文章编号：0253-374X(2019)10-1477-08 DOI:10.11908/j.issn.0253-374x.2019.10.013

改进的固定交通检测器缺失数据综合修复方法

苗 旭[1]，王忠宇[2]，邹亚杰[1]，吴 兵[1]
(1.同济大学 道路与交通工程教育部重点实验室，上海 201804；2.上海海事大学 交通运输学院，上海 201306)

摘要：基于检测器数据的时空相关性，为缺失数据修复模型动态地选择解释变量，在综合考虑检测器数据的周期性趋势和实时变化特性的基础上，提出了一种改进的缺失数据修复方法.对上海市南北高架的线圈流量数据进行数据修复精度测试.结果表明，相较于传统的支持向量回归(SVR)模型，该方法在3个测试检测器上的数据修复平均绝对误差分别减小了3.80%、3.40%、25.23%，并且在数据连续缺失1～10个时平均绝对百分比误差均低于6%.

关键词：交通运输系统工程；缺失数据修复；周期性；支持向量回归(SVR)
中图分类号：U491 **文献标志码**：A

Improved Modification Method of Missing Data for Location-specific Detector

MIAO Xu[1], *WANG Zhongyu*[2], *ZOU Yajie*[1], *WU Bing*[1]
(1. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China; 2. College of Transport and Communications, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Based on the temporal and spatial correlation of detector data, the explanatory variables were dynamically selected for data repair model, and an improved modification method of missing data was proposed considering periodic trend and real-time variability comprehensively. The proposed method was assessed with the data of location-specific detectors in Shanghai, China. Compared with support vector regression (SVR) model, the mean absolute error of three detectors are reduced by 3.80%, 3.40%, 25.23%, and the mean absolute percentage error is less than 6% under different data missing conditions.

Key words: engineering of communications and transportation system; missing data modification; periodic pattern; support vector regression(SVR)

固定交通检测器的数据采集缺失现象对交通数据分析和挖掘等均带来不利的影响，因此有必要进行缺失数据修复.常见的数据修复方法有历史均值法[1-3]、插值法[4-5]、主成分分析法[6-8]、时间序列法[9]及机器学习算法[10-11].历史均值法是最早发展起来的数据修复方法.陆化普等[1]提出了基于历史数据和当前数据加权平均的数据修复方法.姜桂艳等[2]利用相邻时段及路段数据对故障数据进行修复.孙玲等[3]基于缺失数据的时空相关性将相关数据加权重构作为缺失数据的修复值.插值法主要分为指数平滑法、样条插值法及回归方法.Smith等[4]基于相邻时段数据的指数平滑值进行故障数据修复.Boyles[5]比较了简单线性回归模型、多元线性回归模型、局部和全局回归模型、非正态贝叶斯线性回归模型等方法后指出，虽然回归算法简单且容易构建，但是数据修复结果在不同交通状态下不可靠.Qu等[6-7]和Li等[8]提出了概率主成分分析法、贝叶斯主成分分析法及核概率主成分分析法，指出该类方法数据修复精度优于历史均值法及样条插值法.ARIMA(autoregressive integrated moving average model)是常用的时间序列数据修复方法.Ghosh等[9]比较了ARIMA与Holt-Winters指数平滑数据修复方法及随机游走算法，指出ARIMA是一种有效的数据修复方法.近几年，机器学习模型也逐渐应用于缺失数据修复.Tang等[10]提出基于模糊C均值与遗传算法相结合的数据修复方法.Zhang等[11]衡量同一时刻不同地点交通参数的相关性，并提出基于最小二乘支持向量回归的缺失数据修复方法.

对于上述数据修复模型，选择解释变量时的主要依据为交通流数据的时空相关性，所有检测器均采用固定的解释变量，但是不同检测器数据与同一相关序列的相关性存在较大差异，解释变量固定势

收稿日期：2018-12-10
基金项目：国家自然科学基金(51608386)
第一作者：苗 旭(1988—)，女，博士生，主要研究方向为交通数据挖掘、交通拥挤管理. E-mail:miaoxu@tongji.edu.cn
通信作者：吴 兵(1960—)，男，教授，博士生导师，工学博士，主要研究方向为交通控制、交通拥挤管理.
E-mail:wubing@tongji.edu.cn

必影响部分检测器缺失数据的修复精度,而且数据的连续缺失容易导致修复误差的逐步传递和累积.另外,一个有效的数据修复方法既要考虑交通流数据的周期变化特性,又要捕捉复杂交通环境引起的交通流数据的实时变化,这对目前的研究仍具有较大的挑战.为避免连续数据缺失造成的误差累积,基于数据的相关性及连续缺失情况为修复方法动态地选择解释变量,并综合考虑交通流数据的周期性变化趋势和实时变化特性,提出一种改进的数据修复方法.

1 数据来源

本研究选取的数据为2017年3月6日—31日上海市南北高架东侧徐家汇路至大沽路路段20个工作日内固定检测器采集的流量数据.该段快速路长度约为3 km,单向四车道,设计车速为80 km·h^{-1}.主线共布设了7组完好的固定检测器,采集字段为检测器编号、采集时间、流量、平均速度、平均时间占有率等.其中,流量为5 min内经过检测器所处断面的车流量总数.为满足交通管理实时控制的需求,对修复时段t的缺失数据,仅采用历史时段($t-h$)($h\geqslant 1$)的数据进行修复.为方便说明,将分析范围内的检测器重新编号,从南向北方向行驶的车辆依次经过的检测器为1号至7号.检测器空间位置分布如图1所示.

