陈东辉，熊安元，唐为安. 气象灾害风险普查数据质量控制技术研究与应用[J]. 灾害学，2022，37(4)：135 - 142.
[CHEN Donghui，XIONG Anyuan，TANG Weian. Development of Quality Control Methods for Meteorological Disaster Risk Survey Data of China[J]. Journal of Catastrophology，2022，37(4)：135 - 142. doi：10. 3969/j. issn. 1000 - 811X. 2022. 04. 022.]

气象灾害风险普查数据质量控制技术研究与应用*

陈东辉[1]，熊安元[1]，唐为安[2]

（1. 国家气象信息中心，北京 100081；2. 安徽省气候中心，安徽 合肥 230031）

摘 要：全国第一次自然灾害综合风险普查是一项重大的国情国力调查，是提升自然灾害防治能力的基础性工作。普查数据质量是风险普查的“生命线”，直接决定普查工作的质量。全国气象灾害综合风险普查调查对象由传统的“数据表格”向“多维灾体”转变，数据质量控制不仅只是单点极值阈值的质控，还需考虑整个灾害事件致灾因子选取的准确性。针对数据本身属性以及站点空间比对属性异常和重复记录等问题，通过质检规则、管理约束、质量核查分析以及评估结果验证等，建立动态数据质控方法。该方法包括 11 个质控规则 136 个细化项，经过数据质控后，气象致灾因子的填报准确率明显提升，数据重复率显著下降，数据上报完成率和数据质量形式审核通过率均达 100%，为评估区划和未来普查成果高效应用提供有力的基础数据支撑。

关键词：气象灾害；风险普查；数据质量控制；数据审核；质检规则；危险性评估
中图分类号：X43；X915. 5；P429　**文献标志码**：A　**文章编号**：1000 - 811X(2022)04 - 0135 - 08
doi：10. 3969/j. issn. 1000 - 811X. 2022. 04. 022

全国气象灾害综合风险普查是自然灾害综合风险普查的重要内容，对于认识和把握气象灾害发生发展规律、建立综合防灾减灾救灾体系、实现灾害风险管理等具有重要意义[1]。为贯彻落实《国务院办公厅关于开展第一次全国自然灾害综合风险普查的通知(国办发〔2020〕12 号)》精神，全面做好全国气象灾害综合风险普查工作，中国气象局结合气象部门实际，制定并印发《第一次全国自然灾害综合风险普查总体方案》(以下简称总体方案)[2]。总体方案确立了以调查为基础、评估为支撑，客观认识当前全国和各地区主要气象灾害的风险水平，科学预判气象灾害风险变化趋势和特点，形成全国气象灾害风险区划的总体工作思路。

我国周期性普查工作主要有全国人口普查、农业普查、经济普查，各类调查统计数据是普查工作的主要成果，成果能够推广使用的根本是保证数据质量[3-6]。高质量的统计数据的产生依赖于对数据质量的评估[7-8]。国内外权威组织和学者针对数据质量评估方法进行了研究，形成了一系列成果[9-10]。但采用的方法或多或少存在一些局限性，比如有的方法能较大程度地检查出逻辑性错误，但却无法保证数据的准确性；有的为汇总阶段的事后质量评估，而不适合对收集阶段的数据进行质检[11-13]。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。肖心园等[14]针对不同异常数据提出了基于 3 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[15]提出了一种 ETL(Extract - Transform - Load，抽取 - 转换 - 加载)与数据清洗相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。

2020 年开始的全国范围内的分灾种、分区域、长时间序列的气象灾害综合风险普查工作在我国尚属首次。此次灾害调查主要采取自上而下和自下而上两种方法开展，其中在技术层面上采取自上而下形式，国家级和省级技术组承担普查技术规范制定、调查表格设计、普查数据采集信息系统研发等；在实际操作层面上采取自下而上形式，县(区)级根据技术规范和调查表格填报表格，并逐级审核、上报至普查信息收集系统。通过此次普查共收集暴雨、干旱、台风等 10 种气象灾害记录条数 6 245 225 条，其中以低温灾种记录条数最多，达 2 029 730 条，其次为雷电灾种，记录条数为 1 301 675 条，雪灾记录条数相对较少，全国范

* 收稿日期：2022 - 06 - 06　　修回日期：2022 - 08 - 16
基金项目：国家重点研发计划项目(2019YFA0606904)
第一作者简介：陈东辉(1984 -)，男，汉族，河南南阳人，博士，高级工程师，主要从事气象数据分析处理、数据挖掘算法研究. E - mail：chendonghui@ cma. cn
通信作者：唐为安(1980 -)，男，汉族，江苏阜宁人，博士，高级工程师，主要从事气象灾害风险评估与区划、气候变化及其影响评估研究. E - mail：twa1980@ mail. ustc. edu. cn

