DOI:10.13471/j.cnki.acta.snus.2019.04.015

基于 NB 和 DTW 组合模型的稀疏浮动车数据填充*

许伦辉，陈凯勋，郭雅婷
（华南理工大学土木与交通学院，广东 广州 510641）

摘 要：利用浮动车数据对交通状态进行挖掘和分析已经是交通领域的重要研究手段之一，而浮动车数据的稀疏性给数据挖掘工作带来了阻碍。对路网数据缺失特性进行分析后，提出了稀疏浮动车数据的填充模型，模型中对缺失小于30%的偶然性数据缺失采用基于朴素贝叶斯（NB）的分类方法进行数据填充，对缺失大于30%的多发性数据缺失采用基于动态时间规整（DTW）的方法进行二次填充，将两种方法结合并应用于路网实例中，对路段缺失的交通流速度数据进行填充，提升了浮动车数据的路网覆盖率，有效降低了浮动车数据缺失对交通流速度信息的采集、发布以及预测工作带来的影响。

关键词：浮动车；数据缺失；朴素贝叶斯；动态时间规整；速度估计
中图分类号：U491.2 **文献标志码**：A **文章编号**：0529－6579（2019）04－0136－10

Sparse floating car data filling based on NB and DTW combined model

XU Lunhui, *CHEN Kaixun*, *Guo Yating*
(South China University of Technology, School of Civil Engineering and Transportation, Guangzhou 510641, China)

Abstract: It has been one of the essential study methods to analyze traffic condition via the floating car data (FCD) in transportation field. The sparsity of FCD, Nevertheless, is a hindrance to researches. After analyzing the data missing features of road network, this paper put forward a filling model for the sparse FCD, which fills the random data missing based on Naive Bayes classifier and double fills the multiple data missing based on dynamic time warping. By combining the two methods mentioned above and applying them to field cases, filling the roads' traffic flow velocity data missing can substantially increase network coverage rate of FCD and reduce the impact on collecting, launching and forecasting the traffic flow velocity data brought by the FCD missing.

Key words: floating car; data missing; Naive Bayes; dynamic time warping; speed estimation

浮动车数据（floating car data）是指从安装了GPS模块和无线通信模块的出租车、公交车等车辆接收到的数据，数据可包括时间、经纬度、车辆编号、运营状态、速度等信息。浮动车在20世纪90年代初开始投入使用，至今已经被广泛应用于交通领域。浮动车具有安装使用方便、维护成本低、覆盖面广、数据量大、精度高且实时性强的特点。文献［1］以浮动车数据为单一数据来源，提出了基于广义自适应平滑方法（GASM）的在线交通流速度估计模型。文献［2］引入置信点，提出了基于

* **收稿日期**：2018－09－27
基金项目：国家自然科学基金（61263024）；广东省自然科学基金（2015A030313797）
作者简介：许伦辉（1965 年生），男；**研究方向**：交通系统建模与分析；E-mail：lhx_scut@163.com
通信作者：陈凯勋（1993 年生），男；**研究方向**：交通系统建模与分析；E-mail：651423076@qq.com

浮动车大数据的城市道路复杂度分析方法，并将该方法应用于广州市路网的复杂度分析。文献［3］提出了基于浮动车数据的城市不同等级道路通行能力评估方法，以较低的成本全面掌握城市交通状态在空间和时间上的不同特征。文献［4］利用道路实测数据建立了面向其应用的轻型车尾气排放实时测算模型，反映相同平均速度下的排放变化。文献［5］提出一种基于 HMM（Hid-den Markov model）的地图匹配模型，并有效应用于海量浮动车数据地图匹配。但，以上研究多基于完善的浮动车数据，所提出的模型在浮动车数据缺失的情况下并不适用。近年来，各国学者也提出了一些解决浮动车数据稀疏问题的方法。文献［6］对比了 PPCA 和 KPPCA 方法，在适当考虑时空相关性的前提下，证明了 KPPCA 方法显著减少了对缺失数据估算的误差。文献［7］将季节性 ARIMA 交通流预测模型用于计算交通流数据的缺失，改善了对稀疏数据的不适应性。文献［8］将数据丢失事件看作是严重堵塞的情况下防止浮动车进入拥堵区域的状况，以此建立了一种基于新的交通流运行指标的时间序列相似性测量方法，并将此方法作为定义动态交通网络瓶颈的基础。文献［9］提出了一种估计和修正方法对行程时间进行估计与修正，得到了比改进的内插值法更接近实际的估计值。文献［10］建立了一个微仿真模型，该模型考虑了多个误差因素，通过对已有的稀疏浮动车数据驱动进行增强来估计交叉路口的信号相位和时间。文献［11］提出了一个三层神经网络模型，从浮动车的历史数据中获取各路段的时空特性，以时空特性作为输入，建立路段间的关联关系来估计稀疏路段的行程时间。文献［12］提出了一个朴素贝叶斯估计模型对缺失的路网交通流速度数据进行估计，在一定程度上缓解了数据稀疏带来的影响。

上述研究将偶然性和多发性的缺失统一为单一缺失状态进行填充，在实际的应用中会出现对特殊缺失情况的不适应，从而导致填补效率降低。例如：文献［12］所提出的模型在以时间维度的参数进行估计时，会因某个时段历史数据的大量缺失而对估计效果产生很大影响，导致模型不再适用于此类型缺失情况。基于对真实数据缺失情况的分析，本文将数据缺失类型归纳为偶然性缺失和多发性缺失，并针对两种数据缺失类型分别提出了基于朴素贝叶斯（NB）和基于动态时间规整（DTW）的方法对路网交通流速度进行估计，以增强缺失数据的填充效果以及模型在实际应用中的适用性，降低路网浮动车数据缺失对实际应用和理论研究带来的影响。

