张宇, 周燕, 陶邦一, 等. 基于时序相关性分析方法的浮标异常数据识别[J]. 海洋学报, 2020, 42(11): 131–141, doi:10.3969/j.issn.0253−4193.2020.11.013

Zhang Yu, Zhou Yan, Tao Bangyi, et al. Identification of abnormal buoy data based on time series correlation analysis method[J]. Haiyang Xuebao, 2020, 42(11): 131–141, doi:10.3969/j.issn.0253−4193.2020.11.013

基于时序相关性分析方法的浮标异常数据识别

张宇[1], 周燕[2], 陶邦一[1,3*], 顾吉星[4], 赵传高[4], 郝增周[1],
张艺蔚[1,5], 黄海清[1], 毛志华[1]

(1. 自然资源部第二海洋研究所 卫星海洋环境动力学国家重点实验室, 浙江 杭州 310012; 2. 浙江省海洋科学院, 浙江 杭州 310007; 3. 南方海洋科学与工程广东实验室, 广东 广州 511458; 4. 国家海洋局烟台海洋环境监测中心站, 山东 烟台 264006; 5. 中国科学院上海技术物理研究所, 上海 200083)

摘要: 海洋生态浮标异常数据的实时早期监测识别是保证观测数据质量的关键。本研究通过对浙江沿海浮标多年数据的分析, 发现了与传统跳变异常数据不同的渐变异常数据类型。该异常类型呈现出在时序变化过程中连续平稳, 但随时间逐渐偏移, 最后整体偏离正常的分布特征, 并且单一参数的分析方法无法对此异常进行有效识别。因此本研究利用海洋环境参数中酸碱度 (pH)、溶解氧 (DO) 和叶绿素 (Chla) 三者的多参数相关性规律, 提出了在一定时序上两两参数间相关性是稳定甚至是一致的假设, 将 8 天时间窗口的两两相关系数 (R_{8d}) 和前后两天 R_{8d} 之差的绝对值 (ΔR) 作为相关性和稳定性核心指标, 建立了基于相关性的渐变异常数据自动识别方法。为浮标传感器渐变异常的早期识别提供了一个新的思路, 有助于提升海洋生态浮标异常数据的自动化监测能力。

关键词: 生态浮标; 环境监测; 真实性检验; 相关性分析

中图分类号: P715.2 **文献标志码:** A **文章编号:** 0253−4193(2020)11−0131−11

1 引言

海上浮标是获得长时序、高精度海洋环境参数最主要的手段, 确保浮标数据的质量可靠性是开展数据应用所面临的首要问题, 因此开展浮标异常数据的检测识别是其中一项重要工作[1]。异常数据一般指超过正常合理数值范围的以及偏离由海洋环境引起的变化规律的数据[2]。将台州大陈(TZ01)和温州南麂岛(NJ01)的两个浮标叶绿素数据与 Aqua/MODIS、VIIRS 和 GOCI 海洋水色卫星反演的叶绿素产品进行比对(图 1), 研究发现浮标反演的叶绿素产品存在两种异常类型: (1)浮标数据在时序分布上连续且与卫星数据有较好的一致性(图 1a), 但由于海洋随机过程产生如图 1a 中红色方框标记的跳变数据, 属于传统意义上的跳变异常数据类型; (2)红色条带标记的一段浮标叶绿素测量异常数据呈现出: 在时序变化过程中连续平稳, 但随时间逐渐推移, 最后整体偏离正常数据的分布特征(图 1b), 这种异常数据属于一种新的渐变异常数据类型。基于海洋环境要素时序数据分布平稳的假设[3−6], 传统的异常数据统计识别方法仅对跳变异常数据类型的数据检测有效, 而对渐变的质量异常数据类型无法识别[5−12]。主要原因在于异常发生的初始阶段, 其变化特征与由海洋环境变化引起的变化趋势很难在没有先验知识的条件下进行区分, 只

收稿日期: 2019−09−28; **修订日期:** 2020−06−22。

基金项目: 国家重点研发计划(2018YFC0213103, 2016YFC1400901); 第二海洋研究所所基本科研业务费专项(QNYC201602); 民用航天技术预先研究项目(D040401-06); 国家自然科学基金(41876033)。

作者简介: 张宇 (1993—), 女, 辽宁省锦州市人, 从事海洋遥感方向研究。E-mail: 15174094738@163.com

*** 通信作者:** 陶邦一 (1983—), 男, 副研究员, 从事海洋遥感方向研究。E-mail: taobangyi@sio.org.cn

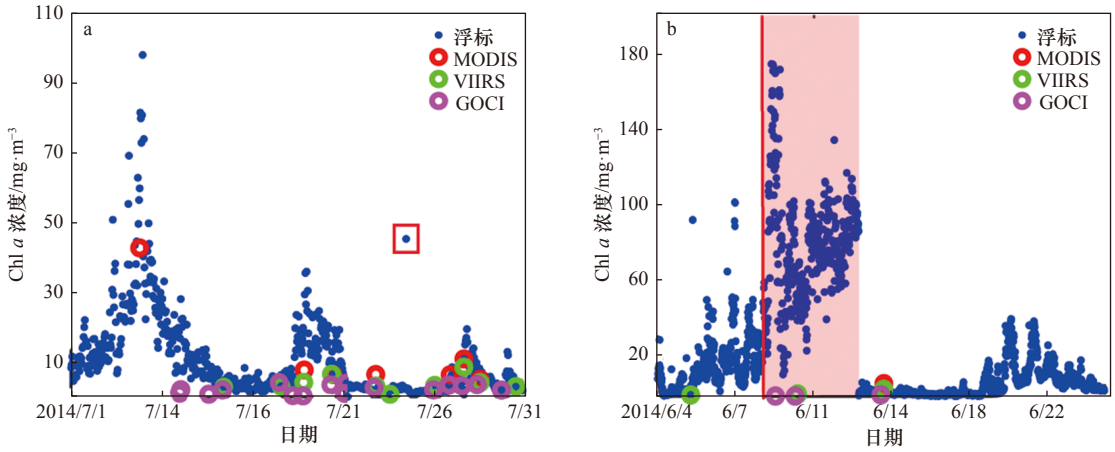

图 1 浮标数据与卫星数据叶绿素 *a* 浓度对比

Fig. 1 Comparison of chlorophyll *a* concentration between buoy data and satellite data

a. 2014 年 7 月 TZ01(台州大陈)浮标叶绿素浓度与卫星数据的对比结果; b. 2014 年 6 月 NJ01(温州南麂)浮标叶绿素浓度与卫星数据的对比结果

a. Comparison of chlorophyll *a* concentration between TZ01 buoy data and satellite data in July 2014; b. comparison of chlorophyll *a* concentration between NJ01 buoy data and satellite data in June 2014

有利用后续正常数据分布特征等后验证知识进行识别。这类渐变异常数据可能与传感器探头受污、供电异常等因素有关。渐变的异常数据类型在长时间观测的浮标数据中时有发生,因此如何在这一类型异常数据发生的初始阶段进行有效识别,对于浮标异常的实时监测、及时维护、保证数据的可靠和连续性具有现实意义。

国内外都已开展海上浮标观测应用工作多年,但实际上实现各类型异常数据的自动检测识别仍有较大难度[3−4],国内对海洋数据的检测主要依赖专家经验、历史资料以及常识形成的海洋环境监测数据检验标准库[5]。目前已有的异常检测方法主要有极值检验、一致性判断、递增性判断、格拉布斯检验、狄克逊检验、拉依达检验、过度梯度检测、尖峰检测和无梯度检测等[1,6−19],这些异常检测方法主要是针对单一参数在某一时间尺度的平稳随机过程中进行统计学的分析处理,在传统跳变异常数据类型识别中取得了较好的检测效果,但对于渐变异常数据类型的自动检测识别研究较少。