围内共收集 141 827 条。高质量数据是普查工作顺利展开的前提，数据质量控制技术水平的高低则是确保普查数据质量的根本，也直接决定了气象灾害致灾危险性评估与区划及综合风险评估与区划结果的质量。气象数据质量控制方法多通过阈值和一致性检验，但对于此次气象灾害风险普查工作的复杂性和致灾因子调查的不确定性，需要结合气象灾害事件客观化识别和空间化验证的属性规则来综合判定。本文拟从数据质控方法、数据质检规则、管理流程、质量核查分析以及评估结果验证等方面来阐述全国气象灾害风险普查数据质量控制技术，其中通过系统质检建立质检规则库保障“事前”，管理流程约束保障“事中”，可疑数据核查分析和致灾危险性评估结果验证来评估“事后”，构建全过程数据质量闭环，从而最大限度保障气象灾害普查数据质量，为成果高效应用提供有力支撑。

1 数据和方法

1.1 数据

根据我国气象灾害种类的分布、影响程度和特征，本次全国气象灾害风险普查的气象灾害包括暴雨、干旱、台风、高温、低温、大风、冰雹、雪灾、雷电、沙尘暴等 10 种。通过调查和科学分析，获取的国、省、市、县 10 类气象灾害致灾因子数据，即以县(区)级行政区为基本单元，开展全国气象灾害的特征调查和致灾孕灾要素分析而获取我国主要气象灾害的致灾因子信息，覆盖空间范围为全国各省、直辖市、自治区和新疆生产建设兵团(不含香港特别行政区、澳门特别行政区和台湾省)，时间范围为 1978 — 2020 年近 40 年数据。

1.2 质控方法研究

借鉴气象观测数据质控方法，根据气象领域对数据质量控制方法的特殊规范和要求，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，分析数据是否合理[16-19]。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。其次建立针对普查数据的涵盖阈值并融合要素一致性以及空间一致性等质控方法。具体处理方法如下。

(1)要素一致性。对某个气象测站历史记录中某观测要素结合气象灾害事件发生过程(时间)中曾出现的最大值(最小值)，判断气象资料要素值是否超出极值作为要素一致性检查。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，各要素不能超出一定时间内的变化范围，超出则判为可疑。

(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[20]。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。

$$\hat{x}_{i,j} = a_j + b_j y_{i,j}。 \tag{1}$$

式中：$y_{i,j}$为第 j 个初步参考站第 i 日要素实测值，$\hat{x}_{i,j}$为被检站第 i 日要素估计值，a_j 和 b_j 为回归系数。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_j^2)：

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

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')：

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

$$s' = \sqrt{n} \Big/ \sqrt{\sum_{j=1}^{n} s_j^{-2}}。 \tag{4}$$

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 $|x_i - x'_i| > f_s'$ 时，表示被检站第 i 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 数据质控流程

2.1 质控总体设计

由于涉及灾害种类多、覆盖面广、时间跨度长，因此对调查数据的质控需考虑信息、技术、流程和管理在内的四大因素，从而构建全过程的数据质量控制闭环，总体设计如图 1 所示。

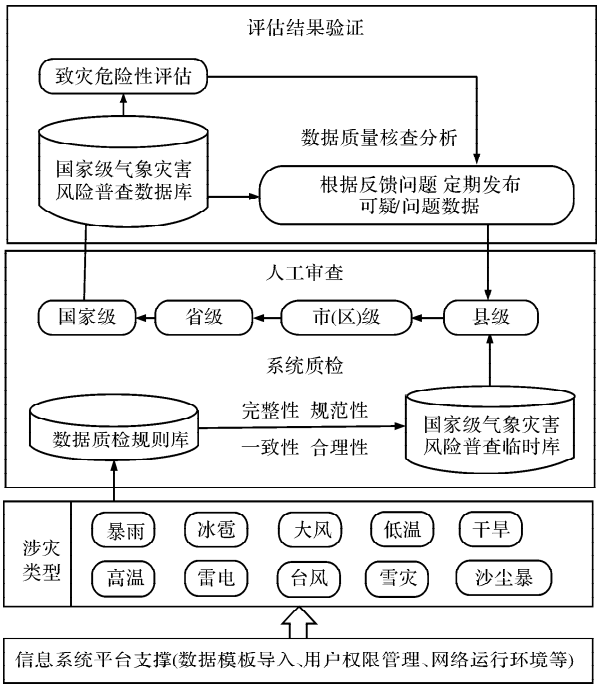


图 1 全国气象灾害调查数据质控总体设计

为满足国、省、市、县四级高效开展气象灾害风险普查工作以及保证数据高一致性，基于“云+端”气象集约化业务布局，建立物理统一、逻辑分布的气象灾害风险普查信息系统平台。基于“云”一级部署形成国家级气象灾害风险普查数据库，提供国、省、市、县四级“端”应用。对上报调查类数据资源采用集中统一管理，使国、省、市、县用户(包括汇交和审查用户)共同操作“一套数据”，避免因分布式架构需要数据频繁同步导致数据不一致。

