环境监测中异常数据识别与修复

郁　辉　沈祠福　王德顺　疏仁宗
浙江交科环境科技有限公司　浙江　杭州　311300

摘　要：环境监测是保障生态环境质量、促进可持续发展的关键环节。在实际监测过程中，由于设备故障、环境因素干扰等多种原因，常常会产生异常数据。这些异常数据若不及时识别与修复，将严重影响环境监测结果的准确性和可靠性。本文探讨环境监测中异常数据的识别方法，包括统计分析、数据挖掘等，并提出相应的修复技术，如插值法、替换法、同化修复等。通过实例分析，验证这些方法的可行性和有效性，为环境监测数据的准确性和可靠性提供有力保障。

关键词：环境监测；异常数据；识别与修复

1 环境监测的基本概念

环境监测的基本概念是指通过对影响环境质量的各种因素的代表值进行测定，并研究其变化及对环境影响的科学活动。这一过程涉及对大气、水体、土壤、噪声、固体废物、放射性物质以及生态系统等各种环境要素进行定期或不定期的监测，以确定环境质量及其变化趋势。环境监测的目的是准确、及时、全面地反映环境质量现状及发展趋势，为环境管理、污染源控制、环境规划等提供科学依据，同时也是评价环境质量、制定环境标准和环境法规、监督环境管理政策实施效果的重要手段。环境监测通常包括现场调查、监测点布设、样品采集、样品运输与保存、实验室分析测试、数据处理与结果评价等多个环节。随着科技的进步，现代环境监测技术正向着自动化、智能化、网络化方向发展，各种高精度的监测仪器和先进的监测方法不断涌现，极大地提高了环境监测的准确性和时效性。

2 环境监测中异常数据的识别方法

在环境监测过程中，异常数据的识别是确保数据质量、提高监测准确性和可靠性的关键环节。异常数据可能源于仪器故障、操作失误、环境因素突变或数据录入错误等多种原因。及时准确地识别这些异常数据，对于维护监测数据的完整性和真实性至关重要。

2.1 统计方法

统计方法是异常数据识别中最基础且广泛应用的一种方法。它基于概率论和数理统计的原理，通过对数据的分布特征、均值、方差、标准差等统计量的分析，来判断数据是否偏离正常范围。常用的统计方法包括：3σ原则：基于正态分布假设，数据点落在均值正负3倍标准差范围内的概率为99.73%，超出此范围的数据被视为异常。这种方法简单直观，但要求数据近似正态分布。Z分数法：通过计算每个数据点与平均值的距离（以标准差为单位），即Z分数，来判断数据的异常程度[1]。Z分数绝对值越大，数据越异常。箱线图法：利用数据的四分位数（Q1, Q2, Q3）和四分位距（IQR = Q3-Q1）来构建箱线图，通过观察数据点是否落在箱线图之外来判断异常。箱线图法不仅适用于正态分布数据，也适用于偏态分布数据。统计方法的优点是易于理解和实施，但前提是数据需满足一定的分布假设，且对于复杂多变的环境监测数据，单一统计方法可能难以全面捕捉异常。

2.2 相对比较法

相对比较法是通过将监测数据与历史数据、同类数据或预期值进行比较，以识别异常。这种方法适用于时间序列数据或具有明确参考标准的数据集。历史对比：将当前监测数据与同一监测点的历史数据进行比较，若当前数据显著偏离历史趋势或范围，则视为异常。同类对比：将同一时间段内不同监测点的数据进行对比，若某点数据与其他点数据存在显著差异，可能表明该点数据异常。标准比对：将监测数据与国家或地方规定的环境质量标准进行比对，超出标准的数据视为异常。相对比较法的优点在于能够直接反映数据的变化趋势和相对位置，但依赖于可靠的历史数据和标准值，且对于突发性的环境变化可能不够敏感。

2.3 模型法

模型法通过建立数学模型来描述环境变量之间的关系，并据此预测正常值范围，从而识别异常数据。多元线性回归：利用多个自变量（如温度、湿度、风速等）来预测因变量（如污染物浓度），通过比较实际观测值与模型预测值的差异来判断异常。神经网络模型：利用神经网络强大的非线性拟合能力，训练模型以识别正常数据模式，并自动标记不符合模式的数据为异常。机器

学习算法：如支持向量机（SVM）、随机森林等，通过训练大量数据样本，学习数据的内在规律和特征，实现对异常数据的精准识别。模型法的优点在于能够处理复杂多变的数据关系，提高异常识别的准确性和鲁棒性，但模型构建和训练过程较为复杂，且对数据的完整性和质量要求较高。

