城市表层土壤重金属污染分析

摘要

本文给出重金属浓度空间分布、各区域污染程度和污染原因，建立神经网络模型用遗传算法求解污染源位置，再将模型拓展以期得到地质环境演变模式。

问题一：通过八种重金属浓度的等值线图来描述其空间分布，利用污染程度指数来衡量不同功能区的污染程度。

以八种重金属在各采样点的浓度为初始数据，利用三角形线性插值法，运用 matlab 编程得到八种重金属浓度值在空间上的分布，并绘制成等值线图，得到了八种重金属的空间分布特点。同时，我们利用污染程度指数作为指标来衡量不同功能区的污染程度，得到结论：污染程度从高到低依次是工业区、交通区、公园绿地区、生活区、山区。

问题二：通过相关分析和关联规则分析来确定重金属污染的主要原因。

我们运用 spss 软件在显著性水平 0.01 的情况下对八种金属浓度进行 person 相关分析，并对不同金属之间的相关性进行双侧检验，结果发现：重金属 Cu、Cr、Ni 具有显著的相关性，其中 Cr、Ni 的相关性极其显著；重金属 Pb 与 Cd、Pb 与 Cu、Zn 与 Pb 有显著的相关性。我们再利用 SAS 进行关联规则分析，其中将功能区作为前项，土壤表层是否收到重金属污染作为后项，根据所得可信度分析出：所有重金属污染都主要来源于工业废水废渣，另外 As 和 Cr 污染还来源于生活垃圾，Pb 污染还来源于汽车尾气的排放。

问题三：建立神经网络模型，用遗传算法确定污染源的位置。

为了确定污染源的位置，我们建立了基于matlab平台的BP人工神经网络模型。结果表明该模型能准确反映重金属污染物的传播特征并得出重金属浓度的空间分布。然后利用遗传算法搜索到污染源的位置。求解结果如下表：

As	Cd	Cr	Cu	Hg	Ni	Pb	Zn
(1200,3006)	(2105,2593)	(3487,5325)	(2101,3390)	(2643,2875)	(3211,5686)	(1991,3329)	(1699,2867)
(18508,10206)	(4753,10875)		(3198,5822)	(13868,2354)	(24001,12366)	(4508,5412)	(3725,5487)
(27680,12111)	(17450,3825)			(15243,9186)			(9583,4512)
	(21301,11467)						(13653,9655)

综合分析所得污染源所在位置，发现不同金属的污染源有同源现象，依据同源性汇聚污染源，绘制了八种重金属的污染源汇总图。

问题四：神经网络模型的优点是具有较强的自组织、自学习能力、泛化能力和充分利用了海拔高度的信息；缺点是训练要求样本点容量较大。可以通过搜集前几年该城区八种重金属浓度的采样数据和近几年工厂分布多少位置的变化、交通路段车流量的变化、人口及生活区分布变化与植被分布多少位置的变化等数据，进一步拓展神经网络模型，得到该城市地质环境的演变模式。

关键词：三角形线性插值 相关分析 关联规则 神经网络 遗传算法

一、 问题重述

随着城市经济的快速发展和城市人口的不断增加，研究人类活动影响下城市地质环境的演变模式，日益成为人们关注的焦点。

不同的区域环境受人类活动影响的程度不同，按照功能划分，城区一般可分为生活区、工业区、山区、主干道路区及公园绿地区，现对某城市城区土壤地质环境进行调查。将所考察的城区划分为间距 1 公里左右的网格子区域，按照每平方公里 1 个采样点对表层土（0~10 厘米深度）进行取样、编号，并用 GPS 记录采样点的位置、海拔高度及其所属功能区等信息。分析获得了每个样本所含的多种化学元素的浓度数据。另一方面，按照 2 公里的间距在那些远离人群及工业活动的自然区取样，将其作为该城区表层土壤中元素的背景值。

我们需要解决如下问题：

(1) 给出八种主要重金属元素在该城区的空间分布图，并建立指标来分析该城区内不同区域重金属的污染程度。

(2) 对数据进行挖掘，分析出重金属污染的主要原因。

(3) 分析重金属污染物的传播特征，由此建立模型，从而确定污染源的位置。

(4) 分析所建立模型的优缺点，收集更多的信息，通过这些信息建立模型来研究城市地质环境的演变模式。

二、 符号说明和模型假设

2.1 符号说明

符号	意义
k	表示不同功能区
i	表示金属的种类
j	表示不同的样本
x_{ij}	表示样本 j 中金属 i 的浓度
x_i	表示金属 i 背景值的平均值
σ_i	表示金属 i 背景值的标准差
Y_{ij}	表示 x_{ij} 标准化后的值
ω_i	表示金属 i 的权重系数
I_k	区域 k 的污染程度指数

2.2 模型假设

① 假设样本值真实可信。
② 假设重金属浓度随空间是连续变化的。

三、 模型的建立与求解

3.1 问题一的求解

3.1.1 八种重金属元素在该区域的空间分布

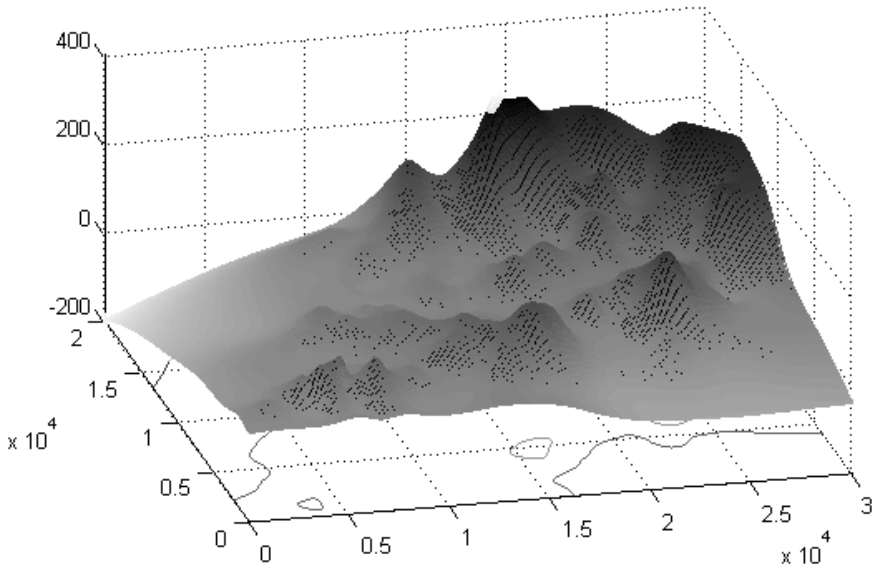

图 1 该城区的地形分布图

首先，我们根据样本点的位置和海拔绘制出该城区的地貌，见图 1。我们运用 matlab 软件，根据各个网格区域中的重金属含量，用三角形线性插值的方法得到各种重金属含量在空间上分布的等值线图。

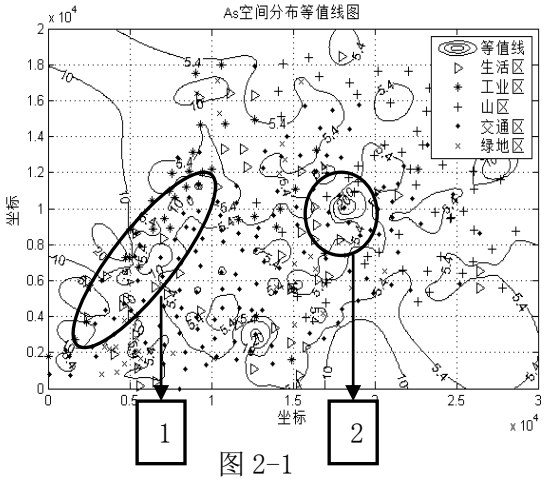

图 2-1

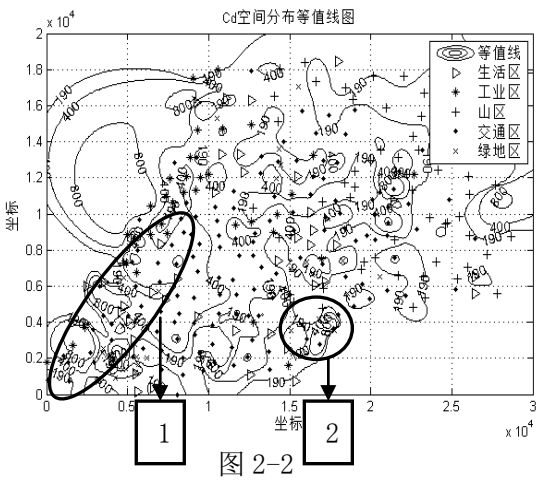

图 2-2

图 2-1 给出了 As 在该区域的空间分布：图中可以观察到 As 有两个明显的高值中心，我们标记为区域 1 和 2。这两个区域都处于工业区分布范围内，并以该两个区域作为中心向外延伸，浓度逐渐减少，同时我们注意到在山区的很多区域

