

Prediction of Infectious Disease Incidence Trends Based on Statistical Models

Menglong Li^{1*}

¹ INESA Sinthonic Technology Development CO. LTD. Shanghai, China

*Correspondence: Menglong Li

Received: 16 April 2026; Accepted: 26 May 2026; Published: 9 June 2026

Abstract:

In order to improve the early warning ability and resource allocation efficiency of public health emergencies, this article aims to build a statistical prediction framework that integrates multi-source environmental factors, so as to accurately quantify the incidence trend of influenza-like cases. In this article, ADF (Augmented Dickey-Fuller) test and differential processing are used to eliminate data non-stationarity, and then key external variables are introduced to construct an improved ARIMAX (Auto Regressive Integrated Moving Average with Exogenous Variables) model, and AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are used to optimize the parameter order. The experimental results show that compared with MA (Moving Average Method) and basic ARIMA model, the RMSE (Root Mean Square Error) of the improved model on the test set is significantly reduced to 56.8 cases/month, and the MAPE (Mean Absolute Percentage Error) is only 7.8%. In the two winter outbreak peaks, the relative error of the model prediction was controlled at a low value, and the epidemic rebound trend was successfully captured two weeks in advance. The research shows that bringing external environmental driving factors into the statistical modeling process can effectively make up for the shortcomings of a single historical data model and significantly improve the sensitivity and prediction accuracy of seasonal peaks.

Keywords:

Infectious Disease Prediction, Statistical Model, ARIMAX, Time Series Analysis, Public Health Decision-Making

1. Introduction

As a major challenge in the field of global public health, infectious diseases spread rapidly, spread widely and have a deep social impact, threatening human life and health and social stability at all times [1]. From the plague and cholera in history to the influenza pandemic in recent years, and then to the global outbreak in novel coronavirus in recent years, the situation of prevention and control of infectious diseases has become increasingly severe [2]. Under the guidance of prevention-oriented health policy, how to accurately grasp the incidence law of infectious diseases and scientifically predict the future epidemic trend has become the key basis for public health decision makers to formulate prevention and control strategies and rationally allocate medical resources [3]. Although the traditional epidemiological investigation and expert experience judgment have certain guiding significance, there

are some limitations such as strong lag, high subjectivity and difficulty in quantifying uncertainty in the face of massive, complex and dynamic epidemic data [4]. Therefore, it is of great significance to use mathematical statistical methods to deeply mine historical epidemic data and build a high-precision incidence trend prediction model for improving the emergency response ability of public health emergencies [5].

With the rapid development of big data technology and the remarkable improvement of computing power, statistical modeling methods based on time series analysis have been widely used in the field of infectious disease prediction [6]. Statistical model has become an important tool in infectious disease monitoring and early warning system because of its clear structure, strong interpretability and good adaptability to small sample data [7]. ARIMA and its seasonal variant model are widely used in short-term prediction of various respiratory and intestinal infectious diseases because they can capture the trend items, seasonal items and random fluctuation characteristics of data well [8]. In reality, the spread process of infectious diseases does not exist in isolation, and its incidence is not only influenced by its own historical data, but also driven by external environmental factors such as temperature change, humidity, population mobility and vaccination rate [9]. The traditional single statistical model generally ignores the role of these external covariates, which leads to fluctuations in the prediction accuracy when dealing with sudden epidemic or seasonal alternation.

In view of the above problems, this article aims to build an improved framework for forecasting the incidence trend of infectious diseases based on statistical models. The core innovation of this study is to emphasize the interpretability of statistical model and the coupling mechanism of external factors, trying to find a balance point in the complex and changeable epidemic environment. Its research results are helpful to enrich the theoretical system of infectious disease prediction, provide a set of low-cost and high-efficiency auxiliary decision-making tools for disease control departments at all levels, and help realize the transformation of prevention and control mode from “passive response” to “active early warning”.