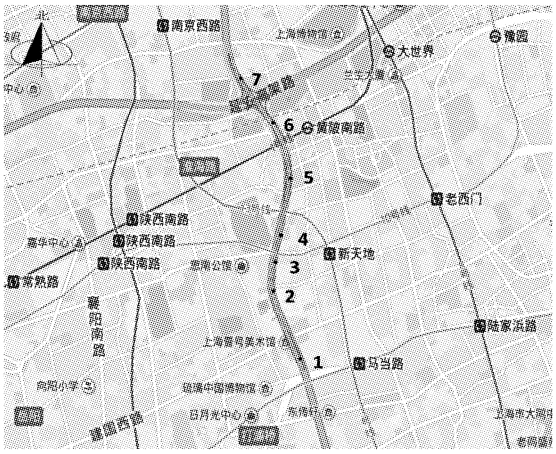


图1 上海市南北高架检测器分布

Fig.1 Detectors on the north-south viaduct in Shanghai

2 综合数据修复方法

所提出的综合数据修复方法将检测器采集的流量数据分成两部分,即周期性变化趋势与实时变化残差值.描述周期性特征的函数主要有三角级数法[12]、简单平均值法(SAM)[13]及双指数平滑法[14].选择简单且常用的简单平均值法进行周期性变化趋势描述,采用动态选择解释变量的支持向量回归(DV-SVR)算法进行实时变化残差值的预测.下文称所提出的综合数据修复方法为SAM-DV-SVR,计算式如下所示:

$$Y(t)=D(t)+R(t) \tag{1}$$

式中:$Y(t)$为t时段检测器采集的流量实际值;$D(t)$为流量数据的周期性部分;$R(t)$为残差值.

2.1 简单平均值法周期分析

图2为3号检测器2017年3月份一周工作日的流量数据分布.可以非常明显地看出,流量数据呈现出以24 h为一个周期的反复特性.计算每个检测器3月份任意2个工作日的数据相关系数,并进一步得到相关系数均值,该均值可以反映检测器的日变化趋势的一致性.计算得出3号、4号、5号检测器的流量数据相关系数均值分别为0.978、0.927、0.944,可以看出3号检测器流量数据的日变化趋势更为相似.假设连续采集N天的工作日数据,每天采集样本数为n,每天采集的流量数据可记为

$$\begin{cases}\boldsymbol{Y}_1=(Y_1(1),Y_1(2),Y_1(3),\cdots,Y_1(n))\\ \vdots\\ \boldsymbol{Y}_N=(Y_N(1),Y_N(2),Y_N(3),\cdots,Y_N(n))\end{cases} \tag{2}$$

简单平均值法计算式为

$$D(t)=\frac{1}{N}\sum_{r=1}^{N}Y_r(t) \tag{3}$$

本研究选取3月6日至3月22日的13个工作日的数据计算周期趋势,因此$N=13$,$n=288$.

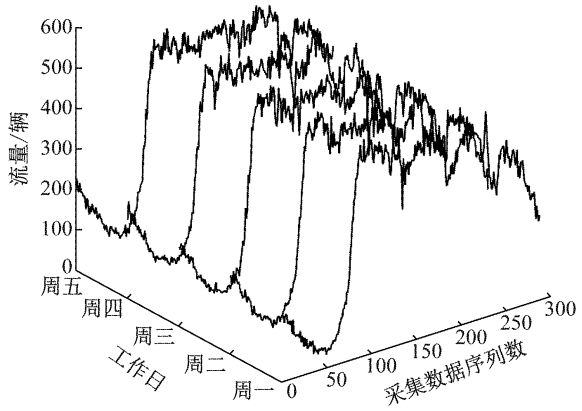


图2 工作日流量的周期性分析

Fig.2 Periodic analysis of flow on weekdays

2.2 动态选择解释变量的支持向量回归模型

2.2.1 备选相关序列构建

每个检测器每天采集的流量数据可组成288维

的向量，将缺失数据所在的向量称为目标向量 $\boldsymbol{S}$，而由相关数据组成的向量称为相关序列. 根据以往研究结论[15]，共选择了 8 个备选相关序列，如表 1 所示. 将目标向量 $\boldsymbol{S}$ 分别与相关序列 $\boldsymbol{S}_1$ 至 $\boldsymbol{S}_8$ 进行相关系数计算，可分别得到目标向量与各相关序列的相关系数，将相关系数的大小作为缺失数据修复模型解释变量的重要选择依据. 相关系数计算式为

$$R_a=\frac{\sum_{j=1}^{n}(S(j)-\overline{S})(S_a(j)-\overline{S_a})}{\sqrt{\sum_{j=1}^{n}(S(j)-\overline{S})^2\sum_{j=1}^{n}(S_a(j)-\overline{S_a})^2}} \quad (4)$$

式中：R_a 为目标向量 $\boldsymbol{S}$ 与相关序列 $\boldsymbol{S}_a$ 的相关系数；$a=1,2,\cdots,8$；$S(j)$ 为目标向量 $\boldsymbol{S}$ 的第 j 个采样数据，$S_a(j)$ 为相关序列 $\boldsymbol{S}_a$ 的第 j 个采样数据；$\overline{S}$ 与 $\overline{S_a}$ 分别为向量 $\boldsymbol{S}$ 与 $\boldsymbol{S}_a$ 的样本数据均值.

