Dean&Francis
ISSN 2959-6157

Application of the ARIMA Model in Predicting Disease Incidence

Xiyu Qiao

Department of Statistics, University of Illinois at Urbana-Champaign, Illinois, 61820, United States
xiyuq2@illinois.edu

Abstract:

The topic of this review paper is the application of the regressive Integrated Moving Average (ARIMA) model in disease prediction. As one of the most popular models in time series prediction, ARIMA is widely used and receives extensive attention. Predicting the incidence of diseases is important, as it can help the public make timely strategies or prepare in advance for the future, such as the number of medical facilities. This article presents successful cases of ARIMA and models that outperform ARIMA. The conclusion is that the ARIMA model can indeed provide reliable prediction results in many situations. However, when dealing with data that has complex nonlinear characteristics or high volatility (such as the incidence rate of certain infectious diseases), more advanced machine learning models (such as LSTM and random forest) or traditional yet adaptable models (such as Holt-Winters) often exhibit lower prediction errors and stronger fitting capabilities. This article holds that the future direction of scientific research should not be limited to the application of a single model but rather should actively explore the "hybrid model" or "integrated model" that combines the advantages of the ARIMA model with those of other models.

Keywords: ARIMA models; Forecasting; public health.

1. Introduction

The outbreak of diseases such as COVID-19 highlights the importance of sound medical capacity planning and preparedness for emerging crises [1]. This not only threatens human health but also has a profound impact on social and economic development. In history, there have been countless cases where many infections were caused by the failure to make timely plans. For example, due to the lack of public health intervention strategies, the hospitalization rate of the flu was as high as 150% in some areas of Hong Kong in 1970 [2]. As a result, predicting the disease development trend at the start of the disease and then formulating effective control methods are viable ways to reduce the risk of spread.

Autoregressive Integrated Moving Average Model (ARIMA) is a combination of the differenced autoregressive model with the moving average model, which was proposed by Box and Jenkins in the 1970s. As one of the most popular models in time se-

ISSN 2959-6157

ries prediction, the ARIMA model can effectively fit past data and help predict future points in the time series. It is widely used in disease prediction [3]. Although the ARIMA model has achieved many successful cases in disease prediction, it still faces certain doubts in practical applications. Some research pointed out that the ARIMA model has limitations. In certain scenarios, other models such as Long Short-Term Memory (LSTM), random forest (RF), and Holt-Winters exhibit significantly better predictive performance compared to the sole ARIMA model. Therefore, it is necessary to clearly present the principles and advantages, as well as the limitations of the ARIMA model, along with the comparison results with other models. Through this review paper, it is expected to provide a systematic reference basis for researchers engaged in disease prediction and prevention, and to offer theoretical support for public health decision-makers in selecting appropriate prediction tools.

The article first presents some successful cases of ARIMA, raises its controversial aspects, and lists specific examples. Then, compare the ARIMA model with other prediction methods, propose a hybrid model, and explore the potential and development direction of this hybrid model in public health forecasting.