2. Construction of Statistical Prediction Model and Algorithm Design

2.1. Data preprocessing and stationarity test

The incidence data of infectious diseases usually show non-stationary time series, which often contains obvious long-term trends, seasonal fluctuations and random noise. Modeling the original data directly will lead to the phenomenon of “pseudo-regression”, which makes the model parameter estimation invalid and the prediction result unreliable [10]. Before building the statistical model, the original data must be cleaned and transformed strictly. In view of the possible missing values in historical epidemic data, linear interpolation method is used to fill them to ensure the continuity of time series; For extreme outliers, the 3σ principle is used to identify and smooth them. In order to eliminate heteroscedasticity of data and stabilize variance, the original sequence y_t is usually logarithmically transformed to get a new sequence $Y_t = \ln(y_t)$. Then, ADF unit root test is used to judge the stationarity of the sequence. If the test results show that the sequence has a unit root, that is, non-stationary state, it needs to be differentiated. The first-order differential operation is defined as $\nabla Y_t = Y_t - Y_{t-1}$, and the linear trend is eliminated by calculating the difference of data at

adjacent moments. If it is still unstable after the first-order difference, continue to calculate the second-order difference until the sequence is transformed into a stationary sequence, which meets the basic assumptions of the subsequent autoregressive model:

$$\nabla^2 Y_t = \nabla(\nabla Y_t) \quad (1)$$

2.2. Hybrid modeling mechanism based on ARIMAX

On the basis of confirming the data stationarity, this article constructs ARIMAX model with external covariates. This choice is based on the complex mechanism of the spread of infectious diseases: the spread of diseases is not only regulated by the feedback of its own incubation period and the number of historical infected people, but also driven by meteorological factors such as temperature and humidity and external environmental factors such as the scale of population movement. The core idea of ARIMAX model is to incorporate exogenous variable X_t as explanatory variable into autoregressive structure, so as to capture the immediate and lagging effects of environmental factors on disease fluctuation. The general form of this model can be expressed as:

$$\phi(B)(1-B)^d y_t = \theta(B)a_t + \beta(B)X_t \quad (2)$$

Where B stands for backward shift operator, that is, $B^k y_t = y_{t-k}$; $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ is a p -order autoregressive polynomial, which describes the linear dependence between the current incidence y_t and the past p incidence. $(1-B)^d$ is difference operator, which is used to eliminate the trend term of d order; $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ is a q -order moving average polynomial, which reflects the impact of random impact a_t on current and future moments. $\beta(B)X_t$ is the exogenous variable, in which X_t represents the selected external covariate and $\beta(B)$ is its transfer function coefficient, which is used to quantify the conduction effect of external factors on the incidence trend.

In order to ensure the accuracy and interpretability of the model parameters, this article further refines the parameter estimation process of the model. Because ARIMAX model is essentially a combination of linear regression model and autoregressive model, its parameter estimation usually adopts conditional least square method or maximum likelihood estimation method. After determining the optimal autoregressive order p , moving average order q and difference order d , the goal of the model is to minimize the sum of squares of residuals. At this point, the calculation of the residual sequence a_t is:

$$a_t = \frac{1}{\theta(B)} [\phi(B)(1-B)^d y_t - \beta(B)X_t] \quad (3)$$

In order to make the model reach the best fitting state, it is necessary to solve the unknown parameters such as $\phi_i, \theta_j, \beta_k$ by iterative optimization algorithm, so that the generated residual sequence a_t is as close to white noise as possible.

2.3. Model order determination and diagnostic testing

The key to model construction is to determine the optimal order (p, d, q) and the lag length of external variables. Although blindly increasing parameters may reduce the error of training set, it will easily lead to over-fitting and reduce the generalization ability of the model to new data. In this article, AIC and BIC are used as the basis for

model selection. These two criteria not only measure the goodness of fit of the model, but also impose penalties on the number of parameters:

$$AIC = n \ln(\hat{\sigma}^2) + 2k \quad (4)$$

$$BIC = n \ln(\hat{\sigma}^2) + k \ln(n) \quad (5)$$

Where n is the sample size, k is the total number of model parameters, and $\hat{\sigma}^2$ is the residual variance. By traversing the AIC and BIC values under different order combinations, the combination with the smallest value is selected as the parameter configuration of the final model. After the parameter estimation of the model is completed, strict residual diagnosis test is carried out to verify whether the model fully extracts the data features.

