Operations Research and Fuzziology 运筹与模糊学, 2022, 12(2), 388-396
Published Online May 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.122040

Hans汉斯

基于SARIMA-GPR模型的短时交通流预测应用研究

王飞云[1]，胡　尧[1,2]

[1]贵州大学数学与统计学院，贵州　贵阳
[2]贵州大学公共大数据国家重点实验室，贵州　贵阳

收稿日期：2022年4月18日；录用日期：2022年5月13日；发布日期：2022年5月19日

摘　要

交通流量数据具有周期性、不平稳性、复杂性等特点，若使用单一模型对其进行预测，则预测效果不是很好，因此提出一种组合的SARIMA-GPR模型。SARIMA (Seasonal Autoregressive Integrated Moving Average)模型与GPR (Gaussian Process Regression)模型分别很好拟合交通流量的线性部分与非线性部分，且GPR模型考虑到数据的噪声，能更好地抓取到数据信息。对原数据进行特征提取与分析，训练SARIMA模型与GPR模型，得到两个预测模型，根据模型的MAE得到两个模型的权重值，得到最终的预测值。将该组合模型与SARIMA、GPR、SVM、SARIMA-SVM组合模型进行预测效果对比，实验结果表明，SARIMA-GPR模型预测效果要优于单一模型，预测结果平均绝对百分比误差(MAPE)减少到4.51%，预测结果更接近真实数据。

关键词

交通流，SARIMA模型，GPR模型，组合模型，预测

Research on the Application of Short-Term Traffic Flow Prediction Based on SARIMA-GPR Model

Feiyun Wang[1], Yao Hu[1,2]

[1]School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
[2]State Key Laboratory of Public Big Data, Guizhou University, Guiyang Guizhou

Received: Apr. 18th, 2022; accepted: May 13th, 2022; published: May 19th, 2022

文章引用: 王飞云, 胡尧. 基于 SARIMA-GPR 模型的短时交通流预测应用研究[J]. 运筹与模糊学, 2022, 12(2): 388-396.
DOI: 10.12677/orf.2022.122040

Abstract

Traffic flow data have the characteristics of periodicity, instability and complexity. If a single model is used to predict it, the prediction effect is not very good. Therefore, a combined SARIMA-GPR model is proposed. The SARIMA (Seasonal Autoregressive Integrated Moving Average) model and the GPR (Gaussian Process Regression) model fit the linear part and the nonlinear part of the traffic flow well respectively, and the GPR model takes into account the noise of the data and can better capture the data information. Perform feature extraction and analysis on the original data, train the SARIMA model and the GPR model, and obtain two prediction models. According to the MAE of the model, the weight values of the two models are obtained, and the final prediction value is obtained. The combined model is compared with SARIMA, GPR, SVM, and SARIMA-SVM combined model. The experimental results show that the prediction effect of SARIMA-GPR model is better than that of a single model, and the mean absolute percentage error (MAPE) of the prediction results is reduced to 4.51%, the prediction effect is closer to the real data.

Keywords

Traffic Flow, SARIMA Model, GPR Model, Combined Model, Prediction

1. 引言

在当今，交通发展给我们生活带来了多方面的好处，因为它提高了人们的出行效率，也提高了社会的生产力。据统计[1]，截至 2020 年初，贵阳市民用车辆有 168.71 万辆，比上一年初增长了 13.0%，其中汽车拥有量有 134.12 万辆，比上一年初增长了 15.9%；私人汽车拥有量有 119.70 万辆，比上一年增长了 16.5%。车辆数持续增长，会使交通问题愈加严重，对于交通拥堵带来的影响，如何有效减少和解决显得尤为重要，短时交通流预测是交管部门有效管理采取实施的主要依据。

2017 年 Nicholas G. Polson 和 Vadim O. Sokolov 开发了一种深度学习模型来预测交通流量，显示出深度学习如何提供精确的短期交通流预测[2]。2018 年 Unsok Ryu 为了提高短期交通流量预测的性能，提出了互信息(MI)构造交通状态向量的方法，并且此方法在短期交通预测中具有良好的预测精度[3]。2020 年 Azadeh Emami 提出了一种基于褪色记忆卡尔曼滤波融合连接车辆和蓝牙传感器数据的短时交通流预测方法[4]。国内对于交通流预测的研究也发展的十分迅速。胡伍生等人[5]为充分利用统计数据，提高交通流预测精度，在 2020 年构建了一种新颖的短期交通流预测的神经网络 BP 模型。温美玲[6]等针对交通拥堵问题，在 2021 年提出了基于轨迹大数据的交通拥堵评估和预测方法，使用深度学习算法建立了交通拥堵预测模型。孙越[7]等人在 2021 年提出一种 ARMA-LSTM 组合模型的对铁路客流量进行预测。温惠英[8]等人在 2021 年提出一种基于时延特性的短时交通流预测研究；张国赟[9]等人在 2022 年提出一种改进 ARIMA 模型对城市轨道交通短时客流进行预测。GPR 是基于统计学习理论和贝叶斯理论发展起来的一种机器学习方法，适于处理非线性复杂回归问题，且泛化能力强，与神经网络、支持向量机相比，GPR 容易实现，因此本文提出一种基于 GPR 模型与 SARIMA 模型相结合的方法来解决道路交通流预测问题，实验结果表明该组合模型的预测效果要优于其他单一预测模型。

