

聚类分析在山西省盐碱土土属 分类中的应用初探*

李文银 张永武
(山西农业大学)

摘 要

应用数学方法解决土壤学及其分类问题在国内正处于探索阶段,本文收集了 40 个盐碱土剖面的分析资料,经过一系列前处理工作,应用聚类分析法进行盐碱土基础分类单元的探讨。因此,计算了样品各变量间的相关系数、样品间的距离系数,根据计算结果,建立了相应的聚类谱系图,形成了盐碱土土属的样品间和变量间的分类系统,反映了样品间和变量间的亲疏序列。经聚类分析后,在 0.08 的距离水平下,把 40 个样品的原 7 个土属聚为 5 个;在 0.6 的相关水平下,将选取的 9 个变量分为 4 群。

聚类分析是多元分析的方法之一,它通过对样品各变量以及样品间的一系列处理,建立聚类谱系图,从而直观清晰地反映了土壤样品以及土壤类型之间的亲疏序列,并能反映变量间的相互关系,以显示土壤属性的多因素综合特征,从而合理准确地区分土壤类型。

本文收集了近二百个盐碱土剖面分析结果,经计算整理,选用 40 个剖面样品进行盐碱土土属的数值分类。40 个样品主要收集于山西北部(大同、怀仁)和中部(太谷、祁县)盆地,海拔高度分别在 960—1100 米和 700—800 米之间,处于干旱草原向半干旱森林草原过渡地带,年均气温分别为 7.4 和 10.1℃,蒸降比分别为 5.14 和 4.40,地带性土壤分别为淡栗钙土和淡褐土,境内河流发育,地下水位多在 0.5—2.5 米之间,地下水矿化度多在 3—15 克/升之间;盐碱土多分布在山前洼地,交接洼地和河流一级阶地局部洼地,成土母质以次生黄土、红黄土和白垩纪冲洪积物为主,由于地理条件和高差不同,两盆地土壤的区域特征仍有较大差别。

盆地四周山地土壤的淋溶和流失,使盐基物质向盆地聚积,为盐碱土的发育和形成提供了物质基础;地下水的高水位和高矿化度,是盐碱土形成的主要因素;年均蒸发量大于年降雨量 4—5 倍,致使土壤内水盐运行较长时期处于上升过程,盐分在地表大量积聚。

一、盐碱土土属各变量间的相互关系

每个样品选取 9 个变量,分别作了数据标准化、相关分析、聚类分析等项处理。显示

* 在本研究工作中,边俊生、李艳琴、赵德龙同志参加了数据计算;赵真工程师提供了部分资料并予以指导;林成谷、亓毓吉、张毓庄教授给予帮助和支持,特此一并致谢。

了在一定盐碱土类型中, 变量间的多元效应和单元效应。因此便于在改良利用过程中抓住主要问题所在, 促使其向预定的便于利用的方向转化, 建立合理的区域性土壤生态系统。现将常用统计值列于表 1, 说明盐碱土各变量数据特征。

表 1 盐碱土土属各变量数据特征统计

Table 1 The statistics of the characteristics of variables of the genera of salt-affected soil

统计值 Statistics values	变量项目 Variables								
	全盐量 (%)	CO ₃ ²⁻ (meq/ 100g)	HCO ₃ ⁻ (meq/ 100g)	Cl ⁻ (meq/ 100g)	SO ₄ ²⁻ (meq/ 100g)	Ca ⁺⁺ (meq/ 100g)	Mg ⁺⁺ (meq/ 100g)	Na ⁺ + K ⁺⁺ (meq/ 100g)	pH (1:1)
最大值 (x max)	1.159	7.613	3.285	8.375	9.378	5.818	4.677	18.93	10.31
最小值 (x min)	0.050	0.000	0.000	0.047	0.020	0.040	0.060	0.70	7.87
极差 (x max-x min)	1.109	7.613	3.285	8.325	9.358	5.778	4.671	18.23	2.44
平均值 (\bar{x})	0.382	1.287	0.857	2.415	2.561	1.436	1.216	4.484	8.29
标准差 (S)	0.333	1.610	0.513	1.760	1.393	1.119	1.034	4.061	0.608
变异系数 (C.V.%)	87.17	125.1	59.86	72.68	54.39	77.92	85.03	90.60	7.33

