Analyzing Regional and National House Price Trends Over Time: A Bayesian Hierarchical Modeling Approach

Serena E. Alvarez
sealvare@uci.edu

Saanvi Shankar
saanvis2@uci.edu

Abstract

This project develops a Bayesian hierarchical model to explore how regional house price trends (Midwest, Northeast, South, West) relate to national trends over time in the U.S., adjusting for inflation for Median and Average House Prices. This analysis provides insights into the relationships and dynamics between national and regional house prices on a quarterly basis.

1 Introduction

The price of housing in America is a significant indicator of the state of the economy, giving us a good representation of economic health. While national housing trends tend to catch the most attention, regional housing prices may offer more nuanced insights. This project aims to investigate how regional price trends (Midwest, Northeast, South, West) relate to national price trends when adjusting for inflation using a Bayesian hierarchical model.

A Bayesian Hierarchical model works well for this type of nested data structure, where regional prices being nested in national prices. This approach will account for region-specific differences in prices while still keeping the national average in mind, which will provide a deeper inference.

The dataset consists of historical quarterly data on average and median house prices, adjusted by the Consumer Price Index (CPI), for the United States as a nation and for four distinct regions: Midwest, Northeast, South, and West. This dataset allows us to provide insight into how national market conditions affect regional market conditions and vice versa.

Through this paper, we will conduct our modeling approach and hope to uncover underlying relationships between regional and national housing prices trends. We hope that our findings can contribute to a better understanding of housing prices being used as an economic indicator.

2 Data Description

We sourced the data for this project from the Federal Reserve Economic Data (FRED), which provides very organized, clean, and reliable datasets. The specific data series we pulled from the site are:

1. Average Sales Price and Median Sales Price of Houses Sold for the Midwest, Northeast, South, West Census Regions, and the United States as a whole.[1-10]
2. Consumer Price Index: All Items for the United States.[11]

The data is complete for all regions beginning in 1975, so we filtered our data to begin on 07-01-1975. We then combined these all and CPI into one 195×11 matrix with 11 columns:

- `DATE`: The date of the recorded data in quarterly intervals from 07-01-1975 to 01-01-2024.

Preprint. Under review.

- ASPMW: Average Sales Price of Houses Sold for the Midwest.
- ASPNE: Average Sales Price of Houses Sold for the Northeast.
- ASPS: Average Sales Price of Houses Sold for the South.
- ASPW: Average Sales Price of Houses Sold for the West.
- ASPUS: Average Sales Price of Houses Sold for the United States.
- MSPMW: Median Sales Price of Houses Sold for the Midwest.
- MSPNE: Median Sales Price of Houses Sold for the Northeast.
- MSPS: Median Sales Price of Houses Sold for the South.
- MSPW: Median Sales Price of Houses Sold for the West.
- MSPUS: Median Sales Price of Houses Sold for the United States.
- CPI: Consumer Price Index for the United States.

To adjust for inflation, we multiplied each price cell by the CPI value corresponding to its date. This is the common way CPI is used and this adjustment was crucial to ensure comparability of prices over time by accounting for inflation. After that, we removed any unnecessary columns and completed the dataset we will be using for the project.

3 Data Preparation

We went through several steps to prepare our data for model fitting and exploratory analysis. The steps included creating time variables, setting the correct predictor column types, and transforming the data into a long format.

Below is a detailed description of our process for each step:

Creating Time Variables Time Variables were necessary in order to capture the effect of time on the response variable. These variables were created by extracting the year and quarter from the date column and then combining into a continuous time variable that reflects both the year and quarter. Specifically, the year was extracted from the date, and the quarter was calculated by dividing the month by three and rounding up to the nearest integer.The continuous time variable was thus:

$$\text{Time} = \text{Year} + \frac{(\text{Quarter} - 1)}{4}$$

This calculation ensures that the Time variable reflects both the year and the specific quarter within that year, allowing us to model the data accurately using the quarters.

Correct column types We made sure the columns of Region, and Price_Type are set to categorical variables (factors) to ensure correct modeling later on.

Transforming Data to Long Format We first loaded the data and ensured it was in correct format. We adjusted the Date column to be formatted in the appropriate 'Date' type which allows for time-based operations to be preformed.

