MEDIAN HOUSE VALUE DETERMINANTS OF DEMOGRAPHIC CHARACTERISTICS, ACCESSIBILITY TO AMENITIES, AND SPATIAL SPILLOVERS IN CORE BASED STATISTICAL AREAS IN THE UNITED STATES, 2010-2015

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Mingxi Pan

August 2017

© 2017 Mingxi Pan

ABSTRACT

This study contributes to basic knowledge of the determinants of house value in the United States by examining factors related to demography, time lag, accessibility to amenities, and house value spillovers using spatial data analysis and statistical learning techniques. Household type, education attainment, unemployment rate (and its time lag variable), industry by occupation, median household income (and its time lag variable), accessibility to transportation and spillovers effect are found to be important determinants of house value in core based statistical areas in the United States.

BIOGRAPHICAL SKETCH

Mingxi Pan, who was born in Hebei, China, is currently a Master Candidate in Regional Science at Cornell University. She also received her Bachelor's Degree in Management in Land Resource Management from Nanjing Agricultural University. Over time, she pursues her interests in socio-economic analysis of urbanization and spatial analysis.

To my grandfather Yulu Jiang

ACKNOWLEDGMENTS

I would first like to express my gratitude to my thesis advisor Professor Kieran Donaghy for advising and inspiring me in the completion of my thesis as well as in my entire education at Cornell; Professor Jennifer Minner for her valuable guidance and patience. I would also like to thank all the staff from Department of City and Regional Planning for the generous support.

I extend special thanks to Professor David G Rossiter of the School of Integrative Plant Science and Professor Yang Ning of the Department of Statistical Science at Cornell University, for their patient help and sharing their insights in developing this work. I would also like to thank my senior colleague Ziye Zhang, who encouraged me by answering my questions and generously sharing his experience of research. Thank my dear friends, Lu Liao, Yijun Yang, Binzhe Wang, Yicong Yang, Meng Yu, Jie Xu, Beijia Zhang, Kaiqi Bian, and my local host family Deborah Schmidle and Tim Schmidle, who shared my happiness and sorrows during the process.

I want to thank my dearest parents for their continuous encouragement and endless love that supports me and empowers me to pursue my dreams. Thank you!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CBSA	Core Based Statistical Area
OMB	Office of Management and Budget
MSA	Metropolitan Statistical Area
PCA	Principal Components Analysis
OLS	Ordinary Least Squares
SAR	Spatial Autoregressive Model

CHAPTER 1

LITERATURE REVIEW

Overview

Householders care more about house value because house values are a life cycled[1] and dominant part of their living costs.

One interesting housing behavior was noted by Harvard economist Glaeser (1998) that people are paid more in larger cities and they are also willing to pay more for housing in those cities.

Although it is acknowledged that house value is influenced by regional economic, political, social, environmental and historical variables, factors of house value are still one of the most challenging topics for social scientists and policymakers to study by monitoring change over regions and time.

From previous studies, governments of developing countries care more about how to provide housing inexpensively that allow the poor to have access to economic opportunity and are more policy oriented (Green, 2014). While developed countries focus more on examining the determinants of the house value. In this study, I examined the impact of population factors, accessibility to amenities and spatial spillovers on house values.

Spatial Structure of Urban Area

The US population has moved from rural to urban areas and from smaller towns to larger cities (Metropolitan) since the country's founding (Boustan, Bunten,

[1] Most of people only buy one house or a few houses during their whole life. Therefore, they pay much attention on housing values.

Hearey, 2013). According to data from World Urbanization Prospects[2], the US urban population reached 81.4% in 2014. Metropolitan areas are geographic units that are defined by the Census Bureau to include one or more contiguous counties anchored by a central city of a sufficient size (Boustan, Bunten, Hearey, 2013). Now the term of Core Based Statistical Area (CBSA) is used to replace the definitions of metropolitan areas by the Office of Management and Budget. The term CBSA[3] is a collective term for both metro and micro areas. A metropolitan area contains a core urban area of 50,000 or more population. A micropolitan area contains an urban core of at least 10,000 but less than 50,000 population. Figure 1 displays the metropolitan areas' and Micropolitan areas' locations for the CBSAs. Figure 2 displays the population density of CBSAs through 2010-2015.

[2] https://esa.un.org/unpd/wup/DataQuery/
[3] http://cber.cba.ua.edu/asdc/metro_micro.html

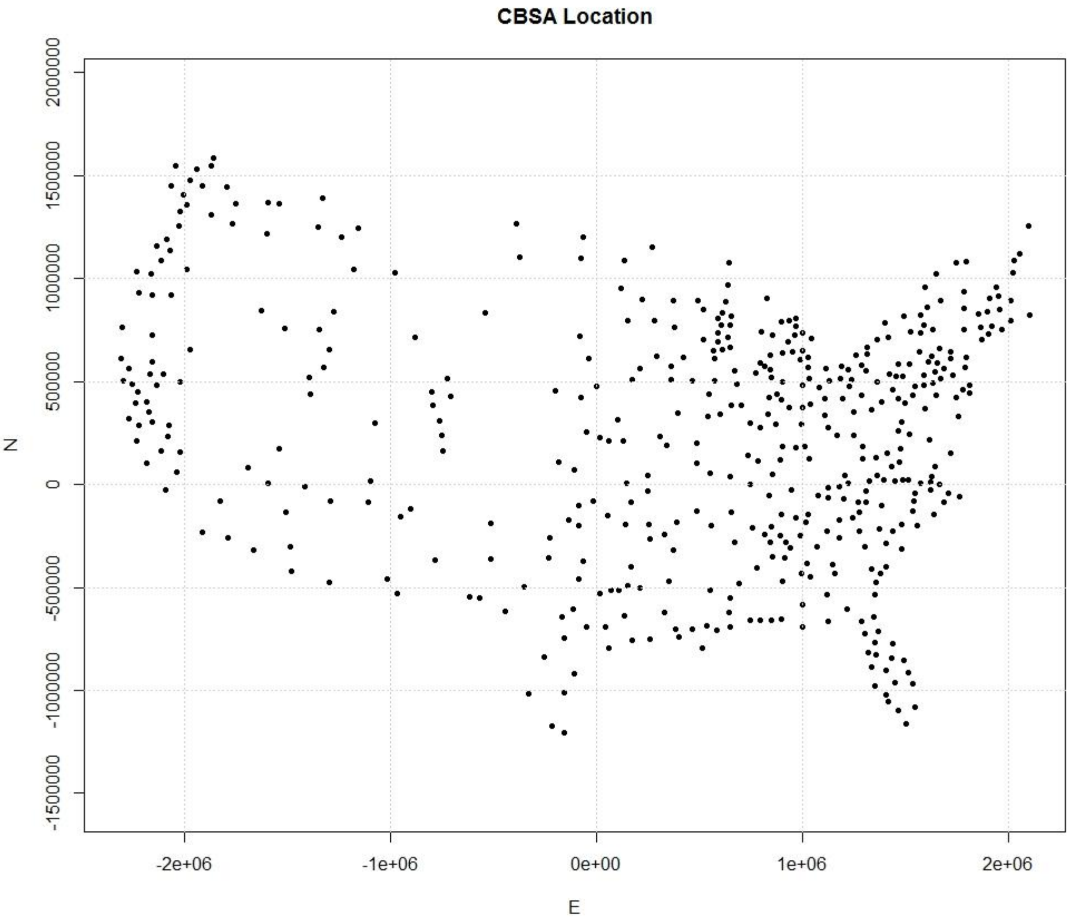

Figure 1 CBSA Location

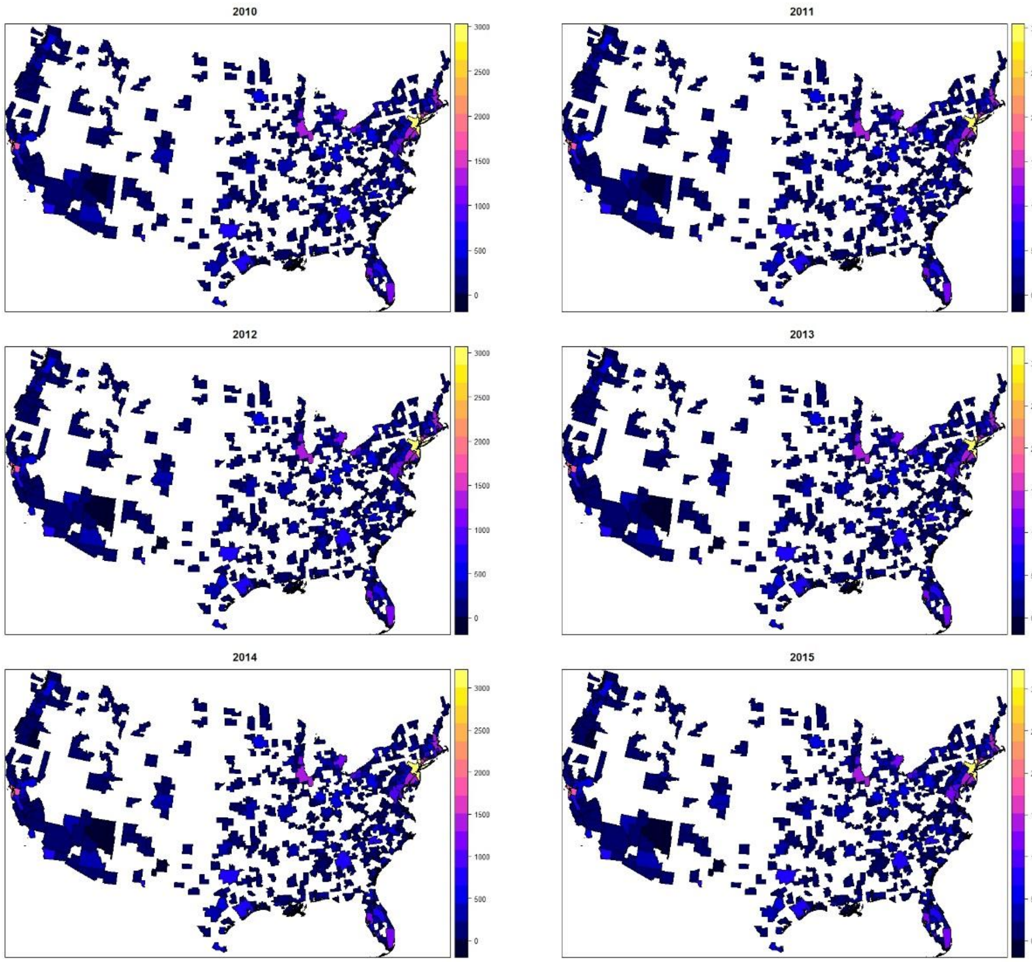

Figure 2 Population Density Trend through 2010-2015, Continental US

Boustan (2013) argued that both households and employment have relocated from the central city to the suburban ring over the twentieth century due to rising incomes and falling commuting costs, which can be explained by the existing monocentric models of urban land use(Alonso, 1964; Muth, 1969; Mills, 1972). These models emphasize the tradeoff between accessibility to the city center and space (distance to the city center). The phenomenon has been illustrated by the population density trend above, i.e., although the general trend is stable through years (2010-2015), Southern California regions such as San Francisco, Los Angeles, have lower population density than other west coastal areas in recent years.

House Value within Metropolitan areas

Although it is accepted that cities are dense collections of people, cities are also made up by housing units, and houses are often more durable than the city's population because people are transient.

Across metropolitan areas, the Rosen-Roback model characterizes the tradeoff between income, amenities and housing costs. The Rosen-Roback framework (Rosen, 1979; Roback, 1983) models character of a group of workers and firms, each free to move between cities with fixed quantity of lands and different amenity levels. Amenity level includes both consumer amenities and productive amenities. Workers in city *i* receive wages and pay for living. Wages and rents adjust until firms and workers are indifferent whether or not moving to other locations.

Based on this framework, Glaeser (2008) mentioned housing prices are a function of exogenous population and exogenous shocks to wages and amenities. Due to the complexity of the housing market within CBSAs, we take the accepted aspects of house value factors (representing population variables, income, and amenity) based on the Rosen-Roback Model and explore related previous studies. Figure3 shows the historical trend of house value from 2010 to 2015 within CBSAs. The figure shows that the highest house value is clustered in CA and northeast regions, and the relatively low house values are in the mid part of the United States. The general trend of house value is increasing over time (the color of the same region became lighter through years).

