道路交通流数据质量控制与评价

谷远利 马韵楠 李巨伟

【摘要】动态交通数据质量评价可为交通管理系统提供可靠的数据支持，是交通管理系统的重要组成部分。本文针对故障数据—缺失、不规则时间点和异常数据这 3 类常见问题，提出了交通流数据质量的控制方法和流程。同时，第一次提出分三阶段分别对不同 ITS 检测器的交通数据质量进行评价，并构建了六维评价指标体系。最后以北京市二环路段交通流数据为例，对本文控制流程和评价方法进行了分析与验证。

【关键词】数据质量控制；故障数据；数据质量评价；ITS 检测器

1、引言

近年，尽管我国各大道路交通基础设施建设取得了较大成就，但是机动车和驾驶员的高速增长使交通管理工作承担了沉重的压力。与此同时，人民群众对交通信息的需求逐步向多元化、便捷化、高效化的模式转变。然而，要提供高质量的信息服务，就必须以高质量的数据为基础。通常检测设备直接获取的交通流数据层次差异大、有效性迥异、周期不同，无法真实反应交通流实际的运行情况。面对庞大而复杂的动态交通数据，如果不进行数据质量控制，既不利于从中找出规律，也不利于数据存储和利用。

为体现多元属性的综合性概念，正确率已不再是衡量数据质量的惟一标准。因此，将数据质量定义为数据能否满足用户需求并较为准确的反应客观事实的程度，这样的相对性概念更能符合实际需求。交通流数据质量好坏直接关系到能否为路网服务水平评价等提供可靠的数据保障，是智能交通系统发挥其应有作用的一个必要前提，也是城市交通管理指挥调度平台的重要依据[1]。因此，对交通数据进行质量控制和评价已经成为非常关键并亟待解决的问题。

2、交通流数据质量控制

2.1 概述

进行质量控制不仅可以提高交通信息服务平台发布数据的精度，也能使交通管理者直观并快速的了解城市道路交通运行状况，以给出正确的决策。

在整个数据处理过程中，质量控制主要针对数据属性，将海量原始数据中的故障数据，即异常、不规则时间点和丢失的进行修正与补齐，必要时还需要对故障检测器的判别。通过数据质量控制可以将原始数据输出为具有正确时间、空间分布和有效参数值的交通流数据，以方便后续研究和应用。

2.2 故障数据处理

对于交通流数据质量控制而言，主要目的就是提高数据精度。将采集到的交通流数据分为正确数据和故障数据，故障数据包括丢失数据、不规则时间点数据和异常数据。

现有的研究中，一般只笼统的给出产生故障数据的几种情况[2]，并未深入的挖掘分析其源头。本文针对浮动车、微波和牌照法三种常用检测器在获取数据的整个过程中可能产生故障数据的原因和优缺点进行了比较，如表 1 所示。

表 1 检测器产生故障数据原因及优缺点比较

	浮动车	微波	牌照法
数据丢失	无法匹配到地图上；浮动车数量不足；	硬件故障、噪声干扰、通讯故障	天气、灯光等影响
不规则时间点数据	GPS 漂移、传输过程	硬件故障	设备不足
数据异常	不正常行驶、急停；定位精度不高；道路条件；	车流密度过大、遮挡、有车辆停在检测点	天气、灯光等影响
优点	全天候、大范围的采集，获取动态数据	精确度较高、性能稳定	提供可视图像
缺点	涉及信息传输和安全等的维护，存在 GIS 系统不稳定	不适用于拥堵以及大型车较多、车型分布不均匀的路段	造价高，易受天气、灯光、阴影等环境因素影响，晚上报误率高

2.2.1 丢失数据处理

数据丢失指某一次上传无法获得应有的理论数据样本，表现为关键交通流参数缺失，可能是因为数据无法获取或操作过程中被遗漏，这种缺失是不可避免的。以北京为例，采集器每 2min 采集一次，即每天每个地点每条车道应产生 720 条数据。丢失数据的处理主要有通过交通流历史记录或者曲线拟合的方式近似复现丢失数据；根据导致丢失数据时间段的长短和补缺丢失数据的数据来源不同进行补缺[3]：①对于 1min 内丢失的，优先采用同一检测器临近时间的数据平均；②对于 1h 内丢失的，优先采用同一时间同一检测器采集的不同同向车道的数据；③对超过 1h 的，采用标准值+扰动因子方法修补历史同期数据。

