第17卷 第4期 土 壤 学 报 Vol. 17, No. 4
1980 年 11 月 ACTA PEDOLOGICA SINICA Nov., 1980

聚类分析在太湖地区水稻土物质分类上的应用*

刘多森 徐 琪 陆彦椿
（中国科学院南京土壤研究所）

近十余年来，数理统计中的多元分析发展很快，并在科学技术的各个领域广泛应用。土壤数值分类，就是多元分析引进土壤分类学的结果。土壤物质分类，可用数值分类为其表达形式。

聚类分析（Cluster analysis）是研究"物以类聚"的数学方法。它属于多元分析的重要分支。数学研究者对聚类分析作过比较详尽的介绍[1]。七十年代以来，随数学研究者对这一方法的逐步完善，聚类分析得到各国土壤学界的重视。美国 Cipra 等[5]、英国 Cuanalo 等[6]、日本 Kyuma 等[7]、苏联 Рожков 等[8]，都曾用聚类分析解决土壤分类问题。

本文拟就太湖地区的水稻土物质类型进行聚类分析的尝试。取水稻土耕层标本 30 个，分别编号为 1, 2, ……, 30。研究的土壤指标有 7 项：有机质含量，全氮含量，全磷含量，全钾含量，粗粉粒(0.01—0.05 毫米)含量，粘粒（<0.001 毫米)含量，阳离子交换量。30 个水稻土耕层标本各指标的分析数据见表 1。

各指标的观测值因量纲不同，或量纲虽然相同而数量水平(量级)不同，均可能对数值分类带来不合理的影响。为了消除这些影响，应对各个指标值（原始数据）进行标准化。本文采用标准差标准化方法，得到相应的 210 个标准化值。这一标准化方法，我们过去已有说明[2]。

所选 30 个水稻土，均可用上述 7 项指标的标准化值构成的 7 维向量表示。为了系统地比较这 30 个 7 维向量的亲疏程度，首先要规定一个表示水稻土两两之间亲疏程度的统计量。按任何一个规定的统计量，都将计算出 30 个水稻土两两之间该统计量的 435 个数值。广义地说，该统计量的个数是 $n(n-1)/2$，其中 n 是样品总个数。

在聚类分析中，表示二样品之间亲疏程度的统计量，常用的有以下两种：

1. 第 i, l 样品之间的绝对值距离 $d_{il}(1)$

$$d_{il}(1)=\sum_{j=1}^{m}|x_{ij}-x_{lj}|$$

$$(i, l=1, 2, \cdots, n)$$

式中 i, l 是样品的编号，j 是指标的编号，n 是样品总个数，m 是指标总个数，x_{ij} 和 x_{lj} 分别是第 i, l 样品的第 j 指标的标准化值。绝对值距离愈小，表示二样品的关系愈亲近；愈大，则愈疏远。

