周 强.农业微气象观测数据清洗和质控技术研究[J].湖北农业科学,2020,59(14):37-40,51.

农业微气象观测数据清洗和质控技术研究

周 强
(山东省气象服务中心,济南 250031)

摘要:基于农田特有气象观测设备和环境属性,建立农业气象数据清洗标准和质控方法,以提升农业气象观测数据质量。针对数据属性异常和重复记录情形,选取Bohn数据清洗模型的空缺值清洗方法和噪声数据清洗方法。通过农业微气象观测站点空间内观测要素历史数据统计,获取清洁数据指标,应用于数据质量动态阈值生成方法,建立农业微气象数据质量控制模型。清洗质控后的数据评估指标表明,经过数据清洗和质控模型后数据准确率和重复性均有明显改善。数据清洗质控方法有助于准确获取农业气象灾害监测信息,为农业的防灾减灾提供有效决策支撑。

关键词:农业微气象;数据质控;Bohn数据清洗模型
中图分类号:P49;TP274 **文献标识码**:A
文章编号:0439-8114(2020)14-0037-04
DOI:10.14088/j.cnki.issn0439-8114.2020.14.006 **开放科学(资源服务)标识码(OSID)**:

Study on cleaning and quality control technology of agricultural micro-meteorological observation data

ZHOU Qiang
(Shandong Meteorological Service Center, Jinan 250031, China)

Abstract: In order to improve the quality of agrometeorological observation data, the cleaning standard and quality control method of agrometeorological data are established based on the unique meteorological observation equipment and environmental attributes of farmland. For the case of abnormal data attributes and repeated records, the method of cleaning the blank value of Bohn data cleaning model and the method of cleaning the noise data are selected. Through the historical data statistics of observation elements in the space of agricultural micro meteorological observation station, the clean data index is obtained and applied to the dynamic threshold generation method of data quality, and the quality control model of agricultural micro meteorological data is established. The data evaluation indexes after cleaning and quality control showed that the accuracy and repeatability of the data are significantly improved after data cleaning and quality control model. The data cleaning quality control method is helpful to obtain the monitoring information of agrometeorological disaster accurately and provide effective decision support for agricultural disaster prevention and reduction.

Key words: agromicro meteorology; data quality control; Bohn data cleaning model

农田气象信息是农业生产管理的重要参考依据,随着物联网监测技术的迅速发展,农业设施微型气象观测站点已大规模布设。数据质量问题伴随农业气象观测数据的急剧增长而日益凸显,从而促使了数据清洗技术在农业气象数据方面的应用。

国内对数据清洗技术的研究还处于初步阶段,通常是在统计回归方法中验证数据进行一些基础研究。于力超等[1]基于关联规则的缺失数据插补和最近邻插补方法,利用挖掘得到的关联规则提升度和支持度乘积的倒数作为权重,解决了最近距离样本单元产生不同插补值的问题。戴明锋等[2]在分析数据缺失机制前提下,通过二分类 Logistic 回归插补法,根据发生概率大小确定插补值。刘燕[3]选取近邻择优补差法继承 Logistic 回归插补法的高精确度和最近邻插补法的单元择优性,通过模拟比较多种回归插补方法发现,基于回归的近邻择优插补法可

收稿日期:2020-05-12
作者简介:周 强(1987-),男,山东潍坊人,工程师,硕士,主要从事应用气象和行业服务,(电话)15666976635(电子信箱)Mars-zq@163.com。

以获得更好的插补效果。

随着气象部门观测手段自动化和数据传输速度持续的提高，在地面自动站观测资料质量控制技术方面也积累了一定的经验[4-6]。肖心园等[7]针对不同异常数据提出了基于 3 次样条插值和皮尔逊相关的光伏数据清洗方法，可以得到更优化的数据利用率和重构正确率。潘腾辉等[8]提出了一种 ETL 与数据清洗结合的分布式数据集成工具，将数据清理的技术引入到 ETL 中，基于统计聚类方法和关联规则的数据清洗算法，清洗数据信息的框架。

气象数据质量控制方法多通过阈值和一致性检验完成，但结合农业特定应用领域，需要用农业和气象并存的属性规则判定。本研究选取符合农业气象特性的数据清洗和质控方法，建立农业微气象数据质控流程，检测并剔除数据文件中所有明显的错误和不一致，同时对比和合并相似重复记录，以期及时高效地为用户提供可靠的农田气象观测信息，提升农业生产效率。

