Modelling Spatial Dependence with Pairwise Correlations

Natalia Bailey, Sean Holly and Hashem Pesaran*
University of Cambridge and University of Southern California
PRELIMINARY, PLEASE DO NOT QUOTE

June 2013

Abstract

An understanding of the spatial dimension of economic and social activity requires methods that can separate out the relationship between spatial units that is due to the effect of common factors from that which is purely spatial even in an abstract sense. The same applies to the empirical analysis of networks in general. We are able to distinguish between cross-sectional strong dependence and weak dependence. Strong dependence in turn suggests that there are common factors. We use cross unit averages to extract both national and regional factors and contrast this to a principal components approach widely used in the literature. We then use thresholding to determine significant bilateral correlations (signifying neighbours) between spatial units and compare this to an approach that just uses distance to determine units that are neighbours. We apply these methods to real house price changes at the level of the Metropolitan Statistical Area in the USA.

JEL Classification: C21, C23
Keywords: House Prices, Exponent of Cross Sectional Dependence, Spatial Dependence.

*We acknowledge financial support under ESRC Grant ES/I031626/1.
1 Introduction

The nature and degree of spatial dependence in economic, geographical, epidemiological and ecological systems has long been the focus of intensive study. Geographers regard the fundamental question in economic geography to be what explains the uneven pattern of economic activity in space. Indeed the New Economic Geography starting with Krugman (1991) addresses exactly this question. But where we have a data rich environment with observations on many spatial units over many time periods there may be obstacles to understanding these uneven patterns in spatial data because of complex dependence between spatial units that reflects both local (clustering) and common factors. This paper describes some methods for reducing these data fields (Quah, 1994) to manageable proportions. Recent developments in spatial econometrics have generated a growing literature on methods for modelling and measuring spatial or cross section dependence in data sets with a panel structure where there are observations over time and over space. This in turn has identified a number of central research questions. What is the source of dependencies in space? To what extent are the observed dependencies between different spatial units due to common factors -for example, aggregate shocks - that effect different units rather than being the result of local interactions that generate spatial spill-over effects? When there are common effects, conditioning on variables specific to the spatial unit alone need not deliver cross section error independence. Moreover, neglecting cross sectional or spatial dependence can lead to spurious inference.

In this paper we describe a new approach to the modelling of weak (spatial) dependence that can be used when the time series dimension of a panel is relatively large and apply the method to house prices in the USA at the level of Metropolitan Statistical Areas (MSA).

Currently, there are two main approaches to modelling cross sectional or spatial dependence in large panels: spatial processes and factor structures. Spatial processes were pioneered by Whittle (1954) and developed further in econometrics by Anselin (1988), Kelejian and Prucha (1999), and Lee (2002), amongst others. Factor models were introduced by Hotelling (1933) and first applied in economics by Stone (1947). They have been applied extensively in finance (Chamberlain and Rothschild 1983, Connor and Korajczyk, 1993, Stock and Watson, 1998, and Kapetanios and Pesaran, 2007), and in macroeconomics as in Forni and Reichlin (1998) and Stock and Watson (2002).

Estimation of panels with spatially correlated errors include the use of parametric methods based on maximum likelihood (Lee, 2004; Lee and Yu, 2008, Yu et al., 2008), or GMM methods (Lin and Lee, 2005; Kapoor et al. 2007; Kelejian and Prucha, 1999, 2009). Non-parametric methods using spatial HAC estimators have been applied by Conley (1999), Kelejian and Prucha (2007) and Bester et al.(2009). For situations in which there are more than one possible factor there is the Common Correlated Effects approach (Pesaran, 2006), the maximum likelihood method of Robertson and Symons (2007) and the principal components technique of Bai (2009).

We show that once common factors - a form of strong spatial dependence - are removed that an analysis of weak spatial dependence in house prices reveals interesting correlations in house price movements across the USA.\footnote{Throughout the paper, $\lambda_1 (A) \geq \lambda_2 (A) \geq \ldots \geq \lambda_N (A)$ will be used for the sorted eigenvalues of the $N \times N$ matrix $A = (a_{ij})$, where $\lambda_1 (A)$ and $\lambda_N (A)$ are the maximum and minimum eigenvalues respectively, $\|A\|_1 = \max_{1 \leq j \leq N} \left\{ \sum_{i=1}^{N} |a_{ij}| \right\}$ for its maximum absolute column sum norm, $\|A\|_\infty = \max_{1 \leq i \leq N} \left\{ \sum_{j=1}^{N} |a_{ij}| \right\}$ for its maximum absolute row sum norm, $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$ for its spectral norm, and $\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2}$ for its Frobenius norm.}
A Hierarchical Spatio-temporal Model of House Price Changes

In this section we consider spatial and temporal relationships in the form of a hierarchy. Let \( \pi_{jrt} \) denote the rate of change of some local feature\(^2\) in the \( j \)th locality in region \( r = 1, 2, \ldots, R \), and consider the following hierarchical factor model

\[
\pi_{jrt} = a_{jr} + \beta_{jr} \bar{\pi}_t + \gamma_{jr} \bar{\pi}_t + \xi_{jrt},
\]

\( j = 1, 2, \ldots, N_r; r = 1, 2, \ldots, R; t = 1, 2, \ldots, T, \)

where \( \bar{\pi}_t = N_r^{-1} \sum_{j=1}^{N_r} \pi_{jrt} \), and \( \bar{\pi}_t = N^{-1} \sum_{r=1}^{R} \sum_{j=1}^{N_r} \pi_{jrt} \), with \( N = \sum_{j=1}^{R} N_r \). Write the above model more compactly as

\[
\pi_t = a + \mathbf{B} \mathbf{R}_N \pi_t + \Gamma \mathbf{P}_N \pi_t + \xi_t,
\]

where \( \pi_t \) is an \( N \times 1 \) vector of the local feature, partitioned by regions with the local feature in region 1 included first, followed by the local feature in the second region and so on, namely

\[
\pi_t = (\pi_{11t}; \pi_{21t}; \ldots \pi_{N_11t}; \pi_{12t}; \pi_{22t}; \ldots \pi_{N_22t}; \ldots; \pi_{1Rt}; \pi_{2Rt}; \ldots; \pi_{N_RRt})^\prime.
\]

Similarly

\[
a = (a_{11}; a_{21}; \ldots a_{N_11}; a_{12}; a_{22}; \ldots a_{N_22}; \ldots; a_{1R}; a_{2R}; \ldots; a_{N_RR})^\prime.
\]

\( \mathbf{B} \) and \( \Gamma \) are an \( N \times N \) diagonal matrices with their ordered elements given by

\[
\beta_{11}; \beta_{21}; \ldots \beta_{N_11}; \beta_{12}; \beta_{22}; \ldots \beta_{N_22}; \ldots; \beta_{1R}; \beta_{2R}; \ldots; \beta_{N_RR},
\]

and

\[
\gamma_{11}; \gamma_{21}; \ldots \gamma_{N_11}; \gamma_{12}; \gamma_{22}; \ldots \gamma_{N_22}; \ldots; \gamma_{1R}; \gamma_{2R}; \ldots; \gamma_{N_RR},
\]

respectively. Finally, \( \mathbf{R}_N \) and \( \mathbf{P}_N \) are \( N \times N \) projection matrices such that \( \mathbf{R}_N \pi_t \) give the regional means and \( \mathbf{P}_N \pi_t \) the national mean of the local feature. More specifically, let \( \boldsymbol{\tau}_{N_r} \) be an \( N_r \times 1 \) vector of ones, and \( \boldsymbol{\tau}_N \) an \( N \times 1 \) vector of ones, then

\[
\mathbf{P}_N = \boldsymbol{\tau}_N (\boldsymbol{\tau}_N^\prime \boldsymbol{\tau}_N)^{-1} \boldsymbol{\tau}_N^\prime,
\]

and

\[
\mathbf{R}_N = \begin{pmatrix}
\mathbf{P}_{N_1} & 0 & \ldots & 0 & 0 \\
0 & \mathbf{P}_{N_2} & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & \mathbf{P}_{N_{R-1}} & 0 \\
0 & 0 & \ldots & 0 & \mathbf{P}_{N_R}
\end{pmatrix},
\]

where \( \mathbf{P}_{N_r} = \boldsymbol{\tau}_{N_r} (\boldsymbol{\tau}_{N_r}^\prime \boldsymbol{\tau}_{N_r})^{-1} \boldsymbol{\tau}_{N_r}^\prime \). It is assumed that \( R \) is fixed, and for each \( r, N_r/N \) tends to a non-zero constant as \( N \to \infty \). \( \mathbf{P}_N \pi_t \), for \( r = 1, 2, \ldots, R \), and \( \mathbf{P}_N \pi_t \) can be

\[\max_{1 \leq t \leq T} \left\{ \sum_{j=1}^{N} |a_{ij}| \right\} \text{ for its maximum absolute row sum norm.}\]

\( \text{In this paper we consider real house prices at the level of Metropolitan Statistical Areas but it can also refer to other economic characteristics such as employment, unemployment, wages, incomes etc.} \)
viewed as regional and national factors that are consistently estimated by simple averages. They also represent the strong form of cross-sectional dependence in the local feature.

To model the remaining weak cross-sectional dependence in the local feature we shall consider the following spatial model for \( \xi_t = (\xi_{1t}, \xi_{2t}, \ldots, \xi_{N_t})' \)

\[
\xi_{it} = a \xi + \sum_{j=1}^{p_i} \lambda_{ij} \xi_{it-j} + \sum_{j=0}^{q_i} \psi_{ij} \xi_{it-j} + \zeta_{it}, \quad \text{for } i = 1, 2, \ldots, N,
\]

where

\[
\xi_{it} = \frac{w_i' \xi}{w_i' \tau_N}, \text{ if } w_i' \tau_N > 0,
\]

\[= 0 \text{ if } w_i' \tau_N = 0,
\]

and \( w_i \) denote the \( i \)th row of the \( N \times N \) spatial matrix \( W \). Here we assume that \( W \) is given and return to a discussion of how best it can be estimated later. Writing the above model in matrix notation we also have

\[
\xi_t = a \xi + \sum_{j=1}^{p} \Lambda_j \xi_{t-j} + \sum_{j=0}^{q} \Psi_j W \xi_{t-j} + \zeta_t,
\]

where \( p = \max(p_1, p_2, \ldots, p_N), \quad q = \max(q_1, q_2, \ldots, q_N), \quad \Lambda_j \) and \( \Psi_j \) are \( N \times N \) diagonal matrices with \( \lambda_{ij} \) and \( \psi_{ij} \) over \( i \) as their diagonal elements, and \( \zeta_t = (\zeta_{1t}, \zeta_{2t}, \ldots, \zeta_{Nt})' \). The above specification can be readily generalized to a spatial (network) model where we draw a distinction between positive (complementary) and negative (substituting) connections, namely

\[
\xi_t = a \xi + \sum_{j=1}^{p} \Lambda_j \xi_{t-j} + \sum_{j=0}^{q^+} \Psi_j^+ W^+ \xi_{t-j} + \sum_{j=0}^{q^-} \Psi_j^- W^- \xi_{t-j} + \zeta_t,
\]

where \( W^+ \) and \( W^- \) are \( N \times N \) network matrices for positive and negative connections. It is common in the literature to think of spatial relationships as involving spillovers from one area to another with the (implicit) assumption that the spillovers are positive. But this need not be the case. Migration across space could raise/lower wages or house prices in one locality and lower/raise them into another locality.

