Lecture 3:
Spatial Analysis with Stata

Paul A. Raschky,

University of St Gallen, October 2017

Today's Lecture

1. Importing Spatial Data
2. Spatial Autocorrelation
   - 2.1 Spatial Weight Matrix
3. Spatial Models
   - 3.1 Identification
   - 3.2 Spatial Models in Stata
   - 3.3 Spatial Model Choice
4. Application
5. Mostly Pointless Spatial Econometrics?
6. Useful Stata commands
7. Zonal Statistics

Introduction

Why?

- Stata includes a number of commands that allows you to import, manipulate and analyze spatial data.
- Sometimes, stata performs better than other GIS software (ArcGIS). For example with large data.
- Spatial models in stata.

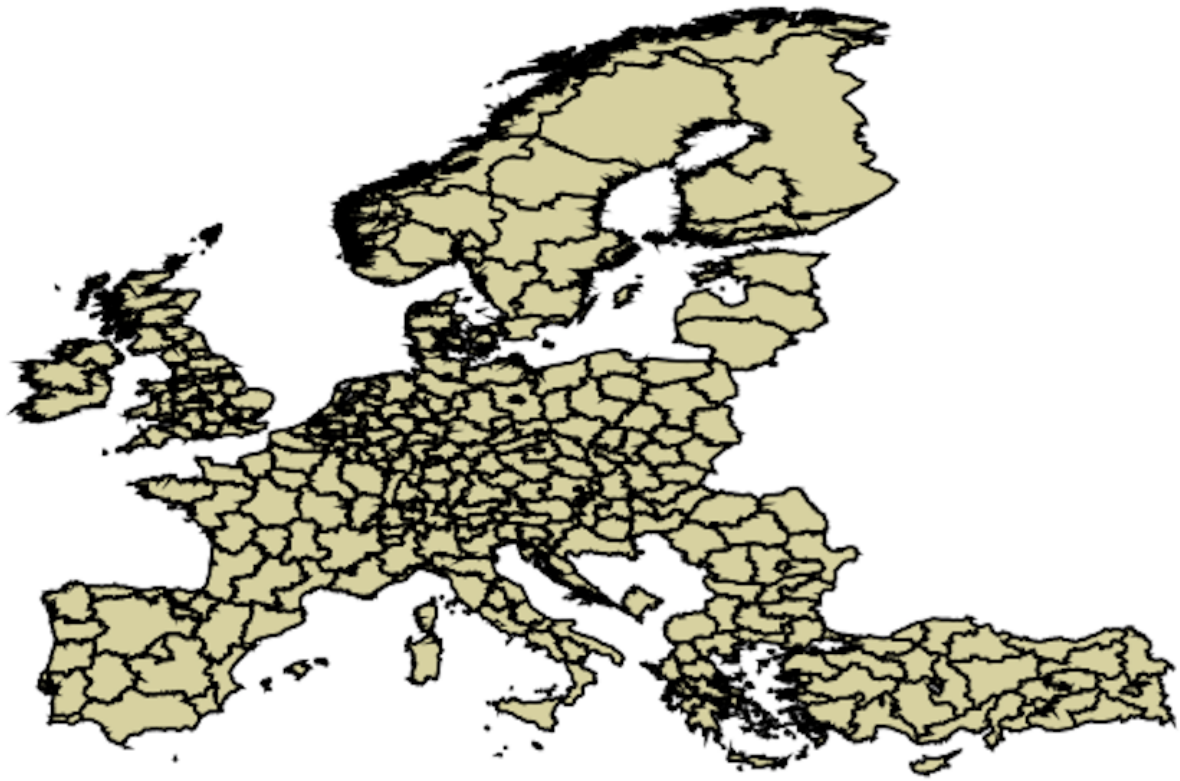
1. Importing spatial data - Vector

- Stata cannot directly load shapefiles (.shp)
- shp2dta imports shapefiles and converts them to .dta
- Syntax: shp2dta using shp. filename, database(lename) coordinates(lename) [options]
- Example:

```
shp2dta using EUR_NUTS2, database(eunuts2) coordinates(eunuts2xy) genid(id) gencentroids(c)
```

- eunuts2.dta: contains information from .dbf file, id, latitude (y) and longitude (x).
- eunuts2xy.dta: contains geometric information from .shp file.

