acreg: Arbitrary Correlation Regression

Fabrizio Colella, Rafael Lalive, Seyhun O. Sakalli, Mathias Thoenig
(UNIL) (UNIL) (King's College) (UNIL)

www.acregstata.weebly.com

(Virtual) Swiss Stata Meeting 2020

Bern, November 2020

Introduction

Motivation I

Modeling the convoluted correlation structures between units improves inference

- Spatial data:
  - Geographical positions of observations
  - Neighborhood structures
- Network data:
  - Social networks
  - Mobile data
  - Co-working relations

Motivation II

But only a few studies offers a flexible theoretical framework (Bester et al., 2011)

Commonly used practices:

- Spatial Data
  - Cluster (Cameron et al., 2011)
  - Conley's Spatial Clustering (Conley, 1999a)
- Network Data
  - Cluster

Motivation III

And the STATA literature on the topic is limited

- Robust (White, 1980) and Two-way clustering corrections (Cameron and Miller, 2015) included in most programs computing OLS and 2SLS regressions.

- In the Spatial literature there are some programs to account for correlation using coordinates
  - Conley, 1999b
  - Hsiang, 2010

- There are no STATA packages available to account for correlation between neighbors or observations in a network

Motivation IV

In a related paper (Colella et al., 2019):

- Building on White (1980), we develop an *Arbitrary Clustering* approach to deal with inference with any type of topological and temporal dependence between observational units
- We perform extensive Monte Carlo simulations for both spatial and network data structures comparing different methods
- We show that commonly used techniques reject the null hypothesis about 110% times more than they should, while with our approach gets close to the true rejection rate. Go
- Provide guidelines for conducting inference in complex settings

This Paper

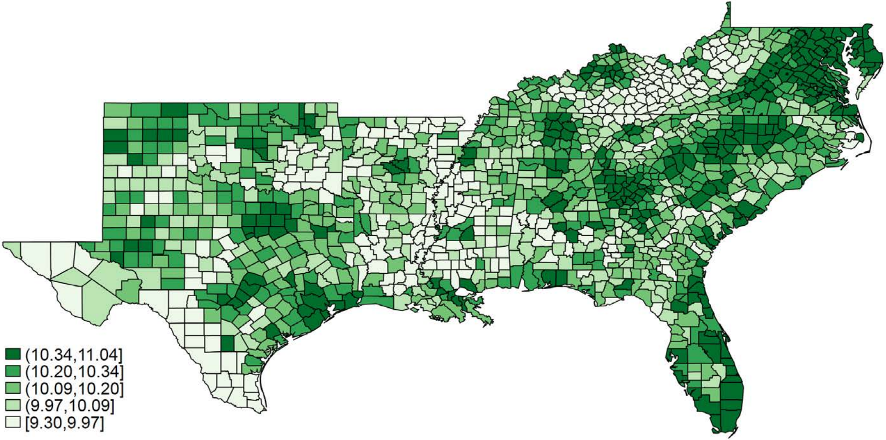
We introduce a new STATA package (and a companion paper) implementing the standard errors correction approach proposed in Colella et al. (2019):

ACREG: Arbitrary Correlation Regression

- Computes adjusted standard errors for:
  - Spatial data (coordinates or contiguity matrix),
  - Network data (adjacency matrix),
  - Multi-way clustering environments (infinite list of clustering variables)
- Suits OLS and 2SLS settings
- Includes temporal correlation for panel data

