THE COST OF LIVING: A ZILLOW HOUSING FORECAST

PATRICK BEAL, TIARA EDDINGTON, MADELYN VINES

Abstract. The purpose of this paper is to provide a method for forecasting changes in housing market trends on multiple levels. We include data on median income, taxation, house price, and demographic breakdown, each at a state level. Our analysis includes K-means time-series clustering, ARIMA and Kalman system forecasting, and Bayesian hierarchical modeling. We hypothesize that using these methods, we can predict future housing market trends (assuming no shocks) and make inferences about the factors influencing shifts in the market.

1. Problem Statement and Motivation

Our project will explore the prediction of housing prices based on Zillow housing price data from 2020 and other demographic information grouped by region. We hope to be able to provide buyers and sellers with relevant information on the future and current states of the housing market that will have an impact on their decision-making process.

One question a buyer or seller might be asking is: What state/area is best to move to given my financial situation and goals? Additionally, how does demographic data influences housing prices and their growth? We will answer this using a Bayesian Hierarchical model. A house that would be the most profitable to buy is one with a low intercept and high slope because that indicates you are buying at a low price relative to how much it will grow, making the profits greatest if you sell later on. Similarly, if you are someone who invests in real estate, where should someone buy a house if they value diversity in their investments. We will answer this by using time series clustering on the percentage change in housing costs to show which states are grouped together. With an accurate cluster, an investor would pick houses in separate groups. Another question might be: What will future housing prices be in a given area based on prior information? Using ARMA modeling and Kalman Filtering will help us test potential predictions and their accuracy.

Previous research for predicting housing prices has included using machine learning techniques such as Random Forests and Gradient boosting. One paper summarizes that Random Forests are the most accurate for training but prone to overfitting, XGBoost and LightGBM are better at generalizing without overfitting, and Hybrid Regression improves accuracy by combining

Date: April 1, 2025.

models but may be less efficient [TNDM20]. Another paper uses data mining techniques, particularly linear regression, to analyze past market trends and predict future house prices [BMM16]. Other research uses economic supply-demand modeling, Hedonistic regression, and empirical evidence on housing cost data, including factors such as immigration, urbanization, changing fertility, and rising life-expectancy, in order to study how demographic shifts impact house prices. They predict that these factors can account for 40.54% of the observed growth in house prices from 1970 to 2010 [GY22]. Suggested future work includes: investigating the coupling affect and finding faster training methods. This research differs from our project mainly by our use of different demographic data, namely percent of population of specific races. Additionally, we use other models such as VARMAX, AutoARIMA, and Bayesian hierarchical modeling.

2. Data

Our main source of data was median housing prices per state obtained from Zillow, a popular online real estate marketplace that helps users buy, sell, and rent homes [zil]. The data was collected of the last day of each month, starting in January, from the year 2000 to 2020. Zillow has extensive experience with pricing and other real-estate information for about 20 years. We had two cases of missing values to address. A few states were randomly missing isolated data points. These were sparse in the dataset, so we used the surrounding data points to average a reasonable estimate for the missing values (using pandas's linear interpolation method). A larger problem was that for some states, measurements started significantly later than others, so the first few years of data were entirely missing. We decided to fill in missing leading values with the first recorded measurement. However, we also added a column for each region to indicate whether leading data was filled in and for how long, so we can keep this in mind when we train and forecast on the data. Lastly, we removed unnecessary features such as region type, size rank, and region ID.

Our remaining data was merged in from the Current Population Survey (CPS) via IPUMS [FKR+24]. CPS is sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS). It is the primary source of labor force statistics for the population of the United States. This data gives observations on the household level on location (state and county), income, property tax, number of people in the household, marital status of parents, race, and population weight. We used weights to find average income, property tax, proportions of each race, and population for each state. Most of the racial categories were combinations of races that had either 0 or very few people in the category, so we dropped these and stuck with white, black, asian, and native American. Our data measured these 4 features per time index per state. Missing values occurred when there wasn't a sample to contribute to the mean. We filled these in with zeros.

