Detecting Price Diffusion Across Natural Geographic Boundaries:
An Application of Vector Autoregressive Models

J. Bradley Eustice
Washington State University

MS Statistics Project
7 July 2016

Abstract

While it has been shown that housing prices diffuse across arbitrary political boundaries, does price also diffuse across natural geographic barriers? This paper develops a vector autoregressive model for a sample data set of Washington housing prices from the 4th quarter of 2007 through the 4th quarter of 2015. Regions are constructed such that the boundaries between them are more substantive (i.e. mountains) than arbitrary political lines. Additional vector autoregressive models with the inclusion of exogenous variables are also estimated. In general, the results are consistent across the models. However, due to small sample size and p-values on the boundary, the null hypothesis is only *weakly* rejected. That is, housing prices in Western Washington weakly Granger-cause housing prices in Eastern Washington.

1. Introduction

It has clearly been shown that location is paramount in determining housing prices (Kiel and Zabel, 2008). Hedonic pricing models decompose an assets value into its constituent parts. Hedonic models that only take into account the physical attributes of residential properties cannot distinguish between a 3-bedroom, 2-bathroom house located in Manhattan, and the same house located in Waco, Texas. The location parameter comprises, in a general sense, the other residential properties, businesses, demographics, and market forces in the neighborhood (Case and Mayer, 1996; Sedgley et al, 2008). When house prices change in one neighborhood, when and how does that affect the house prices in adjacent neighborhoods? This spatial and temporal spill-over effect is the main focus of this paper.

The spill-over effect is not a recent discovery. Basu and Thibodeau (1998), among others, found that spatial autocorrelation is an important explanatory variable of regional housing prices. Homes in the same neighborhood share many of the same location amenities and services. For example, as Basu and Thibodeau point out, local government services such as police and fire departments are common among all residents in the area. The size of one's "neighborhood" has been studied at varies levels. Neighborhood can be defined as an area as small as adjacent homes to the size of a state or country where each of the latter has its own laws, economic conditions, taxes, and demographics (Pollakowski and Ray, 1997). Holly, Pesaran, and Yamagata (2010) found a spatial effect at the State level, in that changes in house prices in one State had a statistically significant impact on house prices in contiguous States. However, there are clearly within State differences that are averaged out when defining the neighborhood at the State level. For example, the population centers of San Francisco and Los Angeles are very different than their rural neighbors in the Central Valley of California, which produces about 25% of the nation's food supply (USGS, 2016). These major differences are lost when looking at California as a whole.

In the example above, the Central Valley is bounded by the Northern and Southern Coast Ranges to the west and the Sierra Nevada Mountain Range to the east. These serve as natural barriers between the Central Valley and the densely populated areas of San Francisco and Los Angeles. While the definition of neighborhood has taken on many forms in the literature, no papers have looked at modeling regional house prices by defining the spatial component as significant "natural" neighborhoods. Tirtiroglu and Clapp (1996) use a similar idea but in a different context. They look at the returns to housing from a financial perspective and the Connecticut River serves as a barrier that slows the diffusion of information. While not explicitly looking at natural barriers, Holly, Pesaran, and Yamagata (2011) find that house prices in New York can predict house prices in London. They attribute this to the common trait these cities share as financial capitals of the world.

When defining the size of a neighborhood as anything larger than a city, it becomes unfeasible to use traditional hedonic pricing models to price individual houses. In these cases, macro-level variables are used to explain changes in the average house prices of the region. For example, McQuinn and O'Reilly (2008) look at the relationship between how much an individual can borrow and the sale price of the house. The amount an individual can borrow is a function of income and interest rates, where the latter is a macro-level variable. They find that a long-run relationship does in fact exist. While hedonic models may work well in the short-run, they do not capture the long-run trends that are assessed by including macro-level variables. Both supply side and demand side factors affect regional house prices. These include new construction, population growth or decline, the unemployment rate, and the inflation rate. In addition to these, many papers have found a link between neighborhood demographics and house prices (Case and Mayer, 1996; Sedgley et al, 2008).

