# Method for Optimal Arrangement of Soil Sampling Based on Neural Networks and Genetic Algorithms

Wu Zongshu[1,a], Ai Jiaoyan[1,b], Deng Chaobing[2,c], Cai Yajuan[1,d], Wei Zongming[1,e]

[1]College of Electrical Engineering, Guangxi University, Nanning 530004, China

[2]Guangxi Autonomous Region Environmental Monitoring Central Station, Nanning 530028, China

[a]wuzongshu1989@163.com, [b]aijy@gxu.edu.cn

**Keywords:** Neural networks; Genetic algorithms; Soil sampling; Optimal arrangement

**Abstract.** In order to explore a new way of optimization for soil sampling layout, in this paper, spatial distribution of heavy metal concentrations which is based on RBF neural network fitting was studied by genetic algorithm, corresponding network structure and algorithm flow were constructed, in addition the network and algorithm parameters settings were analyzed and a sampling experiment using this method was conducted in an abandoned mine located in a county of Guangxi. The results of optimizing the layout point prove that the number of the sampling points can be reduced by about a half under the premise of meeting the fitting accuracy, this will greatly reduce the cost of sampling and analysis and the redundancy of the data, and is expected to promote the relevant application in soil composition analysis and other related fields.

## Introduction

To understand the spatial distribution of heavy metals in soil by soil sampling is a essential part of the environmental assessment. Due to the high cost of consumption of soil sampling, rational distribution of soil sampling point is critical[1].The most commonly used method of soil sampling distribution abroad are mainly simple random sampling, subjective judgment sampling, zoning sampling, regular grid sampling and mixed sampling. These sampling methods often lead to the number of sampling points redundancy, and local sampling density can not meet accuracy[2,3]. Furthermore it is Time-consuming and labor-intensive resulting. In recent years, genetic algorithms showed a huge advantage in solving combinational optimization problems. Its intelligent optimization methods that can automatically obtain and guide to the search space of optimization,and adjust the search direction adaptively, More over,I t does not require determined rules to realize the Soil sampling point layout optimization[4].

## Processing Flow of Soil sampling layout optimization

Figure 1 shows the processing flow of soil sampling layout optimization, which is mainly constructed by RBF neural networks and genetic algorithms. This paper put a heavy metal Pb sampling result of an abandoned mine in Guangxi as the research object, through RBF neural network to fit the space density of Pb, combining with genetic algorithms to optimize the fitting surface. And by comparing the model results established under different parameters to select the optimal solution to provide a reference for optimal soil sampling.

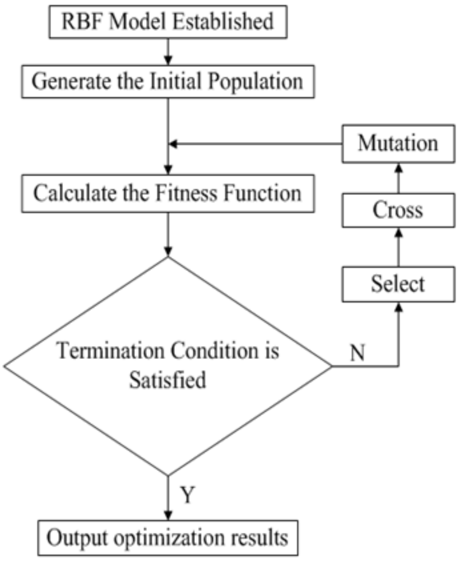
RBF Model Established
Generate the Initial Population
Mutation
Calculate the Fitness Function
Cross
Select
Termination Condition is Satisfied
N
Y
Output optimization results

Fig.1 Processing Flow of Soil sampling layout optimization

## Spatial concentration of heavy metals distribution fitting based on RBF neural network

**Network Structure.** RBF neural network structure of heavy metal concentrations has two inputs,one output and one hidden layer, and use the function Guass as the activation function, shown in Figure 2, the spatial coordinates of sampling points as the network input, heavy metal concentrations as the network output.It establish a link between the spatial coordinates and heavy metal concentration, its function expression is: *Z=f(x,y)*.

Where: *(x,y)*- sampling point spatial coordinates; *Z*- heavy metal Pb concentration[5,6].

