European Journal of Business and Management Research
Vol 9 | Issue 4 | July 2024
ISSN 2507-1076

RESEARCH ARTICLE

Modeling Nairobi Residential Real Estate Prices using ARIMA

Dan Chirchir*, Mirie Mwangi, and Cyrus Iraya

ABSTRACT

The residential real estate market is big and affords investors investment opportunities. The price changes are key in determining the overall return. Structural and atheoretical models are the two main approaches to modeling real estate prices. Structural models link prices to fundamental factors such as economic indicators and property supply, amongst others. Atheoretical models attempt to predict prices by leveraging on the statistical properties of time series data and may be extended to augment fundamental factors. This study focused on time series modeling using ARIMA. The objective of the paper was to identify a suitable ARIMA model that is efficient in predicting house prices in Nairobi. The training data was for the period 2010Q3 to 2019Q2. The out of sample test data was for six quarters: 2019Q3 to 2020Q4. The Box-Jenkins methodology was adopted. Seven ARIMA models and six AR models were identified, estimated, and used in predicting prices using out of sample data. The study found out that AR models outperformed ARIMA models. The paper contributes to knowledge being among the first to apply ARIMA in Nairobi house market using hedonic house prices. The paper may inform investment strategy and portfolio management by investors. It may inform policy since house price forecasts may have social and economic effects.

Keywords: ARIMA, Box-Jenkins Methodology, MAE.

Submitted: October 22, 2023

Published: July 19, 2024

10.24018/ejbmr.2024.9.4.2201

Department of Financ and Accounting, Faculty of Business and Management Sciences, University of Nairobi, Kenya.

Corresponding Author:
e-mail: dchirchir@uonbi.ac.ke

1. Introduction

The residential real estate market is big and affords investors investment opportunities. The price changes are key in determining the overall return. Structural and atheoretical models are the two main approaches to modeling real estate prices. Structural models link prices to fundamental factors such as economic indicators and property supply, amongst others. Such factors may include fundamental economic variables such as economic growth and household incomes. Property characteristics such as the size of the house, location, age, amenities available, etc., may also drive house prices (Brown, 1997). Besides, property supply factors such as approved building plans, new units completed, and financing, amongst others, could also determine house prices. The house prices may also be driven by rent income and land price. Atheoretical models attempt to predict prices by leveraging the statistical properties of time series data, which may be extended to augment fundamental factors (Crawford & Fratantoni, 2003). This study focused on time series modeling using Autoregressive Integrated Moving Average (ARIMA).

The Nairobi residential real estate market has been on a growth trajectory. The government of Kenya has been spearheading the affordable housing project aimed at increasing home ownership. Investors and other players are interested in understanding the evolution of house prices in Nairobi. Atheoretical models have not been studied extensively in the Nairobi market partly due to a lack of suitable data. Crawford and Fratantoni (2003) underscore the efficiency of atheoretical models in determining and forecasting housing prices. The findings from atheoretical studies are varied. The basic premise for these models is the understanding of time series properties of the real estate price indices. The best model is the one that fits the estimate well and is capable of out-of-sample forecasts with minimal errors relative to others. Several studies have found that the ARIMA models outperform other models (Al-Marwani, 2014; Hepsen & Vatansever, 2010; Sklarz *et al*., 1987; Vishwakarma, 2013). Gupta and Das (2010) in their study contend that Bayesian Vector Auto-Regressive Models (BVAR) are efficient compared to the ARIMA models. Other studies found ARIMA models wanting because they do not consider the presence of structural

breaks or even states in the time series (Crawford & Fratantoni, 2003; Barari *et al*., 2014). There is no consensus as to the most suitable time series model since this may vary across markets and property types. There is still room to study this in different contexts.

In summary there are gaps emerging from the empirical review. There is scanty empirical evidence on modeling real estate prices using atheoretical models in Nairobi market. This paper is aimed at addressing this gap by modeling house prices in Nairobi using ARIMA.

