土 壤 (Soils), 2009, 41 (6): 998~1003

不同插值方法对成分数据空间预测结果的影响①
——以土壤连续分类模糊隶属度值为例

檀满枝[1]， 密术晓[1,2]， 李开丽[1,2]， 陈 杰[1,3]

（1 土壤与农业可持续发展国家重点实验室（中国科学院南京土壤研究所），南京 210008；
2 中国科学院研究生院，北京 100049；3 郑州大学水利与环境学院，郑州 450001）

摘 要： 地球科学中成分数据（compositional data）非常普通，其在进行空间插值时必须满足 4 个条件：每一位置各组分之和为常数，每一组分为非负，插值结果无偏最优。本文以土壤连续分类模糊隶属度值为例，数据经对数正态变换、非对称对数比转换、对称对数比转换后进行普通克里格插值结果和成分克里格插值（compositional kriging）结果进行比较。结果表明，对原始数据和经对数正态变换后数据进行插值，每一位置预测结果隶属度之和不能满足常数 1。经非对称对数比转换后，插值结果虽然满足各个位置组分之和为 1，但是预测结果精度较低，且预测结果空间分布连续性不明显。数据经对称对数比转换后插值结果和成分克里格插值结果，都能满足成分数据空间插值的 4 个条件，但二者各有优势。相比较而言，对称对数比转换方法得到的预测结果更能体现土壤空间连续渐变特征，而成分克里格插值结果能保证隶属度本身是最优无偏估计。

关键词： 成分数据；对称对数比转换；非对称对数比转换；成分克里格

中图分类号： S159.9；O159

地球科学中成分数据（compositional data）非常普通，数据常常表达为分数或百分比。例如土壤颗粒组成、岩石的化学组成，沉积物中花粉和有孔虫目的组成等[1]。另外，不明显的成分数据有指示数据和经模糊 *c*-均值分类得到的土壤连续分类模糊隶属度值。各组分之和为常数，且每一组分为非负。因此组分相关性中含由闭合效应引起的伪相关，并且不服从正态分布，使得统计分析存在困难[2]。由于成分数据结构的特殊性，在进行地统计学空间预测之前，必须进行特殊的处理，使得插值结果同时满足以下 4 个条件：每一位置各个组分插值结果均为非负，且之和为常数，估计误差最小化和无偏估计[3-4]。国外进行过土壤颗粒组成空间预测方法的对比研究，结果表明原始数据经对称对数比转换后进行克里格插值取得的结果最理想，而直接对原始数据进行克里格插值的方法不可取[5]。在国内有关成分数据空间预测的实际研究中，定和和非负限制常常被有意或无意地忽略，从而造成插值结果不符合实际情况，例如土壤颗粒组成空间预测常常是这种情况[6-7]。基于离散样点土壤属于每一类别的模糊隶属度值，进行土壤模糊连续制图，国外在这方面的研究较多[4,8-12]，但没有研究对模糊隶属度值空间预测的各种方法进行过系统的比较和分析。因此本文选取区域土壤样点属于不同类别的模糊隶属度值，数据经各种方法转换后插值和成分克里格插值得到的结果进行系统的分析和比较，总结出各种方法的优势和不足，为今后自然界中诸如此类成分数据的空间预测提供方法借鉴。

1 土壤连续分类模糊隶属度值

作为地球表面的自然连续体，土壤的连续性特征不仅表现于地理空间上的分布，同时也表现于属性空间上的变异。土壤样本往往表现出对于不同土壤类型的多重相似性。换言之，土壤样本对于一个特定土壤类型的隶属关系不是非 0 则 1，而是对于一个以上的土壤类型均表现出部分隶属关系，即模糊隶属关系。模糊隶属关系的理论基础是模糊逻辑或模糊集理论[14-15]。 根据部分隶属关系把土壤样本划分入不同的类

①基金项目：中国科学院南京土壤研究所创新前沿项目（ISSASIP0716）和国家自然科学基金项目（40701070 和 40571065）资助。
作者简介：檀满枝（1978—），女，安徽望江人，博士，助理研究员，主要从事土壤资源演变、土壤空间预测及土壤调查制图研究。E-mail: mzhtan@issas.ac.cn