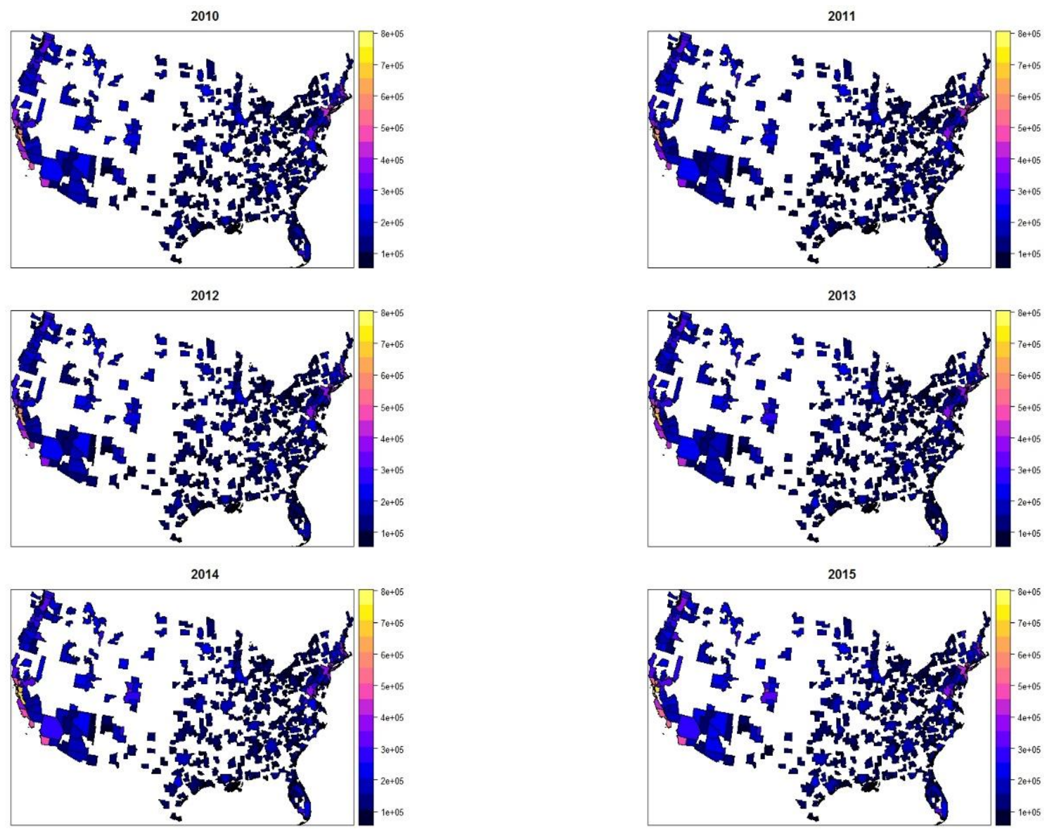

Figure3 House Value Trend, Continental US

Demographic Characteristics and House Value

Demographic change in the United States has been observed by metropolitization in general and suburbanization within metropolitan areas (Hanushek, Yilmaz, 2011).

It is accepted that there is dramatic mobility of the US population. In metropolitan areas, high-income residents tend to occupy houses that are newer, larger, and of higher quality and filter the older units down to lower-income denizens of the city (Boustan, Bunten, Hearey, 2013).

The Tiebout model (Tiebout, 1956) describes population sorting among a set of jurisdictions. In the Tiebout model, mobile households are free to choose amongst a variety of municipalities offering different bundles of public goods and tax rates (Boustan, Bunten, Hearey, 2013). The Tiebout model characterizes the interplay between demands for public goods and households' choice of locations.

Households' choice of locations will influence the housing demand within regions, thus influencing the overall house value. Housing and life-cycle decisions made by these households affect house values and can be directly linked to demographic variables (Reed, 2016).

Miles (2012) develops a model of the housing market with the property that house price rises relative to incomes in the evolution of population density. Albouy and Stuart (2014) used three broad amenity types including quality of life, productivity in tradable sector, and productivity in non-tradable sector to show that (population) density information can provide or refine measures of land value and local productivity. Reed (2016) confirmed that households with specific demographic characteristics are closely associated with a certain level of house prices at the suburban level based on a case study of Melbourne, Victoria, in Australia. Li (2014) found that neighborhood with more homogeneous minority populations commands higher prices in Vancouver, Canada. Brasington (2015) examines the impact that neighborhood race, age, income, and education segregation have on housing prices in seven MSAs in Ohio.

However, few studies have examined whether demographic characteristics models on house value are important across regions. Moreover, few studies have focused on monitoring change over time continuously and including or combining with other factors that would affect the purchasers' decision such as accessibility to amenities and spatial factors.

Accessibility to Amenities and House Value

Let's take an example of two big cities, Philadelphia and Phoenix. With the similar population size of the cities, i.e., Philadelphia and Phoenix, why is house value more expensive in Philadelphia? Philadelphia is well known as "the University City" and also an important traffic hub connecting south, north and west of US. Does this matter?

When one buys a house, one is buying a set of neighborhood amenities, schools, transportation systems, and taxes (Green, 2011).

Green (2011) also mentioned that house value varies a lot because of differences in transportation costs for materials. Pagliara (2011) indicated that transport impacts on residential location considered have significant effects on the attractiveness of residential location and hence on house prices. Hanna (2007) used the multiple-equation, fixed effects model attempts to test whether communities exposed to high levels of pollution will have lower housing prices and poorer residents than less polluted locations. Liu (2013) tracked the attribution of industry growth to housing prices over time. Thorsnes (2015) analyzed the spatial variation in natural amenities with household incomes and house characteristics and suggested large-scale effects of the public housing developments. Cucchiara (2013) pointed out that school quality becomes an urban amenity due to marketization of neighborhoods and the high level of demands for quality education through study in Philadelphia. Jensen (Jensen, Thursby, 2010) presented a theoretical model of faculty consulting in the context of government and industry funding for research within the university.

Few previous studies have included accessibilities to transportation across cities. In addition, few studies show the relationship between accessibility to universities and house value.

Spatial Spillovers of House Value

Most social phenomena are spatially correlated. It is accepted that there are spatial spillovers of housing prices. Meen (1997) interpreted that migration, equity transfer, information asymmetries and the spatial patterns in the fundamentals of housing prices play a key role in the spatial spillovers of house prices. Cho (2012) investigated neighborhood spillover effects with the rezoning of vacant parcels and housing prices in the Knoxville, TN area and indicated that the rise in housing prices in a neighboring location is expected to increase the probability of rezoning local vacant land. Vansteenkiste (2009) estimated a global vector autoregression model to assess the prospects for housing price spillovers in the euro area. Abelson (2013) used three spatial hedonic models to studying the effects of residential density and public transport on the median house prices in 626 suburbs across Sydney. Hui(2016) investigated the spatial spillover effects of urban land scape views and the accessibility to amenities on the property price in Central Business District (CBD) of Guangzhou city via spatial econometric analysis. Pijnenburg (2017) used a panel smooth transition regression model that allows for heterogeneity across time and space in spatial housing price spillovers and for heterogeneity in the effect of the fundamentals of house price dynamics to test the disposition effect.

Other factors could also cause spatial spillovers of house values. It is acknowledged that productivity advantages that accrue to urban firms and workers are influenced by their close proximity to one another (Boustan, Bunten, Hearey, 2013). There are spillover effects between industries which share ideas, inputs and output linkages and workers among regions. Badinger (Badinger, Egger, 2007) used a spatial autoregressive residuals model to model intra and inter industry productivity spillovers and found that intra-industry remainder spillovers turn out economically more significant than R&D spillovers. The productivity spillovers could cause clusters across regions and affect the housing demand and cause spatial spillovers of house values.

The house value reflects the value of land, therefore can reflect the value of location.

Few previous studies have included demographic factors especially demographic of industry factor associated with spatial spillovers of housing value across cities.

Conclusion

Based on the Rosen-Roback model, this research examines impacts of population factors, accessibility to amenities and spatial spillovers on house value. In this research, a case study was used within CBSAs in the Continental US from 2010 to 2015.

The contribution of this study builds upon the traditional location theories. I combined house value with both attributes of population factors and of accessibility to amenities, and explore a time sequence through 2010 to 2015 within CBSAs, which have not been taken into consideration yet. Moreover, I examine the housing value spatial spillover effect of demographic factors, especially industry occupation factors, and accessibility to amenities on house values.

CHAPTER 2

DATA SOURCE AND VARIABLE PRE-SELECTION

This study focus on examining the impact of demographic variables, accessibility to amenity variables, and spatial spillovers on house values across regions. My hypothesis is that house values are a function of exogenous population and exogenous shocks to wages and amenities (Glaeser, 2008). Meanwhile, high house values indicate household location preferences and investment opportunities for housing development. House values are a function of exogenous spatial spillover effect as well. We use quantitative methods to conduct the analysis. Furthermore, we would like to use the evidence to understand the structure of house values across cities and provide policy suggestions.

Demographic Data Source

In this study, demographic statistical data are collected from Social Explore[4], which is a website that can help users visualize dynamic maps and demographic data reports.

I used its 2010-2015 American Community Surveys (1-Year Estimates) data[5] and selected the geographic type by core based statistical areas (CBSA) within the Continental US. I excluded HI, AK and Puerto Rico due to low population.

Spatial Data Source

Shapefiles for CBSAs are downloaded from DATA.GOV[6], which is the U.S. Government's open data website.

---

4 http://www.socialexplorer.com/

5 We chose to use the 1-year estimates because this study focus on analyzing large populations and most current data

6 https://www.data.gov/

Shapefiles for Amtrak station and airport are downloaded from National Transportation Atlas Database.

Shapefiles for college and university are downloaded from ArcGIS official website.

Moreover, we calculated the accessibility to amenities data i.e., Amtrak station counts, airport counts, and university counts in ArcGIS.

Dependent Variable

We chose Median House Value for all owner-occupied housing units during 2010 to 2015 as our dependent variable. The house value is mixed with all the types of houses, i.e., house and lot, mobile home and lot, condominium unit. Owner-occupied means the owner or co-owner lives in the unit (even if it is mortgaged or not fully paid for).

Independent Variables

I chose demographic attributes across six years (2010-2015) as follows. Table 1 shows the attribute categories of the demographic variables:

Population Density (per sq. mile)

This variable was included to understand and control for the effects of population density on housing prices. It has been argued that housing value growth was faster in denser metropolitan areas during the 1980s. However, metropolitan area level population density was negatively associated with price growth from 1996-2006 (Glaeser, 2012). Therefore, it is hard to see the clear relationship between population density and house value potentially due to the suburbanization in recent decades or complexity of the dependent variable (house value).

Age

Different age cohorts affect property values where older households have different housing needs compared to recently formed households (Barrios, 2013). I divided the age cohorts into three groups: young (0-18), adults (18-64), and senior (65+). In previous studies, scholars focus more on the micro perspective of how age affects house value. My hypothesis is that in the macro perspective, adults potentially affect house value the most since they are the main housing consumers.

Race

Due to the long history of racial segregation and the long-term effects of racism and discrimination, black residents as a whole have accrued less capital and have limited access to financing. They were disproportionately affected by predatory lending and the effects of the sub-prime mortgage crisis. Neighborhood characteristics that are associated with lower housing prices must be those that the marginal homebuyer[7] seeks to avoid (Boustan, 2013). A "tipping" phenomenon happened because of the potential for rapid neighborhood transition from majority white to majority black (Schelling, 1971). Growing racial diversity should have impacts on house values. Since the United States is a famous multicultural country, I hypothesize that race may affect house value across regions.

Household Type

Costa and Kahn (1999) explored the location choices of dual-career couples and found college-educated couples are more likely to live in large metropolitan areas as wages increase. I hypothesize that couples or nonfamily population within regions may influence the house values.

Education Attainment (For population 25 years and over)

Increases in educational attainment were found for people in all U.S. Census defined racial groups between 1960 and 2009 (Hanushek, Yilmaz, 2011). Glaeser (2012) concluded that price growth was dramatically higher in less educated cities with higher initial housing values. I included this variable to examine the effect of education attainment on housing values.

[7] Outliers who pay huge premiums over the "consensus price" that experienced agents would come up with based on previous sales and micro market factors. https://www.theglobeandmail.com/real-estate/toronto/marginal-home-buyers-setting-market-value-in-toronto/article34311849/

Unemployment Rate (For civilian population in labor force 16 years and over)

Over the twentieth century, Metropolitan employment has increasingly left central cities (Boustan, Bunten, Hearey, 2013). I hypothesize that unemployment rate has a negative impact on house value.

Industry by Occupation (For employed civilian population 16 years and over)

Information-based industries like finance have not decentralized but the majority of cities have experienced ongoing employment decentralization since at least 1950 (Boustan, Bunten, Hearey, 2013). Firms in related industries can more easily share ideas, inputs and output linkages, and workers which cause the agglomeration (Marshall, 1961). Population from different industries have different mobility and different housing affordability. The spillover effects between industries may also affect house value spillover effects. We included the proportion in certain industry sectors as the variable.

Median Household Income (In 2010 inflation adjusted dollars)

High-income residents tend to occupy houses that are newer, larger, and of higher-quality (Boustan, Bunten, Hearey, 2013). Because people are paid higher wages on average in larger cities, they are also willing to pay more for land and thus housing in those cities (Green, 2011). I hypothesize that income and house value have a strongly positive relationship.

Residence 1 year ago in the United States

Boustan (2013) mainly emphasized the dramatic mobility of the US population. This variable reflects the population mobility across regions directly within one year. I hypothesize that population mobility would affect house value.

Table1
Demographic Attribute Table

Attributes	Population density	Age	Race	Household	Education	Unemployment rate	Industry by Occupation	Income	Residence
	Population density= land area/ total population	Numbers of people: young(0-18) adult(18-64) senior(65+)	Numbers of people in: White alone Black or African American alone Asian alone	Numbers of people in: Married couple family Nonfamily households (male or female householder)	Numbers of people in: Bachelor's degree Master's degree	Unemployment rate=unemployed population/ labor force population	Proportion of People in: Agriculture, forestry, fishing and hunting, and mining Manufacturing Retail trade Finance and insurance, and real estate and rental and leasing Educational services, and health care and social assistance	Median household income (in 2015 inflation adjusted dollars)	Numbers of people in: Same house 1 year age Moved within same county Moved from different county within same state Moved from different state Moved from abroad

I chose accessibility to amenities variables as follows. These data do not have any time sequence:

Accessibility to transportation

Number of Amtrak Station[8] and Number of Airports

With the falling of transport costs, manufacturing establishments no longer needed to locate close to their customer base (Boustan, Bunten, Hearey, 2013). This producer amenity will attract firms to the area with accessibility to transportation and increase the housing demand, therefore, charge higher house value.

Accessibility to transportation also gives people more mobility across cities. Transportation improvements reduce the time cost of travel and households can relocate further.

To represent accessibility, I used counts instead of distance because it is hard to define the exact distance to transportations across cities. Figure 4 and Figure 5 show the Allocation of Amtrak Stations and Airports in CBSAs.