### 3 The Spatial Econometric Model

The standard spatial econometric model (Anselin, 1988) can be written:

\[
x_{ot} = \psi W x_{ot} + u_{ot}, \quad (3)
\]

where \(|\psi| < 1, \quad x_{ot} = (x_{1t}, \ldots, x_{Nt})', \quad u_{ot} = (u_{1t}, \ldots, u_{Nt})', \quad \beta \) is a \( m \times 1 \) vector of coefficients, and \( W \) is the spatial weight matrix. The error terms \( u_{ot} \) are assumed to be identically and independently normally distributed with zero mean and variances collected in matrix \( \Sigma = \text{Diag}(\sigma_o^2) \), where \( \sigma_o^2 = (\sigma_1^2, \ldots, \sigma_N^2)' \). Hence, (3) can be re-written as

\[
x_{ot} = \psi W x_{ot} + \Sigma^{1/2} \varepsilon_{ot}, \quad (4)
\]

where \( \varepsilon_{ot} = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})' \) and \( \varepsilon_{it} \sim IIDN(0,1) \), for \( i = 1, \ldots, N \). This is commonly referred to in the spatial econometrics literature as a spatial autoregressive model, or a
spatially lagged dependent variable model. The usual approach is to specify the $W$ matrix and then estimate equation (3) directly. One possible problem with this is that apparent cross sectional dependence that is meant to be captured by $W$ may actually be due, in part, to common effects from the exogenous variables. We propose a method below for first purging the data of common factors using a factor model and then focusing attention on the resulting residuals to identify possible spatial patterns.

3.1 Cross-sectional and spatial dependence in panels

3.1.1 Spatial dependence - as a form of weak CSD

Conventionally, spatial dependence is characterized by use of a predetermined metric such as space or ‘economic distance’ (Lee and Pesaran (1993), Conley and Dupor (2003), and Conley and Topa (2003), Pesaran, Schuermann and Weiner (2002) and review of spatial econometrics by Ansell (2001)). However, often in economic applications these may not be appropriate measures. In some instances trade flows might be relevant, whilst in the case of inter-industry dependencies input-output matrices might provide the appropriate spatial metric (Holly and Petrella (2012)). Alternatively, there may be dependencies between geographical areas that reflect cultural similarity, and migration or commuting relationships.

Irrespective of the measure used spatial dependence relates to spill-over effects that are not pervasive in nature. In other words it conforms to the notion of cross-sectional weak dependence (CWD) as defined in Chudik, Pesaran and Tosetti (2011). To see why, consider as an example the first-order spatial autoregressive, SAR(1), model

$$\mathbf{x}_t = (\mathbf{I}_N - \psi \mathbf{W})^{-1} \Sigma^{1/2} \varepsilon_{ot},$$

where $|\psi| < 1$, $\mathbf{x}_t = (x_{1t}, ..., x_{Nt})'$, $\varepsilon_{ot} = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$, $\Sigma$ is an $N \times N$ diagonal matrix with $\sigma_i^2 < K < \infty$ on its $i$th diagonal, and $\mathbf{W}$ is the spatial weight matrix. Both $\Sigma$ and $\mathbf{W}$ can depend on time but we assume them to be time invariant. The correlation matrix for this model is given by

$$\mathbf{R} = (\rho_{ij}) = \Sigma^{-1/2} (\mathbf{I}_N - \psi \mathbf{W})^{-1} \Sigma (\mathbf{I}_N - \psi \mathbf{W}')^{-1} \Sigma^{-1/2},$$

or

$$\mathbf{R} = [\mathbf{I}_N - \psi (\mathbf{G} + \mathbf{G}') + \psi^2 \mathbf{G}' \mathbf{G}]^{-1},$$

where

$$\mathbf{G} = \Sigma^{-1/2} \mathbf{W} \Sigma^{1/2}.$$ 

To simplify the exposition we also assume that $\mathbf{W}$ is symmetric and row standardized such that $\|\mathbf{W}\|_1 = 1$. Thus, in this setting

$$\mathbf{R} = [\mathbf{I}_N - 2\psi \mathbf{G} + \psi^2 \mathbf{G}^2]^{-1} = \mathbf{I}_N + 2\psi \mathbf{G} + 3\psi^2 \mathbf{G}^2 + 4\psi^3 \mathbf{G}^3 + \ldots.$$ 

Interactions in social networks can also be ‘spatial’ in an entirely abstract sense. For example Bhattacherjee and Holly (2012) explore interactions among members of a committee using a spatial analogy.
The degree of cross-sectional dependence among the $N$ units can be conveniently summarised by their average cross-correlation (excluding the diagonal elements),

$$\bar{\rho}_N = \frac{\tau' R \tau - N}{N (N - 1)} = \frac{\tau' R \tau}{N (N - 1)} - \frac{1}{N - 1},$$

(7)

where $\tau$ is an $N \times 1$ vector of ones. In the case of cross-sectional independence $\bar{\rho}_N \to 0$ since $\frac{\tau' R \tau}{N (N - 1)} = O\left(\frac{1}{N^2}\right)$. In the case of the spatial autoregressive, SAR(1), model we use (6) in (7) so that

$$\tau' R \tau = N + 2 \psi \tau' G \tau + 3 \psi^2 \tau' G^2 \tau + \ldots,$$

(8)

where

$$\tau' R \tau \leq \tau' \lambda_1 (R).$$

We have that

$$\lambda_1 (R) \leq \sqrt{\|R\|_1 \|R\|_\infty} \leq \|R\|_1,$$

where

$$\|R\|_1 \leq 1 + 2 |\psi| \|G\|_1 + 3 |\psi|^2 \|G^2\|_1 + \ldots$$

$$\leq 1 + 2 |\psi| \|G\|_1 + 3 |\psi|^2 \|G\|_1^2 + \ldots$$

$$= 1 + 2 |\psi| + 3 |\psi|^2 + \ldots$$

$$= \frac{1}{(1 - |\psi|)^2},$$

and using the fact that

$$\|G\|_1 = \|\Sigma^{1/2} W \Sigma^{1/2}\|_1 \leq \|W\|_1 = 1.$$

More precisely, from (8) we get

$$\tau' G^j \tau \leq \tau' \lambda_1 (G^j) < N [\lambda_1 (G)]^j, \text{ for } j = 1, 2, \ldots$$

and setting $\vartheta = \lambda_1 (G)$ we arrive at

$$\frac{\tau' R \tau}{N} \leq \frac{1}{(1 - |\psi \vartheta|)^2},$$

where $|\psi \vartheta| < 1$. Finally,

$$\bar{\rho}_N \to 0,$$

indicating that $x_{st}$ is cross-sectionally weakly correlated. Note however that in practice $\|R\|_1$ will be changing with $N$ and for finite $N$ the value of $\bar{\rho}_N$ can exceed 0 sometimes by a large amount, particularly if $\psi$ is close to unity.

### 3.1.2 Cross-sectional weak and strong dependence

This last result raises the question of how much can $\bar{\rho}_N$ exceed zero before $x_{st}$ ceases to be cross-sectionally weakly dependent and instead is considered a cross-sectionally strongly correlated data set. To this end we draw from the analysis in Pesaran (2013) and consider that $x_{st}$ are generated from the following factor model

$$x_{it} = \gamma_i f_t + \varepsilon_{it},$$

(9)
where \( f_i = (f_{1i}, f_{2i}, \ldots, f_{mi})' \) is the \( m \times 1 \) vector of unobserved common factors (\( m \) being fixed) with \( E(f_i) = 0 \), and \( \text{Cov}(f_i) = I_m \). \( \gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{im})' \) is the associated vector of factor loadings, and \( \varepsilon_{it} \) are idiosyncratic errors that are cross-sectionally and serially independent with variance \( \omega_i^2 \), namely \( \varepsilon_{it} \sim IID(0, \omega_i^2) \). The degree of cross-sectional dependence of \( x_{it} \), is governed by the rate at which the average pair-wise error correlation coefficient, \( \overline{\rho}_N = [2/N(N-1)]\sum_{i=1}^{N}\sum_{j=i+1}^{N} \rho_{ij} \), tends to zero in \( N \), where \( \rho_{ij} = \text{Corr}(x_{it}, x_{jt}) \).

In the case of the above factor model we have, \( \text{Var}(x_{it}) = \sigma_i^2 = \omega_i^2 + \gamma_i' \gamma_i \), \( \rho_{ij} = \delta_i' \delta_j \), for \( i \neq j \), where

\[
\delta_i = \frac{\gamma_i}{\sqrt{1 + \gamma_i' \gamma_i}}. \tag{10}
\]

Then it is easily seen that

\[
\overline{\rho}_N = \left( \frac{N}{N-1} \right) \left( \overline{\delta}_N' \overline{\delta}_N - \frac{\sum_{i=1}^{N} \delta_i' \delta_i}{N^2} \right). \tag{11}
\]

Consider now the effects of the \( j^{th} \) factor, \( f_{jt} \), on the \( i^{th} \) unit, \( x_{it} \), as measured by \( \gamma_{ij} \), and suppose that these factor loadings take non-zero values for \( M_j \) out of the \( N \) cross-section units under consideration. Then following Bailey, Kapetanios and Pesaran (2012 - BKP), the degree of cross-sectional dependence due to the \( j^{th} \) factor can be measured by \( \alpha_j = \ln(M_j)/\ln(N) \), and the overall degree of cross-sectional dependence of units \( x_{it} \) by \( \alpha = \max_j(\alpha_j) \). \( \alpha \) is the exponent of \( N \) that gives the maximum number of \( x_{it} \) units, \( M = \max_j(M_j) \), that are pair-wise correlated. The remaining \( N - M \) units will only be partially correlated. BKP refer to \( \alpha \) as the exponent of cross-sectional dependence. \( \alpha \) can take any value in the range 0 to 1, with 1 indicating the highest degree of cross-sectional dependence. Considering that \( \gamma_i' \gamma_i = O(m) \) where \( m \) is fixed as \( N \to \infty \), the exponent of cross-sectional dependence of \( x_{it} \) units can be equivalently defined in terms of the scaled factor loadings, \( \delta_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{im})' \). Without loss of generality, suppose that only the first \( M_j \) elements of \( \delta_{ij} \) over \( i \) are non-zero, and note that\(^4\)

\[
\tilde{\delta}_{j,N} = \frac{1}{N} \left( \sum_{i=1}^{M_j} \delta_{ij} + \sum_{i=M_j+1}^{N} \delta_{ij} \right) = \frac{M_j}{N} \left( M_j^{-1} \sum_{i=1}^{M_j} \delta_{ij} \right) = N^{\alpha_j-1} \mu_j = O(N^{\alpha_j-1}),
\]

where \( \mu_j = \left( M_j^{-1} \sum_{i=1}^{M_j} \delta_{ij} \right) \neq 0 \), for a finite \( M_j \) and as \( M_j \to \infty \). Similarly, \( N^{-2} \sum_{i=1}^{N} \delta_{ij} \delta_{ij} = O(N^{2\alpha_j-2}) \), and using (11) we have

\[
\overline{\rho}_N = O(N^{2\alpha-2}).
\]

The values of \( \alpha \) in the range \([0, 1/2)\) correspond to different degrees of cross-sectional weak dependence, as compared to values of \( \alpha \) in the range \((1/2, 1]\) that relate to different degrees of cross-sectional strong dependence. Under the SAR(1) model specification of (5) \( \overline{\rho}_N \to 0 \) and \( \|R\|_1 = O(1) \), indicating that \( \alpha \) must reside in the range \([0, 1/2)\).

