ARIMA Modelling to Forecast and Analyze Indian Sectoral Stock Prices

JyothiManoj and Aloysius Edward J.
Kristu Jayanti College, K Narayanapura, Kothanur post
Bengaluru – 560077 Karnataka

Abstract

Time series forecasting is an active research area that has drawn considerable attention for application in a variety of areas. Auto Regressive Integrated Moving Average (ARIMA) models are one of the most important time series models used in financial forecasting over the past three decades. This paper attempts to address the forecasting of stock prices. The forecasting models ARIMAs are applied to forecast the stock prices. The results suggest that ARIMA (1, 1, 0) is the most suitable model to be used for forecasting stock prices. Closer examination suggests that the stock prices are upward trends and could be considered as a worthy investment.

Keywords: Forecasting, Stationary, Estimation, ARIMA, Time Series Modelling, Sectoral Stock Prices

Introduction

Analysis of financial time- series of stock prices has always been attractive for the purpose of forecasting. An efficient forecasting model will be of great significance for investors. The present study concentrates on building models which are sector- specific making use of a wide range of stock prices.Forecasting is a process in management to assist decision making. It is also described as the process of estimation in unknown future situations. In a more general term it is commonly known as prediction which refers to estimation of time series or longitudinal type data.

The most popular model for this method is the Box-Jenkins model introduced by [1]. Box-Jenkins has suggested the time-series autoregressive integrated moving average (ARIMA) model for forecasting. Like any other such methods, it requires historical time series data on the variable under forecasting. It assumes that the future values of a time series have a clear and definite functional relationship with current, past values and white noise. Kumar et al. [2] stated that the ARIMA offers a good technique for predicting the magnitude of any variables. The model has been successfully tested in many forecasting. In fishery industries, Lloret, et al. [3] suggests ARIMA models as the most appropriate to forecast fishery landings in the Hellenic marine waters, since systematic biological time- series data sets from explanatory variables are lacking. This methodology has been used to model and forecast the landings and catch per unit effort of many fish and invertebrate species. In financial forecasting, Fang [4] combines two methods to develop the fuzzy ARIMA model based upon the works of time-series ARIMA (p,d,q) model and fuzzy regression model. He uses the new method Fuzzy ARIMA to forecast the foreign market exchange and get the accurate forecasting value in a short time period. Edigerand Akar [5] used ARIMA model and seasonal ARIMA methods to forecast primary energy demand on fossil fuel in Turkey starts in year 2005 to year 2020. Wood and Dasgupta [6] used regression model, ARIMA's model and neural network model to forecast the MSCI, Capital Market Index of United Stated Amerika. They found that the ARIMA model which was built on the percentage changes in 3-period moving average is performing better than the ARIMA model build on the index itself.

The idea that Box and Jerkin's ARIMA model has predictability in many business activities including stock price is accepted in many researches in various countries. The entire data set used in the paper is divided into two; 70 percent of the data is used to develop the model and the remaining 30%is used to

check the accuracy in forecasting. The papers is arranged in the following order: - Past studies by other researchers in the related field is expressed in the next part of the paper as Review of Literature which is followed by the Methodology adopted in this study and the Results obtained with their Discussion; finally Conclusion and references are attached.

Review of literature

ARIMA model is widely used to analyses the impact of past values in predicting future.

Mohamad As'ad [7]in the study on forecasting daily demand for electricity, develops a model which will help understand how much past data must be used to forecast the peak demand on the days of the week. Four appropriate ARIMA (autoregressive integrated moving average) models based past three, six, nine and twelve months of data are considered. Using RMSE (root mean square error) and MAPE (mean absolute percentage error) to measure forecast accuracy, it is shown that the ARIMA model build based on past three months data is the best model in term of forecasting two to seven days ahead and ARIMA model based on past six months data is the best model to forecast one day ahead. The model is a seasonal ARIMA model that is found to be suitable.