[8] This variable was included to understand and control for the effects of residents' accessibility to amenities on housing prices since Amtrak is a passenger service.

ALLOCATION OF AMTRACK STATION WITHIN CBSA (CONTINENTAL US)

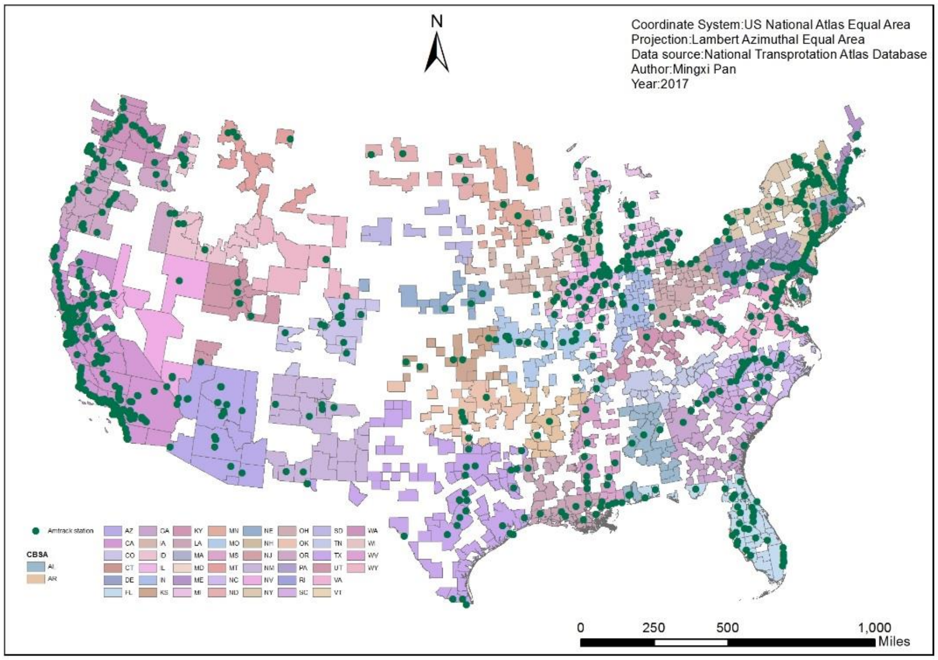

Figure 4 Allocation of Amtrak Stations

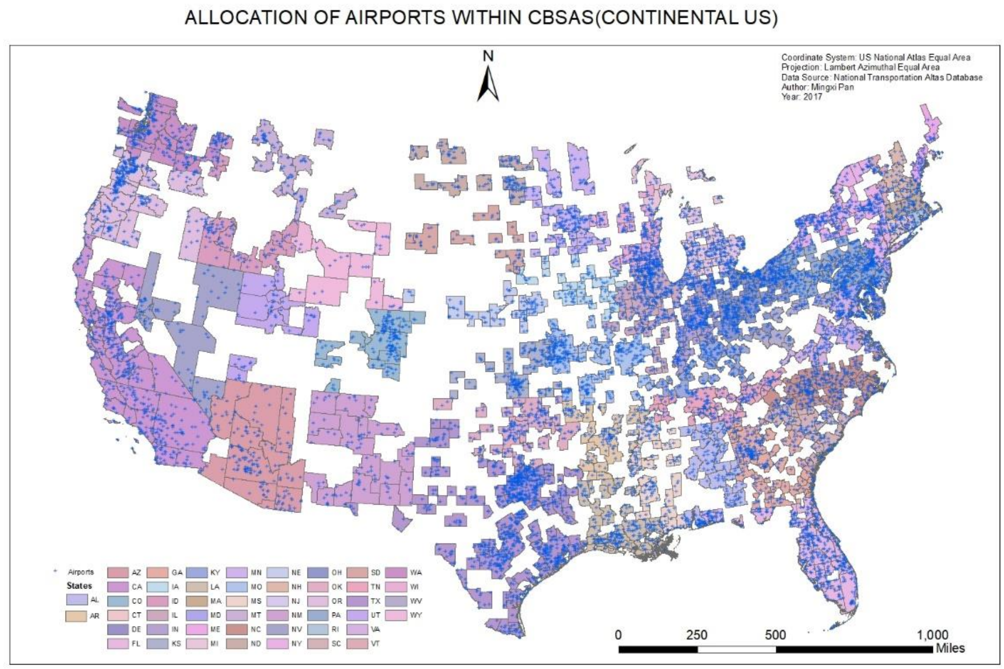

Figure 5 Allocation of Airports

Accessibility to University

Number of Colleges and Universities[9]

The location of central places was reinforced by human investments in supportive infrastructure, to which Cronon (1992) refers as second nature[10]. Luttik (2000) finds that houses with a park (garden) view require an extra premium of 8% in the Netherlands.

Most colleges and universities are open to the public. Residents always use university libraries, attend classes, and enjoy other amenities, i.e., go to the university concerts. Most of the colleges and universities own good view[11]. For example, Cornell University wants to build “Garden Campus” to benefits the residents around. This

[9] In this study, we only include colleges, universities, and professional schools according to NAICS catagolories

[10] Human-constructed nature

[11] There are also utilitarian campuses without much green space

consumption amenity will attract workers to the area and thereby drive up the housing prices. Figure 6 shows the allocation of Colleges and Universities in CBSAs[12].

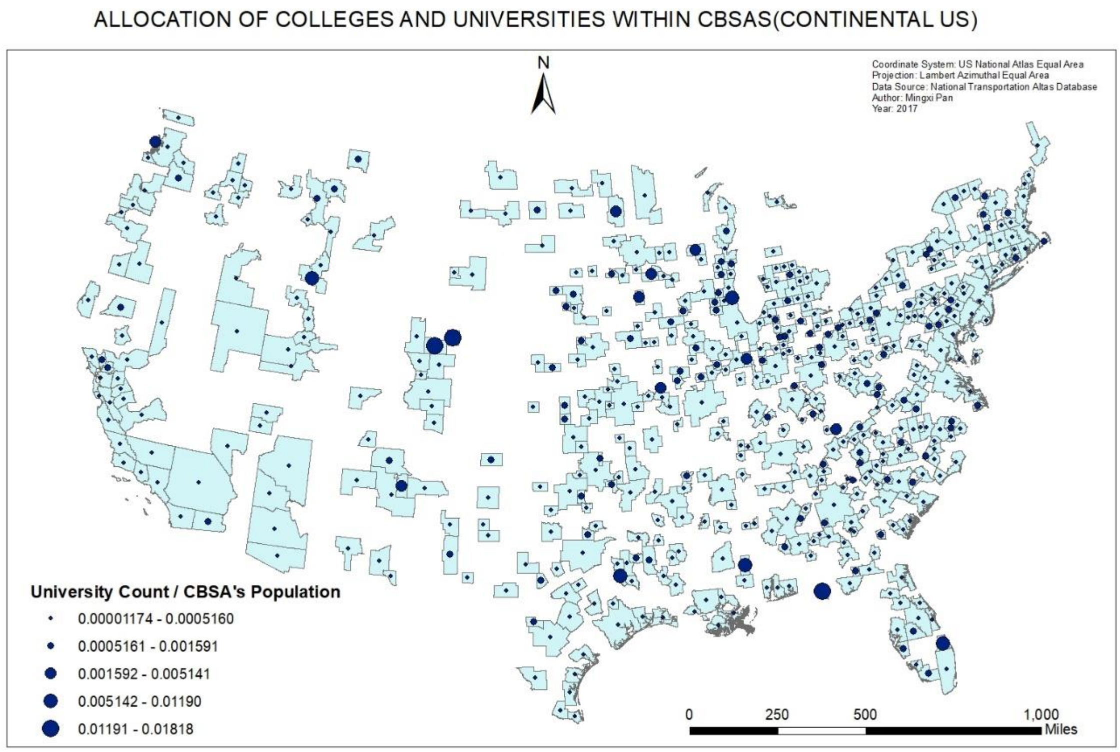

Figure 6 Allocation of Colleges and Universities

12 We showed the location of colleges and universities by count of colleges and universities and normalized by CBSA's population due to the number of colleges and universities will be highly correlated with population

CHAPTER 3

DEMOGRAPHIC VARIABLE SELECTION: PRINCIPAL COMPONENTS ANALYSIS

Multi-collinearity Problem

It is acknowledged that if there are variables that are highly correlated, with other variables we end up with large R square and standard errors of parameter estimates, so it is unlikely that we get statistical significance. This is called multicollinearity.

My dataset contains twenty-three demographic attributes through six years' periods. Some of the attributes with absolute values are facing a multi collinearity problem. For example, the number of people in young age and nonfamily household numbers. In order to solve the collinearity problem, it is accepted that if one can sketch a theoretical model of cause-effect, one might consider Structural Equation Modelling (SEM). However, there is no strong theory on my pre-selected demographic variables. Then I considered principal components analysis (PCA) to solve the problem.

Explain PCA: Correlation Matrix, and Eigen Decomposition

PCA, or Factor analysis, is a popular unsupervised (do not have response variable) tool used for data visualization (James, 2013) and reduction. Therefore, it is a data mining approach and a good method to show the relations between variables.

PCA is a direct calculation of a matrix. PCA projects the original vector space (matrix made up of the original variables) onto a same dimensional space by eigenvectors. In the new (pc) space, the new synthetic variables are called principal components and they are orthogonal to each other, in other words. completely uncorrelated. The synthetic variables are arranged in decreasing order of variance explained. In my research, I scaled the data only to get the correlation matrix.

The key insight is that the eigen decomposition orders the synthetic variables into descending amounts of variance, and ensures they are orthogonal (Hotelling, 1933).

Relevance to Earlier Studies

Reed (2016) used principal components analysis (PCA) to identified social dimensions from a range of demographic variables. Reed (2016) mentioned that the attributes have a high level of multi-collinearity and therefore are unsuitable for direct input into an ordinary least squares (OLS) analysis. He emphasized the powerful ability of PCA to collapse multiple demographic variables into a smaller set of uncorrelated factors.

In this study, I performed PCA on demographic variables by each year and also looked at loadings (eigenvectors) and biplots to select the most representative indicators for each of PC1-7.

Preliminary PCA

In the study, the pre-selected twenty-three demographic attributes were allocated into nine groups (Table 1). Figure 7 shows two-dimensional PCA plots by each year that function as data visualizations. Table 2 lists the Demographic PCA factors with loadings I selected for each of PC1-7.

The PC1/2 biplots and the loadings in Table 2 show that:

In 2010, Age, Household, Education attainment, and Residence are highly correlated and dominate PC1. This set of variables has already explained 65.2% of the total variance. The Unemployment rate, Income, and Retail trade together determine PC2 for the most part. These explain 7% of the variance and are almost independent of PC1. In PC1, the largest absolute loadings are Adult, Married couple family, Bachelor's degree, and Resident one year ago attributes (about 0.257), negative loadings are not strong in PC1. Therefore, PC1 represents an overall strength of Age (adult), Household (married couple family population), Education (bachelor's degree population) and Residence (same house one year ago). In PC2, it represents strength of Income (0.412), Retail trade (0.414), and negative loading Unemployment rate (-0.558).

In 2011, Age, Household, Education attainment, and Residence are highly correlated and dominate PC1. This grouping has already explained 65.1% of the total variance. The Unemployment rate, Income, and Retail trade together determine PC2 for the most part. These explain 6.8% of the variance and are almost independent of PC1. In PC1, the largest absolute loadings are Adult, Married couple family, Bachelor's degree, and Resident one year ago attributes (about 0.257). Negative loadings are not strong in PC1. Therefore, PC1 represents an overall strength of Age (adult), Household (married couple family population), Education (bachelor's degree population) and Residence (same house one year ago). In PC2, it represents strength of Income (0.53), Retail trade (0.39), and negative loading Unemployment rate (-0.55).

In 2012, Age, Household, Education attainment, and Residence are highly correlated and dominate PC1. This grouping has already explained 65.2% of the total variance. The Unemployment rate, Income, and Finance together determine PC2 for the most part. These explain 6.6% of the variance and are almost independent of PC1. In PC1, the largest absolute loadings are Adult, Married couple family, Bachelor's degree, and Resident one year ago attributes (about 0.257). Negative loadings are not strong in PC1. Therefore, PC1 represents an overall strength of Age (adult), Household (married couple family population), Education (bachelor's degree population) and Residence (same house one year ago). In PC2, it represents strength of Income (-0.55), Finance (-0.41), and Unemployment rate (0.59)[13].

In 2013, Age, Household, Education attainment, and Residence are highly correlated and dominate PC1. This grouping has already explained 65.4% of the total variance. The Unemployment rate, Income, and Finance together determine PC2 for the most part. These explain 6.9% of the variance and are almost independent of PC1. In PC1, the largest absolute loadings are Adult, Married couple family, Bachelor's degree, and Resident one year ago attributes (about 0.256). Negative loadings are not strong in PC1. Therefore, PC1 represents an overall strength of Age (adult), Household (married couple family population), Education (bachelor's degree population) and Residence (same house one year ago). In PC2, it represents strength of Income (0.55), Finance (0.51), and Unemployment rate (-0.57).

In 2014, Age, Household are highly correlated and dominate PC1. This grouping has already explained 65.3% of the total variance. The Unemployment rate, Income, and Retail trade together determine PC2 for the most part. These explain 7% of the variance and are almost independent of PC1. In PC1, the largest absolute loadings are Adult, Married couple family attributes (about 0.257), negative loadings are not strong in PC1. So PC1 represents an overall strength of Age (adult), Household (married couple family population). In PC2, it represents strength of +Income (0.57), -Unemployment rate (0.55), - Retail trade (0.39).

