文章编号：2095-2163(2021)10-0012-08 中图分类号：U491 文献标志码：A

# 基于 ARIMAX 的城市道路交通流短期预测模型

袁鹏程[1]，周天乐[2]
（1 上海理工大学 管理学院，上海 200093；2 上海电科智能系统股份有限公司，上海 200063）

**摘 要**：随着出行者对交通信息的预期依赖的增加，构建和发展更为精准的交通流预测模型显得更加具有实际意义。ARIMA模型作为常见的时间序列处理工具被广泛应用于各个领域。然而，ARIMA 预测模型构建是建立在平稳时间序列基础上的，但是其实建立在一元变量的基础上，并且在具体的模型构建过程中一元变量还通常会因差分而造成有用数据信息的丢失，影响最终预测结果。为此，本文考虑通过引入新的参数来弥补传统模型因差分造成信息丢失，构建基于交通流短期预测的 ARIMAX 模型。利用构建的 ARIMAX 模型对 5 天的交通流量进行预测，仿真显示模型的结果误差较小，说明该模型具有一定的实用价值。

**关键词**：交通流预测；ARIMAX 模型；交通流平稳性

## Traffic flow forecasting based on ARIMAX model

YUAN Pengcheng[1], ZHOU Tianle[2]
(1 Business School, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 Shanghai Dianke Intelligent System Co. Ltd., Shanghai 200063, China)

**【Abstract】** With the increase of travelers´ expected dependence on traffic information, it is more practical to build and develop more accurate traffic flow prediction model. As a common time series processing tool, ARIMA model is widely used in various fields. However, ARIMA prediction model is built on the basis of stationary time series, but in fact, it is built on the basis of unary variables. In the process of specific model construction, unary variables usually cause the loss of useful data information due to difference, which affects the final prediction results. Therefore, this paper considers introducing new parameters to make up for the information loss caused by difference in traditional models, and constructs ARIMAX model based on short-term traffic flow prediction. The ARIMAX model is used to predict the 5-day traffic flow. The results show that the model can reduce the error, which shows that the model has a certain practical value.

**【Key words】** traffic flow forecasting; ARIMAX model; stability of traffic flow

## 0 引 言

随着交通拥堵和不确定性逐渐成为新常态，车联网、自动驾驶和大数据技术也得到了不断发展，交通流研究将会进入重要的变革期。而交通流特性主要由交通流速度、密度和流量三个部分组成，其中交通流量尤为重要，并能直接反映交通运行状况。精准短时交通流量预测就可以直观反映调查路段或地区的交通变化状况，为交通控制与管理提供可靠依据。同时，也能为出行者提供准确地道路信息，避免不必要的拥堵。

目前，国内外对于交通流量预测已经做过很多研究[1]。最常见的就是基于统计方法的模型和神经网络模型。自上世纪七十年代末，ARIMA 模型[2]提出以来，即已广泛应用于各个领域[3]。但由于 ARIMA 模型的局限性等因素，往往会结合数据自身特点加以调整[4-5]。例如，针对模型单一的问题，田瑞杰等人[6]提出一种时间序列与人工神经网络相结合的预测模型；基于时间序列分析方法，韩超等人[7]提出一种短时交通流实时自适应预测算法，减小遗忘因子进一步提高预测的性能；针对 ARIMA 模型获取非线性特性的局限性，王晓全等人[8]加入广义自回归条件异方差—均值，相比于 ARIMA-SVR 模型和 ARIMA-GARCH 模型得到了更好的预测精度；通过证实交通流量存在时序上的周期性，祁伟等人[9]引入季节性 ARIMA 模型融合了邻近的交通流观察值和交通流数据的周期性。此外，也有深度学习[10]、基于相空间重构理论的局部预测方法[11]等研究。在上述交通流预测过程中仅仅利用了交通流量自身信息进行预测，并没有加入其他影响因素用

基金项目：国家自然科学基金(71601118)。

作者简介：袁鹏程(1982-)，男，博士，副教授，硕士生导师，主要研究方向：交通系统建模与分析；周天乐(1994-)，男，硕士，工程师，主要研究方向：交通流预测、智能交通系统。

收稿日期：2021-07-22

于提高预测精度，丁永兵等人[12]通过结合路网结构，利用主成分回归建立上下游交通流回归模型，对模型残差进行 ARIMA 建模，得到的 ARIMAX 模型要优于 ARIMA 模型。但在交通领域并没有考虑将影响交通流量的因素（例如：道路占有率等）加入模型进行预测，而在其他的一些研究方向[13-14]就考虑将相关的参数加入模型进行预测，并取得了不错的效果。

构建传统时间序列模型的前提条件就是时间序列的平稳。通常为了达到序列的平稳性会对原序列进行差分处理，但却会丢失了数据信息。本文考虑引入道路占有率等因素来增加原始数据信息提高预测精度。研究中，首先介绍了 ARIMAX 模型的原理，接着对原始数据进行预处理，使其达到平稳的条件，然后通过利用 Python 来搭建 ARIMAX 模型拟合参数，继而对构建的模型加以验证，最后进行交通流预测。通过分析最终评价指标结果可知，模型拟合效果较好，各种误差结果均偏小，达到了预期的效果。

## 1　模型理论

### 1.1　ARIMAX 模型

差分自回归移动平均模型（Autoregressive Integrated Moving Average Model，ARIMA 模型）是通过自回归移动平均模型（Auto Regression Moving Average Model，ARMA 模型）扩展而来的。ARIMA 模型中，先对时间序列进行差分使其达到平稳状态，再对差分后的时间序列建立 ARMA 模型。而 ARMA 模型是将自回归模型（Auto Regression Model，AR 模型）和移动平均模型（Moving Average Model，MA 模型）有机组合而成的。对此拟展开研究分述如下。