* Na⁺ + K⁺ 按差数计。

(一) 将原始数据标准化

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{s_j}$$

$i = 1, 2, \dots, n$ (样品数)

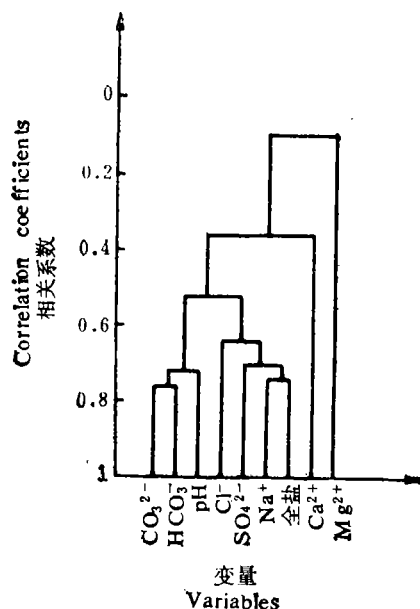


图 1 盐碱土土属变量间聚类谱系图

Fig. 1 Dendrogram of the variables of the genera of the salt-affected soil

$j = 1, 2, \dots, V$ (变量个数)

式中 \bar{x}_j 为 j 变量平均数; s_j 为 j 变量的标准差; x_{ij} 为第 i 样品第 j 个变量原始数据; x'_{ij} 为经标准化后的新数据。

经过这样的变换后,使每个变量均处于同一量度。

(二) 求变量 j 与 k 的相关系数

$$r_{jk} = \frac{1}{n-1} \sum_{i=1}^n x'_{ij} x'_{ik}$$

式中 x'_{ij} 和 x'_{ik} 分别为第 i 个样品的第 j 和第 k 个变量的测定值 (经标准化变换); n 为样品数。计算结果列于表 2。

表 2 盐碱土土属变量间相关阵 (r 值*)

Table 2 The correlation matrix among the variables of the genera of the salt-affected soil (r value*)

变 量 Variables	变 量 Variables								
	全盐量 TSC	CO ₃ ²⁻	HCO ₃ ⁻	Cl ⁻	SO ₄ ²⁻	Ca ⁺⁺	Mg ⁺⁺	Na ⁺ +K ⁺	pH
全盐量(TSC)	1.000								
CO ₃ ²⁻	0.525	1.000							
HCO ₃ ⁻	0.434	0.765	1.000						
Cl ⁻	0.631	0.375	0.301	1.000					
SO ₄ ²⁻	0.691	0.449	0.352	0.750	1.000				
Ca ⁺⁺	0.027	0.188	0.208	0.113	0.373	1.000			
Mg ⁺⁺	-0.271	-0.183	-0.020	-0.297	-0.246	0.202	1.000		
Na ⁺ + K ⁺	0.756	0.578	0.496	0.609	0.652	0.023	-0.252	1.000	
pH	0.597	0.727	0.583	0.448	0.448	-0.346	-0.100	0.467	1.000

* $r = 0.403$ $\alpha = 0.01$; $r = 0.312$ $\alpha = 0.05$; $d.f. = 38$

据此,把各变量的分群组合以直观的谱系图表示。

(三) 讨论

由变量间的相关分析结果看出,除 Ca 和 Mg 与其它变量间的相关关系处于不显著 (即 $r < 0.317$) 水准外,其它各变量间大都处于显著 (即 $r > 0.317$) 或极显著 (即 $r > 0.408$) 水平,盐碱土土属类型与各变量值具有多元函数关系。

1. 各变量与 pH 值的变化关系根据相关系数的绝对值的大小排列如下:

$$\text{CO}_3^{2-} > \text{HCO}_3^- > \text{全盐量} > \text{Na}^+ > \text{SO}_4^{2-} > \text{Cl}^- > \text{Ca}^{++} > \text{Mg}^{++}$$

