

慈溪市表层土壤养分与重金属的多变量统计分析^{*}

刘庆¹, 吕振江²

(1. 滨州学院 黄河三角洲生态环境山东省重点实验室, 山东 滨州 256603; 2. 西北农林科技大学 林学院, 陕西 杨陵 712100)

摘要:地球化学变量在土壤中受到多个因素与过程影响使得它们之间的关系非常复杂。利用直方图、正态 Q - Q 图、数据转换、聚类分析等多变量分析方法,对浙江省慈溪市表层土壤样品中的重金属、养分以及土壤 pH 值进行了调查与分析。结果认为:变量之间存在着显著的变异,其中直方图表现出数据多模态的特点与正态 Q - Q 图的多扭结的特点表明了这些数据中多重总体的存在,正态 Q - Q 图还显示数据存在异常值。大部分变量并没有通过 K - S 正态分布检验或对数正态分布检验,Box - Cox 变换使得数据更接近正态分布,是一种更有效的正态转换方式。聚类分析的结果认为所有变量根据影响因素的不同可分为 3 个组分。在所有的影响因素中,土地利用被认为是最重要的影响因素。

关键词:土壤地球化学变量;数据转换;多变量分析

中图分类号:S153.61

文献标识码:A

文章编号:1005-3409(2009)06-0106-06

Statistical Analysis of Nutrients Variables and Heavy Metals in Surface Soils of Cixi County

LIU Qing¹, L Ü Zhen-jiang²

(1. Key Laboratory of Eco-environmental Science of Yellow River Delta of Shandong Province, Binzhou University, Binzhou, Shandong 256603, China; 2. College of Forestry Science, Northwest A & F University, Yangling Shaanxi 712100, China)

Abstract: Geochemical variables are affected by multiple factors and processes in soils, and the different responses to these factors causing complicated multivariate relationships between them. In this study, the relationships between heavy metals and nutrients variables in top soils of Cixi county were investigated using multivariate analyses, the methods are histograms, normal quantile-quantile (Q - Q) plots, data transformation and Cluster analysis. The results showed that there was strong variation in the values of these variables. Multi-model features in the histograms and multi-kink features in the normal quantile-quantile (Q - Q) plots were observed for these variables implying the existence of multiple populations. Obvious outliers were identified using the normal Q - Q plots. Most of the variables did not pass a Kolmogorov - Smirnov test for either normal or lognormal distribution, and the Box - Cox transformation was effective in transposing data to a form suited to further parametric statistical analyses. Cluster analysis classified the variables into three groups and variables that are affected by different factors such as agricultural activities and industrial pollutions. Land use was found to be the most important influencing factor in studied area.

Key words: soil geochemistry variables; data transformation; multivariate analyses

土壤地球化学变量受多种因素的影响,包括小尺度上的矿物组成以及区域尺度上的地质状况、土壤类型和地形特征等^[1-3]。大多研究认为:地球化学变量大多遵循对数正态分布而很少遵循正态分布^[5-6],并且大多伴有正偏态效应。近年来,有研究者把对数正态分布看作是环境地质领域普遍存在的

正偏态分布的特殊的案例^[7]。但非正态和非对数正态分布在地球化学数据库中亦大量存在^[5-6,8]。由于许多统计分析要求数据符合正态分布或对数正态分布,地球化学变量的这种非正态和非对数正态分布的特征为其进一步的统计分析提出了一个挑战。因此,在对一些地球化学变量进行统计分析之前,有

^{*} 收稿日期:2009-05-26

基金项目:国家科技攻关计划项目“小城镇土地集约利用与生态安全评价研究”(2003BA808A22 - 4);山东省滨州学院“博士研究生奖励基金”(2007 Y07)

