Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 4, No. 1, 2025

Research on the Prediction Problem of Outpatient Income Data Based on the Time Prediction Model

Yifan Ding[1], Min Cai[2], Rongyu Qiao[3], Yangyang Wang[1, *]

[1]School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou 450046, China;

[2]School of Water Conservaney, North China University of Water Resources and Electric Power, Zhengzhou 450046, China;

[3]School of Art, North China University of Water Resources and Electric Power, Zhengzhou 450046, China.

[*] **Corresponding author**: wyybackaroye@icloud.com

Abstract: With the widespread application of internet big data, accurate Medical outpatient data prediction is crucial for improving the efficiency of medical management systems and optimizing resource allocation. This study focuses on the rehabilitation medicine ward of a certain hospital over the past five years, selecting a date, number of patients, total drug revenue, and daily ward income as basic input data. After basic data multiprocessing, an ARIMA time series model is established to predict and fill part of the outpatient income data for the rehabilitation medicine ward. At the same time, under the premise of ADF test stability, the expected data results obtained from the ARIMA model form a feature matrix. Furthermore, a Random Forest Classification Model is constructed to get the feature importance of the predicted data, and the feature importance is used as the weight of the ARIMA model. The weighted average method is used to calculate the predicted value of the overall outpatient income data for the rehabilitation medicine ward. The results show that the established random forest ensemble model prediction method improves the accuracy of medical outpatient prediction by 17.7%, compared to the traditional ARIMA model, it has higher prediction stability and accuracy, and can significantly improve the performance of data mining algorithms on medical datasets.

Keywords: Medical outpatient service data prediction, ARIMA Time Series model, Random Forest Classification Model, Weighted Average Method

1. Introduction

Smart healthcare construction[1], as a new quality of productivity in the field of health and health, includes the outpatient and emergency charge system for outpatient and emergency patients, as well as the inpatient charge system for inpatient cost management. Establishing a smart medical information network platform system allows patients to enjoy safe, convenient, and high-quality diagnostic and treatment services with shorter waiting times and basic medical expenses. Smart healthcare construction has driven innovation-driven development, digital and intelligent, deep integration of industry, optimization of resource allocation, policy support, and environmental construction, high-quality development, improvement of patient experience, and cross-border cooperation in the medical field. It not only promotes the improvement of medical service quality but also provides new directions and momentum for the future development of the medical industry.

In the past, due to the limited improvement of resource utilization efficiency of traditional methods, it was difficult to cope with the occurrence of complex situations. In recent years, establishing a smart medical information network platform system has become a trend, allowing patients to experience high-quality and differentiated medical services in a shorter time. This study establishes an ARIMA time series model, uses appropriate data mining auxiliary software, and realizes the prediction of outpatient income data corresponding to the rehabilitation medicine ward. At the same time, based on the existing prediction, the prediction results of the above time series model are collected, and the original data form a training set and test data set. By establishing a random forest model, further prediction and filling of missing data are realized.

2. The detective values of the ARIMA model

The data of this study comes from https://www.nmmcm.org.cn/match_detail/33.

The ARIMA model is a classic model for predicting single-variable time series, known as the Auto regressive Integrated Moving Average model, denoted as ARIMA(p,d,q), where p, d, and q are the parameters of the model, representing the auto regressive term, the order of conferencing, and the moving average term, respectively[2]. The auto-regressive (AR) part describes the relationship between the current value and historical values. A p-order auto-regressive model can be described as:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \varepsilon_t X_t \quad (1)$$

Where Xt is the value of the time series at time t, c is the constant term, Φ is the auto-regressive coefficient, and ε is the error term. The conferencing (I) is to transform a non-stationary time series into a stationary time series. The number of differences d indicates how many conferencing operations are needed to make the series stationary. For example, the first-order difference is defined as: X'_t; The moving average (MA) part focuses on the accumulation of error terms. A q-order moving average model can be expressed as:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (2)$$

In the formula, μ is the mean of the series, and $\theta 1, \theta 2 ... \theta q$ is the moving average coefficient. Combining the above three parts can form the overall mathematical expression of the

ARIMA model[3]. The ARIMA model mainly focuses on the auto correlation, conferencing nationality, and moving average characteristics of a single time series data. The specific model establishment steps are as follows:

Before modeling research, it is necessary to import the original statistical data to draw a time series chart,like Fig. 1 Overview of raw data. By drawing this time series chart, it can be preliminary judged whether the time series has obvious trends and cycles.