1　稀疏数据填补方法

1.1　数据处理

将地图转化为路网并链接浮动车数据，选择适合区域进行数据提取和分析。对于实时交通系统，5 min 是常用的信息发布时间间隔，本文以 5 min 作为数据采集间隔，将一天分为 288 个时段进行采集与处理分析。以一条浮动车数据为例，其数据格式为：

由表 1 可知，若该路段某一时段的浮动车数据丢失，速度数据也必然随之丢失。以缺失速度数据的路段 ID 代替交通流速度作为索引对数据进行实时监测，可简化缺失数据的查找过程且不影响数据的估计和填补工作，转化过程可用式（1）表示。

$$\boldsymbol{V}_t=\begin{bmatrix} v_{11} & \cdots & v_{1c} \\ \vdots & \ddots & \vdots \\ v_{r1} & \cdots & v_{rc} \end{bmatrix}\Rightarrow \boldsymbol{M}_t=\begin{bmatrix} m_{11} & \cdots & m_{1c} \\ \vdots & \ddots & \vdots \\ m_{r1} & \cdots & m_{rc} \end{bmatrix} \tag{1}$$

表 1　浮动车数据格式

Table 1　Format of floating car data

符号	定义	数据内容	备注
TIME	日期	20170507	2017 年 5 月 7 日
PERIOD	时间段	37	第 37 个时间段,185 ～ 190min
LINKID	路段 ID	380129	路段唯一标识
FROMNODE	来自节点	243022	上一节点标识
TONODE	去向节点	254150	下一节点标识
GOSPEED	路段平均速度	42.936	该路段的交通流平均速度,km/h

式中，$\boldsymbol{V}_t$ 为路网交通流速度矩阵；v_{rc} 为道路 r 在 c 时段的交通流速度数据；$\boldsymbol{M}_t$ 为路网各路段 ID 矩阵；m_{rc} 为道路 r 在 c 时段的路段 ID，若数据缺失，则路段 ID 为 0。

1.2 偶然性缺失填补方法

某一路段某一时刻的数据缺失具有偶然性，即历史数据中该路段这一时段的数据较为完整或仅有少量缺失，实时查找出的缺失是偶然发生的，这样的缺失多发生于较高等级的道路上。朴素贝叶斯分类方法简单、快速，对于此类数据缺失有较好的填充效果。

1.2.1 朴素贝叶斯模型 本文采用朴素贝叶斯方法对缺失数据进行估计，以平衡浮动车数据在空间分布不均的问题。此方法中，速度数据从 0 km/h 开始每 5 km/h 为一组分为 20 个速度组，大于 100 km/h 的计入第 20 组，将速度转化为 20 个离散值。用同一路段历史交通流速度 v_H（同一时段数据平均值）和同一路段上一时段交通流速度 v_S 作为朴素贝叶斯分类方法的两个输入参数，估计当前时段当前路段的交通流速度 v，公式如下：

$$y = f(v) = \arg\max_v P(v \mid v_H, v_S) = \arg\max_v \frac{P(v_H, v_S \mid v)P(v)}{\sum_{v \in V} P(v_H, v_S \mid v)P(v)} \tag{2}$$

式中，v 为研究路段某时段的估计平均交通流速度；$P(v)$ 为研究路段某时段的估计平均交通流速度为 v 时的概率；$P(v \mid v_H, v_S)$ 为研究路段历史交通流速度为 v_H 且上一时段交通流速度为 v_S 的情况下，该路段当前时段估计平均交通流速度为 v 的概率；$P(v_H, v_S \mid v)$ 为研究路段估计平均交通流速度为 v 的情况下同一路段历史交通流速度为 v_H，且同一路段上一时段交通流速度为 v_S 的概率。

由于两个引入的参数相互独立，即其上一时段交通流速度与历史交通流速度无直接联系，而上式中的分母 $\sum_{v \in V} P(v_H, v_S \mid v)P(v)$ 对所有的 v 都相同，在比较后验概率时可以省略其进行计算，仅比较分子大小即可，所以在某时段缺失路段的交通流速度估计公式如下：

$$y = f(v) = \arg\max_v P(v \mid v_H, v_S) = \arg\max_v P(v_H \mid v)P(v_S \mid v)P(v) \tag{3}$$

式中，$P(v_H \mid v)$ 为研究路段估计平均交通流速度为 v 的情况下同一路段历史交通流速度为 v_H 的概率；$P(v_S \mid v)$ 为研究路段估计平均交通流速度为 v 的情况下同一路段上一时段交通流速度为 v_S 的概率。各参数的计算过程如式(4) ~ (6) 所示。

$$P(v_S \mid v) = \frac{\sum_{i=1}^{N} I(v_S, v_i = v)}{\sum_{i=1}^{N} I(v_i = v)} \tag{4}$$

式（4）中，N 为前 30 天内用于研究的所有速度数据量（第一个时段无 v_S，此时仅使用 v_H 作为模型参数）；$I(v_i = v)$ 为判断 $v_i = v$ 的布尔值，若 $v_i = v$ 则 $I(v_i = v) = 1$，否则为 0；$I(v_S, v_i = v)$ 为判断 $v_i = v$ 时 v_S 单个统计量的布尔值，若 $v_i = v$ 时存在 v_S 则 $I(v_S, v_i = v) = 1$，否则为 0。

$$P(v_H \mid v) = \frac{\sum_{i=1}^{N} I(v_H, v_i = v)}{\sum_{i=1}^{N} I(v_i = v)} \tag{5}$$

