Frequentist Inference in Spatial Discrete Choice Models with Endogenous Congestion Effects and Club-Correlated Random Effects

Arnab Bhattacharjee

Spatial Economics & Econometrics Centre (SEEC)
Heriot-Watt University, Scotland, UK
a.bhattacharjee@hw.ac.uk

Rob L. Hicks *College of William and Mary, USA*

Kurt E. Schnier *University of California Merced, USA*

21st London Stata Users Group Meeting, Sept. 2015

SEEC

Spatial Economics
& Econometrics Centre

http://seec.hw.ac.uk/

Spatial structure

- (Almost) all economic activity is spatial
  - Inherently spatial phenomena – Migration, Transport, Trade, etc.
  - Urban economics – Spatial structures of cities
  - Housing economics – Neighbourhood effects, Deprivation, Demand/Price spillovers, Location based policy
  - Industrial economics, Economic growth, Network economics, Regional economics/ Input-Output models, Health economics, etc.
  - Spatial general equilibrium
    – location based policies and shocks, and their spatial transmission
- Spatial structure of economic activity highlight several features
  - High spatial variation – more pronounced at smaller scale
  - Patterns of agglomeration and (importantly) dissociation
  - Spatial spillovers, leading to endogenous evolution of space
    – How is space transformed? – Produced? (*a la* Henri Lefebvre)

Agglomeration (and dissociation) economies

Figure: Innovation activity across Europe (Arbia *et al.*, 2010)

Higher spatial variation at lower spatial scales

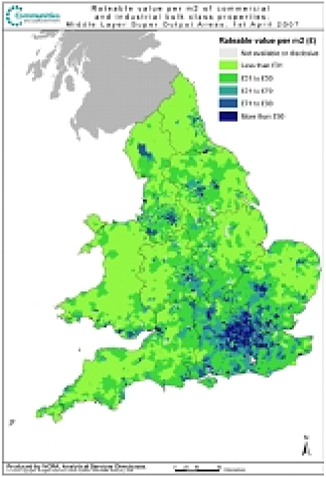


Figure: Land prices in England/Wales (Overman and Rice, 2008)

Endogenous evolution of spatial structure

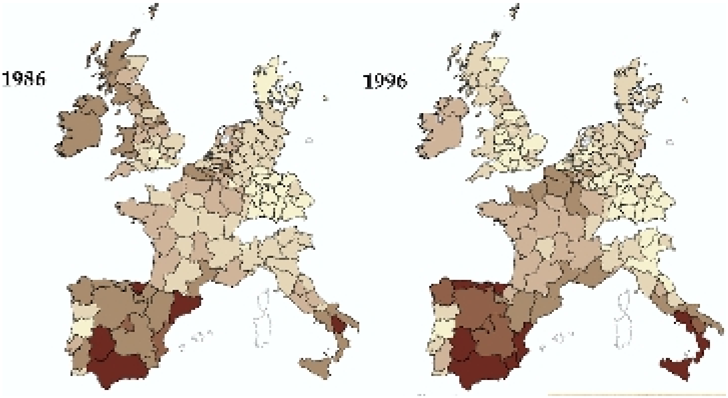


Figure: Unemployment clusters across Europe (Overman and Puga, 2000)

Spatial structure at time 0 $\longrightarrow$ Agglomeration/ Dissociation/ Specialisation $\longrightarrow$ Migration + Space based planning/ shocks $\longrightarrow$ Further agglomeration/ dissociation/ specialisation $\longrightarrow$ Transformed spatial structure at time 1

Urban economics, Geography and New Economic Geography

- Urban economics and geography
  – Alonso-Muth-Mills and onwards
    - Equilibrium location of households
    - Monocentric model of the city – extended to multiple centres (**hotspots**)
    - Tobler's First law of geography – **Distance decay** (Schaeffer)
    - **Checkerboard patterns** – spatial competition and segregation (Schelling)
    - **Spatial heterogeneity** – unique spaces (Hartshorne)

Urban economics, Geography and New Economic Geography

- New Economic Geography
  – Fujita-Krugman-Venables
    - Firms and workers, not regions, as agents
    - Firm-consumer interactions (dd linkages) → Divergent spatial outcomes
    - Transaction costs → Firm relocation → Core-Periphery → **Spatial heterogeneity**
    - I-O relations between firms → Firms linked to suppliers & consumers → **Endogenous evolution** of specialised regional economies
    - Spatial outcomes – Path dependent and contingent
      → **Spatial dependence** (endogeneity)
      → Potentially suboptimal (efficiency, welfare)

Regional economics

- Regions as spatial units
  - Spatial fixity
    - ★ Each region is unique
    - ★ Subject to different economic processes
      → Divergent spatial outcomes across regions
      → **Spatial heterogeneity**
  - Spillovers, spatial externalities and clustering
    → **Spatial interactions** (**endogenous spatial dependence**)
  - Endogenous specialised regional economies
    → **Endogenously produced spatial structure**

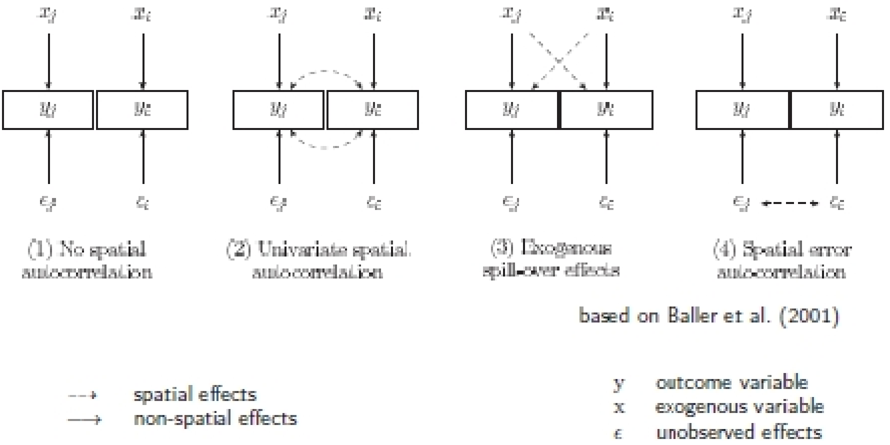
Philosophy: Abstract and Endogenous Space

- The Production of Space (Henri Lefebvre, 1974 [trans. 1991])
  - '*Euclidean space is defined by its "isotopy" (or homogeneity), a property which guarantees its social and political utility. The reduction to this homogenous Euclidean space, first of nature's space, then of all social space, has conferred a redoubtable power upon it. All the more so since that initial reduction leads easily to another – namely, the reduction of three-dimensional realities to two dimensions (for example, a "plan," a blank sheet of paper, something drawn on paper, a map, or any kind of graphic representation or projection).*' (p.285)
  - '*(S)pace is a (social) product . . . the space thus produced also serves as a tool of thought and of action . . . in addition to being a means of production it is also a means of control, and hence of domination, of power.*' (p.26)
  - '*Change life! Change Society! These ideas lose completely their meaning without producing an appropriate space.*' (p.59)

Traditional Spatial Econometrics

- Spatial econometrics (Anselin, Bera, Kelejian, Lee, LeSage, Prucha, etc.)
  - Spatial structure (**spatial heterogeneity**)
  - Spatial interaction (**spatial dependence**)
  - Also **spatial scale**, but largely an empirical (not econometric) issue
  - Spatial heterogeneity
    - *Geographically weighted regressions* (Fotheringham, etc.)
    - With panel data, intercept dummies (location fixed effects) and slope dummies (heterogenous location specific slope effects)
  - Spatial interaction
    - *Spatial regression models* (next)
    - Spatial weights (interaction) matrix, W
    - Known W, symmetric, zero diagonal
    - Based on economic/ geographic distances/ contiguity, fixed a priori
  - New literature (later)
    - Inferences on an unknown weights matrix
      – Substantial uncertainty about true distances/ measurement
    - Endogenous spatial weights
      – Endogenously evolving spatial structure

Spatial interaction (Anselin 1988, etc.)



(1) Linear regression model; (2) Spatial lag model; (3) Spatial Durbin model; (4) Spatial error model.

Spatial regression models (Anselin, 1988, 1999; Cressie, 1991; LeSage and Pace, 2009)

- Consider spillover in prices between two housing markets (**E**dinburgh and **G**lasgow)
  - Prices in Edinburgh (y_E) affected by its own attributes x_E
  - But also there is spatial (spillover) effect from the housing market in Glasgow, through a spatial weight w_{GE}
  - Likewise for Glasgow
  - Different spatial regression models account for the spatial effect in different ways
- **Spatial Lag Model**, or **Spatial Autoregressive Model** (SAR) (Whittle, 1954)
  - Endogenous spatial interaction, Wy is called the spatial lag of y
  - Prices in Edinburgh (y_E) directly affect those in Glasgow (y_G), and vice versa

$$\left.\begin{array}{l} y_E = \rho w_{GE} y_G + x_E \beta + \varepsilon_E \\ y_G = \rho w_{EG} y_E + x_G \beta + \varepsilon_G \end{array}\right\} W = \begin{bmatrix} 0 & w_{GE} \\ w_{EG} & 0 \end{bmatrix}$$

$$\Rightarrow \quad y = \rho W y + X\beta + \varepsilon.$$

Spatial regression models (contd.)

- But what if the interaction effect arises not through prices ...
  - But through the unobservable errors (ε)?
  - **Spatial error model** (Anselin 1988)

$$y = X\beta + \varepsilon, \quad \varepsilon = \lambda W \varepsilon + \eta.$$

- **Combinations/ extensions** (Case 1991; Anselin 1999)
  - Comb. spatial lag and error: $y = \rho W y + X\beta + \varepsilon, \quad \varepsilon = \lambda M \varepsilon + \eta$.
  - Spatial Durbin model: $y = \rho W y + X\beta + MX\gamma + \varepsilon$.
- **Factor based spatial dependence** (Pesaran 2006, Bai 2009)
  - Strong spatial dependence (Pesaran 2006)
  - Heterogenous loadings (γ_i) to common factors (f_t)

$$y_{it} = \gamma_i' f_t + X_{it}\beta + \varepsilon_{it}.$$

  - Plus, spatial lag/ error effects – structural spatial (weak) dependence
- **Conditional autoregressive (CAR) models** (Besag, 1974)
  - Reduced form condl distbn: $y_i | y_{j, j \neq i} \sim N\left(\sum_{j \neq i} w_{ji} y_j + X_i \beta, \sigma_i^2\right)$.
  - Convenient for Bayesian modelling

Specification testing (1)

- Different models have very different interpretations
- For example, endogenous spatial (lag) effects or spatial error (spillover) effects?
- But specification testing is a hard problem (even for known W)
- Very difficult to distinguish between alternate models
- Why? Reduced forms of spatial models are very similar.
- Spatial lag (SAR) and spatial error (SEM) models

$$SAR \quad : \quad y = \rho W y + X\beta + \varepsilon \Rightarrow y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon;$$
$$SEM \quad : \quad y = X\beta + \varepsilon, \;\; \varepsilon = \lambda W \varepsilon + \eta \Rightarrow y = X\beta + (I - \lambda W)^{-1} \eta.$$

  - In practice ρW and λW are small, so that $(I - \rho W) \approx (I - \lambda W) \approx I$
  - Hence, enormously difficult to distinguish between models from data
  - Large literature, only briefly covered here
  - Anselin & Bera (1988); Bera & Yoon (1993); Anselin, Bera, Florax & Yoon (1996); Born & Breitung (2011), etc.

