刘一鸣. 2015. IVSA 模型在双来源自动站小时降水数据实时拼接中的应用. 气象,41(11):1398-1407.

IVSA 模型在双来源自动站小时降水数据实时拼接中的应用*

刘一鸣

国家气象信息中心,北京 100081

提 要: 利用 2012 年 5 月 1 日至 7 月 31 日全国自动气象站两种来源实时上传资料中的“非缺测不一致”小时降水数据,详查了问题的产生原因,为在实时拼接过程中高效判断数据正确性,提出较小尺度时间序列求证比对模型(IVSA):当同一时间同一台站的两种来源小时降水值发生不一致时,首先在较小时间尺度(分钟级)序列上使用内部一致性检查方法分别求证小时降水的正确性;当各自在较小时间尺度序列均无法证伪时,将单元出错概率引入两分钟降水序列的比对过程,并据此竞优遴选得出较为可靠的小时降水数据。指出:(1)产生非缺测不一致问题的原因主要有生成报文时观测数据不全、报文处理环节不一致、台站信息不正确三类。(2)使用 2012 年 5 月 1360 组实例形成 IVSA 模型参数后,模型在 2012 年 6—7 月的 4017 组非缺测不一致数据中取得了 99.65%的判断准确率。通过 IVSA 模型,非缺测不一致的小时降水数据取舍问题可在分钟降水序列比对中找出答案。

关键词: 自动气象站,非缺测不一致,数据源拼接,求证比对模型,相关可信度,地面观测

中图分类号: P413 **文献标志码:** A **doi:** 10.7519/j.issn.1000-0526.2015.11.010

Application of Inter Verification Sequence Alignment Model to Two Data Source Splicing of AWS Hourly Precipitation

LIU Yiming

National Meteorological Information Centre, Beijing 100081

Abstract: Non-default inconsistent hourly precipitation data are an abnormal status in automatic weather station (AWS) observation, which can be often met as hourly precipitation data are transmitted and recorded in 2 sources. Three groups of related instances are listed first, and the direct reasons that deeply hidden are found out manually. To solve this problem, Inter Verification-Sequence Alignment (IVSA) model in smaller time scale is raised in this article. When non-default inconsistent data from same station appears at the same time, verification with smaller time scale data (minute precipitation) is made respectively. If both data cannot be proved wrong with inner verification step, then unit error possibility is added into sequence alignment method. Correlation credibility is calculated and more reliable data can be selected accordingly. After then, monthly data in May 2012 (1360 pairs of instances) are used to train the parameters, and the data (4017 pairs of instances) from 1 June to 31 July 2012 are input to verify the efficiency of applying IVSA in real time data environment, getting an accuracy of 99.65%. It is concluded that IVSA model can eliminate non-default inconsistence in hourly precipitation data under running rules.

* 国家气象信息中心青年科技基金资助项目(NMICQJ201109)和中国气象局气象关键技术集成与应用重点项目(CMAGJ2013Z01)共同资助

2014 年 10 月 8 日收稿; 2015 年 5 月 5 日收修定稿

作者:刘一鸣,主要从事气象信息加工处理与气象资料质量控制工作. Email:yimingliu@aliyun.com

Key words: automatic weather station (AWS), non-default inconsistent data, data source splicing, Inter Verification-Sequence Alignment Model, correlated credibility, surface observation

引　言

我国对降水资料的自动观测始于 20 世纪 90 年代末。2000—2005 年已逐步完成 700 余个基本基准站的自动站安装，并在大部分台站开展过自动与人工对比观测和双轨并行(任芝花等，2007a；刘小宁等，2008；杨萍等，2011)。2005 年发布了全国统一的数据文件上传格式，从那时起多要素站以“地面气象要素数据文件”形式、单要素站以“加密自动雨量站观测资料文件”形式实时上传至国家气象信息中心，实现了国家级自动站降水资料的实时共享服务；也由此构成了小时降水数据的两种数据来源(多要素站、单要素站)，并存在一定数量的单站双数据来源现象(即：同一个站，既发“地面气象要素数据文件”报，也发“加密自动雨量站观测资料文件”报)。随着区域站建设力度的逐年增强，台站数目不断增多，自动站观测在时间频次与空间覆盖率上的优势逐步显现(俞小鼎，2012；许新田等，2012；陈涛等，2013)，截止到 2012 年 6 月，自动站考核站数已达 31819 站，其中包括国家站 2411 个、区域站 29408 个(中国气象局预报网络司，2012)，部分站点的观测频次已提升至分钟级；同时随着观测技术不断优化成熟，部分台站采用分期建设方式，由原来的单要素站升级为多要素站，相应的数据来源也应同步完成切换。近期，随着地面气象资料一体化工作的推进，2012 年开始国家站以新文件形式实时上传国家气象信息中心；同时通过地面气象观测自动化业务综合试点工作的开展，未来地面观测数据可能以更新的形式(BUFR 及消息体)进行描述。

在多要素站与单要素站长期并存的业务现状下，将站数尽可能多、空间覆盖尽可能全的两来源降水数据以尽可能高的时效拼接在一起，会更加贴近预报服务业务用户的需求。然而在尝试实时拼接时，来自两来源的小时降水数据中存在的微小不一致成为拼接过程中不可回避的问题。如：针对单站双数据来源的情况，最初于 2005 年采用“先入为主法”进行快速拼接，即以两来源中最先到达的数据为准提供下游用户使用。后来随着业务应用的不断深入，发现单站双数据来源中存在一个来源数据缺测、而另一来源中数据非缺测的个别情况；此时如采用“先入为主法”，很可能会影响到局部区域的数据完整性。

