ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 4, June 2012

Applicability of Box Jenkins ARIMA Model in Crime Forecasting: A case study of counterfeiting in Gujarat State

Anand Kumar Shrivastav, Dr. Ekata

***Abstract*— For any police organization, detection and prevention of crime incidents is a big challenge. Normally police organization maintains data of crime and criminals for various purposes like investigation, coordination and future strategies. The forecast made with the help of historical crime data may support law enforcement agencies in their decision making activities and tactical operations. The time series is one of the tools for making prediction and autoregressive integrated moving average (ARIMA) model has been successfully used in forecasting econometrics, social science and many other problems. This model has the advantage of accurate forecasting over short-term. In this study, we have utilized the crime data of Gujarat State pertaining to counterfeiting of currency. We have used Box Jenkins ARIMA model for short term crime forecasting.**

***Index Terms*— ARIMA, Box Jenkins, Crime, Forecasting.**

I. INTRODUCTION

Despite the increased emphasis on proactive policing, the core of police work remains that of responding to calls for service, making effective deployment strategies paramount to a well-functioning police department [12]. The crime forecasting is an emerging approach in criminological research. Crime forecasting is not widely practiced by police. While there are numerous econometric studies of crime or incorporating crime in the literature, one is hard-pressed to find police departments or other police organizations making regular use of forecasting for policing. From a more proactive standpoint, problem-oriented policing efforts may be enhanced by a more accurate scanning of areas with crime problems, in that one can examine both distributions of past crimes and predictions of future concentrations [8]. The ability to predict can serve as a valuable source of knowledge for law enforcement agencies, both from tactical as well strategic perspectives. Forecasting can help a police department's performance by strategic deployment efforts and efficient investigation direction. The origin of crime forecasting is in year 1998, when US National Institute of Justice (NIJ) awarded five grants to study crime forecasting for police use as an extension of crime mapping [12]. The purpose of this paper was to probe the applicability of univariate time series model for crime forecasting. This paper, attempts to outline the practical steps which need to be undertaken to use Box Jenkins ARIMA time series models for crime forecasting. The paper is organized as following: In Section 2, brief outline of time series forecasting, ARIMA and ARIMA modeling have been given. In section 3 we have applied ARIMA model on historical crime data of Gujarat State, pertaining to counterfeiting, to make short term forecasting. Section 4 is related to results and conclusions.

***Manuscript received May 15, 2012*.**

***Anand Kumar Shrivastav**, Research Scholar, Department of Computer Science, Mewar University, Chittorgarh, India, (e-mail: shrivastav.anand@gmail.com).*

***Dr. Ekata,** Associate Professor, Department of Applied Science, Krishna Institute of Engineering and Technology, Ghaziabad, India (e-mail: ekata4@rediffmail.com).*

.

II. TIME SERIES FORECASTING

A time series is a set of data pertaining to the values of a variable at different times. A time series has an important property which makes it quite distinct from any other kind of statistical data. Formal examples of time series are the population of India at each successive decennial census; daily business handled by a bank, monthly production statistics of a steel mill; monthly traffic accident fatalities etc. The crime criminal data can be arranged in a regular fashion i.e. monthly, quarterly, half yearly or yearly and can be assumed to be a time series data. Time series analysis is used to detect patterns of change in statistical information over regular intervals of time. We project these patterns to arrive at an estimate for the future. The time series analysis helps us cope with uncertainty about the future [8]. All statistical forecasting methods are extrapolatory in nature i.e. they involve the projection of past patterns or relationships into the future [3]. Time series forecast is one of the most important tools for research in the field of social sciences and engineering. Forecasting is an essential tool in any decision making process. The quality of the forecasts management can make it strongly related to the information that can be extracted and used from past data.