Specification testing (2)

- Spatial lag (SAR) and CAR model
  - no specification tests, but see Wall (2004)
  - very different implications, but SAR and CAR models are observationally very similar
- Factor and non-factor (SAR, SEM) models
  - Pesaran (2013) strong dependence test
  - But not quite sure what it tests
  - When can you say errors have only spatial weak (structural) dependence
- Even harder, if the spatial weights matrix W is unknown
  - Test for weak dependence (Bhattacharjee and Holly, 2013)
  - Based on an estimated W

Spatial weights matrix

- Consider the pure-SAR model $y = \rho W y + \varepsilon$
  - Treatment for mixed regressive SAR model (including $X\beta$) similar
  - Also SEM model and SEM with moving average errors ($\varepsilon = \eta + \lambda W \eta$)
- $W = [w_{ij}]$ is a $n \times n$ constant matrix with zero diagonal
- The element w_{ij} measures the spatial distance between units i and j.
  - can be geographic distance, or contiguity, or economic distances (trade, migration, etc.)
- The spatial weights matrix is often row-normalized
  - The initial matrix W^* is often symmetric by construction
  - The row-normalized matrix W is asymmetric in general: $\sum_{j=1}^{n} w_{ij} = 1$.
- Consider lattice of 5 units with first order (rook) contiguity spatial wts

$$\left.\begin{matrix} 1 & 2 & \\ 3 & 4 & 5 \end{matrix}\right\} W^* = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \Rightarrow W = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 1/2 & 0 \end{bmatrix}.$$

Spatial and time series models

- Define $S_n(\rho) = I - \rho W_n$ with the idea that W changes with sample size n.
- At true value of **spatial autoregressive parameter** ρ_0, $S_n = S_n(\rho_0)$ assumed invertible for identification of reduced form:

$$y_n = S_n^{-1} X_n \beta_0 + S_n^{-1} \varepsilon_n.$$

- Spatial models include stationary time series model as special cases
- For example, AR(1) can be written as a special SAR(1):

$$\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \rho \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

- This shows why Wy is called the **spatial lag**
- Also, for stationarity, we must have $|\rho| < 1$ as well as $\|W\| < 1$.

Broad view on estimation

- Spatial econometrics/ statistics provides estimates (known W)
- SAR/ SEM models:
  - GMM (Kelejian & Prucha, 1988, 1989; etc)
    - Good small sample properties, but inefficiency/ bias?
    - Reasonable asymptotic results
    - Additional regressor X is required
  - maximum likelihood (ML)/ quasi-ML (Anselin 1988; Lee 2004)
    - Poor small sample properties
    - Difficult to get asymptotic results
    - High computation intensity
- CAR/ SAR models: Bayesian estimation (LeSage and Pace, 2009)
- Factor based models
  - Statistical factor analysis (Bai & Ng, 2005; Forni, Hallin, Lippi & Reichlin, 2005; Bai, 2009)
  - Common correlated effects (Pesaran, 2006; Pesaran & Tosetti, 2011)
- Weights typically measured by economic or geographic distances
- Key conceptual issue is the asymptotic setting

Time series versus Spatial econometrics

- What really are the challenges in estimation?
- Autoregressive Distributed Lag model – ADL$(\infty, 0)$
  - Geometric (Koyck) distributed lags

$$Y_t = \alpha + \rho Y_{t-1} + \rho\theta Y_{t-2} + \rho\theta^2 Y_{t-3} + \ldots + \varepsilon_t$$

  - Koyck transformation: $Y_t = \alpha(1-\theta) + (\rho + \theta) Y_{t-1} + (\varepsilon_t - \theta\varepsilon_{t-1})$
  - Main issue: 2 regression coefficients, 3 parameters
  - Solutions: ML or Two stage
    - ★ First stage: Estimate θ consistently, for example as the ratio of coefficients in AR model
    - ★ Second stage: Substitute in Koyck transformed model – similar to feasible GLS
- Are spatial data similar to time series?
  - Spatial units are irregularly spaced
  - No sense of a clear direction of progression
  - Geographical space or attribute space?
  - Most importantly, asymptotic setting – how are units added as $n \to \infty$ (following slide)

Time series versus Spatial econometrics (contd.)

- Spatial autoregressive model

$$Y_i = \alpha + \rho \sum_{j \neq i} w_{ji} Y_j + \varepsilon_i .$$

  - Assume W is known
    - ★ Standard assumption – decreasing function of geog/econ distances
    - ★ Like the θ in ADL, quantifies how interactions decrease with lags (distance)
  - Then, similar to the ADL model
    - ★ Standard ML, quasi-ML or GMM – estimate ρ
    - ★ Large literature in spatial econometrics (Anselin, 1988, 1999; LeSage and Pace, 2009; Elhorst, 2010, 2011; etc.)
    - ★ We will look at some of these methods next
    - ★ But before that, the asymptotic setting!

Asymptotics in spatial stochastic processes

- Quite different from time series
  - complicating factors
  - many open problems
  - no clear sigma field for accumulating evidence as $n \longrightarrow \infty$
- Intuition: Regularity conditions to limit spatial dependence (memory) and heterogeneity
  - Obtain uniform laws of large numbers (consistency) and central limit theorems (asymptotic normality) – easier said than done!
  - Results for dependent stationary sequences do not work – heteroscedasticity
  - Standard moment condns translate into constraints on W and ρ (or λ) (Anselin & Kelejian 1997; Pinkse & Slade 1998; Kelejian & Prucha 1999; etc.)
  - Condns on W are generally satisfied by contiguity based wts, but not necessarily econ distances
    - ⋆ certainly not negative spillovers, or core-periphery relationships

Asymptotics in spatial stochastic processes (2)

- *Increasing domain* or *Infill* asymptotic sampling structures?
  - Increasing domain: New units are added at the edges (boundary)
    - Similar to time series – but what is the boundary here – no clear direction of information flow?
  - Infill asymptotics: spatial domain is bounded, new observations added between existing ones – increasingly denser surface
    - Many "increasing domain" results do not follow with "infill"
    - But, farmer-district kind of models are promising (Robinson, 2010)
    - Increasing domain more popular
  - Promising new development – exponential random graphs
    - *Consistent under sampling* property (Shalizi & Rinaldo, 2013)
    - As sample size increases from n to $n+1$, all ${}^{n+1}C_n = n+1$ combinations have same joint probability law
    - Clustering of large networks (Vu, Hunter & Scheinberger, 2013)
- Need CLT and LLN for triangular arrays
  - A key point is that W_n itself changes as observations are added
  - Translates into bounded row and column norms for W_n and S_n^{-1}, and conditions on ρ (or λ) to ensure S_n is invertible

Software

- Statistics Toolbox for Matlab 1.1, 2.0
  - Spacestatpack for Fortran 90
  - Kelley Pace (of LeSage and Pace, 2009) and Ronald Barry
  - Fantastic sparse matrix functionality – very fast
- Geoda
  - Luc Anselin and colleagues at Arizona State University
  - Very powerful in ML and related computations
- "*A Primer for Spatial Econometrics: With Applications in R*"
  - Giuseppe Arbia, Chairman, *Spatial Econometrics Association*
- Stata
  - Comprehensive user programs by Maurizio Pisati and David M. Drukker
  - Markus Eberhardt/ Mark Schaffer developing more efficient programs
- Inferences on unknown and potentially endogenous W
  - Our methods – currently routines in Matlab
  - Plans to develop a toolbox in Stata
  - Spatial Economics & Econometrics Centre, Heriot-Watt University, Edinburgh, Scotland

Stata programs for spatial regression models (Drukker 2009)

- Syntax: $y = Y\pi + \lambda Wy + X\beta + u, \quad u = \rho Mu + \epsilon$. (endog regressor Y)
- Programs: *spreg* (ML, IV) and *spivreg* (GMM) – Drukker etc.
- Simulated data based on true US counties, 'inspired by' Powers and Wilson (2004) and Levitt (1997)
  - **dui**: dependent variable, alcohol related arrests per 100,000 daily vehicle miles
  - **police**: size of police force (potentially endogenous)
  - **nondui**: non-alcohol-related arrest rate
  - **vehicles**: no of vehicles registered per 1,000 residents
  - **dry**: counties that prohibit alcohol sale
  - **elect**: whether county govt faces an election
  - W, M: county level first order contiguity (row normalized)

```
. spivreg dui nondui vehicles dry (police = elect), id(id) ///
>         dlmat(ccounty) elmat(ccounty) nolog
```

Spatial autoregressive model   Number of obs = 3109
(GS2SLS estimates)

dui	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dui						
police	-1.457429	.0431113	-33.81	0.000	-1.541925	-1.372932
nondui	-.0003766	.0008295	-0.45	0.650	-.0020024	.0012492
vehicles	.0986084	.0017506	56.33	0.000	.0951772	.1020395
dry	.4538402	.0276385	16.42	0.000	.3996698	.5080106
_cons	9.578695	.3646643	26.27	0.000	8.863966	10.29342
lambda						
_cons	.731306	.0132648	55.13	0.000	.7053074	.7573046
rho						
_cons	.2836782	.0718875	3.95	0.000	.1427812	.4245752

Instrumented: police
Instruments: elect

Strong and weak spatial dependence (Pesaran 2006)

- Spatial dependence driven by latent factors
  - Different branch of the literature
  - *Latent factors* represent macroeconomic (global) shocks
  - Affect all spatial units, but to varying degrees (*heterogenous loadings*)
    - ★ Resulting in spatial autocorrelation
  - Structural implications very different from models based on W
    - ★ Spatial weights relate to underlying spatial lattice
    - ★ Create spatial effects in general equilibrium
    - ★ Even if there were no latent factors
    - ★ Structural vs factor-based spatial model (Bhattacharjee and Holly, 2011)
- Spatial panel data factor model:

$$y_{it} = \mu_i + \gamma_i' f_t + X_{it}\beta + \rho W y_{it} + (I - \lambda W)^{-1} \varepsilon_{it}.$$

  - Includes factors (f_t) with heterogenous loadings (γ_i)
  - Time factors or cross-section factors? – a matter of semantics
  - Pesaran (2006): Factor structure needs to be modeled explicitly
    - ★ Otherwise, *spatial granularity (stationarity) condition* may be violated

Strong and weak spatial dependence (2)

- **Spatial weak dependence** (Pesaran 2006)
  1. N and T both go to infinity – relative rates do not matter.
  2. **Spatial granularity**: *W has bounded row and column norms:* $\max\{\|\mathbf{W}\|_1, \|\mathbf{W}\|_\infty\} < 1$. *This is like a spatial stationarity condition.*
  3. *For each i, ε_{it} follows a linear stationary process with absolutely summable autocovariances*

$$\varepsilon_{it} = \sum_{s=0}^{\infty} a_{is}\epsilon_{is},$$

     where $\epsilon_{is} \sim IID(0,1)$ with finite fourth-order cumulants.