2. Literature Review

Al-Marwani (2014) studied modeling and forecasting the real estate residential property market in Manchester City in the United Kingdom. He used both all-UK property indexes and specific property type indices, such as for detached, semi-detached flats and terraces. His main contribution was in forecasting real estate prices for different property types within Manchester City. He found that simple ARIMA models fitted the data well and could be used for forecasting, including factoring seasonality. Sklarz *et al*. (1987) studied the Autoregressive (AR) and ARIMA models using US property data. The study found that AR did a better job than ARIMA in forecasting housing prices. This was the case owing to low forecasting errors in AR compared to ARIMA. Brooks and Tsolacos (2000) researched the UK market and made use of CB Hiller Parker series. They used AR, Vector Autoregressive (VAR), and random walk models. They find AR model to best fit the estimate and does a better job in forecasting compared to the other models.

Birch and Sunderman (2003) employed exponential smoothing when estimating residential prices. It was an improvement to the overreliance on ARIMA models. The forecast based on their technique did not outperform those of traditional hedonic models as was expected. Nevertheless, the model used in the study was able to surmount some of the challenges associated with regression models. The study did not focus on fundamental economic variables.

Hepsen and Vatansever (2010) studied Dubai residential prices. They used the Box Jenkins ARIMA model to forecast prices. They find the model to be appropriate. However, Stevenson (2007) asserts that ARIMA models have a bias on the model specifications, thus giving varied outcomes. They studied the Irish market and deployed ARIMA, VAR, and Ordinary Least Squares (OLS) models. They find ARIMA models to be superior to the other models when the market sets aside the fundamentals.

Vishwakarma (2013) focused on the Canadian market and collected data between 2002 and 2011. He deployed ARIMA models that factored in economic variables such as inflation, exchange rates, interest rates, and GDP. He finds that the ARIMA models, in their simplicity, outperform other previously used methods when it comes to short-term forecasts of prices. The previously used methods were Kalman filter and Vector Error Correction Models (VECM).

Clapp and Giaccotto (2002) studied house prices in Dade County in Florida. They used several AR models, repeat sales, and hedonic models in forecasting real estate prices in that County. They performed one step ahead forecasts. They concluded that the hedonic model outperformed the other models. Guirguis *et al*. (2005) in their study employed six methods using quarterly data between 1975 and 2002 drawn from US housing market. The techniques used included AR, VECM, Kalman filter and Random Walk (KRW), Kalman filter and Autoregressive (KAR), GARCH, and exponential smoothing. They find that GARCH and KAR models outperform the rest when it comes to out of sample forecasts.

Crawford and Fratantoni (2003) studied the US housing market. They used ARIMA, GARCH and Regime Switching models to test their performance in both in sample and out of sample forecasts. The regime switching model was introduced to try and account for structural breaks and cycles in the market. They find that ARIMA family was the best in out of sample forecasts. Regime switching emerged more efficient with in-sample forecasts.

Miles (2008) extended the work of Crawford and Fratantoni (2003). They included additional models namely Generalised Autoregressive (GAR) and bilinear models. They find that GAR was better than ARIMA model a departure from the findings in Crawford and Fratantoni (2003). They concluded that this was the case due to high volatility. Rapach and Strauss (2009) also studied the US market. They find that the AR models and economic models are efficient in forecasting house prices. Gupta and Das (2010) researched twenty US state and used Bayesian Vector Autoregressive (BVAR) and VAR models to forecast house prices. They performed one step and four step ahead forecasts between Q1 2007 and Q1 2008. They found that BVAR model was better than the VAR models as evidenced by the average Root mean squared errors (RMSEs).

Barari *et al*. (2014) studied the US market with a focus on the existence of structural breaks. They first identified the structural breaks in the price indices and then deployed various models to make forecasts on prices. They used data for the period between 1995 and 2010. They find that, indeed, the real estate price series exhibited structural breaks. Upon running the models on the series with the breaks, they conclude that ARIMA models fall short of the reality of real estate dynamics.

