基于 K-L 变换的多维空间数据正态性检验方法及其应用

赵相伟，靳奉祥，王　健，季　民
（山东科技大学 测绘科学与工程学院，山东 青岛 266510）

摘　要：分析了多维数据正态性检验方法及其适用性，研究了适用于多维空间数据正态性检验的基于 K-L 变换的检验法和基于最小生成树的检验法，并应用蒙特卡罗方法对两种方法进行了对比实验，结果证明基于 K-L 变换的检验法具有检验准确度高、鲁棒性强、运算速度快等优点。应用基于 K-L 变换的检验法对某区域连续 6 年的植被指数和降雨量的差异数据进行了多维正态性检验，进而分析了该区域 6 年中植被指数和降雨量变化的随机性。

关键词：多维空间数据；正态性检验；K-L 变换；最小生成树；蒙特卡罗方法

中图分类号：P208　　**文献标志码**：A　　**文章编号**：1672-3767(2011)02-0041-07

Normality Testing Method and Its Application of Multidimensional Spatial Data Based on K-L Transformation

ZHAO Xiangwei，JIN Fengxiang，WANG Jian，JI Min
（Geomatics College，Shandong University of Science and Technology，Qingdao，Shandong 266510，China）

Abstract：The normality testing method of multidimensional data and its applicability were analyzed. The testing methods for the normality tests of multidimensional spatial data based on Karhunen-Loeve (K-L) transformation and on minimal spanning tree (MST) method respectively were studied and the contrast experiments of these methods were made by using Monte-Carlo method. The results showed that the method based on K-L transformation has advantages, such as high precision, strong robustness and fast calculation, etc. The multidimensional normality testing was made for the difference data of vegetable indices and rainfall in some region in continuous 6 years with the method based on K-L transformation and the randomness of vegetable indices and rainfall variation in 6 years in that region was also analyzed.

Key words：multidimensional spatial data；normality testing；Karhunen-Loeve (K-L) transformation；minimal spanning tree (MST)；Monte-Carlo method

随着测绘、遥感、计算机等相关技术的发展，人们采集了大量的空间数据，形成了海量的多维空间数据。从这些多维空间数据中获得有用的信息，分析地物变化的规律一直是亟待解决的问题。为了研究地物随时间变化的规律，首要解决的问题是怎样有效地评判地物在两个时刻是否发生了变化、变化是否为随机的。由于人们对多维空间数据及差值数据缺乏先验的知识，而且通常不能进行大量的重复采样，需要从能够反映地物在两个时刻状态模式的多维空间数据入手，获得两个多维空间数据的差值数据，并对其进行标准化处理，然后对标准化的差值数据进行正态性检验，若检验结果为正态分布且均值为 0，可据此判定地物在两时刻的

收稿日期：2011-01-05
基金项目：国家自然科学基金项目(41074003).
作者简介：赵相伟(1974—)，男，山东嘉祥人，讲师，博士研究生，主要从事 GIS 软件开发与应用、空间数据处理与分析方面的研究. E-mail：tlzxw1696@163.com.
靳奉祥(1962—)，男，山东淄博人，教授，博士，博士生导师，主要从事大地测量与测量工程、空间数据处理与分析方面的研究. E-mail：fxjin@sdust.edu.cn.

状态模式发生的变化是随机的。因此研究适用于多维空间数据的正态性检验方法对分析状态模式的变化特征具有重要意义。

当前，对一维数据的正态性检验方法已比较成熟，主要包括偏度和峰度等统计量检验法[1]、P-P图或Q-Q图法[1-2]、卡方拟合优度检验法等[2]；对多维数据的正态性检验方法的研究仍在进行，尚未得到能够直接地、精确地判断多维数据正态性的有效方法[3-4]。目前国内外学者对多维数据正态性检验方法的研究可归纳为统计图法、基于假设检验的方法，包括基于偏度和峰度等统计量的检验法[5-9]、马氏距离法[10]、投影变换法[3,10-12]、最小生成树法[4,13]等。这些方法各有特点和适用性，需要对它们进行系统地分析和比较，寻找适合于多维空间数据正态性检验的方法。