Abdullah [8] in an attempt to find a model to address the forecasting of gold bullion coin selling prices. The forecasting models ARIMAs are applied to forecast the gold bullion coin prices. The data used is the daily selling prices of Malaysia's own gold bullion coins, KijangEmas Gold Bullion Coins for the year w2002- 2007. They have used the correlogram for parameter estimation. The diagnostic check is carried out by Ljung – Box Q statistic which checks if the residuals are white noise or not..The result suggests that ARIMA (2, 1, 2) is the most suitable model to be used for forecasting gold bullion coin prices.

Adebiyi et al [9] in their paper examines the forecasting the performance of ARIMA and artificial neural networks model with stock data from New York Stock Exchange. The paper presents the experimental results obtained by both ANN and ARIMA and suggests that ANN is equally or more efficient when compared to the traditional econometric models.

Raymond Y.C. Tse, (1997) suggested that the following two questions must be answered to identify the data series in a time series analysis: (1) whether the data are random; and (2) have any trends? This is followed by another three steps of model identification, parameter estimation and testing for model validity. If a series is random, the correlation between successive values in a time series is close to zero. If the observations of time series are statistically dependent on each another, then the ARIMA is appropriate for the time series analysis.(10)

Meyler et al (1998) drew a framework for ARIMA time series models for forecasting Irish inflation. In their research, they emphasized heavily on optimizing forecast performance while focusing more on minimizing out-of-sample forecast errors rather than maximizing in-sample 'goodness of fit'.

Stergiou (1989) in his research used ARIMA model technique on a 17 years' time series data (from 1964 to 1980 and 204 observations) of monthly catches of pilchard (Sardinapilchardus) from Greek waters for forecasting up to 12 months ahead and forecasts were compared with actual data for 1981 which was not used in the estimation of the parameters. The research found mean error as 14% suggesting that ARIMA procedure was capable of forecasting the complex dynamics of the Greek pilchard fishery, which,

otherwise, was difficult to predict because of the year-to-year changes in oceanographic and biological conditions. (11)

Contreras et al (2003) in their study, using ARIMA methodology, provided a method to predict next-day electricity prices both for spot markets and long-term contracts for mainland Spain and Californian markets. In fact a plethora of research studies is available to justify that a careful and precise selection of ARIMA model can be fitted to the time series data of single variable (with any kind of pattern in the series and with autocorrelations between the successive values in the time series) to forecast, with better accuracy, the future values in the series.

Methodology

An Auto regressive (AR) process is a series which is dependent on its own lagged values. The AR(p)model refers to the regression model where no repressors' other than the current and previous values of p-lags of the variable are involved.It may be represented as

$$Y_t = \alpha_{0+} \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p}$$

Moving average (MA) model is relevant if the AR process is not the only mechanism that generates Y , but it also involves the past values of the error terms. An MA (q) process represented as

$$\varepsilon_t = \beta_1 \varepsilon_t + \beta_2 \varepsilon_{t-2} + \beta_3 \varepsilon_{t-3} + \ldots + \beta_q \varepsilon_{t-q}$$

which is a linear combinations of white noise errors.

When Y has both the characteristics of AR and MA, we refer to it as ARMA(p, q) process. (12)

Theobjectiveof ARIMA model which are also known as Box-Jenkins model is to identify and estimate a statisticalmodel which can be interpretedashaving generated thesampledata.Hence stationarity is an important pre-requisiteMost of the financial time series are not stationary but integrated. Differencing the series will yield a stationary time series. If a series becomes stationary when differenced d times, we refer to the series as I(d). therefore , if we apply ARMA(p,q)to a series which is I(d), then the original time series is ARIMA(p, d, q).