Because the survey was conducted on a subsample of the population, the estimates are not completely representative of the true distributions of each value for each state.

In order to merge our data we have to linearly interpolate all of the CPS data to be along the same time index as the Zillow data. Linear interpolation makes more sense than holding values constant and changing once per year, since values such as population and proportion of race change gradually. We could then merge the data on the date and state index. This combined dataset includes 51 states and 301 months of housing price and demographic data. We believe that this amount is sufficient for our project. We split the data by having the training set go up to 2016 and the testing set go from 2016 to 2020, making the data 80% training and 20% testing. This split ensures that the data is more independent between training and testing. Although the data isn't collected in separate sessions, we did make sure that the test data is chronologically after the training data.

3. Methods

To begin analyzing our data, we tested different methods of detrending our data.

3.1. **Classical Decomposition.**

We attempted to detrend via differencing, moving average, and OLS to find the method that produced a nearly covariance stationary time series. Ultimately, we found that differencing gave the best results, and we used this method throughout the rest of our analysis. We then analyzed the different components of our time series using a classical decomposition—we approximated the trend and season with a 12-month period using the `statsmodels.tsa.seasonal` package. We hypothesized that housing data should have a roughly yearly seasonal component, based on the idea that people are more mobile and likely to move during different months of the year. After detrending our data, we wanted to use the time series analysis methods discussed in class to understand which models give the best forecasts of future housing prices.

3.2. **Autoregressive Integrated Moving Average.**

First, we tried fitting and forecasting a model based on the housing price data for a single state using the `pmdarima.auto_arima` model (this is an implementation of an ARIMA model that uses the Kalman filter under the hood in order to find the best-fitting number of autoregressive and moving average terms). For this model, we tested orders of autoregression and moving averages up to a maximum of 12, assuming that each data point should at most depend on the values from the entire previous period/year. Next, we attempted to forecast housing data based on information about other variables, such as housing data from other states with a similar housing market or with the state's own demographic data. We trained and evaluated `VARMAX` models to analyze both of these effects.

3.3. **Time-Series K-means Clustering.** We also clustered the various states according to the observed changes in housing prices, using `TimeSeriesKMeans` from `sklearn.clustering`. This tool allows us to use `KMeans` to cluster, but with an algorithm designed specifically for time series data.

3.4. **Bayesian Hierarchical Model.** Finally, we created a Bayesian hierarchical model to make inferences about the influences on the data and to provide an alternative method of predicting future time series values. We used a difference-of-logs linear parameter model. Let P represent the median house price, tax be the average property tax, pop be the population, $\alpha_{native}, \alpha_{asian}, \alpha_{black}$, and α_{white} be the proportion of the population in each racial category. Our hierarchical model considers state-level parameters for the time slope and intercept, drawing from a population distribution, as well as population parameters for the coefficients on tax, population, and race data. Further, we let i denote the state and t denote the time (in months). The model and prior distributions we use are

$$\begin{aligned}
P_{i,t} &= \beta_i^0 + \beta_i^t\ t + \beta_{tax}\ tax_i + \beta_{pop}\ pop_i + \beta_{native}\ \alpha_{native,i} \\
&\quad + \beta_{asian}\alpha_{asian,i} + \beta_{black}\alpha_{black,i} + \beta_{white}\alpha_{white,i} + \epsilon, \\
\beta_i^0 &\sim \mathcal{N}(\mu, \sigma_\mu), \\
\beta_i^t &\sim \mathcal{N}(\eta, \sigma_\eta), \\
\beta_{tax} &\sim \mathcal{N}(0, 20), \\
\beta_{pop} &\sim \mathcal{N}(0, 1), \\
\beta_{native} &\sim \mathcal{N}(0, 10000), \\
\beta_{asian} &\sim \mathcal{N}(0, 10000), \\
\beta_{black} &\sim \mathcal{N}(0, 10000), \\
\beta_{white} &\sim \mathcal{N}(0, 10000), \\
\epsilon &\sim \mathcal{N}(0, \sigma), \\
\mu &\sim \mathcal{N}(200000, 100000), \\
\eta &\sim \mathcal{N}(1000, 1000), \\
\sigma_\mu &\sim \text{Exp}\left(\frac{1}{50000}\right), \\
\sigma_\eta &\sim \text{Exp}\left(\frac{1}{5000}\right), \\
\sigma &\sim \text{HalfCauchy}(100).
\end{aligned}$$