- One has to ensure spatial weak dependence
  - Closely related to conditions for spatial stationarity in Kelejian and Prucha (1998) and Lee (2004)
  - It is necessary to remove spatial strong dependence
  - In other words, the factor structure has to be modelled
  - After doing this, residuals can be tested for weak dependence
    - ⋆ Pesaran (2013) weak cross-section dependence (CD) test

Strong and weak spatial dependence (3)

- How should one model the factor structure? Two leading methods.
- Statistical factor analysis (Forni, Hallin, Lippi and Reichlin, 2004; Bai and Ng, 2005; Bai, 2009)
  - Dynamic factor model where spatial (cross sectional) dimension is 'not well-ordered'
  - Assumption of 'cross-sectional exchangeability' (Forni et al., 2004)
  - Labelling of time and cross-section dimension largely semantic
    - Unless temporal dynamics is modelled explicitly
    - Convenient to estimate factors along the longer dimension
    - If $T > N$, time factors, otherwise spatial factors
- Common correlated effects (Pesaran 2006; Pesaran and Tosetti, 2011)
  - The effect of factors are encompassed in cross-section averages.
  - Important:
    - Averages of both y and X have to be included
    - The averages capture the factor structure; hence, heterogenous slopes
    - This works only when N and T are both large
  - Very easy to apply

Strong and weak spatial dependence (4)

- Finally, one needs to ensure strong dependence is removed
  - Pesaran (2013) CD test for weak dependence in residuals
  - Bhattacharjee and Holly (2013) test if W is not known
- Once strong dependence is removed
  - Purely spatial structural dependence remains
  - Weak dependence can be modelled as SAR or SEM, or a combination
- Useful framework for modelling nonstationary spatio-temporal processes (Bhattacharjee, Higson and Holly, 2013)
  - Dynamic heterogenous panels (Pesaran and Smith, 1995)

$$y_{it} = \lambda_i y_{i,t-1} + \beta_i x_{it} + \alpha_i + \varepsilon_{it}, \qquad i = 1, \ldots, N; t = 1, \ldots, T.$$

  - Translates into a panel error correction model

$$\Delta y_{it} = \alpha_i + \gamma_i \Delta x_{it} + (1 - \lambda_i)\left(y_{i,t-1} - \theta_i x_{i,t-1}\right) + \varepsilon_{it}$$

Strong and weak spatial dependence (5)

- Add common correlated effects to account for potential strong spatial dependence

$$\Delta y_{it} = \alpha_i + \gamma_i \Delta x_{it} + (1 - \lambda_i)\left(y_{i,t-1} - \theta_i x_{i,t-1} - \delta_i \overline{y}_{t-1} - \mu_i \overline{x}_{t-1}\right) + \varepsilon_{it}$$

  - Potentially multiple long run equilibria
  - Test for strong spatial dependence in the errors

- If errors are not purely structural, add cross-section averages of short run terms as well

$$\begin{aligned} \Delta y_{it} &= \alpha_i + \gamma_i \Delta x_{it} + \tau_i \overline{\Delta y}_t + \psi_i \overline{\Delta x}_t \\ &\quad + (1 - \lambda_i)\left(y_{i,t-1} - \theta_i x_{i,t-1} - \delta_i \overline{y}_{t-1} - \mu_i \overline{x}_{t-1}\right) + \varepsilon_{it} \end{aligned}$$

- Finally, model structural spatial dependence using a SAR/ SEM model

Strong spatial dependence CD test (Pesaran 2013)

- Degree of strong spatial dependence measured by the "cross-section exponent" α (Bailey, Kapetanios and Pesaran, 2012)
  - $0 < \alpha < 1$; 1 being the highest degree of cross section dependence, or strong spatial dependence
  - $\alpha = \max_j \alpha_j ; \alpha_j = \ln\left(M_j\right) / \ln\left(N\right)$, where M_j is the number of spatial units that unit j has strong (significant) spatial correlation with
  - α is the exponent of N that gives the maximum number of x_{it} units that are pairwise spatially correlated
- Pesaran CD test (2013): To test for weak or spatial dependence, denote the sample pair-wise correlations of (i, j) units by

$$\widehat{\rho}_{ij} = \widehat{\rho}_{ji} = \frac{\sum_t \left(x_{it} - \overline{x}_i\right)\left(x_{jt} - \overline{x}_j\right)}{\sqrt{\left[\sum_t \left(x_{it} - \overline{x}_i\right)^2\right]\left[\sum_t \left(x_{jt} - \overline{x}_j\right)^2\right]}}.$$

Strong spatial dependence CD test (2)

- Then, the CD statistic is then defined by

$$CD = \left[TN(N-1)/2\right]^{1/2}\bar{\rho};$$
$$\bar{\rho} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\widehat{\rho}_{ij}.$$

  - As $N \to \infty$, $CD \to N(0,1)$ under the null hypothesis of weak cross section dependence
  - The hypothesis is stated implicitly as $\alpha < (2-\epsilon)/4$, where $T \to \infty$ at the rate $T = \kappa N^{\epsilon}$.
- Very easy to apply. But difficult to intuit.
  - Suppose N and T increase at the same rate. Then $\epsilon = 1$.
  - In that case, weak dependence hypothesis is $H_0 : \alpha < 1/4$.

Unconventional Spatial Econometrics

- New literature
- Unknown/ Endogenous Spatial Weights Matrix W

Endogenous Spatial Weights – Why?

- Motivating example: Kelejian and Piras (2014)
  - Cigarette demand in the US
  - Spatial spillovers across states
  - Consumers may travel across state boundaries for purchase
  - If distance is small (exogenous) and if prices are low (endogenous)
  - Weights matrix is not unknown but endogenous
  - Distance based (exogenous) W used for IV estimation

- Estimated weights matrices are by definition endogenous

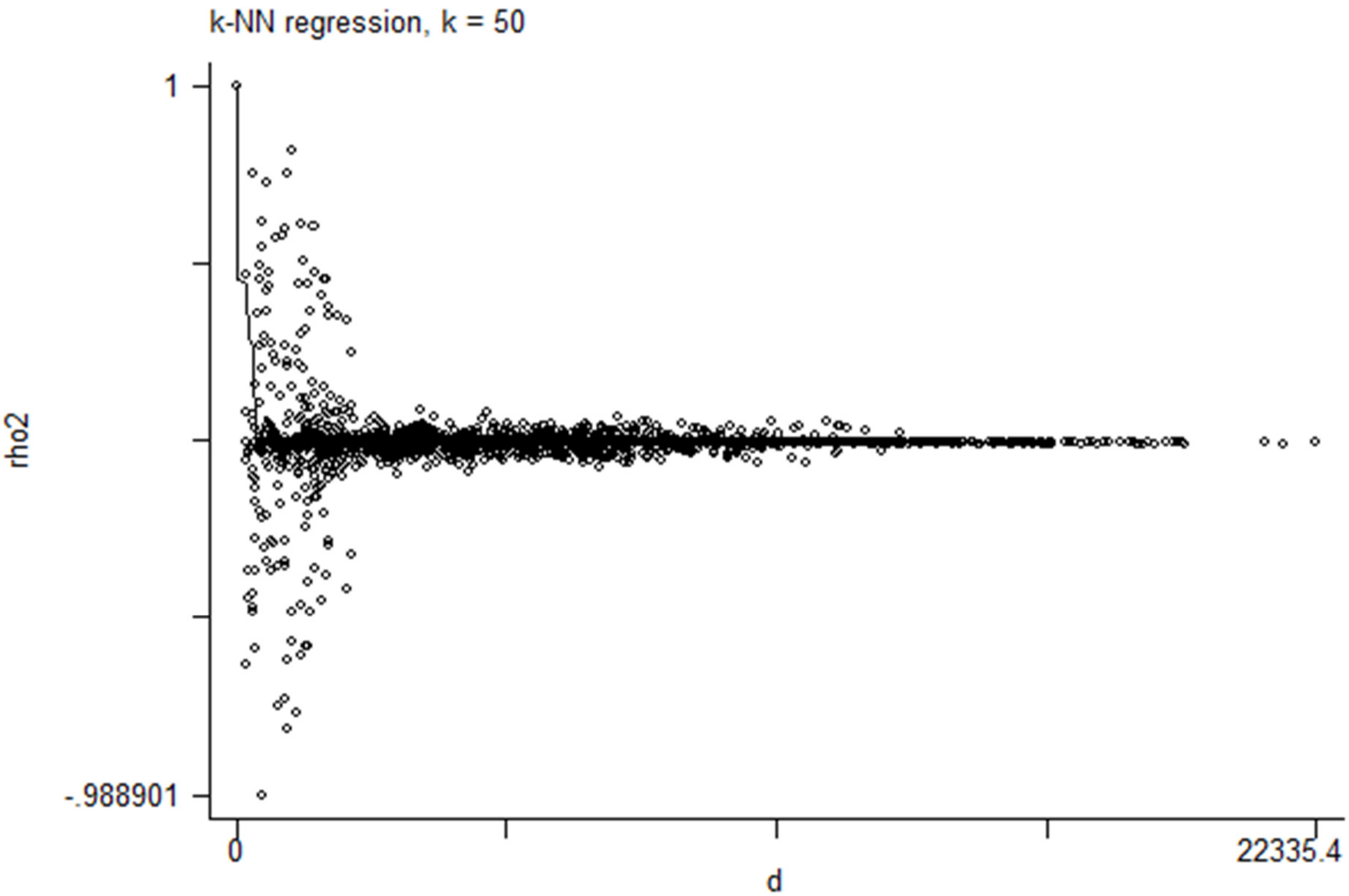
- Lefebvre (1974 [1991]): *The Production of Space*
  - Space is an endogenous product of economic activity
  - *"(Social) space is a (social) product ... the space thus produced also serves as a tool of thought and of action; that in addition to being a means of production it is also a means of control, and hence of domination, of power. ... Change life! Change Society! These ideas lose completely their meaning without producing an appropriate space."*

Why an estimated (unknown) W?

