Time Series Analysis of Ames Housing Data

Brayden Jansen O. Ang, Miguel Antonio H. Germar, Andrea Mikaela S. Zialcita

Abstract

This project explores the Ames Housing dataset as a high quality alternative to the Boston Housing data. Focusing on three-bedroom houses sold in Ames, Iowa between 2006 and 2010, a time series model was constructed based on the average monthly sale price in order to forecast the average sale price for the next two months. The time series was tested for stationarity using the augmented Dickey-Fuller test, which revealed non-stationarity. This was addressed through differencing. Subsequent augmented Dickey-Fuller and Ljung-Box tests revealed stationarity and autocorrelation, respectively. Analysis of the autocorrelation and partial autocorrelation functions, as well as the Akaike information criterion (AIC) and parsimony, suggested that an ARMA$(1, 1)$ model best fit the time series for the differenced data. The chosen model was found to be both causal and invertible. Forecasting using this model predicted a sharp increase in average monthly sale price for three-bedroom houses in the next two months.

1 Introduction

1.1 Background

Traditionally, the Boston Housing dataset has been used for research on housing prices. However, the Ames Housing dataset offers a more modern, detailed, and comprehensive alternative, as it provides information on thousands of property sales in Ames, Iowa. This makes it suitable for time series modeling, which will be carried out to discover temporal dependencies and understand the behavior of housing price movements through statistical diagnostics and model fitting.

1.2 Statement of the problem

This paper aims to model the mean sale price of three-bedroom houses in Ames as a time series and predict the mean sale price of three-bedroom houses in Ames for the next 2 months.

1.3 Scope and limitations

We limit our paper to the investigation of the most appropriate ARMA model for a single kind of house (by number of bedrooms, whichever number is most prevalent in the data). That is, we do not aim to model panel data for multiple house sizes, and we do not aim to regress sale prices on data apart from lagged sale prices, or use models other than ARMA models. The paper is concerned with autocorrelations in short-term fluctuations in house sale prices for the purpose of forecasting the general housing market in the city, not predicting prices for individual houses. We also limit forecasting to two months into the future.

Submission for MATH 62.2 Time Series and Forecasting, AY 2024-2025 Second Semester

2 Methods

2.1 Description of data

The Ames Housing data was created by Dean De Cock in 2011. Compiled by the Ames Assessor's Office, it contains accurate data about housing sales between 2006 and 2010 in the city of Ames, Iowa.

The data has 2930 instances, of which most (1597) are three-bedroom houses. Hence, the data was filtered to three-bedroom houses, so that all price observations used in the project can be expected to come from the same distribution. Furthermore, there are 81 variables that describe the physical and locational characteristics of the residential properties. However, only information of year and month sold, as well as sale price, were used for the project.

2.2 Relevant assumptions

We assume that the dataset contains complete data for the population of three-bedroom house sales in the city of Ames.

We also assume that the discrete time series of mean monthly sale prices generally reflects the sales trends in the entire city, which in reality may change at more granular intervals of time and may be subject to local differences in each district.

Another assumption is that housing markets are inherently volatile, so it is appropriate to use low lag orders when modeling with ARMA in this context. This is based on Dufitinema [2021], where most of the tested models have AR and MA orders of 1 or 2 quarters, corresponding to 3 or 6-month lags, as well as Yilmaz and Kestel [2020], whose models have an AR order of 1 month and an MA order of up to 3 months.

Therefore for our purposes, we only expect to forecast housing prices up to 2 months in advance. Autocorrelation at lag 2 is sufficient for us to model the data, apart from the standard assumption of stationarity. We intend to limit our AR and MA orders to a maximum of 2 months, unless PACF and ACF strongly suggest higher lag orders, in which case a maximum of 6 months is acceptable.