1 数据清洗质控技术介绍

1.1 数据清洗技术

数据清洗目的在于删除重复信息、纠正存在的错误，并提供数据一致性[9-11]。数据清洗的主要内容如图 1 所示，依据数据源种类不同，解决数据属性、完整性和惟一性等方面的问题。

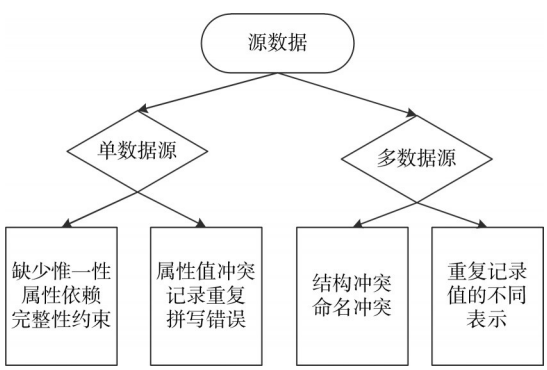


图 1 数据清洗内容

从数据清洗方法上进行分类，结果如图 2 所示。数据清洗原理通常是指利用数理统计、数据挖掘或预定义的清理规则将脏数据转化为满足数据质量要求的数据。

以目前被普遍采用的 Bohn 数据清洗模型为例，首先对源数据进行数据检查，通过统计分析的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归方程的值，利用常识性规则和业务特定规则等简单规则库检查数据值，并使用不同属性间的约束、外部的数据来检测和清理数据。通过聚类分析方法分析数据词法，明确各个字段内不同要素的连贯性，同时确保所有数据字段与已知清单匹配。最后判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并或清除为一条记录。

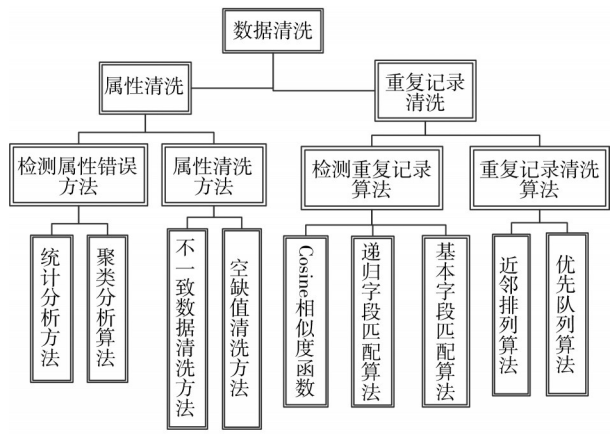


图 2 数据清洗方法分类

1.2 数据质控技术

气象领域对数据质量控制方法有其特殊的规范和要求，主要是要求数据符合天气学、气候学原理，以气象要素的时间、空间变化规律和各要素间相互联系的规律为线索，分析气象资料是否合理[12-15]。常用的处理方法如下。

1)台站气候极值检查。极值是指某个固定测站历史记录中某要素曾出现过的最大值(最小值)，气象资料要素值是否超出极值的检查为极值检查。判断资料的基础是进一步核实超出对应观测站点要素极值的观测资料。

2)时间一致性检查。利用气象要素随时间变化的规律，对气象资料变化进行时间一致性的检查，各要素资料不能超出一定时间内的变化范围，超出的资料为可疑资料。

3)空间一致性检查。根据气象参数具有一定的空间分布特点而进行的检查。通常采用空间回归检验法进行空间一致性检查，其有效性取决于观测站网的密度和被检参数与空间的相关程度[16-19]。

将逐日的观测站要素数据与被检站周边站点相关系数进行显著性检验，找出相关性最好的 5 个站，被检测观测要素与 5 个相关站逐一建立一元线性回归方程。

$$\hat{x}_{i,j} = a_j + b_j y_{i,j} \quad (1)$$

式中，$y_{i,j}$ 为第 j 个初步参考站第 i 日要素实测值，为被检站第 i 日要素估计值。

最后，计算被检站全月要素观测值与各回归方程估计值间的均方根偏差(s^2)。

$$s_j^2 = \frac{1}{m-2}\sum_{i=1}^{m}\left(x_i - \hat{x}_{i,j}\right)^2 \quad (2)$$