### 3.1.3 Identifying the degree of cross-sectional dependence

From the above exposition it is evident that in many applications the data considered might be "contaminated" with factors that cause much stronger cross-sectional dependence than what is accounted for in a spatial setting. In other words, mistaking (9) for a

\(^4\)The main results in the paper remain valid even if \( \sum_{i=M_j+1}^{N} \delta_{ij} = O(1) \). But for expositional simplicity we maintain the assumption that \( \sum_{i=M_j+1}^{N} \delta_{ij} = 0 \).
spatial specification can lead to grossly inaccurate estimates of cross-sectional dependence. This brings us to the question of identifying the strength of cross dependence.

Suppose observations \( x_{it}, i = 1, 2, ..., N \) and \( t = 1, 2, ..., T \) are available and the aim is to model how \( x_{it} \) and \( x_{jt} \) are dependent, across all \( i \) and \( j \), with \( N \) and \( T \) relatively large. The first step in modelling cross-sectional or spatial dependence is to find out if the observations are cross-sectionally weakly or strongly dependent. One should only consider the application of spatial methods if the cross-sectional exponent of the observations, \( \alpha \), is sufficiently small, and particularly not close to unity. Temporal dependence can be modelled through common factors or unit-specific dynamics, using autoregressive distributed lag models, or GVAR specifications (Dees et al., 2007).

A multi-step procedure is as follows:

1. Apply the cross section dependence (CD) test developed in Pesaran (2013) to \( x_{it} \). Only proceed to spatial modelling if the null of weak cross dependence is not rejected.

2. If the null of weak dependence is rejected, model the (semi-) strong dependence by use of factor models or cross section averages, and check that the residuals, denoted by \( e_{it} \), are weakly cross-correlated (now by applying the CD test to \( e_{it} \)).

3. Identify local neighbours for spatial weight matrix \( W \) and apply spatial or network modelling techniques to \( e_{it} \), to model the residual weak dependencies.

To test for weak or spatial dependence, denote the sample pair-wise correlations of \((i, j)\) units by

\[
\hat{\rho}_{ij} = \hat{\rho}_{ji} = \frac{\sum_{t=1}^{T} (x_{it} - \bar{x}_i) (x_{jt} - \bar{x}_j)}{(\sum_{t=1}^{T} (x_{it} - \bar{x}_i)^2)^{1/2} (\sum_{t=1}^{T} (x_{jt} - \bar{x}_j)^2)^{1/2}},
\]

(12)

when the test is applied to \( x_{it} \), and \( \bar{x}_i = N^{-1} \sum_{t=1}^{N} x_{it} \). The CD statistic is then defined by

\[
CD = \left[ \frac{TN(N-1)}{2} \right]^{1/2} \hat{\rho}_N,
\]

(13)

where

\[
\hat{\rho}_N = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \hat{\rho}_{ij}.
\]

(14)

Pesaran (2013) shows that \( CD \to N(0,1) \), under the null hypothesis that cross-sectional exponent of \( x_{it} \) is \( \alpha < (2 - \epsilon) / 4 \), as \( N \to \infty \), such that \( T = \kappa N^\epsilon \), for some \( 0 \leq \epsilon \leq 1 \), and a finite \( \kappa > 0 \). If \( H_0 \) of weak dependence is rejected for \( x_{it} \) in step 1 of the multi-step procedure, then according to BKP the exponent of cross-sectional dependence, \( \alpha \), can be estimated. There are different ways of estimating the exponent of cross-sectional dependence if \( 1/2 \leq \alpha < 1 \). We refer to Appendix I for details of the estimation of \( \alpha \) under BKP. Once step 2 is established then we can be confident that our data have been sufficiently stripped of excess common cross dependence and step 3 can follow.
4 Data-driven spatial weight matrix $W$

We now revert back to (5) and focus our attention on the spatial weight matrix $W$. Typically, this is constructed using extraneous information, such as geodesic, demographic or economic information, not contained in the data set under consideration, $x_{it}$. In economic applications, economic measures, such as commuting times, trade and migratory flows across geographical areas have been used. For example, in GVAR modelling trade weights are used in the construction of cross-sectional averages (Pesaran et al. 2004). Alternatively geographical contiguity can be used as in Holly et al (2011). Such measures are often preferable over geodesic measures - since they are closer to the decisions that underlie the observations, $x_{it}$, and they allow also for possible time variations in the weighting matrix which of course is not possible if we use only measures of distance.

In practice, it is often difficult to obtain appropriate measures of economic distance for the analysis of interdependencies. It is, therefore, desirable to see if $W$ can be constructed without recourse to such extraneous information. In applications where the time dimension is reasonably large (around 50-80), it is possible to identify the non-zero elements of $W$ with those elements of $\hat{\rho}_{ij}$, as expressed in (12), that are different from zero at a suitable significance level. But since there are a large number of such statistical tests, multiple testing procedures that control the overall size of the tests have to be used.

The multiple testing problem arises when we are faced with a number of (possibly) dependent tests and our aim is to control the size of the overall test. Suppose we are interested in a family of null hypotheses, $H_{01}, H_{02}, ..., H_{0m}$ and we are provided with corresponding test statistics, $Z_{1T}, Z_{2T}, ..., Z_{mT}$, with separate rejection rules given by (using a two sided alternative)

$$
\Pr (|Z_{iT}| > CV_{iT} | H_{0i}) \leq p_{iT},
$$

where $CV_{iT}$ is some suitably chosen critical value of the test, and $p_{iT}$ is the observed $p$ value for $H_{0i}$.

Consider now the family-wise error rate (FWER) defined by

$$
FWER_T = \Pr (\bigcup_{i=1}^{m} (|Z_{iT}| > CV_{iT} | H_{0i})），
$$

and suppose that we wish to control $FWER_T$ to lie below a pre-determined value, $p$. Bonferroni provides a general solution, which holds for all possible degrees of dependence across the separate tests. By Boole’s inequality we have

$$
\Pr (\bigcup_{i=1}^{m} (|Z_{iT}| > CV_{iT} | H_{0i})) \leq \sum_{i=1}^{m} \Pr (|Z_{iT}| > CV_{iT} | H_{0i})
$$

$$
\leq \sum_{i=1}^{m} p_{iT}.
$$

Hence to achieve $FWER_T \leq p$, it is sufficient to set $p_{iT} \leq p/m$. However, Bonferroni’s procedure can be quite conservative, particularly when the tests are highly correlated. This means that the procedure does not reject hypotheses as often as it should and therefore lacks power. A step-down procedure is proposed by Holm (1979) which is more powerful than Bonferroni’s procedure, without imposing any further restrictions on the degree to which the underlying tests depend on each other.
If we abstract from the \( T \) subscript and ordered the \( p \)-values of the tests, so that
\[
p(1) \leq p(2) \leq \ldots \leq p(m),
\]
are associated with the null hypotheses, \( H_{(01)}, H_{(02)}, \ldots, H_{(0m)} \), respectively. Holm’s procedure rejects \( H_{(01)} \) if \( p(1) \leq p/m \), rejects \( H_{(01)} \) and \( H_{(02)} \) if \( p(2) \leq p/(m - 1) \), rejects \( H_{(01)}, H_{(02)} \) and \( H_{(03)} \) if \( p(3) \leq p/(m - 2) \), and so on.

In our application, there is also the issue as to whether to apply the multiple testing procedures to all distinct \( N(N - 1)/2 \) non-diagonal elements of \( R = (\rho_{ij}) \) simultaneously, or to apply the procedures row-wise, by considering \( N \) separate families of \( N - 1 \) tests defined by \( \rho_{ij} = 0 \), for a given \( i^0 \), and \( j = 1, 2, \ldots, N, j \neq i^0 \). Given our objective (identification of the neighbours of the \( i \)th unit), the row-wise approach seems more appropriate and less conservative as compared to considering all \( N(N - 1)/2 \) tests simultaneously.

Hence, we introduce the following thresholding procedure for the construction of a suitable spatial weights matrix \( W \). This is applied to the elements of the estimated \( R \), say \( \hat{R} \). Let us take the case of \( i^0 \) MSA and consider the \( N - 1 \) pairwise correlation coefficients, \( \hat{\rho}_{i^0 j} \), for \( i^0 \neq j = 1, 2, \ldots, N \). Under the null that \( i^0 \) and \( j \) are unconnected, \( \hat{\rho}_{i^0 j} \sim N(0, T^{-1}) \). Using the Bonferroni procedure, the size of the \( N - 1 \) separate tests is controlled at \( \alpha \) (which we set at 5\%) if \( \hat{\rho}_{i^0 j} = 0 \) is rejected when
\[
|\hat{\rho}_{i^0 j}| > T^{-1/2} \Phi^{-1} \left( 1 - \frac{\alpha}{2(N - 1)} \right) = \psi(p, N, T).
\]
\( \Phi^{-1}(.) \) represents the inverse of the cumulative distribution function of the standard normal variate. Then,
\[
\hat{w}_{i^0 j} = I( |\hat{\rho}_{i^0 j}| > \psi(p, N, T)),
\]
for \((i^0, j)\) of \( \hat{W} = (\hat{w}_{ij}) \), where \( I(.) \) is an indicator function.

Other multiple testing procedures can also be considered and Efron (2010) provides a recent review. But most of these procedures are not suitable for our row-wise application and tend to place undue restrictions on the dependence of the underlying test statistics. Although the Holm procedure is more powerful, unfortunately it can lead to contradictions if applied row-wise. To see this consider the simple case where \( N = 3 \) and the \( p \)-values for the three rows of \( R \) are given by
\[
\begin{pmatrix}
- & p_1 & p_2 \\
p_1 & - & p_3 \\
p_2 & p_3 & -
\end{pmatrix}.
\]
Suppose that \( p_1 < p_2 < p_3 \). Then \( \rho_{13} = 0 \) is rejected if \( p_2 < p \) when Holm’s procedure is applied to the first row, and rejects \( \rho_{13} = 0 \), if \( p_2 < p/2 \) when the procedure is applied to the third row. The row-wise application of Bonferroni’s procedure is not subject to this problem - since it applies the same \( p \)-value of \( p/(N - 1) \) to all elements of \( R \).

5 US house prices

Interactions between regions in the US have been studied by many including Cromwell (1992), Pollakowski and Ray (1997), Carlino and DeFina (1998, 2004), Carlino and Sill (2001), Del Negro (2002), Owyang, Piger and Wall (2005) and Partridge and Rickman
We opt to study house price data at the level of the Metropolitan Statistical Area. Partly because of the availability of spatially disaggregated data, but also because of the role of housing in household wealth and in the transmission of monetary policy (and more recently as the conduit for a major global shock) there is a large literature on the spatial dimension of, particularly, house prices. House price shocks can spillover into adjacent geographical areas and ripple across the economy Meen (1999), Holly, Pesaran and Yamagata (2011).