#### 1.1.1　自回归模型 AR

$p$ 阶自回归模型，记为 $AR(p)$，是一种处理时间序列的方法，用同一变数如 $x$ 的之前各期，即 $x_t$ 至 $x_{t-p}$ 的值来预测 $x_t$ 的值，并假设各数值之间为线性关系。公式如下：

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t \tag{1}$$

其中，$c$ 为常数项；$\varepsilon_t$ 是均值为零，标准差为 $\sigma$ 的随机误差项。

当引入延迟算子 $B$，即 $B^n x_t = x_{t-n}$，并将 $AR(p)$ 模型中心化后，可简记为：

$$\Phi(B)\, x_t = \varepsilon_t \tag{2}$$

其中，$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$，称为 $p$ 阶自回归系数多项式。

#### 1.1.2　移动平均模型 MA

$q$ 阶移动平均模型，记为 $MA(q)$，是一种简单平滑预测模型，可根据时间序列 $x_t$ 至 $x_{t-p}$ 的平均值，以预测 $x_t$ 的值。其公式如下：

$$x_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{3}$$

其中，$\mu$ 是序列均值，$\theta_1, \cdots, \theta_q$ 是参数，$\varepsilon_t, \cdots, \varepsilon_{t-q}$ 都是白噪声。

当引入延迟算子 $B$，即可得到 $B^n x_t = x_{t-n}$，并将 $MA(q)$ 模型中心化后，可简记为：

$$x_t = \Theta(B)\, \varepsilon_t \tag{4}$$

其中，$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$，称为 $q$ 阶移动平均系数多项式。

#### 1.1.3　ARIMAX 模型

ARIMAX 模型就是带输入变量的 ARIMA 模型，其构造思想是：假设响应序列 $\{y_t\}$ 和输入变量序列（即自变量序列）$\{x_{1t}\}, \{x_{2t}\}, \cdots, \{x_{kt}\}$ 均平稳，首先构建响应序列和输入变量序列的回归模型：

$$y_t = \mu + \sum_{i=1}^{k} \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it} + \varepsilon_t \tag{5}$$

其中，$B$ 为延迟因子，即 $B^n x_t = x_{t-n}$；$\Phi_i(B)$ 为第 $i$ 个输入变量的自回归系数多项式；$\Theta_i(B)$ 为第 $i$ 个输入变量的移动平均系数多项式；$l_i$ 为第 $i$ 个输入变量的延迟阶数；$\{\varepsilon_t\}$ 为回归残差序列。

因为 $\{y_t\}$ 和 $\{x_{1t}\}, \{x_{2t}\}, \cdots, \{x_{kt}\}$ 均平稳，而且平稳序列的线性组合仍然是平稳的，所以残差序列 $\{\varepsilon_t\}$ 为平稳序列，即：

$$\varepsilon_t = y_t - \left( \mu + \sum_{i=1}^{k} \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it} \right) \tag{6}$$

使用 ARMA 模型继续提供残差序列 $\{\varepsilon_t\}$ 中的相关信息，最终得到的模型为：

$$\begin{cases} y_t = \mu + \sum_{i=1}^{k} \dfrac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it} + \varepsilon_t \\ \varepsilon_t = \dfrac{\Theta_i(B)}{\Phi_i(B)} a_t \end{cases} \tag{7}$$

其中，$\Phi(B)$ 为残差序列自回归系数多项式；$\Theta(B)$ 为残差序列移动平均系数多项式；$a_t$ 为零均值白噪声序列。

## 2　参数估计

在选择了拟合模型后，就要利用时间序列的值

确定模型的口径,即估计模型中未知参数的值[15]。ARIMAX 模型可以通过许多不同的方法来估计,包括将模型转换为非线性最小二乘法、GLS 或极大似然估计。由于极大似然估计不需要从样本开始时丢弃观测值,或者需要从后期投射来创建观测值,因此比较适用于模型拟合。未知参数的极大似然估计(Maximum Likelihood Estimation, MLE)就是使得似然函数、即联合密度函数达到最大的参数值[16]。使用极大似然估计必须已知总体的分布函数,而在时间序列分析中,序列总体的分布通常是未知的[17-18]。为了便于分析和计算,通常假设序列服从多元正态分布[19]。

设 $K$ 维随机向量 $\boldsymbol{x} = [x_1, \cdots, x_k]^{-1}$ 的密度函数为:

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{K}{2}}} \cdot \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot e^{\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{8}$$

其中, $K$ 表示向量 $\boldsymbol{x}$ 的维度;均值向量 $\boldsymbol{\mu}$ 是 $K$ 维向量;协方差矩阵 $\boldsymbol{\Sigma}$ 是一个 $K \times K$ 的对称正定阵,则称 $\boldsymbol{x}$ 服从 $K$ 元正态分布,也称 $\boldsymbol{x}$ 为 $K$ 维正态随机向量,简记为: $\boldsymbol{x} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。其似然函数为:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{\frac{-KN}{2}} \cdot |\boldsymbol{\Sigma}|^{\frac{-N}{2}} \cdot e^{\frac{-1}{2}\sum_{n=1}^{N}(\boldsymbol{x}^n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^n-\boldsymbol{\mu})} \tag{9}$$

对数似然函数为:

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = C - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}^n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^n - \boldsymbol{\mu}) \tag{10}$$

其中, $C = \frac{-KN}{2}\ln(2\pi)$ 为一个常数。接着对 $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ 求偏导、整理,最终得到极大似然估计为:

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}^n - \bar{\boldsymbol{x}})(\boldsymbol{x}^n - \bar{\boldsymbol{x}})^{\mathrm{T}} \tag{11}$$

其中, $N$ 为样本个数。

## 3 评价指标

在前文基础上,还要对预测值的优劣进行评价,研究中用到的评价指标主要有:平均绝对百分误差、平均绝对误差、均方误差。这里将给出分析表述如下。

(1)平均绝对百分误差(Mean Absolute Percent Error, *MAPE*),又叫平均绝对离差,是所有单个观测值与算术平均值的偏差的绝对值的平均。平均绝对误差能够避免误差相互抵消的问题,因而可以准确反映实际预测误差的大小。具体数学公式为:

$$MAPE = \sum_{t=1}^{n}\left|\frac{x_i - y_i}{y_i}\right| \times \frac{100}{n} \tag{12}$$

(2)平均绝对误差(Mean Absolute Error, *MAE*),又叫平均绝对离差,是所有单个观测值与算术平均值的偏差的绝对值的平均。平均绝对误差能很好地反映预测值误差的实际情况。具体数学公式为:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{13}$$

(3)均方误差(Mean-Square Error, *MSE*)是参数估计值与参数真值之差平方的期望值。*MSE* 可以评价数据的变化程度。*MSE* 的值越小,预测模型描述实验数据则具有更好的精确度。具体数学公式为:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} \tag{14}$$

式(12)~(14)中, $y_i$ 为预测值, $x_i$ 为真实值。

(4)拟合优度。是指模型的预测值对实际值的拟合程度。度量拟合优度的统计量是可决系数(亦称确定系数) $R^2$。$R^2$ 最大值为 1。$R^2$ 的值越接近 1,说明回归直线对观测值的拟合程度越好;反之, $R^2$ 的值越小,说明回归直线对观测值的拟合程度越差。具体数学公式为:

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \tag{15}$$

其中, $y$ 为模型预测值; $\hat{y}$ 为流量观测值; $\bar{y}$ 为观测值的平均数。

## 4 模型构建

### 4.1 数据

本文采用的数据来自于美国加利福尼亚州交通局的公开数据集(Peformance Measurement System, PeMS),采用的是维克多维尔城市的某一条路从 2018 年 3 月 5 日至 4 月 13 日工作日期间每 5 min 为间隔的交通流数据,共 8 640 组数据,分析可得每天数据的基本统计特征见表 1,截取前一周(即 2018 年 3 月 5 日至 2018 年 3 月 9 日)的数据如图 1 所示。