2.2.2 不规则时间点数据处理

一些数据因为传输问题导致后台数据无法按照反馈频率(2min)或数据编号的顺序获取，即为不规则时间点的数据。为了使得到的数据可以后续同步应用，减少后期不必要的工作量，也为了方便准确的发现缺失数据，需要将所采集到的原始交通流信息样本中的时间记录值字段与正常情况下采集时间点对比。将所有数据的时间点与各自最邻近的标准时间点作差之后，得到左偏和右偏差值[4]，使之符合横向时间序列（同一地点一天 24 小时数据排列）和纵向时间序列（同一地点每天同一时刻数据）。

2.2.3 异常数据

异常数据指不在期望的范围内或不满足已有的原理与规则（如交通流理论），在检测周期中发生突变，不符合客观事实和逻辑的无效数据。

异常数据判别是对原始数据中交通流参数记录值进行检验，由于无效的概念比较模糊，相对不易判别。异常数据判别方法有交通流参数合理阈值原理、交通流机理以及阈值原理和交通流机理结合的 3 种[5]。此外，还需判别异常数据是否为事件数据（事故、道路维修等），若为事件数据则无需修正。

异常数据的修正方法大多以回归、线性插值和加权移动平均等传统时间序列预测方法为主[6]。目前在应用中采用对于不满足阈值理论的数据，用阈值替代异常数据，但是不区分时间段和路段给予统一较高阈值来替换会导致修正速度偏大，不能准确的反应道路实际情况。考虑到浮动车可以获取大量连续数据，所以采用线性插值法或历史数据平均值替换更为准确和简便。

2.2.4 故障检测器的判断

有时路段会出现数据连续周期较大偏离正常值，此时应考虑对检测器进行故障判别，尽量减少错误数据的产生。常用的判断方法有以下几种：①连续交通数据格式错误；②连续交通数据丢失或异常；③连续交通数据重复传入；④所丢失或异常数据样本占总量的比例较高；⑤夜间交通数据拥堵样本占夜间或全天交通样本比例较高[6]。

如果有符合以上情况的数据，则应该根据数据编号对应检查固定检测设备是否故障，根据发回故障数据的车辆信息，由出租车公司检查相应浮动车 GIS 设备。

3、交通流数据质量评价方法

数据质量评价要具有实际可操作性，就必须对数据质量进行量化，量化的第一步是制定出合理的评价指标。旅行时间数据精确度和完整度均可以达到 95%以上，故将其作为评价数据质量的参照。

现有的质量评价方法主要是进行宏观分析[7]，但是考虑到路网中各个路段的指标可能不同，从交通信息发布和交通管理指挥部门的角度来说，微观分析同样重要。本文以相邻两个微波检测器（约 500 米）为一个 Link 进行评价，通过分析引起交通数据质量降低的主要原因，建立评价体系如下：

① 时效性：衡量数据及时反应道路交通状态的能力。定性评价方法如表 2 所示。

表 2 时效性量化

时效性|时间间隔	采集时间间隔	发布时间间隔
良好	0.5-1 min	<5 min
较好	1-2 min	5-10 min
一般	2-5 min	10-15 min
较差	>5 min	>15 min

② 准确度：数据反映实际交通状态的能力，即单个或数据组与正确数据的相符程度。表示为检测数据值/参照数据值的百分比。原则上应考虑节假日、工作日、不同天气等因素的影响。

③ 完整度：数据在时间和空间上反映道路交通状态的程度，即同一个检测器实际获取的样本量占理论应获取样本量的比例。

④ 覆盖度：同种检测器所采集的数据量覆盖对象路段或整个路网的程度。当不满足覆盖度指标时，考虑用历史相同情况下（天气、是否工作日等）的数据替代。

⑤ 有效度：衡量异常数据的比例，即路段或路网所有同种检测器所采集到的有效数据

样本量占总样本量的比例。

⑥ 可信度：面向交通管理部门和交通参与者两种用户[8]，定性衡量是否能满足用户需求，可作为对数据质量评价结果的回访核查。

4、交通流数据质量评价过程

关于交通流数据质量评价的研究目前主要是关于数据预处理或是评价指标的理论探索。本文的创新点在于，首次提出分三次对数据质量分别进行评价，即针对原始数据、质量控制后数据、将多种采集方法融合后的数据，并以旅行时间数据作为质量评价的标准。详细质量控制和评价流程图如图 1 所示：