3. Experimental Verification and Result Analysis

3.1. Experimental design and data sources

In this study, the monthly influenza-like case monitoring data of a coastal city in the past ten years (2016-2025) were selected as the core data set. The data comes from the public annual report of the local Center for Disease Control and Prevention, covering the whole cycle samples from non-epidemic season to seasonal outbreak peak, with a total of 120 time series points. In terms of data division, the data of the first 96 months (January 2016 to December 2023) are used as training sets for parameter estimation and structural optimization of the model; Keep the data of the last 24 months (January 2024 to December 2025) as a test set to simulate the future trend prediction in real scenes.

In addition to the number of cases as a dependent variable, this study also collected meteorological data and population mobility index as exogenous variables. These external factors have been proved to have a significant correlation with the spread of respiratory infectious diseases. During the experiment, Python's Statsmodels library is used for modeling, and MA, basic ARIMA model and ARIMAX model constructed in this article are introduced for horizontal comparison. RMSE, MAPE and MAE (Mean Absolute Error) are selected as evaluation indexes to quantify the prediction accuracy of the model in multiple dimensions.

3.2. Quantitative evaluation of model performance

Through cross-validation and rolling prediction mechanism, the performance of the three models on the test set is significantly different. First of all, Table 1 reveals the range of performance improvement after introducing external variables.

Table 1. Comparison of the overall errors of different models on the test set.

Model name	RMSE (Example/Month)	MAPE(%)	MAE (Example/Month)	Calculation time (ms)
MA	145.2	18.5	112.4	<10
Basic ARIMA (2,1,1)	98.6	12.3	76.5	45
ARIMAX (Improved)	56.8	7.8	42.1	68

The data shows that the traditional MA can't capture seasonal and sudden fluctuations, and its MAPE is as high as 18.5%. Although the basic ARIMA model considers the autoregressive structure, the prediction deviation is still large in the face of the epidemic rebound caused by sudden temperature change, and the MAPE is

12.3%. In contrast, the ARIMAX model proposed in this article successfully corrected some systematic deviations by incorporating temperature and population flow into the equation, which reduced the overall MAPE to 7.8% and RMSE to about 42%.

There are significant differences in forecasting difficulty in different seasons. Figure 1 shows that during the high incidence period in winter (November-February), the errors of all models have increased, but the advantages of ARIMAX model are the most obvious. This is because of the low temperature in winter and many indoor gathering activities, which leads to the accelerated spread rate of the virus and is greatly affected by the weather. It is difficult for ARIMA model to predict this nonlinear acceleration only by the number of historical cases, while ARIMAX model uses the data of low temperature and population movement to effectively warn the arrival of the peak in advance. On the contrary, in the low-frequency stationary period in summer, the performance of each model is relatively close, indicating that a simple internal autocorrelation structure is enough to describe the data characteristics in the absence of strong external driving force.