浓度都在背景值之下。

图 2-2 给出了 Cd 在该区域的空间分布：同样我们也可以在该图中观察到两个明显的高值中心区域 1 和 2，这两个高值中心的地理位置和图 2-1 中高值中心的地理位置很相近，说明金属 As 和 Cd 在该城区的空间上的分布很相近。同时从图中我们还可以看到在整个城区内 Cd 的浓度普遍较大，而且很多区域内 Cd 的浓度都远超过背景值的最大浓度。

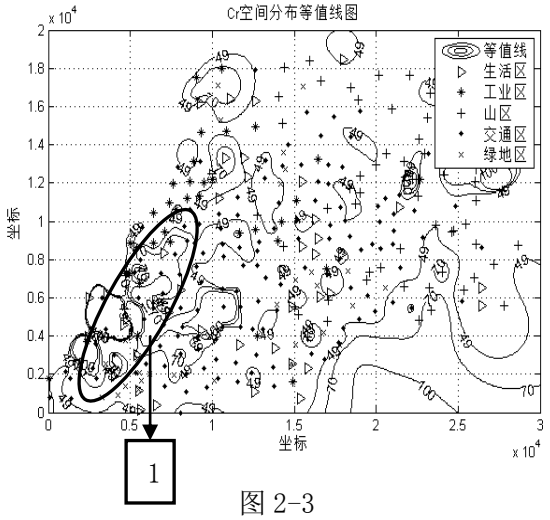

图 2-3

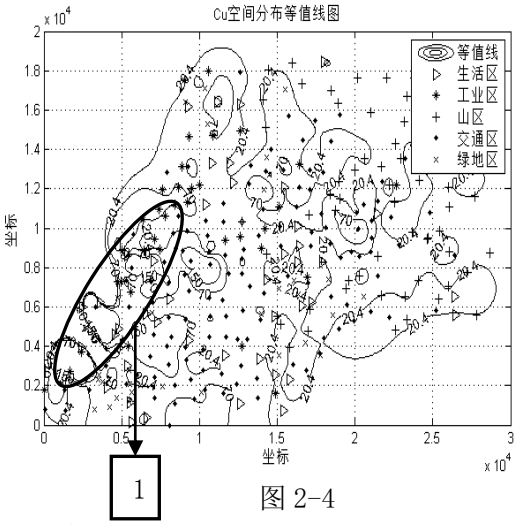

图 2-4

图 2-3 和图 2-4 分别给出了 Cr 和 Cu 在该区域的空间分布：显然，我们可看出 Cr 和 Cu 的浓度高值中心都出现在了西南角的工业带处。但两者的不同之处为：Cr 的浓度在城区的大部分区域都大于背景值的最大浓度，而 Cu 的分布正好相反。

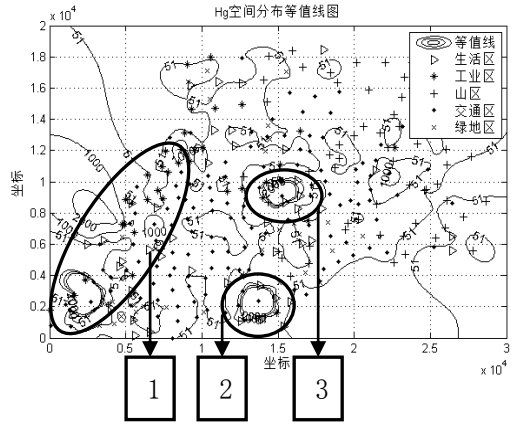

图 2-5

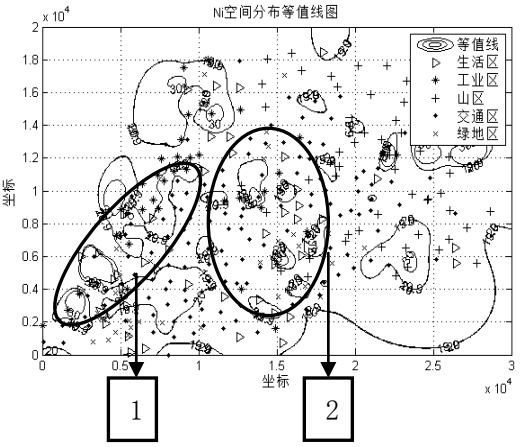

图 2-6

图 2-5 给出了 Hg 在该区域的空间分布：可以看点 Hg 有三个高值中心——区域 1、2 和 3，这三个区域都分布在工业区的附近，并且还可以看出这三个高值中心处的等值线都十分密集。

图 2-6 给出了 Ni 在该区域的空间分布：区域 1 和区域 2 是 Ni 浓度的两个高值中心，大约占该城区面积的 1/3，但在远离工业区的生活区交通区、绿地和山区，Ni 的浓度普遍较小。

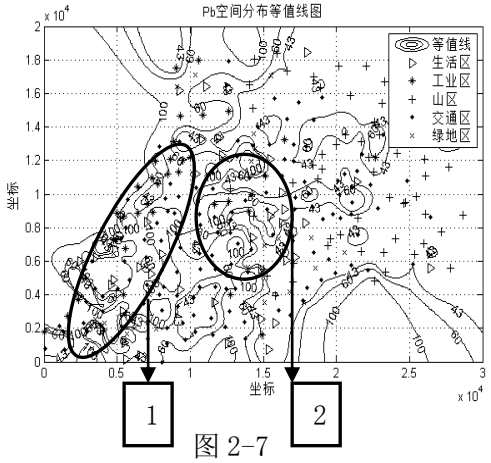

图 2-7

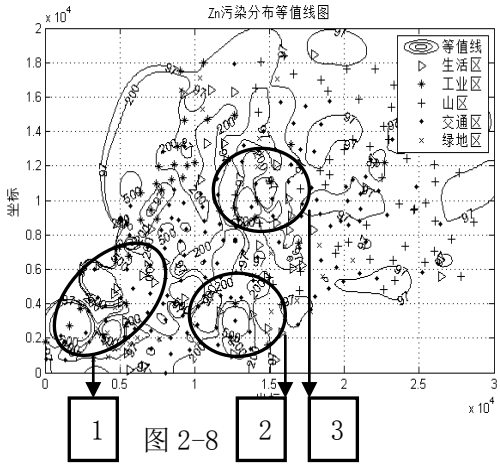

图 2-8

图 2-7 给出了 Pb 在该区域的空间分布：浓度的高值中心区域 1 分布在工业带上而另一个高值中心区域 2 在交通区内。这可能反映了 Pb 污染的来源主要是工业废水废渣和交通工具尾气的排放。

图 2-8 给出了 Zn 在该区域的空间分布：从图中可以看出 Zn 的出现了三个浓度高值中心分别标为区域 1、2、3，并且影响范围较广，在生活区、交通区和绿地的大部分区域内 Zn 的浓度普遍都比较高。

3.1.2 该城区不同区域重金属的污染程度

我们引入一个区域污染程度指数 I_k 来衡量不同区域重金属的污染程度，定义如下：

$$I_k = \frac{1}{m}\sum_{i=1}^{8}(\omega_i \sum_{j=1}^{m} Y_{ij}) \tag{1}$$

其中 k 表示不同功能区，i 表示金属种类，m 表示金属 i 在功能区 k 内的样本个数。$Y_{ij} = \frac{x_{ij} - \overline{x_i}}{\sigma_i}$（$x_{ij}$ 为样本 j 中金属 i 浓度，x_i、σ_i 分别表示金属 i 浓度背景值的平均值和标准差），ω_i 为权重，由公式 $\omega_i = \frac{e_i}{\sum_{i=1}^{8} e_i}$ 求得，中 e_i 值见表 1[1].