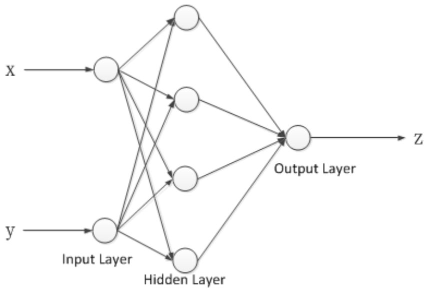

Fig.2 RBF neural network structure

**The spatial distribution of heavy metal concentration FITTING.** Fitting the spatial distribution of heavy metals concentration is implemented on matlab platform by following steps:

1)Normalize coordinates and soil Pb concentration values. Corresponding process using the following formula:

$$x = (x - x_{\min}) / (x_{\max} - x_{\min}) \tag{1}$$

$$y = (y - y_{\min}) / (y_{\max} - y_{\min}) \tag{2}$$

$$Z = (Z - Z_{\min}) / (Z_{\max} - Z_{\min}) \tag{3}$$

2)Use matlab function meshgrid to generate 100*100 interpolation points, and normalize the interpolation points correspondingly.

3)Establish newrb function command on matlab to train the network to obtain the relationship between heavy metal concentration and the plane coordinates. Its command calling format is:*net=(P,T,goal,spread,mn,df)*. Where:*net*- neural network model;*newrb*-matlab radial basis function neural network to invoke the command; *P*- input matrix;*T*- output matrix, the text of concentration;*goal*-mse mean square error function;*spread*- expansion coefficient;*mn*- the number of hidden layer neurons maximum;*df*- iterative frequency.Only *spread* and *mn* in parameters need to be adjusted.In this study, use "trial and error" to determine the required number *mn* of neurons in hidden layer and extended constant *spread* to interpolate the soil properties,and validate by the test sample in order to obtain the best network structure and parameters.

4)Substitute the Interpolation point obtained from meshgrid function command as input to the trained network structure to simulate, forecasting the coordinates of the point corresponding to the range of concentrations of heavy metals in the corresponding value.At last process the data with Anti-normalization formula to get the fitting surface required by optimization.The predicted results stored in txt format and convert to Raster files by Arctoolbox in Arcgis10.0 displaying in Arcmap , as shown below:

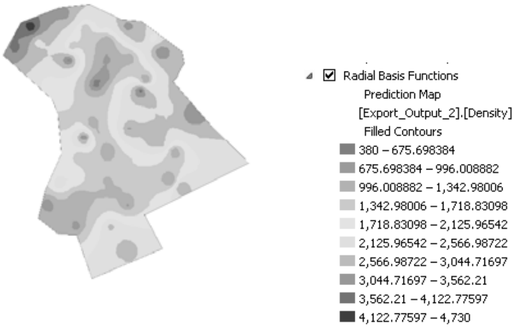

Fig.3 Prediction of the flat distribution of Pb in soil without optimized

## Optimization of sampling points based on Genetic Algorithm

In order to reduce the number of sampling points without affecting the sampling result.this paper use the genetic algorithm which does not require precise mathematical model and can globally search optimization to achieve the optimization of sampling points.The basic optimization steps are as follows:

1) Code: feasible solution in solution space is expressed as the structure of the genotype data string before solving and different combinations of string structure data constitutes different feasible solutions.This paper use binary encoding,Whether each sampling point is selected as parameters, If selected is 1, 0 is not selected, the code length is 45.

2) Generate the initial population: randomly generate 20 or 40 initial string structure data, each string structure data becomes an individual that represents a scheme of heavy metal samples and 20 or 40 to form a group,the group is the initial iterate point of genetic algorithm.

3)Fitness evaluation test：Judge the merits of the individual by calculating the fitness of an individual based on the practical standard, namely, the represented advantages of the feasible solutions. In this paper,we choose the surface fitted by 45 sampling points based on the RBF neural network as a standard,and compare it with the surface which is fitted by sampling points selected by genetic algorithm so that we can get the mean square error as the fitness function. the expression is:

$$f = \sum_{i=1}^{N} std(y_i - o_i). \tag{4}$$

4) The selection operator: Choose individuals from a population with high fitness, while eliminating the low fitness ones. In this paper, we use roulette wheel selection method which based on the ratio of the fitness selection policy. Selection probability $p_i$ for each individual $i$ is:

$$F = \sum_{i=1}^{N} f(X_i); \tag{5}$$

$$p_i = \frac{f(X_i)}{F}, i = 1,2,3\cdots i; \tag{6}$$

in the equation above,*F* represents the total fitness value of the individuals; and *N* represents the individual number in the population.

