潮位数据奇异值判别与处理

余汉学
（中国海洋大学海洋与大气学院，山东 青岛 266100）

摘　要：潮位数据是海洋学、海洋气象学、海洋工程学等领域研究的重要数据，但是潮位数据中可能会存在一些异常值，这些异常值会影响潮位数据的质量和可靠性，因此判别和处理潮位数据中的异常值具有重要意义。本文采取了统计特性检验与连续性检验来判定异常值，并采用了三次样条插值来处理奇异值数据。

关键词：莱因达准则；格林布斯准则；梯度检测；尖峰检测

0 引 言

海洋是地球上最重要的自然资源之一，而潮汐作为海洋运动的重要组成部分，对海洋生态、气候变化和海洋工程等领域都具有重要的影响。因此，对潮汐的研究和监测显得尤为重要。然而，潮汐数据中常常存在一些异常值或奇异值，这些值可能对潮汐数据的分析和应用造成负面影响。因此，对潮汐数据中的奇异值进行判别和处理具有重要的意义。目前，潮位数据奇异值判别与处理方面的研究取得了一些成果，如：潮位数据奇异值判别与处理方面的研究取得了一些成果：基于物理模型的奇异值判别与处理；基于时间序列分析的奇异值判别与处理；基于机器学习和深度学习的奇异值判别与处理；基于多源数据融合的奇异值判别与处理。本文将介绍潮汐数据奇异值的判别和处理方法中的统计特性检验与连续性检验，旨在提高潮汐数据的质量和可靠性，为相关领域的研究和应用提供更加可靠的数据支撑。

1 理论分析

1.1 统计特性检验

统计特性检验是一种常用的判别潮汐数据中奇异值的方法，其基本思想是将潮汐数据的各种统计特征（例如平均值、标准差、偏度、峰度等）与一些预先设定的阈值进行比较，从而判断该数据是否存在异常值。这些统计特征和阈值的设定通常是基于某些先验知识和经验来确定的。

1.1.1 莱因达准则

莱因达准则（Reinhardt criteria）是一种常用于判别潮汐数据中奇异值的方法，它是基于潮汐数据的周期性和连续性来进行判别的。

莱因达准则规定与观测值x_i；相应的剩余误差v_i满足公式(1-1)，否则认为该剩余误差异常，对应的观测值也异常。

$$\mathrm{v_i} \le 3\delta \tag{1-1}$$

式中，观测值的剩余误差$\mathrm{v_i}$由公式(1-2)计算得到;δ是观测值的标准差，由公式(1-3)计算得到。

$$v_i = |x_i - \bar{x}| \tag{1-2}$$

$$\delta = \frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1-3}$$

式中，n是观测值的总数；$\bar{x}$是观测值的平均值,由公式(1-4)计算得到:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1-4)$$

1.1.2 格林布斯准则

格林布斯准则规定，要素观测值需满足公式(1-5)，否则数据异常。

$$|x_i - \bar{x}| \le G(\alpha, n)\delta \qquad (1-5)$$

式中，x_i是观测值;$\bar{x}$是观测值的平均值,计算公式见公式(1-4)；δ是观测值的标准差。计算公式见公式(1-3)；n 是数据序列中样本个数；$G(\alpha, n)$是格林布斯界值，计算公式见公式(1-6)。

$$G(\alpha, n)\delta = \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2(a/n, n-2)}{n-1+t^2(a/n, n-2)}} \qquad (1-6)$$

式中α是显著性水平(α最大为 0.1)，t为自由度为$n-2$、显著性水平为a/n的单边界检验 t 分布的临界值。

1. 2 连续性检验

统计特性检验是一种常用的判别潮汐数据中奇异值的方法，其基本思想是将潮汐数据的各种统计特征（例如平均值、标准差、偏度、峰度等）与一些预先设定的阈值进行比较，从而判断该数据是否存在异常值。这些统计特征和阈值的设定通常是基于某些先验知识和经验来确定的。

