methodology, for 2-D cases, it is recommend to always use ANN-based profiling methodology.

3-D cas**e:** When comparing the RMSE value obtained using the 3-dimensional ANN-based model (with RMSE value of 6.4%) with that obtained via the regression-based model (with RMSE value of 17.4%), it can be observed that the ANN model significantly outperformed the regression model in this 3-D profiling task. Error rate of the regression model is about 270% of that reported for the ANN model. Note that the same 85 data points were used to develop both models. Moreover, knowing that all Surfer-based eight methodologies are only suited for 2-D profiling and can not perform 3-D profiling, makes it very clear that ANN-based methodology is the only one to use for any efficient 3-D profiling tasks.

Table 1: Profiling methods and their corresponding RMSE values for the 7.5 m and 3 m interval cases

#	Method	Error(%RMSE) for 3 m Interval case	Error(%RMSE) for 7.5 m Interval case
1	ANN	3.72	19.17
2	RBF	4.80	38.95
3	Kriging	4.99	23.66
4	MS	4.99	19.78
5	LP	5.15	19.30
6	MC	5.23	20.40
7	NN	8.02	38.95
8	PR	8.71	26.42
9	IDP	10.43	42.39

4.2 Comparison of Contour Maps

Contour maps were generated using the Surfer 8.0 software program. This program was used to produce contamination concentration contour maps for the hypothetical site using the previously mentioned (x, y, V and z when applicable) data banks. Contour maps where generated for the 3 m and 7.5 m interval 2-D cases as well as 3-D (7.5 m interval) case discussed in the previous sections. Due to space limitations, selected contour maps from the 2-D 3-m interval case are shown in this paper. Also, for the 3-D case, only contour maps at z = 7.5 m are shown herein.

2-D Case: For visual comparative purposes, a base line contour map of the pollutant concentration distribution of the site based on the actual 10,201 data points was generated as depicted in Figure 1. This map is used herein as a baseline to compare the profiling accuracy of the nine profiling methods listed in Table 1. When comparing the contour maps of selected profiling

methods (depicted in Figures 2 to 5), the ANN-based contour map (Figure 2) is clearly the one that most closely resemble the base line contour map shown in Figure 1. Note that, as indicated in Table 1, The ANN-based method attained the lowest RMSE value of 3.7% among all nine profiling methods. The remaining methods present lesser degrees of similarities with the baseline map (see Figures 2 to 5 as a representative set). Contour maps produced by the inverse distance to a power method (Figure 5) can be considered as the worst compared with the baseline map shown in Figure 1.

Two of the contour maps produced respectively by the radial basis function and kriging methods (Figures 3 and 3, respectively) show, in the north east region, areas of non-contamination where actual contamination is present. Considering the results listed in Table 1 and the contour maps depicted in Figures 2 to 5, it can be noticed that the performance of the ANN-based method is very consistent. This method is producing the best contour map and attaining the least RMSE value among all nine methods listed in Table 1. Therefore, ANN-based method should be considered as the method of choice for any 2-D site profiling.

One common observation among all models considered herein is that no model was able to accurately characterize the actual (logarithmic) behavior of the variable V at the south and west edges of the site. In order to account for this logarithmic behavior more data points, taken from the south and west edges, are needed to be included in the models' profile development process.

3-D Case: Baseline contour map for the distribution of V variable at z = 7.5 m is shown in Figure 6. This map was generated based on 1,681 actual data points derived directly from Equation 3. The corresponding ANN-based and regression-based contour maps at z = 7.5 m are depicted in Figures 7 and 8, respectively. The RMSE values obtained in this case (6.4 % for ANN method and 17.4% for regression model) are an indicative of the degree of agreement between the profiles presented in Figures 7 and 8 with that shown in Figure 6. ANN-based profile (even though it was developed by utilizing no more that 1% of the available data at the z = 7.5 m level) presents a reasonable agreement with the actual map. The profile generated from the regression model has very low degree of similarity with the actual profile shown in Figure 6. As in the 2-D case, no model was able to accurately characterize the actual (logarithmic) behavior of the variable V at the south and west edges of the site. To address this profiling deficiency, far more data points (taken from the south and west edges) are needed to capture this logarithmic behavior.

5. CONCLUSIONS

The use of ANN methodology for contaminates profiling, demonstrated in this study, provided the most reliable predictions about the location and extent of contamination at the hypothetical site. ANN proved to attain the lowest RMSE in both the 2-D and 3-D comparisons cases. ANN-based profiling models also produced the best contaminant distribution contour maps for the 2-D and 3-D cases considered in this study. Along with the fact that ANN is the only profiling methodology that allows for efficient 3-D profiling, this study demonstrates that ANN-based methodology provides the most accurate data predictions and site profiling contour maps for a contaminated site.