随着技术发展,目前浮标平台上搭载的传感器数目和测量的参数越来越多,而在这些测量参数中存在某些参数相互关联的特征。多元时间序列数据分析方法(如建立矢量自回归(VAR)、多元谱分析,广义自回归条件异方差模型(GARCH)等)被广泛地应用到质量异常数据的检测和识别上[20−23]。Tsay[24] 将 4 种类型单参数时间序列异常数据识别方法拓展到了多元序列数据。此外,异常质量数据检测中也应用到了矢量自相关性系数、差分整合移动平均自回归模型、遗传算法等方法[25−29]。然而,上述方法主要适用于跳变异常数据的识别,而且对数据平稳性要求较高、计算流程复杂,并未在本文发现的渐变异常数据类型上得到应用。其中,在海洋多元长时序数据异常识别方面,Schuckmann 等[12] 提出相关性分析方法成功地识别了叶绿素浓度高而溶解氧浓度低的错误数据类型。在浮标数据质量控制中的应用,仅给出了白天叶绿素浓度高而溶解氧浓度低的错误数据类型的识别案例。窦文洁等[18] 则根据海洋碳酸盐系统中海水 CO_2 分压本身于水体温度盐度存在定量相关性关系的特点,在假设观测参数变化在非常小的时间尺度内为一平稳过程的基础上,提出了基于多参数观测序列差分统计特征的异常点识别方法。虽然该方法由于仅基于参数平稳性假设而无法进一步有效识别渐变异常数据类型,但相比于单一参数分析方法,利用多参数强关联性对异常数据进行检测,会对数据的处理有更加全面、深入的把控。

目前,我国常规生态浮标通常会同时观测酸碱度(pH)、溶解氧(DO)浓度以及叶绿素 *a*(Chl *a*)浓度等数据。大量的研究表明,它们之间虽然具有较紧密的相关性[30−31],特别是在海水藻类生长暴发期间[31],但它们并不存在稳定的相关关系,如谢群等[32] 在流沙湾得出溶解氧浓度与叶绿素 *a* 浓度成正比例关系,尤其是冬季,海水中的溶解氧浓度与叶绿素 *a* 浓度具有极显著正相关,春季次之,夏秋两季两者之间不存在相关性的结论。可见在不同海域、不同季节及不同海洋过程中的参数之间相关性特征具有明显的差异性,并不能类似于 Schuckmann 等[12] 采用事先设定的相关性特征

进行多年长时序数据的处理。Hollinger 和 Richardson[33] 在海洋数据不确定分析时提出了“单塔日变化法”, 其基本假设是相邻日期间在相同或相似的环境条件下数据变化过程相似。因此本研究认为, 浮标观测到的正常多参数数据不仅单一参数在时序变化上平稳连续, 并且两两参数间的相关性在一定时序上稳定甚至一致。

本文基于上述假设, 希望通过对浙江沿岸海域浮标多年的 pH、DO 浓度、Chl *a* 浓度数据相关性进行分析, 了解当某一参数出现渐变异常时, 与其他参数的相关性特征的变化规律, 基于多参数相关性变化提出一种简单、适用的渐变异常数据检测识别方法, 并且探讨该方法在该海域的适用性。

2　数据来源

浮标数据采用浙江省沿岸的温州南麂大沙岙(NJ01)、台州大陈(TZ01)、宁波南韭山(NB01)、宁波渔山(NB03)、舟山嵊泗绿华(ZS03)和舟山普陀东极(ZS04)6 处浮标数据, 浮标分布如图 2 所示。

观测时间在 2012 年 8 月至 2017 年 5 月之间, 数据每 15 min 或 1 h 传输 1 次, 以同一浮标同一时间获取的数据为一组, 共计 183 967 组数据, 其中状态显示异常、故障或维护的数据有 8 662 组, 仪器运行正常状态的原始浮标数据(DO 浓度、Chl *a* 浓度和 pH)有 175 305 组, 占总数据量的 95%, 数据情况如表 1 所示。对仪器运行正常状态的 175 305 组数据进行分析处理, 对其他状态的数据不予处理。

3　基于相关性的渐变异常数据识别方法

本文根据 pH 与 DO 浓度具有正相关关系, Chl *a* 浓度与 pH 和 DO 浓度的关系因藻类生长、季节变换等因素呈现显著正相关或不相关关系等特点, 利用最基本相关性统计学方法来计算 pH 与 DO 浓度、pH

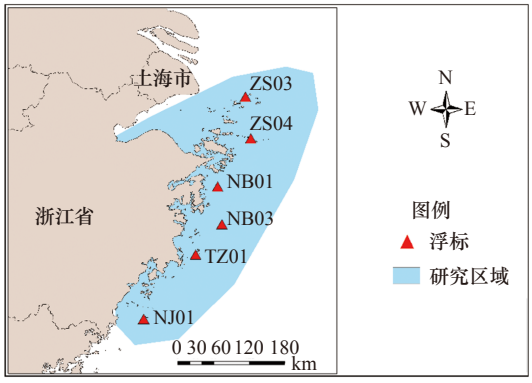

图 2　研究区域浮标分布

Fig. 2　Distribution of buoys in the study area

表 1　浮标数据统计

Table 1　Statistical buoys data

浮标	起止时间	原始数据/组	异常/维护等状态数据/组	正常状态数据/组
NJ01	2013年7月至2017年5月	57 070	5 006	52 064
TZ01	2012年8月至2017年5月	38 359	846	37 513
NB03	2014年7月至2017年5月	24 086	469	23 617
NB01	2013年7月至2017年5月	34 129	994	33 135
ZS04	2015年8月至2017年5月	15 573	556	15 017
ZS03	2015年8月至2017年5月	14 750	791	13 959
合计		183 967	8 662	175 305

与 Chl *a* 浓度、DO 与 Chl *a* 浓度两两相关性系数。在对生态浮标数据进行多参数协同分析后发现, 异常判定方法的关键是相关性计算时所要选取的时间窗口以及基于相关平稳性异常的判定方法。

3.1　时间窗口的确定

由于浙江沿岸海域生化参数日变化动态范围较大, 如以太短或者太长的时间段内的两两相关性来建立成段异常数据方法, 则存在较大的随机与不确定性, 不利于对长时序浮标数据的稳定性研究与渐变异常数据的早期识别。因此, 选择合适的时间窗口, 对于建立相关性分析处理异常数据模型至关重要。本文将浙江沿岸 6 处仪器运行正常状态的 175 305 组浮标数据经过不可能出现的范围和 5S 方法剔除异常数据等预处理后, 剩余 156 305 组浮标数据参与多参数协同分析。其中, 选出 13 620 余组各参数质量较好的浮标数据对其进行两两相关性分析。部分正确的浮标数据序列如图 3 所示。

将图 3 的 pH、DO 浓度和 Chl *a* 浓度数据的两两相关系数(R_{nd})计算的时间窗口逐天扩大, 从 1 d 扩大到 16 d, 结果如图 4a 所示。由图可见, 随着时间窗口的扩大, 相关性逐步提升, 并且当扩大到 8 d 时 R_{nd} 都处于稳定状态, 即当时间窗口大于 8 d 后相关性并未明显增强。同时以 8 d 为时间窗口对图 3 中的多组浮标长时序数据进行 8 d 两两相关系数(R_{8d})的计算(图 4b), 可以看出正常原始浮标数据的 R_{8d} 在一定时期内同样非常稳定, 因此时间窗口选定为 8 d, 同时将 R_{8d} 作为检测渐变异常数据的核心参数。

3.2　基于相关性的异常判定方法

首先, 利用多参数之间相关性程度来进行异常数据判定。如图 4b 所示, 正常数据的 R_{8d} 变化平稳, 状态稳定。为定量化正常 R_{8d} 变化的范围, 利用 6 处浮标多年中的正常数据集, 统计了浙江海域 R_{8d} 的数值

