Spatial analysis and spatial dependence

Kristian Skrede Gleditsch

Department of Government

University of Essex

http://privatewww.essex.ac.uk/~ksg/

Overview

Spatial dependence and spatial statistics

- Some examples of spatial dependence
- Visualization of spatial data
- Specifying dependence structures
- Testing for spatial dependence
- Spatial autoregressive linear model
- Pointers to advanced topics

Examples of spatial dependence

Tale of two neighbors, Promethia and Tragedia

- Promethia does everything right, but below average growth rate
- Neighbor Tragedia has bad policies and disastrous economic performance
- Promethia's low growth rate reflects spill-over effects from Tragedia
- Observations for Promethia and Tragedia clearly not independent
- Easterly & Levine find that Africa term in cross country growth models becomes non-significant when controlling for contagion

Examples of spatial dependence

Civil wars not determined by attributes of states alone, events in other states can influence risk

- Central Africa: Uganda, Rwanda, Zaire, Kashmir conflict

Local character of environmental problems

- Murdoch and Sandler: SO_2 emission reductions in Europe easier to achieve since states more private benefits from reducing omissions
- Less local problems (e.g., NOX and $CO2$) more difficult

Visualization

Visualization may suggest important structure in data

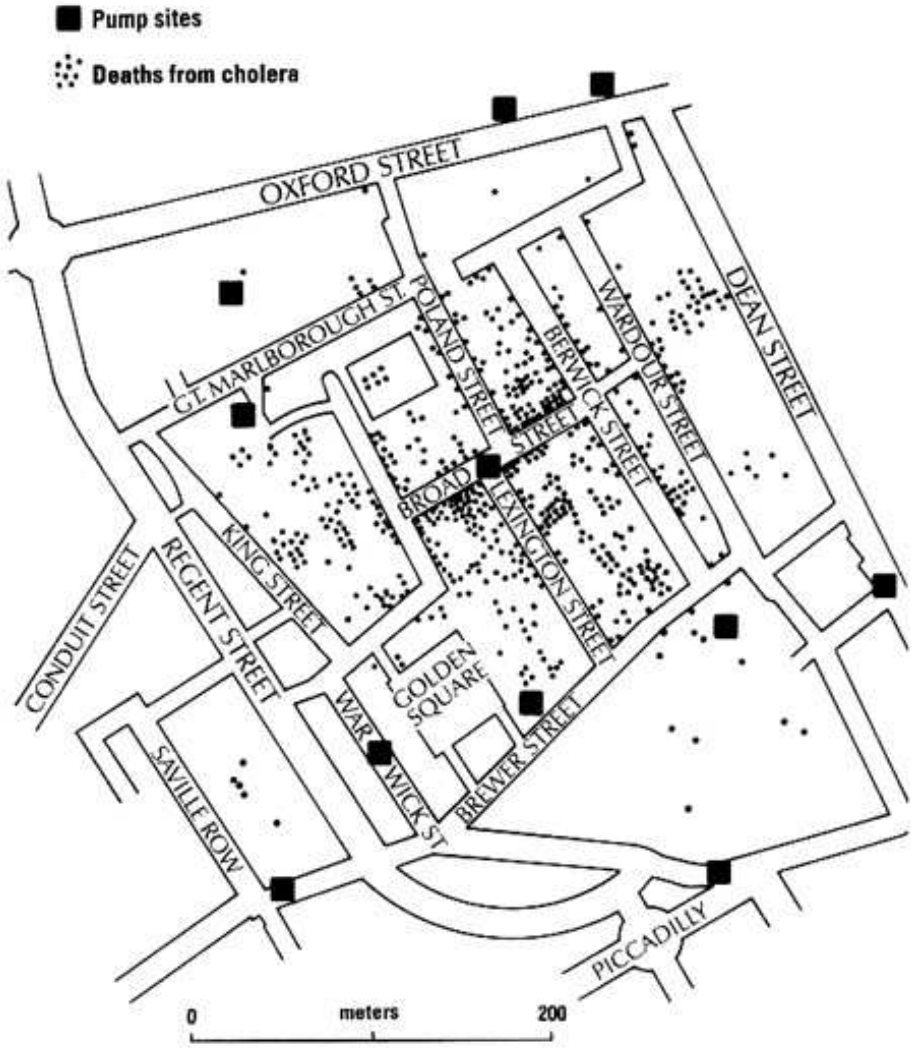
- Tables large, often unwieldy
- Maps can summarize information on one page
- Where are the high and low values located?
- Spatial structure/pattern?