3. Methodology

The paper adopted a quantitative research design to forecast residential housing prices in Nairobi. The paper focused on ARIMA. The paper used hedonic house price index for Nairobi for the forty-two quarters spanning 2010Q3 to 2020Q4. The index was constructed using data from a sample of residential houses in Nairobi over 42 quarters. The data collected for each house included the actual selling price, size measured in square feet, number of bedrooms, location, house type, and the quarter in which the house was sold. The selling periods were assigned dummy time variables, and the cross-sectional regression model was estimated. The exponents of the coefficients of the dummy time variables were used to construct the

index. The hedonic model used was specified as follows (Wolverton & Senteza, 2000; Sirmans *et al*., 2005):

$$Ln\ P_{it} = \propto + \beta_1 Ln\ Size_i + \beta_2\ Hse\ Type_i + \beta_3 Location_i + \beta_4 Bedroom_i + \sum_{t=2}^{T} \theta_{it} D_{it} + e_{it} \quad (1)$$

where $Ln\ P_{it}$ is log of the price of house *i, Size* is measured in square feet, *Hse Type* is value of 1 if apartments and 0 if standalone house, *Location* is value of 1 if the house is in an upmarket area and 0 if located elsewhere, *Bedroom* is number of bedrooms, D_{it} is the dummy variables for time denoting the 42 quarters in the study period, and e_{it} is error term.

The price index data was then split into two. The model training data was for the period 2010Q3 to 2019Q2. The out of sample test data was for six quarters: 2019Q3 to 2020Q4. The Box-Jenkins methodology was adopted (Box & Jenkins, 1976; Brooks, 2019). The methodology generally follows four steps. The first is to identify the models to be estimated. This is achieved by plotting auto-correlation functions (ACF) and autocorrelation functions (PACF). ACF calculates the correlation between past and current observations of a series. PACF measures the correlation between past observations and current observations, having controlled for the observations in between. Information criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also used. The second step is to estimate the models identified. Thirdly, diagnostic tests such as white noise tests and stability tests are carried out to confirm the model's validity. The last step is forecasting the time series and determining the most efficient model. The following is the specification of the models.

In an AR process, the series values depend on the past values of the series. An important assumption is stationarity to avoid explosive models. The AR (p) model was specified as follows:

$$HP_t = u + \sum_{i=1}^{p} \varnothing_i HP_{t-i} + \varepsilon_t \quad (2)$$

where HP_t is the house price series, P is order of the AR model denoting the number of lags, $\varnothing$ is the coefficient of lagged variable, and ε_t is the error term and is assumed to be normally distributed.

A moving average (MA) model is a linear combination of white noise processes. Therefore, the series values depend on the current and previous values of a white noise disturbance term. MA(q) was specified as follows:

$$HP_t = u + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \quad (3)$$

where HP_t is the house price, q is order of the MA model denoting the number of lags, ε_t is the error term and is assumed to have a normal distribution, and θ is the coefficient of the residuals fitted.

ARMA model is a combination of AR and MA models. If the series are differenced, then ARIMA (p, d, q) model is specified as follows:

$$HP_t = u + \sum_{i=1}^{p} \varnothing_i HP_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \quad (4)$$

where HP_t is the Housing Price, p is order of the AR model denoting the number of lags, d is order of integration, q is order of the MA model denoting the number of lags, ε_t is the error term and is assumed to have a normal distribution., θ is the coefficient of the residuals fitted, and $\varnothing$ is the coefficient of lagged variable.

The order level p and q were determined using the Box and Jenkins (1976) procedure. Once the data was fitted, dynamic out-of-sample forecast was done for six quarters (2019Q3–2020Q4). The mean absolute error (MAE) was computed for each model fitted. The model with the least MAE was deemed the most efficient.

4. Results

Box-Jenkins methodology was used and implemented in four steps. The constructed house price index used to identify the model span 36 quarters (2010Q3–2019Q2) is in Fig. 1. The graph pointed to a non-stationary series. As such, we tested for stationarity using an augmented dicky fuller test.

Table I shows the stationary test for price index. The absolute test statistic 1.720 was less than the critical value of 2.975 at 5% significance. Also, the p = 0.4208 was greater than 5%; hence, the null hypothesis of non-stationarity could not be rejected.

However, at the first difference, the price index series becomes stationary ($p < 0.05$ and test-statistic within the rejection region). This is depicted in Table II.

In addition, the stationarity of the differenced series was also confirmed by Fig. 2.

