Kalman Filter in Iron Ore Prediction and Parameter Estimation

School of Agricultural, Computational and Environmental Sciences

A project submitted by

John Worrall

Supervisor:

Dr Enamul Kabir

Master of Science in Mathematics and Statistics

June 2019

1. Introduction

Forecasting the commodity value is a challenging process due to the stochastic non-linear nature of the financial market. While there have been a number studies on statistical and filtering techniques for stock market prediction, this paper improves on the current literature by investigating the novel development of Kalman filters techniques applied to iron ore prices.

The forecasting capabilities of Kalman filters are studied and compared to traditional Autoregressive Integrated Moving Average (ARIMA) models. The comparison was conducted through the efficient modelling and forecasting of monthly commodity prices of iron ore over 30 years.

The study signifies the important role of parameter estimates obtained in both ARIMA and the Kalman process when devising optimal approaches to forecasting long term iron ore prices. Results of the study will show the best performing models obtained by comparison of statistical performance metrics and the potential in developing automatic prediction system for commodity prices, patterns and volatility in the economic sector for applications such as risk management, asset pricing and allocation.

1.2 Background

Stock market forecasting by means of times series analysis and modelling of related variables is of primary importance in the economic and financial sectors. The prediction of stock prices is useful for investors and traders, economists and policy makers in terms of allocating investment correctly leading to less wasted resources and a more prosperous economy (Rigobon & Sack, 2003). Developing and evaluating these predictive models is essential in

macroeconomic strategies, construction of portfolio investments and risk management. Further analysis of time series may reveal the critical characteristics in behavioural and temporal patterns of the financial market.

After oil, iron ore is one of the most important commodities although given relatively little attention. A necessary input for the production of primary steels, iron ore feeds the world's largest metal market and is the backbone for global infrastructure (Lundmark & Nilsson, 2003), estimated at a trillion dollars a year. Western Australia is the world's largest iron ore producer, generating 33 percent of global production in 2017. Iron ore industry is a critical part of the national economy and is of importance when future planning and development (Morandi et al, 2014). Yet iron ore of the commodities has received comparatively little research in time series and predicting techniques.

Statistical techniques are frequently used in stock market prediction. Milosevic (2016) demonstrated how algorithms enables analysts to predict the movement of stocks as well as the ratio of movement over a fixed amount of time. However, to date no known study has been done applying Kalman filters for monthly iron ore price predictions.

1.3 Aim

The project aims are to fill the identified gap by investigating a novel Kalman Filtering (KL) technique for next month prediction of iron ore prices (USD). This research has the following objective:

1) Develop a Kalman Filter model and evaluate performance accuracy against ARIMA model using known performance metric techniques.

The results of the prediction models can determine whether Iron Ore prices can be effectively forecasted and whether the model can be applied in a broader financial context. The frameworks and overall findings can be useful to investors and the larger community, adding to current research.

2. Literature Review

Researchers have faced several difficulties when trying to model complex systems such as the stock market due to its nature. Described as non-linear, chaotic, highly dimensional (Schmidt, 2011) and a complex dynamic system, prediction is challenging. Models that approximate non-stationary financial time series may include noisy error prone features. The relationship between the model input and output is essentially non-linear (Xiao et al., 2012), where stock prices include variables of higher degrees, adding to the complexity of modelling and predicting the stock market.

Kalman filters are effective predictive models, which fuses information of technical and fundamental analysts to forecast price in stock exchange markets (Yan & Guosheng, 2015). The high fluctuations and time varying characteristics of the stock market, suits the Kalman filter dynamic and real time updating characteristics. However, the challenges in Kalman filters are in initiate set of parameters and estimates. Paper (Elliott & Hyndman, 2007) investigate filter-based EM algorithm over standard-based EM algorithm for parameters estimates with Kalman filter, identifying future works in application of Kalman filters models.

3. Methodology

For the project, we split the dataset into a training series for model build and parameter estimation and a testing series for forecast models ARIMA and KF comparison.

3.1 Auto-Regressive Moving Average model (ARIMA)

A commonly used numerical forecasting model for market time series, the ARIMA model systematically characterises past, current and future trends in the data. Generally, the data model consists of an autoregressive (AR), integrated (I) and a moving average (MA) component of varying values, which may be identified in seasonal and/or non-seasonal occurrences (Box et al., 2015). The general form of ARIMA is as follows:

$ARIMA\ (p, d, q) -$ Autoregressive, AR(p) relates to the current value of the time series to past values of order p. Moving average, MA (q) relates to the past forecast errors of order q and differencing (d), and adjusts for non-stationarity.

