适用于不规则DEM数据的粗差探测算法*

杨晓云[1]** 梁 鑫[1] 岑敏仪[2]

1. 广西工学院土木建筑工程系，柳州 545006；
2. 西南交通大学土木工程学院 地理信息工程中心，成都 610031

摘要 粗差的存在会造成数字高程模型(DEM)空间上的严重扭曲，有时能导致DEM及其产品严重失真，甚至完全不能使用. 因而有关DEM的粗差诊断问题已愈来愈引起人们的关注. 文中在对现有的粗差探测算法进行分析的基础上，提出3种针对不规则DEM的粗差检测方法，分别是：基于粗差探测率（即数据点被怀疑为粗差的可能性）算法、基于残差中位数和基于残差均值的算法，并探讨了算法中某些影响粗差检测的因素. 最后文中通过Monte Carlo仿真试验来验证以上3种算法的有效性和可行性.

关键词 数字高程模型 粗差判断率 残差中位数 残差均值

过程质量控制是DEM生产的重要环节之一. 而影响DEM精度的因素是多种多样的，误差的性质可分为3类：系统误差、随机误差和粗差. 与随机噪声相比，粗差对DEM所反映现实空间变化的扭曲更为严重. 因此，设计一些算法检测DEM数据中的粗差并将其消除是完全必要的.

国内外许多学者早已密切注意DEM的粗差探测研究. Hannah[1]在早期的论文中提到过一种针对规则格网数据的粗差探测算法，他以每个点与周围点之间的坡度值及其坡度变化值为限制条件来检测粗差. Östman[2]指出并不存在唯一的标准和量测方法来提高DEM的质量，他认为至少要考虑高程、坡度和曲率的正确性. Felicísimo[3]假设DEM数据误差服从Gauss分布，他以栅格数据中某点高程值与其内插估值(通过其周围的格网点内插)之间的差值建立统计量，用标准t分布检验来分析判断该点是否为粗差. López[4-6]则提出了一种基于主成分分析的方法，他将规则格网数据划分为一系列的带状区域，并将它们看作多元数据序列来处理，通过统计的方法即可查找出高程异常值.

上述算法都是针对规则格网DEM，而它们往往是由不规则DEM数据内插得到. 如果原始DEM含有粗差，在内插过程中就会对多个格网点产生影响，此时采用规则格网的数据去探测粗差，将会增加许多困难，有时甚至会使算法失败. 因此，李志林提出了基于点方式[7]的适用于不规则DEM数据的粗差检测算法，它以检测点P为中心，首先确定一个选取数据点范围的窗口(设半径为R)，然后计算窗口范围内所有点高程的算术平均值(或加权平均值)作为P点的估值(或“真值”)，最后计算P点高程值与估值的高程较差δ_i. 如果δ_i大于另一计算出来的阈值，则认为P点含有粗差.

本文就是在点方式算法的基础上，提出了3种适于不规则DEM数据的粗差探测算法，即基于粗差判断率（即数据点被怀疑为粗差的可能性）算法、基于残差中位数和基于残差均值的算法，并通过Monte Carlo仿真试验来验证以上3种算法的有效性和可行性.

2006-05-08 收稿，2006-09-15 收修改稿

* 国家自然科学基金(批准号：40271092)、香港特别行政区研究基金委员会(编号：香港理工大学5068/99E)和广西工学院硕士基金(批准号：500428)资助项目

** E-mail：ailiou105@163.com

1 基于粗差判断率算法(YXY算法)[8]

YXY算法用邻域数据点的曲面函数拟合残差建立检验统计量. 在计算过程中，每一个DEM的高程点会在不同的数据窗口内参与拟合运算，换句话说，同一个DEM高程点在不同的拟合区域中将会计算出不同的残差值. 当以残差值判释粗差时，不同拟合面内的残差值对同一DEM高程点是否为粗差的判断并不一致，即可能会出现在一个拟合面内认为该点含有粗差，而在另一个拟合面内则认为该点为正确点.