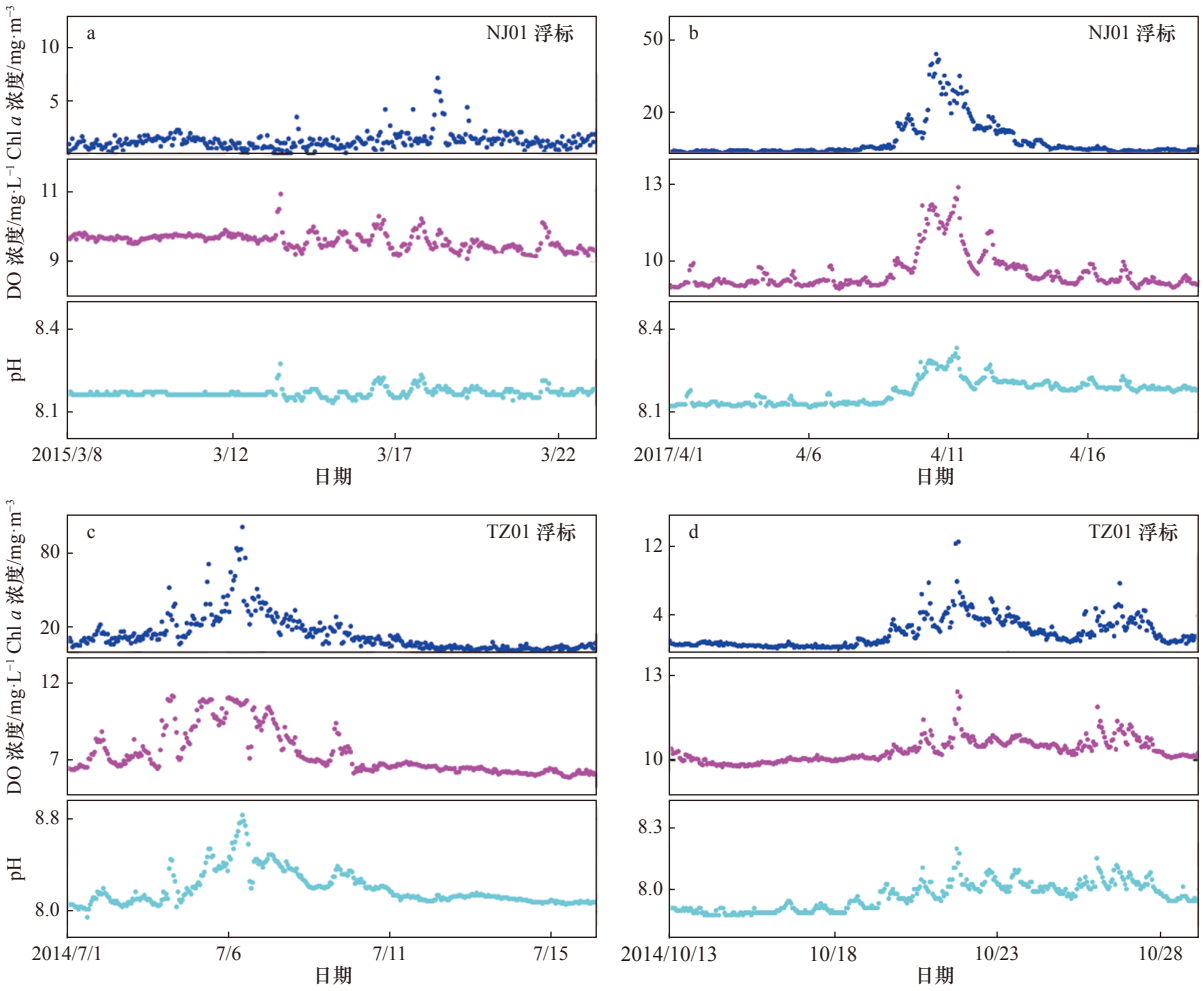

图 3 部分正确的浮标数据序列

Fig. 3 Partially correct buoy data sequence

a. 2015 年 3 月 NJ01 浮标原始数据; b. 2017 年 4 月 NJ01 浮标原始数据; c. 2014 年 7 月 TZ01 浮标原始数据; d. 2014 年 10 月 TZ01 浮标原始数据

a. NJ01 buoy raw data in March 2015; b. NJ01 buoy raw data in April 2017; c. TZ01 buoy raw data in July 2014; d. TZ01 buoy raw data in October 2014

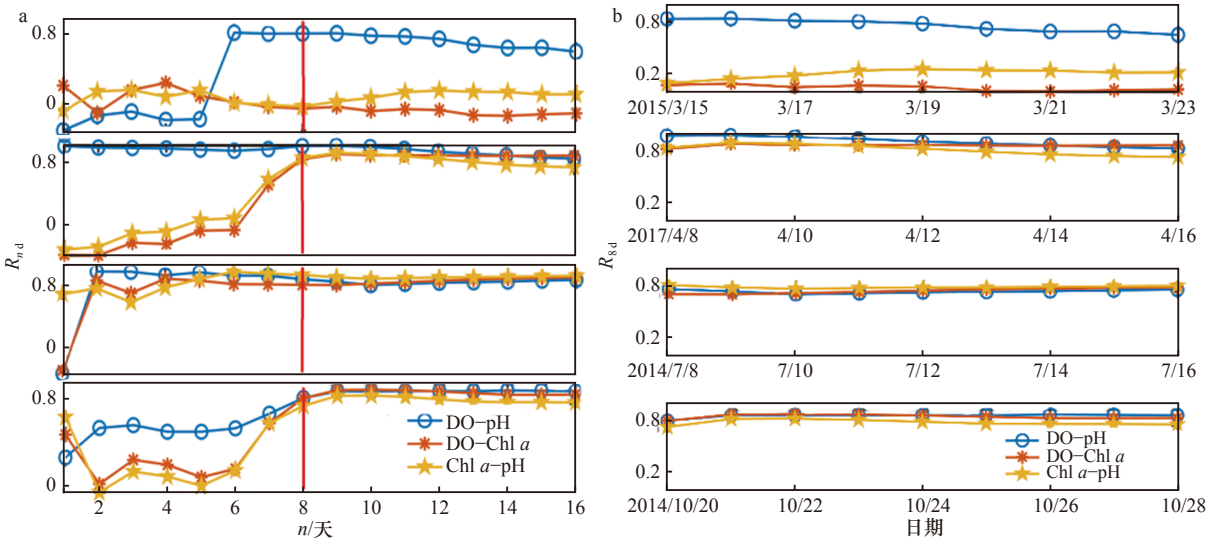

图 4 不同时间窗口的相关系数(a)和基于 8 d 时间窗口的相关系数(b)

Fig. 4 Correlation coefficient for different time windows (a), and correlation coefficient for 8 d time window (b)

分布情况, 如图 5 和表 2 所示。统计结果表明: (1) pH 与 DO 浓度之间正相关性最强, 几乎所有正常数据的 R_{8d}(pH−DO) 都大于 0; (2) DO 浓度与 Chl *a* 浓度之间相关性次之, 其 R_{8d}(DO−Chl *a*) 大于−0.3, 其中大于 0 的数据近 95%; (3) pH 与 Chl *a* 浓度之间相关性变化较大, 但有 92% 数据的 R_{8d}(pH−Chl *a*) 大于 0, 另 6.8% 的 R_{8d}(pH−Chl *a*) 数据在−0.3～0 之间, 并且仅有 1.2% 的 R_{8d}(pH−Chl *a*) 小于−0.3。因此, (1) 正常数据 pH 与 DO 浓度、pH 与 Chl *a* 浓度、DO 浓度与 Chl *a* 浓度明显存在较高的正相关关系, 其判定原则较为简单, 即两项以上的相关性 R_{8d} 都大于 0.5 可以作为数据正常有效标志; (2) 因 pH 与 DO 浓度之间不存在负相关关系, 明显错误数据的判定原则为当 R_{8d}(pH−DO)<0 时为异常值; 另外当 R_{8d}(pH−DO)>0 时, 如果 Chl *a* 浓度与 DO 浓度、pH 之间不存在较强的负相关关系, 即 R_{8d}(pH−Chl *a*)<−0.3 时, 或 R_{8d}(DO−Chl *a*)<−0.3, 可识别为可疑数据。实际上, 单一的 R_{8d} 只能用于识别相对明确的正确及异常数据, 而对于其中某项相关性 R_{8d} 小于 0.5 的浮标数据需要进一步采用其他相关性时序稳定性指标来进行识别。