式（5）中，N 为一天内用于研究的所有时段；$I(v_i = v)$ 为判断 $v_i = v$ 的布尔值，若 $v_i = v$ 则 $I(v_i = v) = 1$，否则为 0；$I(v_H, v_i = v)$ 为判断 $v_i = v$ 时 v_H 单个统计量的布尔值，若 $v_i = v$ 时存在 v_H 则 $I(v_H, v_i = v) = 1$，否则为 0。

$$P(v) = \frac{\sum_{i=1}^{N} I(v_i = v)}{N} \tag{6}$$

式（6）中，N 为前 30 天内用于研究的所有速度数据量；$I(v_i = v)$ 为判断 $v_i = v$ 的布尔值，若 $v_i = v$ 则 $I(v_i = v) = 1$，否则为 0。将式（3）中的参数逐一计算后，代回式（3）中关联并遍历，可得到最终的估计结果。

1.2.2 朴素贝叶斯模型修正 假设选用参数 v_H、v_S 与速度 v 为正相关关系，并在模型可行性分析部分进行验证，此正相关关系使得速度相差较大的范围会出现 $P(v_H \mid v) = 0$ 或 $P(v_S \mid v) = 0$。为了避免出现概率值为 0 而影响最终估计结果的适应性和准确性，本文采用拉普拉斯修正（Laplace correcting）对存在条件概率进行修正处理，如下式（7）~（9）所示。

$$P(v_H \mid v) = \frac{\sum_{i=1}^{N} I(v_H, v_i = v) + 1}{\sum_{i=1}^{N} I(v_i = v) + |M|} \tag{7}$$

$$P(v_S \mid v) = \frac{\sum_{i=1}^{N} I(v_S, v_i = v) + 1}{\sum_{i=1}^{N} I(v_i = v) + |M|} \tag{8}$$

$$P(v) = \frac{\sum_{i=1}^{N} I(v_i = v) + 1}{N + |M|} \tag{9}$$

式中，M 为速度取值数量。

1.3　多发性缺失填补方法

某一路段某一时刻的数据缺失具有多发性，即历史数据中该路段这一时段的数据同样存在大量的缺失，这样的缺失多发生于低等级的道路上。这种情况下历史数据的估计价值不高，估计结果也有较大偏差甚至无法进行估计，朴素贝叶斯分类方法不再适用于这种缺失类型。近年来，已有学者将相似性度量的方法应用于信息补全、短时预测等研究。文献［13］提出了一种基于时间序列相似性搜索的短时交通流量预测方法，并以高速公路交通流量数据证明该方法的预测性能优于传统方法。在相似性度量中，DTW 算法已在语音识别等领域得到广泛应用，此算法支持不同长度的时间序列的相似性度量，定义了序列间的对齐匹配关系[14]。

本文提出了基于 DTW 的相似度算法，求解匹配时累计差值最小的两路段所对应的规整函数，以寻找最相似路段，并对多发性的数据缺失进行填充。在寻找相似路段前，使用 Daubechies（db4）小波获取用于比对的低频基波，4 是小波的阶数，假设观测数据由 2 部分组成：

$$x(t) = s(t) + n(t) \tag{10}$$

式中，$x(t)$ 为观测值；$s(t)$ 为真实值；$n(t)$ 为噪声，即 $n(t) = N(0,\delta^2)$。

在小波分解后的低频信号中调取源路段，并标记为序列 $Q = q_1,q_2,\cdots,q_i,\cdots,q_n$，其中 n 为调取到的该路段交通流速度数据总量。依次调取研究范围内其他所有路段的低频信号，记为 $C_{\text{LINKID}} = c_1, c_2,\cdots,c_j,\cdots,c_m$。其中，LINKID 为当前调取的路段 ID，$m$ 为调取到的其他路段的交通流数据总量。要计算每个搜寻路段与源路段序列的累计差值，首先要对序列实时匹配差值进行计算，具体公式如下：

$$D[i,j] = d(q_i - c_j) + \min\begin{cases} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{cases} \tag{11}$$

式中，$D[i,j]$ 为序列匹配至 q_i，c_j 时的累计差值；$d(q_i - c_j)$ 为两数据点间的欧氏距离。当计算过程执行到序列的最后一个样本时，可得搜寻路段与源路段序列的累计差值：

$$\text{DTW}(Q,C) = D[n,m] \tag{12}$$

使用 27° −45° −63°的局部路径约束，最佳的 DTW 路径会跳过序列中的离群点[15]，使得在计算时省掉了一些不需要计算的部分，如图 1 所示。计算完成后，比较累计差值可得出源路段的相似路段，相似路段的低频信号和随机噪声信号最终组成源路段的缺失时段数据，并可将估计的速度转化为对应的分组数值，填充入数据缺失部分。

