文章编号：1006 - 544X（2009）03 - 0374 - 04

误差分布对 PCA 粗差探测算法的影响

黄世斌[1]，周国恩[1]，梁　鑫[1]，杨晓云[1,2]

（1. 广西工学院 土木建筑工程系，广西 柳州　545006；2. 西南交通大学 土木工程学院，成都　610031）

摘　要：从误差分析的角度出发建立 2 种粗差模型：独立粗差模型和相关粗差模型，采用 MATLAB 随机数发生器生成模拟粗差数据。并利用主成分分析（PCA）算法进行检验，结果表明，误差的空间相关特性降低了 PCA 算法的敏感度，为粗差的探测带来困难，但总体上具有一定的实用性。

关键词：数字高程模型；误差分布；粗差探测；主成分分析法

中图分类号：P207　　**文献标志码**：A

主成分分析（principal component analysis，PCA）是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法，早期主要应用于高维数据的降维处理。此外，PCA 还是一种有效的多维孤立点分析方法，因此常用于异常检测领域[1]。一些学者（Hawkins，1974；López 等，1993）[2-3]尝试采用 PCA 方法对规则格网 DEM 数据进行粗差检测。通常情况下，粗差算法的设计和使用效果受误差分布、地形等因素影响较大。而作为 DEM 产品的最终用户，往往无法获知 DEM 的质量信息，更不要说误差的分布情况。目前，有关误差空间分布的研究微乎其微，这是由于它们总是和生产过程密切相关。由于研究对象通常是单一的 DEM，因此无法对真误差做出合理的分析。为了简化上述问题，本文从误差分析的角度出发，建立了两种不同类型的粗差模型，即独立粗差模型和相关粗差模型。然后，将模拟的粗差数据加载到原始 DEM 中，并采用 PCA 方法对其进行试验分析，最终得到一些有益的结论。

1　基于 PCA 的粗差探测算法

利用 PCA 法进行粗差探测，首先要将原始 DEM（$m \times n$）数据进行分割，以行方向为主划分为长条形状，其大小为 $w \times n$（$w \leqslant n$），这里称为带状区域 DEM（w 为行数，n 为列数），如图 1 所示。带状 DEM 具有空间多元变量的特性，因此可以采用多元统计方法分析，进行粗差探测。

利用 PCA 法进行粗差的探测，首要的问题是如何体现 PCA 与粗差之间的关系。通过大量的试验分析发现带状区域 DEM 主成分得分的分布情况反映了粗差的存在。假设某一带状区域 DEM 含有粗差，那么通常在最后几个主成分得分序列中，粗差数据所对应的得分值会处于边缘分布，同其他值明显区分开。同样，在其他带状区域也可以发现类似的规律，这可能预示这些断面含有粗差。

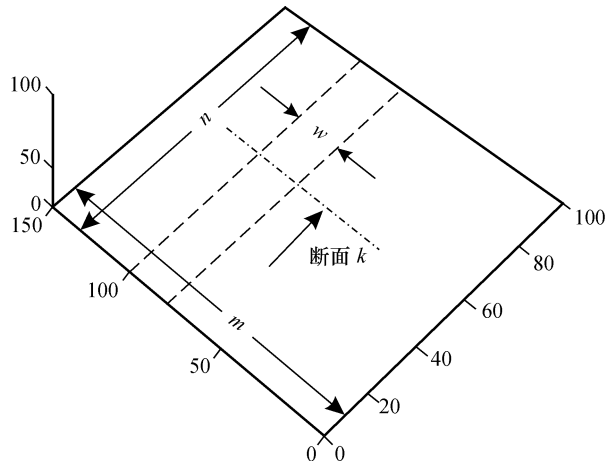


图 1　DEM 数据带状区域分划

Fig. 1　Sketch of the strip notation in DEM

收稿日期：2008 - 11 - 17

基金项目：国家自然科学基金资助项目（40271092）；广西教育厅科研项目（200808LX348）；广西工学院院管基金项目（500634）

作者简介：黄世斌（1974—），男，硕士，讲师，研究方向：大地测量学与测量工程，E-mail：ailiou105@163.com。

以上就是基于 PCA 法进行差异检测的关键思想，即非系统误差对最弱主成分影响较大。

如果通过设定某一阈值来检测每一个主成分得分值，工作量明显增大，Hawkins(1974)[2] 提出一个类似的方法，计算统计量 T_2 来分析得分的分布情况。T_2 是针对断面 k 来定义的，方程式如下：