2. 数据来源及研究方法

2.1. 数据来源

本文的数据是来自贵州省贵阳市观山湖区过车数据，它涉及到观山湖区的长岭南路与阳关大道等71个交叉路口的过车量，包括了2020年4月13日00:00到2020年4月17日19:10每五分钟的过车量，总共有1384个数据，使用15分钟作为间隔时间，最终分析的数据有461个。1小部分数据如下表1所示：

Table 1. Traffic flow section data
表1. 交通流部分数据表

时间	车流量/辆
2020.4.17 17:30	42,301
2020.4.17 17:45	43,058
2020.4.17 18:00	14,297
2020.4.17 18:15	162
2020.4.17 18:30	125
2020.4.17 18:45	107
2020.4.17 19:00	93

根据表1可以看出最后的四个数据很明显为异常数据(一般是指出现不在范围的数据)，倒数第五个数据也出现一定的偏差，而这里交通流数据的异常数据一般指交通流远小于道路通行能力，这里出现异常数据的原因可能有如下方面：一是设备出现故障，二是车流量密度过大造成收集数据不能正常传送到目标位置。异常数据会影响到后面数据分析预测的精度，因此需要对异常数据进行分析处理，使数据具有较高的质量，从而去提高交通流的预测精度。

对于异常数据的处理，在不考虑突发因素(比如交通事故)情况下，一般是先将其删掉再采用近似值修复补齐方式。这里使用相邻时段交通流量的平均值法，公式如下：

$$\tilde{x}(t) = E(X) - 2\sqrt{Var(X)} \tag{1}$$

式中 $E(X)$ 是与 t 时刻相邻的交通流量期望值，$\sqrt{Var(X)}$ 是与 t 时刻相邻的交通流量的标准差，取四个时刻与 t 时刻相邻。由此计算得到最后五个数据修补后的值分别为38,985、37,904、35,555、32,607、30,609。

2.2. 研究方法

2.2.1. SARIMA 模型

季节性差分自回归滑动平均[10] (Seasonal Autoregressive Integrated Moving Average, SARIMA)模型是指序列中的季节效应和其他效应之间的关系具有一定的关系。又根据对季节效性提取的难易程度，将其分为简单季节模型与乘积季节模型。

a) 简单季节模型。当序列中的季节效应与其他效应之间的关系属于加法关系，即：

$$X_t = T_t + S_t + I_t.$$

这个时候的各种效应的信息提取是较容易的，然后使用 d 阶的趋势差分，D 步的季节差分运算，将简单的季节性模型拟合到原始序列的观测值。模型结构如下：

$$\nabla_D \nabla^d x_t = \frac{\Theta(B)}{\Phi(B)} \varepsilon_t. \tag{2}$$

式中：

D 为周期的步长，d 是提取趋势信息所用的差分阶数，$\{\varepsilon_t\}$ 为白噪声序列，并且 $E(\varepsilon_t)=0$，$Var(\varepsilon_t)=\sigma_\varepsilon^2$。

$\Phi(B)=1-\phi_1 B-\cdots-\phi_p B^p$，为 p 阶自回归系数多项式。

$\Theta(B)=1-\theta_1 B-\cdots-\theta_q B^q$，为 q 阶移动平均系数多项式。

该简单季节模型简记为 ARIMA$(p,d,q)\times(0,1,0)[D]$。

在使用 SARIMA 模型预测时，首先需要判断该序列的平稳性与季节性，如果序列为非平稳序列，则要通过差分的方式使之达到平稳序列，然后再用 ARMA 建模进行预测。

b) 乘积季节模型。当序列中的季节效应与其他效应之间的关系属于乘法关系，即：

$$X_t=T_t\cdot S_t\cdot I_t.$$

这个时候使用 d 阶的趋势差分，D 阶以 S 为周期的季节差分运算，将乘积的季节性模型拟合到原始序列的观测值。模型结构如下：

$$\nabla^d\nabla_S^D x_t=\frac{\Theta(B)\Theta_S(B)}{\Phi(B)\Phi_S(B)}\varepsilon_t. \tag{3}$$