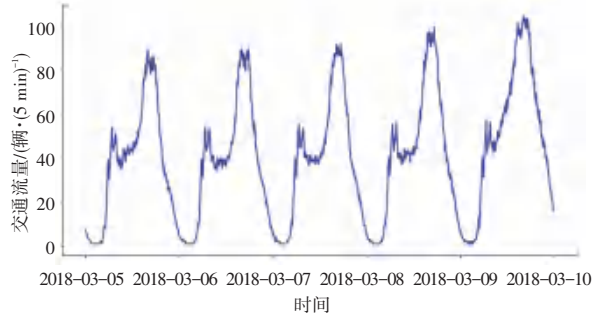
### 4.2 数据的平稳性检验

考虑到现存的虚假回归问题,在模型拟合前就要对各序列的平稳性进行检验。只有当每个序列都平稳时,才能使用 ARIMAX 模型拟合多元序列之间的动态回归关系。

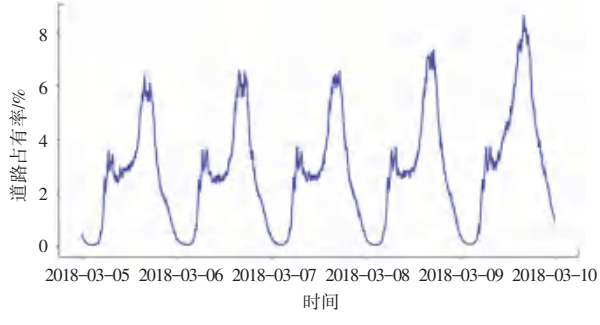
**表 1　交通流量、占有率数据的基本统计特征**

**Tab. 1　Statistical characteristics of traffic flow and occupancy**

| 日期 | 交通流量 | | | 占有率 | | |
|---|---|---|---|---|---|---|
| | 平均值 | 方差 | 标准差 | 平均值 | 方差 | 标准差 |
| 2018/03/05 | 37.010 42 | 666.996 42 | 25.826 27 | 0.024 683 0 | 0.000 314 9 | 0.017 744 7 |
| 2018/03/06 | 37.093 75 | 679.473 85 | 26.066 72 | 0.025 049 7 | 0.000 338 7 | 0.018 403 8 |
| 2018/03/07 | 38.305 56 | 708.837 19 | 26.624 00 | 0.025 581 3 | 0.000 341 7 | 0.018 484 6 |
| 2018/03/08 | 41.173 61 | 823.136 53 | 28.690 36 | 0.027 678 8 | 0.000 419 6 | 0.020 483 2 |
| 2018/03/09 | 49.482 64 | 1 010.499 70 | 31.788 36 | 0.034 541 3 | 0.000 590 6 | 0.024 301 3 |
| 2018/03/12 | 37.010 42 | 666.996 42 | 25.826 27 | 0.024 683 0 | 0.000 314 9 | 0.017 744 7 |
| 2018/03/13 | 37.093 75 | 679.473 85 | 26.066 72 | 0.025 049 7 | 0.000 338 7 | 0.018 403 8 |
| 2018/03/14 | 38.305 56 | 708.837 19 | 26.624 00 | 0.025 581 3 | 0.000 341 7 | 0.018 484 6 |
| 2018/03/15 | 41.173 61 | 823.136 53 | 28.690 36 | 0.027 678 8 | 0.000 419 6 | 0.020 483 2 |
| 2018/03/16 | 49.482 64 | 1 010.499 70 | 31.788 36 | 0.034 541 3 | 0.000 590 6 | 0.024 301 3 |
| 2018/03/19 | 37.010 42 | 666.996 42 | 25.826 27 | 0.024 683 0 | 0.000 314 9 | 0.017 744 7 |
| 2018/03/20 | 37.093 75 | 679.473 85 | 26.066 72 | 0.025 049 7 | 0.000 338 7 | 0.018 403 8 |
| 2018/03/21 | 38.305 56 | 708.837 19 | 26.624 00 | 0.025 581 3 | 0.000 341 7 | 0.018 484 6 |
| 2018/03/22 | 41.173 61 | 823.136 53 | 28.690 36 | 0.027 678 8 | 0.000 419 6 | 0.020 483 2 |
| 2018/03/23 | 49.482 64 | 1 010.499 70 | 31.788 36 | 0.034 541 3 | 0.000 590 6 | 0.024 301 3 |
| 2018/03/26 | 37.010 42 | 666.996 42 | 25.826 27 | 0.024 683 0 | 0.000 314 9 | 0.017 744 7 |
| 2018/03/27 | 37.093 75 | 679.473 85 | 26.066 72 | 0.025 049 7 | 0.000 338 7 | 0.018 403 8 |
| 2018/03/28 | 38.305 56 | 708.837 19 | 26.624 00 | 0.025 581 3 | 0.000 341 7 | 0.018 484 6 |
| 2018/03/29 | 41.173 61 | 823.136 53 | 28.690 36 | 0.027 678 8 | 0.000 419 6 | 0.020 483 2 |
| 2018/03/30 | 49.482 64 | 1 010.499 70 | 31.788 36 | 0.034 541 3 | 0.000 590 6 | 0.024 301 3 |
| 2018/04/02 | 37.010 42 | 666.996 42 | 25.826 27 | 0.024 683 0 | 0.000 314 9 | 0.017 744 7 |
| 2018/04/03 | 37.093 75 | 679.473 85 | 26.066 72 | 0.025 049 7 | 0.000 338 7 | 0.018 403 8 |
| 2018/04/04 | 38.305 56 | 708.837 19 | 26.624 00 | 0.025 581 3 | 0.000 341 7 | 0.018 484 6 |
| 2018/04/05 | 41.173 61 | 823.136 53 | 28.690 36 | 0.027 678 8 | 0.000 419 6 | 0.020 483 2 |
| 2018/04/06 | 49.482 64 | 1 010.499 70 | 31.788 36 | 0.034 541 3 | 0.000 590 6 | 0.024 301 3 |
| 2018/04/09 | 37.010 42 | 666.996 42 | 25.826 27 | 0.024 683 0 | 0.000 314 9 | 0.017 744 7 |
| 2018/04/10 | 37.093 75 | 679.473 85 | 26.066 72 | 0.025 049 7 | 0.000 338 7 | 0.018 403 8 |
| 2018/04/11 | 38.305 56 | 708.837 19 | 26.624 00 | 0.025 581 3 | 0.000 341 7 | 0.018 484 6 |
| 2018/04/12 | 41.173 61 | 823.136 53 | 28.690 36 | 0.027 678 8 | 0.000 419 6 | 0.020 483 2 |
| 2018/04/13 | 49.479 17 | 1 010.194 01 | 31.783 55 | 0.034 544 4 | 0.000 590 8 | 0.024 305 5 |
| 均值 | 40.613 08 | 799.711 05 | 28.279 16 | 0.027 506 9 | 0.000 414 5 | 0.020 355 9 |




(a) 2018 年 3 月 5 日到 3 月 9 日的交通流量




(b) 2018 年 3 月 5 日到 3 月 9 日的道路占有率

**图 1　一周的交通流量、占有率图**

**Fig. 1　Traffic flow and occupancy in a week**

观察图 1 能发现交通流量与占有率的呈周期性变化，为了直观展示其规律，绘制交通流量和道路占有率的自相关图如图 2 所示。从 2 个自相关图中，研究发现序列的自相关系数递减至零的速度相当缓慢，在很长的延迟时期里，自相关系数一直为正，而后又一直为负，显示出明显的三角对称性，这是一种具有单调趋势的非平稳序列。为了将序列达到平稳状态，考虑采用简洁、有效的差分方法。因此，研究中将原序列进行一阶差分，再对差分后的序列检验平稳性。为了检验序列的平稳性，陆续提出了许多

方法,其中应用最多的是单位根检验,而适用范围最广的则是 ADF 检验,即增广 DF 检验(Augmented Dickey-Fuller,ADF)检验。检验时,原假设为序列非平稳,通过构造 ADF 检验统计量:

$$\tau = \frac{\hat{\rho}}{S(\hat{\rho})} \tag{16}$$

其中,$S(\hat{\rho})$ 为参数 $\rho$ 的样本标准差。

通过蒙特卡洛方法,可以得到 $\tau$ 检验统计量的临界值表。当临界值小于 0.05 时,拒绝原假设,认为序列平稳。对一阶差分后的交通流量和道路占有率进行检验,检验结果参见表 2。观察 ADF 检验结果显示,经过一阶差分后的交通流量 $\{\tilde{\mathrm{N}}y_t\}$、占有率 $\{\tilde{\mathrm{N}}x_t\}$ 均达到平稳状态,因此可以用于构建 ARIMAX 模型。

