基于时间序列的高空管制区航班流量预测

倪 超[1],叶博嘉[1],赵 云[1,2],田 勇[1]

(1. 南京航空航天大学 民航学院,江苏 南京 211106; 2. 民用航空中南地区空中交通管理局,广东 广州 510403)

摘 要: 为了采用较少的数据,较为准确地预测高空管制区航班流量,为管制单位制定工作计划提供参考,采用了时间序列方法,对航班流量进行预测。选取了三种时间序列预测方法,分别是基于统计分析的ARIMA模型、基于机器学习的随机森林和人工神经网络方法,并对其原理和特性进行分析。对实际运行数据的平稳性、自相关性和偏相关性等特性进行分析,并进行归一化处理。采用三种时间序列预测方法对高空管制区航班流量时间序列进行拟合和预测,结果表明,时间序列方法在进行航班流量预测时具有较好的效果,其中基于随机森林的时间序列预测方法准确度最高,预测准确率达到94.8%。

关键词: 高空管制区; 流量预测; 时间序列; 机器学习

中图分类号: V355 **文献标识码:** A **文章编号:** 1671-654X(2020)05-0043-05

Flight Flow Prediction in High Altitude Air Control Area Based on Time Series

NI Chao[1], YE Bo-jia[1], ZHAO Yun[1,2], TIAN Yong[1]

(1. *College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106, *China*;
2. *Central and Southern Regional Air Traffic Management Bureau, CAAC, Guangzhou* 510403, *China*)

Abstract: In order to use less data to predict the flight flow in the high-altitude air traffic control area accurately and provide reference for the control unit to make work plans, this study adopts the time series method to predict the flight flow. Firstly, three time series prediction methods are selected, namely, ARIMA model based on statistical analysis, random forest method and artificial neural network method based on machine learning, and their theories and characteristics are analyzed. Then the stability, auto correlation and partial correlation of the actual operation data are analyzed and normalized. Finally, three time series methods were used to fit and predict the flight flow time series in the high-altitude area. The results showed that the time series method had a good effect in the flow prediction, among which the time series prediction method based on random forest had the highest accuracy, with the prediction accuracy reaching 0.948.

Key words: high altitude control area; flight flow forecast; time series; machine learning

引言

随着航空运输业的不断高速发展,空中交通流量不断增加,空域管制负荷也由此不断提高。高空管制区作为空域中重要的组成部分,航班起飞后的大部分时间都在高空管制区飞行,其繁忙程度往往反映出整个空域运行的繁忙程度。

在我国当前的高空管制模式下,管制扇区和管制员的安排在一定时间内是固定的,但是由于一天内的航班流量会有较大波动,也由此造成管制员工作的不平衡。因此,对高空管制区的航班流量进行统计分析并做出准确的预测,将有助于空中交通管理机构提前做出预案和调整以应对航班流量的变化,从而缓解目前管制负荷和压力不断增加的局面。

时间序列是将观测值按一定的时间顺序排列的集合,可用于描述现象随时间发展变化的特征,通过把握其稳定性、趋势、周期和不稳定因素,建立合适的时间序列模型,可用于实现对未来时间内数据的预测。时间序列方法自提出以来[1],得了到广泛的应用和发展,目前,常用的时间序列方法可分为基于统计分析的时间序列方法和基于机器学习的时间序列方法。

收稿日期: 2020-06-11 **修订日期:** 2020-07-23

基金项目: 国家自然科学基金资助项目(61671237; U1933119)

作者简介: 倪 超(1997-),男,江苏常州人,硕士研究生,主要研究方向为空中交通流量管理。

基于统计分析的时间序列方法主要包括AR、MA、ARMA、ARIMA、ARIMAX、Holt-Winter等，这些方法和模型在处理线性时间序列时，具有良好效果。在基于统计分析的时间序列模型的研究与应用中，莫凡等[2]建立ARIMA模型，对机场航班延误时间序列进行预测，结果表明，该模型对机场中长期航班延误预测有良好效果。刘博等[3]同样选择ARIMA模型，根据历史数据，实现对未来几年民航运输量的预测，相对误差位于0.1%～0.9%之间。衡红军等[4]构建了动态回归ARIMAX模型，对航站楼内短时段的值机旅客人数进行了较为准确的预测。

