ANALYSIS OF THE HOUSING MARKET IN THE METROPOLITAN AREAS IN THE UNITED STATES

A Dissertation

by

YARUI LI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	David J. Leatham
Committee Members,	David Bessler
	Ximing Wu
	Qi Li
Head of Department,	C. Parr Rosson III

December 2013

Major Subject: Agricultural Economics

ABSTRACT

The housing market plays a significant role in shaping the economic and social well-being of U.S. households. It helps spur U.S. economic growth when house prices rise, and drags the economic growth when house prices drop. In this dissertation, an analysis is conducted to project the direction of the U.S. housing market and to discover how it interacts with economic fundamentals. New pieces of information are found, which are deemed to facilitate decision making for both policy makers and investors.

In the first part of the dissertation, the groupings of U.S. housing markets are studied using cluster and discriminant analysis. Three clusters are found, which are located in the central, the east coast, and the west coast of US. There are no price signals transmitted among these housing market clusters, nor within each cluster. Thus, the communication of information in the housing market is through the process of utility convergence of marginal residents, and no price convergence across regions is found.

Next, the impact of credit constraint on the house prices is examined with the stochastic components of the price series being considered. Both a simulation technique and a DAG approach are employed. The resulting causal pattern shows that credit constraints affect the house prices directly and positively. Moreover, credit constraints work as an intermediary, passing the influence of the house investor, household income, and user cost onto house prices, which suggests that the credit relaxation policy should be carried out with caution when house inventory and household income send inconsistent signals.

Last, the model selection for house price analysis is discussed from the perspective of large-scale models—dynamic factor (DFM) model and large-scale Bayesian VAR (LBVAR) model. The LBVAR models are found to have superior performance compared to the DFM model throughout the prediction period. Also, it is found that the combined forecasts do not necessarily outperform individual forecasts. Even though independent information from different individual models improves the forecast accuracy, the benefit gained from marginal information is offset by the larger error brought by such combination.

DEDICATION

To my beloved husband,

Zhengxin Zhang

To my dearest father and mother,

Ming Li & Junhua Huang

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Leatham, and my committee members, Dr. Bessler, Dr. Wu, and Dr. Li, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my mother and father for their encouragement and to my husband for his patience and love.

NOMENCLATURE

ADF	Augmented Dickey Fuller Test
AIC	Akaike Information Criterion
AR	Autoregressive Model
DAG	Directed Acyclic Graph
DFM	Dynamic Factor Model
DSGE	Dynamic Stochastic General Equilibrium Model
ECM	Error Correction Model
FAVAR	Factor-Augmented Vector Autoregressive Model
HQC	Hannan and Quinn's Φ Measure
LBVAR	Large-Scale Bayesian Vector Autoregressive Model
MVC	Multivariate Copulas Distribution
PCA	Principal Component Approach
PC	Principal Component
RMSE	Root Mean Square Error
RSQ	R-Square
SBC/SIC	Schwarz-Loss Measure
SPRSQ	Semi-Partial R-Squared
VAR	Vector Autoregressive Model
VARMA	Vector Autoregressive Moving Average Model

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1 INTRODUCTION

The housing market is of great importance to the economy. House construction and renovation boost the economy by increasing in aggregate expenditures, employment and the volume of house sales. They also stimulate the demand for related industries such as household durables. The oscillation of house prices affects the value of asset portfolio for most households for whom a house is the largest single asset. Moreover, price movements influence the profitability of financial institutions and the soundness of the financial system. Recent studies further justify the necessity of house price analysis, concluding that the housing sector plays a significant role in acting as a leading indicator of the real sector of the economy and that assets prices help forecast both inflation and output (Forni, Hallin, Lippi, and Reichlin, 2003; Stock and Watson, 2003; Das, Gupta, and Kabundi, 2009a; Kim, Leatham, and Bessler, 2007). Thus, a comprehensive and systematic analysis for the housing market can provide valuable information to policy makers and help them better control inflation and design more effective policies. Also, these analyses can guide individual market participant to make wise investment decisions.

This dissertation examines the U.S. housing market from three perspectives: the patterns of price movement, the impacts of credit constraint on house price, and the large-scale model selection for house price analysis.

The first essay studies the clustering of U.S. housing markets and the patterns of price movement between and within those clusters. Cluster analysis is used to classify

housing markets into three clusters based on economic fundamentals and housing attributes. Discriminant analysis validates the clustering results and suggests that all the economic and amenity variables contribute to grouping homogenous markets and separating distinct ones. Time series econometric models are used to estimate the interaction of house prices. Both between- and within-cluster analysis are conducted. The error terms derived from these models are further analyzed by a directed acyclic graph (DAG) approach to examine the patterns of price movement. For both between- and within-cluster models, there exist no statistically significant causal flows of innovation among the examined metropolitan areas. The shock in one area due to local factors is not going to cause fluctuation in house price in other areas. Thus, house prices in different regions may move together and converge over time under the effect of macroeconomic fundamentals, but there is no cross-sectional communication of house price.

In the second essay, the interaction between credit constraints and house prices is studied based on inverted demand approach (Duca, Muellbauer and Murphy, 2011(b)). Under this approach, the house price is assumed to be a function of house supply, income, user costs and credit constraint. We model the dependence among the stochastic components of house price, credit constraints, user costs of owning a house and other variables using multivariate copulas distribution (MVC). Based on the simulated data, several quantile values are derived, which provide more information for political or investment decision than single point estimation does. The causation between house price and credit constraint is also examined using a DAG approach, and the resulting

causal pattern suggests that credit constraint not only directly affects house price, it also works as the intermediate passing the influence of other factors onto house price, which complicates the enactment of credit policy.

The third essay focuses on model selection for analyzing house price in US metropolitan areas from the perspective of large-scale models. This study lends support to the superior performance of the LBVAR model compared to DFM model throughout the prediction period. Also, our study suggests that combined forecasts do not necessarily outperform individual forecasts. Even though independent information from different individual models improves the forecast accuracy, the benefit gained from marginal information is offset by the larger error brought by such combination.

2 TRANSMISSION OF PRICE CHANGES BETWEEN AND WITHIN CLUSTERED U.S. HOUSING MARKETS

2.1 Background

In the last three decades, residential house prices in U.S. metropolitan areas exhibit considerable fluctuations over time and across regions. However, these fluctuations follow very different patterns. After examining the house prices of 40 metropolitan areas over the 1980-2004 periods, Himmelberg, Mayer and Sinai (2005) find three patterns exist for U.S. housing market: (1) house price peaked in the late 1980s, fell to a trough in the 1990s, and rebounded by 2004; (2) a "U" shape history--high in the early 1980s and high again by the end of the sample; (3) house prices have declined since 1980 and have not fully recovered. They divide the 40 metropolitan areas into three groups with each group following one of the three patterns. The interesting point is that those areas in the same group are not necessarily geographically adjacent, and the areas adjacent to each other are not always in the same group. For example, the house price of Fort Worth follows the third pattern, while the price of its neighbor--Dallas follows the first one. New Orleans, instead, shares the same house price pattern with Fort Worth.

Therefore, geographical proximity fails to warrant the homogeneity of housing markets. More factors need to be considered in the process of identifying homogeneous housing groups so as to enact suitable policy for each group, to diversify debt and equity

portfolio, as well as to hedge the housing market risk. There are many literatures provide support to the standpoint that economics dominates geography in terms of differentiating housing markets (Gyourko and Voith, 1992; Jud and Winkler, 2002; Chan, Ng and Ramchand, 2012). House prices are found to facilitate classifying homogeneous housing market and so are other elements such as unemployment rate, household income, dwelling size, housing unit quality and neighborhood quality.

The urban economics suggests that house demands and house prices across cities should adjust so that no household will wish to move and marginal residents of all locations receive identical utility (Rosen, 1979; Roback, 1982). Because theoretically it is utility that converges rather than incomes, house prices, or city amenities, there is little theoretical support for the idea that house prices should converge (Kim and Rous, 2012). However, while regional per capita incomes are converging, it is tempting to conjecture that this phenomenon may, in turn, be driving convergence in regional house prices. In addition, other factors like labor and capita mobility may also be contributing to regional house price convergence (Clark and Coggin, 2009). Thus, homogenous housing markets that share similar economic fundamentals and amenities may experience house price convergence among themselves. In order to determine the interrelationship of house prices across regions, it is important to understand the transmission of price signal within groups of homogeneous housing markets as well as between those groups.

The objective of this essay is to study the pattern and strength of price signals transmitted among homogeneous groups of housing markets, as well as within each group. Cluster analysis is conducted to classify twenty-nine U.S. metropolitan areas

(MSAs) into homogeneous groups based on variables capturing housing attributes and economic environment. Discriminant analysis is employed next to validate the grouping results from the cluster method. A directed acyclic graphs (DAG) approach is used last to identify the pattern of price movements across the grouped housing markets, and to infer housing market integration based on the resulting patterns from the graphs.

The contributions of this essay are two-fold. First, previous researches identify the clustering of housing prices in a limited manner (Lu, 2009). By applying cluster analysis and discriminant analysis, this essay investigates the grouping patterns of the U.S. housing market and analyzes whether housing attributes and economic factors contribute to the identification of homogeneous groups for housing markets. Second, this essay extends the understanding of the price movement and convergence between and within homogeneous groups of housing markets based on a DAG approach, which makes no *a priori* assumptions on the causal patterns of the movements. The information obtained from these analyses can be used in investment portfolio construction to reduce the unsystematic risk of the portfolio.

The rest of the essay is organized as follows. Section 2 provides a literature review. Section 3 introduces the cluster method and discriminant analysis techniques, as well as the causal modeling under the DAG approach. Section 4 discusses the data. Section 5 presents results. Section 6 concludes and discusses the limitations of this study.

2.2 Literature Review

This section reviews previous literature from three aspects. First, it reviews the studies examining the fluctuations of house prices across cities. The purpose of these studies were to find the impact of local and national circumstances on the volatility of house prices, and provide support to the hypothesis that patterns of house prices are driven by macroeconomic factors. The second part of this section discusses the application of cluster analysis in economic studies, especially in real estate areas. The third part reviews econometric techniques used to discover the patterns of price movement across regions, and compares traditional models with a DAG approach to justify the use of it in this essay.

The variations in the house prices across regions have been examined by a large body of literature. For example, Fik, Ling and Mulligan (2003) present an interactive variables approach and test its ability to explain price variations in an urban residential housing market. They find that accessibility indices, distant gradients and locational dummies cannot fully account for the influence of absolute location on the market price of housing because there are an indeterminable number of externalities (local and nonlocal) influencing a given property at a given location. They suggest this approach be used when estimating the value of housing for geographic areas where very little is known *a prior* about the neighborhoods or submarkets. McGreal and De La Paz (2013) estimate the role of attributes in asking price formation for housing market. They use hedonic model and apply STAR methodology to avoid the bias generated by

autocorrelation and control for spatial dependence. Their results show that the pricing of attributes varies by geographical region and over time with property size and economic and demographical attributes being the key variables explaining asking price formation.

The paper by Capozza, Hendershott, Mack and Mayer (2002) explores the explanations for momentum and cyclical behavior of house prices. They find the variation in the cyclical behavior of real house prices across metropolitan areas is attributable to more than just variation in local economies. Also, they discover that real house prices react differently to economic shocks depending on such factors as the growth rates of the underlying population and real income in the area, the size of the area, and construction costs. Sutton (2002) employs a small VAR model to examine the extent to which house price fluctuations can be attributed to fluctuations in national incomes, interest rates and stock prices. The author finds that favorable economic developments captured by these variables appear to have played an important role in house price gains.

There is one major finding of the above papers and of many other papers not reviewed in detail here (Gyourko and Voith, 1992; Jud and Winkler, 2002; Abraham and Hendershott, 1996; Lu, 2009). That is, geographical factors are not sufficient to explain the fluctuation in house prices, and economic and demographical attributes help the explanation to a large extent. To summarize, the variables found to contribute to the pattern of house price fluctuation include but are not limited to employment growth, population growth, income growth, construction costs, interest rates, property size and stock price. Thus, in order to classify housing markets into homogeneous groups, these variables should be considered in the group identification process.

Cluster analysis is the most common method to classify data into a set of categories, and it has been applied in a wide variety of fields, such as engineering, computer sciences, life and medical sciences, astronomy and earth sciences, and social sciences (Xu and Wunsch, 2009). There are also a number of applications of cluster analysis in economic area. For example, the San Diego Association of Governments (2002) uses the cluster method to explore the representation of local industry drivers and regional dynamic. Chicago Metropolitan Agency for Planning (2009) applies cluster analysis to identify industries that are geographically concentrated or of a similar nature, and that make use of related buyers, suppliers, infrastructure and workforce. Cunningham and Maloney (2001), based on the results from cluster analysis, try to find the heterogeneity among microenterprises and explain why small firms exist in Mexico. Gupta and Huefner (1972) use cluster analysis to find the correspondence between financial ratios and basic industrial attributes, and Yang and Hu (2008) examine regional disparity in China using cluster analysis.

The application of cluster method to house price analysis is limited but becomes popular in the recent decade. Abraham, Goetzmann and Wachter (1994) use the K-means clustering algorithm to explore the interrelationship of housing market returns using the returns to house price indices data in 30 metropolitan areas. Goetzmann and Wachter (1995b) apply cluster analysis to examine portfolio diversification for 21 metropolitan areas. Case, Clapp, Dubin and Rodrigues (2004) employ a hedonic model that includes homogeneous within-county distincts created on the basis of cluster analysis. Bourassa, Cantoni and Hoesli (2008) use districts defined by the local property

tax assessment office as well as a classification of census tracts generated by principal components and cluster analysis to analyze the impacts of alternative submarket definitions when predicting house prices. Lu (2009) applies cluster method to study how housing attributes impact house price across cities based on variables such as employment rate, household income and neighborhood quality. Shimizu and Watanable (2010) conduct a cluster analysis with Ward's method to observe spatial relationships between house price fluctuations for regions in US and Japan. Besides the studies on U.S. housing markets, cluster method is also applied to the examination of housing markets in many other countries. For example, Chan, Ng and Ramchand's study (2012) for Singapore; Leung, Chow and Han's study (2008) for Hong Kong; Apergis, Simo-Kengne and Gupta's study (2013) for South Africa; Kim and Park's study (2005) for Korea; Hensen and Vatansever's study (2012) for Turkey.

Previous studies employ a variety of methods to identify the patterns in house price fluctuation across regions. For example, Hiebert and Roma (2010) test for price convergence and analyze key factors explaining price differentials in a panel regression framework. Favara and Song (2013) use a user-cost model to study how dispersed information affects the equilibrium house price. Gyourko, Mayer and Sinai (2006) use a simple two-location model allowing for differences in the elasticities of supply across locations to show how inelastic land supply can link the stylized patterns in house price. Capozza, Hendershott, Mack and Mayer (2002) explore the dynamics of real house prices by estimating serial correlation and mean reversion coefficients from a panel data set of 62 metropolitan areas. To examine long-run house price convergence across US

states, Holmes and Otero's modeling strategy (2011) employs a probabilistic test statistic for convergence based on the percentage of unit root rejections among all state house price differentials. Hirata, Kose, Otrok and Terrones (2013) evaluate the roles played by a variety of global shocks, including shocks to interest rates, monetary policy, productivity, credit, and uncertainty, in explaining house price fluctuations using a wide range of factor-augmented vector autoregressive models.

The methods employed by the above papers have two points in common. First, they examine the dispersion or convergence of house price across regions based on the interaction between house price and other variables, such as construction cost, land supply and policies. However, not all the variables impacting house prices are included in their models, so only the part of house price movement related to the examined variables is explained. Modeling price variables across regions and overtime directly may provide more information regarding price discovery of housing market, and such model is free of the concerns about incomplete set of variables. Second, the relationships between house price and other variables are estimated with econometric models, and then tests are conducted to verify the significance of the coefficients and the *a priori* assumed patterns. While such *a priori* assumption models about price movement may serve as a reasonable starting point for analysis, they by no means govern the way that observational data must interact in reality. This is all just to say, simply, that one should be cognizant of the fact that the conclusions which flow from such models are not independent of the *a priori* assumptions inherent in their construction. Insofar that this is the case, the results from this framework can be misleading if this fact is forgotten. Thus,

we employ a DAG approach in this study to overcome such problems inherent in the *a priori* assumption approach in order to estimate the transmission of house price signal across regions.

To sum up, in the examination of the patterns of price movement across regions, geographical factors are not sufficient to explain the flows of price signals, and economic and demographical attributes help the explanation to a large extent. Based on those economic and demographical variables, cluster analysis can efficiently divide housing markets into homogeneous groups. When sorting out the causal flow of price signal across groups of housing markets and within each group of markets, modeling with price variables directly may provide more information and is free of the concerns about incomplete set of variables. Also, in the process of search for patterns, a DAG approach shows innovation over traditional modeling techniques by making no *a priori* assumptions on the price movement pattern and let the data speaks for itself, which is deemed to provide information from a new perspective.

2.3 Methodologies

2.3.1 Cluster Analysis

One of the most important of the myriad of data analysis activities is to classify or group data into a set of categories or clusters (Xu and Wunsch, 2009). A cluster should be described in terms of internal homogeneity and external separation. In other

words, data objects in the same cluster should be similar to each other, while data objects in different clusters should be dissimilar from one another (Gordon, 1999; Hansen and Jaumard, 1997; Jain and Dubes, 1988). Both the similarity and the dissimilarity should be elucidated in a clear and meaningful way.

According to Xu and Wunsch (2009), four basic steps should be followed when carrying out cluster analysis. The first step is feature selection or extraction. In this step, distinguishing features from a set of candidates should be chosen. Generally, ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, and easy to obtain and interpret. In this essay, we select the housing attributes and economic factors which are proved by previous studies to be important in explaining the fluctuation of house price across regions. The second step is clustering algorithm design or selection. This step consists of determining an appropriate proximity measure and constructing a criterion function. Here, Ward's method is used to assess the similarity between clusters. The object of Ward's method is to minimize the increase of the within-class sum of the squared errors,

(1) $$E = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \mathbf{m}_k \right\|^2,$$

caused by the merge of two clusters. In this expression, K is the number of clusters and $\mathbf{m}_k$ is the centroid of cluster C_k defined as $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$, where n_i is the number of data points belonging to the cluster. So, the distance between cluster C_i and C_j can be represented as

(2) $$D(C_i, C_j) = \Delta E_{ij} = \frac{n_i n_j}{n_i + n_j} \left\| \mathbf{m}_i - \mathbf{m}_j \right\|^2 .$$

The distance between a cluster C_l and a new cluster $C_{(ij)}$ formed by the merge of C_i and C_j is then written as

(3) $$D(C_l, (C_i, C_j)) = \frac{n_i + n_l}{n_i + n_j + n_l} D(C_l, C_i) + \frac{n_j + n_l}{n_i + n_j + n_l} D(C_l, C_j) - \frac{n_l}{(n_i + n_j)^2} D(C_i, C_j)$$

The third step is cluster validation. In this essay, discriminant analysis is employed to test the robustness of cluster analysis (Yu, 2009; Hoesli, Lizieri, and Macgregor, 1997). The last step is to interpret the results so as to gain meaningful insights from the original data.

There are two types of clustering, known as partitional and hierarchical clustering. In this essay, agglomerative hierarchical clustering is employed. Agglomerative clustering starts with N clusters, each of which includes exactly one data point. A series of merge operations is then followed that eventually forces all objects into the same group. There are four steps involved in this clustering method. First, one starts with N singleton clusters and calculates the proximity matrix for the N clusters. Second, in the proximity matrix, one searches the minimal distance $D(C_i, C_j) = \min_{l \le m, l \le n, m \ne n} D(C_m, C_l)$, where D(.,.) is the distance function, and combine cluster C_i and C_j to form a new cluster C_{ij}. Third, one updates the proximity matrix by computing the distances between the cluster C_{ij} and the other clusters. Fourth, one repeats steps 2 and 3 until only one cluster remains (Xu and Wunsch, 2009).

2.3.2 Directed Acyclic Graph (DAG) Approach

Empirical studies in economics have primarily relied on economic theory or researchers' intuitions in order to identify the structure and parameters of economic models (Kwon and Bessler, 2011). However, theory is oftentimes too heterogeneous to provide a conclusive causal structure or does not provide sufficient information to identify the underlying causal structure. Moreover, such *a priori* models fail to define the way observational data *must* interact and may provide incorrect causal inference. Distinguished from "Deductive Causation", which arises from either innate ideas or from mathematics on assumed behavior, "Inductive Causation" relies on observational data and infers a causal graph from conditional independencies among variables. As a basis for inductive causal inference in econometrics, the DAG method has been applied to many research topics, e.g., environmental and economic sustainability (Bessler, 2005), market integration and price discovery (Bizimana, Angerer and Bessler, 2012), price dynamics in agricultural markets (Bessler, Yang and Wongcharupan, 2003; Bessler and Akleman, 1998), and interest rate transmission (Oxley, Reale and Wilson, 2009) among others.

A directed graph uses arrows and vertices to illustrate the causal relationships among variables, whose values are measured in non-time sequence. Vertices connected by an edge are said to be adjacent. A directed edge is an edge which has an arrow indicating its causal direction, while undirected edge does not have a causal direction. If

we have a set of vertices {A, B, C, D}, the undirected graph contains only undirected edges, for example A→B. A directed graph contains only directed edges, for example C→D. An acyclic graph is one for which there is no path from any given variable which leads back to that variable. For example, the path A→B→C→A is labeled as "cyclic" because we move from A to B, but then return to A by way of C. A directed acyclic graph is a directed graph that contains no directed cyclic paths. Because cyclic graphs are not identifiable, only acyclic graphs are discussed in this essay. The terms from genealogy are used when referring to variables in causal model. For example, in the path A→B→C, the variables A and B are ancestors of variable C. Variable C is the descendent of variables A and B. Variable A is the grandparent of variable C and parent of variable B.

There are several algorithms discussed in the machine learning literature that can be used to identify DAGs. This study employs the PC algorithm (Bessler, 2003). Three conditions should be satisfied to apply the PC algorithm. First, the causal Markov condition, which states that given its parents, a variable should be conditionally independent of its non-descendants. The second condition requires that no variable is omitted which causes two or more other variables selected for analysis. The last condition requires that a zero correlation between variables should not be the results of cancellations of deeper parameters connecting these variables.

The PC algorithm determines the causal pattern among a set of variables in three steps. First, starting with a completely undirected graph, each variable in the set is connected to every other variable via an undirected edge. Next, edges between variables

are removed if the null hypothesis cannot be rejected that the correlation between any two variables is not significantly different from zero. Edges that remain are said to survive "zero order conditioning", and these edges are subjected to a series of first order conditioning tests with the null hypothesis that the conditional correlation between any two variables on a third variable is not significantly different from zero. Edges are removed if the null hypothesis cannot be rejected. The test of second and higher order conditioning then continues following the same rule. Last, an arrow (direction) is assigned to each of the surviving edges according to the directional separation (d-separation) definition, which is given in Pearl (2000):

> Definition: *X, Y, and Z are three disjoint sets of variables. A path p is said to be d-separated by a set of nodes Z if and only if (1) p contains a chain* $i \rightarrow m \rightarrow j$ *or a fork* $i \leftarrow m \rightarrow j$ *such that the middle node m is in Z, or (2) p contains an inverted fork (or collider)* $i \rightarrow m \leftarrow j$ *such that the middle node m is not in Z and such that no descendant of m is in Z. A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.*

The reasoning of sorting out causal patterns by d-separation can be illustrated by a simplified example. There are four variables {A, B, C, D}, and corr (A, D) =0 and corr (A, C) $\neq 0$. Assume we find that corr (A, D| B) $\neq 0$ and corr(A, C| B)=0, which means variables A and D are d-connected while variables A and C are d-separated. According to the d-separation definition, there exists three possible directed acyclic graphs for variables A and C, which are $A \rightarrow B \rightarrow C$, $A \leftarrow B \leftarrow C$, and $A \leftarrow B \rightarrow C$. Using only this information we cannot determine which graph presents the true causal pattern

between variables A and C, however, when coupled with the unique directed graph for variable A and D ($A \rightarrow B \leftarrow D$), a complete directed graph can be drawn for these four variables as the one shown in Figure 1.

When analyzing real world problems, a large number of variables are tested and the causal patterns are much more complicated. TETRAD IV, a software program developed at Carnegie Mellon University, is employed for the estimation in this essay.

While DAG has gradually demonstrated its usefulness to address such identification issues (Kwon and Bessler, 2011), there are some limitations of the method and the PC algorithm as well. First, DAG may give misleading results when one attempts to infer causal relations among variables where one or more of the variables has an infinite variance (Bessler, 2005). Second, variables used in a DAG model need to follow a multivariate normal distribution for the model to be fully efficient. Third, the PC algorithm result depends on the significance level chosen by the researcher in determining edges. Namely, in order for the algorithm to converge to all correct decisions with probability of 1, the significance level used in making the decisions should decrease as the sample size increases. Thus, the use of higher significance levels may improve performance in small sample sizes (Spires, Glymour and Scheines, 2000).

2.4 Data

The data used in this analysis are obtained from American Housing Survey (AHS) for metropolitan statistical areas (MSAs) in 2011, the latest survey available. The AHS is

sponsored by the Department of Housing and Urban Development (HUD) and conducted by the U.S. Census Bureau. It is the most comprehensive national housing survey in the United States, and provides current information on a wide range of housing subjects, including size and composition of the nation's housing inventory, vacancies, fuel usage, physical condition of housing units, characteristics of occupants, equipment breakdowns, home improvements, mortgages and other housing costs, persons eligible for and beneficiaries of assisted housing, home values, and characteristics of recent movers (AHS, 2011).