经过长期的演化与发展，地面气象资料质量控制领域已形成一套包括气候学极值检查、区域或台站界限值检查、要素之间内部一致性检查、时间一致性检查以及空间一致性检查在内的质量控制方法体系(Igor，2004；熊安元，2003；中国气象局，2010)。长期以来，由于一方面受限于过去人工观测的频次较稀，另一方面有碍于实时服务对时效性的要求较高，这些方法在地面日值、月值、年值等气候资料中得到了深入的研究和广泛的应用(Sciuto et al，2009；刘小宁等，2005；任芝花等，2007b)，而在小时值观测资料中的探索相对有限(窦以文等，2008；任芝花等，2006)。自 2009 年，国家气象信息中心首先以自动站小时降水(任芝花等，2010)为例，尝试将一些成熟的质量控制方法分批应用到逐小时观测的实时资料服务中(鞠晓慧等，2010；赵煜飞等，2011；李志鹏等，2012；周笑天等，2012)。一方面，空间一致性检查方法需要尽可能完整的邻近站数据；另一方面，实时质量控制系统也需要一套高时效、高可靠的数据源。于是 2010 年以来采用“非缺测优先法”进行小时降水数据的实时拼接，即非缺测数据较缺测数据具有更高的优先级，可随时替换缺测数据，但不可被缺测数据所替换。

然而随着近年自动站数目的激增与站网规模的扩大，在近期的实时资料中发现，对于单站双数据来源而言，存在两来源数据均非缺测且不一致的情况。对这种同站同时刻探测要素不一致现象，单从小时降水数据本身无法完成数据有效性的判断。通常的辨识方法(任芝花等，2007a；Igor，2004；熊安元，2003；中国气象局，2010)：可采用内部一致性检测——将分钟降水累计值与同一时间段的小时降水值比较。但笔者在采用此法进行检验时发生过两来源数据各自内部一致性检查均正确，而最终数据仍不一致的情况。因此，内部一致性法难以解决所有问题。

序列分析是基于随机过程理论和数理统计学方法研究随机数据序列所遵从规律的方法，具有准确把握数据相似性与同源性的特点(Fitch，1983；Lip-

man et al,1984;Altschul et al,1985),在金融、地质、生物、医药等领域已取得广泛应用。本文旨在利用序列分析方法,针对 2012 年 5 月 1 日至 7 月 31 日自动站两种来源实时上传资料中的"非缺测不一致"小时降水数据,在逐站次详查问题产生原因的基础上形成模型参数培训集合与结果验证集合,探讨非缺测不一致问题发生的规律性,并尝试提出一套高效便捷的正确数据筛选方法。

1 资料来源与预处理方法

本文使用的资料来源于 2012 年 5 月 1 日至 7 月 31 日(世界时)全国气象资料国内通信系统在全国范围实时收集到的自动站观测资料数据文件。为下文叙述方便,首先对涉及的名词给出如下解释:

来源 1:地面气象要素数据文件(中国气象局监测网络司,2005;中国气象局监测网络司,2008),包括小时降水量在内的气温、气压、湿度、降水、风向、风速等多种要素的自动站数据文件。

来源 2:加密自动雨量站观测资料文件(中国气象局监测网络司,2005;中国气象局监测网络司,2008),包括与降水相关的小时降水量、日累计降水量、分钟降水量等要素的自动站数据文件。

非缺测不一致:指国家气象信息中心接收到的来源 1 与来源 2 数据出现同一站在同一时间小时降水量均为非缺测值,且数值不相等的情况。

为提取非缺测不一致小时降水数据,首先对来源 1 资料与来源 2 资料分别进行解析处理,提取得到 36185 站 74591639 份来源 1 与 5922 站 11518378 份来源 2 中的小时降水值及其对应的分钟降水序列(序列长度为 60),其中同站同时编发两来源数据的有 624 站 2862270 份,占总样本量(86110017 份)的 3.32%,其中同时以两来源编发且数据不一致的有 108 站 43566 份,占总样本量的 0.051%,其中七成为一来源缺测、另一来源非缺测的情况,可采用"非缺测优先法"判断。通过对两数据集合逐小时对比,并由台站级与省级业务人员结合本地人工观测、雷达观测等资料,开展逐站次人工确认核查,筛选得到源自 48 站的 5377 组(10754 份)满足非缺测不一致定义且有明确反馈的样本数据,占总样本量的 0.012%,其中 5068 组来源 1 正确且来源 2 错误,309 组来源 2 正确且来源 1 错误。

2 解决方法

非缺测不一致情况具有发生时间不定、原因各异的特点,问题排查难度较大,后文详述问题原因的查找,是在发现问题之后,由台站级与省级业务人员结合本地人工观测记录及雷达观测等资料,配合电话、信函等方式逐站完成的,耗时较长;这种方法显然不适用于实时观测资料的高时效处理与应用过程。为此,本文借鉴时间序列分析方法(Pearson,1998;Altschul et al,1994;阎继伟,2006),构建了较小尺度时间序列求证比对模型(简记为 IVSA),尝试在现有业务流程与工作模式不变的情况下,对此类问题予以实时判断。

2.1 IVSA 模型的提出

通常情况下,小时(累计)降水量可描述为式(1)的形式:

$$X=\int_{t=0}^{60}x(t)\mathrm{d}t \tag{1}$$

实际业务中,小时降水为 60 段分钟(累计)降水量的累加,可描述为:

$$X=\sum_{i=1}^{60}x_i \tag{2}$$

式中,X 为小时降水量;x_i 为第 i 分钟的分钟降水量;t 为时间;与此同时,考虑到以下两点情况:

