Real Estate Prices, Past and Future

Bob Ding

2020-11-12

Introduction

As a major part of the economy, real estate has been constantly and closely monitored by investors and researchers. Since 1970 [1] when most countries' statistical offices or central banks began to collect data on house prices, interest in predicting and forecasting house prices have gradually augmented, catalyzing out more and more sophisticated modeling techniques. Due to modern society's increased ability to collect and store more data, predicting real estate prices has shifted to data-driven, which further improved modeling precision.

Being such a sophisticated product, real estate prices are impacted by many factors. While most of the factors helpful in predicting the house are observable and descriptive to the house itself, such as house' size, number of bathrooms, and whether it possesses a swimming pool, etc., there are also not observable factors that also impact house prices, such as the underlying real state market economy, periodicity of real estate prices, and so on.

Many previous researchers have already proposed multiple ways of predicting and forecasting house prices. From the most simple regression methods as proposed in Kaggle's advanced regression technique blog posts [2], to those that account for repeated sales of houses [3], and to those that consider temporal effects, such as proposed by Fernando[4] and Nihar et. al[6]. Though these studies are drastically different, and research can incorporate more effects to propose more sophisticated models, each study's focus is different and clear. Thus, it is important to make certain of the research question before creating a model.

We propose our goal of this study. We are interested in building an all-in-one model that considers observable house data while also assuming the unobservable temporal effect from the real estate house market. To further narrow down, we are interested in only modeling houses at Durham, NC. Thus, the only type of house that we will be researching is a house in Durham, NC, due to the terrain's better familiarity. There are 4 research questions including 1) how house specific observable, such as number of beds, number of bathrooms, etc. variables affect housing prices, 2) what periodic temporal effects is presented in the past house market, 3) extract past real estate market temporal effects, and 4) making short term forecast of housing prices.

1. Exploratory Data Analysis

1.1 Data description

The Dataset is scraped from the Redfin official set [5]. Redfin is a real estate brokerage that was founded in 2004. Its website records all of Redfin's historical house purchase in the past 3 years. Therefore, we scraped these 3 years of data recorded for North Carolina, ranging from 2017 April to 2020 May. This dataset contains 6962 observations. Thanks to Redfin's meticulous data management, no missing value in any field was presented. Each observation is a recorded deal of house purchase. Therefore, rather than being subjective such as the seller's one-sided proposed selling price, the price is the real deal price between customer and seller, which is objective enough for us to fit the model on.

The dataset contains many covariates. There are some hard-to-process string information, such as the name of the community, and highly detailed geographical information beyond our interests. Therefore, to simplify our research, we introduce the following covariates of our interest.

Name	Description	Mean	Standard Deviation
`Price`	the deal price (dollar) of the house	319702.000	18926.533
`beds`	number of beds in the house	3.784	0.742
`sold.dat`	the date on which the deal is settled	not meaningful	not meaningful
`baths`	the number of bathrooms the house has	2.825	0.661
`square.feet`	usable area (ft^2) measured in square feet	2437.000	907.523
`lot.size`	total area (ft^2) of the lot	10275.000	1501.712
`house.age`	age (years) of the house when purchased	10.710	9.852
`property.type`	Condo, Townhouse, or Single-Family residential	not available	not available
`latitude`	latitude of the house	35.962	0.054
`longitude`	longitude of the house	-78.861	0.061

Table 1: Variable of Interests

1.2 Exploring Data

The first impression is that the distribution of house types in category Condo, Townhouse, and Single Family Residential is highly uneven. This is shown in Appendix figure 6. Besides, A complete pairwise-plot has also been attached in the Appendix in figure 7. The rest of the following section will include 3 subsections explaining the 3 major concerns about data assumption, including suggestions on how to address them. Besides, an additional subsection is also added to explain the engineering of additional predictors and how we address interactions.

1.2.1 Multi-collinearity

The number of bathrooms of the house has a very strong linear correlation with the total square feet the house has. Notice that in figure 1.2.1 below, strong collinearity is shown between the number of baths and square feet of the house, achieving a correlation of 0.7843. Though they're not high enough for us to concern about identifiability issues, we should still be careful in the final model output for these highly correlated covariates. More pairwise relationships between variables can be found in the pair plot shown in Appendix at figure 7

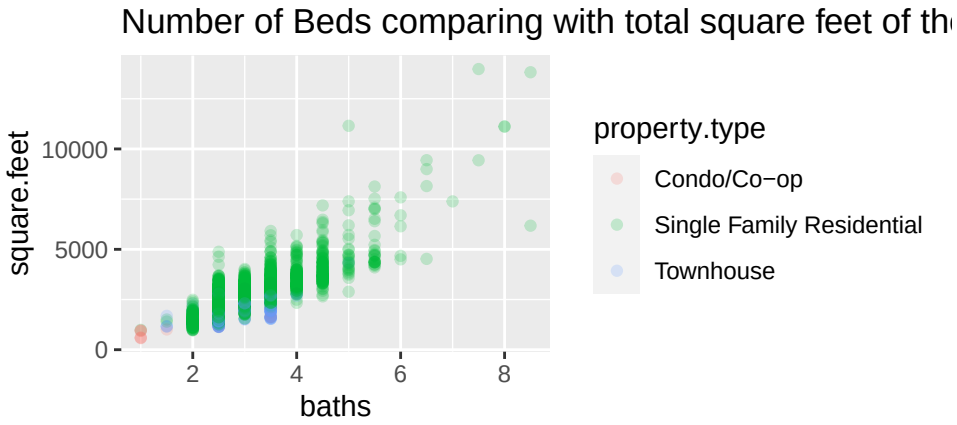


Figure 1.2.1

1.2.2 Heteroscedasticity

Apart from some strong collinearity and high correlation between covariates, heteroscedasticity is also evident in this dataset. We found that the increase in these predictors' values and response value leads to increased variance. Surely enough, bigger and more expensive houses exhibit greater variation in price. This violates the linear regression homoscedasticity assumption, as shown in figure 1.2.2 below. To address this, we perform log-transformation on the response variable and create a final regression response variable `log.price,` which is the logarithm of the house deal price. As shown in the last line of the pair plot, the Heteroscedasticity problem is resolved without harming the response and predictors' linear relationship. It has also made some linear relationships more evident.

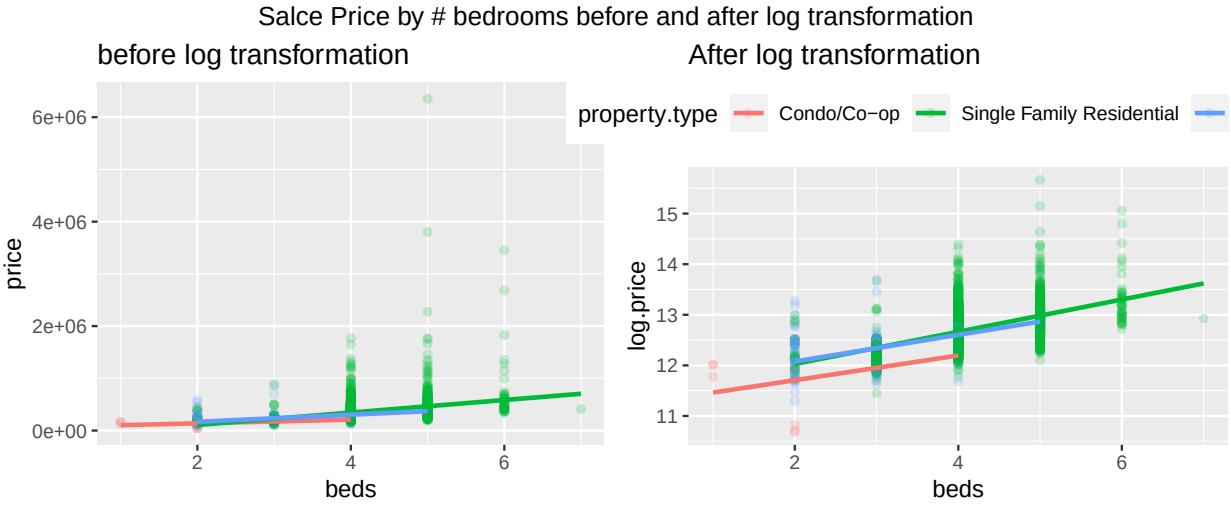


Figure 1.2.2 (b)

Figure 1.2.2 (b)

1.2.3 Stationarity

When determining what model to capture the temporal effect, it is imperative to verify whether the stationarity assumption has been satisfied. In our case, due to the fact it is unlikely that there are houses sold each day, we need to construct the time series by windowing the raw data – that is, we treat all the real estate deals that happened in a time window as deals happening at the same timestep. By specifying our window's width, we can modulate and balance between flexibility in temporal effect and validity of regression coefficients. At this point, we choose window width to be 15 days.