- Different drivers of spatial dependence – substantial uncertainty
  - Geographical distances and contiguity
  - Economic distances – trade, migration
  - Socio-cultural distances
  - Core-periphery relationships – asymmetric spatial weights
  - Negative spatial weights – spatial competition at local scales
  - Different drivers in different places, combinations of interdependencies
- Large and spatially persistent general equilibrium effects
  - Harris et al. (2010); Bhattacharjee, Maiti & Petrie (2013)
  - Even if heterogeneity often provides best explanation (McMillen, 2010)
  - Multiplier effects – interzonal IO-type models (Hewings & Parr, 2007)
- Policy experiments
  - Influence of place based policies, local shocks
  - Different aggregate welfare for policy placed at different places
- Negative spatial autocorrelations
  - Schelling's (1969) model of spatial segregation
  - Patterns of agglomeration and dissociation (Arbia, Espa & Quah, 2008)
  - Spatial competition (Kao & Bera, 2013)
  - Often implies negative spatial weights, particularly at local scales

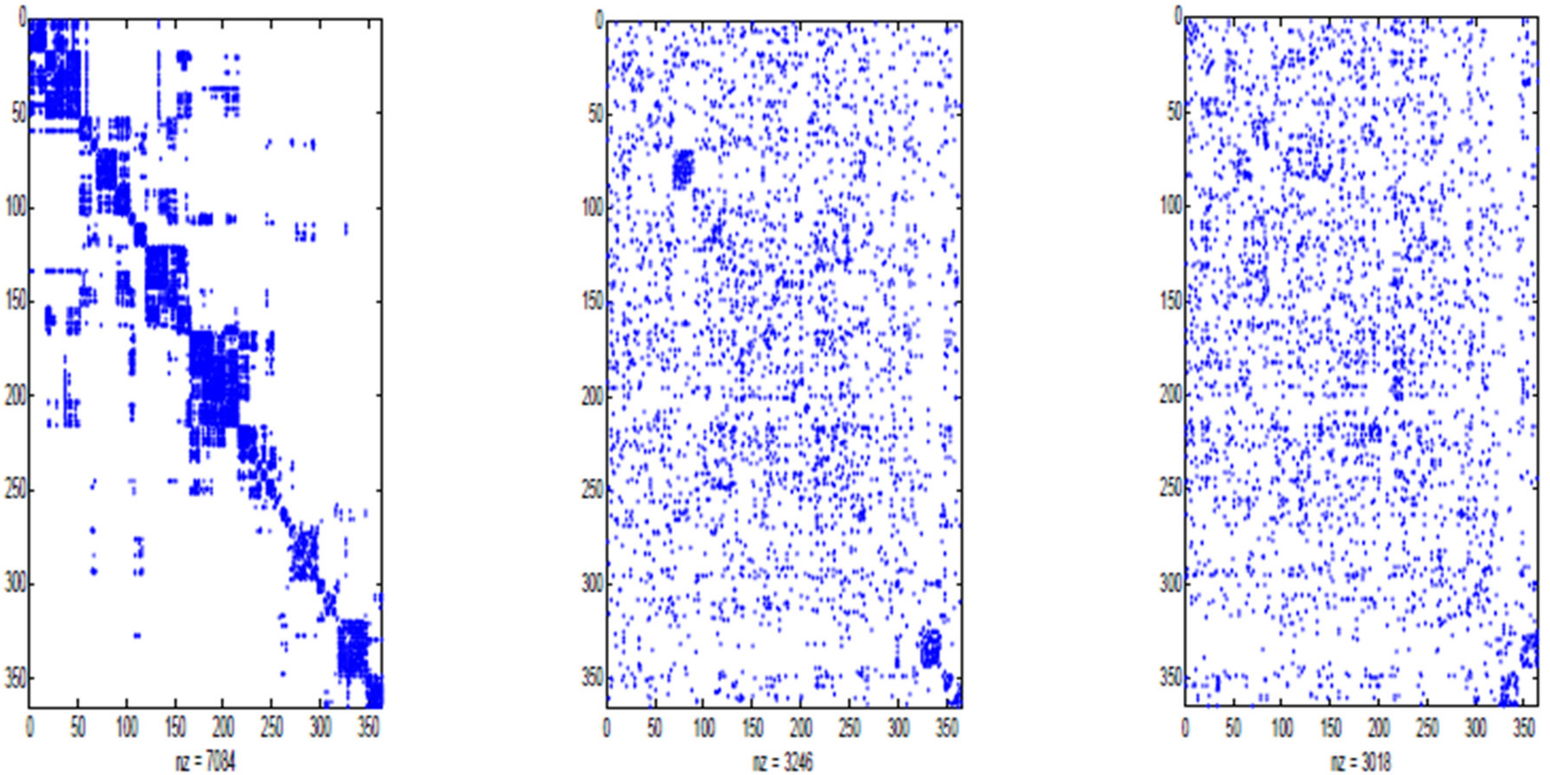
NP regression on irregular domains

(Bhattacharjee, Maiti, Wang and Zhong, 2015)



Spatial Structures: Distance- and Correlation-based Neighbours

W_{200m} $\qquad$ $\hat{W}^{+}$ $\qquad$ $\hat{W}^{-}$



Estimation of W using panel data

[Bhattacharjee and (Late) Jensen-Butler, 2013]

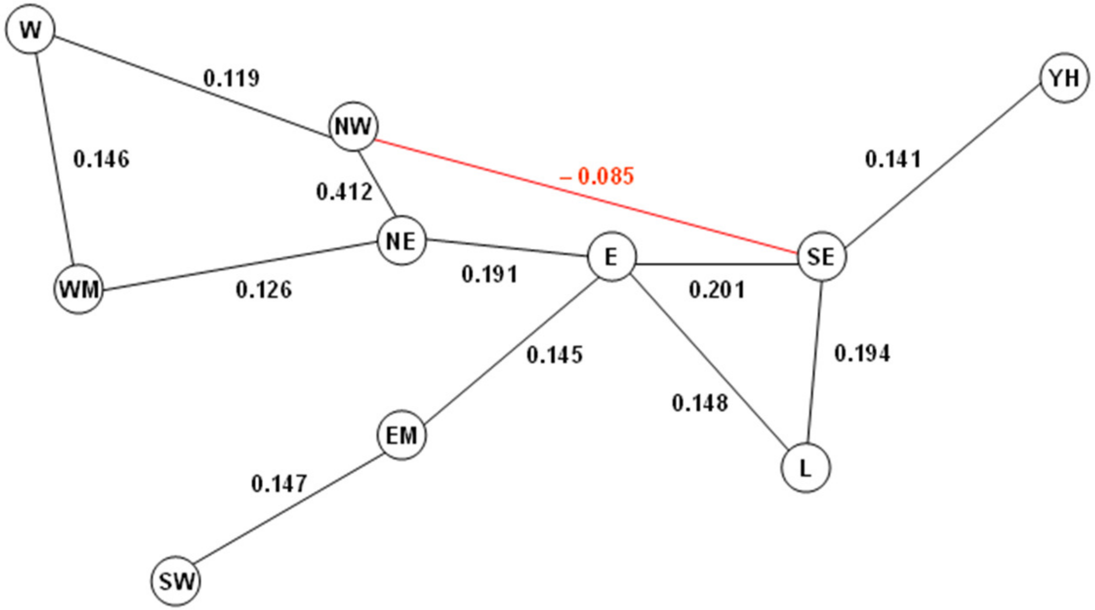
- Spatial error model, fixed N, T increases asymptotically

$$y = X\beta + \varepsilon,\ \ \varepsilon = \lambda W\varepsilon + \eta.$$

  - λ and W are not identified separately; inference on λW=W
  - Meen (1996): Estimate base model, regress resid. for one region on others
  - Does not work because of endogeneity
  - However, spatial autocovariance matrix Σ is (consistently) estimated
- Proposition 1: Non-identification
  - Σ does not fully identify W
  - Intuitively, W has N(N-1)/2 elements, Σ identifies only about half
- Proposition 2: Identification under structural constraints
  - However, a symmetric W is exactly identified, upto proportionality
  - The converse is not true
- Proposition 3: Estimation
  - A (nonlinear) moment estimator minimizes the symmetry metric
  - Consistent and asymptotically Gaussian (under minimal assumptions)
- Proposition 5: Computation
  - Gradient projection algorithm (Jennrich 2001) delivers the above estimator
- Also works for spatial lag model; see Beenstock & Felsenstein (2012)

Abstract space **and housing**

(Bhattacharjee and Jensen-Butler, 2013)



Estimation of spatial weights under *structural constraints*

- Spatial diffusion of housing demand under *symmetry*
- Panel data on govt. office regions in England and Wales
- Negative spatial weights
- Geographic distances or contiguity important
  … but also cultural distances/ core-periphery/gentrification/transport
- Implemented in Matlab

Alternatives to symmetry assumption

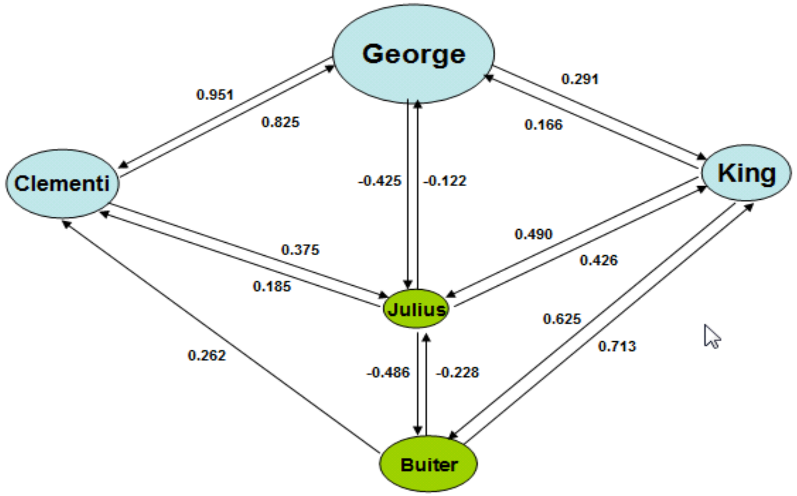
- Symmetric spatial weights not a satisfactory assumption
  - Not appropriate to core-periphery relationships
  - Asymmetric influences in committees, e.g., FOMC Chairman
- Identification under moment conditions (Bhattacharjee & Holly, 2013)
  - Peripheral regions used as instruments for core regions
  - Moment restrictions from higher order spatial lags
  - Analogy with panel data system GMM
  - Additional instruments from temporal lags, plus, lags of exog. regressors
  - Kleibergen-Paap weak instruments tests, plus Hansen-Sargan J test
- Sparseness of W
  - Bailey, Holly & Pesaran (2015): Two contiguity matrices, W^+ and W^-
    - Significant positive & negative autocorrelations, others zero
    - Reconstruct weights matrix as $W = W^+ + W^-$
  - Ahrens and Bhattacharjee (2015): IV Lasso – more later
- Causal graphs (Bhattacharjee and Eibich, ongoing)
  - Can one identify (infer on) triangular systems? $i \rightarrow j$, $(i,j) \rightarrow k$, etc.
  - Lower triangular W will be identified easily, but can we infer this structure?