表 1 八种重金属的毒性系数

重金属	Zn	Cr	Cu	Pb	Ni	As	Cd	Hg
毒性系数 e	1	2	5	5	5	10	30	40

我们运用 matlab 软件，对剔除异常值后的数据进行分析计算，得到的数值如下，见表 2：

表 2 五个功能区的区域污染指数

功能区	生活区	工业区	山区	交通区	公园绿地区
区域污染指数	6.0232	36.5382	0.7511	24.8282	6.5437

通过表 2 中的数值我们分析可以得到：污染程度从高到低依次是工业区、交通区、公园绿地区、生活区和山区，其中工业区的污染程度远大于其他功能区，而山区的污染程度远小于其他功能区，同时，交通区的污染程度也较大，比较而言，公园绿地的污染程度相对较小。

3.2 问题二的求解

根据上面的分析，我们可以发现各种重金属在城区的含量基本都大于背景值中的最大值，可以说明土壤表层重金属的含量受人类活动的影响。而人类活动引起的重金属污染大致可以分为三种类型：工业废水废渣污染、生活垃圾污染、交通废气污染。下面我们对八种重金属做来源解析。

3.2.1 土壤重金属元素之间的相关性

我们可以通过分析不同元素之间的相关性来衡量不同元素同源的可能性大小，一般情况下，若元素之间的相关性显著，说明它们来自同一污染源的可能性较大。Person 相关分析就是研究变量间相关程度的一种统计分析方法，通过计算在一定的显著性水平下的相关系数来衡量不同元素之间的相关性。Person 相关系数取值在-1~1 之间，绝对值越大，说明相关性越显著。

为了分析重金属污染的主要原因，我们可以计算八种重金属元素的 Person 相关系数，可以认为相关性高的重金属元素是同源的。陈怀满等人曾对我国土壤重金属污染的主要来源进行了总结表[2]（见表 3-1），表明在我国工矿生产、农业活动、污灌等人为活动都是造成土壤重金属污染的重要来源。然后进一步分析该类重金属污染的主要原因。

表 3-1 我国土壤重金属污染的主要来源

来源	重金属
矿产开采、冶炼、加工排放的废气、废水和废渣	Cr、Hg、As、Pb、Ni、Mo
煤和石油燃烧过程中排放的飘尘	Cr、Hg、As、Pb
电镀工业废水	Cr、Cd、Ni、Pb、Cu、Zn
塑料、电池、电子工业排放的废水	Hg、Cd、Pb、Ni、Zn
Hg 工业排放的废水	Hg
染料、化工制革工业排放的废水	Cr、Cd
汽车尾气	Pb
农药、化肥	As、Cu、Cd

我们运用 spss 软件，在显著性水平为 0.01 的情况下，对题目所给的八种重金属元素含量数据进行 person 相关分析，并对不同金属之间的相关性进行双侧检验，得到的相关系数如下表 3-2：

表 3-2 不同金属间的相关系数

相关系数	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn
As	1							
Cd	0.255	1						
Cr	0.189	0.352	1					
Cu	0.160	0.397	0.532	1				
Hg	0.064	0.265	0.103	0.417	1			
Ni	0.317	0.329	0.716	0.495	0.103	1		
Pb	0.290	0.660	0.383	0.520	0.298	0.307	1	
Zn	0.247	0.431	0.424	0.387	0.196	0.436	0.494	1

分析表 3-2，我们可以得到如下结论：

① 土壤中重金属 Cu、Cr、Ni 含量的相关性显著，其中 Cr、Ni 的相关性极其显著，表明它们来自同一污染源的可能性很大。

② 土壤中重金属 Pb 与 Cd、Pb 与 Cu、Zn 与 Pb 含量之间达到了显著的正相关，表明它们来自同一污染源的可能性很大。

③ 土壤中重金属 Pb 与 Cd、Cu、Pb 具有显著的相关性，表明土壤中 Pb 的污染源可能比较多、来源途径较多。

④ 土壤中重金属 Hg、As 与其它各重金属之间的相关性不强，表明土壤中 Hg、As 的来源可能比较独特，具有自己特有的污染源。

3.2.2 关联规则分析

若土壤表层中重金属的浓度大于背景值取值范围中最大值的浓度，我们将认为其受到污染，反之，则为无污染。将功能区作为前项，土壤表层某金属是否收到污染作为后项，利用关联规则挖掘，分析不同功能区与受各种金属污染的可信度，见表 4。

表 4 不同功能区中各种金属污染的关联规则

As 的关联规则分析

Rule	Confidence(%)
交通区==>As 污染	50
山区==>无 As 污染	83.33
生活区==>As 污染	68.18
公园区==>As 污染	74.29
工业区==>As 污染	61.11

Cd 的关联规则分析

Rule	Conifence(%)
交通区==>Cd 污染	76.81
山区==>Cd 无污染	74.24
生活区==>Cd 污染	72.73
工业区==>Cd 污染	86.11

Cr 的关联规则分析

Rule	Conifence(%)
山区==>无 Cr 污染	77.27
公园==>无 Cr 污染	77.14
生活区==>Cr 污染	43.18
工业区==>Cr 污染	38.89

Cu 的关联规则分析

Rule	Conifence(%)
交通区==>Cu 污染	81.88
山区==>无 Cu 污染	75.76
工业区==>Cu 污染	91.67
生活区==>Cu 污染	72.73
公园区==>Cu 污染	68.57

Hg 的关联规则分析

Rule	Conifence(%)
山区==>无 Hg 污染	95.45
交通区==>Hg 污染	40.58
工业区==>Hg 污染	52.78
生活区==>Hg 污染	40.91
公园区==>Hg 污染	40

Ni 的关联规则分析

Rule	Conifence(%)
交通==>无 Ni 污染	80.43
山区==>无 Ni 污染	81.82
公园==>无 Ni 污染	85.71
工业区==>Ni 污染	41.67
生活区==>Ni 污染	27.27

Pb 的关联规则分析

Rule	Conifence(%)
交通区==>Pb 污染	63.04
山区==>无 Pb 污染	83.33
工业区==>Pb 污染	80.56
生活区==>Pb 污染	56.82
公园==>无 Pb 污染	60

Zn 的关联规则分析

Rule	Conifence(%)
山区==>无 Zn 污染	87.88
公园==>无 Zn 污染	51.43
交通区==>Zn 污染	71.01
工业区==>Zn 污染	77.78
生活区==>Zn 污染	59.09

通过表 4 我们可以得到每种金属的污染主要原因，结论如下：

① As 污染的主要原因分析：在山区中无 As 污染的可信度为 83.33%，在公园绿地区、生活区和工业区的有 As 污染的可信度均大于 60%，这说明 As 主要来源于工业废水废渣、生活垃圾，由于公园绿地大多与工业区、生活区交叉分布，所以在公园绿地区 As 污染也较为严重。

② Cd 污染的主要原因分析：在山区中无 Cd 污染的可信度为 74.24%，在工业区、交通区和生活区的有 As 污染的可信度均大于 70%，这说明 Cd 主要来源于工业废水废渣、汽车尾气，由于生活区大多在工业区与交通区之间，所以在生活区 Cd 污染也较为严重。

③ Cr 污染的主要原因分析：在山区和公园绿地区中无 Cr 污染的可信度都约为 77%，在工业区和生活区的有 Cr 污染的可信度大约 40%，这说明 Cr 主要来源于工业废水废渣、生活垃圾。

④ Cu 污染的主要原因分析：在山区和公园绿地区中无 Cu 污染的可信度都约为 75.76%，在工业区、生活区和交通区的有 Cr 污染的可信度大约 70%，其中工业区高达 91.67%，这说明，工业废水废渣、汽车尾气、和生活垃圾的堆放都是 Cu 污染的重要原因，其中工业和交通的污染尤为重要。