There are 29 metropolitan areas reported in the 2011 AHS, and thus used in this analysis. Table 1 lists all these metropolitan areas. The selection of variables is based on previous literature which has shown significant interaction between the included variables and fluctuation in house price. All variables used for cluster analysis are reported in Table 2, and they are housing value, unemployment rate, tax payment, mortgage rate, household income, unit size, rooms, crowding, neighborhood quality rating and unit quality rating. Data for all these variables are available in 2011 AHS except for unemployment rate, which is obtained from Bureau of Labor Statistics. Since cluster analysis is sensitive to the scales of variables, data used in the analysis is standardized with a mean of zero and a standard deviation of one.

2.5 Results

2.5.1 Cluster Analysis

Table 3 presents the cluster history of Ward's Minimum Variance Cluster Analysis. The first column of the table lists the number of clusters, and the second column lists the variables or clusters that merge into a new cluster. The *Frequency* column gives the number of elements in the cluster. Semi-partial R-squared (SPRSQ) shows the decrease in the proportion of variance accounted for resulting from joining the two clusters, and equals the between-cluster sum of squares divided by the corrected total sum of squares. SPRSQ is a measure of the homogeneity of merged clusters, so SPRSQ is the loss of homogeneity due to combining two clusters to form a new cluster. Thus, the SPRSQ value should be small to imply that we are merging two homogeneous groups. R-square (RSQ) is the proportion of variance accounted for by clusters. It measures the extent to which clusters are different from each other (so when there is only one cluster, the RSQ value is zero). This RSQ value should be high.

The hierarchical clustering analysis starts with 29 clusters, with each metropolitan area consisting of a cluster. The cluster history shows that Kansas City and St. Louis are the two metropolitan areas closest to each other based on the value of distance function. So these two cities are combined together to form a new cluster, with totally 28 clusters left. The *Frequency* is two because the newly formed cluster has two elements: Kansas City and St. Louis. The SPRSQ is 0.0022, which means the proportion

of variance decreases by 0.22% by joining Kansas City and St. Louis together as a new cluster (CLS 28). Cluster combination continues in the same way until all the cities are in the same cluster. Taking cluster 24 as another example, compared to the value of distances function between Virginia Beach and other MSA or other cluster, the value of distance function between Virginia Beach and cluster 28 is the smallest. This suggests combining Virginia Beach with Kansas City and St. Louis together, the two MSAs in cluster 28, to form a new cluster (CLS 24), with 24 clusters left. By doing this, the proportion of variance decreases by 0.52%.

The number of clusters best summarizing the similarity and dissimilarity of data is determined based on the decrease of SPRSQ, since it is a measure of homogeneity of merged clusters. Figure 2 presents a plot of number of clusters versus SPRSQ. There is not a defined cut-off point suggested by any literature. But, from the plot, we can see that SPRSQ drops fast from one cluster to three clusters, and the curve tends to be flat after three clusters. Thus, three clusters perform best in grouping homogeneous cities together while separating dissimilar ones into different groups.

Figure 3 presents the results of cluster analysis in a tree diagram (dendrogram). The between-cluster sum of squares is plotted on the y-axis. The larger this value is, the more distinct the two MSAs are. For example, Kansas City and St. Louis are most similar to each other compared to other possible combination and have the smallest between-cluster sum of square. Thus, they are grouped first (CLS 28). Then, they are grouped further with Virginia Beach to form a cluster of three MSAs. This dendrogram presents the same information as in the cluster history table, but in a more visual-

convenient way. By dividing MSAs into three groups, we obtain the following group identification. The first cluster contains 19 MSAs: Birmingham, Virginia Beach, Kansas City, St. Louis, Cincinnati, Columbus, Phoenix, Indianapolis, Portland, Buffalo, Pittsburgh, Cleveland, New Orleans, Milwaukee, Atlanta, Denver, Dallas, Fort Worth and Memphis. The second cluster consists of four MSAs: Providence, Sacramento, Riverside and Charlotte. The third cluster is comprised of six MSAs: Oakland, San Diego, Los Angeles, San Jose, San Francisco and Anaheim.

The cluster identification of the 29 MSAs is shown on a U.S. map in Figure 4. The MSAs in the first cluster are marked with red dots. The MSAs in the second cluster are marked with blue dots, and the ones in the third cluster are marked with green dots. The remaining dark grey dots represent the MSAs not in the analysis due to data unavailability. From the figure, several interesting findings can be obtained. First, the MSAs in the first cluster are located in the central area of US. Compared to the MSAs in the other two clusters, these MSAs have the lowest values in household income, unemployment rate, tax payment and housing value, and they have the highest level of interest rate. From the perspective of housing attributes, these MSAs have the largest square footage per unit and the largest number of rooms, and thus they have the lowest level of crowding. However, these MSAs have the worst overall opinions of neighborhood (lowest rating of neighborhood compared to other clusters). These characteristics of the first cluster are consistent with the economic conditions and geographical traits of these MSAs. For example, the majority of heavy manufacturing industries and old-style farming are located in the central US, and these sectors hire less-

educated labor force and provide low income. Also, the pressure on land use is small in the central US, and thus houses are generally larger and cheaper.

Second, the MSAs in the second cluster are close to coasts of US, mostly east coast. They have the highest unemployment rate, the smallest square footage of unit and the best overall opinion of both housing structure and neighborhood. For the rest of the economic and housing attributes, this cluster is between the first and the third clusters. The MSAs in this cluster are more developed than those in the first cluster, and people in these areas have higher level of income and require higher living quality.

Third, the MSAs in the third cluster are located along west coast of U.S. and all in the state of California. This state has very strong performance in financial service, trade, transportation education and manufacturing. With its advantage in high-tech industries, high-educated labor force and convenient transportation, California attracts a large amount of capital and a large number of companies to its markets, which results in keen competition of land use. Thus, the MSAs in this cluster have the highest value in household income, tax payment and housing value. Also expected is the highest degree of crowdedness of their houses.

Fourth, the clustering pattern of the examined MSAs supports that not only economic factors but also geographical factors matter in the formation of homogeneity in U.S. housing market. Moreover, housing attributes are shaped by local geographical and economic factors. Thus, our finding supports a complementary relationship between economy and geography in terms of differentiating housing markets, instead of an economy dominating geography relationship suggested by some of previous studies.

2.5.2 Discriminant Analysis

We employ discriminant analysis to validate the results from hierarchical cluster analysis. The basic purpose of discriminant analysis is to estimate the relationship between a single categorical dependent variable and a set of quantitative independent variables. This analysis is widely used to identify the group to which an object belongs. Its difference from cluster analysis is that the number of clusters is known in discriminant analysis while the number of clusters is unknown in cluster analysis. Since the number of clusters is determined by cluster analysis as three, we try to assign the 29 MSAs into three groups and check whether the membership of each group is the same as indicated by cluster analysis.

Based on Fisher's linear discriminant analysis, we want to derive the linear combinations of the economic factors and housing attributes that will discriminate best between defined groups. Each of the linear combination is known as a discriminant function, which takes the following form:

(4) $$y_i = w_i^T x$$

y_i is a 1×29 row vector of discriminant scores for the i^{th} linear combination, one score for each MSA. w_i is a 10×1 column vector of the discriminant weight for the i^{th} linear combination. x is a 10×29 matrix since we have 10 economic and amenity variables and 29 observations, one for each MSA. For the three-cluster problem we are facing, we

need to seek two linear combinations that maximize the separability of the discriminant scores $y_i's$ (i=1, 2).

Table 4 reports the discriminant weights for the two linear combinations which best separate clusters. Wilks' lambda (P-value<0.0001), Hotelling-Lawley trace (P-value<0.0001), and Pillai's trace (P-value<0.0001) statistics all suggest that the discriminatory power of the discriminant functions are statistically significant at 5% significance level. Error count estimate for clusters is zero, which means the grouping result from the discriminant analysis is the same as that from cluster analysis. Figure 5 presents the results of discriminant analysis graphically. A number (1, 2 or 3) denotes which cluster an observation belongs to in the hierarchical cluster analysis. We can see that the discriminant functions work well in separating the three clusters (no overlapping in the distribution) and the assignment of each MSA is the exactly the same as the assignment from cluster analysis. Thus, we conclude that the results from hierarchical cluster analysis are valid.

2.5.3 Error Correction Model and DAG

The causal flows of price signal are sorted out by DAGs based on the residuals from a vector autoregression (VAR) model. Thus, we estimate the correct form of VAR before conducting the DAG analysis.

There are three clusters found for U.S. housing markets, and we want to examine the transmission of price signal both among the clusters and within each cluster. Thus,

four VAR models are to be estimated. For the between-cluster estimation, three series of quarterly average housing values of MSAs are examined, one for each cluster. The data is from 1991:Q1 to 2013:Q1, for a total of 89 observations. For the within-cluster estimations, the data used are quarterly housing values of each MSA over the same period.

In the estimation of between-cluster model, let X_t denotes a vector of average quarterly housing values. First, we need to determine whether X_t is stationary based on Augmented Dickey-Fuller (ADF) test. The results of ADF test on level and first differences of housing values are reported in

From the results, we can conclude that all the three series of average housing values are *I*(1). Next, loss metrics on lag lengths from VARs on housing values are calculated and reported in Table 6. HQC has the lowest values with three lags and SBC has the lowest value with two lags. Since the lag selections indicated by different loss metrics are not consistent, we adopt the largest indicated lag length (*k*=3) in model specification.

Thus, we can model these three series in an error correction model (ECM) as following:

$$\Delta X_t = \mu + \Pi X_{t-1} + \sum_{i=1}^{2} \Gamma_i \Delta K_{t-i} + e_i \tag{5}$$

where Π and Γ_i are parameter matrices to be estimated, μ is a constant vector and e_i is a vector of white noises. If Π is of full rank, then X_t is stationary in levels and model (5) can be rewritten as a VAR in level model. If Π has zero rank, then model (5) can be

reduced to a VAR in first difference model. If the rank of Π is a positive number but Π is not of full rank, there exist matrices of adjustment coefficient α and matrices of long-run parameter β, such that $\Pi = \alpha\beta'$. Johansen cointegration test is conducted to determine the rank of Π, and its results are reported in Table 7, which indicates that the rank of Π is one. So, an ECM model is appropriate for between-cluster estimation, and there exists one long-run stationary relation in the three clusters. However, under ideal open market conditions, two long-run or cointegrating relationships would have been found (Engle and Granger, 1991). Thus, there are some types of constraints to information flow or market imperfections are preventing full adjustment to long-run equilibrium in these areas (Vitale and Bessler, 2006). The estimated ECM is not reported here. Table 8 reports the contemporaneous correlation matrix between the residual terms from the ECM model. From this table, we notice that all three clusters are positively correlated. The third cluster shows relatively high correlation with innovations from the second cluster (0.6010), but relatively low correlation with the first cluster (0.2702).

PC algorithm is applied to the correlation matrix presented in Table 8, and results are given in figure 5. We can see that there is not price information flowing among the three clusters. A Chi-square test is conducted on the null hypothesis of completely disconnected graph and *p*-value is reported to be 1.0000. Due to the data size, a significance level of 10% is used based on Spires, Glymour and Scheines' (2000) suggestion. Thus, we cannot reject the null hypothesis and conclude that the three clusters are significantly independent of each other in terms of price movement at 10% significance level. The result conforms to our expectation. Even though the housing

markets are integrated gradually under the influence of new transportation technology, developed infrastructure and more mobile resources, especially talent and brains, the house prices are still determined locally by economic and geographical conditions, such as type of major industry, employment and competition for land use. There is no transmission of price signals between regions with different combination of economic and geographical conditions.

While DAG indicates there exists no price signals across clusters, the decomposition of forecast error variance provides support to some price communication among these clusters across time. Table 9 reports the proportion of prediction error covariances by variable. These numbers partition the price uncertainty in each cluster at horizons of zero, one and twelve quarters ahead. Partition results can be provided at any horizon, but to save space we focus just on three periods. The *lead* column shows how many step-ahead the forecast is made for. For example, the uncertainty associated with current house price of cluster 1 is explained by surprises in the current period from its own cluster. No other cluster is responsible for current period innovations in the cluster 1. If we move ahead to one quarter, the uncertainty in the house price in cluster 1 is primarily influenced by its own one period innovation (98.98%) and there are trivial influences from innovations from cluster 2 (1.00%). Finally, at the long horizon of three years, uncertainty in house price in cluster 1 is explained by earlier innovation from cluster 2 (13.23%) and cluster 3 (10.34%), as well as its own previous surprises (76.43%). Overall, cluster 1 is the dominant cluster for price discovery in these three clusters. Innovation in Cluster 2 has greater influence on the uncertainty in house price

of cluster 3 as time go by, and dominates in accounting for price uncertainty in cluster 3 in the long run (77.81%). Thus, there is price signals transmitted over time among these clusters, but no contemporaneous price communication across clusters.

Next, we want to find out whether price movement exists between MSAs within homogeneous clusters. The analysis for cluster 2 and 3 follows the same procedure as the between-cluster analysis. However, cluster 1 contains 19 MSAs (variables), which cause over-parameterization problem based on only 89 observations. Thus, we adopt Bayesian VAR (BVAR) model to analyze cluster 1. The results of ADF tests, loss metrics and Johansen cointegration tests for cluster 2 and 3 are provided in Table 11 and Table 12 respectively. The *p*-values reported in Table 10 indicate that all the housing value series are $I(1)$ no matter which cluster they belong to. The loss metrics reported in Table 11 suggest that a lag length of four is most appropriate for the VAR model of cluster 2 and a lag length of five works best for the model of cluster 3. The results of Johansen cointegration tests reported in Table 12 show that the cointegration rank of cluster 2 is two and the cointegration rank of cluster 3 is four. This suggests that ECM model is appropriate for the estimation of both clusters. Also, it implies that constraints to information flow or market imperfections exist for MSAs in both clusters and prevent full adjustment to long-run equilibrium in these areas. Table 13 reports the contemporaneous correlation matrix for cluster 1 from BVAR model and Table 14 and Table 15 report the contemporaneous correlation matrixes for cluster 2 and 3 from ECM models, respectively.

The results from PC algorithm for the three clusters are given in figure 6 to 8. For these cases, Chi-square tests are conducted on the null hypothesis that the population covariance matrix over all of the measured variables is equal to the estimated covariance matrix over all of the measured variables written as a function of the free model parameters. Again, due to the data size, a significance level of 10% is used based on Spires, Glymour and Scheines' (2000) suggestion. For cluster 1, there exist some causal flows between the innovations of the 19 MSAs, and the patterns of causal flows divide the MSAs in cluster 1 into smaller groups. Several MSAs (New Orleans, Dallas and Cleveland) are not part of any innovation interaction. The *p*-value of the resulting causal patterns is zero, so we reject the null hypothesis and conclude that the resulting patterns are not reliable and fail to warrant the significance of these communications of innovation. There is not adequate information to draw conclusion about the economic significance of these patterns on price movement. But even if they are economic significant, the interaction of innovations for the MSAs are small scale and only between a small number of MSAs. The same situation exists for cluster 2 and 3. From figure 7 and 8, we can see that the communications of innovation are only between two or three MSAs, and the *p*-values of these two causal patterns suggest that the resulting causal patterns are not reliable and fail to warrant the significance of price movement between MSAs from statistical perspective. Thus, just like in the between-cluster analysis, house prices are determined locally by economic and geographical conditions, and there is no transmission of price signal between regions. This is consistent with urban economic theory, which advocates that trend in utility convergence carries information flows in

housing markets while house price convergence across regions is not happening due to such communication of market information.

Forecast error variance decompositions for cluster 2 and cluster 3 are reported in Table 16 and Table 17. Because cluster 1 has 19 variables, its decomposition table is too large to be reported in the essay, we only present here the major findings for cluster 1 along with the findings for the other two clusters. For both cluster 1 and 2, there exists no dominant MSA for price discovery in the short-run and long-run. The uncertainty associated with the house price of each MSA is explained primarily by surprises in its own region. For cluster 3, Los Angeles is the dominant MSA, and San Francisco is the second mover at the long run. Thus, we conclude that there is little price communication over time in among MSAs in cluster 1 and 2. However, there is price signal transmitted among MSAs in cluster 3. As the two largest cities in California, Los Angeles and San Francisco are the two dominant MSAs, whose innovations in house price will contribute to the uncertainty in house price in other MSAs (also in California) in the same cluster.

To sum up, even though the clustered MSAs share similar economic and amenity attributes, there is no statistically significant innovation communication among these areas, no matter between or within clusters. The common moving trends shared by these MSAs are the results of external economic fundamentals, such as income and employment. The price shock in one area due to local factors will not cause price fluctuations in its neighborhood areas. Thus, price information is independent across regions while local economic fundamentals are considered. Also, we find that US housing markets are not integrated and some types of constraints to information flow or

market imperfections are preventing full adjustment to long-run equilibrium in housing market. The communication of housing market information may be a result of utility convergence as suggested by urban economic theory, and such communication will not result in house price convergence across region. Over time, price signals are transmitted between clusters and within cluster 3.

However, because PC algorithm requires that all the input variables follow normal distribution, the results regarding causal flows among innovation need to be interpreted with caution. Jarque-bera test for normality is conducted and only part of the residuals from ECM and BVAR models are found to follow normal distribution. Thus, the correlation matrixes of these residuals fail to convey all the information about the interaction between them. In this case, the derived causal flows are an approximate of the true causal patterns.

2.6 Conclusion

This essay aims to examine the U.S. housing market from the perspective of market clustering and regional price movement. Cluster analysis is conducted to classify 29 U.S. metropolitan areas into three homogeneous clusters based on variables capturing housing attributes and economic environment. Discriminant analysis is employed next to validate the clustering results. It finds that all the economic and amenity variables significantly contribute to the assignment of a MSA into one of the three clusters. Also, the three clusters are separated far away from each other and no overlapping occurs

between clusters. The three clusters are located in the central, east coast and west coast of US respectively and the pattern of clustering is consistent with both economic conditions and geographical traits of the MSAs. Thus, the MSAs share similar economic and geographical characteristics are more likely to have similar attributes of housing market. However, this does not warrant price signal flowing across these MSAs.

A directed acyclic graphs approach is used to identify the pattern of price movements across the clustered housing markets. We find no statistically significant innovation communication among these MSAs, no matter between or within clusters. The price shocks in one area due to local factors will not introduce price fluctuations in other areas. We also find that U.S. housing markets are not integrated well and some types of constraints to information flow or market imperfections are preventing full adjustment to long-run equilibrium in housing market. Thus, the spatial equilibrium proposed in the urban economics does not yet exist in US housing market. However, the trend of utility convergence proved by urban economic theory may be the reason driving information flowing among housing markets, while price convergence across regions is not a result of such information communication.

The findings in this essay have several policy implications. First of all, policy control over central housing market does not have overflow effect on housing markets in coast areas, because there is no causal flow of house price innovations across these regions. Second, house prices in MSAs are cointegrated to some extent, so they are moving together due to convergence of economic fundamentals, investment behavior across regions and intelligence mobility. But the US housing market is not fully

integrated and policy incentives should be put into practice to encourage information and resource flow and fasten the adjustment to long-run equilibrium in housing market as suggested by urban economic theory. These policy incentives may include subsidies to the production in certain regions and tax benefit to building houses in some areas.

Even though the patterns of price movements across the clustered housing markets are not statistically significant, the economic significance of these patterns needs to be examined further. There are some limitations of this essay. First, the metropolitan areas include both central cities and suburbs, and are widely distributed geographically. However, the metropolitan area survey data are not necessarily representative of the whole housing markets (Lu, 2009). Second, there are 47 metropolitan areas in US, but only 29 of them are examined in the essay due to data availability. Thus, the representativeness of the results is weakened. Third, cluster analysis is cross-sectional type of method. Once the economic fundamentals and housing attributes changes significantly over time, the clustering pattern of US housing market is expected to change as well. However, the lack of integration and price movement in housing market is expected to persist over a long time.

3 LINKAGE BETWEEN THE U.S. HOUSING MARKET AND CREDIT STANDARDS

3.1 Background

The roots of the sub-prime mortgage crisis have been investigated a lot in the recent years. The declining real interest rate, lower credit standards, unreliable credit scoring technology, new structured mortgage products and easy monetary conditions are among those to be blamed. All these factors working together fueled the credit boom and housing bubble in U.S. during 2002 to early 2007.

As Greenspan suggested, the housing bubble was fundamentally engendered by the decline in real long-term interest rate. After the federal funds rate was reduced from 3.5% to 3.0% in 2001 after terrorist attack, it was further lowered to 1.0% after the accounting scandals in 2002. This decline to the historical low encouraged the home sales and refinancing. Adjustable rate mortgage (ARM) surged, which has its interest rate adjusted based on the market interest rate. At the same time, new structured mortgage products, such as collateralized debt obligations (CDOs) and mortgage-backed securities (MBSs), became increasingly popular among both domestic and foreign investors. These products securitize a pool of illiquid assets, including mortgage loans, and investors are paid back using proceeds or payments from those assets. The false AAA ratings of those structured products apparently convinced investors of the ability of

the products to meet their financial commitment. Thus, money kept flowing into the market and funding the housing bubble.

The low federal interest rate and increasing popularity of securities backed by subprime mortgages convinced lenders to lower their credit standards and extend loans to many borrowers with low down-payments and poor credit histories. As a result, housing demand got larger and so did the housing bubble. A large portion of subprime loan was ARMs, which at first benefited from the low federal funds rate over 2001-2003. But the fast increase in treasury interest rates starting from the second half of 2004 caused many subprime ARMs be reset at a much higher interest rate, and thus resulted in difficulties for many homebuyers to pay off their mortgages. The home loan default rates rose and it was hard for structured securities to sustain their values. During 2006-2007, more than three-quarters of the AAA-rated CDO bonds were downgraded (Bloomberg report, 2008), which further depressed the structured securities market. As a result, much less CDOs and private-label MBSs were issued, which in turn reduced the demand of outstanding mortgage and housing, and the housing bubble burst.

After the housing bubble, house price collapsed in 2007, and millions of American households became underwater on their mortgage. Because house is the largest single asset for most people, the contraction in housing wealth inevitably had a significant impact on consumer demand and on the aggregate economy. The recovery of the whole housing market seemed to be tied to the recovery of the general economy, so enormous government stimulation, low interest rate, tax credit and other forms of modification of loans were put into practice to get housing market back on track. The

housing market is on its way to recovery, but, because of what happened during the subprime mortgage crisis, lenders are reluctant to make home mortgage loan easy to borrowers. According to the Federal Reserve's April survey of senior officers, officers are not loosening up their tight credit requirements while their banks are seeing stronger demand for home loans.

Credit standards, along with interest rate and investment in structured securities, played an important role in the chaos and recovery of U.S. housing market. So this essay is aimed to discover the interactions between credit constraint and house price.

There are two objectives of this essay. First is to model the dependence among the stochastic components of house price, credit standard and other variables using multivariate copulas distribution (MVC). While correlation is only appropriate in measuring dependence for variables following multivariate normal distribution (Embrechts, McNeil, and Straumann, 1999), copula works well in separating a joint distribution into dependence structure and marginal distributions without normality assumption. The second objective is to forecast and simulate the underlying distribution of house prices based on the modeled dependence, and discover the causal flows between variables based on the simulated error terms. Directed acyclic graphs are used to find the causal patterns.

The reminder of the essay is organized as follows. The second section reviews previous literature. The third section discusses methodologies, and data is explained in the section 4. Models and results are presented and discussed in section 5. Section 6 concludes this essay, and the limitation in the analysis is discussed as well.

3.2 Literature Review

Traditional models analyzing house price usually only account for the impact from treasury interest rate and other economic fundamentals, but fail to consider credit standard in the models. House price models omitting credit constraints perform poorly in the 2000s (Duca, Muellbauer and Murphy, 2011 (b); Gallin, 2006). Magne and Rady (2006) replicate the facts that credit constraints delay some household' first home purchase and identify the ability of young households to afford the down payment on a starter home as a powerful driver of the housing market. Ariccia, Igan, and Laeven (2008) prove that the sharp increase in delinquency rates in the U.S. subprime mortgage market over 2006-2008 is related to the past credit boom and loosening credit standards. The close linkage between housing market and credit standard is also supported by Duca, Muellbauer and Murphy (2011(a), 2011(b)). All these studies lend support to the importance of credit standard in house market analysis.

In this essay, we employ time series econometric model to estimate the interaction of house price with credit constraint, as well as with other variables suggested by the inverted demand approach. What distinguish our estimation from previous literatures is that risk measure is incorporated into the model, and both credit constraint and house price are treated as stochastic random variables. Therefore, a distribution, rather than a point value, of house price for a certain level of credit standard will be derived.

According to Clements, Mapps and Eidman (1971), simulating uncertainty without realistically representing the covariance between related variables may introduce bias and variability into the analysis. They discuss a procedure that can correlate two events in the simulation model. Richardson and Condra (1978) extend the method of Clements, Mapp and Eidman, and report a general procedure for correlating random values of exogenous variables that is not distribution specific. In both of these procedures, the correlation matrix is calculated as the starting point. However, the correlation fails to convey the dependence structure when variables do not follow a multivariate normal distribution. In such cases, copulas offer a more flexible way to model dependence structure by not restricting the underlying uniform marginal distribution to be linearly correlated (Woodard, Paulson, Vedenov, and Power, 2011). Thus, in this essay, multivariate copulas distribution is used to simulate the stochastic components for both house price and credit constraint.

3.3 Models

3.3.1 Multivariate Copulas Simulation

Supported by Sklar's theorem, copulas separate a multivariate distribution function into the marginal distributions and the underlying dependence structure. It follows that

(6) $$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n))$$ or

(7) $$C(u) = C(u_1, u_2, ..., u_n) = F(F_1^{-1}(u_1), ..., F_n^{-1}(u_n))$$

where F(.) is the joint distribution function, $F_i(.)$ is the marginal distribution for the i^{th} variables, and C is the copula distribution.