作者简介:刘庆(1972 -),男,山东菏泽人,博士,主要研究方向为土壤生态与环境。E-mail: qy7271 @163.com

必要了解这些变量的概率分布特征。

地球化学变量在土壤中受到多个过程的影响使得它们之间的关系变得更为复杂^[1-4]。因此,当面对来自于较大区域受不同土壤类型、土地利用类型与人类影响的大量样品的数据进行统计分析时,这种分析就显得特别有意义。目前,多元分析技术在环境科学研究领域已被广泛应用^[7,9-14]。这种应用为揭示多变量之间的关系,相互之间的影响以及化学组分来源提供了一个有效的方法。本研究利用浙江省地球化学调查数据,利用统计工具对其中的养分、重金属以及 pH 等指标的地化特征进行分析,同时对这些指标进行相关分析和聚类分析,以便更好地理解各变量之间的相互关系及其影响因素。

1 研究区概况

慈溪市地处浙东杭州湾南岸,东离宁波 60 km,北距上海 148 km,西至杭州 138 km,位于东经 121°02' - 121°42',北纬 30°21' - 30°24',全境东西长约 55 km,南北最宽处约 30 km,全境总面积 1 700 多 km²,陆域面积约 1 074 km²,是长江三角洲经济圈南翼环杭州湾地区上海、杭州、宁波三大都市经济金三角的中心,是全国经济发展速度最快的地区之一。慈溪市土地资源特点是平原多,山地少,平原面积约占陆域面积的 85%,山地丘陵区面积约占陆域面积的 15%。慈溪市虽然平原广阔,但由于人口增长、城镇等建设用地的快速增加,人均占有耕地面积不多。

2 材料与方法

2.1 数据来源

本研究中所使用的数据为浙江省慈溪市地球化学调查数据。该数据土壤样品采样时采用网格法均匀布点采样,样品全部采集自 0 - 20 cm 表层土,用 GPS 对每一个采样点精确定位,采样密度为 4 km² 一个采样点。土壤样品各属性指标的分析均采用国家标准方法进行。

2.2 分析方法

本研究的统计分析通过统计软件为 SPSS13.0 来实现。所使用的统计分析方法主要包括基础统计、直方图、正态 Q - Q 概率图、数据转换、聚类分析等。需要指出的是,此类研究要求样品采集时随机进行采样,而本研究中的采样方法为网格法系统采样,这可能会对分析结果产生一定的影响。

3 结果与分析

3.1 基础统计

土壤中各元素含量的最低值、中位数、最高值以及百分位数统计结果如表 1 所示。从表 1 可见,各

元素在土壤中含量的原始数据显示了其在土壤的变化范围较大,各元素在土壤中这种较大的变化显示了研究区域内土壤性质的较大差异,由于土壤中各指标因子较大差异的存在,因此,推荐使用中位数来代表某种元素在土壤中的含量。

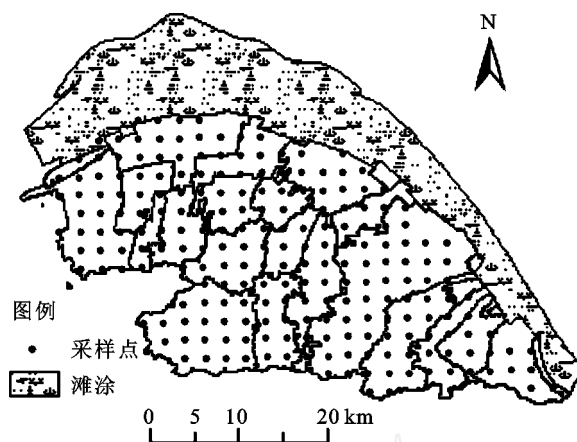


图 1 研究区采样点位图

3.2 直方图

各变量在土壤中含量的直方图见图 2。从其直方图上可以看出,大部分的变量显示出向右倾斜的正偏态分布特征,这种分布特征表明了土壤大多数元素受到自然因素和人类活动的双重影响。从直方图上还可以看出,土壤 pH 值呈现出双峰分布,反映了该区两种截然不同性质的土壤类型的存在,这可能会导致土壤中元素分布的多模式的出现。