2.2.2　解释变量动态选择

为充分考虑流量数据的时空相关性，解释变量的选择至少包括一个时间相关序列向量及一个空间相关序列向量. 解释变量动态选择的依据一是目标向量与相关序列向量相关系数的大小，二是连续缺失数据的数量. 首先，构建相关序列 $\boldsymbol{S}_1$ 至 $\boldsymbol{S}_8$，若数据存在连续缺失现象，如检测器 $(t-1)$ 时段及 t 时段数据均缺失，则由 $(t-2)$ 时段数据作为相关序列 $\boldsymbol{S}_1$，记为 $\boldsymbol{S}_{1,2}$，$(t-3)$ 时段数据作为相关序列 $\boldsymbol{S}_2$，记为 $\boldsymbol{S}_{2,3}$，依次类推；然后，计算相关系数 R_1 至 R_8，根据相关系数大小选择解释变量来进行缺失数据修复. 解释变量选择流程如图 3 所示. 图 3 中，m 为解释变量的数量.

表 1　相关序列描述

Tab.1　Description for correlation sequences

序号	分类	编号	相关序列	序列描述
1	时间相关性	$\boldsymbol{S}_1$	相同检测器 $(t-1)$ 时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_1$ 为检测器 k 在日期 d 的 13∶15 时段的数据
2		$\boldsymbol{S}_2$	相同检测器 $(t-2)$ 时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_2$ 为检测器 k 在日期 d 的 13∶10 时段的数据
3		$\boldsymbol{S}_3$	相同检测器相邻 2 天同时段数据均值组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_3$ 为检测器 k 在日期 $(d-1)$ 与 $(d-2)$ 的 13∶20 时段的数据的平均值
4		$\boldsymbol{S}_4$	相同检测器前 1 周相同时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_4$ 为检测器 k 在日期 $(d-7)$ 的 13∶20 时段的数据
5	空间相关性	$\boldsymbol{S}_5$	相邻前 1 个编号检测器同日期同时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_5$ 为检测器 $(k-1)$ 在日期 d 的 13∶20 时段的数据
6		$\boldsymbol{S}_6$	相邻后 1 个编号检测器同日期同时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_6$ 为检测器 $(k+1)$ 在日期 d 的 13∶20 时段的数据
7		$\boldsymbol{S}_7$	相邻前 2 个编号检测器同日期同时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_7$ 为检测器 $(k-2)$ 在日期 d 的 13∶20 时段的数据
8		$\boldsymbol{S}_8$	相邻后 2 个编号检测器同日期同时段数据组成	假设检测器 k 在日期 d 的 13∶20 时段数据缺失，则 $\boldsymbol{S}_8$ 为检测器 $(k+2)$ 在日期 d 的 13∶20 时段的数据

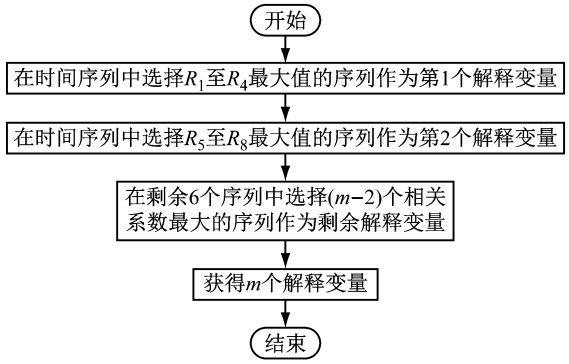


图 3　解释变量选择流程

Fig.3　Flow chart of explanatory variable selection

2.2.3　支持向量回归模型

设训练样本集 $M=\{(y_i,V_{a,i},V_{b,i},V_{c,i},V_{d,i}),i=1,\cdots,l\}$，其中 $V_{a,i}$、$V_{b,i}$、$V_{c,i}$、$V_{d,i}$ 为动态选取的输入变量，y_i 为相应的输出值，本研究中 y_i 为目标检测器的缺失数据，l 为训练样本个数. 支持向量回归模型的基本思想是寻找一个从输入空间到输出空间的非线性映射函数 $\boldsymbol{\varphi}(x)$，通过该函数将训练样本集映射到高维特征空间 P，因此可在空间 P 中对原始问题进行线性回归[16]. 映射关系如下所示：

$$f(x)=(\boldsymbol{w}\cdot\boldsymbol{\varphi}(x))+b,\quad \boldsymbol{w}\in P \quad (5)$$

式中：$\boldsymbol{w}$ 为权重值；$(\cdot)$ 为内积运算；b 为偏置项. $\boldsymbol{w}$ 和 b 通过最小化下列函数进行估计：

$$R(\boldsymbol{w})=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}H_\varepsilon(Y(t),\hat{Y}(t)) \quad (6)$$

式中：$Y(t)$ 为真实值；$\hat{Y}(t)$ 为输出值；ε 为不敏感代价函数定义的误差；$H_\varepsilon(Y(t),\hat{Y}(t))$ 为 ε 不敏感损失函数；C 为惩罚系数. $H_\varepsilon(Y(t),\hat{Y}(t))$ 的定义如下所示：

$$H_\varepsilon(Y(t),\hat{Y}(t))=$$

$$\begin{cases} |Y(t)-\hat{Y}(t)|-\varepsilon, & |Y(t)-\hat{Y}(t)|>\varepsilon \\ 0, & \text{其他} \end{cases} \tag{7}$$

式(7)表明,若SVR输出值$\hat{Y}(t)$与真实值$Y(t)$差的绝对值小于设定的ε时,损失函数值忽略不计;否则,损失函数值大小为$|Y(t)-\hat{Y}(t)|$超出ε的部分.引入核函数,将式(6)与(7)转换为求解下式:

$$\begin{aligned} &\min \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{l}\xi_i+C\sum_{i=1}^{l}\xi_i^* \\ &\text{s. t. } \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(x_i)+b-Y(t)\leqslant\varepsilon+\xi_i \\ &\qquad Y(t)-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(x_i)-b\leqslant\varepsilon+\xi_i^* \\ &\qquad \xi_i,\xi_i^*\geqslant 0,\quad i=1,\cdots,l \end{aligned} \tag{8}$$