In this essay, the Frank copula is used. It belongs to the class of Archimedean parametric copulas. According to Woodard et al. (2011), Archimedean copulas are a relatively flexible class of copulas that can adequately model a wide range of alternative dependence structures, and most have analytical solutions. The Frank copulas take the form

(8) $$C(u_1, u_2, ..., u_n) = \varphi^{-1}(\sum_{i=1}^{n} \varphi(u_i))$$

where $\varphi(t) = -\ln \frac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}$ with $\alpha \in \mathbb{R} \setminus \{0\}$. The Frank copulas can be simulated by first generating independent standard uniformly distributed variables, and then inverting the conditional copula density to generate random draws. The detailed method and algorithm used for Frank copulas simulation follows the one described by Embrechts, McNeil, and Straumann (1999). Let

(9) $$C_i(u_1, u_2, ..., u_i) = C(u_1, ..., u_i, 1, ..., 1), \quad i = 2, ..., n-1$$

represents i-dimensional marginal distribution of $C(u_1, u_2, ..., u_i)$, and write $C_1(u_1) = u_1$ and $C_n(u_1, u_2, ..., u_n) = C(u_1, ..., u_n)$. Suppose that $(U_1, ..., U_n)' \sim C$; the conditional distribution of U_i given the values of the first *i*-1 components of $(U_1, ..., U_n)'$ can be written in terms of derivatives and densities of the *i*-dimensional marginal

$$C_i(u_i \mid u_1,...,u_{i-1}) = P(U_i \le u_i \mid U_1 = u_1,...,U_{i-1} = u_{i-1}) = \frac{\partial^{i-1} C_i(u_1,...,u_i)}{\partial u_1 ... \partial u_{i-1}} \Big/ \frac{\partial^{i-1} C_{i-1}(u_1,...,u_{i-1})}{\partial u_1 ... \partial u_{i-1}} \quad (10)$$

provided both numerator and denominator exist. To simulate a value from $C_i(u_i \mid u_1,...,u_{i-1})$, generally a *u* is simulated from uniform(0,1) and then calculate $C_i^{-1}(u \mid u_1,...,u_{i-1})$. The specific steps for Frank copulas simulation are:

Step 1: simulate a value u_1 from *U*(0,1),

Step 2: simulate a value u_2 from $C_2(u_2 \mid u_1)$,

Step 3: continue in this way,

Step 4: simulate a value u_n from $C_n(u_n \mid u_1,...,u_{n-1})$.

In this essay, the =MVCOPULA( ) Simetar Stochastic function is used, which is programmed to return correlated uniform random numbers generated from an Archimedean copula.

According to the simulation algorithm, U_1 is set as an independent variable by default. U_2 is assumed to be granger caused by U_1, and U_3 is assumed to be granger caused by U_1 and U_2, etc. In the first run of copulas simulation, the house price is set as the n^{th} variable, which is reasonable because it is expected to be granger caused by all the other *n*-1 variables. An arbitrary order is assigned to each of the rest n-1 variables, and their orders will be updated using the causal patterns derived by directed acyclic graph method. Another run of copulas simulation will be conducted to incorporate the

causal information. This process of updating simulation order can repeat several times until a stable causal pattern and simulation order is obtained.

3.3.2 Directed Acyclic Graph (DAG) Method

DAGs uncover contemporaneous causal orderings among variables using arrows and vertices. Arrows represent the direction of information flow between variables, and there is no path that is from a variables and return to that same variables. There are several algorithms discussed by Pearl (2000) that can be used to identify and estimate the casual structure embedded in innovations, and the PC algorithm is used in this article. This algorithm starts with a completely undirected graph, i.e. each variable in the set is connected to every other variable by an undirected edge. Then, correlation and partial correlation are calculated for each pair of variables. If they are not significantly different from zero according to some critical statistic, then no significant relationship is defined for this pair of variables, and the edge between them is removed. Last, the remaining edges are believed to have directions, and an arrow (direction) is assigned to each of the edges according to the directional separation (d-separation) definition, which is given in Pearl (2000). TETRAD IV, software developed at Carnegie Mellon University, is used to determine causal patterns with a correlation based approach.

Empirically, PC algorithm requires a relatively large number of observation ($n \geq 100$) to ensure the reliability of the derived contemporaneous causal patterns.

Simulation can be used to overcome this problem by generating 1000 of data from the estimated distribution.

3.4 Data

According to the inverted demand approach, the variables are house price, house inventory, income, credit standard and imputed rental cost per dollar house price. In order to place these variables on a common scale, we divide each variable by its standard deviation. Quarterly data are used, which cover the time period from 1993:Q1 to 2010:Q4. Detail description of these variables and data is given next.

3.4.1 Housing Prices and Inventory

Median house prices for newly sold single family houses are collected from the census data of U.S. Department of Commerce. We adjust these nominal house prices with the Consumer Price Index (CPI-U), which is compiled by the Bureau of Labor Statistics and is based on a 1982 base of 100. The resulting real house prices are used in the analysis. Housing vacancies is used as the proxy for inventory. The number of year-round vacant, for sale houses (in thousands) is reported by the U.S. Census Bureau.

3.4.2 Income and Credit Standard

Disposable income is used, which is the amount of income left to an individual after taxes have been paid and available for spending and saving. Data for real disposable personal income is reported by the Bureau of Economic Analysis. Loan to value (LTV) ratio is considered as a good proxy for credit standard (Ariccia, Igan and Laeven, 2008; Duca, Muellbauer and Murphy, 2011(a), (b)). The LTV data is obtained from the historical summary data from the Federal Housing Finance Agency's monthly survey of rates and terms on conventional single-family non-farm mortgage loans.

3.4.3 Imputed Rental Cost per Dollar House Price.

Imputed rental cost measures the cost of owning a house, which compares the value of living in that property with the lost income that one would have received if the owner has invested that capital in an alternative investment. According to Himmelberg, Mayer, and Sinai (2005), the computation of imputed rental cost should take into account differences in risk, tax benefits from owner occupancy, property taxes, maintenance expenses, and any anticipated capital gains from owning the home. It can be written as:

(20) Imputed rental cost= $P_t r_{f,t} + P_t \omega_t - P_t \tau_t (r_{m,t} + \omega_t) + P_t \delta_t - P_t g_{t+1} + P_t \gamma_t$

where $r_{f,t}$ and $r_{m,t}$ is the risk-free interest rate and mortgage rate respectively, ω_t is the property tax rate, τ_t represents the marginal income tax rate, and δ_t is the depreciation

rate. g_{t+1} is the expected capital growth rate during the year *t+1*, and the last term γ_t is the risk premium per dollar house price for compensating the higher risk of owning a house instead of renting one.

The first term is the interest that the homeowner could have earned from other investment other than a house. The second term is the cost of property taxes. The third one is the tax shield benefit from property taxes and mortgage interest payment. These two payments can be itemized when filing federal income taxes and they are deductible from total taxable income. The fourth term is the house maintenance costs, estimated as a fraction $\delta_t$of house value. The fifth component is the expected capital gain during the yearg_{t+1}, and the last term is the total risk premium for compensating the higher risk assumed by homeowner by owning a house instead of renting one.

Imputed rental cost per dollar house price is the ratio of imputed rental cost to house price, and can be calculated as

(21) $$uc_t = r_{f,t} + \omega_t - \tau_t(r_{m,t} + \omega_t) + \delta_t - g_{t+1} + \gamma_t$$

The data for risk-free interest rates are the yield on Treasury bill at maturity of one year. The mortgage rate data is the conventional single-family mortgage rate that report by Federal Housing Finance Agency. Marginal tax rate of a typical home buyer is $\tau = 25\%$. The depreciation rate is $\delta = 2.5\%$ and the risk premium is $\gamma = 2\%$ (Himmelberg, Mayer, and Sinai, 2005). The expected capital growth rate is calculated as the average capital growth rate over the previous four periods.

3.5 Results

Based on the inverted demand approach suggested by Duca, Muellbauer and Murphy (2011, (b)), we model house price as a function of house supply, income, user costs and credit standard, i.e. $P_t = f(Inv_t, DPI_t, UC_t, LTV_t)$, where *t* is a time index, P is real house price, *Inv* represents house inventory, and *DPI* is real disposable person income. *UC* represents imputed rental cost per dollar house price. These variables are assumed to follow a multivariate copulas (MVC) distribution. Let $Y_t = (P_t, Inv_t, DPI_t, UC_t, LTV_t)'$, we can write the model as :

(22) $$\tilde{Y}_t = \hat{Y}_t + \varepsilon_t$$

$\tilde{Y}_t$ is stochastic with $\hat{Y}_t$ as its deterministic component and ε_t as its stochastic components. $\hat{Y}_t$ is estimated using econometric model and ε_t is simulated using MVC distribution. From now on, we denote our random variable as Y_{it} with *i* indicating the position (order) of the variable in the Y_t vector. For example, variable P_t is denoted as Y_{1t} for being the first variable in the Y_t vector.

The steps for estimating the parameters for the MVC distribution of $\tilde{Y}_t$ are similar to the steps for parameter estimation for multivariate empirical distribution illustrated by Richardson (2010), which are:

(1) Calculate the best econometric model to predict each of the random variables $\hat{Y}_{it}$.

(2) Calculate the residuals, $\hat{\varepsilon}_{it}$, from the econometric estimates as $\hat{\varepsilon}_{it} = Y_{it} - \hat{Y}_{it}$.

(3) Calculate the 5×5 Kendall's tau concordance matrix using the unsorted residuals.

(4) Simulate a 5×1 vector of correlated uniform standard deviates or CUSD's using Frank copulas.

(5) Calculate the fractional residuals for each variable as $e_{\hat{\varepsilon}_{it}} = \hat{\varepsilon}_{it} / \hat{Y}_{it}$ and then sort these values for each of the random variables. Denote the sorted fractional residuals as $S_{\hat{\varepsilon}_{it}} = sorted(e_{\hat{\varepsilon}_{it}})$. Calculate the pseudo minimums and maximums for each variable using the sorted fractional residuals.

(6) Assign probabilities to each of the sorted fractional residuals including a zero to the pseudo minimum and a one to the pseudo maximum. Denote the resulting CDF of the sorted fractional residuals as $F(S_{\hat{\varepsilon}_{it}})$.

After estimating the parameters for the MVC distribution of $\tilde{Y}_t$, we can simulate $\tilde{Y}_t$ as follow:

(23) $$\tilde{Y}_{it} = \hat{Y}_{it} \times (1 + EMP(S_{\hat{\varepsilon}_{it}}, F(S_{\hat{\varepsilon}_{it}}), CUSD_i))$$

where EMP( ) is a Simetar Simulation function used to generate empirical random variable based on the three inputs. By using the format of S_i as fractional deviates from a forecast, we insure the relative risk of the random variables to remain constant over the simulation period.

Next, we will present the estimation and simulation results step by step.

In the first step of selecting the best econometric model to estimate the deterministic component $\hat{Y}_{it}$, we consider multivariate time series models and Johansen cointegration test is conduct to facilitate model selection. Times series Y_{it} might be

stationary in levels or first differences, i.e. *I*(0) or *I*(1). Rather than pretesting these for unit roots, the Johansen procedure formulates the question within the model. For the model being tested in this essay,

$$\Delta Y_t = c + \Pi Y_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta Y_{t-i} + \varepsilon_t \tag{24}$$

if the cointegration test fails to reject the null of cointegration rank r=0, the inference is that the error-correction coefficient Π is zero and the error correction model (ECM) reduced to a VAR model in first differences. If the cointegration test rejects all the cointegration ranks r less than n (number of random variables), the inference is that Π has full rank and Y_t is stationary in levels which can be modeled with VAR in levels.

Both trace and maximum eigenvalue tests are conducted to determine the rank of Π, and the test statistics are reported in Table 18. Both test statistics indicate that we cannot reject the null of cointegration rank is zero at 5% statistical significance level. Thus, first differences are used in VAR model to estimate the random variables $\hat{Y}_t$.

We also applied the cointegration rank search method discussed in Bessler and Wang (2005), and jointly select the lag length and cointegration rank based on Schwartz information criterion (SIC) and Hannan and Quinn's Φ measures (Φ). The information criterion statistics are listed in Table 19, and they indicate a VAR model in first differences with four lags is the most preferred model. Based on the results of the Johansen cointegration test and the statistics of SIC and Φ, the VAR (4) model in first differences is estimated. The estimated c and Γ_i (i=1,…,4) is reported in Table 20.

Kendall's Tau concordance matrix is calculated based on the residuals from the VAR model, and it is reported in Table 21. Kendall's Tau correlation coefficient is a non-parametric statistic used to measure the association between variables. When the sample fails to follow normal distribution or sample size is small, this measure of rank correlation is a robust alternative, which does not rely on any assumptions on the distribution of variables. For example, assume *X* and *Y* are two joint random variables, any pair of observation (x_i, y_i) and (x_j, y_j) are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. The Kendall's tau coefficient is defined as:

$$\tau = \frac{(number\ of\ \ concordant\ pairs) - (number\ of\ \ discordant\ pairs)}{\frac{1}{2}n(n-1)} \tag{25}$$

and the coefficient must be in the range $-1 \le \tau \le 1$. From Table 21, we can see that LTV is moderately associated with house price (0.1460), house inventory (0.1326) and disposable household income (0.1535). Except for LTV, house price has weak association with other variables. Thus, if house price interacts with other variables, LTV might serve as an intermediary between them.

Following the steps (4)-(6) discussed earlier in this section, we forecast $\Delta\tilde{Y}_t$ for 2011:Q1 based on the last four observations (2010:Q1-2010:Q4) in our data set as:

(26) $$\Delta\tilde{Y}_{i,t} = (\hat{c}_i + \sum_{j=1}^{4}\sum_{k=1}^{5}\hat{\Gamma}_{jk}\Delta Y_{k,t-j}) \times (1 + EMP(S_{\hat{\varepsilon}_{it}}, F(S_{\hat{\varepsilon}_{it}}), CUSD_i))$$

and simulate it for 1000 times.

The summary statistics of the simulated random variables are reported in Table 22. Also reported are the historical mean and standard deviation of the variables. We can see that the historical and simulated means and standard deviations are very close to each other, so we expect the simulation results are valid and capture the major characteristics of historical data. To further validate the simulation a comparison is made between the simulated series and historical series, and the resulting statistics are listed in Table 23. According to these statistics, we fail to reject the null that the mean vectors are equal, and also fail to reject the null that the covariance matrices are equivalent. Thus, the conclusion that the simulated distribution is valid can be made.

For each iteration, random variables $\tilde{Y}_{i,t}$ and residuals $\tilde{\varepsilon}_{it}$ are recovered as

(27) $$\tilde{Y}_{i,t} = Y_{i,t-1} + (\hat{c}_i + \sum_{j=1}^{4}\sum_{k=1}^{5}\hat{\Gamma}_{jk}\Delta Y_{k,t-j}) \times (1 + EMP(S_{\hat{\varepsilon}_{it}}, F(S_{\hat{\varepsilon}_{it}}), CUSD_i))$$ and

(28) $$\tilde{\varepsilon}_{it} = (\hat{c}_i + \sum_{j=1}^{4}\sum_{k=1}^{5}\hat{\Gamma}_{jk}\Delta Y_{k,t-j}) \times EMP(S_{\hat{\varepsilon}_{it}}, F(S_{\hat{\varepsilon}_{it}}), CUSD_i))$$

respectively. Thus, we have 1000 out-of-sample forecast for house price in 2011:Q1. Based on these data, a CDF graph is obtained and presented in figure 9. Different from previous studies only giving a point forecast of house price, the graph specify the distribution of forecasted median house price, based on which we can derive the value of certain quantiles of interest. Some important quantile values are calculated and summarized in Table 24. For example, the median housing prices for 2011:Q1 have 5%

chance of dropping below $211,779.6, and 5% chance of rising above $245,679.2 (95% confidence interval). The median value of the simulated house price is $228,729.4, which is very close to the observed median house price $226,900 for that time period.

With a 5-step-ahead forecast based on the VAR model, we simulate the distribution of forecasted $\tilde{Y}_t$ at 2012:Q1. Its CDF graph is presented in figure 10 and several quantile values are listed in Table 24. We predict that the median housing prices have 5% chance of going below $206,775.3, and 5% chance of shooting up beyond $239,874.2 at 2012:Q1. The median value of the median housing price in that period is forecasted to be around $223,324.8, which is again very close to the observed median house price $ 225,750.

Based on the tables and figures discussed above, the simulation of stochastic components retains all the importation information about the variables and their interactions. Together with the forecast of VAR-in-difference model, it not only gives a close point forecast for median house price, but also describes the underlying distribution of it, which provides more information to both policy makers and housing market investors. For example, policy makers can conduct scenario analysis by changing the value and level of fluctuation of LTV or house inventory to explore the change in the distribution of house price so as to better control the level of house price and its fluctuation. Investors can adjust their investment portfolio based on the risk indicated in the house price distribution and their risk appetite to find their optimal portfolio on the capital market line.

The last thing we consider is the contemporaneous causal patterns between house prices, credit standard and other factors. From the VAR in first difference estimation, we have 68×5 matrix of residuals, which contains the contemporaneous causal information for the five variables. This residual matrix is used as the input of DAG model, and the causal flows sorted out by PC algorithm is presented in Figure 12 Contemporaneous Causal Patterns among the Five Random Variables, 2011:Q1The resulting causal pattern is consistent with the one obtained based on the simulated residuals. This conforms to our expectation because the complete homogeneity test suggests the equivalence of correlation matrixes of estimated residuals and simulated residuals. In other words, the estimated residuals and simulated residuals convey the same contemporaneous causal information among the five variables. A chi-square test is formulated in the PC algorithm to test the null hypothesis that "the population covariance matrix over all of the measured variables is equal to the estimated covariance matrix over all of the measured variables written as a function of the free model parameters" (TETRAD IV User's Manual). Since indicated p-value is 0.915, greater than 0.05, we fail to reject the null and conclude that the derived causal pattern is statistically significant at 5% significance level. The pattern reveals that the credit standard is the only direct causal variable to house price. The other three variables affect house price indirectly, and all of them have credit standard working as their messenger.

The positive causal relationship between LTV and house price confirms that relaxing credit standard will cause an increase in the house price. The easier it is to get a home loan, the more likely people will choose to buy a house instead of renting one.

Thus, the larger demand drives up the house prices. The positive correlation between the DPI and LTV is consistent to the general expectation. People with higher disposable income usually have lower default risk, and thus less strict credit constraints are imposed upon them leading to higher LTV. A possible explanation for the positive causal flow from house inventory to LTV is: as house inventory goes up, house supply may exceed house demand, so LTV may be brought up to encourage house purchase behavior to reduce the vacant houses. Similarly, as house ownership costs increase, LTV may be adjusted upward to make it more attractive for families without a lot of liquid funds to enter the market.

The causal information conveyed in Figure 12 can be very helpful for policy making. For example, when government notices a phenomenon of oversupply of houses, it should consider relaxing credit constraints to encourage the house buying behavior so as to stimulate the housing market and related industries. But, this decision should be made with disposable household income taken into account as well. If household income is decreasing in the presence of oversupply of houses, relaxing credit constraints is not suggested and should be carried out with caution. Low dispensable household income indicates low capability of households to pay off their mortgage on time and higher risk of foreclosure, which increases the expected loss to the economy as a whole. Thus, LTV should be watched closely and assigned a value to balance the oversupply of house inventory and the risk of too many sub-prime mortgages.

The housing bubble that occurred in 2007 is an example that disposable household income was not considered in the presence of thriving housing market.

During those days, value of assets owned by a household, instead of dispensable household income, was the major consideration in loan decision. Loan agents were over-optimistic about the future of housing market, and believed that house prices would keep rising which warranted the ability of household to pay off their mortgage. If disposable household income was considered, loan agents would be cautious regarding how much cash, instead of collaterals, they could get back and whether the payments would be on time. Therefore, the knowledge of causal relationship found here can help set credit standard responsibly and avoid going down the same road of subprime mortgage crisis in 2007.

However, there are two sources of biases in the derived causal patterns. First, PC algorithm requires that no variable is omitted which causes two or more other variable selected for analysis. But, there might exist economic fundamentals we fail to consider which influence two or more of the variables in our model. Second, variables used in a DAG model need to follow a multivariate normal distribution for the model to be fully efficient. But, the result of Jarque-bera Test for normality shows that only parts of the residuals from the model are normally distributed. Therefore, the derived causal pattern is an approximate of the true causal relationship among variables.

3.6 Conclusion

The occurrence of subprime mortgage crisis brought attention to the interaction between relaxing credit standard and housing market bubble. In this essay, we use

multivariate copulas (MVC) distribution to model the dependence among the stochastic components of house price and credit constraint, as well as other variables. Also, we forecast and simulate the underlying distribution of house price based on the modeled dependence. There are two improvements over the previous studies. First, dependence, instead of correlation, among variables is examined with MVC. Correlation is only appropriate in measuring dependence for variables following multivariate normal distribution, while copula works well in separating a joint distribution into dependence structure and marginal distributions with no normality assumption. Therefore, MVC is a better way to examine dependence when variables do not follow multivariate normal distribution. Second, instead of a point value, we forecast the full distribution of house price. To accomplish this, stochastic simulation is applied. Based on the simulated data, several quantile values are derived, which can be useful input for political or investment decision. Moreover, we can change the values for parameters to simulate for alternative scenarios and see how the distribution or risk changes across difference scenarios, which is an interest topic to address in future research.

This essay also improves our understanding about interaction between house price and credit standard by sorting out their causal patterns. Causation usually contains more information than correlation. The derived causal patterns show that, more than just being correlated with house prices, reduced credit standard causes increase in housing price. Furthermore, we find that other factors, such as disposable income, housing inventory and cost of house ownership, impose their impacts on house price through their influence on LTV. In other words, credit standard not only directly affects house

price, it also works as the intermediate passing the influence of other factors onto house price. This finding suggests policy makers take extra caution when considering relaxing credit standard to stimulate housing market and its related industries. If both oversupply of houses and decreasing disposable household income is observed, housing bubble might exist and relaxing credit standard to reduce house oversupply might be devastating.

There is a limitation of this essay regarding the causal pattern analysis. DAG only works to sort out contemporaneous causal flows, so all the data input are for the same time *t* and the resulting causal pattern is valid for that time *t* only. When causal flows vary over time, they have to be re-estimated for each time period.

4 FORECASTING HOUSE PRICES: DYNAMIC FACTOR MODEL VERSUS LBVAR MODEL

4.1 Background

The housing market is of great importance to the economy. Housing construction and renovation boost the economy by increasing in aggregate expenditures, employment and the volume of house sales. They also stimulate the demand for related industries such as household durables. The oscillation of housing prices affects the value of asset portfolio for most households for whom a house is the largest single asset. Moreover, price movements influence the profitability of financial institutions and the soundness of the financial system. Recent studies further justify the necessity of housing price analysis, concluding that the housing sector plays a significant role in acting as a leading indicator of the real sector of the economy and that assets prices help forecast both inflation and output (Forni, Hallin, Lippi, and Reichlin, 2003; Stock and Watson, 2003; Das, Gupta, and Kabundi, 2009a). Thus, a timely and precise forecast for housing prices can provide valuable information to policy makers and help them better control inflation and design more effective policies. Also, these forecasts can direct individual market participants to make wise investment decisions. In light of the economic recession started by the sub-mortgage crisis, analyzing the influence of the burst of the housing price bubble and predicting its future moving trend is more important than ever.

Unlike the financial market, the housing market is illiquid and heterogeneous physically and geographically, which makes forecasting house prices a difficult task. Moreover, the subtle interactions between house prices and other macroeconomic fundamentals further complicate predictions. The change in house prices can either be attributed to a national phenomenon, such as the effect of monetary policy, or to regional factors, such as local taxation. Changing housing prices can also either indicate changes in the real sector, such as labor input and production of goods, or be affected by activities in the nominal sector, such as financial market liberalization (Gupta, Miller, and Van Wyk, 2010).

Many previous studies find empirical evidence supporting the significant interrelationships between house prices and economic variables, such as income, interest rates, construction costs and labor supply (Linneman, 1986; Wheaton, 1999; Quigley, 1999; Tsatsaronis and Zhu, 2004). However, because house price is a leading indicator of inflation and output, they are expected to interact with a much wider range of real and nominal variables. Thus, the quantification of these interrelationships is not enough for a precise estimation or prediction of house prices in a way that small-scale models potentially omit information contained in thousands of variables. In other words, a large number of economic variables help predict the growth of real house prices (Rapach and Strauss, 2009).

This essay aims to discuss the model selection for analyzing recent house price in 40 metropolitan areas in the United States from the perspective of large-scale models, i.e., the Dynamic Factor Model (DFM) and the Large-scale Bayesian Vector

Autoregressive (LBVAR) model. There are three major contributions of this essay. First, the most recent data to 2012:M6 are used for the estimation, which updates the understanding of U.S. housing market and the forecast performance of large-scale models. Second, for the DFM model, a dynamic component approach is used, which has one half of the estimation error as a static component approach. Finally, an encompassing test is conducted, and the forecast combination of DFM and LBVAR models is found to improve forecast accuracy half of the time. In the other half of the time, the results are mixed. This suggests that each of the two models contains marginal information that is not used in the prediction of its counterpart. In other words, a combined forecast may contain more relevant information, which makes it a better forecast alternative to an individual prediction.