从直方图上可以直接观察到的另一个特征是少数元素呈现出正态分布特点。但是在环境地球化学领域,非正态分布占主要地位的认识已得到广泛的认同^[5-7],这可能和元素含量受到自然与人为等多种复杂因素共同影响有关。

3.3 原始数据的正态 Q - Q 图

为了更好地了解数据分布的概率特征,利用原始数据所做的正态 Q - Q 图见图 3。正态概率图中的点是由数据中的每一个样本的观测值(X 轴坐标)与其正态分布的期望值(Y 轴坐标)所组成。这些点落在斜线上的越多,则说明数据分布就越接近正态。如果被检验的变量值的分布与已知分布基本相同,那么在 Q - Q 概率图中的散点应该围绕在一条斜线的周围,如果两种分布完全相同,那么在 Q - Q 概率图中点应该与斜线重合^[15]。从 Q - Q 概率图上看,大部分变量呈现偏离正态分布的特征。需要注意的是,个别变量显示出与众不同的分布特征,如 pH 值显示了多扭结的特点,这表明研究区存在两种不同类型的土壤环境,并且数据存在多总体的特征,但是结点的个数和总体的个数并不一定相同。

表 1 土壤各变量的基础统计分析

变量	最小值	10 %分位数	25 %分位数	中值	75 %分位数	90 %分位数	最大值	标准差	变异系数/ %
Cu/ (mg · kg ⁻¹)	1.000	17.840	21.950	29.200	37.475	46.900	252.800	24.37	73.3
Zn/ (mg · kg ⁻¹)	47.000	64.000	72.000	85.000	96.250	114.300	25500	24.99	28.4
Pb/ (mg · kg ⁻¹)	16.000	19.000	21.000	25.000	33.000	39.000	75.000	8.65	31.1
Cd/ (mg · kg ⁻¹)	0.067	0.108	0.123	0.142	0.164	0.191	0.302	0.04	23.9
Cr/ (mg · kg ⁻¹)	23.000	52.700	63.000	72.000	80.000	86.300	183.000	16.9	23.9
As/ (mg · kg ⁻¹)	3.500	4.900	5.800	6.600	7.700	9.400	15.500	1.88	27.3
Hg/ (mg · kg ⁻¹)	0.026	0.044	0.059	0.082	0.158	0.258	0.967	0.12	95.0
OrgC/ %	0.280	0.540	0.670	0.830	1.153	1.670	2.660	0.46	47.1
K ₂ O/ %	2.020	2.147	2.250	2.420	2.590	2.846	3.360	0.29	11.6
P/ %	0.162	0.537	0.723	0.915	1.111	1.224	1.804	0.28	30.8
pH	4.610	5.653	6.785	8.130	8.350	8.680	10.010	1.18	15.5