The Box Jenkins methodology suggests finding the valuesof p and q for AR and MA respectively by referring to the correlogram. The autocorrelation function graph indicates the value of q while the Partialautocorrelation function graph indicates the valueof p. For an MA (q) model, moving average of order q, ACF Dies Down or Cuts off after lag q while for AR (p), autoregressive of order p PACF Dies Down or Cuts off after lag p. [10] This is further confirmed by least values of Akaike's Information criterion (AIC) value; the least value of AIC is considered most suitable.[13]
Model diagnosis can be carried out by the values of Root mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE). Further, the prediction accuracy is measured by an accuracy measure defined as

Accuracy percent = (1 – residual/actual series value)*100

where residual is the absolute difference between actual and estimated values.[14]

Result and Discussion

The study deals with the opening price of stocks from 6 sectors. The data is collected from websites of Bombay Stock Exchange Ltd., and National Stock Exchange Ltd.

Table 1: Sectors and the companies for each sector under study

Sector	No. of stocks	Companies
Banking	4	Axis Bank, SBI, ICICI, HDFC Bank
Automobiles	5	Bajaj Auto, Maruti Suzuki, Tata Motors, Hero Motorcorp, M&M
Health care	3	Sun Pharma, Dr Reddy's, Cipla
Power	2	NTPC, Tata Power
Oil & Gas	3	ONGC, Coal India, Gail India
IT	3	TCS, Wipro, Infosys

The descriptive statistics for each of the stock price is presented in Table 2. All the stocks except 2 (Coal India and Bajaj) are daily observations of the past 9 years (Feb 2007 – April 2015) with 1996 observations, while for Coal India 1076 observations of 5 years daily data (November 2010 – March 2015) and 1680 observations of Bajaj of past 8 years (May 2008 –March 2015). Among the sectors 'Banking' sector seems to have an overall highest average while the 'Oil& Gas' sector has the least overall average.

Table 2: Descriptive Statistics of each stock price.

Sector	Firm	N	Mean	SD	Skewness	Kurtosis	Jarque-Bera	Probability
Banking	Axis Bank	1996	1014.21	351.51	-0.002	2.61	12.21	0.002
	SBI	1966	1952.47	583.05	-0.579	3.75	155.94	0.000
	ICICI	1966	939.04	278.1	-0.109	3.53	26.78	0.002
	HDFC	1966	1167.89	587.9	0.681	2.26	196.19	0.000
Pharma	Cipla	1966	327.82	110.1	1.091	4.66	615.15	0.000
	Sun Pharma	1966	965.27	419.4	0.744	2.90	182.31	0.000
	Dr. Reddy's	1966	1514.92	792.8	0.5055	2.51	103.91	0.000
Automobiles	Bajaj	1680	1618.24	52.38	-0.566	2.88	90.77	0.00
	Maruthi Suzuki	1966	1375.59	650.1	1.7035	6.11	1741.04	0.000
	Tata Motors	1966	539.597	303.43	0.7954	2.78	210.97	0.000
	Hero	1966	1651.88	626.96	-0.040	2.53	18.27	0.000
	M & M	1966	793.455	231.8	0.3397	3.31	46.20	0.000
Power	NTPC	1966	173.811	29.022	0.5235	3.22	93.38	0.000
	Tata Power	1966	676.325	548.68	0.0569	1.24	255.44	0.000
Oil & Gas	ONGC	1966	665.141	395.67	0.3499	1.47	234.21	0.000
	Coal India	1076	333.883	37.817	-0.253	2.31	32.69	0.000
	Gail	1966	379.767	72.736	-0.416	2.91	57.66	0.000
IT	TCS	1966	1251.57	529.12	0.9414	2.98	290.41	0.000
	Wipro	1966	458.424	107.59	0.2024	3.06	13.68	0.000
	Infosys	1966	2469.20	675.24	0.02	2.43	26.96	0.000

The highest average price is found to be for Infosys with comparatively high standard deviation also while the lowest is found for NTPC at an average 173.81 with SD 29.02. All the series are asymmetric (skewness coefficient ≠ 0) and tend to have kurtosis other than normal, few are leptokurtic (kurtosis coefficient > 3) while few are platykurtic (kurtosis< 3). Jarque- Bera test with hypothesis of normality is rejected for all the series.