(1) 无论在来源 1 还是在来源 2 的数据文件中,小时降水量这一被检数据在更小时间尺度(分钟级)上可以找到支撑性证明数据。

(2) 通常情况下,小时降水量的非缺测不一致情况在数据文件中具有分钟级降水量选取时间段不同的特征,其中一种来源存在提取分钟降水时间段不足或未提取到分钟降水数据的问题。

由此提出一个适用于实时遴选较为可靠数据的模型,以使两数据源的实时拼接流程可在数据准确性尽可能得以保障的前提下高效完成。

2.2 IVSA 模型的描述

IVSA 模型包括内部一致性检测与基础值交叉检测两个主要步骤,当被检值为 Z,参考值为 Z',被检值对应的较小尺度被检序列为 $z_1,z_2,\cdots,z_i$,参考值对应的较小尺度参考序列为 $z'_1,z'_2,\cdots,z'_i$ 时,IVSA 模型可详细描述为:

(1) 步骤 1(内部一致性检测):淘汰被较小时间尺度数据序列证伪的被检值:

如果 $Z=\sum_{i=1}^{60} z_i$,则较小尺度序列证实被检值 Z,进行下一步判断;　　(1.1)

如果 $Z\neq\sum_{i=1}^{60} z_i$,则较小尺度序列证伪被检值 Z,淘汰该被检值 Z。　　(1.2)

(2) 步骤 2(基础值交叉检测):通过较小时间尺度数据验证,两被检值均被证实的情况,采用序列比对方法,择优选取:

将被检序列中的 z_i 与参考序列中的 z'_i 进行比对

当 $z_i=z'_i$ 时,$a_i=1$　　(2.1)

当 $z_i\neq z'_i$ 时,$a_i=1-p(z_i,z'_i)$　　(2.2)

则被检值 Z 的相关可信度 $A=(\sum_{i=1}^{60} a_i)/|z_i|$。　　(2.3)

选取相关可信度 A 较大的 Z,认为其较另一值更为可信。

其中 A 为被检值 Z 的相关可信度;a_i 为被检序列中 z_i 的相关可信度;z'_i 为参考序列中的第 i 个变量;$|z_i|$ 为被检序列 $z_1,z_2,\cdots,z_i$ 的长度,本文中为固定值 60;$p(z_i,z'_i)$ 为当被检序列中值为 z_i、参考序列中对应的第 i 个变量为 z'_i 时,z_i 错误 z'_i 正确的概率,简称单元出错概率,实际应用中为根据各类错误的发生频次做出的统计值。

在 IVSA 模型中来源 1 与来源 2 的小时降水量互为被检值和参考值,分钟降水量互为被检序列和参考序列。即当来源 1 的小时降水量作为被检值时,来源 2 的小时降水量为参考值,来源 1 中的分钟降水为被检序列,来源 2 中的分钟降水为参考序列,此时 $p(z_i,z'_i)$ 为 $p_1(z_i,z'_i)$;当来源 2 的小时降水量作为被检值时,来源 1 的小时降水量为参考值,此时 $p(z_i,z'_i)$ 为 $p_2(z_i,z'_i)$。由于非缺测不一致问题的发生具有隐蔽性高、间歇性强的特点,培训形成 $p(z_i,z'_i)$ 的过程需要足够长时间的数据积累与准确的原因反馈(一般应达数百份),并且培训数据的时间段要尽可能与验证数据的时间段相接近,本文选定 2012 年 5 月的 1360 组数据作为 $p(z_i,z'_i)$ 的培训集合,2012 年 6—7 月的 4017 组数据作为模型运算结果验证集合。

2.3　IVSA 模型的特点

IVSA 模型的整体流程如图 1 所示,当发现一组非缺测不一致的小时降水数据时,首先执行模型步骤 1 的内部一致性检测,未通过检测者说明其分钟降水序列不支持其小时降水值 X,将该被检值 X 淘汰。如两者均通过内部一致性检测,则说明非缺测不一致现象的原因在于两者的基础值,即分钟降水序列存有不一致,则可通过步骤 2,结合单元出错概率 $p(x_i,x'_i)$ 计算相关可信度 A,并选择两者中相关可信度较大者胜出,从而筛选出相对更为可靠的小时降水值 X。

IVSA 模型的计算复杂度为 O(n)级,运算量随样本数据量的增加线性增长,适合在实时业务环境下快速生成运算结果。单元出错概率 $p(x_i,x'_i)$ 的选取可进行本地化调配,在地面观测系统分布式采集数据的业务现状下具有较强的适应性。并且小时降水数据和与其对应的分钟降水序列在同一数据文件的同条记录中(中国气象局监测网络司,2005),使得该模型的应用不受数据存储形式所限。

2.4　IVSA 模型的业务逻辑

为使 IVSA 模型更加适应当前复杂的业务环境,在实现过程中增加了如下业务逻辑:

业务逻辑 1:缺测以零计入累计值

考虑到报文中的实际上报分钟降水值 y_i 为缺测值时,中心站软件存在以零计入小时降水值的业务现行作法,于是通过公式(3)计算分钟降水量 x_i。

$$x_i=\begin{cases}0 & (y_i=\text{缺测值})\\ y_i & (y_i\neq\text{缺测值})\end{cases}\qquad(3)$$