A.) ARIMA (Autoregressive Integrated Moving Average)

ARIMA model was introduced by Box and Jenkins (hence also known as Box-Jenkins model) in 1960. It is an extrapolation method for forecasting and, like any other such method, it requires only the historical time series data on the variable under forecasting. ARIMA models are the most general class of models for forecasting a time series. Normally, the ARIMA model is represented as ARIMA(p,d,q) where p is the number of autoregressive

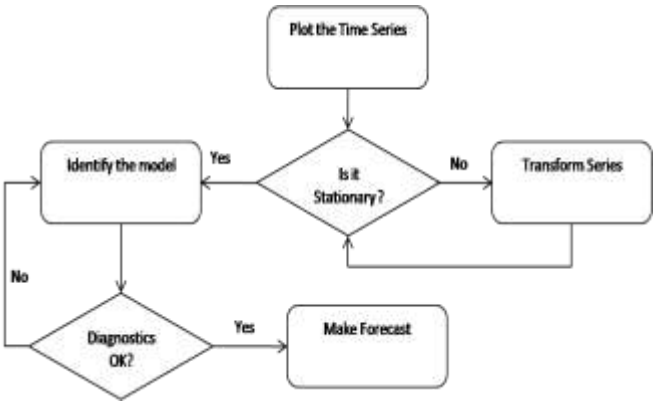
terms, d is the number of non-seasonal differences, and q is the number of lagged forecast errors in the prediction equation. The identification of the appropriate ARIMA model for a time series begins with the process of finding integer, usually very small (e.g., 0, 1, or 2), values of p, d, and q that model the patterns in the data. When the value is 0, the element is not needed in the model. The middle element, d, also known as trend component is investigated before p and q. The goal is to determine if the process is stationary and, if not, to make it stationary before determining the values of p and q. The augmented Dickey–Fuller (ADF) test is most widely used test for checking stationarity of a series. If d = 0, the model becomes ARMA, which is linear stationary model. ARIMA (i.e. d > 0) is a linear non-stationary model. If the underlying time series is non-stationary, taking the difference of the series with itself 'd' times makes it stationary, and then ARMA is applied onto the differenced series. A stationary process has a constant mean and variance over the time period. There are various methods available to make a time series stationary. Normally differencing techniques are used to transform a time series from a non-stationary to stationary by subtracting each datum in a series from its predecessor. If a series is stationary without any differencing it is designated as I(0), or integrated of order 0. On the other hand, a series that has stationary first differences is designated I(1), or integrated of order 1. The term 'shock' is used to indicate an unexpected change in the value of a variable (or error). For a stationary series a shock will gradually die away. In other words, the effect of a shock during time't' will have a smaller effect in time't+1', a still smaller effect in time't+2', etc. The lags of the differenced series appearing in the forecasting equations are called "auto-regressive" terms. The auto-regressive components represent the memory of the process for preceding observations. The value of p is the number of auto-regressive components in an ARIMA (p, d, q) model. The value of p is 0 if there is no relationship between adjacent observations. When the value of p is 1, there is a relationship between observations at lag 1 and the correlation coefficient $\emptyset_1$ is the magnitude of the relationship. When the value of p is 2, there is a relationship between observations at lag 2 and the correlation coefficient $\emptyset_2$ is the magnitude of the relationship. Thus p is the number of correlations we need to model the relationship. The lags of the forecast errors are called "moving average" terms. The moving average components represent the memory of the process for preceding random shocks. The value q indicates the number of moving average components in an ARIMA (p, d, q). When q is zero, there are no moving average components. When q is 1, there is a relationship between the current score and the random shock at lag 1 and the correlation coefficient θ1 represents the magnitude of the relationship. When q is 2, there is a relationship between the current score and the random shock at lag 2, and the correlation coefficient θ2 represents the magnitude of the relationship. When one of the terms is zero, it's usual to drop AR, I or MA component. For example, I(1) model is ARIMA(0,1,0), and a MA(1) model is ARIMA(0,0,1). The autocorrelation function (ACF) and partial correlation function (PACF) are very important for the definition of the internal structure of the analysed series. The models can be identified through patterns in their autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs). The following table summarizes rough guideline for using these parameters in initial model identification.

	AR(p)	**MA(q)**	**ARMA(p,q)**
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

B.) Steps of ARIMA model

(i) Identification of ARIMA (p,d,q) structure
(ii) Estimating the coefficients of the formulation
(iii) Fitting test on the estimated residuals
(iv) Forecasting the future outcomes based on the historical data

The steps of ARIMA model building methodology is presented in a flow chart in figure-1.