Social networks & committees, BoE MPC

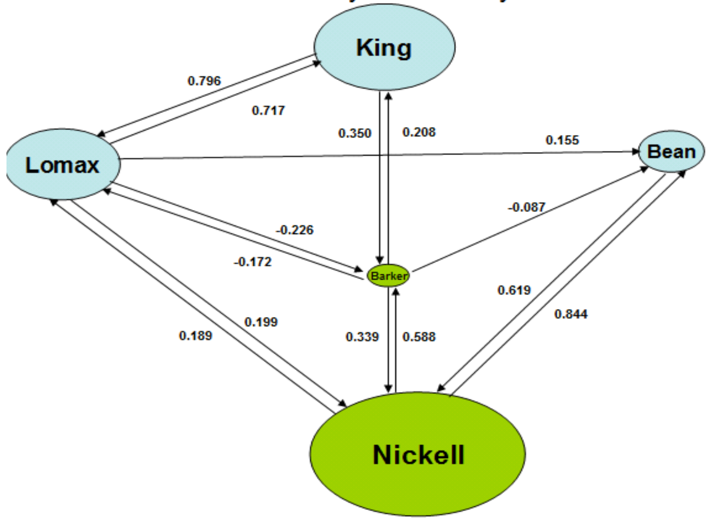
(Bhattacharjee and Holly, 2013)

- Censored regression model
- T >> N $\rightarrow \infty$, moment condns.
- Implementation using *ivreg2*
- Asymmetric network structure
  - *Important implications*
  - *No connection between some members*
- Negative spatial weights
- Large variation in influence

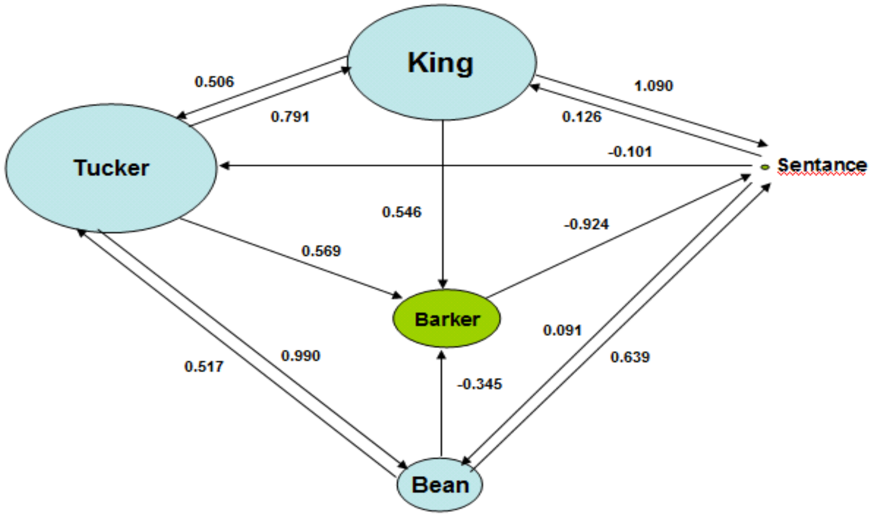
Governor – George
33 months: Sept. 1997 to May 2000



Governor – King
35 months: July 2003 to May 2006



Governor – King
44 months: Oct. 2006 to May 2010



Lasso

Ahrens & Bhattacharjee (2015)

"Two Step Lasso Estimation of the Spatial Weights Matrix"

Here, we consider **estimation of $\mathbf{W}_n$** in the spatial autoregressive panel model

$$y_{it} = \sum_{j=1}^{n} w_{ij} y_{jt} + \mathbf{x}'_{it}\boldsymbol{\beta}_i + e_{it}$$

where y_{it} is the response variable and $\mathbf{x}_{it}$ is a vector of exogenous regressors.

Two major challenges:

- *Endogeneity*, arising from reverse causality.
- *High-dimensionality*: The model is not identified unless the number of parameters, $n(n-1) + Kn$, is smaller than the number of observations, nT, or further assumptions are made.

Our approach:

- assumes that $\mathbf{W}_n$ is approximately sparse (i.e. 'many' zero elements), but allows for $n(n-1) + Kn \gg nT$
- in contrast to previous attempts, requires no further prior knowledge about the network structure.

We propose a two-step estimator based on the Lasso estimator due to Tibshirani (1996).

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,.$$

The ℓ_1-penalization sets some of the coefficient estimates to exactly zero, making the Lasso estimator attractive for **model selection**.

The Lasso is a popular and well-established technique. Recent theoretical contributions include Bickel et al. (2009); Bühlmann and Van de Geer (2011); Belloni et al. (2012).

Important for us: While OLS requires $p \leq n$, the Lasso can deal with $p \gg n$ under the assumption that β is (approximately) sparse, i.e.,

$$s := \|\beta\|_0 \ll n.$$

The two-step Lasso estimator

Recall, the spatial autoregressive model poses two challenges: *Endogeneity* and *high-dimensionality* (i.e., $p \gg n$).

The Lasso allows us to deal with high-dimensionality, but does not address the endogeneity issue.

Inspired by IV/2SLS, **we propose a two-step (post-)Lasso estimator** that allows for endogeneity and high-dimensionality.

Thus, our approach is related to the recently emerging literature that allows for $p \gg n$ and endogeneity; see Fan and Liao (2014, *Ann. Stat.*), Lin et al. (2014, *J. Am. Stat. Assoc.*), Gautier and Tsybakov (2014, *Unpublished*).

Estimating the Spatial Weights Matrix

The spatial autoregressive panel model and its reduced form can be rewritten as

$$\mathbf{y}_i = \sum_{j=1}^{n} w_{ij}^{\star}\mathbf{y}_j + \mathbf{X}_i\beta_i^{\star} + \mathbf{e}_i, \tag{SE}$$

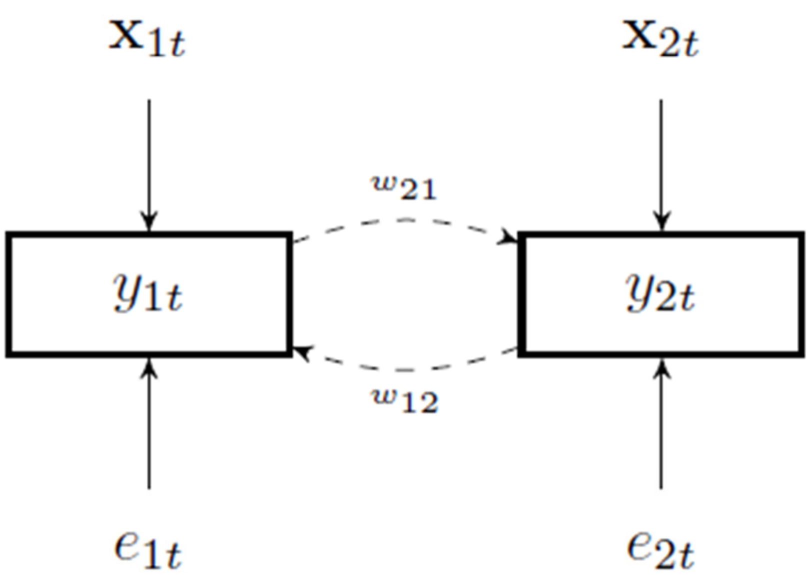
$$\mathbf{y}_j = \sum_{s=1}^{n} \mathbf{X}_s\boldsymbol{\pi}_{j,s}^{\star} + \mathbf{u}_j, \qquad i,j = 1,\ldots,n, \tag{RF}$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, and $\mathbf{X}_i = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')'$ is the $T \times K$ matrix of exogenous regressors.

In the **first step**, the Lasso is applied to (RF), relevant instruments are identified and the predictions $\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n$ are obtained (by Lasso/post-Lasso).

In the **second step**, the Lasso is applied to (SE), but $\mathbf{y}_i$ is replaced with $\widehat{\mathbf{y}}_j$ to account for endogeneity.

Estimating the Spatial Weights Matrix



Intuition for $n = 2$: We exploit $\mathbf{x}_{1t}$ as instruments in order to identify w_{21} which represents the causal effect of y_{1t} on y_{2t}, et vice versa.

Monte Carlo simulation

The data-generating process is given by

$$y_{it} = \sum_{j \neq i} w_{ij} y_{jt} + \eta_i + x_{it}\beta + \varepsilon_{it}$$

with $\beta = 1$, $x_{it} \sim N(0,1)$ and $\varepsilon_{it} \sim N(0, \sigma_{it}^2)$ with

$$\sigma_{it}^2 = \frac{(1 + x_{it}\beta)^2}{\frac{1}{NT}\sum_{i,t}(1 + x_{it}\beta)^2}.$$

We consider two different spatial weights matrices. Specification 1 is given by

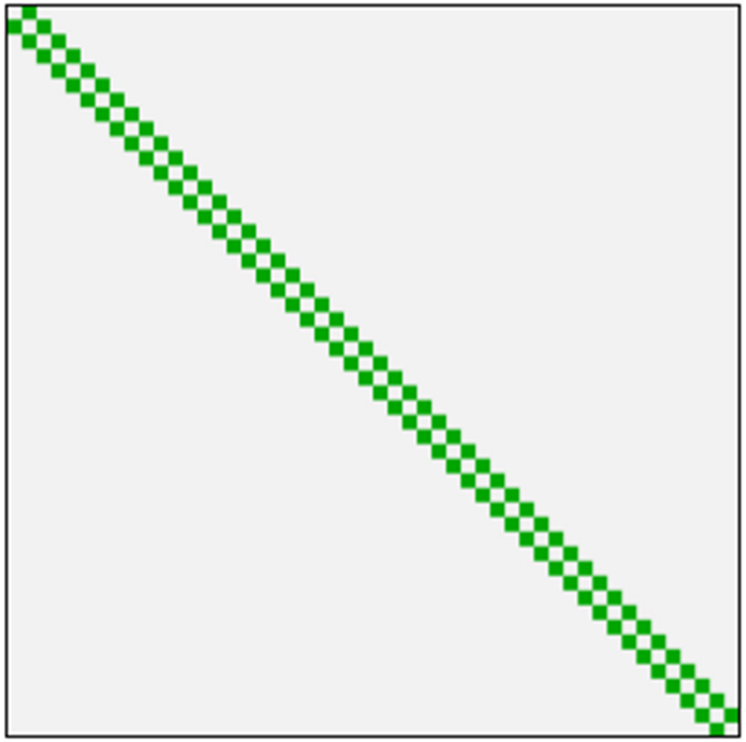
$$w_{ij} = \begin{cases} 1 & \text{if } |j - i| = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i, j = 1, \ldots, n$$

and specification 2 is given by

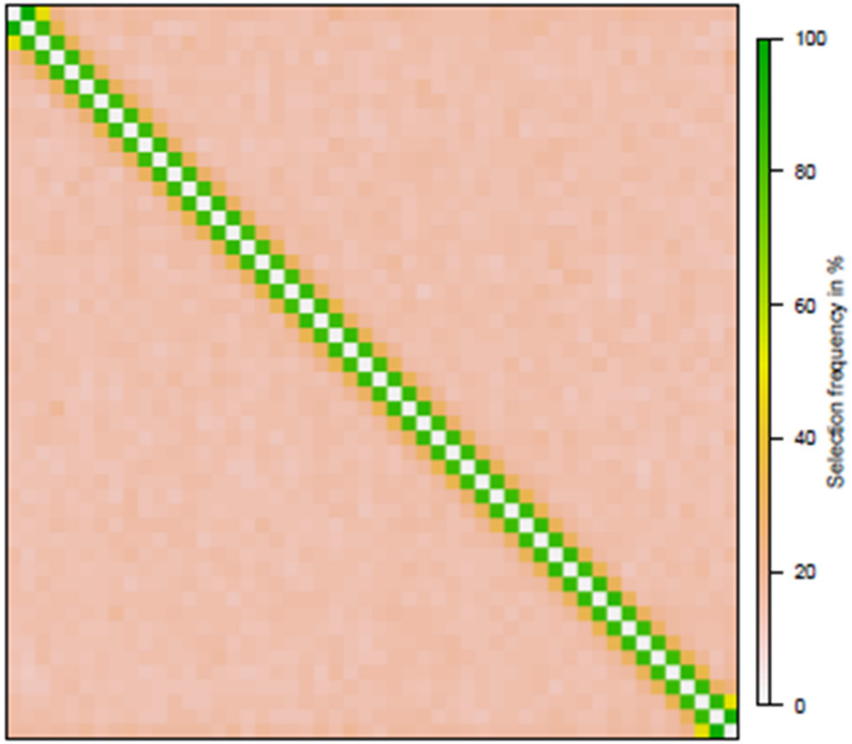
$$w_{ij} = \begin{cases} 1 & \text{if } j - i = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i, j = 1, \ldots, n$$

Monte Carlo simulation

Selection frequency for $n = 50$, $T = 50$.



