Estimating spatial panel models using unbalanced data

Gordon Hughes
University of Edinburgh

Andrea Piano Mortari & Federico Belotti
CEIS, Universita Roma Tor Vergata

12th September 2013

Outline

- Reasons for using spatial panel models?
  - Spatial interactions – e.g. tax & environmental policies
  - Spatial spillovers – migration or relocation of industrial activity
  - Controlling for spatially-correlated omitted variables
- Econometric models, data and software
  - Spatial lags & errors – parallels with time series models
  - Stata, R & Matlab – community routines
- Unbalanced panels
  - Changes in population of countries, states, etc
  - Spatial interactions with missing data
- US electricity demand by state
  - Price effects and regulation

Spatial analysis in Stata

- Variety of special purpose routines written by users and available through SSC
  - Manipulation of spatial data
  - Cross-section spatial regressions
- StataCorp-related routines – also through SSC
  - shp2dta converts ESRI shapefiles to dta files – similar to programs converting to csv or xls files
  - spmat, spreg, spivreg, etc for construction & manipulation of spatial weights and for cross-section spatial regressions

Nature of spatial panel data

- Large N and/or large T?
- Missing data and spatial weights
  - Contiguity vs inverse distance
  - To (row) standardise or not?
- Examples:
  - Energy demand – gasoline, electricity, etc
  - State tax and fiscal policies
  - Cross-country models of economic development
  - Spatial hedonic models & hedonic valuation

Econometric specification

- Fixed or random effects – can we talk about random effects with complete sample of states or countries?
- Lagged dependent variable or within panel serial correlation
- Why are data missing – missing at random assumption

Key models

Spatial auto-regression model (SAR)

$$y_{it} = \rho W y_t + X_{it}\beta + \mu_i + \varepsilon_{it}$$

Spatial Durbin model (SDM)

$$y_{it} = \rho W y_t + X_{it}\beta + WX_t\varphi + \mu_i + \varepsilon_{it}$$

Spatial autocorrelation model (SAC)

$$y_t = \rho W y_t + X_t\beta + \mu + v_t \text{ with } v_t = \lambda M v_t + \varepsilon_t$$

Key models 2

Spatial error model (SEM)

$$y_{it} = X_{it}\beta + \mu_i + v_{it} \text{ with } v_{it} = \lambda W v_t + \varepsilon_{it}$$

Generalised spatial random errors (GSPRE)

$$y_t = X_t\beta + \mu + v_t \text{ with } \mu = \rho_1 W\mu + \eta \text{ and } v_t = \rho_2 M v_t + \varepsilon_t$$

Procedure xsmle - syntax

xsmle varlist [if] [in] [weight], WMATrix(string) [MODel(string) FE RE EMATrix(string) DMATrix DURBin(varlist) ROBust DKRAAY(#) DLAG ERRor(#) NOConstant]

- "varlist" = depvar indvars [required].
- "wmat(WN)", "emat(WE)", "dmat(WD)" refer to an N x N matrices of spatial weights for spatial lags, spatial errors and Durbin variables [at least one of wmat() or emat() is required].
- "model(string)" specifies the type of model to be estimated. The default is "sar" and alternatives are "sdm", "sem", "sac" and "gspre".
- "fe | re" specifies that a fixed or random effects model should be used – the default varies according to the model specified.

Procedure xsmle – syntax 2

- “durbin(varlist)” specifies a set of spatially-weighted regressors.
- “vce()” specified type of variance-covariance estimator – options include likelihood-based and sandwich estimators:
  - hessians from optimization – vce(oim), vce(opg);
  - panel & cluster robust standard error – vce(robust) vce(cluster clusvar);
  - Driscoll-Kraay variant of Newey-West robust standard errors with default or specific lag – vce(dkraay #)
- "dlag" includes the lagged dependent variable in the model. This is only available for model(sar) and model(sdm).
- “err(#)” specifies the error structure for the GSPRE model. The default is the most general version ($\rho 1 \neq \rho 2 \neq 0$).
- "noconstant" specifies that the model should be estimated without adding a constant term.