We select mean-zero with relatively narrow variance for the tax and population coefficients because these values can change a lot, though we don't anticipate huge changes in the house price as a result (relative to the size of the change in these values). The population coefficient prior distributions were chosen with large variance and mean zero because the proportion of the population in each race is very small, so it changes very little, though we

allow for large changes in the median house price as a result of these changes. For our hyperparameters μ and η, we select rough estimates as the means with large variances so as to allow the model to explore and converge to the right population values. Next, σ_μ and σ_η were selected to have exponential distribution with very wide tails to avoid imposing too tight of a prior distribution on the variance of our prior distributions. Before sampling, we first use z-score normalization mapping $x_{i,t}$ to $\frac{x_{i,t}-\overline{x}_i}{\sigma_i}$ to transform the independent variables and standardize the scales to allow for easier sampling. This comes at the cost of precise interpretability, but the general results still hold.

4. Results and Analysis

4.1. **Classical Decomposition.** The most interesting result of applying a classical decomposition to our data set was that the seasonal component of housing prices in each state had two different peaks throughout the year. Though the height of each peak varied from state to state, nearly every state displayed housing prices that peak halfway through the year, decrease for 1-3 months, then increases and peaks again before the end of the year. Additionally, the classical decomposition led us to believe that housing markets over the past 5 years are associated with a different trend and seasonal component than in prior decades. This conclusion comes from the fact observation that the residual component of the housing data for each state is relatively small from 2000 to 2019, but starting in 2020, the residual plot becomes erratic and fluctuates rapidly, with a much larger magnitude than previous years (see Fig. 1). We concluded that this makes sense, because housing prices are subject to random shocks that are heavily dependent on the country's overall economic conditions, as well as changes in economic policy. The start of the COVID-19 pandemic in 2020 drastically affected both of these features, which gives a plausible explanation for the variation in housing prices after 2020. However, this change makes training and forecasting after 2020 based on data from before 2020 very difficult, so in many of our models, we choose to drop data after January 2020, and recompute our train-test split on the remaining previous data.

4.2. **Autoregressive Integrated Moving Average.**

4.2.1. *Univariate ARIMA.* Next, we used a variety of ARIMA models in an attempt to forecast future housing prices. First, we used the `auto_arima` model to determine the optimal order of difference, number of autoregressors, and order of the moving average component. The parameters that gave the lowest AIC is an ARIMA(4, 1, 0) model, shown in Fig. 3, 2. It is surprising that the optimal model does not have a moving average component, and this presents a potential topic of further study. The main issue we encountered when forecasting with ARIMA models is that the variance of the data forecast increases as the forecast continues. For some states, the

FIGURE 1. Housing Prices in Utah, 2000-2024

variance becomes so large that the forecast is almost trivial because the 95% confidence interval spans hundreds of thousands of dollars.