The Snow map

Snow's map of cholera deaths in London

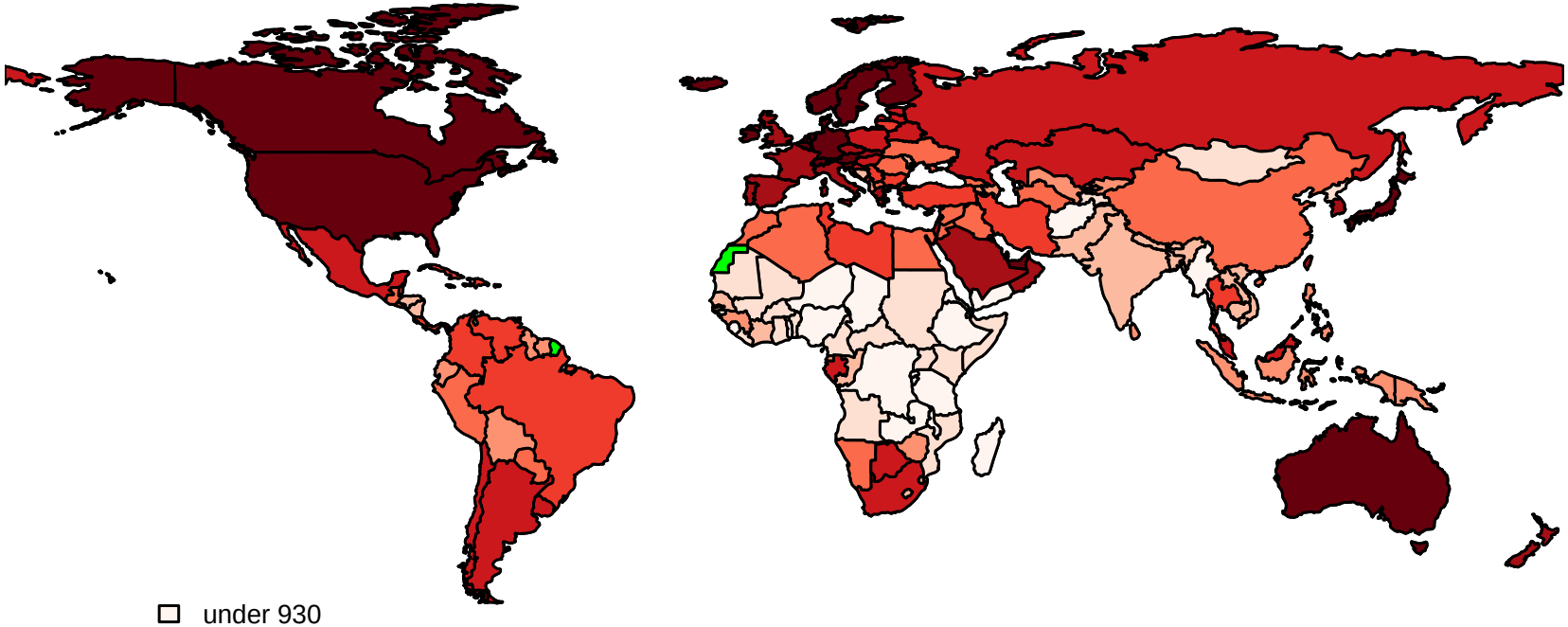
- September 1854 cholera epidemic
- 500 from same section of London died within a ten-day period
- Bacteria not yet know, sources of disease obscure
- Snow had previously published paper on water as possible source of spread of cholera
- Created a map of location of deaths and water pump