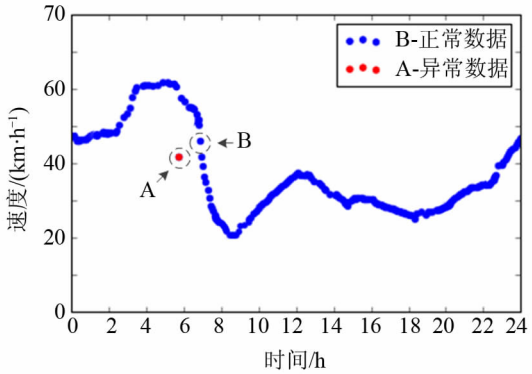

图 1　离群点示意

Fig. 1　Diagram of outliers

2　试验分析

选择深圳市南山区南海大道、滨海大道、创业路及后滨海路所框定的范围作为研究区域，如图 2 所示。所选区域覆盖了各等级道路包括快速路、主干路、次干路、支路。调取的浮动车样本数据均来自深圳市城市交通规划设计研究中心，数据时间范围为 2018 年 5 月 1 日至 5 月 31 日，样本总量约 400 万条，采样间隔为 5 min。在 Matlab 平台下，对朴素贝叶斯算法的参数相关性和 DTW 算法的相似性进行评价，并对组合模型的效率和准确性进行了验证，并将最终的数据填充效果通过地图匹配展示到 TransCAD、GIS 等软件的路网分布图中。

图 2　研究区域路网

Fig. 2　Network of study area

2.1 模型可行性分析

2.1.1 朴素贝叶斯参数评价 以选定范围内快速路的实时交通流速度 v 所处分组为横坐标，同一路段前一时段的交通流速度 v_S 所处分组为纵坐标做相关性分析，结果如图 3（a）所示。图 3（a）中，两个变量呈现出高度的正相关性，说明相邻时段的速度相近甚至相等是大概率事件，用相邻时间段的路网交通流速度数据作为估计缺失数据的一个条件参数是有效且可行的。

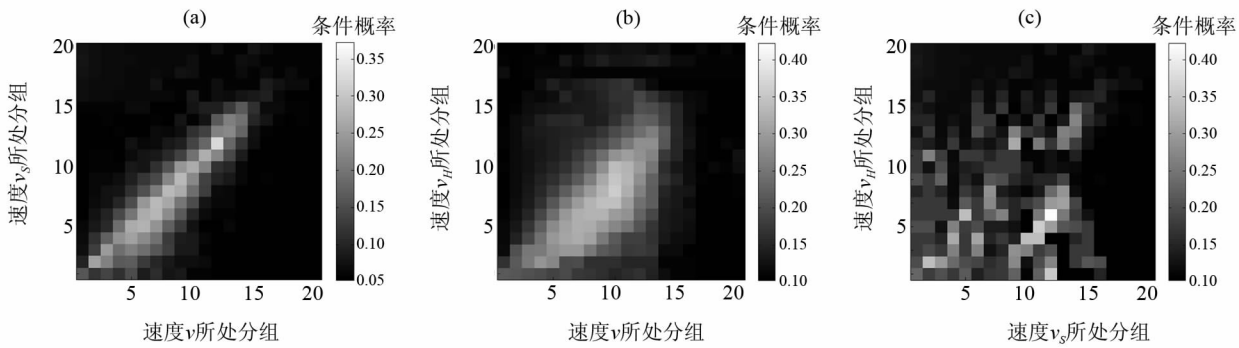

图 3 条件参数与交通流实时速度的相关性

Fig. 3 Relationship between v_S, v_H and v

以选定范围内快速路的实时交通流速度 v 所处分组为横坐标，同一时间段的历史速度 v_H 所处分组为纵坐标做相关性分析，结果如图 3（b）所示。图 3（b）中，高概率色格仍然更多集中于对角线附近，虽与图 3（a）相比更为分散，但仍能看出其正相关性。造成概率集中度不高的原因主要是历史数据提取量有限，仅引入前 30 d 的历史数据进行相关条件的概率计算，对结果产生了轻度的干扰。扩大样本容量进行处理分析，可得出分布更加集中，更加接近正相关的图形。从图 3（c）可以看出：同一路段前一时段的交通流速度 v_S 与同一时间段的历史速度 v_H 相互独立，没有相关关系。

2.1.2 DTW 相似性评价 在使用 DTW 算法前，先调取所有路段前 30 d 的历史交通流速度数据并对其进行小波滤波分析[16]，采用 Db 小波的 4 阶函数进行小波分解，将噪声部分表示为 4 组高频信号，对主流低频变化信号以及高频噪声信号进行重构，得到如图 4 所示的原始序列、低频序列、噪声序列。