Metropolitan Statistical Areas\(^5\) (MSAs) are large urban concentrations. They range in size, according to population in 2008, from the smallest - Carson City - with a population of 55,000, to New York and its environs with a population of 18.97 million. Moreover, there can be considerable distances between MSAs. The pairwise average distance is 1156 miles, though of course this is exaggerated by the relative sparseness of the distribution of MSAs in the Midwest. Indeed, by comparison the study of regional house prices in the UK by Holly, Pesaran and Tamagata (2011b) deals with distances of a much smaller magnitude. Distance is, therefore, likely to be an important factor for the spatial distribution of house prices, though size could also play a role.

Our choice of house prices is motivated by the role that housing plays in spatial equilibrium models (Glaeser et al., 2008, Glaeser and Gottlieb., 2009). The standard approach in urban and regional economics is to assume a spatial equilibrium. At the margin firms and households have to be indifferent between different locations. Firms employ labour up to the point at which the wage is equal to marginal product; construction companies supply housing up to the point at which marginal cost is equal to marginal product. Finally households have to be indifferent about where they are located, taking into account wages, the price of houses and the local availability of amenities (proximity to sea, mountains, temperature, etc). The combination of the labour supply curve, the supply curve for housing and the labour demand determines simultaneously the population of say a locality and then wages and the price of housing. Idiosyncratic differences in space in terms of productivity, particular characteristics of an area and the construction sector determine differences across space in population density, household incomes and the price of houses. There are a number of equilibrating processes. Households will move between areas in response to differences in wages, house prices and characteristics of an area. There can also be agglomeration effects because productivity rises with the size of the city. Clearly this equilibrated process has to be a long run phenomenon. It takes time for households to relocate in response to changing economic circumstances. It also takes time (time to build) for construction companies to increase the supply of housing.

In the spatial literature interactions between spatial units are described by a spatial weights matrix \(W\) which is formed using a metric specified \textit{a priori} by the researcher. We take an alternative approach and seek to discern the weighting matrix from the data itself. A common approach is to use contiguity or geographic proximity. So if a spatial unit is next to another unit the weight is unity, and zero otherwise. However, because of the way in which they are defined, very rarely are MSAs geographically next door. There is the added complication that there may be strong factors common to spatial units that account for a perceived correlation between units. So as to be able to uncover genuine spatial

---

\(^5\)Metropolitan statistical areas are geographic entities delineated by the Office of Management and Budget and are used by Federal statistical agencies when collecting, tabulating, and publishing Federal statistics for spatial units in the USA. The MSAs are defined by a core area with a large population concentration, together with adjacent areas that have a high degree of economic and social integration with that core through commuting and transport links.
correlations we first purge the data of common factors at both the national and regional level. We then threshold the matrix of correlations in house prices after extracting the common factors in order to construct the weights matrix $W$. In what follows we provide a detailed analysis of the ideas involved in the proposed methodology and how these translate into the study of US house prices.

5.1 Spatial dependence in US house prices

We turn to our empirical study of identifying the spatial structure of US house prices where we make use of the methods discussed above. We consider 363 Metropolitan Statistical Areas in total, excluding three MSAs located in Alaska and Hawaii. Note that the District of Columbia is treated as a single MSA. The sample period is 1975Q1 to 2010Q4. We denote house prices in MSA $j$ located in state $s$ in quarter $t$ by $P_{jst}$, for $j = 1, ..., N_s$, $s = 1, ..., S$, $t = 1, ..., T$, where $\sum_{s=1}^{S} N_s = N = 363$ MSAs, $S = 49$ states (comprised of 48 states and the district of Columbia), and $T = 144$ quarters. We then compute real house prices as:

$$ p_{jst} = \ln \left( \frac{P_{jst}}{CPI_{st}} \right), \text{ for } j = 1, 2, ..., N_s; \ s = 1, ..., S; \ t = 1, 2, ..., T, $$

where $CPI_{st}$ is the Consumer Price Index of state $s$ in quarter $t$. Details on the sources of these data can be found in Appendix II. We have ordered the MSAs by State, starting at the East Coast and moving towards the West Coast, following the list in Table 6 of Appendix III.

We continue by regressing $p_{jst} - p_{jst-1}$, the rate of change in real house prices, on an intercept and three quarterly seasonal dummies to obtain the seasonally adjusted series, $\hat{\pi}_{jst}$, as residuals from:

$$ p_{jst} - p_{jst,t-1} = c_{js} + \delta_{1js}d_{1t} + \delta_{2js}d_{2t} + \delta_{3js}d_{3t} + \epsilon_{jst}, \text{ for } j = 1, 2, ..., N_s; \ s = 1, ..., S; \ t = 2, ..., T. \quad (16) $$

5.2 Spatial weights matrix based on distance

First, we use distance between MSAs as a measure of contiguity. Metropolitan Statistical Areas are deliberately defined as one or more large cities at the core which through their effect on commuting patterns also have a substantial influence over the surrounding region. So there is always geographical space between one MSA and another. In order to construct adjacency matrix $W$ we make use of data for geodesic distance which are derived from applying the Haversine formula to data on Latitude-Longitude of zip codes, cross referenced to each of the 363 MSAs (more details about the methodology can be found in Appendix IV). We consider $W$, of dimension $N \times N$, to be symmetric in our analysis. We identify as neighbours for each MSA, $i$ ($i = 1, ..., N$), all MSAs that lie within a radius of $M$ miles. This pattern translates into a value of 1 for elements $(i, j)$ and $(j, i)$ of $W$ if MSA $i$ is a neighbour (falling within the given radius) of MSA $j$, or a value of 0 otherwise. Diagonal entries $(i, i)$ take a value of 0 as well, indicating that MSA $i$ can not be a neighbour of itself.

We study three cases: (i) MSAs within a radius of $M = 100$ miles, (ii) MSAs within the radius of $M = 200$ miles, and (iii) MSAs within a radius of $M = 300$ miles. This gives rise to three $W$ matrices, namely (i) $W_{100m}$, (ii) $W_{200m}$, and (iii) $W_{300m}$, which are
sparse by nature but of a different degree depending on the cut-off point set by the radius. We compare the degree of sparseness of $W_{100m}$, $W_{200m}$, and $W_{300m}$ by the percentage of non-zero elements in each over the total number of elements in the relevant $W$ ($N(N-1)$ elements overall). This gives a total of 1.55% of non-zero elements in $W_{100m}$, 5.38% of non-zero elements in $W_{200m}$, and 10.43% of non-zero elements in $W_{300m}$ (all versions of $W$ are of dimension $363 \times 363$). As expected, the number of non-zero elements increases when the radius within which MSAs are considered to be neighbours, rises. As a visual aid, in figure 1 we depict all three $W$ matrices. In this figure we have ordered the MSAs by State as described in the previous section, from top to bottom and from left to right. The sparseness of the $W$ matrices is captured by white slots when the relevant entries are equal to zero. There is considerable clustering along the diagonal, as we would expect, but because we are using a line to depict a plane, sometimes an MSA may lie at the edge of a State (or region) and fall within the radius of another state or region. Clearly, as the radius is increased from 100 to 200 miles the amount of leaching increases.

**Figure 1: Weighting Matrices Specified by Distance**

<table>
<thead>
<tr>
<th>Radius=100miles</th>
<th>Radius=200miles</th>
<th>Radius=300miles</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="100miles Matrix" /></td>
<td><img src="image2" alt="200miles Matrix" /></td>
<td><img src="image3" alt="300miles Matrix" /></td>
</tr>
</tbody>
</table>

### 5.3 Steps 1 and 2: Examine degree of cross-sectional dependence

As explained in Section 3.1.3, before proceeding with the construction of a data driven spatial weights matrix $\hat{W}$, we need to ensure that any existence of cross-sectional strong dependence in the seasonally adjusted house price data, $\hat{\pi}_{jst}$ has been first stripped out. Ignoring the State within which each MSA is located, we compute the CD test of Pesaran (2013) as defined in (13) and (14) where $\hat{\rho}_{ij}$ is the pair-wise correlation coefficient of $\hat{\pi}_{it}$ and $\hat{\pi}_{jt}$. The resulting CD statistic clearly rejects the null of cross-sectional weak dependence ($CD = 640.46$). Very high cross-sectional dependence may be due not only to spatial factors but also common national and regional factors. Applying the method proposed by BKP we calculate the exponent of cross sectional dependence (standard error in parenthesis) and obtain $\alpha = 0.985 (0.03)$ which suggests cross-sectional strong or semi-strong dependence in real house price changes across MSAs. This implies that it would be inappropriate to apply spatial modelling techniques to $\hat{\pi}_{jst}$ as suggested in step 1 of the multi-step procedure of Section 3.1.3.

The effects of cross-sectional strong dependence to the change in real house prices can be modelled using observed (national / regional income, unemployment and interest rates), or unobserved common factors (using principal components). Alternatively, as
argued in Pesaran (2006), we can use cross-sectional averages at the national and regional level. More specifically, we run the following regressions

\[ \hat{\pi}_{jrt} = a_{jr} + \beta_{jr}\pi_{rt} + \gamma_{jr}\bar{\pi}_t + u_{jrt}, \quad j = 1, 2, ..., N_r; \quad r = 1, ..., R; \quad t = 2, ..., T, \]  

where \( \hat{\pi}_{jrt} \) denotes the rate of real house price changes (seasonally adjusted) in MSA \( j \) located in region \( r \) at quarter \( t \), \( \pi_{rt} \) is an \( R \times 1 \) vector where \( N_r \) corresponds to the number of MSAs in a certain region \( r \), \( \bar{\pi}_t \) is a \( R \times 1 \) vector, and \( N = \sum_{r=1}^{R} N_r \). We identify a total of \( R = 8 \) regions in the US containing an average of approximately 45 MSAs each. These are: (i) New England, (ii) Mid East, (iii) South East, (iv) Great Lakes, (v) Plains, (vi) South West, (vii) Rocky Mountains, and (viii) Far West (see Table 6 of Appendix III for more details).

The de-factored real house price changes are then given by residuals from (17), namely

\[ \hat{u}_{jrt} = \hat{\pi}_{jrt} - \hat{a}_{jr} - \hat{\beta}_{jr}\pi_{rt} - \hat{\gamma}_{jr}\bar{\pi}_t, \quad j = 1, 2, ..., N_r; \quad r = 1, ..., R; \quad t = 2, ..., T. \]

Then, in accordance to step 2 of the multi-step procedure of Section 3.1.3, we include all residuals \( \hat{u}_{jrt} \) from quarter \( t \) in a \( N \times 1 \) vector \( \hat{u}_t = (\hat{u}_{jrt}) \), \( j = 1, ..., N \) and we apply the CD test of Pesaran (2013) on \( \hat{u}_{jt} \) (we ignore subscript \( r \) momentarily). The resulting CD statistic is much reduced, standing at \(-6.05\). At first, this suggests that the de-factoring was not entirely successful in eliminating the effects of cross-sectional dependence. Since the CD test is rejected, if only marginally, we proceed to estimate the exponent of cross-sectional dependence as well. We obtain \( \hat{\alpha}_{cs} = 0.602 \) (0.03) which is close to the borderline between strong and weak dependence of 1/2 and suggests that if some trace of common effects remains then it is relatively insignificant.