基于机器学习的时间序列方法，则是将多种机器学习方法，用于对时间序列的预测，在处理非线性时间序列时具有良好效果。Takashi等[5]使用由多层受限玻尔兹曼机(RBM)构成的深度信念网络(DBN)对时间序列进行特征捕获，可用于近似或短期预测。Rohitash等[6]在对时间序列进行分解后使用Elman神经网络对复杂时间序列进行预测，提高了预测准确性。肖凡等[7]首先使用小波变换对时间序列进行分解，随后采用支持向量机对各组小波系数进行预测和融合。Paulo等[8]为了预测网络流量，采用了神经网络、ARIMA和Holt-Winters 3种方法进行预测并比较。

本研究旨在针对高空管制区航班流量数据建立时间序列模型，对序列的平稳性、周期性等特性进行分析，并通过选取不同的时间序列方法和合适的参数，建立相应的时间序列模型对航班流量数据进行拟合和预测，从而为管制单位提前调整工作计划，平衡管制员工作负荷提供依据和参考。

1 模型建立

时间序列是指，将一系列时刻$t_1,t_2,\cdots,t_i,\cdots,t_n$按照$t_1<t_2<\cdots<t_i<\cdots<t_n$的时间先后次序排列，并按该顺序得到对应的数字组成序列集合$\{x(t_1),x(t_2),\cdots,x(t_i),\cdots,x(t_n)\}$。时间序列预测分析是把观测得到的时间序列数据，利用曲线拟合、参数估计来建立时间序列模型，对未来时间的数据进行预测。本研究采用的时间序列方法包括基于统计分析的ARIMA模型，基于机器学习的随机森林和人工神经网络，用以实现对高空管制区航班流量数据的拟合和预测。

1.1 ARIMA模型

ARIMA(差分整合移动平均自回归)模型是基于统计分析的时间序列方法中较为常用的模型，融合了多种模型的特点，具有只需内生变量，而不需要其他外部变量即可进行拟合和预测的优点。ARIMA模型由ARMA(自回归移动平均)模型及差分过程构成，ARMA模型表达式如下：

$$x_t=\phi_1x_{t-1}+\phi_2x_{t-2}+\cdots+\phi_px_{t-p}+\omega_t-\theta_1x_{t-1}-\theta_2x_{t-2}-\cdots-\theta_qx_{t-q} \tag{1}$$

式中，$\phi_1,\phi_2,\cdots,\phi_p$为自相关模型系数；$x_{t-1},x_{t-2},\cdots,x_{t-p}$为前$p$个值；$\theta_1,\theta_2,\cdots,\theta_q$为移动平均模型系数；$\omega_t$为随机干扰；$x_{t-1},x_{t-2},\cdots,x_{t-q}$为前$q$个值；$x_t$为预测值。

ARIMA模型则是在ARMA模型的基础上增加差分过程，差分表达式为：

$$\nabla x_t=x_t-x_{t-1}=(1-B)x_t \tag{2}$$

式中，B为延迟算子，∇表示差分算子。p,q及差分次数d为ARIMA模型中的主要参数，确定参数后的模型记为ARIMA(p,d,q)。差分次数d的确定其达到平稳所需的差分次数。在这里，常用的平稳性检验方法为ADF(Augmented Dickey-Fuller)检验，目的是检验序列中单位根是否存在，若存在，则说明时间序列不平稳，否则，可以说明时间序列平稳。

模型中p和q值的确定来自于自相关函数和偏相关函数。首先，可通过如下公式：

$$\hat{\gamma}_k=\frac{1}{n}\sum_{i=1}^{n-k}x_ix_{i+k}\qquad k=1,2,\cdots,n \tag{3}$$

获得自协方差的估计值，对自相关函数估计：

$$\hat{\rho}_k=\frac{\hat{\gamma}_k}{\hat{\gamma}_0}\frac{\sum_{i=1}^{n-k}x_ix_{i+k}}{\sum_{i=1}^{n-k}x_i^2}\qquad k=1,2,\cdots,n \tag{4}$$