Compared to the methods discussed herein, ANN-based methodology is characterized by its flexibility and generality. Its flexibility is demonstrated by its potential to accurately predict values of a certain contaminate parameter at a specific location when only supplied with x, y and z (for 3-D cases) coordinates. Its generality lies in its power to capture the mode of change in the spatial distribution of a pollutant's parameter based on all available data. Accordingly, all available data at various spatial locations are nicely utilized by the ANN-profiling model in order to efficiently capture the spatial distribution behavior for the parameter of interest.

REFERENCES

Ali, Hossam and Najjar, Yacoub 1998. Neuronet-based approach for assessing the liquefaction potential of soils. *Transportation Research Records*, No. 1633, pp. 3-8.

Dowla, U. and Rogers, L. 1995. *Solving problems in environmental engineering and geosciences with artificial neural networks.* Cambridge, MA: MIT Press.

Huang, C., Najjar, Y. and Romanoschi, S. 2006. Characterizing the fatigue life of asphalt concrete. *Intelligent Engineering Systems through Artificial Neural Networks*, Vol. 16, pp. 369-374.

Itani, Omar and Najjar, Yacoub 2000. 3-D modeling of spatial properties via artificial neural networks. *Transportation Research Records*, No. 1709, pp. 50-59.

Mandavilli, S., Najjar, Y. and Abu-Lebdeh, G. 2005. Modeling crash rates for the Kansas rural expressway network. *Intelligent Engineering Systems through Artificial Neural Networks*, Vol. 15, pp. 721-730.

Mryyan, S.A. and Najjar, Y. M, 2006. Using neural network to investigate the environmental impact of an abandon landfill. *ASCE World Water and Environmental Resources Congress 2006*, pp. 250-262.

Mryyan, S.A. and Najjar, Y. M. 2005. Investigating the environmental impact of an abandon landfill. *Intelligent Engineering system through Artificial Neural Networks,* Volume 15. pp. 751-761.

Najjar, Y. M. and Basheer, I. A. 1996. A neural network approach for site characterization and uncertainty prediction. *Uncertainty in the Geological Environment: From Theory to Practice, ASCE GSP* No. 58, Vol. 1, pp. 134-148.

Najjar, Yacoub and Felker, Victoria 2003. Modeling the time-dependent roughness performance of Kansas PCC pavements. *Intelligent Engineering Systems through Artificial Neural Networks*, Vol. 13, pp. 883-888.

Surfer User's Guide 2004. *Golden Software*, Inc.

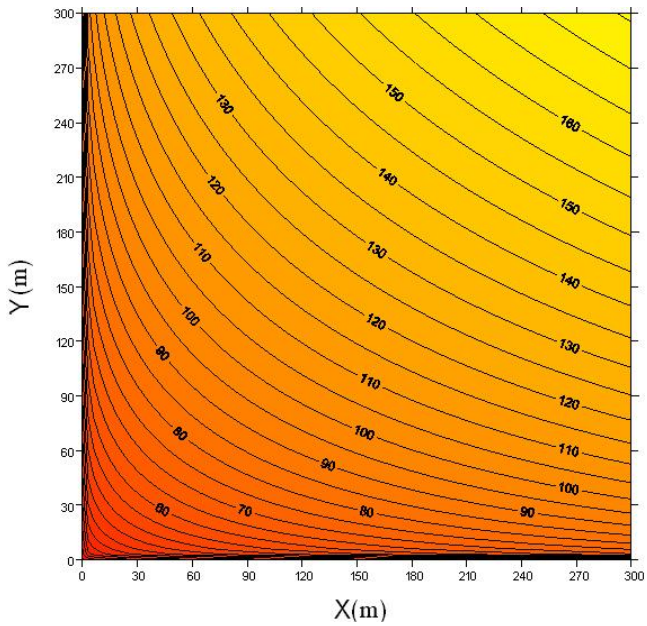

Figure 1 Baseline contour map of the pollutant V (for the 3 m interval case) based on 10,201 actual data points