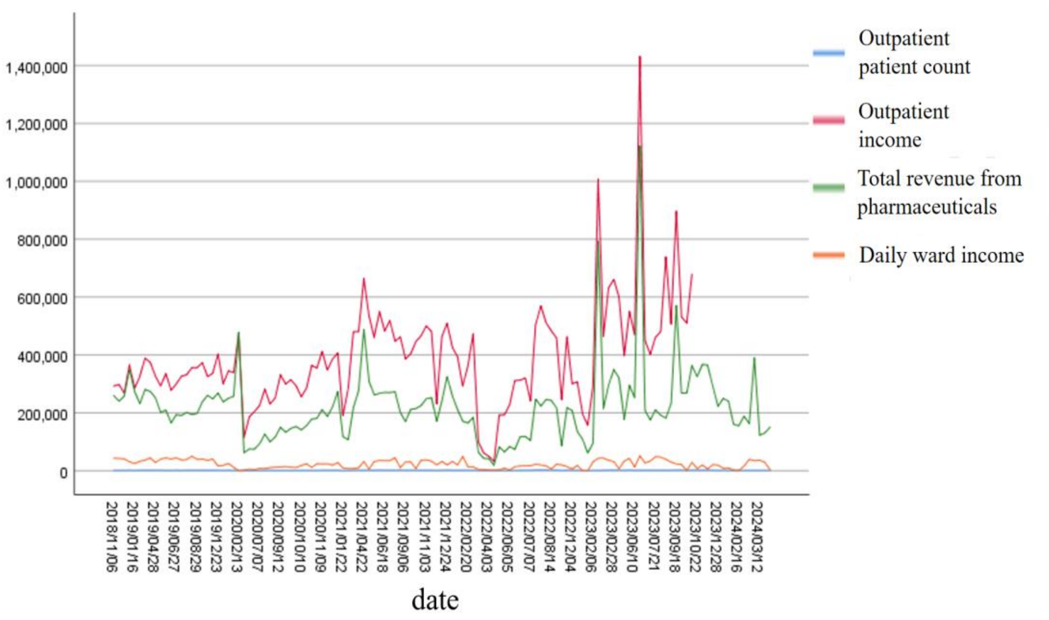

Fig. 1 Overview of raw data

Before establishing the research model, it is necessary to test the stability of the sequence and whether it is white noise. The study uses DFtest or ADFtest to test whether the sequence is stable. Rejecting the null hypothesis indicates stability, and it is necessary to continue through means such as conferencing to stabilize. Only if the null hypothesis of sequence stability is not rejected can modeling continue[4]. The results of this sequence test show that, based on the variable outpatient income, when the conferencing is 0th order, the significance P-value is 0.0003, which is significant at the level, rejecting the null hypothesis, so the sequence is a stable time series. Among them, the obtained ADF Statistic data value is: -4.3733, and the p-value test value is: 0.0003.

By using LBtest and other randomness tests to determine whether the time series is white noise, only rejecting the null hypothesis is acceptable; otherwise, there is no value in continuing modeling if it is white noise. The sequence test results show that the sequence rejects the null hypothesis and is non-white noise. The P-value is 0.0003, and the data test result is non-white noise, which allows for model establishment.

Subsequently, based on the above tests and analysis, the identification of the time prediction model is studied. By analyzing the statistical characteristics of the model's auto-correlation ACF and partial auto-correlation PACF, the model order is determined to be (1,1,1)[5]. It is judged whether the data has seasonality and whether seasonal conferencing should be performed to establish the ARIMA model. The generated auto-correlation ACF chart is as shown in fig 2:

Fig. 2 Auto-correlated ACF plot

By using maximum likelihood estimation, the least squares method, and other methods for model estimation, the model parameters are obtained[6], and the final model formula is as follows:

$$y(t) = 3230.791 + 0.367 * y(t-1) - 0.891 * \varepsilon(t-1) \quad (3)$$

Check whether the residual sequence of the model is white noise to verify the effectiveness of the model. The ARIMA model requires the model to have pure randomness, that is, the model residual is white noise. Check the model test table, according to the P-value of the Q-statistic (P>0.05) for the white noise test of the model, which can also be combined with the information criteria AIC and BIC values for analysis (the lower the better), and can also be analyzed through the model residual ACF/PACF chart. According to the model parameter table, the model formula is combined with the time series analysis chart for comprehensive analysis to obtain the backward prediction order result.