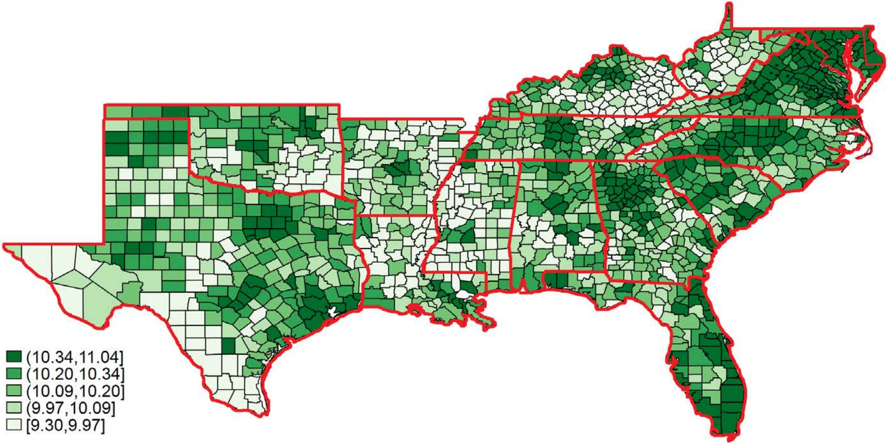
Correlation with Spatial Data

Correlation in Space



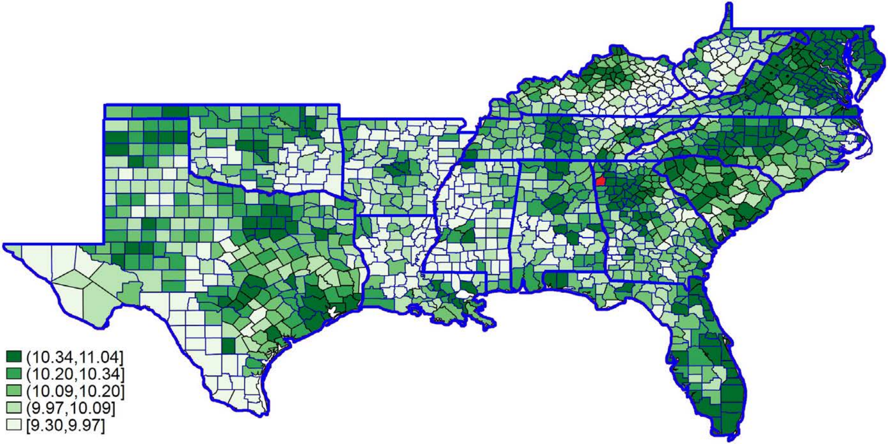
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation in Space - Clustering by State



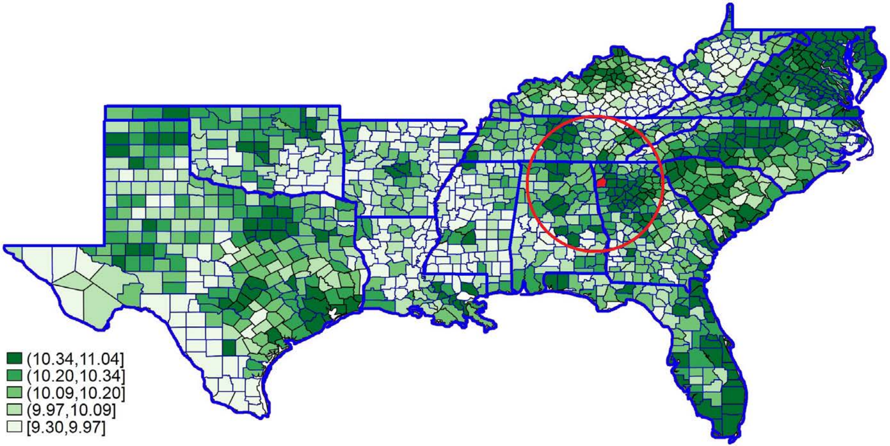
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation in Space - Clustering by State



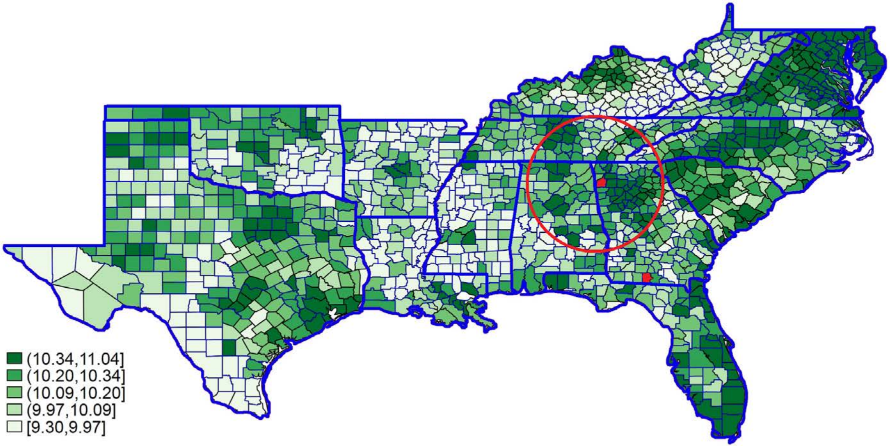
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation in Space - Conley 1999