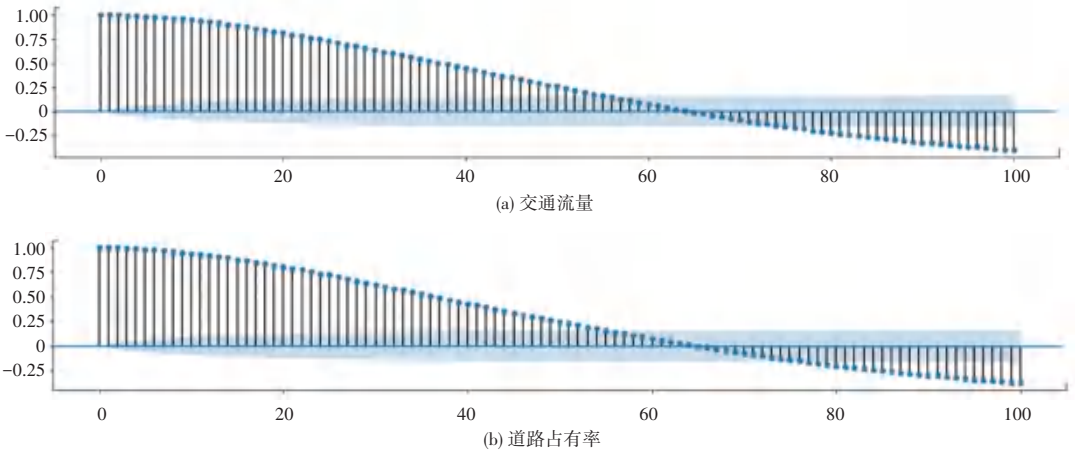



**图 2 交通流量、道路占有率原始数据的自相关图**

**Fig. 2 Autocorrelation of original data of traffic flow and occupancy**

**表 2 交通流量、道路占有率一阶差分后的单位根检验**

**Tab. 2 ADF test after first-order difference of traffic flow and occupancy**

| | Trend and Intercept | | Intercept | | None | |
|---|---|---|---|---|---|---|
| | *t - Statistic* | *Prob.* * | *t - Statistic* | *Prob.* * | *t - Statistic* | *Prob.* * |
| Volume | −12.739 48 | 0.000 0 | −12.738 75 | 0.000 0 | −12.739 47 | 0.000 0 |
| Occupany | −13.244 38 | 0.000 0 | −13.244 95 | 0.000 0 | −13.245 80 | 0.000 0 |
| Test critical values | −3.959 106 | 1% | −3.430 936 | 1% | −2.565 232 | 1% |
| | −3.410 327 | 5% | −2.861 684 | 5% | −1.940 861 | 5% |
| | −3.126 914 | 10% | −2.566 888 | 10% | −1.616 675 | 10% |

### 4.3 模型的建立

经过平稳性检验,一阶差分后的交通流量和车道占有率平稳,可以建立动态回归模型。首先,构建车辆流量 $\{\tilde{\mathrm{N}}y_t\}$ 与占有率 $\{\tilde{\mathrm{N}}x_t\}$ 的回归模型,由此推得数学公式为:

$$\tilde{\mathrm{N}}y_t = \mu + \frac{\Theta(B)}{\Phi(B)} B^k \tilde{\mathrm{N}}x_t + \varepsilon_t \tag{17}$$

接下来,要确定自回归系数 $p$ 与移动平均阶数 $q$ 的值。通过计算使模型的赤池信息准则(Akaike Information Criterion, *AIC*)和贝叶斯信息准则(Bayesian Information Criterion, *BIC*)达到最小值的 $p$、$q$ 值。为此,分别计算各种 $p$,$q$ 组合的 *AIC* 和 *BIC* 值,并绘制 *AIC*、*BIC* 的热力图,如图 3 所示。通过图 3 来寻找 *AIC*、*BIC* 值最小的 $p$ 与 $q$ 的组合为(6,5)。再将差分后的序列带入模型,用极大似然估计进行拟合得到参数值,详见表 3。至此,最终模型可写为如下形式:

$$\nabla y_t = 2.005\ 6 + 0.941\ 7\ \nabla x_{1t} + 0.172\ 3\ \nabla x_{2t} + 0.340\ 8\ \nabla x_{3t} + 0.568\ 6\ \nabla x_{4t} - 0.233\ 4\ \nabla x_{5t} - 0.269\ 9\ \nabla x_{6t} + \frac{1 + 0.187\ 4B + 0.439\ 2B^2 - 0.634\ 4B^3 - 0.156\ 8B^4 + 0.791\ 1B^5}{1 - 0.245\ 7B - 0.474B^2 + 0.578\ 7B^3 + 0.136\ 9B^4 - 0.854\ 7B^5 + 0.059\ 4B^6} a_t \tag{18}$$