式中，n表示时间序列内变量值的个数。获得$\hat{\rho}_k$后，可对偏相关系数$\hat{\alpha}_{k,k}$进行估计：

$$\begin{cases}\hat{\alpha}_{1,1}=\hat{\rho}_1\\ \hat{\alpha}_{k+1,k+1}=\dfrac{\hat{\rho}_{k+1}-\sum_{j=1}^{k}\hat{\rho}_{k+1-j}\hat{\alpha}_{k,j}}{1-\sum_{j=1}^{k}\hat{\rho}_j\hat{\alpha}_{k,j}}\\ \hat{\alpha}_{k+1,j}=\hat{\alpha}_{k,j}-\hat{\alpha}_{k+1,k+1}\end{cases} \tag{5}$$

在获得时间序列的自相关函数和偏相关函数后，可根据自相关系数和偏相关系数的拖尾和截尾特性，确定ARIMA模型中参数p和q。但最优参数仍需通过准则函数进行计算和对比，主要包括以下3种准则：

赤池信息准则(Akaike Information Criterion)：

$$AIC(\gamma)=\ln\hat{\sigma}_\gamma^2+\frac{2\gamma}{n} \tag{6}$$

贝叶斯信息准则(Bayesian Information Criterion)：

$$BIC(\gamma)=\ln\hat{\sigma}_\gamma^2+\frac{\gamma\ln n}{n} \tag{7}$$

汉南-昆信息准则(Hannan-Quinn Information

Criterion):

$$HQIC(\gamma) = \ln \hat{\sigma}_{\gamma}^{2} + \frac{\gamma \ln(\ln n)}{n} \tag{8}$$

式中,$\hat{\sigma}_{\gamma}^{2}$ 为极大似然估计值,n 为参与计算的样本个数,γ 为模型定阶数。在使用函数准则进行拟合效果检验时,准则值越低,表示拟合效果越好。

1.2　随机森林

随机森林是一种基于决策树的机器学习方法,也属于集成学习方法,主要的优点为平衡误差,准确度较高,低过拟合风险以及稳定性较高,可用于解决分类和回归预测问题。本研究主要采用随机森林方法对时间序列进行拟合回归和预测。

随机森林中使用的决策树一般为 CART 决策树,但不同的是,并不加入剪枝操作。CART 决策树采用 Gini 系数来选择和划分属性,Gini 系数定义入下:

$$Gini(P) = \sum_{k=1}^{n} P_k(1 - P_k) = 1 - \sum_{k=1}^{n} P_k^2 \tag{9}$$

式中,P_k 值表示每种类别出现的概率,Gini 系数越低表示决策树越稳定。

随机森林方法是在决策树的基础上加入了 Bagging 算法,在使用时,首先在原数据集中通过 Bootstraping 即有放回抽样的方式,随机从总共的 m 个特征中选取 k 个特征,再根据这 k 个特征建立决策树,重复以上过程 n 次建立 n 棵随机的决策树,最后对各个决策树结果进行记录并投票,获得最终结果,因此随机森林具有样本随机和特征随机的双重随机性,可以有效提高预测精度和避免过拟合。

1.3　人工神经网络

人工神经网络是根据人脑神经元网络进行抽象后建立的运算模型,具有较强非线性映射能力、容错性和泛化能力。其基本结构有输入单元、隐单元和输出单元,输入单元接受外部世界的信号与数据,输出单元实现处理结果的输出,隐单元是处在输入和输出单元之间,不能由系统外部观察的单元。

神经网络的输出取决于网络的结构、网络的连接方式、权重和激活函数,其中激活函数对神经网络的结果影响较大,常见的激活函数有 Sigmoid 函数,tanh 函数和 ReLu 函数。其主要特点为,Sigmoid 函数能够把输入的连续实值变换为 0 和 1 之间的输出,但不是以 0 为中心; tanh 函数是以 0 为中心但未解决梯度消失问题; ReLu 函数解决梯度消失的问题,且只有线性关系,运算速度更快。因此,在实际使用时需要根据数据的特点和不同的需要,选择合适的激活函数。

