I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12584

Inference with Arbitrary Clustering

Fabrizio Colella
Rafael Lalive
Seyhun Orcan Sakalli
Mathias Thoenig

AUGUST 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12584

Inference with Arbitrary Clustering

Fabrizio Colella
HEC University of Lausanne

Rafael Lalive
HEC University of Lausanne and IZA

Seyhun Orcan Sakalli
HEC University of Lausanne

Mathias Thoenig
HEC University of Lausanne

AUGUST 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Inference with Arbitrary Clustering*

Analyses of spatial or network data are now very common. Yet statistical inference is challenging since unobserved heterogeneity can be correlated across neighboring observational units. We develop an estimator for the variance-covariance matrix (VCV) of OLS and 2SLS that allows for arbitrary dependence of the errors across observations in space or network structure, and across time periods. As a proof of concept, we conduct Monte Carlo simulations in a geospatial setting based on US Metropolitan areas; tests based on our estimator of the VCV asymptotically correctly reject the null hypothesis where conventional inference methods, e.g. those without clusters, or with clusters based on administrative units, reject the null hypothesis too often. We also provide simulations in a network setting based on the IDEAS structure of co-authorship and real life data on scientific performance; the Monte Carlo results again show that our estimator yields inference at the right significance level already in moderately sized samples, and it dominates other commonly used approaches to inference in networks. We provide guidance to the applied researcher with respect to (i) including or not potentially correlated regressors and (ii) choice of cluster bandwidth. Finally we provide a companion statistical package (acreg) enabling users to adjust OLS and 2SLS coefficient's standard errors, accounting for arbitrary dependence.

JEL Classification: C13, C23, C26

Keywords: clustering, arbitrary, geospatial data, network data, cluster, spatial correlation, instrumental variables

Corresponding author:
Rafael Lalive
Department of Economics
HEC University of Lausanne
CH-1015 Lausanne
Switzerland
E-mail: Rafael.Lalive@unil.ch

* Our companion statistical package (acreg) can be downloaded at the following address https://acregstata.weebly.com. For helpful comments and valuable feedback on early versions of our command we thank Samuel Bazzi, Nicolas Berman, Richard Bluhm, Johannes Buggle, Mathieu Couttenier, David Drukker, Ruben Durante, Ruben Enikopolov, Elena Esposito, Matthew Jackson, Melanie Krause, Eleonora Patacchini, as well as participants at the Swiss Meeting of Stata Users (Zurich, 2018) and the Workshop on Geodata and Economics (Braunchweig, 2018).

1 INTRODUCTION

Recent years have witnessed a tremendous surge of empirical studies with data endowed with a topology, such as spatial data or network data. In these data, unobserved shocks can be correlated across neighboring observational units, where the neighborhood refers to the physical space or to the network structure. In both settings, inference is challenging because the sampling structure of the data and of the VCV matrix exhibits overlapping clusters–a feature that is vastly ignored by applied econometricians.[1] Indeed, a common practice with spatial data consists of considering non-overlapping clusters (typically administrative units) defined at a level of aggregation that encompasses the scale of the resolution of the data by several orders of magnitude–e.g., standard errors are clustered at the region level, while observational units typically correspond to $0.5^o \times 0.5^o$ grid cells.[2] In addition to the loss of efficiency when it turns to estimation, such a practice is subject to caution for observational units that are located close to the frontier between two clusters, as shown in our analysis below. In the case of network data, the practice is even more radical, as many studies simply do not correct for the potential correlation of unobserved shocks across neighbors.

We propose an approach to obtain asymptotically valid inference in spatial/network settings with any type of topological and temporal dependence between observation units. We also provide the community with a companion statistical package. Our `acreg` command enables Stata users to estimate 2SLS models with panel data and an arbitrary clustering structure. Arbitrary here refers to the way units could be correlated with each other in space/network and time. We impose no restrictions so that our approach can be used with a wide range of data. Our estimator for the variance-covariance (VCV) matrix of the estimated parameters builds on the seminal insight by White (1980) who showed that a sandwich-type VCV can be estimated by constructing a consistent estimator of the VCV of the parameters. Specifically, the estimator uses estimated regression errors and knowledge of the clustering structure to reconstruct estimates of the unknown elements of the sandwich formula. Our approach follows Conley (1999) by specifying a circle around each unit that specifies how distance dependence is likely to reach, allowing for decay or not. This type of clustering structure is well known in spatial data, and statistical packages are available online for ordinary least squares (OLS) estimations. Our contribution is twofold. First, we show how to perform inference in instrumental variables (IV) or two-stage least squares (2SLS) settings allowing for a Conley-type clustering structure. Second, we allow users to define the metric in a flexible way: In addition to spatial distance, our approach can

[1] Multiway clustering is somewhat more flexible, allowing errors to correlate, for instance, within units over time and across time periods (Cameron *et al.*, 2011). However, multiway clustering assumes regularity in the clustering structure that may not hold in real-life settings with spatial/network data.