5)The crossover operator: Firstly, we take out one pair of mating individuals from the mating pool in random, and then according to the string length, we choose two stochastic bits from the mating pair as the cross-bit. At last, we implement crossover operation according to the crossover probability, the pairing individual swap interchangeable content in the cross section to form a new individual. The crossover probability is 0.6.

6)The mutation operator: Randomly select one individual among the middle group, and then alter the value of a certain gene to the size of the mutation probability. Here we define the mutation probability as 0.04.

The basic flow chart is as follow,

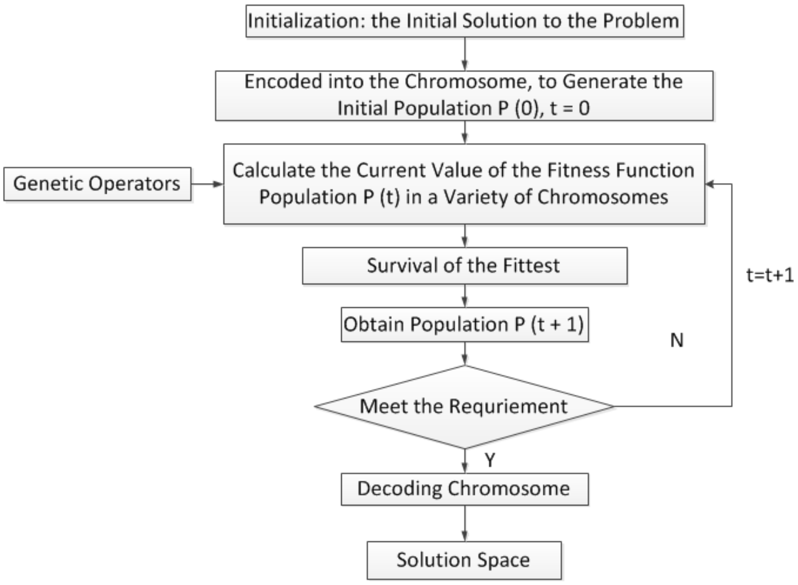
Initialization: the Initial Solution to the Problem
Encoded into the Chromosome, to Generate the Initial Population P (0), t = 0
Genetic Operators
Calculate the Current Value of the Fitness Function Population P (t) in a Variety of Chromosomes
Survival of the Fittest
t=t+1
Obtain Population P (t + 1)
N
Meet the Requriement
Y
Decoding Chromosome
Solution Space

Fig.4 Genetic algorithm flowchart

## Discussion

The optimization was implemented on matlab platform.In this paper, When optimizing we designed six program according to the different number of initial group and iteration,namely when the initial number of groups is 20, 40, each iteration of 20 times, 50 times, 100 times, to obtain the optimal combination of their respective programs, as shown in Table 1 in FIG. 5.The sampling number of these 6 programs are reduced about a half, and solid circle in the chart are selected as the fitting basic point from the 45 sampling points,which means The genetic algorithm can successfully complete the optimization of reduce the sampling point under a certain precision.

Table1 Iterative optimal combination of different times

| The number of population | The number of iterations | Error variance | Optimal combination (sampling point number) | Remark |
|---|---|---|---|---|
| 20 | 20 | 0.5399 | 1 3 5 6 7 10 11 14 15 17 18 19 22 24 26 31 32 41 42 43 45(a total of 21 points) | Plan A |
| | 50 | 0.5299 | 6 7 8 9 11 13 14 15 19 23 24 29 31 33 35 36 38 40 41 42 43 44 45(a total of 23 points) | Plan B |
| | 100 | 0.5257 | 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 21 22 24 26 27 31 32 33 35 36 40 43 44 45(a total of 30 points) | Plan C |
| 40 | 20 | 0.5067 | 5 6 7 10 16 21 22 23 25 27 29 30 31 32 33 36 39 41 43 44 45(a total of 21 points) | Plan D |
| | 50 | 0.5033 | 1 3 4 5 6 7 8 10 11 16 19 23 24 26 28 29 30 31 33 34 35 36 37 38 39 43 44 45(a total of 28 points) | Plan E |
| | 100 | 0.5013 | 1 3 5 7 10 11 14 15 16 18 20 21 24 28 29 30 33 34 35 37 38 40 41 43 45(a total of 25 points) | Plan F |