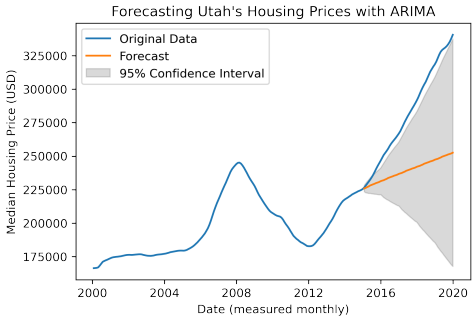

FIGURE 2. Forecasting Housing Prices in Utah, 2000-2020

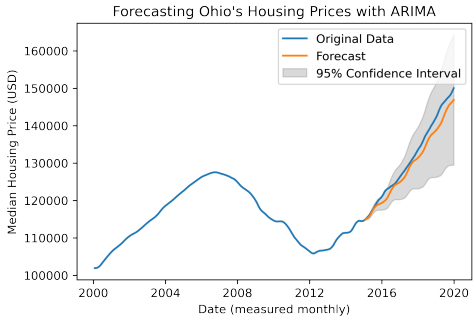

FIGURE 3. Forecasting Housing Prices in Ohio, 2000-2020

4.2.2. *Vector ARIMA.* We also experimented with forecasting future housing prices using `VARMAX` models. First, we hypothesized that the housing price in one state can be more accurately predicted using information about housing prices in other states with similar housing markets. So, we used time series clustering to partition the 50 states into 7 clusters based on the percent change in housing price (this gives us a good idea of which states' housing prices move and react to shocks in similar ways). For each cluster, we initialized and trained a `VARMAX` model using the housing data for each state assigned to that cluster (see Fig. 4). This model resulted in a higher AIC than a single time series ARIMA model, but in many cases still provided more accurate forecasting. It is important to note that this method is not useful in all cases, because the result of the clustering algorithm placed some states in singleton clusters, with no related states to include in combined model. We also trained `VARMAX` models on individual states' data for both housing price and demographic features, hoping that the correlation between housing prices and social and racial demographics would result in more accurate forecasting. Unfortunately, these models did not yield good results. For many states, the model completely failed to converge, and predicted a constant housing price equal to the mean of the time series. For a few states, the model converged and gave usable results, but these cases were scarce and there did not appear to be any pattern or discerning feature between the states for which the model did not converge and the states for which it did.

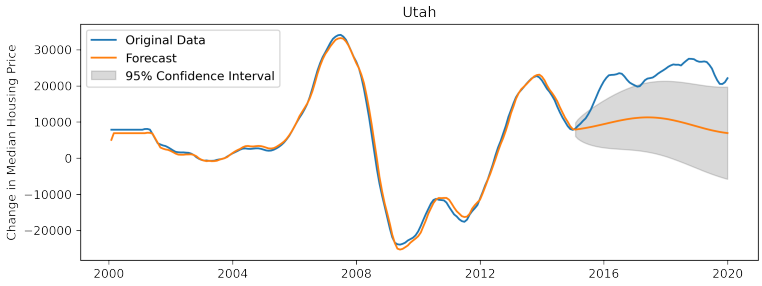

FIGURE 4. `VARMAX` Forecasting on Housing Prices in Utah, 2000-2020

4.3. **Time-Series K-means Clustering.** In order to answer which states to buy real estate in if you want to diversify your investments, we decided to use time series clustering to show which states are most similar to each other. Our KMeans algorithm was trained on the percentage change in housing prices and predicted with 2-10 clusters. We found that the smallest clusters started by dividing the states by those in the middle and those on the coasts. Gradually this middle region group shrunk as we increased the number of clusters 5, 6. This result indicates that if you want the most diverse real estate investment, in terms of percent change in housing price,

then you might want to buy houses in a mixture of middle states and states along the perimeter of the country. States like California and Florida often were clustered either by themselves or in the smallest groups, making them seem to have a more likely difference than other states.

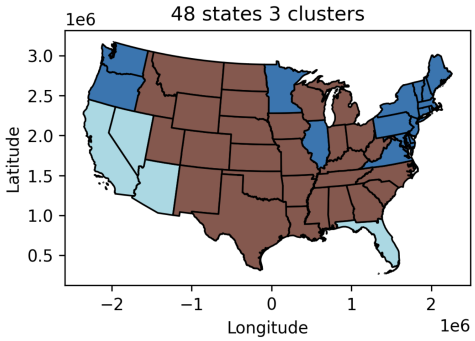

FIGURE 5. Clustered states with 3 groups.