* 本文承熊毅教授斧正，谨致谢意。

表 1　太湖地区 30 个水稻土的理化性质

Table 1 The chemical and physical properties of 30 paddy soils in Tai Hu area

发生类型 Genetic type	土壤编号 Soil No.	有机质 O. M. (%)	全氮 Total N (%)	全磷 Total P_2O_5 (%)	全钾 Total K_2O (%)	粗粉粒 (0.01—0.05mm) Coarse silt (%)	粘粒 (<0.001mm) Clay (%)	阳离子交换量 CEC (m. e. /100g)
侧渗水稻土 Side bleaching paddy soil	1	1.33	0.073	0.061	1.73	56.9	15.1	10.97
	2	1.68	0.070	0.050	1.10	49.5	13.4	9.58
	3	1.09	0.010	0.032	1.58	57.4	10.7	5.91
	4	1.30	0.072	0.044	1.13	56.5	11.7	8.40
	5	1.53	0.080	0.045	1.31	50.7	17.0	8.55
	6	1.36	0.149	0.022	1.48	45.0	9.0	10.18
滞水水稻土 Stagnating paddy soil	7	2.09	0.142	0.054	1.27	52.0	18.0	11.51
	8	2.39	0.143	0.080	1.51	61.4	11.2	12.86
	9	1.81	0.110	0.200	1.41	50.6	19.2	14.90
	10	1.21	0.079	0.190	1.37	46.0	24.0	13.73
	11	1.69	0.101	0.118	1.50	44.5	21.5	17.25
	12	1.82	0.113	0.082	1.75	48.6	12.4	16.50
爽水水稻土 Permeable paddy soil	13	2.84	0.141	0.118	1.58	45.0	19.5	11.43
	14	2.81	0.168	0.136	1.70	41.4	18.4	20.47
	15	2.43	0.167	0.154	1.15	41.6	25.0	22.18
	16	2.68	0.119	0.083	1.32	49.7	22.8	17.36
	17	3.35	0.180	0.214	1.54	32.7	29.7	19.37
	18	2.85	0.151	0.190	1.89	46.5	17.1	19.02
漏水水稻土 Percolating paddy soil	19	1.47	0.093	0.161	2.13	52.0	12.0	10.84
	20	1.85	0.105	0.192	2.04	58.2	15.5	10.90
	21	2.35	0.135	0.164	2.62	36.5	27.0	19.56
	22	3.05	0.146	0.062	1.55	33.0	18.5	12.23
	23	2.14	0.129	0.172	2.13	47.0	23.0	16.66
	24	2.86	0.159	0.150	2.33	38.0	25.0	17.37
囊水水稻土 Waterlogged paddy soil	25	1.46	0.095	0.086	1.49	51.5	17.7	10.72
	26	3.92	0.220	0.276	1.78	45.0	17.5	18.17
	27	2.27	0.124	0.098	1.81	54.5	9.4	14.76
	28	4.30	0.233	0.172	2.12	54.4	7.5	17.10
	29	2.54	0.135	0.136	1.85	55.5	6.2	13.39
	30	4.18	0.237	0.120	2.23	44.1	12.7	20.57

2. 第 i, l 样品之间的欧氏距离 d_{il} (2)

$$d_{il}(2) = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{lj})^2}$$

$$(i, l = 1, 2, \cdots, n)$$

式右各符号意义同前。欧氏距离愈小，表示二样品的关系愈亲近，反之亦然。

按环境科学的一般理解，环境域某点位标准化值较高的有害物，对该点位逆质量的负荷，要比对该点位各有害物含量标准化值的平均值的负荷更大[3]。在这一意义上说，欧氏距离可能比绝对值距离更适用于环境质量评价。因为欧氏距离强调了二点位标准化值之

差较大的有害物对环境质量的影响。但在土壤数值分类中，土壤个体之间标准化值之差较大的指标，对土壤特征的影响不一定更为突出，即不一定需要用 2 次方加以强调。亦即对土壤数值分类而言，欧氏距离的计算步骤虽然比绝对值距离繁琐，但可能并不会因此而带来更好的分类效果。根据我们对太湖地区水稻土进行聚类分析的实践，说明就此研究对象而言，欧氏距离的分类效果略逊于绝对值距离。

在规定了样品之间的距离之后，尚需规定类与类之间的距离，并用规定的类与类的距离对 n 个样品逐步聚类，以达到对样品分类的目的。类与类之间的距离，目前定义方法很多。不同的定义，有不同的聚类方法，并产生不尽相同的分类效果。由于具体的研究对象不同，类与类的距离究竟用什么定义为宜，很难一概而论。有的数学研究者认为，类平均法是比较好的定义方法之一。

我们分别以最短距离、最长距离、平均平方距离（类平均法）定义类与类的距离，对太湖地区 30 个水稻土进行了聚类分析。

假设，第 i，l 样品之间的距离为 d_{il}，距离阵中的最小元素为 D_{pq}，则将类 G_p 和类 G_q 合并为一新类 G_r，类 G_p，G_q，G_r 同其他类 G_k 的距离分别为 D_{pk}，D_{qk}，D_{rk}。

定义两类之间的距离为最短距离，则

$$D_{rk} = \min_{\substack{i \in G_r \\ l \in G_k}} d_{il} = \min \{D_{pk}, D_{qk}\}$$

定义为最长距离，则

$$D_{rk} = \max_{\substack{i \in G_r \\ l \in G_k}} d_{il} = \max \{D_{pk}, D_{qk}\}$$

定义为平均平方距离，则

$$D^2_{rk} = \frac{1}{n_r n_k} \sum_{\substack{i \in G_r \\ l \in G_k}} d^2_{il}$$

$$= \frac{n_p}{n_r} D^2_{pk} + \frac{n_q}{n_r} D^2_{qk}$$

式中 n_r，n_p，n_q 分别为类 G_r，G_p，G_q 包含的样品数，且有 $n_r = n_p + n_q$

用上述三种定义方法，可在任何一种样品距离 d_{il} 的基础上，获得相应的三幅太湖地区水稻土聚类图。比较结果认为，对于太湖地区水稻土而言，由于地域不够广阔，土壤个体之间的变差不够突出，所以使空间浓缩而分类灵敏度低的最短距离法的分类效果不够理想，而类平均法及使空间扩张而分类灵敏度高的最长距离法的分类效果较好，且以类平均法更优，因为分类灵敏度高的方法有时会出现不合理的结果。

