实测潮位异常值判别方法比较

罗俐雅[1]，崔彦萍[1]，甘　敏[2]，陈永平[2]

（1. 江苏省水文水资源勘测局，江苏 南京　210029；2. 河海大学，江苏 南京　210098）

摘要： 针对近岸自动测站实测潮位数据质量偏低的问题，分别采用拉伊达准则、肖维勒准则和方国洪准则对其异常值进行自动判别，利用 T_TIDE 软件对判别后的实测潮位进行调和回报分析，以检验不同准则的判别效果。结果表明，采用拉伊达准则可以对潮位数据序列中异常值密集和连续出现的情况有较好的判别效果，总体上拉伊达准则较肖维勒准则和方国洪准则表现更优。通过异常值判别后，潮位调和分析的回报精度有明显提高。

关键词： 实测潮位；拉伊达准则；肖维勒准则；方国洪准则；异常值

中图分类号： TV12　　**文献标识码：** B　　**文章编号：** 1007-7839（2018）04-0037-05

Comparison of discriminating methods on the measured tide level outliers

LUO Liya[1]，CUI Yanping[1]，GAN Min[2]，CHEN Yongping[2]

（1. *Jiangsu Hydrology and Water Resources Survey Bureau*，*Nanjing* 210029，*Jiangsu*；
2. *Hohai University*，*Nanjing* 210098，*Jiangsu*）

Abstract： In view of the low data quality of the actual measured tidal level in the near-shore automatic station, the Rajda criterion, Chauvenet criterion and Fang Guohong criterion were used to automatically determine the abnormal value, and the T_TIDE software was used to perform the tune-in-return analysis on the measured tide level after the discrimination, which could test the distinguishing effects of different criteria. The results showed that the Rajda criterion could be well used to discriminate the occurrence of dense and continuous outliers in tidal level data series. In general, the Rajda criterion performed better than the Chauvenet criterion and Fang Guohong criterion. The return precision of the tidal harmonic analysis was obviously improved after the discrimination of the abnormal value.

Key words： measured tide level；Pauta criterion；Chauvenet criterion；Fang Guohong criterion；abnormal value

0　引言

在实际的潮位观测中，尤其是长时间的自动观测，由于受到仪器故障、恶劣天气、地理位置制约和观测方式等因素的影响，很难得到从观测初始时刻到结束时刻这段时间内完整的高质量数据资料[1]。如果将含有异常值的数据直接用于潮汐调和分析，有可能带来较大的预报误差[2]。因此，科学合理地判别异常值，对于准确的潮位分析至关重要。

通常有 2 种方式对异常潮位数据进行处理。第 1 种是手工处理，即通过比较数值的大小或分析要素的变化趋势等进行判别处理[3]。这种方式主要取决于操作员的主观判断，可靠性无法保障，当潮位数据多时，工作量将非常大；第 2 种是通过给计算机设定一个判别准则，让计算机自动判别

收稿日期： 2018-02-11
基金项目： 江苏省水利科技项目（2015006）
作者简介： 罗俐雅（1978-），硕士，高级工程师，主要从事水情情报预报工作。

其异常值[3]。方国洪等[4]介绍了2种计算机判别异常值的方法，第1种是利用2次抛物线拟合得到一个拟合值，通过比较实测值与拟合值的差值来判断数据是否异常，当异常值较少时，此方法能较方便地找出异常值，如果异常值周期性出现，此方法不再适用；第2种是根据大误差出现的可能性来判断，该方法基于概率论理论，设定了判别标准，找出异常值的效率较快，为方国洪等的推荐准则。此外，许军等[5]借助余水位曲线的变化趋势来判断，可以很好地判别以离散形式出现的异常点数据，但不太适用异常值数据集中出现时的情况。董玉磊等[6]采用了基于回归分析的方法来判断异常潮位，该方法是通过分析被检测数据所在的验潮站与附近验潮站之间的线性关系来判别异常值，能有效地判别出由验潮仪零点逐渐变化而带来的潮位数据异常等问题，此方法是基于验潮站与附近验潮站之间的回归分析，需要附近有验潮站才能判别。

鉴于上述分析，当前针对实测潮位异常值判别方法存在一定局限性，有必要探讨如何高效准确地识别近岸实测潮位中的异常值。事实上，潮位异常值与真实值之间的误差可以当作粗大误差[7]，熊艳艳等[8]介绍了多种粗大误差的判别方法，并对它们的适用性做了比较，其中拉伊达准则、肖维勒准则[8]适用于样本数较多且服从正态分布异常值的检验，它们被应用到异常波浪的判别中[9]取得了良好效果。本文将上述2种准则引入到潮位异常值的判别中，并与方国洪准则进行比较，通过对潮位资料的调和分析，定量说明3种准则的判别效果。