为确保本次气象灾害风险普查数据的准确性，除人工审查外，提供通过系统质检功能，提高数据的准确率和格检效率。针对 10 类气象灾害设计的每张调查表的所有核心数据项进行格式检查、逻辑合理性检查、关联性检查等，保障灾害风险普查属性信息的准确。

2.2　数据质检规则建立

调查数据在全国气象灾害综合风险普查技术规范要求基础上，采用行政记录检查法、逻辑规则检验法、局部空间自相关检测法等方法，通过系统平台建立数据质检规则库，对用户信息、行政区划信息、气象台站信息、致灾因子等进行完整性、规范性、一致性、合理性的质量检查，如表 1 所示。

(1)数据完整性。包括调查数据上报完整性和数据本身完整性。重点检查填报指标是否必填、选填、缺省值以及重复值等要求。

(2)数据规范性。分为数据格式规范性和文件格式规范性。数据格式规范性包括填写采集数据类型是否符合要求(如字符型、数值型、整型、浮点型、日期型、日期时间型)，数据长度、精度、选项个数的规范性(如单选、多选、选项个数不超过 XX 个)等；文件格式规范性包括上传文件是否符合格式要求等。

(3)数据一致性。分为逻辑一致性、属性一致性、时空一致性。逻辑一致性包括填报致灾因子间逻辑关系约束、致灾因子间逻辑关系等；属性一致性包括致灾因子的量纲一致性等；时空一致性包括填报经纬度是否在本级行政区范围内等。

(4)数据合理性。分为值域合理性、异常值合理性。值域合理性包括致灾因子是否在值域范围内等；异常值合理性包括填报数据的边界范围控制。

基于数据质检规则，以雪灾为例，数据校验方法和对应质检指标项如表 2 所示。

表 1　气象灾害普查系统数据质量检查规则

质检规则	质检指标项	指标说明
A 完整性	A. 1 完整性	检查填报指标是否必填，气象普查系统从设计上除了管理字段选项(如填表人、复核人、审查人信息)可选外，其他数据填报项均为必填，从而减少定义选填就不填。
	A. 2 缺省值	在调查数据中，针对客观存在的无观测、无记录以及上述完整性里提到必填但无法填或短时间不好填的情况，统一用“999999”标识。也是 2. 4 节“数据核查分析”功能设计的核心，纳入数据问题处置闭环机制。
	A. 3 重复值	气象调查数据主要基于辖区内自动站观测数据作为其计算致灾因子的根据，因此针对辖区内可能存在站点重复填报，采取时间、站号、致灾因子值完全匹配则判定为重复记录。考虑到实际操作中对重复值实时质检会严重影响数据填报效率，此检查采用非实时检查，并纳入“数据核查分析”数据问题处置闭环机制。
B 规范性	B. 1 数据格式规范性	包括填报指标数据类型是否符合要求(如字符型、数值型、整型、浮点型、日期型、日期时间型)，数据长度、精度、选项个数的规范性(如单选、多选、选项个数不超过 XX 个)等。
	B. 2 文件格式规范性	包括上传文件是否符合格式要求，主要是基于气象普查系统批量导入数据模板的规范性。
C 一致性	C. 1 逻辑一致性	包括填报指标选项间逻辑关系约束、填报指标间逻辑关系、调查表间逻辑关系等。
	C. 2 时间一致性	包括填报时间与事实一致性、填报时间的范围。
	C. 3 属性一致性	包括填报指标项是否唯一、量纲是否统一等。
	C. 4 空间一致性	包括填报经纬度是否在行政区划范围内、灾害影响范围是否在登记台站信息内或行政区内等。
D 合理性	D. 1 值域合理性	包括填报数据属性值是否在值域范围内。由于气象灾害数据调查本着“边普查、边应用、边发挥效益”的原则，对于值域范围会根据实际情况进行不断优化和调整，也将纳入“数据核查分析”问题处置闭环机制。
	D. 2 异常值合理性	包括填报数据的时空属性是否符合既定时空范围。