以绝对值距离作为样品距离，以平均平方距离作为类与类之间的距离，则得图 1 表示的聚类图。如从该图的平均平方距离 44 处划一垂直于横轴的直线，则可将 30 个水稻土区分为 5 个物质类型。表 2 列出了这 5 个物质类型的水稻土各指标观测值的平均值和标准差。为了使聚类分析得出的物质分类与发生分类相比较，表 3 列出了这些水稻土的地理特征，并在表 4 说明了 30 个水稻土在物质类型—发生类型矩阵各元素上占据的土壤个体数。关于太湖地区水稻土的发生分类[4]，已有阐述。

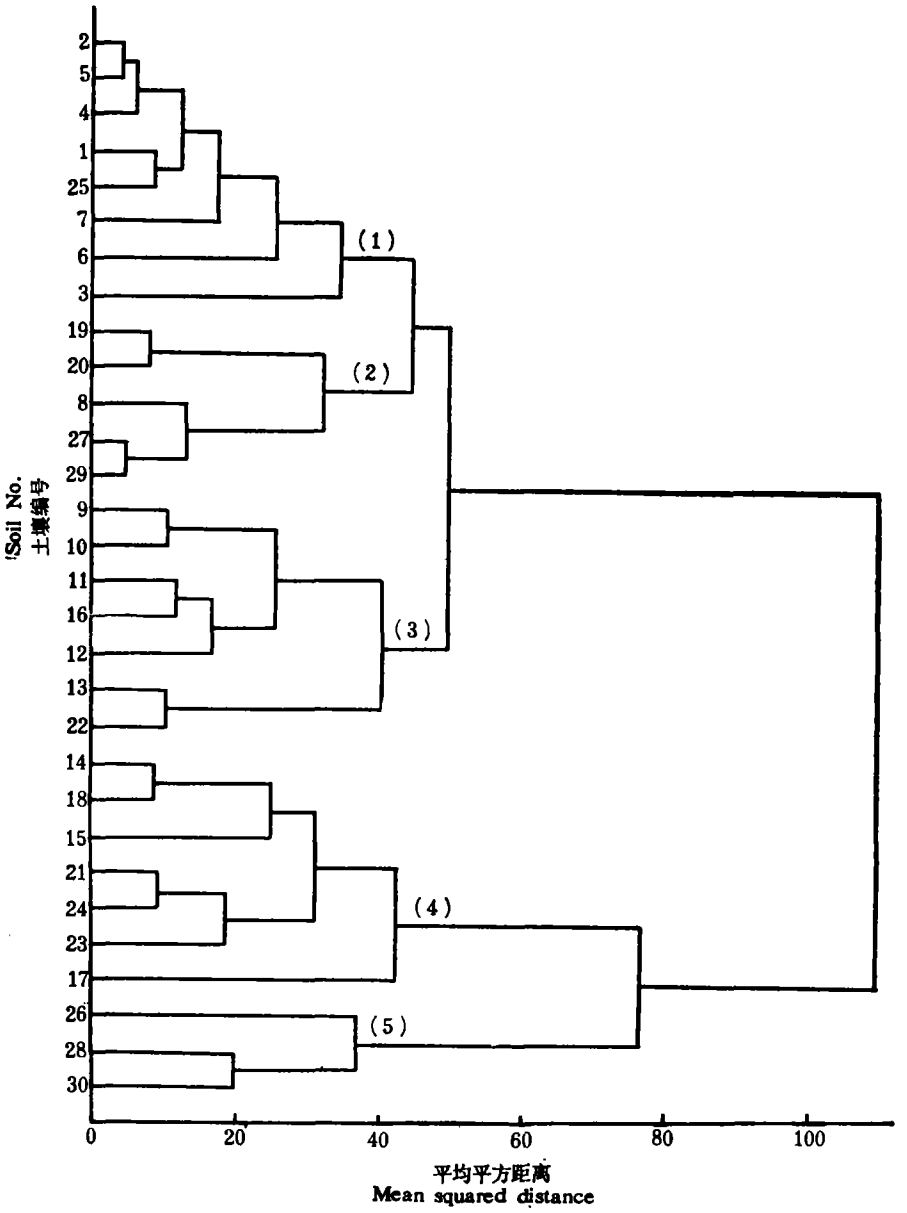

图 1　太湖地区 30 个水稻土聚类图

Fig. 1 The cluster distribution of 30 paddy soils in Tai Hu area

(1) 低肥力型 Low fertility (2) 中肥力粗粉质型 Medium fertility, coarse silty type (3) 中肥力细粉质型 Medium fertility, fine silty type (4) 高肥力型 High fertility (5)潜在肥力型 Medium fertile soils with high potential fertility

如果所取指标不能完全反映相应土壤的发生学特征，则数值分类与发生分类不尽相同。随着仪器分析的进展，我们将可能选择适当的指标，使数值分类充分反映土壤个体之间在发生学上的亲疏关系。另一方面，只要我们所取的物质指标在土壤个体之间存在差异，则由数值分类所表达的土壤物质分类，将必然反映土壤个体之间在所取指标提供的信

表2 太湖地区水稻土5个物质类型土壤特性的平均值和标准差

Table 2 Means and standard deviations of the soil characteristics for the 5 material categories of paddy soil in Tai Hu area

物质类型 Material category	地形 Topography	土壤个数 Soil number	有机质 O. M. (%)	全氮 Total N (%)	全磷 Total P_2O_5 (%)	全钾 Total K_2O (%)	粗粉粒 (0.01—0.05 mm) Coarse silt (%)	粘粒 (<0.001mm) Clay (%)	阳离子交换量 CEC (m.e./100g)
1.低肥力型 Low fertility	丘陵 Hill	8	1.48±0.302	0.086±0.0441	0.049±0.0193	1.39±0.221	52.4±4.29	14.1±3.41	9.48±1.819
2.中肥力粗粉质型 Medium fertility, coarse silty type	平原 Plain	5	2.10±0.438	0.120±0.0208	0.133±0.0456	1.87±0.240	56.3±3.61	10.9±3.42	12.56±1.671