The remainder of this paper is organized as follows. Section 2 provides a literature review. Section 3 introduces and illustrates the DFM and LBVAR models, as well as encompassing tests. Section 4 discusses the data. Section 5 evaluates and compares individual and combined forecasts. Section 6 concludes the paper and discusses the limitations of the employed models.

4.2 Literature Review

The advantages of large-scale models over small-scale counterparts are proved and discussed by many scholars (Forni, Hallin, Lippi, and Reichlin, 2005; Das, Gupta, and Kabundi, 2008, 2009b, 2011; Gupta and Kabundi, 2008a; Gupta, Kabundi, and

Miller, 2009a; Stock and Watson, 2004; Bloor and Matheson, 2010). Thus, this section is placed in the context of research using large-scale models for housing price prediction. The most popular methodologies for analyzing large-scale datasets include DFM, LBVAR (spatial or non-spatial), Factor-Augmented Vector Autoregressive (FAVAR) model, Dynamic Stochastic General Equilibrium (DSGE) model, and forecast combination methods. Their forecasting performances have been examined and compared in many previous studies, but the conclusions vary to a large extent.

First, the forecasting performances between DFM and LBVAR are discussed by Das, Gupta, and Kabundi (2008), Das, Gupta, and Kabundi (2009a), and Gupta and Kabundi (2008a). These three papers examine the housing market in South Africa but with different aggregation levels. Das, Gupta, and Kabundi (2008) and Gupat and Kabundi (2008a) claim that DFM is the better model to base one's forecast, while Das, Gupta, and Kabundi (2009a) obtain the opposite conclusion, i.e., LBVAR outperform DFM. Second, the forecasting performances between FAVAR and LBVAR are discussed in the studies of Das, Gupta, and Kabundi (2009b) and Gupta, Kabundi, and Miller (2009a, 2009b). The housing price growth rate in nine census divisions of the U.S., U.S. real house price index, and the housing prices in twenty U.S. states are studied in these three papers. The first and third papers show evidence supporting that FAVAR is better suited for forecasting house price growth. But the second paper concludes that small-scale BVAR model outperforms both FAVAR and LBVAR in terms of forecasting. Third, the comparison of forecasting power between DSGE and other large-scale models are discussed in the paper by Gupta, Kabundi, and Miller (2009a) and Gupta and

Kabundi (2008b). The first and second papers are conducted under the background of U.S. and South Africa housing markets, respectively. The result of the first paper shows that DSGE model forecast a turning point more accurately than the FAVAR and LBVAR models, while the second paper suggests that DFM performs significantly better than DSGE. Last, forecast combination methods are discussed by Stock and Watson (2003). The authors find that the combination forecasts performed well when compared to forecasts constructed using DFM framework, but they also attribute the poor performance of the DFM forecasts to the relatively small number of series examined.

The contradictive conclusions regarding the forecasting power of these popular large-scale models indicate that there is not a large-scale model that performs consistently better than its other alternatives. The superior forecasting performance of a model is defined with respect to the time period examined and the specific object studied. When the examined time period and study object change, the forecasting power of a model might be strengthened or weakened. This explains why some models are best for U.S. market but not for the South Africa market and why the best-suited models for data of metropolitan level, census division level and states level are different even for the same country. For the same reason, results from the studies using old data are becoming less convincing as time goes by. Since the observations of 2006:Q4 is the most recent data used in the previous papers examining U.S. housing market, the results from those papers obviously can no longer be applied to current housing market, especially after the sub-prime mortgage crisis. Our paper uses the most updated data to 2012:M6, and

examines the housing market by metropolitan areas. Thus, it updates and extends the understanding of U.S. housing market.

In all the reviewed papers which apply DFM framework to U.S. housing market analysis, static principal component approach (PCA) is used, which estimates the common component by projecting onto the static principal components of the data. However, based on contemporaneous covariances only, it fails to exploit the potentially crucial information contained in the leading and lagging relationships between the elements of the panel (Forni *et al*., 2005). In this paper, we use the dynamic component approach proposed by Forni et al. (2005). This approach obtains estimates of common and idiosyncratic variance-covariance matrices at all leads and lags as inverse Fourier transforms of the corresponding estimated spectral density matrices, and thus overcomes the limitation of static PCA.

4.3 Models

Economy-wide forecasting models are generally formulated as Vector Autoregressive (VAR) model or Vector Autoregressive Moving Average (VARMA) models. But the over- parameterization problem embedded in these model results in multi-colinearity and loss of degrees of freedom which can lead to inefficient estimates and large out-of-sample forecasting errors (Dua and Ray, 1995). Therefore, these models are no longer appropriate for cases with a large number of cross-sectional variables, and Dynamic Factor Model (DFM) and Bayesian Vector Autoregressive (BVAR) are

proposed to overcome the over-parameterization problem. The first two parts of this section discuss these two models, and the third part discusses the encompassing test used to combine the forecasts.

4.3.1 Dynamic Factor Model (DFM)

Within DFM framework, each time series in the panel is structured as the sum of two mutually orthogonal components: the common component and the idiosyncratic component. The common component is strongly correlated with the rest of the panel and has reduced stochastic dimension, while the idiosyncratic component is either mutually orthogonal or "mildly cross-correlated" across the panel. In the DFM, multivariate information is used for forecasting the common component, and the idiosyncratic can be predicted reasonably well by means of traditional univariate methods, i.e., AR (4) model.

The DFM used in this paper follows the framework developed by Forni*et al.* (2005), which has three desirable characteristics. First, it adopts the dynamic principal component (PC) method, which has smaller estimation errors than its static counterpart proposed by Stock and Watson (1999). Instead of using only contemporaneous covariances, the dynamic PC method bases its estimation on the common and idiosyncratic variance-covariance matrices at all leads and lags. Second, this DFM method obtains its *h*-month-ahead forecast as the projection of the *h* month observation onto the estimated generalized principal components, which overcome the two-sided filtering problem of the DFM method proposed by Forni, Hallin, Lippi, and Reichlin

(2000). Two-sided filtering is not a problem for within-sample estimation, but it does cause some difficulties in the forecasting context due to the unavailability of future observation. Third, this DFM method allows for cross-correlation among the idiosyncratic components, because orthogonality among these components is an unrealistic assumption.

Consider a double sequence $\{ y_{it}, i \in \mathrm{N}, t \in \mathrm{Z} \}$. Suppose that $\{ x_{it}, i \in \mathrm{N}, t \in \mathrm{Z} \}$ is the standardized version of $\{ y_{it} \}$, i.e. the *n*-dimensional vector process $\mathbf{x}_n = \{\mathbf{x}_{nt}, t \in \mathrm{Z}\}$, where $\mathbf{x}_{nt} = (x_{1t}\ x_{2t}\ \ldots\ x_{nt})'$, is zero mean and stationary for any *n*. According to Forni*et al*. (2005), $\mathbf{x}_{nt}$ can be written as the sum of two orthogonal components:

$$x_{it} = b_{i1}(L)u_{1t} + b_{i2}(L)u_{2t} + \ldots + b_{iq}(L)u_{qt} + \xi_{it} = \chi_{it} + \xi_{it} \tag{29}$$

where $\mathbf{u}_t$ is a $q \times 1$ of dynamic factors and L stands for the lag operator. The variables χ_{it} and ξ_{it} represent the common and idiosyncratic components respectively. χ_{it} is unobservable and needs to be estimated. Forni*et al*. (2000) have shown that the projection of x_{it} on all leads and lags of the first q dynamic principal components of $\mathbf{x}_n$, obtained from the population spectral density matrix $\mathbf{\Sigma}_n$, converges to χ_{it} in mean square as *n* tends to infinity, i.e., $\chi_{it,n} \xrightarrow{p} \chi_{it}$, where $\chi_{it,n}$ denoted this projection. Empirically, we construct the finite-sample counterpart of $\chi_{it,n}$, which is based on the estimated spectral density matrix $\hat{\mathbf{\Sigma}}_n$, call it $\hat{\chi}_{it,n}$. By combining the convergence of $\chi_{it,n}$ to χ_{it} with the fact that $\hat{\chi}_{it,n}$ is a consistent estimator of $\chi_{it,n}$ for any *n* as *T* goes to

infinity, it can be derived that $\hat{\chi}_{it,n}$ is a consistent estimator of χ_{it} as any *n* as *T* tends to infinity. Thus, equation (29) can be re-written as $\hat{x}_{it} = \hat{\chi}_{it,n} + \hat{\xi}_{it}$, where $\hat{\xi}_{it}$ is a consistent estimator of ξ_{it} based on a traditional univariate method.

The *h*-month ahead forecast of $y_{i,T+h|T}$ is computed as follows:

(30) $$\hat{y}_{i,T+h|T} = \hat{\sigma}_i \hat{x}_{i,T+h|T} + \hat{\mu}_i = \hat{\sigma}_i (\hat{\chi}_{i,T+h|T} + \hat{\xi}_{i,T+h|T}) + \hat{\mu}_i$$

where $\hat{\sigma}_i$ and $\hat{\mu}_i$ are the sample variance and sample mean of the i^{th} variable and T is the sample size. $\hat{\xi}_{i,T+h|T}$ can be estimated using a traditional univaritate method, i.e., AR(4). $\hat{\chi}_{i,T+h|T}$ is obtained by the dynamic PC analysis, which starts with the estimation of the sample autocovariance matrix of $\mathbf{x}_{nt} = (x_{1t}\ x_{2t}\ \ldots x_{nt})'$, i.e., $\hat{\mathbf{\Gamma}}_{n,k} = \frac{1}{T-k}\sum_{t=k+1}^{T} \mathbf{x}_{n,t}\mathbf{x}'_{n,t}$.

Then the spectral density matrix of $\hat{\mathbf{\Gamma}}_{n,k}$ is calculated through discrete Fourier transform $\hat{\mathbf{\Sigma}}(\theta_h) = \frac{1}{2\pi}\sum_{k=-M}^{M} w_k \hat{\mathbf{\Gamma}}_{n,k} e^{-i\theta_h k}$, where w_k is Barleet-lag window estimator weight $w_k = 1 - \frac{|k|}{M+1}$, and $\theta_h = \frac{2\pi}{2M+1}h$, *h*=-*M*, …, *M*. To ensure the consistency of results, M is a function of T and should satisfy two conditions that $M(T) \to \infty$ as $T \to \infty$ and $\limsup_{T\to\infty} M^3(T)/T < \infty$ as $T \to \infty$. Empirically, $M = \sqrt{T}$ is usually used.

Then a two-step procedure proposed in the study of Forni*et al*. (2005) follows. First step is to obtain estimates of common and idiosyncratic variance-covariance matrices at all leads and lags as inverse Fourier transforms of the corresponding estimated spectral density matrices. At a given frequency θ, there exists

(31) $$\mathbf{V}(\theta)\hat{\mathbf{\Sigma}}(\theta) = \mathbf{D}(\theta)\mathbf{V}(\theta)$$

where $\mathbf{D}(\theta)$ is a diagonal matrix having the eigenvalues of $\hat{\mathbf{\Sigma}}(\theta)$ on the diagonal and $\mathbf{V}(\theta)$ is the $n \times n$ matrix whose columns are the corresponding row eigenvectors. Based on the central idea of PC analysis which claims that the first few (q) largest PCs (dynamic factors) will account for most of the variation in the original variables (Jolliffe, 2002), the spectral density matrix of the common component have the following relationship with the first q largest eigenvalues and their corresponding eigenvectors:

(32) $$\mathbf{V}_q(\theta)\hat{\mathbf{\Sigma}}_\chi(\theta) = \mathbf{D}_q(\theta)\mathbf{V}_q(\theta) \text{ or } \hat{\mathbf{\Sigma}}_\chi(\theta) = \tilde{\mathbf{V}}_q(\theta)\mathbf{D}_q(\theta)\mathbf{V}_q(\theta)$$

where $\tilde{\mathbf{V}}_q$ denotes the conjugate transpose of $\mathbf{V}_q$. The spectral density matrix of the idiosyncratic component is the estimated as $\hat{\mathbf{\Sigma}}_\xi(\theta) = \hat{\mathbf{\Sigma}}(\theta) - \hat{\mathbf{\Sigma}}_\chi(\theta)$. The covariance matrices of common and idiosyncratic parts are estimated respectively through the inverse Fourier transform of spectral density matrices as following:

(33) $$\hat{\mathbf{\Gamma}}_k^\chi = \frac{2\pi}{2M+1}\sum_{j=-M}^{M} \hat{\mathbf{\Sigma}}_\chi(\theta_h)e^{ik\theta_h} \text{ and } \hat{\mathbf{\Gamma}}_k^\xi = \frac{2\pi}{2M+1}\sum_{j=-M}^{M} \hat{\mathbf{\Sigma}}_\xi(\theta_h)e^{ik\theta_h}$$

The second step is to use these estimates to construct the contemporaneous linear combinations of x_{it}'s that minimize the idiosyncratic-common variance ratio, and the linear combination gives the estimate of $\hat{\chi}_{it,n}$. The resulting aggregates can be obtained as the solution of a generalized principal component problem: $\mathbf{V}_G\hat{\mathbf{\Gamma}}_0^\chi = \mathbf{D}_G\mathbf{V}_G\hat{\mathbf{\Gamma}}_0^\xi$, where D_G is a diagonal matrix having the generalized eigenvalues of the pair ($\hat{\mathbf{\Gamma}}_0^\chi, \hat{\mathbf{\Gamma}}_0^\xi$) on the diagonal and $\mathbf{V}_G$ is the $n \times n$ matrix whose columns are the corresponding row

eigenvectors. The j^{th} generalized PCs are defined as $\hat{\mathbf{P}}^{G}_{t,j} = \mathbf{v}_{G,j}\mathbf{x}_{nt}$, where $\mathbf{v}_{G,j}$ is the j^{th} generalized row eigenvector corresponding to the j^{th} largest generalized eigenvalues. Based on the PC theory, the *r* aggregates $\hat{\mathbf{P}}^{G}_{t,j}$, j=1, …, *r*, preserves most of the information of $\mathbf{x}_n$. Consider a space Δ_r spanned by the *r* aggregates, $\hat{\chi}_{it}$ is the projection of x_{it} onto this space, i.e., $\hat{\chi}_{it} = proj(x_{it} \mid \Delta_r)$. The *h*-month ahead forecast $\hat{\chi}_{i,T+h|T}$ is based on the information available at time *T* and is estimated as the projection of x_{iT} on to the space spanned by the *r* aggregates $\hat{\mathbf{P}}^{G}_{T,j}$, j=1, …, *r*. Thus, the estimates of $\hat{\chi}_{i,T+h|T}$ is:

(34) $$\hat{\chi}_{i,T+h|T} = \hat{\mathbf{\Gamma}}^{\chi}_{h}\tilde{\mathbf{V}}_{Gr}(\mathbf{V}_{Gr}\hat{\mathbf{\Gamma}}_{0}\tilde{\mathbf{V}}_{Gr})^{-1}\mathbf{V}_{Gr}\mathbf{x}_{nT}$$

where $\mathbf{V}_{Gr}$ is the $n \times r$ matrix whose columns are the generalized row eigenvectors corresponding to the *r* largest generalized eigenvalues. The $\hat{y}_{i,T+h|T}$ can be obtained by plugging $\hat{\chi}_{i,T+h|T}$ into equation (3).

4.3.2 Large-Scale BVAR (LBVAR) Model

LBVAR is another alternative to VAR to accommodate large-scale variables and overcome the over parameterization problem. As described in Litterman (1981), Doan, Litterman, and Sims (1984), Todd (1984), Litterman (1986), and Spencer (1993), instead of estimating longer lags and/ or less important variables, the Bayesian technique imposes restrictions on these coefficients by assuming that these are more likely to near zero than the coefficients on shorter lags and/or more important variables. If, however,

there are strong effects from longer lags and/or less important variables, the data can override this assumption. This method supplements the data with prior information on the distribution of the coefficients. With each restriction, the number of observations and degrees of freedom are increased by one in an artificial way. Therefore, the loss of degrees of freedom due to over parameterization associated with a VAR model is not a concern in LBVAR model.

The restrictions are imposed by specifying normal prior distributions with means zero and small standard deviations for all coefficients with decreasing standard deviations on increasing lags. The exception is the coefficient on the first own lag of a variable that has a mean of unity. This prior is called the "Minnesota prior" and takes the form $\beta_i \sim N(1, \sigma_{\beta_i}^2)$ and $\beta_j \sim N(0, \sigma_{\beta_j}^2)$, where β_i represents the coefficients associated with the lagged dependent variables in each equation of the LBVAR and β_j represents any other coefficient. The standard deviation of the prior distribution for lag *m* of variables *j* in equation *i* is specified as $\sigma(i, j, m) = [w \times g(m) \times f(i, j)] \frac{\hat{\sigma}_i}{\hat{\sigma}_j}$, where

$f(i, j) = \begin{cases} 1, & if \ \ i = j \\ k, & other \ wise \end{cases} (0 < k < 1)$, $g(m) = m^{-d} (d > 0)$ and $\hat{\sigma}_i$ is the standard error of an univariate autoregression for variable *i*. The ratio $\hat{\sigma}_i / \hat{\sigma}_j$ scales the variables to account for differences in units of measurement and allows the specification of the prior without consideration of the magnitudes of the variables. The parameter *w* is the standard deviation on the first own lag and describes the overall tightness of the prior. The tightness on lag *m* relative to lag 1 is given by the function *g*(*m*), and is assumed to

have a harmonic shape with decay factor *d*. The tightness of variable *j* relative to variables *i* in equation *i* is represented by the function *f*(*i*, *j*). The value of *f*(*i*, *j*) determines the importance of variable *j* relative to variable *i*, *with* higher values implying greater interaction. A tighter prior occurs by decreasing *w*, increasing *d*, and/or decreasing *f*(*i*, *j*).

In the analysis, both regional and national data are used. Realizing that national variables affect both national and regional variables, and regional variables primarily influence only other regional variables, the LBVAR should be estimated with asymmetric priors. Following Das, Gupta, and Kabundi (2009b), the weight, i.e. $f(i, j)$, of a national variable in a national equation, as well as a regional equation, is set at 0.6. The weight is fixed at 0.1 and 0.01 in other regional and national equations, respectively. Last, the weight of the regional variables in its own equation is 1.0. In the standard Minnesota-type prior, the overall tightness (*w*) takes the values of 0.1, 0.2, and 0.3, while the lag decay (*d*) is generally chosen to be equal to 0.5, 1.0, and 2.0.

4.3.3 Encompassing Test

A linear combination of multiple forecasts may often yield more accurate forecasts than using an individual prediction to the extent that the component forecasts contain useful and independent information (West, 2001; Newbold and Harvey, 2002; Fang, 2003; Wang and Bessler, 2004; Kisinbay, 2007; Costantini and Pappalardo, 2008). In order to further enhance predictive power, Kisinbay (2007) and Costantini and

Pappalardo (2008) suggest reducing the number of available forecasts before combining them, and encompassing tests usually are conducted to fulfill this task.

Following West (2001), the encompassing test used here is one in which the explained variable is regressed on competing out-of-sample predictions. Suppose there are two models, model *i* and *j*. The encompassing regression is written as $y_t = \beta_1 \hat{y}_{it} + \beta_2 \hat{y}_{jt} + \varepsilon_t$, where y_t is the variable being explained by the competing models, $\hat{y}_{it}$ ($\hat{y}_{jt}$) is the forecast of y_t from model *i*(*j*). Model *i* is said to encompass model *j* if $\beta_1 \neq 0$, $\beta_2 = 0$ because model *j* does not contain marginal information helpful in explaining y_t, conditional on model *i*. Similarly, model *j* is said to encompass model *i* if $\beta_1 = 0$, $\beta_2 \neq 0$. In other cases, no conclusion can be drawn regarding which model encompasses the other one.

Usually, root mean square error (RMSE) is the measurement used to judge the forecast performance of competing models. However, a forecast with smaller RMSE does not necessarily contain all the information of the one with larger RMSE. Thus, RMSE and encompassing tests should be working together as complementary forecast criteria (Ericsson, 1992). According to the algorithm described by Costantini and Pappalardo (2009), the encompassing test and forecast combination are carried out as following. First, calculate the RMSE of the out-of-sample forecast for each model using out-of-sample forecasts and observed values. Rank the models according to their performance based on RMSE. Second, pick the model with the lowest RMSE, and test sequentially whether this model encompasses other models, using the West test showed

above. Any model that is encompassed is deleted from the list. Third, repeat step 2 but pick the model with the second lowest RMSE if it is still in the list. Next, continue with the model with the third lowest RMSE, and so on, until no encompassed model remains in the list. Last, using several forecast combining methods with all models previously selected to obtains the combined forecast.

Three well-known forecast combination methods are used to generate alternative combined forecasts: RMSE-weighted combinations, rank-weighted combinations and the thick modeling approach. All these methods calculate the combined forecast $\hat{y}_t^c$ as $\hat{y}_t^c = \sum_{i=1}^{m} \omega_i \hat{y}_{it}$, where ω_i is the weight of the combination for model *i*. The definition of ω_i is what distinguishes one combination method from another. For RMSE-weighted combinations, the weight of model *i* is defined as $\hat{\omega}_i = \frac{(1/RMSE_i)}{\sum_{j=1}^{m}(1/RMSE_j)}$. For rank-weighted combinations, the weight of model *i* is calculated as $\hat{\omega}_i = \frac{(1/rank_i)}{\sum_{j=1}^{m}(1/rank_j)}$, where $rank_i$ is the rank of the i^{th} model based on its RMSE. The thick-modeling approach keeps the top α percent of the best performers in the forecast combination, but there is no theoretical guideline on how to choose α. Thus, in this paper, an arithmetic average of all the forecasts from surviving models is used as the combined forecast for this method.

4.4 Data

The DFM and LBVAR models are estimated based on 162 quarterly series, which comprise of 40 house price index series and 122 macroeconomic series. The quarterly house price index figures for the 40 metropolitan areas are obtained from the Federal Housing Finance Agency (FHFA). The data for macroeconomic indicators are taken from the DRI/McGraw Hill Basic Economics Database provided by IHS Global Insight. Each of the series is listed with details in appendix. Data between 1981:Q1 and 2007:Q4 are used for the in-sample estimation, and the data between 2008:Q1 and 2012:Q2 are used for the out-of-sample forecast of the housing price growth of the 40 metropolitan areas in the US. The out-of-sample forecast is done for one to twelve months ahead. With the motivation to examine the U.S. housing market during and after the sub-prime mortgage crisis and to compare the forecasting power of large-scale models for this time period, the choice of 2008:Q1 as the onset of forecast horizon emerges naturally.,

According to Himmerlberg, Mayer, and Sinai (2005), over the 1980-2004 periods, the 40 metropolitan area house prices have followed one of three patterns: (1) house price peaked in the late 1980s, fell to a trough in the 1990s, and rebounded by 2004;(2) a "U" shape history: high in the early 1980s and high again by the end of the sample;(3) house prices have declined since 1980 and have not fully recovered. The 40 metropolitan areas are divided into three groups with each group following one of the three patterns respectively, and they are reported in Table 25.

The number of dynamic factors (q) in the DFM is determined using the criterion proposed by Forni*et al*. (2000). The criterion suggests there should be a substantial gap between the variances explained by the q^{th} and the $(q+1)^{th}$ principal component. A pre-assigned minimum, such as 5%, for the explained variance, could be used as a practical criterion for the determination of the number of dynamic factors to be retained. A 5% limit is suggested by Forni *et al*. (2000) in an empirical exercise.

4.5 Results

The optimal number of dynamic factors is determined to be 10 based on the criterion discussed at the end of last section. The LBVAR model is estimated with 4 lags to account for seasonality. Given the specifications of DFM and LBVAR models, we estimate them over the period of 1981:Q1 to 2007:Q4, and calculate the out-of-sample 1- through 6-quarter-ahead forecasts for the period of 2008:Q1 to 2012Q2. In the standard Minnesota-type prior, the overall tightness (w) takes the values of 0.1, 0.2, 0.3, and the lag decay (d) is generally chosen to be equal to 0.5, 1.0, and 2.0.Thus there are nine LBVAR models estimated, each with a difference combination of w and d. Together with DFM, we have 10 competing models and the forecast performances of these alternative models are compared. The forecast accuracy is measured with RMSEs, and encompassing tests are employed as a complementary measurement. We consider 1- to 2-quarter-ahead forecast as short term forecast, 3- to 4-quarter-ahead forecast as middle term forecast and 5- to 6-quarter-ahead forecast as long term forecast. Discussion of the

forecasting results is carried out according to the prediction terms. Results from RMSE-comparison are discussed first, and results of encompassing test follows.

Table 26 reports the RMSEs of 1- through 6-quarter ahead forecasts from the 10 competing models for metropolitan (metro) areas in the group 1. Figure 12 graphically present the information in the Table 26. There are several interesting findings for this metro group. First, the DFM model underperforms all the LBVAR models throughout the forecast period. Its RMSE in the 1-quarter ahead forecast is 64.85% higher than that of the best model, and this difference in RMSEs widens as the prediction period increases. For example, the RMSE of DFM model is 152.31% higher than the RMSE of the best model for the 6-quarter-ahead forecast. Second, LBVAR(0.1,2.0) model performs the best in the short to middle-term forecast, i.e. 1- to 3-quarter ahead forecast. Many other LBVAR models outperform it in the long-run (4- to 6-quarter ahead), but the differences in their RMSEs is quiet small (0.0504). Third, there is not a single model consistently dominates other models over the forecast period. The performance of the model doing well in the short-term forecast deteriorates gradually as predicting period gets longer. Fourth, a t-test for two sample assuming unequal variances is conducted to see whether the difference between forecasts from DFM and LBVAR models is significantly different from zero. A significant difference is indicated, which conforms to our expectation based on their large difference in RMSE.

