A Research on Theft Crime Based on SARIMA Model and Improved X13-ARIMA-SEATS Model

Qi Zhang

ABSTRACT

This thesis takes the theft crime data of a city as the research object and analyses the time pattern of theft crime, hoping to be able to provide the public security organs certain decision-making reference for preventing and reducing the theft crime. SARIMA model and improved X13-ARIMA-SEATS model were established based on the time series data. The model with higher precision accuracy is selected to analyse the time pattern of theft crime. Through modelling, we can respectively determine the SARIMA prediction model and the improved X13-ARIMA-SEATS model. After that, two models are used for in-sample prediction. By comparing the prediction accuracy of the two models, it can be found that the prediction accuracy of X13-ARIMA-SEATS Model is higher. Therefore, X13-ARIMA-SEATS Model is chosen to analyse the optimal time model for theft crime. Through this model, we can analyse the time pattern of theft crime data.

KEYWORDS

SARIMA model, Improved X13-ARIMA-SEATS model, Theft crime, Time pattern.

INTRODUCTION

Theft Crime, accounts for about 80% and leads the trend of that. The world's police in fighting crimes gradually change from the passive response to the active prevention. It is increasingly important for crime data storage, analysis and prediction. This establishes the foundation for our analysis and research of crime data. Theft crime has strong time pattern, and studying its time pattern is helpful to analyse and predict the crime of theft. At present, domestic research results on the time pattern of theft crimes mainly focus on the basic descriptive statistical level. Compared with foreign studies, our research is not deeply enough, but some...
scholars use mathematical or statistical models to study the time pattern, such as analysing the time series data of theft crimes based on the X11 seasonal adjustment method[1]. There are many methods commonly used in seasonal time series analysis, including seasonal decomposition method, concentration degree method, circular distribution method, ARIMA model, SARIMA model[2], X12 seasonal adjustment method, X11-ARIMA model[3], X12-ARIMA model[4], X13-ARIMA model, as well as TRAMO SEATS model. paper makes analysis by both SARIMA model and X13-ARIMA-SEATS the effect of the mobile holidays (the Spring Festival, Tomb-sweeping Day, Dragon Boat Festival and Mid-Autumn Festival) has to be considered. PBC (department of investigation and statistics of the people's bank of China)

This X13-ARIMA-SEATS model is an improvement of the X12-ARIMA model. It a combination of X12-ARIMA model and the seasonal adjustment function of TRAMO SEATS model. As for Chinese data, developed the X12-ARIMA module with added Chinese Spring Festival effect based on EViews software, but it is only for internal using[5]. In this thesis, there are two methods, which are SARIMA model and X13-ARIMA-SEATS model, are used to compare and predict the data of theft crimes in r language[5].

DATA PREPROCESSING AND RESEARCH METHODS

DATA SOURCE AND PRE-PROCESSING

The theft crime data is from 2012 to 2018, according to the time shall be carried out in accordance with the in summary form monthly data from 2012 to 2018, with undergoing data cleaning and removing duplicate items. Specifically, because leap year and the length of month will have impact on the number of monthly theft cases, monthly data is divided by the number of days of that month, to show the average of crimes in each day during the month. Then assuming that there are 30 days per month, the average values are multiplied to 30, for the purpose of eliminating the influence of leap year effect and the length of month effect on the quite amount of theft cases.

RESEARCH METHODS

A. SARIMA model. The SARIMA is abbreviated as $ARIMA(p,d,q)(P,D,Q)^s$ model, and general expression is:
\[
\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B)\epsilon_t
\]  

(1)

\(Y_t\) in formula (1) represents the research of time series. \(\epsilon_t\) represents a sequence of the random components in a time series. \(B\) means the backward shift operator; \(p\) and \(q\) are the order of autoregressive and moving average, \(d\) is the difference order, \(P\) and \(Q\) are the order of seasonal autoregressive and moving average, \(D\) is the seasonal difference order; \(S\) is the seasonal period.