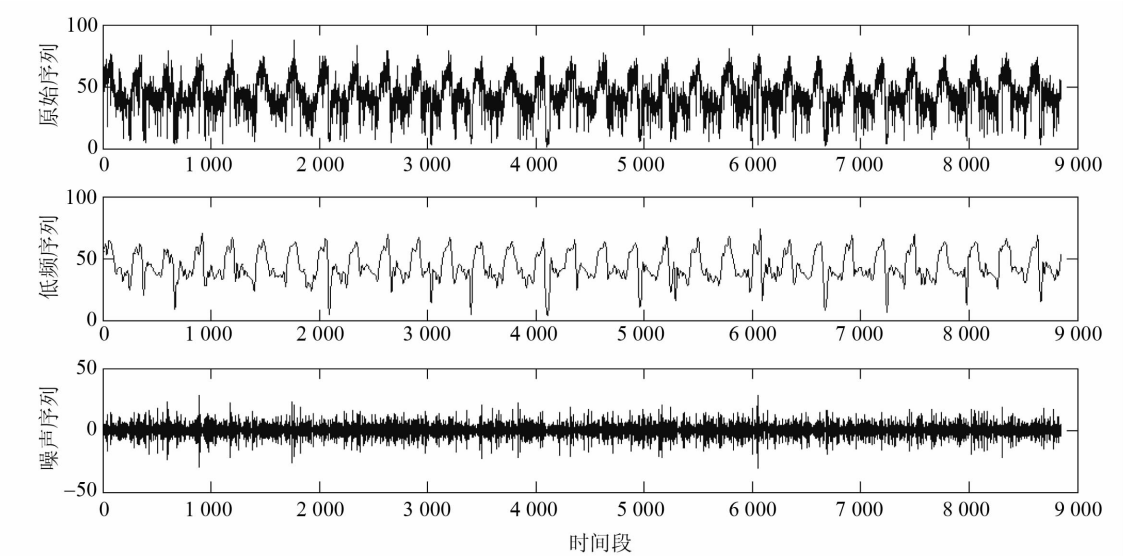

图 4 基于 Db4 小波函数的信号分解结果

Fig. 4 Signal decomposition result based on Db4 wavelet function

图 4 中，原始序列是低频序列和高频噪声序列叠加产生的信号图形。小波分解可看作一个信号滤波的去噪过程[17]，图中的高频信号在应用中可视为主要噪声源，其噪声序列由不相关的随机变量构成，具有很强的随机性。在去噪后，低频序列仍保留了原始序列的周期、振幅、变化趋势等特征，因此使用低频序列进行 DTW 的计算可以有效降低噪声干扰，得到更加平稳且符合实际规律的样本数据[18]。利用低频基波进行相似路段的比对，最终查找到与源路段相似的路段，取源路段和相似路段一周内速度变化进行对比分析，结果如图 5 所示。

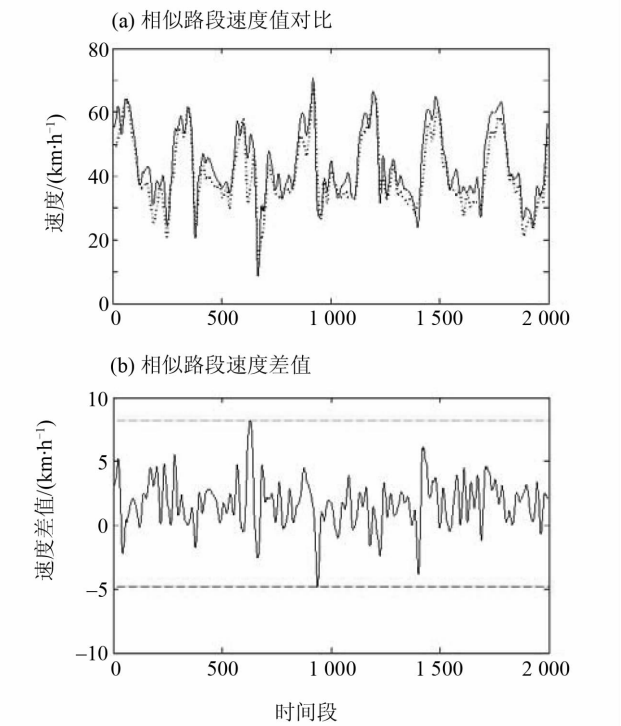

图 5　源路段与相似路段速度变化对比

Fig. 5　Comparison of the speed of the original and similar sections

图 5（a）中，虚线表示源路段，实线表示源路段的相似路段。截取了历史数据中一周的数据进行对比分析，可以看出相似路段的变化趋势、振幅、周期等特性与源路段契合，但相似路段的可靠程度需将两路段的实时偏差、偏差的峰值、波动范围等特性进行进一步比对才能得到验证。图 5（b）中以两路段对应样本的差值组成序列，从这个序列中可以看出两路段的相似程度较高，对应点位的偏差峰值不超过 +8 和 −5，其他差值多在更小的固定范围内波动，得到稳定小范围差值的原因是在执行 DTW 算法时，超出指定范围的样本已在路径约束中作为离群点被剔除，因此以此方法筛选出的相似路段具有一定的可靠性。将研究范围全部源路段与相似路段进行比对，对应点位的偏差峰值不超过 +10 和 −8，偏差峰值在 +5 和 −5 之间的点位占比 83.67%。

2.2　模型评价

2.2.1　数据缺失程度划分　为量化偶然性缺失和多发性缺失的定义对模型的准确性及数据填补率的影响，本文将偶然性缺失分别定义为数据缺失程度小于 10%、20%、30%、40%、50%，不满足偶然性缺失条件的则为多发性缺失。对定义后的缺失进行分别处理，用 67.54% 的低完整率源数据进行验证，考量模型使用后估计结果的效果以确定数据缺失性质的量化区分，具体对比如下图 6 所示。