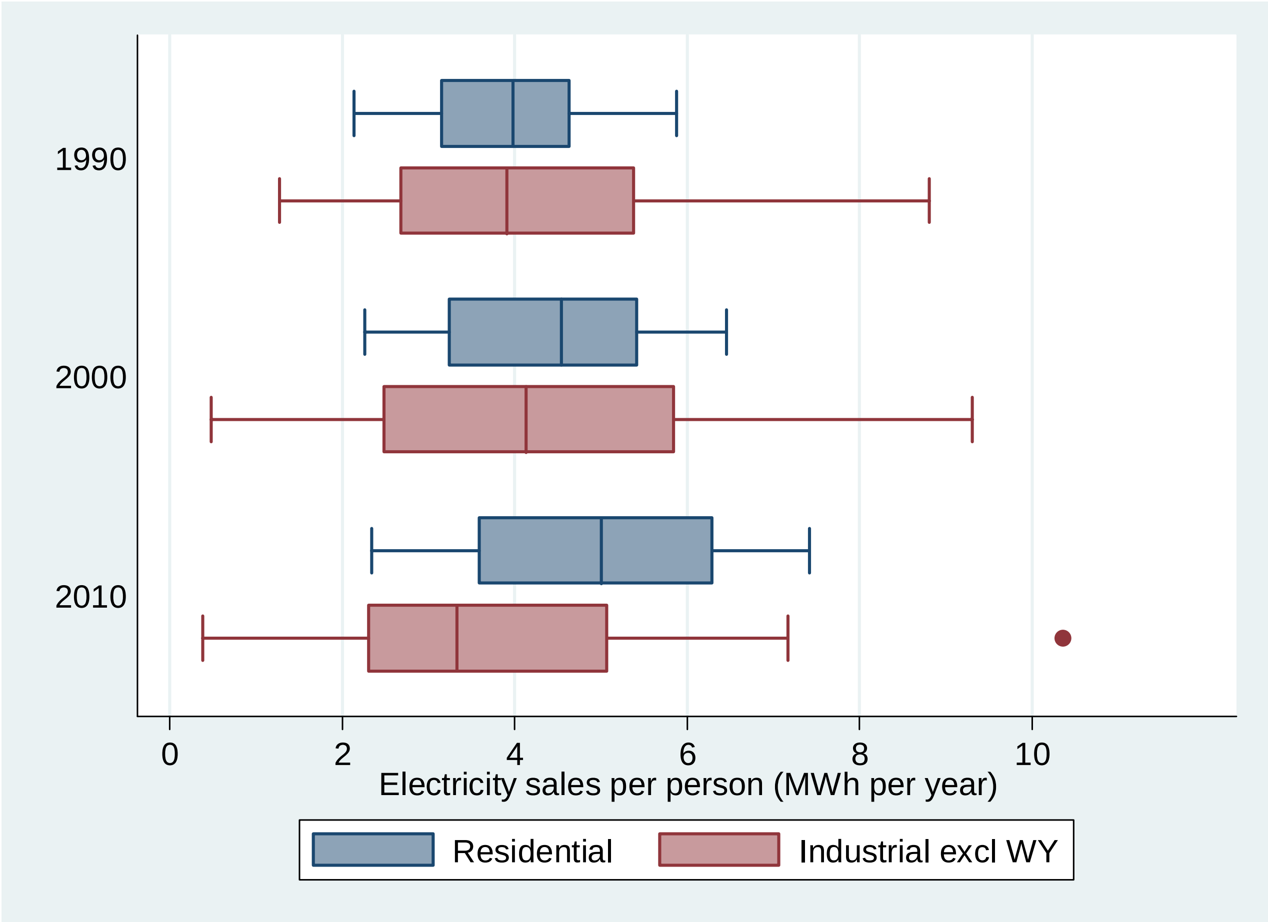
Features of xsmle

- Fast for N ~ 500, copes with N ~ 2000
  - Memory & multiple core processing beneficial
- Full range of Stata options for ML estimation and post-estimation
- Quite general syntax & options
  - Multiple sets of spatial weights for different components
  - Selection of Durbin variables
  - Both individual and time fixed effects permitted
  - Analytical & important weights permitted
- Generates estimates of direct & indirect impacts plus associated standard errors (by Monte Carlo sampling)