2. Literature Review

The application of ARIMA is extensive. Up to now, there have been many successful examples of ARIMA predictions. In the successful cases used in this article, model validation was conducted through various methods first, and then predictions were made for the topics. The following are six cases of using ARIMA for prediction: Using ARIMA to predict the number of diarrhea cases. The conclusion is that the number of diarrhea cases shows a fluctuating trend. In 2016, there were 4,943 cases, and in 2017, it decreased 4,170 cases. In 2018, it increased sharply to 5,730 cases, but in 2020, it dropped again to 2,227 cases. In 2021, it increased to 2,291 cases. According to the prediction results, it is expected that the number of diarrhea cases in Kenadi City will decrease to 2,250 cases in 2022, and the total number is expected to be 1,612 cases in 2023. In this prediction, the month with the highest number of diarrhea cases occurred in October, which is consistent with the unpredictable abnormal weather conditions [4]. The study evaluated how well the ARIMA model predicted confirmed and recovered COVID-19 cases using Kuwait as an example. The conclusion is that the prediction of the ARIMA model remains highly accurate. For all stages except the first one, the actual numbers of confirmed and recovered cases are within the upper and lower limits of the 95% confidence interval predicted by the model. The researchers attribute the differences of the first stage to the inconsistent COVID-19 testing undertaken during the first phase of the Kuwait preventive plan [5]. Using the ARIMA model to predict leptospirosis in the Baltic region. This study comprehensively evaluated the potential of the ARIMA model in predicting the morbidity of leptospirosis across Estonia, Lithuania, and Latvia. The performance of this model was evaluated using the Mean Absolute Percentage Error (MAPE). Lithuania had the highest accurate forecasts, according to the statistics, with a MAPE value of 6.841. In contrast, Estonia and Latvia exhibited low accuracy, which may be a reflection of regional variations in epidemiological patterns and case variability [6]. Using ARIMA to predict the prevalence of dengue hemorrhagic fever (DHF) in Sulawesi Tenggara, Indonesia. The prediction results show that the peak of dengue hemorrhagic fever incidence occurs in March, July, and November, and it shows an upward trend year by year. The MAPE value of the prediction results for DHF cases is 4.41%, indicating a good prediction effect. Public health practitioners can use this model to prevent and eradicate dengue hemorrhagic fever[7]. In the ICU (Intensive Care Unit), short-term (the next hour) predictions of clinical and laboratory parameters for pediatric patients after cardiac surgery are made to provide early warnings of potential deterioration and assist doctors in timely intervention. The conclusion is that the best ARIMA model selected for neonatal patients is (2, 1, 1); the best ARIMA model selected for infant patients is (1, 1, 0); and the best ARIMA model selected for patients on ventilators is (1, 0, 2) [8]. It is concluded that the number of CKD patients in China and the economic burden of CKD will continue to increase by using the ARIMA (1,1,1) model, which has the determination coefficient (0.99), mean absolute percentage error (0.26%), mean absolute error (343,193.8), and root mean square error (628,230.3) all have obtained a suitable value. To be specific, the number of CKD patients in China will increase by an average of 2.6 million (1.6%) each year, and the total economic burden of CKD in China will increase by an average of 3.1 billion US dollars each year from 2020 to 2025 [9].

With the wide application of ARIMA, it has also aroused many controversies. Many people point out that one of the main assumptions of the ARIMA model is that the time series is stable. Therefore, the ARIMA model is suitable for stabilizing time series data with obvious trends, seasonality, or periodicity. It is also suitable for short-term predictions. Therefore, many people compare ARIMA with other models and draw the conclusion that other models are superior. For example, when it comes to predicting infectious diarrhea, due to the limitation of the ARIMAX model, which assumes linear relationships

between the independent and dependent variables, the random forest (RF) model could provide better predictions. Researchers chose the root mean square error (RMSE) and mean absolute percentage error (MAPE) to compare RF and ARIMA. The conclusion was that RF was better than ARIMA because the RMSE (0.04) and MAPE (6.88%) of RF on the training set were lower than those of ARIMA (RMSE 0.08 and MAPE 28.53%), and the RMSE (0.31) and MAPE (20.89%) of RF on the test set were significantly lower than those of ARIMA (RMSE 0.45 and MAPE 28.53%), with smaller prediction errors and higher accuracy [10]. Regarding the prediction of rubella incidence. Due to the wide application of the ARIMA model and Holt-Winters in the prediction of infectious diseases, the Holt-Winters model and the ARIMA model were compared, and it was concluded that the Holt-Winters multiplicative model is superior to ARIMA because its SMAPE is lower (0.65 vs 1.01), and the prediction accuracy is higher [11]. Regarding the prediction of Tuberculosis (TB) cases. Since different models are suitable for different data characteristics, traditional mathematical prediction models perform well in the prediction of infectious diseases, and no researcher has applied the ARIMA, Grey Model First Order One Variable (GM (1,1)), and Long Short-Term Memory (LSTM) models to predict the tuberculosis cases in the Chinese mainland region. To determine the best prediction model to forecast the epidemic trend of tuberculosis cases in the China region from January to December 2021, the three models were compared using mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) in order to produce more accurate prediction results. The research shows that the MAE values for LSTM, GM (1,1), and ARIMA are 2676.08, 8805.39, and 5638.43, respectively. The RMSE values are 16344.92, 37452.98, and 22599.46, respectively. The MAPE values are 0.0368, 0.1210, and 0.0706, respectively. The values of the LSTM model are smaller than those of the GM (1,1) and ARIMA models. The LSTM model can better fit the trend of the actual values [12].

These cases all demonstrate that although ARIMA is a classic model with clear theoretical advantages, its performance heavily depends on the degree of match between the data characteristics and the model assumptions. When dealing with high non-linearity, non-stationarity, or complex seasonal patterns, although ARIMA research can achieve an accurate prediction, machine learning models (such as RF and LSTM) or more flexible traditional time series models (such as Holt-Winters) may have greater advantages.