式中：

$$\Phi(B)=1-\phi_1 B-\cdots-\phi_p B^p,\ \Theta(B)=1-\theta_1 B-\cdots-\theta_q B^q,$$

$$\Phi_S(B)=1-\phi_1 B^P-\cdots-\phi_P B^{PS},\ \Theta_S(B)=1-\theta_1 B^Q-\cdots-\theta_Q B^{QS}$$

该乘积季节模型简记为 ARIMA$(p,d,q)\times(P,D,Q)_S$。

2.2.2. GPR 模型

高斯过程回归[11] (Gaussian Process Regression, GPR)模型，它可以是线性模型，也可以是非线性模型，本文利用它建立非线性回归模型，它是利用 Bayes 思想的一种监督学习方法，也是一种非参数回归。简单来说，就是通过贝叶斯的方法求出参数的后验分布，然后根据参数的后验分布与训练数据来求出测试数据的分布，从而进行贝叶斯估计与区间预测[12]。它与支持向量机相较，模型比较好解释，并且它加入了噪声过程，能够更好拟合数据，抓取数据信息。该回归模型可表示为

$$Y(\boldsymbol{x})=\sum_{j=1}^{p} f_j(\boldsymbol{x})\beta_j+Z(\boldsymbol{x})=f^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\beta}+Z(\boldsymbol{x}). \tag{4}$$

其中 $f_1(\cdot),\cdots,f_p(\cdot)$ 表示知道的回归函数，$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 是未知的回归系数向量，$Z(\cdot)$ 是一个 0 均值的平稳高斯过程。$f^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\beta}$ 描述 $\boldsymbol{x}$ 的长期趋势，而 $Z(\boldsymbol{x})$ 模型则是局部偏离长期趋势。

假设 $\boldsymbol{y}^{tr}=\left(y\left(\boldsymbol{x}_1^{tr}\right),\cdots,y\left(\boldsymbol{x}_{n_s}^{tr}\right)\right)^{\mathrm{T}}$ 表示知道的 n_s 个训练数据，$\boldsymbol{y}^{te}=\left(y\left(\boldsymbol{x}_1^{te}\right),\cdots,y\left(\boldsymbol{x}_{n_s}^{te}\right)\right)^{\mathrm{T}}$ 表示未知的 n_e 个测试数据，$\boldsymbol{Y}^{tr}=\left(Y\left(\boldsymbol{x}_1^{tr}\right),\cdots,Y\left(\boldsymbol{x}_{n_s}^{tr}\right)\right)^{\mathrm{T}}$ 和 $\boldsymbol{y}^{te}=\left(Y\left(\boldsymbol{x}_1^{te}\right),\cdots,Y\left(\boldsymbol{x}_{n_s}^{te}\right)\right)^{\mathrm{T}}$ 分别表示训练数据与测试数据过程模型。假设

$$\left[\begin{pmatrix}\boldsymbol{Y}^{te}\\ \boldsymbol{Y}^{tr}\end{pmatrix}\middle|\boldsymbol{\beta}\right]\sim N_{n_e+n_s}\left(\begin{pmatrix}\boldsymbol{F}_{te}\\ \boldsymbol{F}_{tr}\end{pmatrix}\boldsymbol{\beta},\lambda_z^{-1}\begin{pmatrix}\boldsymbol{R}_{te} & \boldsymbol{R}_{te,tr}\\ \boldsymbol{R}_{te,tr}^{\mathrm{T}} & R_{tr}\end{pmatrix}\right). \tag{5}$$

其中 $\lambda_z^{-1}=\frac{1}{\sigma^2}$ 表示过程精度，$\boldsymbol{F}_{te}$ 和 $\boldsymbol{F}_{tr}$ 分别表示测试集与训练集的回归函数。

$\pi(\boldsymbol{\beta})$表示$\boldsymbol{\beta}$的密度函数，则由贝叶斯公式可得到$\boldsymbol{\beta}$后验分布的密度函数为

$$\pi\left(\boldsymbol{\beta}\left|\boldsymbol{Y}^{tr}\right.\right)=\frac{\pi\left(\boldsymbol{Y}^{tr}\mid\boldsymbol{\beta}\right)\cdot\pi(\boldsymbol{\beta})}{\pi\left(\boldsymbol{Y}^{tr}\right)}\propto\pi\left(\boldsymbol{Y}^{tr}\left|\boldsymbol{\beta}\right.\right)\cdot\pi(\boldsymbol{\beta}). \tag{6}$$