3.中肥力细粉质型 Medium fertility, fine silty type	平原 Plain	7	2.16±0.694	0.116±0.0230	0.122±0.0539	1.50±0.147	45.3±5.93	19.7±3.79	14.77±2.400
4.高肥力型 High fertility	平原 Plain	7	2.68±0.406	0.156±0.0185	0.169±0.0264	1.91±0.498	40.5±5.22	23.6±4.52	19.23±1.847
5.潜在肥力型 Medium fertile soils with high potential fertility	滨湖洼地 Depression nearly the lake	3	4.13±0.194	0.230±0.0089	0.189±0.0794	2.04±0.235	47.8±5.70	12.6±5.00	18.61±1.777

表3 太湖地区水稻土不同发生类型的地理特征

Table 3 Geographic characteristics of different genetic types of paddy soils in Tai Hu area

发生类型 Genetic type	母质 Parent material	地形 Topography	海拔*(米) Altitude (m)	起源土壤 Original soil	土壤发生学特征 Soil genetic characteristics
侧渗水稻土 Side bleaching paddy soil	下蜀黄土 Xia Shu loess	丘陵 Hill	>10	黄棕壤 Yellow-brown earth	水分侧渗，淋溶显著
滞水水稻土 Stagnating paddy soil	黄土状沉积物 Loessial deposit	太湖平原 Tai Hu plain	4.9—10	草甸黄棕壤 Meadow yellow-brown earth	剖面分化明显，滞水
爽水水稻土 Permeable paddy soil	黄土状沉积物 Loessial deposit	太湖平原 Tai Hu plain	4.4—6.9	草甸土 Meadow soil	剖面分化不明显，爽水
漏水水稻土 Percolating paddy soil	沿江冲积物 Alluvial deposit	沿江平原 Riverside plain	4.5—10	石灰性草甸土 Calcareous meadow soil	有不同程度石灰反应，漏水
囊水水稻土 Waterlogged paddy soil	冲积—湖积物 Alluvial-lacustrine deposit	交接洼地：滨湖洼地 Depression between hill land and plain; depression, nearly the lake	3.7—6.9	沼泽土 Marsh soil	潜育作用明显，囊水

* 取吴淞口海面为基准。 Above the average sea level outside the Wusong Kou.