[2] For example, see the GAEZ v3.0 Global Agro-ecological Zones dataset of FAO: http://www.gaez.iiasa.ac.at/

deal with travel distance, travel costs, contiguity and any concept of distance in a network.

A first example of application of our approach relates to a clustering structure allowing for spatial and temporal decays with geocoded data. Indeed, empirical work has been fueled by the growing availability of geocoded data and the integration of geographic information systems (GIS) in the toolkit of economists. From development and urban economics to economic history, big spatial data at a high level of resolution enable researchers to move the analysis within countries and to craft compelling empirical designs (e.g., RDD, DiD), for the purpose of causal analysis, as various endogeneity concerns are alleviated by exploiting fine-grained variations and discontinuities in the variables of interest.[3] A second and broader class of applications relates to all clustering structures that are based on a metric that is not spatial distance (i.e., Euclidean or geodesic) such as contiguity or any type of network topology. More specifically, consider a scholar interested in studying economic outcomes at the county level in the U.S. In such a scenario, it is likely that contiguous counties are affected by common shocks and this should be reflected in the clustering structure. The issue here is that counties have different sizes (much larger in the West; see the map in Figure 3), preventing the researcher from imposing the same spatial kernel (Conley, 1999, as in) across the entire sample. Another setting relates to networks. Consider a scholar interested in violence between rebel groups in Africa. These groups are affected by common shocks not only in the physical space through their location but also in the cultural/social space through their ethnic affiliations. Groups that are ethnically close tend to be affected by similar shocks. Hence, it is important that the clustering structure accounts for ethnic (or genetic or linguistic) relatedness.

We provide results from extensive Monte Carlo simulations based on real-life data for documenting our arbitrary clustering regression approach. A first set of simulations relates to the clustering structure allowing for spatial and temporal decays. We construct environments where OLS or IV regressions with robust standard errors clustered at the administrative level reject the null hypothesis of no effect in approximately 20% of all cases when the significance level of the test is set at 5%. Conventional inference does not improve as the sample size increases, suggesting that the conventional approach produces inconsistent estimates of the variance-covariance matrix. By adopting the arbitrary clustering estimator, we find that the null-rejection rate is approximately 10% for small samples and converges quickly to the true significance level of 5% as the sample size increases. This pattern suggests that the arbitrary clustering correction produces consistent estimates of the VCV, enabling applied econometricians to conduct robust inference in the presence of spatial correlation. Our second Monte Carlo study deals with network data based on coauthorship in Economics from IDEAS. Here, we again find that applied econome-

[3] For a survey, see Michalopoulos and Papaioannou (2017).

tricians adopting conventional inference using robust standard errors that neglect the network correlation in both regressors and outcomes would severely overstate the precision of their estimates. By contrast, inference that allows for arbitrary clustering yields rejection rates close to the correct 5% threshold. Finally, we exploit our Monte Carlo results to provide guidance to the applied researcher with respect to (i) including control variables; (ii) multiple spatially correlated regressors; and (iii) setting the adequate (spatial/network distance) bandwidth for the estimator.

This paper is related to several strands in the literature. First, our approach to conducting inference is inspired by White (1980)'s seminal work on consistent estimation of the VCV. White (1984) also proposed an estimator that allows for robust inference when data are clustered, as, e.g., in random samples of units observed over multiple time periods. Bertrand *et al.* (2004) discuss how clustering affects studies that adopt a differences-in-differences design. Cameron *et al.* (2011) extended this approach to clustering in multiple dimensions. Second, a large body of literature on spatial econometrics discusses inference approaches. Conley (1999) develops robust inference in settings where shocks to spatial units are spatially dependent, also allowing for decays. Kelejian and Prucha (1998, 1999) develop estimators in a spatial setting with spatial dependence in both the dependent variable and the regressors. We complement this literature by allowing for arbitrary forms of clustering. The recent surge in availability of data with complex spatial dependence structures creates unprecedented demand for flexible modeling in order to ensure unbiased inference in complex settings. We also allow users to specify outside instruments, a requirement that is very important for applied papers but that seems overlooked or not discussed in the more theory-driven spatial econometrics literature. Finally, our simulations results, based on real data, qualify the main insights from Kelly (2019), who studies inference problems in spatial studies using artificial data.