2.3 Definitions and theorems

Definition: ARMA process. A time series W_t of observations recorded at discrete times t is said to be an ARMA(p, q) process if it is stationary and satisfies

$$W_t = \alpha + \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $\phi_p \neq 0$, $\theta_q \neq 0$, and $\{ Z_t \} \sim \text{WN}(0, \sigma^2)$ with $\sigma^2 > 0$. An autoregressive process AR(p) of order p is an ARMA$(p, 0)$ process, while a moving average process MA(q) of order q is an ARMA$(0, q)$ process.

Definition: Causality. An ARMA(p, q) model is said to be *causal* if W_t can be written as a one-sided linear process:

$$W_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \psi(B) Z_t,$$

where

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j \quad \text{and} \quad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Definition: Invertibility. An ARMA(p, q) model is said to be *invertible* if the time series W_t can be written as

$$\pi(B) W_t = \sum_{j=0}^{\infty} \pi_j W_{t-j} = Z_t,$$

where

$$\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j \quad \text{and} \quad \sum_{j=0}^{\infty} |\pi_j| < \infty.$$

Definition: AR and MA polynomials. The AR and MA polynomials are defined as

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \quad \phi_p \neq 0,$$

and

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q, \quad \theta_q \neq 0,$$

respectively, where B is a complex number.

Theorem 1. An ARMA(p, q) model is *causal* if and only if $|B| > 1$ whenever $\phi(B) = 0$.

Theorem 2. An ARMA(p, q) model is *invertible* if and only if $|B| > 1$ whenever $\theta(B) = 0$.

2.4 Testing stationarity and autocorrelation

Let $\{ X_t \}$, $t \in \mathbb{Z}$ be the time series of monthly mean sale prices of three-bedroom houses in Ames, Iowa. Each observation is recorded at the end of the corresponding month. Let t denote the number of months since the beginning of year 2006. The dataset contains observations for t from 1 to 55 and is provided in Appendix A.

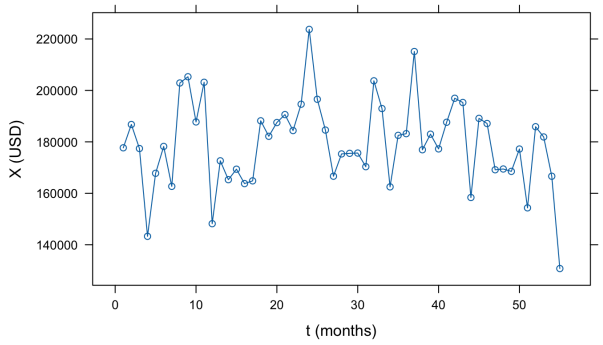

Figure 1: Line plot of observed values of X_t from $t = 1$ to $t = 55$.

Prior to time series modeling, we test assumptions of stationarity and autocorrelation.

We test the stationarity of $\{ X_t \}$ using the augmented Dickey-Fuller (ADF) Test. The null hypothesis is that the roots of the autoregressive (AR) characteristic polynomial include a unit root, that is, the process is not stationary. This gives us p-value $= 0.2192 > 0.05$, so we fail to reject the null hypothesis. That is, we say that $\{ X_t \}$ is not stationary.

Hence we transform $\{ X_t \}$ using a lag-1 difference, defined as

$$Y_t = \nabla X_t = X_t - X_{t-1}.$$

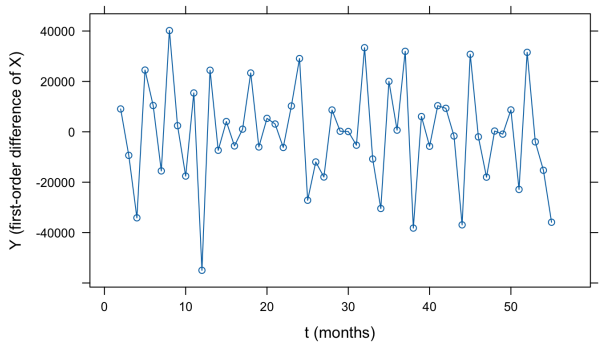

Figure 2: Line plot of observed values of Y_t from $t = 2$ to $t = 55$.

