

海洋环境监测数据集质量控制方法研究^{*}

向先全¹,路文海^{1,2},杨翼¹,付瑞全¹

(1. 国家海洋信息中心 天津 300171; 2. 中国海洋大学 青岛 266100)

摘要:文章以各监测机构上报的海洋环境监测数据为对象,有别于海洋环境监测实验室内的质量控制,提出了监测数据集的一系列质量控制方法,有效提高监测数据集质量水平,为海洋环境保护信息化持续健康发展提供高质量的数据保障。

关键词:海洋环境;监测数据集;质量控制;一致性检验;离群点检验

中图分类号:P76; X321 **文献标识码:**A **文章编号:**1005-9857(2015)01-0088-04

1 引言

海洋环境监测数据质量直接影响海洋环境管理决策的科学性,准确可靠的监测数据是海洋环境科学研究和海洋综合管理的依据^[1]。如何进行海洋环境监测数据集的质量控制,是海洋环境保护及海洋信息化持续健康发展的关键问题。

目前世界上先进的海洋环境观测系统,如美国海军和国家海洋与大气管理局发起的永久的集成海洋观测系统 IOOS,欧盟资助的一项海洋信息标准化项目 MarineXML,以及现在正在全世界建立的 ARGO 计划等,都有自身观测及监测数据的质量控制模式和标准。但这些质控模式和标准是根据各自的观测内容制订的,有很强的针对性和限定性。在海洋环境监测质量保证与控制方面,美国环保署(USEPA)、欧盟环境委员会、加拿大渔业与海洋部(DFO)、日本海上保安厅(JCG)等发达国家及组织都对各自的海洋环境监测制定了相关的制度和规范^[2-4]。

经过近年来国内海洋环境监测工作的大力推进,国内海洋环境监测制度及质量控制规范制定较多,如:国家海洋局“海洋监测规范第 2 部分:数据处理与分析质量控制”,国家环境保护部“全国近岸海域环境监测网质量保证和质量控制工作规定(试行)”。但这些监测质量控制及检验方法大多是针对实验室内质量控制的研究,目前,海洋环境监测数据集质量控制尚未有一个系

统化的检验方法^[5]。

本研究以各监测机构上报的海洋环境监测数据为对象,有别于海洋环境监测实验室内的质量控制,提出了监测数据的一系列质量控制方法。

2 监测数据集质量控制方法

海洋环境监测是为了获取高质量可靠的监测数据,高质量的监测数据应具有准确性、精密性、代表性、完整性及可比性^[6-7]。本研究从以下 7 个方面着重介绍了海洋环境监测数据集的质量控制方法。

2.1 站位基础信息一致性检验方法

对于海洋环境监测数据的质量控制,首先要保证该条记录的基础信息的准确性。通过基础信息一致性检验可解决监测数据记录查重、基础信息错误等问题。一条监测数据记录的基础信息主要包括:监测区域、站位编号、站位经纬度、监测日期等。一般而言,对于同一任务的监测要素,这 4 个信息点唯一确定一条监测数据记录;另外,对于生物质量、生物多样性监测等的海洋环境监测数据,由于每个站位中,每个物种都形成一条记录,因此,对于含有生物种名信息的监测数据记录,利用监测区域、站位编号、站位经纬度、监测日期和生物种类共 5 个信息点来确定该条记录的唯一性。根据不同的监测任务以及监测要素的数据记录特点,分以下 3 种情形对站位

^{*} **基金项目:**海洋公益科研专项项目“海洋生态红线区划管理技术集成研究与应用”(201405007);天津市科技兴海项目“天津沿海风险污染源排查和环境灾害风险评价及应用研究”(KJXH2012-28)。

的基础信息进行质控。

(1) 站位编号、站位经纬度和监测日期。对于水质、沉积物等的海洋环境监测数据,存在 4 类问题:① 站位编号相同,经纬度不同;② 站位编号不同,经纬度相同;③ 监测日期相同,站位编号相同;④ 监测日期相同,经纬度相同。

(2) 站位编号、站位经纬度、监测日期和物种。对于生物质量、生物多样性监测等海洋环境监测数据,主要存在 4 类问题:⑤ 站位编号相同,经纬度不同;⑥ 站位编号不同,经纬度相同;⑦ 监测日期相同,站位编号相同,且物种相同;⑧ 监测日期相同,经纬度相同,且物种相同。