Finally, import the processed data and use computer numerical simulation research to use the well-fitted ARIMA model for the prediction of future outpatient income. The solution result is shown in the following Figure 3 ARIMA model's future outpatient income prediction:

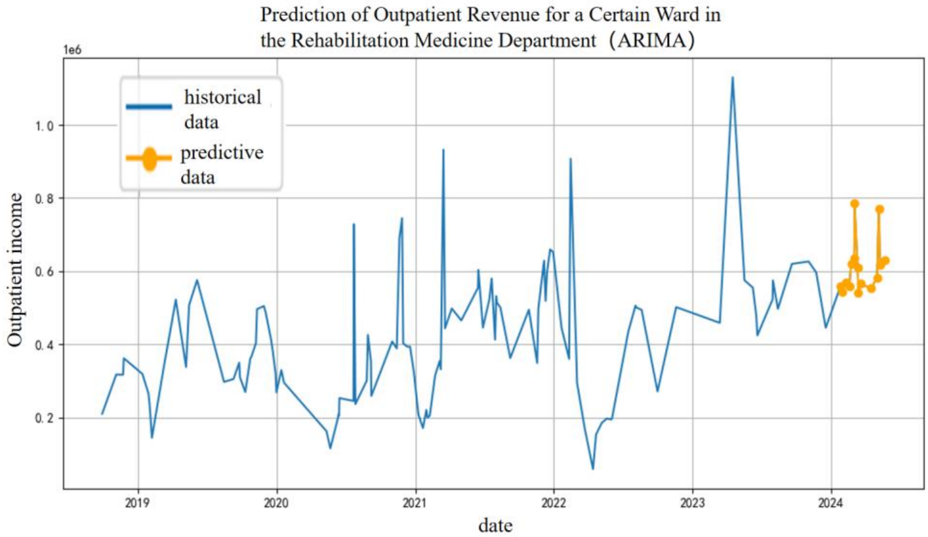

Fig. 3 Prediction of Future Outpatient Revenue Using ARIMA Model

3. Set the model of the random forest algorithm optimization prediction

3.1. Data Multiprocessing

For the original data of outpatient visits, outpatient income, total drug revenue, and daily ward income, data multiprocessing is performed. The Z-value calculation of the normal distribution is generated by SPSS, and the outliers are removed by requiring $|Z| \geq 2$ [7].

3.2. Construction of the Random Forest Model

The data results predicted by the previous ARIMA model are input into the random forest ensemble model to build an ensemble learning model, which is used to improve the prediction performance and stability of the model.

The ensemble learning model is an important concept in the field of machine learning. It refers to a machine learning paradigm that solves problems by combining multiple basic learners. The basic idea can be summarized as "diversity and voting," that is, by constructing multiple basic learners and letting them make independent predictions on input data, and then combining the prediction results of each basic learner in some way, such as voting or weighting, to produce a final prediction result. Random forest is an ensemble learning method composed of multiple decision trees. Each tree makes an independent prediction on the data, and the final decision result is determined by voting or averaging. The specific solution steps are as follows:

First, predict using a single model, use the ARIMA model to predict and obtain the predicted value; use the VAR model to predict and obtain the predicted value; and then construct a feature matrix, by using the predicted results of the ARIMA model and the VAR model as new data features, construct a feature matrix α; finally, perform ADF testing, if the ADF statistic has a large absolute value and is negative, it usually indicates that the data is stable, and the p-value indicates the significance level of the test result. If the p-value is less than 0.05, the null hypothesis can be rejected, indicating that the data is stable. The ADF test for the rehabilitation medicine ward data is as follows: ADF Statistic is -4.2357; p-value is 0.0003.

The above p-value is much less than 0.05, and we have sufficient reason to reject the null hypothesis, proving the nationality of the time series, which means it can be applied to the ARIMA model and the VAR model, because the model assumes that the mean and variance of the data are constant in the time direction, and the assumptions of the ensemble model part are satisfied, so modeling and prediction can be carried out.