2　实例分析

本研究所使用的数据为广州区域管制中心所辖高空管制区运行三室 AC12 扇区 2019 年 12 月 1 日至 12 月 15 日的航班流量数据,按照小时为单位进行统计和排列作为时间序列,并对其特性进行分析。随后采用以上 3 个时间序列模型和方法进行拟合和预测,并选取相应的评价指标对结果进行对比。

2.1　数据处理与指标选取

首先对本研究所用时间序列流量数据进行归一化操作,归一化是将样本数据映射至 [0,1] 或 [-1,1] 的区间,归一化后可以提高数据的表现,加速模型的收敛过程。本研究采用方法为最大最小值归一法,将数值归一化至 [-1,1],其计算公式为:

$$x_{\text{norm}} = -1 + 2\frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

随后,对数据集进行划分,以 2019 年 12 月 1 日至 12 月 14 日的航班流量数据作为训练集; 以 2019 年 12 月 15 日的航班流量数据作为测试集。而对于随机森林方法和人工神经网络方法,训练集标签为 2019 年 12 月 2 日至 12 月 14 日的分时航班流量数据,特征为对应的前 24 个小时的分时航班流量数据及时间(h); 测试集标签为 2019 年 12 月 15 日的分时航班流量数据为标签,特征为对应的前 24 个小时的分时航班流量数据及时间(h)。

模型将选取决定系数 R^2(Coefficient of Determination) 和均方根误差 $RMSE$(Rooted Mean Squared Error) 为指标对拟合和预测的准确度进行评价。决定系数 R^2 的范围为 [0,1],R^2 的值越接近 1,说明估计值对实际值的拟合程度越好; 反之,说明估计值对实际值的拟合程度越差。均方根误差 $RMSE$ 表示估计值与真实值之差平方和的均值算术平方根,可以评价数据的拟合程度,$RMSE$ 的值越小,说明预测模型描述实验数据具有更好的精确度。

2.2　模型训练

首先采用 ADF 检验方法对数据进行平稳性检验,检验结果如表 1 所示。

表 1　平稳性检验

统计量	数值
T 统计量	-16.970 2
p 值	$9.170\ 32 \times 10^{-30}$
1% 临界值	-3.451 02
5% 临界值	-2.870 64
10% 临界值	-2.571 62

表中,p 值接近于 0,且 T 统计量在 99% 的置信水

平下是显著的，拒绝接受原假设，结果满足平稳性要求。不需对数据再进行差分处理，可以确定 ARIMA 模型中的 d 值为 0。

对时间序列各阶的自相关函数(ACF)和偏相关函数(PACF)进行计算，函数结果如图 1 所示。

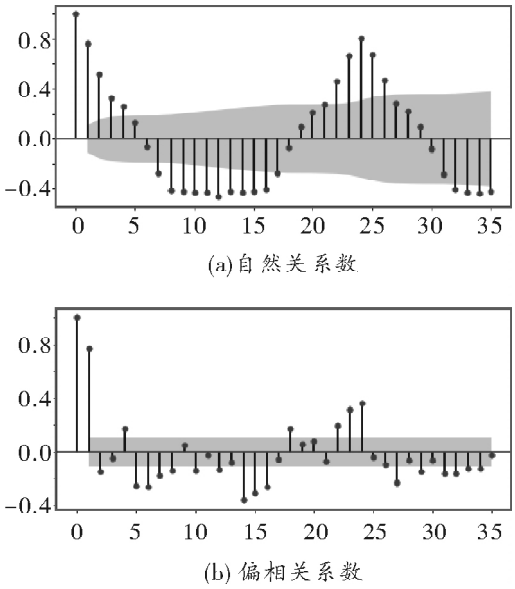

图 1　时间序列自相关系数与偏相关系数

根据图 1，可以看出自相关函数拖尾，故可将 ARIMA 模型中的参数 q 设置为 0；而 PACF 函数的截尾阶数并不明显，但可以看出 2 阶、8 阶、12 阶、16 阶和 24 阶后，偏相关系数落入置信区间，故确定备选阶数为 2，8，12，16 和 24。采用 3 个准则进行评价，见表 2。