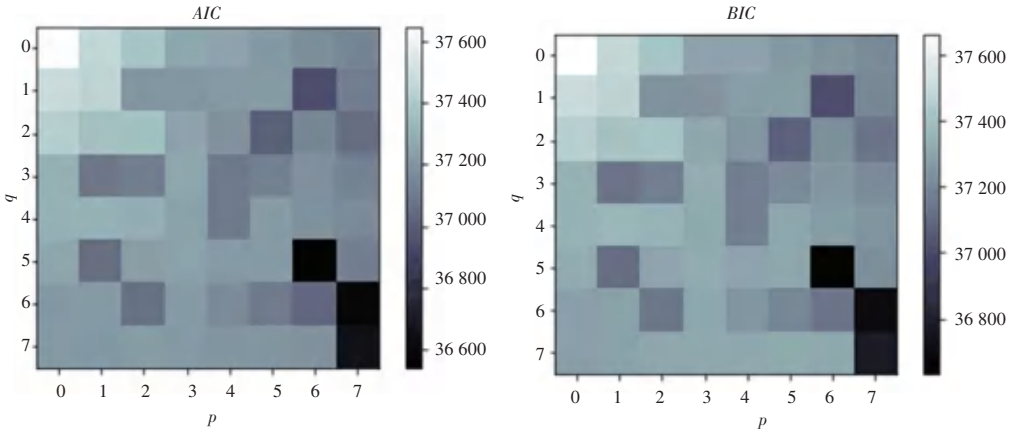



**图 3　*p*、*q* 各种组合的 *AIC*、*BIC* 热力图**

**Fig. 3　*AIC* and *BIC* thermodynamic diagram of various combinations of *p* and *q***

**表 3　*ARIMAX*(6,1,5)模型参数拟合**

**Tab. 3　Parameter fitting of *ARIMAX* (6,1,5)**

| Latent Variable | Estimate | Std Error | *z* | *P* >\| *z*\| | 95%C.I. |
|---|---|---|---|---|---|
| *AR*(1) | 0.245 7 | 0.034 1 | 7.206 3 | 0 | (0.178 9\|0.312 6) |
| *AR*(2) | 0.474 0 | 0.033 2 | 14.272 5 | 0 | (0.408 9\|0.539 1) |
| *AR*(3) | −0.578 7 | 0.009 9 | −58.477 6 | 0 | (−0.598 1\|−0.559 3) |
| *AR*(4) | −0.136 9 | 0.027 4 | −4.993 1 | 0 | (−0.190 6\|−0.083 2) |
| *AR*(5) | 0.854 7 | 0.026 8 | 31.847 9 | 0 | (0.802 1\|0.907 3) |
| *AR*(6) | 0.059 4 | 0.011 5 | 3.429 7 | 0.000 6 | (0.016 9\|0.061 9) |
| *MA*(1) | −0.187 4 | 0.032 5 | −5.758 3 | 0 | (−0.251 2\|−0.123 6) |
| *MA*(2) | −0.439 2 | 0.029 3 | −15.002 | 0 | (−0.496 5\|−0.381 8) |
| *MA*(3) | 0.634 4 | 0.002 3 | 281.049 4 | 0 | (0.629 9\|0.638 8) |
| *MA*(4) | 0.156 8 | 0.029 4 | 5.334 5 | 0 | (0.099 2\|0.214 5) |
| *MA*(5) | −0.791 1 | 0.026 1 | −30.279 9 | 0 | (−0.842 3\|−0.739 9) |
| Beta avg_occup | −1.279 9 | 0.240 7 | −5.318 2 | 0 | (−1.751 5\|−0.808 2) |
| Normal Scale | 2.005 6 | | | | |

考虑到差分的方法对确定性信息的提取可能不充分，因此还要进一步地对残差序列进行检验。如果检验结果显示为残差序列的自相关性不显著，就说明 ARIMAX 模型对信息的提取比较充分。在此基础上，就是对模型的残差序列进行检验，判断是否存在残存有效信息。为此，对其进行 ADF 单位根检验和 Durbin-Watson 检验(D-W 检验)，结果见表 4 以及绘制残差的 Q-Q 图，见图 4。

**表 4　ARIMAX 模型残差检验**

**Tab. 4　Residual test of ARIMAX**

| | D-W 值 | ADF 检验 (*P* 值) |
|---|---|---|
| ARIMAX 模型 | 2.000 37 | 2.022 6e-68 |

从表 4 中可以发现 D-W 值趋近于 2，即接受原假设：残差序列不存在 1 阶自相关性；单位根检验结果 *P* 值远小于 0.05 说明残差显著平稳。从图 4 可以看出，散点基本落在直线两端，故残差满足均值为 0 的正态分布。满足以上条件后，就可用 ARIMAX 模型对此后一周的交通流进行拟合预测。其中，这一周的预测流量与真实流量如图 5 所示。

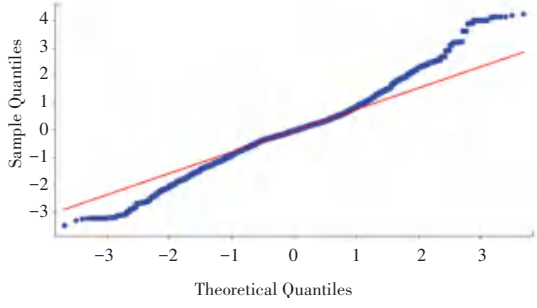



**图 4　ARIMAX 模型的残差 Q-Q 图**

**Fig. 4　Residual Q-Q diagram of ARIMAX model**

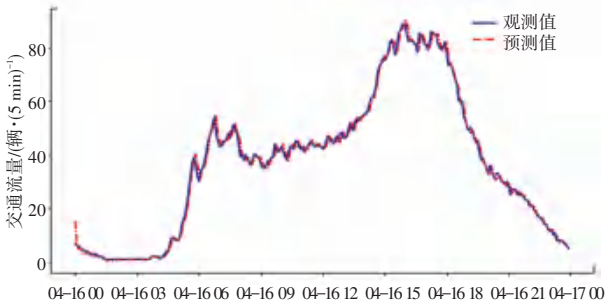
然后用平均绝对误差、均方误差、平均绝对百分比误差来衡量交通流量实际值与 ARIMAX 模型的预测值(见表 5)，并计算模型的拟合优度为 0.876 95。