Similarly, Table 27 and figure 13 present the results of RMSEs comparison for metro group 2. For this group, the DFM model still underperforms all the LBVAR models, and its performance get worse as prediction period gets longer. Its RMSE in the

1-quarter ahead forecast is 19.40% higher than that of the best model, and its RMSE in the 6-quarter ahead forecast becomes 118.12% higher than that of the best model. The LBVAR(0.1,2.0) model still performs best in the short-run, but LBVAR(0.2,2.0) and LBVAR(0.3,2.0) forecast most accurately in the middle-term and long-term respectively. In the long run, the RMSEs of most LBVAR models converge to 2.135. So their difference is very subtle. T-tests for two sample assuming unequal variances are conducted, and all LBVAR forecasts are not significantly different from each other at 6-quarter-ahead forecast. However, forecast from DFM model is significantly different from the forecast from the best model no matter in the short-, or middle- or long-term.

Table 28 and figure 14 show the resulting RMSEs for metro group 3. The findings for this group differ somehow to the findings for the other two groups. First, the forecast from DFM model is not significantly different from the forecasts from LBVAR(0.1,0.5), LBVAR(0.1,1.0), and LBVAR(0.1,2.0), which are the models with relatively small RMSEs. However, forecast performance of DFM model again deteriorates fast as predicting period gets longer. Second, in the group of LBVAR models, LBVAR(0.3,0.5) underperforms others in the short- and middle-run, but it performs well in the long-run. Third, the forecasts from LBVAR for this group are closer to each other than for the other two groups. However, similar to the other two groups, there is no one model consistently dominates other models in either short-term forecast or long-term forecast. Next, the results of encompassing test are discussed.

The results of encompassing tests for metro group 1 are presented in Table 29. For this group, the DFM model is not encompassed by other models only in the 3- and 5-

quarter ahead forecast, indicating that it contains marginal information that LBVAR models do not have for these forecasts. However, the benefit of marginal information from the DFM models is eliminated by the additional errors brought by it, because the combined forecasts do not have lower RMSEs than individual forecasts. In the 6-quarter ahead forecast, LBVAR(0.2,2.0) dominates all the other models, so combined forecasts are the same as the forecast from LBVAR(0.2,2.0).

Table 30 reports the results of encompassing tests for metro group 2. First, in the short- to middle-term prediction (2- to 4-quarter ahead forecast), DFM is not encompassed by any other models and contains additional information to LBVAR models. But, again, the accuracy gained by adding marginal information from DFM model to LBVAR models is offset by the larger error introduced by the DFM model. So, combined forecasts fail to outperform individual forecast throughout the predicting period.

Table 31 shows the results of encompassing tests for metro group 3. From the table, we can see that the DFM model is one of the dominant models in the short-term and middle-term forecast. The combined forecast for the 1-quarter ahead forecast outperforms all the individual models. However, the combined forecasts for other predicting periods underperform individual models due to the larger errors introduced by multiple models when combining them together.

After examining the results from RMSE calculation and encompassing test separately, we now summarize the major findings with the information from both sides. For metro group 1, the DFM model underperforms all the LBVAR models throughout

the forecast period. The difference in RMSE of the DFM model and the RMSE of the best model widens as the prediction period increases. LBVAR(0.1,2.0) model performs the best in the short to middle-term forecast. Some other LBVAR models outperform it in the long run, but the difference in their RMSE is small. No single model consistently dominates other models over the forecast period, and the performance of the model doing well in the short-term forecast deteriorates gradually as predicting period gets longer. Moreover, t-test for two sample assuming unequal variances is conducted and suggests that the difference between forecasts from DFM and LBVAR models is significantly different from zero. The DFM model contains marginal information that LBVAR models do not have in the 3- and 5-quarter ahead forecast. However, the benefit from marginal information is eliminated by the additional errors brought at the same time. So, the combined forecasts do not have lower RMSEs.

For metro group 2, the DFM model still underperforms all the LBVAR models, and its performance get worse as prediction period get longer. The LBVAR(0.1,2.0) model still performs the best in the short run, but LBVAR(0.2,2.0) and LBVAR(0.3,2.0) forecast most accurately in the middle-term and long-term respectively. In the long run, the RMSEs of most LBVAR models converge, so their difference is very subtle. T-tests also prove that all LBVAR forecasts are not significantly different from each other at 6-quarter-ahead forecast. However, forecast from DFM model is significantly different from the forecast from the best model no matter in the short-, or middle- or long-term. In the short- to middle-term prediction (2- to 4-quarter ahead forecast), DFM contains additional information to LBVAR models, but the accuracy gained by adding marginal

information from DFM model to LBVAR models is offset by the larger error introduced by the DFM model. So, combined forecasts fail to outperform individual forecast throughout the prediction period.

For metro group 3, the forecast from the DFM model is not significantly different from the forecasts from LBVAR(0.1,0.5), LBVAR(0.1,1.0), and LBVAR(0.1,2.0), which are the models with relatively small RMSEs. But, the forecast performance of DFM model again deteriorates fast as predicting period gets longer. The forecasts from LBVAR for this group are closer to each other than for the other two groups. However, similar to the other two groups, there is no one model consistently dominates other models in either short-term forecast or long-term forecast. The DFM model is one of the dominant models in the short-term and middle-term forecast. The combined forecast for the 1-quarter ahead forecast outperforms all the individual models. However, the combined forecasts for other predicting periods underperform individual models due to the larger errors introduced by multiple models when combining them together.

4.6 Conclusion

This essay discuss the model selection for analyzing housing prices in 40 metropolitan areas in the United State from the perspective of large-scale models, which are Dynamic Factor Model (DFM) and Large-scale Bayesian Vector Autoregressive (LBVAR) model. These models accommodate a large panel data comprising 162 quarterly series for the U.S. economy, and an in-sample period of 1980:Q1 to 2007:Q4

are used to forecast 1- to 6-quarters-ahead house price growth rate over the out-of-sample horizon of 2008:Q1 to 2012:Q2. The 40 metropolitan areas can be divided into three groups based on their house price moving patterns. The forecast evaluation for the two large-scale models is based on two complementary criteria: RMSE and encompassing test.

Examining both the RMSE measures and the results from encompassing tests, we have several interesting findings. First, the DFM model underperform the LBVAR models most of the time in all three groups, and its forecasting power deteriorates fast as predicting period gets longer. For example, the difference in RMSE of the DFM model and the RMSE of the best model widens as the prediction period increases. T-tests suggest that the differences are significantly different from zero. Second, there is not a single model consistently outperform other models over the whole prediction period. The model forecasting better in the short run performs worse in the long run. However, the forecasts from LBVAR models converge in the long run, and t-test suggests that they are not significantly different from each other. Third, the DFM model is not encompassed by other models in the short-term and long-term. However, the accuracy gained by adding marginal information from DFM model to LBVAR models is offset by the larger error introduced by the DFM model. So, combined forecasts fail to outperform individual forecast. The only exception is the 1-quarter ahead forecast for group 3.

Overall, our study lends support to the superior performance of the LBVAR model compares to DFM model throughout the prediction period. Also, our study suggests that combined forecasts are not necessarily outperform individual forecasts.

Even though independent information from different individual models improves the forecast accuracy, the benefit gained from marginal information is offset by the larger error brought by such combination.

Although DFM has its advantage in the long-term forecast, there are two caveats in its application. First, if there are structural changes of the economy, both the in-the-sample forecast and the out of the sample forecast would be inaccurate. In this essay, Chow tests are conducted to evaluate the stability of the estimated coefficients, and test statistics indicate that structure changes at the end of 2007 exist for half of the variables. However, the data size is too small to conduct estimation for period after 2007. So, the estimation in this easy is the best we can do. In future, when more data are available, an analysis with new data set should be conducted to update the results and to mitigate estimation bias due to data unavailability. Second, the estimation procedures used are linear in nature, and hence, they fail to take into account of the nonlinearities in the data (Das, Gupta, and Kabundi, 2009a). Meanwhile, LBVAR model has two major limitations. First, the forecast accuracy is sensitive to the choice of the priors. So if the prior is not well specified, an alternative model used for forecasting may perform better. Secondly, the selection of the prior based on some objective function for the out-of-sample forecasts may not be 'optimal' for the time period beyond the period chosen to produce the out-of-sample forecasts (Das, Gupta, and Kabundi, 2008, 2009a)

5 SUMMARY

The housing market plays a significant role in shaping the economic and social well-being of U.S. households. It helps spur U.S. economic growth when house price rises, and drags the economic growth when house price drops. In this dissertation, we conduct analysis to project where the U.S. housing market is headed and to discover how it interacts with economic fundamentals. New pieces of information are found, which are deemed to facilitate decision making for both policy makers and investors.

In the first part of the dissertation, the grouping patterns of U.S. housing markets are studied using cluster and discriminant analysis. Three clusters are found, which are located in central US, east coast and west coast of US. There is no price signal transmitted among these housing market clusters, nor within each cluster. Thus, the communication of information in housing market is through the process of utility convergence of marginal residents, and no price convergence across regions is found in this process.

Next, the impact of credit constraint on the house price is examined with stochastic components of series considered. Both a simulation technique and a DAG approach are employed. The resulting causal pattern shows that credit constraint affects the house price directly and positively. Moreover, credit constraints work as an intermediary passing the influence of house inventor, household income, and user cost onto house price, which suggest credit relaxation policy be carried out with caution when house inventory and household income send inconsistent signals.

Last, the model selection for house price analysis is discussed from the perspective of large-scale models—dynamic factor (DFM) model and large-scale Bayesian VAR (BVAR) model. The LBVAR models are found to have superior performance compare to the DFM model throughout the prediction period. Also, it is found that the combined forecasts do not necessarily outperform individual forecasts. Even though independent information from different individual models improves the forecast accuracy, the benefit gained from marginal information is offset by the larger error brought by such combination.

REFERENCES

Abraham, J.M., W.N. Goetzmann and S.M. Wachter. “Homogeneous Groupings of Metropolitan Housing Markets”, *Journal of Housing Economics* 3(September 1994):186–206.

Abraham, J.M., and P.H. Hendershott. “Bubbles in Metropolitan Housing Markets”, *Journal of Housing Research* 7(1996):191-206.

Apergis, N., B.D. Simo-Kengne, and R. Gupta. “Convergence in Provincial-Level South African House Prices: Evidence from the Club Convergence and Clustering Procedure.” Working paper No. 2013-22, Dept. of Econ., University of Pretoria, South Africa, 2013.

Ariccia, G.D., D. Igan, and L. Laeven. “Credit Booms and Lending Standards: Evidence from the Subprime Mortgage Market.” IMF working paper No. WP/08/106, April 2008.

Bessler, D.A. and D.G. Akleman. "Farm Prices, Retail Prices, and Directed Graphs: Results for Pork and Beef." *American Journal of Agricultural Economics* 80(December 1998):1144-49.

Bessler, D.A., J. Yang, and M. Wongcharupan. "Price Dynamics in the International Wheat Market: Modeling with Error Correction and Directed Acyclic Graphs," *Journal of Regional Science* 43(February 2003): 1-33.

Bessler, D.A. *On World Poverty: Its Causes and Effects*. Rome, Italy: Food and Agricultural Organization of the United Nations, Research Bulletin, 2003.

Bessler, D.A. “On Modeling Environmental and Agricultural Interfaces with Directed Acyclic Graphs.” Comments for OECD Conference, Paris France, June 30-July, 2005.

Bizimana, J., J.P. Angerer, and D.A. Bessler. "Cattle Markets Integration and Price Discovery in Three Developing Countries of Mali, Kenya, and Tanzania," Paper presented at AAEA annual meeting, Seattle, Washington, August 12-14, 2012.

Bloor, C., and T. Matheson. “Analyzing Shock Transmission in a Data-Rich Environment: A Large BVAR for New Zealand.” *Empirical Economics* 39(2010): 537-558.

Bourassa, S.C., E. Cantoni, and M. Hoesli. “Predicting House Prices with Spatial Dependence: Impacts of Alternative Submarket Definitions.” Research paper No. 08-01, Swiss Finance Institute, Swiss, 2008.

Capozza D.R., P.H. Hendershott, C. Mack, and C.J. Mayer. “Determinants of Real House Price Dynamics.” Working paper 9262, National Bureau of Economic Research, Cambridge, 2002.

Case, B., J. Clapp, R. Dubin, and M. Rodrigues. “Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models.” *Journal of Real Estate Finance Economics* 29(September 2004):167-191.

Chan, L., H.T. Ng, and R. Ramchand. “A Cluster Analysis Approach to Examining Singapore’s Property Market.” *Property Markets and Financial Stability*. BIS Papers, Bank for International Settlements 64(March 2012):43-53.

Chicago Metropolitan Agency for Planning. “Industry Cluster Analysis: Regional Economic Base Analysis.” Technical Document, Chicago, 2009.

Clark, S.P., and T.D. Coggin. “Trends, Cycles and Convergence in U.S. Regional House Prices.” *The Journal of Real Estate Finance and Economics* 39(May 2009):264-283.

Clements, A.M., H.P. Mapp, and V.R. Eidman. “A Procedure for Correlating Event in Farm Firm Simulation Models.” Oklahoma Agricultural Experiment Station, Technical Bulletin No. T-131, August 1971.

Costantini, M., and C. Pappalardo. “Combiantion of Forecast Methods Using Encompassing Tests.” Economic Series 228, Institute for Advanced Studies, 2008.

Cunnuingham, W.V., and W.F. Maloney. “Heterogeneity among Mexico’s Microenterprises: An Application of Factor and Cluster Analysis.” *Economic Development and Cultural Change* 50(October 2001):131-156.

Das, S., R. Gupta, and A. Kabundi. “Is a DFM Well-Suited in Forecasting Regional House Price Inflation?” Working paper No. 200814, Dept. of Econ., University of Pretonia, 2008.

Das, S., R. Gupta, and A. Kabundi. “Could We Have Predicted the Recent Downturn in the South African Housing Market?” *Journal of Housing Economics* 4(2009a):325-335.

Das, S., R. Gupta, and A. Kabundi. “The Blessing of Dimensionality in Forecasting Real House Price Growth in the Nine Census Divisions of the US.” Working paper No. 200902, Dept. of Econ., University of Pretonia, 2009b.

Das, S., R. Gupta, and A. Kabundi. “Forecasting Regional House Price Inflation: A Comparison between Dynamic Factor Models and Vector Autoregressive Models.” *Journal of Forecasting* 30(2011): 288-302.

Doan, T.A., R.B. Litterman, and C.A. Sims. “Forecasting and Conditional Projection Using Realistic Prior Distributions.” *Econometric Reviews* 3(1984): 1-100.

DRI BASIC Economics. Macroeconomic Database, machine-readable data file. 1946 - present. Lexinton, MA, DRI/McGraw-Hill, 1996.

Duca, J.V., J. Muellbauer, and A. Murphy. “Housing Prices and Credit Constraints: Making Sense of the U.S. Experience.” Federal Reserve Bank of Dallas, working paper No. 1103, April 2011(a).

Duca, J.V., J. Muellbauer, and A. Murphy. “Shifting Credit Standards and the Boom and Bust in U.S. House Prices.” Federal Reserve Bank of Dallas, working paper No. 1104, April 2011(b).

Dua, P., and S.C. Ray. “A BVAR Model for the Connecticut Economy.” *Journal of Forecasting* 14(1995):167-180.

Embrechts, P., A. McNeil, and D. Straunmann, “Correlation: Pitfalls and Alternatives.” A short, non-technical article, *RISK Magazine* (May 1999): 69-71.

Engle, R.F., and C.W.J. Granger. *Long Run Economic Relationships.* Oxford: Oxford University Press, 1991.

Ericsson, N.R. “Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration.” *Journal of Policy Modeling* 14(1992): 465-495.

Fang, Y. “Forecasting Combination and Encompassing Tests.” *International Journal of Forecasting* 19(2003): 87-94.

Favara, G., and Z. Song. “House Price Dynamics with Dispersed Information.” *Journal of Economic Theory* 148(May 2013): in press.

Fik, T.J., D.C. Ling, and G.F. Mulligan. “Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach.” *Real Estate Economics* 31(November 2003):623-646.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. “The Generalized Dynamic-Factor Model: Identification and Estimation.” *The Review of Economics and Statistics* 4(2000): 540-554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. “Do financial variables help forecasting inflation and real activity in the euro area?” *Journal of Monetary Economics* 6(2003): 1243-1255.

Forni, M., M. Hallin, M. Lippi, and L.Reichlin. “The Generalized Dynamic Factor Model, One Sided Estimation and Forecasting.” *Journal of the American Statistical Association* 100(2005): 830-840.

Gallin, J. “The Long-Run Relationship between Home Prices and Income: Evidence from Local Housing Markets.” *Real Estate Economics* 34(2006): 417-38.

Goetzmann, W.N., and S.M. Wachter. “Clustering Methods for Real Estate Portfolios”, *Real Estate Economics* 23(Fall 1995):271–310.

Goetzmann, W. N., and S. M. Wachter. “The Global Real Estate Crash: Evidence from an International Database.” *A Global Perspective on Real Estate Cycles*. S.J. Brown and C.H. Liu, eds. Boston: Kluwer Academic Publishers, 1995b.

Gordon, A. *Classification*, 2nd ed. London, UK: Chapman and Hall/CRC Press, 1999.

Gupta, M.C., and R.J. Huefner. “A Cluster Analysis Study of Financial Ratios and Industry Characteristics.” *Journal of Accounting Research* 10(Spring 1972):77-95.

Gupta, R., A. Kabundi. “Forecasting Macroeconomic Variables Using Large Datasets: Dynamic Factor Model versus Large-Scale BVARs” Working paper No. 200816, Dept. of Econ., University of Pretonia, 2008a.

Gupta, R., A. Kabundi. “A Dynamic Factor Model for Forecasting Macroeconomic Variables in South Africa.” Working Paper No. 200815, Dept. of Econ., University of Pretonia, 2008b.

Gupta, R., A. Kabundi, and S.M. Miller. “Forecasting the US Real House Price Index: Structural and Non-Structural Models with and without Fundamentals.” Working paper No. 200927, Dept. of Econ., University of Pretonia, 2009a.

Gupta, R., A. Kabundi, and S.M. Miller. “Using Large Data Sets to Forecast Housing Prices: A Case Study of Twenty US States.” Working paper No. 200905, Dept. of Econ., University of Pretonia, 2009b.

Gupta, R., S.M. Miller, D.V. Wyk. “Financial Market Liberalization, Monetary Policy, and Housing Price Dynamics.” Working paper No. 201009, Dept. of Econ., University of Pretonia, 2010.

Gyourko, J., and R.Voith. “Local Market and National Components in House Price Appreciation.” *Journal of Urban Economics* 32(July 1992):52-69.

Gyourko, J., C. Mayer, and T. Sinai. “Superstar Cities.” Working paper No. 12355, National Bureau of Economic Research, Cambridge, MA, 2006.

Hansen, P., and B. Jaumard. “Cluster Analysis and Mathematical Programming.” *Mathematical Programming* 79(1997):191-215.

Hepsen, A., and M. Vatansever. “Using Hierarchical Clustering Algorithms for Turkey Residential Market.” *International Journal of Economics and Finance* 4(January 2012):138-150.

Hiebert, P., and M. Roma. “Relative House Price Dynamics across Euro Area and US Cities Convergence or Divergence?” Working paper No. 1206, European Central Bank, Washington DC, 2010.

Himmelberg, C., C. Mayer, and T. Sinai. “Assessing High House Prices: Bubbles, Fundamentals and Misperceptions.” *Journal of Economic Perspectives* 19(Fall 2005):67-92.

Hirata, H., M.A. Kose, C. Otrok, and M.E. Terrones. “Global House Price Fluctuations: Synchronization and Determinants.” Working paper, International Monetary Fund, Washington DC, 2013.

Hoesli, M., C. Lizieri, and B.D. Macgregor. “The Spatial Dimensions of the Investment Performance of UK Commercial Property.” *Urban Studies* 34(August 1997):1475–1494.

Holmes, M.J., J. Otero, and T. Panagiotidis. “Investigating Regional House Price Convergence in the United States: Evidence from a Pair-wise Approach.” Working paper, The Rimini Center for Economic Analysis, Italy, 2011.

Jain, A., and R. Dubes. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

Jud, G.D., and D.T. Winkler. “The Dynamics of Metropolitan Housing Prices.” *The Journal of Real Estate Research* 23(January/April 2002): 29-42.

Jolliffe, I.T. (2002). Principal Component Analysis. New York: Springer-Verlag.

Kim, J.W., D.J. Leatham, and D.A. Bessler “REITs’ Dynamic under Structural Change with Unknown Break Points.” *Journal of Housing Economics* 16(April 2007):37-58.

Kim, K., and J. Park. “Segementeation of the Housing Market and Its Determinants: Seoul and Its Neighboring New Towns in Korea.” *Australian Geographer* 32(2005):221-232.

Kim, Y.S., and J.J. Rous. “House Price Convergence: Evidence from US State and Metropolitan Area Panels.” *Journal of Housing Economics* 21(February 2012):169-186.

Kisinbay, T. "The Use of Encompassing Tests for Forecast Combinations," IMF Working Papers 07/264, International Monetary Fund, 2007.

Kwon, D. and D. Bessler. “Graphical Methods, Inductive Causal Inference, and Econometrics: A Literature Review.” *Computational Economics* 38(2011):85-106.

Leung, F., K. Chow, and G. Han. “Long-term and Short-term Determinants of Property Prices in Hong Kong.” Working paper No. 0815, Hong Kong Monetary Authority, Hong Kong, 2008.

Linneman, P. “An Empirical Test of the Efficiency of the Housing Market.” *Journal of Urban Economics* 20(1986): 140-154.

Litterman, R.B. “Forecasting with Bayesian Vector Autoregression- Five Years of Experience.” *Journal of Business and Economic Statistics* 4(1986):25-38.

Litterman, R.B. “A Bayesian Procedure for Forecasting with Vector Autoregressions.” Working paper, Federal Reserve Bank of Minneapolis, 1981.

Lu, Y. “Analysis of Residential Housing Markets in Large U.S. Metropolitan Areas.” Ph.D. dissertation, University of Missouri, Columbia, December 2009.

Magne, F.O., and S. Rady. “Housing Market Dynamics: On the Contribution of Income Shocks and Credit Constraints.” *Review of Economic Studies* 73(2006):459-485.

McGreal, S. and P.T. De La Paz. “Implicit House Prices: Variation over Time and Space in Spain.” *Urban Studies* 50(February 2013):1-20.

Newbold, P., and D.I. Harvey. “Forecasting combination and encompassing”. *A Companion to Economic Forecasting.* M.P. Clements and D.F. Hendry, eds. Oxford: Blackwells, 2002.

Pearl, J. *Causality*. Cambridge, UK: Cambridge University Press, March, 2000.

Quigley, J.M. “Real Estate Prices and Economic Cycles.” *International Real Estate Reviews* 2(1999): 1-20.

Oxley L., M. Reale, and G.T. Wilson. “Constructing Structural VAR Models with Conditional Independence Graphs.” *Mathematics and Computers in Simulation* 79 (2009): 2910-2916.

Rapach, D.E., and J.K. Strauss. “Difference in Housing Price Forecast Ability across U.S. States.” *International journal of Forecasting* 25(2009): 351-372.

Richardson, J.W., and G.D. Condra. “A General Procedure for Correlating Events in Simulation Models.” Working paper, Department of Agricultural Economics, Texas A&M University, 1978.

Richardson, J.W. “Simulation for Applied Risk Management with an Introduction to SIMETAR.” Working paper, Department of Agricultural Economics, Texas A&M University, 2010.

Roback, J. “Wages, Rents, and the Quality of Life.” *Journal of Political Economy* 90(December 1982):1257-1278.

Rosen, S. *Wage-based Indexes of Urban Quality of Life*. Mieszkowski, P., Straszheim, M. eds. Baltimore, U.S.: Johns Hopkins University Press, 1979:74-104.

San Diego Association of Governments. “Understanding Cluster Analysis.” 2002.

Shimizu, C., and T. Watanable. “Housing Bubbles in Japan and the United States.” *Public Policy Review* 6(March 2010):431-472.

Smith, E.B. “‘Race to Bottom’ at Moody’s, S&P Secured Subprime’s Boom, Bust.” *Bloomberg*, September 2008.

Spencer, D.E. “Developing a Bayesian Vector Autoregression Forecasting Model.” *International Journal of Forecasting* 9(1993): 407-421.

Sprites, P., C. Glymour, and R. Scheines. *Causation, Prediction, and Search*, 2nd. ed. Boston: MIT Press, 2000.

Stock, J.H., and M.W. Watson. “Macroeconomic Forecasting Using Diffusion Indexes.” *Journal of Business and Economic Statistics* 2(2002): 147-162.

Stock, J.H., and M.W. Watson. “Forecasting Output and Inflation: The Role of Asset Prices.” *Journal of Economic Literature* 3(2003):788-829.

Sutton, G.D. “Explaining Changes in House Prices.” *BIS Quarterly Review* (September 2002):46-55.

Todd, R.M. “Improving Economic Forecasting with Bayesian Vector Autoregression.” *Quarterly Revie*w (Federal Reserve Bank of Minneapolis) (Fall, 1984): 18-29.

Tsatasaronis, K., and H. Zhu. “What Drives Housing Price Dynamics: Cross-Country Evidence.” *BIS Quarterly Review* March, 2004:65-78.

U.S. Department of Housing and Urban Development. 2011 *American Housing Survey (AHS)* Washington DC, 2012.

Vitale, J., and D.A. Bessler. “On the Discovery of Millet Prices in Mali.” *Papers in Regional Science* 85(March 2006):139-162.