值，即模糊子集。20 世纪 80 年代末期，模糊逻辑结合地统计学方法开始广泛应用在土壤分类连续制图领域，国内是最近两年才刚刚发展起来的。模糊 *c*-均值算法是土壤科学应用最广泛的土壤连续分类方法，分类结果为每一样点属于不同类别的隶属度之和为 1，且隶属度值均为非负。从模糊隶属度数据结构来看，它是一类典型的成分数据。

2　研究方法

2.1　非对称对数比转换

Aitchison 于 1982 和 1986 年提出成分数据的对数比转换方法，将成分数据变换成其组分的比值对数（称“对数比”），其对数比将近似地服从正态分布，就这样，对数比转换同时解决了成分数据统计分析中的闭合效应和统计分析这两个问题[16-17]。Pawlowsky 等[3]将对数比方法与地质统计学方法相结合，提出了成分数据的区域化统计方法。常用的对数比转换又称为非对称对数比转换[8-9]（asymmetry Logratio transform），某些文献中又称为加和对数比转换[5]（additive logratio transform），具体计算公式如下：

$$\mu'_{ij}(x)=\ln\frac{\mu_{ij}(x)}{\left(\coprod_{j=1}^{c}\mu_{ij}(x)\right)^{1/c}} \tag{1}$$

转回公式为：

$$\mu_{ij}(x)=\frac{\exp\mu'_{ij}(x)}{\sum_{j=1}^{c}\exp\mu'_{ij}(x)} \tag{2}$$

式中，$\mu_{ij}(x)$ 为第 i 个样点的土壤对于第 j 个聚类类别的隶属度，$\mu'_{ij}(x)$ 为第 i 个样点的土壤对于第 j 个聚类类别的隶属度的对称对数比转换值。

2.2　对称对数比转换

对称对数比转换[8-9]（symmetry Logratio transform），某些文献中称改进的加和对数比转换[5]（modified additive logratio transform），公式为：

$$\mu'_{ij}(x)=\ln\frac{\mu_{ij}(x)+\eta_j}{\left(\coprod_{j=1}^{c}(\mu_{ij}(x)+\eta_j)\right)^{1/c}} \tag{3}$$

转回公式为：

$$\mu_{ij}(x)=\left(\frac{\exp\mu'_{ij}(x)}{\sum_{j=1}^{c}\exp\mu'_{ij}(x)}-\frac{\eta_j}{1+\sum_{j=1}^{c}\eta_j}\right)\left(1+\sum_{j=1}^{c}\eta_j\right) \tag{4}$$

式中，$\mu_{ij}(x)$ 为第 i 个样点的土壤对于第 j 个聚类类别的隶属度，$\mu'_{ij}(x)$ 为第 i 个样点的土壤对于第 j 个聚类类别的隶属度的对称对数比转换值，η 为常数，取研究区除 0 外最小隶属度值的一半。

2.3　成分克里格

由于没有一种现成的克里格插值方法考虑过成分数据的特殊性，因此 De Gruijter 等[4]于 1997 年提出的成分克里格是一种专门针对这类数据的插值方法，该方法是在普通克里格插值的基础上发展起来的。然而它与协同克里格插值不同，它不能保证成分数据之间的线性相关性，此外，不能获取交叉-变量图模型。考虑到无偏限制，和普通克里格一样，成分克里格也是最小化估计方差。最小化估计方差通过设置联合拉格朗日乘数一阶偏导数，也就是对于一阶线性方程 a_c，μ_c 或者 β 为 0。

$$\sum_{j=1}^{n_c}\lambda_{jc}C_{ijc}+\mu_c+\alpha_c z_{ic}+\beta z_{ic}=C_{i0c}\forall_i,c \tag{5}$$

$$\sum_{i=1}^{n_c}\lambda_{ic}=1\ \ \forall c \tag{6}$$

$$\sum_{i=1}^{n_c}\lambda_{ic}z_{ic}=0，\text{并且}\ a_c\geq 0\qquad \forall c \tag{7}$$

$$\sum_{c=1}^{k}\sum_{i=1}^{n_c}\lambda_{ic}z_{ic}=1 \tag{8}$$

式中，λ_{jc} 对于隶属类别 c（成分数据）赋值于观察点 j 点的权重值。C_{ijc} 为观察点 i 和 j 隶属类别 c 隶属度的协方差。C_{i0c} 为观察点 i 和预测点隶属类别 c 隶属度的协方差。n_c 为用于预测类别 c 的观察点的数量。因此，特定预测点的隶属度值可以通过公式计算：

$$\hat{z}_c=\sum_{i=1}^{n_c}\lambda_{ic}z_{ic}\quad\forall c \tag{9}$$

估计方差可以通过代数处理经替代的权重更有效地表示为：

$$\sigma^2_{RC}=\sigma^2_c-\sum_{i=1}^{n_c}\lambda_{ic}C_{i0c}-\mu_c-(\alpha_c+\beta)\hat{z}_c\qquad\forall c \tag{10}$$