Then, we transformed the data from wide to long format. By reshaping the data into a long format, we allow each row to represent one observation, and thus we allow for a more flexible and more efficient analysis. We pivot the data, creating new columns of type of price and region, and finally remove the extra columns. These are the new columns for the long format:

- Price: The actual sales price.
- Region: The geographic region (Midwest, Northeast, South, West).
- Price_Type: Indicates whether the price is an average or median sales price.
- Time: A continuous variable that reflects both the year and quarter.
- National_Price: The national average or median sales price corresponding to the date.

Long Format is important when preforming hierarchical modeling and allows for the analysis of the nested structures.

We have now cleaned and prepared our data to be ready for the model fitting.

4 Exploratory Data Analysis

We conducted a preliminary analysis on the data to understand its structure and its quirks. We include visualizations as they are particularly useful in understanding the differences between each region.

Summary Statistics We examine the summary statistics for our dataset to understand the central tendency and variation within the data (Table 1).

Regional and National House Prices This plot (Figure 1) was created too overlay the national average on the region prices to see how well they line up and give a clear visual comparison. We can see that the national and regional prices align very closely, indicating a high correlation between national and regional house price trends.

Distribution of House Prices by Region We generated a histogram to visually summarize how the housing prices are distributed along region (Figure 2). As we can see, all regions are positively skewed. The South shows a higher variability and could be due to a mix of urban and rural housing options which differ greatly.

Correlation This plot (Figure 3) shows the relationship between regions and national prices with scatter plot for each region of region prices vs. national prices. Again, we can see the strong positive correlation between the two prices. Intuitively this makes sense since national pricing trends affect regional pricing trends.

5 Model Specification

In this section, we specify the Bayesian hierarchical model using the `rstanarm` package in R.

5.1 Model Description

The Bayesian hierarchical model is specified with regional prices nested within national prices. The model includes both fixed effects and random effects which can account for both fixed and random variations. The following are the model components:

5.1.1 Fixed Effects

The fixed effects in the model are:

- `Time`: Captures the dynamics of house prices over time.
- `Region`: Represents the geographic regions (Midwest, Northeast, South, West).
- `Price_Type`: Indicates whether the price is an average or median sales price.
- `National_Price`: The national average or median sales price corresponding to the date.

The interaction terms between `Time`, `Region`, and `Price_Type`, as well as between `National_Price` and `Region`, are added to capture the effect of all of these terms being included in the model together.

5.1.2 Random Effects

The random effects structure allows for random intercepts and slopes for each region, capturing the variability within and between regions. In our model (due to time constraints), we only include a variable for random intercept depending on region:$(1|\text{Region})$

5.2 Priors

We give general priors to stabilize the model:

- fixed effects $N(0, 1)$, auto-scaled for the fixed effects.

- intercept $N(0, 1)$, auto-scaled for the intercept.

These priors are relatively non-informative, allowing the data to primarily drive the inference.

5.3 Likelihood

The likelihood function is specified as Gaussian (normal), which is appropriate given that the house price data is continuous: family = gaussian()

5.4 Hyperparameters

The model includes several hyperparameters to ensure proper convergence and efficient sampling:

- `chains= 4`: Four Markov Chain Monte Carlo (MCMC) chains to run.
- `iterations= 1000`:One thousand iterations per chain (extremely time consuming to do 2000)
- `adapt_delta= 0.95, max_treedepth= 15`: Control parameters to improve sampling efficiency and convergence

6 Model Diagnostics

To evaluate our model, we will preform several diagnostics checks including posterior predictive checks, Rhat values, and credible intervals.

Posterior Predictive Checks Posterior predictive checks (PPCs) are used to check the model's fit to the observed data. The PPC plot (Figure 4) compares the distribution of the observed data (y) with the distribution of the data generated by the model (y_{rep}). We can see that distributions of the observed and replicated data are closely aligned, indicating that the model captures the underlying structure of the data to a good enough degree.