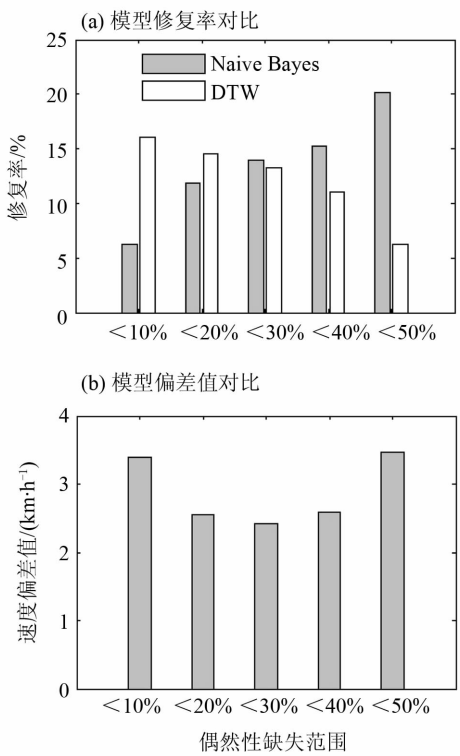

图 6　数据缺失定义对模型影响

Fig. 6　The impact of data loss definition on the model

图 6（a）中，经过二次填充后路网交通流数据的完整程度均能达到 90% 以上，但第一次填充率的增大将会抑制第二次填充的效果。因为第一次填充越多，第二次填充的样本容量将会减小，找到相似路段的几率也会随之减小。由图 6（b）可知，将过多数据定义为偶然缺失，使用朴素贝叶斯法或过度使用 DTW 法都会使模型的偏差值增大，即准确度降低。这是因为：朴素贝叶斯法采用过度缺失

的历史数据作为条件参数，将会影响估计精度；而，DTW 算法查找相似路段的方法采用欧氏距离，对噪声较为敏感，若对计算过程产生过度影响，可能会造成噪声干扰路段与非相似路段的混淆，导致计算结果产生偏差。以同样的方式对 76.33% 和 85.10% 完整率的源数据进行验证，结果表明：在数据修复的完整程度上各分组均表现良好；但，阈值定义为 30% 时，偏差值可维持在 2.2 ～ 2.5 km/h。因此，本文选择将数据缺失程度小于 30% 作为偶然性缺失的定义，在保证模型填充效率的同时提升模型估计结果的准确性。

2.2.2 模型对比 对三种模型进行对比分析。模型一采用文献［10］中的方法，仅使用朴素贝叶斯分类，以同路段的历史交通流速度和前一时刻交通流速度作为输入参数对缺失数据进行估计填充；模型二中路网数据缺失情况不做区分，统一采用 DTW 方法对偶然性缺失和多发性缺失同时进行填充；模型三为本文所采用的方法，先将数据缺失情况进行划分，再配合使用朴素贝叶斯方法和 DTW 方法分别对偶然性缺失和多发性缺失分别进行估计填充。

以路网 30 天内的前 20 天数据作为训练数据，后 10 天数据作为测试数据；测试数据将分别被视作缺失样本进行填充，并将真实数据与填充数据差值的绝对值作为本次填充的估计误差；每个测试数据的填充及误差计算过程相互独立互不影响，最终所有测试样本的平均绝对误差（MAD）和均方误差（MSE）的计算如式(13) ～(14)所示：

$$\mathrm{MAD} = \sum_{i=1}^{n} \frac{|\mathrm{obs}_i - \mathrm{pre}_i|}{n} \tag{13}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{obs}_i - \mathrm{pre}_i)^2 \tag{14}$$

式中，n 为测试数据的样本总数；obs 为真实数据的观测样本；pre 为模型填充的估计样本。最终对模型的填充效果和估计准确度进行比较，结果如表 2 所示。

表 2 三种模型效果比较

Table 2 Comparison of three models

模型类型	NB	DTW	NB-DTW
源数据完整性/%	67.54	67.54	67.54
填充后完整性/%	81.29	90.33	94.37
平均绝对误差 MAD	4.34	5.37	4.83
均方误差 MSE	19.14	29.09	23.63

由上表可知，单独使用 DTW 算法的模型二的平均绝对误差和均方误差均较大，平均绝对误差超出了速度的一个组值范围；而，模型一和模型三的平均绝对误差均在一个组值的范围内。当误差在可接受范围内时，模型三的填充率比模型一的高出 13.08%，填充效果有较显著的提升。因此，本文所提方法在保证估计结果准确性的同时大幅提升了缺失数据的填充率，相比于单一使用朴素贝叶斯方法或 DTW 方法的模型更为优化。在运行时间方面，DTW 方法寻找相似路段的过程耗时较长，但此步骤可提前离线完成，组合模型实时计算的平均时长为 15.6 s。

2.3 试验结果分析

以 2017 年 5 月 31 日第 96 时段（7：55 到 8：00）的数据进行模型的试验。数据填充前，路网交通流速度的原始分布如图 7 所示，虚线表示该时段路网处在数据缺失的状态，颜色由浅至深表示速度由低到高。对路段 ID 矩阵 $\boldsymbol{M}$ 进行统计分析，得到该时段路网的数据覆盖率为 62.68%，较高的数据缺失率和数据密度的分布不均给路网的交通流速度分布研究、交通流速度的预测等带来了较大的困难。

使用朴素贝叶斯分类方法对路网缺失数据进行估计，结果如图 8（a）所示。处理后，路网的数据覆盖率达到了 73.76%，比原始数据提升了 17.68%，但路网中仍有部分路段存在数据缺失的现象。这一现象多存在于低等级道路上，说明低等级道路上历史数据缺失大于 30% 的现象严重，导致朴素贝叶斯方法抛弃对这些路段的估计，进而采用 DTW 算法查找相似路段对估计结果进行补充，结果如图 8（b）所示。经过两个步骤的处理之后，路网的数据覆盖率达到了 92.33%，比原始数据提升了 47.30%。因此，将两个方法搭配使用的填充效果显著，很好地弥补了朴素贝叶斯方法对于多发性数据缺失估计的不足，也在一定程度上降低了使用 DTW 方法进行预测带来的误差。