$$T_2(k) = \sum_{j=1}^{j=w} \frac{1}{\lambda_j} \times a_j(k)^2。 \quad (1)$$

其中：λ_j 为协方差矩阵的特征值；$a_j(k)$ 为第 j 个带状区域断面 k 的主成分得分序列。

假设得分值服从正态分布，则 $T_2(k)$ 的分布函数是混合伽马函数，这一特性并非是严格要求。如 $T_2(k)$ 超过规定的限值，则标记出对应的断面，作为粗差探测的备选项。

通常情况下，$T_2(k)$ 值都较小，但若粗差存在，情况则有所不同。此时，得分序列中的某个值可能远远大于其他值。既然每个得分值都是其对应断面上所有高程数据的线性组合，那么最有可能含有粗差的数值序列应该对 $T_2(k)$ 的影响最大。

因此，对上述标志好的断面 k 进行简单的敏感度分析，那些对 $T_2(k)$ 函数造成巨大影响的列就被划分为 a 类粗差候选序列，同样可以依次选出 b 类、c 类候选序列。同样，还可对以列为主划分的带状区域进行上述操作。将两套备选粗差进行比较，确定最终粗差的位置并进行修正。

由于粗差数据会影响统计量的计算，因而上述过程需要迭代进行。在每一次迭代过程中，都可以得到一组“粗差候选序列”。如果确定为真正的粗差，就对相应的数据点加以修正，并从“候选序列”中去除，接着进行下一步迭代过程，并设置一定的条件来控制程序的进程。

2　试验分析

为了验证 PCA 算法对独立粗差模型和相关粗差模型的检测效果，本文以某一实测 DEM 作为研究区域。首先，利用 MATLAB 随机数发生器模拟粗差数据；然后加载到试验数据中；最后应用 PCA 法对其进行粗差探测，并将试验结果用于统计分析及比较说明。

2.1　数据说明

本文所选试验的 DEM 数据覆盖范围是 1 km × 1 km，位于 ZX 铁路线[4]，格网大小 90 × 90，格网间距 10 m。该区域的高程变化范围 0 ~ 200 m，属于明显的山地地貌，平均高程 178.06 m，高程标准差为 14.918 m。图 2 为该区域地形矩形格网图。

在 R^w 空间中，$n \gg w$ 是保证有充分点位用于判释粗差的先决条件。因此，首先应将原始 DEM 数据划分成一系列的带状区域（图 3）。通过大量的试验分析，选择参数 $w = 9$，即带状 DEM 的维数为 9×90。

2.2　粗差的模拟

采用两种误差模型：独立粗差模型；相关粗差模型。粗差率设定为 2.5%，明显高于 Ostman 提出的粗差率限值 0.5%[5]。

独立粗差模型最早是由 Bethel 和 Mikhail(1984) 等人[6] 提出的：假设 DEM 中的粗差在空间上不相关，即某一格网点的粗差值并不影响周围格网点的质量。如图 4 左侧图形所示，这类粗差在形态上表现为地面单点的突然凸起或凹下，形似山峰，因此又称为峰状误差。本文利用MATLAB中的随