Once again, we test the stationarity of $\{ Y_t \}$ using the ADF Test. This gives us p-value $< 0.01 < 0.05$, so we reject the null hypothesis and say that $\{ Y_t \}$ is stationary.

We then test the autocorrelation of $\{ Y_t \}$ using the Ljung-Box test with lag 2. The null hypothesis is that the ACF is equal to 0 for both lags 1 and 2, that is, the data is not autocorrelated. Since p-value $= 0.02408 < 0.05$, we reject the null hypothesis and say that $\{ Y_t \}$ is autocorrelated.

We note that the Ljung-Box test is commonly performed with a lag approximately equal to $\ln T$ which is equal to $\ln 54 \approx 4$. However, given our assumption that housing markets are volatile and we only expect to forecast house prices up to 2 months in advance, we also assume that autocorrelation at lag 2 is sufficient for modeling the data.

Since the first-order difference $\{ Y_t \}$ satisfies the assumptions of stationarity and autocorrelation, we proceed to modeling.

3 Results and discussion

3.1 Modeling

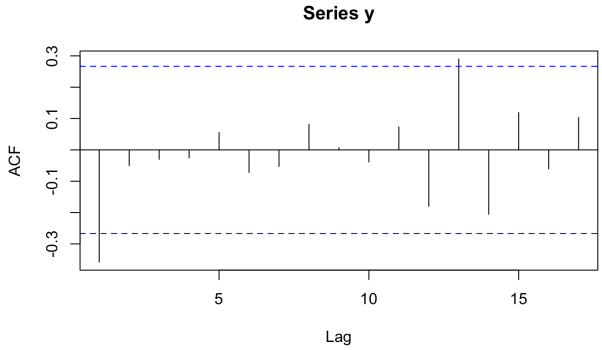

Figure 3: ACF Plot for $\{ Y_t \}$.

The ACF of $\{ Y_t \}$ takes on mostly small values after lag 1. We note that ACF is large at lag 13. However, we ignore lag 13 for reasons of parsimony and because of our assumption about the volatility of the housing market. Hence we conclude that ACF cuts off at lag 1.

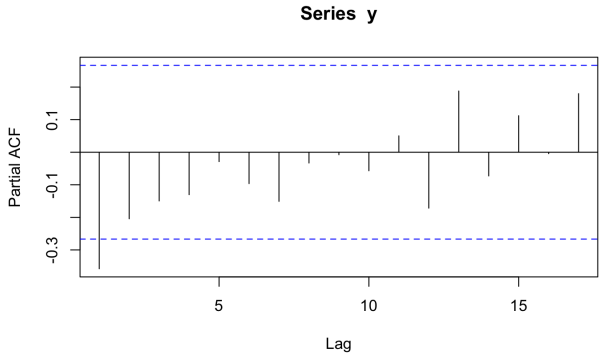

Figure 4: PACF Plot for $\{ Y_t \}$.

Similarly, the PACF of $\{ Y_t \}$ cuts off after lag 1.

In summary, since ACF and PACF both cut off after lag 1, we consider MA(1), AR(1) and ARMA$(1,1)$ as candidate models.

Fitting AR(1) to $\{ Y_t \}$ produced a model with AIC $= 1222.36$, log-likelihood equal to -608.18, and equation

$$\begin{aligned} Y_t &= -741.6501(1 + 0.3735) - 0.3735Y_{t-1} \\ &= -1020.2037 - 0.3735Y_{t-1} + Z_t, \end{aligned} \tag{1}$$

where $\{ Z_t \} \sim \text{WN}(0, \sigma^2)$.