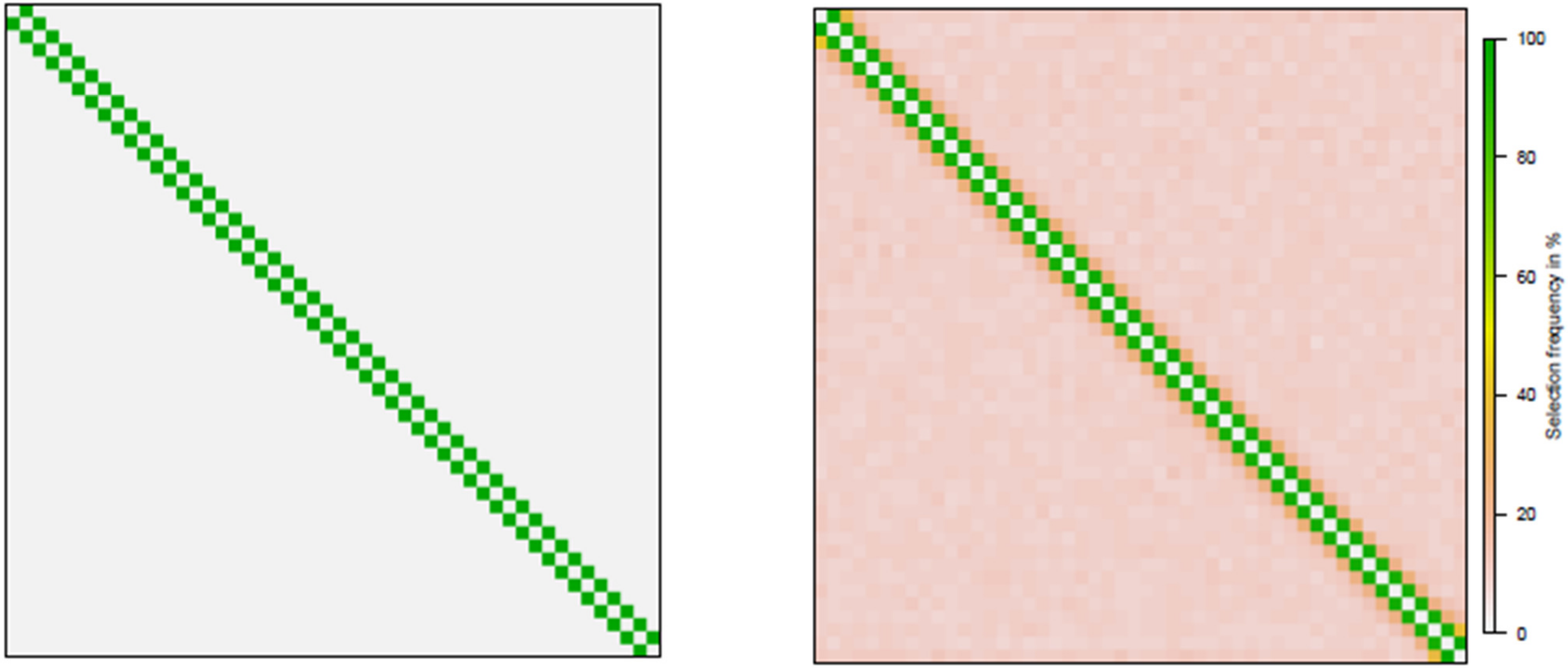
(a) True structure with $\sum_j |w_{ij}| = 0.7$.



(b) Two-step Lasso

Monte Carlo simulation

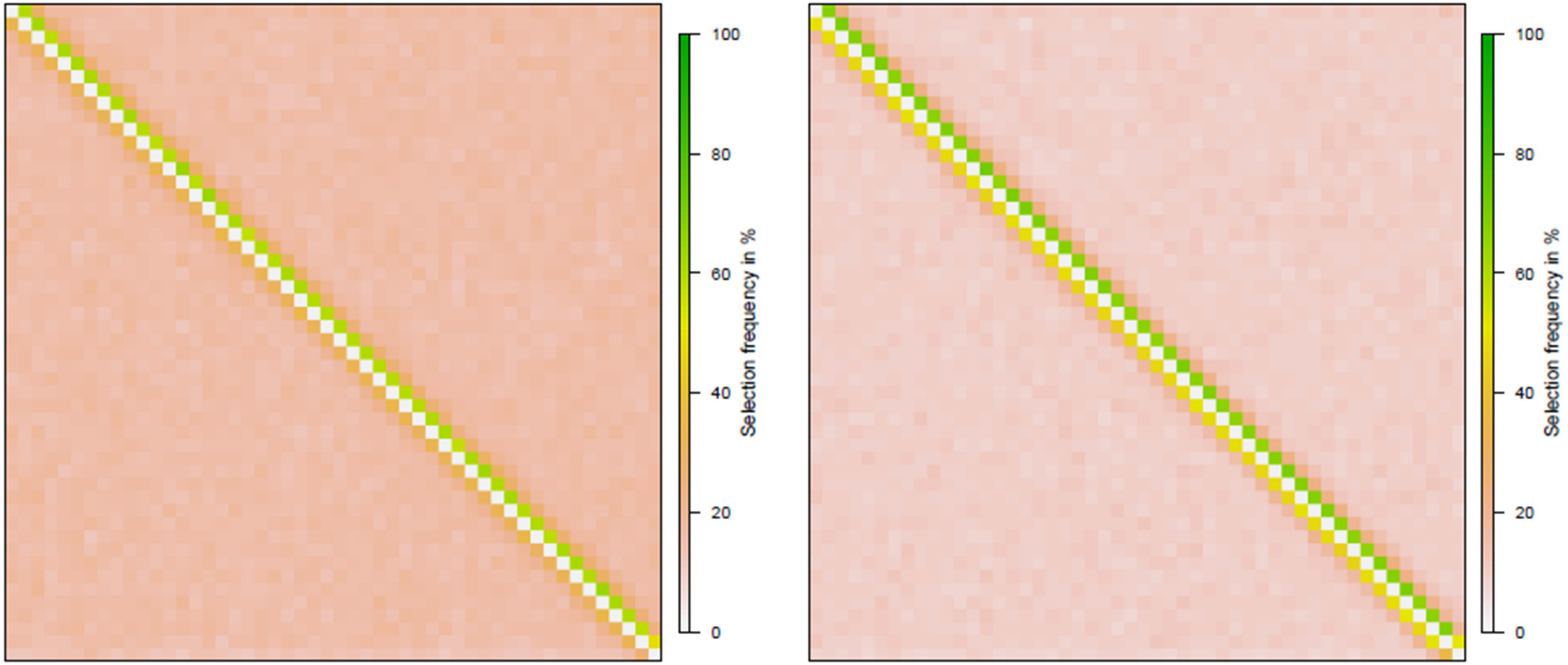
Selection frequency for $n = 50$, $T = 50$.



(a) True structure with $\sum_j |w_{ij}| = 0.7$.

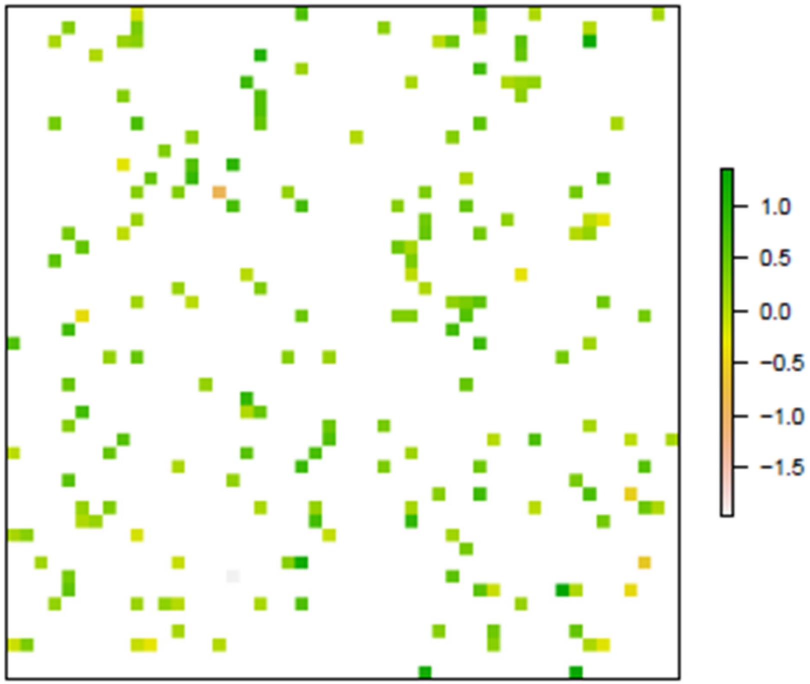
(b) Two-step post-Lasso

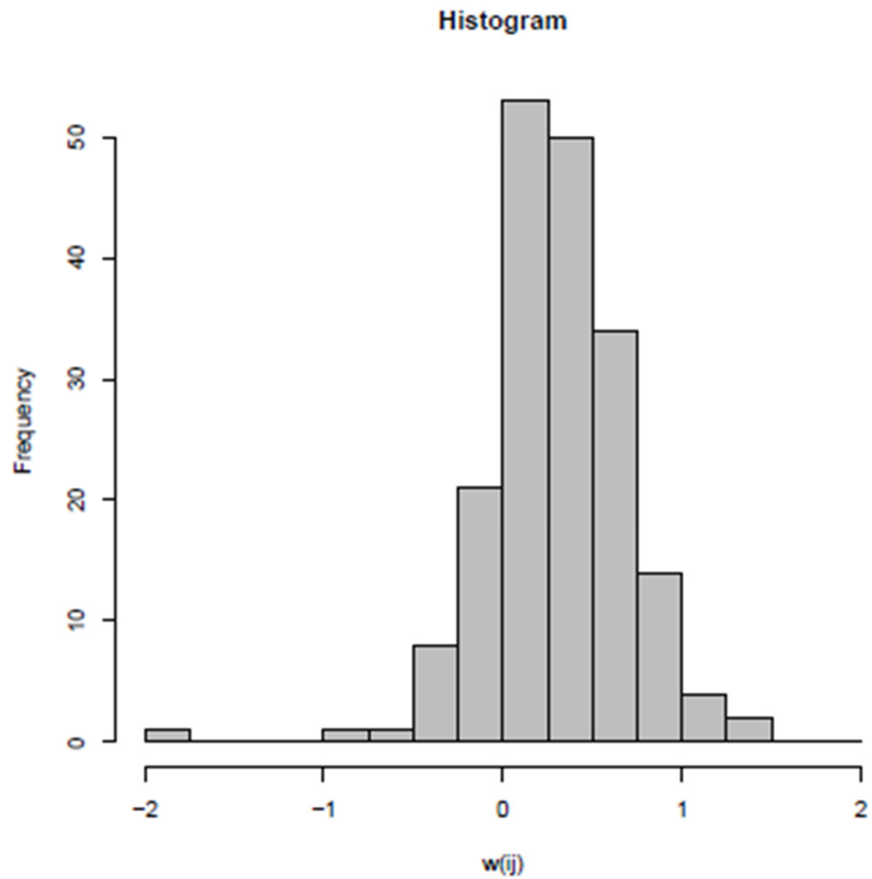
Specification 2
(Lasso and post-Lasso)



Application to US Housing Market

Estimated weights matrix





Application to US Housing Market

Estimated weights matrix

Can geographic distance explain interaction effects?