Having confirmed stationarity, we embarked on the identification of AR and ARIMA models. We plotted the ACF and PACF of the differenced price index to help in identifying the possible models. Fig. 3 depicts the ACF. Based on the chart, only the first lag is significant as it is outside of the 2-sided bands. This may suggest an MA (1) model.

Fig. 4 shows the PACF plot. Lags 1,5,6,10,12, and 14 are significant. This points to AR (1), AR (5), AR (6), AR (10), AR (12) and AR (14) models.

Considering both ACF and PACF plots, Table III depicts the possible models that can be used to fit the house price index.

The correct model should also be identified based on the information criteria. We, therefore, estimated the models

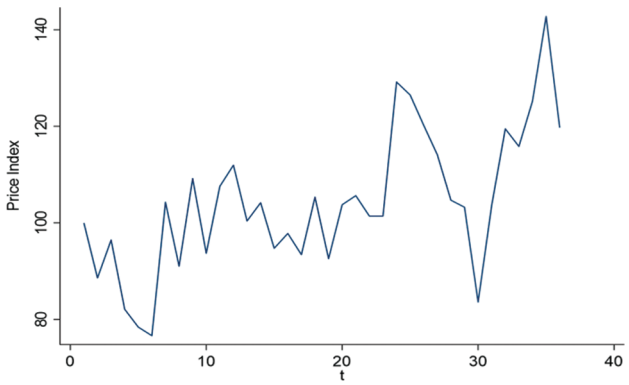

Fig. 1. House price index at levels (2010Q3–2019Q2).

TABLE I: Stationarity Test at Levels—Price Index

Variable: Price index			Number of obs = 34 Number of lags = 1	
H0: Random walk without drift, d = 0				
			Dickey-Fuller Critical value	
	Test statistic	1%	5%	10%
Z(t)	−1.720	−3.689	−2.975	−2.619
MacKinnon approximate p-value for Z(t) = 0.4208.				

Source: Author, 2023.

TABLE II: Stationarity Test at First Difference—Price Index

Variable: D. Price index			Number of obs = 33 Number of lags = 1	
H0: Random walk without drift, d = 0				
			Dickey-Fuller Critical value	
	Test statistic	1%	5%	10%
Z(t)	−4.622	−3.696	−2.978	−2.620
MacKinnon approximate p-value for Z(t) = 0.0001				

Source: Author, 2023.

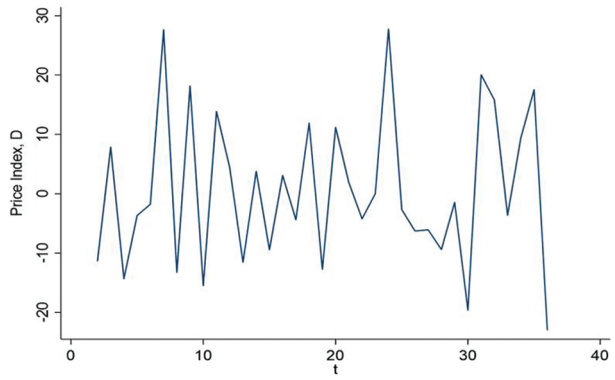

Fig. 2. Differenced house price index (2010Q3–2019Q2).

in Table III, performed diagnostic tests, and evaluated the AIC and BIC to identify the appropriate models. The results for ARIMA models are in Table IV.

The results for AR models are in Table V.

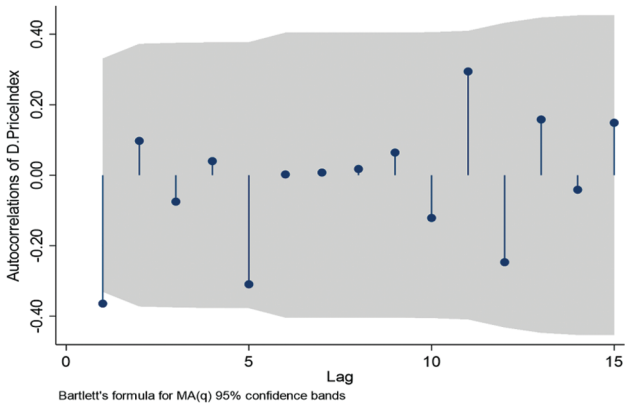

Fig. 3. ACF—differenced price index. Bartlett's formula for MA(q) 95% confidence bands.