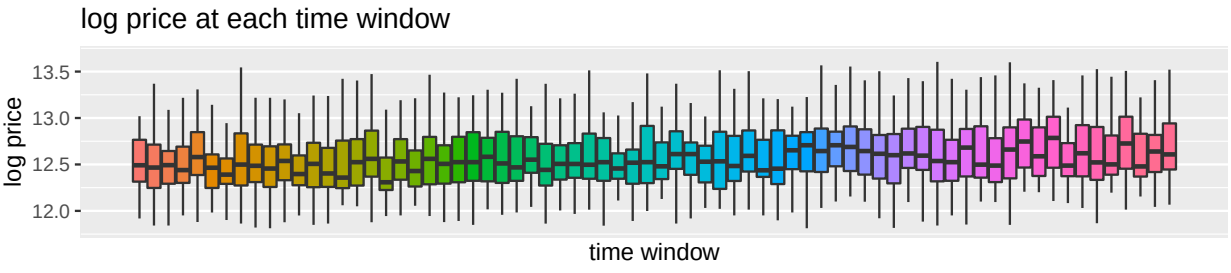
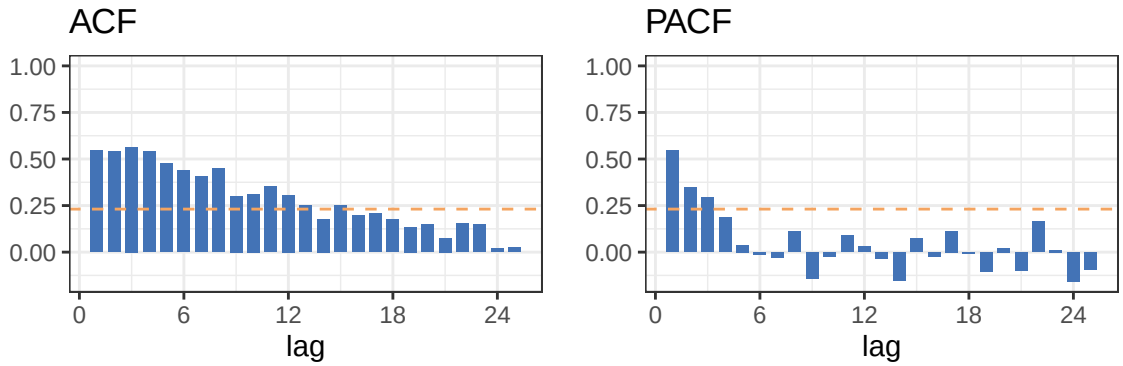


Figure 1.3.2 (a)

Under the 15-day window size, from figure 1.3.2 (a), we learn that there is a slight upward trend in the distribution of the logarithm of house prices through time. Though this figure can only show the distribution of log price marginalized out by all other exogenous variables, it is helpful for us to identify a linearly

increasing time trend. Facing the slightly non-stationarity upward trend, we propose 2 model choices: 1) ignore (no trend model), 2) adding variable `sold.dat`, the sold date, into the linear observation model as a predictor in order to capture the linear trend using the linear model (linear trend model). We will fit these 2 models and determine the final model by checking their performance and validation results.

Besides, it would also help identify the number of lags p by 1) checking autocorrelation and partial autocorrelation (ACF/PACF) plots and 2) making a hypothesis. Below in figure 1.3.2 (b) shows the ACF/PACF plot mentioned. We observe that lag 1,2,3 are statistically significantly correlated. Therefore, we should at least propose have $p \geq 3$. However, as we also hold the hypothesis that house prices might perform cycles of a very long period, we further extend $p \geq 12$ given that a window is now defined to be 15 days. Therefore, we will start by incorporating 12 auto-correlation terms in our model. Due to the excess amount of lags, we will design Bayesian ridge shrinkage for AR(p) coefficients. Details are fully explained in Appendix section A.



1.2.4 Engineered Feature and Interaction

We hypothesize that an excessive number of beds in a house with a comparatively low number of bathrooms can impact house prices. Therefore, we created the variable `room.Diff`, which means how much more bathrooms the house has than beds. We found such feature creates distinct effects across Single Family Residential, Townhouse, and Condo in affecting log of price. In figure 1.2.4 (a), we can observe a different slope of the variable for 3 types of houses. Therefore, together with `room.Diff`, interactions between the bed-bath difference and house type should also be added to our model.

Lastly, we believe that incorporating a simple distance to Duke covariable can increase prediction power, as we are modeling house prices at Durham. Therefore, we engineered the new variable `dist.duke`, indicating the house's road distance to Duke. This is calculated by a package called `mapdist`. It has a full map of the US. . By drawing the shortest route between 2 addresses, the package calculates their road distances. Such an address is inferred from latitude and longitude. The package will locate the street closest to the given longitude and latitude. We use this procedure to calculate road distances between each house and Duke. Then, as figure 1.2.4 (b) shows, aligned with our conjecture, we anticipate a longer distance to Duke, leading to a lower housing price.

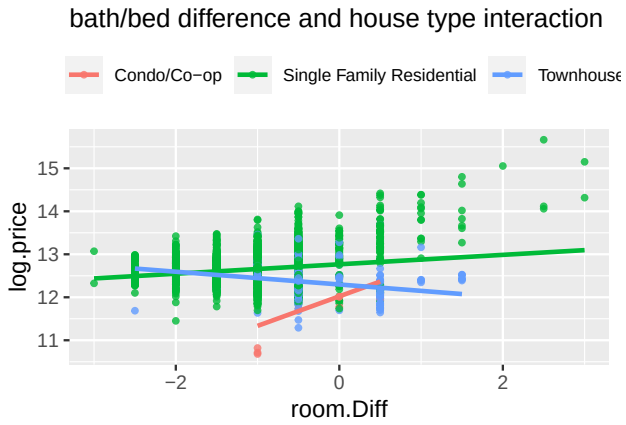


Figure 1.2.4 (a)

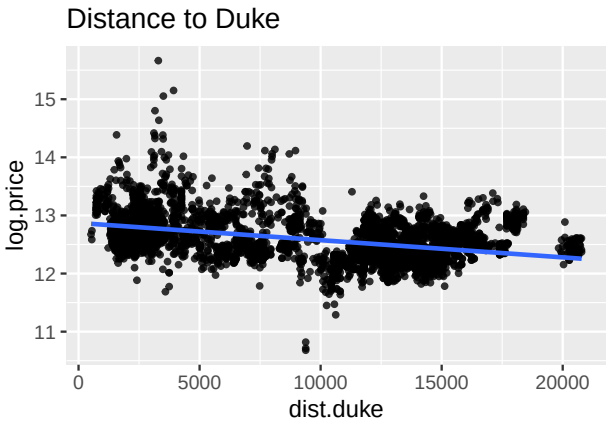


Figure 1.2.4 (b)

2 Model Formulation

We define our model as a regression model on top of an AR(p) model. After EDA, we determine response to be the logarithm of the house deal price. The predictors are the number of beds, the number of bathrooms, total square feet of the house, the entire size of the lot, house's age, the type of the house (condo, single-family residential, or townhouse), the difference between beds and bathrooms, distance to Duke, sold date (plus 1 for 1 day later), and finally interaction between house type and the bed-bath difference. Besides, we define the intercept α_t as the time-varying not observable "market effect." The transition of α_t follows an AR(p) process. Therefore, the full model is defined as below. (**Note:** This model has incorporated sold date as a predictor. This corresponds to the linear trend model mentioned in EDA 1.2.3 about stationarity. To construct the ignore trend model, remove sold date together with its β coefficient)

$$\begin{aligned}
\log(\text{house price})^{(i)} &= y_t^{(i)} && (1)\\
&= \beta_{\text{beds}}\ \mathbf{beds}_t^{(i)} + \beta_{\text{sold date}}\ \mathbf{sold\ date}_t^{(i)} + \beta_{\text{beds}}\ \mathbf{baths}_t^{(i)} + && (2)\\
&\quad \beta_{\text{lot size}}\ \mathbf{lot\ size}_t^{(i)} + \beta_{\text{house age}}\ \mathbf{house\ age}_t^{(i)} + && (3)\\
&\quad \beta_{\text{room difference}}\ \mathbf{room\ Diff}_t^{(i)} + \beta_{\text{dist. to Duke}}\ \mathbf{dist.\ to\ Duke}_t^{(i)} + && (4)\\
&\quad \beta_{\text{square feet}}\ \mathbf{square\ feet}_t^{(i)} + \beta_{\text{property type}}\ \mathbf{property\ type}_t^{(i)} + && (5)\\
&\quad \beta_{\text{interaction}}\ \mathbf{room\ Diff}_t^{(i)} \times \mathbf{property\ type}_t^{(i)} + \alpha_t + \nu_t^{(i)} && (6)\\
\alpha_t &= \sum_{i=1}^{p} \theta_i \alpha_{t-i} + \omega_t && (7)\\
\omega_t &\sim \mathcal{N}(0, w) && (8)\\
\nu_t^{(i)} &\sim \mathcal{N}(0, v) && (9)
\end{aligned}$$

Where logarithm of deal price, denoted as $y_t^{(i)}$, is the response of the i^{th} ($i \in \{1, 2, 3 \cdot n_t\}$) house sold on the t^{th} timestep (window). (note: suppose for each window $t \in \{1, 2, 3, \cdots, T\}$, there are n_t sold houses in the t^{th} window. Then n_t need not be equal for all t). The rests are predictive variables. α_t is an time varying intercept which will be modeled by the AR(p) model described in (5), (6). $\nu_t^{(i)}$ is an additional observation uncertainty and ω_t is an additional evolution uncertainty. To simply our modeling process, we take $\nu_t^{(i)}$, ω to have constant variance at all time.

Through reparametrization, we can simplify the model as the following. A detailed explanation of why we should formulate the model in this compact form is due to inference. A detailed explanation can be found in Appendix B.

$$
\begin{aligned}
\boldsymbol{\alpha}_t &= \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1} + \boldsymbol{W}_t \\
\boldsymbol{y}_t &= \mathbf{1}\boldsymbol{\alpha}_t + \beta \boldsymbol{X}_t + \boldsymbol{\nu}_t \\
\boldsymbol{W}_t &\sim \mathcal{N}(\mathbf{0}, w\boldsymbol{I}) \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I})
\end{aligned}
$$

$$
\mathbf{1} \coloneqq \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \boldsymbol{\alpha}_t \coloneqq \begin{bmatrix} \alpha_t \\ \alpha_{t-1} \\ \alpha_{t-2} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \dots & \theta_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \alpha_{t-2} \\ \alpha_{t-3} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1}
$$

$$
\begin{aligned}
\boldsymbol{\alpha}_0 &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\
\boldsymbol{\beta} \mid (\tau = v^{-1}), \kappa &\sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{I}(\tau\kappa)^{-1}\right) \\
p((\tau = v^{-1}) \mid \kappa) &\propto 1/\tau \\
\kappa &\sim \boldsymbol{G}(1/2, 1/2) \\
\boldsymbol{\theta} | (\phi = w^{-1}), \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}/\phi\right) = \mathcal{N}((\frac{1}{p}, \frac{1}{p}, \cdots, \frac{1}{p})^T, \frac{1}{p}\phi^{-1}\mathbf{I}) \\
(\phi = w^{-1}) | \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathbf{G}\left(a_0 = \frac{v_0}{2}, b_0 = \frac{v_0 s_0^2}{2}\right) = \mathbf{G}\left(\frac{1}{2}, \frac{1}{2}\right)
\end{aligned}
$$

3 Methodology

This section explores the methodology of using MCMC sampling with the forward-backward algorithm to make inference on parameters. Besides, we will also discuss how the methodology and the inferred parameter posterior distribution could answer our pre-stated 4 research questions in the introduction.