The Snow map



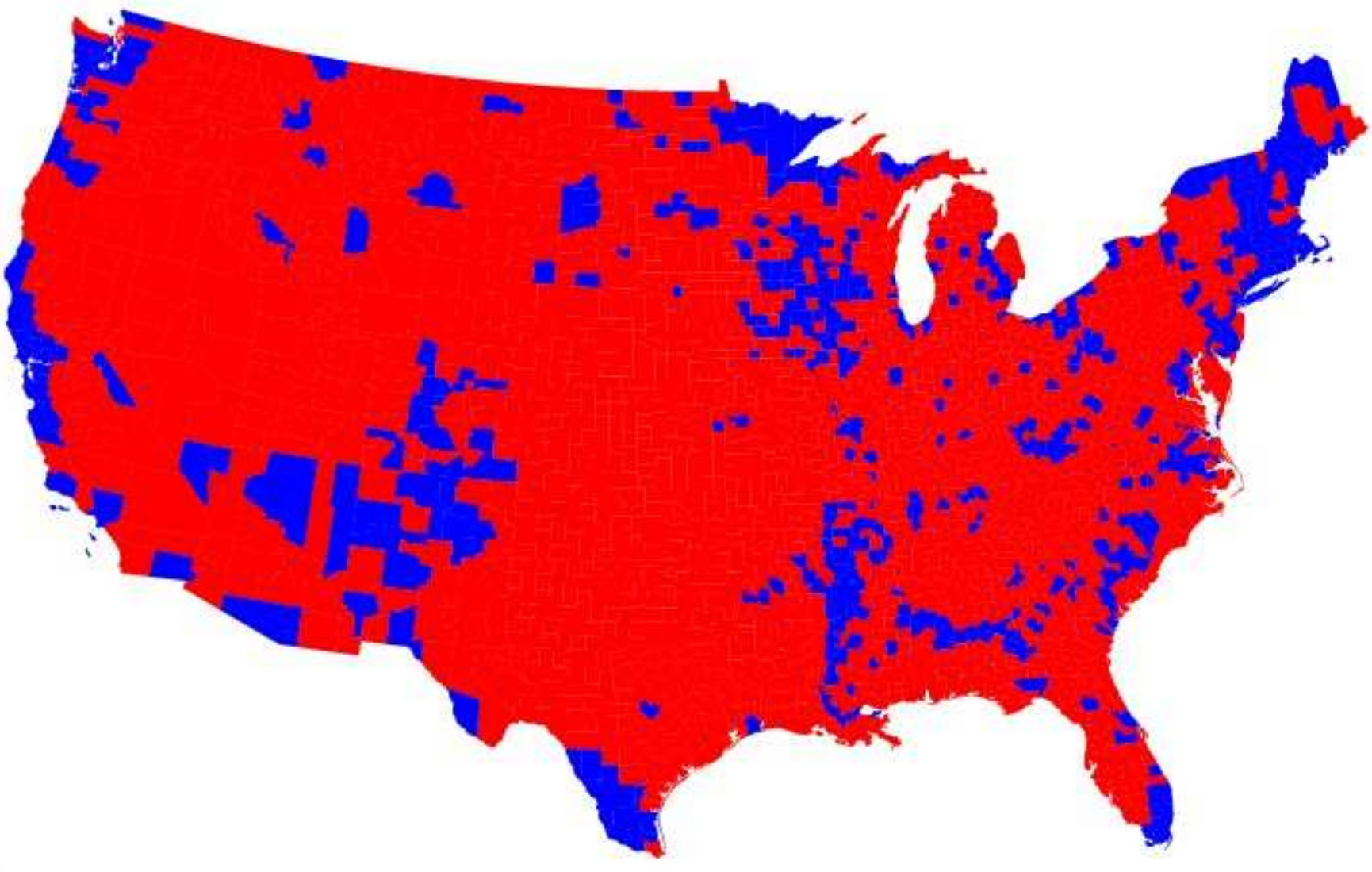
University of Essex

Global income distribution



University of Essex

Red (Bush) and blue (Kerry) counties



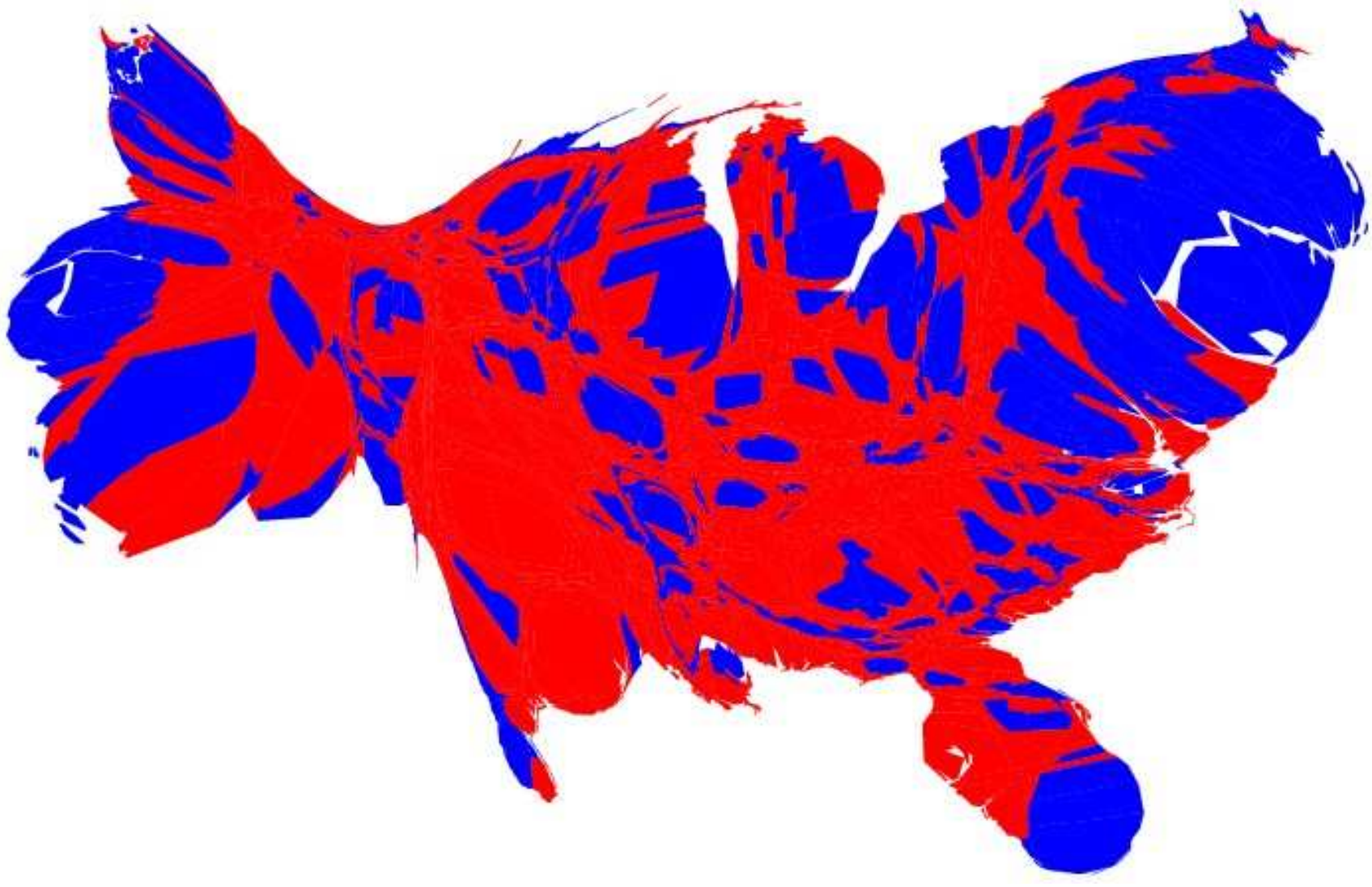