1 多维数据的正态性检验方法及其适用性分析

1)统计图法

统计图法是一种传统的多维数据正态性检验方法，该方法利用马氏距离构造检验统计量 D^2，在数据服从正态分布的前提下，由 D^2 的经验分布函数绘制分位数落散点图(Q-Q图)或概率散点图(P-P图)，散点应分布在一条过原点且斜率为1的直线上。若散点明显偏离该直线，则认为不服从正态分布[3]。该方法用 D^2 的经验分布函数估算值作散点分布图，精度较低，同时需要人眼观察散点偏离该直线程度，具有主观性，因此不宜应用到多维空间数据的正态性检验中。

2)基于偏度和峰度统计量的检验法

基于偏度和峰度统计量的多维正态性检验法主要应用Mardia[5-6]提出的偏度与峰度统计量(记为 $b_{M,1}$ 和 $b_{M,2}$)和Srivastava[7]提出的偏度与峰度统计量(记为 $b_{S,1}$ 和 $b_{S,2}$)分别构造新的统计量 $MJB_M=N\left\{\frac{b_{M,1}}{6}+\frac{(b_{M,2}-p(p+2))^2}{8p(p+2)}\right\}$ 和 $MJB_S=Np\left\{\frac{b_{S,1}}{6}+\frac{(b_{S,2}-3)^2}{24}\right\}$ 作为正态性检验的依据[4-10]。由于 MJB_M 和 MJB_S 的极限分布为卡方分布，而且方差较大[8]，因而该方法适合于结合蒙特卡罗采样方法检验某一总体的正态性，不适用于检验一个多维空间数据的正态性。

3)基于马氏距离的检验法

马氏距离法通过计算某一样本向量与已知正态总体均值向量的马氏距离 D_m，构造统计量 $F=\frac{(n-m)n}{(n^2-1)m}D^2\sim F_\alpha(m,n-m)$ 来判断该样本向量是否来自正态总体[10]，因而该方法也不适用于检验一个多维空间数据的正态性。

4)基于投影变换的检验法

基于投影变换检验法的基本思想是把原数据进行适当的变换，投影到另一个空间中，以变换后的数据为基础进行正态性检验，主要有基于投影寻踪的检验方法和基于K-L变换的检验方法等[3,11-12,14]。

基于投影寻踪的检验方法把数据投影到低维空间中，得到相应的统计量，计算统计量的极限分布，进而实现正态性检验。关于投影寻踪检验方法国内外典型的研究成果有K-V型检验、K-V-S型检验、V-S型检验、多维椭球等高分布的检验和投影偏度和投影峰度法[14]。投影寻踪检验方法得到统计量的极限分布，适合于用蒙特卡罗采样方法检验某一总体的正态性，不适用于检验一个多维空间数据的正态性。

基于K-L变换的检验方法将多维数据通过K-L正交变换投影到正交空间中，由于K-L变换是线性变换，不改变数据的正态性(参考2.1)，因此可通过检验变换后数据的正态性，进而判断原多维数据的正态性。这种方法适用于多维空间数据的正态性检验，其工作原理与实现方法在2中详细阐述。

5)最小生成树法

Smith等[4]在利用多维数据集进行模式识别研究时，提出了针对多维数据集正态性检验的基于最小生成树检验法。该方法把被检验数据与模拟生成的正态数据进行混合，利用混合数据之间的距离构造最小生成树和检验统计量。这种方法也适用于多维空间数据的正态性检验，其工作原理与实现方法在3中详细阐述。