表 4　30 个水稻土在物质类型—发生类型矩阵上的分布

Table 4 Distribution of 30 paddy soils on the matrix of the material and genetic types

物质类型 Material category	发生类型 Genetic type					合计 Total
	侧渗水稻土 Side bleaching paddy soil	滞水水稻土 Stagnating paddy soil	爽水水稻土 Permeable paddy soil	漏水水稻土 Percolating paddy soil	囊水水稻土 Waterlogged paddy soil	
低肥力型 Low fertility	6	1			1	8
中肥力粗粉质型 Medium fertility, coarse silty type		1		2	2	5
中肥力细粉质型 Medium fertility, fine silty type		4	2	1		7
高肥力型 High fertility			4	3		7
潜在肥力型 Medium fertile soils with high potential fertility					3	3
合计 Total	6	6	6	6	6	30

息范围内的发生学特征或肥力特征上的某些差异。本文所取指标，仅是不同发生类型代表土壤的耕层某些理化性质，不能充分反映水稻土的发生学特征。因此，在这一基础上得出的水稻土物质分类，只是反映耕层理化性质上的亲疏关系，而与发生分类必然有一定出入，同时又部分地反映了在发生学上的地位。从表 4 看出，侧渗水稻土及滨湖洼地的囊水水稻土的物质类型，同其在发生分类上的地位相一致；滞水水稻土、爽水水稻土的物质类型，同发生分类大体一致，分别主要出现于中肥力细粉质型、高肥力型；漏水水稻土因系冲积物母质发育而成，其物质类型在中肥力型、高肥力型内不规则变动，同发生分类缺乏趋势性联系。

现在，我们将讨论按图 1 得出的太湖地区水稻土的 5 个物质类型。

1. 低肥力型　主要由侧渗水稻土构成，并包含全磷、全钾、阳离子交换量较低而粗粉粒含量较高的个别滞水水稻土（7 号土），有机质、全氮、全磷、全钾、阳离子交换量较低而粘粒含量较高的位于交接洼地的个别囊水水稻土(25 号土)。本型主要处于丘陵，并包含近丘陵的交接洼地及平原局部较高处。本型质地较粗，养分贮量及阳离子交换量均显著低于其他各型水稻土。

2. 中肥力粗粉质型　此型包含有机质、全氮、粘粒、阳离子交换量较低而粗粉粒含量较高的一部分漏水水稻土（19，20 号土），有机质、粘粒、阳离子交换量较低而粗粉粒含量较高的一部分交接洼地的囊水水稻土（27，29 号土），粗粉粒含量较高而粘粒、阳离子交换量较低的个别滞水水稻土（8 号土）。本型处于平原较高处及交接洼地，养分贮量及阳离子交换量中等，粘粒含量很低，粗粉粒含量则居于各型水稻土之冠。

3. 中肥力细粉质型　主要由滞水水稻土构成，并包含全氮、全磷、阳离子交换量较

低而粗粉粒含量较高的少数爽水水稻土(13, 16 号土),全磷、全钾、粗粉粒含量较低的个别漏水水稻土(22 号土)。本型主要处于平原较高处,养分贮量及阳离子交换量中等,质地介于低肥力型、中肥力粗粉质型与高肥力型之间。

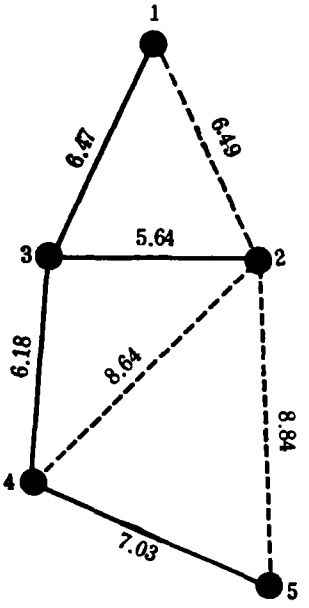

图 2 太湖地区水稻土 5 个物质类型的联接图

Fig. 2 The linkage graph of the 5 material categories of paddy soils in Tai Hu area (1, 2, 3, 4, 5 indicate 5 material categories of paddy soils as same as Fig. 1)

4. 高肥力型 主要由爽水水稻土及一部分全钾、粘粒、阳离子交换量较高的漏水水稻土构成。本型主要处于平原较低处,是各型水稻土中粗粉粒含量最低、而粘粒含量和阳离子交换量最高的土壤,也是除潜在肥力型外养分贮量最高的土壤。高肥力型是太湖地区最肥沃的水稻土物质类型。

