Forecasting crime using the ARIMA model

Conference Paper · November 2008

DOI: 10.1109/FSKD.2008.222 · Source: IEEE Xplore

3 authors, including:

Peng Chen
People's Public Security University of China

Hongyong Yuan
Tsinghua University

Forecasting Crime Using the ARIMA Model

Peng Chen, Hongyong Yuan, Xueming Shu
Department of Engineering Physics, Centre for Public Safety Research, Tsinghua University, Beijing
p-chen07@mails.tsinghua.edu.cn

Abstract

In this paper, time series model of ARIMA is used to make short-term forecasting of property crime for one city of China. With the given data of property crime for 50 weeks, an ARIMA model is determined and the crime amount of 1 week ahead is predicted. The model's fitting and forecasting results are compared with the SES and HES. It is shown that the ARIMA model has higher fitting and forecasting accuracy than exponential smoothing. This work will be helpful for the local police stations and municipal governments in decision making and crime suppression.
***Key words**: ARIMA model, crime forecasting*

1. Introduction

As economy and society developing year by year, many new types of crime appear in China and their total amount are increasing all the time. Now many cities of China are confronted with numerous crimes happening every day, and this have brought a big trouble for crime suppression. So in this way, accurate crime forecasting is desired.

Time series is one of the tools for making prediction. Since autoregressive integrated moving average (ARIMA) model was created by George E.P.Box and Gwilym M.Jenkins in 1970, it has been successfully used in forecasting economic, marketing, industry production, social problems, etc. This model has the advantage of accurate forecasting over short-term for the series $x_1, x_2, \cdots, x_n$. While it has the limitation that at least 50 or more observations should be used.

Using time series model to make short-term forecasting of crime is a new research field appearing recently, and it is relatively rare in the world. In the year 2003, Whilpen Gorr [1] and his colleagues had made short-term crime forecasting for the local police station of Pittsburgh in America. They took the exponential smoothing as the forecasting tool to make monthly prediction and got satisfactory results. However, they didn't explore the method of ARIMA model in further.

While in this paper, the ARIMA model will be taken to make short-term forecasting of property crime for one city of China, and the fitting and forecasting results will be compared with the simple exponential smoothing (SES) and holt two-parameter exponential smoothing (HES). The paper is organized as following: firstly, the principles of ARIMA, SES and HES are reviewed in section 2. Next, the formulation of ARIMA model based on crime data is given in section 3. Then the models are applied to make short-term forecasting in section 4. Finally, the conclusion is given.

2. Review the concept of models

2.1. ARIMA model

ARIMA (***p, d, q***) model is a linear model that fit for dealing with stochastic series. Generally, it is originated from the autoregressive model AR (***p***), the moving average model MA (***q***) and the combination of AR (***p***) and MA (***q***), the ARMA (***p, q***) model [2]. The ARIMA (***p, d, q***) model is usually organized as the following formation:

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ Ex_s\varepsilon_t = 0, \forall s < t \end{cases} \tag{1}$$

Where ***p, q*** are orders of the AR model and MA model, ***d*** is the number of series difference. ***p, d, q*** are all integers. ε_t is the estimated residual at each time period. If the model is optimal, it should be independent and distributed as normal random variables with mean 0. σ_ε^2 is the variance of the residuals.

$$\nabla^d = (1-B)^d$$

$$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

Where the $\Phi(B)$ and $\Theta(B)$ are polynomials in B of degrees ***p*** and ***q***. B are the backward shift operator.

The formulating of ARIMA model is a complicate process, but in summarization it includes four steps [3]:

(1) Identification of the ARIMA (***p, d, q***) structure.
(2) Estimating the coefficients of the formulation.
(3) Fitting test on the estimated residuals.
(4) Forecasting the future outcomes based on the historical data.

During the four steps, the first is the most important. Usually, the autocorrelation (ACF) and partial autocorrelation (PACF) functions are taken to identify the models because they show different features of the

978-0-7695-3305-6/08 $25.00 © 2008 IEEE
DOI 10.1109/FSKD.2008.222

functions. For AR (***p***), the ACF tails off at the order of ***p*** but PACF cutoff; for MA (***q***), the ACF cutoff but PACF tails off at the order of ***q***; for ARMA (***p, q***), none of the ACF and PACF tail off.

The ARIMA model could provide forecasting results with upper limits, lower limits and forecasted values. The upper and lower limits provide a confidence interval of 1-α. The α is the given confidence, which means that any realization within the interval will be accepted.