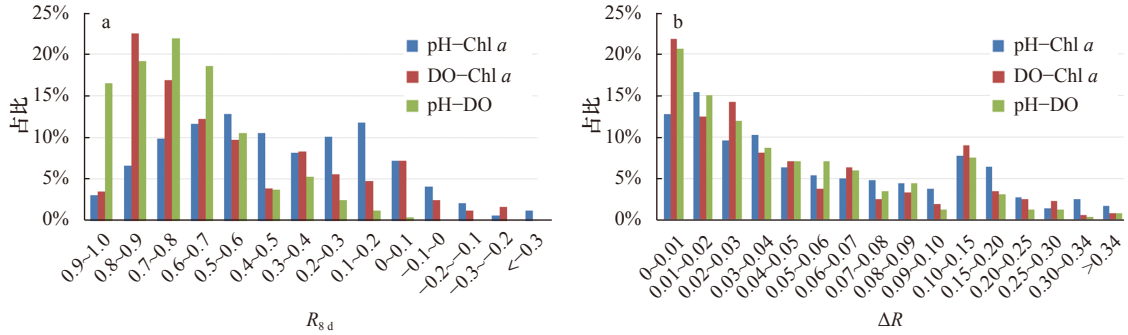

图 5　R_{8d}(a)和 Δ*R*(b)的分布情况

Fig. 5　Distribution of R_{8d} (a) and Δ*R* (b)

表 2　R_{8d} 的分布情况

Table 2　Distribution of R_{8d}

R_{8d}值	R_{8d}(pH−DO)	R_{8d}(DO−Chl *a*)	R_{8d}(pH−Chl *a*)
0.5～1.0	86.80%	64.80%	44.20%
0～0.5	13.20%	29.80%	47.80%
−0.3～0	0%	5.40%	6.80%
<−0.3	0%	0%	1.20%

表 3　Δ*R* 的分布情况

Table 3　Distribution of Δ*R*

Δ*R*	Δ*R*(pH−DO)	Δ*R*(DO−Chl *a*)	Δ*R*(pH−Chl *a*)
0～0.03	47.50%	48.40%	37.70%
0.03～0.06	22.90%	19.00%	21.90%
0.06～0.10	15.20%	14.00%	18.00%
0.10～0.34	13.60%	17.90%	20.70%
>0.34	0.80%	0.70%	1.70%

如前文所述, 本文认为正常浮标多参数数据之间的两两相关性在一定时序上是稳定甚至是一致的, 因此需要建立一个指标来表征 R_{8d} 本身的稳定性。本文利用前后两天 R_{8d} 之差的绝对值(Δ*R*)作为判断相关性时序分布稳定与否的指标。通过统计正常浮标数据的 Δ*R* 分布情况(图 5b, 表 3)可见, 其中约有 60%～70% 数据的 Δ*R*<0.06, 且 Δ*R*<0.1 数据都达到了 77.6% 以上。从数据分布上可以看出, 正常浮标数据的多参数之间相关性变化是稳定或缓变的过程, 符合前文的稳定性假设。由于计算求得 Δ*R* 的标准差为 0.068, 同时从表 3 的统计结果也可看出, 各相关性中 Δ*R*>0.34 的数据仅占 1.0% 左右, 因此选取 5 倍标准差[19] 即 0.34 为判断稳定性阈值, 即当 0<R_{8d}(pH−DO)<0.5, −0.3<R_{8d}(DO−Chl *a*, pH−Chl *a*)<0.5 时, 有一项 Δ*R*>0.34 则判定为异常值。

利用 R_{8d} 和 Δ*R* 两项指标进行渐变异常数据的判断与识别流程如图 6 所示。第一步利用单一指标 R_{8d} 来判定简单的正确数据和异常数据; 第二步则是利用 Δ*R* 作为 R_{8d} 稳定性指标来进一步判定异常数据。

4　结果与讨论

4.1　识别算法实例与结果对比分析

本文利用浙江温州、台州及舟山海域 NJ01 浮标、TZ01 浮标以及 ZS04 处浮标典型数据, 对渐变异常数据的判定方法进行了适用性验证。首先, 选取同样位于台州外海 TZ01 浮标的 2015 年 4–6 月(图 7a–c)和 2014 年 6–7 月(图 7d–f)的两组正常数据。第一组

2015 年的原始数据是十分具有代表性的正确数据，Chl *a* 浓度、DO 浓度和 pH 之间存在非常高的正相关性，两两 $R_{8\,d}$ 大于 0.5，并且相关性的变化平稳，ΔR 都小于 0.34。而第二组 2014 年的正确数据相比于第一组数据变化更加复杂，从 6 月下旬，Chl *a* 浓度与 DO 浓度和 pH 存在极弱的相关性，$R_{8\,d}$(DO−Chl *a*) 和 $R_{8\,d}$(pH−Chl *a*) 接近于 0，但随着 7 月初藻华事件的出现，上述 $R_{8\,d}$ 逐渐升高，并在整个藻华期间处在一个平稳的高相关性时期。虽然在这一过程中 $R_{8\,d}$ 的总体变化很大，但根据 ΔR 的计算结果都小于 0.34，可以说明这个变化过程是稳定的渐变过程。那么根据图 6 的识别方法仍然可以准确判定为正确数据，因此证明了本文方法的适用性。

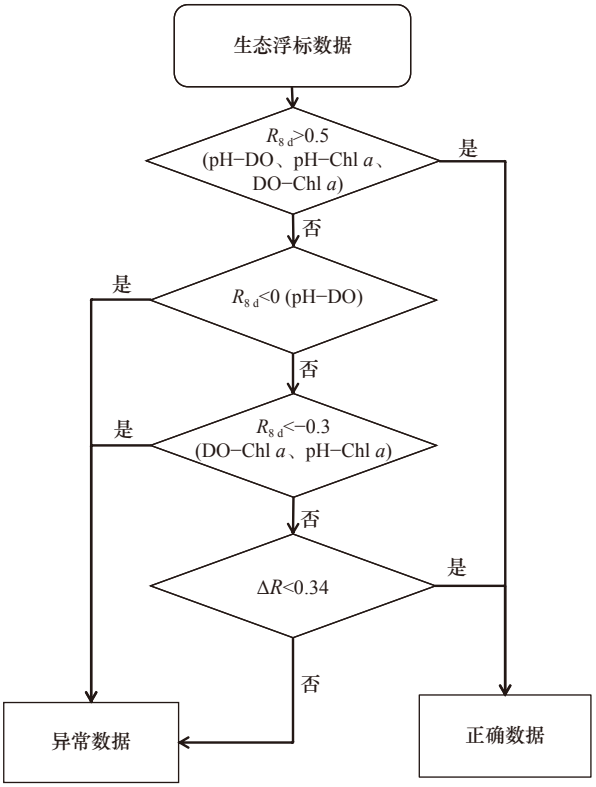

图 6 数据判断流程图

Fig. 6 Flow chart of buoy data processing

本研究同样利用了浙江台州海域 TZ01 浮标 2013 年 5−6 月(图 8a−c)，以及温州海域 NJ01 浮标 2014 年 6 月(图 8d−f)和 2015 年 3−4 月(图 8h−j)的 3 组存在渐变异常的数据集对本文识别方法进行了适用性验证。