(Fig.1: ARIMA model building steps)

III. BUILDING ARIMA MODEL FOR CRIME FORECASTING

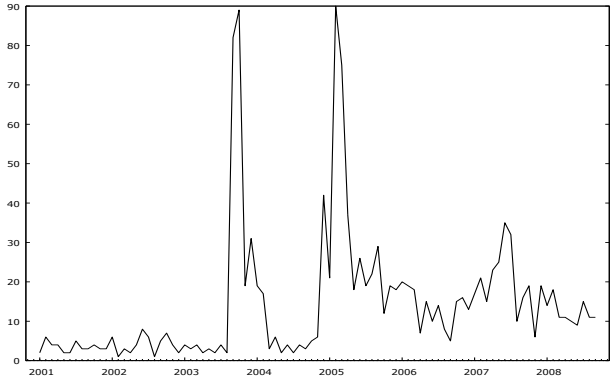
The formulation of ARIMA model depends on the characteristics of the series. In this paper, we have used the Crime data of counterfeit currency for past 8 years (96 months) beginning year 2001 of Gujarat State (Table-1). The crime data has been taken from "Crime in India" , an annual publication of National Crime Records Bureau (www.ncrb.gov.in). We have utilized 93 months data for analysis and 3 months data for validation of our forecasted results. We have used GRETL (Gnu Regression, Econometrics and Time-series Library) software for plotting the graphs and analysis of the data set.

(Table-1: Registered cases of Counterfeit Currency)

Year	Month	No. of Cases	Year	Month	No. of Cases
2001	Jan	2	2005	Jan	21
	Feb	6		Feb	90
	Mar	4		Mar	75
	Apr	4		Apr	37
	May	2		May	18

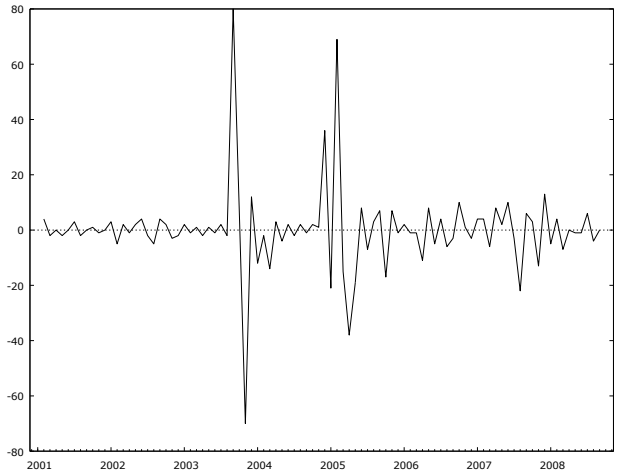
	Jun	2		Jun	26
	Jul	5		Jul	19
	Aug	3		Aug	22
	Sep	3		Sep	29
	Oct	4		Oct	12
	Nov	3		Nov	19
	Dec	3		Dec	18
2002	Jan	6	2006	Jan	20
	Feb	1		Feb	19
	Mar	3		Mar	18
	Apr	2		Apr	7
	May	4		May	15
	Jun	8		Jun	10
	Jul	6		Jul	14
	Aug	1		Aug	8
	Sep	5		Sep	5
	Oct	7		Oct	15
	Nov	4		Nov	16
	Dec	2		Dec	13
2003	Jan	4	2007	Jan	17
	Feb	3		Feb	21
	Mar	4		Mar	15
	Apr	2		Apr	23
	May	3		May	25
	Jun	2		Jun	35
	Jul	4		Jul	32
	Aug	2		Aug	10
	Sep	82		Sep	16
	Oct	89		Oct	19
	Nov	19		Nov	6
	Dec	31		Dec	19
2004	Jan	19	2008	Jan	14
	Feb	17		Feb	18
	Mar	3		Mar	11
	Apr	6		Apr	11
	May	2		May	10
	Jun	4		Jun	9
	Jul	2		Jul	15
	Aug	4		Aug	11
	Sep	3		Sep	11
	Oct	5		Oct	8
	Nov	6		Nov	10
	Dec	42		Dec	16

The Box-Jenkins's methodology for forecasting requires the series to be stationary. The time series plot and correllogram of data provide a strong evidence of a non-stationary series.