The core data for this paper consists of housing prices for each Washington State county from 2007 to 2015. The geography and urbanization of Washington State provides two distinct regions: Eastern and Western Washington. These regions are naturally divided by the Cascade Mountain Range. The diffusion of housing prices across the Cascades is estimated by vector autoregressive models. There may be price diffusion across arbitrary political lines (Brady, 2011; Holly et al, 2010); however, the focus of this paper is whether or not price diffuses across natural geographic barriers.

The rest of the paper is as follows: Section 2 develops the model and estimation technique, as well as describes the data. Section 3 presents and discusses the results. Section 4 concludes.

2. Methods

2.1 Vector autoregressive models

To identify the diffusion of housing prices across time, an autoregressive model is most appropriate. Autoregressive models are used when past values may have an effect on current values. The variable of interest is the median quarterly housing prices of the two regions. To identify the diffusion of housing prices across time and space, the autoregressive model must be extended to the multivariate case. Not only do the past housing prices of region *i* influence current prices of region *i*, but the past prices of the other regions may also have an effect as well. The estimation technique used in this paper is called vector autoregression (VAR). VAR models were first proposed as an alternative to structural models used in macroeconomic modeling by Sims (1980). The basic VAR(*p*) model with *p* time lags is

$$Y_t = c + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \varepsilon_t, \tag{1}$$

where *t* is the time period. Y_t is a *k × 1* vector denoting the variable of interest where *k* represents the number of potential interdependent variables (in our case *k* = 2 regions). Y_{t-1} through Y_{t-p}, c, and ε are *k × 1* vectors denoting the time lags of Y_t, the intercept, and the error, respectively. A_1 through A_p are each *k × k* matrices of unknown parameters to be estimated.

Other variables have been found to influence housing prices beyond the information contained in lagged prices. McQuinn and O'Reilly (2008) found mortgage rates to be significant. The unemployment rate is also vital in determining house prices (Abelson et al, 2005; Xu and Tang, 2014). To incorporate these control variables as well as others, the VAR(*p*) must be slightly modified. The intermediate and full models used in this paper are referred to as the VARX, or the vector autoregressive model with exogenous variables. The VARX(*p*) model with *p* time lags is

$$Y_t = c + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \beta X_t + \varepsilon_t, \tag{2}$$

where the terms and dimensions are identical to the basic VAR(*p*) model with the addition of exogenous variables. X_t is a *n × 1* matrix where *n* is the number of exogenous variables. β is a *k × n* matrix of unknown parameters to be estimated.

2.2 Assumptions

Similar to multivariate regression models, the exogenous variables are the same across the two regions. The only difference is the inclusion of the lagged values of the dependent variables. However, with the inclusion of time-series data, additional assumptions must be met beyond those of multivariate regression. The assumptions of VAR models are: normality, no autocorrelation, and stability. The normality condition states that the errors are normally distributed, i.e. $\varepsilon_t \sim N(0, \sigma^2)$. Autocorrelation states that the residuals are correlated across time. The strongest assumption of VAR models is stability. The stability condition states that a stable process will not diverge to infinity. By theorem, if a stochastic process is stable, then it is also covariance stationary. This means that the first and second moments do not change through time, i.e. $E(x_t) = \mu$ for all t and $E[(x_t - \mu)(x_{t-h} - \mu)'] = \Sigma$ for all t and h. These three conditions must be met before VAR(*p*) and VARX(*p*) models produce valid results

2.3 Data

The core data for this paper is compiled by the Runstad Center For Real Estate Studies at the University of Washington. The data consists of estimated quarterly median housing prices for each Washington State county from the 4th quarter 2007 through the 4th quarter of 2015. Of the 39 counties in Washington State, three are dropped from the dataset. Two of these counties are dropped because of missing values over the sample period (Lincoln and Wahkiakum), one on either side of the Cascades.