式中:ξ_i、ξ_i^*为松弛变量.引入拉格朗日乘子对最优化问题进行求解,可得到权值计算式,如下所示:

$$\boldsymbol{w}=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\boldsymbol{\varphi}(x_i) \tag{9}$$

式中:α_i、α_i^*为拉格朗日乘子,即最小化$R(\boldsymbol{w})$的解.α_i、α_i^*必须满足

$$\sum_{i=1}^{l}\alpha_i=\sum_{i=1}^{l}\alpha_i^*,\quad 0\leqslant\alpha_i,\alpha_i^*\leqslant C,i=1,2,\cdots,l \tag{10}$$

通过式(5)和式(9),可以将$f(x)$表示为

$$\begin{aligned} f(x)&=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,x)+b \\ K(x_i,x)&=\boldsymbol{\varphi}(x_i)\cdot\boldsymbol{\varphi}(x) \end{aligned} \tag{11}$$

式中:$K(x_i,x)$为核函数.核函数不同,模型决策函数的最终形式也不相同.SVR模型支持常见的线性、多项式、径向基(RBF)、Sigmoid等4种核函数,本研究选取最常用的RBF核函数.

在ε-SVR的构建时,常数C作为惩罚系数控制损失的大小,模型求解中C可作为调节参数,影响训练模型的分类性能.此外,RBF核函数中参数g的数值也会明显影响模型的预测性能.在参数设置过程中,采用网格分析法及交叉验证法对支持向量回归中的常数C及RBF核函数参数g进行参数寻优.交叉验证法为:将原始数据均分成3组,对每组子集数据做1次验证集,其中2组子集数据作为训练集,最后得到3个模型,用这3个模型最终验证集的分类准确率平均值作为性能评价指标.网格分析法是通过编程枚举的方式对不同参数下的模型预测效果进行对比.此处以数据缺失一个的情况为例介绍惩罚系数C及核函数参数g的选择对SVR模型的影响.该实验采用均方误差(α_{MSE})作为评价指标,计算公式为

$$\alpha_{\mathrm{MSE}}=\frac{1}{n_1}\sum_{t=1}^{n_1}(Y(t)-\hat{Y}(t))^2 \tag{12}$$

式中:n_1为修复数据个数.

图4为惩罚系数C及核函数参数g对SVR模型预测结果的影响.从图4可以看出,惩罚系数C较小时,SVR处于"欠学习"状态,预测误差并不是最小,随着C的增大,误差减小随后又逐渐增大,说明当C大于某一值后,SVR模型处于"过学习"状态.C在一定的区间内时,不同的取值得到的误差相差不大,说明对于固定的g,存在多个C可以使得SVR模型取得较好的预测能力.同样,随着g的增大,预测均方误差呈现先减小后增大的两边大中间小的趋势,说明当g增大到一定程度之后,SVR模型呈现"过学习"现象.可见,g的变化对模型的预测能力也有非常大的影响.通过网格学习方法,遍历$\log_2 C$及$\log_2 g$ 2个参数在-5到5之间的所有组合,选择最优的参数建立数据修复精度最高的回归模型.另外,针对不同的检测器选择及不同的解释变量输入,SVR模型依据网格分析法及交叉验证法对2个参数进行重新选择.

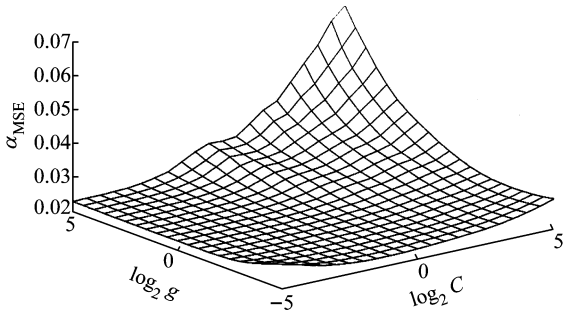


图4 C与g对SVR模型的影响

Fig.4 Influence of C and g on SVR model

3 实际案例及结果分析

选择编号为3号、4号、5号的检测器作为模型测试对象.将3月6日—10日(周一至周五)数据作为相关序列构建的基础数据,如3月13日缺失数据修复时的相关序列$\boldsymbol{S}_4$的构建需要使用3月6日的历史数据.3月13日—22日的8个工作日数据作为模型训练数据,用来进行模型参数的标定.3月23日—31日的7个工作日数据作为模型预测结果的测试数据,用来评价模型的泛化能力.如前所述,数据采集时不仅存在单个数据缺失现象,还存在连续数

据缺失现象. 选取的 3 月 6 日—31 日 3 个检测器数据均为 100%检测无缺失数据，将 3 月 23 日—31 日的 7 天数据随机剔除 10%的数据，分别构建连续缺失 1～10 个数据的场景进行数据修复，进而与采集的真实数据进行比较，从而验证模型的修复精度. 数据修复精度评价指标包括平均绝对误差(β_{MAE})、平均绝对百分比误差(γ_{MAPE})、均方根误差(δ_{RMSE}). 3 个指标的表达式如下所示：

$$\beta_{MAE}=\frac{1}{n_1}\sum_{t=1}^{n_1}|\hat{Y}(t)-Y(t)| \tag{13}$$

$$\gamma_{MAPE}=\frac{1}{n_1}\sum_{t=1}^{n_1}\left|\frac{\hat{Y}(t)-Y(t)}{Y(t)}\right|$$

$$\delta_{RMSE}=\sqrt{\frac{\sum_{t=1}^{n_1}(Y(t)-\hat{Y}(t))^2}{n_1}}$$

首先，基于第 2.1 节所述简单平均值法计算 3 个检测器的周期；其次，根据第 2.2 节所述方法构建 8 个相关序列来计算相关系数，并根据数据缺失情况及相关系数的大小动态选择解释变量；然后，基于支持向量回归模型预测缺失数据的残差值；最后，将预测的残差值与周期值相加组成缺失数据修复值.