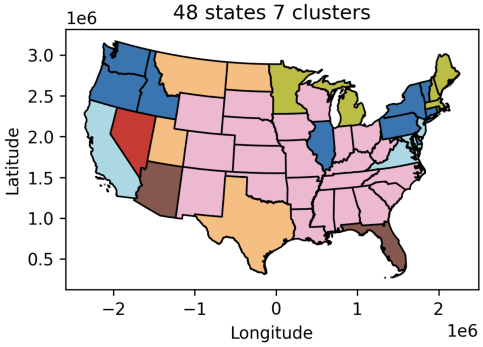

FIGURE 6. Clustered states with 7 groups.

4.4. **Bayesian Hierarchical Model.** To estimate the parameters in the hierarchical model described earlier, we apply the No-U-Turn (NUTs) algorithm, implemented in PyMC3. This algorithm is an extension of the Hamiltonian Monte Carlo Sampling algorithm which improves the efficiency of the sampling. We ran four chains, each tuning with 2000 samples and then iterating for 2000 more samples, obtaining 8000 samples in total for each parameter. As with the other models, this model was only trained on the data up through 2016.

Figure 7 in the appendix shows the trace plot resulting from the sampling, including the overall distributions for the intercept μ and slope η, along with posterior distributions for the slope and intercept for each state, and the stope for the standardized features. The results give reason to believe that the distributions of parameters have converged, with overlapping kernel density estimates of each parameter and erratic plots of the samples, showing that the sampling algorithm did not get stuck in a tail of the distribution.

The strength of the hierarchical model is shown in Figure 8 and Figure 9. We see the mean and a 94% credible interval for the posterior distributions of the slope and intercept parameters for each state, indicating which states had the highest prices in 2000 which have been growing fastest since then. From these, we observe that Washington D.C. (District of Columbia) and Hawaii have been growing the fastest, while Nevada and Michigan have been growing the slowest (or even decreasing on average, from 2000 to 2016).

Although the basic form of the model is a linear regression, because of the unit standardization performed before running the model, the interpretations of the coefficients are ambiguous. However, because each feature

now has the same scale, we can interpret the magnitude of the coefficients as the importance of this feature relative to the other features. For example, as shown in Figure 7, an increase in the native American population tends to lead to a decrease in housing prices, as one example.

5. Ethical Implications and Conclusions

We did not find any privacy issues with the collected data because it was grouped by region and did not reveal personal information about an individual. However, there are still other ethical implications to consider. One thing to watch out for is predicting on demographic features, such as race. If race, or other features that may show personal biases, are important factors in our predictions then our model may be showing that living in areas with certain races will negatively, or positively depending on if you are a buyer or seller, affect the housing prices. This may lead to further bias or unfair treatment. Furthermore, if people find out that a particular state has the best average price to housing price trend, then people may rush to move there, which affects people in that area and may even skew the numbers we have calculated to being inaccurate. In a worst-case scenario, these methods could lead to a destructive feedback loop, where regions which higher proportions of a certain race are labeled as less desirable to live in, and therefore nobody moves there or buys homes there, which means that region is unable to develop, and becomes even less desirable to live in.

In addition to the ethical concerns around race, our results about which housing markets are correlated and diversification of real-estate portfolio's could be extremely beneficial to real-estate developers or wealthy investors seeking to increase their static income. However, this could direct companies and wealthy people to buy up houses in less expensive, developing markets, thus driving up the housing price in that area and making homes less affordable for average people.

Because of the ethical concerns listed above and the inconsistency of some of our models, it is most likely not ethical to use our results as a guide for individuals to advise where to live or buy houses, or to feed opinions on certain regions. Perhaps a more ethical use for these results would be for policy makers to better understand the driving factors behind housing prices and use this information in order to decide on economic policies and allocate funding to regions that would benefit from it the most.