表 2　参数准则检验

p	AIC	BIC	HQIC
2	2 469.35	2 484.62	2 475.44
8	2 403.03	2 441.20	2 418.25
12	2 397.58	2 451.02	2 418.88
16	2 211.31	2 280.02	2 238.70
24	2 113.42	2 212.67	2 152.99

由表 2 可知，当 p 值为 24 时，所有准则数值均为最小，因此，可以确定 ARMA 模型中 p 设为 24，由此，ARIMA 模型中的参数确定为(24，0，0)。

对于随机森林方法参数的确定，采用方法为网格搜索，确定各参数最终结果如下：决策树数量 *n estimators* 设置为 45，分裂点属性数目 *max features* 设置为 10，决策树最大深度 *max depth* 设置为自动。根据以上主要参数对随机森林进行训练。

对于人工神经网络，主要参数调整结果如下，一次训练所用样本数 *Batch_size* 设置为 20，隐含层层数 *dense* 设置为 2，激活函数选择为 ReLu 函数，最大迭代次数 *epochs* 设置为 100。

2.3　结果对比

在使用以上模型对时间序列进行拟合后，对数据进行去归一化操作，使其与原数据对应，将拟合结果与实际值进行对比，结果如图 2～图 4 所示。

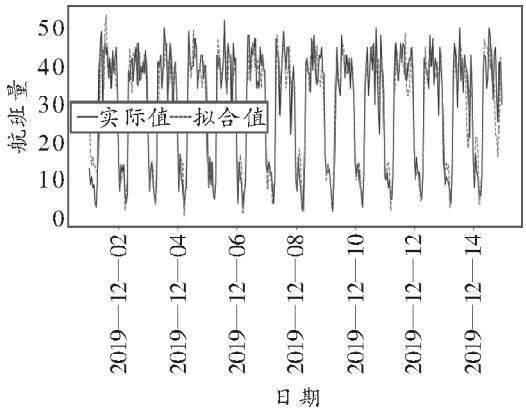

图 2　ARIMA 模型拟合结果

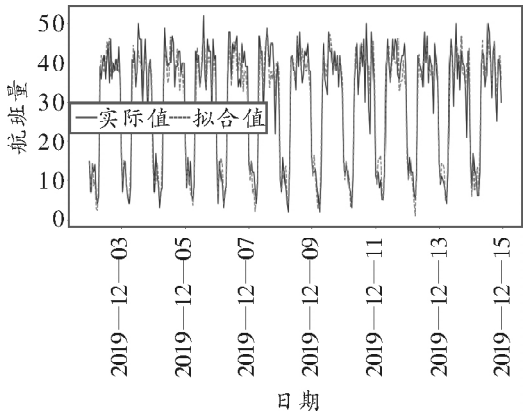

图 3　人工神经网络拟合结果

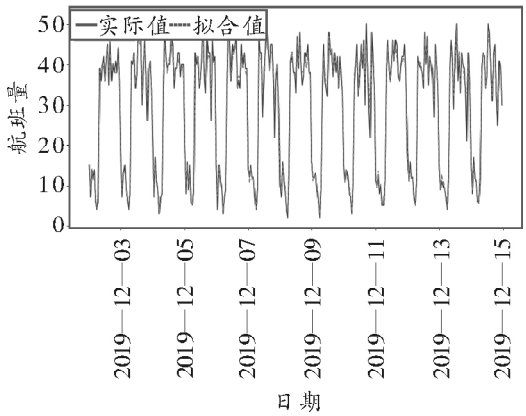

图 4　随机森林拟合结果

从以上对比结果可以看出，随机森林方法对观测值的拟合效果最好，人工神经网络方法次之。其中，随机森林方法中特征重要度最高为时间(h)、前 23 h 和前 24 h 航班量，重要度分别为 0.307，0.238，0.190。