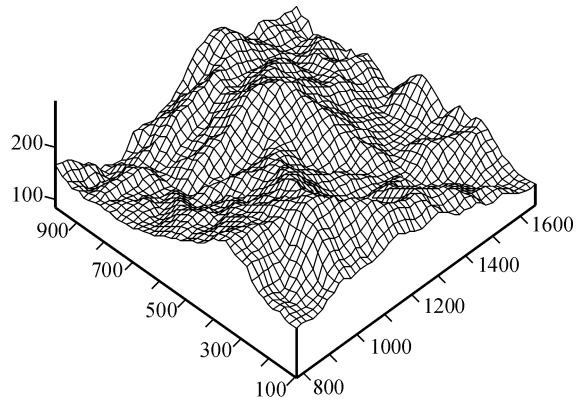


图 2　原始 DEM 数据网格图

Fig. 2　Mesh view of the test digital elevation model

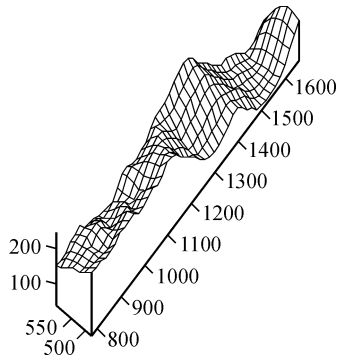


图 3　带状 DEM 数据示意图

Fig. 3　Mesh view of a single strip of DEM

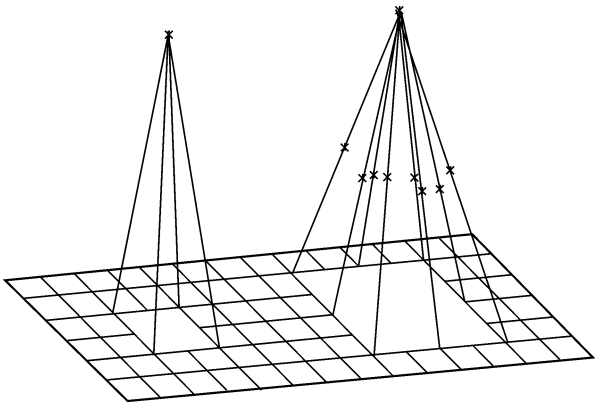


图 4 独立粗差和相关粗差模型

Fig. 4 Sketch of the spike-like and pyramid-like gross error model

机数发生器进行这类粗差的模拟，模拟值为 -10 ~10 m，呈均值分布，其中大于 5 m 的粗差超过总数的 1/2。然后将模拟粗差随机地加入到已知试验数据的高程序列中。

同时尝试采用另一种更为合理的粗差模型——金字塔状模型，如图 4 右侧图形所示。一旦 DEM 某一格网数据点含有 Δ 的粗差，则其周边的 8 个数据点分别含有 0.5Δ 的粗差，Δ 的取值范围 -10 ~10 m。以上就是所谓的金字塔模型，也可称为相关粗差模型，它们在空间上呈相关分布，该模型符合 DEM 生产的实际情况。

2.3 试验结果及分析

假设原始 DEM 数据无粗差存在，采用 PCA 法进行粗差的探测，首先将模拟的粗差（独立粗差和相关粗差）序列随机地加到原始数据中，为了便于分析粗差分布及对原始数据的影响，先给出原始 DEM 加入粗差后的等高线（图 5）。由于相关粗差模型所造成的误差影响范围更大，因此图 5b 等高线的扭曲变形较图 5a 的更为严重。然后，以行为主划分将原始规则格网 DEM 数据划分为带状区域，每一个带状 DEM 大小为 9 × 90。最后，依据 PCA 算法进行检测，得到粗差候选序列。此外，还应对以列为主划分的带状区域进行同样的操作。将两套备选粗差进行比较，确定最终粗差的位置并进行修正。

通过 PCA 法检测，得到修正后的 DEM 等高线（图 6）。可以看出，无论是独立粗差模型，还是相关粗差模型，PCA 算法均可以检测到绝大部分的变形数据，并对其进行修正，原先扭曲的等高线也恢复光滑连续。但从检测效果上分析，图 6a 的效果明显好于图 6b。这是由于空间呈相关分布的粗差数据彼此相互影响，当粗差的绝对值较小时，该高程点和其周边 8 个高程点因粗差所造成的局部地面凸起或凹下被近似认为是自然表面，使得 PCA 法的敏感度降低。因此在图 6b 中，某些粗差并未检测出，呈现为地面的凹点和凸点。也就是说，粗差的空间相关特性通常会给探测算法带来一定的困难，但在总体上具有一定的实用性。