The Model requires making statistical inference of the following parameters: $\{w, v, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\alpha}_{1:T}\}$, where T is the last timestep index. To obtain these, we choose to apply MCMC sampling with the forward-backward algorithm within the MCMC sampler. In Appendix F and below is the algorithm that returns MCMC samples for these parameters. A highly detailed derivation of each step's sampling distribution can be found in appendix A. Notice that to apply MCMC, we must design a prior distribution for all the parameters. Due to the flexibility of prior design and our previous sections stated multicollinearity problem, we have decided to propose Bayesian Ridge regression priors for both the AR(p) coefficients $\boldsymbol{\theta}$ and also the linear regression model $\boldsymbol{\beta}$ coefficients. Details can also be found in Appendix A. From now on, denote $\mathcal{D}_t \coloneqq \{\boldsymbol{X}_{1:t}, \boldsymbol{y}_{1:t}\}$, all the information in the observed data from initial time to time t.

We have proposed 4 research questions at the end of the introduction section. A detailed explanation of how our inferred parameters will address each research question has been included in Appendix F subsection F.1.

Algorithm 1: parameter inference algorithm

Result: Sampling distribution for $\{w, v, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\alpha}_{1:T}\}$
Initialize: $\mathbb{P}((\theta_1, \dots, \theta_p)^\intercal), \mathbb{P}(w), \mathbb{P}(v)$ via pre-set prior distribution
while *not converged* **do**
  Calculate: posterior mean and covariance for $\boldsymbol{\alpha}_t|\mathcal{D}_t$: $m_t, C_t \ \forall t \in 1:T$ via forward filtering algorithm in Appendix
  Sample: $\boldsymbol{\alpha}_t|\mathcal{D}_T$ from $m_t^*, C_t^*, \forall t \in 1:T$ by backward smoothing
  Sample: $\boldsymbol{\theta}, (\phi = w^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}$ by first sampling $(\phi = w^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}$ and then $\boldsymbol{\theta}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, \phi$
  Sample: $\boldsymbol{\beta}, (\tau = v^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\theta}$ by first sampling $(\tau = v^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\theta}$ and then $\boldsymbol{\beta}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\theta}, \tau$
end
Return: samples of $w, v, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:T}$

4 Validation and Results

4.1 Model Validation

To validate the model, we have developed a 3 way splitting of total data, which includes 1) the training set, 2) the in-sample testing set, and 3) the (out-of-sample) forecast testing set. Explicitly, using 15 days as the size of the window, there are 75 windows of data. To obtain the forecast testing set, we take the last 5 windows of data, incorporating 239 observed house deals, as our forecast testing set. After we trained our model using the training set of data, we will forecast these 5 windows of data to test our out-of-sample model forecasting performance. To obtain the in-sample testing set, we randomly pick 20% of house deal observation from each of the first 70 windows, incorporating 537 observed house deals, to form the in-sample testing set. The rest of the data forms the training set. We make inferences on parameters based on the training set and predict the in-sample testing set. Such an error is the in-sample prediction error. In short, we will validate both the in-sample prediction mean absolute error (MAE) and also the out-of-sample forecast MAE, and we expect the latter to be slightly greater than the former. Such a validation result is in Table 5 of Appendix C.4. We observe that in-sample prediction error is smaller for both no trend and linear trend model. Nevertheless, the out-of-sample error increased more for the no trend model compared to the linear trend model, denoting that the no trend model is not doing as well in forecast as and the linear trend model. This is just as anticipated. However, generally, the MAE of both models is still acceptable, as the error of predicting and forecasting house prices is usually not going beyond 500 dollars. An illustration of the data split is shown below.

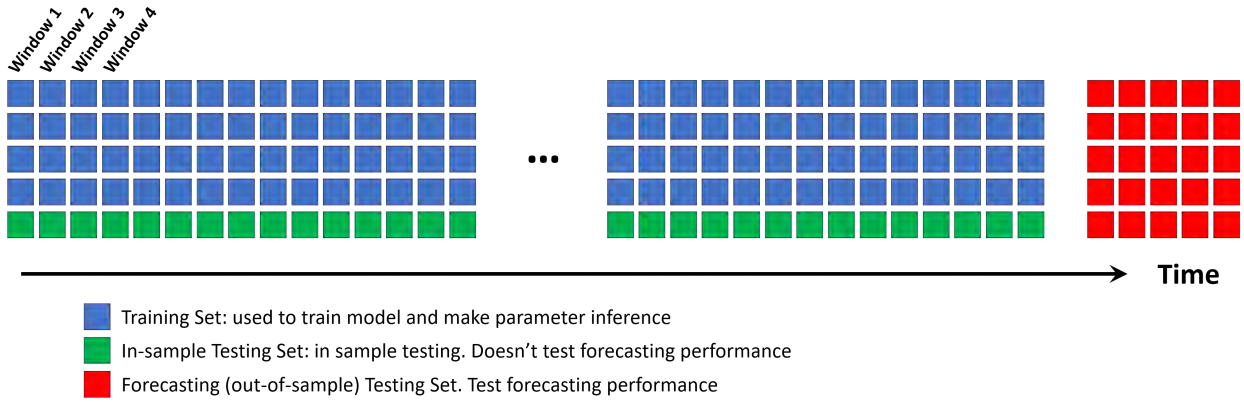


Figure 1: Illustration of 3 way data split

4.2 Model Results

Just as the model itself suggests, we have an abundance of results. For the first research questions, the modeling results are all in Table 3 in Appendix C. Interpretations are the same as that of the linear model. For instance, holding all else constant, having 1 more bed in the same house will make the median of the house price decrease to $e^{-0.13}$ multiplicative factor, which is about 89.34 percent; holding all else constant, having 1 more bathroom will make the median of the house price increase to $e^{0.21}$ multiplicative factor, which is 123.36 percent. Interaction terms can also be interpreted. For example, holding all else constant, if a house is a townhouse, then having even one more bed than bathroom will make the median of the house to decrease to $e^{0.415-0.599}$ multiplicative factor, which is reduced to the surprising 0.5627 percent. However, if the house is a single-family house, then having even one more bed than the bathroom will make the house's median to decrease to $e^{0.415-0.550}$ multiplicative factor, which is reduced to 87.37 percent. Lastly, for the linear trend model, we want to speak for the `sold.dat` coefficient. For the same house, regardless of market periodicity, we expect the house to be sold at $e^{0.0001387}$ = 100.0139% its price the day before. To provide a

better idea of annual house market inflation, we expect the same house to be sold 5.17% more expensive than it was in the last year (calculated without taking consideration of amortization. Aka. not taking account of the increase of the house age variable `house.age`), which is than twice of the federal reserve rate. For houses in a city like Durham, this surely is a very reasonable estimate. Due to the page limit, we only show some examples in this section. Complete interpretation can be found in Appendix C.1 below Table 3.

The second model result is about house market periodicity. From Table 4 in Appendix C.2, we learn that the housing market shows 2-month, 4-month, annual, and seasonal periodicity, with the order from the most significant to the least significant. I conjecture that 4 months is the Duke university semester cycle, with the spring semester, summer, and fall semester nearly equally partitioning the entire school year. Therefore, it is reasonable that house prices may adjust with this 4-month cycle. The 2-month cycle might also be part of the 4-month periodicity as the periodicity need not be of a standard sine wave. The annual and seasonal periodicity is very intuitive as it is common sense that there could be seasonal business cycles and the annual economic cycle that impacts the house price.

Our result for research question 3, the market movement, can be found in Appendix C.3, and our final forecasts are in Appendix C.4 Table 5, which has already been discussed in model validation section 4.1.

5 Diagnostic and Sensitivity Analysis

5.1 Model Diagnostics

Our diagnostic procedure for model diagnostics is 2-fold since the model is essentially a combination of 2 models, the linear observation model, and the time-series evolution model. Thus, diagnostics should evaluate whether assumptions for both models are met.

For the linear observation model, we have plotted the residual plot of linear regression. In Appendix D.1 figure 3 and D.2 figure 5, the x-axis represents index of the houses observations. We observe that all of them have constant variance, centered at zero, homoskedastic, and are Gaussian distributed. Thus, the assumption for the linear regression model is satisfied.

Besides, by checking the posterior distribution of Table 3 in Appendix C.1, it is recognizable that all the variables have a 95% credible interval not intersecting 0. Thus, we are confident that all the linear regression parameters show some predictive power to the house price's logarithm.

For the evolution model, due to its latency, we cannot diagnose the autoregressive residuals. However, as the parameters are inferred from MCMC, it is necessary to check the convergence of MCMC. In Appendix D.1 and figure 2 and in D.2 figure 4, we surely observe the MCMC has well mixed. Hence our model satisfies all the required assumptions.