1 三种判别准则

1.1 拉伊达准则

拉伊达准则基于样本服从正态分布的假定，认为被检验值与平均值之间差值的绝对值超过3倍样本的标准差时被检验值数据异常，需要舍弃，然后重新生成样本继续判断。

对于潮位异常值检验，首先假设所有实测潮位值都是正常的，第 i 个实测潮位值为 x_i，用实测潮位值进行回报的对应潮位为 h_i，误差 r_i 计算如下：

$$r_i=x_i-h_i \tag{1}$$

潮位资料通常为1年的逐日每小时数据，设有 N 个，将误差作为样本，假设误差服从正态分布，计算样本的平均值 $\bar{r}$ 和标准差 S 为：

$$\bar{r}=\frac{\sum_{i=1}^{N} r_i}{N} \tag{2}$$

$$S=\sqrt{\frac{\sum_{i=1}^{N}(r_i-\bar{r})^2}{N-1}} \tag{3}$$

$$|r_i-\bar{r}|>3S \tag{4}$$

当 r_i 与 $\bar{r}$ 的差值超过 $\pm 3S$ 时，认为 r_i 为异常值，即对应的潮位 h_i 异常。

1.2 肖维勒准则

肖维勒准则假设样本服从正态分布，认为在 N 个数据点中，出现概率小于 $1/2N$ 的数据点，可认为是异常值，应该舍弃然后重新生成样本继续判断。设 Z 为某个大于0的值，当 $|r_i-\bar{r}|\leqslant ZS$ 时，满足如下函数关系：

$$P\left(\frac{|r_i-\bar{r}|}{S}\leq Z\right)=2\frac{1}{\sqrt{2\pi}}\int_0^Z e^{-\frac{t^2}{2}}dt \tag{5}$$

设存在某个特定的 Z_c，称其为肖维勒准则数，当 $\frac{|r_i-\bar{r}|}{S}>Z_c$ 时，认为出现了概率小于 $1/2N$ 的数据点，则：

$$P\left(\frac{|r_i-\bar{r}|}{S}>Z\right)=\frac{1}{2N} \tag{6}$$

联合式(5)、式(6)可得：

$$2\frac{1}{\sqrt{2\pi}}\int_0^Z e^{-\frac{t^2}{2}}dt=1-\frac{1}{2N} \tag{7}$$

N 已知时，可以根据式(7)解得 Z_c，若 $|r_i-\bar{r}|>Z_cS$，即可认为 r_i 为异常值，即对应的潮位 h_i 异常。

1.3 方国洪准则

方国洪准则假设误差服从正态分布，且其平均值为零，方差为 υ_r。实际分析时，先假设所有数据正常，当选用了 J 个分潮用于调和分析自报时，观测误差平方值 r_i^2 为：

$$r^2{}_i=\frac{N}{N-2J-1}(x_i-h_i)^2 \tag{8}$$

方差 υ_r 为：

$$\upsilon_r=\frac{1}{N-2J-1}\sum_{i=1}^{N}(x_i-h_i)^2 \tag{9}$$

在这个假设下，某个被检验值误差小于 Z 的概率 P 为：

$$P=\frac{2}{\sqrt{2\pi}}\int_0^{\frac{Z}{\sqrt{v_r}}}e^{-\frac{t^2}{2}}dt \tag{10}$$

所有点误差均小于 Z 的概率 P 为：

$$P_0=P_N \tag{11}$$

如果给定 P_0，则可以求出 Z，假设为 μ，称它为临界系数，使得：

$$\left[\frac{2}{\sqrt{2\pi}}\int_0^{\mu}e^{-\frac{t^2}{2}}dt\right]^N=P_0 \tag{12}$$

μ^2 近似按照下式计算：

$$\mu^2=a+b\ln N+c\ln^2 N \tag{13}$$

式中 a、b、c 为系数，取值见表 1。

表 1　对应 P_0 下 a、b、c 系数取值表

P_0	a	b	c
0.80	1.40	1.680	0.0126
0.90	2.56	1.738	0.0096
0.95	3.75	1.776	0.0078
0.99	6.59	1.837	0.0045

若有某个值的 $r_i^2>\mu^2 v_r$ 时，认为相应的观测值异常。当第 1 次将所有异常潮位判断出来后，用回报值替代异常值，然后进行第 2 次判别，2 次判别出来的异常值作为最后的判别结果。一般情况下通常取 P_0=0.9，后续采用方国洪准则进行异常值判别时取 P_0=0.9。

2　异常点判别结果

本文选取了江苏浒浦闸和万福闸下 2 个代表潮位站点的资料进行分析。按照上述 3 种准则发现了下面几类异常值：（1）某个区间段潮位突然被抬升，见图 1；（2）实测高潮位在一个时间段内保持不变，见图 2；（3）高潮位异常大，见图 3。异常值点数识别统计结果汇总在表 2 中。