表 2 雪灾过程及危险性因子调查数据质检规则

序号	字段描述	数据类型	长度	是否必选	校验方法	对应指标项
1	区域名称	VARCHAR	18	否	验证区域名称是否在行政区划信息表(字典)	A. 1/B. 1/C. 4
2	行政区划代码	VARCHAR	12	否	验证行政区划代码是否在行政区划信息表(字典)，不足 12 位末尾补零	A. 1/B. 1/C. 4/D. 2
3	开始时间	VARCHAR	8	否	年月日：日期范围验证 day 格式为 yyyyMMdd，且开始时间≤结束时间	A. 1/A. 2/B. 1/C. 1/C. 2/D. 2
4	结束时间	VARCHAR	8	否	年月日：日期范围验证 day 格式为 yyyyMMdd，且开始时间≤结束时间	A. 1/A. 2/B. 1/C. 1/C. 2/D. 2
5	站号	VARCHAR	6	否	验证站号是否在台站信息表(字典)	A. 1/A. 3/B. 1/D. 2
6	累积降雪量/mm	NUMERIC	7	否	验证浮点数且精度为小数点后 1 位	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
7	最大积雪深度/cm	INTEGER		否	正整数验证	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
8	积雪日数/d	NUMERIC	7	否	验证浮点数且精度为小数点后 1 位，且进行值域判断：0 < 该项值	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
9	降雪日数/d	NUMERIC	7	否	验证浮点数且精度为小数点后 1 位	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
10	最低气温/℃	NUMERIC	7	否	验证浮点数且精度为小数点后 1 位	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
11	最大日降水量/mm	NUMERIC	7	否	验证浮点数且精度为小数点后 1 位	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
12	日最大风速/(m/s)	NUMERIC	7	否	验证浮点数且精度为小数点后 1 位，且进行值域判断：0 < 该项值 < 70	A. 1/A. 2/A. 3/B. 1/C. 1/D. 1
13	备注	VARCHAR	500	是	判断是否缺省值	A. 2
14	填表人	VARCHAR	10	是	系统匹配填表人、复核人、审查人实名信息以及系统时间进行自动补充，减少填报成本和人为错误。	A. 1
15	复核人	VARCHAR	10	是		
16	审查人	VARCHAR	10	是		
17	联系电话	VARCHAR	11	是		
18	填写单位	VARCHAR	100	是		
19	填写日期	VARCHAR	8	是		

表 3 风险普查数据汇交与审核人员信息登记表

序号	行政区划代码	单位名称	姓名	联系方式(绑定微信的手机号)	用户角色(填汇交或审核)

2.3 管理流程约束

2.3.1 用户认证管理

为加强气象灾害普查数据安全与数据质量管理，启用实名用户证书登录方式。各级气象部门根据普查工作的实际情况，报送本辖区内普查和审核人员的名单信息，具体格式如表 3 所示。其中，单位行政区划代码、单位名称、姓名、联系方式、用户角色为必填项；单位行政区划代码为国普办发布的县级及以上行政区划代码；单位名称为普查人员所在气象部门单位；姓名必须为气象部门证书(实名证书)对应的姓名；联系方式为本人微信绑定手机号，方便后续开通微信推送填报任务提醒服务；用户角色按规定要求填写，包括对应行政区的汇交或审核人员。

通过“证书 + 白名单”双重认证方式实现用户的初始化与权限分配。用户实名认证管理一方面，避免传统使用用户名密码登录方式可能存在账户泄露导致数据风险。另一方面，由于用户实名实现数据责任划分和电子留痕，减少数据胡填乱报，一定程度上保障了数据质量。

2.3.2 人工审核机制

国、省、地(市)、县(区)四级气象部门依据《气象灾害调查与风险评估技术规范》[21]和《气象灾害综合风险普查成果汇交和质量审核管理办法》[22]，对本级或其下级部门线上汇交的数据成果进行人工质量审核。国、省、地(市)三级行业部门，应对下级部门汇交的数据与成果进行质量审核。上级气象部门应及时向下级气象部门反馈质量审核结果，对未通过审核的应要求在规定时限内完成修改更新和再次汇交。

在国家级通过人工审核后，即转入人工抽查阶段。人工抽查工作由国家级和省级分别负责完

成。国家级成立国家级气象灾害普查数据核查组，负责对各省上报的数据进行抽查，抽查数据要求覆盖各省，各省被抽查数据占该省调查对象的比例不低于3%。省级气象部门应成立本省气象灾害普查数据核查组，负责对本省各地(市)、县(区)上报的数据进行抽查，抽查数据应具有地域代表性，抽样比例不低于本省调查对象总数的5%。相比于气象灾害风险普查信息系统平台的数据质检，人工审核和人工抽查不仅需要对填报数据的完整性、规范性、一致性、合理性对进行审核，还应通过气象月报表、气象志、地方志等多源数据来重点核查填报数据的真实性，即某填报数据是否为真实的发生值，而不是由于观测系统错报、观测员错误记录或填报人员错误填报等问题而产生的。