5.2 Sensitivity Analysis

Thanks to Bayes Ridge, we do not need to perform sensitivity analysis on the shrinkage coefficient κ, as we've placed an objective prior on it. Besides, all model coefficients have shown significance, therefore not sensitivity analysis is needed. All priors we set for parameters are objective and flat. Thus, sensitivity check for priors are also not necessary. In short, the model doesn't have any strong assumptions, whose impact on modeling results should be tested.

In this section, the major sensitivity concern is about the window size previously mentioned in section 1.2.3. We preset the window size to be 15 days. However, we are concerned about whether increasing the window size to 30 days will improve our model. Although the 30-day window is a more reasonable time-varying window size for real estate prices in our context, increasing the window size can drastically decrease the effective sample size for AR(p) process as the total number of windows decreased. Besides, MCMC does not mix well when the effective sample size becomes small, which pillaged the model's ability to make a reasonable inference. Below is the validation error of 30 days of windows size. Comparing to the error of

15 days window size model included in Appendix C.4, a longer window does not perform well. Therefore it should not be used.

Table 2: 1 step forecast MAE for 2 models using 30 days as window size (out-of-sample error)

no trend	linear trend
6065	16041

6 Discussion

Our model shows superior flexibility in integrated modeling observable factors impacting house prices and the not observable market's impact on house prices. Prior could be designed specifically for all parameters to enable shrinkage or even other functionalities. By incorporating the house's sell date in the linear observation model, we can overcome the stationarity restriction of the AR(p) model without causing identifiability issues. Therefore, this model is properly functional and can provide interval estimates for all coefficients. Besides, more variables and effects can be constantly added easily to the model to boost predictive power further.

One limitation is that the model currently only assumes a linear trend of all time. This need not be true. A way to solve this is by applying the DLM framework and adding locally linear dynamics into the $\mathbf{\Theta}$ matrix. Higher orders polynomials to approximate local trend can also be considered, such as locally quadratic or cubic dynamics. Therefore, we need not worry about the model not capturing the complex trend, as, by Stone Weierstrass Theorem, any continuous functions (trend) defined on a compact interval can be approximated by countable numbers of polynomials. Thus, any trend can be approximated as long as it is continuous.

Another limitation is that the time series data we currently have is not long enough to analyze even longer terms of market periodicity. The ideal length would be to have more than 50 years of house deal data, which is apparently achievable. Thus, to retain enough degrees of freedom, we are forced to narrow house deal window size to only 15 days, therefore providing useful estimates.

The final limitation is that the model could have taken in more linear predictors. Community and other geographical information are vital in determining the house price. To further improve this case study, we should have incorporated more of that information. However, due to the time limit and our limited resources, we ignored this potential improvement and decided to focus on the modeling process.

Bibliography

[1] Schularick, M., & Steger, T. (2014). No Price Like Home: Global House Prices, 1870–2012. Federal Reserve Bank of Dallas, Globalization and Monetary Policy Institute Working Papers, 2014(208). doi: 10.24149/gwp208

[2], Manisaurabh. (2020, September 14). House-Prices-Advanced-Regression-Technique. Retrieved October 18, 2020, from https://www.kaggle.com/manisaurabh/house-prices-advanced-regression-technique

[3], Ghysels, E., Plazzi, A., Valkanov, R., & Torous, W. (2013). Forecasting Real Estate Prices. Handbook of Economic Forecasting, 509-580. doi:10.1016/b978-0-444-53683-9.00009-8

[4]. Aguilar, F. (2019, July 15). Time Series Analysis on US Housing Data. Retrieved October 18, 2020, from https://medium.com/@feraguilari/time-series-analysis-modfinalproyect-b9fb23c28309

[5]. Redfin. (2020). Real Estate, Homes for Sale, MLS Listings, Agents | Redfin. https://www.redfin.com/

[6]. Bhagat, N., Mohokar, A., & Mane, S. (2016). House Price Forecasting using Data Mining. International Journal of Computer Applications, 152(2), 23-26. doi:10.5120/ijca2016911775

Appendix

A Parameter Inference

A.1 Forward Filtering

$$
\begin{aligned}
\boldsymbol{\alpha}_t &= \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1} + \boldsymbol{W}_t \\
\boldsymbol{y}_t &= \mathbf{1}\boldsymbol{\alpha}_t + \beta \boldsymbol{X}_t + \boldsymbol{\nu}_t \\
\boldsymbol{W}_t &\sim \mathcal{N}(\mathbf{0}, w\boldsymbol{I}) \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I}) \\
\mathbf{1} &:= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

First, denote that

$$
\begin{aligned}
\boldsymbol{\alpha}_t | \mathcal{D}_{t-1}, - &\sim \mathcal{N}(\boldsymbol{\Theta} m_{t-1}, \boldsymbol{\Theta} C_{t-1} \boldsymbol{\Theta}^T + w\boldsymbol{I}) = \mathcal{N}(a_t, R_t) \\
\boldsymbol{y}_t | \boldsymbol{\alpha}_t, \mathcal{D}_{t-1}, - &\sim \mathcal{N}(\mathbf{1}\boldsymbol{\alpha}_t + \boldsymbol{X}_t\beta, v\boldsymbol{I}) \\
\mathbb{P}(\boldsymbol{\alpha}_t | \mathcal{D}_t) &\propto \mathbb{P}(\boldsymbol{\alpha}_t | \mathcal{D}_{t-1}, -)\mathbb{P}(\boldsymbol{y}_t | \boldsymbol{\alpha}_t, \mathcal{D}_{t-1}, -) \\
&\propto \exp\left\{ -\frac{1}{2}\left[\boldsymbol{\alpha}_t^T (R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1})\boldsymbol{\alpha} - 2\boldsymbol{\alpha}_t^T (R_t^{-1} a_t + v^{-1}\mathbf{1}^T(\boldsymbol{y}_t - \boldsymbol{X}_t\beta)) \right] \right\} \\
&\sim \mathcal{N}\left(\left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1} (R_t^{-1} a_t + v^{-1}\mathbf{1}^T(\boldsymbol{y}_t - \boldsymbol{X}_t\beta)), \left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1} \right) \\
&= \mathcal{N}(m_t, C_t)
\end{aligned}
$$

$$
\begin{aligned}
a_t &= \boldsymbol{\Theta} m_{t-1} \\
R_t &= \boldsymbol{\Theta} C_{t-1} \boldsymbol{\Theta}^T + w\boldsymbol{I} \\
m_t &= \left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1} (R_t^{-1} a_t + v^{-1}\mathbf{1}^T(\boldsymbol{y}_t - \boldsymbol{X}_t\beta)) \\
C_t &= \left(R_t^{-1} + v^{-1}\mathbf{1}^T\mathbf{1}\right)^{-1}
\end{aligned}
$$

Use this equation to update

A.2 Backward Smoothing

Suppose we already know that

$$
\mathbb{P}(\boldsymbol{\alpha}_{t+1} | \mathcal{D}_T) \sim \mathcal{N}(m_{t+1}^*, R_{t+1}^*)
$$

Let's look at log likelihood of $\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1} | \mathcal{D}_T$. Using conditional independence, we have

$$
\begin{aligned}
-\frac{1}{2}\ell(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+1}; \mathcal{D}_T) &= \log \mathbb{P}\left(\boldsymbol{\alpha}_{t+1} \mid \boldsymbol{\alpha}_t\right) + \log \mathbb{P}\left(\boldsymbol{\alpha}_t \mid \mathcal{D}_t\right) - \log \mathbb{P}\left(\boldsymbol{\alpha}_{t+1} \mid \mathcal{D}_t\right) + \log \mathbb{P}\left(\boldsymbol{\alpha}_{t+1} \mid \mathcal{D}_T\right) \\
&= (w)^{-1}(\boldsymbol{\alpha}_{t+1} - \boldsymbol{\Theta}\boldsymbol{\alpha}_t)^T(\boldsymbol{\alpha}_{t+1} - \boldsymbol{\Theta}\boldsymbol{\alpha}_t) + (\boldsymbol{\alpha}_t - m_t)^T(C_t)^{-1}(\boldsymbol{\alpha}_t - m_t) - \\
&\quad (\boldsymbol{\alpha}_{t+1} - a_{t+1})^T(R_{t+1})^{-1}(\boldsymbol{\alpha}_{t+1} - a_{t+1}) + \\
&\quad (\boldsymbol{\alpha}_{t+1} - m^*_{t+1})^T(C^*_{t+1})^{-1}(\boldsymbol{\alpha}_{t+1} - m^*_{t+1}) + \text{constant} \\
&= \boldsymbol{\alpha}^T_{t+1}({C^*_{t+1}}^{-1} + w^{-1}\boldsymbol{I} + R^{-1}_{t+1})\boldsymbol{\alpha}_{t+1} + \boldsymbol{\alpha}^T_t(w^{-1}\boldsymbol{\Theta}^T\boldsymbol{\Theta} + C^{-1}_t)\boldsymbol{\alpha}_t + \\
&\quad 2\boldsymbol{\alpha}^T_{t+1}(-w^{-1}\boldsymbol{\Theta})\boldsymbol{\alpha}_t - 2\boldsymbol{\alpha}^T_t(C^{-1}_t m_t) - 2\boldsymbol{\alpha}^T_{t+1}(R^{-1}_{t+1}a_{t+1} + {C^*_{t+1}}^{-1}m^*_{t+1}) + \text{constant}
\end{aligned}
$$

One eternity of calculation later, we end up with:

$$
\begin{aligned}
J_t &= C_t\boldsymbol{\Theta}^T(\boldsymbol{\Theta}C_t\boldsymbol{\Theta}^T + w\boldsymbol{I})^{-1} \\
m^*_t &= m_t + J_t(m^*_{t+1} - \boldsymbol{\Theta}m_t) \\
C^*_t &= C_t + J_t(C^*_{t+1} - \boldsymbol{\Theta}C_t\boldsymbol{\Theta}^T - w\boldsymbol{I})J^T_t
\end{aligned}
$$