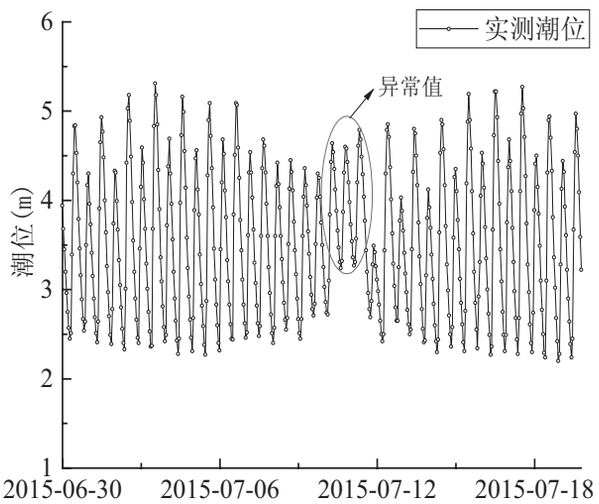

图 1　浒浦闸部分实测潮位过程图

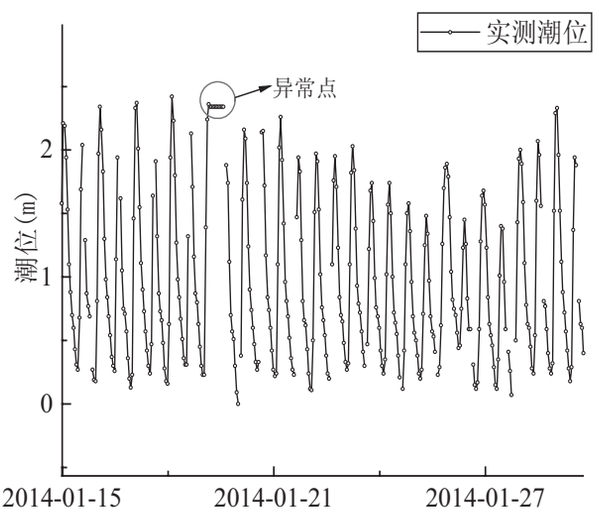

图 2　万福闸下部分潮位过程

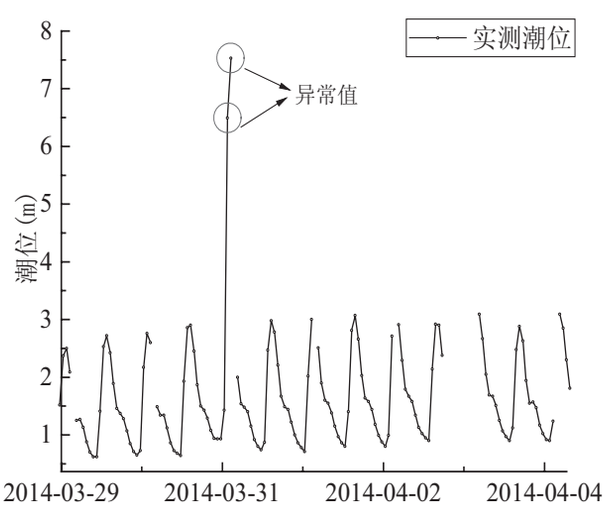

图 3　万福闸下部分实测潮位过程图

表 2 不同准则识别潮位异常点数表

	浒浦闸 2014	浒浦闸 2015	万福闸下 2014
拉伊达准则	346	383	362
肖维勒准则	299	249	315
方国洪准则	20	26	10

从表 2 可以看出，拉伊达准则识别异常点数多于肖维勒准则、方国洪准则。对比 3 种准则识别的异常点位置，拉伊达准则找出的异常值包含肖维勒准、方国洪准则找出来的所有异常值，肖维勒准则找出的异常值点包含方国洪准则找出来的点。对于图 1、图 2 中这种异常值密集出现的异常点，拉伊达准则好于肖维勒准则，而肖维勒准则又优于方国洪准则，后 2 种准则对异常点连续段边缘点存在漏判现象。

3 调和分析结果

为了定量比较 3 种准则的判别效果，本文采用了 T_TIDE[10] 软件对判别后的潮位数据序列进行调和分析，通过潮位的回报精度来说明 3 种准则的相对优劣。为了确保用于分析潮位的数据可靠，采用自报值循环逼进法 [11] 对判别出来的异常数据进行修正或插补，然后对修正或插补后的数据序列进行调和分析。