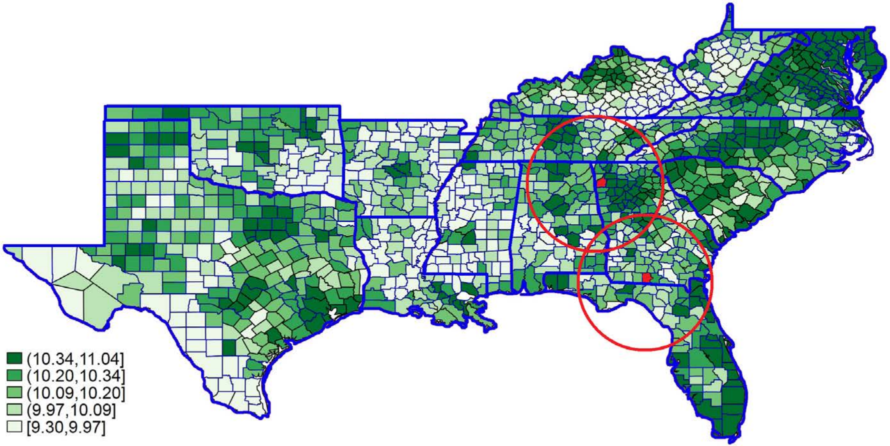
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation in Space - Conley 1999



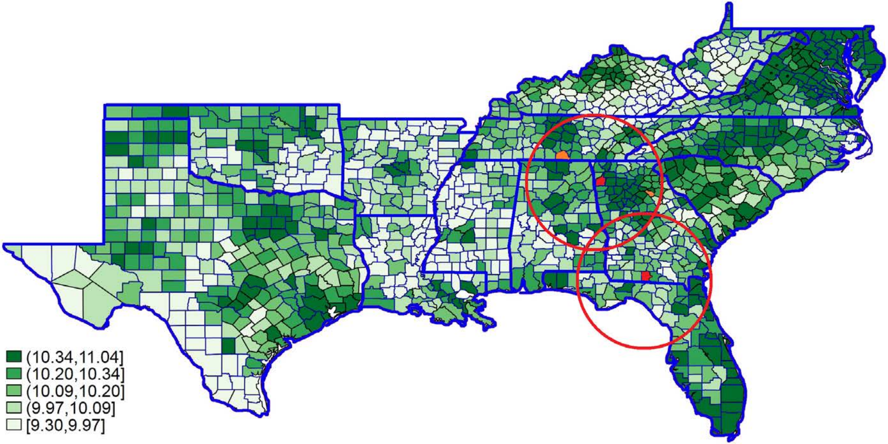
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation in Space - Conley 1999



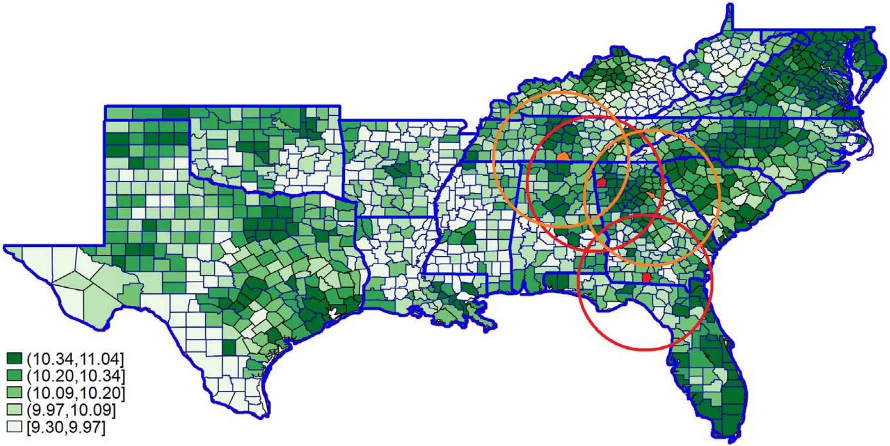
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation in Space - Conley 1999



Income in 1990 for southern U.S. counties - Messner et al. (1999)