业务逻辑 2:更正报数据优于非更正报数据

如果非缺测不一致的两个小时降水量数据中,存在一个来源于更正报,那么更正报具有更高的相关可信度 A。

3　实例分析

3.1　问题的发现

非缺测不一致问题的发现,始于将两来源下小时降水数据实时拼接使用的业务需求。从业务数据中提取到的三组非缺测不一致报文实例中(图 2),下划线标记部分代表以0.1 mm为单位的小时降水

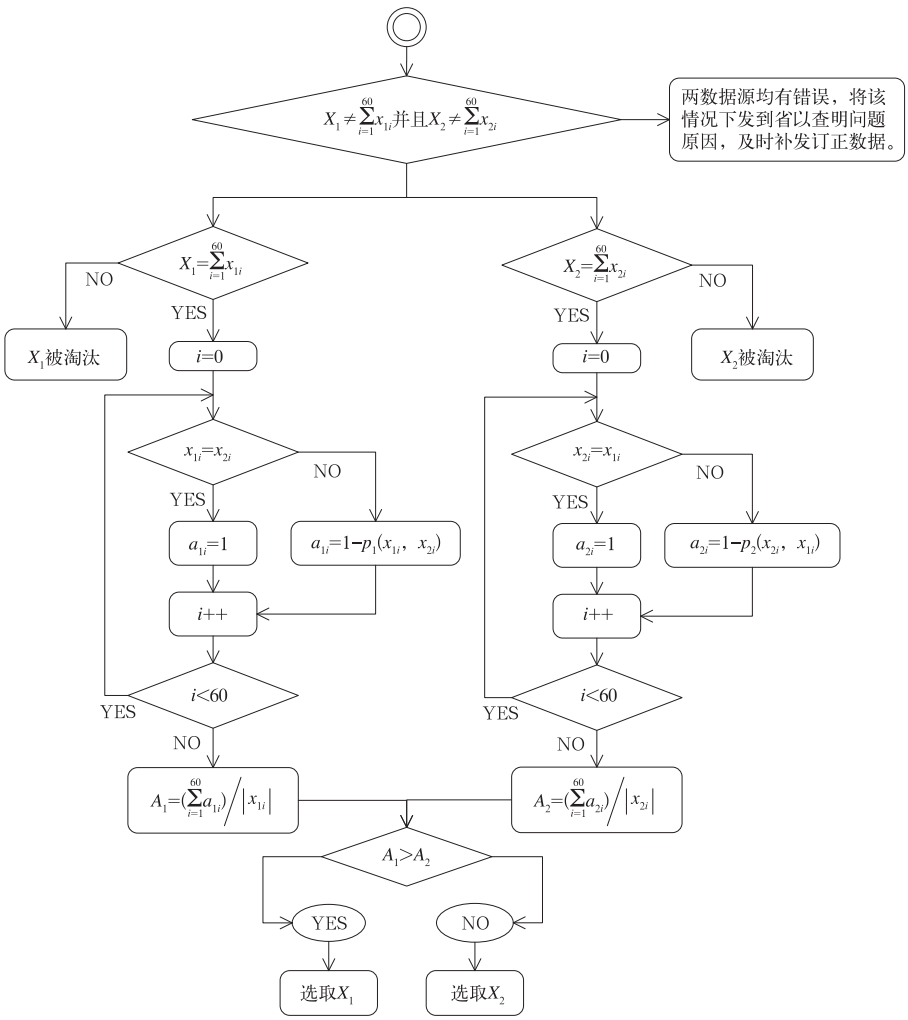

图 1 IVSA 模型流程图

Fig. 1 Flow chart of IVSA model

量；虚线框标记部分代表 60 段分钟降水量，每两位字符对应 1 个以 0.1 mm 为单位的分钟降水量（其他字段的格式说明详见自动站文件格式的详细规定（中国气象局监测网络司，2005））。

3.2 原因详查与错报特征

图 2 所示的三组非缺测不一致实例分别代表产生非缺测不一致问题的三类情况，经与台站核实并进行多方调查，该问题发生的直接原因有以下三类：

（1）生成报文时观测数据不全。第一组中 III44 站来源 1 编报小时降水 3.1 mm（实例 1.1），相同时间相同站的来源 2 编报小时降水却为 1.9 mm（实例 1.2），其数据流程为来源 1 与来源 2 均由省级数据库在对原始观测数据解析入库后生成（图 3 中实例 1）。区域站与中心站间的通信受 GPRS 信号延迟影响，在生成来源 2 文件时后 29 min 的数据并未接收到，而在随后进行的生成来源 1 文件操作时，后 29 min 的数据已收全，由分钟雨量累加小时雨量使用的有效数据时间段不尽相同造成了两来源数据的不一致。发生此类错误的数据特征为错报的分钟降水序列未收齐，部分数据编报缺测标志。

（2）报文处理环节不一致。第二组中 III46 站来源 1 编报 0.0 mm（实例 2.1），对应来源 2 编报则为 1.6 mm（实例 2.2）。因来源 2 中包括日累计雨量，须将之前若干小时的小时降水计算在内，所以该省在此环节引入数据库完成来源 2 数据生成（图 3 中实例 2）。而当台站发现前报有误，补发订正报时，由于生成来源 2 报文的流程中未包括处理订正报环节，所以造成订正信息仅以来源 1 上传，未对来源 2 的数据进行更新，数据差异由此发生。发生此类错误的数据特征为两分钟降水序列中存有数据不一致。

