DOI:10.14075/j.jgg.2015.05.015　　**文章编号**:1671-5942(2015)05-0793-04

顾及偏态的 IQR 法则在 GPS 时间序列异常值探测中的应用

杨凯钧[1]　袁　鹏[2]　秦昌威[1]

1　武汉大学测绘学院，武汉市珞喻路 129 号，430079
2　武汉大学卫星导航定位技术研究中心，武汉市珞喻路 129 号，430079

摘　要：引入非正态数据分布偏态因子 MC 值对 GPS 时间序列的偏态性进行度量，并构造相应的函数，改进标准 IQR 法则中异常值的探测区间。利用 Sopac 提供的 CHAN 站时间序列进行实验，剔除异常值后进行插补，人工添加不同量级、不同密度的异常值。分别对比两种方案的探测结果，证实顾及偏态的 IQR 法则较标准 IQR 法则算法更严谨、探测效果更好。

关键词：时间序列；IQR；异常值；偏态；分位值
中图分类号：P228　　**文献标识码**：A

对 GPS 时间序列进行处理之前，需要对其异常值信息进行处理。常用的处理方法是基于统计学稳健孤立值的探测法——标准 IQR 法则[1-2]，其具有很好的抗差性。然而 IQR 法则针对的是具有很强对称性的正态分布数据，而 GPS 时间序列不一定构成标准的正态分布，往往表现出部分偏态的性质。再加上天线变更、迁站、地震等因素会导致 GPS 时间序列发生突变，其偏态性则更加明显。所以，标准的 IQR 法则在 GPS 时间序列探测上的准确性和可靠性受到质疑。为改进标准 IQR 法则的不足，本文引入顾及偏态的 IQR 法则对 GPS 时间序列进行粗差探测。

1　标准 IQR 法则

标准 IQR 法则[3]是在假定数据符合标准正态分布情况下，对数据从小到大排列，取其中位值和标准化的 IQR 参数进行数据集散性分布的度量。数据从小到大排列后，取离 1/4 处最近的值为下分位值 Q_1，离 3/4 处最近的值为上分位值 Q_2，四分位间距 IQR 表示为：

$$\mathrm{IQR}=Q_2-Q_1 \tag{1}$$

构造 IQR 准则下的异常值探测区间：

$$[Q_1-1.5\cdot\mathrm{IQR},Q_2+1.5\cdot\mathrm{IQR}] \tag{2}$$

式中变量均在同一个时间窗口 w 内构造，w 为所取时间序列的窗口宽度(一般取 1 a)。设 v 为待判断时间序列的值，若 v 在上述区间外，则判定为异常值(置信度 99%)。循环遍历每个数据值，以当前值为中心构造若干个长度为 1 a 的时间窗口，计算其分位值，构造判断区间，直到判断完时间序列中所有的数据值[4]。

2　顾及偏态的 IQR 法则

引入 Brys 提出的偏态度量因子 MC[3,5]：

$$\mathrm{MC}=\underset{p_i\leqslant m\leqslant p_j}{\mathrm{med}}\,h(p_i,p_j) \tag{3}$$

med 为取中值函数。当 $p_i\neq p_j$ 时，

$$h(p_i,p_j)=\frac{(p_j-m)-(m-p_i)}{p_j-p_i} \tag{4}$$

时间序列从小到大排列后，中间若有若干值相等，即当 $p_i=p_j=m$ 时，$p_{n_l}=m$，$l=1,2,\cdots,k$。

$$h(p_{n_i},p_{n_j})=\begin{cases}-1, i+j<k\\ 0, i+j=k\\ +1, i+j>k\end{cases} \tag{5}$$

其中，m 为时间序列窗口的中位数，p 为时间序列数据值，i、j 为时间序列序列号。可以看出，MC 的取值范围是[−1,1]。MC>0，时间序列右偏；MC<0，时间序列左偏；MC=0，时间序列对称分布。Brys 还论证了 MC 值既对偏态敏感，又具有较强的抗差性，符合探测 GPS 时间序列异常值的

收稿日期:2014-08-12
项目来源:国家自然科学基金(41374033)。
第一作者简介:杨凯钧，硕士生，主要从事 GNSS 精密数据处理研究，E-mail：apple_hacker@163.com。

需要[3,5]。

为将偏态信息引入探测区间，将 IQR 法则的探测区间改造为：

$$[Q_1 - f_m(\mathrm{MC}) \cdot \mathrm{IQR}, Q_2 + f_n(\mathrm{MC}) \cdot \mathrm{IQR}] \quad (6)$$

如需指数函数能较好地满足区间改造的符合度，还得满足 MC=0 时改造区间与原区间一致，即

$$\begin{cases} f_m(\mathrm{MC}) = 1.5\mathrm{e}^{a\mathrm{MC}} \\ f_n(\mathrm{MC}) = 1.5\mathrm{e}^{b\mathrm{MC}} \end{cases} \quad (7)$$