Under the linear ARIMA models, it is required that the time series be free of any deterministic structures such as level shifts, local time trends and seasonal pulses (Harvey et al., 1999). Assumptions of ARIMA models are that the series have constant error variance and that the parameters of the proposed model remain constant over the course of time.

More appropriate ARIMA models may be determined through the three stages of model fitting: identification estimation and diagnostic check (Box et al., 2015). Stage one, identification, includes the selection of a set of more appropriate models through an examination of the ACF and PACF distribution of the time series. The recommendation or most optimally performing model is based on the combined testing for the minimum value of the Akaike Information Criterion (AIC), defined (Akaike, 1974) as follows:

$$AIC = -2\mathrm{k} - 2\ln(\hat{\mathrm{L}})$$

Eq.1 Akaike information criterion, where k is the number of parameters and $\hat{\mathrm{L}}$ is the maximum value of likelihood function.

3.2 Kalman Filter (KF)

Under the assumption that our data is in the form of a Gaussian distribution we will apply an iterate measurement update and motion prediction through linear equations. Kalman filters can combine measurements from one state and system dynamics to give better estimates of both the unmeasured and measured states.

The first step obtains apriori estimates of probability of states, then Bayes rules updates to calculate a better estimate of probability. Bayes rules are describe below,

Consider two normal probability distributions given by,

$$f(X|\mu_1,{\sigma_1}^2) = \frac{1}{\sqrt{2\pi{\sigma_1}^2}} e^{-\frac{(X-\mu_1)^2}{2{\sigma_1}^2}}$$
$$f(X|\mu_2,{\sigma_2}^2) = \frac{1}{\sqrt{2\pi{\sigma_2}^2}} e^{-\frac{(X-\mu_2)^2}{2{\sigma_2}^2}}$$

The Bayesian probability is therefore given by,

$$f\big((X|\mu_1,{\sigma_1}^2)\big|(X|\mu_2,{\sigma_2}^2)\big) = \frac{f(X|\mu_1,{\sigma_1}^2)f(X|\mu_2,{\sigma_2}^2)}{\int_{-\infty}^{\infty} f(X|\mu_1,{\sigma_1}^2)f(X|\mu_2,{\sigma_2}^2)dX}$$

This step gives a better estimate of price position and velocity, even though are only interested in the estimate of price alone.

$$x = \begin{bmatrix} p \\ v \end{bmatrix}, \qquad mean\ stat\ vector, p = Distance\ v = Velocity$$

The final form of the Kalman filter state propagation (prediction) and measurement update with dynamic system in the following form (Kalman, 1960),

$$x(t+1) = \emptyset x(t) + \Gamma w(t) \quad (1)$$
$$y(t) = Hx(t) + v(t) \quad (2)$$

Where $\emptyset = state\ transition\ matrix$, H = measurement matrix, $w(t) = model\ noise$
$v(t) = measurement\ noise, y(t) = measurement\ vector, x(t) = state\ vector$

Lemma,

$$\hat{x}(t+1|t+1) = \hat{x}(t+1|t+1) + K(t+1)\varepsilon(t+1) \quad (3)$$

$$\hat{x}(t+1|t) = \emptyset\hat{x}(t|t) \quad (4)$$

$$\varepsilon(t+1) = y(t+1) - H\hat{x}(t+1|t) \quad (5)$$

$$K(t+1) = P(t+1|t)H^T[HP(t+1|t)H^T + R]^{-1} \quad (6)$$

$$P(t+1|t) = \emptyset P(t|t)\emptyset^T \Gamma \emptyset \Gamma^T \quad (7)$$

$$P(t+1|t+1) = [I - K(t+1)H]P(t+1|t) \quad (8)$$

where initial values defined

$$\hat{x}(0|0) = Ex(0) = \mu_0$$

$$P(0|0) = E[(x(0) - \mu_0)(x(0) - \mu_0)^T] = P_0 \quad (9)$$

Recursive chain from by equation $3 - 9$.