[13] In 2012, the sign of unemployment rate is negative but the relative relationship to other variables are the same as previous years.

In 2015, Age, Household, and Education are highly correlated and dominate PC1. This grouping has already explained 65.3% of the total variance. The unemployment rate and Income together determine PC2 for the most part. These explain 7.1% of the variance and are almost independent of PC1. In PC1, the largest absolute loadings are Adult, Married couple family, Bachelor's degree attributes (about 0.257), negative loadings are not strong in PC1. So PC1 represents an overall strength of Age (adult), household (married couple family population), and Education (bachelor's degree population). In PC2, it represents strength of +Income (0.52), -Unemployment rate (0.57).

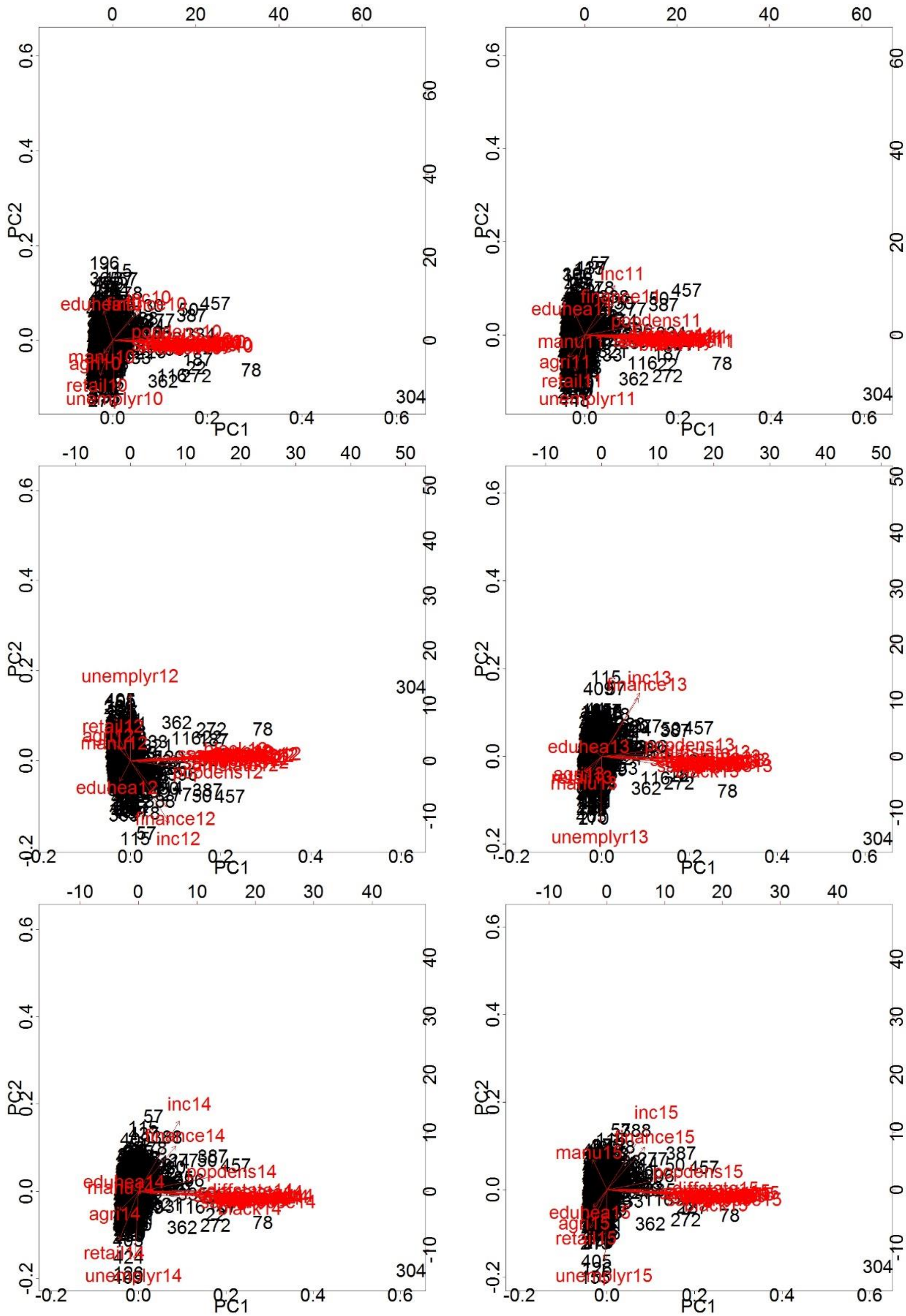

Figure 7 Principle Component Analysis Biplots from 2010 to 2015

Table 2
PCA Factors

	2010		2011		2012		2013		2014		2015	
	Factor name	Total Variance	Factor name	Total Variance	Factor name	Total Variance	Factor name	Total Variance	Factor name	Total Variance	Factor name	Total Variance
PC 1	Adult(0.257) Married Couple family(0.257) Bachelor's degree(0.257) Same house one year age(0.257)	65.3%	Adult(0.257) Married Couple family(0.257) Bachelor's degree(0.257) Same house one year age(0.257)	65.1%	Adult(0.257) Married Couple family(0.257) Bachelor's degree(0.256) Same house one year age(0.256)	65.2%	Adult(0.256) Married Couple family(0.257) Bachelor's degree(0.256) Same house one year age(0.256)	65.4%	Adult(0.257) Married Couple family(0.257)	65.3%	Adult(0.257) Married Couple family(0.257) Bachelor's degree(0.256)	65.3%
PC 2	Median household income(0.412) Unemployment rate(-0.558) Retail trade(0.414)	72.1%	Median household income(0.53) Unemployment rate(-0.55) Retail trade(0.39)	71.9%	Median household income(-0.55) Unemployment rate(0.59) Finance and insurance, and real estate and rental and leasing(-0.41)	71.9%	Median household income(0.55) Unemployment rate(-0.57) Finance and insurance, and real estate and rental and leasing(0.51)	72.3%	Median household income(0.57) Unemployment rate(-0.55) Retail trade(-0.39)	72.3%	Median household income(0.52) Unemployment rate(-0.57)	72.3%

PC 3	Educational services, and health care and social assistance(-0.61)	77.5%	Manufacturing (0.7)	77.4%	Retail trade(0.63) Manufacturing (0.6)	77.6%	Manufacturing(0.76)	77.9%	Manufacturing(0.76)	77.9%	Manufacturing(-0.59)	77.9%
PC 4	Agriculture, forestry, fishing and hunting, and mining(0.50) Manufacturing(-0.64)	82.6%	Educational services, and health care and social assistance(0.77)	82.5%	Agriculture, forestry, fishing and hunting, and mining(-0.68)	82.8%	Agriculture, forestry, fishing and hunting, and mining(-0.73)	83.2%	Agriculture, forestry, fishing and hunting, and mining(-0.76)	83.3%	Agriculture, forestry, fishing and hunting, and mining(-0.72)	83%
PC 5	Retail trade(-0.52)	87.5%	Agriculture, forestry, fishing and hunting, and mining(0.69)	87.6%	Educational services, and health care and social assistance(0.8)	87.9%	Educational services, and health care and social assistance(0.66)	87.8%	Educational services, and health care and social assistance(0.62)	88.2%	Educational services, and health care and social assistance(0.81)	87.9%
PC 6	Unemployment rate(-0.73)	90.9%	Unemployment rate(-0.72)	91.1%	Unemployment rate(0.69)	91.4%	Unemployment rate(-0.67)	91.5%	Unemployment rate(-0.74)	91.5%	Unemployment rate(-0.68)	91.4%
PC 7	Finance and insurance, and real estate and rental and	93.5%	Retail trade(0.63)	93.7%	Median household income(0.51)	93.9%	Median household income(0.58)	94.1%	Retail trade(0.41)	93.8%	Finance and insurance, and real estate and rental and	93.9%

leasing(- 0.51)	leasing(- 0.40)

PCA Model Result

The number of factors per year varied across years.

It is because the dependent variable (Median house value) is mixed with different types, i.e., house, condominium unit. It is hard to find a general trend between population density and house value or it is due to the unobserved social factors such as decentralization. I finally excluded population density in the PCA model.

I also excluded the race attributes due to low loading in PC. It is also because race attributes are more important in the inner city level of house values, i.e., racial segregation in a city causing higher housing values in white residential clusters than black residential clusters. However, it is hard to tell whether race attributes would affect the housing values across regions.

CHAPTER 4

ORDINARY LEAST SQUARES ANALYSIS

Time Lag Effect

In a previous study, Glaeser (2006) presented a dynamic model of market the housing. His model is consistent with time-lag price changes due to incomplete information or price rigidity of the market.

In my study, I hypothesize that house value factors may have time lag effects on house values. The response of house value is hysteretic.

In this study, I only lagged the income attribute and unemployment rate attribute due to the institutional effect. i.e., Laborers tend to be abler to buy houses if they have more surplus savings since last year. Consumers are likely to be less confident to buy expensive houses if the unemployment rate increased last year.

I created a time lag variable model:

$$Y_t = \beta_0 + \sum \beta_i \, x_{it} + \sum \beta_n \, x_{nt-1} + \varepsilon$$

Where i represents the attributes, n represents the attributes with time lag effects within t and t-1 time period, $n \in i$.

Model Diagnostics, Heteroscedasticity test, and Variable transform

I used a linear model to test my 2011-2015 house value models with time lag variables based on the PCA models. Figure 8 shows the clear tendency for positive residuals at low fitted values, negative residuals at medium. And the fitted values are unevenly distributed. There are only a few large values and most of them are small to medium. Figure 8 and 9 also shows that there are some extreme residuals. This may due to spatial correlation of the residuals. I test the models for spatial autocorrelation in later sections. Figure 8 and Figure 9 confirmed that the linear model is not justified due to the clusters and skewed distribution of residuals.

