

对数比统计分析在黄土剖面化学组成中应用的研究

韦玉春¹, 张建明², 倪绍祥¹

(1. 南京师范大学地理科学学院, 南京 210097)

(2. 兰州大学地理系, 兰州 730000)

[摘要] 利用对数比统计分析方法, 对陕西洛川剖面化学组成数据进行了分析. 结果表明, 由于化学组成中存在的数据闭和关系, 经典统计学方法难以获得各组成之间的真实关系. 洛川黄土剖面中 SiO_2 和 CaO 之间的负相关关系是虚假的. 由对数比度主成分分析方法获得的化学组成主成分载荷, 与黄土剖面的淋溶和积聚序列相同, 具有明确的地球化学意义. 获得的得分图, 能够较好地表达样品之间的过渡关系.

[关键词] 化学组成; 黄土剖面; 对数比统计分析

[中图分类号] B151; [文献标识码] A; [文章编号] 1001-4616(2001)02-0114-05

地球化学数据是一类典型的组分数据. 组分数据由于各个样品具有的定和的特点, 歪曲了各个成分之间的关系. 利用经典的统计学方法难以发现成分之间的真实的关系, 这给数据分析带来了困难. 本文利用 Aitchison(1986)提出的对数比分析方法, 对陕西洛川剖面的各化学组成的相关关系进行了分析讨论.

1 组分数据的闭合效应

组分数据指每行(代表一个样品)中的各个变量(代表各个成分)的值都是正数并且每行的和为一个常数的数据, 其闭合效应主要表现在 3 个方面.

1. 无论各成分之间的真实相关性如何, 由于每个样品中各个成分的和为常数, 组分数据中至少有两个成分之间派生出负的相关关系.

2. 组分数据可带来伪相关, 符号可正可负. 研究表明, 派生的负相关和伪相关会极大地歪曲成分之间的真实相关性, 歪曲的方式随数据而变, 难以估计.

3. 组分数据的取值范围是固定的, 所在的空间为单形(Simplex), 所以, 基于实数空间正态分布假设的经典统计学方法不再适用.

对数比(logratio)统计分析是 Aitchison(1982, 1986)提出的用来解决组分数据闭合效应问题的一整套统计分析方法:

设含 D 种成分的组分数据为:

$$X = [x_{ij}] \quad i = 1, 2, \dots, m \quad j = 1, 2, \dots, D$$

则对数比变换为:

$$Y = [Y_{ij}] = [\log(x_{ij}/x_{iD})] \quad i = 1, 2, \dots, m \quad j = 1, 2, \dots, D$$

可以证明, 对数比变换后, 成分之间的关系不受闭合效应的影响, 所以能代表成分之间的真实

收稿日期 2000-07-06

基金项目: 国家自然科学基金资助项目(49871072)

作者简介: 韦玉春, 1965—, 南京师范大学地理科学学院博士研究生, 从事地理信息系统与土地科学的研究.

的关系.此外,变换后数据的取值范围为 $(-\infty, \infty)$,并且常常符合多元正态分布,因而可以使用经典的多元统计分析方法进行分析.

对数比变换带来了统计学上的“不对称性”问题.为此,Aitchison(1986)提出了中心化对数比变换:

$$Z = [Z_{ij}] = [\log(x_{ij}/g(x_i))],$$

其中

$$g(x_i) = (x_{i1} \cdot x_{i2} \cdots x_{iD})^{1/D}, i = 1, 2, \dots, n, j = 1, 2, \dots, D$$

变换后的协方差矩阵满足了“对称性”的要求,但有一个缺点,即矩阵是奇异的.但在多数情况下,这种奇异性是组分数据的统计学标志.

数据采用中心化对数比变换后,形成的协方差矩阵适用于主成分分析,由此形成的方法,称为对数衬度主成分分析(logcontrast PCA,简称为LPCA).

2 分析方法和数据特征

数据的分析使用LPCA方法^[2,4]程序参照文献[4]设计,并用文献[2]附录中数据集1和数据集9进行了验证.

化学组成数据是典型的组分数据.就黄土剖面而言,以陕西省的洛川剖面研究的比较深入^[5],使我们有可能进行这种对比性分析.洛川剖面中,化学组成包括了 SiO_2 、 Al_2O_3 、 Fe_2O_3 、 FeO 、 CaO 、 MgO 、 Na_2O 、 K_2O 、 P_2O_5 、 TiO_2 和 MnO .本文使用了73个样品,其中包括马兰黄土8个,离石黄土60个,午城黄土5个.按黄土层与古土壤层划分,包括了32个古土壤层,41个黄土层.