Wang, Z., and D.A. Bessler. “Forecasting Performance of Multivariate Time Series Models with Full and Reduced Rank: An Empirical Examination.” *International Journal of Forecasting* 20(2004): 683-695.

Wang, Z., and D.A. Bessler. “A Monte Carlo Study on the Selection of Cointegrating Rank Using Information Criteria.” *Economic Theory* 21(2005): 593-620.

West, K.D. "Tests For Forecast Encompassing When Forecasts Depend On Estimated Regression." *Journal of Business and Economic Statistics* 19(January 2001): 29-33.

Wheaton, W.C. “Real Estate ‘Cycle’: Some Fundamentals.” *Real Estate Economics* 27 (1999): 209-230.

Woodard, J.D., N. Paulson, D. Vedenov, and G. Power. "Efficiency in the Modeling of Dependence Structures: An Application of Alternative Copulas to Agricultural Insurance Rating," *Agricultural Economics* 42-IS1 (November 2011): 101-112.

Xu, R., and D.C. Wunsch. *Cluster*. New Jersey: IEEE Press, 2009.

Yang, Y., and A. Hu. “Investigating Regional Disparities of China’s Human Development with Cluster Analysis: A Historical Perceptive.” *Social Indicators Research* 86(2008): 417-432.

APPENDIX

In the chapter 4, 122 data series are used to compare the forecasting performance between DFM and LBVAR models. These data series are taken directly from DRI/McGraw Hill Basic Economics Database. Each of the series is listed with following details: series number; series mnemonic in the database; data span; transformation code and series description in the database. Format follows Stock and Watson (2002) paper. The transformation codes are: 1- no transformation; 2- first difference; 4- logarithm; 5- first difference of logarithm.

	OUTPUT ---------- real output and income			
1	IPS11.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL
2	IPS299.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS
3	IPS12.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS
4	IPS13.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS
5	IPS18.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS
6	IPS25.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT
7	IPS32.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - MATERIALS
8	IPS34.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS
9	IPS43.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)
10	IPS67.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - MINING NAICS=21
11	IPS68.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - ELECTRIC AND GAS UTILITIES
12	IPS10.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX
13	IPS307.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES
14	IPS316.M	1980:1 - 2012:6	5	INDUSTRIAL PRODUCTION INDEX - BASIC METALS
15	PMI.M	1980:1 - 2012:6	5	PURCHASING MANAGERS' INDEX (SA)
16	PMP.M	1980:1 - 2012:6	5	NAPM PRODUCTION INDEX (PERCENT)
17	YPR.M	1980:1 - 2012:6	5	PERS INCOME CH 2005 $,SA-US
18	YP@V00C.M	1980:1 - 2012:6	5	PERS INCOME LESS TRSF PMT CH 2005 $,SA-US

EMP ---------- employment and hours

19	LHEM.M	1980:1 - 2012:6	4	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)
20	LHNAG.M	1980:1 - 2012:6	5	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)
21	LHUR.M	1980:1 - 2012:6	5	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)
22	LHU680.M	1980:1 - 2012:6	1	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)
23	LHU5.M	1980:1 - 2012:6	1	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)
24	LHU14.M	1980:1 - 2012:6	1	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)
25	LHU15.M	1980:1 - 2012:6	1	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)
26	LHU26.M	1980:1 - 2012:6	1	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)
27	CES0000000001.M	1980:1 - 2012:6	5	TOTAL NONFARM EMPLOYMENT.SA
28	CES0500000001.M	1980:1 - 2012:6	5	TOTAL PRIVATE EMPLOYMENT.SA
29	CES0600000001.M	1980:1 - 2012:6	5	GOODS PRODUCING EMPLOYMENT.SA
30	CES1000000001.M	1980:1 - 2012:6	5	MINING AND LOGGING EMPLOYMENT.SA
31	CES2000000001.M	1980:1 - 2012:6	5	CONSTRUCTION EMPLOYMENT.SA
32	CES3000000001.M	1980:1 - 2012:6	5	MANUFACTURING EMPLOYMENT.SA
33	CES3100000001.M	1980:1 - 2012:6	5	DURABLE GOODS MANUFACTURING EMPLOYMENT.SA
34	CES3200000001.M	1980:1 - 2012:6	5	NONDURABLE GOODS MANUFACTURING EMPLOYMENT.SA
35	CES0700000001.M	1980:1 - 2012:6	5	SERVICE PROVIDING EMPLOYMENT.SA
36	CES4000000001.M	1980:1 - 2012:6	5	TRADE,TRANSPORTATION, AND UTILITY EMPLOYMENT.SA
37	CES4200000001.M	1980:1 - 2012:6	5	RETAIL TRADE EMPLOYMENT.SA
38	CES4142000001.M	1980:1 - 2012:6	5	WHOLESALE TRADE EMPLOYMENT.SA
39	CES5500000001.M	1980:1 - 2012:6	5	FINANCIAL ACTIVITIES EMPLOYMENT.SA
40	CES0800000001.M	1980:1 - 2012:6	5	PRIVATE SERVICE PROVIDING EMPLOYMENT.SA
41	CES9000000001.M	1980:1 - 2012:6	5	GOVERNMENT EMPLOYMENT.SA
42	CES3000000009.M	1980:1 - 2012:6	1	AVG WEEKLY OT,PROD WORKERS: MFG,SA-US
43	PMEMP.M	1980:1 - 2012:6	1	NAPM EMPLOYMENT INDEX (PERCENT)

HSS ---------- housing starts and sales

44	HSFR.M	1980:1 - 2012:6	4	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA)
45	HSNE.M	1980:1 - 2012:6	4	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.
46	HSMW.M	1980:1 - 2012:6	4	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.
47	HSSOU.M	1980:1 - 2012:6	4	HOUSING STARTS:SOUTH (THOUS.U.)S.A.
48	HSWST.M	1980:1 - 2012:6	4	HOUSING STARTS:WEST (THOUS.U.)S.A.
49	HSBR.M	1980:1 - 2012:6	4	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)
50	HMOB.M	1980:1 - 2012:6	4	MOBILE HOMES: MANUFACTURERS' SHIPMENTS (THOUS.OF UNITS,SAAR)

INV ---------- real inventories and inventory-sales ratios

51	PMNV.M	1980:1 - 2012:6	1	NAPM INVENTORIES INDEX (PERCENT)
	ORD ---------- orders and unfilled orders			
52	PMNO.M	1980:1 - 2012:6	1	NAPM NEW ORDERS INDEX (PERCENT)
53	PMDEL.M	1980:1 - 2012:6	1	NAPM VENDOR DELIVERIES INDEX (PERCENT)
54	MOCMQ.M	1980:1 - 2012:6	5	NEW ORDERS (NET) - CONSUMER GOODS & MATERIALS, 1996 DOLLARS (BCI)
55	MSONDQ.M	1980:1 - 2012:6	5	NEW ORDERS, NONDEFENSE CAPITAL GOODS, IN 1996 DOLLARS (BCI)
	SPR ---------- stock prices			
56	FSPCOM.M	1980:1 - 2012:6	5	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)
57	FSPIN.M	1980:1 - 2012:6	5	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)
58	FSDJ.M	1980:1 - 2012:6	5	COMMON STOCK PRICES: DOW JONES INDUSTRIAL AVERAGE
	EXR ---------- exchange rates			
59	EXRSW.M	1980:1 - 2012:6	5	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)
60	EXRJAN.M	1980:1 - 2012:6	5	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)
61	EXRUK.M	1980:1 - 2012:6	5	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)
62	EXRCAN.M	1980:1 - 2012:6	5	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)
	INT ---------- interest rates			
63	FYFF.M	1980:1 - 2012:6	1	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)
64	FYGM3.M	1980:1 - 2012:6	1	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)
65	FYGM6.M	1980:1 - 2012:6	1	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)
66	FYGT1.M	1980:1 - 2012:6	1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)
67	FYGT5.M	1980:1 - 2012:6	1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)
68	FYGT10.M	1980:1 - 2012:6	1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)
69	RMMBCAAANS.M	1980:1 - 2012:6	1	YIELD ON MOODY'S AAA CORP BONDS-US
70	RMMBCBAANS.M	1980:1 - 2012:6	1	YIELD ON MOODY'S BAA CORP BONDS-US
71	SFYGM3	1980:1 - 2012:6	1	SPREAD FYGM3-FYFF
72	SFYGM6	1980:1 - 2012:6	1	SPREAD FYGM6-FYFF
73	SFYGT1	1980:1 - 2012:6	1	SPREAD FYGT1-FYFF
74	SFYGT5	1980:1 - 2012:6	1	SPREAD FYGT5-FYFF
75	SFYGT10	1980:1 - 2012:6	1	SPREAD FYGT10-FYFF
76	SBCAAA	1980:1 - 2012:6	1	SPREAD BCAAA-FYFF
77	SBCBAA	1980:1 - 2012:6	1	SPREAD BCBAA-FYFF
	MON ---------- money and credit quantity aggregates			
78	FM1.M	1980:1 - 2012:6	5	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)

79	FM2.M	1980:1 - 2012:6	5	MONEY STOCK: M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP), BIL$
80	FMNC2.M	1980:1 - 2012:6	5	MONEY STOCK: NONTRANSACTION COMPONENTS IN M2 (BIL$,SA)
81	MNY2@00.M	1980:1 - 2012:6	5	MONEY SUPPL-M2 IN 2005 $,SA-US
82	FMFBA.M	1980:1 - 2012:6	5	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)
83	FMRRA.M	1980:1 - 2012:6	5	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)
84	FCLBMC.M	1980:1 - 2012:6	1	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)
85	CCINRV.M	1980:1 - 2012:6	5	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)
86	ALCIBL00.M	1980:1 - 2012:6	5	COML&IND LOANS OUTST IN 2000 $,SA-US
	PRI ---------- price indexes			
87	PMCP.M	1980:1 - 2012:6	1	NAPM COMMODITY PRICES INDEX (PERCENT)
88	PWFSA.M	1980:1 - 2012:6	5	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)
89	PWFCSA.M	1980:1 - 2012:6	5	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)
90	PWIMSA.M	1980:1 - 2012:6	5	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)
91	PWCMSA.M	1980:1 - 2012:6	5	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)
92	PUNEW.M	1980:1 - 2012:6	5	CPI-U: ALL ITEMS (82-84=100,SA)
93	PU83.M	1980:1 - 2012:6	5	CPI-U: APPAREL & UPKEEP (82-84=100,SA)
94	PU84.M	1980:1 - 2012:6	5	CPI-U: TRANSPORTATION (82-84=100,SA)
95	PU85.M	1980:1 - 2012:6	5	CPI-U: MEDICAL CARE (82-84=100,SA)
96	PUC.M	1980:1 - 2012:6	5	CPI-U: COMMODITIES (82-84=100,SA)
97	PUCD.M	1980:1 - 2012:6	5	CPI-U: DURABLES (82-84=100,SA)
98	PUXF.M	1980:1 - 2012:6	5	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)
99	PUXHS.M	1980:1 - 2012:6	5	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)
100	PUXM.M	1980:1 - 2012:6	5	CPI-U: ALL ITEMS LESS MIDICAL CARE (82-84=100,SA)
101	PUH.M	1980:1 - 2012:6	5	CPI-U: HOUSING (82-84=100,SA)
102	PU803.M	1980:1 - 2012:6	5	CPI-U:ENERGY (82-84=100,SA)
	AHE ---------- average hourly earnings			
103	CES0500000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: TOTAL PRIV,SA-US
104	CES4422000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: UTILITIES,SA-US
105	CES4300000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: TRNSPRT&WHSE,SA-US
106	CES4000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: TRADE,TRNSPRT,&UTILITIES,SA-US
107	CES4200000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: RETAIL TRADE,SA-US
108	CES6000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: PROF&BUS SVC,SA-US
109	CES0800000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: PRIV SVC,SA-US
110	CES8000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: OTH SVC,SA-US
111	CES3200000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: NON-DUR,SA-US
112	CES1000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: MINING&LOGGING,SA-US

113	CES3000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: MFG,SA-US
114	CES7000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: LEIS&HOS,SA-US
115	CES5000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: INFO,SA-US
116	CES0600000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: GDS PRODUCING,SA-US
117	CES5500000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: FIN ACT,SA-US
118	CES6500000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: ED&HEALTH SVC,SA-US
119	CES3100000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: DUR,SA-US
120	CES2000000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: CONSTR,SA-US
121	CES4142000030.M	1980:1 - 2012:6	5	AWE,PROD WORKERS: WSALE,SA-US
	OTH ---------- miscellaneous			
122	U0M083.M	1980:1 - 2012:6	1	BUSINESS CYCLE INDICATORS, CONSUMER EXPECTATIONS,NSA-US (COPYRIGHT,UINV. OF MICHIGAN)

Figure 1 The Directed Acyclic Graph for the Simple Example

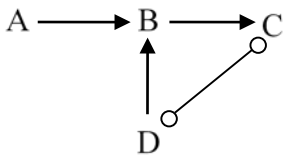
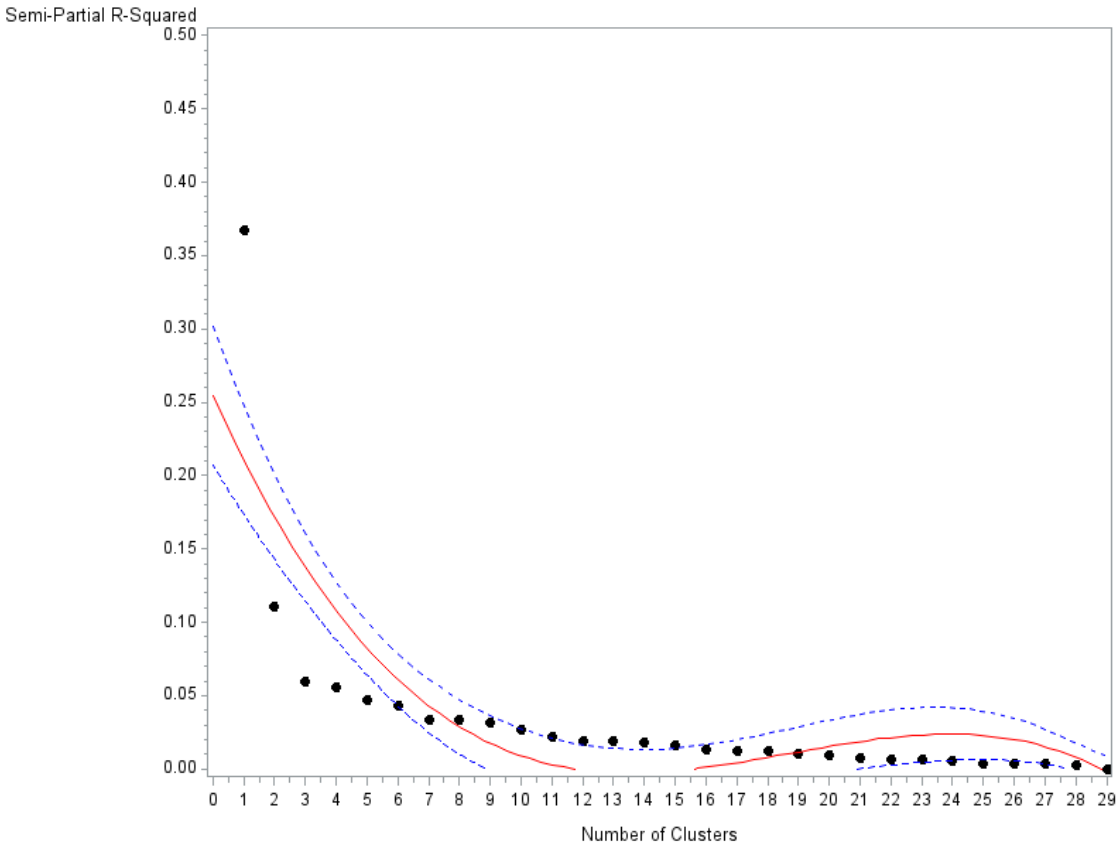
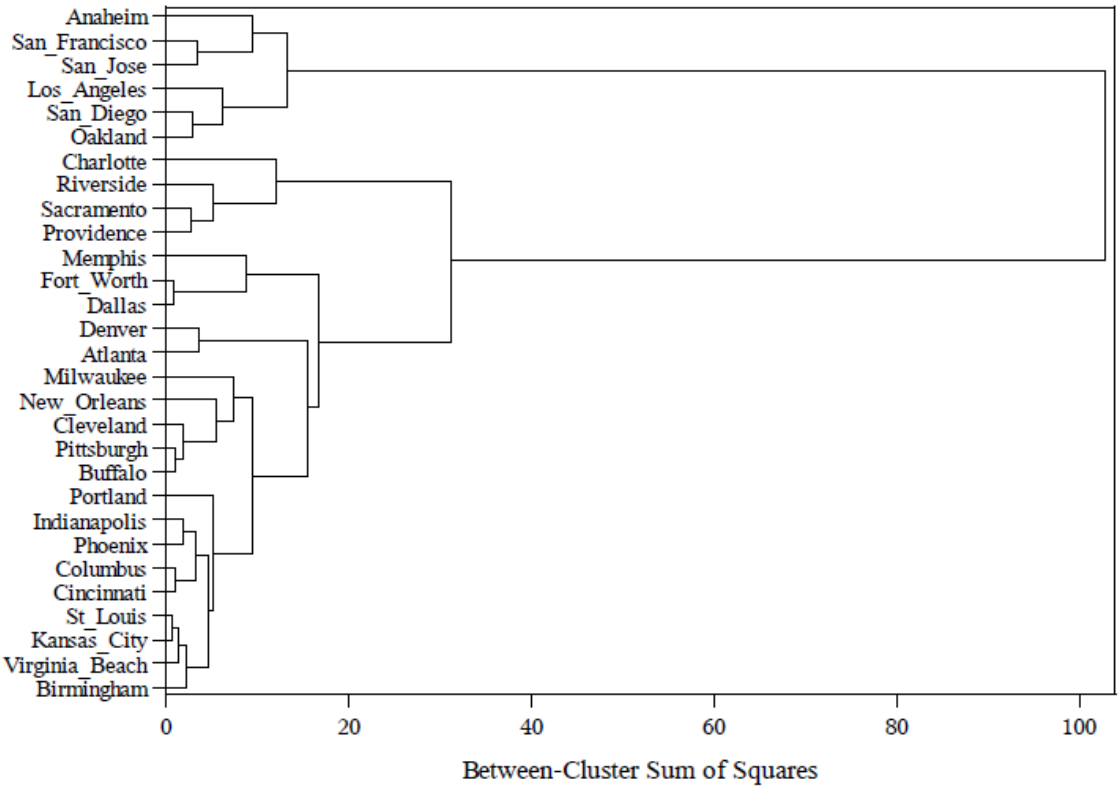


Figure 2 Plot of Number of Clusters Versus Semi-Partial R-Square Using Ward's Cluster Analysis



Note: A dot represents the level of semi-partial R-square with the corresponding number of clusters. The red line is a contour line and the two blue lines are the confidence limit lines.

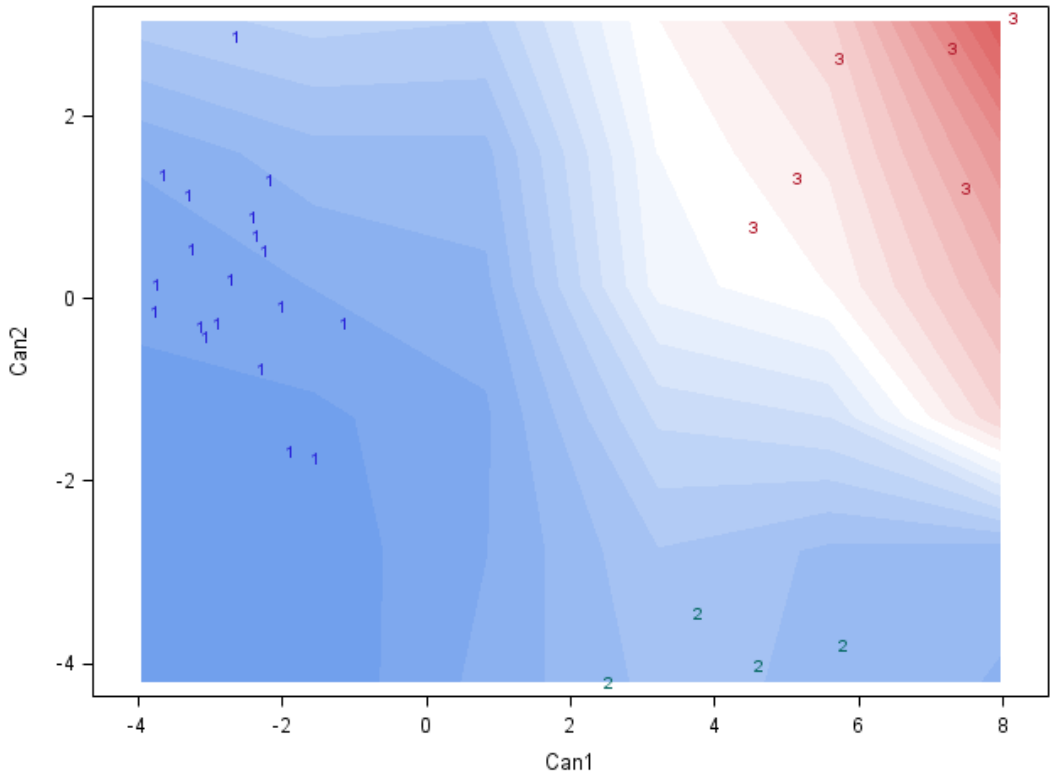
Figure 3 Results of Cluster Analysis-Tree Diagram



Note: The root node of the tree diagram represents the whole data set, and each leaf node is read as a data point. The intermediate nodes describe the extent to which the objects are proximal to each other, and the length of the tree diagram expresses the distance between each pair of data points or clusters, or a data point and a cluster (Xu and Wunsch, 2009).Three clusters are obtained by cutting the tree diagram at an appropriate level.

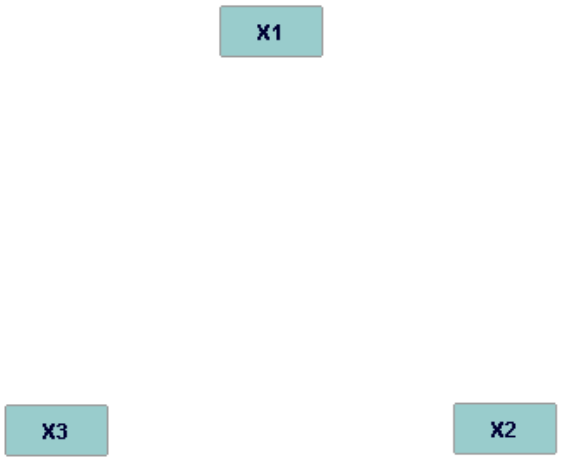
Figure 4 Results of Cluster Analysis-Metropolitan Map

Figure 5 Results of Discriminant Analysis



Note: a number (1, 2 and 3) indicates the cluster to which a MSA is assigned in cluster analysis. Can1 and Can2 are the two linear discriminant analysis (LDA) projections which maximize the separability of the discriminant scores of MSAs. We can see that the two LDA projections work well and there is no distribution overlapping between the three clusters. Also, the clustering pattern is consistent with the results from cluster analysis.

Figure 6 Graph from DAG Approach for Between-Cluster Analysis



Note: *X1*, *X2* and *X3* represent cluster one, two and three, respectively.