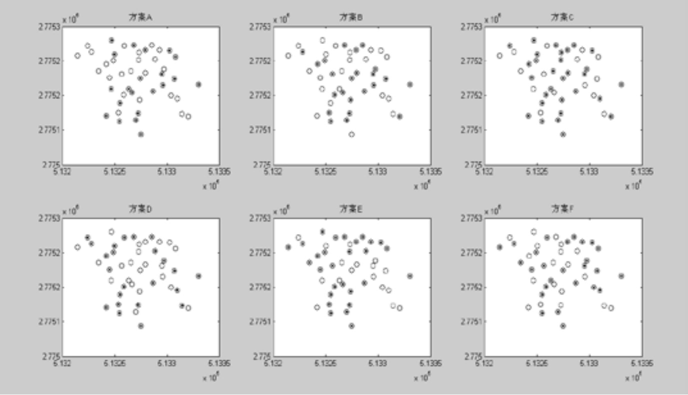

Fig.5 Optimal combination of sampling points distribution of the different number of iterations

Meanwhile, with the increasing number of iterations,the error variance will gradually decrease. The error variance in program A、B、C decrease from 0.5399 to 0.5299 and finally fell to 0.5257, which indicates that more number of iterations more chances to find an optimal combination. Besides, by comparing program A with program D、program B with E and program C with program F, it is obvious that the increase of population number can decrease the error variance to its current optimal.Now we can get figure 6 from Arcmap by fitting the Pb concentration of heavy metal with program F. It is evident from the figure that the concentration in the range of 380 to 1460 concentrated in the central and Pb concentration of heavy metal in the east where was heavy polluted is probably close to 2500-3700,the concentration value in the southeastern part lie in the range of 1850-3200 while Lead contamination which located from 750 to 1500 in the west is comparatively lighter. The concentration distribution of the two figures is approximately the same in trend when compares with Figure 3 and the interpolation points mean square error of these two figure is 0.5013 which is within the allowable range of control accuracy. This provide a further evidence that genetic algorithm can optimize the number of the number of sampling points successfully.

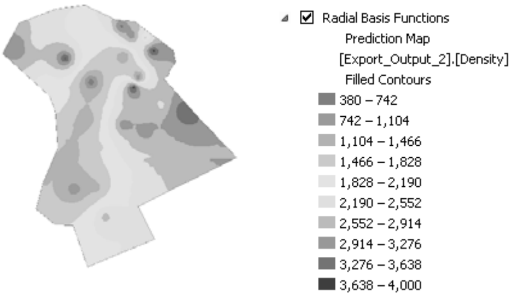

Fig.6 Prediction of the flat distribution of Pb in soil after optimized

## Conclusion

It requires not only a lot of manpower and resources but also long sampling period to collect the information of soil heavy metal pollution, so the purpose of optimize soil heavy metals sampling points is that reduce the sampling point as much as possible under certain precision or obtain the optimal accuracy by optimal permutation and combination under certain number of sampling points.In this study, genetic algorithm is used to optimize the spatial distribution of the heavy metal concentration sampling points which is based on RBF neural network fitting.The experimental results show that the genetic algorithm can reduce nearly half of the sampling point in the case of ensuring accuracy of fitting, this reduce the cost of sampling and analysis thus a feasible method to optimize the sampling point layout can be provided.

## References

[1] Li Qiquan, Wang Changquan, Yue Tianxiang, Li Bing, Yang Juan. Method for spatial variety of

soil organic matter based on radial basis function neural network[J].Transactions of the CSAE, 2010,26(1):87-93.(in Chinese with English abstract)

[2] Shen Zhangquan, Shi Jiebin, Wang Ke, et al. Spatial variety of soil properties by BP neural network ensemble[J].Transactions of the CSAE, 2004, 20(3): 35-39. (in Chinese with English abstract)

[3] Kerry R, Oliver M A. Forest soil acidification assessment using principal component analysis and geostatistics[J].Geoderma, 2007, 140(4): 374-382.

[4] Li Qiusheng, Zhang Ce, Liu Zhenghua.Intelligent Controller Based on Neural Network and Genetic Algorithm[J].Computer Measurement&Control,2007,15(5):610-612.(in Chinese with English abstract)

[5] CHEN Feixiang, CHENG Jiachang, HU Yueming,ZHOU Yongzhang, et al. Spatial Prediction of Soil Properties RBF Neural Network[J].Scientia Geographical Sinica,2013,33(01)：69-74.(in Chinese with English abstract)

[6] Shen Zhangquan, Zhou Bin, Kong Fansheng, et al. Study on spatial variety of soil properties by means of generalized regression neural network[J]. Acta Pedologica Sinica, 2004,41(3): 471-475.(in Chinese with English abstract)