对于站位编号和经纬度不一致的 4 类情况——①、②、⑤、⑥,属于位置精度问题。对于监测日期相同的 4 类情况——③、④、⑦、⑧,判断两条记录的监测参数数值是否完全一致,若完全一致则认为是重复记录,若不完全一致,可认为是平行样记录,但需进一步核实。

2.2 值域一致性检验方法

在海洋环境监测中,每个监测参数有其对应的经验值域范围,通过值域检测规则检验该参数是否超出值域上下限。判断监测参数的正常值域范围可通过参考以下几种方法进行检验。

2.2.1 根据监测参数的理化性质

如盐度、水温、pH、透明度、溶解氧浓度等。对于该类指标,其数据值域变化范围差异不大,且有相关的理论依据作为指导。因此需根据各自参数指标的理化性质及中国近海分布规律给出各自的量级范围。

2.2.2 参照最低检出限及最大污染倍数检验

利用该监测参数在海洋环境监测中规定的监测方法、监测仪器等,确定该监测参数的最低检出限,利用最低检出限指导该参数值域的最小值。根据《GB17378.2 海洋监测规范第 2 部分,数据处理与分析质量控制》中的规定:低于检出限的测试结果,应报未检出,但在区域性监测检出率占样品频数的 $1/2$ 以上(包括 $1/2$)或不足 $1/2$ 时,未检出部分可分别取检出限的 $1/2$ 和 $1/4$ 量参加统计运算。因此在进行量级检验中,可将值域下限取其检出限的 $1/4$ 。

另外,利用该监测参数最大污染等级的倍数,指导该监测参数值域的最大值。重金属类及

有毒污染指标,一般情况下在海水中含量较低,除部分陆源排污口、特殊海洋工程用海等附近海域,不应该出现劣四类污染,因此可将值域上限取其最大污染等级标准的浓度值;化学需氧量、活性磷酸盐、溶解无机氮、总氮、石油类、硫化物等营养状况及一般污染指标,在值域检验时,可将该部分监测指标量级范围的阈值上限取最高污染等级的 10 倍。

2.2.3 参照历史监测参数范围检验

对该监测参数在该区域中近 10 年或更长时间的值域范围进行统计,给出该监测参数在海洋环境监测的最大值、最小值、分位值、平均值、标准差等统计信息,用于判别该参数浓度值的可靠性,并进一步指导该参数值域范围的划分。

2.2.4 区域环境污染特征及相关资料文献检验

利用区域环境污染特征以及相关资料文献,分析特定监测区域内监测参数的污染特征,给出其值域范围。超过该参数值域范围的参数值作为可疑值,需联系监测机构进行沟通解决。

2.3 逻辑一致性检验方法

某些监测参数间存在一定的逻辑关系,即监测参数与监测参数间存在某种相关关系,有些关系具有一定的规律性,可利用这些相关性和规律性,对监测参数的数据集进行质量控制。参数逻辑一致性检验主要有以下几类方法。

(1) 参数不同形态的关系检验。① 总氮(TN)是有机氮(TON)和无机氮(TIN)的总和,因此总氮含量应高于无机氮含量;② 总磷(TP)是水中正磷酸盐、聚合磷酸盐、可水解磷酸盐及有机磷等的总和,因此,总磷含量应高于溶解态磷含量;③ 溶解态氮含量应高于溶解无机氮含量;④ 溶解态磷含量应高于活性磷酸盐含量;⑤ 铬的化合物常见价态有三价和六价,两者可以相互转化,因此总铬含量大于六价铬含量。

(2) 同一物质在不同监测参数之间的关系检验。① 由于一般水体有一部分有机物质不能在 COD_{Mn} 法中被氧化而可在 COD_{Cr} 法中被氧化,因此 COD_{Cr} 含量高于 COD_{Mn} ;② 在 BOD_5 的测定条件下,许多有机物不能被微生物分解,但能为 $K_2Cr_2O_7$ 分解,因此 COD_{Cr} 含量高于 BOD_5 ;③ 大肠菌群和粪大肠菌群都是表征与粪便污染有关的细菌指数,大肠菌群数应高于粪大肠菌

群数。

(3)不同参数内在联系的检验。① 溶解氧在海水中的溶解度,随温度的升高而降低,随盐度的增加而减少;② pH 值与总碱度相关,pH 值越大,总碱度越小;③ 溶解氧含量正常的海水中,硝酸盐含量一般应高于亚硝酸盐含量。