6. Conclusion

The clearest conclusion we can make from our analysis and modeling of data is that housing prices are very difficult to forecast. Though some methods outperformed others, we did not find a model and set of inputs that reliably produced accurate forecasting results. We attribute this to the volatility of the economic situation in recent years, and the fact that housing prices are largely determined by random shocks that come from a

number of different factors. For example, we saw drastic changes in the trend of housing prices across the U.S. surrounding both the recession in 2008 and the beginning of the COVID-19 pandemic in 2020. Because of this, we reasoned that we do not have enough data features to represent all of the inputs that affect housing prices. We did see a correlation between race proportions, population, income, and housing prices, but housing prices are also very likely correlated to things such as national economic policies, political leadership within a certain region, the overall economic condition in the U.S., and even natural disasters. A more thorough dive into the factors involved in housing prices, as well as a further attempt to model and forecast using those factors, would be valuable topic for further research.

Though our forecasting results were not very reliable, we found interesting results in describing the distribution of housing prices in the U.S. We expected to find that different regions of the U.S. (e.g. the midwest, the New England region, the Pacific northwest), would have similar housing markets within themselves. However, the results of time series clustering showed that the majority of states fall into two main clusters, with the rest of the clusters only containing a handful of outliers (see Fig. 5, 6). This conclusion is supported by the results of our `VARMAX` modeling, which showed that for the states that belong to significant clusters (i.e. clusters that contain more than one state), forecasting by considering data from the rest of the states in the cluster produced better results than forecasting solely on one states data.

From the Bayesian model, we can draw inferences on the impact that race, taxation, and population have on the state housing price. We learned that greater populations of native Americans, Asians, and whites tends to lead to lead to decreased housing prices (up to 2016) while a greater proportion of black people leads to increasing housing prices. The hierarchical structure provides many benefits, such as effective estimation of both overall population parameters and state-specific parameters as draws from the larger distribution, along with shrinkage to reduce the variance of estimates from states with few samples. However, the simple linear model used in this case fails to capture the intricacies of the data and account for shocks to the market. Future work will include Bayesian hierarchical estimations of more robust economic models that account for shocks in the market more effectively.

Overall, we have learned that predicting the future is extremely difficult. Various factors within states can complicate the ability to predict. Frequentist models such as ARIMA and Kalman filters can sometimes provide meaningful estimates on some states, but not consistently. The Bayesian methods show promise for being able to simultaneously capture both large-scale trends and individual state movements through shrinkage effects. In future work, we believe that a combination of the autoregressive components and the Bayesian hierarchical model can more effectively and accurately predict market trends, including the effects of shocks, while accounting for the

different impacts that the shocks have on different areas as a result of demographic breakdown and other factors, providing a robust and informative posterior distribution for housing market trends.

References

[BMM16] Nihar Bhagat, Ankit Mohokar, and Shreyash Mane. House price forecasting using data mining. *International Journal of Computer Applications*, 152:23–26, 10 2016.

[FKR+24] Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry. Integrated public use microdata series, current population survey: Version 12.0, 2024.

[GY22] Yifan Gong and Yuxi Yao. Demographic changes and the housing market. *Regional Science and Urban Economics*, 95:103734, 2022.

[TNDM20] Quang Truong, Minh Nguyen, Hy Dang, and Bo Mei. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174:433–442, 2020. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

[zil] Housing data - zillow research.

7. Appendix

Here, we include the large plots from the Bayesian model that require their own page. These should not count towards the page count but are still fun for people who want to see them.