B. X13-ARIMA-SEATS Model. X13-ARIMA-SEATS model is the improvement of X12-ARIMA model, and it mainly includes X12 and regARIMA module. The X13 module uses the iterative moving average method to decompose the sequence into long-term trend, seasonal component and irregular component. The regARIMA module is mainly used to correct and remove the influences of various outliers, holidays and other factors to extend the sequence forward and backward. The model adjustment process is that, firstly, returning various regression factors, and getting rid of significant effects, and preceding the sequence extension by regARIMA module. Then use the X13 module to adjust seasonality. Each component in the data was adjusted and separated iteratively for many times. Finally, the three components were predicted separately and the total sequence was predicted based on the addition model.

RESULTS ANALYSIS OF SARIMA MODEL AND X13-ARIMA-SEATS MODEL

ANALYSIS OF DATA CHARACTERISTICS OF THEFT CRIME DATA

The monthly number of theft crimes in a city from 2012 to 2017 is shown in Figure 1 below.
According to Figure 1, the characteristics of the data can be preliminarily judged. The data has certainly tendency and seasonal periodicity. First, the overall sequence shows an upward trend, but the trend is not obvious, which needs further verification. Secondly, the series has seasonal periodicity with year as the unit. From January to February of each year, it is in the low point of case occurrence, and April is the peak. Next, the theft crime data (setting up $y_t$; regard “$t$” as the time series) will be analysed and predicted by SARIMA model and X13-ARIMA-SEATS model.

MODELLING OF SARIMA MODEL

STATIONARY TEST

By observing the autocorrelation diagram and partial correlation diagram of theft crime data ($y_t$) (Figure 2). The autocorrelation has a certain feature of trailing, and the 12th order autocorrelation is not 0 significantly, so that the sequence is unstable and has seasonal periodicity. As the further, the unit root test which the t statistic $P=0.093>0.05$ proves through this test. Therefore, the sequence needs 1-order difference and 1-order-12-step difference by observing the autocorrelation of the original sequence diagram (Figure 2). Observing again after a difference of autocorrelation figure (Figure 3 left), and carrying on the unit root test $p = 0.01$, it can be considered that after a 1-order-12-step and 1-order difference sequence is stationary series, and the influence of the season basically is eliminated.
MODEL ORDERING

After stationary test, the seasonal cycle length \( s = 12, d = 1, D = 1 \). The Fig.3 shows that after 1-order difference and 1-order-12-step difference, within 12 steps of less than 1-order autocorrelation coefficients are truncated and partial autocorrelation coefficients are tailed. The preliminarily conclusion is that \( p = 1, 2, 3, q = 1, 2, 3 \). However, for the difference sequence in \( k = 12 \), the autocorrelation coefficient is not 0 significantly. Partial autocorrelation coefficient in \( k = 12, 24 \) are both falling into in random interval. So we must consider \( P = 0, Q = 1 \). Therefore, the preliminary alternative models are 9 groups in all.
PARAMETER ESTIMATION AND MODEL TEST

After repeated debugging of modelling by the maximum likelihood estimation method in R language, all the models in the 9 groups have statistical significance in formula (2). But AIC=1041.51 in ARIMA(1,1,2)(0,1,1) model is the smallest. Therefore, this model is selected as the optimal model. The specific data expression is:

\[(1 - B)(1 - B^2)(1 + 0.738B)(1 - 0.429B)(1 + 0.571B^2)(1 + 0.999B^2)\epsilon_i \quad (2)\]

In formula (2), \(B\) is delay operate. \(\epsilon_i\) is residual error sequence. For alternative models, t-statistic was used to test the significance of all parameters, and all are passed, and then Box-Ljung residual test was carried out. The results showed p-value=0.9459(>0.05). So through ARIMA(1,1,2)(0,1,1) model, the residual of the predicted column and the original sequence is white noise, so the model fitting is sufficient. This model can be used to predict the theft crime data from January to June in 2018. The actual value and the fitting value as well as the predicted value in the sample are shown in Figure 4. The average absolute percentage error of the prediction is MAPE = 20.3%.