第一组：

```
III44 LAT521 LON2051 00790 ///// 4                                          实例 1.1：来源 1
20100601020000 /// /// /// /// /// /// //// /// /// /// /// //// 0031 //// //// //// //// //// /// /// //// /// // // ///// ///// //// ///// //// //// //// //// //// ////
//// //// //// //// //// //// //// //// //// //// //// //// //// //// ///// ///// ///// ////
0100010001000100020001000100010001000101000100010101010100000101//010100010001000200000100000100000100
000100000000010000=
```

```
III44 LAT521 LON2051 00790                                                  实例 1.2：来源 2
20100601020000 00019 00161
01000100010001000200010001000100010001010001000101010101000001/////////////////////////////////////////////////////////////=
```

第二组：

```
III46 LAT447 LON2634 01020 ///// 4                                          实例 2.1：来源 1
20100602040000 /// /// /// /// /// /// //// /// /// /// /// //// 0000 //// //// //// //// //// /// /// //// /// // // ///// ///// //// ///// //// //// //// //// //// ////
//// //// //// //// //// //// //// //// //// //// //// //// //// //// ///// ///// ///// ////
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000=
```

```
III46 LAT447 LON2634 01020                                                  实例 2.2：来源 2
20100602040000 00016 00016
000000000000000000000000000000000000000000000000000000000000000000001600000000000000000000000000
0000000000000000000=
```

第三组：

```
III08 LAT718 LON2426 28500 ///// 4                                          实例 3.1：来源 1
20100619020000 /// /// /// /// /// /// //// /// /// /// /// //// 0238 0248 0249 0101 0248 0103 /// /// //// /// // // ///// ///// //// ///// //// ////
//// //// //// //// //// //// //// //// //// //// //// //// //// //// //// //// //// //// ///// ///// ///// ////
000001000001000001000102020203040609121011080906081313151412080402020101010001010001010100010100010001000
1010205060810080403=
```

```
III08 LAT718 LON2426 02850                                                  实例 3.2：来源 2
20100619020000 00000 00000
////////////////////////////////////////////////////////////////////////////////////////////////////=
```

图 2　三组非缺测不一致实例

Fig. 2　3 Groups of non-default inconsistent instances

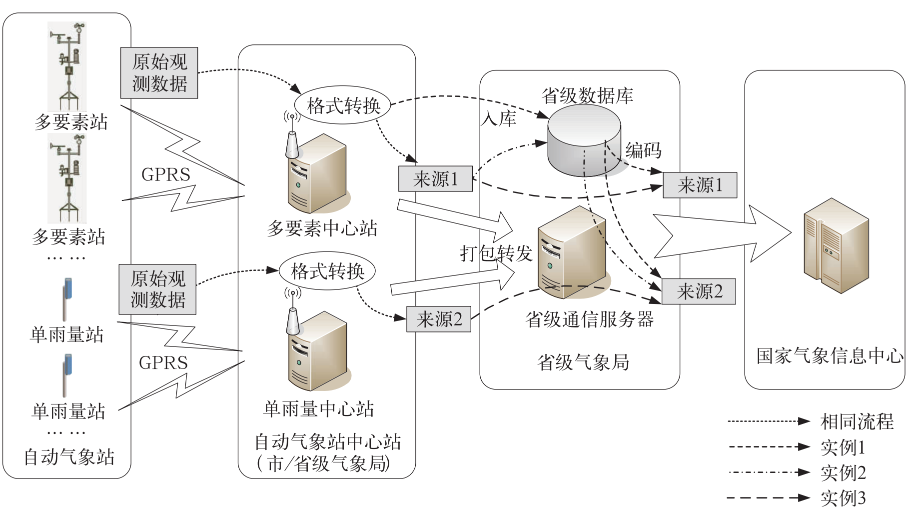

图 3　区域自动站数据流程图

Fig. 3　Flow chart of regional AWS data

（3）台站信息不正确。第三组中 III08 站来源 1 编报 23.8 mm（实例 3.1），对应来源 2 编报则为

0.0 mm（实例 3.2），数据流程对应图 3 中实例 3。当测站由单要素观测升级到多要素观测时，须在多要素中心站增加该站，并在单雨量中心站删除该站，当删除单雨量中心站中该站信息的操作未做时，第三组问题就此产生：单雨量中心站仍认为该站（已升级为多要素站）为单雨量站，在收不到该站观测数据的情况下，将该站来源 2 报文中的每条分钟降水均标为缺测（有时为零值），并将该站的小时降水量置为 0 mm。发生此类错误的数据特征为错报的分钟降水序列全部或大部数据上报缺测或零值。

4 检验效果

4.1 检验过程

为验证 IVSA 模型在数据源拼接过程中判断结果的正确性，本文引入模型判断准确率（M_R）与模型判断错误率（M_W），作为检验 IVSA 模型有效性的重要指标：

$$M_R = \frac{N_R}{N_0} \times 100\% \tag{4}$$

$$M_W = \frac{N_W}{N_0} \times 100\% \tag{5}$$

式中，N_0 为参与验证过程的非缺测不一致样例个数；N_R 为模型判断与反馈结果一致的样例个数，N_W 为模型判断与反馈结果不一致的样例个数，并且满足：

$$N_R + N_W = N_0 \tag{6}$$

$$M_R + M_W = 100\% \tag{7}$$

在 2012 年 5 月 1 日至 7 月 31 日共计三个月的两来源自动站观测资料中，非缺测不一致情况的日出现频次如图 4 所示，由于该情况在无降水时一般不会显现，这也增加了样本搜集的难度。通过大量逐站确认核查，得到台站反馈的明确确认结果，使模型的准确性判断具有足够的参考依据。去除 5377 组样例中不满足内部一致性的数据（步骤 1 未通过），共得到 5273 组非缺测不一致且有反馈的样本数据，分钟降水在这 5273 组数据中的分布情况为 0 mm 共出现 316025 次，占总数的 49.94%；分钟降水>0 mm 共出现 48516 次，占总数的 7.67%；分钟降水缺测共出现 268219 次，占总数的 42.39%。降水为 0 mm 与缺测的情况在其中占有很大比例，而有降水（>0 mm）的情况是不容忽视的，于是可将分钟降水数据按无降水（=0 mm）、有降水（>0 mm）、降水缺测三类加以划分。