其中 pH 值与 CO₃²⁻ 的回归方程为: $\text{pH} = 0.219 + 0.090\text{CO}_3^{2-}$

2. 全盐量与各个变量的相关系数绝对值排序如下:

$$\text{Na}^+ + \text{K}^+ > \text{SO}_4^{2-} > \text{Cl}^- > \text{CO}_3^{2-} > \text{HCO}_3^- > \text{Mg}^{++} > \text{Ca}^{++}$$

(Na⁺ + K⁺)/(Ca⁺⁺ + Mg⁺⁺) = 1.5 - 2.0 (按平均值计),在两盆地中,Na⁺ 是导致和加重土壤盐渍化程度的首要因素,因其以相当的数量存在,使土壤理化、生物性状濒于恶化。

消除过量的 Na^+ 及其危害是调节当地盐碱土各变量间数量平衡的主要手段。

3. 从九个变量的聚类谱系图看出, 在 0.6 的相似水平下把 9 个变量分为 4 群, 即 CO_3^{2-} 、 HCO_3^- 、 pH ; Na^+ 、 SO_4^{2-} 、 Cl^- 、全盐量; Ca^{++} 和 Mg^{++} 各自成一群。以此, 也反映了盐碱土土属各变量间的亲疏关系。

二、聚类分析在盐碱土土属分类中的应用

除进行盐碱土变量间相关关系讨论外, 着重作了盐碱土样品间的聚类分析, 以谋求定量地划分盐碱土土属类型。在考虑处理变量和其它条件相同的情况下, 拟比较原分类法和聚类分析法的结果。

(一) 求每个样品 0—30 cm 各个变量的加权平均值

$$\bar{x} = \frac{t_1 V_1 + t_2 V_2 + t_3 V_3}{t_1 + t_2 + t_3}$$

式中 V_1, V_2, V_3 分别为第一、二、三层变量值; t_1, t_2, t_3 分别为第一、二、三土层的厚度。

(二) 求样品间的距离系数

$$d_{ik} = \sqrt{\frac{\sum_{j=1}^V (x_{ij} - x_{kj})^2}{V}}$$

$i, k = 1, 2, \dots, n;$

式中 x_{ij} 和 x_{kj} 分别为第 i 个和第 k 个样品的第 j 个变量的测定值(经标准化变换); n 为样品数; V 为变量数。

根据计算结果, 说明盆地盐碱土土属样品间的亲疏关系。

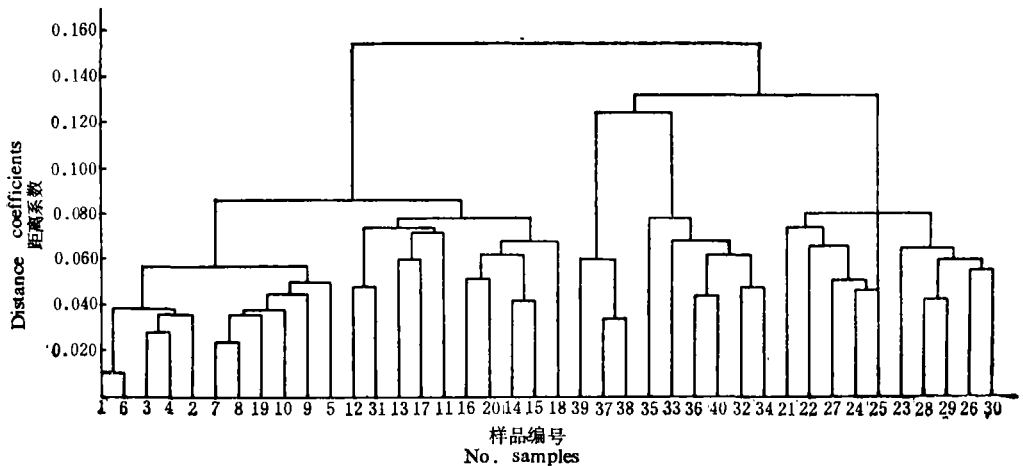


图 2 盐碱土土属样品间聚类谱系图

Fig. 2 Dendrogram of the samples of salt-affected soil

(三) 建立聚类谱系图

在距离系数中,数据愈小者,样品间关系愈相近,如 $d_{1,6} = 0.010$, 故在聚类中优先成群,然后由小到大依次联结,直到所有样品都分群完毕为止,形成一个样品间分类系统,并将分类结果以图表示(图 2)。