⑤ Hg 污染的主要原因分析：在山区中无 Hg 污染的可信度高达为 95.45%，在工业区有 Hg 污染比较严重，另外在交通区，生活区和公园绿地区都有污染但较轻，这说明，Hg 的主要污染来源于工业。

⑥ Ni 污染的主要原因分析：在山区、交通区和公园绿地区中无 Ni 污染的可信度都大于 80%，工业区中有 Ni 污染的置信度为 41.67%，这说明 Ni 污染主要来源于工业废水废渣。

⑦ Pb 污染的主要原因分析：山区中无 Pb 污染的的可信度大于 80%，公园区无 Pb 污染的可信度为 60%，交通区和生活区有 Pb 污染的可信度均大于 55%，工业区中有 Pb 污染的可信度大于 80%，这说明 Pb 污染的主要来源于工业废水废渣和交通工具尾气的排放，而生活区主要分布在交通区的附近，所以生活区 Pb 污染也较为严重。

⑧ Zn 污染的主要原因分析：山区中无 Zn 污染的可信度高达 87.88%，在交通区和工业区中有 Zn 污染的可信度均大于 70%，生活区中有 Zn 污染的可信度约

为 60%，这说明 Zn 污染主要来源于工业废水废渣、汽车尾气的排放，还有一部分来源于生活垃圾。

综合各金属污染主要原因，我们可以进一步得到更多结论：

①无论何种金属在山区中无污染的可信度都很高，都大于 70%，说明山区受人类活动影响的程度最小；八种金属在工业区中有污染的可信度都很高，说明工业区受人类活动影响的程度最大，工业的废水废渣是污染的重要原因。

②Pb 和 Cu 在交通区和生活区的污染可信度均比较大，说明这两类金属同源，主要来源于工业废水废渣和交通工具尾气的排放。这与 3.2.1 中相似系数分析的结果一致。

③Pb、Zn 和 Cd 的污染均在山区和公园绿地去较少，而在其他几个区较高。说明这三类金属同源，主要来源于工业废水废渣、汽车尾气，还有一部分来源于生活垃圾的堆放。这与 3.2.1 中相似系数分析的结果一致。

④ Cu、Cr、Ni 的污染均在工业区和生活区较多，说明这两类金属同源，主要来源于工业废水废渣和生活垃圾的堆放。这与 3.2.1 中相似系数分析的结果一致。

3.3 问题三的求解

3.3.1 模型的求解

由于定性分析重金属污染物的传播特征需要较强的专业知识背景，而所给数据的变量不多（变量仅坐标、高度与功能区），统计学方法又不足以完全揭示变量与污染物传播特征的关系。一方面，人工神经网络具有较强的自组织、自适应与自学习能力，能够在未完全了解重金属污染物传播机理的情况下，完成自变量、变量间与重金属污染物浓度之间的非线性映射；另一方面，将重金属污染传播过程看作网络输入与输出的一类非线性映射，但是仅通过网络的输入输出数据难以准确寻找重金属污染物浓度的极值（即污染源），而遗传算法具有全局的非线性寻优能力。综合上述考虑，我们建立神经网络结合遗传算法求解重金属污染源的数学模型。

1、BP神经网络的建立污染传播模型

1.1 BP神经网络原理简介

BP神经网络是一种多层前馈神经网络，该网络的主要特点是信号前向传递，误差反向传播。在前向传递中，输入信号从输入层经隐含层逐层处理，直至输出层。每一层的神经元状态只影响下一层神经元状态。如果输出层得不到期望输出，则转入反向传播，根据预测误差调整网络权值和阈值，从而使BP神经网络预测输出不断逼近期望输出[3]。

1.2数据预处理

原始的319组数据显然不能满足神经网络的训练要求，因此，我们从问题一中插值过后的60501组数据中随机选择10000组数据。前9900组作为BP网络的训练数据集，后100组数据作为BP网络的检验数据集。

数据归一化可以方便后面数据的处理，并保证程序运行时收敛加快。我们将数据集按如下公式进行归一化处理：

$$y=\frac{(y_{\max}-y_{\min})(x-x_{\min})}{x_{\max}-x_{\min}}+y_{\min} \tag{2}$$

其中，$y_{\max}=1$，$y_{\min}=-1$,x为需要归一化的数据集。

matlab中函数“mapminmax”实现了上述归一化方式。

命令为[inputn,inputps]=mapminmax(input_train)；其中，inputn为归一化后的数据集，inputps为原始数据集信息，input_train为原始数据集。

1.3建立BP网络

虽然在问题一中将原有数据经过插值可以比较准确的反应重金属浓度随地理位置的变化，但是插值过程中并没有考虑地形对重金属污染物传播的影响。所以，我们建立3—N—1的BP网络结构，其中，3表示输入项（分别为坐标x，坐标y，高度z）；N为隐藏层神经元个数；1表示输出项（重金属污染物浓度）。结构图如下：

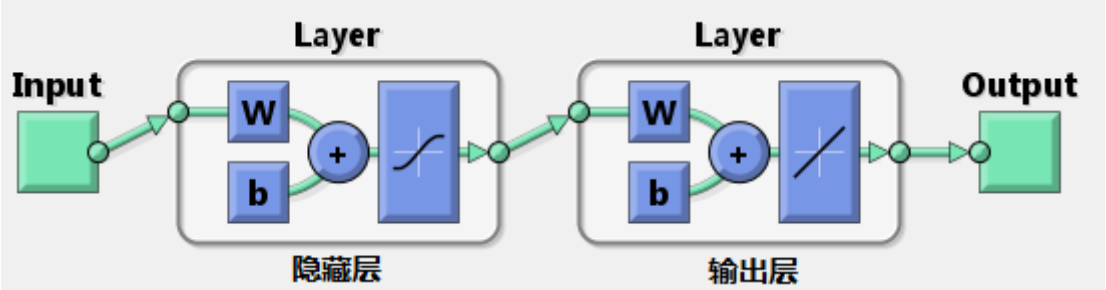

图3 BP神经网络结构图

隐藏层传输函数选择双曲正切S形函数：$a=\frac{e^{n}-e^{-n}}{e^{n}+e^{-n}}$

输出层传输函数采用线性函数：$a=n$

隐藏层神经元个数对BP神经网络预测精度有显著的影响，节点数太少,网络不能很好地学习,需要增加训练次数,训练的精度也受影响；节点数太多,训练时间增加,网络容易过拟合[3]。我们参考如下公式来确定最适隐藏层神经元个数。

$$l<n-1 \tag{3}$$

$$l<\sqrt{(m+n)}+a \tag{4}$$

$$l=\log_2 n \tag{5}$$

式中,n 为输入层节点数;l为隐含层节点数;m 为输出层节点数;a 为0~10之间的常数。在实际问题中,隐含层节点数的选择首先是参考公式来确定节点数的大概范围,然后用试凑法确定最佳的节点数。经过多次试验，我们选择N=100，此时BP神经网络达到了较高的精度。

学习速度同样对BP神经网络具有重要影响作用，学习速度太小，网络学习缓慢，需要增加训练次数；学习速度太大，网络学习迅速，但是容易导致网络不收敛，影响训练的精度。我们最终决定学习速度为0.01，训练次数为300。

BP神经网络的采用梯度修正法作为权值和阈值的学习算法,从网络预测误差

的负梯度方向修正权值和阈值,没有考虑以前经验的积累,学习过程收敛缓慢。对于这个问题,可以采用附加动量方法来解决,带附加动量的权值学习公式为:

$$w(k)=w(k-1)+\Delta w(k)+a\,[w(k-1)-w(k-2)] \tag{6}$$

matlab神经网络工具箱中函数“traingdm”即实现了上述学习方式。

1.4结果输出

训练好的BP神经网络还只能输出归一化后的浓度数据，为了得到真实的数据值，我们还必须对输出数据进行反归一化。反归一化过程可以利用归一化过程中的信息，通过函数“mapminmax”来实现。具体如下：