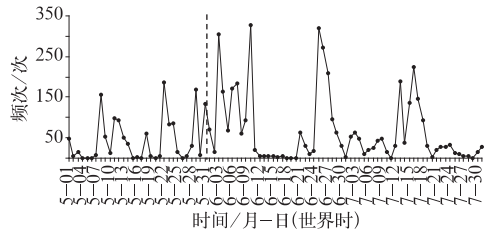

图 4 2012 年 5 月 1 日至 7 月 31 日非缺测不一致情况日出现频次

Fig. 4 Daily frequency of non-default inconsistence from 1 May to 31 July 2012

使用 2012 年 5 月数据作为对模型参数化的培训数据（图 4 中虚线之前的数据），随后运用 2012 年 6—7 月的数据作为模型参数化结果的验证数据（图 4 中虚线之后的数据，$N_0 = 4017$）。去除 2012 年 5 月非缺测不一致样例中来源 1 min 降水与来源 2 min 降水相等的情况后，得到的分钟级降水不一致情况见表 1 第 1 列。根据搜集得到的反馈，在这些不一致的分钟降水中来源 1 正确且来源 2 错误及来源 1 错误且来源 2 正确情况的分布情况分别对应表 1 第 2 列与第 3 列。如果用 $C_{i,j,k}$ 表示表 1 第 i 行第 j 列中的第 k 个数据，则模型参数表表 2 中第 i 行第 j 列错误发生的频率 $P_{i,j}$ 可由式（8）得出，模型参数表表 3 中第 i 行第 j 列错误发生的频率 $P'_{i,j}$ 可由式（9）得出，例如 $P_{1,2}$ 的计算过程如式（10）所示。

$$p_1(x_i, x'_i) = P_{i,j} = \frac{C_{i,j,3}}{C_{i,j,1}} \tag{8}$$

$$p_2(x_i, x'_i) = P'_{i,j} = \frac{C_{i,j,2}}{C_{i,j,1}} \tag{9}$$

$$p_1(x_i = 0\ \text{mm}, x'_i > 0\ \text{mm}) = P_{1,2} = \frac{C_{1,2,3}}{C_{1,2,1}} = \frac{42}{1285} = 3.27\% \tag{10}$$

使用 2012 年 5 月的错误发生频率作为模型中单元出错概率 $p(x_i, x'_i)$，可得单元出错概率参数表 2 与表 3。由于降水测量仪器的准确度与降水强度有关，在降水量≤5 mm 时为±0.1 mm；在降水量>5 mm 时为±2%（中国气象局，2011）。与此同时，各地降水强度差别显著：华南 $R_{24} \geqslant 80$ mm 称为暴雨；西北 $R_{24} \geqslant 25$ mm 就称暴雨了（王秀明，2011）。国际上对暴雨的界定也有所不同，美国将小时降水量 $R_1 > 50$ mm 定为暴雨（Met Office，2011；Glossary of Meteorology，2010）。所以，在业务应用时表 2

中被检值 $x_i>0$ mm 且参考值 $x'_i>0$ mm 参数设置部分可根据实际情况调配，但目前的数据样本中较为罕见（不一致的仅有 3 例），所以不作为本文讨论的重点。

表 1　分钟降水量 x_i 不一致频次分布表

（不一致数|来源 1 正确来源 2 错误数|来源 1 错误来源 2 正确数）

Table 1　Distribution of minute precipitation x_i inconsistence frequency

来源 1	来源 2								
	=0 mm			>0 mm			缺测		
	不一致	1 对 2 错	1 错 2 对	不一致	1 对 2 错	1 错 2 对	不一致	1 对 2 错	1 错 2 对
=0 mm	0	0	0	1285	1243	42	41526	41526	0
>0 mm	3125	3125	0	3	3	0	7265	7265	0
缺测	1357	1030	327	47	0	47	0	0	0

表 2　$p_1(x_i,x'_i)$ 值的确定（单位：%）

Table 2　$p_1(x_i,x'_i)$ values (unit: %)

x_i	x'_i		
	x'_i=0 mm	x'_i>0 mm	x'_i=缺测
=0 mm	0.00	3.27	0.00
>0 mm	0.00	0.00	0.00
=缺测	24.10	100.00	0.00

表 3　$p_2(x_i,x'_i)$ 值的确定（单位：%）

Table 3　$p_2(x_i,x'_i)$ values (unit: %)

x_i	x'_i		
	x_i=0 mm	x'_i>0 mm	x'_i=缺测
=0 mm	0.00	96.73	100.00
>0 mm	100.00	100.00	100.00
= 缺测	75.90	0.00	0.00

4.2　实际问题解决效果

在使用 2012 年 5 月数据形成 IVSA 参数（表 2）的基础上，首先将 3.1 部分的三组实例代入模型，计算结果如表 4 所示，较大的相关可信度 A 以下划线标记。其中样例 1.1 与样例 1.2 给出了相关可信度 A 的详细计算步骤。IVSA 计算结果表明：实例 1 两来源的相关可信度 A 分别为 100.00% 和 53.33%，IVSA 判断来源 1 较为准确。实例 2 两来源的相关可信度 A 分别为 99.95% 和 98.39%，并且由于前者来源为订正报，根据业务逻辑 2 以来源 1 为准。实例 3 的两来源的相关可信度 A 分别为 100.00% 和 0.00%，模型判断来源 1 较为准确。IVSA 计算结果与第 3.2 节所述三组问题报文的实际数据情况均保持一致。