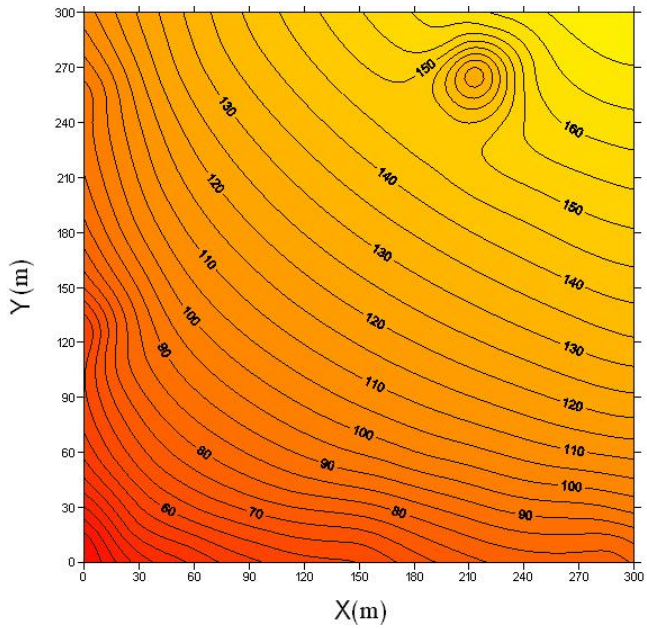

Figure 3 Contour map based on Radial Basis Function method for the 3 m interval case

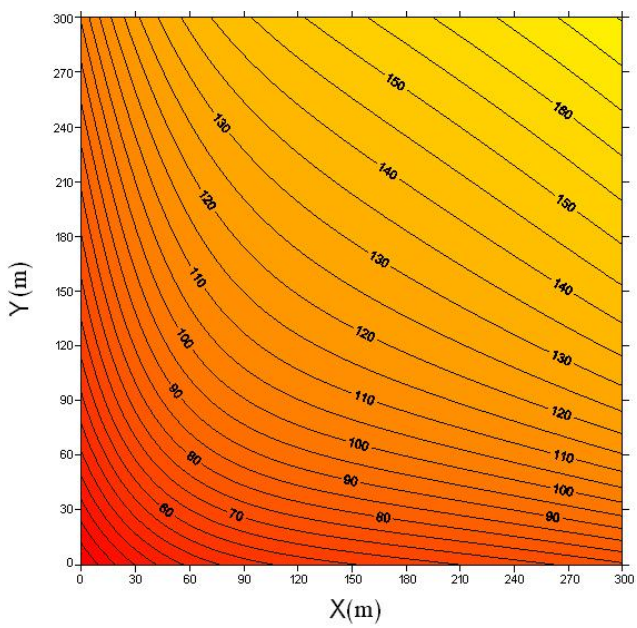

Figure 2 Contour map based on ANN model 2-2-1 for the 3 m interval case

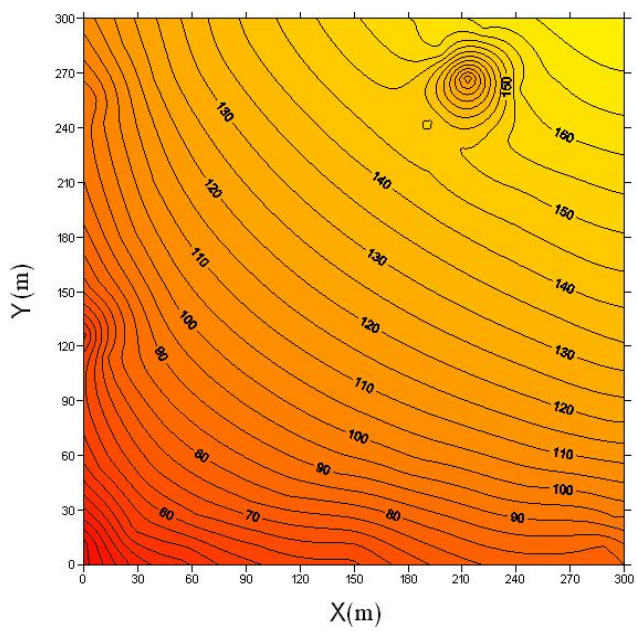

Figure 4 Contour map based on Kriging method for the 3 m interval case

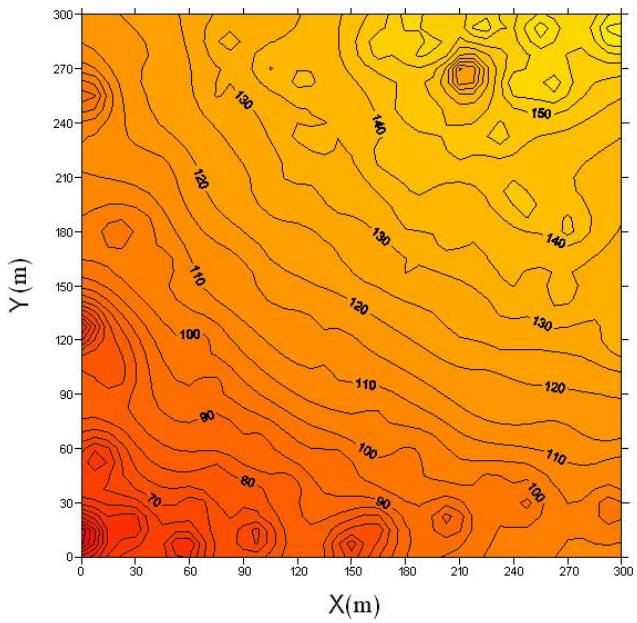

Figure 5 Contour map based on Inverse Distance to a Power method for the 3 m interval case