1.2.1 梯度检验

海洋观测数据在一定的时间或空间范围内具有连续性，时间接近或者位置邻近的观测要素变化值应该在一定范围内，否则认为数据异常。

1.2.2 尖峰检验

海洋观测数据在一定的时间或空间范围内变化是有限的，若出现较大的突变，这一突变值与周围观测值明显不同，则判定其为异常值。

1.2.3 恒定检验

在观测仪器灵敏度和精度足够的情况下,海洋观测要素受流体动力因素的影响,在一定时间或空间范围内不会恒定不变，若恒定不变则数据可能异常

2 实验系统及测量结果

2.1 数据介绍

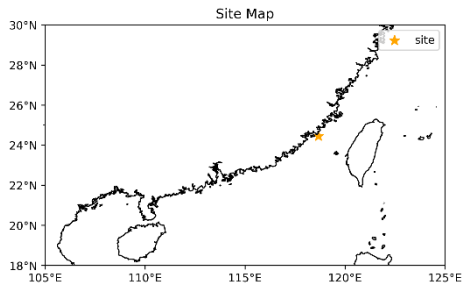


Figure 1 Site Map

本数据选取位置在 118.67° E， 24.45° N（Figure 1）的站点（sta. XM）与 1997-08-01~1997-08-31 (GMT) 记录的潮位数据。数据源是文本文件，可以使用 Python 进行文本读入、处理与数据数组化。

Python 数据分析环境为 Python 3.9.15。使用 Jupyter 3.5.3 作为 IDE，数据源文件为 WORK1.ipynb。

读取的原始数据绘制如下：

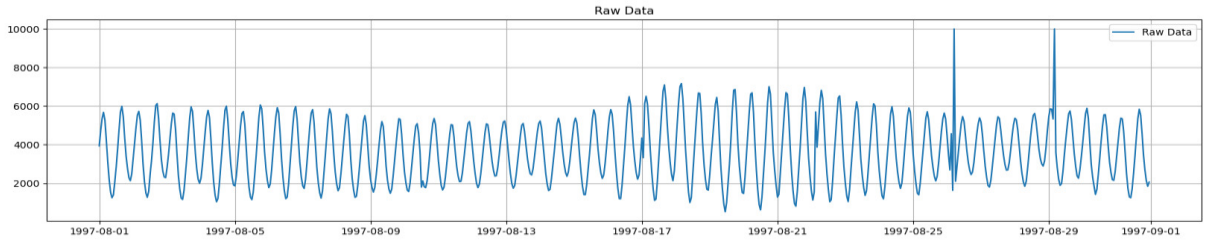


Figure 2 Raw Data

2.2 分析步骤

2.2.1 莱因达准则

使用莱因达准则判断异常值：

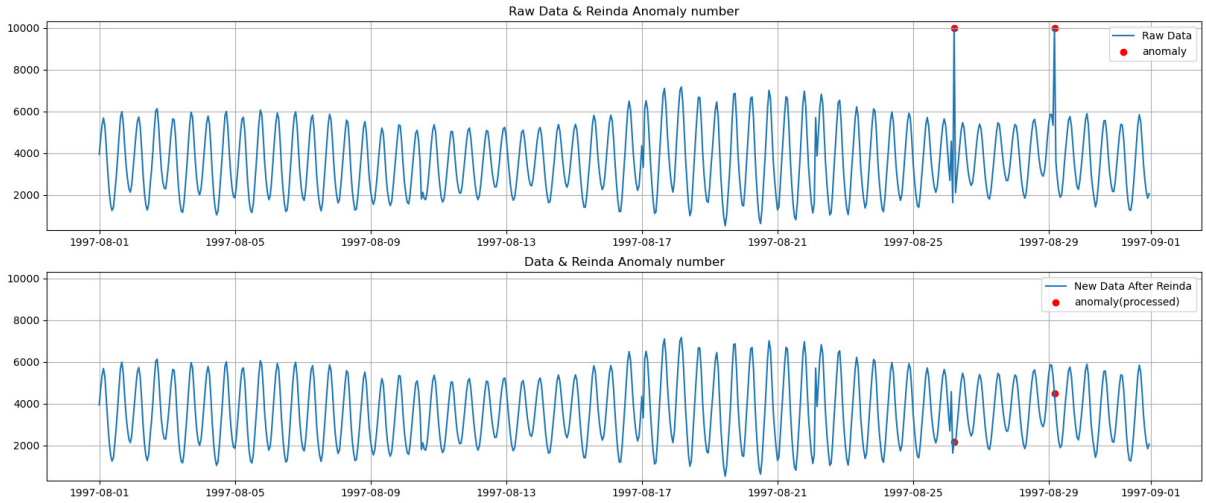


Figure 3 **莱因达准则去除异常值**

2.2.2 格林布斯准则

使用格林布斯准则判断数据（已经通过莱因达准则去除了异常），无异常值输出。如果带入原始数据则会输出与莱因达准则相同的异常值，故不在展示。

2.2.2 梯度检测（变率检测）& 尖峰检测

根据 GB/T 14914，梯度检测的梯度检验参数根据要素类型、观测时间间隔、空间距离、观测时间和区域等因素确定。在所给条件中无法获得，因此采用将梯度数据再带入统计特性检验的方式。（GB/T 14914.6-2021, 海洋观测规范）。