And therefore

$$
\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t+1}, \mathcal{D}_T \sim \mathcal{N}\Big(m_t + J_t(\boldsymbol{\alpha}_{t+1} - \boldsymbol{\Theta}m_t), C_t - J_tR_{t+1}J^\top_t\Big)
$$

A.3 Dynamic model sampling: $(\theta_1, \cdots, \theta_p)^\top$, $w = \phi^{-1}$

This is a simple linear regression $\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\theta} + w_t$ with design matrices as

$$
\boldsymbol{y} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \vdots \\ \alpha_T \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \alpha_{1-1} & \cdots & \alpha_{1-p} \\ \alpha_{2-1} & \cdots & \alpha_{2-p} \\ \alpha_{3-1} & \cdots & \alpha_{3-p} \\ \alpha_{4-1} & \cdots & \alpha_{4-p} \\ \alpha_{5-1} & \cdots & \alpha_{5-p} \\ \vdots & \vdots & \vdots \\ \alpha_{T-1} & \cdots & \alpha_{T-p} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_p \end{bmatrix}
$$

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta}, \mathbf{X}) &\propto \phi^{\frac{T}{2}} \exp\{-\frac{1}{2}\phi(\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})^\intercal(\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})\} \\
\boldsymbol{\theta}|\phi, \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}/\phi\right) = \mathcal{N}((0.5, 0.5, 0.5)^T, 1/3\phi^{-1}\mathbf{I}) \\
\phi|\mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathbf{G}\left(a_0 = \frac{v_0}{2}, b_0 = \frac{v_0 s_0^2}{2}\right) = \mathbf{G}\left(\frac{1}{2}, \frac{1}{2}\right) \\
\boldsymbol{\mu}_n &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{y}\right) \\
\boldsymbol{\Lambda}_n &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0\right) \\
a_n &= a_0 + \frac{T}{2} \\
b_n &= b_0 + \frac{1}{2}(\boldsymbol{y}^\intercal\boldsymbol{y} + \boldsymbol{\mu}_0^\intercal\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^\intercal\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n) \\
\boldsymbol{\theta}|\phi, \boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathcal{N}\left(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n^{-1}/\phi\right) \\
\phi|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathbf{G}\left(a_n, b_n\right)
\end{aligned}
$$

A.4 Observation Model Sampling $(\beta_1 \cdots)^\top$, $v = \tau^{-1}$

Very similar as above, this is also a linear model. Besides, it is possible to apply Bayesian Ridge here. let's create the Bayesian ridge model

$$
\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{\alpha}_t \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \nu_t \\
\boldsymbol{z}_t = (\boldsymbol{y}_t - \boldsymbol{\alpha}_t \mathbf{1}) \mid \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \tau &\sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \mathrm{l}_n/\tau\right) \\
\boldsymbol{\beta} \mid \tau, \kappa &\sim \mathrm{N}\left(\mathbf{0}, \mathrm{l}(\tau\kappa)^{-1}\right) \\
p(\tau \mid \kappa) &\propto 1/\tau
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{y}_t - \boldsymbol{\alpha}_t \mathbf{1} | \boldsymbol{X}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}) &\propto \tau^{\frac{n}{2}} \exp\{-\frac{\tau}{2}(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})\} \\
\mathbb{P}(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{z}_t, \tau, \boldsymbol{\alpha}) &\propto \mathbb{P}(\boldsymbol{z}_t | \boldsymbol{X}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}) \mathbb{P}(\boldsymbol{\beta} | \tau, \kappa, \boldsymbol{\alpha}) \mathbb{P}(\tau | \kappa, \boldsymbol{\alpha}) \\
&\propto \tau^{\frac{n}{2}} \exp\{-\frac{\tau}{2}(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{z}_t - \boldsymbol{X}\boldsymbol{\beta})\} (\tau\kappa)^{\frac{p}{2}} \exp\left\{-\frac{\tau\kappa}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \tau^{-1} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^T(\tau\boldsymbol{X}^T\boldsymbol{X} + \tau\kappa\mathbf{1})\boldsymbol{\beta} - 2\tau\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{z}_t\right]\right\} \\
&\sim N\left((\boldsymbol{X}^T\boldsymbol{X} + \kappa\mathbf{1}_p)^{-1}\boldsymbol{X}^T\boldsymbol{z}_t, \tau^{-1}(\boldsymbol{X}^T\boldsymbol{X} + \kappa\mathbf{1}_p)^{-1}\right) \\
\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{z}_t, \tau, \boldsymbol{\alpha} &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mathbb{P}(\tau | \boldsymbol{\beta}, \kappa, \boldsymbol{z}_t, \boldsymbol{\alpha}) &\propto \tau^{\frac{n+p}{2}-1} \exp\left\{-\frac{\tau\kappa}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \\
\tau | \boldsymbol{\beta}, \kappa, \boldsymbol{z}_t, \boldsymbol{\alpha} &\sim \boldsymbol{G}(\frac{n+p}{2}, \frac{\kappa}{2}\boldsymbol{\beta}^T\boldsymbol{\beta})
\end{aligned}
$$

B Methodology

We define our model as a regression model on top of a AR(P) model.

B.1 Regression (Observation) Model

After EDA, we determines to use response as `log.price`: the logarithm of the house deal price. The covariate predictors are `beds`, `sold.dat`, `baths`, `square.feet`, `lot.size`, `house.age`, `property.type`, `room.Diff`, `dist.Duke`, and finally `room.Diff` interact with `property.type`. The regression model is

$$
\begin{aligned}
y_t^{(i)} = \beta_{\text{beds}}\ \mathbf{beds}_t^{(i)} + \beta_{\text{sold date}}\ \mathbf{sold\ date}_t^{(i)} + \beta_{\text{beds}}\ \mathbf{baths}_t^{(i)} + \beta_{\text{square feet}}\ \mathbf{square\ feet}_t^{(i)} + & \quad (10) \\
\beta_{\text{lot size}}\ \mathbf{lot\ size}_t^{(i)} + \beta_{\text{house age}}\ \mathbf{house\ age}_t^{(i)} + \beta_{\text{property type}}\ \mathbf{property\ type}_t^{(i)} + & \quad (11) \\
\beta_{\text{room difference}}\ \mathbf{room\ Diff}_t^{(i)} + \beta_{\text{dist. to Duke}}\ \mathbf{dist.\ to\ Duke}_t^{(i)} + & \quad (12) \\
\beta_{\text{interaction}}\ \mathbf{room\ Diff}_t^{(i)} \times \mathbf{property\ type}_t^{(i)} + & \quad (13) \\
\alpha_t + \nu_t & \quad (14) \\
\nu_t \sim \mathcal{N}(0, v) & \quad (15)
\end{aligned}
$$

Where $y_t^{(i)}$ is the response of the i^{th} house sold on the t^{th} window date, which is its logarithm of deal price. (notice that suppose for each window $t \in \{1, 2, 3, T\}$, there are n_t sold houses in the t^{th} window. Then n_t need not equal for all t). The rests are predictive variables. α_t is an time varying intercept which will be modeled by the AR(P) model described in the next session. ν_t is an additional observation uncertainty. To simply our modeling process, we take ν_t to have constant variance. Also, to simplify our notation, we write vectorized equation by merging line $(1), (2), (3), (4)$. The compact form is denoted as

$$
\begin{aligned}
\boldsymbol{y}_t &= \alpha_t \mathbf{1}_{n_t} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\nu}_t \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I}_{n_t})
\end{aligned}
$$

B.2 Time Series Model

We construct the AR(p) model to model the underlying intercept α_t as described above in the regression model. As we've already indicated in EDA section, we'll choose $p = 7$ for our

$$
\begin{aligned}
\alpha_t &= \sum_{i=1}^{p} \theta_i \alpha_{t-i} + \omega_t \\
\omega_t &\sim \mathcal{N}(0, w)
\end{aligned}
$$

However, the above parametrization requires us to take-in many timesteps value to predict α_t, we may simply it by vectorizing the expression into the following:

$$
\boldsymbol{\alpha}_t = \begin{bmatrix} \alpha_t \\ \alpha_{t-1} \\ \alpha_{t-2} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \dots & \theta_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \alpha_{t-2} \\ \alpha_{t-3} \\ \vdots \\ \alpha_{t-p} \end{bmatrix} = \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1}
$$

In this way, transition becomes easy, as dependency rely on only the past one timestep.