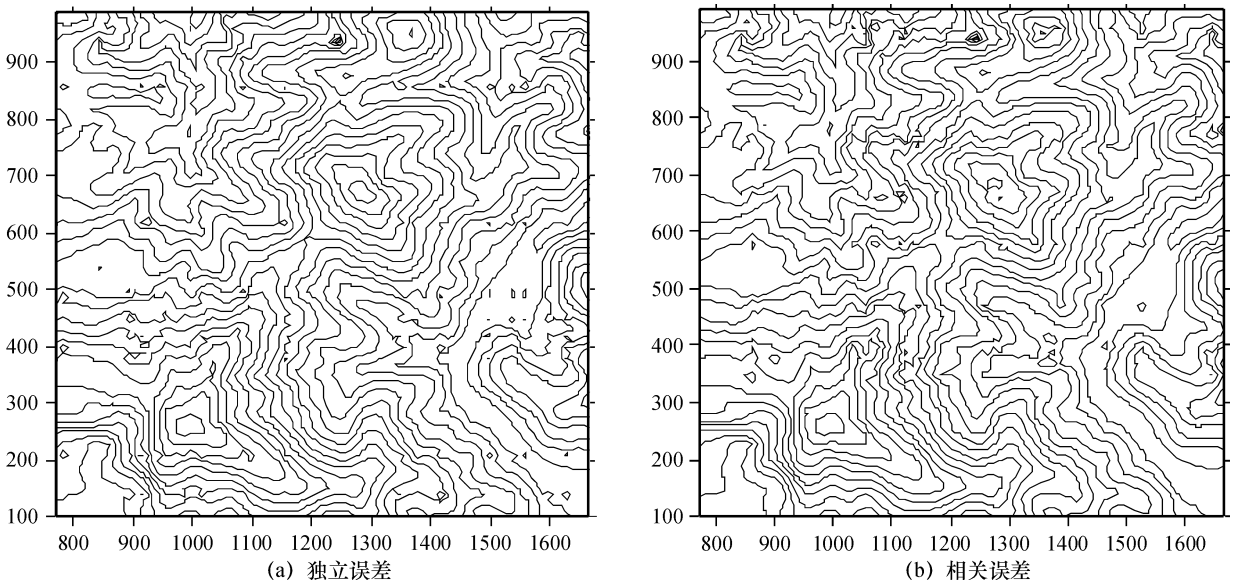
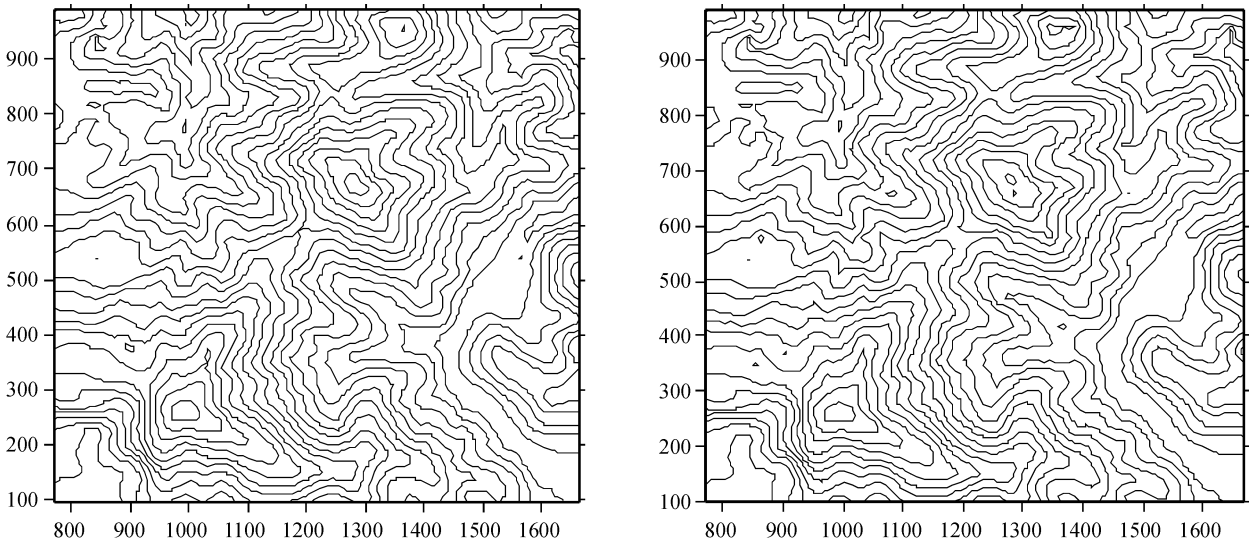