(1) 解释变量动态选择

图 5 为 3 个目标检测器仅缺失一个数据且相邻检测器的相关数据完整时构建的 8 个相关序列. 可以看出，不同的检测器与同一个相关序列的相关系数差异较大. 3 号检测器与时间相关序列 $\boldsymbol{S}_1$ 至 $\boldsymbol{S}_4$ 的相关性明显高于空间相关序列 $\boldsymbol{S}_5$ 至 $\boldsymbol{S}_8$. 与 4 号和 5 号检测器相关性最强的序列均为空间相关序列，4 号检测器与 $\boldsymbol{S}_6$、$\boldsymbol{S}_7$ 相关序列的相关性最大，5 号检测器与 $\boldsymbol{S}_5$、$\boldsymbol{S}_8$ 相关序列的相关性最大. 可以看出，为所有的检测器动态选择不同的解释变量是非常有必要的.

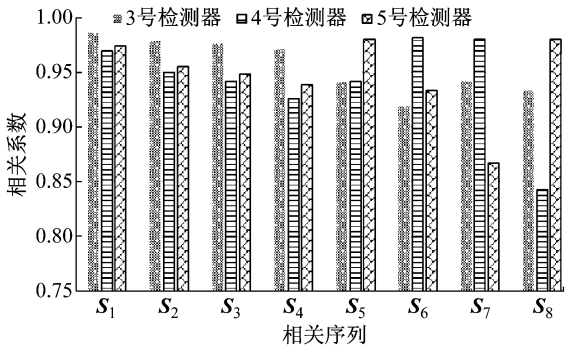


图 5　相关序列的相关系数

Fig.5　Correlation coefficients of correlation sequences

图 6 为 3 个检测器的自相关系数. 横坐标 1 至 9 代表的是($t-1$)至($t-9$)时段，纵坐标为 t 时段分别与($t-1$)至($t-9$)时段数据的相关系数. 可以看出，随着时间距离的增加自相关系数逐渐减小. 3 号检测器数据的自相关系数明显大于 4 号与 5 号检测器的自相关系数.

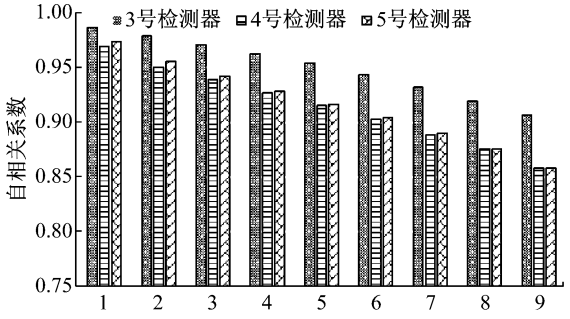


图 6　检测器数据的自相关系数

Fig.6　Autocorrelation coefficients of detector data

表 2 为目标检测器连续缺失 1～10 个数据且相邻检测器数据完整、历史日期数据完整时解释变量的选择方案. 因相邻检测器数据缺失或者历史日期数据缺失时解释变量的选择方案较多，故此处不予列出. 可以看出，对于不同的检测器，解释变量的选择存在较大差异. 其中，$\boldsymbol{S}_{i,k}$表示选取的($t-k$)时段数据作为相关序列 $\boldsymbol{S}_i$，$\boldsymbol{S}_1\boldsymbol{S}_2\boldsymbol{S}_3\boldsymbol{S}_7$ 表示选择 4 个相关序列作为解释变量，分别为相关序列 $\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{S}_3$、$\boldsymbol{S}_7$.

(2) 支持向量回归模型

根据表 2 中连续缺失 1～10 个数据的条件下解释变量的选择方案来动态选择模型的输入数据，如 3 号检测器某个需要修复的数据连续缺失数为 1 时，则选择 $\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{S}_3$、$\boldsymbol{S}_7$ 4 个相关序列的数据作为模型的输入数据，输出数据为缺失数据的残差值，再加上该时段对应的周期值得到缺失数据的修复值. 表 3 为 3 号检测器根据表 2 选择不同解释变量时模型的惩罚系数 C 及核函数参数 g 的选择方案以及残差预测结果的平均绝对误差. 可以看出，解释变量的动态选择，避免了预测误差随着连续缺失个数的增多而导致的误差累积现象.

(3) 数据修复结果

将以往研究中提出的数据修复方法与本研究提出的综合修复方法 SAM-DV-SVR 进行修复精度对比. 参与对比的修复方法包括双指数平滑(DES)方法、常规 SVR 模型、历史数据平均方法(HDAM)、多元线性回归(MLR)方法、反向传播神经网络(BPNN)模型、仅考虑周期趋势的 SVR(SAM-SVR)模型、仅考虑解释变量动态选择的SVR(DV-SVR)