University of Essex

Political maps and cartograms

Displaying geographical shapes/units alone can be misleading

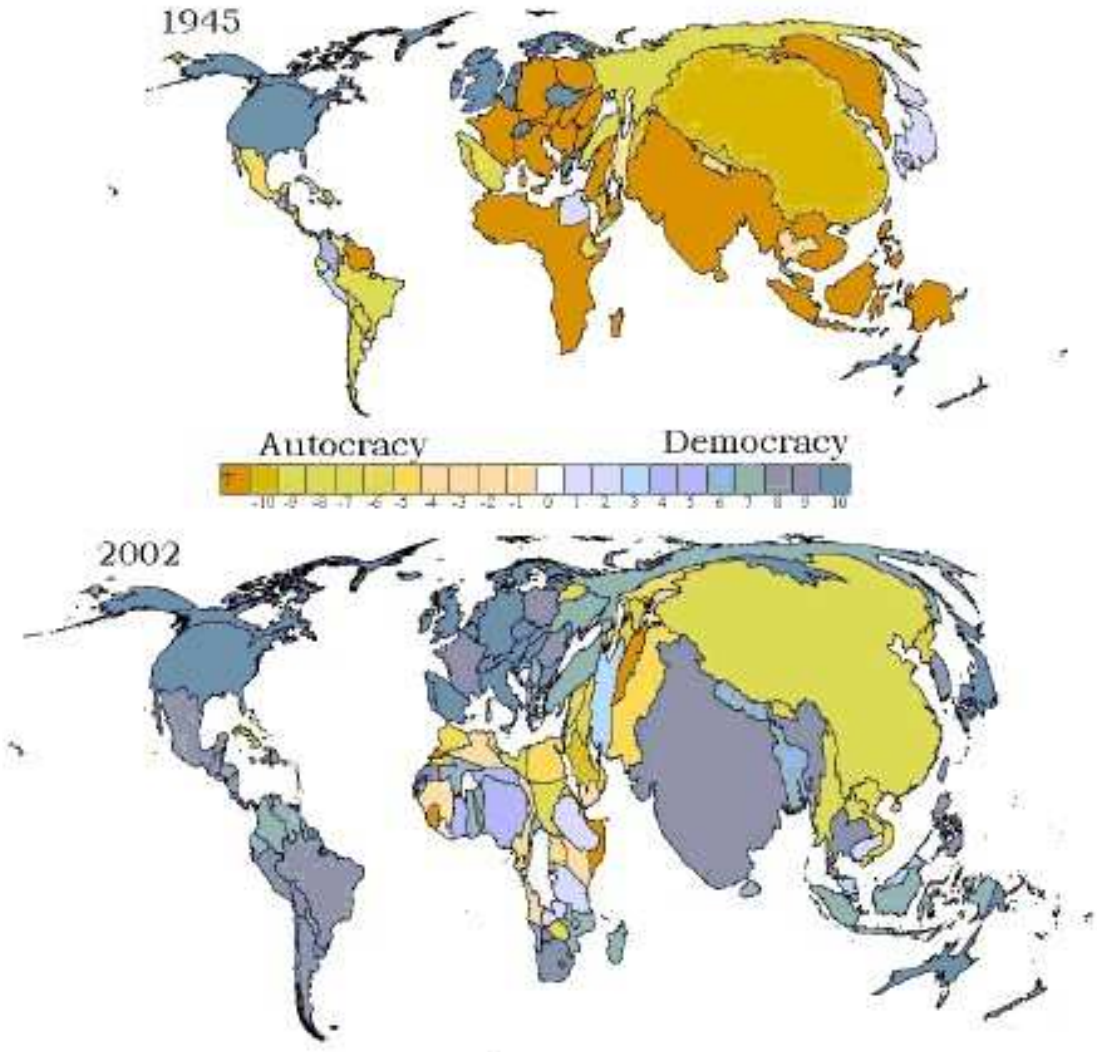
- Counties have very different population size
- More rural than urban counties, associated with differences in voting patters
- Map of blue and red populations should weight by population
- Cartogram scaling each county by population size