Test for Stationarity

Stationarity of the series is the pre- requisite of any time-series to develop any forecasting model. In this study we have used Augmented Dickey- Fuller test to test for stationarity. The results are displayed in table 3.

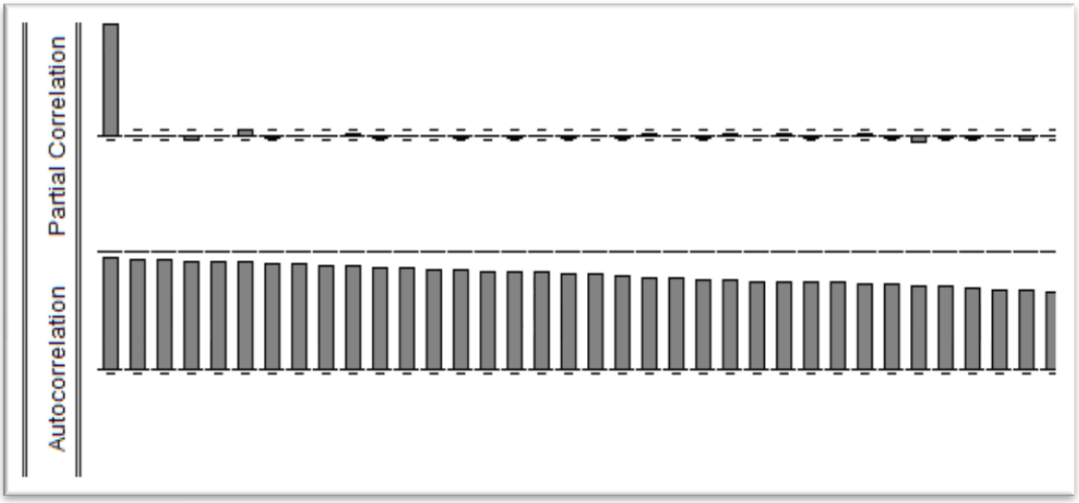
Table 3: ADF test result at level and first difference

ADF at level		t- Statistic	p-value	Conclusion	At level 1	p- value	Conclusion
Sector	Firm						
Banking	Axis Bank	-3.004	0.054	Not stationary	-44.667	0.0001	Stationary
	SBI	-2.531	0.108	Not stationary	-44.333	.0010	Stationary
	ICICI	-2.8711	0.069	Not stationary	-44.675	0.0001	Stationary
	HDFC	-2.077	0.254	Not stationary	-46.320	0.0001	Stationary
Health care	Sun Pharma	-2.584	0.096	Not stationary	-46.516	0.0001	Stationary
	Dr. Reddy's	0.7662	0.993	Not stationary	-43.618	0.0001	Stationary
	Cipla	1.245	0.995	Not stationary	-46.653	0.0001	Stationary
Automobiles	Bajaj	-1.833	0.365	Not stationary	-43.482	0.0001	Stationary
	Maruthi Suzuki	2.1526	0.9999	Not stationary	-44.490	0.0001	Stationary
	Tata Motors	-1.6368	0.4635	Not stationary	-46.494	0.0001	Stationary
	Hero	-1.1833	0.6838	Not stationary	-49.301	0.0001	Stationary
	Mahindra	-1.5271	0.5198	Not stationary	-48.329	0.0001	Stationary
Power	NTPC	-2.8866	0.0471	Not stationary	-51.118	0.0001	Stationary
	Tata Power	-1.0648	0.7315	Not stationary	-44.929	0.0001	Stationary
Oil & Gas	ONGC	-1.4797	0.5439	Not stationary	-46.399	0.0001	Stationary
	Coal India	-2.9584	0.0393	Not stationary	-35.743	0.000	Stationary
	Gail	-2.5498	0.1039	Not stationary	-35.694	0.000	Stationary
IT	TCS	0.7222	0.9926	Not stationary	-47.974	0.0001	Stationary
	Wipro	-1.7123	0.4249	Not stationary	-51.338	0.0001	Stationary
	Infosys	-2.3029	0.1711	Not stationary	-44.094	0.0001	Stationary

ADF test indicates that the series are stationary at first difference, i.e, the series are I(1).Once the stationarity is obtained, the parameters(p, q) od AR and MA models are located by inspecting the correlogram.