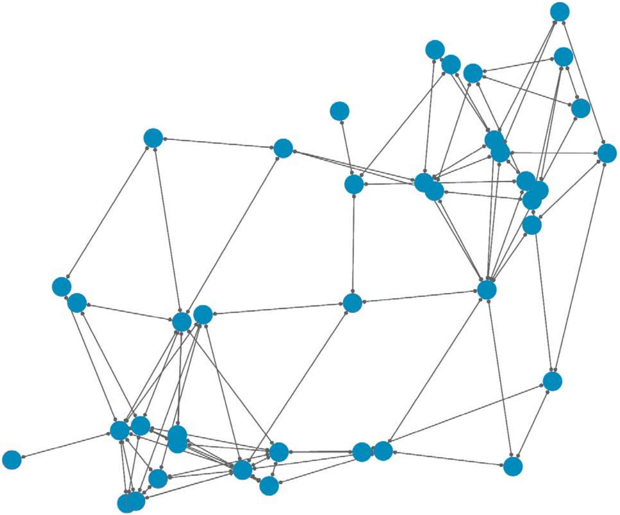
Correlation in Space - Conley 1999



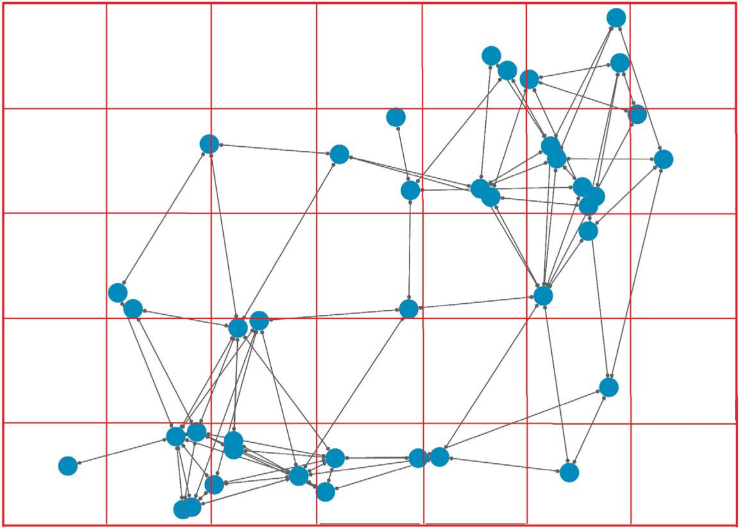
Income in 1990 for southern U.S. counties - Messner et al. (1999)

Correlation with Network Data

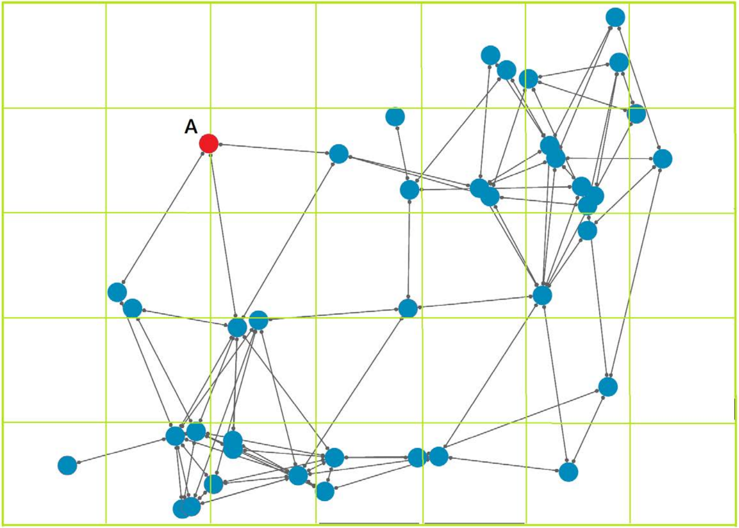
Correlation in Network



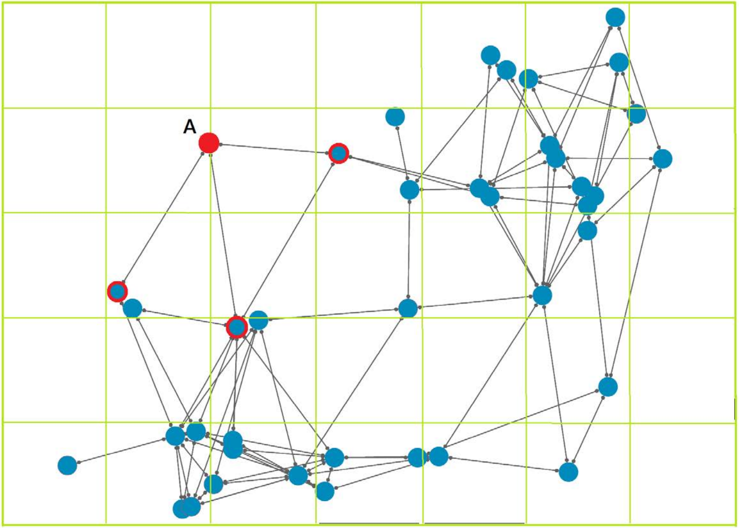
Correlation in Network - One way clustering



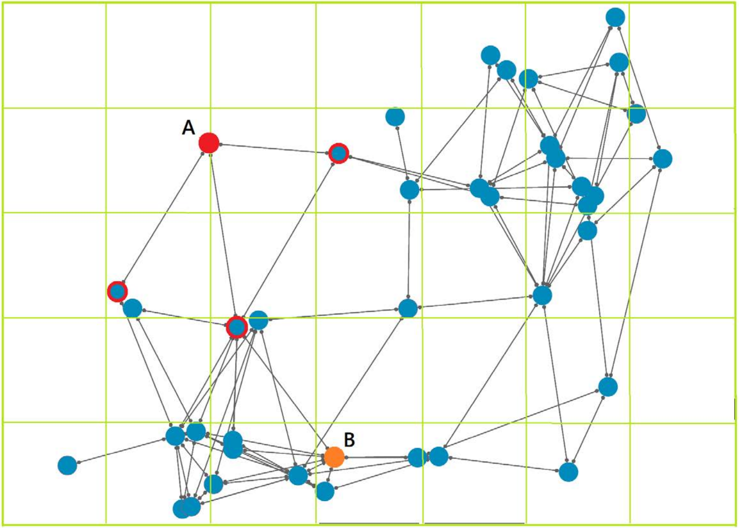
Correlation in Network - One way clustering



