Comparative Analysis of SARIMA and ETS Models in Forecasting Influenza Cases in China

Yuhan Zhang [1, a, *]

[1]*Department of Statistics, University of Washington, Washington, US*
a. email, yuhanz15@uw.edu
** Corresponding author*

Abstract: Accurate prediction of infectious disease cases has always been a research focus in the field of medicine and health. This paper studies the application of time series models in the forecasting of influenza cases in China using the SARIMA and ETS models. The dataset used was collected from the China Public Health Science Data Center from January 2012 to December 2018. The COVID-19 period was excluded to avoid anomalies. Both models were trained on the historical data and tested to see their performance in doing predictions for the year 2018. Based on the evaluation metrics, RMSE and MAPE, the ETS model, one of the best models selected for comparison, was better at modeling the seasonal patterns of influenza, in comparison with SARIMA (0, 1, 0) (1, 0, 0) (12). These findings indicate that the ETS model can make more precise forecasts with wider confidence limits, which may be useful for public health planning and management in China regarding influenza.

Keywords: Influenza prediction; SARIMA; ETS.

1. Introduction

In this perspective, it is highly relevant to note that events of influenza pandemics throughout history, such as the 1918 Spanish flu and the 2009 H1N1 outbreak, have placed enormous importance on the enhancement of forecasting and intervention methods [1]. This is because such events serve as the benchmark for defining serious health consequences concerning influenza and a suitably designed predictive model would enhance public health preparedness and response endeavors. Effective infectious disease forecasting is the backbone of well-managed public health administration. It allows for the implementation of timely interventions, efficient use of resources, and, eventually, the mitigation of adverse health consequences and the economic repercussions of outbreaks. Certain diseases exhibit regular seasonal patterns: examples include influenza, which causes a considerable health burden each year. This reiterates the need for reliable forecasts. Severe influenza epidemics can result in tens of thousands of deaths each year. Indeed, accurate prediction of outbreaks would better prepare the authorities and, hence, reduce the burden of illness and death. Since infectious diseases remain a global threat, predictive model development becomes increasingly imperative.

Various models have been used in recent works for predicting infectious diseases. For example, some research has used the SIR model to predict swine flu outbreaks [2]. Other methods, such as two-strain influenza models, explore how one strain of the virus has its dynamic altered once a vaccination is developed for another strain [3]. Many different statistical models have also been tried in the analysis of influenza, but often examine how it would spread given certain conditions. For instance,

the application of SARIMA models has been done to forecast influenza trends in the U.S. [4]. Another similar example is that of Upadhya et al. (2008); the authors did propose a statistical mathematical transmission model based on the data of a one-week outbreak in India in the year 2006. The results came out satisfactory while it was validated through simulations [5]. Besides this, because influenza can be associated with the death of patients without being the direct cause, some models have been developed to estimate mortality associated with the disease [6].

Both the models are first trained with the historical data. In this study, both the SARIMA and ETS models were applied to forecast monthly influenza cases in China from 2012 to 2018. The SARIMA model captures the trend and seasonality through autoregressive, differencing, and moving average components. The ETS model is integrated into a more flexible structure, incorporating error, trend, and seasonality. The results demonstrate that the ETS model outperforms in predicting influenza cases. This study will add to public health preparedness by providing more dependable forecast tools that would guide health intervention and policy decisions.

2. Data

The data used in this study is collected and released by the authority, and the data can be found on the Datacenter of China Public Health Science [7]. The whole data contains the number of cases of influenza for every province from 2004 to 2020. The monthly data of influenza cases of the entire country is obtained from this website.

Due to the Covid-19 beginning at the end of 2019, the data that is used to construct and select model is only from January 2012 to December 2018. The overall view of the influenza case data is shown in Figure 1.