式中,x_i为被检站第 i 日的实测值;m 为全月日数。

分别计算被检站被检要素第 i 日加权估计值 x'_i 及要素估计值的加权标准差(s')。

$$x'_i = \sqrt{\sum_{j=1}^{n}\hat{x}_{i,j}^2 s_j^{-2}} \Big/ \sqrt{\sum_{j=1}^{n}s_j^{-2}} \quad (3)$$

$$s' = \sqrt{n} \Big/ \sqrt{\sum_{j=1}^{n}s_j^{-2}} \quad (4)$$

式中,j 为第 j 个最终参考站;n 为最终参考站的总数,在这里 n=5。

当$\left|x_i - x'_i\right| > f'_s$时,表示被检站第 i 日的实测值 x_i 未通过空间一致性检查。f'_s 为控制系数,取值范围为 3.0~5.0。

2　农业微气象数据质控方法

本研究中的数据治理方法主要分为数据清洗和质量控制两方面。首先根据农业微型气象观测站设备特性,建立适用于数据清洗流程的农业气象数据属性标准。针对数据属性异常和重复记录情形,选取高效的辨识算法以及相应的空缺值清洗方法和噪声数据清洗方法。基于农业微气象观测站点空间内观测要素历史数据,应用数据质量动态阈值生成方法,建立气象数据质量控制模型。

2.1　基于 Bohn 的数据清洗模型

对于大多数农业气象观测数据来说,数据格式较为固定,常规数据或者特定数据都是进行专门的定义,比如气温为连续数字,日照可以用 0、1 表示,但对于挖掘或者提取到的数据来说,字段的类型格式、长度及语义都可能存在差异,这就需要对数据清洗重新设定规范格式。

基于 Bohn 模型建立的数据清洗流程如图 3 所示。按照数据清洗需求建立农业气象数据标准,采用关联规则方法中效率较高的 FP-树频集算法辨识数据属性质量。基于空缺值清洗方法和噪声数据清洗方法,将判断出的异常属性数据进行筛除分离;通过递归字段匹配算法,在适当的位置使用间隙,允许不匹配字符的缺失,识别字符串缩写的情形,检测出标识同一个数据实体的重复记录。最后利用多趟近邻排序法,将数据库中的记录排序,比较邻近记录,来判识排除重复记录。

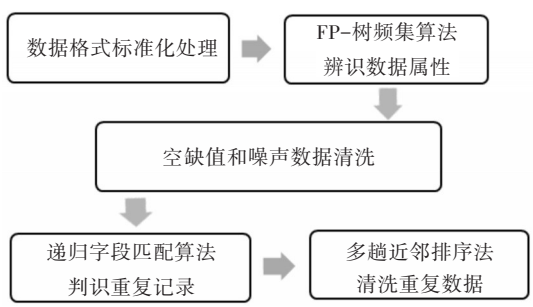


图 3　数据清洗流程

2.2　农业微气象数据的质量控制模型

借鉴气象观测数据质量控制方法,建立针对微气象数据的涵盖阈值、时空一致性以及要素一致性等标准检查的质量控制模型(图 4)。模型重点包括基于站点回归模型的动态阈值生成技术,开展基于动态质控阈值标准的微气象时空一致性检验;基于空间回归方法的空间一致性检验,通过异构异源观测数据辅助的要素一致性检验。