We also compute the CD test on the residuals obtained from de-factoring using Principal Component Analysis. This entails running the following regressions

\[ z_{it} = a_i + \beta'_i\hat{f}_t + u_{it}, \quad i = 1, 2, ..., N; \quad t = 2, ..., T, \]

where \( \hat{f}_t \) is an \( m \times 1 \) vector where \( m \) corresponds to the number of principal components selected using the testing procedure developed by Bai and Ng (2002). \( \beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{iN})' \) are the factor loadings. The Bai and Ng test gives little guidance as to the number of principal components we should use. We would expect at most 4 strong factors to play an important role in the behaviour of house prices at an MSA level. Thus we set the maximum number of factors in the Bai-Ng test to 2, 3, 4 and 6 (in order to include all options) and it chooses the respective maximum in each case. We implement all six information criteria suggested in Bai and Ng (2002).

We repeat the de-factoring analysis using 2, 3, and 4 principal components as all alternatives are plausible. We obtain the residuals \( \hat{u}_{jt} \) and we compute the CD statistic in each instance. This gives 53.39, 10.21, and 2.73, respectively, when 2, 3 and 4 principal components are used. The corresponding exponents of cross-sectional dependence are \( \hat{\alpha}_{2pc} = 0.933 \) (0.04), \( \hat{\alpha}_{3pc} = 0.774 \) (0.04) and \( \hat{\alpha}_{4pc} = 0.726 \) (0.03) respectively.

Finally, whether we include national and regional cross-sectional averages or 2 to 4 factors in the de-factoring regressions, the resulting residuals have weak enough cross-sectional dependence to consider them amenable to spatial modelling. We can now carry on to identify local neighbours for all MSAs.

---

\(^6\)We also considered using State level averages, but there were only a few MSAs in some States.
5.4 Step 3: Identify local neighbours

We wish to construct a spatial weights matrix which is drawn from information contained in the house prices data set, in line with the procedure displayed in Section 4. To do this, first we obtain the sample correlation matrix of $\hat{u}_t, \hat{R}_u$, from the residuals of regression (17). $\hat{R}_u$ is calculated as

$$\hat{R}_u = \hat{D}_u^{-1/2} \hat{V}_u \hat{D}_u^{-1/2},$$

where $\hat{D}_u = diag(\hat{\sigma}_{u,ii}, i = 1, 2, \ldots, N)$, $\hat{V}_u = (\hat{\sigma}_{u,ij})$, $\hat{R}_u = (\hat{\rho}_{u,ij})$, with elements

$$\hat{\sigma}_{u,ij} = T^{-1} \sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt},$$

and

$$\hat{\rho}_{u,ij} = \hat{\rho}_{u,ji} = \frac{\hat{\sigma}_{u,ij}}{\sqrt{\hat{\sigma}_{u,ii}\hat{\sigma}_{u,jj}}}, i, j = 1, 2, \ldots, N, t = 2, \ldots, T.$$ 

Next, we apply the regularisation procedure of Section 4. Let us take the case of MSA $i$ and consider the $N - 1$ pairwise correlation coefficients, $\hat{\rho}_{u,ij}$, for $i \neq j = 1, 2, \ldots, N$. We apply thresholding-by-row as in (15) and obtain

$$\hat{w}_{u,ij} = I(\hat{\rho}_{u,ij} > \psi (p, N, T)),$$

for each $(i, j)$ of $\hat{W}_u = (\hat{w}_{u,ij})$.

We look at the degree of sparseness of $\hat{W}_u$ measured by the percentage of non-zero elements over the total number of elements in $\hat{W}_u$ ($N(N-1)$ elements overall). We have a total of 4.75% non-zero elements. In as far as the degree of sparseness is concerned, it seems that data-driven $\hat{W}_u$ is similar to $W_{200m}$, generated using the predetermined distance metric of a radius of 200 miles. Recall that both $\hat{W}_u$ and $W_{200m}$ are of dimension $363 \times 363$. However, this does not necessarily imply that the non-zero elements appearing in $\hat{W}_u$ are also identified in $W_{200m}$. A more formal comparison of the two versions of the spatial weights matrix is displayed in the next sub-section.

In Figure 2 below we graph $\hat{W}_u$. As in the cases of $W_{100m}, W_{200m},$ and $W_{300m}$, the sparseness of the matrix is captured by white slots when the relevant entries are equal to zero. Recall that the MSAs are ordered by State, moving from the East to West Coast, from top to bottom and left to right. Using the thresholding method, it appears that neighbours of each MSA extend well beyond geographical neighbours, though distinct clusters are evident especially in the West Coast and parts of the East Coast. Divisions of neighbourhoods into the East, the Middle and West of the country are also visible.

Figure 2: Thresholding Using National and Regional Cross-sectional Averages
Therefore, geographical proximity is not the only factor driving spatial connections between MSAs. There are significant correlations well away from the diagonal, with a number of clusters suggesting connections at considerable distances. We analyse these clusters in more depth in the next section.

We repeat thresholding this time on the residuals obtained from de-factoring using (19). As before, we measure sparseness by the percentage of non-zero elements over the total number of elements in \( \hat{W}_u \). The percentage of non-zero elements is 6.43%, 4.96% and 4.42%, respectively, for \( \hat{W}_{u,2pc} \), \( \hat{W}_{u,3pc} \) and \( \hat{W}_{u,4pc} \), where \( \hat{W}_{u,jpc} \), \( j = 2, 3, 4 \) correspond to the matrices generated by de-factoring using 2, 3 or 4 principal components. We compare these to the percentage of non-zero elements found in the distance matrices \( W_{100m} \), \( W_{200m} \) and \( W_{300m} \). It appears that either \( \hat{W}_{3pc} \) or \( \hat{W}_{4pc} \) are best approximations of \( W_{200m} \) using this metric. The matrices are plotted in figure 3 below. In this case, clustering is even more clear at the two edges of the country and to a lesser extend in different parts along the diagonal (equivalent to the middle of the U.S.). More interesting are patterns that emerge on the off diagonal. Again, division of the country into three distinct "neighbourhoods" is evident using this approach. Further, a small cluster is also evident at the bottom left (top right) edge of the matrices, indicating that the East and West Coast of the U.S. appear to be connected even though very far apart geographically.

**Figure 3: Thresholding using principal components**

![Thresholding using principal components](image)

5.4.1 Comparison of \( \hat{W}_u \) with \( W 

We assess the true closeness of the data-driven \( \hat{W}_u \) with the distance-based weight matrix \( W_M \) more formally by quantifying how many of the neighbours identified using one
method are also identified using the other approach. The assessment is complicated by the fact that these matrices are by nature sparse, and hence the probability of a zero realisation in both adjacency matrices $\tilde{W}_u$ and $W_M$ is higher than obtaining a common entry of 1. We concentrate on the upper triangular elements of $\tilde{W}_u$ and $W$ to avoid double-counting. Recall that both $\tilde{W}_u$ and $W_M$ have either entries of zero when MSA $i$ is not a neighbour of MSA $j$, and entries of one when MSA $i$ is a neighbour of MSA $j$.

We create contingency tables from these off-diagonal, upper-triangular elements of the form

$$
\begin{bmatrix}
n_{11} & n_{10} \\
n_{01} & n_{00}
\end{bmatrix},
$$

where:

- $n_{11}$ corresponds to the number of times $\tilde{W}_u$ displays entry of 1 when $W_M$ displays 1.
- $n_{00}$ corresponds to the number of times $\tilde{W}_u$ displays entry of 0 when $W_M$ displays 0.
- $n_{01}$ corresponds to the number of times $\tilde{W}_u$ displays entry of 0 when $W_M$ displays 1.
- $n_{10}$ corresponds to the number of times $\tilde{W}_u$ displays entry of 1 when $W_M$ displays 0.

Then, $n_{11} + n_{00} + n_{01} + n_{10} = N(N-1)/2$, and Pearson’s chi-squared statistic is given by

$$
\chi^2 = \frac{1}{2} N (N-1) \left[ \sum_{i,j=0}^{1} \frac{n_{ij}^2}{(n_i + n_j)} - 1 \right].
$$

We set the significance level at 5%. We compare $\tilde{W}_u$ with each version of $W_M$, namely $W_{100m}$, $W_{200m}$, and $W_{300m}$, and we get the following contingency tables

**Table 1: Contingency tables: Data-driven versus distance-based spatial weights matrices**

<table>
<thead>
<tr>
<th></th>
<th>$W_{100m}$</th>
<th></th>
<th>$W_{200m}$</th>
<th></th>
<th>$W_{300m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{W}_u$</td>
<td>1</td>
<td>139</td>
<td>2993</td>
<td>3132</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>881</td>
<td>61690</td>
<td>62571</td>
<td>0</td>
</tr>
<tr>
<td>$\sum_{rows}$</td>
<td>1020</td>
<td>64683</td>
<td>65703</td>
<td>$\sum_{rows}$</td>
<td>3542</td>
</tr>
</tbody>
</table>

Matrix $W_u$ results from defactoring real house price changes of each MSA by using regressions of these on regional and national cross-sectional averages.

The $\chi^2$ statistics are 179.18, 250.36, and 208.87 for $\tilde{W}_u$ versus $W_{100m}$, $W_{200m}$, and $W_{300m}$ respectively. The comparisons are highly significant at the 5% significance level (critical value is 3.841). $\tilde{W}_u$ is closest to $W_{200m}$.

Similar results are also obtained when using principal component analysis for defactoring. These are shown in Tables 8.5, 8.5, and 8.5 of Appendix V when two, three, or four principal components are used in the defactoring regressions (19).
An alternative approach that takes into account any existing serial correlation in the data is proposed in Pesaran and Timmermann (2009). This entails collecting all off-diagonal, upper-triangular elements of \( W_M \), \( M = 100, 200, 300 \) miles in one vector, \( y_i \), \( i = 1, ..., N(N-1)/2 \), and the corresponding elements of \( \hat{W}_u \) in another vector, \( x_i \), \( i = 1, ..., N(N-1)/2 \). Then, we run the following regression:

\[
y_i = \alpha + \beta x_i + \xi_i,
\]

where \( y_i = \text{vec}(W_{100m}), \text{vec}(W_{200m}) \), and \( \text{vec}(W_{300m}) \), and \( x_i = \text{vec}(\hat{W}_u) \) from the regression-based defactoring using cross-sectional averages. We investigate the independence of \( y_i \) and \( x_i \) in each case by testing \( \beta = 0 \). The resulting t-statistics are 13.74, 16.25, and 14.83 respectively. As before all comparisons are highly significant at the 5% significance level, with \( \hat{W}_u \) being closest to \( W_{200m} \).

### 5.5 Clusters in house prices

A number of studies have been made of the spatial patterns in US house prices. For example, Gupta and Miller (2012a) find that house prices in Los Angeles temporally cause house prices in Las Vegas and furthermore house prices in Las Vegas temporally cause house prices in Phoenix; but Los Angeles house prices do not cause Phoenix prices directly. Las Vegas house prices do not cause Los Angeles house prices and Phoenix house prices do not cause house prices in Las Vegas or Los Angeles. Los Angeles house prices, therefore, are weakly exogenous for Phoenix and Las Vegas. In a further study Gupta and Miller (2012b) examine causality between house prices in Southern California. Measured in terms of house prices at the MSA level Los Angeles is weakly exogenous for Bakersfield, Riverside, San Diego, and San Luis Obispo, the two inland MSAs and the most distant coastal MSAs. Moreover, Oxnard, San Diego, San Luis Obispo, Santa Ana, and Santa Barbara are weakly exogenous for Los Angeles. In other words, each coastal MSA is weakly exogenous for Los Angeles. However, this form of analysis ignores the possibility that there may be common national and regional factors that account for these correlations and failing to condition on the common factors may bias the inferences that can be drawn.