By dividing the input feature matrix into a training set and a test set in a 7:3 ratio and inputting it into the random forest model, the top-down greedy algorithm is used to calculate the feature importance of the two models' predicted data[8], as shown in Table 1 Importance of Features in Predicting Data of Two Models below:

Table 1 Importance of Features in Predicting Data of Two Models

Feature Name	Importance of Features
ARIMA	45.80%
VAR	54.20%

In addition, feature importance can help explain the model's prediction results, measuring the contribution of each feature in the model's prediction. High-importance features have a greater impact on the model's decision, while low-importance features have less impact. Based on this importance, the feature importance of the two models is further used as their respective weights, and the final overall income data of the rehabilitation medicine ward is calculated using Python code to calculate the missing values and predicted values, and compared with the predicted data of the other two models. The weighted average prediction value process uses Python code implementation, and the specific code solution result is shown in the following Table 2 Weighted Average Prediction Results:

Table 2 Weighted Average Prediction Results

Index	Combined_Predictions result
0	434586.639554
1	389031.181043
2	402102.745101
3	371054.899278
4	376686.384705
5	360849.193710
6	363407.649504
7	355426.544896

By importing the predicted results of the previous ARIMA model into the data set for comparison with the predicted values, the following Table 3 Comparison of Prediction Results between Two Models can be obtained:

Table 3 Comparison of Prediction Results between Two Models

Date	ARIMA depicted Revenue Value (Yuan)	Random forest prediction income value (yuan)
2023-07-13	448106.3931	423162.1983
2023-07-28	369884.7283	405210.2869
2023-07-30	413098.7853	392810.8883
2023-08-15	357286.3881	382689.5452
2023-08-17	378382.3563	375253.2574
2023-08-28	350348.4203	369722.541
2023-09-02	360634.7463	365750.8038
2023-09-07	346548.8286	362928.379

Subsequently, the comparison of the two methods is shown in Figure 4 Predicted Overall Trend Chart:

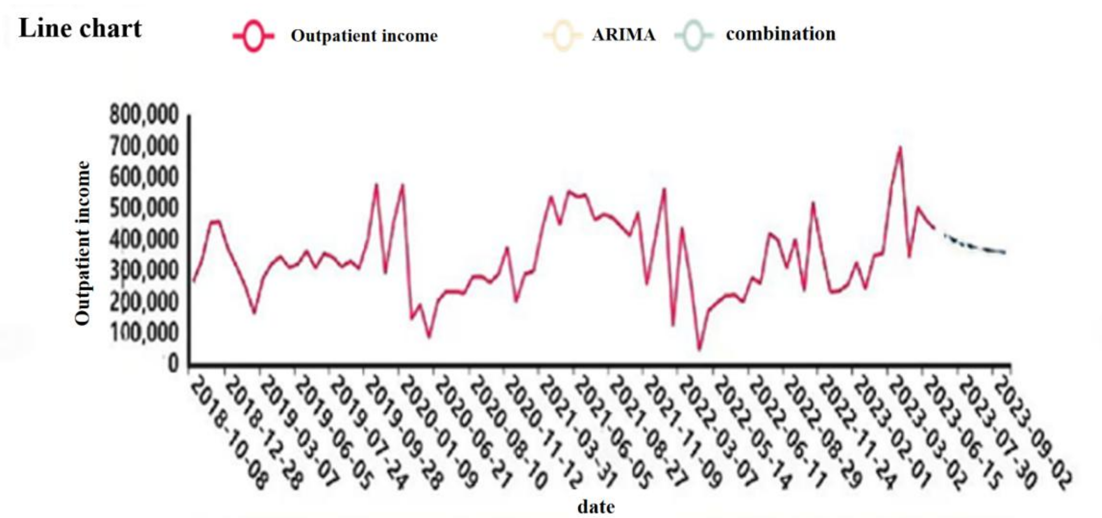

Fig. 4 Predicted Overall Trend Chart

The predicted value fitting chart obtained by the two methods is shown in Figure 5 Prediction value fitting chart as follows. The statistical generation of numerical solution results indicates that compared with the ARIMA prediction scheme, the established random forest ensemble model prediction method improves the accuracy of medical outpatient prediction by 17.7%.