将变率绘制出来并带入莱因达准则进行判别可得：

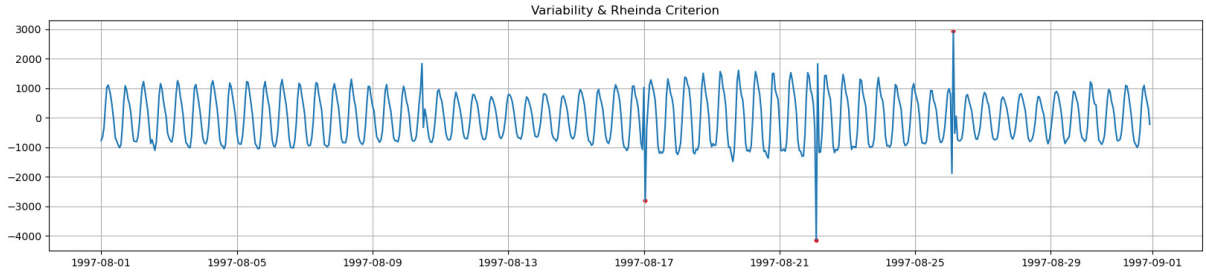


Figure 4 **变率 & 莱因达准则**

将异常值处理后再重新进行上述检测，直至无异常值。再对变率求绝对值重新带入莱因达准则进行异常值处理。

最终无异常值的变率如下：

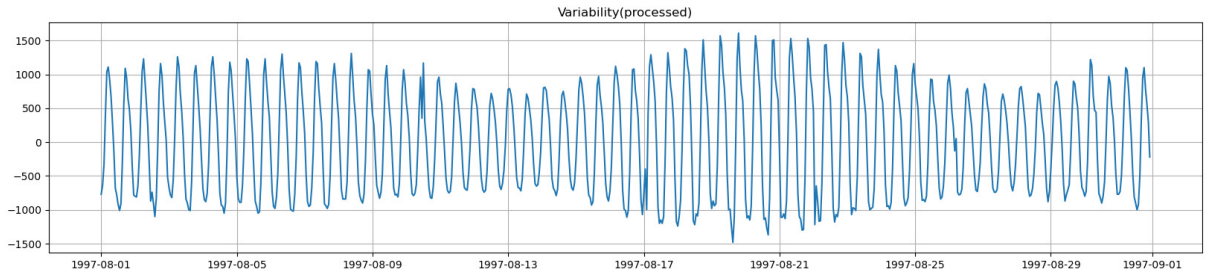
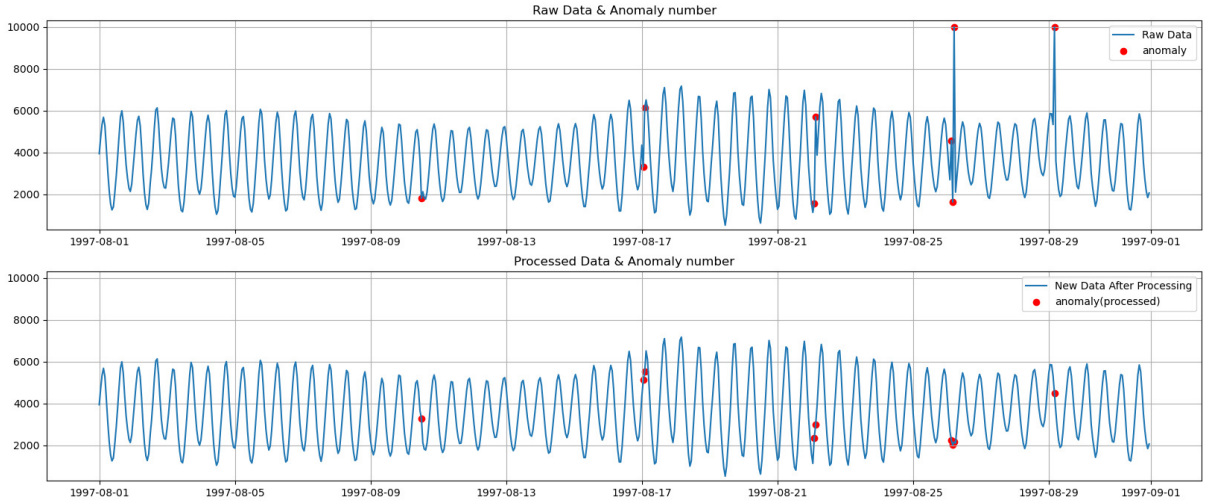