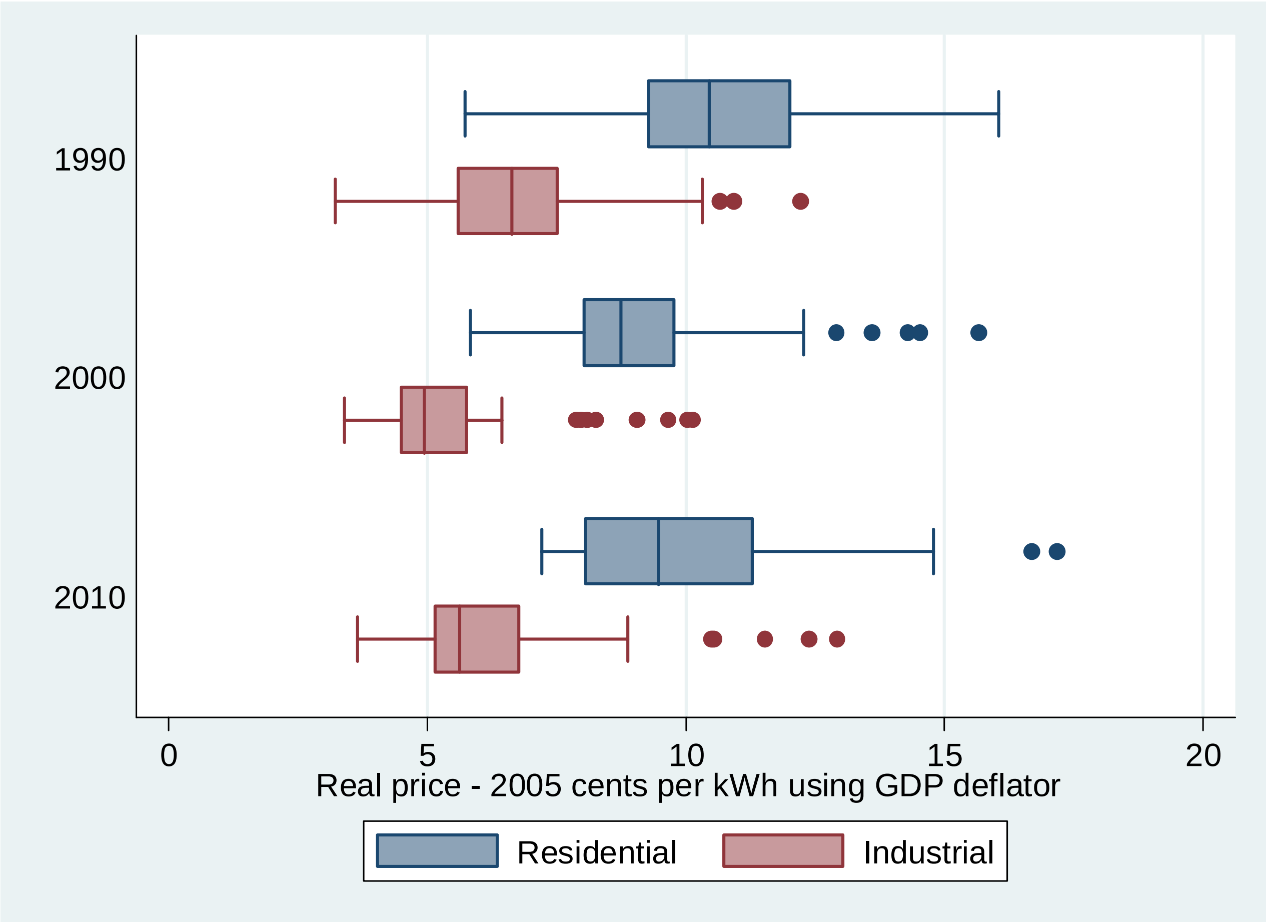
Illustration – US electricity demand

- State data – continental US, 1990-2011
  - Electricity demand by sector
  - Regressors - prices, weather (heating & cooling days)
- Focus on price elasticities and weather impacts
- Likely to be spatial interactions due to
  - Common factors in unobserved variables
  - Competition between states for industry and/or movement of households