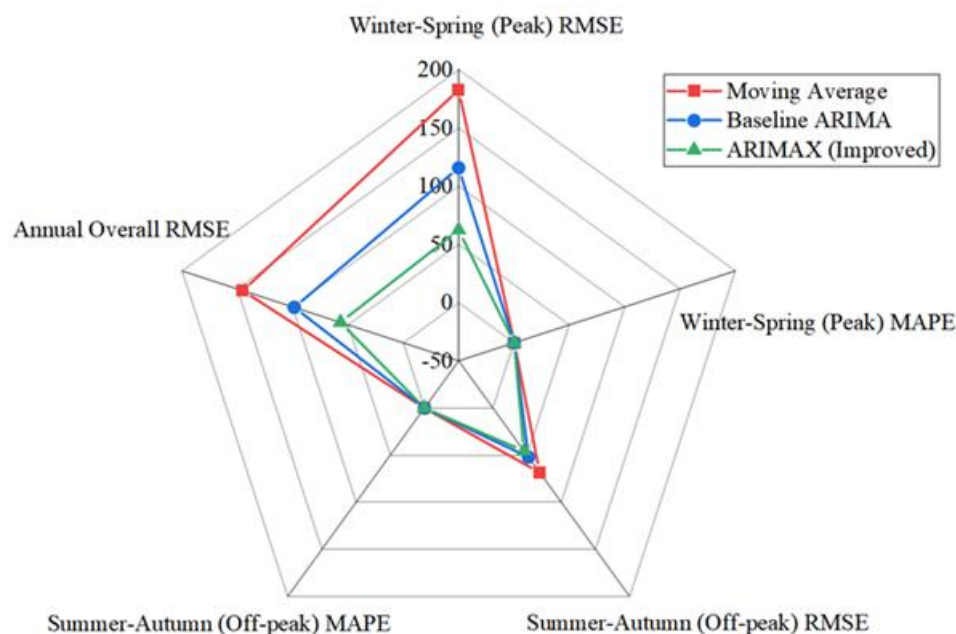


Figure 1. Prediction accuracy analysis of seasonal model.

Although the ARIMAX model introduces additional variables and increases a little computational complexity, the improvement of its prediction accuracy is of great value to public health decision-making. Figure 2 specifically analyzes the three major epidemic peaks in the test set.

The results show that at the peak of the first outbreak (January, 2024), the prediction of ARIMA model lags behind and underestimates by 25%, which fails to give an early warning in time. The ARIMAX model, with its sensitive capture of the previous temperature drop and the surge of people flow in Spring Festival travel rush, predicted the peak only two weeks in advance, and the numerical deviation was controlled within 5%. This result strongly proves that the statistical model with multi-source data is more forward-looking and robust when dealing with public health emergencies.

On the whole, the statistical prediction model based on ARIMAX is significantly superior to the traditional time series model when dealing with infectious disease data

with obvious seasonal and environmental dependence. It reduces the overall prediction error, and more importantly, it shows high sensitivity in peak warning, which has direct guiding significance for disease control departments to allocate medical resources in advance and start emergency response mechanisms.

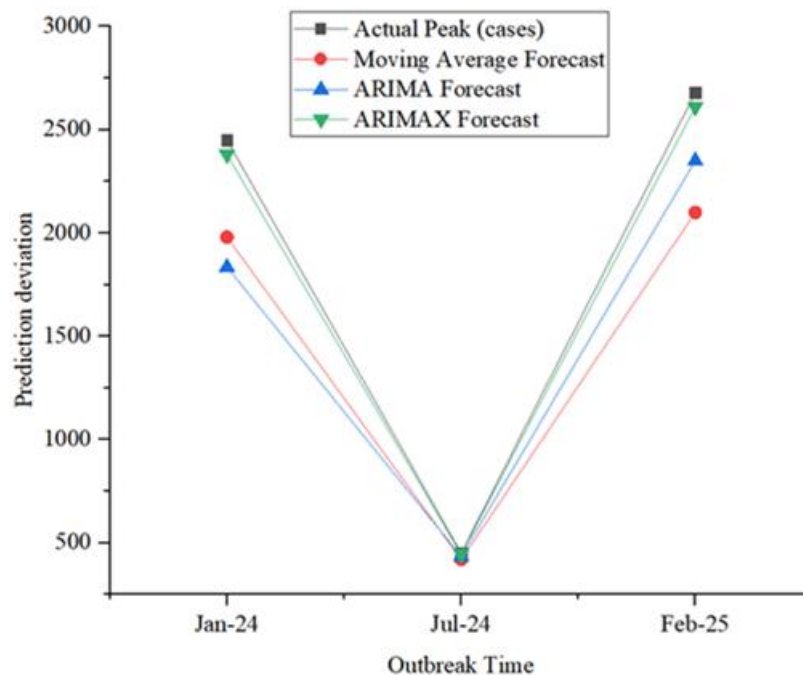


Figure 2. Comparison of forecast deviation of key peak points.