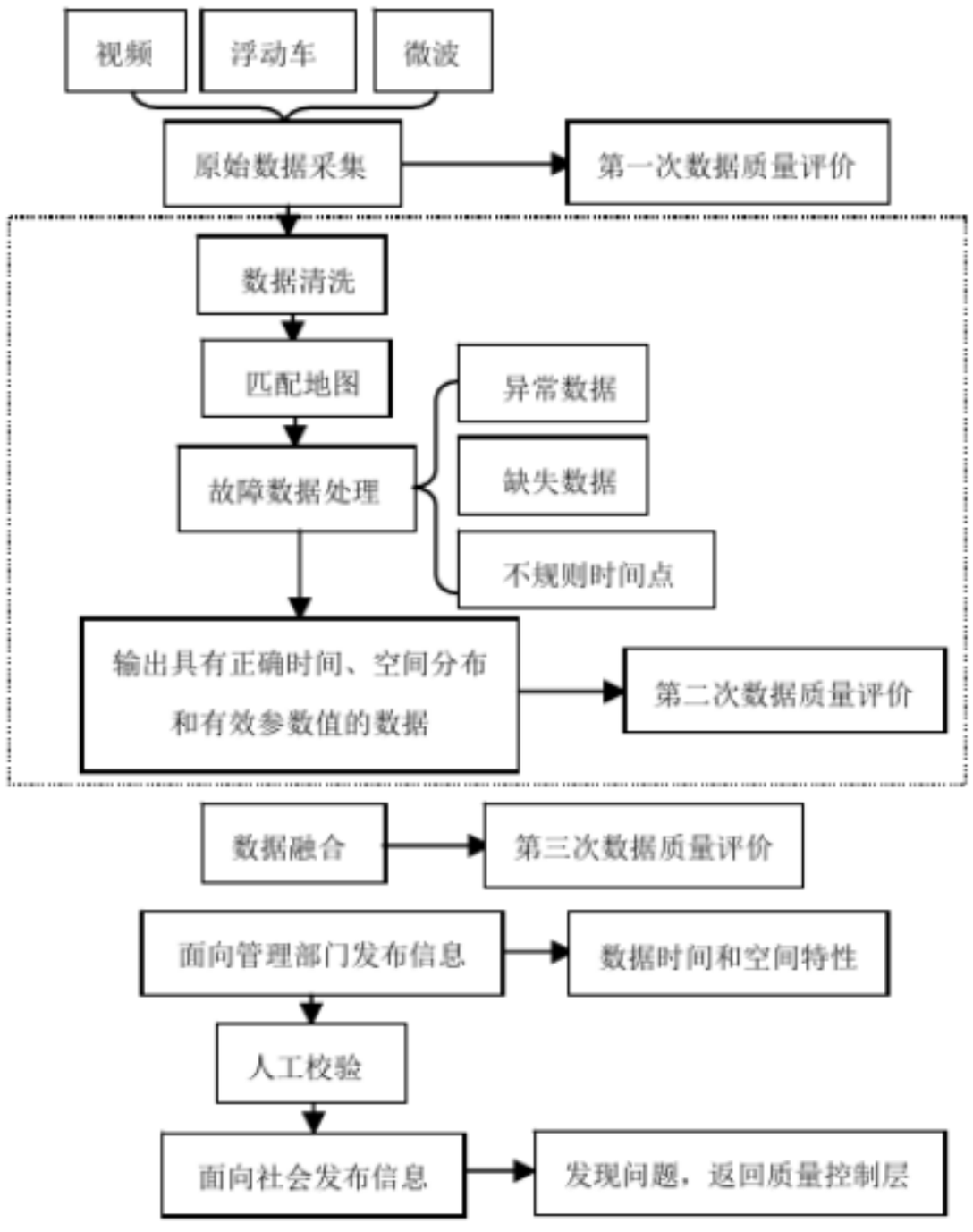

图 1 交通流数据质量控制和评价总流程图

本文对数据进行全面的评价分析，其意义在于通过原始数据精度可以从检测器和传输过程等源头改善数据来源，质量控制后的数据精度可以检验控制方法是否有效，而融合后的精度便于信息发布和后续应用，可以为管理部门实现更智能化和科学的交通管理提供信息支撑。

