Inference with Arbitrary Clustering

Fabrizio Colella,[*] Rafael Lalive,[*]
Seyhun O. Sakalli,[*] Mathias Thoenig[*]

Swiss Stata Users Group Meeting, October 2018

[*]University of Lausanne

Introduction

Motivation

A tremendous surge of empirical analysis with spatial data:

- Growing availability of geocoded data
- Integration of geographic information systems (GIS) in the toolkit of economists

Network relations among individuals known and easily accessible

Need for econometric methods to obtain asymptotically valid inference in settings with varying types of spatial, network, and temporal dependence between observation units

Absence of Stata commands, especially in the 2SLS setting

This paper

Proposes an approach to obtain asymptotically valid inference in the presence of arbitrary correlation (spatial or within a network) in both OLS and 2SLS settings

Provides a package, **acreg**, for the statistical software Stata

Performs Monte Carlo simulations (using spatial data on U.S. towns and counties) to show the properties and performance of the proposed estimator

- Generate random variables and check how close we get to 5% null-rejection rate at 5% test level, following Bertrand, Duflo, and Mullainathan (2004)

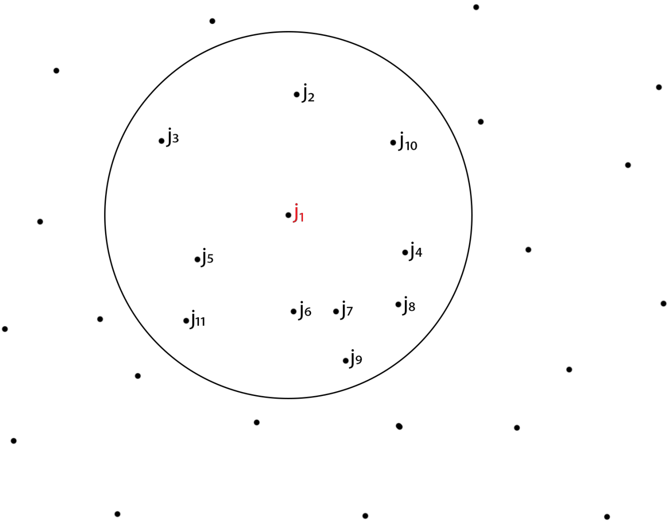
Stata command: *acreg*

What is new in *acreg* compared to existing packages?

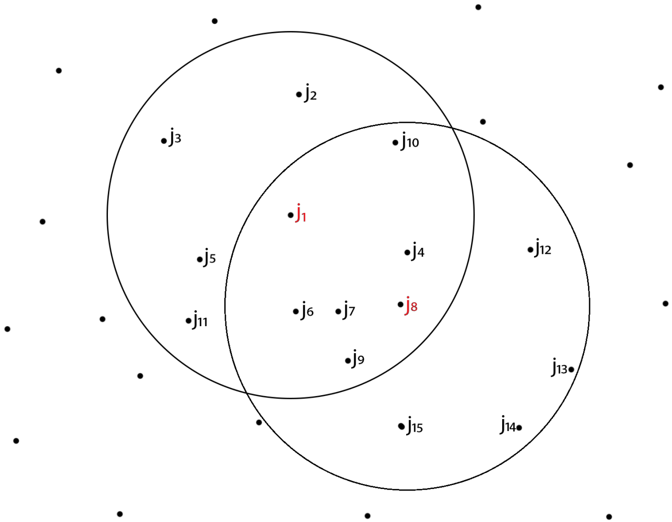
- Performs standard error correction in both OLS and 2SLS settings following White (1980)
- Correlation weights can be given as input or computed from spatial or network relations or multi-way clustering (Cameron et al., 2011)
- Spatial relations can be defined both with a distance cutoff and a contiguity/distance matrix (neighboring observations only)
- Network relations can be defined both with a matrix of links or a distance matrix or with any arbitrary cluster structure that user defines
- Allows for observation i in time t to be correlated with observation j in its cluster in time $t + s$
- HAC standard errors and distance decays are optional
- Fixes some bugs that exist in Conley (1999) and Hsiang (2010)