第一组案例是 2013 年 5−6 月 TZ01 浮标数据(图 8a)，其中，5 月初有一次藻华事件，3 个参数变化同步浮标数据正常，而渐变异常数据实际上出现在 5 月 24 日前后，pH 上升发生偏离，后续在 5 月 30 日前后恢复正常。图 8b 和图 8c 分别给出了对应的 $R_{8\,d}$ 和 ΔR 数据，图中红色为异常值区间(5 月 24−29 日)，灰色部分为相关性计算受异常值影响区间。可以看出 5 月 15 日出现 ΔR>0.34 的情况，但是根据图 6 判断流程，5 月 15 日 3 组 $R_{8\,d}$ 都升高到 0.5 以上，因而仍然判定

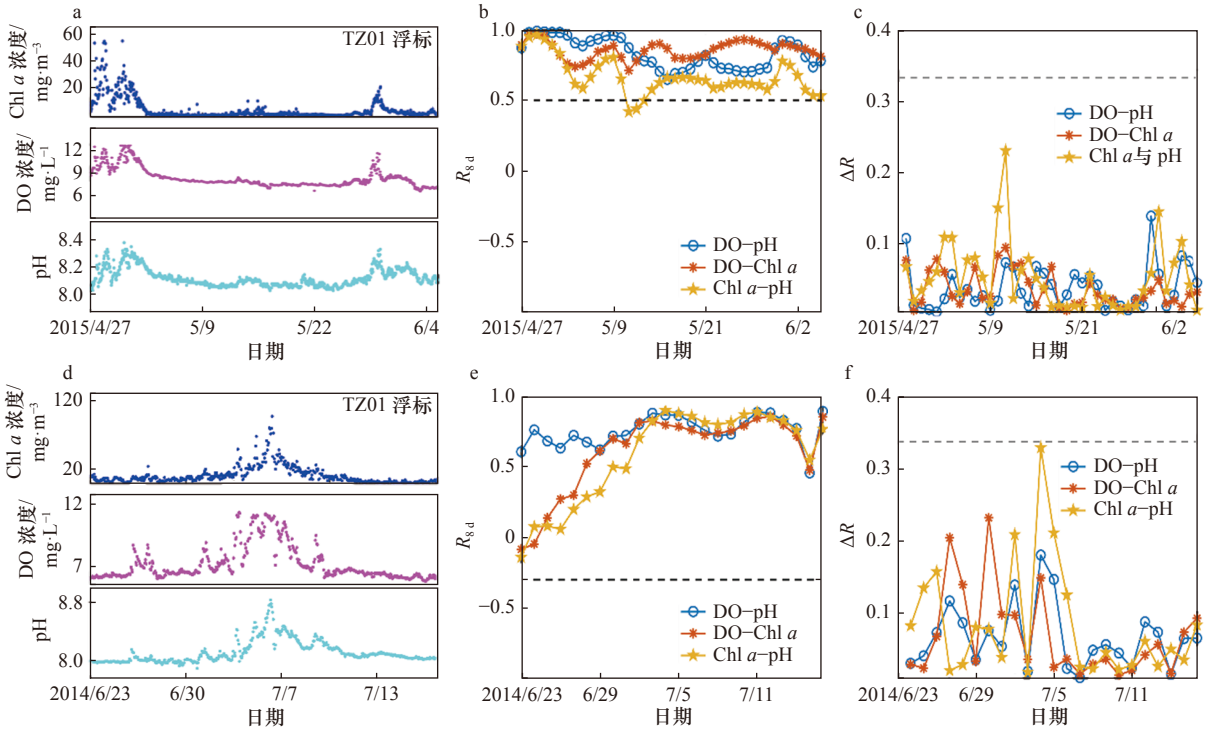

图 7 2015 年 4−6 月(a−c)和 2014 年 6−7 月(d−f)TZ01 浮标的原始数据(a, d)、$R_{8\,d}$(b, e)和 ΔR(c, f)

Fig. 7 TZ01 buoy raw data (a, d), $R_{8\,d}$ (b, e), and ΔR (c, f) in April to June, 2015 (a−c) and June to July, 2014 (d−f)

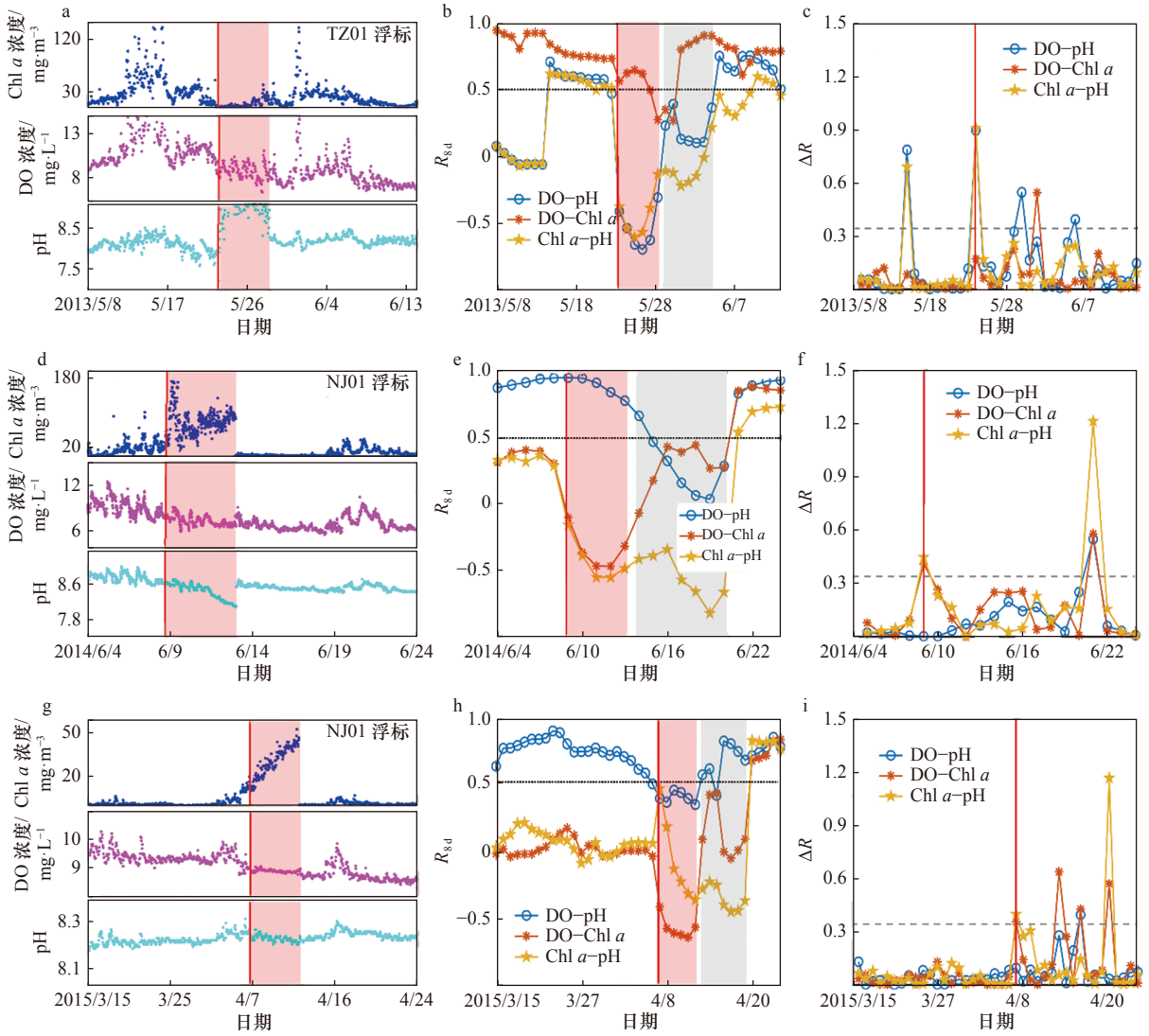

图 8　2013 年 5−6 月 TZ01 浮标原始数据(a)、R_{8d}(b)和 ΔR(c), 2014 年 6 月(d−f)和 2015 年 3−4 月(g−i)NJ01 浮标原始数据(d, g)、R_{8d}(e, h)和 ΔR(f, i)