为此，统计各个数据点在各个拟合面内被怀疑为粗差的百分数($percent_i$)，并确定最佳限值 a，当 $percent_i \geqslant a$ 时，则认为点 i 含有粗差. 这就是基于粗差判断率算法(YXY算法)的理论基础. 其中，$percent_i$ 的计算公式如下：

$$percent_i = \frac{\text{点 } i \text{ 被怀疑为粗差的次数}}{\text{点 } i \text{ 出现在局部区域中的总次数}} \times 100\% \tag{1}$$

2 基于残差中位数和残差均值的粗差探测算法

从YXY算法分析可知，如果某一DEM高程数据含有粗差，它所对应的不同局部区域拟合残差超过阈值的可能性就较大，那么从这些残差值所获得的某些统计信息，如算术平均数和中位数，也应在一定程度上反映粗差的位置和大小. 由此可以设想：计算各个数据点在不同局部曲面内的拟合残差，并统计各个数据点所对应的残差均值(或残差中位数)；当以这些均值(或中位数)作为粗差检测的对象时，若它们超过另一计算阈值，则认为其对应的数据点含有粗差；一旦发现粗差立即改正，以保证数据的准确性. 本文将上述算法称为基于残差均值和残差中位数的粗差探测算法.

下面就以均值算法为例，简要说明其实施步骤：

(1) 首先，以移动二次曲面对局部区域进行的最小二乘平差，获得不同窗口的高程残差值列向量 $\boldsymbol{V}=[\boldsymbol{V}_A^{\mathrm{T}} \quad \boldsymbol{V}_B^{\mathrm{T}} \quad \cdots \quad \boldsymbol{V}_W^{\mathrm{T}}]^{\mathrm{T}}$，$W$ 为窗口数目；

(2) 然后从残差列向量 V 中统计出每一个数据点在不同区域中的拟合残差序列，

$$\boldsymbol{V}_i = [v_{i1}, v_{i2}, v_{i3}, \cdots, v_{im}] \tag{2}$$

其中 $\boldsymbol{V}_i$ 对应第 i 个数据点，m 为数据点 i 参与平差运算的局部区域个数；

(3) 计算各个数据点在不同拟合面内拟合残差的均值，

$$\overline{\boldsymbol{V}}_i = \frac{v_{i1} + v_{i2} + \cdots + v_{im}}{m} \tag{3}$$

(4) 一方面以残差均值向量 $\overline{\boldsymbol{V}}$ 作为粗差检测的对象，另一方面又以它为基础计算统计量 δ_{mean}，并以 $k\delta_{\mathrm{mean}}$ 作为粗差检测的阈值；

$$\delta_{\mathrm{mean}} = \pm\sqrt{\frac{\overline{\boldsymbol{V}}^{\mathrm{T}}\overline{\boldsymbol{V}}}{W-1}} \tag{4}$$

其中，W 为局部窗口数目；

(5) 当 $|\overline{\boldsymbol{V}}_i| > k\delta_{\mathrm{mean}}$，则认为 i 点为粗差点. 一旦发现粗差立即改正，以保证数据的准确性.

和均值算法类似，只是中位数算法在步骤(3)中需要计算各个数据点在不同局部区域内拟合残差的中间值，并以此作为(4)式中统计量的计算单元. 中位数的计算，通常的处理方法是首先要将各变量值按其大小顺序排序，然后提取位于中点位置的那个变量值.

3 算例试验

实验选取3组不规则DEM数据，根据地形平均坡度可划分为平原、丘陵和山地，图1为3类DEM数据地形表面图.

为检验上述3种算法对粗差大小的敏感度，以YXY算法为例，设计仿真试验过程如下：

(1) 用Monte Carlo法仿真模拟有 n 维粗差的数据两套，一套为大粗差($>7\bar{\sigma}_0$)，另一套为小粗差($4.5\bar{\sigma}_0$—$7\bar{\sigma}_0$)；

(2) 分别将大、小粗差随机加到DEM数据点上，根据仿真试验得到YXY算法的最佳限值 a(如表1所示)[8]，分别记录大、小粗差试验的正确判断粗差数目，误判粗差数目；