为求参数 a 和 b 的值，满足改造后的区间也符合 0.7%的超探测区间概率，即

$$\begin{aligned} Q_1 - 1.5\mathrm{e}^{a\mathrm{MC}}\mathrm{IQR} &= Q_{0.003\,5} \\ Q_2 + 1.5\mathrm{e}^{b\mathrm{MC}}\mathrm{IQR} &= Q_{0.996\,5} \end{aligned} \quad (8)$$

其中，$Q_{0.003\,5}$ 表示时间序列的 0.003 5 分位值，$Q_{0.996\,5}$ 表示时间序列的 0.996 5 分位值。则：

$$\begin{aligned} a\mathrm{MC} &= \ln\left(\frac{2 \cdot (Q_1 - Q_{0.003\,5})}{3 \cdot \mathrm{IQR}}\right) \\ b\mathrm{MC} &= \ln\left(\frac{2 \cdot (Q_{0.996\,5} - Q_2)}{3 \cdot \mathrm{IQR}}\right) \end{aligned} \quad (9)$$

通过对时长 1 a 的移动窗口依次探测，可利用改造后的顾及偏态的 IQR 法则探测出各异常值[6-7]。

3 实验及结果比较

为比较标准 IQR 法则和顾及偏态的 IQR 法则在 GPS 时间序列中的探测能力，选取由 Sopac 提供的长春 IGS 站点 CHAN 从 2004 年第 343 d 到 2010 年第 283 d(部分空缺)共2 000个观测日的时间序列进行实验。整体计算发现，N 方向 MC 值为−0.021 6，E 方向 MC 值为−0.051 0，U 方向 MC 值为−0.070 4，3 个方向的数据偏态均向左偏。具体实例化操作过程为：将 GPS 时间序列的 N、E、U 3 个维度依次进行处理，每一维度从当前历元开始往前后各递推长度为 182 d(一共 365 d)的数据作为处理窗口，在窗口范围内对当前维度值进行从小到大的冒泡排序，然后再依次求得各分位值与 MC、IQR 值。处理过程中发现，3 个方向均有左偏与右偏窗口，偏态性在各窗口内表现参差不齐。分别用两种方法探测出粗差并予以剔除后，采用 Sopac 给出的 IGS 基准站改进函数模型，采用正弦、余弦函数描述的测站年、半年周期变化，将 CHAN 站 N、E、U 方向分别建模，并进行插补[8]：

$$\begin{aligned} y(t_i) = a + bt_i + c\sin(2\pi t_i) + \\ d\cos(2\pi t_i) + e\sin(4\pi t_i) + f\cos(4\pi t_i) + \\ \sum_{j=1}^{n_g} g_j H(t_i - T_{gj}) + \nu_i \end{aligned} \quad (10)$$

其中，t 表示日坐标解历元，a 为测站位置，b 为测站线性速度，c、d 是描述测站年周期运动的参数，e、f 是描述测站半年周期运动参数。H 为海维西特阶梯函数，用来表示时间序列的突变，突变前其值为 0，突变后为 1，g 为跳变量大小，n_g 为跳变个数，ν_i 为观测噪声。

图 1 给出了 CHAN 站点采用时间序列改进函数模型在 N、E、U 3 个方向拟合的结果。采用两种方法去除异常值后的点都利用式(10)进行拟合插补，从而得到一组没有粗差的 GPS 时间序列。在这组没有粗差的时间序列中，等间隔地加入不同大小的粗差，并使用两种方法对其探测。

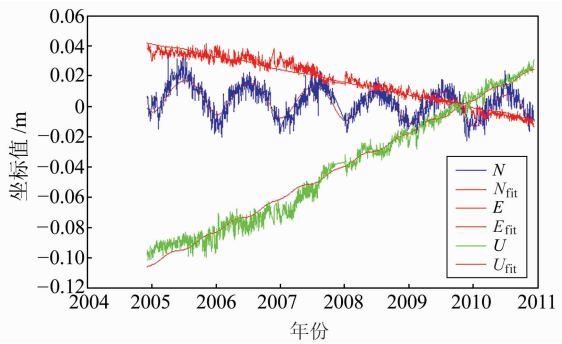

图 1 CHAN 站点时间序列及拟合结果

Fig. 1 The picture of CHAN's time series and the fitting results

在 2 000 个观测日中每隔 50、75、100 d 增加 1、2、3 cm，以测试两种方法分别对不同比例和不同量级异常值的探测能力，结果见图 2～4。为进一步验证顾及偏态的 IQR 法则的异常值探测成功率，并探究异常值间隔与探测成功率的关系，本文选取中国及其周边部分 IGS 站点的长期时间序列，采用如同 CHAN 站的处理方式，利用顾及偏态的 IQR 法则探测异常值并统计各站点 N、E、U 3 个方向的探测成功率(表 1)。

