2023 5th International Conference on Decision Science & Management (ICDSM) | 979-8-3503-2501-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDSM59373.2023.00012

Application of ARIMA Model in Infectious Disease Prediction

Zhangying Luo
Qingdao Huanghai university
Qingdao, China
LZYLZY7214@163.com

Yihao Zhang
Qingdao Huanghai university
Qingdao, China
1097157632@qq.com

Chengbo Yin
Qingdao Huanghai university
Qingdao, China
ycb1004@126.com

Meirong Yang
Qingdao Huanghai university
Qingdao, China
2362791025@qq.com

Jing Li*
Qingdao Huanghai university
Qingdao, China
lijing811021@163.com

***Abstract*—By understanding the epidemic trend of infectious diseases, grasping the epidemiological characteristics and laws of infectious diseases, and exploring the application of time series summation autoregressive moving average model to predict the incidence of infectious diseases. This paper discusses the use of the ARIMA model to model the monthly incidence of notifiable infectious diseases in China from January to December 2021, and applies statistical software such as SPSS26.0 to perform parameter estimation, model diagnosis, and model evaluation, and select the optimal model. . ARIMA prediction analysis was performed on the incidence data of infectious diseases in the first half of 2022. The research results show that the high incidence seasons of common infectious diseases are from March to July and from November to January 2022. The actual incidence trend and forecast curve in the first half of 2022 It is highly consistent, indicating that the fitting accuracy and prediction effect of the ARIMA model are both good. Using the ARIMA model to predict the trend of infectious diseases can timely issue early warning of infectious diseases to the emergency system through the relevant network according to the trend prediction information, which has the practical engineering application value of significantly strengthening the emergency prevention and control of infectious diseases.**

Keywords- ARIMA model; infectious disease prediction; infectious disease component;

I. INTRODUCTION *(HEADING 1)*

The ARIMA model system is also one of the most important, most basic and feasible quantitative prediction model systems in the field of modern clinical disease time series prediction quantification and analysis and prediction. Quantitative statistics and prediction of the impact of horizontal changes, especially because it is a quantitative statistic that is mainly researched on the random laws of clinical onset time series with strong seasonal changes in the scope of the disease. Modeling, analysis, and prediction techniques are increasingly being used by clinical academics in various major disciplines of neurobiological research, especially those related to the medical field and other professional branches of life sciences. , it takes into account the random changes of various random and trend factors, the process of periodic and random changes and various random disturbances that occur in time series changes through the comprehensive model analysis of the main variables, and at the same time uses various model parameters. expressed quantitatively. The entire practical and application design process of ARIMA model design does not require any prior analysis and verification of the theoretical development basis and design mode of the time series model itself. Identify or modify it until a completely satisfactory model can be finally obtained. The specific design and application process can be said to be very simple, convenient and easy to operate with any computer software.

Infectious diseases are a class of diseases caused by various pathogens that can be transmitted between humans, animals and animals or between humans and animals. For some infectious diseases, the epidemic prevention department must keep abreast of its incidence and take timely countermeasures. Therefore, after discovery, it should be reported to the local epidemic prevention department in a timely manner, which is called a statutory infectious disease. In today's society, due to the high concentration of life and work, the masses need close communication and contact, and the mobility of personnel in major places is large. Due to the sudden change of people's long-term living environment factors, the increasing intensity of learning content or skills and tasks, and the fatigue of the body due to excessive work, the overall resistance of the population decreases significantly. By establishing the ARIMA model, we can accurately and objectively predict the main trends and characteristics of infectious disease outbreaks, formulate more targeted response strategies for relevant disease prevention and control institutions in a timely manner, and provide objective and scientific evidence for effective prevention and control measures. Data is used as a support, so as to achieve the transformation from passive prevention to active prevention[1].

Source of data The National Health Commission of China compiled and reported 40 cases of infectious diseases including tuberculosis, chickenpox, measles, rubella, mumps, and influenza from January 2015 to December 2021[4]. The monthly incidence of cases, including the basic situation, age distribution, diagnosis and death, was classified and summarized. The collected information is strictly collated

979-8-3503-2500-3/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDSM59373.2023.00012

and verified to ensure the authenticity and reliability of the data source.