Arbitrary Clustering

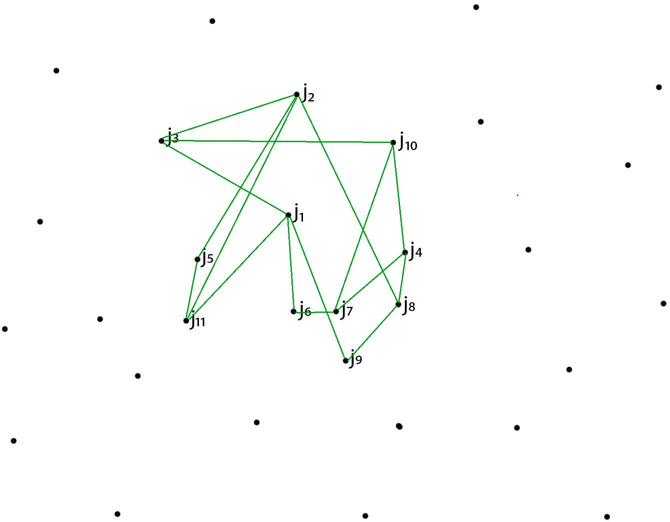
Spatial - 1 Cluster



Spatial - 2 Overlapping clusters



Network



Network - Adjacency matrix

	j_1	j_2	j_3	j_4	j_5	j_6	j_7	j_8	j_9	j_{10}	j_{11}
j_1	**1**	0	**1**	0	0	**1**	**1**	0	0	0	**1**
j_2	0	**1**	**1**	0	**1**	0	0	**1**	0	0	**1**
j_3	**1**	**1**	**1**	0	0	0	0	0	0	**1**	0
j_4	0	0	0	**1**	0	0	**1**	**1**	0	**1**	0
j_5	0	**1**	0	0	**1**	0	0	0	0	0	**1**
j_6	**1**	0	0	0	0	**1**	**1**	0	0	0	0
j_7	0	0	0	**1**	0	**1**	**1**	0	0	1	0
j_8	0	**1**	0	**1**	0	0	0	**1**	**1**	0	0
j_9	**1**	0	0	0	0	0	0	**1**	**1**	0	0
j_{10}	0	0	**1**	**1**	0	0	**1**	0	0	**1**	0
j_{11}	**1**	**1**	0	0	**1**	0	0	0	0	0	**1**

Conceptual Framework

Theoretical VCV of the 2SLS estimator

Standard IV Estimator

$$b_{2SLS} = (\hat{X}'\hat{X})^{-1}(\hat{X}'y)$$

With Variance

$$VCV(b_{2SLS}) = (\hat{X}'\hat{X})^{-1}\hat{X}'\Omega\hat{X}(\hat{X}'\hat{X})^{-1}$$

Where:
y is the Dependent Variable
X is the Matrix of Regressors (exogenous and endogenous)
Z is the Matrix of Instruments (excluded and included)
$\hat{X} = Z(Z'Z)^{-1}(Z'X)$ is the fitted values from the First Stage Regression
Ω is the VCV of errors

Estimating the VCV of the 2SLS estimator

Proposed Estimator for $\hat{X}'\Omega\hat{X}$ is:

$$\hat{X}'(S. \times (uu'))\hat{X} = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T}\hat{x}_{it}u_{it}u_{js}\hat{x}_{js}\mathbf{s_{itjs}}$$

Where:
$u \equiv y - \hat{X}\hat{\beta}_{2SLS}$ are the estimated residuals

- Each *itjs*-th component of **s** is a *correlation weight* [0,1]
- The *correlation weight* can be arbitrarily set
- The *correlation weight* should reflect the dependence of the error of observation *it* on the error of observation *js*

Asymptotics of the proposed estimator (work in progress)

Equivalence with multi-way clustering

- Any bilateral links structure can be represented by a multi-way clustering structure.
- $V\hat{C}V(\hat{\beta}_{2SLS})$ in a multi-way cluster environment can be represented as sum of one-way cluster-robust matrices (Cameron et al. 2011)
- The *sandwich estimator* of the $V\hat{C}V(\hat{\beta}_{2SLS})$ in a one-way cluster environment is consistent as $G \to \infty$ (White 1984; Arellano 1987; Rogers 1993; Hansen 2007)