式中，σ^2_{RC} 为隶属于类别 c 预测误差的方差，σ^2_c 为隶属于类别 c 的方差。

3　案例研究

3.1　研究区土壤模糊连续分类

研究区位于江苏省南京市东郊麒麟镇东流村附近，面积约为 1 km^2。在对研究区进行野外调查的基础上，在不同母质来源、地形部位和土地利用方式下开挖土壤剖面 31 个[18]，钻取土壤样点 85 个，深度为 120 cm（或至基岩）。对土壤剖面形态特征进行观察、描述与记录，并分层采集土壤样品，同时记载样点的地理位置及其周围的景观信息。结合土壤样品实验室分析数据，在对研究区主要发生层进行细分、归整的基础上，划分出 9 个具有重要土壤发生学意义与分类典型性的特征土层 [19]。

基于研究区 116 个样点发育的 9 种特征土层的厚度数据[20]，建立样点对应特征土层类型厚度数据矩阵，样点对应缺失的特征土层类型厚度值用“0”表示。应用模糊 *c*-均值算法（FCM），定量化确定最佳分类参数，把研究区土壤自动分为 4 类，用 A、B、C、D 表示。FCM 输出结果包括类别质心值和样点属于每一类别的模糊隶属度值。

3.2 各种数据转换方法空间预测结果比较

直接用模糊隶属度值进行普通克里格插值，4 种类别隶属度栅格图加和平均值虽然为 1，但是有 40.7% 的栅格之和＞1，有 59.3% 的栅格隶属度之和＜1，没有一个栅格单元之和＝1 的情况出现（图 1I）。单一类别隶属度图最小值都为负值，最大值为 1.53。采用对数正态变换，4 种类别隶属度插值栅格图加和，结果有 13% 的栅格隶属度之和＞1，有 87% 的栅格隶属度之和＜1，没有一个栅格隶属度之和＝1 的情况出现（图 1II）。这显然与隶属度实际情况不符。因此结果说明，不考虑成分数据的特殊性，对原始数据直接进行插值，或经正态变换后进行插值，都会存在必然的不确定性或不可靠性。直接对原始数据进行克里格插值的均方根误差虽然较小（表 1），按理来说这应该是理想的预测结果，但实际证明对原始数据进行插值不可靠，这点同时说明依靠一种方法进行预测精度验证也是不可靠的问题。

经非对称对数比转换后进行普通克里格插值生成的单一类别隶属度图，每一栅格单元的隶属度之和虽然都均为 1，但是预测结果精度很低（表 1）。单一类别隶属度预测结果图渐变过渡特征不明显，空间分布格局不合理（图 2）。

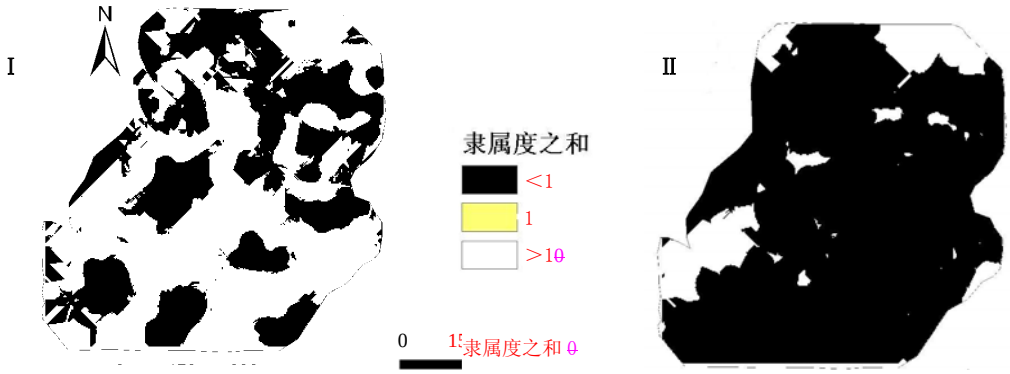

图 1　4 个类别隶属度栅格加和图（I 基于原始数据，II 基于自然对数变换数据）

Fig.1　Sum of the membership values produced by ordinary kriging of the untransformed soil memberships and transformed

