基于时间序列 ARMA 模型的车内污染物浓度预测

陈立新 王志强
天津商业大学 机械工程学院，中国·天津 300134

摘 要：本研究旨在利用时间序列 ARMA 模型对车内污染物浓度进行预测，为车内空气质量的监测和改善提供科学依据。为了预测车内污染物浓度，实验选取车内的甲苯和甲醛作为本次研究的主要目标，通过实验测得车内污染物浓度变化并作为原始数据集，随后建立时间序列 ARMA 对未来 6 小时车内污染物浓度进行预测。结果表明，时间序列 ARMA 模型在车内污染物浓度预测具有较高的精度和可靠性，为车内污染物浓度的控制和监测提供了有效的手段。该研究成果说明时间序列分析方法可大大地改善参数的估计精度，提高模型的预测效果。

关键词：车内污染物；时间序列；ARMA 模型；浓度预测

Prediction of Pollutant Concentration in Vehicles based on Time-series ARMA Model

Lixin Chen Zhiqiang Wang
School of Mechanical Engineering, Tianjin University of Commerce, Tianjin, 300134, China

Abstract: The purpose of this study is to use the time series ARMA model to predict the concentration of pollutants in vehicles, and to provide a scientific basis for the monitoring and improvement of air quality in vehicles. In order to predict the concentration of pollutants in the vehicle, toluene and formaldehyde in the vehicle were selected as the main objectives of this study, and the changes in the concentration of pollutants in the vehicle were measured through experiments and used as the original data set, and then the time series ARMA was established to predict the concentration of pollutants in the vehicle in the next 6 hours. The results show that the time series ARMA model has high accuracy and reliability in predicting the concentration of pollutants in vehicles, and provides an effective means for the control and monitoring of pollutants in vehicles. The research results show that the time series analysis method can greatly improve the estimation accuracy of parameters and improve the prediction effect of the model.

Keywords: in-vehicle pollutants; time series; ARMA model; concentration prediction

0 前言

自 2009 年以来，中国连续 5 年被誉为世界第一的汽车生产大国和消费大国。汽车已成为继家庭和办公室之后，人们的第三个工作场所。数据显示，2022 年中国的汽车保有量已达 4.15 亿辆。2018 年，中国民用与私人汽车分别达到 2.3 亿和 2.1 亿辆，汽车驾驶员数量达 3.4 亿人 [1]。由此可见，中国具有汽车保有量大和驾驶人员多的特点。当汽车发动机启动时，燃料燃烧产生的 CO、SO_2、NOx 以及细小颗粒物等污染物会通过密封不良的车窗、车身拼接处和通风口进入车内，导致车内空气污染。尤其是在交通堵塞的情况下，这种污染会更加严重。早在 2013 年，国家质量检测总局就已经公布，除了发动机、变速器、离合器和气囊等质量问题外，车内异味已成为车主投诉的主要问题之一。

时间序列分析是一种重要的数学方法，用于处理相互关联的动态数据集合，它可以对随时间变化的数据进行分析和预测 [2]。时间序列数据具有独特的特点，即它们随时间的推移而变化，并能够反映过去所有的因果关系 [3]。在空气质量研究 [4]、股票价格预测 [5]、建筑能耗预测 [6] 和电力系统负荷预测 [7] 等多个领域都有应用。与纯 AR 模型或纯 MA 模型相比，ARMA 模型在捕捉序列特征时需要的参数较少，从而减少了过拟合的风险。因此，ARMA 模型在短期预测中具有较好的效果，它能够有效捕捉时间序列的短期波动特征。

尽管车内空气质量问题日益突出，然而现有的研究主要集中在室外空气污染的监测和控制上，对于车内污染物浓度的预测研究相对较少。传统的研究方法往往依赖于定点监测和定时采样，无法实时反映车内污染物浓度变化的动态特征。因此，亟须一种能够实时、准确预测车内污染物浓度的模型，以便及时采取相应的控制措施，保障乘客的健康和安全。基于此，论文利用时间序列 ARMA 模型对未来 6 小时车内污染物浓度进行预测分析，以促进该方法在车内污染物浓度监测中的应用。

1 数据收集

1.1 车内污染物采样准备

本次实验的所有样本均于 2023 年冬季在中国天津采集。天津是中国四大直辖市之一，地处华北地区，海河下游，毗邻渤海。四季分明，年平均气温 13.1℃，平均降雨量

500~700 毫米[8]。天津工业园区因快速城市化而受到污染，长期以来，颗粒污染物一直是天津的主要环境空气污染物。重悬浮粉尘[9]、燃煤和汽车尾气是 PM 污染的主要来源。实验用车为市场上某 2023 款主流车型，使用年限为 5 年。为了避免影响市场，省去了车型和品牌的具体信息。