Dimensionality with arbitrary clustering (work in progress)

Command

acreg - Syntax: baseline

```
acreg depvar [varlist1 ] [(varlist2 =varlist_iv )] [if] [in] [pweight ]
      [, id(idvar) time(timevar ) spatial network
      latitude(latitudevar ) longitude(longitudevar )
      links_mat(varlist_links ) dist_mat(varlist_distances )
      dist(distcutoff ) lag(TIMEcutoff ) lagdist(distTIMEcutoff )
      storeweights storedistances weights(varlist_weights )
      cluster(varlist_cluster ) hac correctr2 nuclust(n_clusters )
      pfe1(fe1var) pfe2(fe2var)]
```

depvar is the dependent variable.

varlist1 is the list of exogenous variables.

varlist2 is the list of endogenous variables.

varlist_iv is the list of exogenous variables used with *varlist1* as instruments for *varlist2* .

acreg - Syntax: Spatial 1

```
acreg depvar [varlist1] [(varlist2 = varlist_iv)] [if] [in] [pweight]
      [, id(idvar) time(timevar) spatial network
      latitude(latitudevar) longitude(longitudevar)
      links_mat(varlist_links) dist_mat(varlist_distances)
      dist(distcutoff) lag(TIMEcutoff) lagdist(distTIMEcutoff)
      storeweights storedistances weights(varlist_weights)
      cluster(varlist_cluster) hac correctr2 nuclust(n_clusters)
      pfe1(fe1var) pfe2(fe2var)]
```

spatial specifies that the environment is a spatial environment.

- **latitudevar** is the variable containing the latitude of each observation, decimal degrees: [-180,180].
- **longitudevar** is the variable containing the longitude of each observation, decimal degrees: [-180,180].
- **distcutoff** specifies the distance cutoff beyond which the correlation between error term of two observations is assumed to be zero.

acreg - Syntax: Spatial 2

acreg *depvar* [*varlist1*] [(*varlist2* =*varlist_iv*)] [*if*] [*in*] [*pweight*]
[, **id**(*idvar*) **time**(*timevar*) **spatial** **network**
latitude(*latitudevar*) **longitude**(*longitudevar*)
links_mat(*varlist_links*) **dist_mat**(*varlist_distances*)
dist(*distcutoff*) **lag**(*TIMEcutoff*) **lagdist**(*distTIMEcutoff*)
storeweights **storedistances** **weights**(*varlist_weights*)
cluster(*varlist_cluster*) **hac** **correctr2** **nuclust**(*n_clusters*)
pfe1(*fe1var*) **pfe2**(*fe2var*)]

spatial specifies that the environment is a spatial environment.

- **varlist_distances** is the list of N variables containing bilateral distances between observations. In the spatial environment, bilateral distance is the spatial distance between observations, i.e., physical distance between two locations.

- **distcutoff** specifies the distance cutoff beyond which the correlation between error term of two observations is assumed to be zero.

acreg - Syntax: Network 1

```
acreg depvar [varlist1] [(varlist2 =varlist_iv)] [if] [in] [pweight]
      [, id(idvar) time(timevar) spatial network
      latitude(latitudevar) longitude(longitudevar)
      links_mat(varlist_links) dist_mat(varlist_distances)
      dist(distcutoff) lag(TIMEcutoff) lagdist(distTIMEcutoff)
      storeweights storedistances weights(varlist_weights)
      cluster(varlist_cluster) hac correctr2 nuclust(n_clusters)
      pfe1(fe1var) pfe2(fe2var)]
```

network specifies that the environment is a network environment.

- **varlist_links** is the list of N dummy variables Specifying the links between observations, i.e., the adjacency matrix. If distcutoff>1 only the first observation in time of each individual will be used as input.