注 :样本数均为 226。

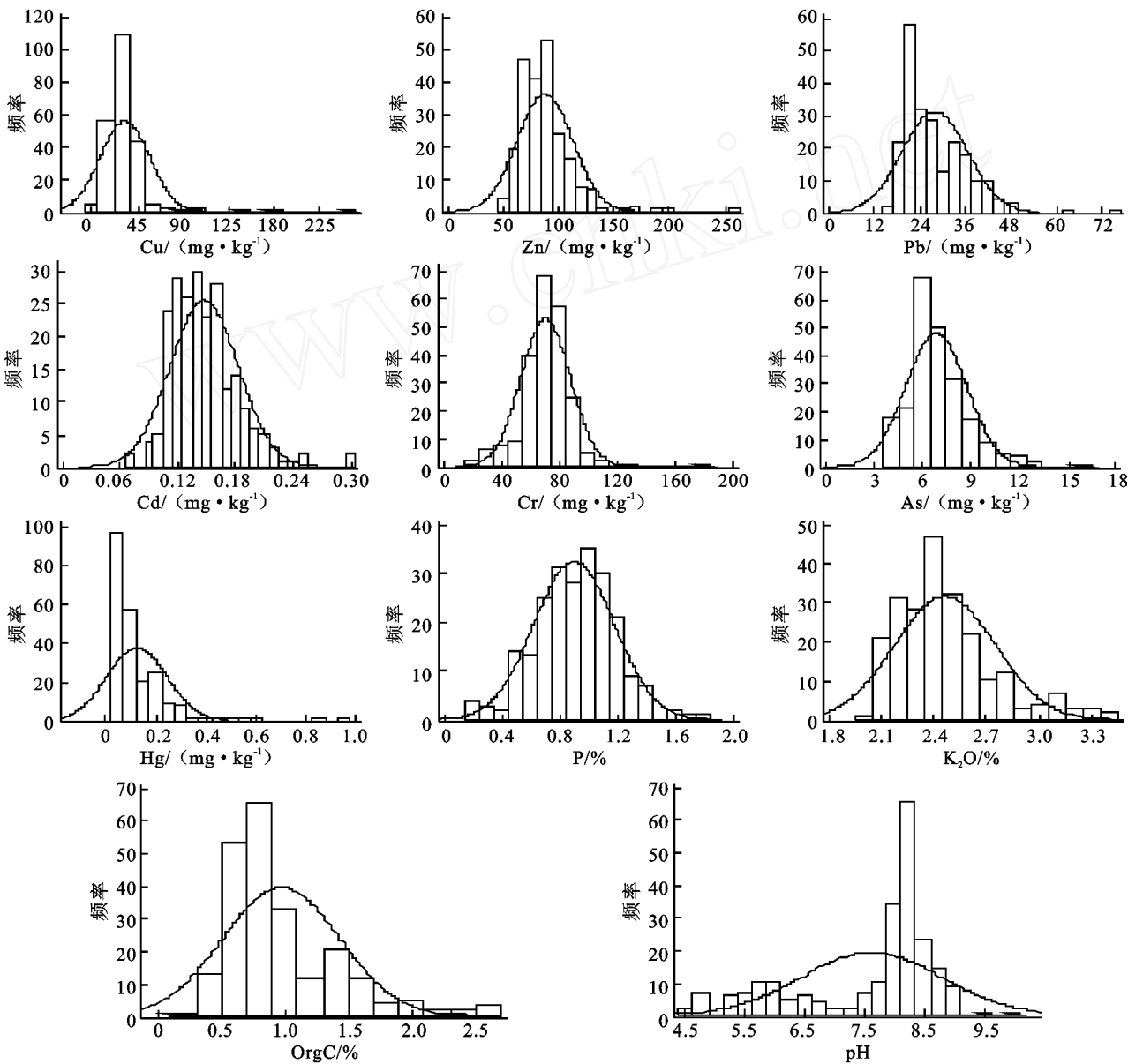


图 2 原始数据的频数分布图

正态 Q - Q 图的另一个特点是从图上可以非常直观的看到数据中所存在极端最大值。如 Cu、Hg、Pb 等。对于数据集中的极值,在统计分析时应从数据中除去,因为这些极值对数据的整体特征影响较

为强烈,并可能对一些较为敏感的、反映数据最终的统计结果的指标带来偏差。如平均值、变异系数以及相关系数等很容易受到极值的影响。

3.4 数据正态转换与检验

通过直方图和正态 Q - Q 图两种方法对数据的表征结果,发现各变量大多并不符合正态分布,因此,需要将数据进行适当的转换以便进行进一步统计分析。本研究利用对数变换和 Box - Cox 两种数据变换方式进行处理,分析结果见表 2。