则由式(6)和多元正态分布的条件分布可得到

$$\left[\boldsymbol{Y}^{te}\left|\boldsymbol{Y}^{tr}=y^{tr},\boldsymbol{\beta}\right.\right]\sim N_{n_e}\left(\boldsymbol{F}_{te}\boldsymbol{\beta}+\boldsymbol{R}_{te,tr}\boldsymbol{R}_{tr}^{-1}\left(y^{tr}-\boldsymbol{F}^{te}\boldsymbol{\beta}\right),\lambda_z^{-1}\left(\boldsymbol{R}_{te}-\boldsymbol{R}_{te,tr}\boldsymbol{R}_{tr}^{-1}\boldsymbol{R}_{te,tr}^{\mathrm{T}}\right)\right). \tag{7}$$

$$\pi\left(y^{te}\left|y^{tr}\right.\right)=\int\pi\left(y^{te},\boldsymbol{\beta}\left|y^{tr}\right.\right)\mathrm{d}\boldsymbol{\beta}=\int\pi\left(y^{te}\left|\boldsymbol{\beta},y^{tr}\right.\right)\pi\left(\boldsymbol{\beta}\left|y^{tr}\right.\right)\mathrm{d}\boldsymbol{\beta}. \tag{8}$$

假设$\boldsymbol{\beta}$的先验信息分布服从一个多元正态分布，也就是$[\boldsymbol{\beta}]\sim N_p\left(\boldsymbol{b}_\beta,\lambda_\beta^{-1}\boldsymbol{V}_\beta\right)$，从而可以得$\boldsymbol{\beta}$的后验分布为

$$\left[\boldsymbol{\beta}\left|\boldsymbol{Y}^{tr}=y^{tr}\right.\right]\sim N_p\left(\boldsymbol{\mu}_{\beta|tr},\boldsymbol{\Sigma}_{\beta|tr}\right). \tag{9}$$

$$\boldsymbol{\mu}_{\beta|tr}=\left(\lambda_z\boldsymbol{F}_{tr}^{\mathrm{T}}\boldsymbol{R}_{tr}^{-1}\boldsymbol{F}_{tr}+\lambda_\beta\boldsymbol{V}_\beta^{-1}\right)^{-1}\times\left(\lambda_z\boldsymbol{F}_{tr}^{\mathrm{T}}\boldsymbol{R}_{tr}^{-1}y_{tr}+\lambda_\beta\boldsymbol{V}_\beta^{-1}\boldsymbol{b}_\beta\right)$$

$$\boldsymbol{\Sigma}_{\beta|tr}=\left(\lambda_z\boldsymbol{F}_{tr}^{\mathrm{T}}\boldsymbol{R}_{tr}^{-1}\boldsymbol{F}_{tr}+\lambda_\beta\boldsymbol{V}_\beta^{-1}\right)^{-1}$$

从而得到测试数据$\boldsymbol{Y}^{te}$的预测分布为

$$\left[\boldsymbol{Y}^{te}\left|\boldsymbol{Y}^{tr}=y^{tr}\right.\right]\sim N_{n_e}\left(\boldsymbol{\mu}_{te|tr},\boldsymbol{\Sigma}_{te|tr}\right). \tag{10}$$

$$\boldsymbol{\mu}_{te|tr}=\boldsymbol{F}_{te}\boldsymbol{\mu}_{\beta|tr}+\boldsymbol{R}_{te,tr}\boldsymbol{R}_{tr}^{-1}\left(y^{tr}-\boldsymbol{F}_{tr}\boldsymbol{\mu}_{\beta|tr}\right),$$

$$\boldsymbol{\Sigma}_{te|tr}=\lambda_z^{-1}\left\{\boldsymbol{R}_{te}-\left(\boldsymbol{F}_{te}\quad\boldsymbol{F}_{tr}\right)\begin{bmatrix}\boldsymbol{V}_\beta^{-1}&\boldsymbol{F}_{tr}^{\mathrm{T}}\\\boldsymbol{F}_{tr}&\boldsymbol{R}_{tr}\end{bmatrix}\begin{pmatrix}\boldsymbol{F}_{te}^{\mathrm{T}}\\\boldsymbol{R}_{te,tr}^{\mathrm{T}}\end{pmatrix}\right\}$$

得到预测数据$\hat{\boldsymbol{y}}^{te}=E\left[\boldsymbol{Y}^{te}\left|\boldsymbol{Y}^{tr}\right.\right]=\boldsymbol{\mu}_{te|tr}$。对于训练数据与测试数据之间的协方差我们利用核函数进行求解，也就是$k(t,s)=Cov\left[\boldsymbol{Y}^{te},\boldsymbol{Y}^{tr}\right]$。