BPoutput=mapminmax('reverse',an,outputps);

其中，BPoutput为反归一化后的数据，an为神经网络预测输出，outputpa为原始输出数据集信息。

2BP神经网络结果分析

以Cu元素为例，训练结束的神经网络性能图如下：

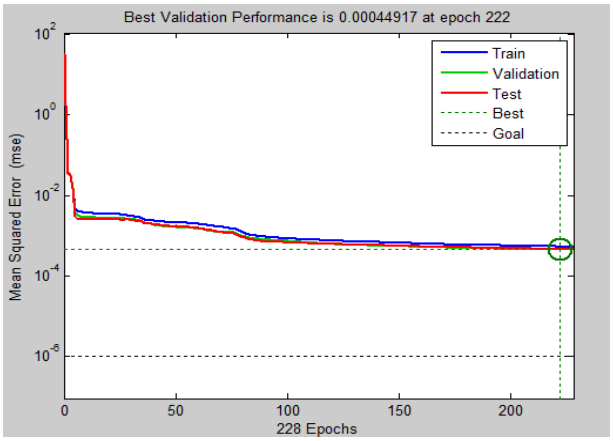

图4 BP神经网络训练性能

如图4，训练在第222次迭代过程达到均方误差最小，MSE=0.00044917。此时，训练结束。

用训练好的神经网络预测重金属污染，如图5。

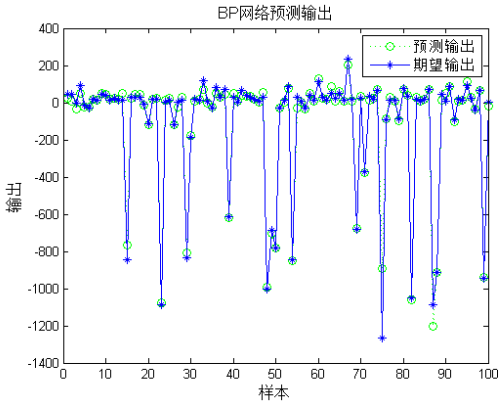

图5-1BP网络预测输出

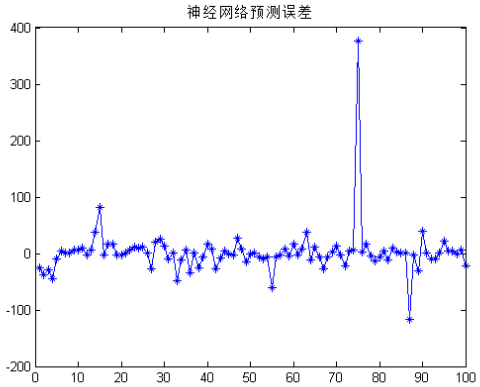

图5-2 神经网络预测误差

可见，预测输出与期望输出相差不大，基本达到了我们要求的效果。

下图为神经网络的预测误差。

结合图5-1分析，神经网络在个别点上误差较大。但由于污染物浓度随空间变化总是渐变的，这类误差不会影响我们寻找污染源。

进一步，我们用训练好的神经网络模拟出了整个城区各种金属的污染传播特征图，如图6：

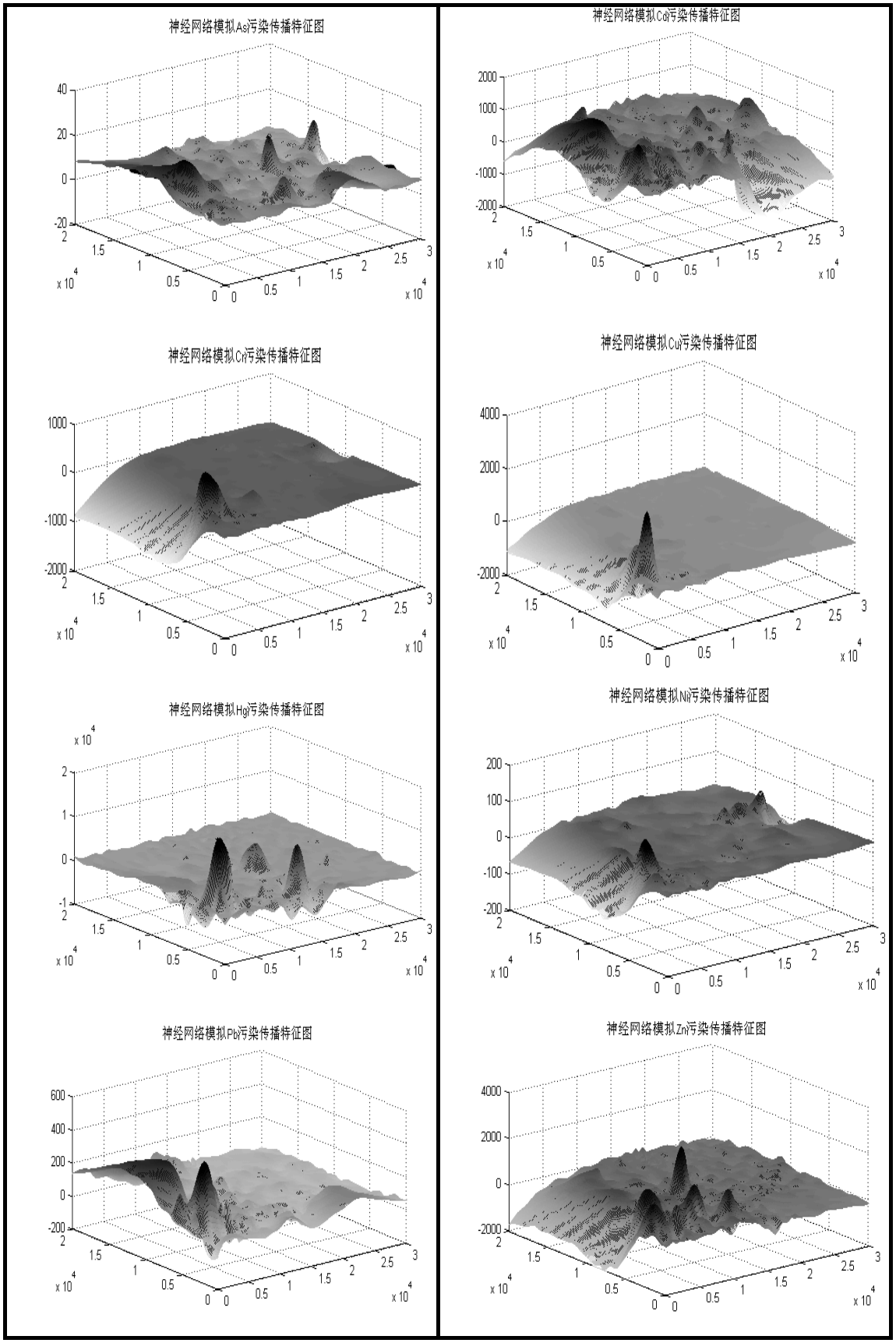

图6 神经网络模拟八种重金属污染分布特征

2、遗传算法求解模型

训练好的神经网络可以准确地模拟各种重金属污染传播的特征，但是由于神经网络是一个“黑箱”，仅通过网络的输入输出数据难以准确寻找重金属污染物浓度的极值（即污染源），我们在此引入遗传算法，通过区域分割的方法(某些污染物可能不止一个污染源)寻找污染源。算法流程图[3]如下：