偏度系数和峰度系数是反映数据图形形状的两个重要参数。其计算结果见表 2。偏度系数的大小反映是分布图形的不对称程度,而峰度系数反映了数

据分布峰相对于正态分布峰的平坦或陡峭程度。大部分的原始数据,其分布规律是正偏态分布,表明高于平均值的数据占较大的比例,而较陡的分布峰表明,多数样本的数据分布集中在相对较低值的区域。

由表 2 可见,对数变换的结果显示出比原始数据更低的偏态值,利用 Box - Cox 变换使得其偏度系数更接近于 0,数据更接近正态分布。由于样本的多总体、不同的检出限、多结点与土壤样品较多等原因,经 Box - Cox 转换以后,仍有许多的变量没有通过正态性检验($P > 0.05$)。然而,通过对比表明,Box - Cox 变换使原始数据更接近正态分布,改善原始数据的对称性方面,是更有效的数据变换方式。

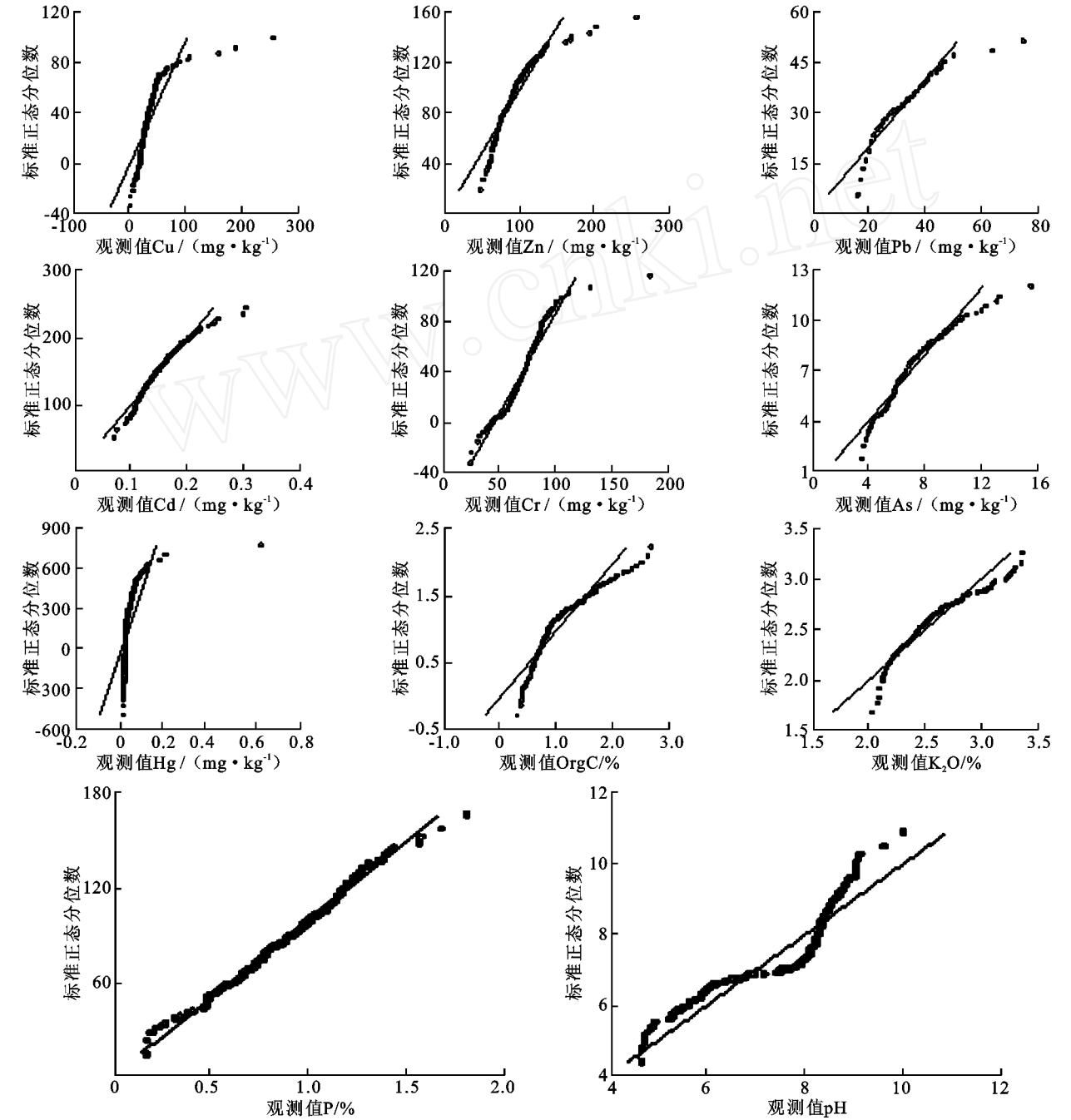


图 3 原始数据的正态 Q - Q 图

表 2 数据转换及其统计分析结果

变量	样本数	原始数据			对数转换			Box - Cox 转换		
		Skewness	Kurtosis	K - Sp	Skewness	Kurtosis	K - Sp	Skewness	Kurtosis	K - Sp
Cu	226	5.211	37.948	0.000	- 1.206	9.120	0.002	- 0.065	6.435	0.009
Zn	226	2.498	11.282	0.001	0.874	2.236	0.186	0.150	0.487	0.358
Pb	226	1.479	4.023	0.000	0.559	- 0.106	0.006	- 0.008	- 0.842	0.025
Cd	226	1.091	2.515	0.098	0.146	0.704	0.832	0.146	0.704	0.833