通过模型可以发现，平均绝对误差、均方误差分别为 1.47 和 3.74，效果较好，并且一般认为 *MAPE* 的值低于 10% 时预测精度较高，本文中 *MAPE* 仅为 6.87，说明 ARIMAX 模型预测效果较好。

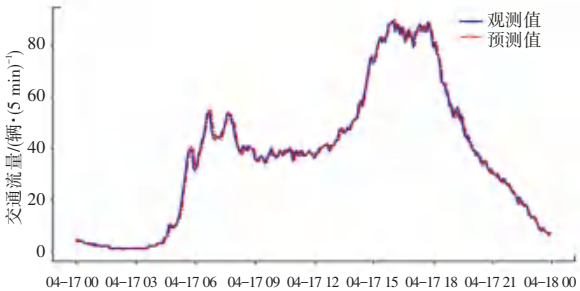
表 5 预测与实际值的 *MAE*,*MSE*,*MAPE*

Tab. 5 *MAE*,*MSE* and *MAPE* of predicted and actual data

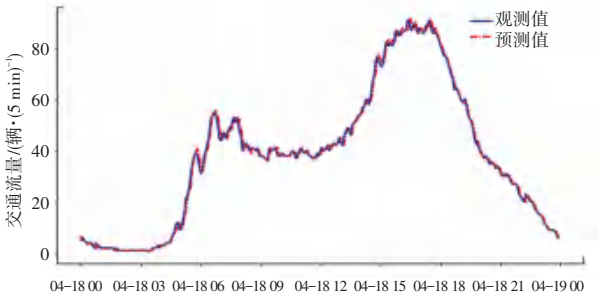
| 日期 | *MAE* | *MSE* | *MAPE* |
|---|---|---|---|
| 2018/04/16 | 1.367 834 | 3.649 851 | 8.086 658 |
| 2018/04/17 | 1.373 786 | 3.577 881 | 6.121 108 |
| 2018/04/18 | 1.295 282 | 3.215 208 | 6.870 525 |
| 2018/04/19 | 1.418 723 | 3.883 052 | 5.364 511 |
| 2018/04/20 | 1.613 909 | 4.410 740 | 7.920 212 |
| 平均值 | 1.413 907 | 3.747 346 | 6.872 603 |



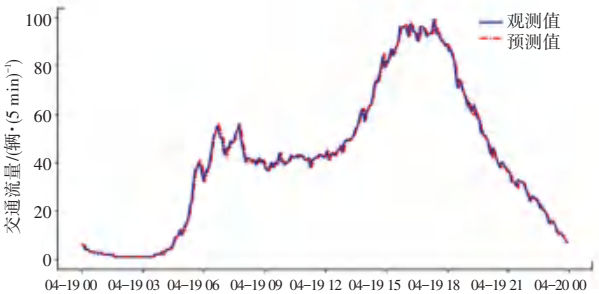

(a) 2018 年 4 月 16 日交通流量预测值与实际值



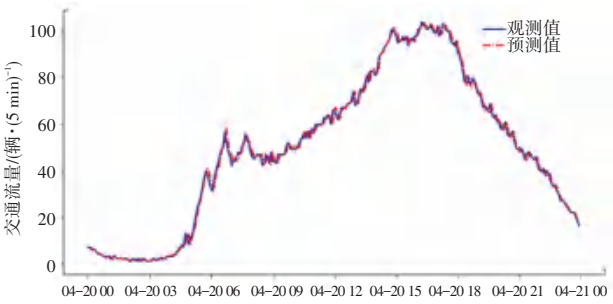

(b) 2018 年 4 月 17 日交通流量预测值与实际值




(c) 2018 年 4 月 18 日交通流量预测值与实际值




(d) 2018 年 4 月 19 日交通流量预测值与实际值




(e) 2018 年 4 月 20 日交通流量预测值与实际值

**图 5 预测流量与真实流量值**

**Fig. 5 Comparison between predicted and actual data**

## 5 结束语

构建 ARIMAX 模型的过程与传统的 ARIMA 模型类似,但与 ARIMA 模型相比,丰富了数据信息,从而提高了预测的精度。将交通流量、道路占有率作为输入序列,先要确保其序列的稳定性,为此采用差分来提取确定性信息。经过一阶差分后,通过单位根检验,序列达到了平稳形态。接下来,就是构建 ARIMAX 模型,以及确定 ARIMAX 模型的阶数。在模型定阶过程中,相比于直接观察绘制的自相关、偏自相关图确定 $p$、$q$ 值的办法,本文通过计算所有 $p$ 与 $q$ 组合的 *AIC* 和 *BIC*,寻找使得 *AIC* 和 *BIC* 最小的那一组数值。如此一来,既提高了精确度,又节省了调参的时间,预测效率明显提高。在模型阶数确定后,利用极大似然估计的方法来拟合参数,得到了一个 ARIMAX 模型。虽然拟合求出了参数模型,但是并不能保证差分的方法能够充分提取确定性信息,因此还要进一步来检验残差。经过 D-W 等方式检验、并发现残差不存在自相关性后,就可以用得到的 ARIMAX 模型进行交通流量预测。为了防止偶然事件的产生,研究中预测了接下来一周的交通流量,并运用多种评价指标进行验算。最终结果显示,采用道路占有率作为外生变量的交通流量 ARIMAX 模型能够很好地拟合流量序列的变化规律,也有着良好的预测精度。而且作为统计类的模型其未知参数对比于神经网络要少得多,具有更快的预测速度,既满足了交通流预测的实效性,也得到了很高的预测精度。

本次研究中,虽然利用道路占有率作为外生变量加入到了交通流量的预测中来减少序列预处理时差分所减少的有效信息量,但是并没有研究道路占有率的加入对预测精度具体提高了多少的百分比,以及道路占有率的加入能否弥补因差分所带来的有

限信息量的丢失，这些都是未来课题的有效考察重点。而且作为交通流参数，还有如速度、车头时距等，若将其也加入交通流量的预测模型中，能否提高预测的精度以及弥补因差分丢失的信息内容，也是下一步需要深入探讨的研究方向。

## 参考文献


[1] 王进，史其信. 短时交通流预测模型综述[J]. 中国公共安全(学术卷)，2005 (1)：92-98.