Rhat Values The Rhat statistic, or the potential scale reduction factor, measures the convergence of the MCMC chains, with a Rhat value close to 1 indicating good convergence. Our Rhat values (Table 2) for all parameters are close to 1, giving us evidence that the chains have converged well and the model's estimates can be trusted to be reliable. We also notice that no Rhat value is above 1.1 and thus the chains are mixed 'well-enough'.

Credible Intervals Credible intervals provide a range of possible values for the model parameters that assist in our understanding of the uncertainty associated with the parameter estimates. We see this referenced in Table 3, gone into more details in the next section.

7 Interpretation, Results, and Conclusion

The main objective of this analysis is to understand the relationship between regional housing prices and national housing prices over time, while accounting for factors such as region, price type, and their interactions terms. We will go over key finding in this section.

7.1 Model Summary Interpretations

Table 3 provides a summary of the model's fixed and random effects.

The following is an interpretation of the summary for the key parameters:

- **Time:** The positive posterior mean suggests that house prices have increased over time. The credible interval does not include zero, indicating a significant positive predictor.
- **Region:**
  - **Baseline Region: Midwest**: The Midwest is our baseline region against which all other regions are compared. Its effect is captured in the intercept term.

  - **Northeast:** The negative estimate for the Northeast region indicates that house prices are on average lower compared to the Midwest region.
  - **South:** The positive estimate for the South region indicates higher prices compared to the Midwest region.
  - **West:** The wide credible interval for the West region tells us that there is a large posterior standard error. However, 0 is included in the model indicating that this is not a strong predictor.
- **Price_Type:** The positive estimate for `Price_TypeMedian` suggests that median prices tend to be higher than average prices.
- **National_Price:** The strong positive estimate for `National_Price` indicates a close relationship between national and regional house prices, with regional prices closely mirroring national trends. This is known and intuitive.
- **Interaction Terms:** The interaction terms reveal how the effects of time and national prices are different across regions and price types. For example, the `Time:RegionNortheast` interaction suggests that the rate of price change over time varies for the Northeast region when comparing to the Midwest region.

7.2 Result Analysis

Our analysis reveals that `Time` and `National_Price` are significant predictors of house prices. The strong positive associations indicate an overall increase in house prices over time and a very close relationship between national and regional trends.

However regional differences do exist. Specifically, the `RegionNortheast` posterior estimate shows significantly lower prices compared to the `RegionMidwest`; while the `RegionSouth` and `RegionWest` regions show variability that is not statistically significant.

The interaction terms further prove the influence of national prices varies across regions, with the South region's house prices the closest to mirroring national trends.

7.3 Conclusions

Our analysis successfully demonstrates the interplay between regional and national house prices over time, emphasizing the importance of using statistical approaches like Bayesian hierarchical modeling for economic indicators. Our results confirm significant variability influenced by regional and economic factors, suggesting that economists should consider both regional and national indicators in their analysis and interpretation.

7.4 Further Research

Future research may include understanding and modeling for the huge increase in housing prices across all regions beginning in 2020. It would be interesting to note it any particular region experienced it harder than another or unintentionally started the huge housing price increase trend we see today.

8 References, Figures, and Tables

References

1. Federal Reserve Economic Data (FRED), Average Sales Price of Houses Sold for the Midwest Census Region, Quarterly, Not Seasonally Adjusted.
2. Federal Reserve Economic Data (FRED), Average Sales Price of Houses Sold for the Northeast Census Region, Quarterly, Not Seasonally Adjusted.
3. Federal Reserve Economic Data (FRED), Average Sales Price of Houses Sold for the South Census Region, Quarterly, Not Seasonally Adjusted.
4. Federal Reserve Economic Data (FRED), Average Sales Price of Houses Sold for the United States, Quarterly, Not Seasonally Adjusted.