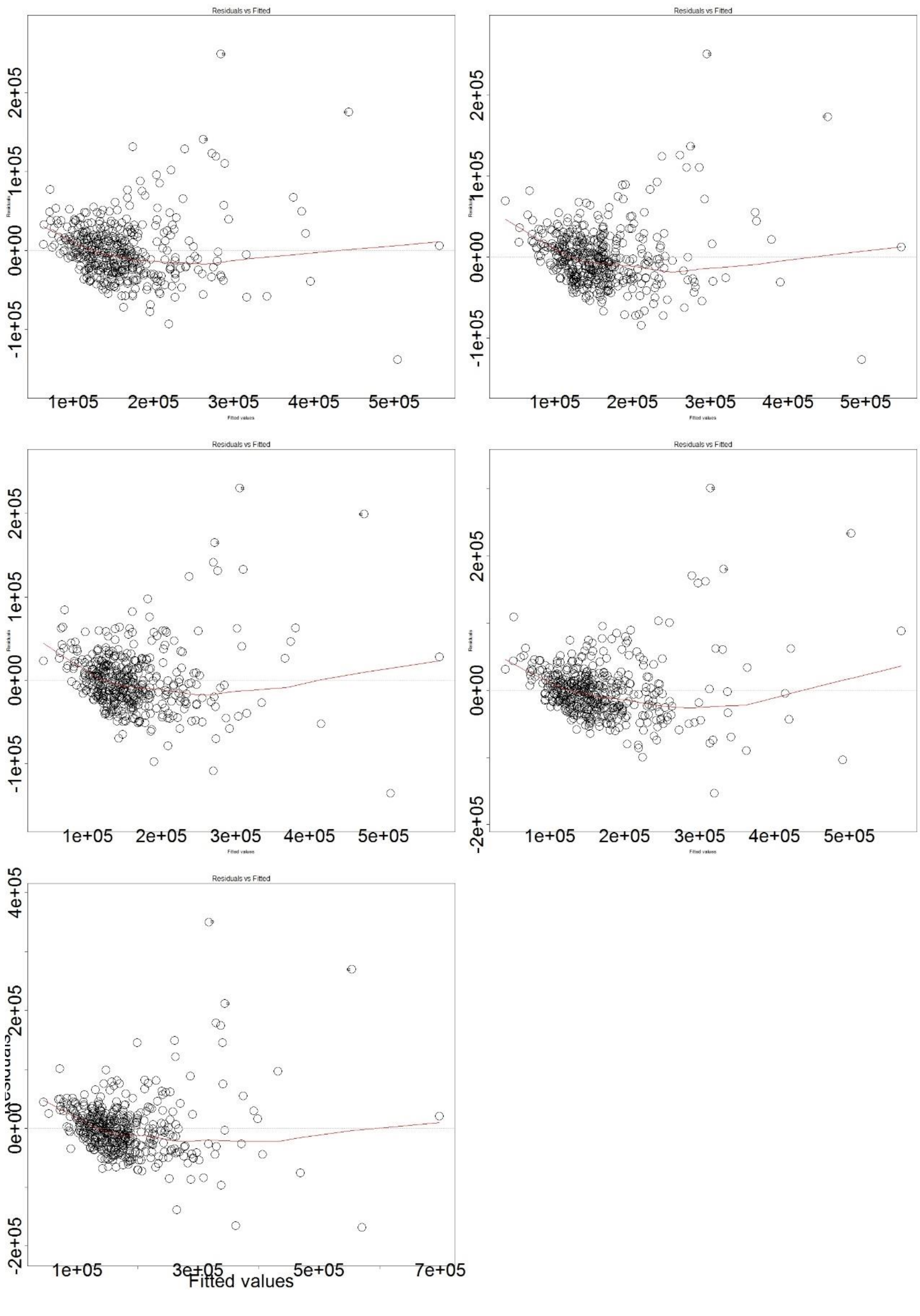

Figure 8 Residual vs Fitted Value Plots

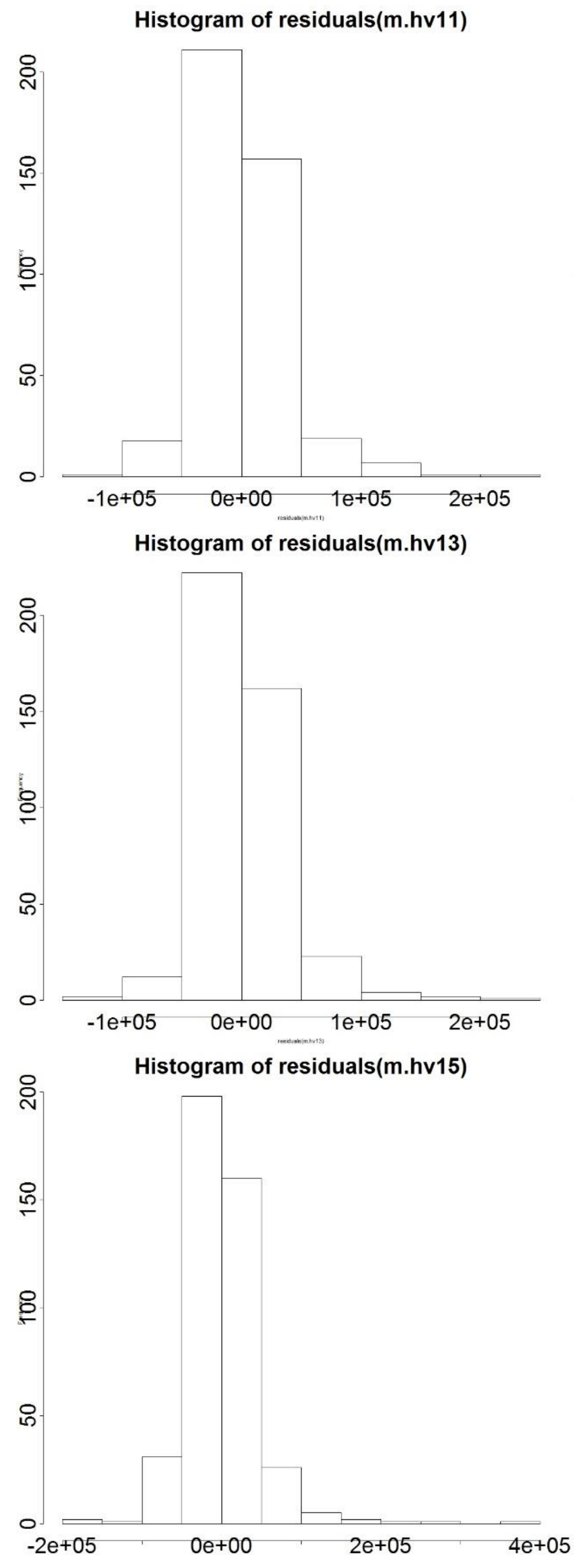

Figure 9 Histogram of Residuals

I tested the heteroscedasticity of each model using White's test and all of the models' p-value<0.05. Table3 shows the results of the White's test for each year model.

Table3

Results of the White's test for heteroskedasticity

	2011	2012	2013	2014	2015
p-value	1.165734e-14	6.994405e-15	0	0	1.110223e-16

Thus we could reject the null hypothesis (there is no heteroscedasticity in residuals). Therefore, heteroscedasticity exists in the models.

In order to reduce heteroscedasticity, I used log transformation on the dependent variable.

Ordinary Least Squares Analysis (OLS)

Based on the conceptual model and using the earlier definitions, I estimated a regression model of the following form:

$$Log(hv_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_n x_{nt} + \beta_{inc\ t-1}\ x_{inc\ t-1} + \beta_{unemplyr\ t-1}\ x_{unemplyr\ t-1} + \varepsilon$$

Where hv_t is the median house value in year t, x represents the variables selected from PCA in year t, x_{inc} and $x_{unemplyr}$ are the time lag variables in year t-1, and ε is the error term.

Then I ran the OLS model in R with robust standard error for heteroscedasticity. The models in Table 3 shows the results of OLS.

Table 4

OLS Results

	2011		2012		2013		2014		2015	
(Intercept)		10.27000 (***)		10.60000 (***)		10.59000 (***)		10.49000 (***)		11.13000 (***)
Factor1	amtrksta	0.02423 (***)	amtrksta	0.02581 (***)	amtrksta	0.03014 (***)	amtrksta	0.03360 (***)	amtrksta	0.03884 (***)
Factor2	airport	-0.00144	airport	-0.00117	airport	-0.00149	airport	-0.00212 (.)	airport	-0.00190 (*)
Factor3	university	-0.00075	university	-0.00089	university	-0.00083	university	-0.00063	university	0.00005
Factor4	agea11	0.00000	agea12	-0.00000	agea13	0.00000	agea14	0.00000	agea15	0.00000
Factor5	married11	-0.00000 (.)	married12	-0.00000 (.)	samehouse13	0.00000	married14	0.00000	married15	-0.00000
Factor6	bachelor11	0.00000 (*)	bachelor12	0.00000 (*)	married13	0.00000 (.)	unemplyr14	0.11180	bachelor15	0.00000
Factor7	unemplyr11	2.05000 (**)	unemplyr12	-0.52640	bachelor13	0.00000	unemplyr13	0.95800	unemplyr15	-1.59700 (.)
Factor8	unemplyr10	-1.0280	unemplyr11	1.95900 (**)	unemplyr13	-0.79410	agri14	-0.07395	unemplyr14	1.52400 (*)
Factor9	agri11	-0.0372	agri12	-0.24390	unemplyr12	1.80700 (**)	retail14	0.48480	agri15	-0.26770
Factor10	manu11	-1.2000 (***)	manu12	-1.40900 (***)	agri13	-0.43460	manu14	-1.25700 (***)	manu15	-1.74900 (***)
Factor11	retail11	0.79000	retail12	-0.38800	manu13	-1.23700 (***)	eduhea14	-0.28520	finance15	-1.73800 (*)
Factor12	eduhea11	-0.2202	finance12	-0.68860	finance13	-0.28410	inc14	0.00001 (**)	eduhea15	-1.04200 (***)
Factor13	inc11	0.00001 (**)	eduhea12	-0.19940	eduhea13	-0.35260	inc13	0.00002 (***)	inc15	0.00002 (***)
Factor14	inc10	0.00002 (***)	inc12	0.00002 (***)	inc13	0.00002 (***)			inc14	0.00001 (*)
Factor15	samehouse11	0.00000	inc11	0.00001 (**)	inc12	0.00001 (*)				
Factor16			samehouse12	0.00000						

Residual standard error	0.2025	0.1951	0.1962	0.2039	0.2117
R^2	0.6934	0.7033	0.711	0.7038	0.7005
Adjust R^2	0.6811	0.692	0.7005	0.6946	0.6904
F-statistic	56.27	62.22	67.59	76.94	69
Sample size	415	437	428	435	428

Significant level. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Age

Number of population in adult age variable does not contributed to an explanation of the variation in house values.

Household Type

Married couple family was excluded in 2013, 2014, 2015 model and was slightly negatively related to house value in 2011, 2012.

Education Attainment (For population 25 years and over)

Population of Bachelor's degree variable contributed to 2011, 2012 and 2013 models. It is slightly positively associated with house value.

Unemployment Rate (For civilian population in labor force 16 years and over)

The unemployment rate and its time lag variables were correlated with the house value through most of the years. In the 2011 model, the unemployment rate was positively associated with house value, and it was negative during 2015. The unemployment rate time lag variable contributed to 2012 and 2015 model and indicated a positive relationship with house value.

Industry by Occupation (For employed civilian population 16 years and over)

The proportion of the population employed in manufacture industry was observed in 2011- 2015 models. It was strongly negatively correlated with the house value. In 2015 model, the proportion of population employed in finance and the proportion of population employed in education and health were strongly negatively related to house value.

Median Household Income (In 2010 inflation adjusted dollars)

As expected the median household income variable and its time lag variable were strongly correlated with the house value for each year. The median household income variable was slightly positively related to house value for each year. The time lag variable was slightly positively associated with house value for each year. This suggests that a high income is linked to more savings to obtaining a high-value house.

Residence 1 year ago in the United States

The population of the same house one year ago was observed in 2011, 2012 and 2013 model. This variable does not contribute to an explanation of the variation in house values.

Accessibility to transportation

Accessibility to an Amtrak station was strongly correlated with the house value for each year. More Amtrak station is linked to accessibility to transportation and obtain a higher value house. The positive trend increases year by year. The accessibility to airport contributes to three models (2014-2015) and is negative related to house values.

Accessibility to University

This variable does not contribute to an explanation of the variation in house value.

CHAPTER 5

SPATIAL EFFECT ANALYSIS

Estimation issues

Previous studies have shown that house values are likely to have spillover effects in the spatial dimension. That is, if there are more links across regions (more neighbors), the house values are more likely be similar to each other. i.e., a high (low) value house is linked to another high (low) value house. This suggests that there may exist spatial effects rather than only social economic factors that contribute to house value. Spatial dependence is a violation of the independence of errors assumption of OLS regression analysis (because the OLS predictive factors may share the same spatial patterning). This may raise model specification issues. Moreover, location can represent the endogeneity of the house value. Therefore, in this chapter, I took consideration of the spatial effect on house value based on previous OLS models.

Spatial neighbors and spatial weights

Bivand (2013) mentioned that creating spatial weights can help check that there is no remaining spatial patterning in residuals due to different weights for different spatial allocation. The first step is to define which relationships between observations are to be given a nonzero weight, which is to choose the neighbor criterion to be used; the second is to assign weights to the identified neighbor links.

In this study, I tried to examine the spillover effect of house value within CBSAs. CBSAs are represented as polygons on the map. Each polygon is either linked with other polygons or with no links. I defined the polygons sharing boundary points as neighbors.

In the first step, I identified the contiguity neighbor links of CBSAs in R. By default, the contiguity condition is met when at least one point on the boundary of one polygon is within the snap (short) distance of at least one point of its neighbor (Bivand, 2013). Neighbor relationships between CBSAs are represented by Figure 10.

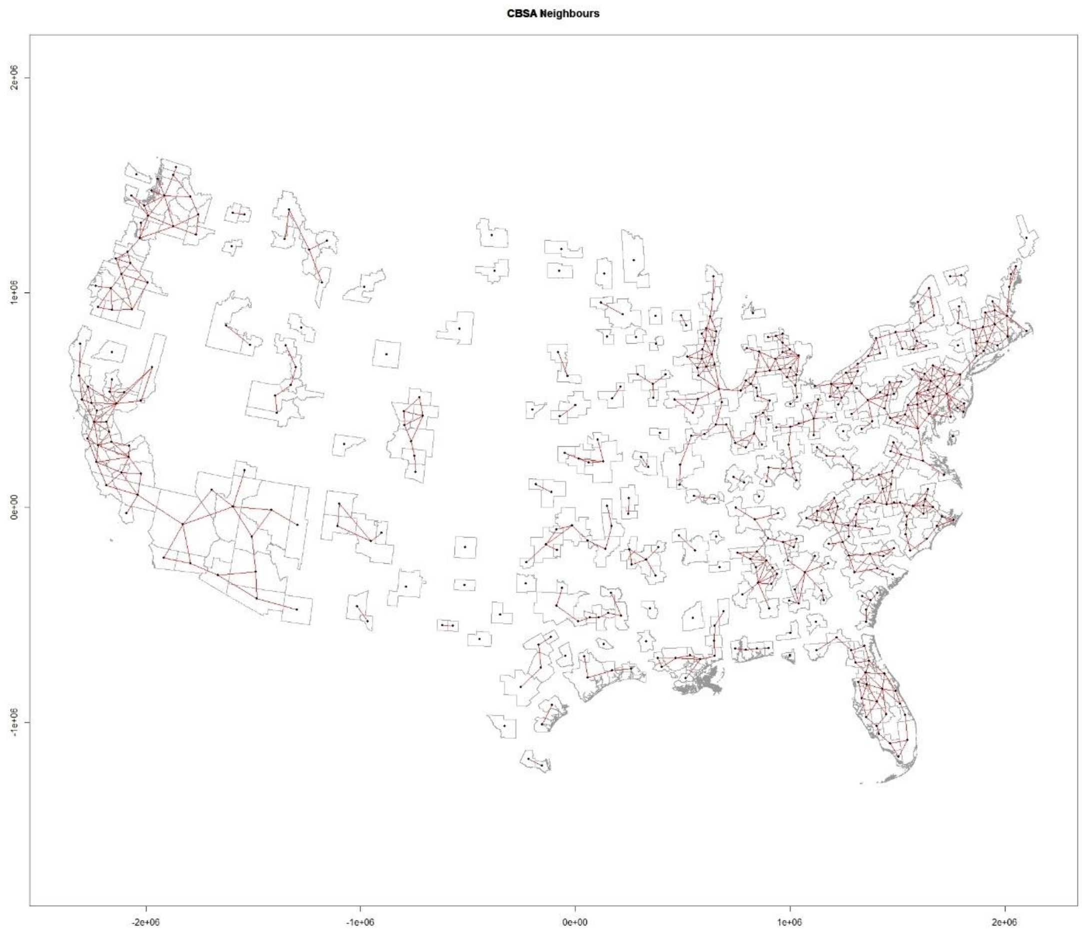

Figure 10 Neighbor Relationships between CBSAs

There are 1184 total links among the 479 polygons; the average number of links is 2.5; 43 of the polygons have no links and 110 of the polygons have only one link – these may be on the periphery of the area.