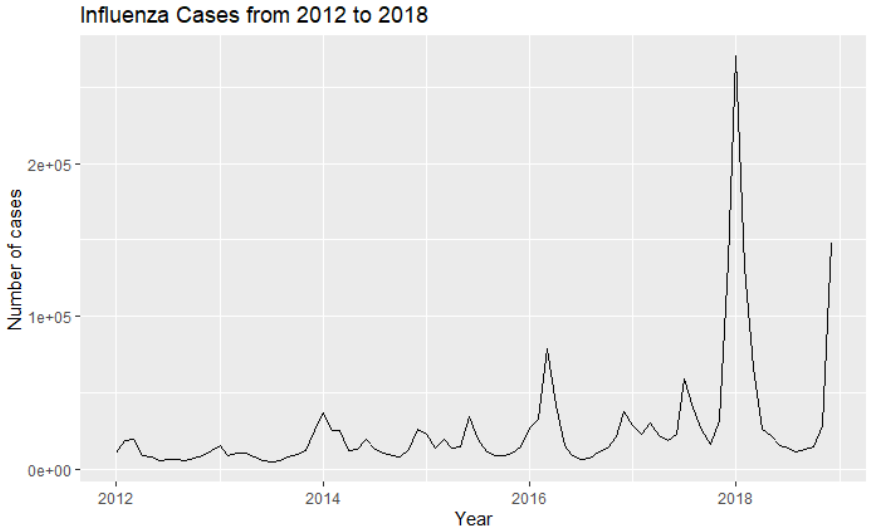

Figure 1: Overall view of data

In Figure 1, although the data changes over time, the data has a pattern. The peaks always show up at the end of a year, while the overall number of cases increases over the years.

3. Model Specification

3.1. SARIMA

The SARIMA model, which stands for Seasonal AutoRegressive Integrated Moving Average, is a powerful time series forecasting method used to model data that exhibit trends or non-stationary

behavior. This model is particularly useful for time series that exhibit both trend and regular periodic fluctuations over a fixed interval, such as monthly or quarterly data. The SARIMA model operates through six primary components: non-seasonal autoregression (AR), seasonal autoregression (SAR), non-seasonal differencing (I), seasonal differencing (SD), non-seasonal moving average (MA), and seasonal moving average (SMA).

Autoregression accounts for the relationship between an observation and a few of its lagged values, differencing is used to make the data stationary by removing trends, and the moving average component models the relationship between an observation and the residual errors from previous time steps. The three seasonal components account for the respective relationship in a timely pattern.

The SARIMA (p, d, q) $(P, D, Q)(s)$ model can be mathematically represented by the following Equation (1) and the explanations are shown in Table 1.

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{I=1}^{P} \Phi_I y_{t-sI} + \sum_{J=1}^{Q} \Theta_J \epsilon_{t-sJ} \quad (1)$$

Table 1: Explanations of parameters of ARIMA model

Parameters	Explanation
y_t	Value of the time series at time t
c	Constant
ϕ_i	Coefficient for the ith autoregressive (AR) term
θ_j	Coefficient for the $j$the moving average (MA) term
Φ_I	Coefficient for the Ith seasonal autoregressive (SAR) term
Θ_J	Coefficient for the Jth seasonal moving average (SMA) term
p	Non-seasonal autoregressive order
d	Non-seasonal differencing order
q	Non-seasonal moving average order
P	Seasonal autoregressive order
D	Seasonal differecing order
Q	Seasonal moving average order
s	Seasonal period

Here, p is the number of lagged values incorporated into the autoregressive part, and q is the number of lagged forecast errors incorporated into the moving average component. Differencing order d is essentially the number of times the time series needs to be different for it to turn stationary. If $d > 0$, then there is differencing that cleans the data from trends and makes the data stationary before the application of the AR and MA parts.

For time series data with seasonality, the SARIMA includes the seasonal components of P, D, and Q, which are seasonal autoregressive, differencing, and moving average, respectively. The seasonal period, s, determines the length of the seasonal cycle [8]. For example, 12 for monthly data showing yearly seasonality. While P denotes the parameter of how the observation relates to its seasonal lags, Q denotes the forecast errors from the previous seasonal periods.

3.2. ETS

ETS refers to a method of time series data forecasting that actually combines the trend, seasonality, and error components. The structure is quite simple and, therefore, easy to understand, and it carries less mathematical complexity. This makes reaching a starting point with the model rather fast, even for non-experts. Secondly, the model has fewer parameters, a condition that reduces overfitting risk

and increases its stability and reliability generally. Also, ETS models are flexible in providing users with choices for trend and seasonality components, depending on the nature of the data, which facilitates the ETS model to be integrated within a broad range of data sets and forecast applications.