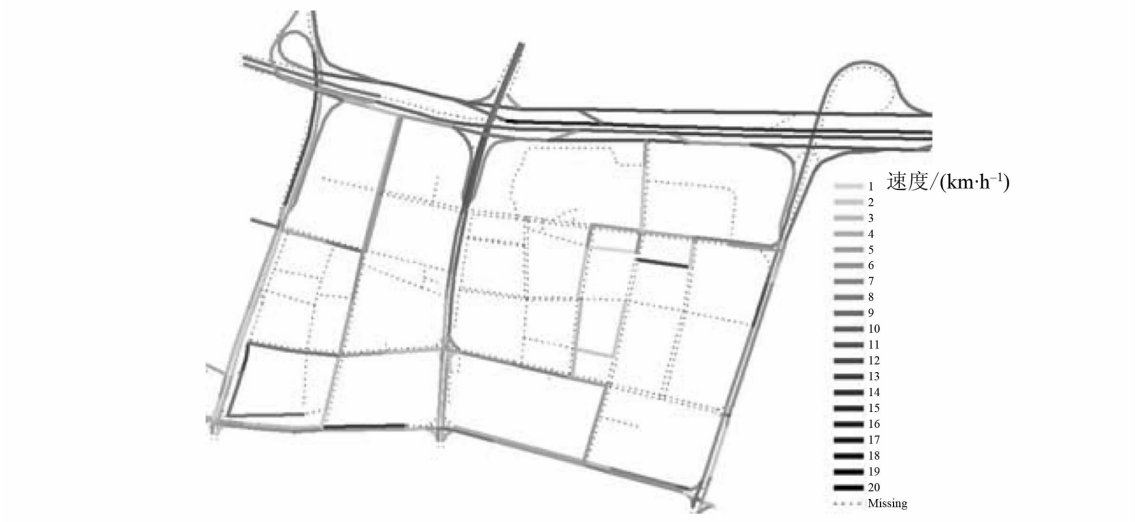

图 7　路网交通流数据缺失现状

Fig. 7　The current situation of data loss of traffic flow

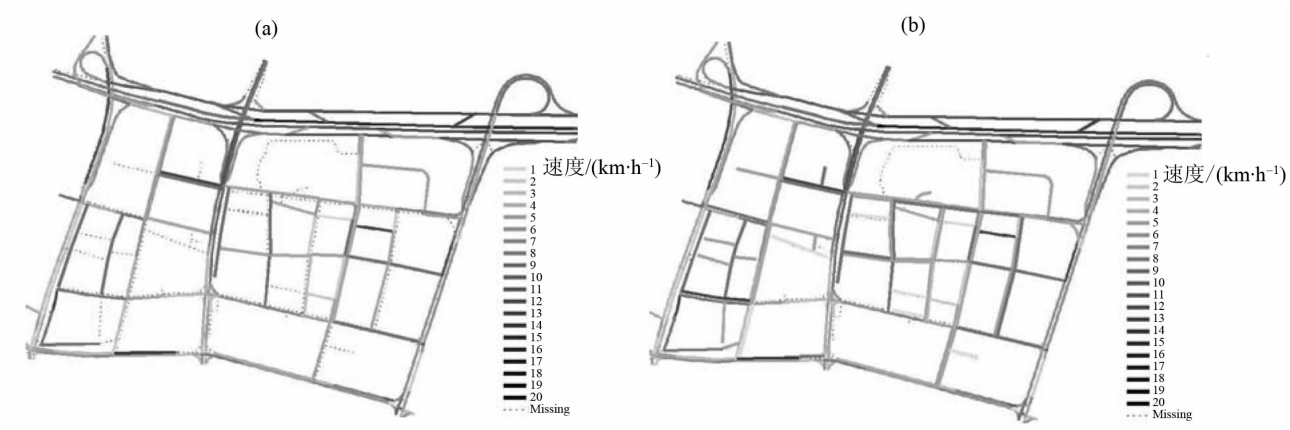

图 8　路网交通流速度数据修复对比

Fig. 8　Comparison of repair of traffic flow velocity data

3　结　论

本文构建了一个稀疏数据的填充模型，将数据缺失的类型分为偶然性和多发性缺失，并配合使用朴素贝叶斯分类方法和 DTW 算法对缺失数据进行填充。试验结果表明：该模型可以使路网的数据覆盖率得到显著的提升，并且混合模型在填充效果上明显优于仅使用朴素贝叶斯方法进行数据填充；填充数据的准确率相较于仅使用 DTW 算法也得到了一定程度的提升。填充后的浮动车数据具有连续性强、准确度高和数据密度分布均匀的特点。后续的工作仍需考虑以下情况以提升模型效率：

（1）为减少模型的偶然性偏差，保证填充结果的准确性和稳定性，应尽可能增加历史数据样本量，提升以历史数据作为输入参数之一的模型的估计准确性；

（2）不同路网的数据缺失情况不同，应在模型使用前重新对偶然性缺失和多发性缺失的不同划分情况所造成的影响进行评估，确定划分范围，以确保模型填充效果得到保障；

（3）应进一步考虑路网复杂性对模型效率和准确性带来的影响，寻找样本量、准确性和计算效率之间的平衡点。

参考文献：

[1] REMPE F, FRANECK P, FASTENRATH U, et al. Online freeway traffic estimation with real floating car data [C]//IEEE International Conference on Intelligent Transportation Systems. 2016.