B.3 Combined Model

We end up with the model as the following

$$
\begin{aligned}
\boldsymbol{\alpha}_t &= \boldsymbol{\Theta}\boldsymbol{\alpha}_{t-1} + \boldsymbol{W}_t \\
\boldsymbol{y}_t &= \mathbf{1}\boldsymbol{\alpha}_t + \beta \boldsymbol{X}_t + \boldsymbol{\nu}_t \\
\boldsymbol{W}_t &\sim \mathcal{N}(\mathbf{0}, w\boldsymbol{I}) \\
\boldsymbol{\nu}_t &\sim \mathcal{N}(\mathbf{0}, v\boldsymbol{I}) \\
\mathbf{1} &\coloneqq \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \\
\boldsymbol{\alpha}_0 &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\
\boldsymbol{\beta} \mid (\tau = v^{-1}), \kappa &\sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{I}(\tau\kappa)^{-1}\right) \\
p((\tau = v^{-1}) \mid \kappa) &\propto 1/\tau \\
\kappa &\sim \boldsymbol{G}(1/2, 1/2) \\
\boldsymbol{\theta} | (\phi = w^{-1}), \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}/\phi\right) = \mathcal{N}((\frac{1}{p}, \frac{1}{p}, \cdots, \frac{1}{p})^T, \frac{1}{p}\phi^{-1}\mathbf{I}) \\
(\phi = w^{-1}) | \mathcal{D}_T, \boldsymbol{\beta}, v &\sim \mathbf{G}\left(a_0 = \frac{v_0}{2}, b_0 = \frac{v_0 s_0^2}{2}\right) = \mathbf{G}\left(\frac{1}{2}, \frac{1}{2}\right)
\end{aligned}
$$

C Current Results

C.1 Answer to question 1

Table 3: Estimate for Beta

	No Trend Model			Linear Trend Model		
	mean	0.025%	0.975%	mean	0.025%	0.975%
baths	0.2163182	0.2023575	0.2298846	0.2167585	0.2033552	0.2308176
beds	-0.1336671	-0.1472986	-0.1198309	-0.1340896	-0.1481884	-0.1207154
dist.duke	-0.0000148	-0.0000148	-0.0000148	-0.0000148	-0.0000148	-0.0000148
house.age	-0.0061420	-0.0061518	-0.0061320	-0.0061436	-0.0061528	-0.0061347
itr_single	-0.5499421	-0.5528564	-0.5471058	-0.5501222	-0.5528907	-0.5471864
itr_townhouse	-0.5992212	-0.6021851	-0.5962845	-0.5993334	-0.6021205	-0.5962204
lot.size	0.0000019	0.0000019	0.0000019	0.0000019	0.0000019	0.0000019
room.Diff	0.4150515	0.4012283	0.4291071	0.4147748	0.4002657	0.4284451
single	0.1953111	0.1935420	0.1970604	0.1951812	0.1934433	0.1970871
square.feet	0.0002385	0.0002383	0.0002387	0.0002385	0.0002383	0.0002387
townhouse	0.0973085	0.0955664	0.0990432	0.0971897	0.0953844	0.0990924
sold.dat1	NaN	NaN	NaN	0.0001388	0.0001350	0.0001433

- Holding all else constant, including time. If a house's number of beds increases by 1, the house price's median is expected to decrease to $e^{-0.134}$ times the original price, which is 87.459 percent of the original price.
- Holding all else constant, including time. If a house's number of bathrooms increases by 1, the house price's median is expected to increase to $e^{0.216}$ times the original price, which is 124.110 percent of the original price.
- Holding all else constant, including time. If a house's total square feet increase by 1, the house price's median is expected to increase to $e^{0.0002385}$ of the original price, which is 100.023 percent of the original price.
- Holding all else constant, including time. If a house's lot size increases by 1 square foot, the house price's median is expected to increase to $e^{0.0000019}$ of the original price, which is 100.00019 percent of the original price.
- Holding all else constant, including time. If a house's age increases by 1 year, the house price's median is expected to decrease to $e^{-0.000614}$ times the original price, which is 99.938 percent of the original price.
- Holding all else constant, including time. Comparing to a Condo, If a house is a Single Family Residential, the house price's median is expected to increase to $e^{0.195}$ times the original price, which is 121.531 percent of the original price.
- Holding all else constant, including time. Comparing to a Condo, If a house is a Townhouse, the house price's median is expected to increase to $e^{0.0971}$ times the original price, which is 110.197 percent of the original price.
- Holding all else constant, including time. If a house's distance to Duke increases by 1 meter, the house price's median is expected to decrease to $e^{-0.0000148}$ of the original price, which is 99.998 percent of the original price.

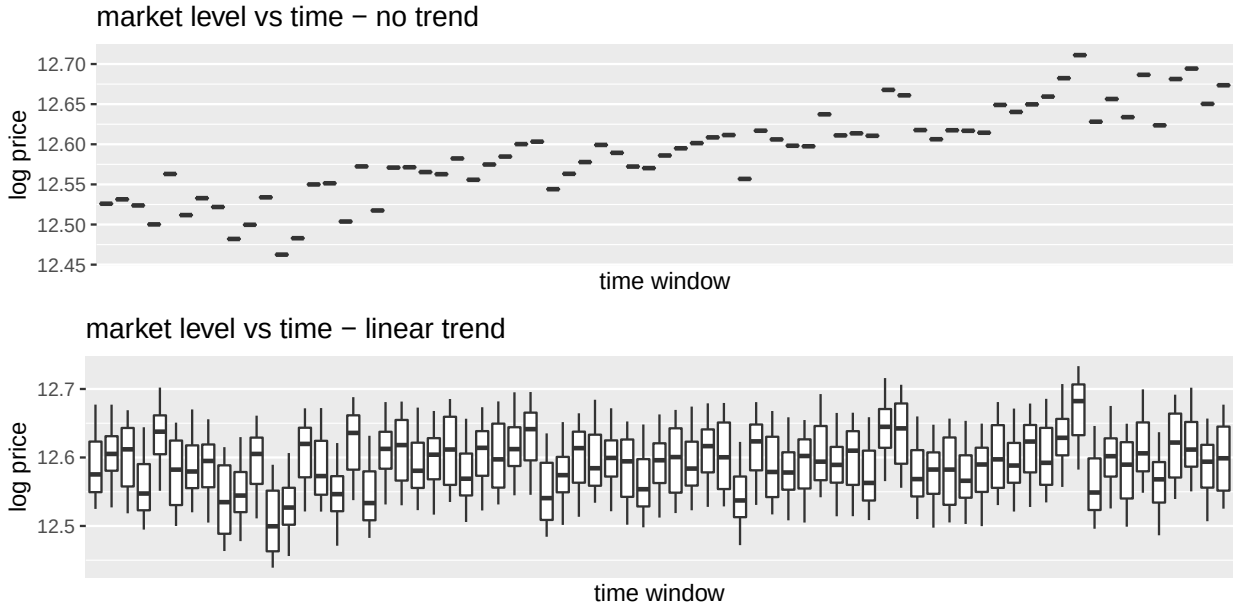
- Holding all else constant, including time. If a house is a Townhouse, having 1 even more bed than bathrooms will make the house price median to decrease to $e^{0.414-0.599} = e^{-0.185}$, which is 83.110 percent of the original price.
- Holding all else constant, including time. If a house is a Single Family Residential, having 1 even more bed than bathrooms will make the house price median to decrease to $e^{0.414-0.550} = e^{-0.136}$ of the original price, which is 87.284 percent of the original price.
- For the same house, regardless of market periodicity, we expect the house to be sold at $e^{\cdot 0001387} = 100.0139\%$ its price the day before. To provide a better idea of annual house market inflation, we expect the same house to be sold 5.17% more expensive than in the last year.

C.2 Answer to question 2

Table 4: Linear Trend Model Periodicity with Moduli (Magnitude)

modulus	periods	period_in_day
1.2789166	4.130674	61.96012
1.0961516	8.831979	132.47969
1.0737059	27.547151	413.20726
0.9499517	6.648033	99.72049
0.6148616	5.246420	78.69630

C.3 Answer to question 3



C.4 Answer to question 4

Table 5: 1 step prediction MAE for 2 models. First row is in-sample error, and second row is out-of-sample error