1.2 车内污染物测定

利用 ppb RAE 3000 VOCs 检测仪对车内挥发性有机物进行测定，本次实验采用随机采样法对车内污染物进行独立采样，在早上、中午和下午用车高峰时段，观察车内污染物的浓度随时间变化的情况，汽车准备完成之后，将采样口置于驾驶位头部呼吸口处，将车窗和车门关闭保证实验与实际情况相符。

1.3 车内污染物浓度

通过现场对车内甲醛浓度和甲苯浓度进行采样，并通过对现场检测的结果分析，甲醛和甲苯的检出率 100%，意味着车内污染的主要来源为甲苯和甲醛。车内甲苯浓度中位数（0.2445mg/m^3）远大于车内甲醛的浓度中位数（0.0645mg/m^3），因此需要控制车内甲苯的浓度。

2 时间序列 ARMA 模型的建立

2.1 ARMA 模型构建

自回归移动平均（ARMA）由两部分组成，一个是自回归模型 AR，另一个是移动平均模型 MA。自回归模型 AR（p）是一种适用于短期预测的模型，它的作用是建立未来数据点和过去数据点的一个关系，表达式如公式(1)所示。

$$x_t = \emptyset_1 x_{t-1} + ... + \emptyset_p x_{t-p} + \varepsilon_t \quad (1)$$

式（1）中 x_t 为时间序列中 x 在 t 时刻的值；$\emptyset$为自回归系数；p 为阶数；ε_t 为独立同分布的随机变量。自回归型是一种未来数值与过去数值的多元线性回归模型，当独立同分布的随机变量满足正态分布的时候，则称时间序列 x 服从 p 阶自回归模型。移动平均模型 MA（q）是建立一个误差关系的线性回归模型，当 t 时刻的误差与之前产生的误差满足公式（2）的关系，那么就称时间序列 x 服从 q 阶移动平均模型。

$$x_t = \theta_1 x_{t-1} + ... + \theta_q x_{t-q} + \varepsilon_t \quad (2)$$

在公式(2)中，θ 表示移动平均数；ε_t 表示 t 时刻的误差；q 为模型阶数。对比 AR 模型和 MA 模型，可以发现 MA 模型对误差项进行自回归。AR 模型通过正序建立线性回归关系，在建立模型时会产生一些误差；MA 模型则描述当前值与自回归部分误差累积的关系。在这种情况下，MA 模型既能对未来进行预测，又能对预测产生的误差进行回归拟合。因此，将两种模型结合起来形成 ARMA 模型。根据公式（1）和公式（2），如果时间序列 x 满足公式（3），则称时间序列 x 服从（p，q）阶自回归移动平均模型 ARMA（p，q）。

$$\theta_1 x_{t-1} + ... + \theta_q x_{t-q} + \varepsilon_t = \emptyset_1 x_{t-1} + ... + \emptyset_p x_{t-p} + \varepsilon_t \quad (3)$$

2.2 ARMA 模型参数设计

要拟合未来数据与历史数据的关系，利用式中的方式滚动构造模型，需要优化的只有 $\emptyset$ 的取值。MA（q）模型的参数估计需要利用公式（4）实际时间序列数据，求得自协方差函数 r_k 的估计值 $\hat{r}_k$ 并代入公式（4）。估计步骤为：①先利用自相关函数 p_k 的估计值 $\hat{\rho}_k$，代入 Yule-Walker 方程组，求得模型参数的估计值 $\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_p$ 。②改写 ARMA 模型，求解估计值 $\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_p$。其中，Yule-Walker 方程如公式（5）所示，ARMA（p，q）模型预测公式如公式（6）所示，预测方差如公式（7）所示。

$$r_k = \begin{cases} \sigma_\varepsilon^2\left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right) & k = 0 \\ \sigma_\varepsilon^2\left(-\emptyset_k + \emptyset_{k+1}\emptyset_1 + \emptyset_{k+2}\emptyset_2 + ... + \emptyset_{q-k}\emptyset_q\right) & 1 \ll k \ll q \\ 0 & k > q \end{cases} \quad (4)$$

$$\begin{cases} \rho_1 = \emptyset_1 + \emptyset_2 p_1 + ... + \emptyset_p \rho_{p-1} \\ \rho_2 = \emptyset_1 \rho_1 + \emptyset_2 + ... + \emptyset_p \rho_{p-2} \\ \cdots \\ \rho_p = \emptyset_1 \rho_{p-1} + \emptyset_2 \rho_{p-2} + ... + \emptyset_p \end{cases} \quad (5)$$

$$\hat{x}_k(l) = \begin{cases} \hat{x}_k(l) & l \gg 1 \\ x_{k+1} & l \ll 0 \end{cases} \quad (6)$$

$$Var[e_k(l)] = (1 + {G_1}^2 + ... + {G_{t-1}}^2)\sigma_\varepsilon^2 \quad (7)$$