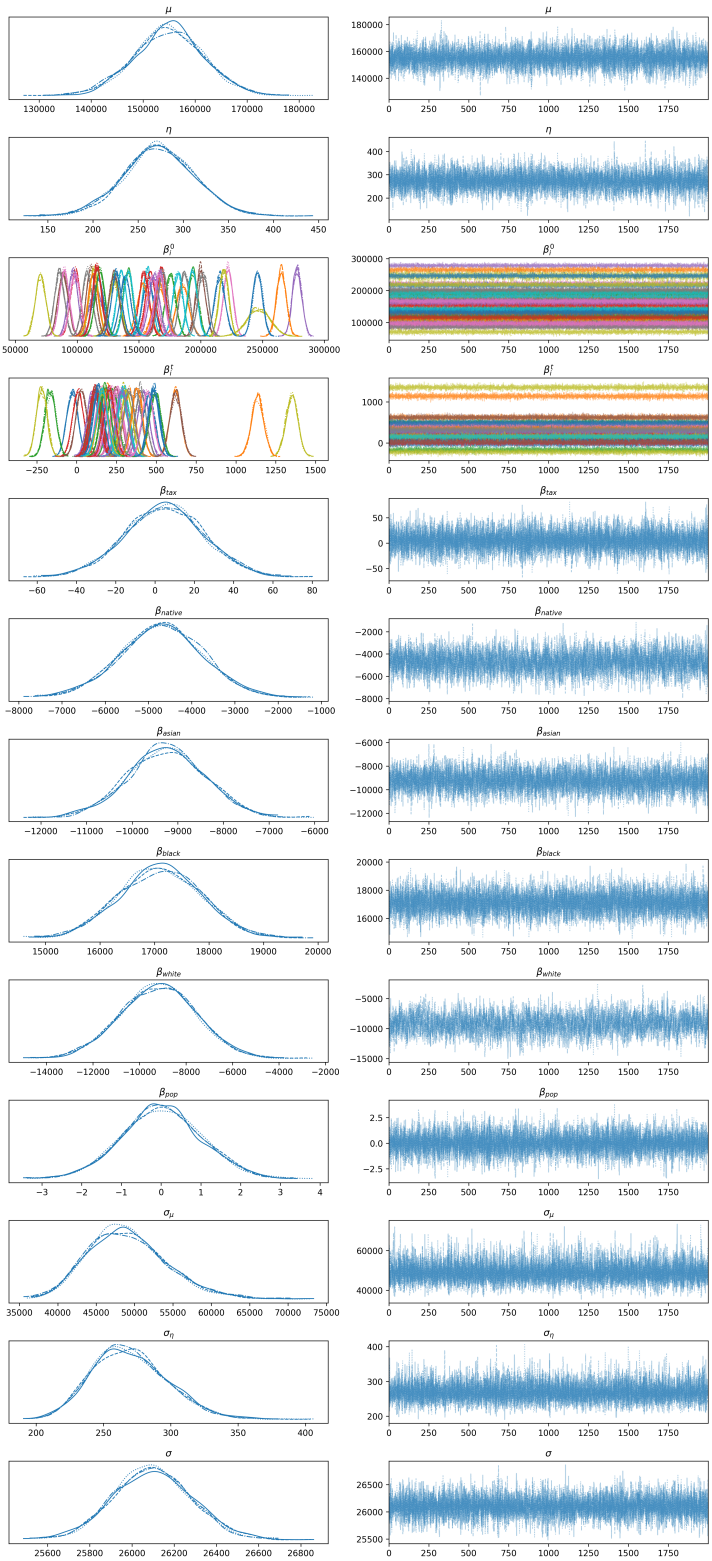

FIGURE 7. Trace Plot of the Posterior Distributions from Bayesian Estimation

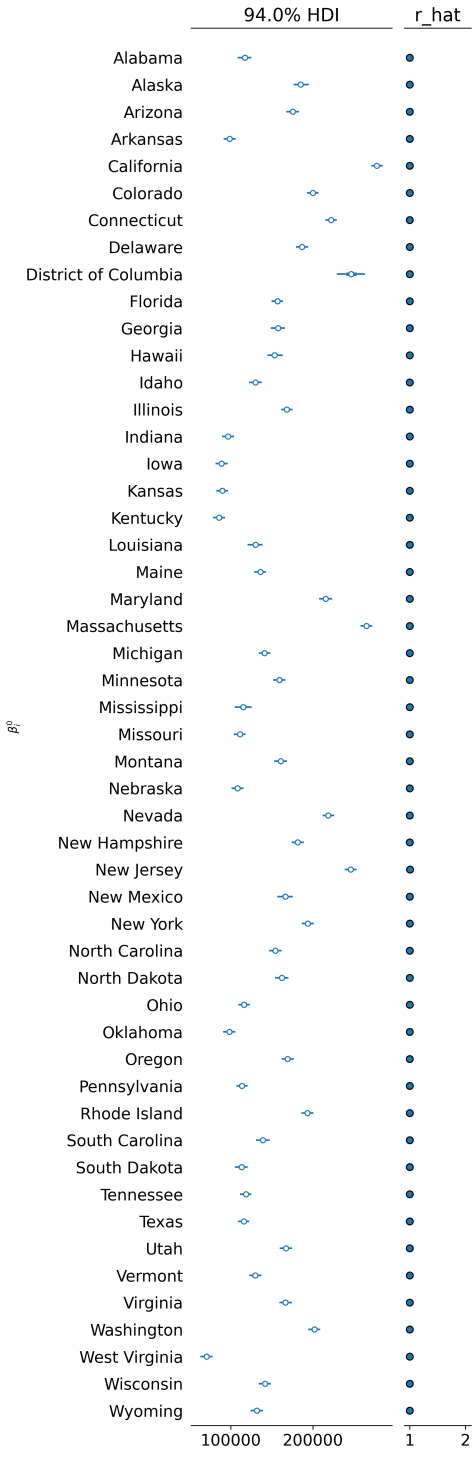