no trend	linear trend
501.8759	81.77021
996.4305	110.60610

D Model Validation Figures

D.1 No Trend Model

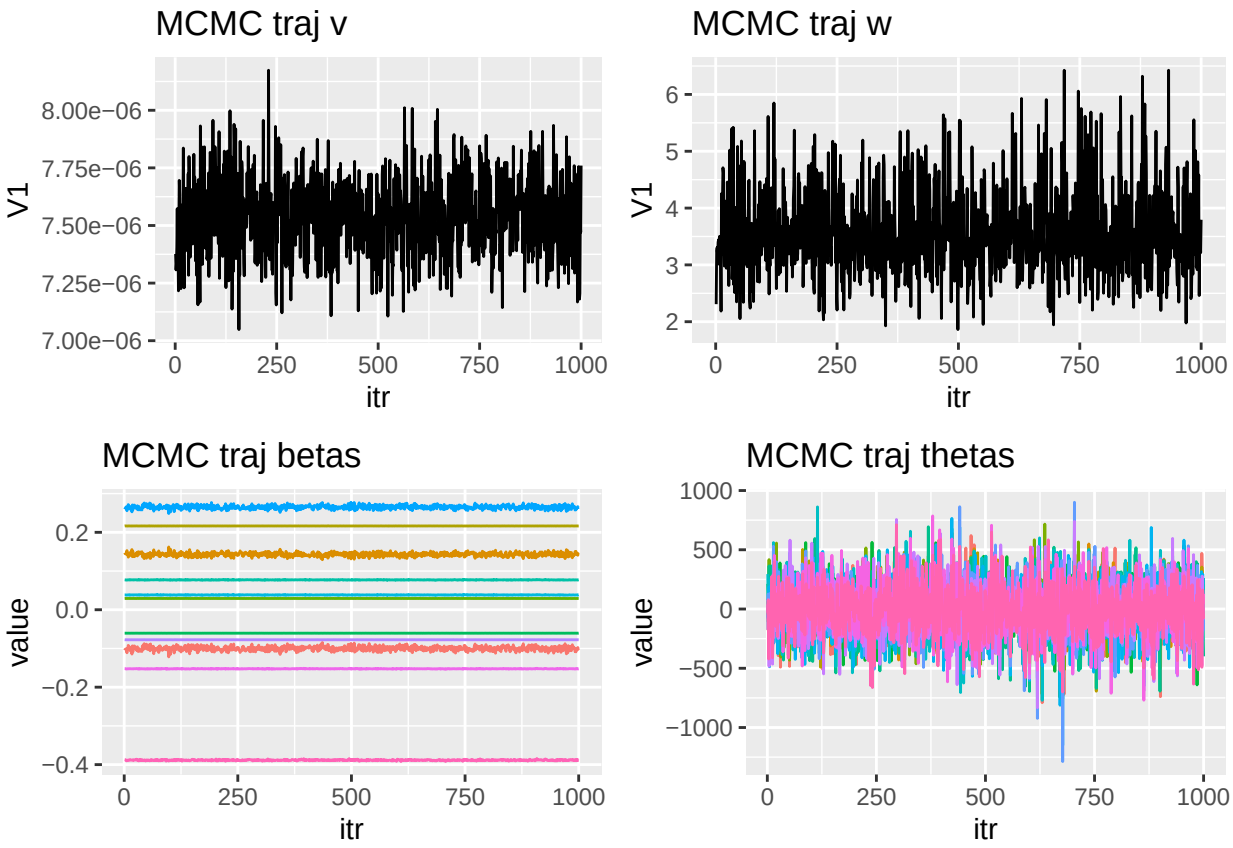


Figure 2: MCMC trace plot no trend model

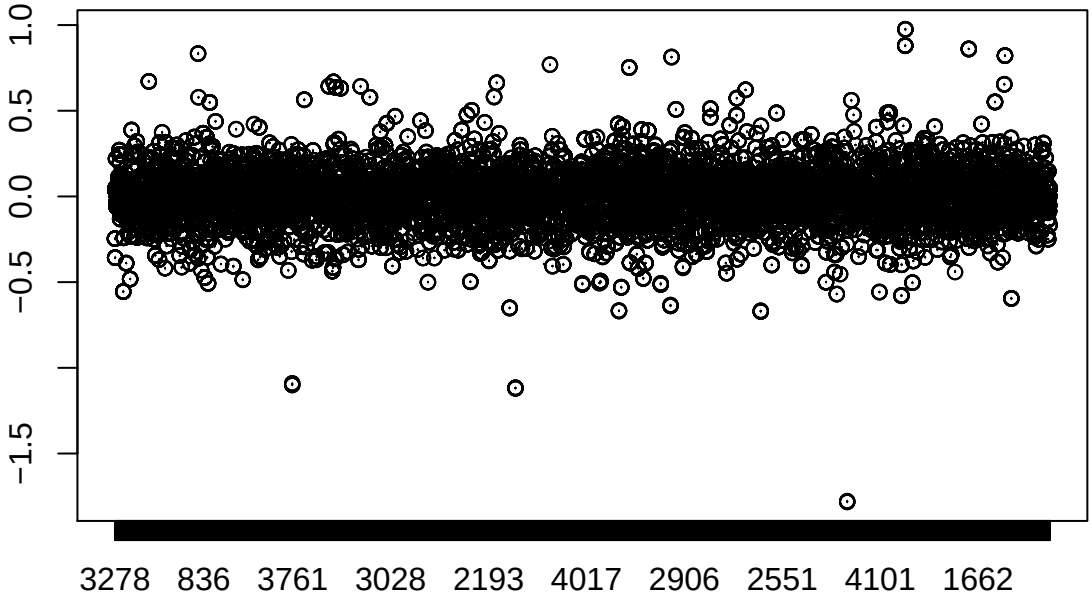


Figure 3: boxplot residual for no trend model

D.2 Linear Trend Model

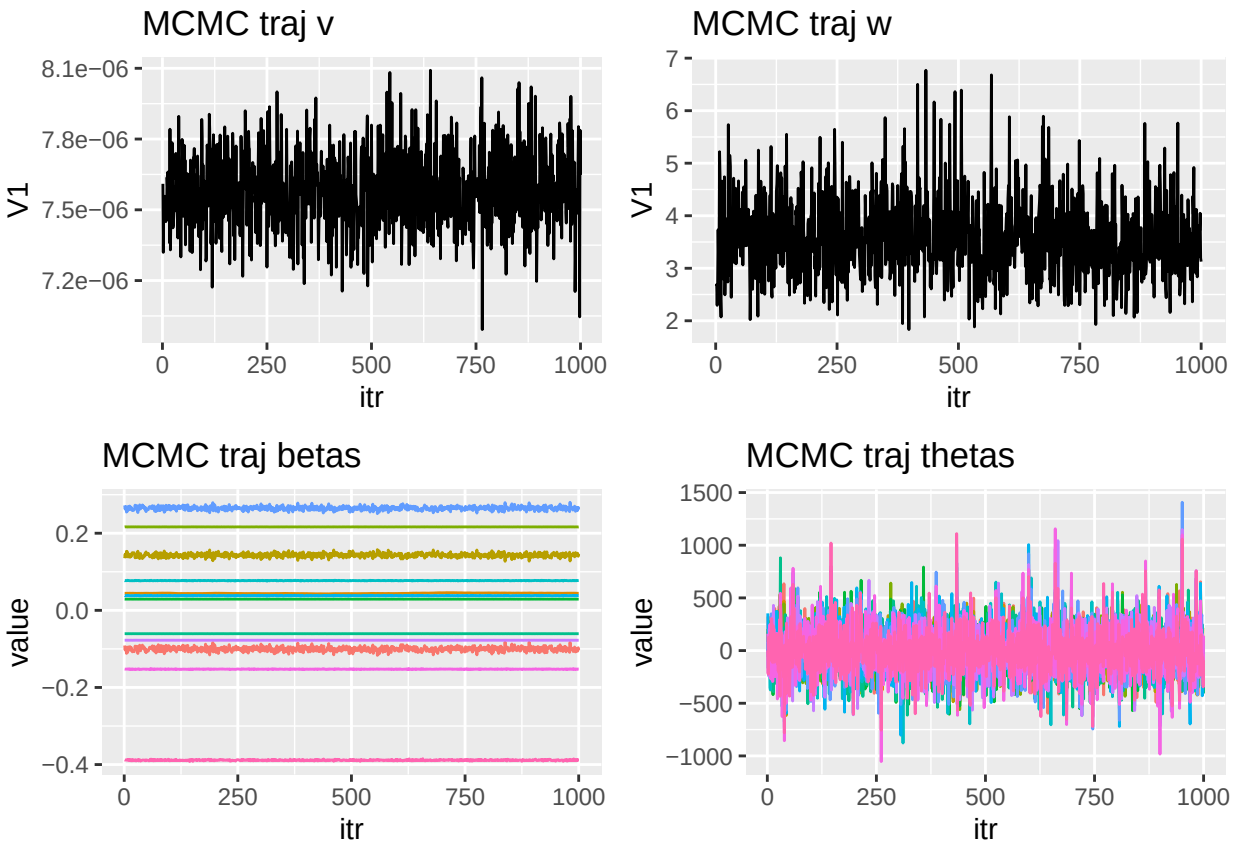


Figure 4: MCMC trace plot linear trend model

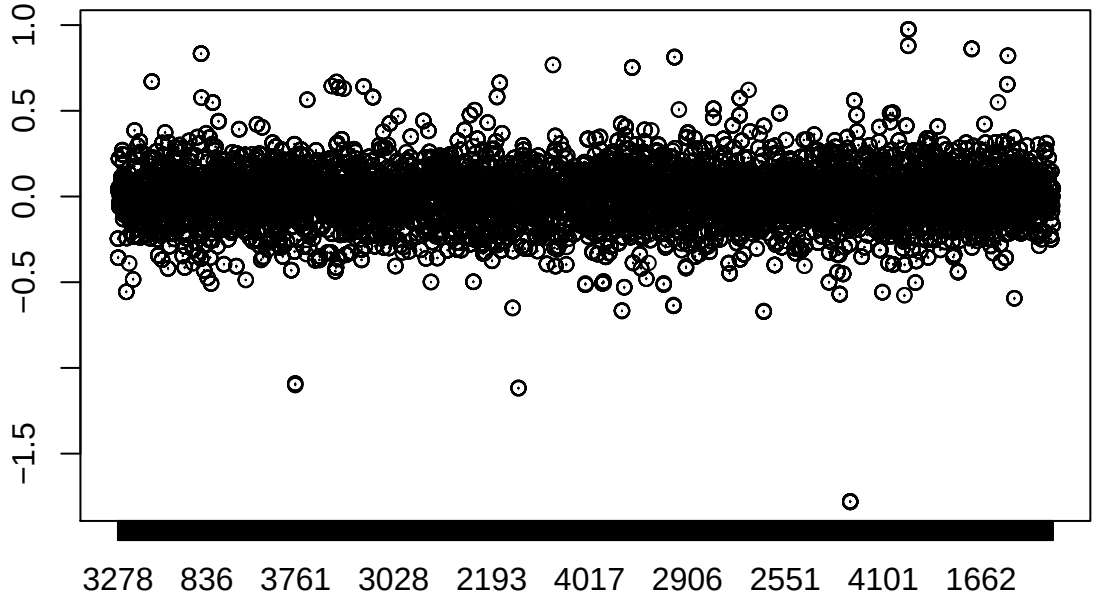


Figure 5: boxplot residual for linear trend model

E Random Figures

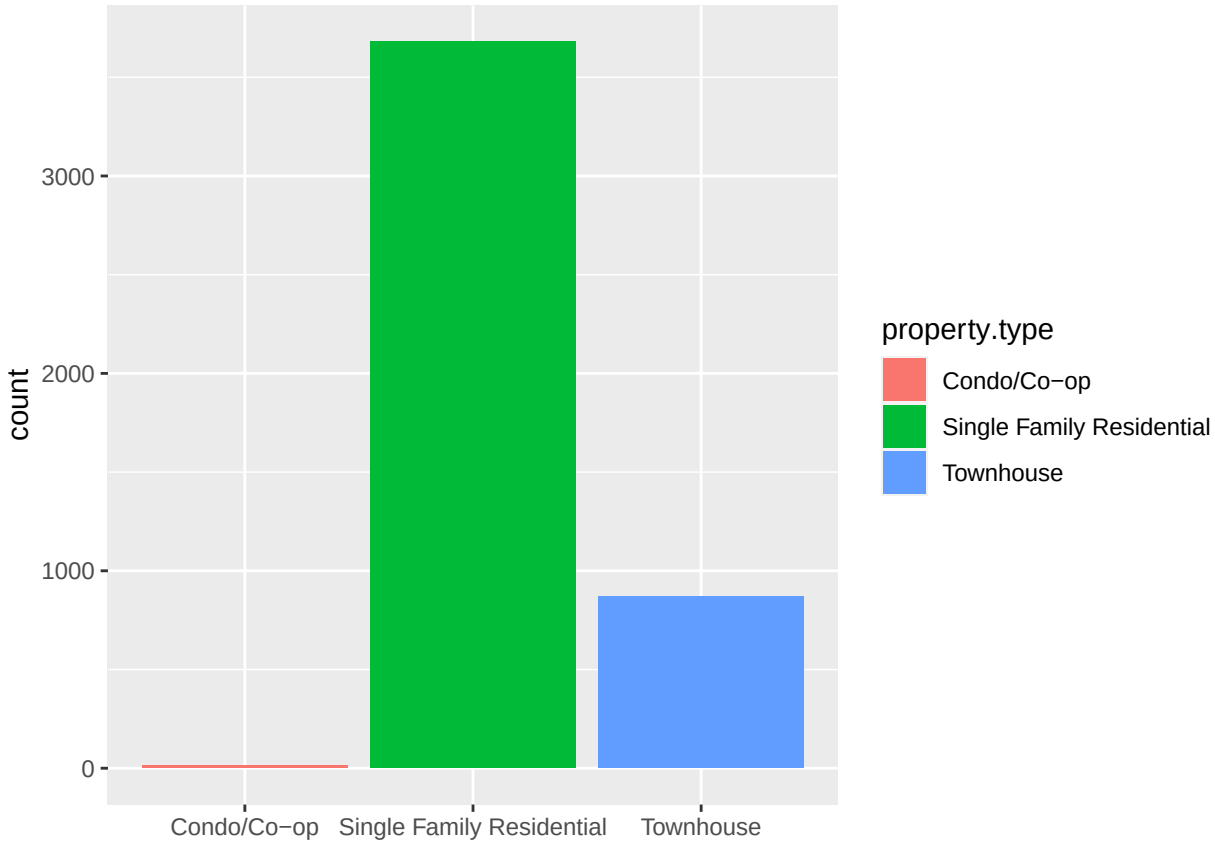


Figure 6: uneven distribution of classes

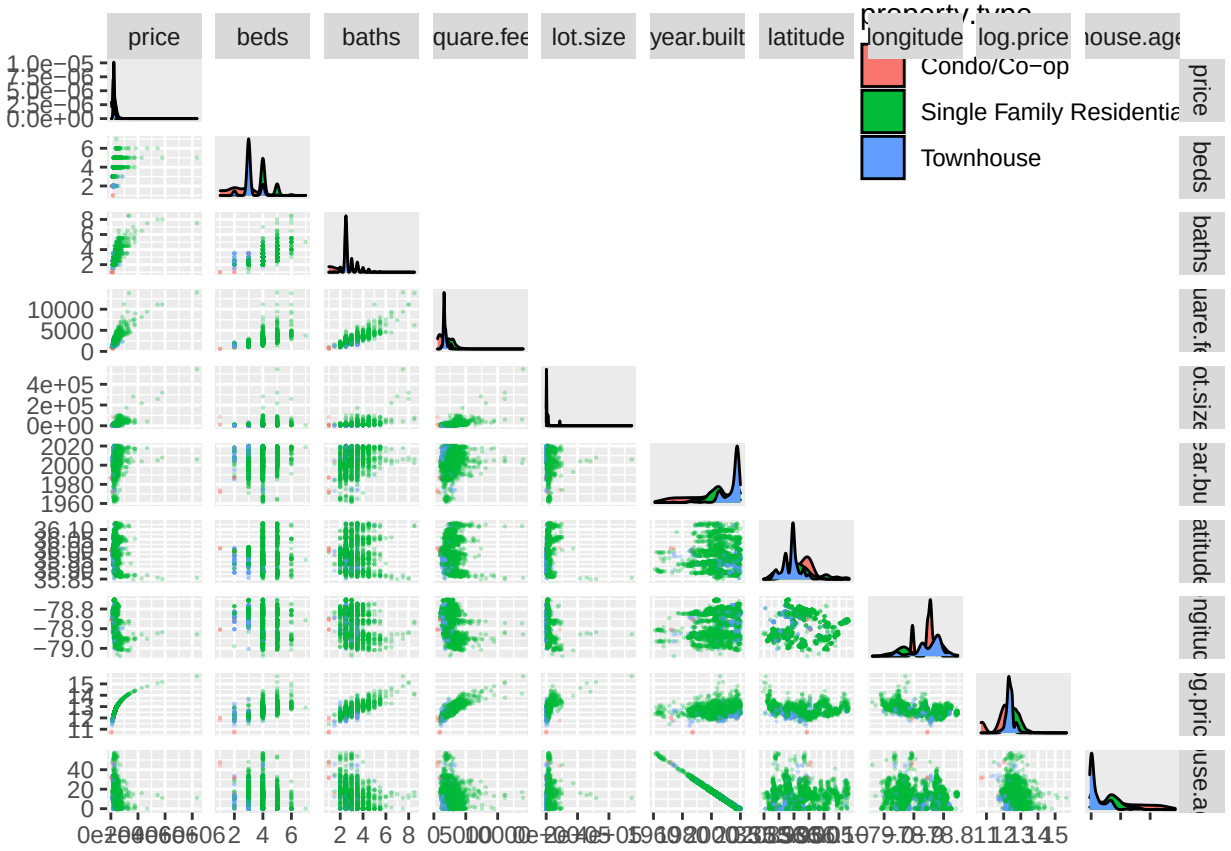


Figure 7: pair plot

F Algorithm

Algorithm 2: parameter inference algorithm

Result: Sampling distribution for $\{w, v, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\alpha}_{1:T}\}$
Initialize: $\mathbb{P}((\theta_1, \ldots, \theta_p)^\intercal), \mathbb{P}(w), \mathbb{P}(v)$ via pre-set prior distribution
while *not converged* **do**
  Calculate: posterior mean and covariance for $\boldsymbol{\alpha}_t|\mathcal{D}_t$: $m_t, C_t \ \forall t \in 1:T$ via forward filtering algorithm in Appendix
  Sample: $\boldsymbol{\alpha}_t|\mathcal{D}_T$ from $m_t^*, C_t^*, \forall t \in 1:T$ by backward smoothing
  Sample: $\boldsymbol{\theta}, (\phi = w^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}$ by first sampling $(\phi = w^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}$ and then $\boldsymbol{\theta}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\beta}, \phi$
  Sample: $\boldsymbol{\beta}, (\tau = v^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\theta}$ by first sampling $(\tau = v^{-1})|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\theta}$ and then $\boldsymbol{\beta}|\boldsymbol{X}, \mathcal{D}_T, \boldsymbol{\theta}, \tau$
end
Return: samples of $w, v, \boldsymbol{\beta}, \boldsymbol{\theta}_{1:T}$

F.1 Answering Research Question

To answer question 1) – how non-temporal descriptive variables affect housing prices, as we've conducted MCMC sampling, we've obtain a distribution of parameter $\boldsymbol{\beta}$. Interpretation on how $\boldsymbol{\beta}$ impacts the house price is identical to the interpretation in a basic Bayesian linear regression. We'll provide point and interval estimate for all the βs.