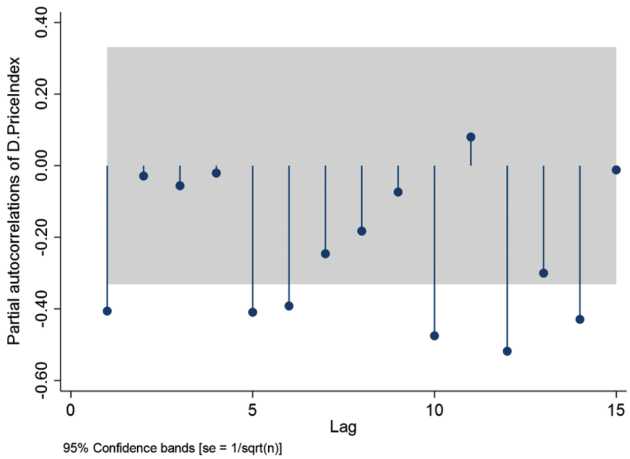

Fig. 4. PACF—differenced price index. 95% Confidence bands [se = 1/sqrt(n)].

TABLE III: Identified Models for Estimation

Sno	Model	Sno	Model
1	ARIMA (1,1,1)	7	AR (1)
2	ARIMA (5,1,1)	8	AR (5)
3	ARIMA (6,1,1)	9	AR (6)
4	ARIMA (10,1,1)	10	AR (10)
5	ARIMA (12,1,1)	11	AR (12)
6	ARIMA (14,1,1)	12	AR (14)

The foregoing AR and ARIMA models were estimated using data between 2010Q3 and 2019Q2. Our out-of-sample forecasts covered the period between 2019Q3 and 2020Q4. Besides, we computed the mean absolute errors (MAE) to determine the efficient model. Table VI shows the results of MAE. AR (6) had the lowest forecasting error based on MAE of 6.1 followed by AR (10). Therefore, it seems the best model is AR (6).

TABLE IV: Results of the Estimated ARIMA Models

Model (ARIMA)	(1,1,1)	(5,1,1)	(6,1,1)	(10,1,1)	(14,1,1)
AIC	277.519	281.642	278.164	283.898	280.341
BIC	283.624	293.853	290.375	302.214	304.763
Number of significant coefficients	0	0	2	4	10
Loglikelihood	−134.759	−132.821	−131.082	−129.949	−124.17
Diagnostic tests					
White noise (Portmanteau test)	0.8858	0.8793	0.7657	0.9464	0.9689
Stability test					
AR parameters	Satisfy	Satisfy	Satisfy	Satisfy	Satisfy
MA parameters	Do not satisfy	Satisfy	Satisfy	Do not satisfy	Satisfy

Note: ARIMA (1,1,1) and ARIMA (5,1,1) were dropped since they did not have significant coefficients with the former also failing the stability test. ARIMA (10,1,1) was also dropped for failing the stability test. ARIMA (1,12,1) was not feasible and therefore dropped. ARIMA (6,1,1) had the lowest AIC (278.164) and BIC (290.375), hence the most likely correct model. However, we still considered forecast ARIMA (14,1,1) in our forecast.

TABLE V: Results of the Estimated AR Models

AR Models	AR (1)	AR (5)	AR (6)	AR (10)	AR (12)	AR (14)
AIC	290.208	288.846	288.085	285.341	287.107	286.582
BIC	294.787	299.531	300.296	303.658	308.476	311.004
Number of significant coefficients	1	1	3	5	3	8
Loglikelihood	−142.104	−137.423	−136.043	−130.671	−129.553	−127.291
Diagnostic tests						
White noise (Portmanteau test)	0.0296	0.8813	0.7929	0.9701	0.9321	0.986
Stability						
AR parameters	Satisfy	Satisfy	Satisfy	Satisfy	Satisfy	Satisfy

Note: AR (1) was dropped since it is not a white noise process (p 0.0296 < 0.05). Considering the remaining models, AR (10) had the lowest AIC, while AR (5) had the lowest BIC. However, we went ahead to forecast all the AR models except AR (1).