1. Importing spatial data - Vector



1. Importing spatial data - Raster

- Stata can read raster in the ASCII format, with ras2dta ado (Muller 2005)
- Each cell becomes one row in Stata
- To export raster as the ASCII format, use Raster to ASCII

2. Spatial Autocorrelation

- Waldo Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970 p. 234)
- $\rightarrow$ **Spatial autocorrelation:** Two or more objects that are spatially close tend to be more similar to each other with respect to a given attribute Y than are spatially distant objects.
- $\rightarrow$ **Spatial custering:** Sub-areas of the study area where the attribute of interest Y takes higher than average values (hot spots) or lower than average values (cold spots)

2. Spatial Autocorrelation

General Ord (1975) (based on work by Whittle (1954)) model for spatial autoregressive process:

$$y_i = \rho \sum_{j=1}^{J} W_{ij} y_j + \epsilon_i \tag{1}$$

- $\sum_{j=1}^{J} W_{ij} y_j$: **Spatial lag** that represents a linear combination of values of y constructed from observations/regions that neighbor i.
- W_{ij} : $n \times n$ **Spatial weight matrix**
- ρ: Scalar of parameters that describe the strength of the **spatial dependence.**

2.1. Spatial Weight Matrix

Example: 7×7 spatial weight matrix W using the first- order contiguity relations for the seven regions

$$C = \begin{array}{c} \\ R1 \\ R2 \\ R3 \\ R3 \\ R5 \\ R6 \\ R7 \end{array} \begin{array}{c} \begin{array}{ccccccc} R1 & R2 & R3 & R4 & R5 & R6 & R7 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

2.1. Spatial Weight Matrix

W Row sum normalized matrix of C.

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

2.1. Spatial Weight Matrix

- Spatial weighting matrices paramterize the spatial relationship between different units.
- Often, the building of W is an ad-hoc procedure of the researcher. Common criteria are:

1. Geographical:
   - Distance functions: inverse, inverse with threshold
   - Contiguity
2. Socio-economic:
   - Similarity degree in economic dimensions, social networks, road networks.
3. Combinations between both criteria.

2.1. Spatial Weight Matrix

- Restricting the number of neighbors that affect any given place reduces dependence.
- Contiguity matrices only allow contiguous neighbors to affect each other.
  - This structure naturally yields spatial-weighting matrices with limited dependence.
- Inverse-distance matrices sometimes allow for all places to affect each other.
  - These matrices are normalized to limit dependence
  - Sometimes places outside a given radius are specified to have zero affect, which naturally limits dependence

2.1. Spatial Weight Matrix

- Geographic distance and contiguity are exogenous, but often used as *proxies* for the true mechanism.
- Row standardization allows us to interpret w_ij as the fraction of the overall spatial influence on country i from country j.
  - This is "practical" but can lead to misspecified models (Kelejian & Prucha 2010; Neumayer and Plümper 2015).
  - It imposes the restriction that if one unit has fewer ties to other units, then each tie is assumed to be more important.
- Scaling of the connectivity variable that enters into $\mathbf{W}$ does not necessarily match the relative relevance of j on i.
  - Taking the log?
- Alternative approach: Kelejian and Prucha (2010).

2.1. Spatial Weight Matrix

In Stata:

- Contiguity:
  - spmat contiguity CONT using eunuts2xy, id($_ID$) norm(row)
- Inverse Distance Matrix (200km cut-off):
  - spmat idistance IDISTVAL200 using eunuts2xy, id($_ID$) norm(row) vtruncate(1/200)

2.2 Global Spatial Statistics

Morans I is a correlation coefficient that measures the overall spatial autocorrelation of your data set.
In Stata:

- spatgsa lngdp, w(IDISTVAL200) moran geary two

3. Spatial Models

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \beta + \epsilon \tag{2}$$

- $\mathbf{y}$ is the $N \times 1$ vector of observations on the dependent variable
- X is the $N \times k$ matrix of observations on the independent variables
- W and is $N \times N$ spatial-weighting matrices that parameterize the distance between regions.
- ρ is a parameter scalars of spatial dependence.