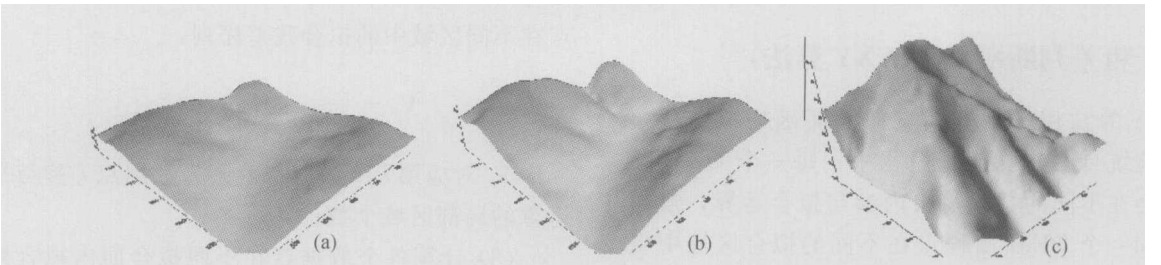

图1 3类DEM数据地形表面图

(a) 平原; (b) 丘陵; (c) 山地

表1 YXY算法针对不同地形的最佳 a 值

地形属性	小粗差	大粗差
平原	0.2	0.2
丘陵	0.3	0.3
山地	0.4	0.5

(3) 重复(1)和(2),共做200次试验. 用下式计算大、小粗差试验的正确检测率(p_d)与误判率(p_c):

$$p_d = \frac{\Sigma(\text{正确判断粗差数目}/n)}{200} \tag{5}$$

$$p_c = \frac{\Sigma(\text{误判粗差数目}/M)}{200} \tag{6}$$

式中 n 为粗差的维数, M 为原始DEM高程点的个数.

(4) 选取不同的粗差维数 n, 重复(1)—(3)的试验, 获得不同粗差率(模拟的粗差维数 n 除以总的DEM数据点数 M)的 p_d 与 p_c.

本文试验设计的粗差维数分别为DEM数据点数的5%, 7.5%和10%, 3种地形类型DEM(平原、丘陵和山地)的试验方案相同. 均值算法和中位数算法的试验过程与YXY算法基本相同, 在探测过程中同样记录这两种算法对应不同粗差率的 p_d 与 p_c, 所不同的是它们避免了限值 a 的选择. 为更好地比较这些新算法的优点, 对相同的实验数据采用点方式算法也进行200次试验.

本文采取直方图的形式对比点方式算法及3种新算法的试验结果, 如图2(a)—(d)所示, 其中横坐标对应粗差率, 纵坐标为粗差检测率或误判率.

由图分析可知, 地形的起伏程度是影响粗差探测的一个重要因素. 在上述测试的4种算法中, 粗差检测率都会随着地形起伏的增大而呈递减趋势, 其中, 在对小粗差的探测中, YXY算法的变化幅度明显大于中位数算法和均值算法, 这说明前者更易受地形变化的影响. 另外, 粗差率也会影响检测效果, 表现为粗差率越高, 粗差检测率 p_d 越低. 这是因为当粗差点增多时, 局部数据窗口含有粗差的可能性增大, 因而对曲面最小二乘拟合参数的影响增强, 使准确探测粗差的可靠性大大降低, 从而导致真正的粗差数据未能判释, 某些正确的数据点却被怀疑, 即误判率 p_c 增加.

本文算法采用移动二次曲面进行局部窗口的最小二乘拟合运算,这较点方式算法在精度上有很大的提高. 此外,新算法还考虑到粗差分布的空间相关性,以各个DEM高程点在不同拟合面内的拟合残差或残差统计量(如均值和中位数)作为粗差判释的对象,使整个算法的设计更加合理,理论上也趋于完善. YXY算法、中位数算法和均值算法用邻域数据点的曲面函数拟合高程值取代点方式算法简单的算术平均值来建立检验统计量, 使得 $\hat{\delta}_{YXY}$, $\hat{\delta}_{中位数}$, $\hat{\delta}_{均值}$ 均小于相应点方式算法的 $\hat{\delta}_0$, 因此前三者较后者有更强的粗差敏感度,尤其是针对于小粗差,如图2(a),(b)所示.

图2(c),(d)对比了上述4种算法对大、小粗差的误判率 p_c,可以看出它们同样受到地形起伏程度和粗差率的影响,通常表现为地形变化越复杂,粗差率越高,则误判率 p_c 就越大. 此外,同其他3种算法相比较,点方式算法虽然检测率 p_d 较低,但它有效地控制了错误判断的情况. 究其原因,这是由于点方式算法以邻域点高程均值作为中心点的内插值,模型误差较大,即 $\hat{\delta}_0$ 大于DEM的实际精度. 这使得该算法对粗差的检测并不是很敏感,因而无论是 p_d 还是 p_c 都偏低. 但从总体上说,YXY算法、中位数算法和均值算法的误判率 p_c 并不是高得不可接受. 从图

2(c),(d)分析中可以看出,3种新算法的 p_c 值最大不超过0.005,而检测效率 p_d 却较点方式算法有很大的提高,因此具有较好的理论意义和实用价值.