Figure 5 **变率（处理后）**

2.3 结果

把原始数据处理完可得到如下结果：



需要说明的是因为变率检测再本数据中的体现是$x_i - x_{i\pm1}$，因此一个异常值可能会导致上图有两个异常点，故在下表中进行了删除。

异常值序号	异常时间
1	1997-08-10 12:00:00
2	1997-08-17 01:00:00
3	1997-08-22 03:00:00
4	1997-08-26 04:00:00
5	1997-08-26 05:00:00
6	1997-08-29 04:00:00

其中与原始数据相比，均值从 3656.22 变为 3636.69，标准差从 1563.81 变为 1525.29。

5 结　论

经过对潮位数据的奇异值判别和处理，本文发现：

- 奇异值判别和处理是潮位数据预处理中的重要环节，能够有效地去除异常值和噪声干扰，提高数据的质量和可靠性。
- 莱因达准则和格林布斯准则是常用的统计特性检验奇异值判别方法，两种方法均可有效地鉴别潮位数据中的异常值。
- 连续性检验是判别数据的连续性的有效方法，能够去除时间序列中的断裂和突变，使得数据更加平滑和连续。
- 综合运用多种方法进行奇异值判别和处理，可以得到更加准确和可靠的潮位数据，为后续的数据分析和应用提供了基础和保障。

因此，对于潮位数据的处理，我们应该采用多种方法相结合的方式，对数据进行全面和细致的检查和处理，以确保数据的质量和可靠性。

参考文献

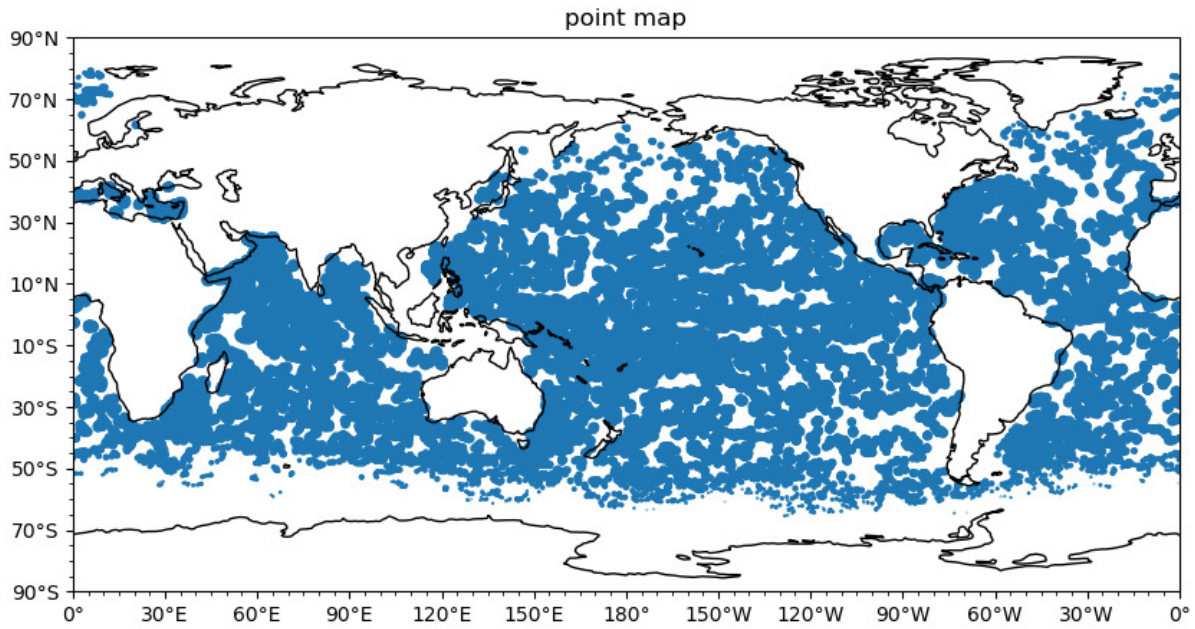
[1]刘永玲.杜凌.李静凯.翟方国 海洋要素计算上机实验指导书[M] 青岛：中国海洋大学出版社，2021

[2]GB/T 14914.6-2021, 海洋观测规范 第6部分：数据处理与质量控制[S].

拓展作业在下一面↓

拓展作业

原始格点数据空间分布如下：



使用公式：

$$Z = \frac{\sum_{i=1}^{n} \frac{1}{(D_i)^p} Z_i}{\sum_{i=1}^{n} \frac{1}{(D_i)^p}}$$

进行空间格点插值并于ERSST V5进行比较如下：