在本文中,我们用“原始数据”代表未经处理的数据,用“单形数据”代表经过中心化对数比变换后的数据.原始数据的主成分分析简称为PCA,单形数据的主成分分析简称为LPCA.

3 结果和讨论

在洛川剖面中,虽然各样品的和并不等于100,但是由于数据本身已具有的闭合性质,所以,这并不影响统计的结果.我们的分析也表明,将数据严格闭合到100,与原来数据的结果是一致的.下面,我们主要从成分之间的关系和样品的分类进行讨论.

3.1 成分之间的关系

相关系数和主成分的载荷可以用来分析各个成分之间的关系.相关系数见表1,其中,显著性水平0.001以下的临界相关系数为3.799.

表1中,右上角部分是数据经中心化对数比变换后的相关系数,与左下角原始数据的相比,相关系数和符号变化大的是 Na_2O 、 SiO_2 和 MgO 三个组分.其中, Na_2O 与 Al_2O_3 、 Fe_2O_3 的负相关明显减弱; SiO_2 与 Al_2O_3 、 Fe_2O_3 、 MgO 、 K_2O 、 MnO 的正相关明显增强; MgO 与 SiO_2 、 Al_2O_3 、 Fe_3O_3 、 TiO_2 的正相关明显增强,与 CaO 、 P_2O_5 的负相关明显增强.

我们以方差累计贡献率80%为标准确定主成分,PCA和LPCA均得到了两个主成分,对应的主成分载荷图见图1. LPCA的方差累计贡献率为88.6%,第一主成分方差贡献率是71.9%,根据土壤化学原理分析,该主成分实际上反映了黄土剖面中的两个基本过程:粘粒的积累过程和 CaCO_3 积累过程.第二主成分方差贡献率是16.7%,主要是 P_2O_5 ,它似乎一个独立的组分.在相关系数表中, P_2O_5 除了与 MgO 有较明显的负相关外,与其余组分几乎没有相关关系.

PCA的方差累计贡献率为97.3%,第一主成分方差贡献率为88.8%,主要反映的是 SiO_2

和 CaCO_3 之间的关系.第二主成分方差贡献率为 8.5% ,主要反映的是 Fe_2O_3 、 Al_2O_3 与 SiO_2 之间的关系.

从数值上看 ,PCA 比 LPCA 的方差累计贡献率要高的多 ,似乎能够更全面地反映数据的特征.但是 ,综合各方面的资料 ,PCA 第一主成分所提供的 SiO_2 和 CaCO_3 之间的关系已经证明是虚假的^[5] ,即第一主成分提供的信息是错误的 ,在此基础上进行的数据解释将不可避免的带来错误.反之 ,LPCA 的结果更为合理 ,图 1 中的化学组成关系可以较好的用黄土剖面的主要地球化学过程来解释 ,反映了黄土形成过程中的主要矛盾.

表 1 洛川黄土剖面化学组分相关系数比较

原始数据	单形数据										
	SiO_2	Al_2O_3	Fe_2O_3	FeO	CaO	MgO	Na_2O	K_2O	P_2O_5	TiO_2	MnO
SiO_2	1.000	0.902	0.793	-0.435	-0.783	0.650	0.076	0.489	-0.157	0.900	0.561
Al_2O_3	0.483	1.000	0.938	-0.555	-0.865	0.752	-0.050	0.631	-0.132	0.977	0.717
Fe_2O_3	0.454	0.879	1.000	-0.660	-0.812	0.711	-0.128	0.632	-0.088	0.925	0.718
FeO	-0.487	-0.578	-0.685	1.000	0.518	-0.561	-0.095	-0.675	-0.113	-0.554	-0.629
CaO	-0.860	-0.826	-0.767	0.624	1.000	-0.558	-0.156	-0.703	-0.223	-0.873	-0.727
MgO	-0.431	0.113	0.051	-0.077	0.130	1.000	0.006	0.571	-0.388	0.753	0.664
Na_2O	-0.246	-0.393	-0.444	0.158	0.215	0.191	1.000	0.434	-0.185	-0.022	0.177
K_2O	0.098	0.474	0.407	-0.470	-0.464	0.398	0.443	1.000	-0.146	0.642	0.756
P_2O_5	-0.052	-0.081	-0.127	0.169	0.079	-0.049	0.014	-0.126	1.000	-0.133	-0.183
TiO_2	0.545	0.917	0.836	-0.551	-0.845	0.138	-0.328	0.487	-0.111	1.000	0.730
MnO	0.108	0.535	0.479	-0.416	-0.445	0.408	0.167	0.684	-0.119	0.553	1.000