——经过对称对数比转换后的研究区 4 种类别土壤隶属度空间预测图进行栅格加和，结果每一栅格隶属度之和均为 1，且每一类别隶属度值都在 0～1 之间。同时预测精度也较合理（表 1）。因此，采用对称对数比对土壤隶属度值进行转换，预测结果较理想。

表 1　各种形式数据空间插值精度比较

Table 1　Assessment of spatial interpolation precision of different kinds of data

类别	均方根误差				平均预测误差			
	原始	自然对数	对称对数比	非对称对数比	原始	自然对数	对称对数比	非对称对数比
A	0.168	1.952	1.017	42.77	-0.001	0.024	0.001	-0.105
B	0.261	2.292	1.484	27.31	0.001	0.026	0.002	0.041
C	0.254	2.335	1.863	291.40	-0.010	-0.053	-0.068	-3.656
D	0.373	2.196	2.268	359.40	0.010	0.080	0.050	2.615
平均	0.264	2.194	1.658	180.22	0	0.019	-0.004	-0.276

3.3　~~3.3~~ 对称对数比转换插值结果和成分克里格插值结果比较

数据经对称对数比转换和成分克里格插值结果均满足成分数据插值的 4 个条件。对称对数比转换比成分克里格插值结果理想，空间上连续渐变特征明显（图 3）。但对称对数比转换由于不是直接对原始数据进行预测，因此不能进行估计误差方差评价，不能保证转回的隶属度值一定满足最优无偏估计，成分克里格隶属度插值结果不太理想可能是成分克里格程序由代数控制所引起的，并且成分克里格是基于原始隶属度数据进行插值的，数据可能不像希望中的那样总是有序的（即进行插值的前 3 个类别与最后插值的那个类别顺序的不同会影响到最后进行插值的那个类别），而对称对数比转换后的数据总是有序的。但是成分克里格的优势在于直接对原始数据进行插值，因此是对隶属度值无偏最优估计，估计误差的方差图显示（图 4），插值结果比较理想。