3.5 Performances Measures

To evaluating the performance of the non-linear and linear based models the following mathematical assessment metrics are recommend (Dawson et al., 2007; Deo et al., 2016; Legates and McCabe, 1999; Willmott, 1981; Willmott, 1982; Willmott, 1984).

Where Sim *is simulated and* Obs *are observed values.*

Coefficient of Determination $(\boldsymbol{R^2})$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(Obs_i - \overline{Obs})\,(Sim_i - \overline{Sim})}{\sqrt{\sum_{i=1}^{n}(Obs_i - \overline{Obs})^2}\sqrt{\sum_{i=1}^{n}(Sim_i - \overline{Sim})^2}} \right)^2$$

The R^2 can be interpreted as the ratio of variation of the observed data explained by the simulated values. Although widely used, the measurement is overly sensitive to high values and insensitive to new values of proportional differences.

Root Mean Square Error $(\boldsymbol{RMSE})$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Sim_i - Obs_i)^2}$$

When explaining the goodness of fit the RMSE can be utilised. RMSE is more sensitive than mean absolute error (MAE), where extreme values are involved.

Relative Root Mean Square Error ($\boldsymbol{RRMSE}$)

$$RRMSE = 100 \times \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{n}(Sim_i - Obs_i)^2}}{\frac{1}{N}\sum_{i=1}^{n}(Obs_i)}$$

RRMSE is expressed relative to the variability of measurements of the goodness of fit. Values are expressed in percentages as small as possible to indicate small deviations between observed and predicted data.

Mean Absolute Error ($\boldsymbol{MAE}$)

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|Sim_i - Obs_i|$$

Not weighting the higher or lower values, MAE evaluates all deviations from observed data. Like $RMSE$, smaller values indicate less deviations.

Mean Absolute Percentage Error ($\boldsymbol{MAPE}$)

$$MAPE = 100 \times \frac{1}{N}\sum_{i=1}^{n}\left|\frac{(Sim_i - Obs_i)}{Obs_i}\right|$$

MAPE are error percentages summarised without regard to the sign.

For the purpose of modelling results, particularly where errors are concerned, both RMSE and MAE are ideal in that they calculate the aggregation of residuals of both observed and simulated data. The weaknesses of both metrics are that they are expressed in their absolute units. When relative errors are required, RRMSE and MAPE can both be explored.

3.6 Dataset

The study uses monthly average prices of iron ore (US dollars per metric ton) from January 1980 to March 2019 that have been sourced from Yahoo finance API (62% Fe, CFR). A total

of 472 observations will be divided into two parts one training set consisting of 459 records and the remaining 12 records (1 year) are used for testing set.

Index	Min	Max	Median	1st Quartile	3rd Quartile	Mean	Std. Dev	Variance
Iron Ore	10.51	187.18	13.82	12.15	58.8	38.38	43.2051	1866.681

Table. 3.2 Iron ore price statistics

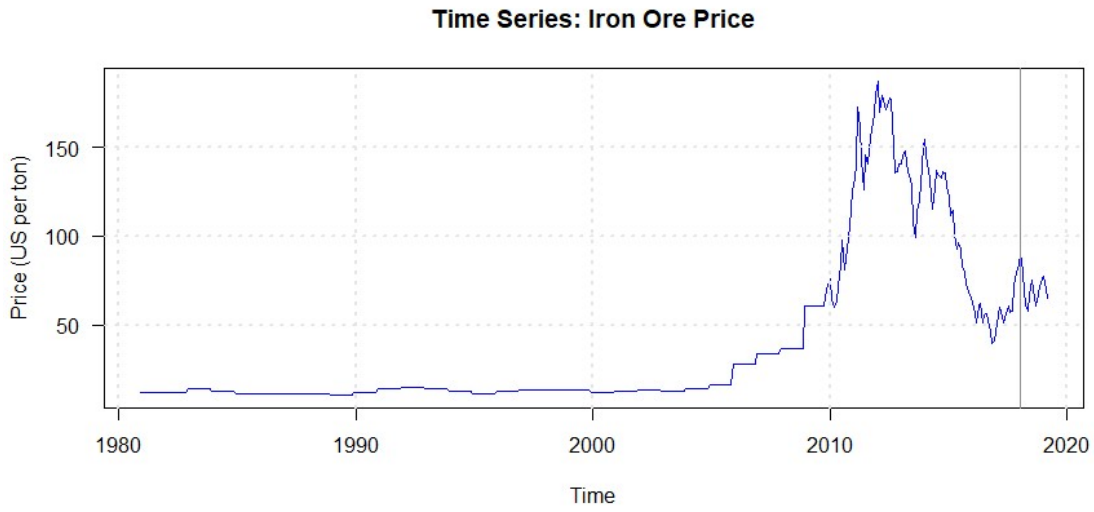

Figure. 1 Price returns of iron ore (January 1980 – March 2019), grey line indicate test and train split.