在 LPCA 图中(图 1)按照由小到大的顺序排列第一主成分的载荷 ,我们得到了如下的序列 CaO , FeO , Na_2O , MgO , P_2O_5 , SiO_2 , Al_2O_3 , TiO_2 , MnO , K_2O , Fe_2O_3 .其中 , $\text{MgO} < \text{P}_2\text{O}_5 < \text{SiO}_2 < \text{Al}_2\text{O}_3 < \text{TiO}_2 < \text{MnO} < \text{K}_2\text{O} < \text{Fe}_2\text{O}_3$ 与黄土风化形成古土壤过程中的积聚系列相同 ,而 $\text{CaO} > \text{FeO} > \text{Na}_2\text{O}$ 则与淋溶系列相同 .LPCA 得到的结果 ,与黄土剖面中粘土矿物分析、微形态分析、粒度分析等获得的结论相符合^[5 6 7] .所以 ,LPCA 能够较好的反映数据的本质特征.

3.2 样品的分类

图 2 是样品的主成分得分图.我们根据文献 5 的分类 ,将样品分为黄土、古土壤和粉砂层三类 ,并用符号表示在图中.

与原始数据相比 ,单形数据具有良好的可分性 ,而且更为合理.在原始数据的得分图中 ,古土壤样品的范围内包括了部分强风化的黄土样品.从粉砂层样品在图中的位置看 ,我们只有对图中的坐标轴进行旋转后才能进行解释.否则 ,如果按第一主成解释 ,粉砂层的环境似乎与强风化黄土层的环境是类似的 ;如果按第二主成解释 ,粉砂层又与部分的古土壤样品相混杂 ,这都与实际情况不符合.在单形数据的得分图中 ,不存在这样的问题.得分图可以按照主成分载荷图产生的模型进行解释 ,黄土样品中以 CaO 的积累为明显的特征 ,而在古土壤样品中 ,以粘土矿物的积累为明显的特征.

同时 ,在得分图中可以看到 ,从古土壤到黄土表现为较明显的过渡特征 ,并存在着混杂.在黄土样品中 ,混杂有部分具有较高的 CaO 含量的古土壤样品.混杂的存在表明 ,就化学组分而言 ,不足以完全区别黄土和古土壤.单一的数据源提供的分类信息毕竟是有限的 ,只有通过多种数据源的相互验证 ,才能得到正确的分类结果.