3 ARMA 模型预测分析

3.1 数据预处理

本次数据集为车内甲苯的时间序列浓度共 236 个，其中前 186 个为训练集数据也就是构造模型的历史数据，后 50 个为测试集数据。

首先，利用 adftest 函数对时间序列进行平稳性检验。检验结果会生成一个标志位，其中 1 表示序列平稳，0 表示序列不平稳。如果标志位显示序列不平稳，则需要进行差分处理。具体步骤是对时间序列进行一阶差分，即将时刻 t 的数据减去时刻（t–1）的数据，构造出新的差分序列。如果一阶差分后的序列仍不平稳，则继续对其进行二阶差分处理，即对一阶差分序列再进行一次差分操作。通过这种方式，可以逐步得到平稳的时间序列。未进行处理之前的序列取值是 [0.057,0.243]，经过差分处理之后的取值范围是 [–0.005,0.016]，判定原时间序列的数据为不平稳，因此使用差分后的数据代替原数据。

3.2 模型定阶

数据平衡性处理后，进行判断适合时间序列的模型并将模型定阶。模型的阶数可以通过自相关图和偏相关图判断模型阶次。自相关是序列经过与自身阶数的滞后形成序列之间的某种相关性，偏相关是在其他序列给定的情况下两个序列相关性的函数。其中拖尾的含义是函数逐渐衰减为 0，根据相关性特征，可利用自相关函数和偏相关函数的截尾性来识别模型，如果自相关函数是拖尾的，那么就可以使用 AR（p）模型；如果偏相关函数是拖尾的，那么就可以使用 MA（q）模型。输入数据后，我们可以得到自相关图和偏相关图，如图 1、图 2 所示。

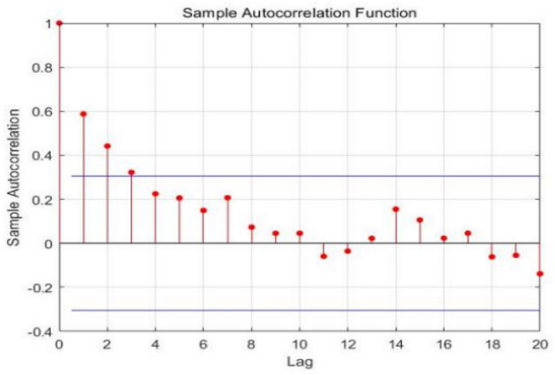

图 1 样本自相关性函数

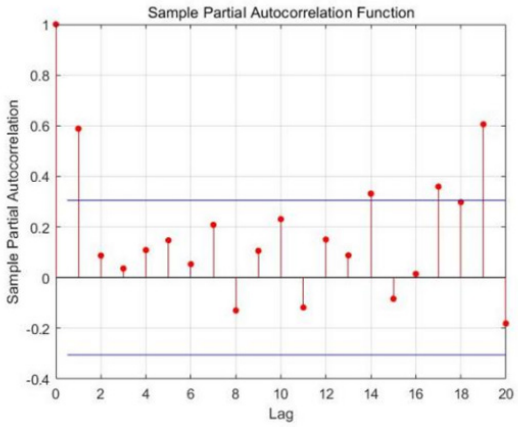

图 2 样本偏相关性函数

图 1 中横坐标为间隔，纵坐标样本自相关性；图 2 中横坐标为间隔，纵坐标为偏相关系数。从自相关性图可以看出，在四阶之后函数有一种靠近 0 的趋势，通常认为它是一个拖尾的状态。但是从图 2 中的偏相关图不太容易发现其中的规律，所以在单一模型不好判断的时候就会选取 ARMA 模型。