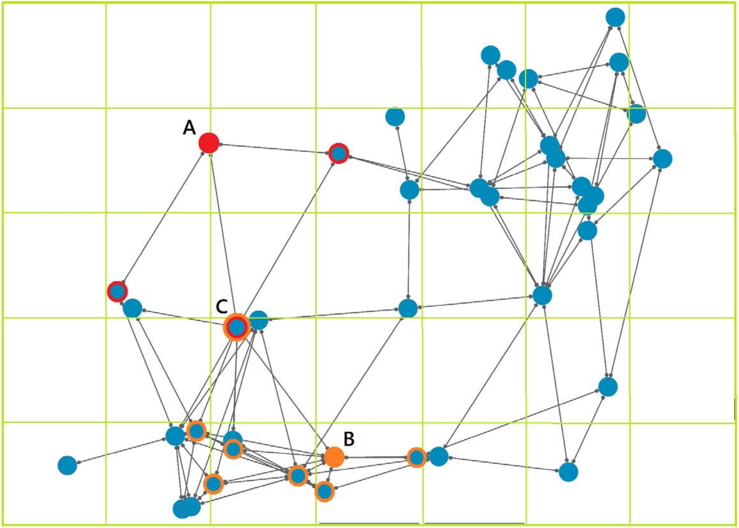
Correlation in Network - Network Clusters



Correlation in Network - Network Clusters



Correlation in Network - Network Clusters



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 OLS estimator

Linear Model

$$y = X\beta + \epsilon$$

Standard OLS Estimator

$$b_{OLS} = (X'X)^{-1}(X'y)$$

With Variance

$$VCV(b_{OLS}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

Where:
y is the Dependent Variable
X is the Matrix of Regressors (exogenous and endogenous)
Ω is the VCV of errors

Estimating the VCV of the OLS estimator

Proposed Estimator for $X'\Omega X$ is:

$$X'(S \times (uu'))X = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T} x_{it} u_{it} u_{js} x_{js} \mathrm{s}_{\mathrm{itjs}}$$

Where:
$u \equiv y - X\beta_{OLS}$ are the estimated residuals

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

Syntax

Syntax - Baseline

`acreg` *depvar* [*varlist1*] [(*varlist2* = *varlist_iv*)] [*if*] [*in*] [*fweight pweight*]

- *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

Syntax - Time Dimension

acreg *depvar varlist1* (*varlist2* = *varlist_iv*),

id(*idvar*) time(*timevar*) lag(#)

- *idvar* is the cross-sectional unit identifier
- *timevar* is the time unit variable
- lag(#) specifies the time lag cutoff for observations with the same idvar

Syntax - Spatial I

acreg *depvar varlist1* (*varlist2* = *varlist_iv*), spatial latitude(*latitudevar*) longitude(*longitudevar*) dist(#)

- spatial specifies the spatial environment
- *latitudevar* is the variable containing the latitude of each observation in decimal degrees: range[-180.0, 180.0]
- *longitudevar* is the variable containing the longitude of each observation in decimal degrees: range[-180.0, 180.0]
- dist(#) specifies the distance cutoff beyond which the correlation between error term of two observations is assumed to be zero, in km

Syntax - Spatial II

`acreg` *depvar varlist1* (*varlist2* = *varlist_iv*), `spatial` `dist_mat(`*varlist_distances*`)` `dist(#)`

- `spatial` specifies the spatial environment
- *varlist_distances* is the list of N variables containing bilateral spatial distances between observations in any meaningful metric, e.g., physical or travel distance between two locations.
- `dist(#)` specifies the distance cutoff beyond which the correlation between error term of two observations is assumed to be zero, in the same metric as *varlist_distances*

Syntax - Network I

acreg *depvar varlist1* (*varlist2* = *varlist_iv*), `network` `links_mat(`*varlist_links*`)` `dist(#)`

- `network` specifies that the network environment
- `varlist_links` is the list of N binary variables specifying the links between observations, e.g., the adjacency matrix. The links between two units can change over time.
- `dist(#)` specifies the distance cutoff (geodesic paths) beyond which the correlation between error term of two observations is assumed to be zero. If it is greater than 1, acreg computes the bilateral distance between two nodes.