Fitting MA(1) produced a model with AIC $= 1214.18$, log-likelihood equal to -604.09, and equation

$$Y_t = -187.6784 - 0.8411Z_{t-1} + Z_t. \tag{2}$$

Fitting ARMA$(1,1)$ produced a model with AIC $= 1213.50$, log-likelihood equal to -602.75, and equation

$$\begin{aligned} Y_t &= -114.4641(1 - 0.2786) + 0.2786Y_{t-1} + Z_t \\ &= -82.5768 + 0.2786Y_{t-1} + Z_t - 0.9999985Z_{t-1}. \end{aligned} \tag{3}$$

We find that ARMA$(1,1)$ has the lowest AIC of the three candidate models.

Casting a wider net for candidate models given our assumption that we may forecast housing prices up to 2 months in advance, we performed a grid search shown in Appendix B. Here, we find that ARMA$(1,1)$ has the lowest AIC of all ARMA(p,q) processes with $p, q \in \{ 0, 1, 2 \}$.

Since ARMA$(1,1)$ has the lowest AIC and highest log-likelihood, we choose this as our model for $\{ Y_t \}$.

We note that the ARMA$(1,1)$ model (3) satisfies causality and invertibility. The proof is provided in Appendix C. So, the AR part of the model does not depend on future values, and the MA part of the model is unique.

3.2 Forecasting

We use this model to forecast future values of Y_t, from which we derive forecasts for X_t.

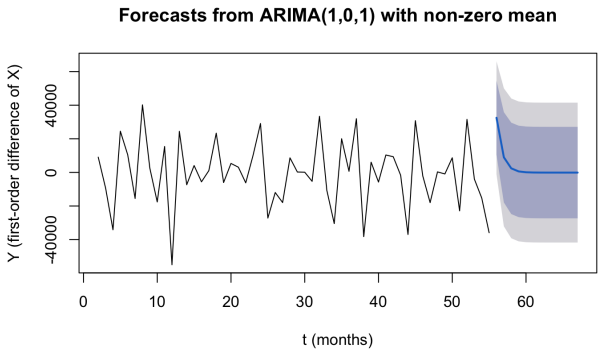

Figure 5: Forecasts for differenced data Y_t for $t = 56$ to $t = 67$.

An increase is predicted from $X_{55} = 130,750.00$ to $X_{56} = 163,270.20$, followed by a smaller increase to $X_{57} = 172,249.00$.

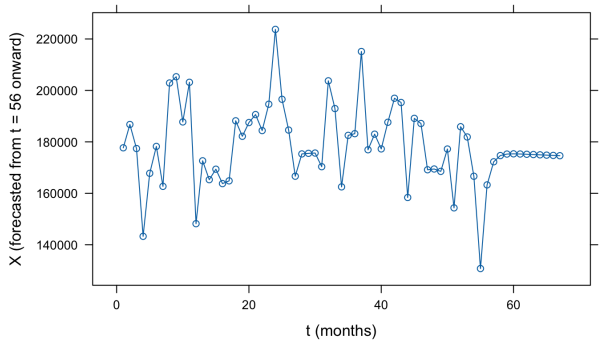

Figure 6: Forecasts for undifferenced data $\{ X_t \}$.

Note that we forecasted 12 months into the future for the purpose of observing the long-term behavior of our model in the plot. This is a secondary observation to be discussed. However, under our initial assumptions about housing market volatility, we only intend to make predictions using the forecasts 1 and 2 months into the future (i.e., $X_{56} = 163,270.20$, $X_{57} = 172,249.00$). Our discussion of results primarily focuses on these short-term forecasts as well as the interpretation of our model coefficients.

3.3 Discussion of results

Interpreting the model coefficients of (3), we see that the coefficient of the AR term Y_{t-1} is much smaller than both the coefficient of the present white noise Z_t and the (absolute value of) the coefficient of previous white noise Z_{t-1}. This implies that housing prices depend more on white noise, which could represent shocks, than previous values of housing prices, which aligns with the volatility of the housing market.