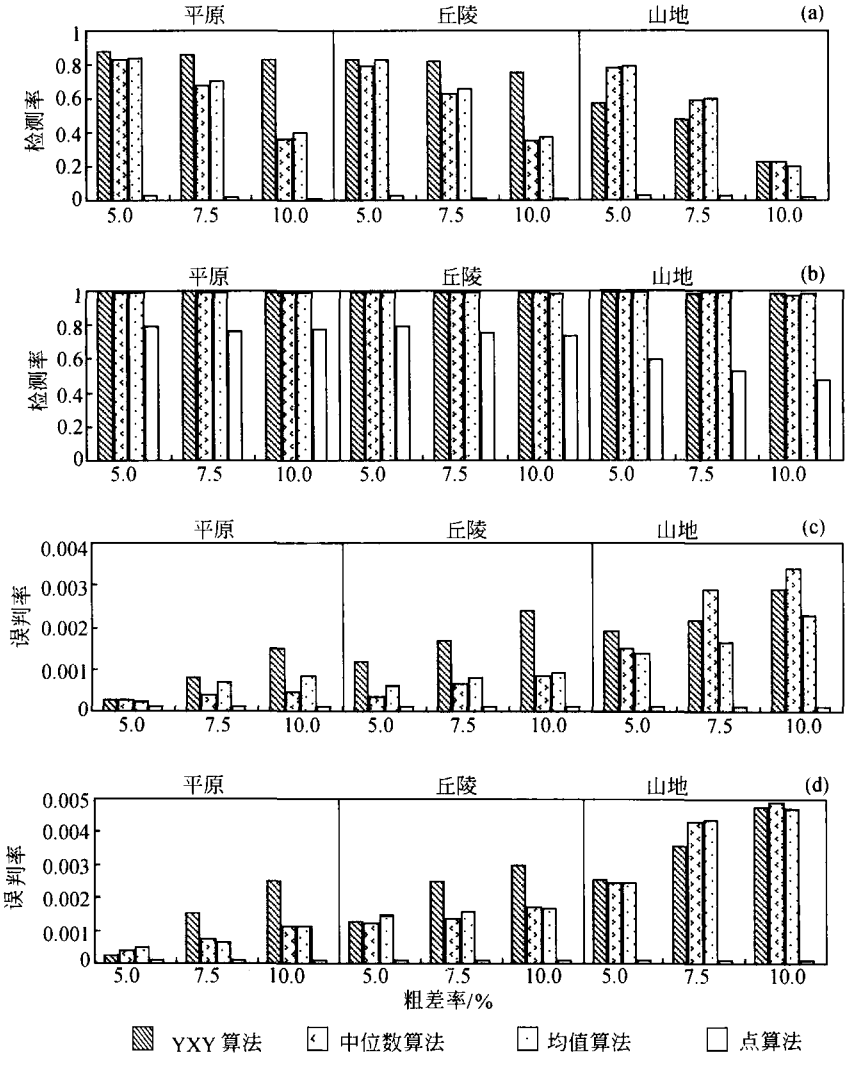

图2 4种算法的试验结果

(a) 4种算法对小粗差探测的检测率对比结果;(b) 4种算法对大粗差探测的检测率对比结果;(c) 4种算法对小粗差探测的误判率对比结果;(d) 4种算法对大粗差探测的误判率对比结果

4 结束语

本文详细论述了3种适于不规则DEM的粗差探测算法,即YXY算法、中位数算法和均值算法,并通过仿真试验的方式验证了它们的有效性和可靠性,其试验结果具有一定的参考价值. 此外,从一定意义来讲,中位数算法和均值算法是对YXY算法的一种改进,因为它们都是对同一问题的不同解决方式,只是前两者避免了限值 a 的选择,实施过程更为简单.