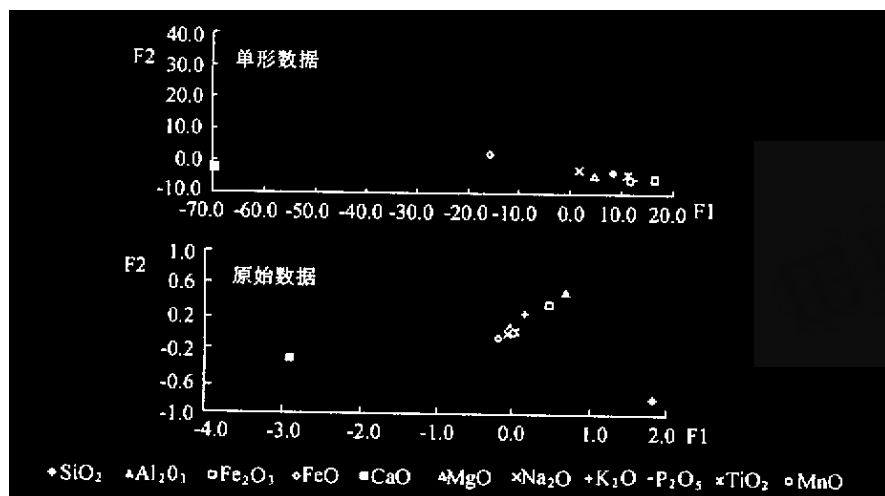


图1 洛川黄土剖面主成分载荷图

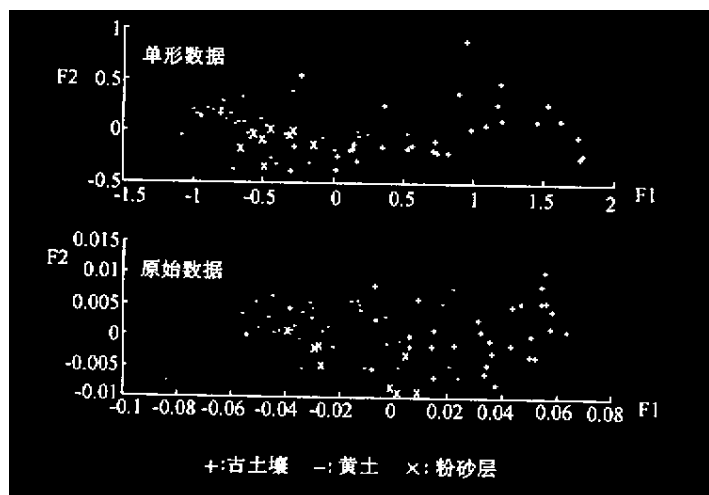


图2 洛川黄土剖面主成分得分图

4 结论

本文通过 LPCA 分析与 PCA 分析的对比,表明在黄土化学组成数据的分析中,传统的 PCA 分析由于数据之间存在的闭合,难以获得数据中的真实关系。洛川黄土剖面化学组成中 SiO_2 和 CaO 之间的负相关关系是虚假的。由 LPCA 获得的化学组成主成分载荷,与黄土剖面的淋溶和积聚序列相同,具有明确的地球化学意义。LPCA 得分图较好地表达了样品之间的过渡关系。

我们认为,对于组分数据使用 LPCA 进行分析是非常必要的。通过分析,一方面我们可以避免成分之间的虚假关系,更好地认识成分之间的关系;另一方面,可以帮助我们建立起合理的认知模型,获得更符合实际的数据分类。现在,随着全球变化研究的深入,获得的黄土剖面化学组成数据越来越多,为了进行更深入的分析,更有必要对这个问题给予充分的注意。

[参考文献]

- [1] Aitchison J. The statistical analysis of compositional data(with discussion) [J]. J Roy Stat Soc Ser ,1982.
- [2] Aitchison J. The statistical analysis of compositional data. Monographs on Statistics and Applied Probability[M]. London :Chapman and Hall ,1986.
- [3] 周蒂. 组分数据的对数比统计方法简介[J]. 地质科技情报 ,1988 ,7(2) :107—114.
- [4] Richard reyment K. Applied factor analysis in the natural sciences[M]. Cambridge university press ,1996. 1221—128.
- [5] 刘东生. 黄土与环境 [M]. 北京 :科学出版社 ,1985. 240—260.
- [6] 文启忠. 中国黄土的地球化学环境 [C]. 地球化学文集(1966—1986) ,1986. 204—206.
- [7] 文启忠. 黄土堆积过程中元素的演化与古气候的关系[J]. 中国第四纪研究 ,1985 ,7(2) :34—40.

Logratio Statistics Analysis and Its Use in the Study of the Chemical Components in Loess Profile

Wei Yuchun¹ ,Zhang Jianming² ,Ni Shaoxiang¹

(1. College of Geographical Science ,Nanjing Normal University ,Nanjing ,210097 ,PRC)

(2. Geography Department in Lanzhou University ,Lanzhou ,730000 ,PRC)

Abstract :In this paper ,an analysis on the chemical component data of the Luochuan loess profile using the logratio statistics (LS) method was conducted. The results show that the classical statistics method result in the false correlation between component and can not to be used as a model to the sample classification. The study also shows that LS ,especially logcontrast principal component analysis ,is a more reliable method for the analysis of compositional data.

Key words :chemical component ;loess profile ;logratio statistics

[责任编辑 陆炳新]

(上接 113 页)

The Analyses and Countermeasures of Problems of Land Use of Urban – Rural Crisscross Area

Zhang Zengfeng ,Huang Kelong

(Geographic College of Nanjing Normal University ,Nanjing ,210097 ,PRC)

Abstract :Crisscross area of urban and rural is a special and microcosmic zone of economic ,and it' s land type is complicated. Along with the rapid development of economic and urbanization ,the economy of the crisscross area of urban and rural develops rapidly ,but there are many problems in the land use of this region ,especially the decrease of the resources of plantation. This essay analyzes the character of land use and the reason that cause above problems in the crisscross area of urban and rural. In order to do some good for the reasonable land use of crisscross area of urban and rural ,the author puts forward many corresponding countermeasures.

Key words :crisscross area of urban and rural ;Land use ;problems ;countermeasures

[责任编辑 陆炳新]