The analysis above strips out such effects so in what follows we analyse in more detail the correlation patterns in US house prices that are free of common factors. From Section 5.1 it is clear that proximity measured by distance is a useful metric for constructing a weighting matrix. However, the correlation analysis with thresholding also suggests that there are patterns that appear unrelated to geographical proximity. We focus here on the correlations obtained by regression (17).

The correlation matrix provides us with a very large number of correlations even after thresholding. We are interested in the connections between spatial units so one measure of connectiveness is the sum of the columns (or rows as the matrix is symmetric). We therefore took column sums of the squares of the correlation coefficients as a measure of the connectiveness of each MSA. In Table 2 we list the 20 most related MSAs and the 20 least related. Note that none of the MSAs that are most connected are the big conurbations (measured by population in 1993). Indeed if we compute Spearman’s rank correlation test we obtain a value of -0.2348 \( (p = 0000)^7 \). So there appears to be a significant negative relationship between population size and the strength of the

\(^7\)p-value in brackets.
relationship between MSAs. What this suggests is that because of the way in which the boundaries of MSAs are drawn, a lot of the connectiveness between areas has already been internalised to the MSA, especially for the largest. The borders of an MSA are drawn so that the connection between an adjacent but not contiguous MSA is likely to be weak. Many of the most connected by correlation are therefore smaller MSAs by population.

Table 2: Most and Least Connected MSAs

<table>
<thead>
<tr>
<th>Region</th>
<th>MSA</th>
<th>State</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>South West</td>
<td>Victoria</td>
<td>TX</td>
<td>104,672</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Holland</td>
<td>MI</td>
<td>200,647</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Fond du Lac</td>
<td>WI</td>
<td>92,556</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Sheboygan</td>
<td>WI</td>
<td>106,362</td>
</tr>
<tr>
<td>Mid East</td>
<td>Lebanon</td>
<td>PA</td>
<td>116,534</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Danville</td>
<td>IL</td>
<td>87,858</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Oshkosh</td>
<td>WI</td>
<td>147,006</td>
</tr>
<tr>
<td>South East</td>
<td>Hattiesburg</td>
<td>MS</td>
<td>111,705</td>
</tr>
<tr>
<td>Mid East</td>
<td>Johnstown</td>
<td>PA</td>
<td>162,015</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Eau Claire</td>
<td>WI</td>
<td>141,304</td>
</tr>
<tr>
<td>Far West</td>
<td>Salem</td>
<td>OR</td>
<td>300,132</td>
</tr>
<tr>
<td>Plains</td>
<td>Mankato North</td>
<td>MN</td>
<td>83,176</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Appleton</td>
<td>WI</td>
<td>181,405</td>
</tr>
<tr>
<td>Rocky Mountains</td>
<td>Lewiston</td>
<td>ID</td>
<td>54,264</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Bloomington</td>
<td>IN</td>
<td>163,251</td>
</tr>
<tr>
<td>South West</td>
<td>Prescott</td>
<td>AZ</td>
<td>122,699</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>La Crosse</td>
<td>WI</td>
<td>120,198</td>
</tr>
<tr>
<td>South East</td>
<td>Palm Coast</td>
<td>FL</td>
<td>35,051</td>
</tr>
<tr>
<td>South East</td>
<td>Alexandria</td>
<td>LA</td>
<td>144,814</td>
</tr>
<tr>
<td>Mid East</td>
<td>Atlantic City</td>
<td>NJ</td>
<td>233,339</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>MSA</th>
<th>State</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>South West</td>
<td>Farmington</td>
<td>NM</td>
<td>96,223</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Glens Falls</td>
<td>NY</td>
<td>122,078</td>
</tr>
<tr>
<td>South West</td>
<td>Austin Round</td>
<td>TX</td>
<td>934,622</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Canton Massillon</td>
<td>OH</td>
<td>400,863</td>
</tr>
<tr>
<td>Mid East</td>
<td>Savannah</td>
<td>GA</td>
<td>271,403</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Battle Creek</td>
<td>MI</td>
<td>137,524</td>
</tr>
<tr>
<td>Plains</td>
<td>Grand Forks</td>
<td>ND</td>
<td>104,666</td>
</tr>
<tr>
<td>Mid East</td>
<td>Cumberland</td>
<td>MD</td>
<td>102,253</td>
</tr>
<tr>
<td>Rocky Mountains</td>
<td>Grand Junction</td>
<td>CO</td>
<td>99,259</td>
</tr>
<tr>
<td>Mid East</td>
<td>Allentown</td>
<td>PA</td>
<td>706,676</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Memphis</td>
<td>TN</td>
<td>1,103,853</td>
</tr>
<tr>
<td>South West</td>
<td>Decatur</td>
<td>IL</td>
<td>118,115</td>
</tr>
<tr>
<td>Far West</td>
<td>Fresno</td>
<td>CA</td>
<td>721,014</td>
</tr>
<tr>
<td>Rocky Mountains</td>
<td>Amarillo</td>
<td>TX</td>
<td>203,198</td>
</tr>
<tr>
<td>South West</td>
<td>Killeen</td>
<td>TX</td>
<td>279,644</td>
</tr>
<tr>
<td>Mid East</td>
<td>Rochester</td>
<td>NY</td>
<td>1,030,518</td>
</tr>
<tr>
<td>South West</td>
<td>Longview</td>
<td>TX</td>
<td>184,389</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Steubenville</td>
<td>OH</td>
<td>140,549</td>
</tr>
<tr>
<td>Rocky Mountains</td>
<td>Salt Lake City</td>
<td>UT</td>
<td>836,649</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Akron</td>
<td>OH</td>
<td>672,505</td>
</tr>
</tbody>
</table>

5.6 Connections at a Distance

5.6.1 New York

We have noted above that while distance provides a good guide to spatial dependence there is also considerable evidence of systematic correlations that appear independent of distance. For example we report the significant correlations of the New York/Northern New Jersey/Long Island Metropolitan Statistical Area in Table 3. Although there are significant positive correlations with other MSAs close by in the Mid East region, there are also significant correlations with the Far West, positive in the case of California, and negative with Oregon and Washington State.

5.7 Networks in House Prices

In Figure 4 we display a network structure constructed from the correlation matrix for the whole of the USA but restricted to correlations greater than or equal to 0.45 (equivalent to a threshold significance level of 0.0025%). Because the figure is rendered as a network it does not of course correspond to the coordinates of a map as would, for example, be provided by a geographical information system (GIS). Nevertheless, it does help to reveal some interesting features. For example to the left of the structure there is a rim of MSAs that are effectively detached from the central hub. In the figure the size of the ellipsoid is proportional to the degree of connectiveness (measured by the number of connections). The ellipsoid labeled 1296 in red is San Francisco. The connected MSAs are listed in...
Table 3: Connections of New York/Northern New Jersey/Long Island Metropolitan Statistical Area

<table>
<thead>
<tr>
<th>Region</th>
<th>MSA</th>
<th>State</th>
<th>Correlation</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid East</td>
<td>Trenton Ewing</td>
<td>NJ</td>
<td>0.645</td>
<td>56</td>
</tr>
<tr>
<td>Mid East</td>
<td>Poughkeepsie Newburgh Middletown</td>
<td>NY</td>
<td>0.441</td>
<td>60</td>
</tr>
<tr>
<td>Mid East</td>
<td>Johnstown</td>
<td>PA</td>
<td>-0.444</td>
<td>257</td>
</tr>
<tr>
<td>Mid East</td>
<td>Reading</td>
<td>PA</td>
<td>-0.333</td>
<td>108</td>
</tr>
<tr>
<td>South East</td>
<td>Panama City Lynn Haven Panama City Beach</td>
<td>FL</td>
<td>0.341</td>
<td>981</td>
</tr>
<tr>
<td>South East</td>
<td>Lafayette</td>
<td>LA</td>
<td>-0.385</td>
<td>1,246</td>
</tr>
<tr>
<td>Plains</td>
<td>Mankato North Mankato</td>
<td>MN</td>
<td>0.333</td>
<td>1,043</td>
</tr>
<tr>
<td>Far West</td>
<td>Modesto</td>
<td>CA</td>
<td>0.357</td>
<td>2,500</td>
</tr>
<tr>
<td>Far West</td>
<td>Napa</td>
<td>CA</td>
<td>0.335</td>
<td>2,388</td>
</tr>
<tr>
<td>Far West</td>
<td>Salinas</td>
<td>CA</td>
<td>0.376</td>
<td>2,557</td>
</tr>
<tr>
<td>Far West</td>
<td>Vallejo Fairfield</td>
<td>CA</td>
<td>0.384</td>
<td>2,549</td>
</tr>
<tr>
<td>Far West</td>
<td>Corvallis</td>
<td>OR</td>
<td>-0.328</td>
<td>2,488</td>
</tr>
<tr>
<td>Far West</td>
<td>Eugene Springfield</td>
<td>OR</td>
<td>-0.366</td>
<td>2,490</td>
</tr>
<tr>
<td>Far West</td>
<td>Portland Vancouver Hillsboro</td>
<td>OR</td>
<td>-0.368</td>
<td>2,458</td>
</tr>
<tr>
<td>Far West</td>
<td>Salem</td>
<td>OR</td>
<td>-0.376</td>
<td>2,468</td>
</tr>
<tr>
<td>Far West</td>
<td>Bellingham</td>
<td>WA</td>
<td>-0.354</td>
<td>2,408</td>
</tr>
<tr>
<td>Far West</td>
<td>Yakima</td>
<td>WA</td>
<td>-0.321</td>
<td>2,332</td>
</tr>
</tbody>
</table>

Table ?? The network includes the Los Angeles MSA and interestingly - through Santa Rosa Petaluma - there is a connection to Denver Aurora Broomfield in Colorado which is some 1300 miles away.

Figure 4: Network Structure for House Prices: 0.0025% significance level.