Cr	226	0.969	8.868	0.051	- 1.250	4.164	0.001	- 1.250	4.164	0.001
As	226	1.208	2.500	0.002	0.233	0.470	0.182	0.224	0.430	0.178
Hg	226	9.876	121.539	0.000	1.075	2.009	0.008	- 0.044	- 0.447	0.278
OrgC	226	1.392	1.869	0.000	0.307	- 0.091	0.073	0.307	- 0.091	0.073
K ₂ O	226	1.083	0.962	0.013	0.800	0.319	0.086	0.124	- 0.655	0.451
P	226	- 0.060	0.335	0.964	- 1.601	4.269	0.036	- 0.060	0.335	0.964
pH	226	- 1.049	- 0.037	0.000	- 1.269	0.492	0.000	- 0.590	- 0.400	0.000

3.5 聚类分析

由于对含有较多变量的数据总体进行相关分析的结果相对复杂,所以为了更清晰地表示各变量之间的相互关系,本研究通过 R 型聚类的方法对所有变量进行进一步的分析,分析结果见图 4。据聚类分析的结果,将参与聚类的所有变量归为 3 类。

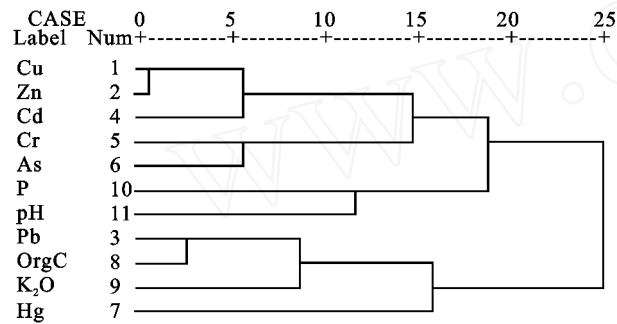


图 4 聚类分析树形结构图

第一类变量主要包括 Cu、Zn、Cd、Cr、As 5 种重金属元素,据调查,这些变量在研究区土壤中的含量均有部分样点超出当地背景含量。这说明,该 5 种元素在当地表层土壤中的含量除受土壤母质中元素含量的影响之外,还可能受当地工业生产中的废气、废液的排放所导致的点源污染有关。因为 5 种元素均为重金属元素,并且该组变量中没有养分变量,因此综合分析认为,5 种元素受农业生产的影响不是很大。