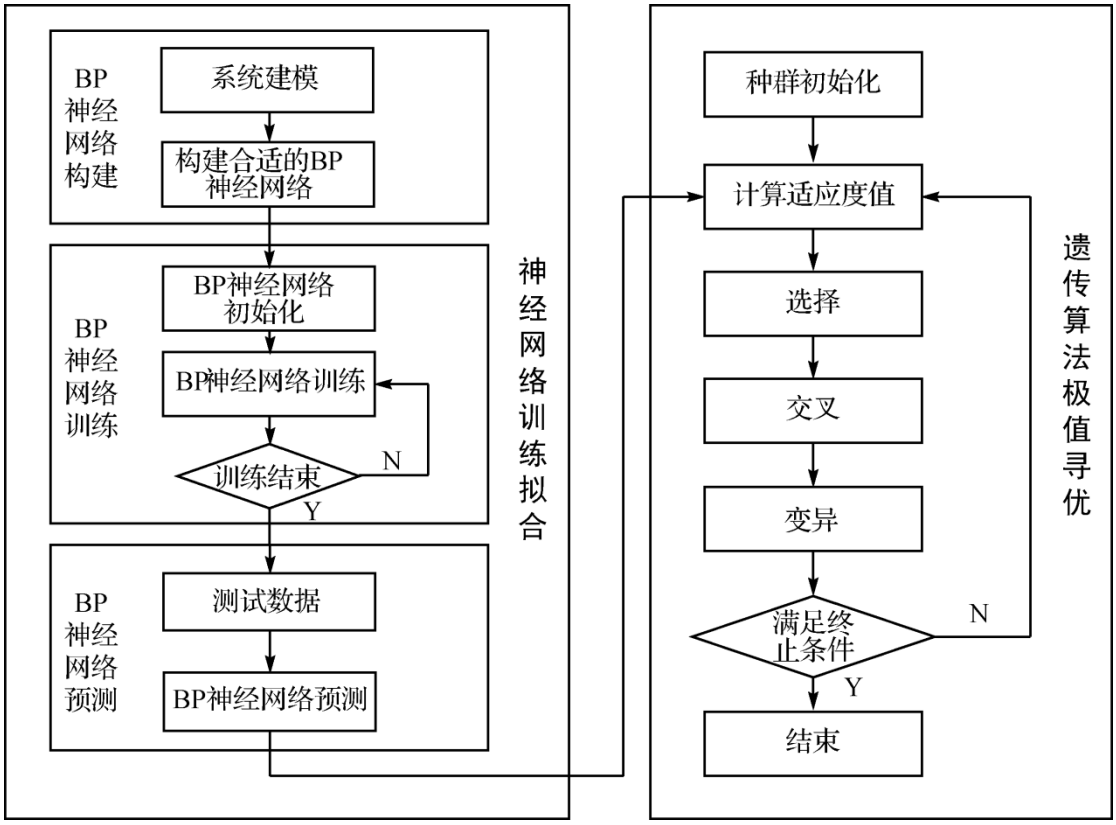

图7 算法流程图

2.1区域划分

不同的重金属污染源可能不同，同时重金属的污染源也可能不止一个。基于遗传算法只能寻找到全区污染最严重的区域的考虑，我们将功能区分布图划分为9块，分别在每块图中利用遗传算法寻找污染源，然后对比表五神经网络模拟出来的污染传播特征图，确定污染源的位置。

区域划分图如下所示：

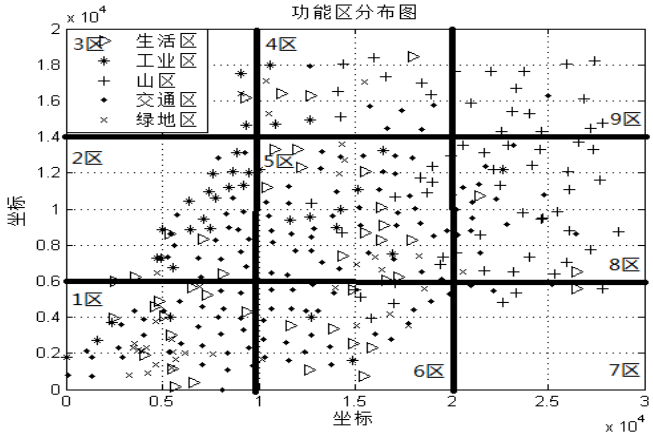

图8 功能区分布图

2.2遗传算法编码

神经网络要求有3个输入项，分别为坐标x，坐标y，高度h。但是对于某一固定点（x，y），此处的高度是不能随机更改的，因此，我们在遗传算法编码的时候仅对（x，y）进行编码，编码采用二进制方式，长度为10。对于已经生成的个体（x，y），我们通过取整将其固定到问题一插值生成的网格点之上，用该网格点的高度值作为此个体的高度。

2.3 程序中使用的遗传算法算子

以下算子均来自英国设菲尔德大学编写的GA工具箱[4]

- 选择算子：采用随机遍历抽样，其调用函数为“sus”；
- 交叉算子：采用单点交叉，其调用函数为“xovsp”；
- 变异算子：采用一定几率简单变异，其调用函数为“mut”；
- 重插算子：设置代沟为0.9，每次将适应度最高的90%保存下来，其余10%通过重插生成，其调用函数为“reins”；

2.4遗传算法结果分析

下面给出了遗传算法在区域一中，查找的污染源的性能图像，算法最后返回最大浓度Max_density=2.701940704296362e+003；坐标x=2101.7；坐标y=3390.0。

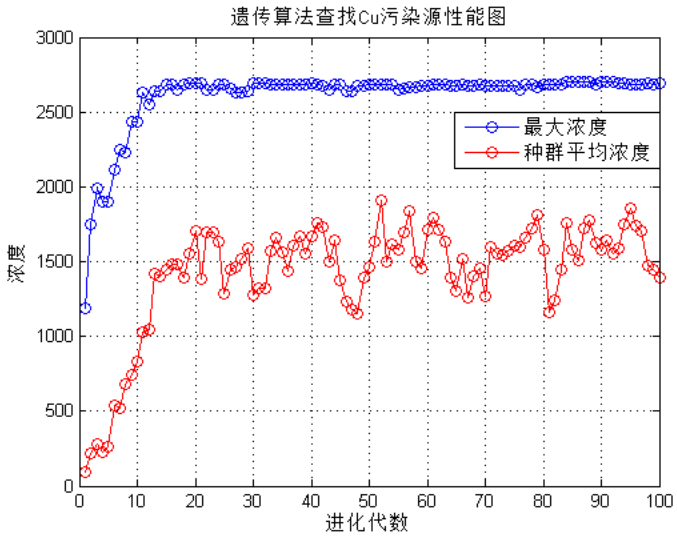

图9 遗传算法查找Cu污染源性能图

结合图8，我们可以确定Cu在区域1中的污染源在（2101,3390）附近，其浓度为2702ug/g。分析样本点，我们发现样本点中浓度最大值出现在（2383,3692）处，并且最大值为2528ug/g。由此能说明该模型能较为准确地反映污染源位置。分析其他重金属，

下表为遗传算法的结果并结合图6综合分析得到的污染源中心位置，见表5

表5 土壤中重金属污染源中心坐标和浓度

As			Cd			Cr			Cu		
x	y	浓度	x	y	浓度	x	y	浓度	x	y	浓度
1200	3006	13.2	2105	2593	1267	3487	5325	868	2101	3390	2701
18508	10206	19.5	4753	10875	1247	–	–	–	3198	5822	812
27680	12111	17.5	17450	3825	1034	–	–	–	–	–	–
–	–	–	21301	11467	813	–	–	–	–	–	–

（续表）

Hg			Ni			Pb			Zn		
x	y	浓度	x	y	浓度	x	y	浓度	x	y	浓度
2643	2875	15620	3211	5686	109	1991	3329	481	1699	2867	1403
13868	2354	10530	24001	12366	50	4508	5412	307	3725	5487	1558
15243	9186	5703	–	–	–	–	–	–	9583	4512	1439
–	–	–	–	–	–	–	–	–	13653	9655	2064

对该结果我们进行进一步验证，观察样本点中最大值点的位置和浓度，比较发现八种重金属通过模型得到的污染源的位置基本处于样本点中最大值点的附近，反映该模型对污染源位置的寻找具有较好的准确性；同时，计算得到的Cu和Pb在污染源位置的浓度高于样本点浓度的最大值，效果较佳；但是As在污染源位置的浓度低于样本点浓度的最大值，效果略差。

综合分析所得污染源所在位置，发现不同金属的污染源有同源现象，依据同源性汇聚污染源，得到了五个主要污染源，其中一个污染源呈带状分布，绘制了八种重金属的污染源汇总图。如下：