Figure 4. Comparison of predicted value, fitting value and actual value of the model.

MODELLING OF IMPROVED X13-ARIMA-SEATS MODEL

DATA PRE-PROCESSING

1. The regARIMA module in X13-ARIMA-SEATS can be used to fix and eliminate all kinds of outliers, such as the calendar effect including mobile holidays, fixed holiday, leap year effect, the length of month effect and the
weekday effect, but the model in R language is based on the US calendar. Weekday effect and trading day effect impact on data is so slight that no need considering. The leap year effect and the length of month effect have been eliminated by averaging above; the fixed holiday effect is the component of seasonal factors in the month. So only the moving holiday calendar effects need to consider. Due to the time of the tomb-sweeping day, dragon-boat festival and Mid-autumn festival are too short to influence theft crime cases, so they can be neglected. So only the effect of the Spring Festival needs to consider. The X13-ARIMA-SEATS model was improved based on these factors. The first step is to eliminate the influence of Spring Festival factors on the data and then to do X13-ARIMA-SEATS process. Assuming that the impact of the Spring Festival reaches n days, including before the festival, during the festival and after the festival, and the influence on the number of theft cases is equal in daily life. If Spring Festival affects $n_1$ days in January and $n_2$ days in February, the variable values is $n_1 / n$ in January, $n_2 / n$ in February, and 0 in the other months (if affects $n_3$ day in March, the value of the variable is $n_3 / n$). By regressing the theft crime variable with Spring Festival variable[5], t-statistics shows that the influence of Spring Festival factors on theft crime data is significant. Theft crime data excluding the influence of the Spring Festival factors is set as $y_t$.

RE-ADJUSTMENT RESULTS OF REGARIMA MODULE

According to the identification of X13-ARIMA-SEATS model, the data excluding Spring Festival doesn’t need logarithmic transformation and which is suitable for addition model. In other words, the theft crime data (excluding the influence of Spring Festival) $y_t$ can be decomposed into the sum of trend composition sequence $y_t^T$, seasonal composition sequence $y_t^S$ and irregular composition sequence $y_t^I$. In regARIMA module, to eliminate the Spring Festival effect the theft of data sequences to construct regARIMA module (ARIMA model combination regression). Through the model identification, it will recognize 11 outliers in total and according to the t statistic test of significance are passed. They are automatically removed and replaced. Building ARIMA model to establish regression residual and ensure $ARIMA(1,0,0)(1,0,0)^{12}$ as the structural model of residual after regression, the regARIMA mathematical expression is:

$$
(1 - 0.676B)(1 - 0.245B)^{12}(y_t - 10563.3) = \varepsilon_t 
$$
F statistics show that there is no seasonal component in the residual sequence of equation (3). RegARIMA module comes to the X11 seasonal adjustment stage after regARIMA expands the sequence.

X13 SEASONAL ADJUSTMENT RESULTS

Data from the X13 module are decomposed into three components: season, trend and irregularity (Figure 5). (1) The data are smoother after seasonal adjustment, which more clearly reflects the change of theft crimes in 2012-2017. According to the Q statistic test, the seasonal adjustment effect of X13-ARIMA-SEATS model is sufficient ($Q=0.89<1$). (2) Upper right in Fig. 5 shows that theft crimes without effect of the Spring Festival have certain seasonal characteristics. From February to May of each year is the peak, which reaches the peak in April, and the level from June to January of the next year is low, and a small peak appears in August and December respectively. Kruskal-wallis chi-square test and F statistic test show that the sequence has identifiable seasonality, but there is no moving seasonality. (3) The figure on the left bottom confirms further the changing trend of seasonal adjustment data. (4) The figure on the right bottom shows that the irregular component is the result of other factors, and its influence is not very great in most months.