Syntax - Network II

acreg *depvar varlist1* (*varlist2* = *varlist_iv*), `network` `dist_mat(`*varlist_distances*`)` `dist(#)`

- `network` specifies that the network environment
- `varlist_distances` is the list of N variables containing bilateral distances between observations in the network, i.e., the number of links along the shortest path between two nodes.
- `dist(#)` specifies the distance cutoff (geodesic paths) beyond which the correlation between error term of two observations is assumed to be zero. If it is greater than 1, acreg computes the bilateral distance between two nodes.

Syntax - Multiway Clustering

`acreg` *depvar varlist1* (*varlist2* = *varlist_iv*), `cluster`(*varlist_cluster*)

- *varlist_cluster* is the list of variables identifying the different clusters. Each variable identify a specific cluster dimension and its clusters.

Syntax - Arbitrary Clustering

`acreg` *depvar varlist1* (*varlist2* = *varlist_iv*), `weights`(*varlist_weights*)

- *varlist_weights* is the list of N ($\times$ T if a time dimension is specified) variables containing the S matrix weights. The N $\times$ T variables need to follow the same order of the observations.

Syntax - Options

Correlation Structure

- `hac` reports Heteroskedasticity and Autocorrelation Corrected (HAC) standard errors; lagcutoff will be the temporal decay, requires id, time, and lagcutoff.
- `bartlett` imposes a distance linear decay between observations within the cutoff in the correlation structure.
- `nbclust(#)` is the number of clusters used to compute the Kleibergen-Paap statistic in case of arbitrary cluster correction; default is 100.

Syntax - Options II

High-Dimensional Fixed Effects

- *fe1var* identifies the first high-dimensional fixed effects variable to be partialled out.
- *fe2var* identifies the second high-dimensional fixed effects variable to be partialled out.
- `dropsingletons` drops singleton groups when pfe1 (and pfe2) is (are) specified.

Storing

Storing Options

- `storeweights` stores the computed weights used to correct the VCV for arbitrary cluster correlation as a matrix under the name weightsmat, which may be used as input for the option varlist_weights; optional only if spatial option, network option, or varlist_cluster is specified.

- `storedistances` stores the computed distances used to correct the VCV for arbitrary cluster correlation as a matrix under the name distancesmat, which may be used as input for the option varlist_distances; optional only if spatial option or network is specified and varlist_distances is not specified.

Saved Values

Scalars

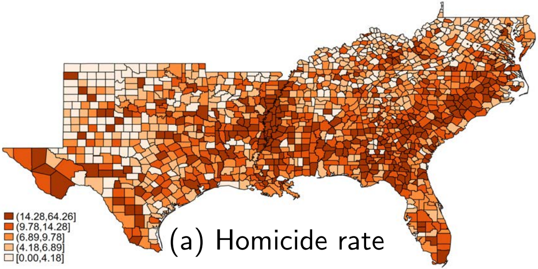
`e(N)`	number of observations
`e(mss)`	model sum of squares (centered)
`e(mssu)`	model sum of squares (uncentered)
`e(rss)`	residual sum of squares
`e(tss)`	total sum of squares (centered)
`e(tssu)`	total sum of squares (uncentered)
`e(r2)`	centered R2 (1-rss/tss)
`e(r2u)`	uncentered R2
`e(widstat)`	Kleibergen-Paap Wald rk F statistic

Matrices

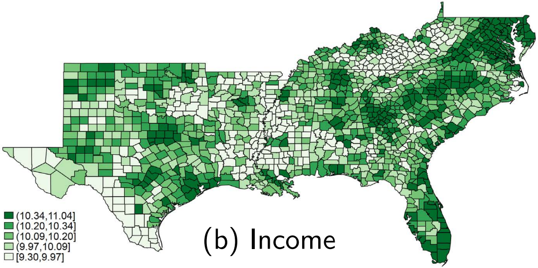
`e(b)`	coefficient vector
`e(V)`	corrected variance-covariance matrix of the estimators

Examples

Income and homicide rate I



(a) Homicide rate



(b) Income

Messner et al., 1999

Income and homicide rate II - Setting

We want to estimate the following equation accounting for potential spatial correlation when computing the SEs.

$$homicidesrate_{it} = \alpha_i + \beta logincome_{it} + \gamma X_{it} + \epsilon_{it}$$

Where i is a county in south-est US and t is one of the four years included in the sample. X_{it} includes log-population, and average age.