Fig. 8　TZ01 buoy raw data (a), R_{8d} (b), and ΔR (c) in May to June, 2013, and NJ01 buoy raw data (d, g), R_{8d} (e, h), and ΔR (f, i) in June 2014 (d−f) and March to April 2015 (g−i)

为正确数据。而在 5 月 24 日(表 4), R_{8d}(pH−DO)下降到−0.42, R_{8d}(Chl a−pH) 下降到−0.39, 并且 ΔR(pH−DO)为 0.89, 大于 0.34, 多项指标符合本文渐变异常数据的判定标准, 因此成功判定为异常数据, 实现了数据异常早期识别。另外 DO−Chl a 的 R_{8d} 虽然有所下降, 但仍然保持在 0.5 以上。如果利用排除法, 本方法可以进一步判定是 3 个参数中 pH 值出了问题。

第二组案例是 2014 年 6 月 NJ01 浮标的数据, 如图 8d 所示渐变异常数据出现在 6 月 9 日前后。根据图 8e 和图 8f 对应的 R_{8d} 和 ΔR 结果(表 4), 在 6 月 9 日虽然 R_{8d}(DO−Chl a)和 R_{8d}(pH−Chl a)的下降并未超过−0.3, 但是 DO−Chl a 和 pH−Chl a 的 ΔR 分别高

表 4　浮标出错日期的 R_{8d} 和 ΔR 情况

Table 4　R_{8d} and ΔR of buoy error date

组别	R_{8d}(pH−DO)	R_{8d}(DO−Chl a)	R_{8d}(pH−Chl a)	ΔR(pH−DO)	ΔR(DO−Chl a)	ΔR(pH−Chl a)
第一组(5月24日)	−0.42	0.56	−0.39	0.89	0.17	0.90
第二组(6月9日)	0.96	−0.10	−0.16	0.003	0.41	0.44
第三组(4月7日)	0.38	−0.41	0.45	0.09	0.35	0.40

达 0.41、0.44, 皆大于 0.34, 故可判断出在 6 月 9 日数据出现了异常。根据后续 Chl *a* 浓度的变化也可看出 6 月 9 日是异常数据出现较早时期, 证明了本文方法早期识别的有效性。

第三组案例是 2015 年 3 月和 4 月 NJ01 浮标数据(图 8g–i), 同样可以看出渐变异常数据开始出现在 4 月 7 日的前后。然而与前两组数据不同的是, 本组不同的 R_{8d} 出现相反的变化趋势(表 4), 其中 R_{8d}(DO−Chl *a*)降到了−0.41, 而 R_{8d}(pH−Chl *a*)却上升到了 0.45, 但两者的 Δ*R* 都大于 0.34, 最终 4 月 7 日被判定为异常数据起始点。通过上述多组案例的验证, 本文的识别方法能够适用于浙江沿海多参数浮标数据的渐变异常类型识别。

4.2 水体生化参数变化带来滞后现象的影响

海洋水体生化特性变化并不完全同步, 一些参数相比于其他参数具有滞后现象。部分大型赤潮发生时, 如图 9a 所示的舟山 2015 年 9 月 25 日前后的一次赤潮事件, 藻类暴发导致叶绿素峰值的出现往往早于 DO 浓度或 pH。这就导致在初期, 叶绿素浓度与 DO 浓度或 pH 的 R_{8d} 数值相对较小(图 9b), 同时随着赤潮的发展, R_{8d} 迅速升高, 导致 Δ*R* 可能大于阈值 0.34(图 9c)。此种情况下, 如果 R_{8d} 都高于 0.5 呈现极高正相关性亦可判定为正确数据, 但是如果小于 0.5 则很容易被错误识别为异常数据。在这种剧烈而快速的海洋过程中, 如何有效识别正常海洋规律现象与渐变异常数据是十分重要但又极具挑战性的问题。实际上, 本文采用的方法是时间同步相关性分析。如果利用时间延迟模式或许可以有效避免此类异常数据的错误识别。然而, 目前对于该类滞后现象产生的海洋学机制并不十分明了。因此, 在识别方法中如何具体引入时间延迟模式(如延迟区间等)还需进一步研究。

4.3 季节性(冬季)相关特征差异的影响

本文数据主要集中在上半年, 而季节性变化(特别是冬季)对近海海域的海洋现象有着重要影响。由于浙江沿岸受河流冲淡水、季风和各类水团影响较大, 冬季浙江沿海海域受浙闽沿岸流影响, 水中悬浮泥沙含量高, 限制了藻类生长。在这一时期, 水体 Chl *a* 浓度、pH 和 DO 浓度的相关性比春、夏和秋 3 个季节要弱很多。图 10 为 2014 年冬季(2014 年 11 月至 2015 年 2 月)TZ01 浮标的观测结果, 从图 10a 可看出, 3 个参数的时序变化依然连续平稳, Δ*R* 也基本在

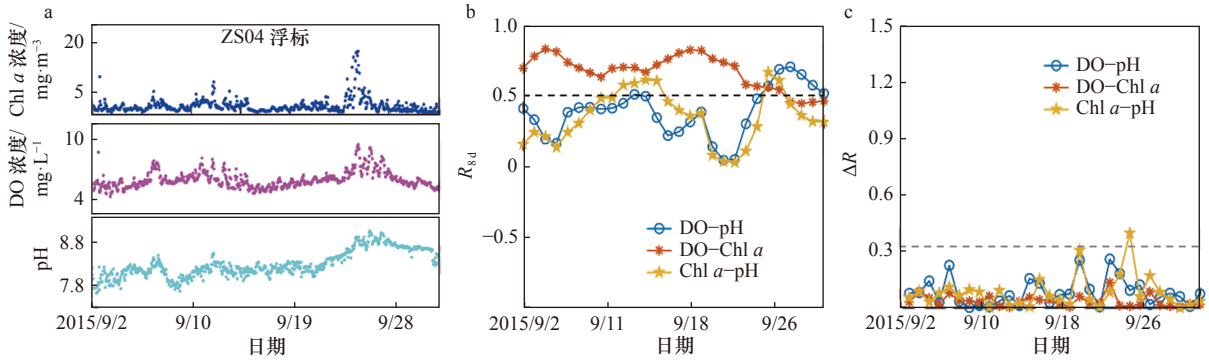

图 9 2015 年 9 月 ZS04 浮标原始数据(a)、R_{8d}(b)和 Δ*R*(c)

Fig. 9 ZS04 buoy raw data (a), R_{8d} (b), and Δ*R* (c) in September, 2015

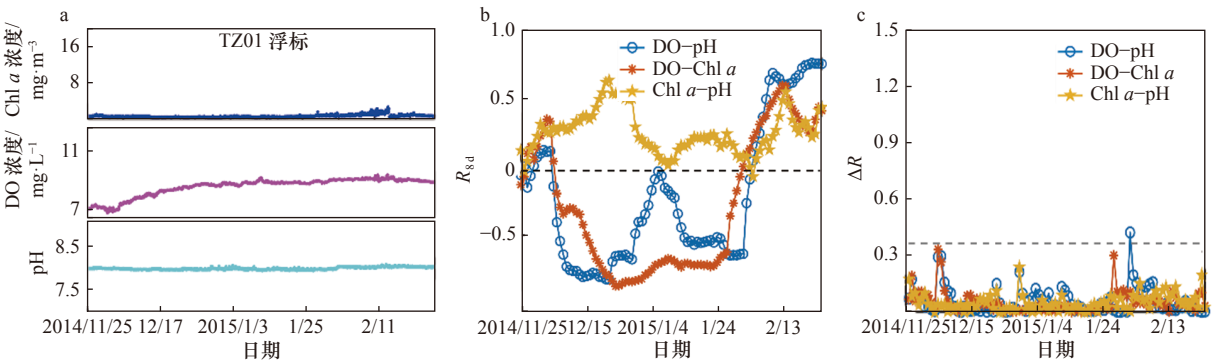

图 10 2014 年冬季 TZ01 浮标原始数据(a)、R_{8d}(b)和 Δ*R*(c)