Skamania County is excluded because it resides on the Cascade Mountains, directly between the two regions of interest. These three counties comprise less than one half of one percent (0.38 percent) of the total population of Washington State between 2007 and 2015. Since the median housing prices are at the county level, a weighted average based on population is used to generate the quarterly median housing prices for the two regions. The yearly county population estimates are computed by the Office of Financial Management of Washington State.

To supplement the dataset, a few other variables are compiled: the seasonally-adjusted unemployment rate, the 30-year fixed mortgage rate, and real per-capita personal income. The seasonally-adjusted monthly unemployment rate for Washington State is calculated by the Bureau of Labor Statistics. The monthly 30-year fixed mortgage rate is obtained from FreddieMac. Both of these are converted to quarterly data by simple averaging. Lastly, personal income is calculated by the Bureau of Economic Analysis. Personal income is the total income attributed to the household sector in Washington State. Real per-capita personal income is calculated using the population statistics made available by the Office of Financial Management.

Two dummy variables are also constructed. The first dummy variable indicates seasonality experienced in the real estate market. Since most real estate transactions take place during the summer months, a dummy variable indicating quarters two and three is created. The second dummy variable indicates the housing crisis. The sample period covers the Great Recession, which the National Bureau of Economic Research defines as the fourth quarter of 2007 through the second quarter of 2009. However, prices for both the West and East regions of Washington State did not bottom out until the first quarter of 2012 (Figure 1).

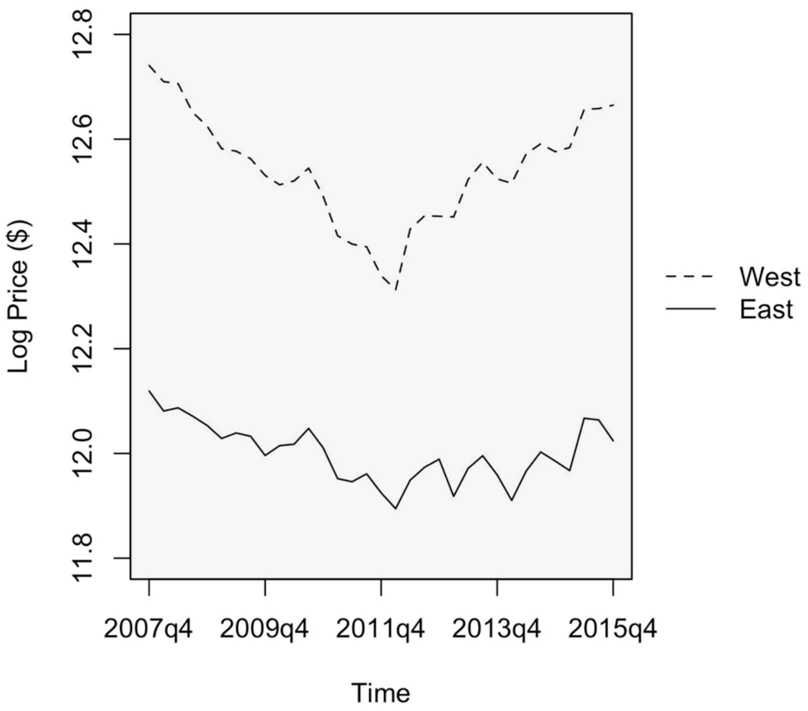

Figure 1. Log of the median housing prices for Western and Eastern Washington

2.4 Estimation method

The dependent variable is the log of the real quarterly median home price for each region. The two regions are Western and Eastern Washington, which are divided by the Cascades. The real quarterly median home price is constructed using the inflation rate, which is, by definition, calculated from the consumer price index (CPI). Using the real instead of the nominal home price helps diminish the trends that is experienced over the sample period. Since the CPI includes all other goods, the cross-price elasticities are effectually incorporated into the dependent variable. Adding the average price of food and energy, for example, as explanatory variables are less relevant when the median home price already incorporates this information through the CPI. The CPI West Region is available from the Bureau of Labor Statistics.