当自相关图和偏相关图难以判断时，可以用 AIC 准则求出最好阶数。赤池信息量准则（AIC，Akaike information criterion）是评估统计模型的复杂度和衡量统计模型“拟合”资料之优良性（Goodness of fit）的一种标准，可以将 p 和 q 的取值进行量化评估。增加自由参数的数目提高了拟合的优良性，AIC 鼓励数据拟合的优良性但尽量避免出现过度拟合（Overfitting）的情况，所以优先考虑的模型应是 AIC 值最小的那一个，但是阶数如果定的过长模型就会变得比较复杂且冗余，因此通常研究过程中阶数限定最大不超过总长度的 1/10。AIC 准则一般情况下的表达式为公式（8）。

$$ALC=2k-ln2ln(L) \quad (8)$$

式中，k 为自由系数的数量；L 为似然函数。一般而言，似然函数 L 会随着模型复杂度提高（k 增大）而增大，如果当 AIC 变小，但是 k 过大时，似然函数增速减缓，导致 AIC 增大，模型过于复杂容易造成过拟合现象。为了 AIC 要提高模型拟合度（极大似然），在设置中引入了惩罚项，使模型参数尽可能少，有助于降低过拟合的出现可能性。本次研究是将 p 和 q 的阶数从 1 开始循环，一直到 AIC 求出最小值是跳出循环，此为一种定性分析方法。得到的 AIC 定阶热力图如图 3 所示。

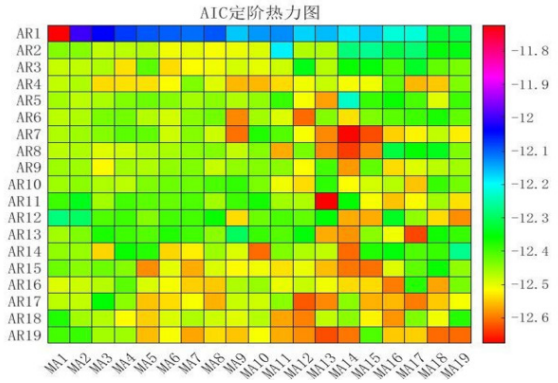

图 3 AIC 定阶热力图

根据热力图可知，AIC 在 AR11 和 MA13 的位置取最小值为 -12.68，所以 ARMA 的 p 和 q 的取值分别是 11、13。

3.3 时间序列预测结果及分析

实验使用平均绝对百分比误差（MAPE）和相关系数 R^2 来评估时间序列模型的精度，作为两种不同的评价指标，MAPE 的值越小模型精度越高，预测效果更好。而相关系数越接近 1 表明拟合效果更好。结果显示，ARMA 模型的 MAPE 为 0.0312，对估计数据的拟合为 96.12%，预测结果如图 4 所示。图 4 下半部分是后 50 个数据与 ARMA 预测数据的比较，能够看出真实值和预测值之间的差距比较小，通过结果数据来看，模型的 FPE 为 3.141×10^{-6}，MSE 为 2.098×10^{-6}，拟合误差在接受范围之内。

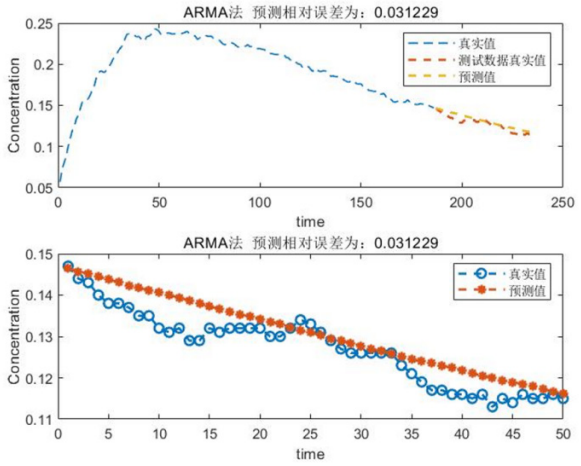

图 4 ARMA 时间序列预测

3.4 ARMA 模型的应用

为了验证模型的实用性，我们将利用之前建立的 ARMA 时间序列模型预测未来的车内污染物浓度值。首先

确定需要预测的数量，该数量与之前划分的测试集长度一致。接着，将未经过差分处理的训练集作为输入存储起来，并将数据输入到训练好的模型中。通过函数工具箱对数据进行差分处理，记录所有的差分值。当数据呈现非平稳状态时，通过累加差分值进行还原处理。本次实验的数据集为 2024 年 4 月 20 日上午 8 点到下午 2 点，每分钟记录一个数据，总共六个小时所采集到的 360 组车内甲苯浓度时序数据。最终，我们从 360 个数据中预测了后 40 个数据，预测结果如图 5 和图 6 所示。