2.4　数据核查分析

在气象灾害普查数据上报过程中，通过以上系统质检和管理约束可基本解决调查数据不规范问题，包括数据格式不对、存在异常值、逻辑性错误等，因为不规范的数据是无法进入普查系统。但是对于可疑/疑似错误的数据，即数据内容本身的科学甄别通过系统质检是不容易发现，并且近千万条上报数据靠人工审核和抽检也会“漏网”。针对以上问题，需要对已上报数据进行“事后”的数据质量核查分析，该工作也应基于系统代替人工去完成。因此在现有数据质检流程上增加数据质量核查分析功能。

数据质量核查分析功能区别于数据质检规则，本质是因为数据核查分析是“事后”，数据质检规则是“事前”，即数据核查是针对已经上报数据的质量筛查，在“事前”考虑当时质检效率和客观填报事实以及当时条件下并未预见的质检规则等(具体见表1的A.2、A.3、D.1内容)。比如，地方在上报数据时存在由于当时未找到相关史料或确实某时段气象观测数据缺测而将其作为缺省值(999999标识)，或者某些数据项由于当时设定阈值过高或过低导致数据已经“入库”需要重新筛查出来让普查填报人员进行再次确认。

从图2可以看出，低温调查表3的“过程累积降水量”为缺省值，因需要普查人员再次确认其数据是否为缺测，如果确实缺测，基于气象灾害普查系统将数据导出后再导入。如果该数据项已经通过重新统计获取过程降水量，则导出结果并更新后再导入系统。通过上述更新导入后，次日将不再作为问题或错误数据发布。基于不断滚动发布最新问题数据，实现问题数据逐一解决完毕，从而保障国家级气象灾害风险普查数据库的数据质量保持最佳。

3　质控效果验证

利用上报且经过质量核查分析的普查数据，以山东省雪灾致灾危险性评估为例，对其上报的普查数据质量以及危险性评估结果图件的合理性、准确性进一步核验。山东省上报1978 — 2020年雪灾调查类数据2 086条，根据《雪灾调查与风险评估技术规范》[21]，以每次过程的累积降雪量、最大积雪深度和降雪日数作为雪灾危险性评估的致灾因子。为了消除各致灾因子量纲可能对评估结果的影响，对各致灾因子进行归一化处理：

$$D_{ij}=0.5+0.5\times\frac{x_{ij}-x_{i\min}}{x_{i\max}-x_{i\min}}。\tag{5}$$

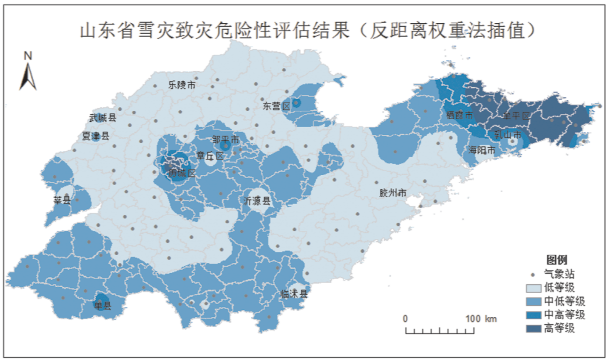
式中：D_{ij}表示第j次雪灾过程的第i个致灾因子的归一化值，以下类同；x_{ij}表示第j次雪灾过程的第i个致灾因子的原始值；$x_{i\min}$表示所有雪灾过程中第i个致灾因子的最小值；$x_{i\max}$表示所有雪灾过程中第个致灾因子的最大值。

在各致灾因子经归一化处理后，采用加权综合，计算得到每次雪灾过程的致灾强度指数。

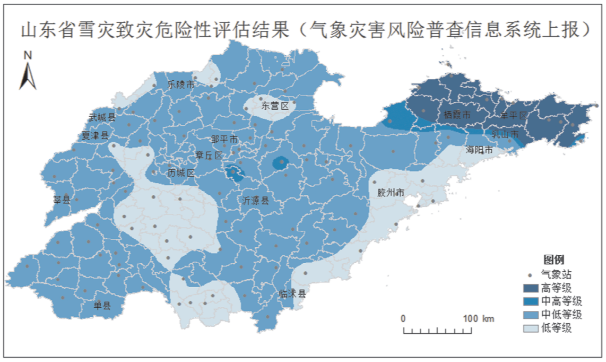
$$V_j=\sum_{i=1}^{3}w_i\times D_{ij}。\tag{6}$$