On the right of the figure there are also some instances of star networks. For example, Niles Benton Harbor (identifier 1237) in Michigan appears to be largely at the centre of a network. Chicago is one of the spokes, which itself has just one further connection to far off (870 miles) Rocky Mount in North Carolina.
Table 4: San Francisco Network

<table>
<thead>
<tr>
<th>MSA</th>
<th>State</th>
<th>ID</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denver Aurora Broomfield</td>
<td>CO</td>
<td>1091</td>
<td>1,797,681</td>
</tr>
<tr>
<td>Los Angeles Long Beach Santa Ana</td>
<td>CA</td>
<td>1199</td>
<td>11,623,575</td>
</tr>
<tr>
<td>Napa</td>
<td>CA</td>
<td>1231</td>
<td>115,321</td>
</tr>
<tr>
<td>Oxnard Thousand Oaks Ventura</td>
<td>CA</td>
<td>1250</td>
<td>688,461</td>
</tr>
<tr>
<td>Riverside San Bernardino Ontario</td>
<td>CA</td>
<td>1280</td>
<td>2,861,221</td>
</tr>
<tr>
<td>San Francisco Oakland Fremont</td>
<td>CA</td>
<td>1296</td>
<td>3,832,170</td>
</tr>
<tr>
<td>San Jose Sunnyvale Santa Clara</td>
<td>CA</td>
<td>1297</td>
<td>1,580,016</td>
</tr>
<tr>
<td>Santa Barbara Santa Maria Goleta</td>
<td>CA</td>
<td>1300</td>
<td>378,889</td>
</tr>
<tr>
<td>Santa Cruz Watsonville</td>
<td>CA</td>
<td>1301</td>
<td>234,318</td>
</tr>
<tr>
<td>Santa Rosa Petaluma</td>
<td>CA</td>
<td>1303</td>
<td>408,982</td>
</tr>
</tbody>
</table>

If we move to the hub of the network there is a complex sheen of connections. In Table 5.7 we list those MSAs that are coloured red in the Figure. It is clear that most of the concentrations are in the Great Lakes and South East regions. In Figure 5 we zoom in on the central hub.

Figure 5: Network Structure of the Core.

Finally in Figure 6 we plot the adjacency matrix for a threshold significance level of 0.0025%. Again the correlations that are loosely connected to geographical proximity are extensive. For example there is a marked ribbon of significance along the row just above 300 on the left axis. This is the Metropolitan Statistical Area of Victoria in Texas which has one of the largest number of connections by number and is the most strongly connected measured by the column norm. Why this should be so remains the subject of further examination.

Figure 6: Adjacency Matrix: 0.0025% significance level.
Table 5: Largest number of connections. Plotted in red in Figure 5

<table>
<thead>
<tr>
<th>Region</th>
<th>MSA</th>
<th>State</th>
<th>MSA ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>New England</td>
<td>Norwich New London</td>
<td>CT</td>
<td>1239</td>
</tr>
<tr>
<td>Mid East</td>
<td>Salisbury</td>
<td>MD</td>
<td>1291</td>
</tr>
<tr>
<td>Mid East</td>
<td>Atlantic City Hammonton</td>
<td>NJ</td>
<td>1019</td>
</tr>
<tr>
<td>Mid East</td>
<td>Trenton Ewing</td>
<td>NJ</td>
<td>1335</td>
</tr>
<tr>
<td>Mid East</td>
<td>Elmira</td>
<td>NY</td>
<td>1104</td>
</tr>
<tr>
<td>Mid East</td>
<td>Johnstown</td>
<td>PA</td>
<td>1164</td>
</tr>
<tr>
<td>Mid East</td>
<td>Lebanon</td>
<td>PA</td>
<td>1189</td>
</tr>
<tr>
<td>South East</td>
<td>Fort Smith</td>
<td>AR</td>
<td>1119</td>
</tr>
<tr>
<td>South East</td>
<td>Crestview Fort Walton Beach</td>
<td>FL</td>
<td>1080</td>
</tr>
<tr>
<td>South East</td>
<td>Palm Coast</td>
<td>FL</td>
<td>1252</td>
</tr>
<tr>
<td>South East</td>
<td>Port St. Lucie</td>
<td>FL</td>
<td>1264</td>
</tr>
<tr>
<td>South East</td>
<td>Owensboro</td>
<td>KY</td>
<td>1249</td>
</tr>
<tr>
<td>South East</td>
<td>Alexandria</td>
<td>LA</td>
<td>1005</td>
</tr>
<tr>
<td>South East</td>
<td>Gulfport Biloxi</td>
<td>MS</td>
<td>1136</td>
</tr>
<tr>
<td>South East</td>
<td>Hattiesburg</td>
<td>MS</td>
<td>1142</td>
</tr>
<tr>
<td>South East</td>
<td>Jackson</td>
<td>MS</td>
<td>1157</td>
</tr>
<tr>
<td>South East</td>
<td>Jacksonville</td>
<td>NC</td>
<td>1160</td>
</tr>
<tr>
<td>South East</td>
<td>Greenville Mauldin Easley</td>
<td>SC</td>
<td>1135</td>
</tr>
<tr>
<td>South East</td>
<td>Myrtle Beach North Myrtle Beach Conway</td>
<td>SC</td>
<td>1230</td>
</tr>
<tr>
<td>South East</td>
<td>Johnson City</td>
<td>TN</td>
<td>1163</td>
</tr>
<tr>
<td>South East</td>
<td>Huntington Ashland</td>
<td>WV</td>
<td>1150</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Danville</td>
<td>IL</td>
<td>1084</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Kankakee Bradley</td>
<td>IL</td>
<td>1168</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Bloomington</td>
<td>IN</td>
<td>1038</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Kokomo</td>
<td>IN</td>
<td>1175</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Michigan City La Porte</td>
<td>IN</td>
<td>1215</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Holland Grand Haven</td>
<td>MI</td>
<td>1145</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Niles Benton Harbor</td>
<td>MI</td>
<td>1237</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Appleton</td>
<td>WI</td>
<td>1015</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Eau Claire</td>
<td>WI</td>
<td>1099</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Fond du Lac</td>
<td>WI</td>
<td>1117</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Oshkosh Neenah</td>
<td>WI</td>
<td>1248</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>Sheboygan</td>
<td>WI</td>
<td>1308</td>
</tr>
<tr>
<td>Plains</td>
<td>Manhattan</td>
<td>KS</td>
<td>1207</td>
</tr>
<tr>
<td>Plains</td>
<td>Mankato North Mankato</td>
<td>MN</td>
<td>1208</td>
</tr>
<tr>
<td>Plains</td>
<td>Columbia</td>
<td>MO</td>
<td>1073</td>
</tr>
<tr>
<td>Plains</td>
<td>Bismarck</td>
<td>ND</td>
<td>1036</td>
</tr>
<tr>
<td>South West</td>
<td>Prescott</td>
<td>AZ</td>
<td>1268</td>
</tr>
<tr>
<td>South West</td>
<td>Victoria</td>
<td>TX</td>
<td>1344</td>
</tr>
<tr>
<td>Rocky Mountains</td>
<td>Lewiston</td>
<td>ID</td>
<td>1190</td>
</tr>
<tr>
<td>Far West</td>
<td>San Francisco Oakland Fremont</td>
<td>CA</td>
<td>1296</td>
</tr>
<tr>
<td>Far West</td>
<td>Salem</td>
<td>OR</td>
<td>1289</td>
</tr>
</tbody>
</table>
6 Impulse Response Analysis of a Spatio-temporal Model

In this section we turn to an impulse response analysis of the complete spatio-temporal model for house prices. To simplify the exposition, and without loss of generality, we set \( p = q^+ = q^- = 1 \) and write the full model of house price changes as (also setting \( a_\xi = 0 \) again without loss of generality)

\[
\begin{align*}
(I_N - BR_N - \Gamma P_N) \pi_t &= a + \xi_t, \text{ for } t = 1, 2, \ldots, T \\
(I_N - \Psi_0^+ W^+ - \Psi_0^- W^-) \xi_t &= (\Lambda + \Psi_1^+ W^+ + \Psi_1^- W^-) \xi_{t-1} + \zeta_t,
\end{align*}
\]

or in terms of \( \pi_t \)

\[
A_0 \pi_t = A_1 \pi_{t-1} + c
\]

where

\[
A_0 = (I_N - \Psi_0^+ W^+ - \Psi_0^- W^-) (I_N - BR_N - \Gamma P_N)
\]

\[
A_1 = (\Lambda + \Psi_1^+ W^+ + \Psi_1^- W^-) (I_N - BR_N - \Gamma P_N)
\]

and

\[
c = (I_N - \Lambda) a - (\Psi_0^+ + \Psi_0^-) W^+ a - (\Psi_1^+ + \Psi_1^-) W^- a.
\]

The above system of equations now yields the following \( VAR(1) \) in the \( N \times 1 \) vector of price changes

\[
\pi_t = d + \Phi \pi_{t-1} + u_t \quad \text{for } t = 1, 2, \ldots, T,
\]

where

\[
d = A_0^{-1} c, \quad \Phi = A_0^{-1} A_1, \quad u_t = A_0^{-1} \zeta_t
\]

Suppose now that we are interested in the effects of a unit change (one standard error) to the composite shock, \( \eta_{tw} = \omega' \zeta_t \) where \( \omega \) is an \( N \times 1 \) vector given weights (in our application they could be relative population weights), which add up to unity, namely \( \omega' \omega = 1 \). The unit shock is defined by \( \sigma_\omega = \sqrt{\omega' Var(\zeta_t) \omega} = (\sum_{i=1}^{N} \omega_i^2 \sigma_i^2)^{1/2} \), where \( \sigma_i^2 = Var(\zeta_{it}) \). Different types of composite shocks can be defined, depending on whether the focus of the analysis is the effects of shocks originating from a given MSA, or from a given region or State. In the case where the interest is in the effects of shocks to a particular MSA, say the \( i^{th} \) MSA, then we set \( \omega \) to an \( N \times 1 \) vector with all elements equal to zero except for its \( i^{th} \) element which is set to unity. If we interested in the effects of a unit shock to a particular region, then we set all elements of \( \omega \) to zero, except for the elements associated with the MSA that belongs to the region under consideration, with the weights of these MSAs set to their relative population within the region.

Using familiar results in Koop, Pesaran and Potter (1996), it is now easily seen that the desired impulse response function is given by

\[
\psi_x(h, \sigma_\omega) = E(\pi_{t+h} | \eta_{tw} = \sigma_\omega, \Omega_{t-1}) - E(\pi_{t+h} | \Omega_{t-1}) = \frac{\Phi^h A_0^{-1} \Sigma \omega}{\sqrt{\sum_{i=1}^{N} \omega_i^2 \sigma_i^2}}, \quad h = 0, 1, 2, \ldots
\]

where \( \Omega_{t-1} = (\pi_{t-1}, \pi_{t-2}, \ldots) \), \( \Sigma \) is a diagonal matrix with \( \sigma_i^2 \) on its \( i^{th} \) diagonal element, and \( \psi_x(h, \sigma_\omega) \) is an \( N \times 1 \) vector that shows the effects of the composite shock, on all the
MSAs. Note that since $\Sigma_\zeta$ is a diagonal matrix then $(\Sigma_\zeta \omega)' = (\omega_1 \sigma_1^2, \omega_2 \sigma_2^2, ..., \omega_N \sigma_N^2)$. The above result further simplifies if we consider the effects of a shock to a particular MSA. In the case of a unit shock to the $i^{th}$ MSA we have

$$\psi_x (h, \sigma_i) = \sigma_i \Phi^h A_0^{-1} e_i, \ h = 0, 1, 2, ...,$$

where $e_i$ is an $N \times 1$ selection vector of zeros except for its $i^{th}$ element which is unity.

Note that the rate of decay of the impulse responses is controlled by the largest eigenvalue (in absolute terms) of $\Phi$ which depends in a complicated non-linear manner on the parameters of the factor and spatiotemporal model and the spatial weight matrices $W^+$ and $W^-$. 