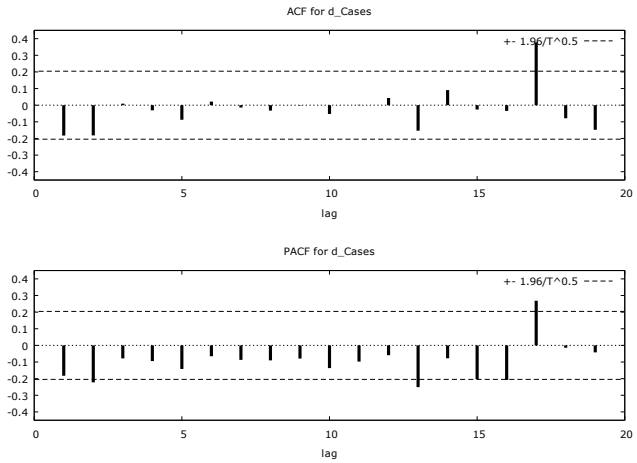


(Fig.2a: Time series plot of original series)

The non-stationarity was also confirmed with the help of Augmented Dickey Fuller (ADF) test. The first order differencing transformation was made to make the series stationary. The time series plot of original and differenced series are shown at figure-2a and Figure-2b respectively. The ACF and PACF plot of differenced series is shown at figure-3.



(Fig.2b: Time series plot of differenced series)



(Fig.3: Correllogram of differenced series)

A.) Model Parameters Estimation

There are several model selection criteria like AIC (Akaike information criteria), HQ (Hannan-Quinn criteria) and SIC(Schwarz information criteria). Among the several methods studied in the literature to judge the fitness of the models, we used Akaike information criterion (AIC) [2]. According to this the model with least AIC value will be selected. We entertained six tentative ARMA models and chose that model which has minimum AIC (Akaike Information Criterion). We have used GRETL(Gnu Regression, Econometrics and Time-series Library) software for model identification and forecasting. Total 118 function evaluations and 35 gradient evaluations were performed by GRETL to find the model parameters. The models and corresponding AIC, HQ and BIC values are given in table-2.

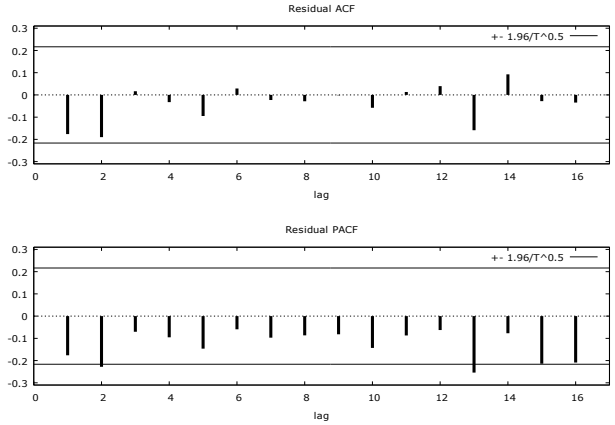
It is obvious from table-2, that the ARIMA(1,1,1) is the best model.

(Table-2: Alternative ARIMA model)

ARIMA Model	Akaike criterion	Hannan-Quinn criterion	Schwarz criterion
101	762.8011	766.8915	772.9315
010	770.0137	771.0315	772.5354
110	770.9250	773.9785	778.4904
111	757.0627	761.1339	767.1498
211	758.9649	764.0539	771.5738
210	768.3624	772.4336	778.4495

B.) Model verification

The model verification is concerned with checking the residuals of the model to see if they contain any systematic pattern which still can be removed to improve on the chosen ARIMA. This is done through examining the autocorrelations and partial autocorrelations of the residuals of various orders. For this purpose, the various correlations upto 16 lags were computed and the same along with their significance were tested by Q-statistics. It was observed that, none of these correlations is significantly different from zero at a reasonable level. The ACF and PACF of the residuals (Figure-4) also indicate good fit of the model. This proves that the selected ARIMA model is an appropriate model.