4. Results

The model results of ARIMA and Kalman development for Iron Ore price forecast will be determined and discussed. First the parameters results configured to training set for both ARIMA and Kalman, followed by the comparison of models predictions with the testing set.

4.1 ARIMA parameters

Prior to building a suitable model, figure 1(plotted time series) is examined and investigated for fitted parameters. To determine trends and reasonable stationarity we take the first difference (figure 4.2) d =1, where d is time differencing in making time series stationary. The PACF (figure 4.1, left) plots the iron ore prices correlation of time x(t) and next month,

there is a significant spike at q indicating a AR(1) or p =1. The slow decay of ACF (figure 4.1, right) further suggesting that the process requires a differencing.

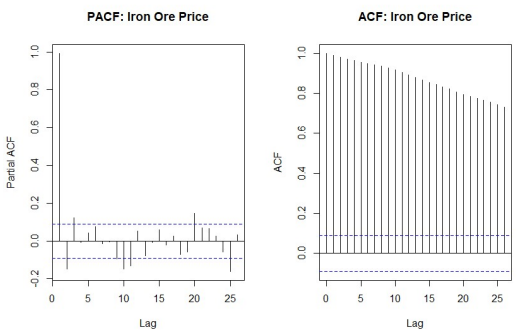

Figure. 4.1 PCAF and ACF of Iron Ore value

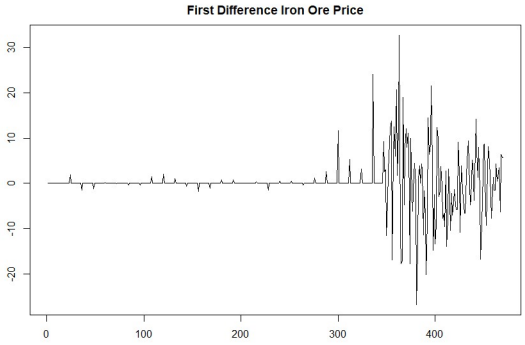

Figure. 4.2 First differences of Iron Ore prices

In order to generate the ARIMA models associated statistic (e.g AIC, loglikelihood) and predictions with libraries 'forecast' under R, version 3.3.3 was used. The results in the Table 4.1 and Table 4.2 illustrate the prediction of the top ten ARIMA and eight seasonal ARIMA rank predicted models. The results for this analysis demonstrate that the best model (lowest AIC) for monthly prediction is the ARIMA (1, 1, 1), and for seasonal the optimal model is ARIMA(1,1,1) (0,0,2) [12].

Figure 4.3 illustrates the model fitting against actual training values, with error histogram and scatter plot both showing a closely fitted model build.