TABLE VI: Forecast Errors—MAE

Forecast: 2019Q3–2020Q4 (Quarter 36 to Quarter 42)	MAE
AR (6)	6.10
AR (10)	8.43
ARIMA (6,1,1)	8.77
AR (5)	9.04
ARIMA (14,1,1)	10.60
AR (12)	11.18
AR (14)	11.43

The outcome of MAE is corroborated by forecasts vs. actual plots. Fig. 5 is for AR models shows the forecasts vs. actual values for 2019Q3 to 2020Q4 depicted in the graph as Q37–Q42 for ARIMA models.

The forecasts for ARIMA (6,1,1) were good, especially for quarters 37, 38, and 39. It predicted a better performance in quarters 41 (2020Q3) and 42 (2020Q4) than actual but that was at the height of corona pandemic. ARIMA (10,1,1) was generally inefficient except for quarter 41 (2020Q3) where it matched the actual value. Fig. 6 shows the forecasts for the AR models over the same six quarters.

The forecasts for AR (6) were accurate, especially for quarters 37, 38, and 39. It predicted a better performance in quarters 41 (2020Q3) and 42 (2020Q4) than the actual one, which could be explained by corona pandemic. AR

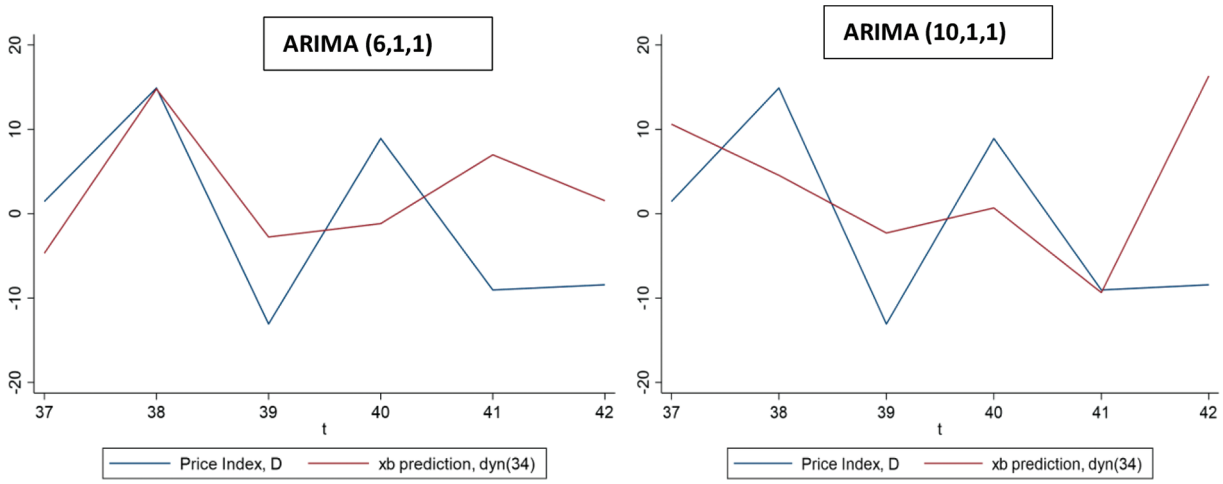

Fig. 5. ARIMA forecasts vs. actuals.