式中：V_j表示第j次雪灾过程的致灾强度指数；w_i表示第i个致灾因子的权重系数，采用专家打分法确定，取等权重，即均取1/3。

图2　利用质量核查分析定期发布此可疑/疑似错误数据



(a) 反距离权重法插值结果



(b) 气象灾害风险普查信息系统上报结果

图 3 山东省雪灾致灾危险性评估结果

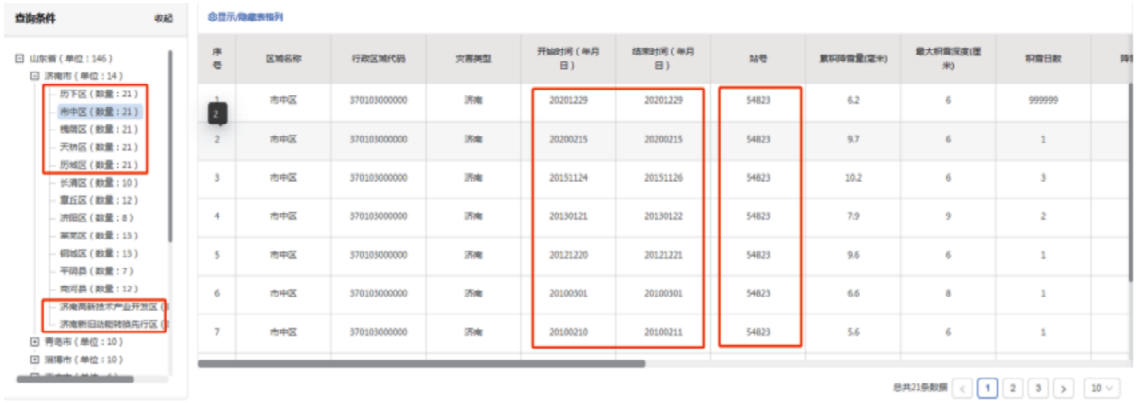


图 4 气象灾害风险普查信息收集系统中 54823 及周边 7 个区填报数据

然后，将山东省 1978 — 2020 年所有雪灾过程的致灾强度指数升序排列，采用百分位数法[23]，分别取 50% 、80% 和 90% 百分位所对应的值，将致灾强度指数划分为弱、较弱、较强、强四个等级，统计每一个国家级气象观测站各等级范围内的致灾强度指数的发生次数。各国家级气象观测站致灾危险性指数由 1 ~ 4 级致灾强度指数的发生次数的归一化值加权综合得到，按强度越强权重系数越大的原则，1 ~ 4 级的权重系数分别取 0. 1、0. 2、0. 3 和 0. 4。基于各国家级气象观测站致灾危险性指数，在 GIS 中采用反距离权重法[24]插值得到了山东省雪灾致灾危险性评估结果(图 3a)。

对比气象灾害风险普查信息收集系统中山东省上报的雪灾致灾危险性评估结果(图 3b)，可以发现：位于济南市历城区附近两者的等级差别较大。尽管插值算法、分类阈值不同可能会造成两者结果在局部地区等级不一致，但是如果在同一区域的分级结果相差两个及以上等级，一般认为是两者使用的数据源不一致造成的。经过对山东省填报数据查询发现，济南市 7 个区上报数据都是使用 54823 这一个国家级气象观测站的数据上报，且上报数据内容一样(图 4)，这可能是导致在 54823 这个站及周边求出的致灾危险性等级偏高的原因。针对 54823 数据重复上报造成评估等级过高的问题，采用删除 54823 重复数据，然后在此基础上重新计算致灾危险性评估结果(图 5)，进一步分析可以发现两者等级分布形态大体一致。

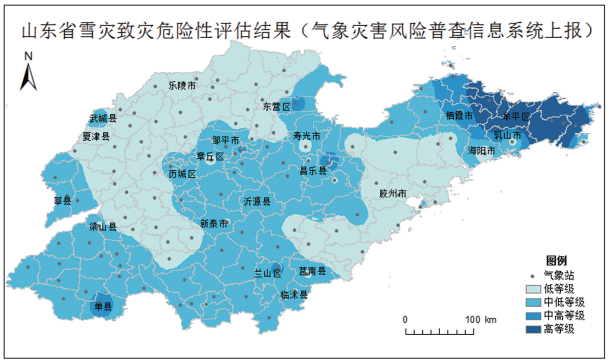


图 5 山东省雪灾致灾危险性评估结果(异常值处理后的结果)

4 讨论和结论

4. 1 讨论

高质量的气象灾害致灾调查数据是气象灾害风险评估、灾害风险管理的基础，也是筑牢气象防灾减灾第一道防线的根本，同时还是科学决策依据的先决性条件。全国第一次气象灾害综合风险普查重点是对 1978—2020 年精细到县(区)级的 10 种气象灾害致灾要素调查，由于涉及范围广、灾害种类多，虽经过省、地(市)、县(区)三级审