Fig. 10 TZ01 buoy raw data (a), R_{8d} (b), and Δ*R* (c) in the winter of 2014

0.34 以内(图 10c), 也说明了数据的平稳性, 但是相关性系数 R_{8d} 时高时低(图 10b), 甚至出现极强的负相关。主要原因是在冬季水温低, 藻类丰富度较小, 导致叶绿素浓度变化很小, 对 pH 和 DO 浓度的影响作用有限。反而在这一时期, DO 浓度的变化受温度影响较大, 有个缓慢上升过程。因此, 冬季浙江海域的 3 个参数在机理上并不存在明确的相关性, 而本文方法也仅基于同一年份前序时间数据的相关性进行渐变异常数据识别, 所以在冬季可能会失效。

在上述不适用的情况下, 我们需对时序相关性的概念进一步拓展, 可利用同一海域季节性数据存在物候等现象, 依靠历史同一时期观测数据集等, 对浮标渐变异常数据进行有效地识别。如刘增宏等[34]采用历史水文观测资料集得到的温−盐度关系对 Argo 剖面浮标盐度资料进行校正, 王辉赞等[35]也同样通过寻找 Argo 浮标不同剖面位置与其“最佳匹配”历史剖面资料对比判别的途径, 对 Argo 浮标盐度偏移现象进行有效甄别。上述方法虽然用的是连续深度剖面数据, 但是替换成连续时间序列数据同样适用。如图 11 所示, 为台州大陈浮标在 2014−2015 年冬季与 2015−2016 年冬季的 pH、DO 浓度和 Chl *a* 浓度数据对比结果, 可以看出这两年冬季同一时期的 pH、DO 浓度和 Chl *a* 浓度数据变化有较好的一致性趋势。因此利用与多年历史数据的相关性, 可对 pH、DO 浓度和 Chl *a* 浓度数据进行异常识别。这或许是本文识别方法在冬季失效问题的一种有效解决方式, 但需要大量历史数据的积累, 目前还无法实现, 需要进一步研究。

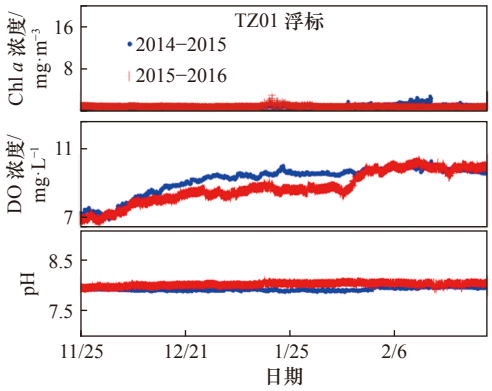

图 11　TZ01 浮标冬季原始数据

Fig. 11　TZ01 buoy raw data in winter

5　结论

本研究通过对浙江沿岸 6 处浮标多年多参数观测数据进行分析, 发现了与传统跳变异常数据不同的渐变异常数据类型。该异常数据类型呈现出在时序变化过程连续平稳, 但随时间逐渐偏移, 最后整体偏离正常数据的分布特征; 并且在异常发生的初始阶段, 其变化特征与由海洋环境变化引起的变化趋势很难在没有先验知识的条件下进行区分。因此本文提出了一种假设: 浮标观测到的正常多参数数据不仅单一参数在一定时序上的变化是平稳连续的, 并且两两参数间的相关性在一定时序上是稳定甚至是一致的。根据上述假设, 本文建立了基于 pH、DO 浓度、Chl *a* 浓度数据两两相关性的渐变异常数据类型自动识别方法, 确定了以 8 d 时间窗口的两两相关系数(R_{8d})作为核心相关性表征指标, 并将前后两天 R_{8d} 之差的绝对值(ΔR)作为判断相关性时序分布稳定性指标, 形成了利用 R_{8d} 和 ΔR 两项指标进行渐变异常数据判断与识别的流程。

本文提出的方法重点突出了多元参数间相关性系数时间序列上的变化特征, 各判别指数计算过程简单、直观, 易于实际浮标监测工作人员的理解和掌握。通过浙江沿海浮标实际测量数据案例检验, 证明了该方法可以用于渐变异常数据类型的实时监测, 对浮标的传感器渐变异常做到早期识别, 特别是由生物污垢导致传感器测量值持续增加而引起的假赤潮现象, 有较好的识别效果, 可解决由此带来的赤潮预报虚警等问题。因此, 在指导浮标日常检查与维护、确保数据的准确性和完整性方面有实际意义。本文根据单参数自校方法无法识别渐变异常数据类型, 提出了一种简单, 实用的有效解决方法。此方法为渐变异常值的自动识别及处理提供了新的思路。由于所处海域的不同, 可能相关性稳定的时间窗口有所不同, 需因地制宜, 考虑季节性差异等因素的影响。因此在后续研究中应当针对在多参数变化不同步、冬季数据和非高斯分布数据等情况下, 识别精度不高等局限性, 可利用多年历史数据对其进行物候特征分析, 提高相关性识别方法精度。

参考文献:

[1]　黄谟涛, 翟国君, 王瑞, 等. 海洋测量异常数据的检测[J]. 测绘学报, 1999, 28(3): 269−277.
Huang Motao, Zhai Guojun, Wang Rui, et al. The detection of abnormal data in marine survey[J]. Acta Geodaetica et Cartographica Sinica, 1999, 28(3): 269−277.

[2] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB 17378.2−2007, 海洋监测规范 第2部分: 数据处理与分析质量控制[S]. 北京: 中国标准出版社, 2008.
General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of China. GB 17378.2−2007, The specification for marine monitoring—Part 2: data processing and quality control of analysis[S]. Beijing: China Standard Press, 2008.

[3] Sivareddy S, Paul A, Sluka T, et al. The pre-Argo ocean reanalyses may be seriously affected by the spatial coverage of moored buoys[J]. Scientific Reports, 2017, 7: 46685.

[4] Dong Guozhong, Chen Dongying. Quality control algorithm for marine meteorological data based on interest degree association rules[J]. Journal of Coastal Research, 2019, 94(S1): 173−176.

[5] 黄冬梅, 康培红, 张明华, 等. 一种基于ULDB的海洋环境监测数据管理系统[P]. 中国: 201110004234. X, 2011−06−01.
Huang Dongmei, Kang Peihong, Zhang Minghua, et al. ULDB (Databases with Uncertainty and Lineage)-based marine environmental monitored data management system[P]. CN: 201110004234. X, 2011−06−01.

[6] ARGO. Argo quality control management, Version 2.4, Argo data management[Z]. 2009.

[7] D'Ortenzio F, Thierry V, Eldin G, et al. White book on oceanic autonomous platforms for biogeochemical studies: instrumentation and measure (PABIM), version 1.3[Z]. 2010.

[8] Fu Wenju, Huang Guanwen, Yang Yuanxi, et al. Multi-GNSS combined precise point positioning using additional observations with opposite weight for real-time quality control[J]. Remote Sensing, 2019, 11(3): 311.

[9] Cosoli S, Grcic B, De Vos S, et al. Improving data quality for the Australian high frequency ocean radar network through real-time and delayed-mode quality-control procedures[J]. Remote Sensing, 2018, 10(9): 1476.

[10] Duan Boheng, Zhang Weimin, Yang Xiaofeng, et al. Assimilation of typhoon wind field retrieved from scatterometer and SAR based on the Huber norm quality control[J]. Remote Sensing, 2017, 9(10): 987.