Red and blue counties revisited

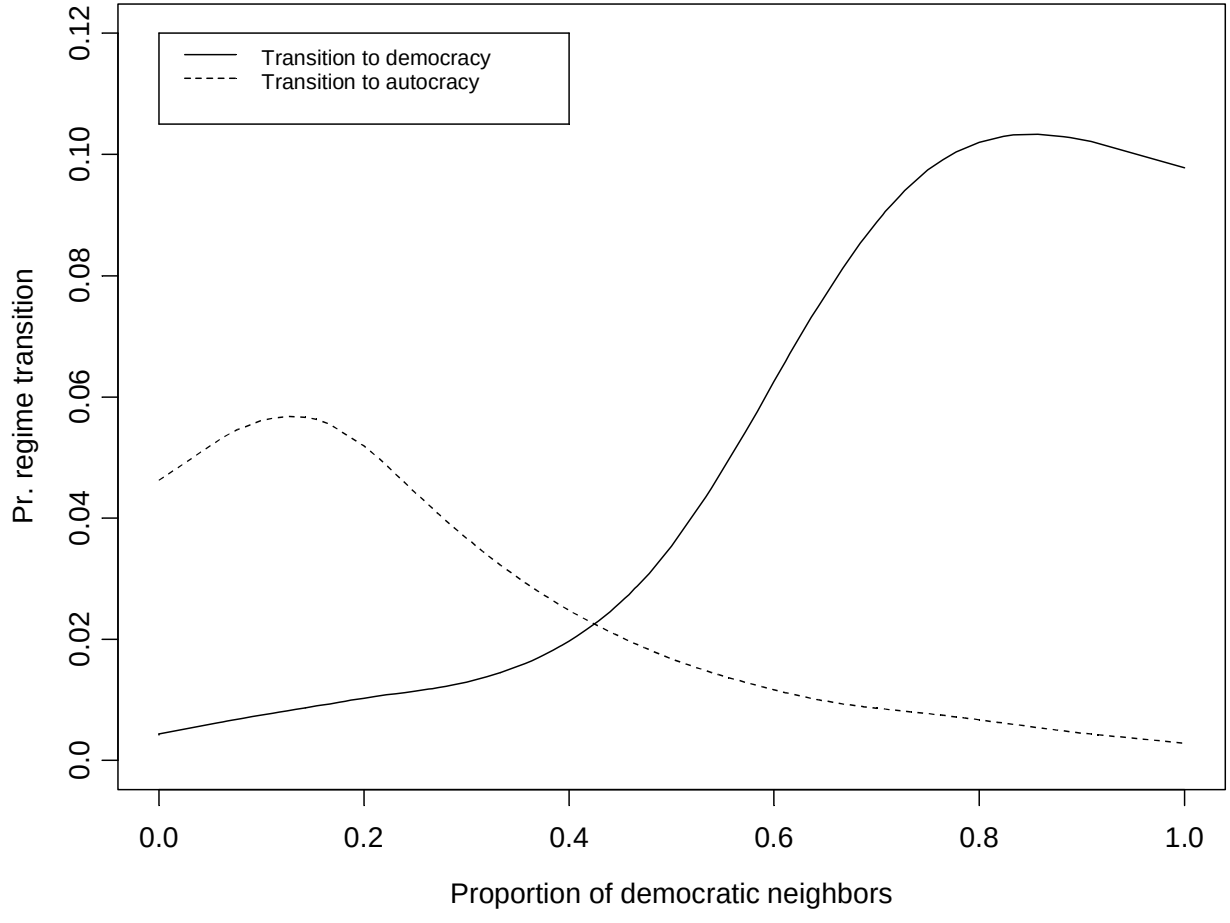


University of Essex

Democratization in the World



Democratization in the World



University of Essex

Mapping data

- Most software proprietary and expensive (*ESRI ArcView, ArcInfo*)
- However, possible to use *ESRI* shapefiles to generate maps in the free package *R*
- `http://www.r-project.org`
- Shapefiles for many areas freely available on web
- For utilities to create diffusion based cartograms, see `http://www-personal.umich.edu/~mgastner/`

Spatial dependence

- Values y_i and y_j often similar for neighbors i and j
- Many diagnostic tests for spatial dependence available
- Global dependence: Moran's I
- Tests for spatial dependence presumes a pre-specified form of dependence
- Hence, tests only able to reject/not reject particular specification

University of Essex

Connectivity

Pattern of connectivities between observations

- Connectivity matrix $\mathbf{W}$
  - $N \times N$ matrix where entries $w_{i,j}$ acquire non-zero values if units i and j connected
  - Alternatively, inverse distance weights based on distance d_{ij} where $w_{ij} = \frac{1/d_{ij}}{\sum_{j=1}^{n} 1/d_{ij}}$
- Criteria for connectivity?
- Time has inherent ordering, but $N \times (N-1)$ possible relationships in space for N units

Connectivity

Many possible forms of distance/connectivities

- Geography-distance link often suggests geographical dependence (e.g., contiguous states, proximate states)
- Observed interaction patterns (e.g., trade, communications)
- Cultural distance (e.g., links between languages)
- Reachability indices vs. geographical distance (e.g., time to travel between points)

Must be justified in each application

Moran's I

How similar is i to its neighbors J?

$$I = \frac{N \sum_i \sum_j c_{ij} (y_i - \bar{y}) (y_j - \bar{y})}{\left(\sum_i \sum_j c_{ij} \right) \sum_i (y_i - \bar{y})^2} \quad (1)$$