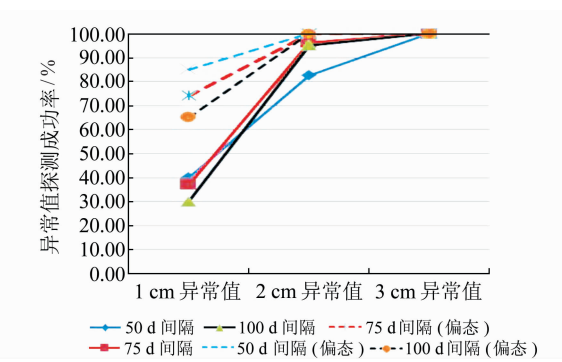

图 2 N 方向探测结果

Fig. 2 The detection results of N direction

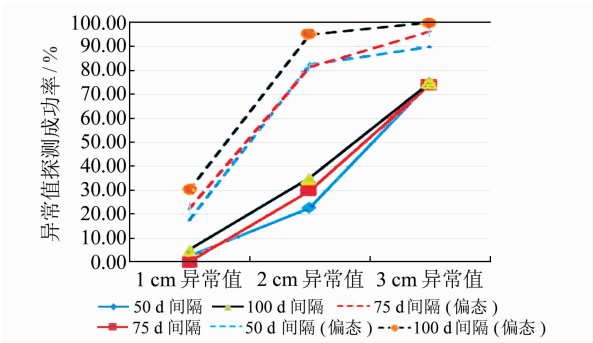

图 3　E 方向探测结果

Fig. 3　The detection results of E direction

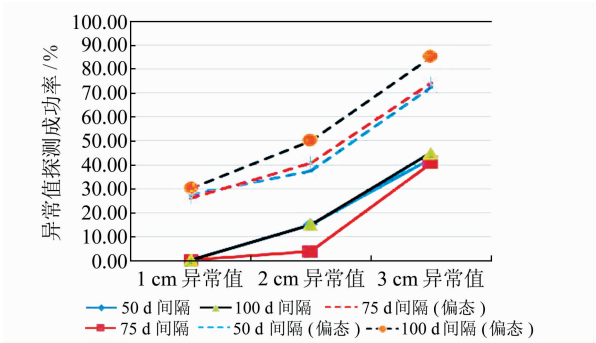

图 4　U 方向探测结果

Fig. 4　The detection results of U direction

表 1　部分 IGS 站点时间序列顾及偏态方法探测异常值的成功率

Tab. 1　The success rate of some IGS sites

	N 方向粗差间隔/d			E 方向粗差间隔/d			U 方向粗差间隔/d		
	50	75	100	50	75	100	50	75	100
CHAN	100.00%	100.00%	100.00%	82.50%	81.48%	95.00%	37.50%	40.74%	50.00%
URUM	82.35%	88.24%	100.00%	70.59%	73.53%	79.41%	47.06%	64.71%	76.47%
KUNA	81.82%	88.64%	90.91%	84.09%	93.18%	97.73%	63.64%	75.00%	81.82%
BJFS	67.44%	83.72%	97.67%	76.74%	88.37%	100.00%	48.84%	58.14%	76.74%
SHAO	76.67%	83.33%	96.67%	66.67%	70.00%	91.67%	55.00%	63.33%	68.33%
POL2	85.42%	89.58%	100.00%	93.75%	100.00%	100.00%	79.17%	85.42%	89.58%
WUHN	92.73%	90.91%	100.00%	87.27%	90.91%	96.36%	80.00%	87.27%	92.73%
KIT3	63.46%	65.38%	73.08%	55.77%	59.62%	65.38%	46.15%	50.00%	55.77%
TNML	70.00%	82.50%	90.00%	95.00%	100.00%	100.00%	42.50%	52.50%	62.50%
LHAZ	68.75%	79.17%	87.50%	91.67%	97.92%	100.00%	54.17%	60.42%	77.08%
USUD	60.00%	62.68%	82.86%	71.43%	77.14%	88.57%	31.43%	37.14%	45.71%