II. Related Work

Coronavirus disease 2019 (COVID-2019) has been identified as a global threat, and several studies are being conducted using various mathematical models to predict the likely evolution of the epidemic. These mathematical models based on various factors and analyses are potentially biased. Here, a simple econometric model is proposed for predicting the spread of COVID-2019, and using the Autoregressive Integrated Moving Average model to predict the prevalence and incidence of COVID-2019 epidemiological trends[3]. According to the current development trend of the epidemic, the autoregressive integrated moving average model was used to predict the incidence of tuberculosis in China from 2018 to 2019, to provide a reference for the prevention and control of tuberculosis[4]. The results show that the model can predict the incidence of tuberculosis well, and can be used for short-term prediction and dynamic analysis of tuberculosis in my country, and has good application value. Attributes related to COVID-19 are investigated using public data to develop dynamic mixture models based on SEIRD and the rate of certainty for automatically selected parameters. The model consists of two parts: the modified SEIRD dynamic model and the ARIMA model[5]. The model can analyze input data in real time and provide long-term and short-term forecasts with confidence intervals that meet current trends in epidemic predictions. The aim of this study was to test the accuracy of the ARIMA best-fit model predictions against the actual values reported throughout the lapse of time[6]. The parameters of the optimized model are found by examining the autocorrelation function and partial autocorrelation function plots and different measures of accuracy to find the best fit. The ARIMA model is a well-known econometric forecasting model capable of producing accurate forecasts when applied to wavelet-decomposed time series [7]. The input dataset consisted of daily deaths from the five countries most affected by COVID-19, which were fed to the mixed model for validation and predicted deaths a month in advance. Comparing these forecasts with those obtained from the ARIMA model will help different countries take measures against Covid-19. The aim of this study was firstly to develop a prediction model for daily confirmed COVID-19 cases based on multiple covariates [8], and secondly, to select the best prediction model based on a subset of these covariates. The results show that ARIMA models with optimally selected covariates are useful tools for monitoring and predicting trends in infectious diseases such as COVID-19 cases. Infectious disease forecasting aims to predict all aspects of seasonal and future epidemics. However, it is very likely that a single model will not capture all the patterns and qualities of the dataset. Ensemble learning combines multiple models to obtain a single prediction using the quality of each model. The autoregressive ensemble moving average, exponential smoothing, and neural network autoregression were separately applied to the disease dataset[9]. The gradient boosting model combines the regression values of the above three statistical models to obtain an ensemble model. The results show that the prediction accuracy of the proposed stacked ensemble model is better than that of the standard gradient boosting model. The results show that this method is able to reduce prediction errors in predicting infectious diseases. Predicting the seasonality and trends of infectious diseases is of great significance for the rational allocation of health resources. In this study, we predict the incidence of tuberculosis by building an autoregressive integrated moving average model and provide support for tuberculosis prevention and control during the COVID-19 pandemic[10]. The results show that this model is the optimal model for predicting the trend of infectious disease incidence. According to the current development trend of infectious diseases, an ensemble model is proposed and its applicability to specific disease datasets is proposed. The proposed fusion model is compared with dynamic ensemble modelsfor different error metrics, namely time series dynamic ensemble, quorum dynamic ensemble, and random forest [11]. It can be found that the proposed ensemble model has good accuracy in all infectious disease datasets. In order to accurately predict the development trend of infectious diseases, so as to take measures to deal with it. An integrated spatiotemporal model based on epidemiological differential equations (SIR) and RNNs [12]. The former, simplified and discretized, is a compact model of the temporal infection trend of a region, while the latter simulates the effects of nearest neighbors. The results show that the spatiotemporal infectious disease model of recurrent neural network and differential equation and its good prediction effect in COVID-19.

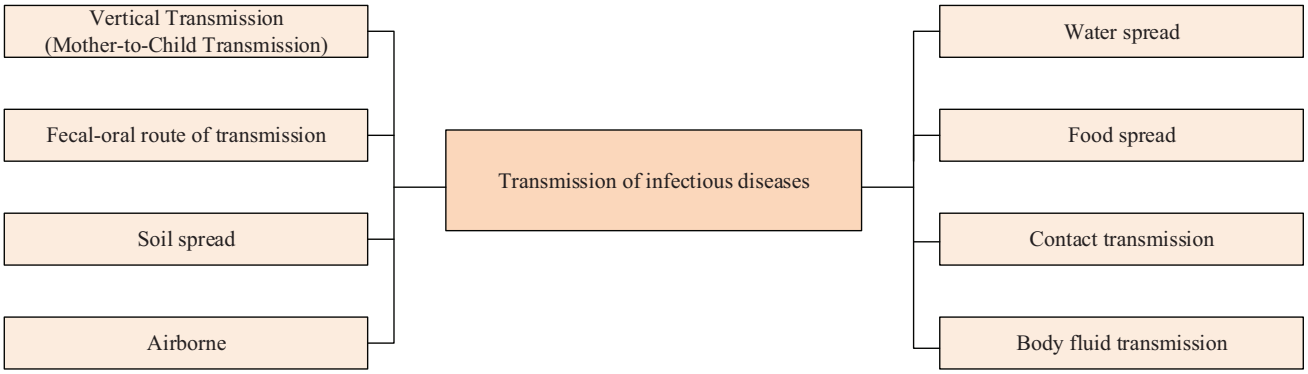

Figure 1. Routes of transmission of infectious diseases.