where c_{ij} is an element in a binary connectivity matrix $\mathbf{C}$

$Z = I/SE(I)$ indicates whether clustering is "significant"

E.g., Moran's $I = 0.47$,$Z > 10$ for spatial clustering in democracy, reject $H0$ of no clustering

University of Essex

Spatial dependence

Usually assume that dv $Y = f(X)$

- Looking at unconditional spatial dependence problematic since the ivs X also likely to be non-randomly distributed
- Test whether residuals from a conditional model displays residual spatial clustering
- Ex: residuals from regression of expected democracy conditional on development still shows spatial dependence
- Moran's $I = 0.40, Z > 8$ for spatial clustering in residuals from regression, reject $H0$ of no clustering, even after controlling for GDP per capita

Spatial dependence

Why is spatial dependence a problem in a regression model?

- Regression assumptions imply that errors of individual observations should be independent and unrelated: $E[\epsilon_i, \epsilon_j] = 0$ for $i \neq j$.
- Spatial dependence violates this assumption, since $Cov(y_i, y_j)$ tends to be positive
- Standard properties do not hold
- More fundamentally, a model assuming that observations are independent is fundamentally incorrect

Spatial regression models

Introduce spatial structure into regression model

Spatial autoregressive (SAR) model: add a right hand side "spatial lag" $\mathbf{W}Y$, where $\mathbf{W}$ is row-normalized

$$E(y_i|x_i, Y) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\rho} w_i.Y, \tag{2}$$

Estimation more difficult since y now on both sides of

equation, but can be estimated by MLE or IV

Spatial regression models

	(1)	(2)	(3)
Variable	OLS	Spatial autoregressive estimates	
Constant	-19.71 (3.66)	-12.96 (3.35)	-19.39 (3.39)
Ln(GDPPC)	2.66 (0.42)	1.72 (0.41)	2.22 (0.41)
ρ_1 (distance)		0.48 (0.09)	
ρ_2 (trade)			0.51 (0.14)
N	170	170	170