表2 解释变量选择结果

Tab.2 Selection results of explanatory variable

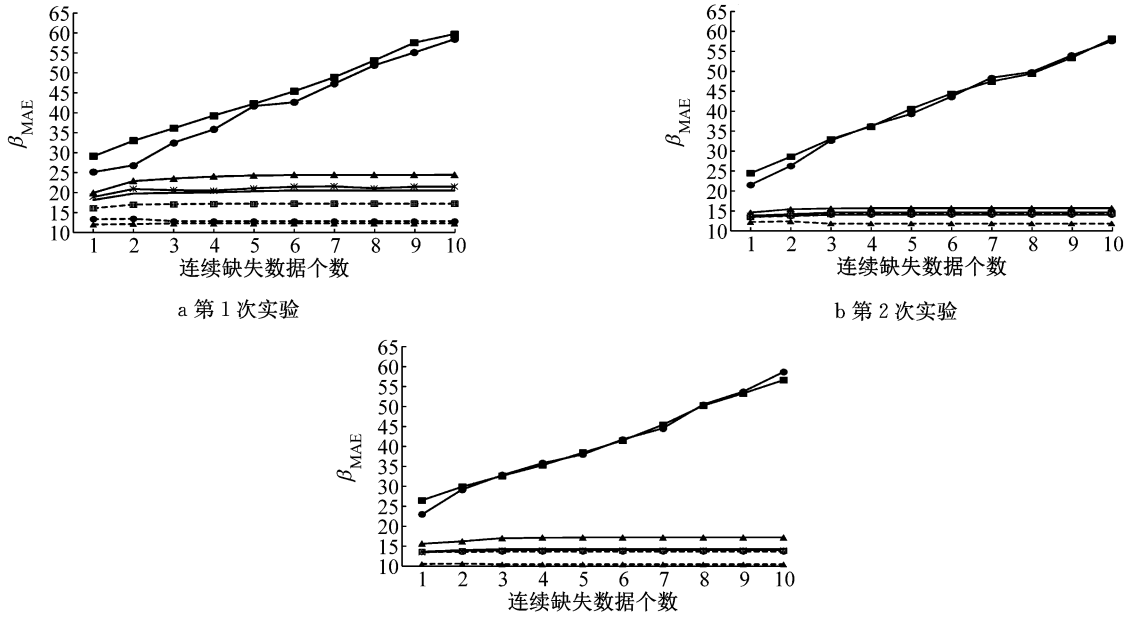
检测器编号	不同连续缺失数据个数下解释变量选择方案									
	1	2	3	4	5	6	7	8	9	10
3号	$S_1S_2S_3S_7$	$S_2S_3S_4S_7$	$S_3S_4S_7S_{1,3}$	$S_{1,4}S_3S_4S_7$	$S_{1,5}S_3S_4S_7$	$S_{1,6}S_3S_4S_7$	$S_3S_4S_5S_7$	$S_3S_4S_5S_7$	$S_3S_4S_5S_7$	$S_3S_4S_5S_7$
4号	$S_1S_2S_6S_7$	$S_2S_3S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$	$S_3S_5S_6S_7$
5号	$S_1S_2S_5S_8$	$S_2S_3S_5S_8$	$S_3S_5S_8S_{1,3}$	$S_3S_5S_8S_{1,3}$	$S_3S_5S_8S_{1,5}$	$S_3S_5S_8S_{1,5}$	$S_3S_5S_8S_{1,3}$	$S_3S_5S_8S_{1,3}$	$S_3S_5S_8S_{1,3}$	$S_3S_5S_8S_{1,3}$

表3 SVR模型参数选择结果及数据修复平均绝对误差

Tab.3 Selection results of parameters and β_{MAE} of SVR Model

连续缺失数据	3号检测器		4号检测器		5号检测器	
	$(\log_2 C,\log_2 g)$	β_{MAE}	$(\log_2 C,\log_2 g)$	β_{MAE}	$(\log_2 C,\log_2 g)$	β_{MAE}
1	(−2.5,−2.0)	14.92	(4.0,−4.5)	15.34	(4.0,0.5)	11.70
2	(−3.0,−0.5)	16.17	(2.0,−0.5)	16.32	(2.0,−1.5)	11.72
3	(−4.0,−1.5)	16.57	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
4	(−3.0,0)	16.93	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
5	(−1.0,0.5)	17.13	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
6	(−2.5,2.0)	17.53	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
7	(−1.0,−0.5)	17.16	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
8	(−1.0,−0.5)	17.16	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
9	(−1.0,−0.5)	17.16	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43
10	(−1.0,−0.5)	17.16	(0.5,−5.0)	16.83	(1.5,−0.5)	11.43

模型及本研究提出的综合数据修复模型SAM-DV-SVR.其中,历史数据平均法为同一检测器前4个时段值均值.常规SVR模型及MLR方法选取常用的4个解释变量作为预测模型输入,分别为目标检测器前2个时段数据(S_1,S_2)及前后断面同时刻数据(S_5,S_6).为保证模型的可对比性,本研究提出的综合模型同样选择4个解释变量.为排除模型预测结果的偶然性,随机剔除10%的数据并对结果验证过程进行了3次重复实验.图7为5号检测器3次重复实验的平均绝对误差.可以看出,3次数据修复平均绝对误差虽然数值大小有所差异,但各模型数据修复精度的排名基本保持一致.从图7还可以看出,HADM及DES方法因仅考虑了交通流数据的时间相关性,只采用本身检测器的历史数据作为解释变