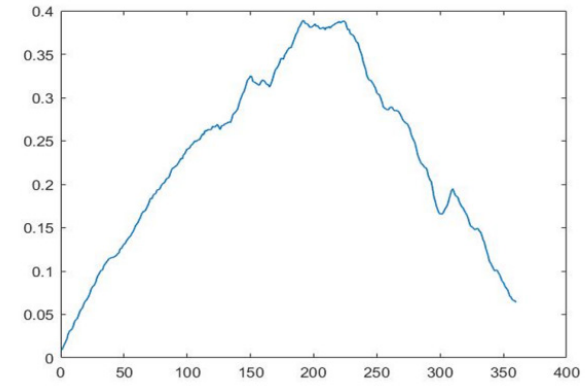

图 5 原始数据

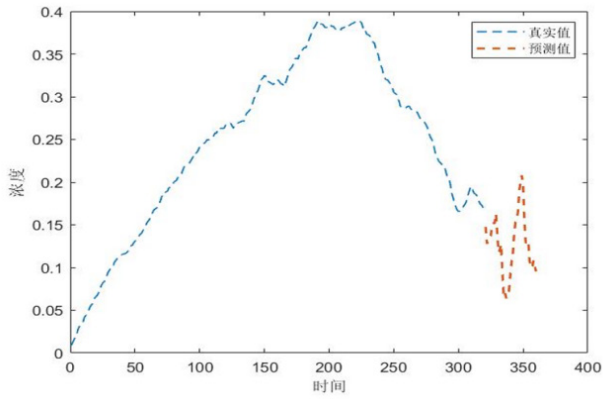

图 6 预测结果

根据图 6 可以看到预测结果虽然会出现一些波动，但其数值上的差距不是特别大，预测结果能够基本预测出浓度的变化趋势。根据量化结果模型的均方根误差为 3.841×10^{-6}，训练数据的预测目标准确率达到 98.11%，模型的预测数据和真实数据之间的误差为 0.068，所以本次实验模型的建立具有实用价值。

4 结语

①通过车内污染物浓度测试实验，得出车内污染的主要来源为甲苯和甲醛。并测得车内甲苯浓度中位数（0.2445mg/m^3）远大于车内甲醛的浓度中位数（0.0645mg/m^3），因此需要控制车内甲苯的浓度。

②对测得的数据进行差分处理后，如果判定序列依然不平稳，则使用处理后的数据集建立模型。首先，选择时间序列模型为 ARMA 模型；然后，根据 AIC 标准，通过定阶热力图确定模型的阶数，得到自回归阶数（p）为 11，移动平均阶数（q）为 13。接下来，设置模型为 A(z)y(t)=C(z)e(t)，并输入训练数据进行计算。结果显示，ARMA 模型的平均绝对百分比误差（MAPE）为 0.0312，模型对估计数据的拟合度为 96.12%。

③通过应用 MRMA 模型进行预测车内污染物浓度，计算得到均方根误差（RMSE）为 3.841×10^{-6}，训练数据的预测目标准确率达到 98.11%，模型的预测数据和真实数据之间的误差为 0.068。结果表明，本次实验中所建立的 ARMA 模型具有较好的预测效果。

参考文献：

[1] 李慧,李世雄,张学敏,等.车内可吸入颗粒物的研究现状与发展趋势[J].环境工程,2015,33(S1):438-442.

[2] Moon J, Hossain MB, Chon K H.AR和ARMA模型阶数选择用于使用ImageNet分类进行时间序列建模[J].信号处理,2021(183):108026.

[3] 许有俊,秦浩斌,李文博,等.浅埋暗挖隧道近距离平行上跨对既有盾构隧道的变形影响分析[J].现代隧道技术,2022,59(3):118-127.

[4] 艾洪福,石莹.基于BP人工神经网络的雾霾天气预测研究[J].计算机仿真,2015,32(1):402-405+415.

[5] 胡聿文.基于优化LSTM模型的股票预测[J].计算机科学,2021,48(S1):151-157.

[6] 曾国治,魏子清,岳宝,等.基于CNN-RNN组合模型的办公建筑能耗预测[J].上海交通大学学报,2022,56(9):1256-1261.

[7] 陈振宇,刘金波,李晨,等.基于LSTM与XGBoost组合模型的超短期电力负荷预测[J].电网技术,2020,44(2):614-620.

[8] 于佳卉,苗芮,孙玫玲.气候变化背景下天津地区采暖期变化特征分析[J].内蒙古气象,2024(4):51-56.

[9] 陈颖,王罗轩,周文迪.公共安全事件多发背景下对公交车司机工作现状的调查与应对思考——以上海市为例[J].时代汽车,2019(17):22-24.

作者简介：陈立新（2000-），女，中国河北承德人，在读硕士，从事空气品质研究。