We discuss the econometric background that allows for arbitrary clustering in the next section. Section 3 presents Monte Carlo evidence on placebo policy shocks for a spatial setting, the U.S. counties. We also document how arbitrary clustering provides reliable evidence in network settings. Section 4 concludes.

2 A Model with cross-section and time dependence

In this section, we present the model and discuss an estimator of the variance-covariance (VCV) matrix of the parameters. Our key focus is on inference with arbitrary dependence of error terms across observations and over time. Arbitrary in this sense conveys that each observation's error term at a particular point in time may depend on other observations' error terms with a certain strength. All this information is collected in a matrix that we call S. In the spatial context, S is

normally built from information on the geographic distance between spatial units, e.g., regions, cities, and countries. In a social network context, S reflects the direct links of each person, also called an adjacency matrix. Note that S allows for varying link strength, such that entries could range from 0 to 1, and S may change over time t. We also always include self-links in S, so its main diagonal contains ones. We begin with a review of the multiway clustering environment (Cameron *et al.*, 2011), we show the main differences between the two approaches, and we conclude the section by extending our setting to the 2SLS case.

Consider n observations at each t instant on time T from the following linear model:

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

where we observe each individual i several times in different periods t. y is a dependent variable, and X is a matrix of k linearly independent components that could include a long list of dummies for each unit, in case we are interested in the within estimates. We can write the OLS estimator as:

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

and the theoretical VCV of the b_{OLS} is:

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

where $\Omega \equiv E(\epsilon\epsilon'|X)$ is the unknown VCV of ϵ.

Building on the seminal insight from White (1980) and following the multiway cluster-robust estimator structure designed by Cameron *et al.* (2011), we propose the following sandwich estimator for the VCV based on the estimated residuals $e \equiv y - Xb_{OLS}$:

$$\widehat{VCV}(b_{OLS}) = (X'X)^{-1}X'(S \times (ee'))X(X'X)^{-1}$$

where S is the matrix capturing how each observation's error term depends on other observation's error terms. The key element of this estimator is the "meat" in the sandwich:

$$X'(S \times (ee'))X = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T} x_{it}e_{it}e_{js}x_{js}s_{itjs}$$

The Cameron *et al.* (2011) setting can be embedded in this framework. The peculiarity of their environment is the presence of D dimensions of clustering with G_d non-overlapping groups in each dimension, where each observation belongs to D groups with one in each dimension. In

their structure, the $it\,js$-th observation can be zero or one. It is equal to one if observation it and observation js share any cluster g_d and equal to zero otherwise. They show that it is a consistent estimator of the theoretical VCV providing that few regularity conditions hold.

Multiway clustering assumes a particular *regularity condition* in the clustering structure. For example, a sufficient condition for a single entry $it\,js$-th of the matrix S to be one is that observations it and js share a cluster in at least one dimension of clustering. This means that if we want to allow observation's i error terms to be correlated with the error terms of both observations j and l, then the error terms of observations j and l must also be correlated. In addition, if the observation i depends on observation j at time t, then they must also be dependent at time s. However, in many real-life settings, this particular clustering structure may not hold.

Conversely, our arbitrary cluster setting allows the units to be correlated with each other in any possible way, without any kind of imposed structure. Simply, the $it\,js$-th component of the matrix S can be zero, one or any other number between the two, depending on the strength of the dependence of the error of observation i on the error of observation j. The flexibility of our structure allows accounting for not only cross-section dependence and time dependence but also interactions between the two, capturing changes in the strength of the correlation that can be due to alterations in the link structure over time or any kind of decay between two moments in time t and s. This allows our estimator to be more suitable for many applications. In a social network context, S reflects the direct links of each person, while in the spatial context, S is built from information on the geographic distance between spatial units.

The framework described above could also be used in the presence of endogeneity. We consider the linear two-stage least squares with more instruments than endogenous regressors: once the endogeneity is taken into account and the causal effect of the explanatory variable on the dependent variable is uncovered through instruments, the procedure to estimate the VCV is qualitatively equivalent to the OLS case.

We consider the same linear model as before, where we add that m of the k components of X are endogenous and a set of $o > m$ excluded instruments for a total of $p > k$ exogenous variables that form the matrix Z. We can write the first stage as:

$$\hat{X} = (Z'Z)^{-1} Z'XZ$$

Then, the 2SLS estimator is:

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

Under standard regularity conditions b_{2SLS} is asymptotically normal with the following theoret-

ical estimated variance matrix:

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

Moreover, the core part $\hat{X}'\Omega\hat{X}$ can be estimated as before by the following:

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

where the estimated residuals now refer to the 2SLS estimator: $u \equiv y - Xb_{2SLS}$.