通过对目前存在的5种代表性的多维数据正态性检验方法的对比研究，发现基于K-L变换的方法和基

于最小生成树法最适合应用于多维空间数据的正态性检验中，下面对两种方法用于多维空间数据正态性检验的原理、检验方法和流程进行对比分析，并用蒙特卡罗方法进行对比实验。

2 基于 K-L 变换的多维空间数据正态性检验方法

2.1 理论基础

假设 $\boldsymbol{X}_{m,n}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m)^{\mathrm{T}}$ 为含有 n 个样本的 m 维空间数据，其期望估值为 $\hat{\boldsymbol{u}}=\frac{1}{n}\sum_i^n(\boldsymbol{X}_{1,i},\boldsymbol{X}_{2,i},\cdots,\boldsymbol{X}_{m,i})^{\mathrm{T}}$，协方差阵估值为 $\hat{\boldsymbol{\Sigma}}=\frac{1}{n-1}\sum_i^n((\boldsymbol{X}_{1,i},\boldsymbol{X}_{2,i},\cdots,\boldsymbol{X}_{m,i})^{\mathrm{T}}-\hat{\boldsymbol{u}})((\boldsymbol{X}_{1,i},\boldsymbol{X}_{2,i},\cdots,\boldsymbol{X}_{m,i})^{\mathrm{T}}-\hat{\boldsymbol{u}})^{\mathrm{T}}$，若 $\hat{\boldsymbol{\Sigma}}$ 为正定阵，即 $\hat{\boldsymbol{\Sigma}}>0$，则 $\hat{\boldsymbol{\Sigma}}$ 可对角化，且存在特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0$。假设 $\lambda_1,\lambda_2,\cdots,\lambda_m$ 对应的单位特征向量为 $\xi_1,\xi_2,\cdots,\xi_m$，记 $\boldsymbol{T}=(\xi_1,\xi_2,\cdots,\xi_{\mathrm{m}})$，由 K-L 变换的性质知 $\boldsymbol{T}$ 为标准正交矩阵，即 $\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}}=\boldsymbol{I}$（单位矩阵），$\boldsymbol{T}$ 可构成 K-L 变换矩阵[15]。记 $\boldsymbol{Y}=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m)^{\mathrm{T}}$，则 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 相互正交，即 $\boldsymbol{Y}$ 的 m 个分量互不相关，此时 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 也称为 $\boldsymbol{X}_{m,n}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m)^{\mathrm{T}}$ 的主成分。

假设 $\boldsymbol{Y}$ 服从正态分布，由多维正态分布的定义[16] 和几个性质[15-16] 可得到下面的结论。

1）由性质 1（线性变换）“多维正态分布向量的线性变换仍为正态分布”，可推出 $\boldsymbol{X}$ 也服从正态分布。因为 $\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}}=\boldsymbol{I}$，所以 $(\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}})^{-1}=(\boldsymbol{T}^{\mathrm{T}})^{-1}\boldsymbol{T}^{-1}=\boldsymbol{I}$，故 $\boldsymbol{X}=(\boldsymbol{T}^{\mathrm{T}})^{-1}\boldsymbol{Y}$。

2）由性质 2（独立性）“不相关与相互独立等价”，可推出 $\boldsymbol{Y}$ 的 m 个分量 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 相互独立。

3）由性质 3（边际分布）“多维正态分布的边际仍为正态分布”，可推出 $\boldsymbol{Y}$ 的 m 个分量 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 的分布为正态分布。

4）由多元正态分布的定义“独立标准变量 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$ 的有限个线性函数 $\boldsymbol{Y}=\boldsymbol{A}(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)^{\mathrm{T}}+\boldsymbol{\mu}\hat{\boldsymbol{\Sigma}}$ 服从 n 维正态分布”，可推出若 $\boldsymbol{Y}$ 的 m 个分量 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 服从正态分布，则 $\boldsymbol{Y}$ 服从正态分布。

由上述结论可知，若检验出 $\boldsymbol{Y}$ 的 m 个一维分量服从正态分布，则可得知 $\boldsymbol{Y}$ 服从正态分布，进而证明原数据 $\boldsymbol{X}$ 也服从正态分布。因此多维正态性检验问题化为了一维正态性检验问题。

2.2 检验方法与流程

多维空间数据集具有时空特性，数据类型较多，需要对空间数据集进行预处理形成多维空间矩阵数据，在此基础上进行正态性检验，检验方法与流程归纳如下。

1）应用矩阵变换等方法生成多维空间矩阵数据 $\boldsymbol{X}$。

2）计算 $\boldsymbol{X}$ 的均值估值 $\hat{\boldsymbol{u}}$ 和协方差阵估值 $\hat{\boldsymbol{\Sigma}}$，判断 $\hat{\boldsymbol{\Sigma}}$ 是否正定，若正定继续下面的操作。

3）对 $\hat{\boldsymbol{\Sigma}}$ 特征值分解，求得 K-L 正交变换矩阵 $\boldsymbol{T}$。

4）对 $\boldsymbol{X}$ 进行 K-L 正交变换，$\boldsymbol{Y}=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m)^{\mathrm{T}}$，得到各主成分分量 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$。

5）对每个分量 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 分别进行一维正态性检验，对于小样本数据可应用 Lilliefors 检验方法进行检验，对于大样本数据可应用 Jarque-Bera 拟合优度检验方法进行检验。