a 第1次实验 b 第2次实验

c 第3次实验

BPNN模型 DES方法 DV-SVR模型 HDAM MLR方法 SAM-DV-SVR模型 SAM-SVR模型 SVR模型

图7 5号检测器3次重复实验平均绝对误差

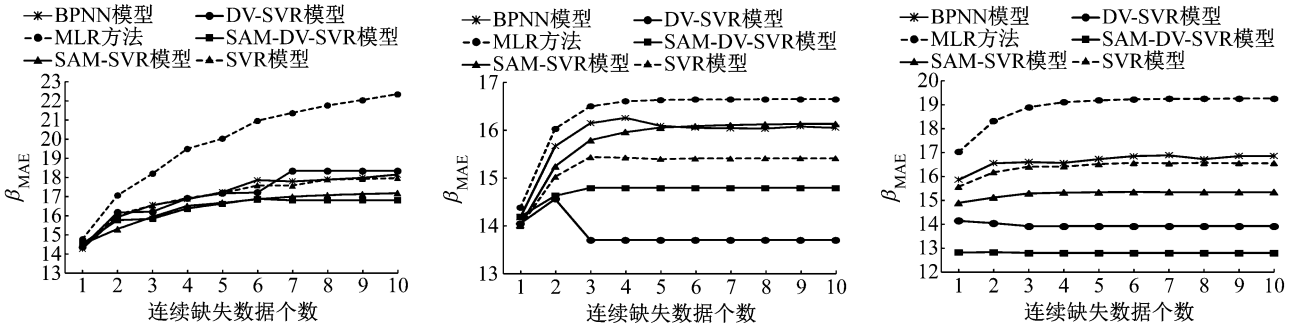
Fig.7 β_{MAE} of 3 repeated experiments on No.5 detector

量，数据修复精度明显低于其他几种模型，并且随着数据缺失个数的增加，修复误差均明显增加. 因此，在下面的讨论中，仅对其他 6 种模型的数据修复结果取平均值进行深入分析.

图 8～10 分别为 6 种模型的数据修复平均绝对误差、平均绝对百分比误差及均方根误差. 分析 3 个检测器的数据修复结果，可以看出：

(1) 相较于传统的 SVR 模型，SAM-DV-SVR 模型对缺失数据修复的精度显著提升.

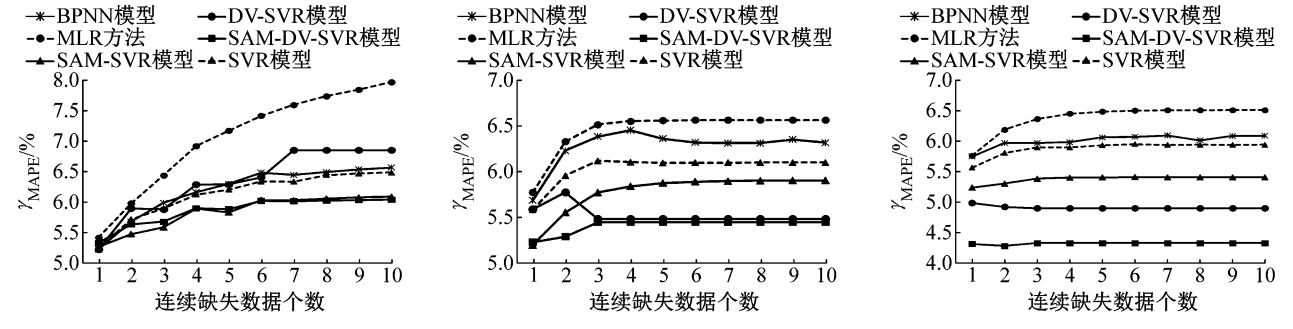
(2) 3 号检测器中 SAM-SVR 模型预测精度明显优于 DV-SVR 模型，而 4 号及 5 号检测器则呈现相反的结论. 原因为 3 号检测器工作日每天流量的周期性变化趋势更为一致，考虑周期性的 SAM-SVR 模型可充分利用流量数据的周期性更好地进行缺失数据的修复. 同时，3 号检测器的时间相关序列的相关系数明显大于空间相关序列的相关系数，采用 DV-SVR 模型在数据连续缺失达到 7 个时会选择空间相关序列进行数据修复，数据修复精度明显较低. 4 号和 5 号检测器空间相关序列的相关性大于时间相关序列的相关性，采用动态变量的 DV-SVR 模型可选择相关性强的空间相关序列作为输入变量以提升缺失数据修复精度.



a 3 号检测器　　b 4 号检测器　　c 5 号检测器

图 8　不同连续缺失数据个数下 6 种模型修复平均绝对误差

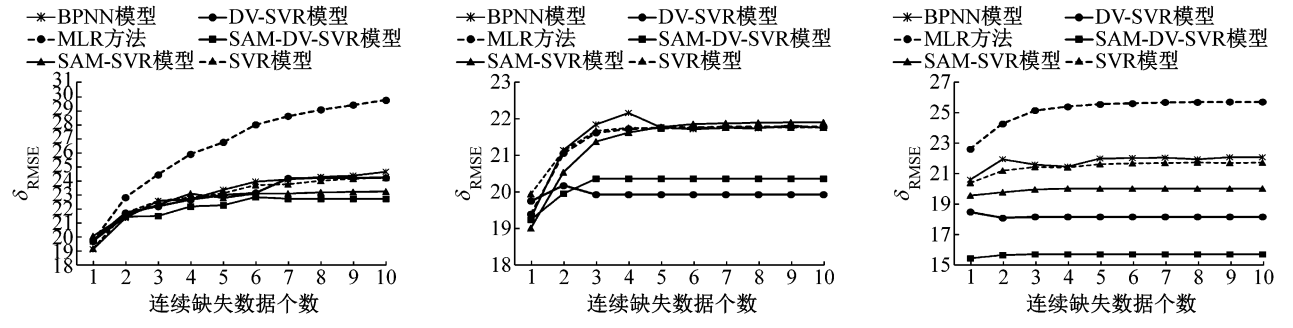
Fig. 8　β_{MAE} of 6 models for different numbers of continuous missing data