In the second step, I used the default style= “W”, in which the weights for each areal entity must sum to unity along rows of the weights matrix; this is the inverse of the number of neighbors. (Bivand, 2013) For example, in my study, the most connected regions are 107 and 227 (row number), there are three neighbors of 107: 41,114 and 295, and the weights are 0.333333 for each. There are four neighbors of 227, 207, 225, 366, and 424, the weights are 0.25 for each.

Spatial autocorrelation

When we have neighbors and their weights, we can determine whether there is any spatial autocorrelation. In my study, I tried test the extent to which house values in neighboring polygons similar. Moran's I statistic measures spatial autocorrelation based on both feature locations and feature values simultaneously (Li, 2007). Therefore, I used Moran's I to test the hypothesis.

Global tests

As a global statistic, Moran's I captures the existence of a homogeneous pattern of spatial association over the entire study area (Anselin, 1995). In my study, I used global Moran's I to test whether house value is or not spatially independent. Figure 11-15 shows the spatial relation of house value with a grey-scale plot for each year (the intensity of the gray proportional to the maximum proportion of house values).

Figure 11-15 Spatial relation of house value, CBSAs

From the figures, highest house values are in the California[14] for each year.

Then I computed global Moran's I to test the hypothesis that house values are more spatially clustered for each year, I used the default weighting: inversely proportional to the number of neighbors.

In 2011, the expectation of Moran's I is $-1/(n-1)$ =-0.002298851; the actual value (0.555142997) is of opposite sign and much larger in absolute value. The probability of incorrectly rejecting the null hypothesis of no spatial association (Type I error) is 2.2e-16.

In 2012, the expectation of Moran's I is $-1/(n-1)$ =-0.002298851; the actual value (0.546988171) is of opposite sign and much larger in absolute value. The probability of incorrectly rejecting the null hypothesis of no spatial association (Type I error) is 2.2e-16.

In 2013, the expectation of Moran's I is $-1/(n-1)$ =-0.002298851; the actual value (0.552099278) is of opposite sign and much larger in absolute value. The probability of incorrectly rejecting the null hypothesis of no spatial association (Type I error) is 2.2e-16.

In 2014, the expectation of Moran's I is $-1/(n-1)$ =-0.002298851; the actual value (0.570229613) is of opposite sign and much larger in absolute value. The probability of incorrectly rejecting the null hypothesis of no spatial association (Type I error) is 2.2e-16.

In 2015, the expectation of Moran's I is $-1/(n-1)$ =-0.002298851; the actual value (0.574919007) is of opposite sign and much larger in absolute value. The probability of incorrectly rejecting the null hypothesis of no spatial association (Type I error) is 2.2e-16.

[14] They are San Jose-Sunnyvale-Santa Clara CA Metro Area and San Francisco-Oakland-Fremont CA Metro Area.

Therefore, house value is probably spatially auto correlated with the default weighting.

Local tests

Anselin's (1995, 2003) "Moran scatter plot" plots a variable of interest against the spatial weighted component of that variable. Thus, a good way to visualize the relation between the global and local measures is to plot a Moran scatterplot. Figure 16-20 shows the Moran scatterplot for each year. The regression line is the global Moran's I. Points with high influence are identified by a special symbol and their CBSAFP number in the original dataset.

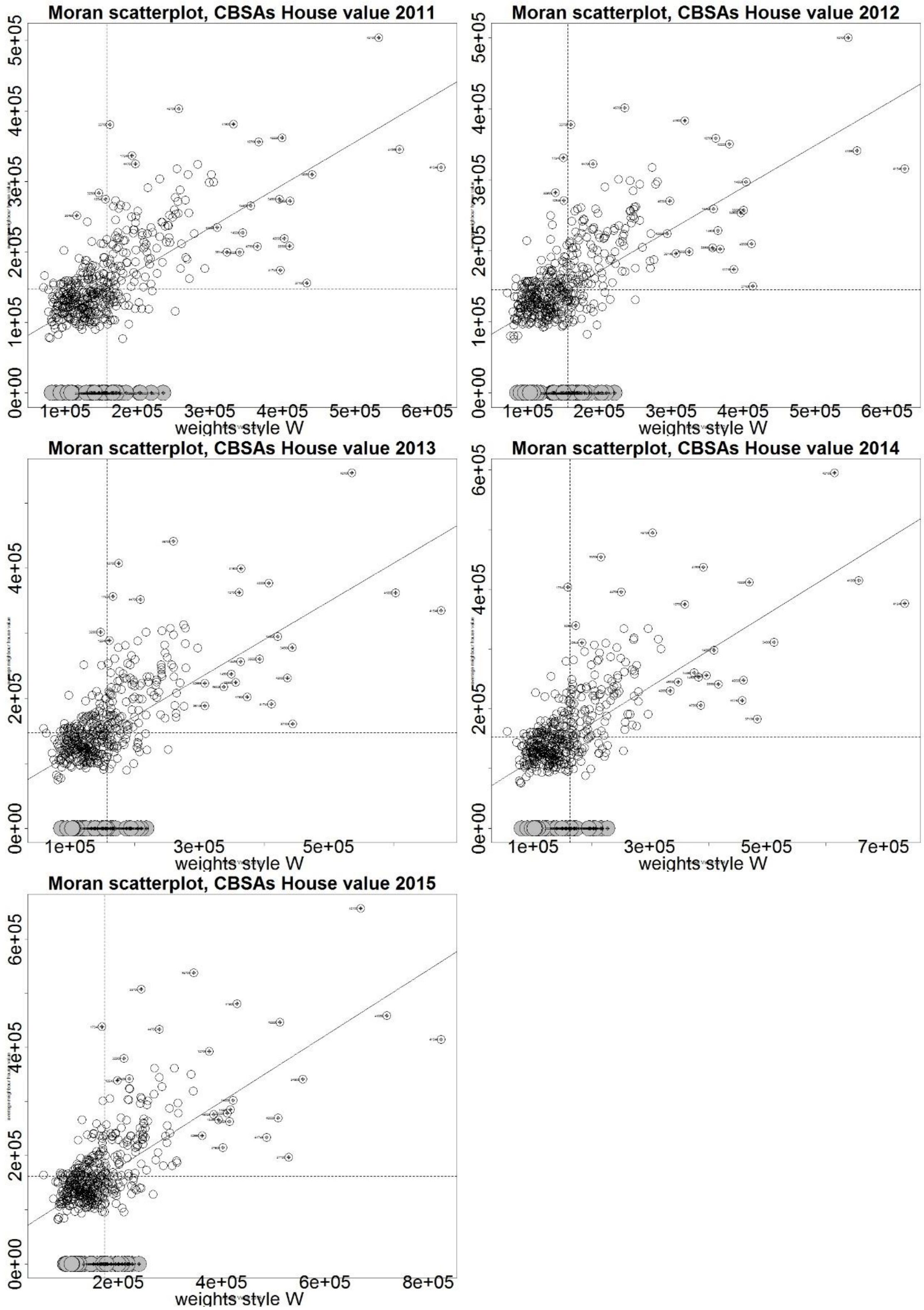

Figure16-20 Moran scatterplot for house values, CBSAs

In 2011, the highest-leverage area is 41940 (San Jose-Sunnyvale-Santa Clara, CA Metro Area), it has the highest house value and a moderately high weighted spatially-lagged proportion, this supports the hypothesis of autocorrelation. The area 41940 is adjacent to areas 32900, 41860, 42100, 41500, 33700, 23420. 42100 (Santa Cruz-Watsonville, CA Metro Area), 41860 (San Francisco-Oakland-Fremont, CA Metro Area) also have high proportion. Areas 32900, 41500, 33700, 23420 have moderately low house value, but high spatially-lagged proportion, these are the low house value neighborhoods adjacent to high house value neighborhoods. However, they have little influence on the slope, because they are almost directly above the centroid. The same trend happened in 2012, 2013, 2014, and 2015.

I plotted these as shaded polygons, with both high (HH), no influence (LL), low proportion neighbors (HL), and the reverse (LH) plots. Figure 21-25 shows the HH LL influence plots.

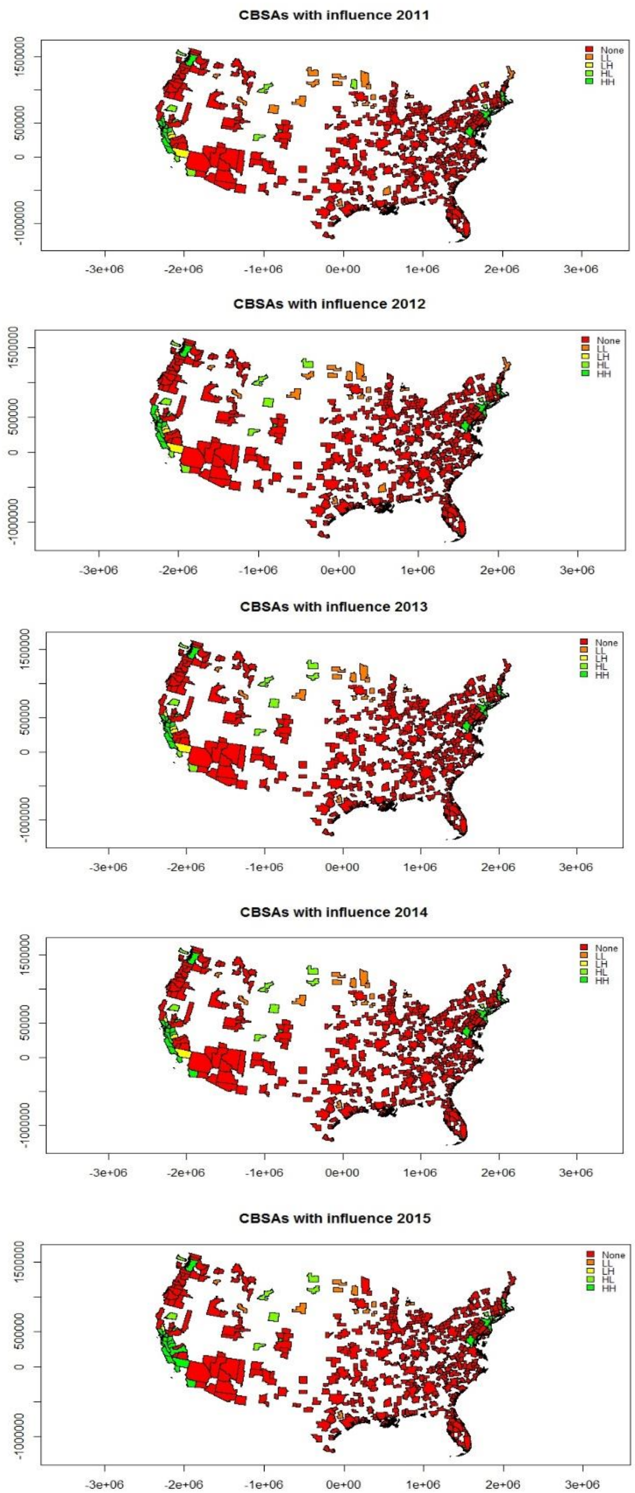

Figure 21-25 HH-LL influence plots for each year

From these HH LL influence plot figures, it is obvious that most of the global Moran's I significance comes from the local Moran's I, high house values associated with high house value, in the mid-CA area.

Then I computed the local Moran's I to test the hypothesis of spatial autocorrelation. For 2011, there are 46 areas with local Moran's I sufficiently high to reject the null hypothesis with less than a 5% chance of Type I error. For 2012 there are 48 areas, 2013 and 2014 45 areas, 2015 44 areas. Most of them are highlighted in the Moran scatterplots. i.e., 41940, 41860, 42100. These results show the evidence of local clustering.

Spatial correlation of residuals

I applied the Moran's test of the residuals for each year. The probability that we would be wrong to reject the null hypothesis of no spatial correlation is only 8.629e-09, 9.05e-11, 2.941e-11, 7.906e-15, 1.15e-13 for year 2011, 2012, 2013, 2014, and 2015.

Thus, we should refit the model as an autoregressive model. This accounts for spatial autocorrelation of the residuals by a regression on the values from adjacent areas. (Bivand, 2013) Figure26-30 shows the residual maps of house value for each year.

Figure 26-30 Residual maps of house value 2011-2015

There seems to be a cluster of high residuals near mid-CA, and another near north east and somewhere in the middle of United States. There are several areas of near zero clusters, i.e., it seems that they have fewer neighbors. This suggests that the linear model is not complete: we should refit the model as an autoregressive model and account for spatial autocorrelation of the residuals.

Spatial autoregressive model

Spatial autoregressive model (SAR) accounts for spatial autocorrelation of the residuals by a regression on the values from adjacent areas. The errors are modeled to depend on each other (Bivand, 2013). I employed SAR model to correct for spatial dependence in this paper:

$$\boldsymbol{Y = \lambda\, W\,(Y - X\beta) + X\beta + \varepsilon}$$

Where λ is the strength of this auto regression term, Y denotes an n*1 vector of the dependent variable (house value for each year), X represents an n*k matrix containing the significant determinants from OLS results of house value, and W is a spatial weight. I use the default style="W" I have already built. This is the inverse of the number of neighbors. We do not need to consider the edge effect according to this study[15].

I ran the expanded model in R and got the spatial results for the model. Table 4 shows the SAR model results.

Table 5

SAR model results

[15] It is because in this study, the periphery areas are the boarders of the country, they do not have edge effects from other regions.

