文章编号：1006-4354 (2007) 04-0041-02

陕西省气象科学数据集及质量控制

张 红 娟

(陕西省气象信息中心，西安　710014)

中图分类号：P413　　文献标识码：B

1 数据集的建立

1.1 资料收集和检查

陕西省气象科学数据集包含的资料为1951—2005年的基本地面气象观测资料数据文件(A0文件)。共整理了陕西99个台站的气压、气温、水汽压、相对湿度、降水量、蒸发量、风、日照时数、雷暴、沙尘暴、雾、扬沙、浮尘等要素资料。对A0文件全部进行了格式检查，并对检查出的错误逐一改正。

1.2 资料加工

数据集的主要资料来源为A0文件，A0文件中存储的数据资料为定时观测数据，数据集所使用的数据为日、旬、月、年等时段的统计资料，因此需对A0文件进行统计加工。依据《全国地面气象观测规范》，解决的主要问题：日、旬、月、年等时段资料的统计方法；缺测数据的处理方法；3次站中无自记仪器时02时（北京时）记录的处理方法等。

1.3 数据集的格式标准

完整的数据集由数据集实体、数据说明文档、附加文档和元数据等4部分组成。数据集所用的主要标准：《气象资料的分类编码及命名规范》、《气象数据集元数据格式标准》、《气象数据集说明文档格式标准》、《气象数据集组织及命名规定》。

1.4 数据集的制作

1.4.1 数据实体　由一系列数据文件组成，是数据或图形的集合，也是数据集的主体。首先依据时值资料统计生成各气象要素的日数据集实体，依托该数据实体逐步生成旬、月、年数据集实体。

1.4.2 数据集说明文档及元数据　数据集说明文档是数据集实体的说明性、标注性文件。描述内容：数据来源、数据集内容、时空属性、数据加工处理方法、数据质量状况和其它有关数据特征的信息。元数据是关于数据与信息资源的数据，即关于数据与数据集的内容、质量、状况和其他特性的信息。数据集说明文档及元数据由相应的编辑器进行录入、编辑和更新。

2 数据质量控制

数据集的数据来源A0文件大都已经过质量控制，但20世纪90年代以前的A0文件为录入员根据纸质报表录入的信息化资料，难免有录入错情，加之数据统计处理过程中仍有可能产生错情，这些错情会直接影响数据集的数据质量，因此有必要在数据处理过程中对相关要素进行质量控制。

2.1 数据实体格式检查

数据集中的数据实体是有一定格式的资料文件，其存储格式有统一的标准，通过检查其文件分类编码是否准确，文件是否为空及数据存储格式是否统一、标准等，确保数据格式正确。

2.2 数据集质量检查

2.2.1 气候学界限值、要素允许值范围检查　气候学界限值是指从气候学角度不可能出现的临界值。要素允许值范围是指气象要素值允许出现的规定范围，如风向只能是0～360°。没有通过该项检查的数据被视为错误数据。各要素的界限值和允许值见表1。

收稿日期：2007-03-07

作者简介：张红娟 (1966-)，女，陕西高陵人，学士，工程师，从事气象报表审核工作。

基金项目：国家科技基础条件平台工作项目“气象资料共享系统建设”(2005DKA31700)

2.2.2 气候极值检查 气候极值检查是检查某要素值是否超过该要素历史上出现过的最大值和最小值。要素极值是随地理区域和季节的不同而有变化的，对于不同站点、不同月份的大多数参数都有已知的气候极值，超过极值的数据都应提出疑问并进行严格审核。

表1 数据集各要素的气候学界限值或允许值

要 素	界限值或允许值范围
本站气压	300～1 100 hPa
气温	−50～60 ℃
水汽压	70 hPa
相对湿度	0～100%
日降水量	≤800 mm
日蒸发量	≤50 mm
日照时数	0～24 h
风向	0～360°
风速	0～65 m/s

2.2.3 内部一致性检查 要素项目间一致性检查是依据一定的气象学原理，对观测资料中某些物理特性关联的气象要素或项目之间是否符合一定规律进行的检测。内部一致性检查可分为三种：单一要素在同一观测时段应有如下逻辑关系，即最高值≥平均值≥最低值，如气压，当同一时段的气压出现与上述逻辑关系有矛盾的情况，则数据中包含有错误数值；同一时刻相同要素不同项目间的一致性检查，如风向和风速的一致性检查，风速还应特别注意，极大风速≥最大风速；同一时刻不同要素之间的一致性检查。各种气象要素从不同侧面描述一个测站的天气气候特征，因此同一时刻不同要素之间存在不同程度的相关，如气温与湿球温度的一致性检查气温应大于等于湿球温度。

2.2.4 合计值检查 对各要素重新计算合计值，与资料中原有的合计值比较，可检查出资料中有无明显的错误。

2.3 数据质量分析及处理

经过质量控制后所发现的错误，基本上由4种原因造成。一是信息化时录入错。由于历史资料中有相当长年代的资料开始时是以纸质报表存储，随后由录入员录入，因各种原因会导致录入错。这种错误陕西省1980年以前的资料中特别多，对于此类错误，应对照纸质报表逐一改正。二是原始数据错误导致数据集中的数据产生错误。此类错误出现频率非常低，错误分布较为分散，需要对错误逐例分析，从原始数据开始查找错误出现的基本位置并加以改正。三是统计方法使用不当或程序编码有误导致数据集出现错误。此类错误出现较为集中，会有大批数据受到影响，需从错误点反推，查找统计方法或程序编码错误并改正。四是由于仪器性能造成数据矛盾。这种矛盾应根据具体情况具体分析，不能一概而论，如很多站1980年以前使用达因测风仪，在风速较小时经常出现极大风速小于最大风速的现象。经分析，达因测风仪正是存在这种性能上的缺陷，后来被淘汰。对这种情况，在说明文档中说明即可。不论哪种错误，均须对错误本身更正，同时，分析错误的传递性，对所涉及的数据集重新制作，以消除对其他数据集数据质量的影响。

3 数据集简介

陕西省气象科学数据集包括常规资料数据集和特色资料数据集两部分。数据集总数据量770 MB，是陕西省相对较为完整的共享资料。

3.1 常规资料数据集

全省99个地面站建站至2005年，包括气压、气温、水汽压、相对湿度、降水量、蒸发、风和日照8个要素的定时、日、旬、月、年地面资料。常规资料数据集按照用户的级别分为4个等级。

零级：2个数据集，包括榆林、延安、西安、汉中、安康5个站2005年当年的月值和年值资料。

一级：2个数据集，包括榆林、延安、西安、汉中、安康5个站建站至2005年逐年的月、年值资料。

二级：5个数据集，除榆林、延安、西安、汉中、安康5个站外，其余94个站建站至2005年的时、日、旬、月、年值资料。

三级：5个数据集，包括99个站建站至2005年的时、日、旬、月、年值资料。

3.2 特色资料数据集

日值资料数据集：陕西96个地面站（除黄陵、秦岭、三原3站）1981—2005年雷暴、沙尘暴、雾三种天气现象日值资料数据集。

月值资料数据集：陕西96个地面站（除黄陵、秦岭、三原3站）1981—2005年陕西逐月沙尘暴、扬沙、浮尘的天气日数资料数据集。