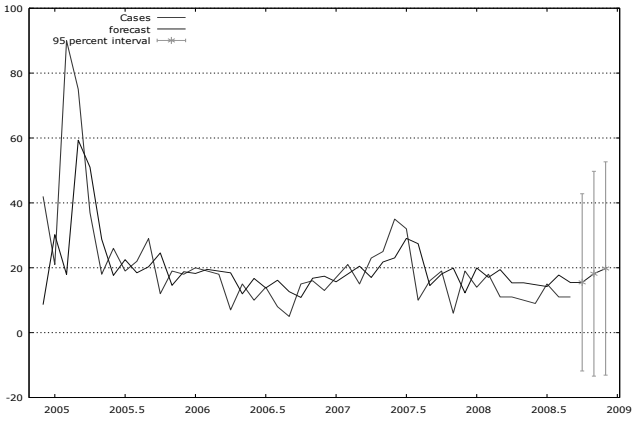
(Fig.4: Corellogram of residuals)

C.) Forecast

The results show in table-3 and Figure-5 indicate that the forecasting model selected is appropriate, therefore the model can be used for forecasting crime in India.

(Table-3: Forecast data)

For 95% confidence intervals, z(0.025) = 1.96				
Obs	Cases	prediction	std. err.	95% interval
2008:10	undefined	15. 50	13.93	(-11.81, 42.81)
2008:11	undefined	18.17	16.10	(-13.39, 49.73)
2008:12	undefined	19.78	16.77	(-13.09, 52.64)



(Fig.5: Plot of actual and forecast data)

IV. CONCLUSION

Crime forecasting is an interesting application area of research and ARIMA model offers a good technique for predicting the magnitude of any variable. Its strength lies in the fact that the method is suitable for any time series with any pattern of change and it does not require the forecaster to choose a priori the value of any parameter. Its limitations include its requirement of a long time series. In our study the developed model for crime forecasting was found to be ARIMA (1,1,1). The validity of the forecasted values was checked with the actual data for the lead periods and it is established that the ARIMA model can be used by researcher for short time forecasting of crime in India.

REFERENCES

[1] Ajoy K. Palit and Dobrivoje Popovic; *"Computational Intelligence in Time Series Forecasting Theory and Engineering Applications"*, 2005, Springer

[2] Akaike, H., 1974, *"A new look at the statistical model identification,"* IEEE Transactions on Automatic control, 19(6), 716-723.

[3] Alan Pankratz, *"Forecasting With Univariate Box- Jenkins Models, Concepts and cases"*, John Wily & Sons 1983.

[4] Bowerman, B. L., Connell, R. T., and Koehler, A B; *"Forecasting, Time Series, and Regression: An Applied Approach",* Thomson, Belmont, CA 2005.

[5] Box G E, Jenkinks G M; *"Time series analysis: Forecasting and control";* Holden Day, San Francisco

[6] C.B. Gupta, Vijay Gupta; *"An Introduction to Statistical Methods"*, Vikas Publishing House; 295-336.

[7] Dickey D A, Fuller W A; *"Distribution of the estimators for autoregressive time series with a unit root"*; Journal of the American Statistical Association 74, 427-431.

[8] Groff, Elizabeth R., and La Vigne, Nancy G., : "*Forecasting The Future Of Predictive Crime Mapping, Crime Prevention Studies",* vol. 13, pp. 29-57 (2002)

[9] Loganathan, Nanthakumar and Yahaya Ibrahim; *"Forecasting International Tourism Demand in Malaysia using Box Jenkins Sarima Application"*; South Asian Journal of Tourism and Heritage (2010), Vol. 3, Number 2

[10] Richard I Levin, David S Rubin; *"Statistics for Management"*; Pearson Prentice Hall, 861-919.

[11] S. Fan, L. Chen and W. J. Lee, *"Short-term load forecasting using comprehensive combination based on multimeteorological information,"* IEEE Trans. Industry Applications, Vol. 45, No. 4, July/Aug. 2009.

[12] Wipen Gorr, Richard Harries; *"Introduction to Crime forecasting"*; International Journal of Forecasting 19, 2003, 551-555.