5. 潜在肥力型 由滨湖洼地的囊水水稻土构成。本型的质地与中肥力型相近,阳离子交换量与高肥力型相近,全磷、全钾含量较高,有机质和全氮含量显著高于其他各型。由于本型处于滨湖洼地,排水不畅,故养分贮量虽高而其有效性甚差。

运用图论的方法处理上述 5 个物质类型,可得出各物质类型之间的联接图(图 2)。该图直观地表明了各物质类型之间的演化关系。图中的实线,表示联接图的最小支撑树。该最小支撑树表明了 5 个物质类型的总体亲缘关系最近的演化图式。图 2 启示,肥力较低的各物质类型,均可按一定途径人为定向培育成高肥力的物质类型——高肥力型。

按聚类分析得出的 5 个物质类型,反映了水稻土某些肥力特征上的差异。耕层某些性质的差异,既与水稻土发生特征有关,而更重要的是决定于耕作施肥及客土改良的影响。不同物质类型的水稻土,肥力特征不同,在农业生产上的改良利用途径也不同;而同一物质类型的水稻土,则具有大体一致的肥力特征和农业利用特点。

在水稻土发生分类的基础上所进行的物质分类,可以反映耕层肥力状况与农业生产的相关性。质言之,不同发生类型的水稻土,通过人为耕作熟化,可以出现某些类同的性状,甚至成为同一物质类型,从而体现了人为定向培育对土壤的影响。因此,上述物质分类可作为划分分类单元、特别是划分基层分类单元的主要依据。

本文对太湖地区水稻土进行的物质分类,只是初步尝试。随着选用指标的不断完善,揭示其发生学内容的可能性也必将增加,进而为研究水稻土发生分类提供更为科学的基础。

参 考 文 献

[1] 方开泰, 1978: 聚类分析 (I), (II)。数学的实践与认识, 1 期, 66—80 页; 2 期, 54—62 页。

[2] 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（以水稻土为例）。土壤学报，16卷2期，173页。
[3] 刘多森，1980：环境质量的数值计算。环境科学研究与进展，科学出版社。
[4] 徐琪、陆彦椿、朱洪官，1980：太湖地区水稻土的发生分类。土壤学报，17卷2期，120—132页。
[5] Cipra, J. E., Bidwell, O. W., and Rohlf, F. J., 1970: Numerical taxonomy of soils from nine orders by cluster and centroid-component analyses. Soil Sci. Soc. Amer. Proc., 34(2): 281—287.
[6] Cuanalo, de la C. H. E., Webster, R., 1970: A comparative study of numerical classification and ordination of soil profiles in a locality near Oxford. J. Soil Sci., 21(2): 340—352.
[7] Kyuma, K., and Kawaguchi, K., 1976: Soil material classification for paddy soils in Japan. Soil Sci. Plant Nutr., 22(2), 111—124.
[8] Рожков, В. А., Прошина, Н. В., 1977: Опыт численной таксономии почв. Почвоведение, №8, 106—116.

THE APPLICATION OF CLUSTER ANALYSIS ON THE MATERIAL CLASSIFICATION FOR PADDY SOILS IN TAI HU AREA

Liu Duo-shen, Xu Qi and Lu Yan-chun
(Institute of Soil Science, Academia Sinica, Nanjing)

Summary

As far as classification of paddy soils in Tai Hu area is concerned, the character difference between soil individuals i and l

$$d_{il}(1) = \sum_{j=1}^{m} |x_{ij} - x_{lj}|$$

is somewhat better than their Euclidean distance. The mean squared distance between soil classes G_r and G_k

$$D^2_{rk} = \frac{1}{n_r n_k} \sum_{\substack{i \in G_r \\ l \in G_k}} d^2_{il}$$

is better than their minimum and maximum distance. In these equations x_{ij} and x_{lj} are the standardized values of character j for soils i and l, m is the number of character, n_r and n_k are the numbers of soils in classes G_r and G_k.

The distance between soil individuals was defined as the character difference and the distance between soil material classes was defined as the mean squared distance. Base on this definition, 30 paddy soils in the area of Tai Hu were examined and divided into 5 material categories, i.e. low fertility; medium fertility, coarse silty type; medium fertility, fine silty type; high fertility and medium fertile soils with high potential fertility.