[2] 何兆成, 庄立坚, 杨文臣,等. 基于大规模浮动车数据的城市道路网复杂度分析[J]. 公路交通科技, 2013, 30(6):120 – 126.
HE Zhaocheng, ZHUANG Lijian, YANG Wenchen, et al. Analysis of urban road network complexity based on large-scale floating car data[J]. Journal of Highway and Transportation Research and Development, 2013, 30(6): 120 – 126.

[3] 张希瑞,方志祥,李清泉,等. 基于浮动车数据的城市道路通行能力时空特征分析[J]. 地球信息科学学报, 2015,17(3):336 – 343.
ZHANG Xirui, FANG Zhixiang, LI Qingquan, et al. Spatial and temporal analysis of urban road capacity based on floating vehicle data[J]. Journal of Geo-information Science,2015,17(3):336 – 343.

[4] 郝艳召,宋国华,邱兆文,等. 基于浮动车数据的机动车排放实时测算模型[J]. 中国环境科学,2015,35(2): 396 – 402.
HAO Yanzhao, SONG Guohua, QIU Zhaowen, et al. Real-time calculation model for vehicle emissions based on floating car data[J]. China Environmental Science,2015, 35(2):396 – 402.

[5] 王晓蒙,池天河,林晖,等. 一种面向海量浮动车数据的地图匹配方法[J]. 地球信息科学学报,2015,17(10): 1143 – 1150.
WANG Xiaomeng, CHI Tianhe, LIN Hui, et al. A map matching method for mass floating vehicle data[J]. Journal of Geo-information Science, 2015,17(10):1143 – 1150.

[6] LI L, LI Y, LI Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence [J]. Tramsportation Research Part C:Emerging Technologies, 2013, 34(9):108 – 120.

[7] 祁伟, 李晔, 汪作新. 季节性 ARiMA 模型在稀疏交通流下的预测方法[J]. 公路交通科技, 2014, 31(4): 130 – 135.
QI Wei, LI Ye, WANG Zuoxin. Utilizing seasonal ARiMA model for traffic estimation in sparseness[J]. Journal of Highway and Transportation Research and Development, 2014, 31(4):130 – 135.

[8] QI H, LIU M, WANG D, et al. Spatial-Temporal congestion identification based on time series similarity considering missing data [J]. Plos One, 2016, 11 (9): e0162043.

[9] 曲鑫,林赐云,杨兆升,等. 采用低频浮动车数据的行程时间估计[J]. 哈尔滨工业大学学报,2016,48(9):30 – 34.
QU Xin, LIN Ciyun, YANG Zhaosheng, et al. Travel time estimation using low-frequency floating car data[J]. Journal of Harbin Institute of Technology, 2016,48(9): 30 – 34.

[10] AXER S , FRIEDRICH B. Estimating signal phase and timing for traffic actuated intersections based on low frequency floating car data [C]//2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE,2016:2059 – 2064.

[11] 张发明,朱欣焰,闪维,等. 利用浮动车大数据进行稀疏路段行程时间推断[J]. 武汉大学学报(信息科学版) 2017,42(1):56 – 62.
ZHANG Faming, ZHU Xinyan, GUO Wei, et al. Infer the travel time of sparse sections by floating car data[J]. Geomatics and Information Science of Wuhan Unicersity, 2017,42(1):56 – 62.

[12] 王晓蒙,彭玲,池天河. 基于稀疏浮动车数据的城市路网交通流速度估计[J]. 测绘学报,2016,45(7):866 – 873.
WANG Xiaomeng, PENG Ling, CHI Tianhe. A method of urban traffic flow speed estimation using sparse floating car data[J]. Acta Geodaetica et Cartographica Sinica, 2016,45(7):866 – 873.

[13] FOWE A J, CHAN Yupo. A microstate spatial-inference model for network-traffic estimation[J]. Transportation Research Part C: Emerging Technologies, 2013, 36:245 – 260.

[14] WANG Jiawei, WANG Yinsong, YUN Meiping, et al. Development of urban road network traffic state dynamic estimation method[J]. Mathematical Problems in Engineering, 2015(5):1 – 10.

[15] YANG Zhaosheng, BING Qichun, LIN Ciyun, et al. Research on Short-Term traffic flow prediction method based

on similarity search of time series[J]. Mathematical Problems in Engineering,2014(7):1-8.

[16] HANG C, ZHANG R, CHEN Z, et al. Dynamic gesture recognition method based on improved DTW algorithm[C]//International Conference on Industrial Informatics-computing Technology. IEEE Computer Society, 2017.

[17] SHEN J, ZHU D, HUANG W, et al. A novel similarity measure approach for time series based on PLA and DTW[C]//Control Conference. IEEE, 2016.

[18] 杨正理，黄其新. 小波变换在行驶车辆检测器中的应用[J]. 公路交通科技，2013，30(2):104-108.
YANG Zhengli, HUANG Qixin. Application of wavelet transform in running-vehicle detectors[J]. Journal of Highway and Transportation Research and Development, 2013, 30(2):104-108.

[19] YUAN G S, GENG L Q. ECG Signal denoising algorithm based on optimization wavelet transform[J]. Advanced Materials Research, 2013(706/708):4.

[20] 胡婕,杨马英. 基于小波变换和动态时间规整算法相结合的故障识别算法[J]. 软件导刊，2007(17):140-142.
HU Jie, YANG Maying. Fault recognition algorithm based on wavelet transform and dynamic time regularization algorithm[J]. Software Guide, 2007(17):140-142.

（责任编辑　王海蓉）