在监测参数逻辑一致性检验中常见的问题主要是:无机氮、溶解态氮、总氮以及活性磷酸盐、溶解态磷、总磷之间的关系错误。根据对这些监测参数的分析方法的研究发现,除了铵盐由于实验室环境污染容易对分析产生正偏差以外,其余参数在分析过程中还原过程和氧化过程的化学反应效率一般小于 100%,也就是分析结果一般产生负偏差。基于这个原理,根据国内外相关文献的研究成果,在对这些有问题的数据进行处理时,遵循以下原则:① 溶解态氮或总氮含量低于溶解无机氮含量的,判断为溶解态氮或总氮数据有误;② 溶解态磷或总磷含量低于活性磷酸盐含量的,判断为溶解态磷或总磷数据有误;③ 总氮含量低于溶解态氮含量的,因这两个要素仅仅是形态差别,分析原理完全一样,无法判断哪个要素数据有误,则同时标注为“可疑”;④ 总磷含量低于溶解态磷含量的,因这两个要素仅仅是形态差别,分析原理完全一样,无法判断哪个要素数据有误,则同时标注为“可疑”;⑤ 亚硝酸盐含量明显高于硝酸盐含量的,核实溶解氧、pH 值等要素,并结合站位所在位置进行具体判别。

2.4 参数值时间分布检验方法

某些监测参数在不同监测时间上,存在一定的值域关系。监测参数随时间序列变化关系比较复杂,需要专家经验结合监测区域环境特征对历年同一监测时段监测参数的浓度变化、年际参数浓度的均值变化、年内各监测月份浓度的变化趋势做出定性判断,从而判断监测参数值的合理性。

2.5 参数值空间分布检验方法

海洋环境在空间上有较强的分布规律,一般认为,海洋环境近岸污染程度大于远海。① 悬浮物的含量决定着海水的水色和透明度,大洋中悬浮物含量只有几微克/L,粒度微小,水色深蓝;近岸和河口海区的悬浮物含量达到 100 mg/L 左右,而且颗粒较粗,水色多呈浅蓝、绿以至于黄;② 根

据重金属污染来源和迁移转化的特点,一般认为重金属污染物在海洋环境中的分布规律河口及沿岸水域高于外海。

有些调查要素,如沉积环境中的大部分污染要素,仅从数据上很难发现系统性的偏差,特别是偏差后的数据也属于正常范围内时。在这种情况下,可检查要素的平面分布图,当发现某一区块要素含量明显高于或低于周边时,应追溯原始记录进行核实,一般可能在结果计算、计量单位与数据匹配、数据转移等过程发生错误。对同时时间某一监测值在某海域空间分布上出现指标值突变的情形,可作为可疑数据。

2.6 离群点检验方法

正常的监测数据集应具有一定的分布规律。若发现部分数据与正常数据有显著性差别的数据,此类数据称为离群数据或异常值。出现此类离群点数据,可能是采样、实验室分析、填报等过程中因失误产出的噪声数据,亦可能是某些地区、时段内的真实数据^[8-9]。离群点检验的目的是为了消除噪音或发现潜在的、有价值的海洋环境现象。对于检测出的离群点可视为可疑数据,其处理方法:分析具体时间具体区域的海洋环境异常情况,联系地方监测机构进行核实。

2.6.1 绘制散点图法

散点图是确定两组数据属性之间是否有联系、模式或趋势的最有效、最简便的二维图形方法。为构造散点图,每个数据对可视为一个代数坐标对,画在二维直角坐标系中。对于海洋环境监测参数数据而言,当两条坐标轴均为监测参数时,可反映两个监测参数之间是否存在某些相关关系;当其中一条坐标轴为时间或空间属性时,可快速地反映该监测参数属性的时间或空间分布规律。通过绘制监测参数数据的散点图,还可直观地查找监测参数的离群点数据。

2.6.2 全局离群点检验——探索性数据分析检验法

由于海洋环境监测参数的理化性质差异,且采样的空间分布不均匀,时间分布零散,因此很难界定每个样本是否为独立样本,且很难判断具体到每个监测参数的数据集符合哪个数理统计分布,故不宜采用基于样本独立或正态分布假设的数理统计检验方法进行离群点检验。