图 5 加入粗差数据后的 DEM

Fig. 5 DEM with simulated gross error



(a) 独立误差　　(b) 相关误差

图 6　修正粗差数据后的 DEM

Fig. 6　DEM with gross error corrected

3　结　论

在充分考虑实际生产中误差分布的情况下，通过独立粗差模型和相关粗差模型的建立，验证了 PCA 算法的有效性。试验表明，由于空间呈相关分布的粗差数据彼此相互影响，当粗差绝对值较小时，误差的影响被近似认为是自然表面的起伏，PCA 算法的敏感度降低。因此，PCA 法对独立粗差的检测效果更好。本文所设计的粗差率达到 2.5%，远远高于 Ostman 提出的粗差率限值 0.5%，具有一定的实用性。

此外，地形因子也是影响粗差探测算法效率的重要因素之一，粗糙的地面、狭窄的河道、陡峭的山峰和小面积的水域，这些地貌均有可能掩盖粗差的存在，这将是下一步研究的主要内容。

参考文献：

[1] Jackson Donald A, Chen Yong. Robust principal component analysis and outlier detection with ecological data[J]. Environmetrics, 2004, 15(2): 129 – 139.

[2] Hawkins D M. The detection of errors in multivariate data using principal components[J]. Journal of the American Statistical Association, 1974, 69: 340 – 344.

[3] López C, Kaplan E. Principal component analysis applied for outlier location and the missing value problem using surface wind and pressure data[DB/OL]. [2008 – 11 – 15]. http://www.fing.edu.uy/ ncarlos/papers/rep93_5/ viento.htm.

[4] 杨晓云，岑敏仪，梁鑫. DEM 粗差探测算法在铁路线路测量中的应用[J]. 桂林工学院学报，2007，27(3)：539 – 541

[5] Ostman A. Quality control of photogrammetrically sampled Digital Elevation Model[J]. Photogrammetric Record, 1987, 12: 333 – 341.

[6] Bethel J S, Mikhail E M. Terrain surface approximation and online quality assessment[J]. International Archives of Photogrammetry and Remote Sensing, Commission Ⅲ, 1984, 25(A3a): 23 – 32.

Effect of Error Distribution on PCA for Detecting Gross Errors in DEM

HUANG Shi-bin[1], ZHOU Guo-en[1], LIANG Xin[1], YANG Xiao-yun[1,2]

(1. *Department of Civil Engineering, Guangxi University of Technology, Liuzhou* 545006, *China*;
2. *School of Civil Engineering, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract: Two error models are built in the test: spike-like error model; pyramid-like error model, which are simulated in MATLAB. Principal component analysis is used to detect the two gross errors. The result shows that the space correlation of error reduces the efficiency of the PCA. Some data polluted by error can not be found, which results in the distortion of original data.

Key words: digital elevation model(DEM); error distribution; gross error detection; principal component analysis (PCA)