6）若检验出一个分量不服从正态分布则结束检验过程，输出 $\boldsymbol{X}$ 不服从正态分布的结果；否则循环对每个分量 $\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_m$ 进行正态性检验，若每个分量都服从正态分布则输出 $\boldsymbol{X}$ 服从正态分布的结果。

2.3 算法复杂度

该方法的实现算法时间复杂度为 $O(n)$，空间复杂度为 $O(n\times m)$，相比之下占用 CPU 和内存资源较少，运算速度快，宜于实际应用。

3 基于最小生成树的多维空间数据正态性检验方法

3.1 理论基础

基于最小生成树的多维空间数据正态性检验法把被检验的数据和用蒙特卡罗方法生成的与其具有相同均值和方差、容量为 m' 的多维正态数据相混合，计算混合数据集中所有样本间的欧氏距离。该方法应用最小生成树的最优化原理，按欧氏距离利用 Prim 提出的最小生成树构造方法创建最小生成树，再用最小生成树

构造统计量 $T'=\frac{T-E[T]}{\sqrt{\mathrm{Var}([T\mid C])}}$。其中：$T'$ 的极限分布为标准正态分布[4,13]；T 为最小生成树中被检验的空间数据和生成的多维正态数据点之间相连接的边数；$E(T)=\frac{2mm'}{L}$；$\mathrm{Var}[T\mid C]=\frac{2mm'}{L(L-1)}(\frac{2mm'-L}{L})+\frac{C-L+2}{(L-2)(L-3)}[L(L-1)-4mm'+2]$，$L=m+m'$，$C$ 为最小生成树中每个节点的度 D_i 之和，即 $C=\frac{1}{2}\sum_{i=1}^{L}d_i(d_i-1)$。

应用假设检验方法，对统计量 T' 进行标准正态分布双边检验，若 $|T'|>Z_{(\alpha)}$，则拒绝原假设，认为原空间数据不服从正态分布；反之，接受原假设，认为原空间数据服从正态分布。

3.2 检验方法与流程

在进行正态分布前同样按 2.2 中所述的方法对数据进行处理，然后进行基于最小生成树的正态分布检验，检验方法与流程归纳如下。

1) 应用矩阵变换等方法生成多维空间矩阵数据 $\boldsymbol{X}$。

2) 应用蒙特卡罗方法生成与其具有相同均值和方差的多维正态数据 $\boldsymbol{Y}$。

3) 生成混合矩阵 $\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}$，计算 $\boldsymbol{Z}$ 中每两个样本间的欧氏距离，生成距离矩阵 $\boldsymbol{D}$。

4) 由距离矩阵 $\boldsymbol{D}$ 生成最小生成树 MST，然后计算 MST 中连接被检验的数据点和用蒙特卡罗方法生成的多维正态数据点之间的边数 T，最后计算每个节点深度之和 C。

5) 应用公式 $T'=\frac{T-E[T]}{\sqrt{\mathrm{Var}([T\mid C])}}$ 计算统计量 T'。

6) 对统计量 T' 进行标准正态分布双边检验，输出检验结果。

3.3 算法复杂度

由于该方法需要利用欧氏距离计算最小生成树，因此实现算法的时间复杂度为 $O(4n(n+1))$，空间复杂度为 $O(n(8n+m))$。该方法占用的 CPU 和内存资源很大，不宜于实际应用。

4 对比实验

4.1 实验方案

为了比较上述两种多维空间数据正态性检验方法的准确度、鲁棒性和适用性，在 Matlab7.6 平台中应用蒙特卡罗方法模拟生成了容量 $n=30,50,100,500,1\,000,2\,000,5\,000,10\,000$，维数 $m=2,3,5,10$ 的正态分布数据和维数 $m=2,3,5,10,20$ 的非正态分布数据。其中，应用正态分布数据能够对比两种检验方法“纳真”的能力。非正态分布数据选择了 T 分布数据，因为当自由度 m 比较小时 T 分布数据与正态分布数据差别较明显，当 m 逐渐变大时 T 分布数据逐渐地接近正态分布数据，用这种逐渐地接近正态分布的数据进行实验能够对比两种检验方法“弃伪”的能力。应用基于 K-L 变换的检验方法和基于最小生成树的检验法对每一种情况进行了 10 000 次对比实验，实验结果如表 1 和表 2 所示。表中的准确度为在 10 000 次检验中正态分布数据检验结果为正态分布的频率和 T 分布数据检验结果为非正态分布的频率。