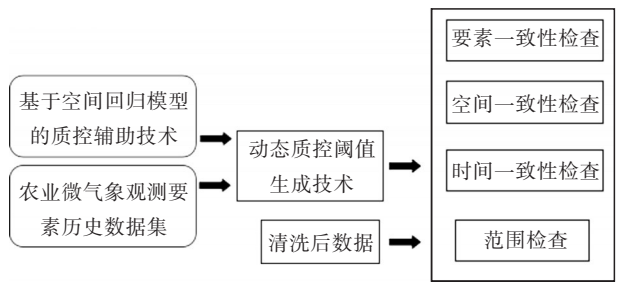


图 4　数据质量控制流程

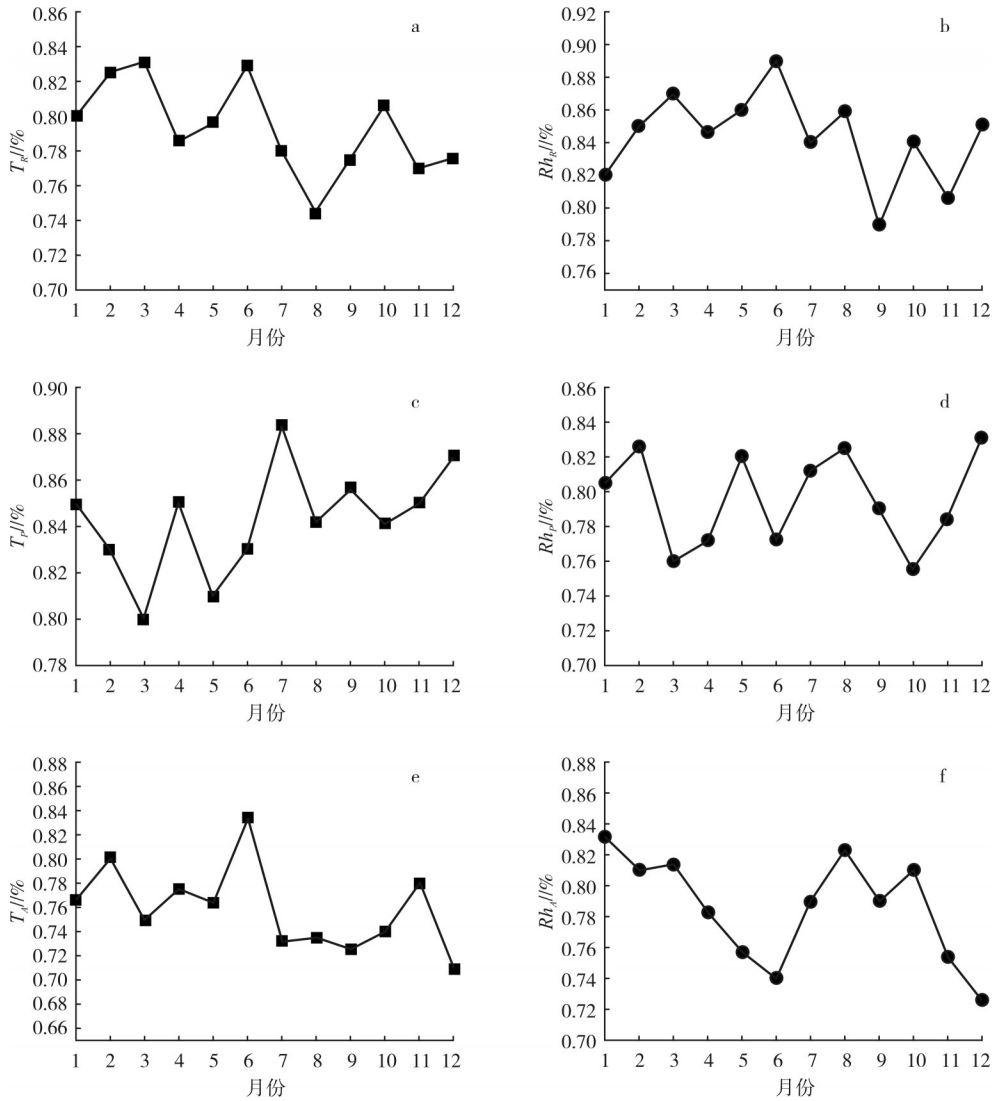
3　农业微气象数据质控模型评估

为评估上述数据清洗和质控方法的效果,引入查准率、精确度和查重率 3 项指标分别检测数据样本。选取 10 个具有订正站的农田小气候气象观测站点,分别以使用率较高的气温和相对湿度要素为例,利用 2019 年全年逐小时的观测数据作为整体样本评估数据。

以订正站数据为标准,将样本数据划分为真实正确样本(*TP*)、真实错误样本(*FP*)、清洗正确样本(*TN*)、清洗错误样本(*FN*)4 种情形,令 *TP*、*FP*、*TN*、*FN* 分别表示其对应的样本数,则本次被清洗数据总数=*TP*+*FN*,识别样本总数=*TP*+*FP*+*TN*+*FN*。

查准率 *P*=*TP*/(*TP*+*FP*)表示为正确数据占清洗后真实总样本的比率。精确度 *A*=(*TP*+*TN*)/(*TP*+*FN*+*FP*+*TN*)则是清洗质控后正确的样本数占样本总数的比例。查全率 *R*=*TP*/(*TP*+*FN*)是正确识别样本和被清洗数据总数的百分比。

选取气温和相对湿度两类气象要素,分别计算其评估指数的逐月变化情况,结果如图 5 所示。从清洗质控后的结果来看,不同月份的数据质量存在一定差异,其中两类要素的查全率和查准率都在 80% 左右,其中相对湿度的查全识别效果较好,而



a、c、e分别为气温的查全率(T_R)、查准率(T_P)、精确度(T_A);b、d、f分别为相对湿度的查全率(Rh_R)、查准率(Rh_P)、精确度(Rh_A)