2.4 聚类方法

聚类方法将相似的数据点归为一类，通过比较数据点与所属类别的中心或特征，识别出偏离类别的异常数据。K均值聚类：预先设定聚类数目K，通过迭代优化每个聚类的中心位置，将数据点分配到最近的聚类中心，远离所有聚类中心的数据点被视为异常[2]。层次聚类：创建数据点的层次结构，通过合并或分裂聚类来寻找最佳聚类数目，异常数据通常位于聚类边缘或自成一类。DBSCAN（基于密度的空间聚类应用噪声）：根据数据点的局部密度来识别聚类，并将低密度区域的数据点标记为噪声（即异常）。聚类方法的优点在于能够自动发现数据的内在结构，对异常数据的识别具有较高的灵活性，但聚类效果的好坏依赖于聚类参数的选择和数据分布的特性。

2.5 时间序列分析

时间序列分析专注于时间序列数据的特性，通过分析数据的趋势、季节性、周期性等，识别数据中的异常点。移动平均法：通过计算一定窗口内的数据平均值来平滑数据，比较原始数据与移动平均线的差异，识别异常点。指数平滑法：利用前一期的平滑值和新观测值的加权和来预测当前值，通过比较预测值与实际观测值的差异来识别异常。ARIMA模型：自回归积分滑动平均模型，能够捕捉时间序列数据的趋势、季节性和随机波动，通过模型残差分析来识别异常。时间序列分析方法的优点在于能够充分利用时间序列数据的内在规律，提高异常识别的准确性，但要求数据具有时间序列特性，且模型构建和参数优化过程较为复杂。

3 环境监测中异常数据的修复技术

3.1 删除法

删除法是最直接的一种异常数据修复技术，其核心思想是将识别出的异常数据直接删除，以确保剩余数据的准确性和一致性。这种方法适用于异常数据数量较少，且删除后对整体数据影响不大的情况。需要注意的是，删除法可能导致数据量减少，从而影响数据的统计特性和分析结果的准确性。因此在使用删除法时，应谨慎评估其对数据分析结果的影响，并确保删除后的数据仍然能够满足分析需求。

3.2 替换法

替换法是通过某种方式将识别出的异常数据替换为合理值的一种修复技术。这种方法通常基于数据的上下文信息和历史数据，利用插值、回归预测等数学方法，为异常数据提供一个合理的替代值。替换法的优点在于能够保持数据量的完整性，避免因删除异常数据而导致的信息损失。然而替换法的准确性依赖于所选择的替代方法和替代值的合理性。在使用替换法时，应确保替代方法能够准确反映数据的真实情况，并避免引入新的误差。

3.3 剔除法

剔除法与删除法类似，但更侧重于对异常数据的细致分析和处理。在剔除法中，不仅识别并删除异常数据，还会对删除后的数据进行必要的填补或调整，以确保数据的完整性和连续性。这种方法适用于异常数据数量较多，且删除后会对整体数据产生较大影响的情况。剔除法的关键在于如何选择合适的填补方法，以及如何调整剩余数据以保持数据的整体一致性。在使用剔除法时，应仔细评估填补方法和调整策略对数据分析结果的影响，并确保最终数据的准确性和可靠性。

3.4 挖掘法

挖掘法是一种基于数据挖掘技术的异常数据修复方法。这种方法通过挖掘数据中的潜在规律和模式，识别并修复异常数据。挖掘法通常包括数据预处理、特征提取、模式识别和异常修复等步骤。在数据预处理阶段，对数据进行清洗和规范化处理；在特征提取阶段，提取数据的关键特征；在模式识别阶段，利用机器学习算法等识别数据的正常模式和异常模式；在异常修复阶段，根据识别出的异常模式，利用数据挖掘技术为异常数据提供合理的替代值。挖掘法的优点在于能够充分利用数据中的信息，提高异常数据修复的准确性和可靠性，挖掘法的实现过程较为复杂，需要较高的计算资源和专业知识。

3.5 统计法

统计法是一种基于统计学原理的异常数据修复技术。这种方法利用统计模型对数据进行分析和预测，识别并修复异常数据。统计法通常包括数据分布分析、统计检验和异常修复等步骤。在数据分布分析阶段，分析数据的分布特征；在统计检验阶段，利用统计检验方法判断数据是否异常；在异常修复阶段，根据统计检验结果和数据的分布特征，为异常数据提供合理的替代值。统计法的优点在于能够充分利用数据的统计特性，提高异常数据修复的准确性和可靠性。统计法的有效性依赖于数据的分布特征和统计模型的准确性。在使用统计法