4.2 实验结果分析

1) 基于 K-L 变换的检验结果分析

从表 1 可以看出，基于 K-L 变换的检验法对正态数据检验的准确度受样本容量 n 的影响较小，而随着样本维数 m 的增大而明显减小，这是因为这种方法对每一维都要进行正检验，但 m 不超 10 时准确度达到 77.50% 以上，能够满足一般应用的需要，说明该方法“纳真”的能力较强；对自由度（$m\leqslant 10$）比较小的 T 分布数据非正态数据检验的准确度非常高，说明该方法“弃伪”的能力很强。

2) 基于最小生成树的检验结果分析

从表 2 可以看出，基于最小生成树的检验法对正态数据检验的准确度随着样本容量 n 的增大而增大，随

着样本维数 m 的增大而减小，当 m 为 10 时已经不能进行正态性检验了，说明该方法对于大样本正态分布数据检验的“纳真”能力较强，但对于小样本正态分布数据检验的“纳真”能力较弱；对小样本高维的 T 分布数据进行非正态性检验的准确度较低，说明该方法“弃伪”能力非常差，即犯第二类错误的概率较大；另外当 $m \geqslant 10$ 时，检验准确度没有明显的规律性，没有检验出服从 T 分布的数据随自由度增大逐渐接近正态分布的特征，说明方法的鲁棒性较差。

表 1　基于 K-L 变换检验法的准确度表

Tab. 1　Accuracy of normality testing method based on K-L transformation　%

容量 n	正态数据 m				T 分布数据 m				
	2	3	5	10	2	3	5	10	20
30	95.11	92.63	88.15	77.49	84.18	76.25	65.49	58.96	56.80
50	94.95	92.59	88.19	77.71	96.09	91.58	82.76	71.98	66.69
100	95.12	92.69	88.45	77.75	99.61	99.44	97.88	90.22	82.22
500	95.38	93.27	88.23	77.55	99.99	99.52	99.03	98.99	98.43
1 000	95.08	93.42	88.08	78.24	100	100	100	99.45	99.20
2 000	95.24	93.20	88.14	78.02	100	100	100	100	100
5 000	95.27	93.74	89.33	77.89	100	100	100	100	100
10 000	95.19	93.03	89.45	77.86	100	100	100	100	100

表 2　基于最小生成树检验法的准确度表

Tab. 2　Accuracy of normality testing method based on MST　%

容量 n	正态数据 m				T 分布数据 m				
	2	3	5	10	2	3	5	10	20
30	92.62	87.21	78.93	32.80	22.90	11.22	9.52	36.30	81.58
50	93.24	90.09	84.24	41.26	34.75	13.86	10.05	31.29	84.91
100	94.58	91.70	85.67	58.65	64.29	42.04	12.28	11.67	68.39
500	96.12	94.35	92.22	85.78	98.31	85.43	41.02	5.25	17.00
1 000	98.77	99.50	95.34	89.26	99.68	98.71	91.99	19.29	18.34
2 000	99.00	99.96	97.83	93.26	100	99.97	99.05	29.80	32.12

注：当样本容量 $n > 2\,000$ 时，基于最小生成树算法出现内存溢出，所以只列出了容量 $n \leqslant 2\,000$ 时的结果。

3）两种方法对比分析

基于 K-L 变换的检验方法与基于最小生成树的检验方法相比较，前者受数据分布类型和样本容量的影响较小，鲁棒性强；前者检验准确度较高，尤其“弃伪”能力远比后者强；前者算法的时间复杂度和空间复杂度都比后者小一个量级，运算速度更快。

5　应用案例分析

5.1　用例数据及预处理

为研究某区域植被生长状况随降雨变化的特性，收集了该区域 2004—2009 年 1 km 分辨率的 MODIS 月成品数据和月平均降雨量数据。MODIS 成品数据区域范围经度跨度为 1.581 0°，纬度跨度约 1.490 5°（跨越两景影像的边界）；月平均降雨量数据为从 16 个气象站获取的月平均降雨量数据，气象站分布较均匀。分别对每月的 MODIS 月成品数据和月平均降雨量数据进行了如下预处理过程，获得每年相邻两月的差异数据和相邻两年同月的差异数据。