5、以北京市交通流数据质量评价为例

现以北京市二环西直门桥至阜成门桥 2013 年 4 月 22 日（周一、晴天）的数据为例，展现本文关于交通流数据质量评价的效果。该路段全长 2.8 公里，双向 6 车道，设有微波检测器 4 个（HI2105a、HI7053a、HI7064a、HI7030a），视频检测器 1 个，对应旅行时间路段为

LD00701。由于晚 10 点至早 6 点数据采集易出错且交通畅通，故不列入考虑。6 点至 22 点的视频、浮动车和微波原始数据如图 2、图 3 和图 4 所示：

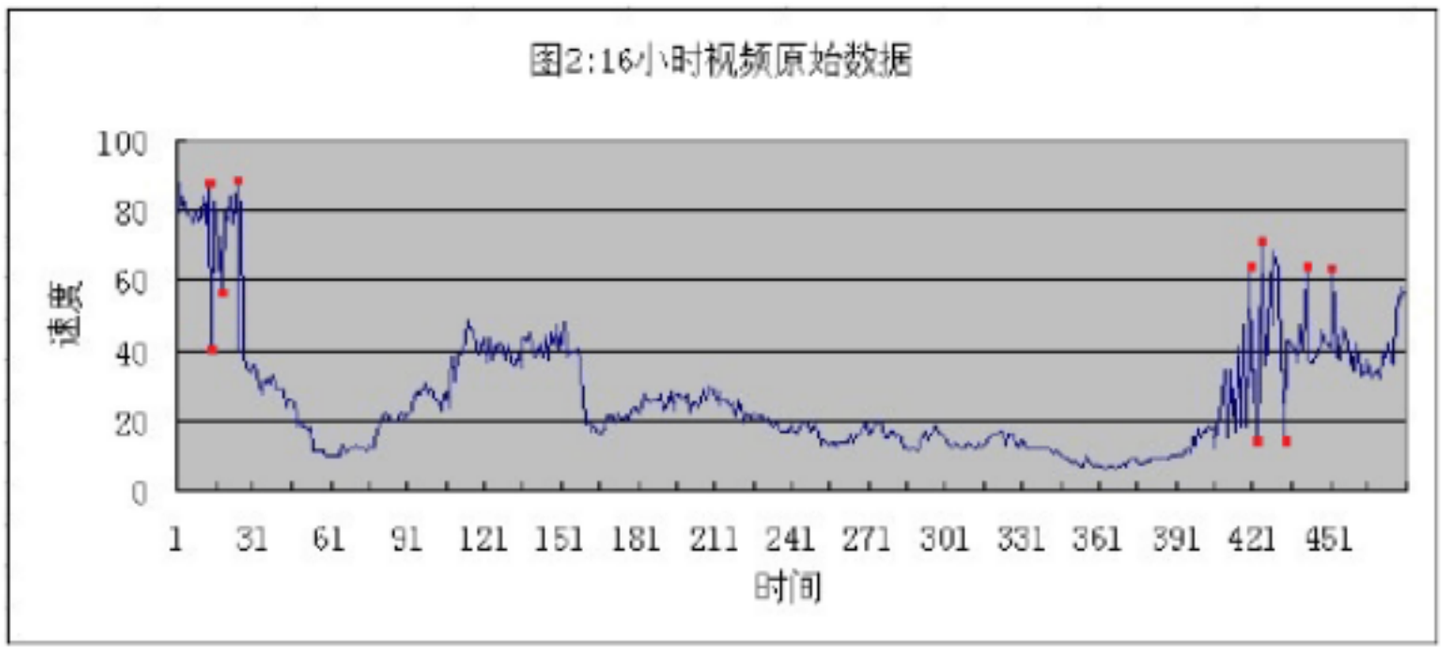

图 2 16 小时视频原始数据

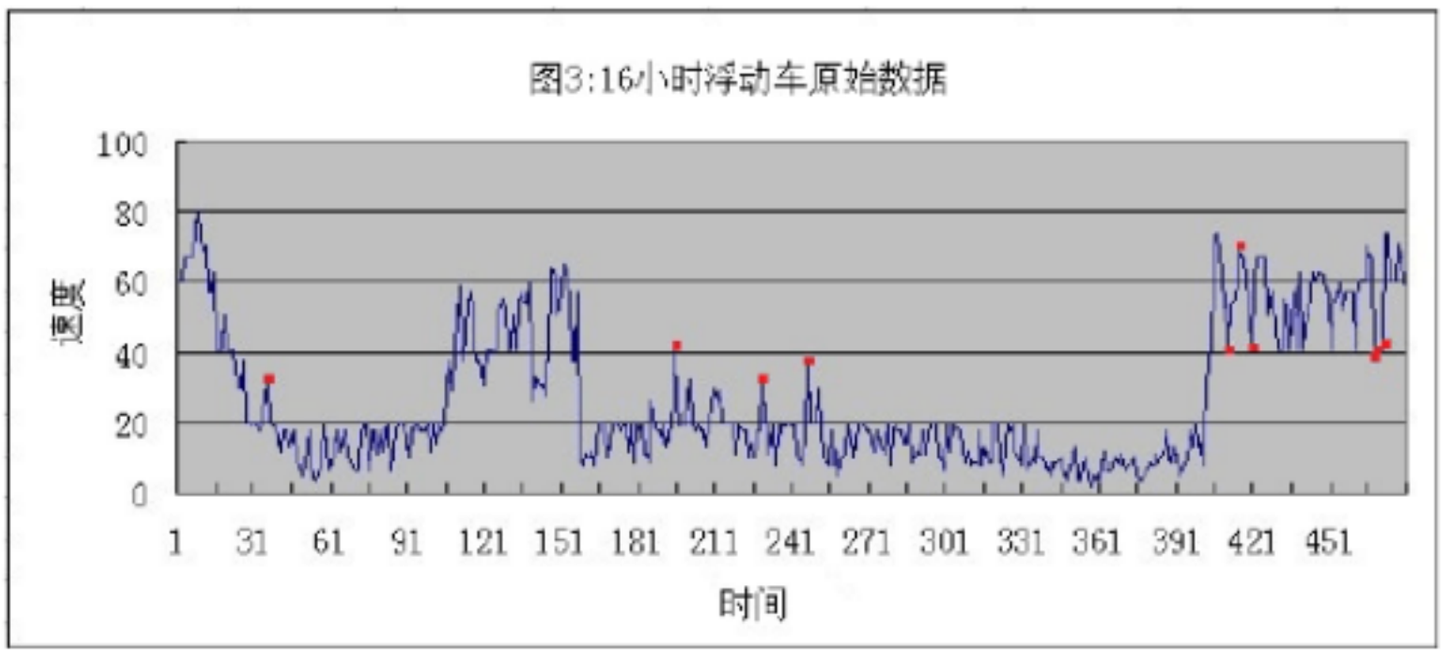

图 3 16 小时浮动车原始数据

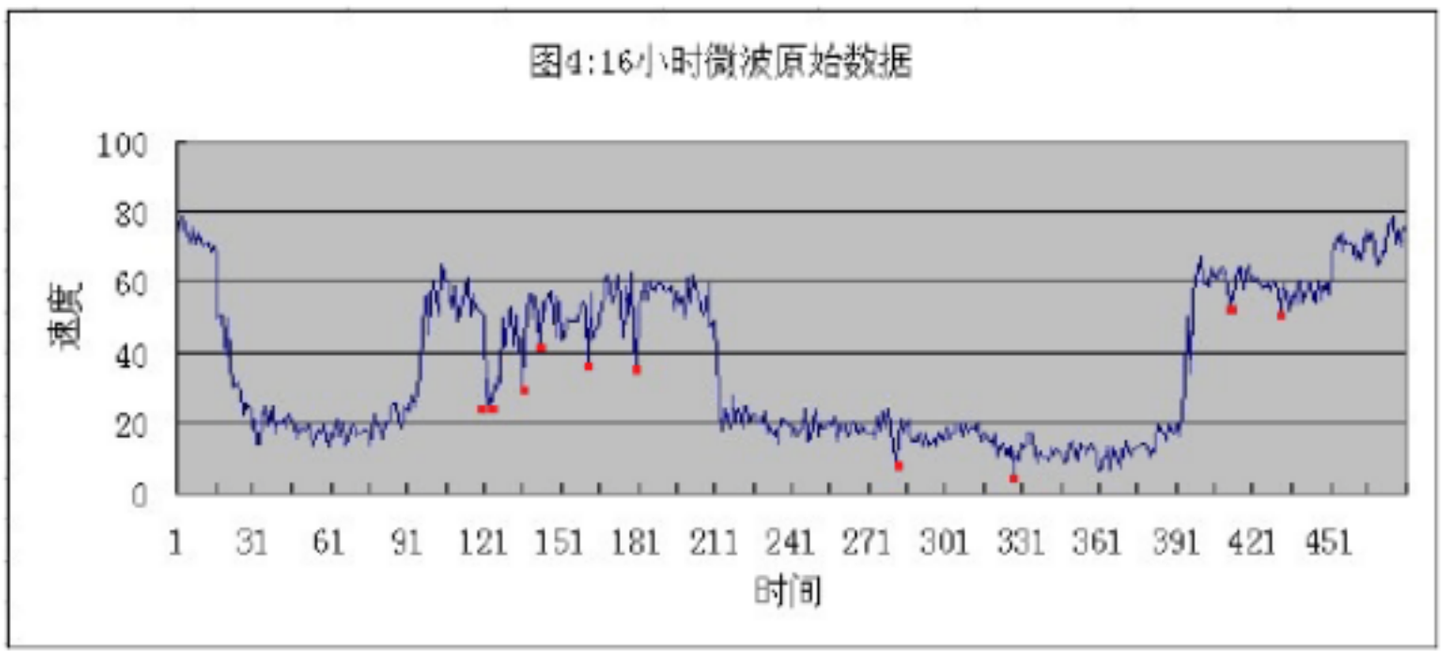

图 4 16 小时微波原始数据

根据质量控制的流程和方法来估计缺失值和正确的数据，并把视频数据作为参照，对原始数据和质量控制后的数据评价结果如表 3 所示：

表 3 评价结果（括号内为原始数据的结果）

	视频（参照）	浮动车	微波
准确度	100%（94.3%）	72.53%（70.02%）	73.48%（71.87%）
完整度	100%	100%（98.61%）	100%（79.3%）
覆盖度	100%	40%	100%
有效度	100%（96.87%）	94.16%（88.95%）	98.33%（97.5%）