三、分类结果比较及其讨论

40 个盐碱土样品,经聚类处理后,在 0.08 的距离水平下,将原来的 7 个土属聚为五个,并将两种分类结果列表如下:

表 3 两种方法分类结果比较

Table 3 Comparison between results of different classification methods

原分类法(主、次离子命名法) Original classification		聚类分析法 Classification by cluster analysis		
样品号 Sample No.	土属名称 Genera name	样品号 Sample No.	土属名称 Genera name	土属距离水平值 The distance level value of genera
⊗1,2,3,4,5, 6,7,8,9,10	碱化草甸土 Alkalized meadow soil	⊗1,2,3,4,5, 6,7,8,9,10,19	$\text{CO}_3^{2-}\text{-Na}^+, \text{HCO}_3^-$ Na^+ 或 CO_3^{2-} 、 HCO_3^- - Na^+ 型	0.056
⊗11,12,13,14,15	硫酸盐苏打盐化草甸土 Sulphate-soda salinized meadow soil	⊗11,12,13,17, 31,16,18,	$\text{Cl}^-(\text{SO}_4^{2-})\text{HCO}_3^-$ $\text{Na}^+(\text{Mg}^{++})$ 型	0.074
⊗16,17,18,19,20	氯化物苏打盐化草甸土 Chloride-soda salinized meadow soil	14,15,20		
⊗21,22,23,24,25	氯化物硫酸盐盐化草甸土 Chloride-sulphate salinized meadow soil	⊗21,22,23,24, 25,26,27,	$\text{SO}_4^{2-}(\text{Cl}^-)\text{-Na}^+$ 型	0.08
⊗26,27,28,29,30	硫酸盐氯化物盐化草甸土 Sulphate-chloride salinized meadow soil	28,29,30		
⊗31,32,33, 34,35,36	硫酸盐盐化草甸土 Sulphate-salinized meadow soil	⊗32,33,34, 35,36,40	$\text{SO}_4^{2-}\text{-Na}^+$ 型	0.078
⊗37,38,39,40	草甸盐土 Meadow solonchak	⊗37,38,39	$\text{SO}_4^{2-}\text{Cl}^-$ - Na^+ 型	0.060

上述图表表明:

1. 在原有的 7 个土属中,经处理后聚为 5 个土属,即把原来的氯化物硫酸盐盐化草甸

土和硫酸盐氯化物盐化草甸土两个土属聚为一个土属。在山西的盐化土壤中, Cl^- 和 SO_4^{2-} 在大多数情况下都具有伴生关系共同存在于土壤中。从原始资料来看, 在两个土属中, Cl^- 和 SO_4^{2-} 互为主、次离子, 但两者的数值较相当, 按平均数计: $\text{Cl}^- = 2.415\text{meq}/100\text{g}$, $\text{SO}_4^{2-} = 2.561\text{meq}/100\text{g}$, 且具有一定的相关性, 其相关系数为 $r_{\text{Cl}^- \cdot \text{SO}_4^{2-}} = 0.750$, 处于极显著水平。在 CO_3^{2-} 和 HCO_3^- 含量极微的情况下, 阳离子中 Mg^{++} 和 Ca^{++} 的含量也很低, 多在 $1.5\text{meq}/100\text{g}$ 以下。以 Na^+ 为主, 多在 $4.5\text{meq}/100\text{g}$ 以上, 这样 $\text{SO}_4^{2-}(\text{Cl}^-)-\text{Na}^+$ 型土和 $\text{Cl}^-(\text{SO}_4^{2-})-\text{Na}^+$ 型都是强酸, 碱生成的强电解质, 性质也较为相似。另一个是把原分类法的硫酸盐苏打盐化草甸土和氯化物苏打盐化草甸土聚为一类, 其主要成分是 NaHCO_3 (或 Na_2CO_3) 两者的相关系数为 $r_{\text{CO}_3^{2-} \cdot \text{HCO}_3^-} = 0.756$, 次要成分分别是 Na_2SO_4 和 NaCl 另三个土属中, 除个别样品发生变动外, 土属类型仍分别是碱化草甸土、硫酸盐盐化草甸土和草甸盐土。