3. Methodology

This review paper has utilized 13 articles. These articles were selected through a search on Google Scholar using the keywords “ARIMA/disease prediction”. All the searched literature was published between 2020 and 2025 to ensure that the data remains up to date. During the literature search process, those lacking empirical validation or theoretical contribution were excluded to ensure that the literature searched is based on scientific evidence and can be referred to. The selected articles have different purposes. This paper mainly intends to convey that although ARIMA can perform accurate predictions, in certain circumstances, there exist more suitable and accurate models. Therefore, the searched literature is all related to this topic. Some of them aim to assess the accuracy of the ARIMA model in specific situations. Some of them aim to use the ARIMA model for prediction and obtain data. Others aim to compare ARIMA with other models to determine which one is more effective. However, all the selected articles include validating model accuracy and making predictions. All of them are supported by data, and the obtained data is analyzed.

4. Results

Table 1. Model comparison between ARIMA and alternative forecasting methods

Data object	Comparison model	Evaluation index	The results of ARIMA	The results of the comparison model	The model with better performance
Infectious diarrhea in Jiangsu Province, China	Random forest (RF) model and Autoregressive Integrated Moving	root mean square error (RMSE) and mean absolute percentage	Testing set: RMSE:0.45 MAPE:28.53 Training set: RMSE:0.08 MAPE:13.78	Testing set: RMSE:0.31 MAPE:20.89 Training set: RMSE:0.04 MAPE:6.88	RF

ISSN 2959-6157

	Average Model (ARIMA)	error (MAPE)			
Rubella incidence	ARIMA model and Holt-Winters	symmetric mean absolute percentage error (SMAPE)	1.01	0.65	Holt-Winters
Tuberculosis (TB) in Homa Bay and Turkana Counties, Kenyain Homa Bay and Turkana Counties, Kenya	ARIMA, Grey Model First Order One Variable (GM (1,1)) and Long Short-Term Memory (LSTM)	Mean absolute error (MAE), RMSE and MAPE	MAE:5638.43 RMSE: 22599.46 MAPE: 0.0706	MAE:2676.08, 8805.39 RMSE: 6344.92, 37452.98 MAPE: 0.0368, 0.1210	LSTM

Table 1 compares the performance of the ARIMA model with other prediction models on different infectious disease data, mainly analyzing from the perspective of evaluation indicators (such as RMSE, MAPE, SMAPE, and MAE). Data on infectious diarrhea in Jiangsu Province: The RMSE (root mean square error) of the Random Forest (RF) model on the test set (0.31) is lower than that of the ARIMA model (0.45), indicating that the RF model provides more accurate predictions. Rubella incidence data: The SMAPE value of the Holt-Winters model is 0.65, which is better than 1.01 of the ARIMA model, indicating that its prediction error is smaller. TB data: the LSTM model significantly outperformed the ARIMA and GM (1,1) models in all three indicators (MAE, RMSE, and MAPE). Particularly, the differences in MAE (LSTM: 2676.08, ARIMA: 5638.43) and RMSE (LSTM: 6344.92, ARIMA: 22599.46) were notably significant.

5. Discussion

ARIMA, as a widely used forecasting model, has a high accuracy rate for short-term predictions and is trusted by many research scholars. However, with the development of technology, more people have proposed models that are better than ARIMA. However, most of these models are not applicable to a wide range of situations but are more suitable for specific cases. Therefore, adopting a hybrid model is a feasible strategy. For instance, the hybrid ARIMA-ANN model demonstrated significantly higher prediction accuracy compared to the seasonal ARIMA model. Using the Diebold-Mariano (DM) test, there was a significant difference in prediction accuracy between the ARIMA-ANN model and the ARIMA (0,0,1,1,0,1,12) model, with $p < 0.001$. Therefore, it is recommended to prioritize the use of the hybrid ARIMA model (such as ARIMA-ANN) in this research topic [13].

This review paper has certain limitations. Firstly, in terms of the scope of model comparison, this paper mainly compared ARIMA with a limited number of forecasting models and failed to cover some other models (such as Transformer). Therefore, it may not fully reflect the current technological development level in the field of time series forecasting. Secondly, due to the limitations of the literature search strategy and the selection of databases, it is impossible to completely rule out the possibility of missing some high-quality studies that meet the inclusion criteria. This article states that ARIMA can accurately make predictions and has a better model. However, it does not deny ARIMA. Therefore, the perspective might not be comprehensive. Given the wide application of ARIMA, it is suggested that future research should focus more on integrating ARIMA with other models to create hybrid models, while maintaining its wide applicability and improving its accuracy.