用一周工作日（4 月 15 日至 19 日）的数据进行 BP 神经元训练[9]，训练图如图 5 所示：

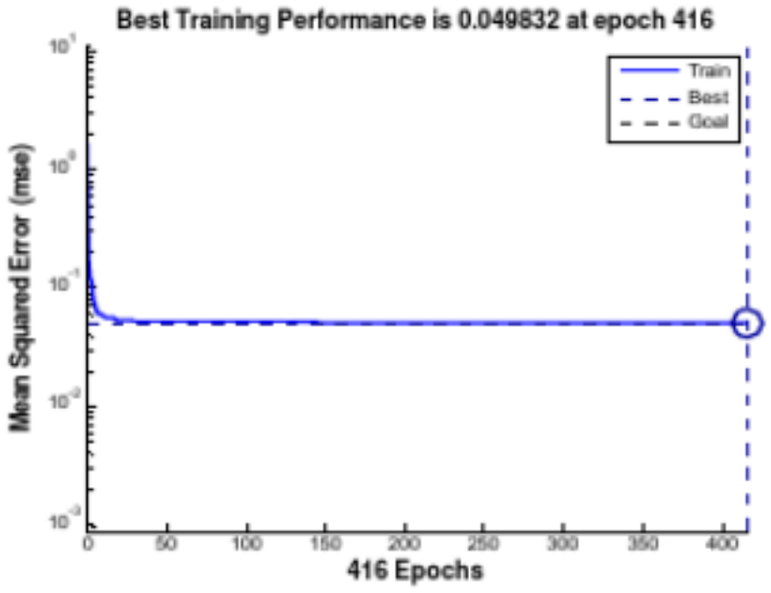

图 5 神经元训练图

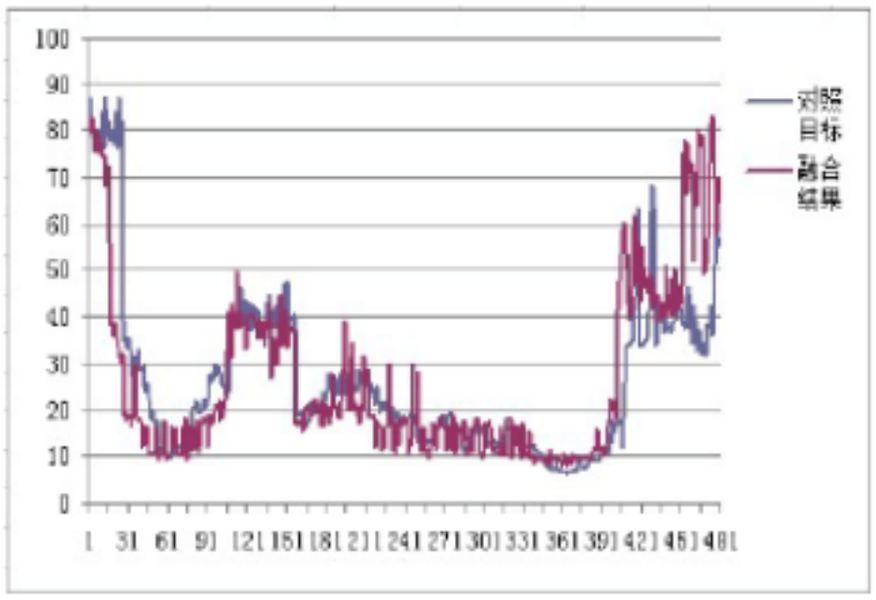

图 6 4 月 22 日融合结果

把微波和浮动车数据作为输入，以视频数据作为融合目标，4 月 22 日的数据融合结果如图 6 所示。

通过本文案例可知，在第一阶段，进行质量控制可以使数据完整有序；融合后能使数据非常接近真实的数据，平均误差仅 3.779%。此外，视频数据在夜间容易出现缺失和高误差，但是在高峰期可以匹配到更多车辆以获得准确的速度。微波和浮动车数据普遍偏高。在未来的研究应该寻找更简单和快捷的方法来进行质量控制和数据融合。

6、小结

动态交通数据质量评价可为交通管理系统提供可靠的支持。本文针对故障数据，提出了交通流数据质量的控制方法和流程。同时，第一次提出分三阶段分别对不同 ITS 检测器的交通数据质量进行评价，并构建了六维评价指标体系。最后以北京市二环路段交通流数据为例，对本文控制流程和评价方法进行了分析与验证。结果表明，该方法对评价指标的提高有明显的改进效果，并且可以很容易地直接应用于实际工作。

【参考文献】

[1] Turner S.M. -Archived Intelligent Transportation System Data Quality: Preliminary Analyses of San Antonio Trans Guide data [J]Transportation Research Record, 2000:13-19.

[2] 牛世峰，姜桂艳.交通数据质量宏观评价与控制方法 [J]《公路》2012,(12):119-123.

[3] 孙亚，朱鲤.ITS 检测器交通流数据质量控制系统研究 [J]《测控技术》2008,(7):35-39.

[4] 耿彦斌，于雷，赵慧.ITS 数据质量控制技术及应用研究[J]《中国安全科学学报》2005,(1):82-87.

[5] 姜桂艳，江龙晖，张晓东，王江锋．动态交通数据故障识别与修复方法[J]《交通运输工程学报》2004,(1):121—125.

[6] 韩卫国,王劲峰,胡建军.交通流量数据缺失值的插补方法[J]《交通与计算机》2005,23(1): 39-42.

[7] 孙亚.定点采集信息数据质量控制理论与方法研究[C]《第十六届海峡两岸都市交通学术研讨会论文集》2008:646-650.

[8] 常永.交通流数据质量预控制的实施对策研究[J]《科技传播》2011,(12):41-44.

[9] 聂庆慧,夏井新,张韦华.基于多源 ITS 数据的行程时间预测体系框架及核心技术[J]《东南大学学报(自然科学版)》2011.(1):199-204.

【作者简介】

谷远利，男，博士，北京交通大学交通运输学院,副教授。电子信箱：ylgu@bjtu.edu.cn

马韵楠，女，北京交通大学交通运输学院，硕士研究生。电子信箱：791053264@qq.com

李巨伟，男，硕士，北京市公安局公安交通管理局，高级工程师