采用以上时间序列模型对 12 月 15 日的扇区分时流量数据进行预测，对比结果如图 5 所示。

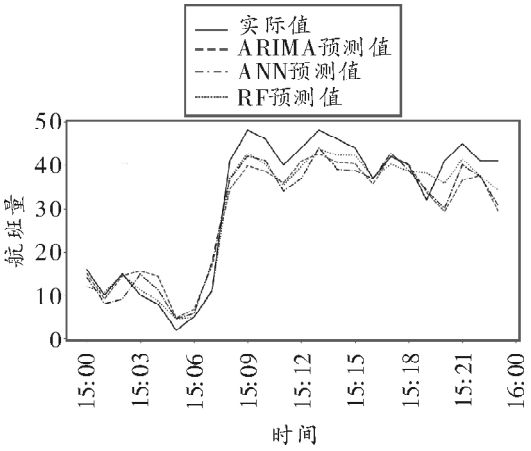

图 5　时间序列预测结果对比

从图 5 中可以看出，随机森林方法的整体预测效果较为稳定，人工神经网络与 ARIMA 模型预测值准确性波动较大。此外，还可以通过指标计算，定量地分析各模型在训练集和测试集的表现，各模型在训练集与测试集的决定系数 R^2 以及均方根误差 *RMSE* 的对比如表 3 所示。

表 3　指标评价对比

预测模型	决定系数		均方根误差训练集	
	训练集	测试集	训练集	测试集
ARIMA	0.835	0.818	5.31	5.89
随机森林	0.981	0.948	2.00	3.63
人工神经网络	0.910	0.894	4.46	5.18

表 3 中，基于随机森林的时间序列方法预测效果最好，训练集决定系数为 0.981，测试集决定系数为 0.948，均方根误差的值分别为 2.00 和 3.63，模型精确度较高; 基于人工神经网络的时间序列方法预测效果也较好，训练集与测试集的决定系数分别为 0.910 和 0.894，均方根误差分别为 4.46 和 5.18; ARIMA 模型预测效果最差，但也达到较为令人满意的水平，训练集与测试集的决定系数分别为 0.835 和 0.818，均方根误差分别为 5.31 和 5.89。

3　结论

1) 本研究依据时间序列分析方法，对航班量时间序列进行了平稳性、自相关系数、偏相关系数等特性进行了分析。

2) 本研究采用基于统计分析的 ARIMA 模型，基于机器学习的随机森林和人工神经网络方法对高空管制区航班量时间序列进行拟合和预测，并得到了较为精确的预测结果，最高预测准确率可达 94.8%。

参考文献:

[1] 乔治·博克斯，格威利姆·詹金斯，格雷戈里·莱茵泽尔. 时间序列分析: 预测与控制 [M]. 王成璋，尤梅芳，郝杨，译. 北京: 机械工业出版社，2011.

[2] 莫凡，赵征，尹韬然. 基于 ARIMA 模型的机场航班延误预测技术研究 [J]. 航空计算技术，2018，48(3): 68 - 73.

[3] 刘博，赵璐，单曲轶. 基于时间序列数据挖掘的我国民航运输量预测分析 [J]. 中国民航飞行学院学报，2019，30(5): 46 - 50.

[4] 衡红军，任鹏. 基于时间序列的机场短时段值机客流量预测 [J]. 计算机仿真，2020，37(2): 26 - 32.

[5] Takashi K，Shinsuke K，Kunikazu K，et al. Time Series Forecasting using a Deep Belief Network with Restricted Boltzmann Machines [J]. Neurocomputing，2014，137: 47 - 56.

[6] Rohitash C，Mengjie Z. Cooperative Coevolution of Elman Recurrent Neural Networks for Chaotic Time Series Prediction [J]. Neurocomputing，2012，86: 116 - 123.

[7] 肖凡，马捷中，任岚昆. 基于小波分析与支持向量机的时间序列预测 [J]. 航空计算技术，2011，41(6): 49 - 52.

[8] Paulo C，Miguel R，Miguel R，et al. Multi - scale Lnternet Traffic Forecasting using Neural Networks and Time Series Methods [J]. Expert Systems，2012，29(2): 143 - 155.