The model contains three basic elements: a) error, accounting for random variation in the series, which usually is not supposed to be due to either trend or seasonal patterns; b) trend component, which describes the long-term course taken by the data; and c) seasonality, intended for regular variations within a fixed period [9].

Figure 2 presents the data processing workflow of the ETS model. First, model identification is carried out aiming to identify the presence of trend and seasonality in the time series, and to choose an appropriate error type: additive or multiplicative. Then, model fitting is performed by estimation of parameters from historical data and their application for fitting the model to the series. Finally, after the model is fitted, it is used for forecasting future values. Lastly, the prediction error is calculated in order to measure how good the predictive accuracy of the model is and to compare the predicted values with actual data.

Figure 2: Procedure of ETS model

4. Result

The training data used in this article is from January 2012 to December 2017, and the testing data is from January 2018 to December 2018. All data is recorded in monthly matter. The ETS and SARIMA models are trained on the training data set, and tested on the testing data set, with the accuracy statistics calculated to determine which model is better. Specifically, the values of MAPE are used to determine which model is better.

After being trained in the training data set, SARIMA (0, 1, 0) (1, 0, 0) (12) and ETS (M, A, M) are selected. For the SARIMA model, there is no autoregressive or moving average term, and the differencing order is $d = 1$, indicating that the data is stationary after being differentiated once. In the seasonal part of SARIMA model, the autoregressive part has order 1 with no seasonal differencing or the seasonal moving average, and the seasonal period is 12, which indicates annual seasonality. For the ETS model, this model includes a multiplicative error component, an additive trend, and a multiplicative seasonal component. With the forecast package in R [10], the fitting results of SARIMA and ETS models are shown in Figure 3 and 4, respectively.