70 percent of the data is only selected for the construction of the model. For all the seriescorrelogram observed suggested, AR (1) to be best with no MA. This is confirmed by observing the AIC as the minimum with various combinations of (1, 1,0), (1,1,2),(2,1,0), (2,1,1),(2,1,2) . No other values of p and q were used as Box- Jenkins method recommends total number of parameters to be less than 3.E-views software is used for data analysis in this paper.

Figure 1: Correlogram of Closing price of Axis Bank



The correlograms of all the series more or less resembled Figure1.the figure suggests AR with lag 1 and no MA. Still the model accuracy was confirmed with different combinations of (1, 1,0), (1,1,2),(2,1,0), (2,1,1),(2,1,2)

The model is appropriateness is also confirmed by Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE). This is reassured by analysingthe percent of accuracy which is determined as Accuracy percent is (1 – residual/actual series value)*100.

Table 4: Parameter estimation for ARIMA model in the 70% of test data

Sector	Firm	p	d	α_1	AIC	RMSE	MAPE	Percent of Accuracy	Sector-wise % of Accuracy
Banking	Axis Bank	1	1	0.9991	10.17	45.50	2.26	80.03	82.23
	SBI	1	1	0.9993	11.54	77.52	1.99	78.23	
	ICICI	1	1	0.9993	10.22	10.02	2.17	88.62	
	HDFC	1	1	0.9990	10.83	54.57	1.72	80.32	
Health care	Sun Pharma	1	1	0.9989	10.61	48.76	1.64	85.95	88.56
	Dr. Reddy's	1	1	0.9989	9.46	27.44	1.44	89.56	
	Cipla	1	1	0.9988	6.53	6.35	1.48	90.10	
Automobiles	Bajaj	1	1	0.9992	10.61	48.78	1.81	90.54	93.21
	Maruthi Suzuki	1	1	0.9983	9.58	29.23	1.78	89.66	
	Tata Motors	1	1	0.9994	9.01	21.85	2.25	95.98	
	Hero	1	1	0.9993	10.03	36.45	1.64	90.33	
	Mahindra	1	1	0.9992	9.09	22.8	1.99	96.05	
Power	NTPC	1	1	0.9996	5.74	4.27	1.54	96.15	95.93
	Tata Power	1	1	0.9994	9.74	3.50	2.02	94.33	
Oil&Gas	ONGC	1	1	0.9999	9.48	27.81	1.79	84.32	89.53
	Coal India	1	1	0.9995	6.67	6.79	1.44	92.31	
	Gail	1	1	0.9995	7.32	9.39	1.73	90.23	
IT	TCS	1	1	0.9989	9.51	28.10	1.78	83.64	84.95
	Wipro	1	1	0.9995	7.79	11.93	1.69	85.63	
	Infosys	1	1	0.9995	11.29	68.67	1.48	82.90	

The coefficient of the AR model implies a slow convergence of the series. Moreover it is interesting to notice that the prediction to a great extent can be dependent on only the variable value with unit lag and is not significantly influenced by the error terms. This mentions the immediate future prediction power the values of the stock prices.

The prediction accuracy is checked for the remaining 30% of test data; the result displayed in Table 5. The results are confirming the appropriateness of the ARIMA models developed.

Table 5: Result of accuracy check on the 30% of test data.