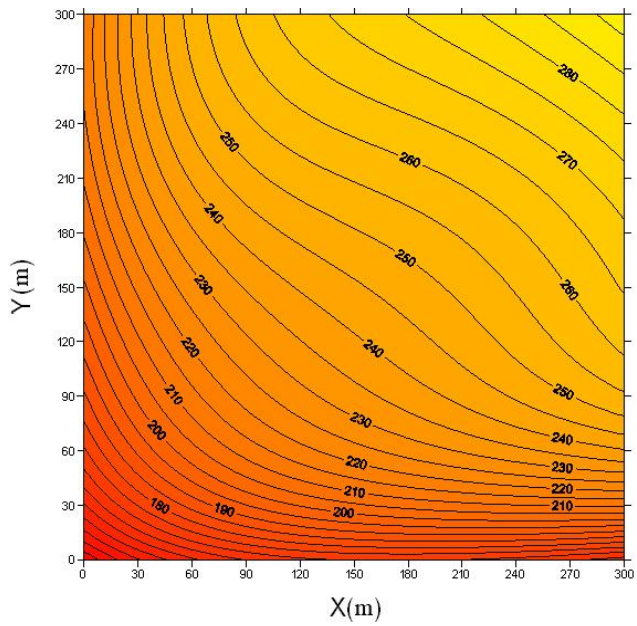

Figure 7 Contour map based on ANN model at z = 7.5 m

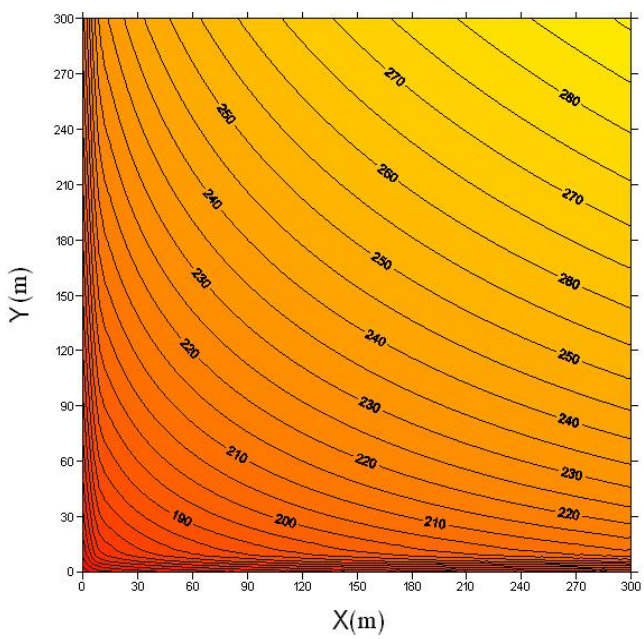

Figure 6 Baseline contour map of the pollutant V (at z = 7.5 m) based on 1,681 actual data points

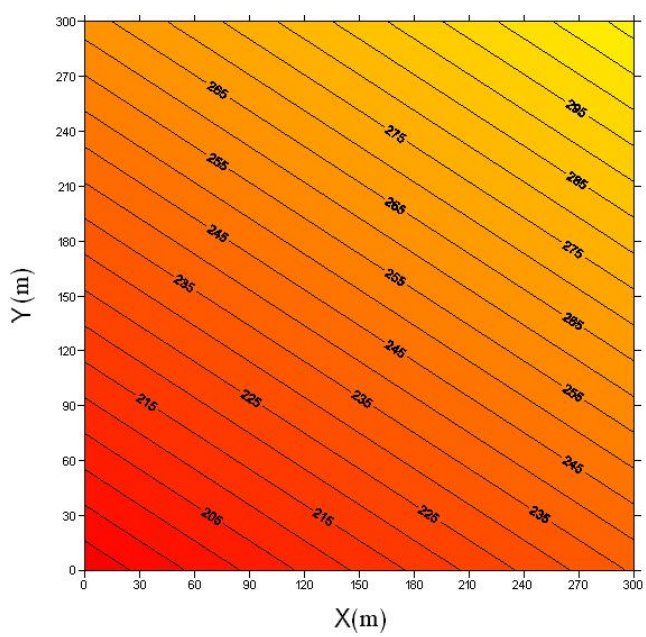

Figure 8 Contour map based on Regression model at z = 7.5 m