5. Federal Reserve Economic Data (FRED), Average Sales Price of Houses Sold for the West Census Region, Quarterly, Not Seasonally Adjusted.
6. Federal Reserve Economic Data (FRED), Median Sales Price of Houses Sold for the Midwest Census Region, Quarterly, Not Seasonally Adjusted.
7. Federal Reserve Economic Data (FRED), Median Sales Price of Houses Sold for the Northeast Census Region, Quarterly, Not Seasonally Adjusted.
8. Federal Reserve Economic Data (FRED), Median Sales Price of Houses Sold for the South Census Region, Quarterly, Not Seasonally Adjusted.
9. Federal Reserve Economic Data (FRED), Median Sales Price of Houses Sold for the United States, Quarterly, Not Seasonally Adjusted.
10. Federal Reserve Economic Data (FRED), Median Sales Price of Houses Sold for the West Census Region, Quarterly, Not Seasonally Adjusted.
11. Federal Reserve Economic Data (FRED), Consumer Price Index: All Items: Total for United States (CPALTT01USM657N).

8.1 Figures, Tables

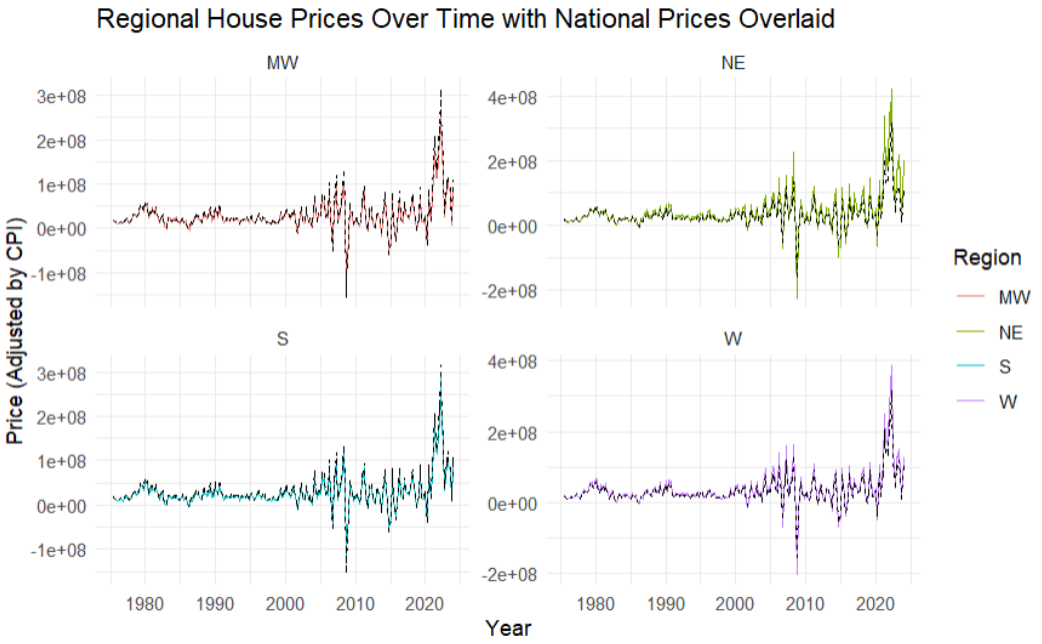

Figure 1: Regional House Prices Over Time with National Prices Overlaid