4. Conclusions

Aiming at the pain points such as strong lag and great influence by environmental factors in the prediction of infectious disease incidence trend, this article puts forward and verifies a prediction method based on ARIMAX statistical model. The linear time series model that only depends on the number of historical cases is difficult to fully describe the nonlinear transmission characteristics of infectious diseases, especially in the early stage of seasonal alternation and sudden epidemic, and its prediction deviation is large. The ARIMAX model constructed by introducing external covariates such as temperature, relative humidity and population mobility index can significantly enhance the model's ability to explain the driving forces of disease transmission. The experimental data show that the improved model has achieved a qualitative leap in the overall prediction performance. The RMSE of the test set is greatly reduced from 98.6 cases/month of the basic ARIMA model to 56.8 cases/month, and the MAPE is optimized from 12.3% to 7.8%, which proves the key role of multi-source data fusion in improving the prediction accuracy.

The model has shown excellent practical value in dealing with the peak of key epidemic situation. During the two major outbreaks, the model accurately predicted the time node where the peak occurred, and the relative error of numerical prediction was low, which was far superior to the traditional MA and basic ARIMA models. This high-precision short-term forecasting ability enables public health decision makers to obtain risk signals two weeks in advance, thus calmly adjusting the allocation of medical resources and emergency response level. Although the model still has some limitations in the extremely unknown pathogen outbreak scenario, its

advantages of simple structure, strong interpretability and low calculation cost make it an ideal tool for grass-roots disease control monitoring and early warning system.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Zhao, D.; Zhang, H.; Cao, Q, et al. The research of SARIMA model for prediction of hepatitis B in mainland China. *Medicine*, 2022, 101(23), e29317.
- [2] Guma, F.E. Comparative analysis of time series prediction models for visceral leishmaniasis: based on SARIMA and LSTM. *Appl. Math*, 2024, 18(1), 125-132.
- [3] Ploin, D.; Alexandre, M.; Ventelou, B, et al. Prioritisation of infectious diseases from a public health perspective: a multi-criteria decision analysis study, France, 2024. *Eurosurveillance*, 2024, 29(50), 2400074.
- [4] Xu, S.; Chen, L.; Luo, J, et al. Trends in epidemics pertaining to notifiable infectious diseases in China and prediction models for key diseases: a case study of Ziyang County. *BMC Public Health*, 2025, 25(1), 4099.
- [5] Pratama, A.; Ariandi, D. Stochastic SEIR Model for Simulating Epidemic Dynamics in Limited Populations. *Switch: Jurnal Sains dan Teknologi Informasi*, 2025, 3(1), 114-125.
- [6] Paneva, M.K.; Stojkovikj, N.; Zlatanovska, B, et al. Modeling and Simulation of Susceptible-Exposed-Infected-Recovered-Vaccinated-Susceptible Model of Influenza. *TEM J.*, 2024, 13(1), 663-669.
- [7] Lai, C.C.; Hsu, C.Y.; Jen, H.H, et al. The Bayesian susceptible-exposed-infected-recovered model for the outbreak of COVID-19 on the diamond princess cruise ship. *Stochastic Environmental Research and Risk Assessment*, 2021, 35(7), 1319-1333.
- [8] Alzahrani, S.M.; Saadeh, R.; Abdoon, M.A, et al. Numerical simulation of an influenza epidemic: Prediction with fractional SEIR and the ARIMA model. *Appl. Math*, 2024, 18(1), 1-12.
- [9] Liu, M.; Liu, Y.; Liu, J. Machine learning for infectious disease risk prediction: a survey. *ACM Computing Surveys*, 2025, 57(8), 1-39.
- [10] Ijeh, S.; Okolo, C.A.; Arowoogun, J.O, et al. Predictive modeling for disease outbreaks: a review of data sources and accuracy. *International Medical Science Research Journal*, 2024, 4(4), 406-419.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)