探索性数据分析是在一组数据中寻求重要信息的过程,无须借助于先验的理论或假设,直接探索隐藏在数据中的关系、模式和趋势等。其特点是对数据来源的总体不作假设,并且假设检验也经常排除在外。探索性数据分析技术中的箱线图可以清晰地表示数据的分布特征。

2.6.3 空间局部离群点检验

空间离群点是与其空间邻域中其他空间对象的非空间属性值存在明显差异的空间对象。空间离群点挖掘是空间数据挖掘的一个重要分支,其目的是为了消除噪音或发现潜在的、有意义的知识。目前在交通控制、遥感图像分析、气象预报和人口统计数据分等分析等领域,已经有广泛的应用。

空间离群点的挖掘首先出现在空间统计学中,主要方法可分为图形检测和定量检测两类。图形检测基于空间数据的可视化,在图形中以醒目方式突出空间离群点,例如变差云图,其优点是形象直观,但只适合低纬数据^[10-11]。定量检测是一种精确检测方法,通过数学计算将空间离群点与其他数据区分开来。根据海洋环境监测数据的空间特性,将海洋环境监测数据分为空间属性和非空间属性。监测站位经、纬度为监测数据的空间属性,监测参数值为非空间属性。

2.6.4 专家经验检验

利用专家经验对某些监测参数进行审核确认。如海洋大气监测分干沉降和湿沉降,若地方

上报时未进行干湿比换算,则很难判断该部分数据的异常,需结合专家经验进行检验。

2.7 生物种名检验方法

对于海洋生物种名的填报,需区分学名和俗名。同时要保证生物种名的学名和中文名的对应。避免填报出现“地中海辐杆藻”和“地中海辐秆藻”“钝头盒形藻”和“钝头盒型藻”“斯氏小泉虫戎”和“思氏小泉虫戎”“光滑河蓝蛤”和“光滑河篮蛤”“李氏(鱼衔)”和“李氏鱼衔”等同一物种填写多个名称的问题。通过对海洋生物种中文名和学名规范化命名及编码,可解决此类问题,同时可解决其属种的归类问题。

3 结束语

目前,海洋环境监测数据质量控制尚未有一个系统化的检验方法,且绝大多数检验方法是针对实验室内质量控制的研究。本研究从海洋环境监测数据管理的角度,以各监测机构上报的海洋环境监测数据为对象,研究了监测数据集的质量控制方法,包括站位基础信息一致性检验、值域一致性检验、逻辑一致性检验、参数值时间分布检验、空间分布检验、离群点检验、生物种名检验等。这一质控方法体系的运行将会大大提高监测数据质量审查、控制的效率和水平,在我国海洋环境监测业务体系中发挥作用,为海洋环境保护信息化持续健康发展提供高质量的数据保障。

参考文献

[1] 国家海洋局《海洋监测质量保证手册》委员会. 海洋监测质量保证手册[Z]. 北京:海洋出版社,2000.

[2] AGGARWAL C. Re-designing Distance Functions and Distance-based applications for high dimensional Data[J]. SIGMOD Record Date,2001,30(1):13-18.

[3] ANGIULLI F,BASTA S,PIZZUTI C. Distance-based detection and prediction of outlier[J]. IEEETrans. Knowledge and Data Eng,2006,2(18):145-160.

[4] YU Dantong,SHEIKHOLESAMI G,ZHANG Aidong. Findout:finding outliers invery large datasets[J]. Knowledge and Information Systems,2002,4(4):387-412.

[5] 许自舟,宋德瑞,赵辉,等. 海洋环境监测数据质量计算机控制方法研究[J]. 海洋环境科学,2009,28(3):320-323.

[6] 陈上及,马继瑞. 海洋数据处理分析方法及其应用[M]. 北京:海洋出版社,1991.

[7] 曹家新. 监测数据的审核及结果的分析与判断[J]. 四川环境,2002,21(2):19-26.

[8] 陆声链. 孤立点挖掘及其内涵知识发现的研究与应用[D]. 南宁:广西大学,2005.

[9] 魏黎,宫学庆,钱卫宁,等. 高维空间中的离群点发现[J]. 软件学报,2002,13(2):280-290.

[10] 薛安荣. 空间离群点挖掘技术的研究[D]. 镇江:江苏大学,2008.

[11] 薛安荣,鞠时光,何伟华,等. 局部离群点挖掘算法研究[J]. 计算机学报,2007,30(8):1455-1463.