We instrument income with the unemployment rate. First stage:

$$logincome_{it} = \alpha_{2i} + \beta_2 unemployment_{it} + \gamma_2 X_{it} + \epsilon_{2it}$$

Income and homicide rate III - Syntax

```
acreg hrate ln_population age (ln_income = unemployment),
spatial latitude(_CX) longitude(_CX) dist(100)
id(_ID) time(_ID) lagcut(30)
pfe1(_ID)
```

- `dist(100)` states that spatial correlation is assumed to vanish after 100 Km
- `lagcut(30)` states that temporal correlation among observations from the same individual is assumed to vanish after 30 time periods (years)
- `pfe1(_ID)` includes individual Fixed Effects in the model through dummies, and partial them out to save time

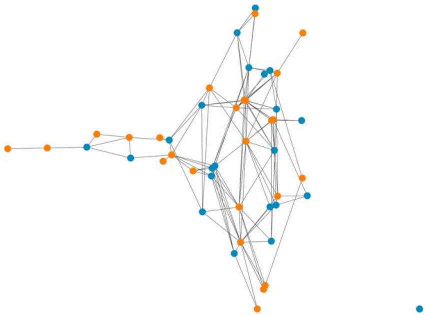
Income and homicide rate IV - Output

```
SPATIAL CORRECTION
DistCutoff: 100
LagCutoff:  30
No HAC Correction
Absorbed FE: _ID
Included instruments: ln_population age
Instrumented: ln_income
Excluded instruments: unemployment
Kleibergen-Paap rk Wald F statistic:   49.605
                                                  Number of obs =      5648
Total (centered) SS     =  144755.2058            Centered R2   =    0.0175
Total (uncentered) SS   =  144755.2058            Uncentered R2 =    0.0175
Residual SS             =  142223.0274
```

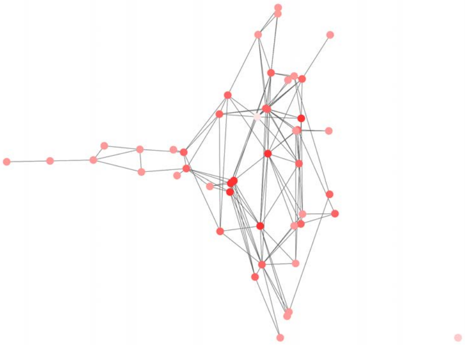
hrate	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ln_income	.2588154	1.149746	0.23	0.822	-1.994645	2.512276
ln_population	-1.630949	1.740873	-0.94	0.349	-5.042997	1.781099
age	.1466193	.2006033	0.73	0.465	-.2465559	.5397944
_cons	-1.31e-17	.1743959	-0.00	1.000	-.3418097	.3418097

```
nb: total SS, model and R2s are after partialling-out.
To get the corrected ones use the option correctr2
```

Gang Network I



(a) Arrest



(b) Ranking

Grund and Densley, 2012

Gang Network II - Setting

We want to estimate the following equation accounting for potential spatial between linked individuals in the network when computing the SEs.

$$arrest_i = \alpha + \beta ranking_i + \gamma X_i + \epsilon_i$$

Where i is an individual, $arrest_i$ indicates the number of times that an individual was arrested and $ranking_i$ is the position in the gangs internal hierarchy. X_{it} includes age, place of residence and four binary variables identifying the birthplace.

Gang Network III - Syntax

```
acreg Arrest Ranking Age Residence i.Birthplace,
network links_mat(_net2_*) dist(1)
```

- `links_mat(_net2_*)` declares that the network structure is defined by the variables `_net2_1` ... `_net2_54`
- `dist(1)` states that network correlation is assumed to vanish after the first degree link

Gang Network IV - Output

```
NETWORK CORRECTION
DistCutoff: 1
LagCutoff:  0
No HAC Correction
No Absorbed FEs
Included instruments: Ranking Age Residence 1b.Birthplace 2.Birthplace
                      3.Birthplace 4.Birthplace
                                                   Number of obs =      54
Total (centered) SS     =  2196.537037             Centered R2   =  0.2442
Total (uncentered) SS   =         7497             Uncentered R2 =  0.7786
Residual SS             =  1660.198039
```

Arrests	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Ranking	-2.168476	.7132431	-3.04	0.002	-3.566407	-.7705455
Age	.7665194	.3730319	2.05	0.040	.0353904	1.497648
Residence	-1.534665	1.618858	-0.95	0.343	-4.707568	1.638239
Birthplace						
Caribbean	0	(empty)				
East Africa	-.2523035	2.258789	-0.11	0.911	-4.679449	4.174842
UK	.7012659	2.984775	0.23	0.814	-5.148785	6.551317
West Africa	.8171717	2.260143	0.36	0.718	-3.612627	5.24697
_cons	2.317286	7.825902	0.30	0.767	-13.0212	17.65577

Conclusion

Conclusion

We built `acreg`: a new user-written Stata routine allowing for standard error correction in OLS and 2SLS estimation of models with complex correlation structure.