![Figure 5. Comparison of data before and after seasonal adjustment (upper left), seasonal composition (upper right), trend composition (bottom left) and irregular composition (bottom right).]
TOTAL SEQUENCE PREDICTION BASED ON DECOMPOSED COMPONENTS

The modelling of trend, season and irregular components based on X13 decomposition is conducive to the rational analysis of the influence of trend, season and irregular factors. In view of the ARIMA model of short-term prediction accuracy is higher. We establish the ARIMA model to forecast of each component and through the model test and parameter estimation. The trend component \((y_t^T)\), seasonal component \((y_t^S)\) and irregular composition \((y_t^I)\) prediction model are:

\[
(1 - B)(1 - 2.414B)(1 + 2.135B^2)(1 - 0.675B^4)(1 + 0.82B^{12})(1 + 0.489B^{24}) y_t^T = (1 - 0.262B) \epsilon_t
\]

\[
(1 - B^2)(1 - 0.858B^2) y_t^S = (1 - 0.251B^2) \epsilon_t
\]

\[
y_t^I = (1 + 0.233B^{12})(1 + 0.303B^{24}) \epsilon_t
\]

The parameter significance (t statistic) and the residual sequences (Box-Ljung) of the three models in formula (4) all passed the test. P-values were all greater than 0.05, which means the residual values of the three models were all white noise, so the fitting effect of the three models was good. The above three models can be used to predict the theft crime from January to June 2018 to get trend component, seasonal component and irregular component respectively predicted value. Then according to the additive model, theft crime in January to June 2018 (assuming 30 days per month) without the Spring Festival factor can be predicted. On the basis, the value which include Spring Festival factor can be calculated, and then revert to the actual days of month forecast. Thus the predicted value in January to June 2018 of theft crime can be calculated. By comparing the actual value with the predicted value, the average absolute percentage error of the prediction is that \(MAPE = 18.15\%\)

COMPARATIVE ANALYSIS FROM PREDICTION RESULTS OF TWO MODELS

In conclusion, the prediction accuracy of X13-ARIMA-SEATS model is higher than that of the SARIMA model, which also shows that the combination of the seasonal adjustment process and the ARIMA model can adjust data better and
predict more accurately. Therefore, the improved X13-ARIMA-SEATS model is chosen as the trend extrapolation prediction model.

CONCLUSIONS AND PROSPECTS

Through the comparative study, the theft crime data of the city from 2012 to 2018 have the following laws:

Firstly, the number of theft crimes in the city has a certain upward trend, but the upward trend is not obvious. In the SARIMA model, that the first-order difference was needed to ensure the stability of the trend which indicated that the data had a certain tendency is found through autocorrelation graph and unit root test. However, in the seasonal adjustment of X13-ARIMA-SEATS model, there is no need to make first-order difference.

Secondly, the data of theft crimes in the city have identifiable seasonality. In the process of X13-ARIMA-SEATS, the data without the factors of the lunar New Year sequence exists in stable and recognized seasonality, but there is no mobile seasonality. Embodying in: the occurring of cases keep the high level from February to May each year, peaked in April; from June to January the following year remain at low levels; August and December appears two small peaks respectively. Public security mechanism should consider seasonal influence effect adequately when fighting theft crime.

Thirdly, the Spring Festival factors of in the theft crime data have a significant impact. Through the modelling regression of the original data and the Spring Festival variables, the influence of the Spring Festival factors on the theft crime data is significant, which is determined by the population characteristics of the city. The city has a large number of floating populations, and the leaving of floating populations leads to fewer theft crimes during the Spring Festival.

This thesis is only a preliminary exploration of the analysis of crime data by using the time series method. In the next step, the further study will be made in many directions. For example, we will analyse the time series of different types of crime data and crime data in different regions to find deeper laws behind the data. Using the new model to study the crime phenomenon; Spatial analysis of crime data and so on.

REFERENCES