2.2. Exponential smoothing

Exponential smoothing method is a tool which is often used to fitting the trend of the series. Usually it is sorted by two types [4]: SES and HES. Their formulations are presented in the Eqs (2) and (3):

SES $$\hat{x}_t = \alpha x_t + (1-\alpha)\hat{x}_{t-1} \tag{2}$$

HES $$\begin{cases} \hat{x}_t = \alpha x_t + (1-\alpha)(\hat{x}_{t-1} + r_{t-1}) \\ r_t = \gamma(\hat{x} - \hat{x}_{t-1}) + (1-\gamma)r_{t-1} \end{cases} \tag{3}$$

Where the α and γ are smoothing coefficients. The two parameters are very important to the forecasting results. It is suggested that the parameters should lie in the interval between 0.05～0.3 so as to get ideal fitting and forecasting outcomes. The determination of initial value of smoothing series is also an important question, usually it is taken by $\hat{x}_0 = x_1$. r_t is the variable of trend series, and its initial value is often presented as following:

$$r_0 = \frac{x_{n+1} - x_1}{n} \tag{4}$$

3. The ARIMA model formulation

The formulation of ARIMA model depends on the characteristics of the series. In this paper, the time series of property crime data comes from the 110 computer aided dispatch (CAD) recordings of the local police station. The property crime includes the robbery, theft and burglary. The three crimes are taken as the research object because they make up more than 90% of the total amount of the crimes. The time series sequence graph of the total property crime is shown in the Fig. 1. From the picture, it can be seen that although some oscillations exist in the series, its trend is rising and keeping steady overall through the time span.

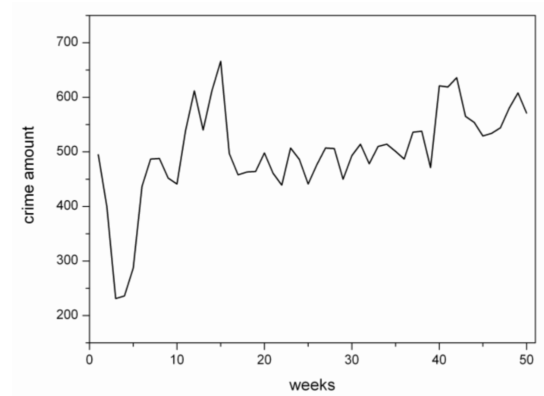

Fig.1. Time series of crime data (by week)

The work of ARIMA model formulating is finished with the statistical software package SPSS.

Firstly, ACF and PACF functions are determined to obtain statistical summary of the series at a particular lag. As shown in Fig.2, it is found that the ACF tails off and PACF cutoff at the order of 1. So according to the mode identification principle in 2.1, the rough model is determined as AR (1). The coefficients of the model are solved with the method of conditional least-square estimation.

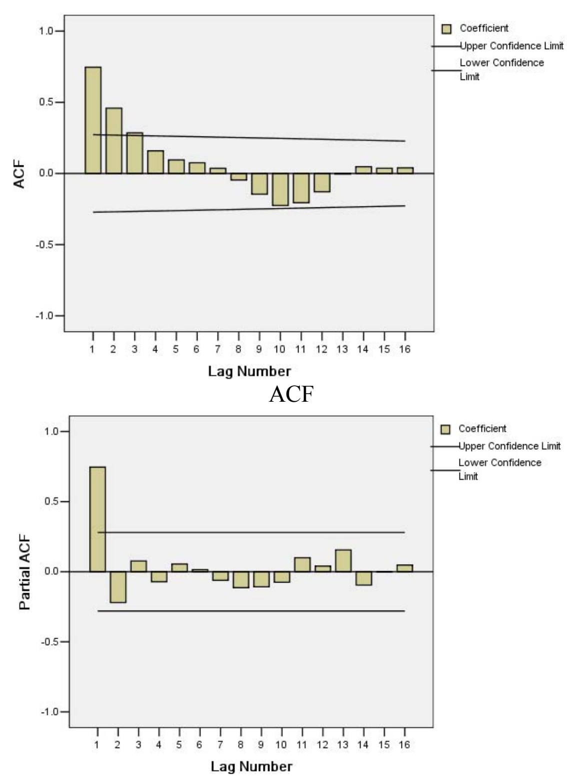

Fig.2 ACF and PACF of the series

The ARIMA procedure fits the model with a certain number of coefficients which are significant [5]. That means a test needs to be done to determine whether the parameters are zero (null hypothesis) or different from zero (alternative hypothesis). Thus two statistics to test are conducted to test the significance of the parameters considered in the model which are T-statistics and p-value. The T statistics is not very informative by itself, but is used to determine the p-value. P-value is determined automatically by the software as 0.05, a-level corresponding 95% of confidence interval. If the p-value is less than this value, the null hypothesis is rejected and the parameter is proved to be significant. If the p-value is more than this value, the null hypothesis is accepted and the parameter will be abandoned.