	2011		2012		2013		2014		2015	
(Intercept)		10.56200		10.66800		10.73300		10.71000		10.76800
		(< 2.2e-16)[16]		(< 2.2e-16)		(< 2.2e-16)		(< 2.2e-16)		(< 2.2e-16)
Factor1	Amtrksta	0.01440	amtrksta	0.01771	amtrksta	0.01676	amtrksta	0.017543	amtrksta	0.01081
		(0.005679)		(0.0002977)		(0.0003611)		(0.00003)		(0.0048249)
Factor2	married11	-0.00000	married12	-0.00000	married13	-0.00000	airport	-0.00125	airport	-0.00051
		(0.002251)		(0.0010629)		(0.2750521)		(0.003584)		(0.2120314)
Factor3	bachelor11	0.00000	bachelor12	0.00000	unemplyr12	0.10742	unemplyr13	-0.95127	unemplyr15	-1.55830
		(0.003190)		(0.0019957)		(0.7651664)		(0.0184753)		(0.0094700)
Factor4	unemplyr11	1.02230	unemplyr11	0.57918	manu13	-1.02340	manu14	-1.15590	unemplyr14	-0.34055
		(0.009679)		(0.1095099)		(0)		(0)		(0.5060525)
Factor5	manu11	-1.17200	manu12	-1.1028	inc13	0.00002	inc14	0.00001	manu15	-0.96322
		(0)		(0)		(0)		(0.00006)		(0)
Factor6	inc11	0.00002	inc12	0.00002	inc12	0.00001	inc13	0.00002	finance15	-0.53937
		(0.006478)		(0)		(0.2056521)		(0)		(0.3098812)
Factor7	inc10	0.00001	inc11	0.00001					eduhea15	0.68222
		(0)		(0.0129696)						(0.0007776)
Factor8									inc15	0.00002
										(0.00000)
Factor9									inc14	0.00001
										(0.0008161)
λ		0.42541		0.43169		0.50207		0.50418		0.66355
LR test value		57.377		63.587		88.969		105.87		138.06
p-value		3.5971e-14		<2.22e-16		<2.22e-16		<2.22e-16		<2.22e-16
Sample size		419		437		428		435		428

[16] P-value were showed in the parentheses

The LR test value represents the likelihood ratio test, which compares the models with and without spatial autocorrelation. The p-value is the probability that rejecting the null hypothesis that the two models are equally likely. For each year, p-values are low, thus we confirm the impression from the map of the residuals that they are spatially auto- correlated.

From the results, we can also see that λ increases through years. This means the spatial dependence increasingly affects the results as time passed by.

Household Type

Number of married couple family population was significant and slightly negatively related to house value in 2011, 2012. Married couple family variable is more likely to be stable within a region and negatively linked to house value spillovers. Meanwhile, this variable was not largely contributed to the explanation of house values due to its small coefficient value.

Education Attainment (For population 25 years and over)

Population of Bachelor's degree was observed in 2011 and 2012 model. It is slightly positively associated with house value. However, education attainment structure has changed during recent years. Bachelor's degree is now less important to house value spillovers.

Unemployment Rate (For civilian population in labor force 16 years and over)

In the 2011 model, the unemployment rate was positively associated with house value, and it was negative during 2014-2015. In 2013, the unemployment rate time lag variable was slightly positive related to house value but indicated a negative relationship with house value in 2014 and 2015. It shows evidence that people are more confident in buying higher value houses if a CBSA and its neighbor CBSAs had lower unemployment rates last year.

Industry by Occupation (For employed civilian population 16 years and over)

Proportion of manufacture industry population was observed in 2011-2015 models. It was negatively correlated with the house value. In 2015 model, proportion of education and health population was positively related to house value. It reflects that the increase housing affordability of service industry than traditional agricultural and manufacture industry. It is evident that industry is strongly related to house value spillovers.

Median Household Income (In 2010 inflation adjusted dollars)

As expected the median household income variable and its time lag variable were strongly correlated with the house value for each year. The median household income variable was slightly positively related to house value for each year. This is also true of time lag variables. This suggests that a high income is linked to more savings to obtaining a high-value house spillover.

Accessibility to transportation

The accessibility to Amtrak station was strongly correlated with the house value for each year. Having more Amtrak stations is linked to accessibility to transportation and obtain a higher value house. The accessibility to airport contributes to two models (2014-2015) and is negatively related to house values. It is potentially due to the noisy surroundings around an airport or more expensive air transportation costs is negatively linked with house value spillovers.

CHAPTER 6

SENSITIVITY ANALYSIS

Sensitivity analysis can perform selective context-sensitive analysis that applies the context-sensitivity when and where doing so is likely to improve the precision that matters for the analysis ultimate goal. (Hakjoo oh, 2015) In this chapter, I used sensitivity analysis to examine the models in order to see whether a sensitive change in context will largely affect the results.

PCA factors reselection

In Chapter 3, I applied PCA to select demographic variables from PC1-7. Here I want to see whether considering fewer PCs will affect the results. In this chapter, I only chose factors for each of PC1-4., which already explain over 80% of the total variance.

Models rerun

Table 5 and Table 6 shows the new OLS and SAR model results based on the new PCA models.

Table 6

New OLS model results

	2011		2012		2013		2014		2015	
(Intercept)		10.27000 (***)		10.51000 (***)		10.45000 (***)		10.36000 (***)		10.67000 (***)
Factor1	amtrksta	0.02423 (***)	amtrksta	0.02582 (***)	amtrksta	0.02988 (***)	amtrksta	0.03380 (***)	amtrksta	0.03933 (***)
Factor2	airport	-0.00144	airport	-0.00117	airport	-0.00156 (.)	airport	-0.00210 (**)	airport	-0.00209 (*)
Factor3	university	-0.00075	university	-0.00089	university	-0.00079	university	-0.00060	university	0.00002
Factor4	agea11	0.00000	agea12	-0.00000	agea13	0.00000	agea14	0.00000	agea15	0.00000
Factor5	married11	-0.00000 (.)	married12	-0.00000 (.)	samehouse13	0.00000	married14	-0.00000	married15	-0.00000
Factor6	bachelor11	0.00000 (*)	bachelor12	0.00000 (*)	married13	0.00000 (.)	unemplyr14	0.06279	bachelor15	0.00000
Factor7	unemplyr11	2.05000 (**)	unemplyr12	-0.4867	bachelor13	0.00000	unemplyr13	1.09200	unemplyr15	-1.26500
Factor8	unemplyr10	-1.0280	unemplyr11	1.93700 (**)	unemplyr13	-0.76570	agri14	0.05172	unemplyr14	1.68500 (*)
Factor9	agri11	-0.0372	agri12	-0.18330	unemplyr12	1.93000 (**)	retail14	0.6868	agri15	-0.40490
Factor10	manu11	-1.2000 (***)	manu12	-1.35600 (***)	agri13	-0.24730	manu14	-1.1800 (***)	manu15	-1.39300 (***)
Factor11	retail11	0.79000	finance12	-0.6630	manu13	-1.1520 (***)	inc14	0.00001 (**)	inc15	0.00002 (***)
Factor12	eduhea11	-0.2202	eduhea12	-0.14160	finance13	-0.08490	inc13	0.00002 (***)	inc14	0.00001 (*)
Factor13	inc11	0.00001 (**)	inc12	0.00002 (***)	inc13	0.00002 (***)				
Factor14	inc10	0.00002 (***)	inc11	0.00001 (**)	inc12	0.00001 (*)				
Factor15	samehouse11	0.00000	samehouse12	0.00000						
Residual standard error		0.2022		0.1949		0.1964		0.2039		0.2117
R^2		0.6937		0.7030		0.7098		0.703		0.7005
Adjust R^2		0.6821		0.6924		0.7		0.6946		0.6904

F-statistic	60.23	66.43	72.17	83.26	69
Sample size	415	437	428	435	428

Table 7
New SAR model results

		2011		2012		2013		2014		2015
(Intercept)		10.56200		10.66800		10.69200		10.8350		10.89400
		(< 2.2e-16)		(< 2.2e-16)		(< 2.2e-16)		(< 2.2e-16)		(< 2.2e-16)
Factor1	Amtrksta	0.01440	amtrksta	0.017713	amtrksta	0.01674	amtrksta	0.012786	amtrksta	0.01383
		(0.005679)		(0.0002977)		(0.0003556)		(0.005965)		(0.0006582)
Factor2	married11	-0.00568	married12	-0.00000	airport	-0.00159	airport	-0.00163	airport	-0.00110
		(0.002251)		(0.0010629)		(0.0028854)		(0.001718)		(0.0091386)
Factor3	bachelor11	0.00000	bachelor12	0.00000	married13	0.00000	married14	0.00000	unemplyr14	-0.94096
		(0.003190)		(0.0019957)		(0.3947723)		(0.090887)		(0.0243979)
Factor4	unemplyr11	1.02230	unemplyr11	0.57918	unemplyr12	0.02617	unemplyr13	-0.74724	manu15	-1.13070
		(0.009679)		(0.1095)		(0.9415616)		(0.054271)		(0)
Factor5	manu11	-1.17200	manu12	-1.1028	manu13	-1.06440	manu14	-1.1299	inc15	0.00002
		(0)		(0)		(0)		(0)		(0)
Factor6	inc11	0.00002	inc12	0.00002	inc13	0.00002	inc14	0.00001	inc14	0.00001
		(0.006478)		(0)		(0)		(0.00002)		(0.0098026)
Factor7	inc10	0.00001	inc11	0.00001	inc12	0.00001	inc13	0.00001		
		(0)		(0.0129696)		(0.1349071)		(0.00000)		
λ		0.42541		0.43169		0.48507		0.54722		0.59649
LR test value		57.377		63.587		82.062		104.04		122.0902
p-value		0		1.5543e-15		<2.22e-16		<2.22e-16		<2.22e-16
Sample size		419		437		428		435		428

Results analysis

According to the model results, model variables are stable relative to the model variables discussed in the previous chapters. One interesting thing is that the proportion of population employed by manufacture variable was highlighted in the spatial autoregressive model for each year. This suggests that the population employed by manufacture is an important variable in this study and it is strongly correlated with house values spillover effects. The population employed by manufacture industry is more likely to be located in lower house value areas. The model results showed that PCA method is accurate and stable within selected range.

CHAPTER 7

DISCUSSION

Implications

In this study, the analysis of house value is expanded to include both nonspatial and spatial factors, including demographic variables, time lag variables (representing income and unemployment rate), accessibility to amenities (representing accessibility to transportation, accessibility to universities) and spatial spillover effect between 2011 and 2015. Approximately 68-70 percent of the variation in house value can be explained by these variables. There is strong evidence of spatial interaction of house value across CBSAs, say, λ is between 0.42 and 0.66 through years. There are both high-high autocorrelation and low-low autocorrelation within CBSAs.

I used the PCA method to select demographic variables in order to avoid the multi- collinearity problem. It was evident that household type, education attainment, unemployment rate and median household income for each year supported the earlier Reed (2016) model. However, age and race are not high loading factors. This may due to the focus of multi region study areas, whereas age and race are more likely to be micro affects within one region. Population density is not high loading factor potentially due to mix types of the dependent variable (house value).

I applied both of the OLS and SAR methods in this study, the differences between OLS results and SAR results are showed in Table 8.

Table 8
Differences of OLS and SAR models

	2011			2012			2013			2014			2015		
		SAR	OLS		SAR	OLS		SAR	OLS		SAR	OLS		SAR	OLS
(Intercept)		10.56200 (< 2.2e-16)[17]	10.27000 (***)		10.66800 (< 2.2e-16)	10.60000 (***)		10.73300 (< 2.2e-16)	10.59000 (***)		10.71000 (< 2.2e-16)	10.49000 (***)		10.76800 (< 2.2e-16)	11.13000 (***)
Factor1	Amtrksta	0.01440 (0.005679)	0.02423 (***)	amtrksta	0.01771 (0.0002977)	0.02581 (***)	amtrksta	0.01676 (0.0003611)	0.03014 (***)	amtrksta	0.017543 (0.00003)	0.03360 (***)	amtrksta	0.01081 (0.0048249)	0.03884 (***)
Factor2	married11	-0.00000 (0.002251)	-0.00000 (.)	married12	-0.00000 (0.0010629)	-0.00000 (.)	married13	-0.00000 (0.2750521)	0.00000 (.)	airport	-0.00125 (0.003584)	-0.00212 (.)	airport	-0.00051 (0.2120314)	-0.00190 (*)
Factor3	bachelor11	0.00000 (0.003190)	0.00000 (*)	bachelor12	0.00000 (0.0019957)	0.00000 (*)	unemplyr12	0.10742 (0.7651664)	1.80700 (**)	unemplyr13	-0.95127 (0.0184753)	0.95800	unemplyr15	-1.55830 (0.0094700)	-1.59700 (.)
Factor4	unemplyr11	1.02230 (0.009679)	2.05000 (**)	unemplyr11	0.57918 (0.1095099)	1.95900 (**)	manu13	-1.02340 (0)	-1.23700 (***)	manu14	-1.15590 (0)	-1.25700 (***)	unemplyr14	-0.34055 (0.5060525)	1.52400 (*)
Factor5	manu11	-1.17200 (0)	-1.2000 (***)	manu12	-1.1028 (0)	-1.40900 (***)	inc13	0.00002 (0)	0.00002 (***)	inc14	0.00001 (0.00006)	0.00001 (**)	manu15	-0.96322 (0)	-1.74900 (***)
Factor6	inc11	0.00002 (0.006478)	0.00001 (**)	inc12	0.00002 (0)	0.00002 (***)	inc12	0.00001 (0.2056521)	0.00001 (*)	inc13	0.00002 (0)	0.00002 (***)	finance15	-0.53937 (0.3098812)	-1.73800 (*)
Factor7	inc10	0.00001 (0)	0.00002 (***)	inc11	0.00001 (0.0129696)	0.00001 (**)							eduhea15	0.68222 (0.0007776)	-1.04200 (***)

[17] P-value were showed in the parentheses

Factor8					inc15	0.00002 (0.00000)	0.00002 (***)
Factor9					inc14	0.00001 (0.0008161)	0.00001 (*)
λ	0.42541	0.43169	0.50207	0.50418		0.66355	

This study has highlighted the demographic trends with reference to industry by occupation and population mobility (representing residence one year ago in the United States). This analysis has also introduced the accessibility to universities as an index of accessibility to amenities. It appears that proportion population employed in manufacturing is relatively important in the models and was negatively correlated with the house value. Possibly this is due to:

- Manufacture industries are more likely to have an agglomeration economic effect neighbors. These industries share labor, capital, resources, and technology. This behavior causes a spillover effect and thus this variable is important in the spatial house value model when considering house value spillover effect; and
- Manufacture industries occupation is relatively immobile and the population should have stable house demand. The potential for relatively lower income than employment in service occupations also leads them to afford homes in areas with lower housing values. There is some evidence to compare with education and health industry occupation in 2015. It shows that education and health industry occupation (mobile and high income) is strongly positively related to house value.