FIGURE 8. Posterior Distributions on the Intercept for each State

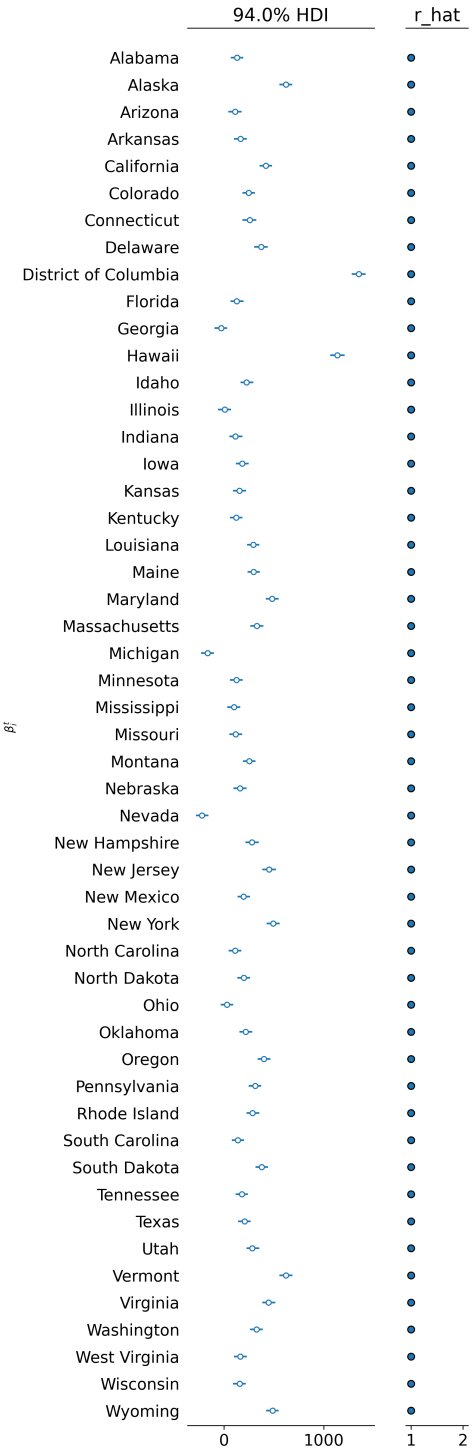

FIGURE 9. Posterior Distributions on the Slope for each State