事实上,影响DEM粗差检测的因素是多种多样的,本文仅讨论了地形属性(平原、丘陵和山地)和粗差率对粗差探测的影响,除此之外,还应考虑的因素有DEM离散点的密度、局部区域选点个数、

局部区域拟合函数的选择、临界系数 k 的设定等等. 在算法的设计中，应综合考虑上述因素，采取合理的取值，这将是一个值得进一步探讨的领域.

参 考 文 献

1 Hannah MJ. Error detection and correction in digital terrain models. Photogrammetric Engineering and Remote Sensing, 1981, 47: 63—69

2 Östman A. Quality control of photogrammetrically sampled digital elevation models. Photogrammetric Record, 1987, 12: 333—341

3 Felicisimo A. Parametric statistical method for error detection in digital elevation models. ISPRS Journal of Photogrammetry and Remote Sensing, 1994, 49: 29—33

4 López C. Locating some types of random errors in digital terrain models. International Journal of Geographic Information Science, 1997, 11: 677—698

5 López C. On the improving of elevation accuracy of digital elevation models: a comparison of some error detecting procedures. Transactions in GIS, 2000, 4: 43—64

6 López C. An experiment on elevation accuracy improvement of photogrammetrically derived DEM. International Journal of Geographical Information Science, 2002, 4: 361—375

7 李志林，朱 庆. 数字高程模型. 武汉：武汉测绘科技大学出版社，2000，95—99

8 杨晓云，顾利亚，岑敏仪，等. 基于不同大小窗口的移动曲面拟合法探测不规则 DEM 粗差的一种方法. 测绘学报，2005，34(2)：148—153

《自然科学进展》投稿须知

《自然科学进展》是国家自然科学基金委员会和中国科学院共同主办的综合性学术月刊，刊登自然科学各学科领域的基础研究和应用基础研究方面的高水平、有创造性和重要意义的最新研究成果论文，以促进国内外学术交流. 中文版由各地邮局公开发行，英文版由英国 Taylor & Francis Ltd 总代理，在世界各地发行.

本刊中文版为《中国科技期刊引证报告》的源期刊，并被《中文核心期刊要目总览》、“生物学文摘”等数据库和检索系统收录；英文版(*Progress in Natural Science*)被 SCI Expanded, Chemical Abstracts (CA), Engineering Index (EI), 俄罗斯《文摘杂志》, 美国《数学评论》和日本《科技文献速报》等多种国际检索系统收录.

请直接登录本刊网站(http://pub.nsfc.gov.cn)投稿. 请使用国标(GB3100～3102-93)规定的法定计量单位. 所含曲线图、示意图和照片要尽量精选，原则上总数不超过 6 幅；图题、图注和纵横坐标参数以及图内说明文字均用中文，参数采用国标规定符号；彩版需额外支付制作印刷费. 表格均采用三线表，易引起含混时，可加辅线，对表中所列诸项需特殊说明时，可在表下用 a), b)等注示. 插图和表格排在正文提及后的适当处. 资助项目需在首页脚注中说明.

投稿时请提供如下材料和信息：(i) 申明稿件无泄密之处，未曾正式发表过，也未同时投往他刊；所有作者都了解文章的内容，并同意署名；简要介绍研究工作的背景及成果的意义；明确所投栏目及学科分类. (ii) 作者的所有联系方式. 通讯地址，邮政编码，电话，传真及 E-mail 地址. (iii) 推荐 5—7 名非本单位的具有正高级职称同行评审专家及其单位、通讯地址，也可提出要求回避的专家，供稿件送审时参考.

稿件经同行专家评议后由编辑部做出取舍决定. 不拟刊登的来稿，编辑部将及时通知作者；对于录用的稿件需酌收版面费，论文刊出的当月同时上网，并赠送 1 本样刊.

论文撰写格式请严格遵循本刊的相关要求. 所列文献按正文中引用的先后排序. 文献的作者不多于 3 位时，需全部列出，文献的作者多于 3 位时，只列前 3 位作者，其余用“等”或“et al.”代替.

联系地址：100085 北京海淀区双清路 83 号 国家自然科学基金委员会《自然科学进展》编辑部

联系电话：(010) 62326952, 62327202； **传真**：(010) 62326921；

本刊网址：http://pub.nsfc.gov.cn； **E-mail**：progress@mail.nsfc.gov.cn