从图 2～4 可以看出,顾及偏态的 IQR 法则对异常值的探测更敏感,其探测效果比标准 IQR 法则有所提高,并且对小异常值的探测效果明显优于标准 IQR。随着异常值的增大,两种方法的探测能力都在增加,并逐步趋近,但标准 IQR 法则的探测能力始终未能超过顾及偏态的 IQR 法则。为保证异常值的探测成功率,建议采用估计偏态的 IQR 法则处理异常值量级 1 cm 以上的时间序列。常规的 GPS 时间序列绝大部分异常值量级均处于 1 cm 以上,因此顾及偏态的 IQR 法则在 GPS 时间序列异常值探测中完全适用。从表 1 可以看出,对于间隔量(密度),两种方法的探测能力在 N、E、U 方向上均随异常值密度的减小而逐步增强,只有个别站点在探测成功率很高时,可能出现稀疏的异常值的探测成功率略低于稍密集的异常值。顾及偏态的 IQR 法则引入偏态因子 MC 值,有效地改正了异常值的探测区间,使 IQR 法则能更准确地探测一般性的时间序列,而不仅是符合标准正态分布的时间序列。

4　结　语

本文顾及 GPS 时间序列的偏态性,引入数据偏态度量因子 MC 值,改进了标准 IQR 法则的探测区间,并通过实验研究标准 IQR 法则与顾及偏态的 IQR 法则对不同量级、不同密度的异常值探测的情况。结果表明,两种方法均对异常值的增大表现出更高的敏感性,而对异常值密度的改变未表现出明显的规律。但无论何种情况,顾及偏态的 IQR 法则的探测能力总是优于标准 IQR 法则,而且在量级较小的异常值探测时比标准 IQR 法则具有更大的优势。顾及偏态的 IQR 法则逻辑上更加严密,算法更为严谨,适用于一般性情况,而标准 IQR 法则仅适用于符合正态分布的 GPS 时间序列,是顾及偏态 IQR 法则的一种特殊情况(MC=0)。

参考文献

[1]　黄立人. GPS 基准站坐标分量时间序列的噪声特性分析[J]. 大地测量与地球动力学, 2006, 26(2):31-33(Huang Liren. Noise Properties in Time Series of Coordinate Component at GPS Fiducial Stations[J]. Journal of Geodesy and Geodynamics, 2006, 26(2): 31-33)

[2]　张恒璟, 程鹏飞. 基于 GPS 高程时间序列粗差的抗差探测与插补研究[J]. 大地测量与地球动力学, 2011, 31(4):71-75(Zhang Hengjing, Cheng Pengfei. Study on Robust Detection and Interpolation from Gross Errors of GPS Height

Time Series[J]. Journal of Geodesy and Geodynamics, 2011, 31(4):71-75

[3] Brys G, Hubert M, Struyf A. A Robust Measure of Skewness[J]. Journal of Computational and Graphical Statistics, 2004, 13(4):996-1 017

[4] 赵超. 一种降雨异常值探测的新方法[J]. 水利水电技术, 2012, 43(7): 13-16 (Zhao Chao. A New Method for Detection of Abnormal Rainfall Data[J]. Water Resources and Hydropower Engineering, 2012, 43(7): 13-16)

[5] Hubert M, Vandervieren E. An Adjusted Boxplot for Skewed Distributions[J]. Computational Statistics & Data Analysis, 2008, 52(12): 5 186-5 201

[6] Hubert M, Stephan V D V. Outlier Detection for Skewed Data[J]. Journal of Chemometrics, 2008, 22(3-4):235-246

[7] Brys G, Hubert M, Rousseeuw P J. A Robustification of Independent Component Analysis[J]. Journal of Chemometrics, 2005, 19(5-7): 364-375

[8] 黄立人. 海面变化的动态预测[J]. 海洋通报, 1991, 10(1): 1-6(Huang Liren. Dynamic Prediction of Sea Level Variation[J]. Bulletin of Oceanography, 1991, 10(1): 1-6)

Consider the IQR Law in the Detection of Outlier in Time Series Considering Skewness

YANG Kaijun[1] *YUAN Peng*[2] *QIN Changwei*[1]

1 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China

2 Research Center of GNSS, Wuhan University, 129 Luoyu Road, Wuhan 430079, China

Abstract: In order to analyze the skewness of GPS time series data, a skewness factor MC is introduced for measurement as well as construction of the corresponding function in order to improve the detection range of the IQR law. We interpolate after rejecting the outliers, using the time series of the CHAN IGS station provided by Sopac experimental sequence, artificially adding different orders of magnitude and different densities of outlier. We confirmed that the IQR law considering skewness is more rigorous in algorithm and more effective in outlier detection than IQR law alone.

Key words: time series; IQR; outlier; skewness; fractile

(上接第 792 页)

and the observed data from nine stations. By comparing the two models, we find that the new model has a higher correction rate with respect to the broadcast model. Moreover, using the new model to forecast ionospheric delay within one week also has a better correction effect than the broadcast model. Statistical data further shows that the average correction precision improvement on grid points is from 67.89% to 78.44%, and the observed data of different stations from 69.81% to 82.34%.

Key words: ionosphere delay model; new Klobuchar model; relax iterative search; Beidou satellite navigation system