3. Spatial Models

1. Spatial Error Model (SEM)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{3}$$

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \epsilon \tag{4}$$

- **Wu** reflects spatial dependence in the disturbance process.

3. Spatial Models

2. Spatial Lag Model (SLM)

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \mathbf{u} \tag{5}$$

- **Wy** reflects spatial dependence in y.

3. Spatial Models

3. Spatial Durbin Model (SDM)

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \gamma\mathbf{W}\mathbf{X} + \epsilon \tag{6}$$

- **WX** spatial lags of the explanatory variables

3.1. Spatial Models - Identification 1

$$y_i = \rho \sum_{j=1}^{J} W_{ij} y_j + X_i \beta + \gamma \sum_{j=1}^{J} W_{ij} X_j + \epsilon_i \tag{7}$$

- Problem: y_i and y_j are determined simultaneously.
- Applying a standard ordinary least squares (OLS) estimator would yield inconsistent estimates and biased coefficients ρ.

3.1. Spatial Models - Identification 1

"Potential solutions:"

- Maximum Likelihood (ML) Estimator as proposed by Anselin (1988)
- Generalized method of moments instrumental variables (GMM2SLS) approach (Kelejian and Prucha (1998; Kelejian and Robinson 1993).
  - Basic idea: Use of X_j, W X_j, or a combination of W X_j with higher order lags W^2X_j, W^3X_j as (internal) instruments for y_j

3.2. Spatial Models - In Stata

1. **SDM:** spreg ml lngdp inv pop w_inv w_pop, id(id) dlmat(IDISTVAL200)
2. **SLM:** spreg ml lngdp inv pop, id(id) dlmat(IDISTVAL200)
3. **SEM:** spreg ml lngdp inv pop, id(id) elmat(IDISTVAL200)

For GMM2sls

- Replace "ml" by "gs2sls"

3.3. Spatial Models - Choice

$$y_i = \rho \sum_{j=1}^{J} W_{ij} y_j + X_i \beta + \gamma \sum_{j=1}^{J} W_{ij} X_j + \epsilon_i \tag{8}$$

Elmhorst (2010)

1. Estimate equation (11) by using OLS and setting $\rho = 0$ $\gamma = 0$.
2. Lagrange multiplier (LM) tests if the spatial lag model or spatial error model is more appropriate (Anselin 1988).
3. If the nonspatial model is rejected in favor of the spatial autoregressive model or spatial error model or both, the spatial Durbin model should be taken.

4. Application - Borsky & Raschky (2014)

- International environmental agreement (IEA)are a popular tool to overcome the free-rider problem associated with global environmental problems.
- However, the voluntary character of many IEAs prevents neither free-riding from nonsignatory countries.
- Interestingly, numerous IEAs are ratified each year, although their benefits are nonrival and nonexcludable.
- Possible explanation: Intergovernmental interaction among the signatory countries.

4. Application - Borsky & Raschky (2014)

- Cross-sectional data sets to measure signatory countries compliance with the obligations of Article 7 of the 1995 UN Code of Conduct for Responsible Fisheries (CoC)
- Data from 53 signatory countries:
  - Compliance effort, marine protected areas, marine biodiversity.
  - GDP, Costs, Inst. Quality, Competition, Env. Performance, Fish Export.

4. Application - Borsky & Raschky (2014)

Spatial Weighting matrix 1 - Inverse Geographic Distance

- The degree of (non)compliance is more easily observed for countries that are geographically closer to each other.
- Intergovernmental interdependence is defined by the degree of interaction due to economic transactions and political relations, which decrease in distance.
- We assume each country to be affected by all other countries in our sample, where closer countries are stronger weighted $\rightarrow$ no distance cut-off
- Row normalized.

4. Application - Borsky & Raschky (2014)

Spatial Weighting matrix 2 - Political Distance

- Based on Rose and Spiegel (2009)
- Mutual involvement in 443 IEAs of each country pair as a distance measure.