然后，将 2012 年 6 月 1 日至 7 月 31 日数据使用相同验证方法代入 IVSA 模型，计算结果（表 5）表明：在参加验证的 4017 组数据中，共发现 104 组未能通过内部一致性检查；在通过内部一致性检查的 3913 组数据中，模型步骤 2 判断计算值与反馈情况一致的有 3899 组，IVSA 模型的判断错误率 M_W 为 0.35%，判断准确率 M_R 高达 99.65%。

IVSA 模型较高判断准确率 M_R 的取得源于设计中（2.2 节）对产生非缺测不一致的三类情况（3.2 节）均有准确的表达。当生成报文观测数据不全时，在先生成的数据来源（实例 1.2）中分钟序列数据不全，未到的若干分钟以缺测标志补齐并以零计入小时降水值，在步骤（2.2）的计算时，缺测相对于其他数值的出错概率 $p(x_i,x'_i)$ 较大，得到的 a_i 会较小，在步骤（2.3）累计相关可信度时得到的 A 也偏小，从而筛选得出后生成的分钟序列（实例 1.1）更为可靠。当报文处理环节不一致时，订正报的流程不一致会造成数据差别的出现，业务逻辑 2 将保证模型判断的有效性。当台站信息不正确时，由异常中心站编发的数据，分钟降水均标为缺测或为 0 值，小时降水量置为 0 mm，在步骤（2.2）计算时也会得到较小的 a_i，确保了步骤（2.3）基于相关可信度 A 的判断与实际情况一致。

IVSA 模型计算结果与反馈情况不一致的有 14 组，模型判断错误率 M_W 为 0.35%。经核查问题发生场景为：自动站在由单要素升级到多要素后，单要素设备在运输过程中未卸下电源仍继续发报造成了异常数据的产生。这是一个违反自动站仪器更换操作规程的极少发生的小概率事件，由于雨量观测设备工作原理（Principal Hazards in U. S. doc，2010；张霭琛，2006；National Weather Service Office，2009）所限，单从数据角度模型无法给出准确的判断，业务应用中可适当引入查询反馈机制做出更为有效的判断。

表 4 三组实例结果分析表

Table 4 Result analysis of real instances of the 3 groups

样例编号	数据来源	小时降水量段（0.1 mm）	分钟降水序列（每两位为 1 min 降水数据，单位 0.1 mm）	相关可信度 A
1.1	来源 1	0031	0100010001000100020001000100010001000101000100010101010100000101//0101000100010002000001000001000001000001000000000010000=	100.00%
样例 1.1 的 60 min a_i			111111111111111111111111111111111111111111111111111111111111	A=60/60
1.2	来源 2	00019	0100010001000100020001000100010001000101000100010101010100001//=	53.33%
样例 1.2 的 60 min a_i			111111111111111111111111111111111111111010000000000000000000000000000000000000000000000000000000000000	A=32/60
2.1	来源 1	0000	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000=	99.95%
2.2	来源 2	00016	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000=	98.39%
3.1	来源 1	0238	0000010000010000010001020202030406091210110809060813131514120804020201010100010100010101000101000100010102050608100804 03=	100.00%
3.2	来源 2	00000	//=	0.00%

表 5 2012 年 6 月 1 日至 7 月 31 日数据验证结果

Table 5 Data validation results of real instances from 1 June to 31 July 2012

观测时间	样例总数	未通过步骤 1 数目	经步骤 2 检查数目	步骤 2 判断准确数	步骤 2 判断错误数
2012.6.1—30	2622	21	2601	2587	14
2012.7.1—31	1395	83	1312	1312	0
合计：	4017	104	3913	3899	14

5 结论和讨论

利用从 2012 年 5 月 1 日至 7 月 31 日全国自动气象站两种来源实时上传资料中提取得到的“非缺测不一致”小时降水数据，通过较小尺度时间序列求证比对模型辅助开展数据分析，结果表明：

（1）小时降水数据分布于两数据来源之下，为非缺测不一致问题的产生创造了可能，但也提供了发现现有业务系统中存在问题、进而加以解决的机会。产生非缺测不一致问题的原因主要包括生成报文时观测数据不全、报文处理环节不一致、台站信息不正确三类，与此对应错误上报的分钟序列具有部分数值缺测、两序列数据不一致、全部或大部分数据上报缺测或零值的数据特征。

（2）运用较小尺度时间序列求证比对模型，基于 2012 年 5 月 1360 组数据统计形成了模型参数，使用 2012 年 6 月 1 日至 7 月 31 日两个月的 4017 组数据对模型的有效性加以验证，99.65%的情况下 IVSA 计算结果与反馈情况保持一致。

（3）非缺测不一致问题发生具有隐蔽性高、间歇性强的特点。IVSA 模型虽然利用小时-分钟降水数据的错报特性，将此类系统误差对实时数据准确性的影响控制在较小范围内。但产生问题的根源实为设备特性、系统环境、业务流程等复杂因素共同作用下引入的系统性误差，在现行运行体系下，这样的误差已存在于历史数据中，并且不排除被继续引入到实时数据中的可能，所以可将数据源拼接与质量控制、查询反馈等流程节点相结合，综合保障数据正确性。