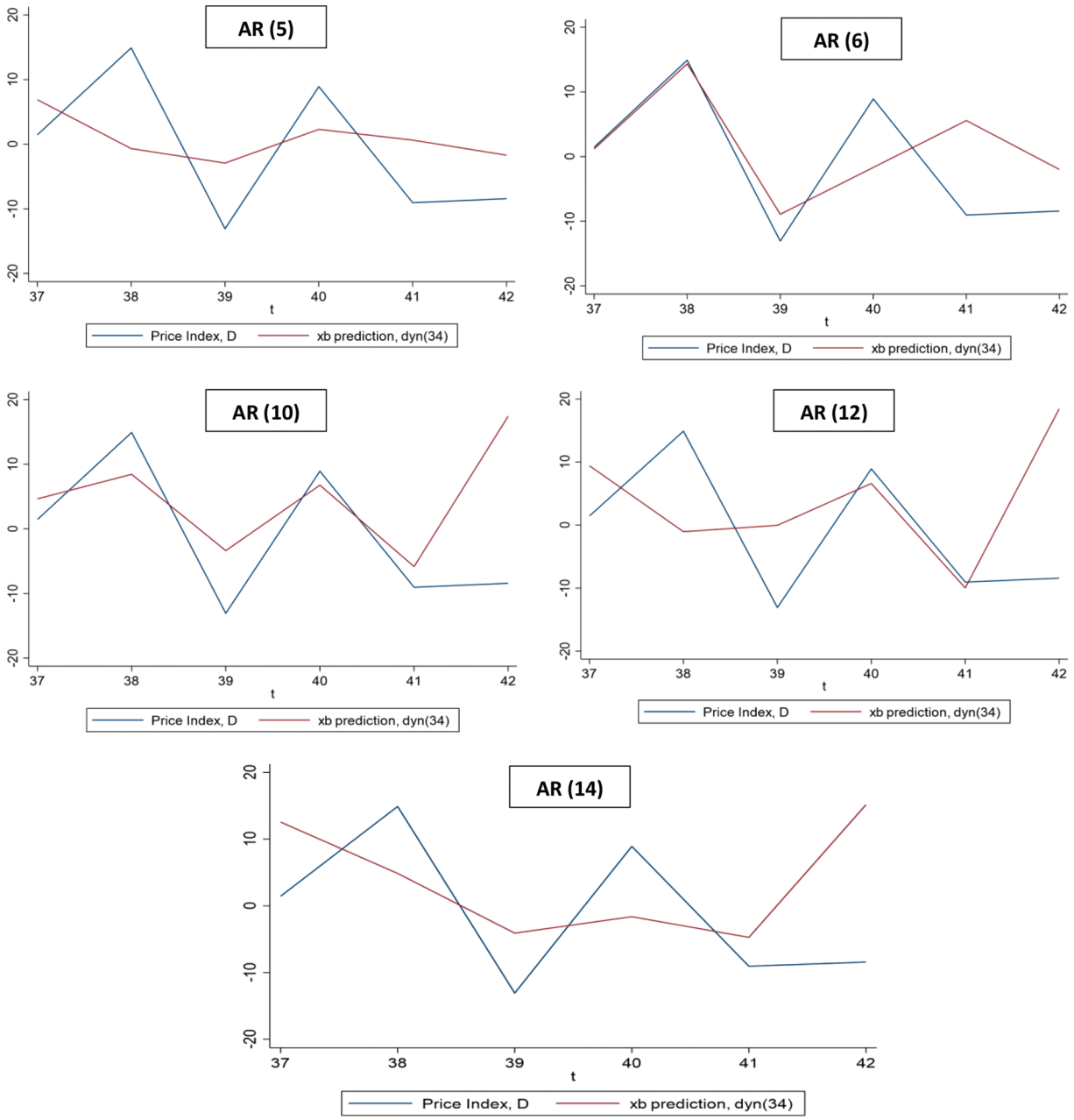

Fig. 6. AR forecasts vs. actuals.

(10) emerged as the second-best model. The rest were generally not efficient.

5. Discussion and Conclusion

The objective of the paper was to model residential house prices in Nairobi using ARIMA. The prices were operationalized by a hedonic house index constructed. The paper followed the Box-Jenkins methodology. Using MAE, the most efficient model was AR (6), followed by AR (10) and ARIMA (6,1,1), respectively. Brooks and Tsolacos (2000) found AR outperformed ARIMA in forecasting house prices in the United Kingdom. Sklarz *et al.* (1987) also concluded that AR outperformed ARIMA in the US market. Al-Marwani (2014), Jadevicius and Huston (2015), and Vishwakarma (2013) confirmed the suitability of ARIMA in forecasting real estate prices in the United Kingdom, Lithuania, and Canada, respectively. Crawford and Fratantoni (2003) found that ARIMA was efficient when forecasting with out-of-sample while regime switching was efficient with in-sample.