第二类主要包括 P 和 pH 值两个变量。这两个变量的主要影响因素可能和地质条件与人类活动有关。由前面的分析亦可看出,慈溪市土壤 pH 值正态 Q - Q 图上显示的多扭结特点显示当地有不同 pH 值土壤存在,而土壤 pH 值的发育主要受土壤母质的影响。因此,综合分析认为,这两个指标主要是受土壤母质的影响,农业生产对土壤中磷的影响次于土壤母质,而工业生产中的排污对其影响不大。

第三类主要包括 Pb、OrgC、K₂O 和 Hg 四个变

量,其中的 OrgC 和 K₂O 是土壤肥力的指示指标,其含量应和农业生产的投入分不开的。而土壤中 Pb 和 Hg 的来源可能与大气沉降有关外。除此之外,从 Pb 和 Hg 与土壤 OrgC 和 K₂O 的关系分析,土壤中 Pb 和 Hg 的含量还与农业生产中的污灌以及施用含 Pb 和 Hg 的有机肥或生物有机肥有关。

4 结论

(1) 由于受土壤类型、土地利用类型以及各种自然和人为因素的综合影响,慈溪市表层土壤中养分和重金属含量具有不同的变异特征,大部分变量既不符合正态分布也不符合对数正态分布,其含量的中值被认为最具代表性的取值。

(2) 通过正态 Q - Q 图可很容易地鉴别出特异值,从而在做进一步的统计分析时将其除去。数据的转换有可能使数据的分布特征由正偏态变为负偏态,而 Box - Cox 转换可能更有效地对数据实现向正态分布的转换,并可以增强数据分布的对称性。

(3) 聚类分析较相关分析更为清楚地反映变量之间的相互关系,通过对所有变量的聚类分析,并综合考虑各变量的地球化学特征,将所有变量归为 3 类。第一类元素主要包括 Cu、Zn、Cd、Cr、As 5 种重金属元素,其在土壤中的含量受土壤母质与工业生产的“三废”排放关系密切。第二类主要包括 P 和 pH 值,这两个变量的主要影响因素可能和地质条件有关,人为活动对这两个指标的影响相对较弱。第三类主要包括 Pb、OrgC、K₂O 和 Hg 4 个变量,其中的 OrgC 和 K₂O 两个变量和当地的农业生产关系密切。土壤中 Pb 和 Hg 的来源可能是受点源和非点源污染的共同影响,其具体的影响因子有待进一步调查研究。

(4)通过对所有元素的分析可以判断,慈溪市表层土壤中元素的含量受母质与人类活动的综合影响。但综合分析后认为,土地利用应该是最重要的影响因素。

参考文献:

- [1] 朱立新. 中国东部沿海平原区土壤地球化学调查方法研究[D]. 长春: 吉林大学, 2003.
- [2] 龚子同. 土壤地球化学的进展和应用[M]. 北京: 科学出版社, 1985: 11.
- [3] 胡以铿. 地球化学中的多元分析[M]. 北京: 中国地质大学出版社, 1991: 8.
- [4] 曾道明, 纪宏金, 陈满, 等. 胶东山城金矿地质与地球化学变量的关系[J]. 吉林大学学报, 2006, 36(4): 511-515.
- [5] Reimann C, Filzmoser P. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data[J]. Environmental Geology, 2000, 39(9): 1001-1014.
- [6] Zhang C S, Manheim, F T, Hinde J, et al. Statistical characterization of a large geochemical database and effect of sample size[J]. Applied Geochemistry, 2005, 20(10): 1857-1874.
- [7] Zhang C S, Selinus O. Statistics and GIS in environmental geochemistry-some problems and solutions[J]. Journal of Geochemical Exploration 1998, 64(1): 339-354.
- [8] McGrath S P, Loveland O J. The Soil Geochemical Atlas of England and Wales, (SGAEW) [M]. London: Blackie Academic & Professional, 1992.
- [9] Gallego J L R, Ordonez A, Loredó J. Investigation of trace element sources from an industrialized area (Avilés, northern Spain) using multivariate statistical methods[J]. Environment International, 2002, 27(7): 589-596.
- [10] Reimann C, Filzmoser P, Garrett R G. Factor analysis applied to regional geochemical data: problems and possibilities [J]. Applied Geochemistry, 2002, 17(3): 185-206.
- [11] Zhang C S, Lalor G. Multivariate relationship and spatial distribution of geochemical features of soils in Jamaica[J]. Chemical Speciation and Bioavailability, 2002, 14(1): 57-65.
- [12] Lee C S, Li X D, Shi W Z, et al. Thornton, I. Metal contamination in urban, suburban, and country park soils of Hong Kong: a study based on GIS and multivariate statistics[J]. Science of The Total Environment, 2006, 356(1/3): 45-61.
- [13] Micó C, Recatalá L, Peris M, et al. Assessing heavy metal sources in agricultural soils of an European Mediterranean area by multivariate analysis [J]. Chemosphere, 2006, 65(5): 863-872.
- [14] Zhang C S. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway Ireland[J]. Environmental Pollution, 2006, 142(3): 501-511.
- [15] 余建英, 何旭宏. 数据统计分析与 SPSS 应用[M]. 北京: 人民邮电出版社, 2003: 4.

(上接第 105 页)

参考文献:

- [1] 汤国安, 杨勤科, 张勇, 等. 不同比例尺 DEM 提取地面坡度的精度研究[J]. 水土保持通报, 2001, 21(1): 53-56.
- [2] 汤国安, 刘学军, 房亮, 等. DEM 及数字地形分析中尺度问题研究综述[J]. 武汉大学学报: 信息科学版, 2006, 31(12): 1059-1066.
- [3] 孔德树. “3S”技术在水土保持工作中的应用及展望[J]. 中国水土保持, 2005(5): 40-42.
- [4] 朱红春, 汤国安, 张友顺, 等. 基于 DEM 提取黄土丘陵区沟沿线[J]. 水土保持通报, 2003, 23(5): 114-117.
- [5] 赵帮元, 喻权刚, 郭玉涛. 建立黄土丘陵区 DEM 的方法探讨[J]. 人民黄河, 2002, 24(4): 33-34.
- [6] 陈晖, 张红, 徐高洪, 等. 基于遥感、DEM 技术的西汉水水土流失变化分析[J]. 人民长江, 2006, 37(12): 12-15.
- [7] Valeo C, Moins S M A. Grid - resolution effects on a model for integrated urban and rural areas[J]. Hydrological Process, 2000, 14: 2505-2525.
- [8] Kenward T, Lettenmaier D P, Wood E F, et al. Effects of digital elevation on model accuracy on hydrologic predictions [J]. Remote Sensing of Environment, 2000(3): 432-444.
- [9] 吴险峰, 刘昌明, 王中根. 栅格 DEM 的水平分辨率对流域特征的影响分析[J]. 自然资源学报, 2003, 18(2): 148-154.
- [10] 赵帮元, 汤国安, 马安利, 等. 不同地貌类型区 1:25 万比例尺 DEM 的建立方法[J]. 水土保持通报, 2002, 22(2): 45-48.
- [11] Quinn P F, Beven K J, Lamb R. The $\ln(a/p)$ index: how to calculate it and how to use it with in the TOPMODEL framework[J]. Hydrological Processes, 1995(9): 161-182.
- [12] Bruneau P, Gascuel - Odoux C, Robin P, et al. Sensitivity to space and time resolution of a hydrological model using digital elevation data [J]. Hydrological Processes, 1995(9): 69-81.