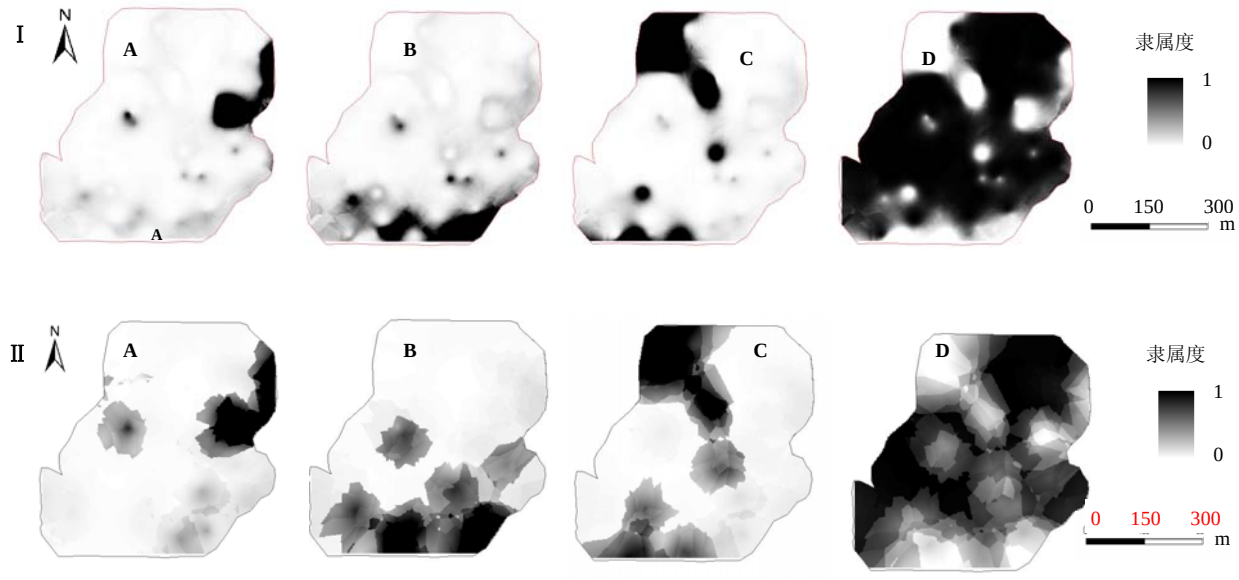

图 3 单一类别隶属度图（Ⅰ数据经对称对数比转换后进行普通克里格插值，Ⅱ成分克里格插值）

Fig 3 Interpolation results of membership values transformed by symmetry log-ratio (Ⅰ) and compositional kriging result (Ⅱ)

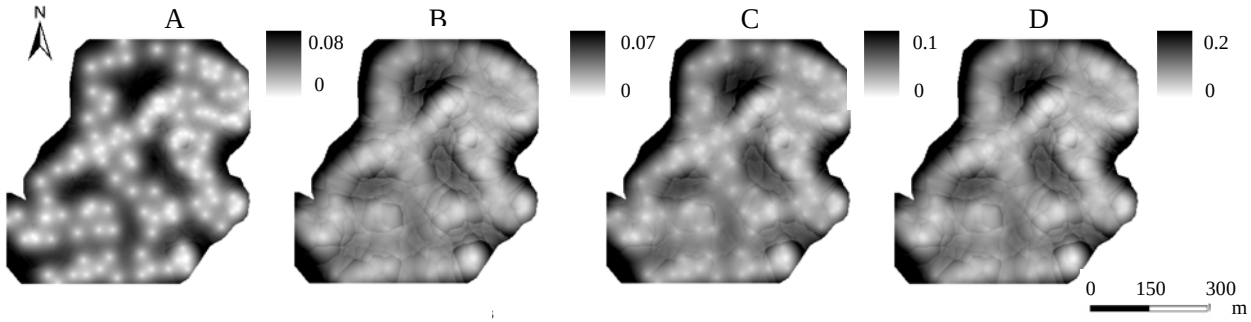

图 4 成分克里格插值估计误差方差图

Fig.4 Maps of variance of prediction error in compositional kriging

~~4 结论~~

4 结论

本文以区域土壤模糊连续分类隶属度值为例，进行成分数据各种空间预测方法的对比分析。成分数据直接进行插值和经对数正态变换后插值都会存在必然的不确定性或不可靠性。

虽然非对称对数比转换、对称对数比转换和成分克里格插值结果均能满足成分数据插值的~~四~~4 个条件，但三者之间又有差异。从预测精度来看，非对称对数比转换预测精度较低，因此首先被排除。而对称对数比转换和成分克里格插值各有优势，成分克里格插值结果空间分布图连续性特征不如对称对数比转换。而对称对数比转换由于不是直接对原始数据进行预测，因此不能对估计误差方差进行评价，不能保证转回的隶属度值一定满足最优无偏估计，而成分克里

格是直接对隶属度值进行插值，可以进行估计误差方差的评价，且能保证隶属度本身是最优无偏估计。

成分数据的结构分析和区域化预测对土壤科学家提出了特定的问题，为了保证得到更确切的预测结果，使用恰当的方法非常重要，例如对称对数比转换后进行普通克里格插值或成分克里格插值。而不是直接对原始数据进行插值或对原始数据进行正态变换后插值，目前已经成为实事。

参考文献：

[1] Walvoort DJJ, de Gruijter JJ. Compositional kriging: A spatial interpolation method for compositional data. Mathematical Geology, 2001, 33: 951−966

[2] 周蒂. 地质成分数据统计分析－困难和探索. 地球科学－中国地质大学学报, 1998, 23: 147−152

[3] Pawlowsky V, Olea RA, Davis JC. Estimation of regionalized compositions: A comparison of three methods. Mathematical Geology, 1995, 27(1): 105−127

[4] De Gruijter JJ, Walvoort DJJ, Van Gaans PFM. Continuous soil maps – A fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma, 1997, 77: 169−95

[5] Odeh IOA, Todd A-J, Triantafilis J. Spatial prediction of soil particle-size fractions as compositional data. Soil Science, 2003, 168(7): 501−514