7 Conclusions

An understanding of the spatial dimension of economic and social activity requires methods that can separate out the relationship between spatial units that is due to the effect of common factors from that which is purely spatial even in an abstract sense. We are able to distinguish between cross-sectional strong dependence and weak or spatial dependence. Strong dependence in turn suggests that there are common factors. We have proposed the use of cross unit averages to extract both national and regional factors and contrast this to a principal components approach widely used in the literature. We then use thresholding to determine significant bilateral correlations (signifying neighbours) between spatial units and compare this to an approach that just uses distance to determine units that are neighbours. In a very data rich environment with observations on many spatial units over long periods of time a way of filtering the data to uncover spatial connections is crucial. We have applied these methods to real house price changes at the level of the Metropolitan Statistical Area. Although there is considerable overlap between neighbours determined by distance and those by thresholding, there is also considerable correlation between MSAs across the United States that suggests other forces at work.

8 Appendix

8.1 Appendix I: Estimation of $\alpha$

Assume (9) for $m = 1$ (without loss of generality) and $1/2 < \alpha \leq 1$. Then, we have

$$Var(\bar{x}_t) = \sigma_f^2 \mu_\nu^2 [N^{2\alpha-2}] + N^{-1} c_N + O(N^{\alpha-2}),$$

(22)

where $\sigma_f^2$ is the variance of the factor process, $\mu_\nu$ is the mean of the factor loadings, and $c_N$ is a bias term. Setting $\sigma_f^2 = 1$ as normalising condition, the population value of $\alpha$ is then given by

$$\alpha \approx 1 + \frac{1}{2} \ln(\sigma_f^2) - \frac{1}{2} \ln(\mu_\nu^2) - \frac{c_N}{2 [N \ln(N)] \sigma_f^2}.$$ 

(23)

Accordingly, Bailey et al (2012) propose the following bias-corrected estimator

$$\hat{\alpha} = \hat{\alpha} (\hat{\mu}_\nu^2) = 1 + \frac{1}{2} \ln(\hat{\sigma}_f^2) - \frac{\ln(\hat{\mu}_\nu^2)}{2 [N \ln(N)] \sigma_f^2},$$

(24)
where firstly
\[ \hat{\sigma}_2^2 = \frac{1}{T} \sum_{t=1}^{T} (\bar{x}_t - \bar{x})^2, \]  
(25)

with \( \bar{x} = T^{-1} \sum_{t=1}^{T} x_t \).

For the estimation of \( \hat{\mu}_v \) the following algorithm is considered:

**Proposed Algorithm**

1. Regress \( x_{it} \) on a constant and \( \bar{x}_t \) and denote the estimated coefficient of \( \bar{x}_t \) by \( \hat{\delta}_i \), for \( i = 1, 2, ..., N \).

2. Compute the \( t \)-ratio associated with the \( i \)-th coefficient, \( \hat{\delta}_i \), \( i = 1, 2, ..., N \), as
   \[ z_{i}^{\hat{\delta}} = \frac{\hat{\delta}_i}{\text{s.e.} \left( \hat{\delta}_i \right)}. \]

3. Construct
   \[ \bar{x}_t(c_{p,N}) = \frac{\sum_{i=1}^{N} x_{it} I \left( \left| z_{i}^{\hat{\delta}} \right| \geq c_{p,N} \right)}{\sum_{i=1}^{N} I \left( \left| z_{i}^{\hat{\delta}} \right| \geq c_{p,N} \right)}, \]
   where \( c_{p,N} = \Phi^{-1} \left( 1 - \frac{p}{N} \right) \) and \( \Phi^{-1} (.) \) is the inverse of the cumulative distribution function of the standard normal variate, and \( p \) is the overall size of the test, which we set to 10%.

4. Estimate \( \mu_v \) by
   \[ \hat{\mu}_v = \hat{\mu}_v (c_{p,N}) = \sqrt{\frac{1}{T} \sum_{t=1}^{T} [\bar{x}_t(c_{p,N}) - \bar{x}(c_{p,N})]^2}, \]
   where \( \bar{x}(c_{p,N}) = T^{-1} \sum_{t=1}^{T} \bar{x}_t(c_{p,N}) \).

Finally,
\[ \hat{\sigma}_N^2 = \hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} \omega_i^2, \quad \hat{\omega}_i^2 = \frac{1}{T} \sum_{t=1}^{T} e_{it}^2, \]
where \( e_{it} = x_{it} - \hat{\delta}_i \bar{x}_t \), and \( \hat{\delta}_i \) denotes the OLS estimator of the regression coefficient of \( x_{it} \) on \( \bar{x}_t \). When \( \Sigma_{\epsilon} \) is not a diagonal matrix, an alternative estimator of \( \hat{\sigma}_N \) is derived.

### 8.2 Appendix II: Data sources

Monthly data for house prices from January 1975 to December 2010 are taken from Freddie Mac. These data are available at:
- http://www.freddiemac.com/finance/cmhpi
  The quarterly figures are arithmetic averages of monthly figures.
- Annual CPI data at State level are obtained from the Bureau of Labor Statistics:
  http://www.bls.gov/cpi/
  The quarterly figures are interpolated using the interpolation technique described in the appendix of GVAR toolbox 1.1 user guide.
The annual population data at MSA level are obtained from the Regional Economic Information System, Bureau of Economic Analysis, U.S. Department of Commerce:
http://www.bea.gov/regional/docs/footnotes.cfm?tablename=CA1-3
Again the quarterly figures are interpolated using the interpolation technique described in the appendix of GVAR toolbox 1.1 user guide.

8.3 Appendix III: Geographical classification of the United States

The table below provides the broader geographical breakdown used in the analysis of US house prices. We identify 8 regions which contain an average of 6.1 States, each of which contains an average of 45.38 Metropolitan Statistical Areas. The classifications are shown in Table 6 together with the number of MSAs included in each State. The details of the exact MSAs used are provided upon request.

Table 6: Geographical Classification of Regions, States and MSAs in the United States

<table>
<thead>
<tr>
<th>Regions</th>
<th>States</th>
<th>No of MSAs</th>
<th>Regions</th>
<th>States</th>
<th>No of MSAs</th>
<th>Regions</th>
<th>States</th>
<th>No of MSAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>New England</td>
<td>Connecticut (CT)  4</td>
<td></td>
<td>Great Lakes</td>
<td>Illinois (IL)  9</td>
<td></td>
<td>South West</td>
<td>Arizona (AZ)  6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maine (ME)  3</td>
<td></td>
<td></td>
<td>Indiana (IN)  13</td>
<td></td>
<td></td>
<td>New Mexico (NM)  4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Massachusetts (MA)  5</td>
<td></td>
<td></td>
<td>Michigan (MI)  14</td>
<td></td>
<td></td>
<td>Oklahoma (OK)  3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New Hampshire (NH)  1</td>
<td></td>
<td></td>
<td>Ohio (OH)  13</td>
<td></td>
<td></td>
<td>Texas (TX)  25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rhode Island (RI)  1</td>
<td></td>
<td></td>
<td>Wisconsin (WI)  12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vermont (VT)  1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mid-East</td>
<td>Delaware (DE)  1</td>
<td></td>
<td>Plains</td>
<td>Iowa (IA)  8</td>
<td></td>
<td>Rocky Mountains</td>
<td>Colorado (CO)  7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>District of Columbia (DC)  1</td>
<td></td>
<td></td>
<td>Kansas (KS)  4</td>
<td></td>
<td></td>
<td>Idaho (ID)  5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maryland (MD)  4</td>
<td></td>
<td></td>
<td>Minnesota (MN)  5</td>
<td></td>
<td></td>
<td>Montana (MT)  3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New Jersey (NJ)  4</td>
<td></td>
<td></td>
<td>Missouri (MO)  8</td>
<td></td>
<td></td>
<td>Utah (UT)  5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New York (NY)  12</td>
<td></td>
<td></td>
<td>Nebraska (NE)  2</td>
<td></td>
<td></td>
<td>Wyoming (WY)  2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pennsylvania (PA)  14</td>
<td></td>
<td></td>
<td>North Dakota (ND)  3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>South Dakota (SD)  2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South East</td>
<td>Alabama (AL)  11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Far West</td>
<td>California (CA)  26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arkansas (AR)  6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Nevada (NV)  3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Florida (FL)  120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Oregon (OR)  6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Georgia (GA)  14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Washington (WA)  10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kentucky (KY)  15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Louisiana (LA)  8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mississippi (MS)  4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>North Carolina (NC)  14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>South Carolina (SC)  8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tennessee (TN)  10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virginia (VA)  9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>West Virginia (WV)  5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8.4 Appendix IV: Calculation of distance

The original data used were Latitude-Longitude of zip codes, cross referenced with each of the 366 Metropolitan Statistical Areas (MSAs). Any missing Latitude-Longitude coordinates were coded manually from Google searches. The geodesic distance between a pair of latitude/longitude coordinates was then calculated using the Haversine formula:

\[
a = \sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat_1)\cos(lat_2)\sin^2\left(\frac{\Delta long}{2}\right),
\]

\[
c = 2a\tan\left(\frac{\sqrt{a}, \sqrt{1-a}}{2}\right),
\]

\[
d = Rc,
\]

where \(R\) is the radius of the earth in miles and \(d\) is the distance. \(\Delta lat = lat_2 - lat_1\), and \(\Delta long = long_2 - long_1\).

8.5 Appendix V: Additional contingency tables

The contingency tables below refer to comparisons of \(\hat{W}_u\) constructed using two or more principal components in the defactoring of the house price equations, and \(W_M\) constructed using distance as metric.

Table 7: Contingency tables: Data-driven versus distance-based spatial weights matrices

<table>
<thead>
<tr>
<th></th>
<th>(W_{100m})</th>
<th></th>
<th>(W_{200m})</th>
<th></th>
<th>(W_{300m})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>(\sum_{rows})</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(\hat{W}_u)</td>
<td>1020</td>
<td>64683</td>
<td>65703</td>
<td>1007</td>
<td>3527</td>
</tr>
<tr>
<td>0</td>
<td>285</td>
<td>3949</td>
<td>4234</td>
<td>707</td>
<td>3527</td>
</tr>
<tr>
<td>(\sum_{rows})</td>
<td>1020</td>
<td>64683</td>
<td>65703</td>
<td>1007</td>
<td>3527</td>
</tr>
</tbody>
</table>

Matrix \(\hat{W}\) results from defactoring house prices of each MSA by using the two principal components corresponding to the two largest eigenvalues of the house prices data set.
The χ² statistics stand at 794.17, 1134.48, and 942.69 for $\hat{W}_u$ versus $W_{100m}$, $W_{200m}$, and $W_{300m}$ when two principal components are used in the defactoring of the house price equations, 611.68, 907.62, and 762.36 when three principal components are used, and 480.18, 618.63, and 489.15 when four principal components are used.

We also implement the Pesaran and Timmermann (2009) approach and run regression (20) and we obtain the following t-statistics for $\hat{W}_u$ versus $W_{100m}$, $W_{200m}$, and $W_{300m}$: (i) for the 2 principal components case these stand at: 29.31, 35.13, and 31.97, (ii) for the 3 principal components case they stand at: 25.49, 31.12, and 28.49, and (iii) for the 4 principal components case they stand at: 22.50, 25.56, 22.71 respectively. Comparison of $\hat{W}_u$ versus $W_{200m}$ outperforms in all principal components set ups, as in the case of cross-sectional averages.

**References**