Figure 7 Graph from DAG Approach for Cluster 1

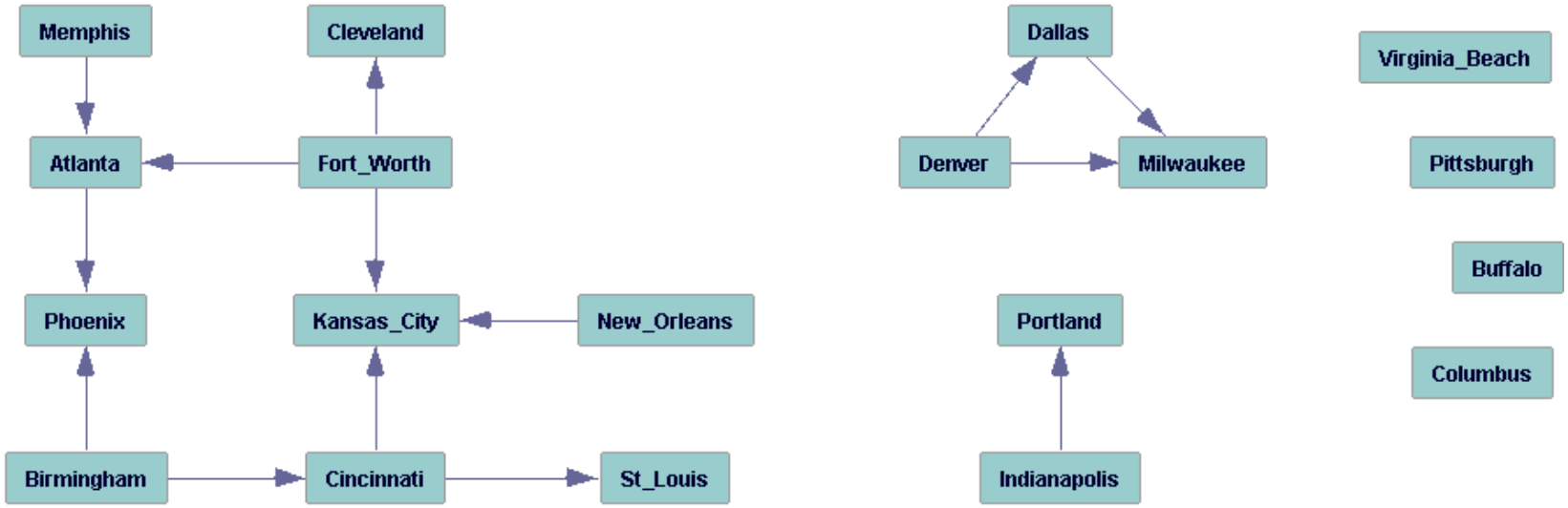


Figure 8 Graph from DAG Approach for Cluster 2

Figure 9 Graph from DAG Approach for Cluster 3

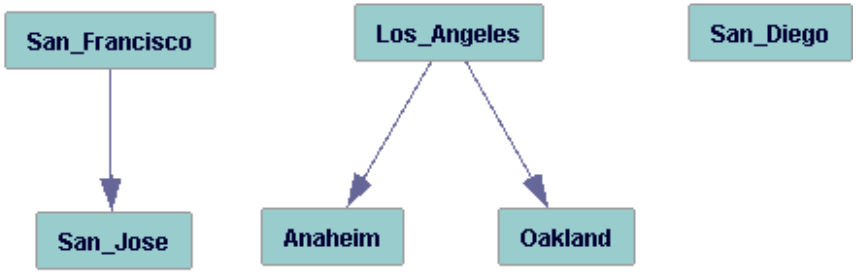


Figure 10 CDF Graph of the Forecasted Housing Price, 2011:Q1

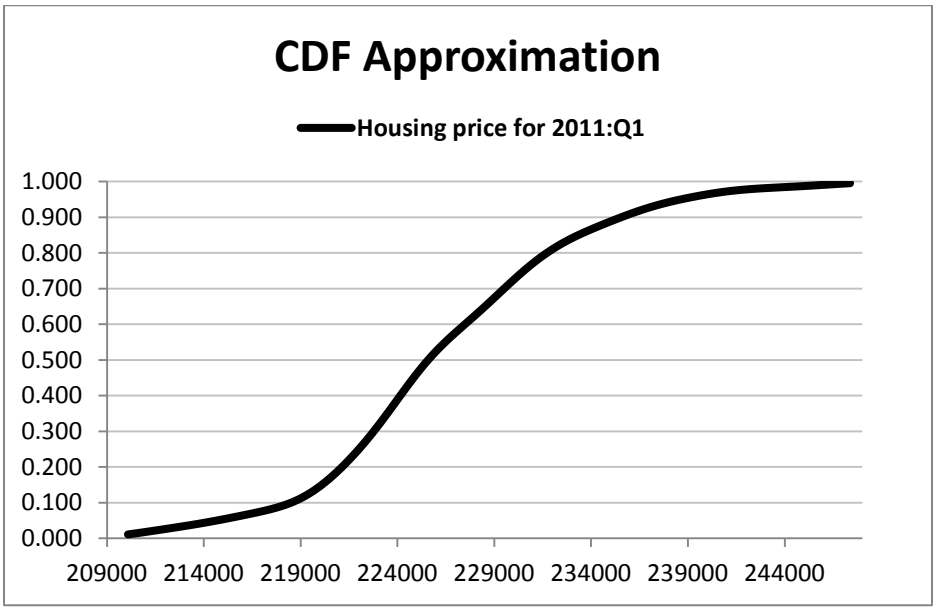


Figure 11 CDF Graph of the Forecasted Housing Price, 2012:Q1

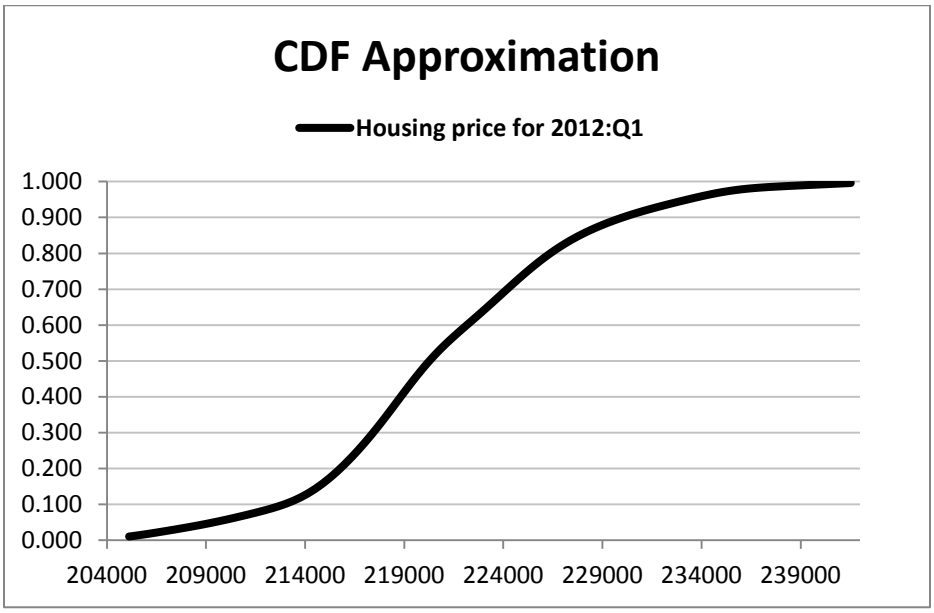
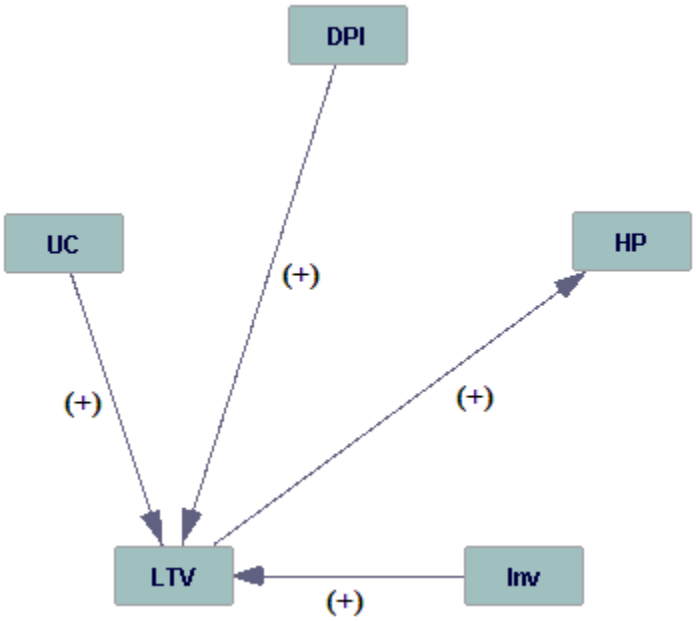
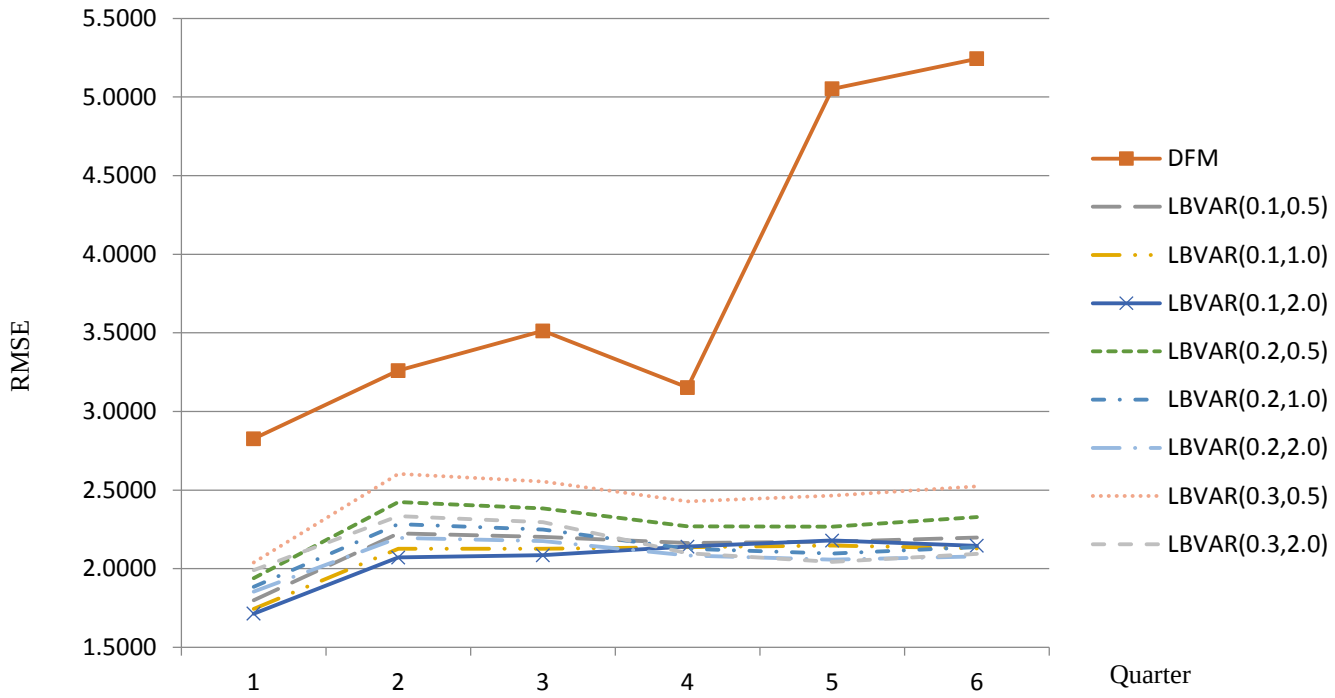


Figure 12 Contemporaneous Causal Patterns among the Five Random Variables, 2011:Q1



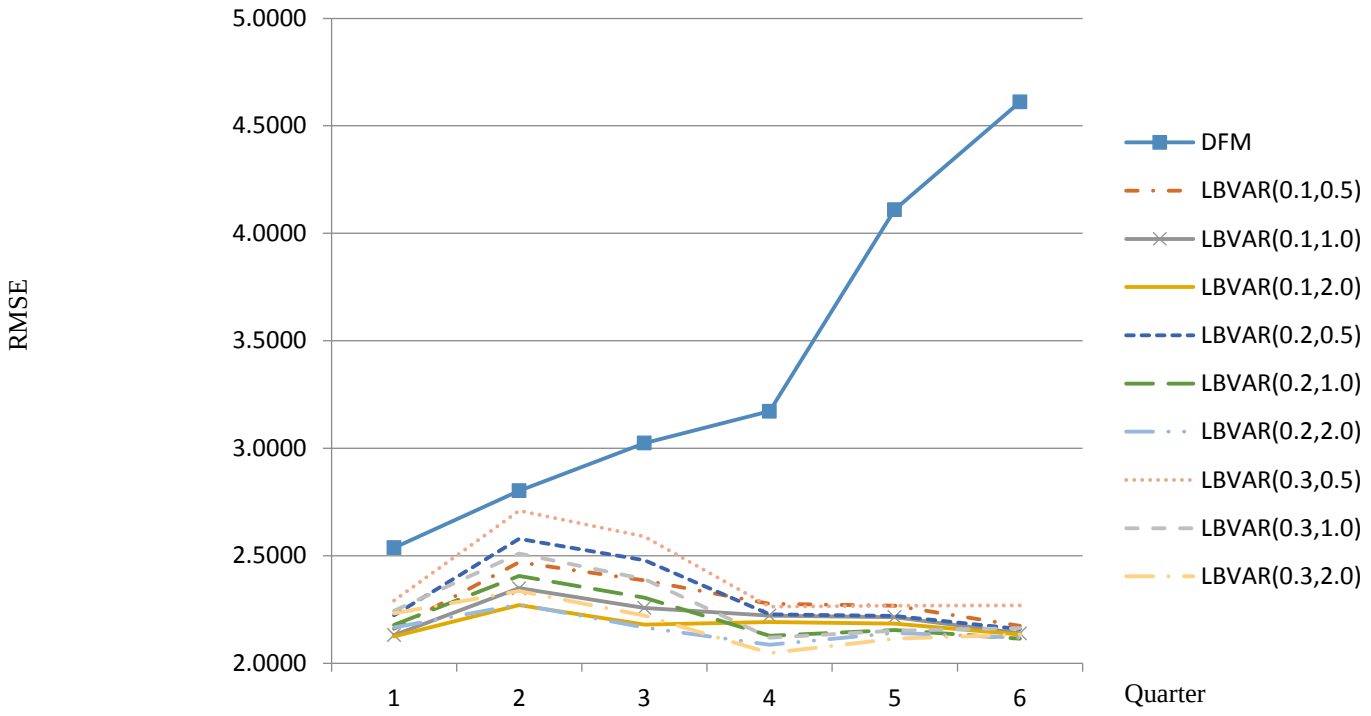
Note: *HP*, *Inv*, *DPI*, *UC* and *LTV* are house price, house inventory, disposable household income, user cost and loan-to-value, respectively.

Figure 13 Root Mean Square Errors (RMSEs) of 1- through 6-Quarter-Ahead Forecasts from DFM and LBVAR Models for Metropolitan Areas in the Group 1



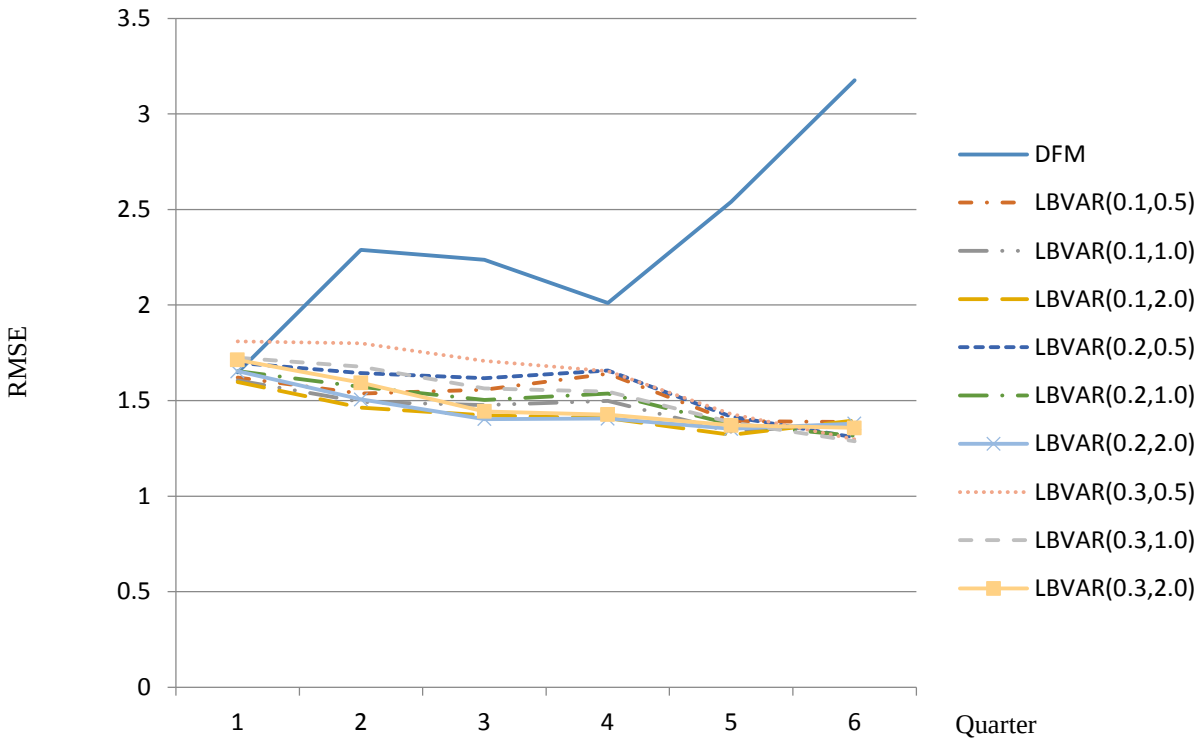
Note: group one includes metropolitan areas with housing price that peaked in the late 1980s, fell to a trough in the 1990s, and rebounded by 2004.

Figure 14 Root Mean Square Errors (RMSEs) of 1- through 6-Quarter-Ahead Forecasts from DFM and LBVAR Models for Metropolitan Areas in the Group 2



Note: group two includes metropolitan areas with housing price that were high in the early 1980s and were high again by the end of 2004.

Figure 15 Root Mean Square Errors (RMSEs) of 1- through 6-Quarter-Ahead Forecasts from DFM and LBVAR Models for Metropolitan Areas in the Group 3



Note: group three includes metropolitan areas with housing prices that declined since 1980 and did not fully recovered by the end of 2004.

Table 1 Metropolitan Statistical Areas Examined

Anaheim	Atlanta	Birmingham	Buffalo	Charlotte
Cincinnati	Cleveland	Columbus	Dallas	Denver
Fort Worth	Indianapolis	Kansas City	Los Angeles	Memphis
Milwaukee	New Orleans	Oakland	Phoenix	Pittsburgh
Portland	Providence	Riverside	Sacramento	San Diego
San Francisco	San Jose	St. Louis	Virginia Beach	

Table 2 Variables Used in Cluster Analysis

Variables	Descriptions
Housing value	Current value of unit
Unit size	Size of the unit (in square feet)
Rooms	Number of rooms in the unit (including bedrooms, bathrooms, living rooms, kitchens, family rooms, office, and other rooms)
Crowding	Number of persons per room
Unit quality rating	Rating of unit as a place to live (scale from 1(worst) to 10(best))
Neighborhood quality rating	Rating of neighborhood as a place to live (scale from 1(worst) to 10(best))
Unemployment rate	Rate of unemployment
Tax payment	Yearly real estate taxes payment
Mortgage rate	Current interest rate on primary mortgage (in %)
Household income	Expected household income in next twelve months

Table 3 Cluster History

Number of Clusters (CL#)	Clusters Joined		Frequency	Semipartial R-Square	R-Square
28	Kansas City	St. Louis	2	0.0022	0.998
27	Dallas	Fort Worth	2	0.0033	0.995
26	Buffalo	Pittsburgh	2	0.0035	0.991
25	Cincinnati	Columbus	2	0.0038	0.987
24	Virginia Beach	CL28	3	0.0052	0.982
23	CL26	Cleveland	3	0.0064	0.976
22	Phoenix	Indianapolis	2	0.0068	0.969
21	Birmingham	CL24	4	0.0076	0.961
20	Providence	Sacramento	2	0.0097	0.952
19	Oakland	San Diego	2	0.0104	0.941
18	CL25	CL22	4	0.0120	0.929
17	San Jose	San Francisco	2	0.0122	0.917
16	Atlanta	Denver	2	0.0128	0.904
15	CL21	CL18	8	0.0164	0.888
14	CL20	Riverside	3	0.0185	0.869
13	CL15	Portland	9	0.0186	0.851
12	CL23	New Orleans	4	0.0194	0.831
11	CL19	Los Angeles	3	0.0220	0.809
10	CL12	Milwaukee	5	0.0263	0.783
9	CL27	Memphis	3	0.0315	0.751
8	CL17	Anaheim	3	0.0337	0.718
7	CL13	CL10	14	0.0340	0.684
6	CL14	Charlotte	4	0.0431	0.641
5	CL11	CL8	6	0.0472	0.594
4	CL7	CL16	16	0.0558	0.538
3	CL4	CL9	19	0.0597	0.478
2	CL3	CL6	23	0.1112	0.367
1	CL2	CL5	29	0.3669	0.000

Note: CL# is the cluster formed when there are # clusters remain (number of cluster equals to #). For example, CL28 comprises Kansas City and St. Louis. With this cluster formed, there are totally 28 clusters remain.

Table 4 Discriminant Weight Vectors

Variable	Weight 1	Weight 2
Housing value	1.4737	2.8016
Unemployment	2.0061	-2.1986
Taxes	0.2677	0.2568
Interest rate	0.2644	-0.2440
Income	1.1047	-1.7720
Unit size	-0.9206	0.6096
Rooms	-0.3772	0.7309
Crowd	-0.1940	1.3158
Unit Quality	0.5460	0.0014
Neighborhood	0.1597	-0.5708

Table 5 Results of Augmented Dickey-Fuller (ADF) Tests on Levels and First-Differences for Between-Cluster Analysis

	Levels		First Differences	
Cluster	t-stat	p-value	t-stat	p-value
1	-1.31	0.1945	-2.07	0.0414
2	-1.54	0.1288	-2.15	0.0351
3	-1.3	0.1992	-2.12	0.0373

Note: the *p*-values for ADF tests on levels are larger than 0.05, which indicates that the null hypothesis of existence of unit root cannot be rejected. In other words, the series of housing values for all three clusters are not stationary. However, the *p*-values for ADF tests on first-difference are smaller than 0.05, so the null hypothesis can be rejected. We can conclude that these series of housing values are $I(1)$.

Table 6 Loss Metrics on Lag Length from VARs on Housing Values for Between-Cluster Analysis

	Information Criteria	
Lag Length k	HQC	SBC
1	4.6509	4.8526
2	3.0792	3.4348*
3	3.0101*	3.5217
4	3.0491	3.7191
5	3.2759	4.1066

Note: $SBC = \log(|\Sigma| + (3k+1)(\log T))/T$ and $\Phi = \log(|\Sigma| + 2(3k+1)\log(\log T))/T$. Σ is the error covariance matrix estimated with 3k+1 regressors in each equation. *T* is the number of observations on each series. The symbol "| |" denotes the determinants operator and log is the natural logarithm. The asterisk ("*") indicates minimum. Thus lag length of three is chosen.

Table 7 Tests of Cointegration among Housing Values for Between-Cluster Analysis

H0: Rank=r	H1: Rank>r	Eigenvalue	Trace	5% Critical Value
0	0	0.2123	38.4435	34.8
1	1	0.1531	18.1571	19.99
2	2	0.0464	4.036	9.13

Note: *r* is the number of cointegrating vectors. We fail to reject the null hypothesis when trace statistic is smaller than its 5% critical value. Thus, there exists one cointegrating vector among the housing values of the three clusters.

Table 8 Correlation Matrix of Innovations from ECM Model for Between-Cluster Analysis

Variable	Cluster1	Cluster2	Cluster3
Cluster1	1.0356	0.4398	0.2702
Cluster2	0.4398	2.0781	0.6010
Cluster3	0.2702	0.6010	5.7988

Table 9 Forecast Error Variance Decomposition for Between-Cluster Analysis

Horizon	Variable	Cluster 1	Cluster 2	Cluster 3
0	Cluster 1	1.0000	0.0000	0.0000
	Cluster 2	0.1934	0.8066	0.0000
	Cluster 3	0.0730	0.2882	0.6388
1	Cluster 1	0.9898	0.0100	0.0001
	Cluster 2	0.1577	0.8346	0.0077
	Cluster 3	0.1660	0.4421	0.3919
12	Cluster 1	0.7643	0.1323	0.1034
	Cluster 2	0.0858	0.8166	0.0976
	Cluster 3	0.1909	0.7781	0.0310

Note: Forecast error variance decompositions are based on observed innovations from the estimated error correction model. The entries sum to one in any row. The interpretation of any row is as follows: looking ahead at the horizon, given in the left-hand-most column (0, 1, 12-period-ahead), the uncertainty in house prices of the cluster in *variable* column is attributed to variation in innovations arising in each cluster in each column heading.

Table 10 Results of Augmented Dickey-Fuller Tests on Levels and First-Differences for MSAs in Each Cluster

Cluster1	Levels	First Differences	Cluster2	Levels	First Differences	Cluster3	Levels	First Differences
Atlanta	0.7407	0.0129	Charlotte	0.8738	0.0008	Los Angeles	0.1593	0.0170
Birmingham	0.8371	0.0008	Providence	0.8684	0.0432	Oakland	0.4638	0.0110
Buffalo	0.9782	0.0008	Riverside	0.9012	0.0157	San Diego	0.6040	0.0176
Cincinnati	0.7716	0.0008	Sacramento	0.9089	0.0117	San Francisco	0.9297	0.0008
Cleveland	0.6483	0.0034				San Jose	0.8462	0.0008
Columbus	0.8294	0.0039				Anaheim	0.8358	0.0480
Dallas	0.9501	0.0008						
Denver	0.8973	0.0008						
Fort Worth	0.9369	0.0008						
Indianapolis	0.8378	0.0001						
Kansas City	0.8356	0.0008						
Memphis	0.7966	0.0008						
Milwaukee	0.8320	0.0153						
New Orleans	0.8976	0.0008						
Phoenix	0.1415	0.0232						
Pittsburgh	0.9635	0.0008						
Portland	0.7873	0.0034						
St. Louis	0.8590	0.0127						
Virginia Beach	0.8323	0.0050						

Note: the *p*-values for ADF tests on levels are larger than 0.05, which indicates that the null hypothesis of existence of unit root cannot be rejected. In other words, the series of housing values are not stationary. However, the *p*-values for ADF tests on first-difference are smaller than 0.05, so the null hypothesis can be rejected. We can conclude that these series of housing values are $I(1)$.

Table 11 Loss Metrics on Lag Length from VARs on Housing Values for Cluster 2 and 3

	Information Criteria			
	Cluster 2		Cluster 3	
Lag Length k	HQC	SBC	HQC	SBC
1	9.0023	9.2712	15.99496	16.60012
2	7.2534	7.7951	14.42922	15.64823
3	7.4091	8.2276	14.09721	15.93897
4	7.0862*	8.1856*	13.57872	16.05239*
5	7.1487	8.5331	13.08026*	16.19523
6	7.1544	8.8282	13.27063	17.03654
7	7.1024	9.0699	13.18086	17.60763

Note: $SBC = \log(|\Sigma| + (4k+1)(\log T))/T$ and $\Phi = \log(|\Sigma| + 2(4k+1)\log(\log T))/T$ for cluster 2. $SBC = \log(|\Sigma| + (6k+1)(\log T))/T$ and $\Phi = \log(|\Sigma| + 2(6k+1)\log(\log T))/T$ for cluster 3. Σ is the error covariance matrix estimated with 4k+1 regressors in each equation for cluster 2 and with 6k+1 regressors in each equation for cluster 3. *T* is the number of observations on each series. The symbol "| |" denotes the determinants operator and log is the natural logarithm. The asterisk ("*") indicates minimum. Thus, lag length of four is chosen for cluster 2 and lag length of 5 is chosen for cluster 3.