- **distcutoff** specifies the distance cutoff beyond which the correlation between error term of two observations is assumed to be zero.

acreg - Syntax: Network 2

```
acreg depvar [varlist1 ] [(varlist2 = varlist_iv )] [if] [in] [pweight ]
      [, id(idvar) time(timevar) spatial network
      latitude(latitudevar ) longitude(longitudevar )
      links_mat(varlist_links ) dist_mat(varlist_distances )
      dist(distcutoff ) lag(TIMEcutoff ) lagdist(distTIMEcutoff )
      storeweights storedistances weights(varlist_weights )
      cluster(varlist_cluster ) hac correctr2 nuclust(n_clusters )
      pfe1(fe1var) pfe2(fe2var)]
```

network specifies that the environment is a network environment.

- **varlist_distances** is the list of N variables containing bilateral distances between observations.

  In the network environment, it is the network distance between observations,i.e., the number of links along the shortest path between two nodes.

- **distcutoff** specifies the distance cutoff beyond which the correlation between error term of two observations is assumed to be zero.

acreg - Syntax: Multiway clustering

```
acreg depvar [varlist1 ] [(varlist2 =varlist_iv )] [if] [in] [pweight ]
      [, id(idvar) time(timevar) spatial network
      latitude(latitudevar ) longitude(longitudevar )
      links_mat(varlist_links ) dist_mat(varlist_distances )
      dist(distcutoff ) lag(TIMEcutoff ) lagdist(distTIMEcutoff )
      storeweights storedistances  weights(varlist_weights )
      cluster(varlist_cluster ) hac  correctr2 nuclust(n_clusters )
      pfe1(fe1var) pfe2(fe2var)]
```

- **varlist_cluster** is the list of variables to use for multi-way clustered SEs.

acreg - Additional Options

- Panel Dimension and optional HAC standard errors
- Allows for sampling weights (*pweights*)
- Allows for ‘if’ and ‘in’ statements
- Allows for partialling out up to 2 high-order fixed effects
- Produces output similar to Stata’s native commands
- Allows for storing distance matrix and weights matrix
- Stores main results in e()

acreg - Output: Spatial

```
. acreg hrate ln_population age (ln_income=unemployment) ,    id(_ID) time(year)
> latitude(_CX) longitude(_CY)   dist(50) lag(50)   spatial

SPATIAL CORRECTION
DistCutoff: 50
LagCutoff:  50
LagDistCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  46.357
                                                       Number of obs =      5648
Total (centered) SS      =  286387.1082                Centered R2   =   -0.0447
Total (uncentered) SS    =  781008.6785                Uncentered R2 =    0.6169
Residual SS              =  299188.6495
```

hrate	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ln_income	3.83872	1.487876	2.58	0.010	.9225371	6.754904
ln_populat~n	-.4411802	.3418671	-1.29	0.197	-1.111227	.228867
age	-.4626917	.1173426	-3.94	0.000	-.692679	-.2327043
_cons	-7.265041	7.926913	-0.92	0.359	-22.8015	8.271422

acreg - Output: Network

```
. acreg hrate ln_population age (ln_income=unemployment)    ,  id(_ID) time(year)
>  links_mat(linksmat*) network  dist(1) lag(50)

NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  50
LagDistCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:  46.357
                                                      Number of obs =      5648
Total (centered) SS      =  286387.1082               Centered R2   =   -0.0447
Total (uncentered) SS    =  781008.6785               Uncentered R2 =    0.6169
Residual SS              =  299188.6495
```

hrate	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ln_income	3.83872	1.487876	2.58	0.010	.9225371	6.754904
ln_populat~n	-.4411802	.3418671	-1.29	0.197	-1.111227	.228867
age	-.4626917	.1173426	-3.94	0.000	-.692679	-.2327043
_cons	-7.265041	7.926913	-0.92	0.359	-22.8015	8.271422

Simulations

Simulations

In each Monte Carlo draw:

1. Generate random variables Y and X_1, and random shocks ε_Y and ε_{X_1} for each observation Go
2. Distribute the random shocks to "linked observations" Go
   - Spatial Environment: kernel around Counties in U.S. Illustration
   - Network Environment: coauthors in economics (RePEc)
3. Introduce the correlation in the model by adding the common shocks to Y and X_1 Go
4. Regression of Y on X_1 and a constant. Go