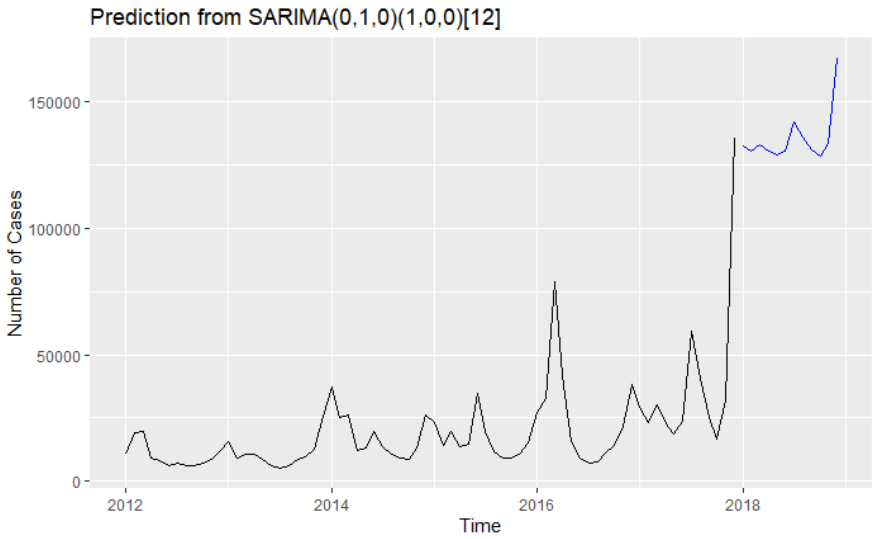

Figure 3: Prediction from SARIMA model

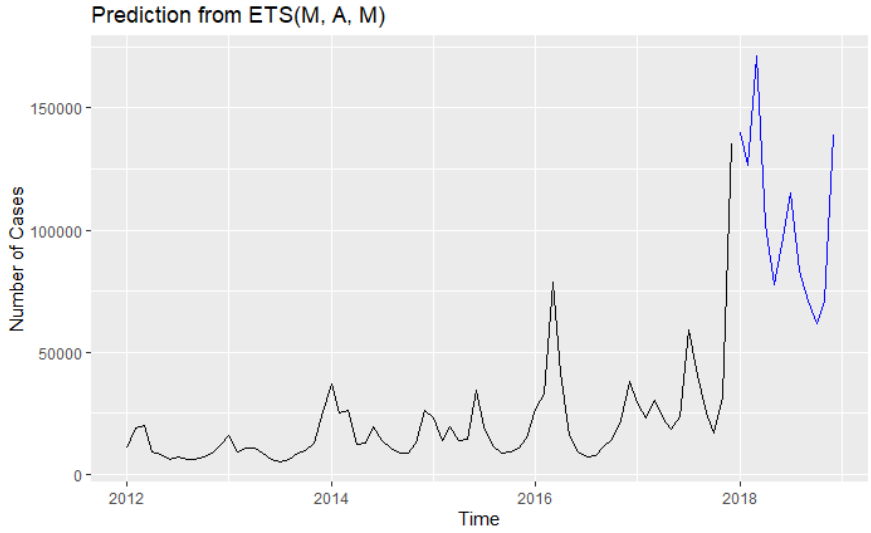

Figure 4: Prediction from ETS model

It is clear that the mean number of cases of the prediction of the two models are nearly the same, while the ETS model captures more about the overall seasonal pattern of the data, which makes it better than the SARIMA model. This is also indicated in the test accuracy of these two models. The two models are used to predict the number of influenza cases from January to December in 2018, and compared with the test data, which is monthly data also from January to December, to get the test accuracy. The Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) are used to determine which model is better [11] (See Table 2).

Table 2: Test accuracy of two models

Model	RMSE	MAPE
SARIMA model	103432.9	474.62981
ETS model	73851.87	286.76616

Based on the accuracy regarding the test set, the ETS model has both a lower RMSE and MAPE, which indicates that the performance of the ETS model is better than that of the SARIMA model.

5. Conclusion

This paper compares the performance of two time series models, SARIMA and ETS, for monthly influenza cases in China from 2012 to 2018. The best-performing SARIMA model, SARIMA (0, 1, 0) (1, 0, 0) (12), and ETS, which is ETS (M, A, M), were trained and evaluated by RMSE and MAPE. Both models capture seasonal trends well, but the results clearly show the superiority of the ETS model over SARIMA because its error rates were lower. The results reflect that the ETS model is a very reliable tool for the projection of influenza progression and can be used as an important guide for public health decision-making in China.

Further research can be done to explore how additional variables can be integrated, which also includes climatic variables and immunization rates, to enhance the predictive accuracy. Also, higher level machine learning models on influenza datasets can be evaluated to see whether they generate better predictive results than the traditional time series methods like SARIMA and ETS.

References

[1] *Hsieh, Y.-C., Wu, T.-Z., Liu, D.-P., Shao, P.-L., Chang, L.-Y., Lu, C.-Y., Lee, C.-Y., Huang, F.-Y., & Huang, L.-M. (2006). Influenza Pandemics: Past, Present and Future. Journal of the Formosan Medical Association, 105(1), 1–6.*

[2] *Coburn, B. J., Wagner, B. G., & Blower, S. (2009). Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). BMC Medicine, 7(1).*

[3] *Ashrafur Rahman, S. M., & Zou, X. (2011). Flu epidemics: a two-strain flu model with a single vaccination. Journal of Biological Dynamics, 5(5), 376–390.*

[4] *Kandula, S., & Shaman, J. (2019). Near-term forecasts of influenza-like illness. Epidemics, 27, 41–51.*

[5] *Upadhyay, R. K., Kumari, N., & Rao, V. S. H. (2008). Modeling the spread of bird flu and predicting outbreak diversity. Nonlinear Analysis: Real World Applications, 9(4), 1638–1648.*

[6] *Thompson, W. W., Weintraub, E., Dhankhar, P., Cheng, P.-Y., Brammer, L., Meltzer, M. I., Bresee, J. S., & Shay, D. K. (2009). Estimates of US influenza-associated deaths made using four different methods. Influenza and Other Respiratory Viruses, 3(1), 37–49.*

[7] *The Data-center of China Public Health Science. Retrieved from: https://www.phsciencedata.cn/Share/index.jsp*

[8] *Box, G. E. P., & Al, E. (2015). Time series analysis: forecasting and control. John Wiley & Sons. 645-660*

[9] *Jain, G., & Mallick, B. (2017). A Study of Time Series Models ARIMA and ETS. SSRN Electronic Journal.*

[10] *Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The Forecast Package for R. Journal of Statistical Software, 27(3).*

[11] *Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250.*