Moreover, the evidence suggests that population mobility (representing residence one year ago in the United States) and accessibility to universities are not important in my models. Though the high mobility of population in the United States, as well as the worldwide reputation for its cutting edged high standard universities infrastructures, the results shows the difference between investment and opportunism. People are more likely to care about long-term stable benefit of investment instead of short-term real-estate speculation. In addition, accessibility to universities is more important to house value in the inner city level, i.e., the nearer to the university center, the higher the house value is. It is hard to determine the general influence of this variable on the house value across regions.

Furthermore, I created time lag variables (representing median household income and unemployment rate) and a five-year timeframe was used in my study. Therefore, it is possible to identify the time trends:

- Household Type: Number of married couple family variable was slightly negatively related to house value from 2011 to 2013. The coefficient reflects that the increase in number of married couple families is predicted to have very slightly percentage decrease in house value.
- Education Attainment: Population of Bachelor degree was observed in 2011 and 2012 models. It was slightly positively related to house value. Over time, this has decreased in importance relative to higher house values.
- Unemployment Rate: This variable was observed in 2011 and 2015, and its time lag variable was observed in 2014. Unemployment rate was positively related to house value in 2011 whereas was negatively related to house value in 2015. This does not reflect the real relationship between unemployment rate and house value since the relationship may be hysteretic within the same year. The time lag variable was negatively linked to house value in 2014 and supported the argument that consumers will have more confidence in housing when the unemployment rate was lower. The variable does not contributed to an explanation of the variation in house values through other years.

- Industry by Occupation: Proportion of population employed by manufacturing was negatively associated with house value over time. Proportion of population employed by education and health was positively related with house value in 2015 whereas finance population was negatively related with house value in 2015 but not significant at 5% significance level. This reflects the increased wealth of the service industry population relative to higher house value than agriculture and manufacture industry.
- Median Household Income: Over time, there is strong evidence that this variable and its time lag variable are positively related to house value. This can partly reflect the relationship between wealth status and house value.
- Accessibility to transportation: Accessibility to Amtrak station is positively linked to higher house value and reflects residents' demands of transportation amenities with housing. One unit increases in Amtrack station is predicted to increase house value by 1%. Therefore, accessibility to Amtrack station is one important indicator of house values especially in mega-regions with high house values. Accessibility to airport was observed in 2014 to 2015 model. This variable was negative to house value potentially due to the high transportation costs and noise surroundings around the airport are decreasingly linked to higher house value.
- λ (representing how much spatial dependence affects the results): Over time, this has increased the importance relative to house value and reflects the increasing trend of spatial dependency from the house value spillover effect.

Conclusions

This study examined the relationships between demographic variables, time lag variables, accessibility to amenities, spatial spillover effect and house value. This analysis assists in developing a better understanding of how house values are affected by both spatial and non-spatial factors. There is strong evidence the household type, education attainment, unemployment rate and its time lag variable, industry by occupation, median household income, and its time lag variable, accessibility to transportation as well as spatial spillover effect have stable significant effects on house value during 2011 to 2015.

Important contributions were made by industry by occupation variable. From 2011 to 2015, proportion of population employed by manufacturing was negatively linked to house value, reflecting a lower proportion of residents in this occupation who live in a higher value housing. In 2015, proportion of population in education and health was strongly positively related to house value. The trend reflects the emergence of the importance of service industry and its increasing wealth and housing affordability relative to traditional agriculture and manufacturing industries over time. The importance of including this occupation variable in the spatial model also confirms that it may strengthen house value spillover effects over time. Policy makers should care about the tradeoff between developing special industries within each CBSAs, which could help them strengthen the links and cooperation among industries thus coordinate house values among different CBSAs. One possible implication pattern is to develop mixed special industry chains, such as “education + manufacturing”, “traveling + agriculture”, “rehabilitation + culture” in order to realize the linkage functions. Furthermore, entrepreneurs should pay much attention to the house values to find the best locations, i.e., lower house value manufacturing industries are more likely to adjacent to lower house value manufacturing industries, services industries are more likely to locate within the higher house value areas.

Moreover, this study showed the time lag effects (representing income time-lag variable and unemployment rate time-lag variable) have significant impacts on house values. Thus, it is possible to provide a new perspective to predict the house value by referencing the income and unemployment rate data in the previous years.

This research has also highlighted the accessibility to transportation variable. This result suggests that regions with better accessibility to transportation may have larger real estate investment potentiality. On the other hand, a good way to attract people and stimulate a region's vitality is to improve local transportation accessibilities.

It is acknowledged that house value is influenced by numerous economic, political and social factors (Reed, 2016). Most of the previous study focused on the micro level of house values, i.e., within a city. The perspective of this study focused on the macro level of house values, i.e., across cities. Thus, this study can provide contributions made by industries, transportations and other macro social influences on house values. The evidence of spatial clusters of house value across CBSAs revealed the allocation pattern of house value. The house values were extremely high in California. It has been proved that the high house values show the investors' preference for locations and potentiality of an area. Figure 31 shows the top 10 global cities for real estate investors.

Figure 31 The top 10 global cities for real estate investors[18]

The index includes Los Angeles, Boston, Chicago, New York, and Houston within CBSAs in the U.S. Most of them have the highest house values which as same

[18] The latest release of the index shows that Los Angeles (California) takes the top spot. Source from http://www.schroders.com/en/schrodersglobalcities/blog/blog/new-release-best-global-cities/

as this study showed. Therefore, this study provides an inventory of CBSAs for real estate investors.

The research showed little change across the years (2011-2015). Further research is required to monitor changes over a longer time span. It is also suggested that the models are applied to other countries to determine to what extent these trends are unique to the U.S. Moreover, this study has excluded the policy factors that would affect consumers' decision and housing supply such as land use and building regulations, school performance, and taxes differences.

REFERENCES

Abelson, P., Joyeux, R., & Mahuteau, S. (September 01, 2013). Modelling House Prices across Sydney. Australian Economic Review, 46, 3, 269-285.

Albouy, D. Y., Stuart, B., & National Bureau of Economic Research. (2014). Urban population and amenities. Cambridge, Mass: National Bureau of Economic Research.

Alonso, W. (1964). Location and land use: Toward a general theory of land rent. Cambridge, Mass: Harvard University Press.

Anselin, L. (April 01, 1995). Local Indicators of Spatial Association—LISA. Geographical Analysis, 27, 2, 93-115.

Anselin, L. (2003). GeoDa 0.9 user's guide, Spatial Analysis Laboratory. University of Illinois, Urbana-Champaign.

Barrios,V.E., Colom,C. and Moles,C.(2013), Lifecycle and housing decisions: a comparison by age cohorts. Applied Economics, Vol. 45, Iss. 32.

Bivand, R., Pebesma, E. J., & Gómez-Rubio, V. (2013). Applied spatial data analysis with R. New York, NY: Springer.

Boustan, L. P., & National Bureau of Economic Research. (2013). Racial residential segregation in American cities. Cambridge, Mass: National Bureau of Economic Research.

Brasington, D. M., Hite, D., & Jauregui, A. (June 01, 2015). HOUSE PRICE IMPACTS OF RACIAL, INCOME, EDUCATION, AND AGE NEIGHBORHOOD SEGREGATION. Journal of Regional Science, 55, 3, 442-467.

Bunten, D., Boustan, L. P., & Hearey, O. (2013). Urbanization in the United States, 1800-2000.National Bureau of Economic Research. doi:10.3386/w19041

Costa, D. L., Kahn, M. E., & National Bureau of Economic Research. (1999). Power couples: Changes in the locational choice of the college educated, 1940-1990. Cambridge, MA: National Bureau of Economic Research.

Coulson, N. E., Liu, C. H., & Villupuram, S. V. (November 01, 2013). Urban economic base as a catalyst for movements in real estate prices. Regional Science and Urban Economics, 43, 6, 1023-1040.

Cronon, W. (1992). Nature's metropolis: Chicago and the Great West. New York: W.W. Norton.

Cucchiara, M. B. (2013). Marketing schools, marketing cities: Who wins and who loses when schools become urban amenities.

Glaeser, E. L. (1998). Are cities dying? The Journal of Economic Perspectives, 12(2), 139-160. doi:10.1257/jep.12.2.139

Glaeser, E. L., Gyourko, J. E., & National Bureau of Economic Research. (2006). Housing dynamics. Cambridge, Mass: National Bureau of Economic Research.

Glaeser, E. L., & John F. Kennedy School of Government. (2008). The economic approach to cities. Cambridge, Mass: John F. Kennedy School of Government, Harvard University.

Glaeser, E. L., Gottlieb, J. D., Tobio, K., & National Bureau of Economic Research. (2012). Housing booms and city centers. Cambridge, Mass: National Bureau of Economic Research.

Green, R. (2011-12-15). Housing Markets, Prices, and Policies. In The Oxford Handbook of Urban Economics and Planning. : Oxford University Press. Retrieved 2 Jul. 2017, from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195380620.001.0001/oxfordhb-9780195380620-e-19.

Hanna, B. G. (January 01, 2007). House values, incomes, and industrial pollution. Journal of Environmental Economics and Management, 54, 1, 100-112.

Hiebert, P., & Vansteenkiste, I. (2009). Do house price developments spill over across euro area countries?: Evidence from a Global VAR. Frankfurt am Main: European Central Bank.

Hui, E. C. M., & Liang, C. (January 01, 2016). Spatial spillover effect of urban landscape views on property price. Applied Geography, 72, 26-35.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Baltimore: Warwick & York.

In Nancy, B., In Kieran, D., In Gerrit‐Jan, K., Eric, A. H., & Kuzey, Y. (December 15, 2011). Urban Education: Location and Opportunity in the United States.

Jensen, R., Thursby, J., & Thursby, M. C. (2010). University-Industry Spillovers, Government Funding, and Industrial Consulting. Cambridge, Mass: National Bureau of Economic Research.

James, G. (2013). An introduction to statistical learning: With applications in R.

Li, H., Calder, C. A., & Cressie, N. (October 01, 2007). Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. Geographical Analysis, 39, 4, 357-375.

Li, Q. (September 01, 2014). Ethnic diversity and neighborhood house prices. Regional Science and Urban Economics, 48, 21-38.

Luttik, J. (2000). The value of trees, water and open space as reflected by house prices in the Netherlands. Landscape and Urban Planning, 48(3), 161e167.

Marshall, A., In Guillebaud, C. W., & Royal Economic Society (Great Britain).
(1961). Principles of economics. London: Macmillan for the Royal Economic Society.

Meen, G. P., & University of Reading. (1997). Regional house prices and the ripple effect: A new interpretation. Reading: University of Reading.

Miles, D. (November 01, 2012). Population Density, House Prices and Mortgage
Design. Scottish Journal of Political Economy, 59, 5, 444-466.

Mills, E. S. (1972). Urban economics. Glenview, Ill: Scott, Foresman.

Muth, R. F. (1969). Cities and housing: The spatial pattern of urban residential land use. (pp. xxii+355-xxii+355)

Oh, H., Lee, W., Yang, H., Heo, K., & Yi, K. (December 01, 2015). Selective X-sensitive analysis guided by impact pre-analysis. Acm Transactions on Programming Languages and Systems, 38, 2.

Pagliara, F., Preston, J., & Kim, J. H. (September 26, 2011). Trade-offs between Transportation, House Price and Amenities in Residential Location Choice in Oxfordshire, UK. International Journal of Urban Sciences, 10, 1, 1-18.

Pijnenburg, K. (February 01, 2017). The spatial dimension of US house prices. Urban Studies, 54, 2, 466-481.

Intra- & Inter-Industry Productivity Spillovers in OECD Manufacturing: A Spatial Econometric Perspective. [S.l.]: [s.n.].

Reed, R. (October 03, 2016). The relationship between house prices and demographic variables. International Journal of Housing Markets and Analysis, 9, 4, 520-537.

Richard, K. Green. (2014). Towards an urban housing policy. Marron Institute of Urban Management, New York University.

Roberts, R. K., Roberts, R. K., Roberts, R. K., & Kim, S. G. (January 01, 2012). Neighborhood spillover effects between rezoning and housing price. The Annals of Regional Science, 48, 1, 301-319.

Roback, J. (1983). Wages, rents, and the quality of life. New Haven, Conn: Yale University, Economic Growth Center.

Rosen, S. (1979). Wage-based indexes of urban, quality of life. Current issues i urban economics.

Schelling, T. C. (July 01, 1971). Dynamic models of segregation. The Journal of Mathematical Sociology, 1, 2, 143-186.

Thorsnes, P., Alexander, R., & Kidson, D. (February 01, 2015). Low-income housing in high-amenity areas: Long-run effects on residential development. Urban Studies, 52, 2, 261-278.

Tiebout, C. M. (1956). A Pure theory of local expenditures. Place of publication not identified: publisher not identified.