[2] BOX G E P，JENKINS G M，REINSEL G C. Timeseries analysis：forecasting and control rev. ed[J]. Journal of Time，1976，31(4)：238-242.

[3] 李凯，曹阳. 基于 ARIMA 模型的网络安全威胁态势预测方法[J]. 计算机应用研究，2012，29(8)：3042-3045.

[4] 谭满春，李英俊，徐建闽. 基于小波消噪的 ARIMA 与 SVM 组合交通流预测[J]. 公路交通科技，2009，26(7)：127-132，138.

[5] 谈苗苗，成孝刚，周凯，等. 基于 ARIMA 和灰色模型加权组合的短期交通流预测[J]. 计算机技术与发展，2016，26(11)：77-81，85.

[6] 田瑞杰，张维石，翟华伟. 基于时间序列与 BP-ANN 的短时交通流速度预测模型研究[J]. 计算机应用研究，2019 (11)：1-8.

[7] 韩超，宋苏，王成红. 基于 ARIMA 模型的短时交通流实时自适应预测[J]. 系统仿真学报，2004，16(7)：1530-1532.

[8] 王晓全，邵春福，尹超英，等. 基于 ARIMA-GARCH-M 模型的短时交通流预测方法[J]. 北京交通大学学报，2018，42(4)：79-84.

[9] 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.

[10] 罗向龙，焦琴琴，牛力瑶，等. 基于深度学习的短时交通流预测[J]. 计算机应用研究，2017，34(1)：91-93，97.

[11] 宗春光，宋靖雁，任江涛，等. 基于相空间重构的短时交通流预测研究[J]. 公路交通科技，2003 (4)：71-75.

[12] 丁永兵，胡尧，沈齐，等. 基于多元时间序列的交通流预测模型[J]. 贵州大学学报(自然科学版)，2017，34(1)：123-127.

[13] 崔和瑞，彭旭. 基于 ARIMAX 模型的夏季短期电力负荷预测[J]. 电力系统保护与控制，2015，43(4)：108-114.

[14] 程燕. ARIMAX 模型方法及其应用—重庆城市居民可支配收入与消费支出[J]. 重庆工商大学学报(自然科学版)，2015，32(11)：80-85.

[15] POLLOCK D S G. Chapter 22-Maximum-likelihood methods of ARMA estimation[M]// Handbook of time series analysis，signal processing，and dynamics. Farnborough，UK：Academic Press，1999：667-695.

[16] 郭艳鹏. 多维 ARMA 模型的谱估计及预测方法[D]. 成都：西南交通大学，2008.

[17] MCLEOD A I，ZHANG Y. Faster ARMA maximum likelihood estimation[J]. Computational Statistics & Data Analysis，2008，52(4)：2166-2176.

[18] 刘次华. ARMA 模型变化点的极大似然估计[J]. 应用数学，1992 (3)：111-113.

[19] OLIVEIRA P J，STEFFEN J L，CHEUNG P. Parameter estimation of seasonal Arima models for water demand forecasting using the harmony search algorithm[J]. Procedia Engineering，2017，186：177-185.


(上接第 11 页)


[12] 许伟，熊卫华，姚杰，等. 基于改进 YOLOv3 算法在垃圾检测上的应用[J]. 光电子·激光，2020，31(9)：928-938.

[13] 魏铖磊，南新元，李成荣，等. 一种多尺度感受视野注意力机制的生活垃圾单阶段目标检测方法[J/OL]. 环境工程：1-12[2021-06-22]. http://kns.cnki.net/kcms/detail/11.2097.X.20210622.1048.010.html.

[14] ZHAO L，PAN Y，WANG S，et al. Skip-YOLO：Domestic Garbage Detection Using Deep Learning Method in Complex Multi-scenes[EB/OL]. [2021-10-28]. https://www.researchsquare.com/article/rs-757539/v1.

[15] DING X，ZHANG X，MA N，et al. RepVGG：Making VGG-style ConvNets great again[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Montreal：IEEE Computer Society，2021：13733-13742.

[16] TAN Mingxing，PANG Ruoming，LE Q V. Efficientdet：Scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle，WA，USA：IEEE，2020：10778-10787.

[17] LIN T Y，DOLLáR P，GIRSHICK R，et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition. Honolulu，HI，USA：IEEE Computer Society，2017：936-944.

[18] YANG Maoke，YU Kun，ZHANG Chi，et al. Denseaspp for semantic segmentation in street scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT，USA：IEEE Computer Society，2018：3684-3692.

[19] HE Kaiming，ZHANG Xiangyu，REN Shaoqing，et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(9)：1904-1916.

[20] YU F，KOLTUN V，FUNKHOUSER T. Dilated residual networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu，HI，USA：IEEE Computer Society，2017：636-644.

[21] LIU S，QI L，QIN H，et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C，USA：IEEE Computer Society，2018：8759-8768.