2.2.3. SARIMA-GPR 模型

在实际应用中，由于交通流数据的复杂性与多变性，在做预测的时候可能会伴随着许多不确定的影响因素。并且对于交通流预测问题，可以采用不同的预测方法去建立多种模型进行预测，但是没有哪一个模型能够完全用于所有的交通状况，因此利用多个单一的模型来进行组合预测。组合模型能够有效的利用单一模型的优点，按照一定的规律将各种单一的模型组合起来。

在该组合模型当中，最重要的就是权值的选择，这是因为用不同的模型对同一个序列提取到的信息是不同的，因此，权值的选择会决定着模型预测的好坏。这里使用 MAE 权重系数[13]来构建组合模型，对于 MAE 小的模型所占的权重要大，从而提升预测效果。

观测数据为$X=(x_1,x_2,\cdots,x_n)$，SARIMA 与 GPR 的预测模型$\{\varphi_1,\varphi_2\}$在t时刻预测值记为$\{\varphi_1(t),\varphi_2(t)\}$，设其权重为$\{\lambda_1,\lambda_2\}$，因此可以得到组合预测模型为

$$\varphi(t)=\lambda_1\varphi_1(t)+\lambda_2\varphi_2(t) \tag{11}$$

得到该组合模型的 MAE 为：$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\varphi(t)-x(t)\right|,t=1,2,\cdots,n$。

$\lambda_1=\frac{d_2}{d_1+d_2},\lambda_2=\frac{d_1}{d_1+d_2}$，$d_1,d_2$分别是这两个模型的 MAE。

3. 实例分析

3.1. 基于 SARIMA-GPR 模型的道路车流量预测

把处理后的这 461 个数据按照 8:2 的比例划分为训练集与测试集。即是使用 384 个数据(2020 年 4 月 13 日 00:00 到 2020 年 4 月 16 日 23:45 来训练模型，77 个数据(2020 年 4 月 17 日 00:00 到 2020 年 4 月 17 日 19:00)来测试模型的好坏。通常评价一个模型最优的预测评价方法就是将预测误差尽可能降到最小。预测误差的计算方法有很多种[14]：有常见的均方误差根(RMSE)、平均误差(ME)、平均绝对误差(MAE)、平均百分比误差(MPE)、平均绝对百分比误差(MAPE)等。

本文对模型的预测效果的好坏评估采取均方误差(RMSE)、平均绝对误差(MAE)以及平均绝对百分比误差(MAPE)。其计算公式分别为：

$$\mathrm{RMSE}(y,\hat{y})=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2},$$

$$\mathrm{MAE}(y,\hat{y})=\frac{1}{n}\left(\sum_{i=1}^{n}|y_i-\hat{y}_i|\right),$$

$$\mathrm{MAPE}(y,\hat{y})=\frac{1}{n}\left(\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\right)\times 100\%$$

其中，y_i 表示真实值，$\hat{y}_i$ 表示预测值，n 表示数据的个数。判断为：RMSE、MAE 和 MAPE 的值越小，就证明预测的模型具有更好的精度。