Including the lag values of the real quarterly median home price for each region yield the basic VAR(*p*) model to be estimated. Additionally, including the exogenous control variables described above yield the VARX(*p*) model for estimation. Table 1 reports summary statistics for the variables used in the VAR(*p*) and VARX(*p*) models. The question of interest is: Do house price shocks in one region jump across natural geographic barriers and affect house prices in the adjacent region? Own-lag effects are the past prices of region *i* when estimating the housing price in region *i*, while cross-lag effects are the past prices of the other regions. The testable hypothesis center on the values for A in Eq. (1) and Eq. (2). If $A = 0$ when $i \neq j$ (i.e. cross-lag effects), then housing prices in region *j* do not affect housing prices in region *i*. This would mean that housing prices do not jump natural geographic barriers.

Table 1
Summary Statistics

Names	Variable descriptions	Mean	Std. Dev.	Min.	Max
$lnWEST$	Log of the real median sale price for Western Washington (dollar)	12.54	0.11	12.31	12.74
$lnEAST$	Log of the real median sale price for Eastern Washington (dollar)	12.00	0.06	11.89	12.12
$MORTG$	30-year mortgage rate (%)	4.54	0.82	3.36	6.32
$UNEMP$	Washington State unemployment rate (%)	7.49	1.78	4.80	10.37
$PERS\_INC$	Average yearly income for households in Washington State ($10,000)	4.32	0.16	4.08	4.66
$CRISIS$	Dummy variable indicating housing crisis			0	1
$SUMMER$	Dummy variable indicating quarters two and three			0	1

2.5 Diagnostic Tools

Vital to VAR model analysis is choosing the appropriate number of lags, *p*. According to the efficient market hypothesis, all relevant information is incorporated into an asset's price. In line with this hypothesis, only the first 4 lags need to be analyzed (1 year), since it is reasonable to assume that anything outside a year would violate the efficient market hypothesis. There are multiple criteria to determine the optimal number of lags, but most are variations of the Akaike information criterion (AIC). The criteria used in this analysis is the Hannan-Quinn information criterion (HQIC). Choosing *p* based on the HQIC has been shown to be a consistent estimate of the true lag order (Lütkepohl, 2005).

The no autocorrelation assumption is tested by a simple multivariate Lagrange multiplier test, which is distributed as a chi-squared with four degrees of freedom. The null hypothesis is that there is no autocorrelation at lag *j*. To test the normality of the errors, the Jarque–Bera test is employed. The Jarque–Bera test statistic is distributed as a chi-squared with four degrees of freedom. The null hypothesis states that the errors follow a multivariate normal distribution.

Since VAR(*p*) models are functions of past lags, Eq. (1) can be written in lag operator notation as $\left(I - \mathrm{A}_1 L - \mathrm{A}_2 L^2 - \cdots - \mathrm{A}_p L^p\right) y_t = c + u_t$. A VAR(*p*) is stable if the roots of $det\left(I - \mathrm{A}_1 z - \mathrm{A}_2 z^2 - \cdots - \mathrm{A}_p z^p\right) = 0$ lie outside the complex unit circle (i.e. have modulus greater than one). Alternatively, Lütkepohl (2005) and Hamilton (1994) determine that a stochastic process is stable if the modulus of each eigenvalue of the companion matrix is less than one, where the companion matrix is defined as

$$F = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & I & 0 \end{bmatrix}.$$

Determining the eigenvalues of the companion matrix is used in the analysis.