(4) 为适应复杂业务环境下的业务逻辑,IVSA 模型按现行规则引入了部分业务逻辑。但在模型辅助检查过程中发现的一些问题,如缺测按 0 值上传、60 个分钟点的降水量未收齐时中心站能否编发小时降水量等,应从制度上加以规范并在业务实现上精准地执行,才能将非缺测不一致异常的发生可能性有效控制在较低水平,而非事后补救。

(5) IVSA 模型在业务应用中可以灵活地由两数据源拼接推广到多数据源比对与拼接过程中,方法具有良好的扩展性。如果中国观测系统在数据源不断丰富的历史大背景下,未来的发展趋势为向美国单点三套设备同步观测靠近,则非缺测不一致情况的发生概率将大大提升,此模型的实现作为先探性理论储备,在多套数据源实时拼接方面将具更为广阔的应用空间。

参考文献

陈涛,代刊,张芳华. 2013. 一次华北飑线天气过程中环境条件与对流发展机制研究. 气象,39(8):945-954.

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

鞠晓慧,任芝花,张强. 2010. 自动站小时气压的质量控制方法研究. 安徽农业科学,38(27):15130-15133.

李志鹏,张玮,黄少平,等. 2012. 自动气象站数据实时质量控制业务软件设计与实现. 气象,38(3):371-376.

刘小宁,任芝花. 2005. 地面气象资料质量控制方法研究概述. 气象科技,33(3):199-203.

刘小宁,任芝花,王颖. 2008. 自动观测与人工观测地面温度的差异及其分析. 应用气象学报,19(5):554-563.

任芝花,熊安元. 2007a. 地面自动站观测资料三级质量控制业务系统的研制. 气象,33(1):19-24.

任芝花,熊安元,邹风玲. 2007b. 中国地面月气候资料质量控制方法的研究. 应用气象学报,18(4):516-523.

任芝花,许松,孙化南,等. 2006. 全球地面天气报历史资料质量检查与分析. 应用气象学报,17(4):412-420.

任芝花,赵平,张强,等. 2010. 适用于全国自动站小时降水资料的质量控制方法. 气象,36(7):123-132.

王秀明. 2011. 台风、暴雨、强对流. 中国气象局培训中心,12.

熊安元. 2003. 北欧气象观测资料的质量控制. 气象科技,31(5):314-320.

许新田,刘瑞芳,郭大梅,等. 2012. 陕西一次持续性强对流天气过程的成因分析. 气象,38(5):533-542.

阎继伟. 2006. 时间序列的数据挖掘研究. 上海:上海交通大学:12-18.

杨萍,刘伟东,仲跻芹,等. 2011. 北京地区自动气象站气温观测资料的质量评估. 应用气象学报,22(6):706-715.

俞小鼎. 2012. 2012 年 7 月 21 日北京特大暴雨成因分析. 气象,38(11):1313-1329.

张霭琛. 2006. 现代气象观测. 北京:北京大学出版社,166.

赵煜飞,任芝花,张强. 2011. 适用于全国气象自动站正点相对湿度资料的质量控制方法. 气象科学,31(6):687-693.

中国气象局. 2010. 地面气象观测资料质量控制. 北京:气象出版社,8.

中国气象局. 2011. 地面气象观测规范. 北京:气象出版社,126.

中国气象局监测网络司. 2005. 关于进行加密自动气象(雨量)站资料传输试验的函(附:加密自动气象(雨量)站数据文件格式、加密自动气象(雨量)站观测资料传输规定等). 中国气象局预报网络司,17.

中国气象局监测网络司. 2008. 自动站观测资料传输文件名调整方案. 中国气象局预报网络司,3.

中国气象局预报网络司. 2012. 2012 年自动站资料考核台站表. 中国气象局预报网络司.

周笑天,褚希,姚志平. 2012. 一种基于 k-means 聚类的实时气温动态质量控制方法. 气象,38(10):1295-1300.

Altschul S F, Boguski M S, Gish W, et al. 1994. Issues in searching molecular sequence databases. Nature Genet, 6:119-129.

Altschul S F, Erickson B W. 1985. Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol, 2(6):526-538.

Fitch W M. 1983. Random sequences. J Molecular Biology, 163:171-176.

Glossary of Meteorology. 2010. Rain. American Meteorological Society. http://amsglossary.allenpress.com/glossary/search?id=rain1.

Igor Zahumensk. 2004. Guidelines on Quality Control Procedures for Data from Automatic Weather Stations. Expert Team on Requirements for Data from Automatic Weather Stations, Third Session, WMO.

Lipman D J, Wilbur W J, Smith T F, et al. 1984. On the statistical significance of nucleic acid similarities. Nucleic Acids Research, 12:215-226.

Met Office. 2011. Fact Sheet No. 3: Water in the Atmosphere. Crown Copyright, 6. http://www.metoffice.gov.uk/media/pdf/4/1/No._03_-_Water_in_the_Atmosphere.pdf.

National Weather Service Office. 2009. 8 Inch Non-Recording Standard Rain Gauge. Northern Indiana. http://www.crh.noaa.gov/iwx/program_areas/coop/8inch.php.

Pearson W R. 1998. Empirical statistical estimates for sequence similarity searches. J Molecular Biology, 276:71-84.

Principal Hazards in U.S.doc. 2010. Chapter 5 - Principal Hazards in U.S.doc, 128. http://training.fema.gov/EMIWeb/edu/docs/fem/Chapter 5-Principal Hazards in U.S.doc.

Sciuto G, Bonaccorso B, Cancelliere A, et al. 2009. Quality control of daily rainfall data with neural networks. J Hydro, 364:13-22.