对数据先进行 SARIMA 信息提取，观察数据的时序图以及自相关系数图进行建模。

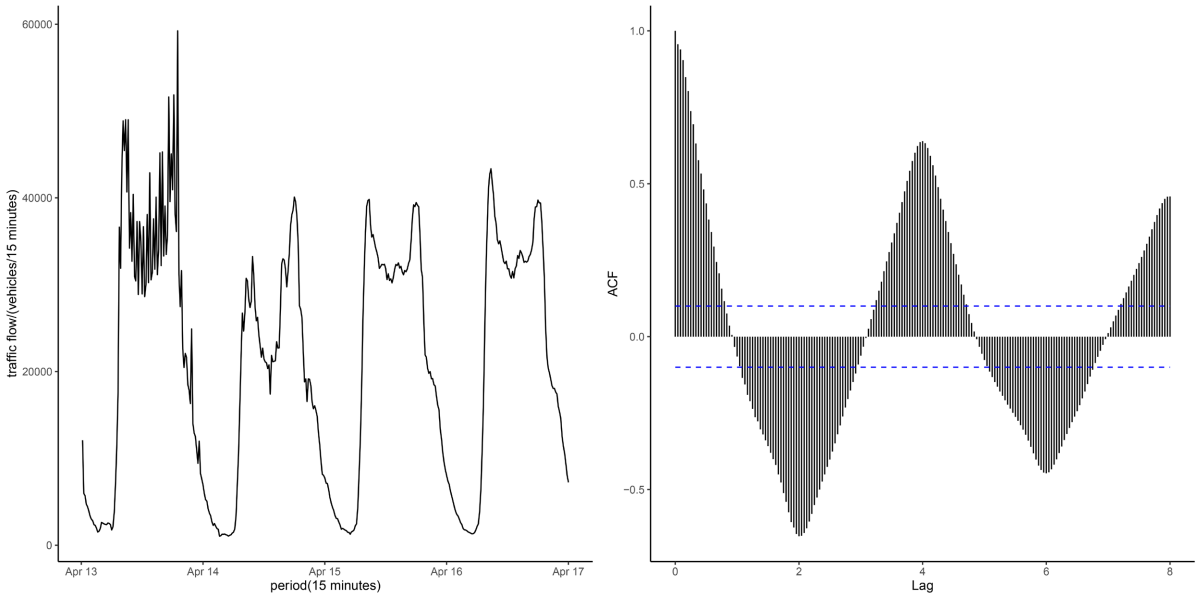

Figure 1. Traffic flow time series diagram (left) and autocorrelation coefficient diagram (right)
图 1. 车流量时序图(左)与自相关系数图(右)

看出车流量数据伴随着一定的季节性(周期性)，有着明显的非线性趋势，属于非平稳数据。季节波动的振幅不受到趋势变动的影响，那么季节与趋势之间通常没有交互影响关系，因此采用简单季节模型。根据图 1 (右)，以及 *R* 中 auto.arima 函数自动定阶，根据模型定阶原则，再根据 AIC 与 BIC 准则[15]，建

立了 ARIMA$(3,2,3)(0,1,0)[96]$ 模型。

得到该模型残差检验结果如下表 2 所示。当延迟期数为 6 期的时候，模型假设检验的 p 值为 0.1187 显著大于 0.05，即是认为在 0.05 的显著性水平下不能拒绝原假设，因此，该残差序列为白噪声序列，即表明残差序列中相关信息已经被提取出来了，所以所建立的模型有效。

Table 2. Model residual series test results
表 2. 模型残差序列检验结果

模型	残差序列 LB 统计量	期数	p-value
ARIMA$(3,2,3)(0,1,0)[96]$	10.144	6	0.1187

对于高斯过程回归模型，令 $\boldsymbol{b}_\beta = 0$，核函数使用最常见的高斯径向基核函数与能对周期性建模的正弦平方内核函数进行组合使预测效果更好。高斯径向基核函数和正弦平方内核函数公式分别为

$$k(x_i, x) = \exp\left\{-\frac{\|x - x_i\|^2}{2\delta^2}\right\},$$
$$k(x_i, x) = \exp\left\{-\frac{2\sin^2\left(\pi\|x - x_i\|^2\right)/p}{\delta^2}\right\}. \tag{12}$$

则可以使用组合的核函数

$$k(t, s) = \exp\left\{-\frac{\|t - s\|^2}{2\delta^2}\right\} + \exp\left\{-\frac{2\sin^2\left(\pi\|t - s\|^2\right)/p}{\delta^2}\right\} \tag{13}$$

其中 δ^2 表示长度的缩放系数，p 表示核函数的周期。对于模型参数的求解，利用对数边际似然方法，进行求解，见(6)式。通过计算得到参数 $\boldsymbol{\omega} = \{\delta^2, p\} = \{45, 8\}$。

基于以上分析，可以知道 SARIMA 模型的平均绝对误差(MAE) $d_1 = 1153$，GPR 模型的平均绝对误差(MAE) $d_2 = 2705$，可以得到组合模型的权重值分别为 $\lambda_1 = \frac{2846}{3999}, \lambda_2 = \frac{1153}{3999}$。