图5 气温和相对湿度在不同评估指数下的逐月变化分析

温度的逐月查准率均优于相对湿度;气温和相对湿度的最低精确度分别是71.0%和72.6%。经过数据清洗和质控模型后数据准确率和重复性均有明显改善,农业微气象数据清洗质控方法可以有效提升观测数据质量。

5 小结与讨论

本研究将农业系统特有气象观测数据与气象行业传统监测数据深度融合,建立农业微气象数据属性标准,采用FP-树频集和多趟近邻排序等算法,实现清洗模型和质量控制模型在农业微气象数据治理中的应用。

基于回归模型的农业微气象观测历史数据集以及动态检测阈值的生成技术,建立了可以实现异源异构观测数据辅助的要素一致性检验的农业微气象数据质量控制模型。评估表明数据清洗质控方法可以准确获取农业气象灾害监测信息,科学防治农业气象灾害,为农业的防灾减灾、应急决策提供有效的支持服务和技术手段,为实现农业生产的安全、优质、高效运行发挥积极作用。

参考文献:

[1] 于力超,金勇进,王 俊. 缺失数据插补方法探讨——基于最近邻插补法和关联规则法[J]. 统计与信息论坛,2015,30(1):35-40.

[2] 戴明锋,金勇进,查奇芬,等. 二分类Logistic回归插补法及其应用[J]. 数学的实践与认识,2013, 43(21):162-167.

[3] 刘 燕. 基于Logistic回归的近邻择优插补法[D]. 天津:天津财经大学,2013.

[4] 俞荣华,田增平,周傲英. 一种检测多语言文本相似重复记录的综合方法[J]. 计算机科学,2002, 29(1):118-121.

[5] 赵一凡,卞 良,丛 昕. 数据清洗方法研究综述[J]. 软件导刊,2017,16(12):222-224.

[6] OTHMAN L B,YAHIA S B. GBARMVC: Generic basis of association rules based approach for missing values completion[J]. International journal of computing & information sciences,2011,9(1):16-22.

(下转第51页)

来作物生长季气温维持在较高水平,与黑龙江省变化趋势一致[12]。

3)积温呈显著上升趋势,以每 10 年 70 (℃·d)的幅度缓慢上升。1992—1999 年维持在较低水平,自 2000 年以来积温维持在较高水平。从年代际变化分析来看,20 世纪 60、70 年代维持在较低水平,20 世纪 80 年代开始增加,2000 年以来处于较高水平。总体呈现波动上升趋势,近 30 年增加明显,1993 年是积温突变年份。积温上升使热量资源表现出总体增加趋势[13]。

4)降水量呈现不显著下降趋势。小波分析生长季各月及总降水量表现出 4~6 年的高频周期变化,长周期方面表现出 16~22 年的变化周期,近 10 年平均降水量较 60 年平均呈增加趋势。自 2000 年开始降水量变化波动较前 40 年明显增大,变化趋势与黑龙江省变化趋势不同[10]。

5)无霜期日数呈现周期变化。1959—1990 年表现出准 6 年的高频变化周期、准 16 年的低频变化周期,自 1990 年开始高频变化周期变长,为准 8 年,低频变化周期开始变长,为 22 年。初霜日变化趋势是逐年延后,终霜日变化趋势是逐年提前,无霜期日数总体呈现显著增加趋势,目前以每 10 年 3.4 d 的趋势增加,霜冻灾害逐年呈减少趋势。

农业气候资源的变化对农业的影响是综合的,气温升高及积温上升,总体增加了北安市的农业热量资源,从而使北安市的积温带向北移动,也使作物种植期提前,生长期缩短。无霜期与积温的变化使北安市的主要作物总产量潜力增加。降水变化影响光照的变化,它们共同作用对作物生产潜力的影响将大于热量资源的变化,且其影响具有不确定性[14-16]。本研究将为评估北安市气候变化对农业生产的影响、调整种植结构及农产品布局建议及措施的提出提供理论参考。

参考文献:

[1] 秦大河. 进入21世纪的气候变化科学——气候变化的事实、影响与对策[J]. 科技导报,2004(7):4-7.

[2] 李　阔,何霄嘉,许吟隆,等. 中国适应气候变化技术分类研究[J]. 中国人口·资源与环境,2016,26(2):18-26.