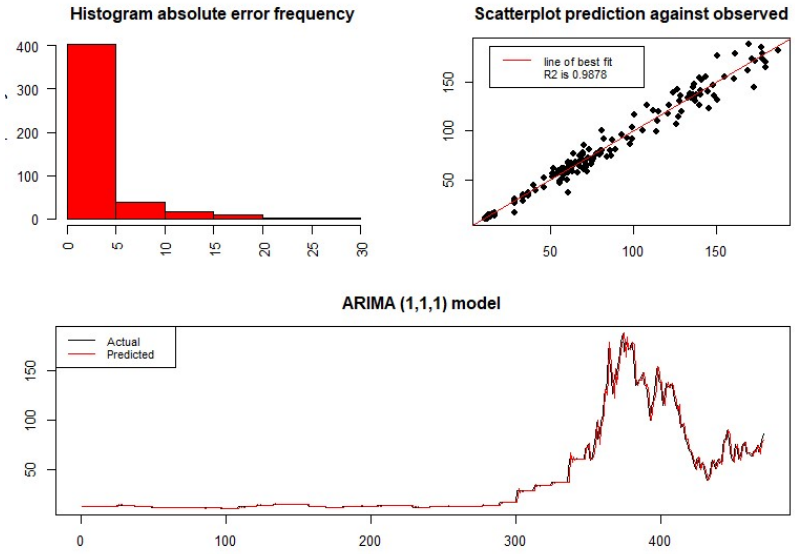

Figure. 4.3 ARIMA model with parameter fit

	Type of ARIMA	Statistics			Parameters					
Model	Structure (p,d,q)	AIC	Log likelihood	Sigma^2	AR1	AR2	AR3	MA1	MA2	MA3
1	ARIMA(0,1,0)	2839.24	-1418.62	24.5						
2	ARIMA(1,1,0)	2821.14	-1408.57	23.48	0.2047					
3	ARIMA(0,1,1)	2813.62	-1404.81	23.1:				0.2796		
4	ARIMA(2,1,0)	2812.07	-1403.03	22.93	0.2356	-0.1526				
5	ARIMA(0,1,2)	2809.64	-1401.82	22.81				0.2449	-0.1124	
6	ARIMA(1,1,1)	**2810.76**	-1402.38	22.86	-0.2832			0.5398		
7	ARIMA(0,1,3)	2811.48	-1401.74	22.8				0.2461	-0.1213	-0.0189
8	ARIMA(3,1,0)	2813.01	-1402.5	22.87	0.2430	-0.1636	0.0475			
9	ARIMA(2,1,1)	2811.7	-1401.85	22.81	-0.1671	-0.0706		0.4148		
10	ARIMA(1,1,2)	2811.25	-1401.62	22.79	0.3199			-0.0722	-0.2020	

Table 4.1 Ten best performing ARMIA monthly models with characteristics

	Type of ARIMA Seasonal	Statistics			Parameters					
Model	Structure (p,d,q)(P,D,Q)(freq)	AIC	Log likelihood	Sigma^2	AR1	sAR1	sAR2	MA1	sMA1	sMA2
1	ARIMA(1,1,1)(0,0,1)[12]	2817.62	-1404.81	23.1	-0.2805			0.2801	0.2801	
2	ARIMA(1,1,1)(0,0,2)[12]	**2811.25**	-1401.62	22.79	0.3199				-0.0722	-0.2020
3	ARIMA(1,1,1)(1,0,0)[12]	2812.3	-1402.15	22.84	-0.1363	-0.1363		0.5231		
4	ARIMA(1,1,1)(1,0,1)[12]	2814.62	-1402.31	22.86	-0.3544	-0.3544		0.4832	0.4832	
5	ARIMA(1,1,1)(1,0,2)[12]	2815.25	-1401.62	22.79	0.3187	0.3187		-0.3197	-0.0702	-0.2016
6	ARIMA(1,1,1)(2,0,0)[12]	2813.69	-1401.85	22.81	0.0637	-0.2149	-0.0876	0.3991	-0.0702	-0.2016
7	ARIMA(1,1,1)(2,0,1)[12]	2815.61	-1401.8	22.81	0.0197	-0.3323	-0.1185	0.2805	0.2805	
8	ARIMA(1,1,1)(2,0,2)[12]	2816.93	-1401.47	22.77	0.3416	0.1643	0.0753	-0.1921	-0.0683	-0.2737

Table 4.2 Eight best performing ARMIA Seasonal models with characteristic

4.2 Kalman filter parameters

For the Bayesian inference process of Kalman filters, the parameters were optimised with Maximum Likelihood Estimation (MLE) method. R package 'DLM' was employed for model build of state spaces and forecasting. Initial parameters (1,1,1,1) were set and with MLE method converging to the following matrix parameters:

F_t	G_t	V_t	W_t
-7.057784e-07	2.151007e+01	-1.428634e-02	-7.057784e-07

Where dynamics linear model state errors and observed components are normally distributed, parameters defined

$$x_0 \sim N(m_0 C_0)$$
$$x_t | x_{t-1} \sim N(G_t x_{t-1}, W_t)$$
$$y_t | x_t \sim N(F_t x_t, V_t)$$

Where y_t is observed data and x_t is unobserved data.