时，应确保数据的分布特征和统计模型的选择符合实际情况，并避免引入新的误差[3]。

4 异常数据识别与修复技术案例分析

4.1 案例分析一

在某城市的空气质量监测项目中，监测站通过先进的传感器设备，持续采集了包括PM2.5、PM10、二氧化硫等关键污染物的浓度数据。这些数据对于评估城市空气质量、制定环境保护政策具有重要意义。在数据分析过程中，项目团队发现某监测站在某一时段内的PM2.5浓度数据异常偏高，远超历史平均水平，且与其他监测站的数据存在显著的不一致性。这一异常现象引起了团队的警觉，他们立即启动了异常数据识别程序。经过深入调查，项目团队发现该异常数据是由于监测站周边临时施工活动导致的扬尘污染。施工活动产生的大量尘土被风吹散到空气中，导致监测站采集到的PM2.5浓度数据异常升高。为了修复这一异常数据，项目团队采用替换法。他们首先排除异常数据，然后利用该监测站前后时段的数据，结合相邻监测站的数据进行插值计算，得出了合理的替代值。通过这一修复过程，不仅恢复了数据的准确性和一致性，还确保了后续数据分析的可靠性。

4.2 案例分析二

在某河流的水质监测项目中，监测点通过定期采集水样，对水温、溶解氧、pH值等多项水质参数进行了全面监测。这些水质参数对于评估河流健康状况、制定水环境保护措施至关重要。在数据分析过程中，项目团队发现某监测点的溶解氧数据在某一时段内突然下降，且与其他监测点的数据存在显著差异。这一异常现象引起团队的关注，他们迅速展开异常数据检测工作[4]。经过现场调查，项目团队确认该异常数据是由于监测点附近的水生植物大量死亡导致的溶解氧下降。水生植物的死亡导致水体中的氧气消耗增加，而氧气补充不足，从而导致溶解氧数据异常下降。为了修复这一异常数据，项目团队采用插值法。他们根据该监测点前后时段的数据，以及相邻监测点的数据，构建溶解氧的时空分布模型。然后，利用该模型为异常数据提供合理的替代值。

4.3 案例分析三

在某地区的遥感环境监测项目中，项目团队利用卫星遥感数据对地表温度、植被覆盖等关键环境参数进行了全面监测。这些数据对于评估地区生态环境状况、制定环境保护规划具有重要意义。在数据分析过程中，项目团队发现某区域的地表温度数据异常偏高，与周围区域存在显著差异。这一异常现象引起了团队的重视，他们立即启动异常数据识别程序。经过仔细分析，项目团队确认该异常数据是由于卫星传感器在该时段内受到云层遮挡导致的。云层遮挡导致传感器无法准确获取地表温度信息，从而产生了异常数据。为了修复这一异常数据，项目团队采用同化修复技术。他们结合地面观测数据和相邻时段的遥感数据，构建了地表温度的时空同化模型。利用该模型对异常数据进行同化修复。通过同化修复，不仅恢复异常数据的准确性，还提高整个数据集的一致性和可靠性。这一案例表明，在遥感环境监测中，同化修复技术是一种有效的异常数据修复方法，能够显著提高遥感数据的应用价值。

结束语

环境监测数据的准确性和可靠性对于环境保护和可持续发展具有重要意义。异常数据的存在不仅会影响数据质量，还可能误导环境管理和决策。因此加强异常数据的识别与修复工作至关重要。本文介绍的识别方法和修复技术为环境监测数据的处理和分析提供了新的思路和方法。未来，随着技术的不断进步，相信会有更多高效、智能的异常数据识别与修复方法涌现，为环境监测事业提供更加精准、可靠的技术支持。

参考文献

[1]景永志,艾自东,田相臣.环境监测中异常数据识别与修复[J].环境工程技术学报,2024,14(3):1098-1104.DOI:10.12153/j.issn.1674-991X.20230717.

[2]陆秋琴,魏巍,黄光球.环境监测系统中异常数据的识别和修复方法[J].安全与环境学报,2021,21(3):1300-1310.DOI:10.13637/j.issn.1009-6094.2020.0529.

[3]尹卫萍,侯鹏,徐亮,宋兴伟,胡玲.环境监测数据异常值的判定及处置[J].环境监控与预警,2020,12(06):63-66.

[4]吕建猛.提高环境监测数据质量的对策研究[J].河南建材,2021(06):108-109.