The forecast suggests that the house prices will increase sharply over the next month and increase less sharply thereafter. It is reasonable to conclude that there will be a decrease in the quantity of three-bedroom houses demanded in Ames, assuming no changes in other non-price determinants of demand.

Figure 5 depicts the differenced data $\{ Y_t \}$ exhibiting mean convergence (close to zero), which explains the mean convergence of the undifferenced data $\{ X_t \}$ as shown in Figure 6. This is because time series models in general tend to converge to a mean or explode with sudden increases and decreases, so time series models can only be used to reliably forecast two or three time steps in the future. Since the coefficient of the previous value Y_{t-1} in (3) is small, it makes sense that our model exhibits mean convergence, as the value of the Y_t's will inevitably shrink.

Another noteworthy observation is that when performing grid search for the ARMA models by increasing AIC, an MA(2) model has the second lowest AIC, while an AR(1) has the highest AIC and is the worst performing model. While this goes against what the ACF and PACF plots suggest, this is likely due to the combined ARMA$(1,1)$ model being the best fit for the data. For a pure MA process, the PACF is expected to gradually decay, and for a pure AR process, the ACF is expected to gradually decay. Both ACF and PACF cut off instead of gradually decaying, which, aside from noise, could explain the poor performance of the pure AR(1) and pure MA(1) models.

References

D. De Cock. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*, 19(3), 2011. doi: https://doi.org/10.1080/10691898.2011.11889627.

J Dufitinema. Forecasting the Finnish house price returns and volatility: A comparison of time series models. *International Journal of Housing Markets and Analysis*, 15(1):165–187, 2021. doi: https://doi.org/10.1108/IJHMA-12-2020-0145.

B. Yilmaz and A. S. S. Kestel. Forecasting house prices in Turkey: GLM, VAR and time series approaches. *Pressacademia*, 9(4):274–291, 2020. doi: https://doi.org/10.17261/pressacademia.2020.1310.

A Appendix

Appendix A. Table 1 contains the data for X_t, the monthly mean sale price of three-bedroom houses in Ames, Iowa. Note that t refers to the number of months since the beginning of year 2006.

Appendix B. Table 2 contains the results of grid search performed on the ARMA(p, q) models, ordered by increasing AIC.

Appendix C. Proof that our ARMA$(1, 1)$ model for $\{ Y_t \}$ is causal and invertible.

The model given in (3) can be written as

$$\phi(B)Y_t = -82.5768 + \theta(B)Z_t,$$

where the AR polynomial is

$$\phi(B) = 1 - 0.2786B,$$

and the MA polynomial is

$$\theta(B) = 1 - 0.9999985B.$$

The solution of $\phi(B) = 0$ is $B = \frac{1}{0.2786} \approx 3.5894$, so $|B| > 1$. The solution lies outside the unit circle. Therefore, the model is causal by Theorem 1.

The solution of $\theta(B) = 0$ is $B = \frac{1}{0.9999985}$, so $|B| > 1$. The solution lies outside the unit circle. Therefore, the model is invertible by Theorem 2.

Thus, our final model for $\{ Y_t \}$, which is the first-order difference of the average monthly sale price X_t, is a causal and invertible ARMA$(1, 1)$ model.

Table 1: Data for X_t.

t	X_t
1	177,689.60
2	186,760.30
3	177,416.00
4	143,279.50
5	167,785.10
6	178,223.90
7	162,703.40
8	202,892.60
9	205,316.30
10	187,748.20
11	203,165.60
12	148,193.80
13	172,636.00
14	165,302.50
15	169,380.10
16	163,777.50
17	164,842.20
18	188,188.00
19	182,171.10
20	187,510.40
21	190,607.20
22	184,397.00
23	194,642.30
24	223,725.60
25	196,557.40
26	184,589.10
27	166,652.40
28	175,310.50
29	175,542.30
30	175,658.50
31	170,349.80
32	203,730.40
33	192,946.90
34	162,504.90
35	182,528.80
36	183,187.60
37	215,132.80
38	176,928.60
39	182,996.70
40	177,264.30
41	187,625.40
42	196,956.30
43	195,302.70
44	158,370.10
45	189,139.50
46	187,140.00
47	169,150.00
48	169,455.20
49	168,528.80
50	177,217.90
51	154,348.10
52	185,889.70
53	181,904.00
54	166,640.10
55	130,750.00

Table 2: Results of grid search performed on ARMA(p, q), ordered by increasing AIC.