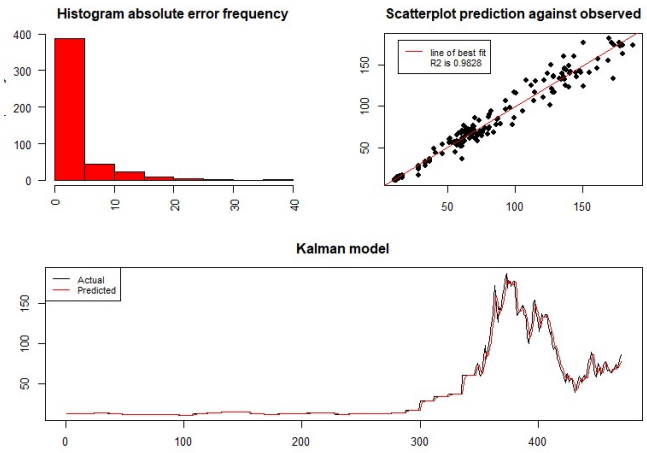

Figure. 4.4 ARIMA model with parameter fit

Training Performance Matrix	Models	
	ARIMA (1,1,1)	**Kalman Filter**
r	0.9878	0.9828
RMSE	4.7763	5.6666
rRMSE	12.446	14.7448
MAE	2.0280	2.4480
MAPE	2.8191	3.3122

Table 4.3 Performance ARIMA versus Kalman Filter parameters fitting.

Figure 4.4 illustrates the model fitting against actual training values. From error histogram there is a high frequency of lower values and the scatter plot shows good correlation with only slightly sparse points at the larger values. Parameters for ARIMA model show a slightly better fitting model.

4.3 ARIMA against Kalman Filter model forecast

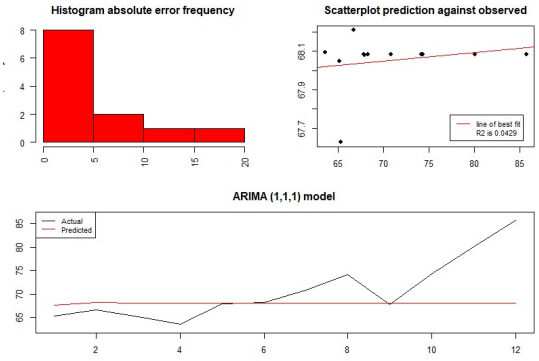

Figure. 4.5 ARIMA model forecasting performance

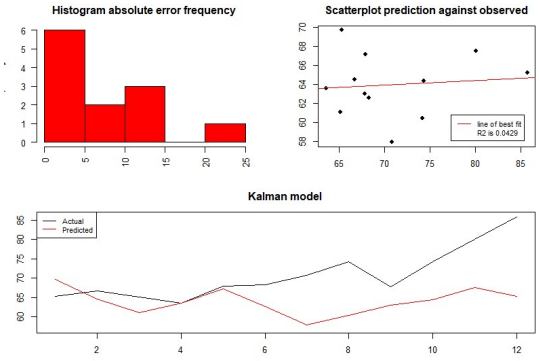

Figure. 4.6 Kalman model forecasting performance

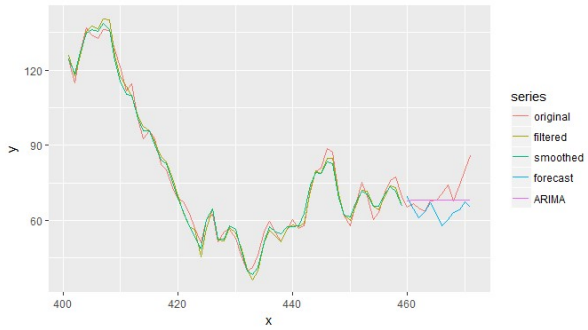

Figure. 4.7 ARIMA & KF forecasting plot

Testing Performance Matrix	Models	
	ARIMA (1,1,1)	**Kalman filter**
r	0.042854	0.009436
RMSE	6.928314	9.705071
rRMSE	9.784488	13.70595
MAE	4.722669	7.626441
MAPE	6.205535	10.20503

Table 4.4 Performance ARIMA versus Kalman Filter forecasting

This section presents the analyses of the ARIMA and Kalman filter applied to testing set. The ARIMA models outperforms the Kalman filter as seen Table 4.4, all metrics indicate better performance on the ARIMA (1,1,1) model. However, on the timeseries plot (Figure 4.5) we can see prediction of ARIMA converges to a value, while the Kalman shows similar movement (Figure 4.6) albeit offset.