III. DATA PREPROCESSING

From 2015 to 2021, the number of deaths from notifiable infectious diseases in China will show an increasing trend. In 2020, the number of deaths from notifiable infectious diseases in China will be 26,400, an increase of 1,100 compared with 2019; 20,700 people, a decrease of 5,700 people from 2020. From January to December 2021, there were a total of 6.8104 million patients with legal infectious diseases in China. As shown in **错误!未找到引用源。**.

TABLE 1： *Number of cases of notifiable infectious diseases in China from January to December 2021 (10,000).*

Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Case	52.11	38.96	51.87	60.16	69.95	68.34
Month	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Case	62.86	48.21	50.09	52.3	56.5	69.69

SPSS 26.0 was used for data processing and analysis. The modeling process of ARIMA model is carried out in four stages:

1. Sequence stabilization: the application of ARIMA requires time series to meet the requirements of stationarity;
2. Model identification: further analysis is mainly based on the characteristics of ACF and PACF graphs;
3. Model Parameter estimation and model diagnosis: perform parameter estimation and diagnosis on the proposed model. If the model is inappropriate, go back to the second stage and re-select the model;
4. Prediction application: The data from January to December 2021 is used to build the model.

IV. ESTABLISHMENT OF ARIMA MODEL

In this paper, the monthly infectious disease data from January to December 2021 is used as the input to construct an ARIMA time series model, and the ACF and PACF graphs are analyzed. 1. The difference value (d) is 0, and the moving average value (q) is 7. Perform the model simulation to obtain the simulation diagram as shown in the following figure, and analyze the simulation results to obtain the overall fitting of the predicted value curve (purple) and the true value curve (blue). The trend is good. The first three points have large errors due to fewer input samples for model construction, and the fitted data curve of the latter nine points is closer to the true value curve, indicating that the model has better prediction accuracy. Based on this, this paper uses the model to predict the data for two months backward, as shown by the red line in the figure, and obtains that the data in January 2022 is 620,000, and the data in February is 628,260.

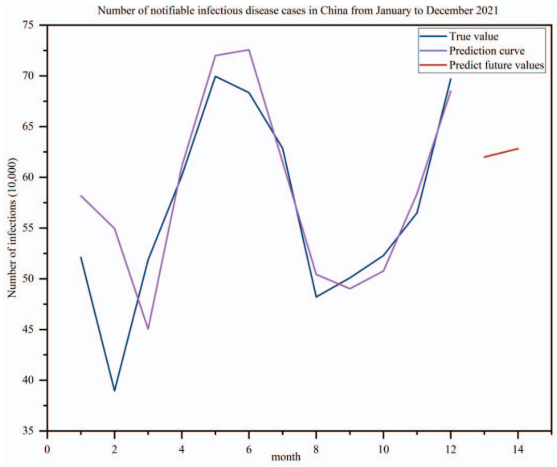

Figure 2. The monthly forecasted incidence of notifiable infectious diseases in China from January to December 2021 by the ARIMA model

V. ERROR ANALYSIS

The average absolute value error of the ARIMA model is 4.166%, and the accuracy of the model is as high as 95%, which is highly consistent with the actual incidence trend and prediction curve in the first half of 2022. It shows that the predicted value of the ARIMA model in the next 6 months has not changed significantly in the incidence of these infectious diseases, and it can better predict the future trend of these infectious diseases. It can predict short-term time series data and has better simulation. Ability, good generalizability.