SAR results

Interpretation of SAR

- ρ estimate of spatial dependence, influence of neighbors
- Model implies feedback:
  - Change in IV in one state will first affects its DV
  - Then affect other states
  - Feed back to country itself
- Hence, coefficients for ivs $\mathbf{X}$ not fully comparable to OLS, must consider implied “equilibrium effect”

SAR results

$$
\begin{aligned}
Y &= \rho \mathbf{W} Y + \mathbf{X}\beta + \epsilon \\
\epsilon &= (I - \rho \mathbf{W}) Y - \mathbf{X}\beta \\
Y &= (I - \rho \mathbf{W})^{-1} \mathbf{X}\beta + (I - \rho \mathbf{W})^{-1} \epsilon \\
E(Y) &= (I - \rho \mathbf{W})^{-1} \mathbf{X}\beta
\end{aligned}
$$

I.e., the “equilibrium effect” of a change in IV will depend upon connectivities with other states, and vary from country to country

SAR results

SAR with geographical distance:

- effect of a unit increase in log of GDP per capita ranges from a 1.75 to a 2.04 point increase in level of democracy, average effect 1.80

SAR with trade connectivities:

- a one unit change in ln GDP per capita on average leads to a 2.24 point increase in democracy, effects for individual countries ranging from 2.22 to 2.52

SAR with two matrices

$$y_i = \mathbf{x}_i\beta + \rho_1 \mathbf{w}_i^A \mathbf{y} + \rho_2 \mathbf{w}_i^B \mathbf{y} + \varepsilon_i$$

SAR with two matrices

Table 1: Democracy and Social Requisites, 1998

	(1)	(2)	(3)	(4)
Variable	OLS	Spatial autoregressive estimates		
Constant	-19.71 (3.66)	-12.96 (3.35)	-19.39 (3.39)	-13.24 (3.11)
Ln(GDPPC)	2.66 (0.42)	1.72 (0.41)	2.22 (0.41)	1.53 (0.37)
ρ_1 (distance)		0.48 (0.09)		0.89 (0.19)
ρ_2 (trade)			0.51 (0.14)	0.59 (0.43)
N	170	170	170	170

Binary dependent variables

- SAR less appropriate for categorical dv's such as conflict, autologistic possible alternative
- Assume a locally dependent Markov field
  - $Pr(y_i \mid y_j, j \neq i)$ depends only on $y_j \iff j$ is a neighbor of i or $w_{i,j} = 1$

In the autologistic, condition likelihood of $y_i = 1$ on conflict among neighbors $\widetilde{y_i}$

$$Pr[y_i = 1|\widetilde{y_i}] = \frac{e^{\alpha + \mathbf{X}_i' \beta_k + \gamma \widetilde{y_i}}}{1 + e^{\alpha + \mathbf{X}_i' \beta_k + \gamma \widetilde{y_i}}}$$

- If $\gamma = 0$ this is a standard logistic, observations not conditional on one another

University of Essex

Example of autologistic

Ward and Gleditsch (2002), N=138

			Spatial Parameters	
	Intercept	Democracy	Democracy	Conflict
Estimator	α	β_1	β_2	γ
Logistic	*-1.309*	-0.022	-0.015	–
MPL $\hat{\psi}$	*-1.840*	-0.020	0.013	*0.298*
MCMC ML $\hat{\theta}$	*-1.712*	*-0.053*	-0.003	*0.261*
Logistic SEs	0.218	0.033	0.048	–
MPL SEs	0.333	0.033	0.051	0.126
MCMC SEs	0.060	0.006	0.010	0.013

Example of autologistic

Prediction based on 1988 estimates for 1989 to 1998 period, $\hat{\pi}$ =0.35 threshold

	Observed	
Predicted	No	Yes
No	60	17
Yes	29	33