表 3 统计了浒浦闸 2014 年、2015 年和万福闸下 2014 年潮汐回报均方差的比较结果。从表中可以看出，虽然基于 3 种准则都可以在一定程度上提高浒浦闸和万福闸下的潮位回报精度，但相比较而言，拉伊达准则表现最好，肖维勒准则次之。

表 3 代表潮站潮位回报均方差统计表

项目	浒浦闸（cm）		万福闸下（cm）
	2014 年	2015 年	2014 年
实测数据	16.4	20.1	16.9
拉伊达准则	12.8	14.7	12.8
肖维勒准则	13.6	16.5	13.6
方国洪准则	15.9	19.6	15.9

4 结果分析

3 种准则的识别效果差异，可以根据 3 种准则的判别原理进行解释。由于含有缺测值，调和分析采用的潮位值小于 8759 个。取 N=8759 的话，肖维勒准则数 Z_c=4.02。拉伊达准则是误差超出 3 倍标准差则判断数据点异常，而肖维勒准则是误差超出 4.02 倍标准差才算异常值。拉伊达准则是一个固定的判别标准，而肖维勒准则数 Z_c 随 N 变化，N 越大则 Z_c 越大，N 不同时它的判别标准会改变。Z_c=3 时，N 为 190，当样本数 N 小于 190 时，肖维勒准则判别异常值比拉伊达准则更为严格，会判断出更多的异常值；当样本数 N 超过 190 时，拉伊达准则比肖维勒准则会判别出更多的异常值。本次样本数 N 远远超过 190，因此拉伊达准则判别出来的异常值数目比肖维勒准则多。

拉伊达准则和肖维勒准则分别认为误差超过 3 倍样本标准差和 4.02 倍样本标准差时数据异常，本次样本中方国洪准则的 μ 值约为 4.37，由于样本 N 较大，可以近似认为方国洪准则下的标准差与前两种准则的标准差 S 相等，相当于方国洪准则认为超出 4.37 倍标准差才算异常值，所以拉伊达准则和肖维勒准则判别出来的异常值数目比方国洪准则多。此外，拉伊达准则、肖维勒准则每剔除一个异常值后重新生成样本，标准差会随着异常值的剔除逐渐变小，判别标准会逐渐变严格，虽然方国洪准则判断了 2 次，但是每次都是一次性判别所有异常值，当异常值较多时，方国洪准则的标准差会较大，判别界限值的差异也导致了拉伊达准则和肖维勒准则的判别标准比方国洪准则更严格。本次样本中很多异常值数目较多且与真实值差别较大，方国洪准则一次性判别所有数据的方法，导致有较多异常值被漏判。

5 结论

利用 3 种准则对潮位异常值进行判别，以探讨潮位异常值对潮位调和分析的影响和 3 种准则判别潮位异常值的准确性，得到以下几个主要结论：

（1）采用拉伊达准则可以对异常值密集和连续出现的情况进行较好的判别，而方国洪准则易发生异常值漏判现象。

（2）通过异常值判别后，潮位调和分析的回报精度有明显提高，相对而言拉伊达准则表现最好，肖维勒准则次之。

潮位异常值的自动判别对于资料的高质量整编和潮位的高精度预报具有重要的实际应用价值。后续将进一步对拉伊达准则中的判别参数进行优化，最大程度地减小漏判或误判数据的比例，有效提高自动测站实测潮位的资料质量。

参考文献：

[1] 张凤烨，魏泽勋，王新怡，等．潮汐调和分析方法的探讨[J]．海洋科学．2011，35(06)：68–75．

[2] 陈宗镛．潮汐学[M]．北京：科学出版社，1980:127．

[3] 黄谟涛，翟国君，王瑞，等．海洋测量异常数据的检测(英文)[J]．测绘学报，1999(03)：269–276．

[4] 方国洪，郑文振，陈宗镛，等．潮汐和潮流的分析和预报[M]．北京：海洋出版社，1986: 90–93．

[5] 许军，刘雁春，暴景阳，等．基于余水位的水位粗差探测与数据修复[Z]．成都：2009．

[6] 董玉磊，曲萌．一种基于回归分析的海上定点验潮站异常数据处理方法[Z]．北京：2015．

[7] 费业泰．误差理论与数据处理[M]．北京：机械工业出版社，2010:4．

[8] 熊艳艳，吴先球．粗大误差四种判别准则的比较和应用[J]．大学物理实验．2010(01)：66–68．

[9] 王红川，左其华．海洋资料中异常值的分析和判别[J]．水利水运科学研究，1998，12(4)：364–365．

[10] Pawlowicz R，Beardsley B，Lentz S．Classical tidal harmonic analysis including error estimates in MATLAB using T_TIDE[M]．Pergamon Press，Inc. 2002．

[11] 吴俊彦，张亚彪．潮位观测资料缺失的补足应用研究[Z]．广西：2008．