核后再上报全国气象灾害风险普查软件系统，但上报的数据质量仍存在一定程度上的不确定性。不确定性可能主要的来自于两个方面：

(1)区级行政单元无气象站点数据填报造成的不确定性，这主要与我国国家级气象观测站点的空间布局有关，我国大多数省份的地市级往往布设一个国家级气象观测站，大多数市辖区没有国家级气象观测站点，而本次普查工作的分辨率至少为县(区)级，这就导致没有国家级气象观测站的市辖区的填报数据要么是缺测，要么是基于该市有国家级气象观测站点的市辖区的气象数据进行填报。通过上文分析可知，数据重复填报在很大程度上会导致致灾危险性评估的不确定性。

(2)异常值判定造成的不确定性，这与气象要素随时间的演变有关。像温度类观测要素，其随时间是正态的、连续渐变的，而降水类的却是偏态的、不连续突变的。如果温度出现异常时，可以通过判断该温度与之前和之后时刻的温度值差异程度来判断该温度是否为异常值，而降水量却不能通过该方法来判定，特别是 20 世纪 90 年代以来，因全球变暖，极端性降水事件的发生愈加频繁，降水异常值出现概率偏大，这些都在一定程度上增加了数据质控的难度和不确定性。

4.2　结论

文中以气象灾害风险普查数据为研究对象，将气象观测数据与气象灾害致灾因子数据深度融合，通过质检规则、管理约束、质量核查分析以及评估结果验证，建立动态数据质控方法，使得数据上报完成率和数据质量审核通过率均达到100%，有力推动全国气象灾害致灾危险性调查任务全面完成，为全国气象灾害风险评估业务应用提供很好支撑。通过本研究获得以下主要结论：

(1)利用常识性规则和业务特定规则等建立数据质检规则库检查数据值，并使用不同属性间的约束、外部的数据来检测和清理数据。

(2)通过用户实名认证管理和人工核查机制既实现数据责任划分和电子留痕，又减少数据胡填乱报，一定程度上保障了数据质量。

(3)运用统计分析等数据核查分析方法识别可能的可疑值或异常值，如偏差分析、识别不遵守分布或回归方程的值，定期发布和反馈问题数据，实现问题数据逐一解决。

(4)利用上报且通过质量核查分析的灾害调查数据，对其上报的致灾危险性评估结果进一步验证质控效果的准确性。

参考文献：

[1]　王国复. 气象灾害调查与风险评估[J]. 城市与减灾，2021(2)：5－9.

[2]　郭娜，林伟，陈红兵，等. 气象灾害风险普查工作制度研究[J]. 农业灾害研究，2021，11(9)：63－64..

[3]　王银辉. 浅谈统计质量和统计安全[J]. 经济视野，2014，3(19)：1.

[4]　张凌宇. 灾害数据质量评估研究[D]. 南昌：南昌大学，2017.

[5]　徐文强. 大数据环境下应急信息质量评估体系研究[D]. 南昌：南昌大学，2019.

[6]　闫爱莲. 浅议统计下数据质量的重要性[J]. 河北煤炭，2009，32(4)：69－70.

[7]　TEE S W，BOWEN P L，DOYLE P，et al. Factors influencing organizations to improve data quality in their information systems[J]. Accounting & Finance，2007，47：335－355.

[8]　BATINI C，CAPPIELLO C，FRANCALANCI C，et al. Methodologies for data quality assessment and improvement[J]. ACM Computing Surveys，2009，41(3)：1－52.

[9]　Geographic information－Data quality：ISO 19157－2013[S]. Switzerland，2013.

[10]　国家统计局. 国家统计质量保证框架(2021)[N]. 中国信息报，2021－06－18(3).

[11]　胡冉冉. 基于人口统计分析的人口普查质量评估研究[D]. 重庆：重庆工商大学，2020.

[12]　陶然. 周期性普查数据质量评估方法与适用性研究[J]. 统计研究，2014，31(8)：66－72.

[13]　耿修林. 普查数据质量的两种检查方法[J]. 中国统计，2006，54(6)：10－11.

[14]　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.

[15]　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，2017，41(10)：133－136.

[16]　陈奕隆. 美国自动地面观测系统[J]. 气象科技，1994，22(3)：48－54.

[17]　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)：56－63.

[18]　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.

[19]　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.

[20]　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.

[21]　中国气象局全国气象灾害综合风险普查工作领导小组办公室. 关于印发气象灾害调查与风险评估技术规范的通知(气普领发〔2021〕4 号)[Z]. 北京：中国气象局，2021.

[22]　中国气象局全国气象灾害综合风险普查工作领导小组办公室. 关于印发成果气象灾害综合风险普查成果汇交和质量审核管理办法的通知(气普领发〔2021〕2 号)[Z]. 北京：中国气象局，2021.

[23]　迟潇潇，尹占娥，王轩，等. 我国极端降水阈值确定方法的对比研究[J]. 灾害学，2015，30(3)：186－190.