AR Order (p)	MA Order (q)	AIC
1	1	1213.5020
0	2	1213.9340
0	1	1214.1820
2	2	1215.4480
1	1	1215.4510
2	2	1217.4630
2	0	1221.3560
1	0	1222.3590

Appendix D. The following are R commands used to execute the project.

```
library(TSA)
library(tseries)
library(forecast)
library(astsa)
library(AmesHousing)

ames = as.data.frame(make_ames())

amesf = ames[ames$Bedroom_AbvGr == 3, c("Year_Sold", "Mo_Sold",
"Sale_Price")]

years_as_months = (amesf$Year_Sold - 2006) * 12
amesf$t_months = years_as_months + amesf$Mo_Sold

ames_ma = amesf[c("Sale_Price", "t_months")] %>%
  dplyr::group_by(t_months) %>%
  dplyr::summarise(Sale_Price = mean(Sale_Price)) %>%
  as.data.frame()

lattice::xyplot(
  Sale_Price~t_months, ames_ma, type=c("l", "p"),
  xlab = "t (months)", ylab = "X (USD)"
)

x = ts(ames_ma$Sale_Price)

adf.test(x)

y = diff(x)
adf.test(y)
Box.test(y, type = "Ljung", lag = 2)

lattice::xyplot(y~2:(n+1), ames_ma, type=c("l", "p"),
                xlab = "t (months)", ylab = "Y (first-order
                difference of X)")

acf(y)
pacf(y)

ar1 = Arima(y, order = c(1, 0, 0))
ar1

ma1=Arima(y, order = c(0, 0, 1))
ma1
```

```
arma = Arima(y, order = c(1, 0, 1))
arma

p_orders = 1:2
q_orders = 1:2

grid_search_aic = function(y, p_orders, q_orders, use_bic=FALSE) {
  arma_aic_values = c()

  for (p in p_orders) {
    for (q in q_orders) {
      result = Arima(y, order = c(p, 0, q))
      if (use_bic) {
        aic_value = result$bic
      }
      else {
        aic_value = result$aic
      }
      arma_aic_values = c(arma_aic_values, aic_value)
    }
  }

  table = cbind(p_orders, q_orders, arma_aic_values) %>% as.data.frame()
  table = table[order(table$arma_aic_values, table$p_orders,
  table$q_orders),]
  return(table)
}

gs_results_AR_only = grid_search_aic(y,p_orders,0)
gs_results_AR_only

gs_results_MA_only = grid_search_aic(y,0,q_orders)
gs_results_MA_only

gs_results_ARMA = grid_search_aic(y,p_orders,q_orders)
gs_results_ARMA

all_results = rbind(gs_results_AR_only, gs_results_MA_only, gs_results_ARMA)
all_results[order(all_results$arma_aic_values, all_results$p_orders,
all_results$q_orders),]

arma.p = forecast(arma, h = 12)
plot(arma.p, xlab = "t (months)", ylab = "Y (first-order difference of X)")

deltas = arma.p$mean

x_pred = c()
prev_x_value = x[n]

for (delta in deltas) {
  next_x_value = prev_x_value + delta
  x_pred = c(x_pred, next_x_value)
  prev_x_value = next_x_value
}

x_with_pred = c(x, x_pred)
t_with_pred = 1:length(x_with_pred)
```

```
lattice::xyplot(x_with_pred~t_with_pred, type=c("l", "p"),
xlab = "t (months)", ylab = "X (forecasted from t = 56 onward)")
```