However, Barari *et al.* (2014) studied the US market and found breaks that may render ARIMA inefficient. Also, ARIMA may have biased forecasts, as was reported by Stevenson (2007), who studied the Irish market. Temür *et al.* (2019) studied the Turkish market. They find a hybrid of ARIMA, and recurrent neural network was suitable for forecasting real estate prices in Turkey.

6. Contributions and Recommendations

The paper sought to model house prices in Nairobi using atheoretical models. The paper finds the AR model suitable for forecasting house prices. This has contributed

to empirical evidence in foresting real estate prices in Kenya. Investors, lenders, regulators, and academicians may find the findings of this study useful. The findings may inform investment strategy and portfolio management by investors. It may also inform policy since house price forecasts may have social economic effects.

Atheoretical models have not been studied extensively in the Kenyan market. The major limitation is availability of data. Kenya lacks a suitable database for real estate transactions, especially in relation to actual purchase prices and individual house characteristics. We hope that this paper will motivate future studies that may include other models beyond ARIMA, such as regime switching.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

Al-Marwani, H. (2014). Modelling and forecasting property types price changes and correlations within the city of Manchester, UK. *Studies in Business and Economics*, *18*(2), 5–15.

Barari, M., Kundu, N. S. S., & Chowdhury, K. B. (2014). Forecasting house prices in the United States with multiple structural breaks. *International Econometric Review*, *6*(1), 1–23.

Birch, J. W., & Sunderman, M. A. (2003). Estimating price paths for residential real estate. *Journal of Real Estate Research*, *25*(3), 277–300.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*. San Francisco, California: Holden-Day.

Brooks, C. (2019). *Introductory Econometrics for Finance*. 4th ed. Cambridge: Cambridge University Press.

Brooks, C., & Tsolacos, S. (2000). Forecasting models of retail rents. *Environment and Planning*, *32*(10), 1825–1839.

Brown, G. (1997). Reducing the dispersion of returns in U.K. real estate portfolios. *The Journal of Real Estate Portfolio Management*, *3*(2), 129–140.

Clapp, J., & Giaccotto, C. (2002). Evaluating house price forecasts. *Journal of Real Estate Research*, *24*, 1–26.

Crawford, G., & Fratantoni, M. (2003). Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. *Real Estate Economics*, *31*, 223–243.

Guirguis, H. S., Giannikos, C. I., & Anderson, R. I. (2005). The US housing market: Asset pricing forecasts using time varying coefficients. *The Journal of Real Estate Finance and Economics*, *30*(1), 33–53.

Gupta, R., & Das, S. (2010). Predicting downturns in the US housing market: A Bayesian approach. *The Journal of Real Estate Finance and Economics*, *41*(3), 294–319.

Hepsen, A., & Vatansever, M. (2010). Forecasting future trends in Dubai housing market by using box-jenkins autoregressive integrated moving average. *International Journal of Housing Markets and Analysis*, *4*(3), 210–223.

Jadevicius, A., & Huston, S. (2015). ARIMA modelling of Lithuanian house price index. *International Journal of Housing Markets and Analysis*, *8*(1), 135–147.

Miles, W. (2008). Boom-bust cycles and the forecasting performance of linear and non-linear models of house prices. *The Journal of Real Estate Finance and Economics*, *36*(3), 249–264.

Rapach, D. E., & Strauss, J. (2009). Differences in housing price forecastability across US states. *International Journal of Forecasting*, *25*(2), 351–372.

Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 3–43.

Sklarz, M. A., Miller, N. G., & Gersch, W. (1987). Forecasting using long order autoregressive processes: An example using housing starts. *Real Estate Economics*, *154*(4), 374–388.

Stevenson, S. (2007). A comparison of the forecasting ability of ARIMA models. *Journal of Property Investment & Finance*, *25*(3), 223–240.

Temür, A, Akgün, M., & Temür, G (2019). Predicting housing sales in Turkey using ARIMA, LSTM and hybrid models. *Journal of Business Economics and Management*, *20*, 920–938.

Vishwakarma, V. K. (2013). Forecasting real estate business: Empirical evidence from the Canadian market. *Global Journal of Business Research*, *7*(3), 1–14.

Wolverton, M., & Senteza, J. (2000). Hedonic estimates of regional constant quality house prices. *The Journal of Real Estate Research*, *19*(3), 235–253.