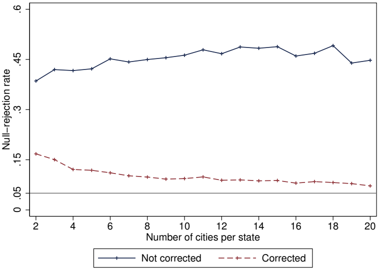
Test: as the number of Monte Carlo draws approaches infinity, the null hypothesis that $\hat{\beta} = 0$, in a test with $\alpha = 0.05$, will be rejected 5% of the times only if spatial correlation is accounted for.

Results

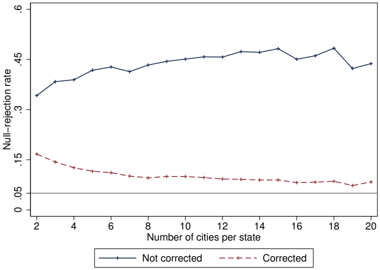
Spatial setting: Null-rejection rates

Data generating process:				Bartlett kernel		
Unit:				U.S. towns		U.S. counties
Sample size:				N=101	N=1001	N=3141
				(1)	(2)	(3)
Spatial correlation	**Correction**	**Endogeneity**	**Estimator**	**Null-rejection rate**		
Panel A: Cross section, $t = 1$						
			OLS	5.9%	5.0%	5.0%
		✓	2SLS	5.6%	5.1%	5.2%
✓			OLS	37.8%	50.2%	28.2%
✓		✓	2SLS	33.4%	48.3%	26.5%
✓	✓		OLS	16.8%	7.2%	5.6%
✓	✓	✓	2SLS	16.7%	8.4%	5.5%
Panel B: Panel, $t = 5$						
			OLS	5.8%	5.1%	5.3%
		✓	2SLS	5.3%	5.0%	4.6%
✓			OLS	39.1%	46.1%	17.9%
✓		✓	2SLS	37.3%	44.3%	15.5%
✓	✓		OLS	19.4%	11.2%	10.1%
✓	✓	✓	2SLS	19.0%	11.1%	9.6%

Spatial setting: Null-rejection rates by sample size, cross section, t=1

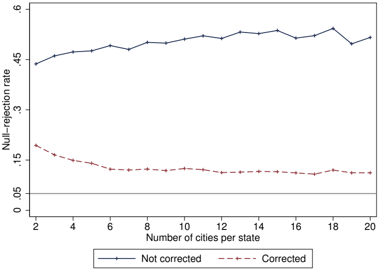


(a) OLS

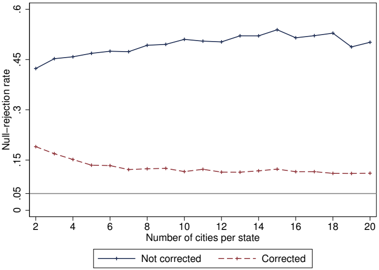


(b) 2SLS

Spatial setting: Null-rejection rates by sample size, panel, t=5



(c) OLS



(d) 2SLS

Network setting: Null-rejection rates

Data generating process:				First-degree friends			
Unit:				Top of the distribution		Random sample	
Sample size:				N=1000	N=2500	N=1000	N=2500
				(1)	(2)	(3)	(4)
Network correlation	**Correction**	**Endogeneity**	**Estimator**	**Null-rejection rate**			
			OLS	5.1%	4.7%	4.7%	5.1%
		✓	2SLS	5.3%	4.9%	5.4%	4.7%
✓			OLS	64.9%	59.0%	26.9%	36.2%
✓		✓	2SLS	63.0%	58.2%	25.4%	35.4%
✓	✓		OLS	13.2%	9.2%	7.5%	8.1%
✓	✓	✓	2SLS	13.4%	9.7%	7.2%	8.4%