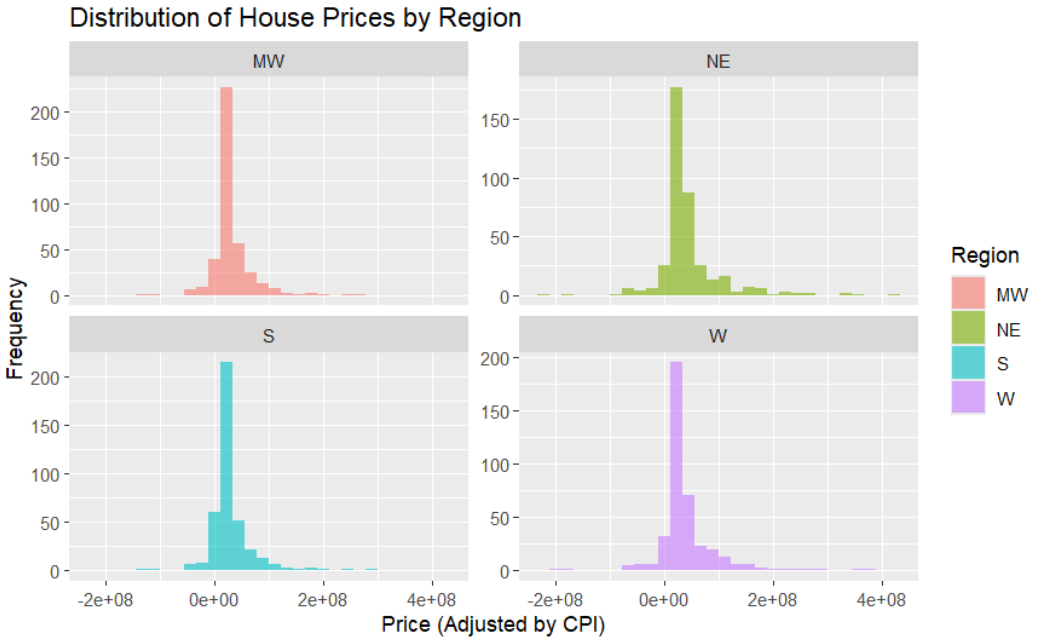

Figure 2: Distribution of House Prices by Region

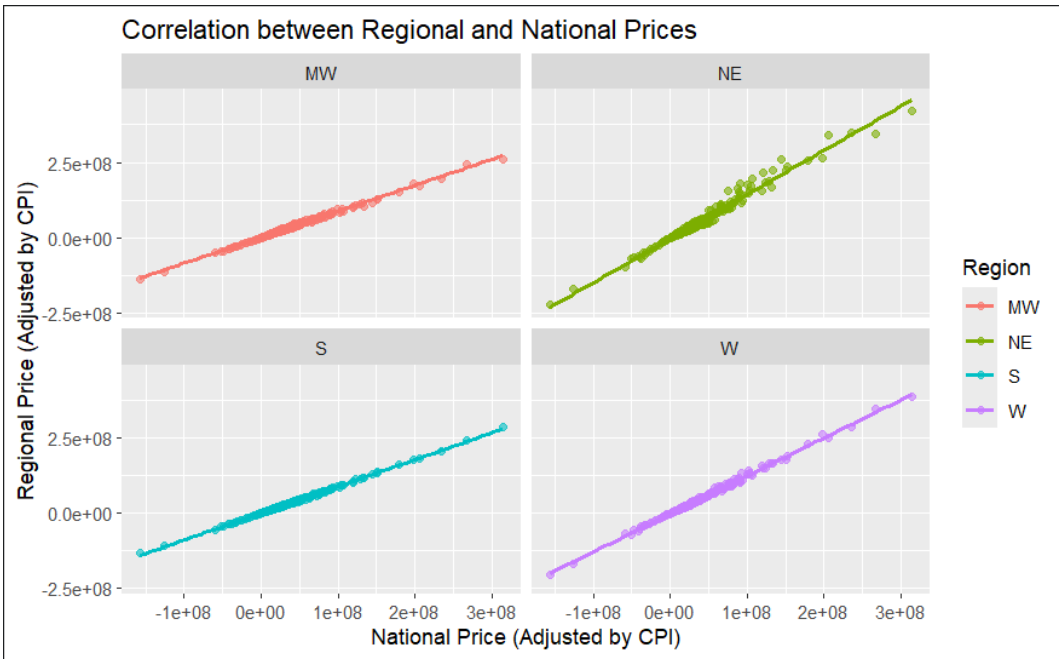

Figure 3: Correlation between Regional and National Prices

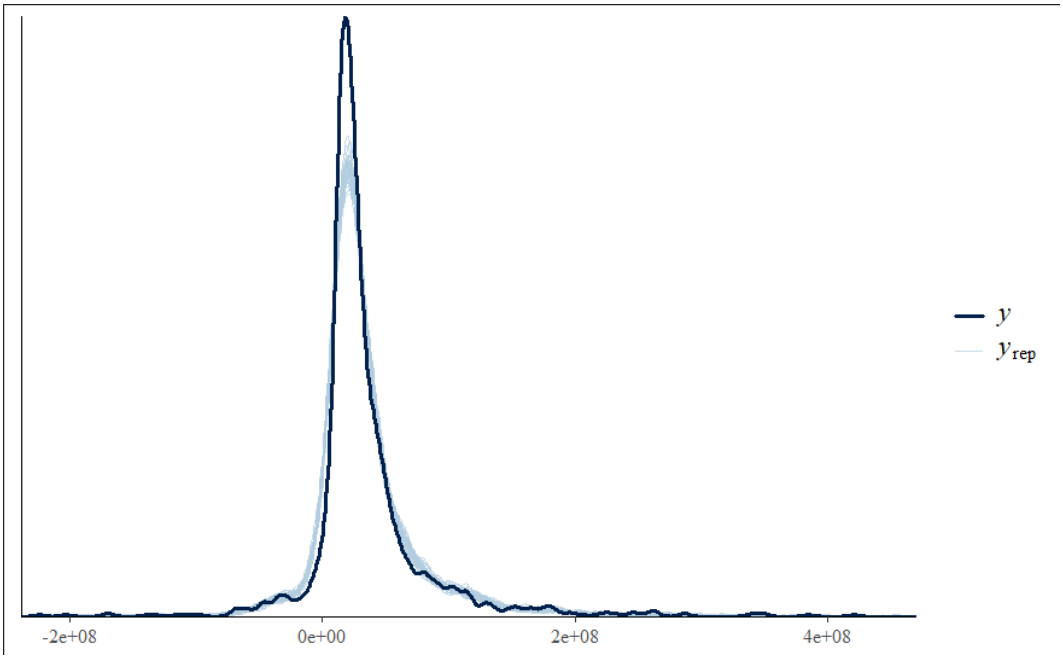

Figure 4: Posterior Predictive Checks

Statistic	Value
Price	
Min	-223,789,784
1st Qu.	15,097,122
Median	24,088,542
Mean	34,653,865
3rd Qu.	42,664,381
Max	421,025,598
Region	
Midwest	780
Northeast	780
South	780
West	780
National	0
Price_Type	
Average	1,560
Median	1,560
Time	
Min	1976
1st Qu.	1988
Median	2000
Mean	2000
3rd Qu.	2012
Max	2024
National_Price	
Min	-156,669,841
1st Qu.	14,360,341
Median	23,063,279
Mean	31,740,257
3rd Qu.	39,139,274
Max	314,131,546

Table 1: Summary Statistics

Parameter	Rhat
(Intercept)	1.0698862
Time	1.0063236
RegionNortheast	1.0515595
RegionSouth	1.0119696
RegionWest	1.0176474
Price_TypeMedian	1.0030157
National_Price	1.0000692
Time:RegionNortheast	1.0127974
Time:RegionSouth	1.0079018
Time:RegionWest	1.0034823
Time:Price_TypeMedian	1.0030157
RegionNortheast:Price_TypeMedian	1.0030157