a 3 号检测器　　b 4 号检测器　　c 5 号检测器

图 9　不同连续缺失数据个数下 6 种模型修复平均绝对百分比误差

Fig. 9　γ_{MAPE} of 6 models for different numbers of continuous missing data



a 3 号检测器　　b 4 号检测器　　c 5 号检测器

图 10　不同连续缺失数据个数下 6 种模型修复均方根误差

Fig. 10　δ_{RMSE} of 6 models for different numbers of continuous missing data

(3) SAM-DV-SVR 模型对 5 号检测器的数据修复精度提升最为明显,相较于传统的 SVR 模型,在数据连续缺失 1～10 个的情况下,平均绝对误差平均减小了 25.23%,而且平均绝对百分比误差均低于 5%.原因为 5 号检测器的流量数据既具有较为一致的日变化趋势,又与相邻检测器的空间相关序列具有较强的相关性.因此,相较于传统的 SVR 模型,考虑周期性的 SAM-SVR 模型可提升数据修复精度,动态选择解释变量的 DV-SVR 模型在数据连续缺失时也可利用相关性强的空间相关序列进行数据修复以保证缺失数据的修复精度.SAM-DV-SVR 模型将上述 2 种因素进行综合考虑,因此可较大幅度地提升 5 号检测器的数据修复精度.

4 结语

SAM-DV-SVR 模型不仅为数据修复模型选择了最佳的解释变量,还综合考虑了交通流数据的周期性变化趋势和实时变化特征.与常用的几种数据修复模型在数据连续缺失 1 至 10 个的条件下数据修复精度的对比结果可以看出,SAM-DV-SVR 模型体现了更高的数据修复精度.

目前仅验证了快速路交通流数据中的流量数据修复,未对普通道路的间断交通流数据进行模型应用验证,在后期研究中予以考虑.另外,本研究采集的数据为断面交通流数据,因此在空间相关序列选择时未考虑同一断面相邻车道情况,后续研究可补充该数据以进行模型的验证.

参考文献:

[1] 陆化普,屈闻聪,孙智源.基于 S-G 滤波的交通流故障数据识别与修复算法[J].土木工程学报,2015(5):123.
LU Huapu, QU Wencong, SUN Zhiyuan. Detection and repair algorithm of traffic erroneous data based on S-G filtering[J]. China Civil Engineering Journal,2015(5):123.

[2] 姜桂艳,江龙晖,张晓东,等.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.
JIANG Guiyan, JIANG Longhui, ZHANG Xiaodong, *et al*. Malfunction identifying and modifying of dynamic traffic data [J]. Journal of Traffic and Transportation Engineering, 2004 (1):121.

[3] 孙玲,刘浩,牛树云.考虑时空相关性的固定检测缺失数据重构算法[J].交通运输工程学报,2010(5):121.
SUN Ling, LIU Hao, NIU Shuyun. Reconstructive method of missing data for location-specific detector considering spatio-temporal relationship[J]. Journal of Traffic and Transportation Engineering, 2010(5):121.

[4] SMITH B L, SCHERER W T, CONKLIN J H. Exploring imputation techniques for missing data in transportation management system [J]. Transportation Research Record: Journal of the Transportation Research Board, 2003,1836(1): 132.

[5] BOYLES S. A comparison of interpolation methods for missing traffic volume data [C]// Proceedings of the 90th Annual Meeting of the Transportation Research Board. Washington DC: Transportation Research Board,2011:23-27.

[6] QU L, LI L, ZHANG Y, *et al*. PPCA-based missing data imputation for traffic flow volume: a systematical approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(3):512.

[7] QU L, LI L, ZHANG Y, *et al*. A BPCA based missing value imputing method for traffic flow volume data[C]// Intelligent Vehicles Symposium. Eindhoven: IEEE, 2008:985-990.

[8] LI L, LI Y, LI Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence [J]. Transportation Research, Part C: Emerging Technologies, 2013,34:108.

[9] GHOSH B, BASU B, O'MAHONY M. Time-series modeling for forecasting vehicular traffic flow in Dublin[C]// Proceedings of the 84th Annual Meeting of Transportation Research Board. Washington DC: Transportation Research Board ,2005:1-22.

[10] TANG J, ZHANG G, WANG Y H, *et al*. A hybrid approach to integrate fuzzy *C*-means based imputation method with genetic algorithm for missing traffic volume data estimation [J]. Transportation Research, Part C: Emerging Technologies, 2015,51: 29.

[11] ZHANG Y, LIU Y. Missing traffic flow data prediction using least squares support vector machines in urban arterial streets [C]// Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining. Nashville: IEEE, 2009: 76-83.

[12] ZOU Y, HUA X, ZHANG Y, *et al*. Hybrid short-term freeway speed prediction methods based on periodic analysis [J]. Canadian Journal of Civil Engineering, 2015,42(8):570.

[13] CHEN C, WANG Y, LI L, *et al*. The retrieval of intra-day trend and its influence on traffic prediction[J]. Transportation Research, Part C: Emerging Technologies,2012,22:103.

[14] TANG J, WANG H, WANG Y, *et al*. Hybrid prediction approach based on weekly similarities of traffic flow for different temporal scales[J]. Transportation Research Record: Journal of the Transportation Research Board, 2014, 2443 (1): 21.

[15] 陆百川,郭桂林,肖汶谦,等.基于多尺度主元分析法的动态交通数据故障诊断与修复[J].重庆交通大学学报:自然科学版,2016(1):134.
LU Baichuan, GUO Guilin, XIAO Wenqian, *et al*. Fault diagnosing and modifying of dynamic traffic data based on MSPCA[J]. Journal of Chongqing Jiaotong University: Natural Science, 2016(1): 134.

[16] 向昌盛.基于支持向量机的时间序列组合预测模型[D].长沙:湖南农业大学,2011.
XIANG Changsheng. Time series combination prediction model based on support vector machine [D]. Changsha: Hunan Agricultural University,2011.