[11] Fichot C G, Downing B D, Bergamaschi B A, et al. High-resolution remote sensing of water quality in the San Francisco Bay–Delta Estuary[J]. Environmental Science & Technology, 2016, 50(2): 573−583.

[12] Schuckmann K, Garau B, Wehde H, et al. MyOcean: real time quality control of temperature and salinity measurements[R]. 2010.

[13] 张明, 张韧, 王辉赞, 等. 基于Argo浮标数据的Aquarius数据产品质量评估[J]. 海洋信息, 2015(3): 21−28.
Zhang Ming, Zhang Ren, Wang Zanhui, et al. Quality evaluation of Aquarius data products based on Argo buoy data[J]. Marine Information, 2015(3): 21−28.

[14] 史静涛. 海洋环境实时观测数据质量控制方法研究与软件实现[D]. 天津: 国家海洋技术中心, 2010.
Shi Jingtao. The data quality control method research and software realization for marine environment real-time observation[D]. Tianjin: National Marine Technology Center, 2010.

[15] Ishii M, Fukuda Y, Hirahara S, et al. Accuracy of global upper ocean heat content estimation expected from present observational data sets[J]. SOLA, 2017, 13: 163−167.

[16] Shulski M, Cooper S, Roebke G, et al. The Nebraska Mesonet: technical overview of an automated state weather network[J]. Journal of Atmospheric and Oceanic Technology, 2018, 35(11): 2189−2200.

[17] Jiang Jingang, Sun Lu, Fan Zhongya, et al. Outlier detection and sequence reconstruction in continuous time series of ocean observation data based on difference analysis and the Dixon criterion[J]. Limnology and Oceanography Methods, 2017, 15(11): 916−927.

[18] 窦文洁, 蒋锦刚, 周斌, 等. 基于多参数差分相关的海洋时序观测数据滤波算法[J]. 海洋学报, 2012, 34(5): 50−58.
Dou Wenjie, Jiang Jingang, Zhou Bin, et al. An algorithm for the difference correlation filter for multi-parameter marine timing observation data[J]. Haiyang Xuebao, 2012, 34(5): 50−58.

[19] Oguma S, Nagata Y. Skewed water temperature occurrence frequency in the sea off Sanriku, Japan, and intrusion of the pure Kuroshio water[J]. Journal of Oceanography, 2002, 58(6): 787−796.

[20] Mardia K V, Kent J T, Bibby J M. Multivariate analysis[J]. Probability and Mathematical Statistics, 1979, 37(1): 123−131.

[21] Wei W W S. Multivariate Time Series Analysis and Applications[M]. New York: John Wiley & Sons Inc., 2019.

[22] Bartholomew D J. Time series analysis forecasting and control[J]. Journal of the Operational Research Society, 1971, 22(2): 199−201.

[23] Olson D A, Riedel T P, Long R, et al. Time series analysis of wintertime O_3 and NO_x formation using vector autoregressions[J]. Atmospheric Environment, 2019, 218: 116988.

[24] Tsay R S. Time series model specification in the presence of outliers[J]. Journal of the American Statistical Association, 1986, 81(393): 132−141.

[25] Razi M A, Athappilly K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models[J]. Expert Systems with Applications, 2005, 29(1): 65−74.

[26] Martínez-Álvarez F, Troncoso A, Riquelme J C, et al. Discovery of motifs to forecast outlier occurrence in time series[J]. Pattern Recognition Letters, 2011, 32(12): 1652−1665.

[27] Cucina D, Di Salvatore A, Protopapas M K. Outliers detection in multivariate time series using genetic algorithms[J]. Chemometrics and Intelligent Laboratory Systems, 2014, 132: 103−110.

[28] Tsay R S. Outliers, level shifts, and variance changes in time series[J]. Journal of Forecasting, 1988, 7(1): 1−20.

[29] Harlé F, Chatelain F, Gouy-Pailler C, et al. Bayesian model for multiple change-points detection in multivariate time series[J]. IEEE Transactions on Signal Processing, 2016, 64(16): 4351−4362.
[30] Boto K G, Bunt J S. Dissolved oxygen and pH relationships in northern Australian mangrove waterways[J]. Limnology and Oceanography, 1981, 26(6): 1176−1178.
[31] Wallace J, Champagne P, Hall G. Time series relationships between chlorophyll-*a*, dissolved oxygen, and pH in three facultative wastewater stabilization ponds[J]. Environmental Science: Water Research & Technology, 2016, 2(6): 1032−1040.
[32] 谢群, 张瑜斌, 孙省利, 等. 流沙湾溶解氧的分布特征及其相关因素的探讨[J]. 环境科学与技术, 2009, 32(9): 39−44.
Xie Qun, Zhang Yubin, Sun Shengli, et al. Distribution characteristics of dissolved oxygen and correlating factors analysis in Liusha Bay[J]. Environmental Science & Technology, 2009, 32(9): 39−44.
[33] Hollinger D Y, Richardson A D. Uncertainty in eddy covariance measurements and its application to physiological models[J]. Tree Physiology, 2005, 25(7): 873−885.
[34] 刘增宏, 许建平, 修义瑞, 等. 参考数据集对Argo剖面浮标盐度观测资料校正的影响[J]. 海洋预报, 2006, 23(4): 1−12.
Liu Zenghong, Xu Jianping, Xiu Yirui, et al. The effect of reference dataset on calibration of Argo profiling float salinity data[J]. Marine Forecasts, 2006, 23(4): 1−12.
[35] 王辉赞, 张韧, 王桂华, 等. Argo浮标温盐剖面观测资料的质量控制技术[J]. 地球物理学报, 2012, 55(2): 577−588.
Wang Zanhui, Zhang Ren, Wang Guihua, et al. Quality control of Argo temperature and salinity observation profiles[J]. Chinese Journal of Geophysics, 2012, 55(2): 577−588.

Identification of abnormal buoy data based on time series correlation analysis method

Zhang Yu[1], Zhou Yan[2], Tao Bangyi[1,3], Gu Jixing[4], Zhao Chuan'gao[4], Hao Zengzhou[1], Zhang Yiwei[1,5], Huang Haiqing[1], Mao Zhihua[1]

(1. *State Key Laboratory of Satellite Ocean Environment Dynamics*, *Second Institute of Oceanography*, *Ministry of Natural Resources*, *Hangzhou* 310012, *China*; 2. *Zhejiang Academy of Marine Sciences*, *Hangzhou* 310007, *China*; 3. *Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou)*, *Guangzhou* 511458, *China*; 4. *Yantai Marine Environmental Monitoring Center Station*, *State Oceanic Administration*, *Yantai* 264006, *China*; 5. *Shanghai Institute of Technical Physics*, *Chinese Academy of Sciences*, *Shanghai* 200083, *China*)

Abstract: The identification of abnormal marine ecological buoy data is the key to ensure the quality of buoy data. In this study, we found that the gradual abnormal data type is different from the traditional jump abnormal data through analysis of the coastal buoy data in Zhejiang for many years. With a single parameter analysis method, it is difficult to work out accurately the new gradual abnormal data type of stable and gradual deviation from the normal data. Therefore, multiple parameters correlation coefficient method is proposed based on the relationships between pH, dissolved oxygen and chlorophyll *a* on the condition of that the correlation between two parameters is stable or even consistent at a certain time series. There are two simple statistical parameters of the cross-correlation coefficient of 8-day time window ($R_{8\,\mathrm{d}}$) and the difference of $R_{8\,\mathrm{d}}$ (ΔR) in this method. Those could be used to automatically detect the gradual abnormal buoy data and do very well. The multiple parameters correlation coefficient method provides a new idea for the gradual abnormal data identification, and also improves the automatic monitoring capability of marine ecological buoy abnormal data.

Key words: ecological buoy; environmental monitoring; validation; correlation analysis