Electricity sales per person



Electricity prices by state
- adjusted by state GDP deflator



Residential demand - FE models

Variables	Non-spatial panel	SAR	SDM
	(1)	(2)	(3)
W*Y		0.388***	0.456***
		(0.056)	(0.050)
ln(Real personal income per person)	0.381***	0.179***	0.198***
	(0.042)	(0.046)	(0.044)
ln(Real average residential price)	-0.243***	-0.246***	-0.294***
	(0.037)	(0.034)	(0.035)
ln(Housing units per person)	1.039***	0.756***	0.658***
	(0.123)	(0.106)	(0.110)
ln(Cooling degree days)	0.0718***	0.0527***	0.0523***
	(0.013)	(0.011)	(0.010)
ln(Heating degree days)	0.189***	0.139***	0.126***
	(0.025)	(0.027)	(0.026)
W*ln(Real average residential price)			0.190***
			(0.044)

Unbalanced panels - options

- Listwise deletion
  - Can mean loss of all or most of sample
- ML estimation of joint model
  - Pfaffermayr for GSPRE model
- Treating panel as pooled cross-section
- Imputation
  - Single imputation can be useful for spatial lags but see Cameron & Trivedi
  - Multiple imputation using Monte Carlo chain approach

ML estimation

- See Pfaffermayr – Spatial Economic Analysis 2009
- GSPRE model – spatially correlated random effects + spatial autocorrelation
- Implemented in Mata code – works on simple test runs with 1 or 2 exogenous variables
- Poor performance in practical cases
  - Failure to converge is very common – non-concave objective function
  - Very sensitive to starting values
  - Not recommended

Pooled cross-section estimation 1

- See Baltagi et al – Journal of Econometrics 2007 & Egger et al – Economics Letters 2005
- Pool cross sections with different sets of panel units (countries) for each period
  - Create spatial weights W_t for each t by row/col deletion and (perhaps) standardisation
  - Full matrix of spatial weights is block diagonal with W_1 .. W_T as the diagonal elements
- Estimate using cross-section spatial procedure such as –spreg- including panel unit dummies for fixed effects

Pooled cross-section estimation 2

- Implemented in Mata with –spmat- and –spreg-
  - Good execution speed and seems robust
- Conceptual issues
  - How to interpret time-varying spatial interactions?
    - Reasonable when the population is changing – e.g. units splitting up or merging
    - Arbitrary exclusion when driven by missing data
    - Should the W_t be row-standardised?
  - Missing data leads to islands with contiguity weights
- Tests: coefficients are severely biased with potentially serious impact on hypothesis tests

Multiple imputation

- -xsmle- has been set up to permit use with –mi-
- Care is needed in specifying the method of imputation that is used – tests use regression imputation controlling for state effects
- Significant cost of setting up & testing the imputation framework
- After this the computational cost is reasonable so advice is to use M > % of missing data
  - Less expensive than bootstrap standard errors – at least with a proper number of repetitions

Comparison of methods 1
Missing y's: coefficient estimates

	No missing data		10% missing data		25% missing data		50% missing data	
	XSMLE - FE	Pooled	Pooled	MI	Pooled	MI	Pooled	MI
Real income	0.105***	0.105***	0.351***	0.107***	0.375***	0.0874**	0.393***	0.147**
	(0.0235)	(0.0235)	(0.0185)	(0.0257)	(0.0187)	(0.0330)	(0.0224)	(0.0532)
Real prices	-0.248***	-0.248***	-0.240***	-0.243***	-0.235***	-0.225***	-0.228***	-0.227***
	(0.0120)	(0.0120)	(0.0138)	(0.0130)	(0.0153)	(0.0155)	(0.0183)	(0.0219)
Housing per person	0.628***	0.628***	1.014***	0.645***	1.063***	0.661***	1.002***	0.708***
	(0.0584)	(0.0584)	(0.0619)	(0.0635)	(0.0688)	(0.0795)	(0.0839)	(0.126)
Cooling index	0.0499***	0.0499***	0.0686***	0.0438***	0.0649***	0.0348***	0.0510***	0.0264**
	(0.00593)	(0.00593)	(0.00655)	(0.00644)	(0.00728)	(0.00743)	(0.00847)	(0.01000)
Heating index	0.127***	0.127***	0.186***	0.118***	0.178***	0.0926***	0.162***	0.0789**
	(0.0147)	(0.0147)	(0.0163)	(0.0159)	(0.0182)	(0.0185)	(0.0223)	(0.0266)
Spatial lag	0.540***	0.540***	0.0642***	0.539***	0.00903	0.569***	0.00430	0.474***
	(0.0352)	(0.0351)	(0.0159)	(0.0386)	(0.0103)	(0.0474)	(0.0127)	(0.0794)