[24]　贾悦，崔宁博，魏新平，等. 基于反距离权重法的长江流域参考作物蒸散量算法适用性评价[J]. 农业工程学报，2016，32(6)：130－138.

[25]　熊安元，陈东辉，梁中军，等. 气象灾害综合风险普查数据质量审核工作细则(气普领发〔2021〕5 号)[Z]. 北京：中国气象局全国气象灾害综合风险普查工作领导小组办公室，2021.

Development of Quality Control Methods for Meteorological Disaster Risk Survey Data of China

CHEN Donghui[1], XIONG Anyuan[1], TANG Weian[2]
(1. *National Meteorological Information Center*, *Beijing* 100081, *China*;
2. *Anhui Climate Center*, *Hefei* 230031, *China*)

Abstract: The first national natural disaster survey programme is a great effort of China to collect data regarding natural hazards, exposure, vulnerability, disaster reduction capacity of major disasters, including data for nine types of meteorological disasters. The forms of survey dataset in this programme range from tabular data, GIS data to unstructured data. Data quality is of great importance which directly influences the reliability of risk assessment results. Due to the large volume and complexity of the survey data, it is of great challenge to implement quality control, considering correctness, completeness, consistency, and standardization of data. Besides the capacity of checking ordinary static data, inspection rules shall also be able to check and ensure that extreme events are extracted properly or not with station dataset. We proposed a set of data quality control methods, including automatic quality inspection rules, management operation requirements, quality check analysis, and verification of assessment results. For inspections, we developed 11 quality control rules with 136 detailed items, which improved both the reliability and speed of data quality review. After the quality control, data revision and resubmission, the ratios of completeness and correctness for final data both achieved 100%. The data quality improved and assured through quality control process can provide a consolidate basis for further risk analysis in the future.

Keywords: meteorological disasters; risk survey; data quality control; quality inspection rule; hazard assessment

(上接第 128 页)

[11] 龚会莲, 胡胜强. 公共危机预警策略的选择逻辑与比较分析[J]. 行政论坛, 2019, 26(03): 138 - 144.
[12] 倪永贵, 许峰, 朱国云. 重大突发公共危机预警: 过程、困境及其应对策略——基于信息空间理论视角[J]. 电子政务, 2021(7): 101 - 112.
[13] 李瑞昌. 技术赋能城市综合应急管理的路径[J]. 求索, 2021(3): 118 - 125.
[14] 王文, 张志, 张岩, 等. 自然灾害综合监测预警系统建设研究[J]. 灾害学, 2022, 37(2): 229 - 234.
[15] 黎江平, 姚怡帆, 叶中华. TOE 框架下的省级政务大数据发展水平影响因素与发展路径——基于 fsQCA 实证研究[J]. 情报杂志, 2022, 41(1): 200 - 207.
[16] Jane E FOUNTAIN. Building the virtual state: Information technology and institutional change[M]. Washington D. C: Brookings Institution Press, 2001: 3 - 14.
[17] 黄冬娅. 压力传递与政策执行波动——以 A 省 X 产业政策执行为例[J]. 政治学研究, 2020(6): 104 - 116, 128.
[18] 文宏, 李风山. 组态视角下大气环境政策执行偏差的生成机理与典型模式——基于 61 个案例的模糊集定性比较分析[J]. 中国地质大学学报(社会科学版), 2021, 21(5): 70 - 81.
[19] 王欢明, 陈佳璐. 地方政府治理体系对 PPP 落地率的影响研究——基于中国省级政府的模糊集定性比较分析[J]. 公共管理与政策评论, 2021, 10(1): 115 - 126.

Efficiency Analysis of Disaster Public Early Warning Based on Qualitative Comparative Analysis of Fuzzy Sets

WANG Guoqiao, ZHAO Lemeng, LI Yaoyuan
(*School of Public Management*, *Northwest University*, *Xi'an* 710127, *China*)

Abstract: By selecting 40 flood disaster cases from 2014 to 2021, using the fuzzy set qualitative comparative analysis method, and based on the TOE analysis framework and existing literature, six conditional variables that affect the efficiency of public disaster early warning are summarized from the three levels of technology, organization and environment, including risk monitoring facilities, forecast release and reception facilities, attention intensity, plan perfection, organizational resource endowment, and public participation. The combination of factors that affect the efficiency of disaster public early warning is empirically analyzed. The research shews that: ① Single condition does not constitute a necessary condition for the efficiency of public disaster early warning, but improving the perfection of the plan plays a more universal role in the efficiency of public disaster early warning; ② High efficiency disaster public early warning generation mode can be summarized into two types: attention oriented mode and plan oriented mode.

Keywords: disasters; public early warning; configuration analysis; TOE framework; influence factor