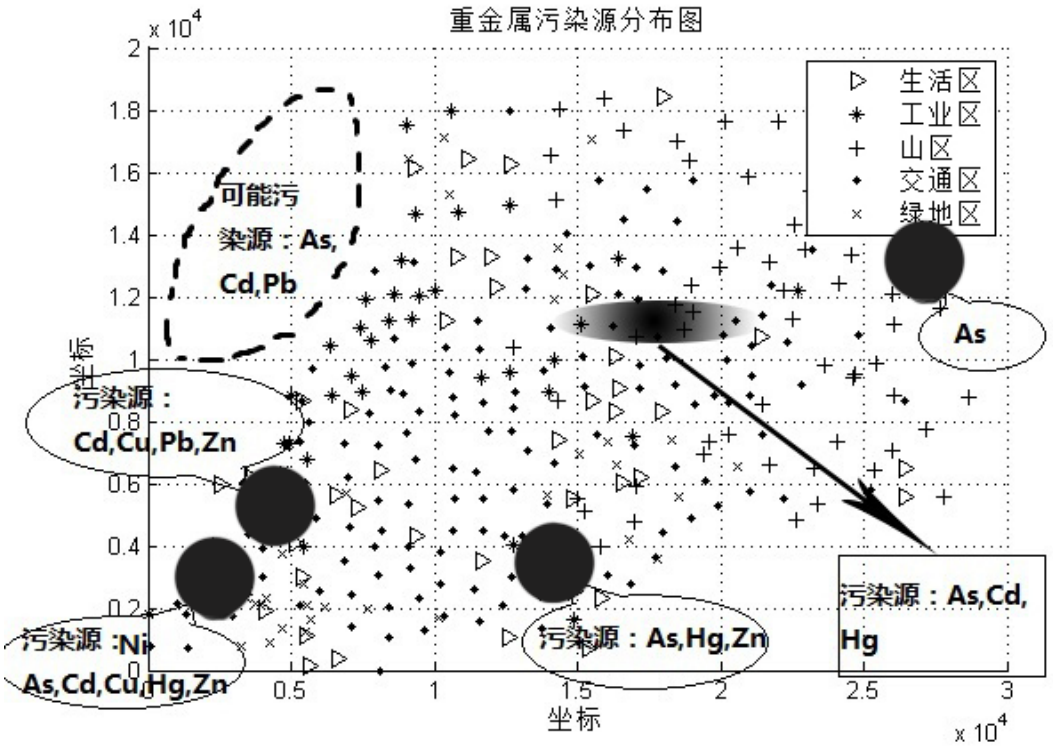

图10 重金属污染源分布图

另外，我们在图6中的神经网络模拟污染传播特征图中发现重金属As、Cd和Pb在城区外同一地区出现污染浓度高值中心。因为城区边界的原始样本值浓度向城区外方向是逐渐增大的，所以，我们推测这里可能存在As，Cd，Pb的污染源，见图10中虚线所围区域。

3.4 问题四的求解

3.4.1 模型的优缺点

模型的优点：

①第一问中虽然用三角形线性插值法也能拟合出各重金属的空间分布情况，但三角形线性插值法只利用了原始数据的 x、y 坐标来做插值拟合，忽略了各样本点高度的信息，由于重金属浓度分布也是受地形影响的（如地形影响地下水径流，将地势高处的水溶重金属带到地势低处，从而在地势低处堆积），所以分布结构不够精确。而神经网络模型利用了采样点的位置坐标（x、y）和高度坐标 h，能更精确地拟合出重金属的空间分布情况。

②由于浓度分布是多种因素共同影响的结果，很难准确地给出各种因素共同作用的公式。神经网络很好地解决了这个问题。神经网络具有较强的自组织、自适应与自学能力，能够在未完全了解重金属污染物传播机理的情况下，完成自变量、因变量之间的非线性映射。

模型的缺点：

由于神经网络构建所需数据量较大，而题目所给原始数据只有 300 多个，此数据量较小，所以不能更加精确地拟合出重金属的空间分布情况。

3.4.2 可能相关的有效信息

研究城市地质环境的演变模式，即在本文章的背景下就是研究八种重金属在土壤表层中浓度分布的演变模式。其中我们主要研究的是人类活动影响下城市地质环境的演变模式。

为了更好地研究此模式我们还需要搜集前几年该城区八种重金属浓度的采样数据。而人类影响表层土壤重金属浓度的因素主要有工厂分布多少位置的变化、交通路段车流量的变化、人口及生活区分布变化和植被分布多少位置的变化等。所以我们还应搜集近几年工厂分布多少位置的变化、交通路段车流量的变化、人口变化、生活区分布变化和植被分布多少位置的变化的信息来研究城市地质环境的演变模式。

3.4.3 模型的建立

首先，我们利用前几年该城区八种重金属浓度的采样数据，运用第三问所建立的神经网络的模型求解出该城区每年八种重金属的污染源。

然后，利用近几年工厂分布多少位置的变化、交通路段车流量的变化、人口变化、生活区分布变化和植被分布多少位置的变化的信息来分析人类活动对该城区八种重金属污染源的影响。具体做法可以是：把各污染源的年分布变化与上述人类影响因子在该污染源附近的年分布变化一一进行相似性分析，相似系数高的因素对该污染源变化影响贡献大。

最后，对影响每种重金属污染源大的人类影响因子进行汇总整合，从而得到该城市地质环境的演变模式。

参考文献：

[1] 徐争启，潜在生态危害指数法评价重金属毒性系数分析[J]，地球科学与技术报，第 31 卷，第 2 期：2008.
[2] 林艳，基于地统计学与 GIS 的土壤重金属污染评价与预测[D]，中南大学，2009.
[3] matlab 中文论坛编注，matlab 神经网络 30 个案例分析，北京：北京航天航空大学出版社，2010.
[4]雷英杰，matlab 遗传算法工具箱及应用，西安：西安电子科技大学出版社，2005.
[5] 夏立红，土壤污染及其防治，上海：华东理工大学出版社，2001.