Comparison of methods 2
Missing x's: coefficient estimates

	No missing data	10% missing data		25% missing data		50% missing data	
	XSMLE - FE	Pooled	MI	Pooled	MI	Pooled	MI
Real income	0.105***	0.384***	0.104***	0.383***	0.104***	0.382***	0.167***
	(0.0235)	(0.0173)	(0.0241)	(0.0202)	(0.0259)	(0.0311)	(0.0317)
Real prices	-0.248***	-0.235***	-0.240***	-0.256***	-0.222***	-0.280***	-0.141***
	(0.0120)	(0.0141)	(0.0129)	(0.0176)	(0.0151)	(0.0260)	(0.0220)
Housing per person	0.628***	0.983***	0.601***	1.066***	0.559***	1.067***	0.348***
	(0.0584)	(0.0634)	(0.0605)	(0.0754)	(0.0661)	(0.110)	(0.0846)
Cooling index	0.0499***	0.0745***	0.0494***	0.0665***	0.0487***	0.0533***	0.0507***
	(0.00593)	(0.00697)	(0.00611)	(0.00804)	(0.00638)	(0.0118)	(0.00681)
Heating index	0.127***	0.177***	0.122***	0.181***	0.121***	0.145***	0.123***
	(0.0147)	(0.0165)	(0.0150)	(0.0195)	(0.0156)	(0.0289)	(0.0170)
Spatial lag	0.540***	0.0326**	0.552***	0.00243	0.572***	-0.00724	0.585***
	(0.0352)	(0.0119)	(0.0359)	(0.0105)	(0.0366)	(0.0197)	(0.0411)

Comparison of methods 3
Missing y's - absolute bias as % of full se

	No missing data	10% missing data		25% missing data		50% missing data	
	Pooled	Pooled	MI	Pooled	MI	Pooled	MI
Real income	0%	1047%	9%	1149%	77%	1226%	179%
Real prices	0%	67%	42%	108%	192%	167%	175%
Housing per person	0%	661%	29%	745%	57%	640%	137%
Cooling index	0%	315%	103%	253%	255%	19%	396%
Heating index	0%	401%	61%	347%	238%	238%	333%
Spatial lag	0%	1352%	3%	1509%	82%	1523%	188%

Comparison of methods 4
Missing x’s - absolute bias as % of full se

	10% missing data		25% missing data		50% missing data	
	Pooled	MI	Pooled	MI	Pooled	MI
Real income	1187%	4%	1183%	4%	1179%	264%
Real prices	108%	67%	67%	217%	267%	892%
Housing per person	608%	46%	750%	118%	752%	479%
Cooling index	415%	8%	280%	20%	57%	13%
Heating index	340%	34%	367%	41%	122%	27%
Spatial lag	1443%	34%	1528%	91%	1534%	128%

Comparison of methods: lessons

- Be careful about use of either ML estimation or pooled cross section unless
  - The model specification is simple and convergence is reliable for ML
  - In cases of a changing population of panel units for which pooled cross section may be appropriate
- When using multiple imputation
  - Test several different methods of imputation
  - Use as many imputations as you can afford to run

Why spatial analysis matters: results for US electricity

- Clear evidence of spatial spillovers in electricity demand – especially for residential use
  - Coefficients on spatial lag in range 0.3-0.45
  - Allowing for spatial effects significantly reduces the coefficients on real income & housing
  - Higher electricity prices in one state associated with higher consumption in neighbouring states
- Policy: State renewable portfolio standards (RPS)
  - Potential price increases to 2020 up to 40%
  - How much effect on consumption and CO_2 emissions?