To address question 2) – what periodic temporal effects is presented in the past house market, we utilize eigen-decomposition of $\boldsymbol{\Theta}$ matrix to shed light on periodicity of house market price. Notice that the MCMC sampler above returns a series of $\boldsymbol{\theta}$ samples. Thus, we've obtained the posterior distribution of $\boldsymbol{\theta}$ without a point estimate. To obtain the point estimate, we can make a Bayes estimator for $\boldsymbol{\theta}$. As we aim to find an estimator $\hat{\boldsymbol{\theta}}$ that minimizes the posterior mean absolute error, we simply take the mean of all $\boldsymbol{\theta}$ samples to obtain this estimator $\hat{\boldsymbol{\theta}}$. Thus, a reasonable point estimate for $\boldsymbol{\theta}$ can be constructed. Through transformation, we can get point estimate of $\boldsymbol{\Theta}$ matrix. Therefore we can perform eigen-decomposition on this point estimate.

Once we extract eigenvalues together with its eigenvectors, we observe several pairs of complex eigenvalues, which generate the periodicity. For example, suppose we observe a pair of eigenvalues $a \pm bi$, this correspond to the α_t, the real estate market, having a sine wave cycle of $Period = \frac{2\pi}{\arcsin(b/\sqrt{a^2+b^2})}$ measured in timestep, due to Euler's formula. To re-measure the perdiod in days, simply times the window size, which is 15, with $Period$. Therefore, we obtain the cycle effect. As there could be multiple cycle effects, each of their contributions are measured in magnitude of the sine wave, which can be calculated by the eigenvalue's modulus $\sqrt{a^2 + b^2}$

To address question 3) – how to extract the past real estate market α_t. In our MCMC sampler, we have also backward sampled the α_t trajectory. Thus, this task can be addressed by taking mean (expectation) and quantiles (uncertainty) of all the α_t samples from MCMC.

To address question 4) – how to forecast short term house prices, we're calcuating closed form distribution of α_T in the MCMC iteration. To forecast, linear transform this sampled closed form distribution via the sampled $\boldsymbol{\Theta}$ in that MCMC iteration to get closed form distribution of y_{T+1}, for example. Sample the y_{T+1} from this closed form distribution, transform it back to normal price using exponent, and record it. After all iterations are over, take the sample mean of all the recorded samples (which is the predictive posterior distribution). We can obtain the predictive posterior expectation of house prices at time $T + 1$. This is our prediction. Model performance (MAE) will be calculated from this mean with real data.

{Last Revised: November 12, 2020}