Predicting Local Housing Markets with Bayesian Vector Autoregression

Walter Coleman, Timur Guler, and Stephen Whetzel

INTRODUCTION: Residential housing markets trajectories are notoriously hard to predict. Prices vary drastically from region to region, and while they are driven to some extent by local and national economic factors, pricing trends cannot be completely explained by these factors. Our team would like to understand these relationships and the uncertainty surrounding them and believe that a Bayesian time series approach using vector autoregression will achieve both goals. We implemented this approach in four metro areas across the US using data from the Federal Reserve Economic Data (FRED) – Atlanta, New York, Tulsa and Topeka.

DATA: FRED collects data on hundreds of economic, demographic, and housing market features by Metro Statistical Area (MSA). We accessed these datasets through the FRED API **(footnote).** The dataset on median home listing price begins in July 2016, so our analysis is limited to this period. Additionally, we chose to exclude data from 2020 and 2021, due to the unprecedented and anomalous nature of the Covid-19 global pandemic. While this phenomenon warrants further exploration, it is beyond the scope of our project.

We began with a set of 12 predictor variables to consider, and from here narrowed our input space to average hourly earnings, unemployment rate, proportion of workers employed in leisure and hospitality, and median days on market for active listings. We will discuss the intuition behind this down-selection as well as the process behind it in our methodology section. Our team faced some data massaging challenges due to the varying nature of these data series, including frequency of collection (i.e. monthly vs annual) and the absence or presence of seasonality (**Figure 1**).

METHODOLOGY: Our approach consisted of three subprocesses: variable selection, model training, and prediction. A preliminary discussion of the autoregression model will inform each of these.

General autoregression is a methodology which predicts future points in a time series based on previous points – it is essentially linear regression with the observations of previous periods as the feature set.

**{equation}**

Vector autoregression expands this to allow for multiple time series predictors – the future states of each of the predictor variables (earnings, unemployment, etc.) are determined by the past states of those individual variables, and the future states of the target variable (median listing price) is determined by the past and predicted future states of *all* variables, including itself.

**{equation}**

The major tuning parameter in this type of model is order, which refers to the number of previous periods incorporated into the model. While including more periods can enhance predictive capabilities, it can also lead to overfitting, especially when the total number of autoregressive parameters approaches or exceeds the number of observations. Given our relatively small observation set for each series, we used and order of **[n]** for our final model.

The next major aspect of model architecture was variable selection. Mindful of the need to control the total number of autoregressive parameters, we limited ourselves to a subset of the 12 initially explored predictors. We first attempted to find the strongest predictors with non-autoregressive linear methods, specifically hierarchical models and un-pooled models for individual cities. These methods did not provide informative results, so we instead used visual analysis of listing price vs. variable scatterplots to select our autoregressive feature space (**Figure 2**). We selected a balance between tightness of correlation and spread across topic (i.e. economy, demographics, housing market). This resulted in the following model:

**{equation}**

In order to better understand the level of uncertainty in both the relationships between variables and the accuracy of final predictions, we used a Bayesian approach. We chose uninformed priors **{N~(0, 10000)}** in order to reflect our lack of initial belief or understanding, resulting in the following final architecture:

**{Bayes rule as applies to problem}**

**How we used MAP to make predictions - Stephen**

Results: questions to answer:

1. What are coefficients of predictors and resulting uncertainty?
2. How do these change by city?
3. Are the actual values of predictions within our confidence intervals of uncertainty distributions?

Figure 1:

Graphical user interface, application, calendar

Description automatically generated

Figure 2:

A picture containing text, different, shelf

Description automatically generated