Conclusions

Conclusions

- We propose a variance-covariance matrix (VCV) estimator, accompanied with a companion statistical package **acreg** for Stata, that allows researchers to obtain cluster-robust inference in OLS and 2SLS settings with arbitrary dependence across observations and over time

- We show that arbitrary clustering correction produces consistent estimates of the VCV by means of Monte Carlo simulations

- **Next step:** Facing theoretically the dimensionality problem (sufficient number of clusters) in the arbitrary clustering environment and produce guidelines for the users

Thank You

Appendix

Data Generating Process (DGP) - Baseline

For each observational unit we generate two iid random variables Y and X_1

$$X_1 \sim N(\overline{X_1}, \sigma_{X_1})$$

$$Y \sim N(\overline{Y}, \sigma_Y)^1$$

For each observational unit we also generate two random shocks ε_Y and ε_{X_1} that are independent and identically distributed (iid):

$$\varepsilon_{X_1} \sim (0, \sigma_{\varepsilon_{X_1}})$$

$$\varepsilon_Y \sim (0, \sigma_{\varepsilon_Y})$$

[1] $\overline{Y}$ and $\overline{X_1}$ can be any number. Given that Y and X_1 are iid, statistical theory predicts that if we regress Y on X_1, the null hypothesis that the β coefficient is equal to 0 at a 5% level, will be rejected with 5% probability.

Back

Data Generating Process (DGP) - Correlation

Spatial Environment

We take each Town/County in US as an observational unit and we dissipate the shocks $\varepsilon_{X_{1i}}$ and ε_{Y_i} to all observations j_s that are within a spatial distance from observation i. We impose a bartlett kernel such that the effect is lower as the spatial distance between observations i and j increases.

The total common shock an observation receives are ς_ξ, with $\xi = \varepsilon_{X_1}, \varepsilon_Y$:

$$\varsigma_{\xi_i} = \xi_i + \sum_{j \neq i}^{N} [1 - (dist_{ij}/distcut)] \times \xi_j$$

Network Environment

We take each author registered at RePEc as an observational unit and we dissipate the shocks $\varepsilon_{X_{1i}}$ and ε_{Y_i} to all her coauthors registered at RePEc. Each coauthor j receives a fraction, ρ, of each shock.

The total common shock an observation receives are ς_ξ, with $\xi = \varepsilon_{X_1}, \varepsilon_Y$:

$$\varsigma_{\xi_i} = \xi_i + \sum_{j \neq i}^{N_i} \rho \times \xi_j \ ; \quad \rho > 0$$

Back

DGP - correlation in the model

We introduce the correlation created into the model by adding the sum of common shocks to the variables, X_1 and Y:

$$\hat{X}_{1i} = X_{1i} + \varsigma \varepsilon_{X_{1i}}$$

$$\hat{Y}_i = Y_i + \varsigma \varepsilon_{Y_i}$$

Endogeneity Panel dimension Back

DGP - regression

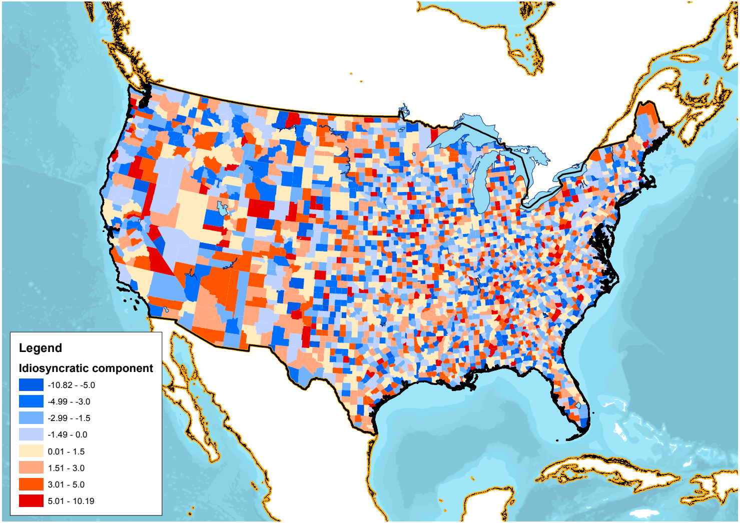
We estimate the following equation both correcting and not correcting for the presence of spatial/network correlation using OLS:

$$
\begin{aligned}
\hat{Y}_i &= \alpha_{3i} + \hat{\beta}\hat{X}_{1i} + \upsilon_i \\
&= \alpha_{3i} + \hat{\beta}(X_{1i} + \varsigma\varepsilon_{X_{1i}}) + (\upsilon'_i + \varsigma\varepsilon_{Y_i})
\end{aligned} \tag{1}
$$