The testable hypothesis center on the parameter values for A in Eq. (1) and Eq. (2). Instead of testing each cross-lag parameter separately (i.e. $A_{t-k} = 0$ when $i \neq j$), all the cross-lag parameters are tested jointly at the 5% level, which leads to more conservative results. The formal empirical hypothesis is H_0: $A_{t-1} = \cdots = A_{t-p} = 0$ when $i \neq j$. Colloquially, the null hypothesis states that housing prices do *not* diffuse across natural geographic boundaries. The hypothesis is tested via Wald tests. If the formal hypothesis is rejected, then housing prices in region *j* contain information that helps predict, or Granger-cause (Granger, 1969), housing prices in region *i*.

References

Abelson, Peter, Roselyne Joyeux, George Milunovich, and Demi Chung. 2005. "Explaining House Prices in Australia: 1970-2003." *Economic Record* 81, S96-103.

Basu, Sabyasachi, and Thomas G. Thibodeau. 1998. "Analysis of Spatial Autocorrelation in House Prices." *Journal Of Real Estate Finance And Economics* 17, no. 1: 61-85.

Brady, Ryan R. 2011. "Measuring the Diffusion of Housing Prices across Space and over Time." *Journal Of Applied Econometrics* 26, no. 2: 213-231.

Branch, William E., Nicolas Petrosky-Nadeau, and Guillaume Rocheteau. 2016. "Financial frictions, the housing market, and unemployment." *Journal of Economic Theory* 164, 101-135.

Case, Karl E., and Christopher J. Mayer. 1996. "Housing Price Dynamics within a Metropolitan Area." *Regional Science And Urban Economics* 26, no. 3-4: 387-407.

Granger, C. W. J. 1969. "Investigating Causal Relations By Econometric Models And Cross-Spectral Methods". *Econometrica* 37, no. 3: 424-438.

Hamilton, James D. 1994. *Time Series Analysis*. Princeton, N.J.: Princeton University Press.

Holly, Sean, M. Hashem Pesaran, and Takashi Yamagata. 2010. "A Spatio-temporal Model of House Prices in the USA." *Journal Of Econometrics* 158, no. 1: 160-173.

Holly, Sean, M. Hashem Pesaran, and Takashi Yamagata. 2011. "The Spatial and Temporal Diffusion of House Prices in the UK." *Journal Of Urban Economics* 69, no. 1: 2-23.

Kiel, Katherine A., and Jeffrey E. Zabel. 2008. "Location, Location, Location: The 3L Approach to House Price Determination." *Journal Of Housing Economics* 17, no. 2: 175-190.

Lütkepohl, Helmut. 2005. *New Introduction To Multiple Time Series Analysis*. New York: Springer.

McQuinn, Kieran, and Gerard O'Reilly. 2008. "Assessing the Role of Income and Interest Rates in Determining House Prices." *Economic Modelling* 25, no. 3: 377-390.

Pollakowski, Henry O., and Traci S. Ray. 1997. "Housing Price Diffusion Patterns at Different Aggregation Levels: An Examination of Housing Market Efficiency." *Journal Of Housing Research* 8, no. 1: 107-124.

Sedgley, Norman H., Nancy A. Williams, and Frederick W. Derrick. 2008. "The Effect of Educational Test Scores on House Prices in a Model with Spatial Dependence." *Journal Of Housing Economics* 17, no. 2: 191-200.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48, no. 1: 1-48.

StataCorp. 2015. *Stata Statistical Software: Release 14.1.* College Station, TX: StataCorp LP.

Tirtiroglu, Dogan, and John M. Clapp. 1996. "Spatial Barriers and Information Processing in Housing Markets: An Empirical Investigation of the Effects of the Connecticut River on Housing Returns." *Journal Of Regional Science* 36, no. 3: 365-392.

USGS, 2016. "California's Central Valley". *USGS California Water Science Center*. http://ca.water.usgs.gov/projects/central-valley/about-central-valley.html.

Xu, Lu, and Bo Tang. 2014. "On the Determinants of UK House Prices." *International Journal Of Economics And Research* 5, no. 2: 57-64.