	$\lvert w_{ij} \rvert$	$1\{w_{ij} \neq 0\}$	
	OLS	*OLS*	*Probit*†
	(1)	(2)	(3)
$log(d_{ij})$	−0.015***	−0.057***	−0.052***
	(0.005)	(0.010)	(0.009)
n_{ij}	0.040***	0.066***	0.034
	(0.011)	(0.023)	(0.024)
R^2	0.021	0.038	

Note: *p<0.1; **p<0.05; ***p<0.01

† Average marginal effects shown.

Spatial Discrete Choice Model

Competing club and congestion effects

Uncovering club and congestion effects

[Bhattacharjee, Hicks and Schnier, 2015]

- Utility for vessel *i* of visiting a site *s* is:

$$U_{ist} = \delta_{st} + \beta X_{ist} + v_{ist} + \varepsilon_{ist}$$

$$\delta_{st} = \gamma_t + \lambda Z_{st} + \mu_{st}$$

$$v_{ist} = u_{ist} + \sum_{j \neq i} w_{ij} u_{jst} \;\; ; \;\; \underline{v}_{st} = \underline{u}_{st} + W_C \underline{u}_{st}$$

- Error assumptions

$$\varepsilon_{ist} \sim iid \;\; GEV$$

$$u_{ist} \sim iid \;\; N(0, \sigma_u^2)$$

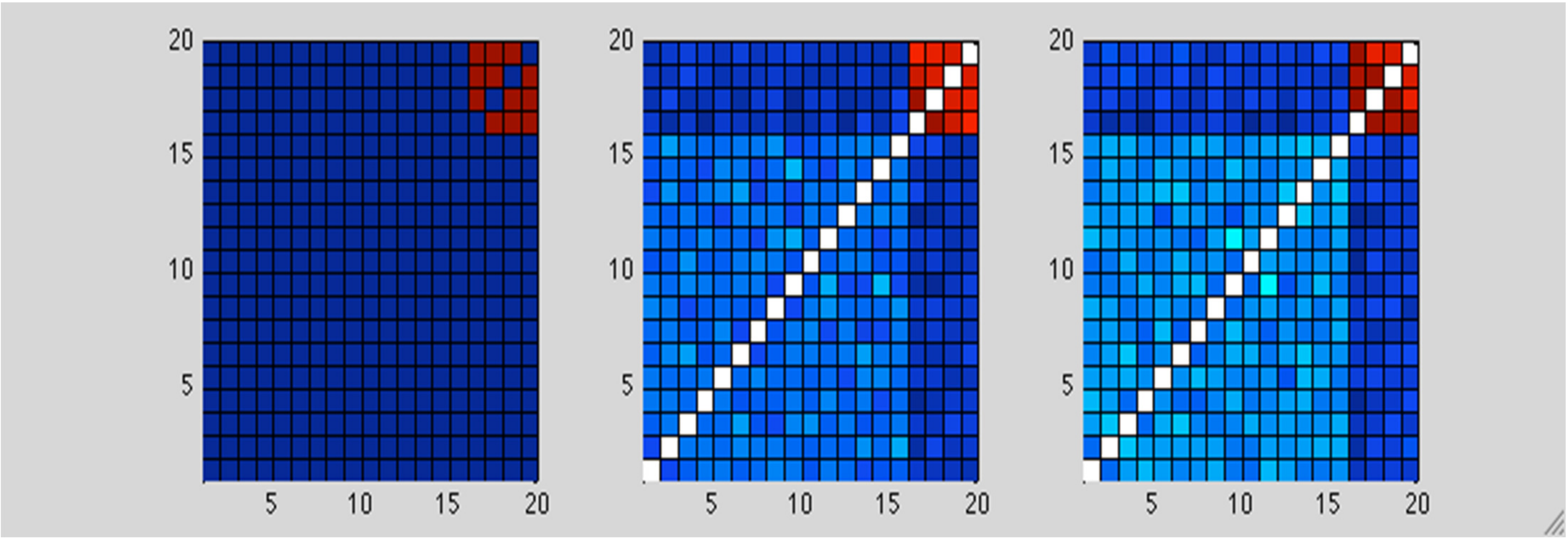
- Estimation of β's Berry-Levinson-Pakes (BLP) and Random Utility model (RUM)
- Symmetric W (Bhattacharjee Jensen-Butler, 2013)

Estimation

- Estimation of *β*'s
  - Berry-Levinson-Pakes: Stata *blp* (Vincent, 2013)
  - Random Utility model: Stata *mixlogit* (Hole, 2007)
  - Very poor estimates - correlated REs cannot be ignored!
  - However, estimates of club effects very good!
- Try Stata *femlogit* (Pforr, 2014)
  - Estimates of *β* improved, not brilliant
  - But W-estimates poor
  - Random effects cannot be extracted from fixed effects
- Finally, try Stata *gllamm*
  - Rabe-Hesketh, Skrondal and Pickels (2003, 2005); Rabe-Hesketh and Skrondal (2003, 2012)
  - Very difficult to implement, but excellent results

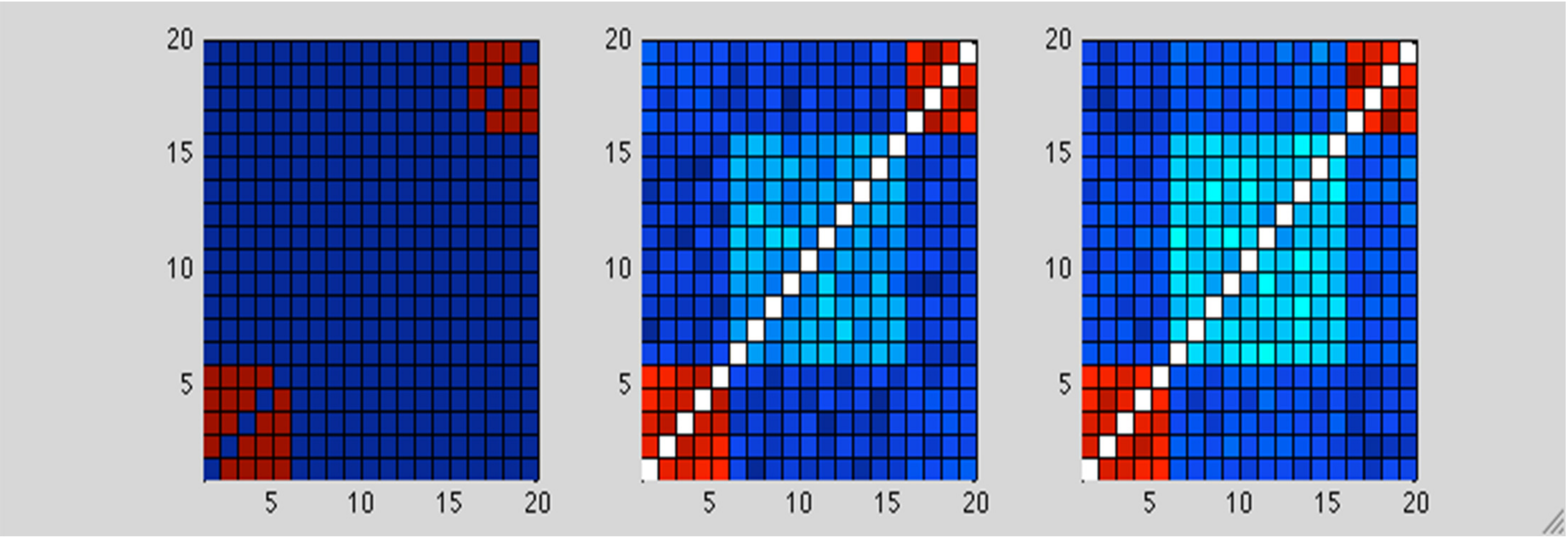
Monte Carlo 1

Model	β_1		β_2		β_3	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
RUM	0.0291	0.0009	-0.1437	0.0206	---	---
Club/Congestion	0.0238	0.0006	0.2088	0.0442	0.5154	0.3707



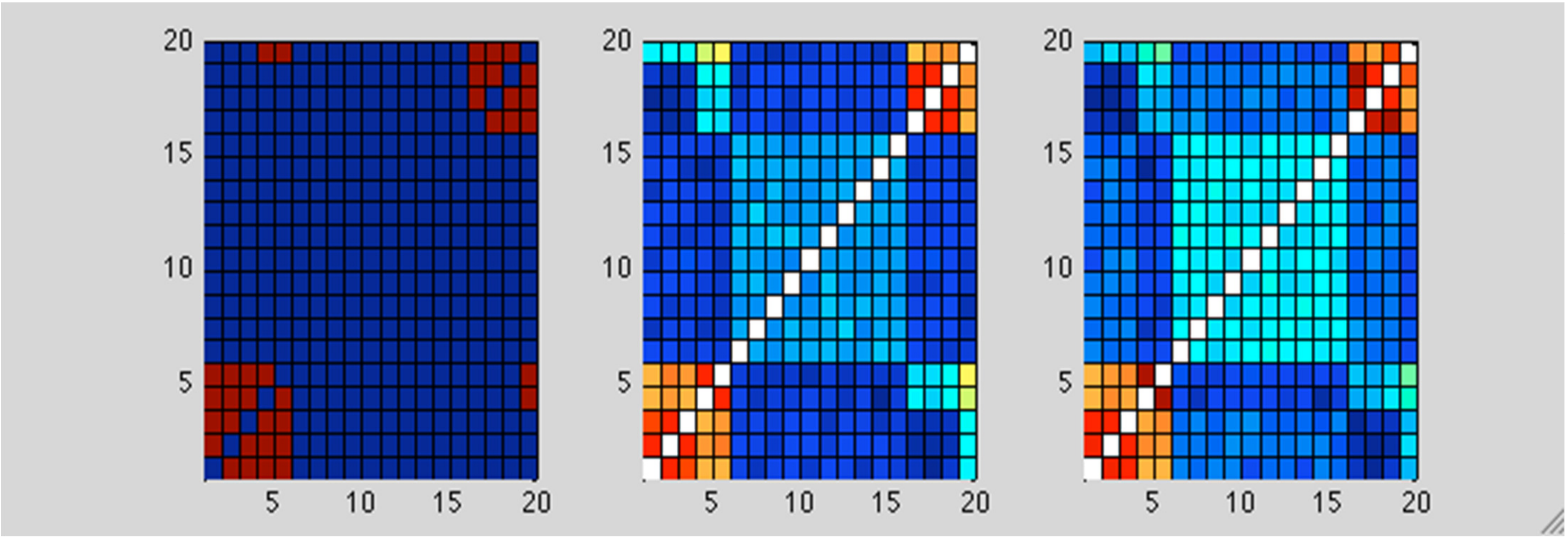
Monte Carlo 2

Model	β_1		β_2		β_3	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
RUM	0.0346	0.0012	-0.1485	0.0220	---	---
Club/Congestion	0.0254	0.0006	0.1813	0.0337	0.9760	1.0905