3.2. 模型预测对比图

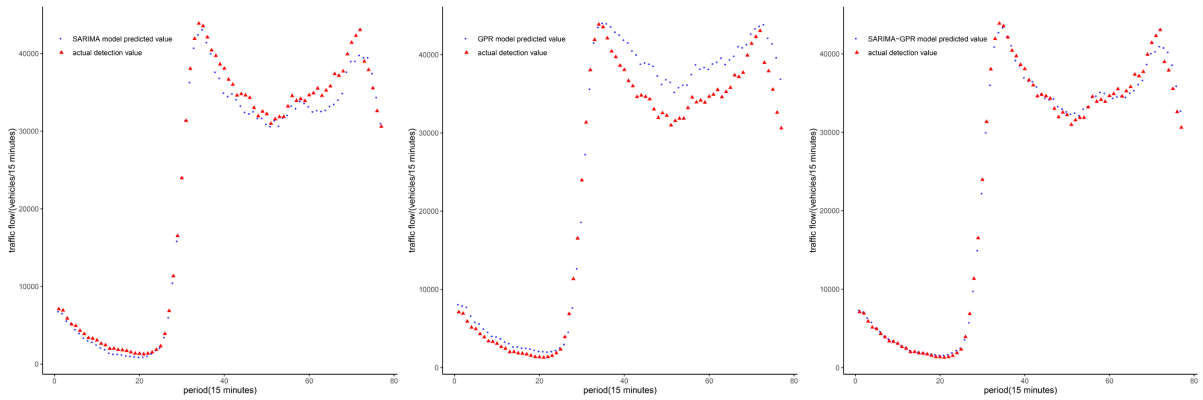

Figure 2. Comparison of predicted values and true values of three single models
图 2. 三种模型预测值与真实值对比图

从图 2 可以看出 SARIMA 模型(左)、GPR 模型(中)和 SARIMA-GPR 模型(右)在整体上的拟合效果还是不错的，预测值与真实值大致走势一样，但明显看出 SARIMA-GPR 模型预测值更接近真实值，表明所提出的模型更有效。

3.3. 模型预测效果对比分析

为了进一步说明 SARIMA-GPR 模型对该道路交通流数据的预测效果，采用相同的数据训练方式对 SVM 模型进行训练，SVM 也是对数据进行非线性拟合较好的一个模型，以及用同样的权值选择方式构建一个 SARIMA-SVM 模型并对其进行训练，对它们进行预测效果评估(表 3，图 3)。

Table 3. Comparison of the prediction effects of the five models
表 3. 五种模型的预测效果对比结果

模型	RMSE	MAE	MAPE (%)
SARIMA	1498.91	1152.61	9.90
SVM	3229.54	2846.05	31.25
GPR	3260.75	2705.21	19.05
SARIMA-SVM	**1057.44**	750.40	4.64
SARIMA-GPR	1071.15	**749.72**	**4.51**

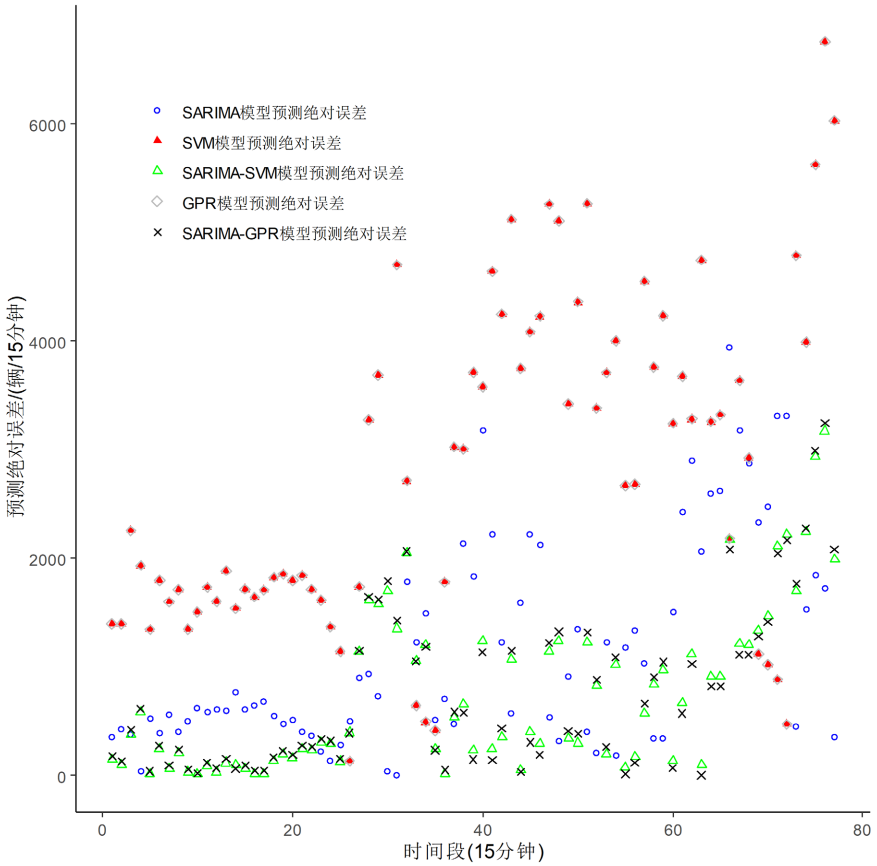

Figure 3. Comparison diagram of absolute error prediction of five models
图 3. 五种模型预测绝对误差对比图