5 Conclusion

Overall, the study highlights the appropriateness of the ARIMA and Kalman approaches to modelling for forecasting a year of iron ore prices. Although findings showed traditional methods of ARIMA outperforming Kalman, it did highlight the advantages of the Kalman in modelling the change in future prices. Furthermore, this paper provides a baseline relevant to ARIMA vs Kalman iron ore forecasting, which have not yet studied. The study illustrates models ability to forecast commodities, providing insightful information on the phenomena of market events for investors, policy makers and other stakeholders to make more informed and profitable decisions.

5.1 Expected Outcomes and Future Works

The outcomes of this project may lead to the enhancement of financial modelling to attain more accurate forecasting capability of models. Specifically, from parameter estimation implementation of Kalman filter vs ARIMA on iron ore market price prediction.
While the results of this study provides interesting insights the recommended future research could include the implementing improved algorithms in Kalman filter such as extended, unscented and ensemble Kalman on other market data such as the indexes and other commodities prices.

References

Akaike, Hirotugu. "A new look at the statistical model identification." IEEE transactions on automatic control 19.6 (1974): 716-723.

Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Elliott, R.J. and Hyndman, C.B., 2007. Parameter estimation in commodity markets: A filtering approach. Journal of Economic Dynamics and Control, 31(7), pp.2350-2373.

Fang, Z., Luo, G., Fei, F. and Li, S., 2010, October. Stock forecast method based on wavelet modulus maxima and kalman filter. In 2010 International Conference on Management of e-Commerce and e-Government (pp. 50-53). IEEE.

Gultekin, S. and Paisley, J., 2017. Nonlinear Kalman filtering with divergence minimization. IEEE Transactions on Signal Processing, 65(23), pp.6319-6331.

Harvey, A., Jan Koopman, S. and Penzer, J., 1999. Messy time series. In Messy Data (pp. 103-143). Emerald Group Publishing Limited.

Javaheri, A., Lautier, D. and Galli, A., 2003. Filtering in finance. Wilmott, 3, pp.67-83.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of basic Engineering, 82(1), pp.35-45.

Kuzin, D., Yang, L., Isupova, O. and Mihaylova, L., 2018, July. Ensemble Kalman Filtering for Online Gaussian Process Regression and Learning. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 39-46). IEEE.

Lautier*, D. and Galli, A., 2004. Simple and extended Kalman filters: an application to term structures of commodity prices. Applied Financial Economics, 14(13), pp.963-973.

Li, J., 2013. An unscented Kalman smoother for volatility extraction: Evidence from stock prices and options. Computational Statistics & Data Analysis, 58, pp.15-26.

Liao, L., 2004. Stock option pricing using Bayes filters. tech. report.

Lundmark, R. and Nilsson, M., 2003. What do economic simulations tell us? Recent mergers in the iron ore industry. Resources Policy, 29(3-4), pp.111-118.

Malanichev, A.G. and Vorobyev, P.V., 2011. Forecast of global steel prices. Studies on Russian Economic Development, 22(3), p.304.

McMillan, D.G., 2001. Nonlinear predictability of stock market returns: Evidence from nonparametric and threshold models. International Review of Economics & Finance, 10(4), pp.353-368.

Morandi, M.I.W.M., Rodrigues, L.H., Lacerda, D.P. and Pergher, I., 2014. Foreseeing iron ore prices using system thinking and scenario planning. Systemic Practice and Action Research, 27(3), pp.287-306

Rigobon, R. and Sack, B., 2003. Measuring the reaction of monetary policy to the stock market. The quarterly journal of Economics, 118(2), pp.639-669.

Schmidt, A.B., 2011. Financial markets and trading: an introduction to market microstructure and trading strategies(Vol. 637). John Wiley & Sons.

Wu, C.J., 1983. On the convergence properties of the EM algorithm. The Annals of statistics, 11(1), pp.95-103.

Xiao, Y., Xiao, J. and Wang, S., 2012. A hybrid forecasting model for non-stationary time series: An application to container throughput prediction. International Journal of Knowledge and Systems Science (IJKSS), 3(2), pp.67-82.

Yan, X. and Guosheng, Z., 2015, June. Application of kalman filter in the prediction of stock price. In 5th International Symposium on Knowledge Acquisition and Modeling (KAM 2015). Atlantis Press.