Table 12 Tests of Cointegration among Housing Values for Cluster 2 and 3

		Cluster 2			Cluster 3		
H0:	H1:			5% Critical			5% Critical
Rank=r	Rank>r	Eigenvalue	Trace	Value	Eigenvalue	Trace	Value
0	0	0.4514	83.3266	47.21	0.5581	206.8011	93.92
1	1	0.2229	32.2891	29.38	0.5112	138.2092	68.68
2	2	0.0972	10.8511	15.34	0.3646	78.0799	47.21
3	3	0.0251	2.1584	3.84	0.2622	39.9883	29.38
4	4				0.1492	14.4485	15.34
5	5				0.0104	0.8803	3.84

Note: *r* is the number of cointegrating vectors. We fail to reject the null hypothesis when trace statistic is smaller than its 5% critical value. Thus, there exist two cointegrating vectors among the housing values of cluster 2 and four cointegrating vectors among the housing values of cluster 3.

Table 13 Correlation Matrix of Innovations from Bayesian VAR Model for Cluster 1

	Atlanta	Birmingham	Buffalo	Cincinnati	Cleveland	Columbus	Dallas	Denver	Fort Worth	Indianapolis	Kansas City	Memphis	Milwaukee	New Orleans	Phoenix	Pittsburgh	Portland	St Louis	Virginia Beach
Atlanta	1.0000																		
Birmingham	0.4057	1.0000																	
Buffalo	0.1210	0.0894	1.0000																
Cincinnati	0.3557	0.4448	0.0969	1.0000															
Cleveland	0.4569	0.3032	0.1168	0.2940	1.0000														
Columbus	0.1581	0.2762	-0.0737	0.2817	0.3759	1.0000													
Dallas	0.3686	0.1771	0.3213	0.2333	0.2796	0.1990	1.0000												
Denver	0.4413	0.1904	0.0972	0.3383	0.3397	0.1237	0.3730	1.0000											
Fort Worth	0.5099	0.3772	0.0521	0.4392	0.4948	0.3480	0.2532	0.3981	1.0000										
Indianapolis	0.4702	0.4257	-0.0886	0.3818	0.2758	0.3209	0.2458	0.2359	0.3962	1.0000									
Kansas City	0.3716	0.3899	-0.0007	0.4705	0.3786	0.4796	0.3399	0.2920	0.5481	0.3375	1.0000								
Memphis	0.4393	0.3117	0.0865	-0.0425	0.2423	0.1473	0.2358	0.2909	0.0658	0.1946	0.1740	1.0000							
Milwaukee	0.4264	0.2389	0.2198	0.4430	0.1711	-0.0200	0.4041	0.4146	0.3969	0.3812	0.3648	0.0823	1.0000						
New Orleans	0.0806	0.2045	-0.1271	0.3505	0.2517	0.2307	-0.0543	0.1162	0.2914	0.0606	0.3782	-0.0992	0.0655	1.0000					
Phoenix	0.4856	0.4722	0.2939	0.3174	0.0061	-0.0174	0.2239	0.1762	0.2769	0.2782	0.3065	0.2074	0.3533	0.0392	1.0000				
Pittsburgh	0.3395	-0.0122	0.3672	0.3120	0.0507	-0.0818	0.1475	0.1546	0.2361	0.1574	0.0749	-0.0378	0.3031	0.0648	0.1810	1.0000			
Portland	0.6439	0.4170	0.1098	0.2020	0.3224	0.1450	0.3014	0.2703	0.3113	0.5071	0.3044	0.3216	0.2907	-0.0822	0.3439	0.0871	1.0000		
St. Louis	0.4880	0.4401	0.1639	0.4854	0.2287	0.3077	0.2634	0.2730	0.4431	0.3249	0.4305	0.2087	0.3653	0.3115	0.1951	0.2774	0.4033	1.0000	
Virginia Beach	0.1665	0.2708	0.1592	0.2097	-0.1344	0.0283	0.0796	0.1247	0.0092	0.1706	0.1365	0.1117	0.1623	0.2252	0.4171	0.0286	0.3386	0.2360	1.0000

Table 14 Correlation Matrix of Innovations from ECM Model for Cluster 2

	Charlotte	Providence	Riverside	Sacramento
Charlotte	1.0000			
Providence	0.0428	1.0000		
Riverside	0.1948	0.0652	1.0000	
Sacramento	-0.0108	0.2463	0.5359	1.0000

Table 15 Correlation Matrix of Innovations from Bayesian VAR Model for Cluster 3

	Los Angeles	Oakland	San Diego	San Francisco	San Jose	Anaheim
Los Angeles	1.0000					
Oakland	0.4432	1.0000				
San Diego	0.1487	0.3191	1.0000			
San Francisco	0.1684	0.2397	0.1582	1.0000		
San Jose	0.3156	0.3629	0.0549	0.3693	1.0000	
Anaheim	0.6470	0.2912	0.3752	-0.0057	0.3364	1.0000

Table 16 Forecast Error Variance Decomposition for Cluster 2

Lead	Variable	Charlotte	Providence	Riverside	Sacramento
0	Charlotte	1.0000	0.0000	0.0000	0.0000
	Providence	0.0030	0.9970	0.0000	0.0000
	Riverside	0.0002	0.0600	0.9601	0.0000
	Sacramento	0.0002	0.0600	0.2873	0.6526
1	Charlotte	0.9919	0.0041	0.0005	0.0004
	Providence	0.0216	0.9552	0.0006	0.0226
	Riverside	0.0379	0.0022	0.8675	0.0923
	Sacramento	0.0006	0.0197	0.0952	0.8845
12	Charlotte	0.8647	0.0463	0.0714	0.0175
	Providence	0.0152	0.8918	0.0062	0.0869
	Riverside	0.0699	0.0581	0.8374	0.0345
	Sacramento	0.0528	0.0530	0.0966	0.7976

Note: Forecast error variance decompositions are based on observed innovations from the estimated error correction model. The entries sum to one in any row. The interpretation of any row is as follows: looking ahead at the horizon, given in the left-hand-most column (0, 1, 12-period-ahead), the uncertainty in house prices of the cluster in *variable* column is attributed to variation in innovations arising in each cluster in each column heading.

Table 17 Forecast Error Variance Decomposition for Cluster 3

Lead	Variable	Los Angeles	Oakland	San Diego	San Francisco	San Jose	Anaheim
0	Los Angeles	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Oakland	0.1902	0.8098	0.0000	0.0000	0.0000	0.0000
	San Diego	0.0219	0.0831	0.8950	0.0000	0.0000	0.0000
	San Francisco	0.0284	0.0286	0.0072	0.9359	0.0000	0.0000
	San Jose	0.0969	0.0422	0.0044	0.0914	0.7651	0.0000
	Anaheim	0.4183	0.0000	0.0855	0.0200	0.0515	0.4248
1	Los Angeles	0.9270	0.0282	0.0178	0.0266	0.0000	0.0005
	Oakland	0.2842	0.6904	0.0000	0.0067	0.0186	0.0000
	San Diego	0.1805	0.2409	0.5401	0.0243	0.0079	0.0064
	San Francisco	0.0173	0.1304	0.0064	0.9708	0.1408	0.0344
	San Jose	0.0827	0.0792	0.0027	0.0939	0.7374	0.0040
	Anaheim	0.6091	0.0547	0.0330	0.0336	0.0186	0.2511
12	Los Angeles	0.4123	0.0317	0.1998	0.3096	0.0147	0.0319
	Oakland	0.2113	0.0904	0.3152	0.2749	0.0980	0.0102
	San Diego	0.0939	0.0330	0.5526	0.2643	0.0555	0.0007
	San Francisco	0.0623	0.0572	0.2654	0.0676	0.4615	0.0861
	San Jose	0.2108	0.0628	0.1194	0.0999	0.4006	0.0664
	Anaheim	0.2848	0.0487	0.3313	0.2830	0.0432	0.0089

Note: Forecast error variance decompositions are based on observed innovations from the estimated error correction model. The entries sum to one in any row. The interpretation of any row is as follows: looking ahead at the horizon, given in the left-hand-most column (0, 1, 12-period-ahead), the uncertainty in house prices of the cluster in *variable* column is attributed to variation in innovations arising in each cluster in each column heading.

Table 18 Statistics for Trace Test and Maximum Eigenvalue Test

	Trace		Max Eigenvalue	
rank	statistic	p-value	statistic	p-value
0	20.5892	0.7174	44.8178	0.8364
1	14.9627	0.7523	24.2349	0.9374
2	4.496	0.9978	9.2722	0.9889
3	4.1824	14.2644	4.7762	0.8320
4	0.5938	0.5987	0.5938	0.5987

Note: based on *p*-values, the null hypothesis of rank=0 cannot be rejected by either trace or eigenvalue test.

Table 19 Schwarz Information Criterion and Hannan and Quinn's Φ on VAR Model in First Differences and ECM Model

	Lag 1		Lag 2		Lag 3		Lag 4		Lag 5	
Rank	SIC	Φ	SIC	Φ	SIC	Φ	SIC	Φ	SIC	Φ
r=0	-94.2824	-94.3742	-60.0524	-60.2376	-60.1825	-60.4626	-94.5801*	-94.9567*	-59.4289	-59.9037
r=1	-59.8731	-60.0583	-60.3530	-60.5382	-60.6799	-60.9600	-89.0565	-89.4331	-60.3077	-60.7825
r=2	-60.1746	-60.2664	-59.9919	-60.0838	-60.3686	-60.6486	-61.0643	-61.4409	-61.1545	-61.6293
r=3	-60.2984	-60.3902	-91.8459	-92.0311	-61.1949	-61.4750	-89.5318	-89.9084	-60.9447	-61.4195
r=4	-60.5460	-60.6378	-60.0597	-60.2449	-61.9640	-62.2441	-62.0282	-62.4048	-88.9113	-89.3861

Note: $SIC = \log(|\Sigma| + (5k+1)(\log T))/T$, and $\Phi = \log(|\Sigma| + 2(5k+1)\log(\log T))/T$. Σ is the error covariance matrix estimated with 5k+1 regressors in each equation. *T* is the number of observations on each series. The symbol "| |" denotes the determinants operator and log is the natural logarithm. The asterisk ("*") indicates minimum.

Table 20 Parameter Estimation of c and Γ_i (i=1,…, 4) for the VAR(4) Model in First Differences (Equation 24)

		Γ_1					Γ_2				
	c	P	Inv	DPI	UC	LTV	P	Inv	DPI	UC	LTV
P	-0.0056	-0.2675	-0.0417	0.1460	-0.0133	-0.0692	0.0707	0.0086	0.4243	0.0058	-0.3035
Inv	-0.0116	-0.2431	-0.1320	-0.1512	0.0144	2.5641	0.0701	-0.1472	-0.3650	-0.0061	0.4494
DPI	0.0020	0.0100	0.0344	-0.3415	0.0034	-0.0030	-0.1013	-0.0052	0.0339	0.0031	0.0848
UC	-0.0284	1.7067	0.3189	5.4613	-0.4950	-3.8404	-3.2651	0.9069	3.0805	0.0368	8.1705
LTV	-0.0024	-0.0024	0.0151	-0.0086	-0.0004	-0.1149	0.0787	0.0149	-0.0991	0.0016	0.0946

	Γ_3					Γ_4				
	P	Inv	DPI	UC	LTV	P	Inv	DPI	UC	LTV
P	0.2233	0.0581	0.2651	0.0098	-0.0191	0.2911	0.1206	-0.7692	-0.0067	0.2318
Inv	-0.0260	-0.0332	1.3452	-0.0178	-1.2384	-0.4085	0.1531	0.9514	-0.0018	-0.1771
DPI	-0.1001	0.0015	0.1277	0.0044	0.2689	-0.0287	0.0174	-0.0066	0.0039	0.3546
UC	-4.7185	2.0143	6.6380	-0.0464	-0.5456	1.9972	-0.8143	4.1447	-0.2246	-8.0833
LTV	0.0591	0.0192	-0.0966	0.0018	-0.1306	0.0021	0.0196	-0.0975	0.0003	0.1061

Note: *HP*, *Inv*, *DPI*, *UC* and *LTV* are house price, house inventory, disposable household income, user cost and loan-to-value, respectively.

Table 21 Kendall's Tau Concordance Matrix Estimated Based on the Residuals from the VAR(4) Model in First Differences (Equation 24)

	P	Inv	DPI	UC	LTV
P	0.9946	0.0580	0.0957	0.2551	0.1460
Inv		1.0000	0.0105	-0.0422	0.1326
DPI			1.0000	0.0825	0.1535
UC				1.0000	-0.0317
LTV					0.9996

Note: *HP, Inv, DPI, UC* and *LTV* are house price, house inventory, disposable household income, user cost and loan-to-value, respectively.

Table 22 Summary of Statistics of Simulated and Historical Data for the Five Random Variables ($\Delta\tilde{Y}_{i,t}$) as Calculated in Equation 25

		HP	Inv	DPI	UC	LTV
Simulated	Mean	0.00904	0.01177	0.00432	0.49404	-0.00009
	StDev	0.03048	0.05949	0.00882	4.17988	0.01335
	CV	337.14260	505.47815	204.30024	846.06331	-15521.97574
	Min	-0.06338	-0.13350	-0.02516	-7.41774	-0.04492
	Max	0.10287	0.15328	0.02464	36.24186	0.03138
Historical	Mean	0.00889	0.01179	0.00426	0.49519	-0.00009
	StDev	0.03060	0.05927	0.00884	4.39360	0.01337

Note: *HP*, *Inv*, *DPI*, *UC* and *LTV* are house price, house inventory, disposable household income, user cost and loan-to-value, respectively.

Table 23 Comparison of the Simulated and Historical Distributions of the Five Random Variables

	Test Value	Critical Value	P-Value
2 Sample Hotelling T2 Test	0.01	11.15	1.000
Box's M Test	9.24	25.00	0.864
Complete Homogeneity Test	9.60	31.41	0.975

Note: Test confidence level is 95%.

Table 24 Some Quantile Values for Forecasted Median Housing Price, 2011:Q1 and 2012:Q1

	Median Housing Price	
Quantile	2011:Q1	2012:Q1
0.05	211779.61	206775.30
0.15	215546.24	210453.00
0.25	219312.86	214130.60
0.50	228729.42	223324.80
0.75	238145.98	232518.90
0.85	241912.61	236196.50
0.95	245679.23	239874.20

Note: The observed median house price is $226,900 in 2011:Q1, which is very close to the median value (50th perentaile). The observed median house price is $225,750 in 2012:Q1.

Table 25 Metropolitan Areas with Three Price Patterns

Group One				
Markets where house prices peaked in the late 1980s and had a trough in the 1990s:				
Atlanta, GA	Dallas, TX	Oakland, CA	Raleigh-Durham, NC	San Francisco, CA
Austin, TX	Jacksonville, FL	Philadelphia, PA	Richmond, VA	San Jose, CA
Baltimore, MD	Los Angeles, CA	Phoenix, AZ	Sacramento, CA	Seattle, WA
Boston, MA	New York, NY	Portland, OR	San Diego, CA	
Group Two				
Markets where house prices were high in the early 1980s and rebounded in the 2000s:				
Charlotte, NC	Columbus, OH	Indianapolis, IN	Milwaukee, MN	St. Louis, MO
Chicago, IL	Denver, CO	Kansas City, KS	Minneapolis, MN	Tampa, FL
Cincinnati, OH	Detroit, MI	Memphis, TN	Orlando, FL	
Cleveland, OH	Fort Lauderdale, FL	Miami, FL	Pittsburgh, PA	
Group Three				
Markets where house prices have declined since the early 1980s and never fully rebounded:				
Fort Worth, TX	Houston, TX	New Orleans, LA		

Table 26 Root Mean Square Errors (RMSEs) of 1- through 6-Quarter-Ahead Forecasts from DFM and LBVAR Models for Metropolitan Areas in the Group 1

	1-Quarter-Ahead	2-Quarter-Ahead	3-Quarter-Ahead	4-Quarter-Ahead	5-Quarter-Ahead	6-Quarter-Ahead
DFM	2.8260	3.2595	3.5125	3.1514	5.0514	5.2435
LBVAR(0.1,0.5)	1.7984	2.2235	2.2015	2.1628	2.1711	2.1972
LBVAR(0.1,1.0)	1.7446	2.1270	2.1276	2.1367	2.1472	2.1285
LBVAR(0.1,2.0)	1.7143	2.0713	2.0850	2.1402	2.1793	2.1453
LBVAR(0.2,0.5)	1.9388	2.4229	2.3830	2.2690	2.2667	2.3276
LBVAR(0.2,1.0)	1.8842	2.2840	2.2491	2.1261	2.0958	2.1365
LBVAR(0.2,2.0)	1.8545	2.1959	2.1750	2.0836	2.0569	2.0782
LBVAR(0.3,0.5)	2.0394	2.6033	2.5543	2.4285	2.4654	2.5242
LBVAR(0.3,1.0)	1.9990	2.4483	2.3916	2.2193	2.1900	2.2542
LBVAR(0.3,2.0)	1.9907	2.3342	2.2954	2.0991	2.0428	2.0949

Note: (1) For each month-ahead-forecast, the model with the smallest RMSE is denoted with red shadow. (2) Group one includes metropolitan areas with housing price that peaked in the late 1980s, fell to a trough in the 1990s, and rebounded by 2004.

Table 27 Root Mean Square Errors (RMSEs) of 1- through 6-Quarter-Ahead Forecasts from DFM and LBVAR Models for Metropolitan Areas in the Group 2

	1-Quarter-Ahead	2-Quarter-Ahead	3-Quarter-Ahead	4-Quarter-Ahead	5-Quarter-Ahead	6-Quarter-Ahead
DFM	2.5377	2.8026	3.0240	3.1722	4.1099	4.6109
LBVAR(0.1,0.5)	2.1604	2.4706	2.3861	2.2761	2.2680	2.1736
LBVAR(0.1,1.0)	2.1314	2.3491	2.2580	2.2222	2.2149	2.1397
LBVAR(0.1,2.0)	2.1254	2.2705	2.1795	2.1919	2.1846	2.1345
LBVAR(0.2,0.5)	2.2238	2.5785	2.4787	2.2279	2.2203	2.1585
LBVAR(0.2,1.0)	2.1792	2.4061	2.3044	2.1282	2.1552	2.1140
LBVAR(0.2,2.0)	2.1669	2.2740	2.1672	2.0864	2.1419	2.1192
LBVAR(0.3,0.5)	2.2922	2.7100	2.5892	2.2635	2.2682	2.2692
LBVAR(0.3,1.0)	2.2445	2.5098	2.3921	2.1187	2.1515	2.1639
LBVAR(0.3,2.0)	2.2315	2.3358	2.2223	2.0482	2.1152	2.1341

Note: (1) For each month-ahead-forecast, the model with the smallest RMSE is denoted with red shadow. (2) Group two includes metropolitan areas with housing price that were high in the early 1980s and were high again by the end of 2004.

Table 28 Root Mean Square Errors (RMSEs) of 1- through 6-Quarter-Ahead Forecasts from DFM and LBVAR Models for Metropolitan Areas in the Group 3

	1-Quarter-Ahead	2-Quarter-Ahead	3-Quarter-Ahead	4-Quarter-Ahead	5-Quarter-Ahead	6-Quarter-Ahead
DFM	1.647588	2.288181	2.237336	2.010204	2.53924	3.176274
LBVAR(0.1,0.5)	1.620543	1.536548	1.556344	1.641213	1.392705	1.388699
LBVAR(0.1,1.0)	1.604821	1.492624	1.477015	1.499096	1.337302	1.3796
LBVAR(0.1,2.0)	1.597208	1.462635	1.422795	1.408373	1.32015	1.39663
LBVAR(0.2,0.5)	1.697171	1.644292	1.617239	1.658264	1.415927	1.30713
LBVAR(0.2,1.0)	1.654114	1.572179	1.503271	1.536133	1.376871	1.315256
LBVAR(0.2,2.0)	1.653304	1.506692	1.40308	1.4057	1.351962	1.380111
LBVAR(0.3,0.5)	1.810195	1.799779	1.707888	1.653442	1.430466	1.297837
LBVAR(0.3,1.0)	1.72393	1.67696	1.563224	1.546582	1.38678	1.286885
LBVAR(0.3,2.0)	1.712892	1.593261	1.44394	1.426769	1.370468	1.357489
	1.597208	1.462635	1.40308	1.4057	1.32015	1.286885

Note: (1) For each month-ahead-forecast, the model with the smallest RMSE is denoted with grey shadow. (2) Group three includes metropolitan areas with housing prices that declined since 1980 and did not fully recovered by the end of 2004.

Table 29 Results of Encompassing Test for Metropolitan Areas in the Group 1

	1-Quarter-Ahead	2-Quarter-Ahead	3-Quarter-Ahead	4-Quarter-Ahead	5-Quarter-Ahead	6-Quarter-Ahead
DFM			X		X	
LBVAR(0.1,0.5)			X	X		
LBVAR(0.1,1.0)			X	X		
LBVAR(0.1,2.0)	X, 1.7142	X, 2.0713	X, 2.0850	X	X	
LBVAR(0.2,0.5)			X			
LBVAR(0.2,1.0)			X			
LBVAR(0.2,2.0)	X			X, 2.0836		X, 2.0782
LBVAR(0.3,0.5)			X	X		
LBVAR(0.3,1.0)				X	X	
LBVAR(0.3,2.0)	X	X			X, 2.0428	
RMSE-weighted	1.8220	2.1614	2.2308	2.1205	2.1981	2.0782
Rank-weighted	1.7496	2.0918	2.1397	2.0932	2.1031	2.0782
Thick-modeling	1.8292	2.1692	2.2602	2.1237	2.4402	2.0782

Note: (1)For each month-ahead-forecast, the models that are not encompassed by other models are marked with 'X'; (2) the smallest RMSE from the 10 competing models is reported after 'X' in the convenience of comparing to the RMSEs from three encompassing tests; (3) the smallest RMSE among those from both individual and combined forecasts is denoted with grey shadow. (4) Group one includes metropolitan areas with housing price peaked in the late 1980s, fell to a trough in the 1990s, and rebounded by 2004.

Table 30 Results of Encompassing Test for Metropolitan Areas in the Group 2

	1-Quarter-Ahead	2-Quarter-Ahead	3-Quarter-Ahead	4-Quarter-Ahead	5-Quarter-Ahead	6-Quarter-Ahead
DFM		X	X	X		
LBVAR(0.1,0.5)				X		X
LBVAR(0.1,1.0)	X				X	
LBVAR(0.1,2.0)	X, 2.1254	X, 2.2705			X	
LBVAR(0.2,0.5)		X	X		X	
LBVAR(0.2,1.0)					X	X, 2.1140
LBVAR(0.2,2.0)			X, 2.1672			X
LBVAR(0.3,0.5)		X	X			
LBVAR(0.3,1.0)					X	
LBVAR(0.3,2.0)				X, 2.0482	X, 2.1152	
RMSE-weighted	2.1405	2.2816	2.1967	2.1994	2.1646	2.3914
Rank-weighted	2.1364	2.3210	2.2633	2.1323	2.1588	2.2217
Thick-modeling	2.1406	2.2821	2.1979	2.2730	2.1653	2.6941

Note: (1)For each month-ahead-forecast, the models that are not encompassed by other models are marked with 'X'; (2) the smallest RMSE from the 10 competing models is reported after 'X' in the convenience of comparing to the RMSEs from three encompassing tests; (3) the smallest RMSE among those from both individual and combined forecasts is denoted with grey shadow. (4) Group two includes metropolitan areas with housing price high in the early 1980s and high again by the end of 2004.

Table 31 Results of Encompassing Test for Metropolitan Areas in the Group 3

	1-Quarter-Ahead	2-Quarter-Ahead	3-Quarter-Ahead	4-Quarter-Ahead	5-Quarter-Ahead	6-Quarter-Ahead
DFM	X	X	X	X		
LBVAR(0.1,0.5)	X	X	X	X	X	
LBVAR(0.1,1.0)	X	X	X	X	X	
LBVAR(0.1,2.0)	X	X, 1.4626	X	X	X, 1.3202	
LBVAR(0.2,0.5)	X	X	X	X	X	X
LBVAR(0.2,1.0)	X	X	X	X	X	X
LBVAR(0.2,2.0)	X	X	X, 1.4031	X, 1.4057	X	
LBVAR(0.3,0.5)	X	X	X	X	X	X
LBVAR(0.3,1.0)	X	X	X	X	X	X, 1.2869
LBVAR(0.3,2.0)	X	X	X	X	X	
RMSE-weighted	1.5686	1.5529	1.4875	1.4577	1.3507	1.2912
Rank-weighted	1.5514	1.5000	1.4411	1.4288	1.3321	1.2879
Thick-modeling	1.5708	1.5652	1.4960	1.4571	1.3514	1.2913

Note: (1)For each month-ahead-forecast, the models that are not encompassed by other models are marked with 'X'; (2) the smallest RMSE from the 10 competing models is reported after 'X' in the convenience of comparing to the RMSEs from three encompassing tests; (3) the smallest RMSE among those from both individual and combined forecasts is denoted with grey shadow. (4) Group three includes metropolitan areas with housing prices declining since 1980 and not fully recovered at the end of 2004