通过表 3 看出 SARIMA-GPR 模型预测结果的 MAE 与 MAPE 是这五种模型里面是最小的，SARIMA-GPR 模型预测平均绝对误差(图 3)要小于其他几个模型，预测效果也要优于其他模型。SARIMA 模型能够很好拟合数据线性部分，GPR 模型考虑到数据的噪声，能够更好抓取数据信息，将两者结合，对原序列的预测效果更好，因此说 SARIMA-GPR 模型测效果是最好的，能够很好提高预测准确度。

4. 结论

通过对贵阳市观山湖区的交通流数据进行分析，可以知道该交通流数据属于非平稳的时间序列数据，具有一定的季节特征，也有非线性特征。针对于交流流量的特征，提出了一种基于机器学习的 SARIMA-GPR 模型对该道路交通流进行预测。实验结果表明，该模型综合了单一模型的优势，使预测效果更好，提高了预测精度。本文的数据是五天工作日的数据，在以后可以考虑获取非工作日的数据再次进行建模，使预测效果达到最好。

基金项目

国家自然科学基金资助项目“道路交通数据的统计模型诊断理论与应用研究”(12161016)。

参考文献

[1] 2019 年贵阳市国民经济和社会发展统计公报[Z]. 贵阳市统计局, 2020. http://tjj.guiyang.gov.cn/2020_zwgk

[2] Polson, N.G. and Sokolov, V.O. (2017) Deep Learning for Short-Term Traffic Flow Prediction. *Transportation Research Part C*: *Emerging Technologies*, **79**, 1-17. https://doi.org/10.1016/j.trc.2017.02.024

[3] Ryu, U., Wang, J., Kim, T., *et al*. (2018) Construction of Traffic State Vector Using Mutual Information for Short-Term Traffic Flow Prediction. *Transportation Research Part C*: *Emerging Technologies*, **96**, 55-71. https://doi.org/10.1016/j.trc.2018.09.015

[4] Emami, A., Sarvi, M. and Bagloee, S.A. (2020) Short-Term Traffic Flow Prediction Based on Faded Memory Kalman Filter Fusing Data from Connected Vehicles and Bluetooth Sensors. *Simulation Modelling Practice and Theory*, **102**, Article ID: 102025. https://doi.org/10.1016/j.simpat.2019.102025

[5] 胡伍生, 吕楚男, 夏晓明. 基于神经网络的短期交通流预测模型[J]. 现代测绘, 2020, 43(5): 10-13

[6] 温美玲, 路鹏远, 蔡林, 等. 基于轨迹大数据的交通拥堵评估和预测[J]. 数字制造科学, 2021, 19(1): 77-80.

[7] 孙越, 宋晓宇, 金莉婷, 刘童.基于 ARMA-LSTM 组合模型的铁路客流量预测[J]. 计算机应用与软件, 2021, 38(12): 262-267+273.

[8] 温惠英, 曹正. 基于时延特性的短时动态交通流预测模型研究[J]. 计算机仿真, 2021, 38(6): 93-97.

[9] 张国赟, 金辉. 基于改进 ARIMA 模型的城市轨道交通短时客流预测研究[J]. 计算机应用与软件, 2022, 39(1): 339-344.

[10] Yaya, O.S. and Fashae, O.A. (2015) Seasonal Fractional Integrated Time Series Models for Rainfall Data in Nigeria. *Theoretical and Applied Climatology*, **120**, 99-108. https://doi.org/10.1007/s00704-014-1153-8

[11] Thomas, J., Brian, J. and William, I. (2018) The Design and Analysis of Computer Experiments. Springer Science + Business Media, LLC, Berlin, 115-118.

[12] Sugasawa, S. (2020) Robust Empirical Bayes Small Area Estimation with Density Power Divergence. *Biometrika*, **107**, 467-480. https://doi.org/10.1093/biomet/asz075

[13] Liu, Z., Jiang, P., Zhang, L., *et al*. (2020) A Combined Forecasting Model for Time Series: Application to Short-Term Wind Speed Forecasting. *Applied Energy*, **259**, Article ID: 114137. https://doi.org/10.1016/j.apenergy.2019.114137

[14] 李键, 牛峰, 黄晓艳, 等. 永磁同步电机有限控集模型预测电流控制预测误差分析[J]. 电机与控制学报, 2019, 23(4):1-7.

[15] Dridi, N. and Hadzagic, M. (2018) Akaike and Bayesian Information Criteria for Hidden Markov Models. *IEEE Signal Processing Letters*, **26**, 302-306. https://doi.org/10.1109/LSP.2018.2886933