附录

```
%由于程序较多，仅给出 Cu 元素的相关程序，其余程序相应修改即可！
function readdata()
data=xlsread('data','sheet1','B4:E322');
save data data  %坐标信息
data1=xlsread('data','sheet2','B4:I322');
save data1 data1 %浓度信息
beij=xlsread('data','sheet3','B4:C11');
save beij beij %背景信息


%**************************************************
%程序功能：画 Cu 污染分布等值线图
%建立日期：2011/9/9
%程序员：Hec1990
%**************************************************
function drawcu()

clc,clear
load data
load data1
Cu=data1(:,4);
X=data(:,1);
Y=data(:,2);
Cu_XY=[X Y Cu];
[x y]=meshgrid(0:100:3e4,0:100:2e4);
z=griddata(Cu_XY(:,1),Cu_XY(:,2),Cu_XY(:,3),x,y,'v4');
mesh(x,y,z)
%ezsurfc(z,x,y)
figure
[c h]=contour(x,y,z);
clabel(c,h)
hold on

[c1 d1]=find(data(:,4)==1);
x1=data(c1,1);
y1=data(c1,2);

plot(x1,y1,'r*');
```

```
[c2 d2]=find(data(:,4)==2);
x2=data(c2,1);
y2=data(c2,2);

hold on
plot(x2,y2,'k*')
[c3 d3]=find(data(:,4)==3);
x3=data(c3,1);
y3=data(c3,2);

plot(x3,y3,'c*')
[c4 d4]=find(data(:,4)==4);
x4=data(c4,1);
y4=data(c4,2);

plot(x4,y4,'y*')
[c5 d5]=find(data(:,4)==5);
x5=data(c5,1);
y5=data(c5,2);

plot(x5,y5,'g*')
grid on
legend('等值线','生活区','工业区','山区','交通区','绿地区');
cu=reshape(z,301*201,1);
a=reshape(x,301*201,1);
b=reshape(y,301*201,1);
save cu a b cu  %训练神经网络数据集
save dist c1 c2 c3 c4 c5   %不同区域

%**************************************************
%程序功能：比较不同区域污染指数
%建立日期：2011/9/9
%程序员：Hec1990
%**************************************************
function comp()
load data1
load beij
for i=1:8
    a(:,i)=(data1(:,i)-beij(i,1))/beij(i,2);
end

As=a(:,1);Cd=a(:,2);Cr=a(:,3);Cu=a(:,4);Hg=a(:,5);Ni=a(:,6);Pb=a(:,7);Zn=a(:,8)
;
```

```
load dist

As1=As(c1,1);Cd1=Cd(c1,1);Cr1=Cr(c1,1);Cu1=Cu(c1,1);Hg1=Hg(c1,1);Ni1=Ni(c1,1);P
b1=Pb(c1,1);Zn1=Zn(c1,1);
As2=As(c2,1);Cd2=Cd(c2,1);Cr2=Cr(c2,1);Cu2=Cu(c2,1);Hg2=Hg(c2,1);Ni2=Ni(c2,1);P
b2=Pb(c2,1);Zn2=Zn(c2,1);
As3=As(c3,1);Cd3=Cd(c3,1);Cr3=Cr(c3,1);Cu3=Cu(c3,1);Hg3=Hg(c3,1);Ni3=Ni(c3,1);P
b3=Pb(c3,1);Zn3=Zn(c3,1);
As4=As(c4,1);Cd4=Cd(c4,1);Cr4=Cr(c4,1);Cu4=Cu(c4,1);Hg4=Hg(c4,1);Ni4=Ni(c4,1);P
b4=Pb(c4,1);Zn4=Zn(c4,1);
As5=As(c5,1);Cd5=Cd(c5,1);Cr5=Cr(c5,1);Cu5=Cu(c5,1);Hg5=Hg(c5,1);Ni5=Ni(c5,1);P
b5=Pb(c5,1);Zn5=Zn(c5,1);

aa=[10 30 2 5 40 5 5 1];
w=[aa(1)/sum(aa)  aa(2)/sum(aa)  aa(3)/sum(aa)  aa(4)/sum(aa)  aa(5)/sum(aa)
aa(6)/sum(aa) aa(7)/sum(aa) aa(8)/sum(aa)];

y1=[mean(As1)  mean(Cd1)  mean(Cr1)  mean(Cu1)  mean(Hg1)  mean(Ni1)  mean(Pb1)
mean(Zn1)];
y2=[mean(As2)  mean(Cd2)  mean(Cr2)  mean(Cu2)  mean(Hg2)  mean(Ni2)  mean(Pb2)
mean(Zn2)];
y3=[mean(As3)  mean(Cd3)  mean(Cr3)  mean(Cu3)  mean(Hg3)  mean(Ni3)  mean(Pb3)
mean(Zn3)];
y4=[mean(As4)  mean(Cd4)  mean(Cr4)  mean(Cu4)  mean(Hg4)  mean(Ni4)  mean(Pb4)
mean(Zn4)];
y5=[mean(As5)  mean(Cd5)  mean(Cr5)  mean(Cu5)  mean(Hg5)  mean(Ni5)  mean(Pb5)
mean(Zn5)];
qu1=sum(w.*y1);qu2=sum(w.*y2);qu3=sum(w.*y3);qu4=sum(w.*y4);qu5=sum(w.*y5);



%**************************************************
%程序功能：训练 BP 网络并测试网络性能
%建立日期：2011/9/10
%程序员：Hec1990
%修改时间：2011/9/11
%**************************************************
clc,clear;
load cu a b cu
load h h
data=[a';b';h'];
data=data';
data1=cu;
k=rand(1,60501);
```

```
[m n]=sort(k);

%获取训练数据与预测数据
input_train=data(n(1:9900),1:3)';
output_train=data1(n(1:9900),1)';
input_test=data(n(9901:10000),1:3)';
output_test=data1(n(9901:10000),1)';

%数据归一化
[inputn,inputps]=mapminmax(input_train);
[outputn,outputps]=mapminmax(output_train);
%% bp 训练

%初始化网络结构
net=newff(inputn,outputn,100);

net.trainParam.show=30;
net.trainParam.epochs=300;
net.trainParam.lr=0.01;
net.trainParam.goal=1e-6;

%网络训练
net=train(net,inputn,outputn);

%% bp 预测
%预测数据归一化
inputn_test=mapminmax('apply',input_test,inputps);

%网络预测输出
an=sim(net,inputn_test);

%网络输出反归一化
BPoutput=mapminmax('reverse',an,outputps);

%% 结果分析

figure
plot(BPoutput,':og');
hold on
plot(output_test,'-*');
legend('预测输出','期望输出','fontsize',12)
title('BP 网络预测输出','fontsize',12)
xlabel('样本','fontsize',12)
ylabel('输出','fontsize',12)
```

```
%预测误差
error=BPoutput-output_test;

figure
plot(error,'-*')
title('神经网络预测误差')


errorsum=sum(abs(error))
save data2 net inputps outputps


%************************************************
%程序功能：BP 网络模拟 Cu 传播
%建立日期：2011/9/11
%程序员：Hec1990
%************************************************
load data2 net inputps outputps
[x y]=meshgrid(0:100:3e4,0:100:2e4);
load h h
%h=reshape(h,201,301);
x=reshape(x,201*301,1);
y=reshape(y,201*301,1);
s=[x y h];
s=s';
input_test=mapminmax('apply',s,inputps);

a=sim(net,input_test);

ObjV=mapminmax('reverse',a,outputps);

x=reshape(x,201,301);
y=reshape(y,201,301);
ObjV=reshape(ObjV,201,301);

mesh(x,y,ObjV);
title('神经网络模拟 Cu 污染传播特征图')


%************************************************
%程序功能：遗传算法寻找局部极值（污染源）
%建立日期：2011/9/11
%程序员：Hec1990
```

```
%**********************************************
clc
clear
NVAR=2;
NIND=40;                    %种群大小
MAXGEN=100;                 %最大遗传代数
LIND=10;                    %个体长度
GGAP=0.9;                   %代沟
trace=zeros(2,MAXGEN);
FieldD=[LIND LIND ;
        0    0   ;
        1e4   0.6e4   ;
        1    1    ;
        0    0    ;
        1    1    ;
        1    1    ];
Chrom=crtbp(NIND,LIND*NVAR);          %初始种群，二进制编码
unit=bs2rv(Chrom,FieldD);       %二进制转为十进制
ObjV=fun(unit);
for gen=1:MAXGEN
    FitnV=ranking(-ObjV);       %分配适应度
    SelCh=select('sus',Chrom,FitnV,GGAP);   %选择 sus:随机遍历抽样
    SelCh=recombin('xovsp',SelCh,0.7);      %交叉（重组）概率 0.7,xovsp:单点交叉
    SelCh=mut(SelCh);                       %变异
    unit=bs2rv(SelCh,FieldD);               %二进制十进制转换
    ObjVSel=fun(unit);                      %目标函数值
    [Chrom ObjV]=reins(Chrom,SelCh,1,1,ObjV,ObjVSel); %重插入子代的新种群
    [Y(gen),I(gen)]=max(ObjV);              %Y 为最优个体下的目标函数值,I 为最优
个体在种群的下标
    trace(1,gen)=max(ObjV);                 %遗传算法性能追踪
    trace(2,gen)=sum(ObjV)/length(ObjV);

end

unit=bs2rv(Chrom,FieldD);

figure
plot(trace(1,1:gen),'b-o');
hold on
plot(trace(2,1:gen),'r-o');
grid on

max(Y)
```

```
%************************************************
%程序功能：遗传算法目标函数程序
%建立日期：2011/9/11
%程序员：Hec1990
%************************************************
function ObjV=fun(unit)
%下载训练好的神经网络
load data2 net inputps outputps

load cu a b   %下载格点资料
load h h

xy=floor(unit(:,1:2)/100)*100;
cc=[a b];
for i=1:length(xy)
    for j=1:length(cc)
        if xy(i,:)==cc(j,:);
            unit(i,3)=h(j);
        end
    end
end

unit=unit';
input_test=mapminmax('apply',unit,inputps);

a=sim(net,input_test);
ObjV=mapminmax('reverse',a,outputps);

ObjV=ObjV';
```