2. 在 5 个土属中, 可归纳为三个亚类, 即盐化草甸土、碱化草甸土和草甸盐土。两个土类, 即草甸土和盐土。在分群中, 只有 19、40 号共两个样品发生了亚类间的变动, 即分别由原来的(苏打)盐化草甸土、草甸盐土变为碱化草甸土和盐化草甸土, 占样品总数的 5%。由此可见, 作为一个完整的分类体系, 各分类单元间应该是互检的(此类问题尚需进一步探讨)。

3. 如前所述, 盐碱土土属类型取决于各变量数据特征及其综合效应。只有综合各变量在质和量上的差异, 才能揭示其内在的本质规律。如原分类法根据土属类型的盐分组成, 以某阴离子占总阴离子的毫克当量数的百分数为依据, 采取主、次离子进行分类命名。从其反映的离子毫克当量数的数量上来说, 主、次离子命名法在分类中只能体现全部离子毫克当量总数的 70% 左右。分类结果具有某种局限性和片面性, 难以反映盐碱土土属类型的多变量综合特征。聚类分析正是适应盐碱土土属分类的这一特征, 根据各变量的分析结果, 聚多种变量为一体, 达到准确、合理、定量地区分土壤盐渍化类型。

应用聚类分析进行盐碱土土属分类的过程中, 有关确定土壤类型的距离水平值、命名等问题有必要进行更深入的研究。

参 考 文 献

- [1] 席承藩等著, 1965: 山西省的盐渍土与盆地土壤。科学出版社。
- [2] 中国科学院计算中心概率统计组编著, 1979: 概率统计计算。科学出版社。
- [3] 南京大学教学系计算数学专业编, 1979: 概率统计基础和概率统计方法。科学出版社。
- [4] 刘多森, 1979: 主组元分析在分辨土壤类型及风化—成土过程上的应用。土壤学报, 第 16 卷 2 期, 172—182 页。
- [5] Richards, L. A., 1954: Diagnosis and improvement of saline and alkali soil. U.S. Department of Agriculture.
- [6] Webster, B.R., 1977: Quantitative and numerical method in soil classification and survey. clarendon press, Oxford.

THE PRELIMINARY INVESTIGATION ON THE APPLICATION OF CLUSTER ANALYSIS IN THE LOWER CATEGORY CLASSIFICATION OF SALINE-ALKALI SOIL IN SHANXI PROVINCE

Li Wenyin and Zhang Yongwu
(Shanxi Agricultural University)

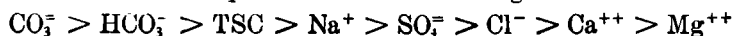
Summary

The present paper deals the application of the cluster analysis method in the basic categories of classification for salt-affected soils on the basis of the analytical data of forty profiles. The results obtained are:

1. According to the coefficients of distance matrix among the samples, the seven genera originally classified can be reclassified into five genera at 0.08 distance level through cluster analysis.

2. Under 0.6 distance level, the nine variables can be divided into the four groups, i.e. (1) CO_3^- , HCO_3^- , pH; (2) Na^+ , SO_4^- , Cl^- , TSC; (3) Ca^{++} and (4) Mg^{++} .

3. The correlation analysis shows that the absolute values of correlation coefficients between ions and pH values can be arranged as follows:



《中国科学院院刊》创刊

《中国科学院院刊》为中国科学院的机关刊物。主要任务是宣传党和政府关于发展科学技术事业的方针政策,评述科技发展趋势,增进社会各界对科学院的了解,沟通科学院与社会的联系,加强国际科学交流。

《中国科学院院刊》为季刊,主要栏目设有“科学发展”、“工作研究”、“科学与社会”、“科学家”、“院所介绍”、“书刊评介”、“国际交流”、“档案库”等。

《中国科学院院刊》由叶笃正任主编。

创刊号将于2月20日由科学出版社出版发行。