[6] 冯娜娜, 李延轩, 张锡洲, 王永东, 廖贵堂. 不同尺度下低山茶园土壤颗粒组成空间变异性特征. 水土保持学报, 2006, 20(3): 123−128

[7] 刘付程, 史学正, 潘贤章, 王洪杰. 苏南典型地区土壤颗粒的空间变异特征. 土壤通报, 2003, 34(4): 247−249

[8] McBratney AB, De Gruijter JJ, Brns DJ. Spatial prediction and mapping of continuous soil classes. Geoderma, 1992, 54(12): 39−64

[9] Triantafilis J, Ward WT, Odeh IO, McBratney AB. Creation and interpolation of continuous soil layer classes in the Lower Namoi Valley. Soil Sci. Soc. Am. J., 2001, 65, 403− 413

[10] Bragato G. Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. Geoderma, 2004, 118: 1−16

[11] Burrough PA, van Gaans PFM, Hootsmans R. Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma, 1997, 77: 115−135

[12] Odeh IOA, McBratney AB, Chittleborough DJ. Fuzzy-c-means and Kriging for mapping soil as a continuous system. Soil Sci. Soc. Am. J., 1992a, 56: 1848−1854

[13] Tan MZ, Xu FM, Chen J, Zhang XL, Chen JZ. Spatial prediction of heavy metal pollution for soils in Peri-urban Beijing, China based on Fuzzy Set theory. Pedosphere, 2006,16(5): 545−554

[14] Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981: 256

[15] McBratney AB, deGruijter JJ. A continuum approach to soil classification by modified fuzzy k-means with extragrades. Journal of Soil Science, 1992, 43: 159−175

[16] Aitchison J. The statistical Analysis of Compositional Data. London: Chapman ＆Hall, 1986

[17] Aitchison J. The statistical analysis of compositional data. J. Royal Stat. Soc. B, 1982, 44:139−177

[18] 檀满枝, 詹其厚, 陈杰. 基于信息熵原理的土壤 pH 影响因素空间相关性分析. 土壤, 2007, 39(6): 953−957

[19] 江宁县土壤普查办公室, 南京市土壤普查办公室. 江苏省江宁县土壤志. 江苏省土壤普查办公室, 1985

[20] 檀满枝, 陈杰. 模糊 *c*-均值算法在区域土壤预测制图中的应用. 土壤学报, 2009, 46(4): 572−577

Influences of Different Interpolation Methods on Spatial Prediction of Compositional Data

—A Case of Fuzzy Membership Values of Soil Continuous Classification

TAN Man-zhi[1], MI Shu-xiao[1,2], LI Kai-li[1,2], CHEN Jie[1,3]

(1 *State Key Laboratory of Soil and Sustainable Agriculture (Institute of Soil Science, Chinese Academy of Sciences), Nanjing* 210008, *China*;

2 *Graduate School of the Chinese Academy of Sciences, Beijing* 100049, *China*;

3 *School of Water Conservancy and Environment Engineering, Zhengzhou University, Zhengzhou* 450001, *China*)

Abstract: Compositional data is very common in geosciences, which must meet four conditions in spatial interpolation, including ensuring positive definiteness and a constant sum of interpolated values at a given position, error minimization and lack of bias. This study took a case of fuzzy

membership values of soil continuous classification, applied three methods of data transformation prior to kriging, i.e., logarithm transformation (LN), asymmetry Logratio transformation (ALR) and symmetry Logratio transformation (SLR). The performance of the transformed values by ordinary kriging was compared with the spatial prediction of the untransformed data using ordinary kringing (UT_{ok}), compositional kriging (CK). The results showed that the sum of interpolated values at a given position wasn't equal to constant 1 by UT_{ok} and LN. Obviously, the above predictive result was theoretically unauthentic. Contrarily, membership values of all the spatial predicted sites summed to 1 when the membership values of the known soils were transformed by asymmetry Logratio and symmetry Logratio approaches and compositional kriging. Comparatively, symmetry Logratio transform could lead to a better spatial continuous distribution pattern. Interpolation results by compositional kriging could keep membership values either unbiased predictions or minimum prediction error variances.

Key words: Compositional data, Asymmetry Logratio transform, Symmetry Logratio transform, Compositional kriging