[3] 赵　锦,杨晓光,刘志娟,等. 全球气候变暖对中国种植制度的可能影响X. 气候变化对东北三省春玉米气候适宜性的影响 [J]. 中国农业科学,2014,47(16):3143-3156.

[4] 赵俊芳,穆　佳,郭建平. 近50年东北地区≥10 ℃农业热量资源对气候变化的响应[J]. 自然灾害学报,2015,24(3):190-198.

[5] 徐文龙,王丽新,李西磊,等. 北安市近60年气候变化特征分析[J]. 农村经济与科技,2018,29(19):52-53.

[6] 徐文龙,南极月,冷宏杰,等. 北安气候变化特征分析及其对农业生产的影响[J]. 安徽农业科学,2010,38(15):8063-8064.

[7] 邱海军,曹明明,曾　彬. 基于小波分析的西安降水时间序列的变化特征[J]. 中国农业气象,2011,32(1):23-27.

[8] 王　颖,施　能,顾骏强,等. 中国雨日的气候变化[J]. 大气科学,2006,30(1):162-170.

[9] 魏凤英. 现代气候统计诊断与预测技术[M]. 北京:北京气象出版社,1999.

[10] 方丽娟,陈　莉,覃　雪,等. 近50年黑龙江省作物生长季农业气候资源的变化分析[J]. 中国农业气象,2012,33(3):340-347.

[11] 赵　东,罗　勇,高　歌,等. 1961年至2007年中国日照的演变及其关键气候特征[J]. 资源科学,2010,32(4):701-711.

[12] 鄢　波,夏自强,黄　峰,等. 黑龙江流域气温突变诊断及极值重现期分析[J]. 水电能源科学,2016,34(10):5-8.

[13] 胡　琦,潘学标,邵长秀,等. 1961—2010年中国农业热量资源分布和变化特征[J]. 中国农业气象,2014,35(2):119-127.

[14] 潘华盛,徐南平,张桂华. 气候变暖对黑龙江省农作物结构调整影响及未来50年农业情景对策[J]. 黑龙江气象,2004(1):13-15,27.

[15] 王石立,庄立伟,王馥棠. 近20年气候变暖对东北农业生产水热条件影响的研究[J]. 应用气象学报,2003, 14(2):152-164.

[16] 刘志娟,杨晓光,王文峰,等. 气候变化背景下我国东北三省农业气候资源变化特征[J]. 应用生态学报,2009,20(9):2199-2206.

(上接第 40 页)

[7] 肖心园,江　冰,任其文,等. 基于插值法和皮尔逊相关的光伏数据清洗[J]. 信息技术,2019(5):19-22,28.

[8] 潘腾辉,林金城,郑细烨,等. 面向数据库清洗的数据质量控制设计[J]. 信息技术,2017(10):133-136.

[9] 李昌华,卜亮亮,刘　欣. 基于聚类和神经网络对建筑节能气候数据清洗的算法[J]. 计算机应用,2018,38(S1):83-86,111.

[10] 窦以文,屈玉贵,陶士伟,等. 北京自动气象站实时数据质量控制应用[J]. 气象,2008,34(8):77-81.

[11] SHAFER M A,FIEBRICH C A,ARNDT D S,et al. Quality assurance procedures in the oklahoma mesonetwork[J]. Journal of atmospheric & oceanic technology,2000,17(4):474-494.

[12] 陈奕隆. 美国自动地面观测系统[J]. 气象科技,1994(3):48-54.

[13] 廖　捷,周自江. 全球常规气象观测资料质量控制研究进展与展望[J]. 气象科技进展,2018,8(1):56-62.

[14] 任芝花,张志富,孙　超,等. 全国自动气象站实时观测资料三级质量控制系统研制[J]. 气象,2015, 41(10):1268-1277.

[15] 韩海涛,李仲龙. 地面实时气象数据质量控制方法研究进展[J]. 干旱气象,2012,30(2):261-265.

[16] JEFFERY S R,ALONSO G,FRANKLIN M J,et al. Declarative support for sensor data cleaning[A]. Proceedings of 4th international conference on pervasive computing [C]. Springer, New York, 2006.83-100.

[17] GILL S, LEE B. A framework for distributed cleaning of data streams[J]. Procedia computer science,2015,52(1):1186-1191.

[18] 李良富,王汉杰,刘金玉,等. 基于黑板模型的地面气象数据质量控制[J]. 气象科技,2006,34(2):199-204.

[19] 范文波. 地面气象观测数据综合质量控制方法研究与实现[D]. 南京:南京信息工程大学,2016.