Sector	Firm	RMSE	Percent of Accuracy	Sector-wise % of Accuracy
Banking	Axis Bank	44.72	82.23	85.71
	SBI	78.58	76.55	
	ICICI	10.33	90.28	
	HDFC	65.47	86.55	
Health care	Sun Pharma	54.56	88.78	91.01
	Dr. Reddy's	27.01	92.45	
	Cipla	10.32	91.25	
Automobiles	Bajaj	45.75	93.25	90.58
	Maruthi Suzuki	30.23	87.2	
	Tata Motors	26.85	89.74	
	Hero	40.45	92.84	
	Mahindra	25.85	94.54	
Power	NTPC	10.37	97.46	96.02
	Tata Power	3.87	95.66	
Oil&Gas	ONGC	29.48	89.23	94.49
	Coal India	4.79	97.45	
	Gail	7.95	94.19	
IT	TCS	26.50	84.75	85.15
	Wipro	10.73	85.81	
	Infosys	60.86	84.65	

Conclusion

ARIMA model is developed on the stock prices of 6 sectors viz. Banking, Healthcare, Automobiles, Power, Oil & Gas and IT. A series of stock prices are used to analyses the sectors. The data is partitioned into two- 70% of observations were utilized for model development while the remaining 30% for confirmation of the accuracy of the model developed. The developed models all have common characteristic that they are all integrated at first order and are Autoregressive models with lag 1 having no MA characteristics. The prediction accuracy is also highly acceptable (> 75% accuracy). Since the series are highly correlated to the immediate past values forecast accuracy will be more. This phenomenon uniformly across various sectors of the stock prices can be made use of for prediction and investment decisions.

References

1. G.E.P., and G. M. Jenkins (1970). *Time series analysis: forecasting and control*, Holden Day, San Francisco.
2. M., Kumar, A. Kumara, N.C., Mallik, C. and R.K. Shuklaa.(2009).“Surface flux modelling using ARIMA technique in humid subtropical monsoon area”, *Journal of Atmospheric and Solar-Terrestrial Physics*, Vol.71, pp. 1293-1298.
3. J.Lloret, ,J.Lleonart, , I.Sole.(2000).“ Time series modeling in landings in North Mediterranean sea”, *ICES Journal of Marine Science*, Vol. 57, pp.171 -184.

4. F.M. Tseng, G.H. Tzeng, H. C. Yu, J.C. Yuan.(2001). "Fuzzy ARIMA model for forecasting the foreign exchange market", *Fuzzy Sets and Systems,* 118, 1-11.
5. S.A Ediger (2006).ARIMA Forecasting of Primary Energy Demand By Fuel in Turkey, *Energy Policy*, Vol. 35, pp.1-8, 206.
6. D. Wood, and B.Dasgupta.(1996)."Classifying Trend movements in the MSCI USA capital market index – A Comparison of Regressions, ARIMA And Neural Network Method", *Computers & Operators Research*, Vol.23, pp. 611-622.
7. Mohamad As'ad. (2012). Finding the best ARIMA model to forecast peak electricity demand; Applied Statistics Education and Research collaboration- Conference paper. Research online; http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1011
8. Lazim Abdullah. (2012). ARIMA model for Gold Bullion Coin Selling Prices Forecasting; International Journal of Advances in Applied Sciences (IJAAS) Vol. 1, No. 4, December 2012, pp. 153~158 ISSN: 2252-8814
9. William H Green (2008).Econometric Analysis; Pearson Education
10. Raymond Y.C. Tse.(1997). An application of the ARIMA model to real-estate prices in Hong Kong, *Journal of Property Finance*, Vol. 8, No. 2, pp.152 – 163.
11. Stergiou, K. I.(1989). Modeling and forecasting the fishery for pilchard (Sardinapilchardus) in Greek waters using ARIMA time-series models, *ICES Journal of Marine Science*, Volume 46, No. 1, pp. 16-23.
12. Gujarathi, Porter, Gunasekar. (2012). Basic Econometrics, McGraw Hill Pvt. Ltd.
13. David Raymond Anderson. (2008). Model based inference in the life sciences: a primer on evidence, New York, Springer.
14. Mondal, Shit, Goswami. (2014).Studyof effectiveness of time seriesmodelling (ARIMA)in forecasting stockprices; *International Journal of Computer Science, Engineering and Applications (IJCSEA)*Vol.4, No.2.