In this paper, the test result is that p<0.0001, which means the parameters are all significant. So the final formulation of the model is:

$$x_t - 502 = 0.752 x_{t-1} + \varepsilon_t \tag{5}$$

In order to check the accuracy of the model, the root mean squared error (RMSE), and mean absolute percent error (MAPE) [6] are taken as the error measurements. The RMSE and MAPE are presented as Eqs (6) and (7). RMSE is a more objective measure in absolute magnitude because MAPE can be easily affected by the magnitude of series. But MAPE does provide information about the relative magnitude of the forecast error. So in this study, the RMSE is used as primary performance measure and MAPE as the supplementary measure.

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x_t-\hat{x}_t\right)^2} \tag{6}$$

$$\text{MAPE}=\frac{100}{n}\sum_{t=1}^{n}\left|\frac{x_t-\widehat{x}_t}{x_t}\right| \tag{7}$$

4. Result and discussion

In this section, the ARIMA model (5) is used to fit the crime data and make short-term forecasting. The results are compared with the other two exponential smoothing models. The 50 weeks' property crime recordings are chosen as sample series in order to meet the basic requirement of ARIMA model.

The outcomes of fitting of the three models are shown in Fig. 2. As the picture shows, it is found that ARIMA model fits the series better than SES and HES. The SES and HES depict the trend of the series well, but they fail to reflect the fluctuations of the series. In contrast, the ARIMA model fits the series well and even picks up the turning points of the series.

The RMSE and MAPE of the models are tabulated in second and third column in Table 1. Since the numbers represent the goodness of each model, so smaller the number better the model will be. The result shows that ARIMA model fits the series better than SES and HES.

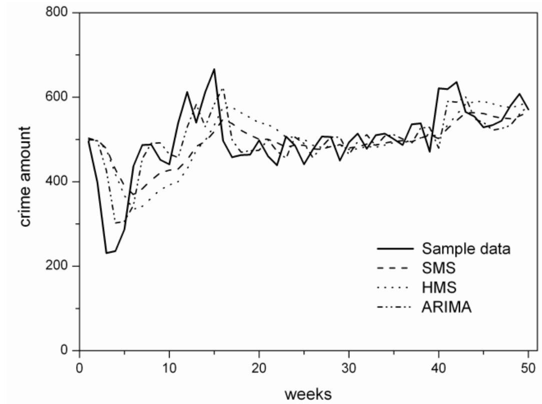

Fig.2. Forecasting performance of three models

With the fitting result, it is believed that ARIMA model makes better accurate forecasting. To verify this point, crime recordings of 1 week ahead is predicted with the three models and their outcomes are compared with the reality. The results are shown in the forth and fifth columns of Table 1. From the table it is seen that ARIMA model approximates the real observation closest. This result is very important in crime suppression for the local police stations and municipal governments. Because with the accurate forecasting crime amount by the ARIMA, some emergency measurements, such as car patrolling, could be prepared in advance and the limited police resources could be deployed rightly to suppress the crime incidents. In this way, the efficiency of working and decision making could be improved greatly for the local police stations and municipal governments.

Table 1 RMSE, MAPE and forecasting result of the models

model	RMSE	MAPE	Forecasting (51^{st} week)	Observation (51^{st} week)
ARIMA	56.94	9.48	567	
SES	75.75	12.82	590	563
HES	84.32	15.02	553	

5. Conclusion

In this paper, time series model of ARIMA is used to make short-term forecasting of property crime for one city of China. The fitting and forecasting results are compared with the other two forecasting tool – the SES and HES. The result shows that ARIMA model fits the data well and makes higher accurate forecasting than the other two models. This work is proved to be very helpful to the local police stations and municipal governments in improving the efficiency of decision-making and emergency management.

6. References

[1] Gorr, W. L., Olligschlaeger, A. M., & Thompson, Y. (2003). Short-term time series forecasting of crime. *International Journal of Forecasting.* 19(2003), pp. 579-594

[2] George E.P.Box and Gwilym M.Jenkins & Gregory C. Reinsel. Time series analysis forecasting and control. *Prentice-Hall, Inc*.4, (1994)

[3] Feng-Mei Tseng, Gwo-Hshiung Tzeng. A fuzzy seasonal ARIMA model for forecasting. *Fuzzy Sets and Systems*. 126 (2002), pp. 367-376

[4] Wang Yan. *Application of time series*. Renmin university of China publisher. 2005

[5] Volkan S. Ediger, Sertac Akar. ARIMA forecasting of primary energy demand by fuel in Turkey. *Energy Policy*. 35 (2007), pp. 1701-1708

[6] H.Brian Hwarng, H.T.Ang. A simple neural network for ARIMA (*p.q*) time series. *Omega*. 29 (2001) pp. 319-333