Monte Carlo 3

Model	β_1		β_2		β_3	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
RUM	0.0356	0.0013	-0.1494	0.0223	---	---
Club/Congestion	0.0268	0.0007	0.1760	0.0318	1.0654	1.2791

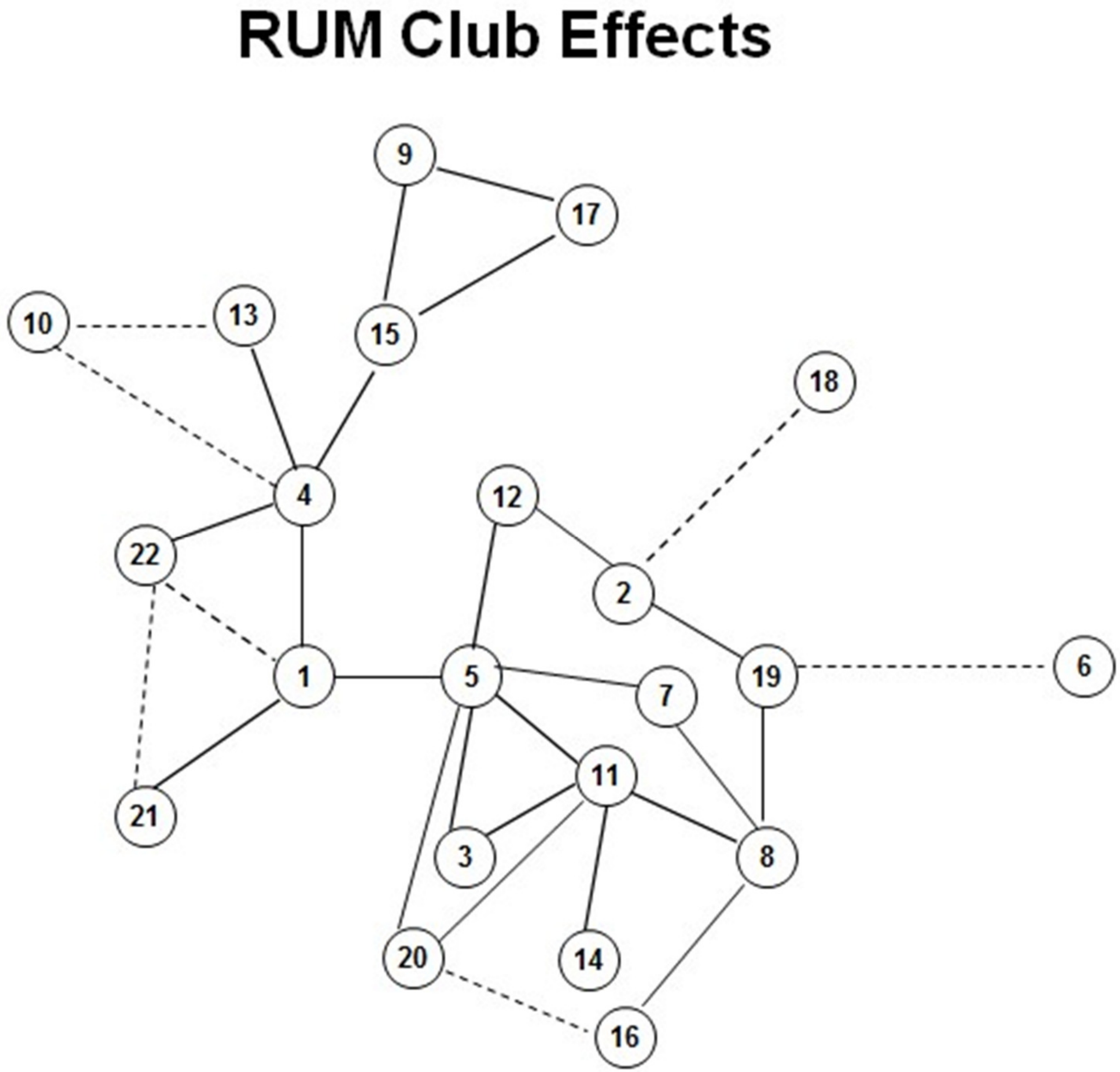


Adaptive quadrature random effects multinomial discrete choice – *gllamm*

- PRELIMINARY
- True parameter values
  - Average catch at location (x): +0.25
  - Congestion (p): –5
  - Distance from yesterday's location (d): –1
- Monte Carlo design
  - 20 vessels, 3x3 grid locations, 12 days
- Preliminary results (average estimates)
  - *femlogit*: $\beta_x = 0.084$; $\beta_p = -14.217$; $\beta_d = -0.249$
  - *gllamm* (non-adaptive): $\beta_x = 0.482$; $\beta_p = -9.862$; $\beta_d = -1.999$
  - *gllamm* (adaptive quad.): $\beta_x = 0.323$; $\beta_p = -2.934$; $\beta_d = -1.133$
- Estimates of club structure similar to BLP

Fishing vessels in the Bering Sea

(Bhattacharjee, Hicks and Schnier, 2015)



Thank you very much!!!

SEEC

Spatial Economics

& Econometrics Centre

Heriot-Watt University
Edinburgh

Exciting work continues ...

... Will be great to have you on board

26th $(EC)^2$ Conference

"Theory and Practice of Spatial Econometrics"

Spatial Economics & Econometrics Centre (SEEC)

Heriot-Watt University, Edinburgh, 18-19 December 2015

SEEC

Spatial Economics & Econometrics Centre

http://seec.hw.ac.uk/ Heriot-Watt University, Edinburgh, UK

Invited Speakers:

Lung-Fei Lee (Ohio State University, USA)
Taps Maiti (Michigan State Univ., USA /Heriot-Watt Univ.)
Daniel McMillen (Univ. of Illinois Urbana-Champaign, USA)
Ingmar Prucha (University of Maryland, USA)
Peter Robinson (London School of Economics)

Scientific Committee:

Sean Holly, Chair (University of Cambridge)
Giuseppe Arbia (Univ. Cattolica del Sacro Cuore, Italy)
Anil Bera (University of Illinois Urbana-Champaign, USA)
Bernie Fingleton (University of Cambridge)
Jesus Mur (University of Zaragoza, Spain)
Mark Schaffer (Heriot-Watt University)

Organising Committee:

Arnab Bhattacharjee, Chair (Director, SEEC, Heriot-Watt Univ.)
David Cobham (Heriot-Watt University)
Gary Koop (Strathclyde University)
Dimitris Korobilis (Univ. Glasgow)
Rod McCrorie (Univ. St Andrews)
Andy Snell (Univ. Edinburgh)

Secretariat: *Jõao Marques* (Univ. Aveiro, Portugal); *Achim Ahrens, Yuka Scott* (Heriot-Watt Univ.)

Venue: Postgraduate Centre, Heriot-Watt University, Riccarton Campus, Edinburgh, Scotland, UK

***Paper submission*:** Deadline 2 Oct. 2015; Acceptance 1 Nov. 2015; URL https://easychair.org/conferences/?conf=ec2seec

Information: a.bhattacharjee@hw.ac.uk

About $(EC)^2$

EC-squared (https://sites.google.com/site/ecpower2/) is a series of annual international conferences on research in quantitative economics and econometrics, launched in 1990. The acronym $(EC)^2$ stands for European Conferences of the Econom[etr]ics Community. The $(EC)^2$ is one of the oldest and most prestigious scholarly society in econometrics. Each annual conference focuses on a specific area of theoretical and/or applied econometrics and usually attracts about 50 top academics worldwide. About 5 of these are selected as invited speakers. From the contributory submissions, about 15 are selected for plenary presentations, and another 30 or thereabouts for poster presentations. There are no parallel sessions.

The conference is in its 26th year and this is the third time it is being organised in the UK (first time in Scotland). Previous occasions were Oxford University (1993, Programme Chair Professor Sir David Hendry, Organiser Professor Neil Shephard) and Centre for Microdata Methods and Practice, University College London (2003, Programme Chair Professor Geert Dhaene, Organiser Professor Andrew Chesher).

The programme chair this year will be Dr Sean Holly, University of Cambridge. Following Scottish tradition in economics, local organisation will be shared across the Scottish universities. The invited speakers are among the topmost in the field worldwide, and the scientific committee likewise. This year, we propose to increase the posters by about 10, and will use this to promote the work of doctoral students and early career researchers. We also intend to give a token prize to the best poster by an early career researcher.

About SEEC

There is a fairly established view, within the discipline of spatial econometrics and outwith, that empirical work in the area does not relate closely to economic theory. Further, in the main, economic theory provides very little insights into how spatial structure should be modeled in applied spatial economic problems. This growing concern led to the establishment of a new research centre, The Spatial Economics & Econometrics Centre (SEEC), in Heriot-Watt University in 2013. The primary objectives of SEEC are to develop, support and conduct high-impact theoretical, applied and policy-relevant research on spatial econometrics, focusing on economics as the core discipline, but building upon research synergies with other subjects: statistics, demography, urban studies, regional studies, fluid mechanics, management sciences, geography, history and philosophy. The idea was to develop SEEC into a world-leading research centre focusing on theoretical and empirical research on spatial social sciences, with impact on business, society and policy. Actively engaging academics and practitioners, and engaging in inter-disciplinary and multi-disciplinary research programmes, the aims of SEEC were not only to provide research leadership at the frontier, but also to develop itself as a repository of domain knowledge for all researchers in the area, across universities and research institutions in the UK, but also through visitor and outreach programmes and a working paper series, for researchers working in similar areas anywhere in the world. For more information, see http://seec.hw.ac.uk/

Substantial progress has been made on the above ambitious objectives. In the two years of its existence, SEEC has emerged as a world leading centre for research on unknown and endogenous spatial structure. Applied contexts have been drawn from a range of areas: monetary policy committees, housing and urban economics, health economics, demographic change, conflict, migration, supply chain management, and regional economic growth. Substantial policy implications have been generated. Outreach has been promoted by organisation of major conferences, such as the $(EC)^2$ Conference on '*Theory and Practice of Spatial Econometrics*' (Heriot-Watt University, 2015) and the ENHR (European Network for Housing Research) Annual Conference (Lisbon 2015), and invited sessions at the American Statistical Association Joint Statistical Meetings (Boston, 2014) and Spatial Econometrics Association Annual Conference (Washington DC, 2013). Close collaboration and exchange programmes, for doctoral students and researchers, have been developed with several institutions, including: Center for Business and Social Analytics (Michigan State University, USA), Centre for International Money and Finance (University of Cambridge, UK), Regional Economics Applications Laboratory (University of Illinois at Urbana-Champaign, USA), Spatial Economics Research Centre (London School of Economics, UK) and University of Aveiro (Portugal). Leading researchers from across the world are visiting SEEC, together with junior researchers from several leading institutions, working in collaboration and disseminating their work through the SEEC Working Paper Series.

SEEC has developed a rich agenda for future research in the spatial sciences. While some progress has been made on achieving its ambitious objectives, much more is required. Many exciting research problems have been identified, both theoretical and applied. SEEC is above all a space to facilitate the meeting of minds interested in the analysis of spatial phenomena, whatever the discipline or application area. Please engage with SEEC in whatever way you can. There are lots of spatial issues that we together can help address.