4. Application - Borsky & Raschky (2014)

Empirical Strategy:

$$c_i = \rho \sum_{j=1}^{J} \omega_{ij} c_j + X_i \beta + \gamma \sum_{j=1}^{J} \omega_{ij} X_j + \varepsilon_i$$

- Following Elmhorst (2010) and Anselin (1988) we use a SDM.
- Estimate the model using ML (and GMM2SLS in robustness exercise).

4. Application - Borsky & Raschky (2014)

Main Results:

	Overall		Behavior		Intention	
	X	WX	X	WX	X	WX
	(1)	(2)	(3)	(4)	(5)	(6)
ρ	.375***		.350***		.253**	
	(.107)		(.114)		(.115)	

- Following Elmhorst (2010) and Anselin (1988) we use a SDM.
- Estimate the model using ML (and GMM2SLS in robustness exercise).

4. Application - Borsky & Raschky (2014)

Political Distance:

Political Distance	
X	WX
(7)	(8)
.386***	
(.106)	

GMM2SLS:

	2SLS
	X
	(1)
ρ	.531***
	(.146)

4. Application - Borsky & Raschky (2014)

Results:

- Intergovernmental interaction has a systematic impact on a countrys level of compliance.
- Other countries compliance behavior acts as a strategic complement, and the effect decreases in distance.
- Indirect spatial spillover effects (w LeSage and Pace 2009):
  - Variables that improve the allocation of property rights in one country seem to generate spatial spillover effects on compliance effort.

5. Mostly Pointless Spatial Econometrics?

Gibbons & Overman (2012) (in a nutshell)

- Estimates of spatial lag, ρ, are endogenous.
- Both ML and GMM2SLS do not adequately solve this problem.
- Proposed Solutions:
  - Natural experiments: Redding and Sturm (2008): Reunification changes **W**.
  - Using exogneous IVs: Luechinger (2009): Pollution and wind directions
  - Discontinuities: Greenstone et al. (2010): Plant relocations and comparing first with second ranked preferences.

6. Other useful commands

Spatial panel models

- tsset data first
- xsmle lngsp pop, wmat(IDISTVAL200) model(sar)

6. Other useful commands

Spatial errors in stata

- Fetzer (2015): reg2hdfespatial
- Hsiang (2010): ols_spatial_HAC

6. Other useful commands

Distance between points in stata

- Vincenty
- Calculating geodesic distances between a pair of points on the surface of the Earth.
- Basic Syntax: vincenty lat1 lon1 lat2 lon2 inkm

6. Other useful commands

Spatial Join in stata

- geoinpoly
  - geoinpoly Y X using eunuts2xy.dta
  - merge m:1 _ID using eunuts2.dta

7. Zonal Statistics

- Zonal Statistics combines information from vector and raster data to calculate summary statistics of raster data value within each polygon
  - Mean / Standard deviation
  - Min / Max / Range
  - Sum
  - Count

7. Zonal Statistics

Polygon examples:

- Countries
- Subnational regions
- Ethnic homelands
- FAO zones
- Grid cells (this lecture)

Raster data examples:

- Population
- Temperature
- Agricultural suitability
- Elevation
- Land use
- Nighttime light

7. Zonal Statistics

Application in Economics

- Nighttime lights at the
  - Country (Henderson et al. 2012)
  - ADM1/ADM2 Level (Hodler & Raschky 2014)
  - City boundaries (Soteygard 2016)
  - Grid Level (Henderson et al. 2017)
  - Ethnic homelands (Michalopoulos & Papaioannou 2013 / 2014, Alesina et al. 2016)

7. Zonal Statistics

Application in Economics

- Temperature/Drought
  - ADM1/ADM2 Level (Hodler & Raschky 2014)
  - Grid Level (Dell et al. 2017)
  - Ethnic homelands (Michalopoulos & Papaioannou 2013 / 2014, Alesina et al. 2016)
- Crop Cover
  - Grid Level (Harari & La Ferrera 2017)
- Elevation & Slope
  - Duflo & Pande (2007)

7. Zonal Statistics

Exercise 3 - Zonal Statistics