6. Conclusion

Time series prediction holds significant importance in social development. By predicting the incidence of diseases, decision-makers can achieve early warnings, optimize resource allocation, and formulate targeted intervention measures. For instance, it helps the medical system prepare for drug and bed resources in advance and can effectively reduce the disease burden and minimize socioeconomic losses. This review reveals that although ARIMA is widely used and has a high accuracy rate, there are always more accurate models in different specific situations than ARIMA. However, these models are usually tailored for specific situations and cannot be widely applied. To make the time prediction results more accurate and easier to

achieve, future research can develop a combined model that integrates ARIMA with other models, striving to create a model that not only retains the wide applicability of ARIMA but also is more stable and accurate. It is expected to have a combined model that is easier to use and performs more stably and accurately in various situations, truly helping people better predict future trends.

References

[1] Grøntved S, Kirkeby MJ, Johnsen SP, Mainz J, Valentin JB, Jensen VB. Towards reliable forecasting of healthcare capacity needs: A scoping review and evidence mapping. *International Journal of Medical Informatics.* 2024;189:105527. doi:10.1016/j.ijmedinf.2024.105527

[2] Yildirim M, Serban N, Shih J, Keskinocak P. Reflecting on prediction strategies for epidemics: Preparedness and public health response. Ann Allergy Asthma Immunol. 2021 Apr;126(4):338-349. doi: 10.1016/j.anai.2020.11.017. Epub 2020 Dec 9. PMID: 33307158; PMCID: PMC7836303.

[3] V. Kotu and B. Deshpande, Data Science Concepts and Practice, 2nd ed. Cambridge, MA, USA: Morgan Kaufmann, 2019. ISBN: 9780128147610.

[4] Tosepu, R., & Ningsi, N. Y. Forecasting of diarrhea disease using ARIMA model in Kendari City, Southeast Sulawesi Province, Indonesia. Heliyon, 2024,10(22), e40247.

[5] Alabdulrazzaq H, Alenezi MN, Rawajfih Y, Alghannam BA, Al-Hassan AA, Al-Anzi FS. On the accuracy of ARIMA-based prediction of COVID-19 spread. *Results in Physics.* 2021;27:104509. doi:10.1016/j.rinp.2021.104509

[6] M. Butkevych and D. Chumachenko, "Time series analysis of leptospirosis incidence forecasting in the Baltic countries using the ARIMA model," Radioelectronic and Computer Systems, vol. 2024, no. 4, pp. 5–19, 2024. ISSN: 2663-2012.

[7] Mistawati, M., Yasnani, Y., & Lestari, H. Forecasting prevalence of dengue hemorrhagic fever using ARIMA model in Sulawesi Tenggara Province, Indonesia. Public Health of Indonesia, 2021,7(2), 75–86.

[8] Sharwardy SN, Sarwar H, Hasan MNA, Rahman MZ. Assessing the Predictive Capabilities of Autoregressive Integrated Moving Average and Linear Regression Models for Acute Changes in Clinical and Selected Laboratory Parameters in Children After Cardiac Surgery in the ICU. Children. 2024; 11(11):1312.

[9] Jian Y, Zhu D, Zhou D, Li N, Du H, Dong X, Fu X, Tao D, Han B. ARIMA model for predicting chronic kidney disease and estimating its economic burden in China. *BMC Public Health.* 2022;22:2456. doi:10.1186/s12889-022-14959-z

[10] Fang X, Liu W, Ai J, He M, Wu Y, Shi Y, Shen W, Bao C. Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. *BMC Infectious Diseases.* 2020;20:222. doi:10.1186/s12879-020-4930-2

[11] Zhu Z. Trend prediction of rubella incidence based on ARIMA model and Holt–Winters multiplicative model. *Proceedings of SPIE.* 2024;12924:129240M. doi:10.1117/12.3012918

[12] Zhao, D., Zhang, H., Cao, Q., Wang, Z., He, S., Zhou, M., & Zhang, R. The research of ARIMA, GM (1,1), and LSTM models for prediction of TB cases in China. PLoS ONE, 2022,17(2), e0262734.

[13] Siamba S, Otieno A, Koech J. Application of ARIMA, and hybrid ARIMA Models in predicting and forecasting tuberculosis incidences among children in Homa Bay and Turkana Counties, Kenya. PLOS Digit Health. 2023 Feb 1;2(2): e0000084.