1）应用专门处理 MODIS 数据的 MRT 软件把数据从正弦坐标系转换到地理坐标系，把两景影像进行合并，然后按照区域范围进行裁切。

2）对降雨量数据应用 ArcGIS 软件进行了克里金插值处理，生成与 MODIS 同范围、同分辨率的栅格数

据，因为 NDVI、EVI 为归一化的数据(乘以系数 Scale，减去偏移量 Offset)，所以降雨量也乘以系数 10^{-4}，取 0～1 之间的数。

3) 从裁切出的 MODIS 数据中提取出 NDVI 和 EVI 两个植被指数波段的数据，根据一定的阈值滤去非植被数据，设置为 NoData。

4) 在 NDVI 或 EVI 的值为 NoData 位置处把月平均降雨量数据栅格值也设置为 NoData。

5) 对每年的相邻两月的 NDVI、EVI 和月平均降雨量数据相减，获得相邻两月的 NDVI、EVI 和月平均降雨量差异数据，分别把 NDVI、EVI 和月平均降雨量差异数据排列为向量，组合在一起构成三元差异矩阵。

6) 同样地，对相邻两年的同月的 NDVI、EVI 和月平均降雨量数据相减，获得同月的 NDVI、EVI 和月平均降雨量差异数据，分别把 NDVI、EVI 和月平均降雨量差异数据排列为向量，组合在一起也构成了三元差异矩阵。

5.2 差异的随机性分析

鉴于 K-L 变换的检验方法具有鲁棒性强、检验准确度较高、计算复杂度低等优点，本文应用该方法对得到的两种差异矩阵数据进行正态性检验。由于 1—3 月份和 11—12 月份植被覆盖率非常低，考虑到数据的有效性，只对每年 4—10 月份的数据进行检验。

同年相邻两月差异矩阵数据检验结果为 2004，2006，2007 和 2009 年的 6，7 月份差异数据为正态数据，据此分析这 4 年 6，7 月份的 NDVI、EVI 和月平均降雨量没有规律性的变化，是随机性变化。其它相邻月份均检验为非正态数据，从 K-L 变换后数据分布拟合图的总体情况上看，前两个成分与正态分布拟合图相似，但峰度都偏大；后一个成分与均匀分布的拟合图相似，但有两峰或多峰。由于分布拟合图较多而且都不完全相同，限于篇幅没有列出分布拟合图，现把每年 5，6 月份和 6，7 月份差异数据的偏度与峰度列于表 3 中，分别表述检验中的非正态数据和正态数据的统计特征。

表 3 K-L 变换后的同年相邻两月差异数据的偏度和峰度表

Tab. 3 Skewness and kurtosis of difference data in adjacent 2 months in same year after K-L transformation

月份	2004 年		2005 年		2006 年		2007 年		2008 年		2009 年	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
	−1.321	2.746	0.298	−0.09	−0.672	0.945	0.701	1.704	0.535	−0.615	−0.615	0.357
5—6	0.351	1.189	−2.646	8.53	−0.839	1.266	−1.757	3.588	−1.279	0.353	0.353	1.270
	0.353	1.375	0.516	1.278	−0.123	0.167	0.006	0.444	0.336	0.204	0.204	0.773
	−0.023	−0.048	0.084	−0.755	0.049	0.041	−0.033	−0.043	−0.383	−0.710	−0.049	0.064
6—7	0.034	0.005	−0.156	−0.371	0.017	−0.032	−0.018	−0.078	−0.058	0.840	0.033	0.025
	−0.002	−0.045	0.044	−0.262	−0.010	0.064	0.020	0.054	−0.310	0.292	−0.031	−0.078

连续 6 年中相邻两年的同月差异数据的正态性检验结果均为非正态分布的数据，但从 NDVI、EVI 和月平均降雨量三个成分数据的分布拟合图上看，都与正态分布的曲线图偏离程度不太大，大多数峰度值偏大。限于篇幅，仅把每相邻两年中 5 月份差异数据的偏度与峰度列于表 4 中，代表性地表述相邻两年同月差异数据的统计特征。

表 4 相邻两年同月差异数据 K-L 变换后的偏度和峰度表

Tab. 4 Skewness and kurtosis of difference data in same month in adjacent 2 years after K-L transformation