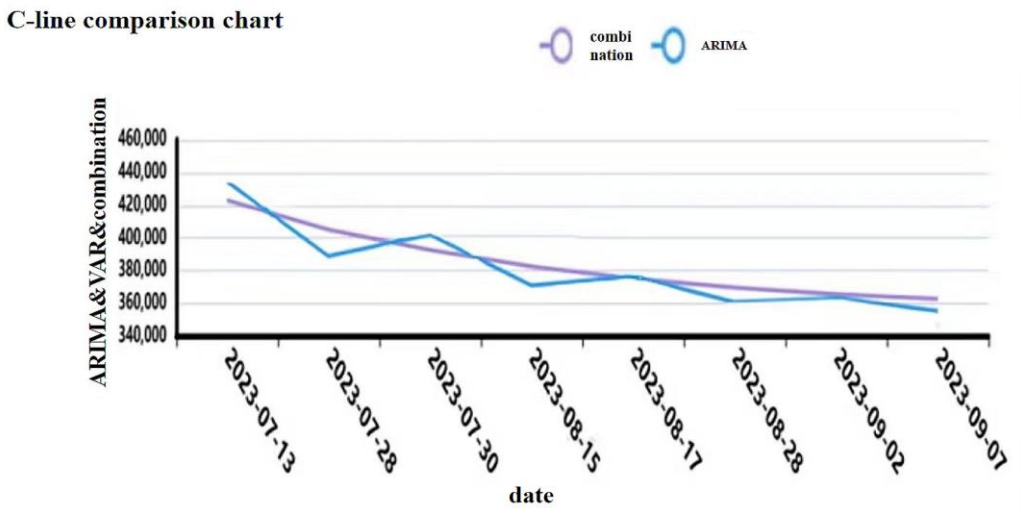

Fig. 5 Prediction value fitting chart

In summary, from the above line charts and numerical solutions, it can be seen that although the outpatient income value of the rehabilitation ward will decline in future development, the overall trend is still upward. Considering the overall background of the times, the development of smart healthcare is an innovative and gradual process. Considering the popularization and promotion of emerging technologies such as 5G and artificial intelligence require time, and the local policy guarantee is not yet perfect, the hospital can increase more humanized care and services while gradually promoting the smart healthcare system, optimize the service process, and improve the medical experience[9]. In the future, smart healthcare will continue to maintain a rapid development momentum. With the continuous advancement of analysis technology and the application of smart models, smart healthcare will become more intelligent, convenient, and personalized, and smart healthcare will also pay more attention to data security, privacy protection, and other issues, providing safer and more reliable medical services for patients.

4. Conclusions

This paper provides a research approach and framework applied to the field of data statistical analysis related to rehabilitation medicine and combines the use of ARIMA and random forest models. By drawing time series charts and observing the trend of each variable over time, a comparative experiment is conducted with the traditional ARIMA time series prediction method on this medical data set. The results show that the random forest prediction has higher prediction stability and accuracy compared to the traditional ARIMA model, and the generated results are easy to visualize, which can significantly improve the performance of data mining algorithms on medical datasets. At the same time, this ensemble model helps medical workers understand the logic and basis behind it, indirectly proving the feasibility of time prediction models for outpatient income data prediction. In the future, the promotion of medical formalization and intelligence will promote data mining and machine learning to shine brilliantly in the rehabilitation medicine department. Technological innovation and algorithm optimization are expected to bring breakthrough progress to rehabilitation medicine and achieve more personalized and efficient rehabilitation treatment.

References:

[1] Fang Mengya, Zhang Yu, Gu Xiaosong, etc Intelligent Medicine and Smart Healthcare [J] Transportation Medicine, 2019, 33 (6): 548-550554.2019.06.003

[2] Zhang Xuepei, Zhou Lin, Liu Min, etc Application of ARIMA Product Seasonal Model in Predicting Monthly Incidence of Foodborne Diseases [J] Public Health and Preventive Medicine, 2024, 35 (5): 6-9

[3] Zhang Xinge, Luo Yue Analysis of Traditional Chinese Medicine Constitution Prediction Based on ARIMA Time Series Model [J] Asia Pacific Traditional Medicine, 2024, 20 (4): 156-162

[4] Wu Yaning, Meng Xiangyu, Shi Fumin Research on Forecasting Shandong Scallion Production Based on Time Series Model [J] Modernization of shopping malls, 2024 (20): 5-7

[5] Liu Zhangjun, Fan Yingfei, Jiang Yunmu, etc Time domain dimensionality reduction expression of seismic motion using filtered white noise model [J] Journal of Vibration Engineering, 2024, 37 (9): 1493-1500

[6] Li Shunjing Maximum likelihood estimation method based on incomplete data - EM algorithm [J] Journal of Chongqing Technology and Business University (Natural Science Edition), 2014, 31 (5): 29-33

[7] Wang Shanshan, Mi Ying Maximum likelihood estimation of multivariate mixed normal distribution parameters [J] Journal of Liaoning University of Technology (Natural Science Edition), 2024, 44 (3): 199-205

[8] Lang Weimin, Chen Zhe, An Haiyan, etc A Scale Exam Planning System Based on Priority Greedy Algorithm [J] Telecommunications Express, 2024 (3): 5-9

[9] Xu Jinnuo, Li Lu, Ma Biao, etc Visualization analysis of smart healthcare graph based on CiteSpace [J] Intelligent IoT Technology, 2024, 56 (3): 1-5