RegionSouth:Price_TypeMedian	1.0036768
RegionWest:Price_TypeMedian	1.0099994
RegionNortheast:National_Price	1.0030157
RegionSouth:National_Price	1.0099984
RegionWest:National_Price	1.0030157
Time:RegionNortheast:Price_TypeMedian	1.0030157
Time:RegionSouth:Price_TypeMedian	1.0030157
Time:RegionWest:Price_TypeMedian	1.0030157
b[Intercept] Region:Midwest	1.0030157
b[Intercept] Region:Northeast	1.0030157
b[Intercept] Region:South	1.0030157
b[Intercept] Region:West	1.0030157
sigma	1.0054389

Table 2: Rhat Values for Key Parameters

Parameter	Estimate	Std. Error	2.5% CI	97.5% CI
(Intercept)	-2.096169e+08	7.251607e+07	-3.765529e+08	-9.676551e+07
Time	7.427872e+04	2.105630e+04	3.328713e+04	1.150234e+05
RegionNortheast	-4.821044e+08	1.134853e+08	-6.455479e+08	-2.215758e+08
RegionSouth	2.973680e+07	6.771198e+07	-7.743078e+06	6.771198e+07
RegionWest	-3.082204e+07	9.190735e+07	-1.526059e+08	1.190735e+07
Price_TypeMedian	2.226181e+07	3.029521e+07	1.369724e+07	3.029521e+07
National_Price	8.519425e-01	7.650520e-03	8.366944e-01	8.668397e-01

Table 3: Summary of Model Output