- `acreg` can accommodate in a flexible way dependence of the errors between units in space or in a network and across time.
- `acreg` includes most of the standard options present in previous commands to estimate regression coefficients.
- The correlation structure can be introduced by the user in a matrix form or built from information on the geographic distance between spatial units or from the links between observations.

Thank You

www.fabcol.weebly.com

www.acregstata.weebly.com

Bester, C Alan, Timothy G Conley, and Christian B Hansen (2011). "Inference with dependent data using cluster covariance estimators". In: *Journal of Econometrics* 165.2, pp. 137–151.

Cameron, A., Jonah Gelbach, and Douglas Miller (2011). "Robust Inference With Multiway Clustering". In: *Journal of Business and Economic Statistics* 29.2, pp. 238–249. URL: https://EconPapers.repec.org/RePEc:bes:jnlbes:v:29:i:2:y:2011:p:238-249.

Cameron, Colin A. and Douglas L. Miller (2015). "A Practitioner's Guide to Cluster-Robust Inference". In: *Journal of Human Resources* 50.2, pp. 317–372. DOI: 10.3368/jhr.50.2.317. eprint: http://jhr.uwpress.org/content/50/2/317.full.pdf+html. URL: http://jhr.uwpress.org/content/50/2/317.abstract.

Colella, Fabrizio, Rafael Lalive, Seyhun Orcan Sakalli, and Mathias Thoenig (2019). "Inference with arbitrary clustering". In: *IZA Discussion Paper*.

Conley, T. G. (1999a). "GMM estimation with cross sectional dependence". In: *Journal of Econometrics* 92.1, pp. 1–45. URL: https://ideas.repec.org/a/eee/econom/v92y1999i1p1-45.html.

Conley, Timothy (1999b). "GMM estimation with cross sectional dependence". In: *Journal of Econometrics* 92.1, pp. 1–45. URL: https://EconPapers.repec.org/RePEc:eee:econom:v:92:y:1999:i:1:p:1-45.

Grund, Thomas U and James A Densley (2012). "Ethnic heterogeneity in the activity and structure of a Black street gang". In: *European Journal of Criminology* 9.4, pp. 388–406.

Hsiang, Solomon M. (2010). "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America". In: *Proceedings of the National Academy of Sciences* 107.35, pp. 15367–15372. ISSN: 0027-8424. DOI: 10.1073/pnas.1009510107. eprint: https://www.pnas.org/content/107/35/15367.full.pdf. URL: https://www.pnas.org/content/107/35/15367.

Manson, Steven, Jonathan Schroeder, David Van Riper, and Steven Ruggles (2017). *IPUMS National Historical Geographic Information System: Version 12.0 [Database]*. Minneapolis: University of Minnesota. http://doi.org/10.18128/D050.V12.0.

Messner, Steven F, Luc Anselin, Robert D Baller, Darnell F Hawkins, Glenn Deane, and Stewart E Tolnay (1999). "The spatial patterning of county homicide rates: An application of exploratory spatial data analysis". In: *Journal of Quantitative criminology* 15.4, pp. 423–450.

White, Halbert (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". In: *Econometrica* 48.4, pp. 817–38. URL: `https://EconPapers.repec.org/RePEc:ecm:emetrp:v:48:y:1980:i:4:p:817-38`.

Appendix

Colella et al., 2019 - Simulations Result

(a) Space - U.S. counties

	Spatial corr.	Estimator	Correction	Null-rejection rate
(1)		OLS	robust	5.2%
(2)	✓	OLS	robust	9.1%
(3)	✓	OLS	cluster	6.8%
(4)	✓	OLS	acreg	5.5%

(b) Network - Coauthors in Economics

	Network corr.	Estimator	Correction	Null-rejection rate
(1)		OLS	robust	4.8%
(2)	✓	OLS	robust	9.8%
(3)	✓	OLS	cluster, affiliation [N = 611]	9.6%
(4)	✓	OLS	cluster, degree city [N = 135]	12.0%
(5)	✓	OLS	acreg	5.6%

Back

Colella et al., 2019 - Findings

- Commonly used methodologies reject the null hypothesis about 110% times more than they should
- With our estimator we get close (no statistical difference) to the test level
- Our estimator asymptotically converges to the true value
- The bias in the SEs emerges only if both the outcome and the v.o.i. follow a topology
- Adding covariates helps in addressing the issue only if they are likely to affect both the outcome and the topology