月份	2004—2005 年		2005—2006 年		2006—2007 年		2007—2008 年		2008—2009 年	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
	−1.578	1.792	0.149	−0.524	−0.290	−0.340	1.114	3.321	−0.966	1.403
5	−0.003	−0.448	0.242	0.254	−1.260	3.939	−1.74	5.479	1.588	−1.917
	0.250	−0.869	−0.231	−0.387	0.226	0.465	0.436	0.421	−0.121	1.176

对于检验结果为非正态分布的数据情况，说明变化不是完全随机性的，而是具有一定规律性的，因此需要在此基础上对差异数据进一步分析，度量差异数据的信息量，分析 NDVI、EVI 随月平均降雨量变化的模式。

6 结论

分析了当前多维空间数据正态性检验方法中存在的问题，发现了两种适用于多维空间数据正态性检验的方法——基于 K-L 变换的检验法与基于最小生成树的检验法。通过对两种方法进行对比实验，证明基于 K-L 变换的检验法具有检验准确度高、鲁棒性强、运算速度快等优点，而最小生成树法虽然也可以用于检验多维空间数据的正态性，但检验准确度受样本容量和数据分布类型的影响，算法的时间复杂度和空间复杂度较大。最后以植被差异数据的正态性检验为案例，探索了基于 K-L 变换的检验法在多维空间数据正态性检验中的应用方法与过程，为分析地物状态模式变化的随机性提供了有效的方法。

参考文献：

[1]李洪成. 数据的正态性检验方法及其统计软件实现[J]. 统计与决策，2009(12)：155-156.

[2]潘振宇，赵耐青. 一般分布和多元正态分布的检验[J]. 数理医药学杂志，2006，19(1)：53-55.
PAN Zhenyu，ZHAO Naiqing. A new test on random distribution and multivariate normal distribution[J]. Journal of Mathematical Medicine，2006，19(1)：53-55.

[3]汪政红，周清志. 两种多元正态性检验方法的应用和比较[J]. 中南民族大学学报：自然科学版，2009，28(3)：99-103.
WANG Zhenghong，ZHOU Qingzhi. Application and comparison of two methods for multivariate normality test[J]. Journal of South-Central University for Nationalities：Natural Science Edition，2009，28(3)：99-103.

[4]SMITH S P，JAIN A K. A test to determine the multivariate normality of a data set[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1988，10(5)：757-761.

[5]MARDIA K V. Measures of multivariate skewness and kurtosis with applications[J]. Biometrika，1970，57(3)：519-530.

[6]MARDIA K V. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies[J]. Sankhyā，Series B，1974，36(2)：115-128.

[7]SRIVASTAVA M S. A measure of skewness and kurtosis and a graphical method for testing multivariate normality[J]. Statistics & Probability Letters，1984，2：263-267.

[8]KOIZUMI K，OKAMOTO N，SEO T. On Jarque-Bera tests for assessing multivariate normality[J]. Journal of Statistics：Advances in Theory and Applications，2009，1：207-220.

[9]MALKOVITCH J F，AFIFI A A. On tests for multivariate normality[J]. Journal of the American Statistical Association，1973，68：176-179.

[10]纪宏金. 多元正态总体假设检验在矿化带识别中的应用[J]. 长春地质学院学报，1991，21(3)：321-326.
JI Hongjin. An application of test of hypothesis of multivariate normal population to mineralization zone recognization[J]. Journal of Changchun University of Earth Science，1991，21(3)：321-326.

[11]BERAN R，MILLAR P W. Confidence sets for a multivariate distribution[J]. The Annals of Statistics，1986，14(2)：431-443.

[12]BARINGHAUS L，HENZE H. Limit distribution for measure of multivariate skewness and kurtosis based on projections [J]. Journal of Multivariate Analysis，1991，38(1)：51-69.

[13]FRIEDMAN J H，RAFSKY L C. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests[J]. The Annals of Statistics，1979，7(4)：697-717.

[14]陈广雷. 基于投影偏度和投影峰度的投影寻踪自助法的正态性检验[J]. 数学杂志，2006，26(2)：147-154.
CHEN Guanglei. The testing for normality based on pp-skewness and pp-kurtosis with bootstrap method[J]. Journal of Mathematics，2006，26(2)：147-154.

[15]何晓群. 多元统计分析[M]. 北京：中国人民大学出版社，2006.

[16]张尧庭，方开泰. 多元统计分析引论[M]. 北京：科学出版社，1982.