Null hypothesis that $\hat{\beta} = 0$ will be rejected 5% of the time at 5% level if spatial correlation in the model is accounted for.

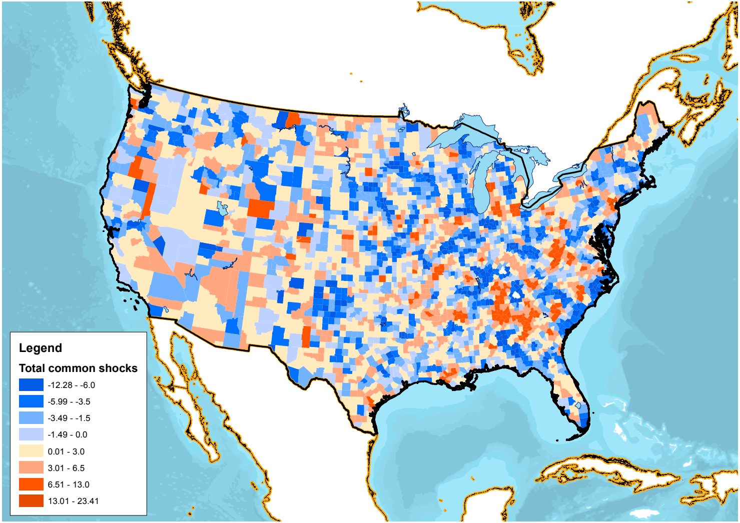
▸ Back

Illustration 1: Idiosyncratic shocks



Correlated Shocks
Back

Illustration 2: Spatially correlated shocks



Back

Data Generating Process, endogeneity

We introduce endogeneity to the model by adding an endogenous variable, *End*, as a regressor:

$$Y_i = \alpha_{1i} + \delta_1 X_{1i} + \delta_2 End_i + \mu_i \tag{2}$$

We generate a random variable *IV*, which is independent and identically distributed (iid) to Y and X_1:

$$IV = \overline{IV} + \epsilon_{IV},\ \epsilon_{IV} \sim N(0, \sigma_{\epsilon_{IV}});$$

We define End_i as:

$$End_i = F(X_{1i}, IV_i) + \epsilon_{Y_i}$$

We introduce correlation to the 2SLS model by adding the sum of common random shocks, $\varsigma\varepsilon_{IV_i}$, to the variable *IV* and computing *End* as a function of correlated variables and common shocks:

$$\hat{IV}_i = IV_i + \varsigma\varepsilon_{IV_i}$$

$$\hat{End}_i = F(\hat{X}_{1i}, \hat{IV}_i) + \epsilon_{Y_i} + \varsigma\varepsilon_{Y_i}$$

Back

Data Generating Process, panel dimension

Before introducing correlation to the model, we introduce auto-correlation of degree 1 by adding a fraction of the random common shock an observation receives in time $t-1$ to the random common shock it receives in time t:

$$\varepsilon_{Y_{it}} = \varepsilon_{Y_{it}} + \phi\varepsilon_{Y_{it-1}};$$
$$\varepsilon_{X_{1it}} = \varepsilon_{X_{1it}} + \phi\varepsilon_{X_{1it-1}};$$
$$\varepsilon_{IV_{it}} = \varepsilon_{IV_{it}} + \phi\varepsilon_{IV_{it-1}};$$
$$\phi > 0$$

This ensures that observation i in time t affect observation j in time $t+1$ if i and j are in the same arbitrary spatial cluster, i.e., $dist_{ij} \leq distcut$.

Back