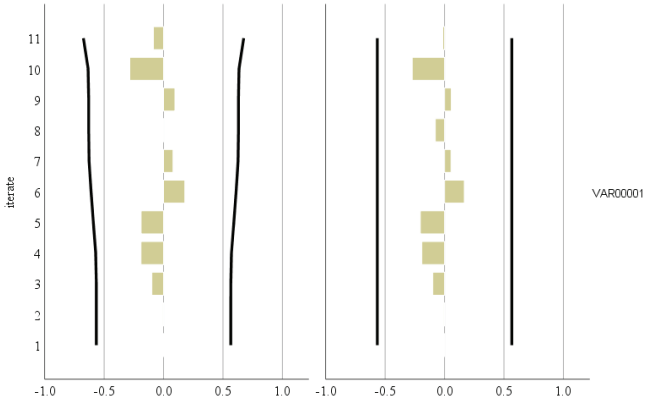

Figure 3. Residual PCF and PACF test

VI. CONCLUSION

In this paper, an infectious disease prediction model based on ARIMA is proposed. By predicting the number of infectious disease cases in China from January to December 2021, and analyzing the computer simulation results, it predicts that the epidemic trend of infectious diseases will change with time and season. As time goes by, there is a more seasonal concentration trend. Winter and spring are the high incidence seasons, which are from March to July and from November to January 2022. The prediction of the infectious disease studied is suitable for the fitting accuracy and prediction of the infectious disease by the ARIMA model. The results are satisfactory. The experimental results show that the infectious disease prediction model proposed in this paper comprehensively considers the characteristics of the changing trend, periodicity and sudden occurrence of infectious diseases, and has high prediction accuracy and reliable results.

VII. DISCUSS

The new technology for the analysis and prediction of the future development and evolution of infectious diseases is to use a technical means according to the current evolutionary characteristics of various key infectious diseases and their possible development of the objective environment, such as the occurrence, development and changes, and the main domestic and foreign related predictable infectious disease factors. Through the use of macro analysis and quantitative judgment theory research and systematic and rational application of mathematical model analysis and calculation and other predictive diagnostic methods, it is necessary to analyze the possible occurrence and potential occurrence of various major preventive infectious diseases and their development and changes, and the possible development trend characteristics of the development and change process. A comprehensive scientific and quantitative judgment on the level of development and change of the law and its possible future development trend and its change mechanism, for us to make a correct decision to formulate the prevention and control of infectious diseases in my country and the rational government guidance and control on important emergent infectious diseases and their occurrence mechanisms. Related technology and coping strategy planning and strategy evaluation research provide important basis.

VIII. ACKNOWLEDGMENT

This paper is one of the research results of "Prediction an d Analysis of COVID-19 income based on time series Model " of Shandong University Students' Innovation and En trepreneurship Training Program. （Project No: S202113320 027)

REFERENCES

[1] Maeda H, Sando E, Toizumi M, et al. Epidemiology of Coronavirus Disease Outbreak among Crewmembers on Cruise Ship, Nagasaki City, Japan, April 2020. Emerging infectious diseases, 2021, 27(9):2251-2260.

[2] Triacca M, Triacca U. Forecasting the number of confirmed new cases of COVID-19 in Italy for the period from 19 May to 2 June 2020. Infectious Disease Modelling, 2021, 6(2).

[3] Wang Y, Shen Z, Jiang Y. Comparison of ARIMA and GM (1, 1) models for prediction of hepatitis B in China. PloS one, 2018, 13(9): e0201987.

[4] Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, 2020, 29: 105340.

[5] Moftakhar L, Mozhgan S, Safe M S. Exponentially increasing trend of infected patients with COVID-19 in Iran: a comparison of neural network and ARIMA forecasting models. Iranian Journal of Public Health, 2020, 49(Suppl 1): 92.

[6] Wang L, Liang C, Wu W, et al. Epidemic situation of brucellosis in Jinzhou city of China and prediction using the ARIMA model. Canadian Journal of Infectious Diseases and Medical Microbiology, 2019, 2019.

[7] He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. International Journal of Infectious Diseases, 2018, 74: 61-70.

[8] Wang K W, Deng C, Li J P, et al. Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network. Epidemiology & Infection, 2017, 145(6): 1118-1129.

[9] Fang L, Wang D, Pan G. Analysis and estimation of COVID-19 spreading in Russia based on ARIMA model. SN Comprehensive Clinical Medicine, 2020, 2(12): 2521-2527.

[10] Pourghasemi H R, Pouyan S, Farajzadeh Z, et al. Assessment of the outbreak risk, mapping and infection behavior of COVID-19: Application of the autoregressive integrated-moving average (ARIMA) and polynomial models. Plos one, 2020, 15(7): e0236238.

[11] Meng P, Huang J, Kong D. Prediction of Incidence Trend of Influenza-Like Illness in Wuhan Based on ARIMA Model. Computational and Mathematical Methods in Medicine, 2022, 2022.

[12] Zhai M, Li W, Tie P, et al. Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province, China: a time series predictive analysis. BMC Infectious Diseases, 2021, 21(1): 1-12.