3 Simulation Study

We conduct Monte Carlo simulations to illustrate how correlation across units within an arbitrary cluster, e.g., spatially close units or friends in a network, affect the rejection rate of the null hypothesis if such correlation is not accounted for while estimating the standard errors. We implement Monte Carlo simulations using real-life data to construct arbitrary clusters, i.e., geocoded data on U.S. counties for the spatial setting and data on coauthors from IDEAS RePEc for the network setting. In each Monte Carlo iteration, we generate policy shocks that are based on variables that are randomly drawn from a normal distribution. These policy variables and outcome variables are independent and identically distributed (iid).[4] Statistical theory predicts that the null hypothesis that two iid variables are uncorrelated will be rejected 5% of the time as the number of Monte Carlo draws approaches infinity if the significance level is 5%. We confirm this prediction. Then, we introduce within-cluster correlation to the randomly generated policy variables and show that the null-rejection rate exceeds 5%, indicating a higher rate of Type 1 error if within-cluster correlation is not accounted for while estimating the standard errors. We show that our proposed estimator asymptotically reduces the null-rejection rate to the theoretical benchmark of 5% in the presence of within-cluster correlation in both OLS and 2SLS settings.

3.1 Spatial Setting

To illustrate how correlation across spatial units leads to an over-rejection of the null hypothesis if such correlation is not accounted for, we use data at the U.S. county level. We extract tabular information on median earnings, education level, age, race, and gender aggregated at the county

[4] This statement is true for the simulations in the OLS setting. To introduce endogeneity to the model in the 2SLS setting, we generate policy variables that are correlated with outcome variables.

level for 2000 from the National Historical Geographic Information System (NHGIS) database (Manson *et al.*, 2017). The NHGIS is a part of the Integrated Public Use Microdata Series (IPUMS) project of the University of Minnesota and provides tabular U.S. Census data and GIS boundary files. There are 3,141 counties in total in our sample.

3.1.1 The Data Generating Process and Hypotheses

To quantify the problem induced by spatial correlation in policy variables, we generate two sets of policy shocks: random and spatially correlated shocks. We first explain how we generate the random policy shocks, i.e., those that affect some counties and not others in each Monte Carlo draw. We generate a random variable from a normal distribution. Then, we select the counties that are in the top quartile of the distribution of this random variable as counties that receive a "placebo" policy shock. Panel (a) of Figure 1 visualizes an example of the distribution of the policy shock variable we draw at random.

We estimate the following equation using OLS:

$$Y_c = \alpha_{1c} + \beta_1 Policy_c + \delta_1 X_c' + \epsilon_c \tag{1}$$

where Y_c is the natural log of median earnings in county c in 2000. $Policy_c$ is a dummy variable indicating whether county c receives a policy shock. X_c is a vector of county-level controls, which comprises the share of population with tertiary education, share of females, share of blacks, median age and its square, and natural log of total population in 2000. Given that Y and $Policy$ are independent and $Cor(Policy_c, \epsilon_c) = 0$, statistical theory predicts that the null hypothesis that $\beta_1 = 0$ will be rejected 5% of the time at a 95% confidence interval as the number of Monte Carlo draws approaches infinity.

Next, we generate the spatially correlated policy shocks by adding spatial correlation to the randomly drawn placebo policy shocks. Specifically, we use the coordinates of the centroid of each county to compute the bilateral distance between counties. We define a distance cutoff such that there are on average five counties in spatial clusters; the cutoff is 56 kms in our baseline analysis.[5] For each county, we compute the share of neighboring counties within its spatial cluster that are affected by the policy shocks. We define the spatially correlated policy shock as the sum of the share of neighboring counties that are affected and the dummy variable indicating whether the county itself receives an idiosyncratic placebo policy shock. Therefore, the spatially correlated policy shocks are the sum of the idiosyncratic policy shocks and the policy shocks that are shared by all counties located within an arbitrary spatial cluster.

[5] Note that we adopt a uniform spatial decay kernel in our simulations. We have explored Bartlett-type kernels as well and find that results are fairly comparable to those we present here.

Panel (b) of Figure 1 visualizes the distribution of policy shocks that are spatially correlated across counties within arbitrary spatial clusters. The distribution of policy shocks in panel (a) is idiosyncratic and does not follow any spatial pattern, whereas that in panel (b) is marked with spatial correlation across counties that are in close proximity.

Figure 2 visualizes the distribution of log median earnings in 2000 at the county level, i.e., the outcome variable in equation 1. The distribution of the log median earnings exhibits a degree of spatial correlation across counties, i.e., spatial clustering of counties with high and low values of earnings, that are nearby. To quantify how introduction of spatial correlation to the main variable of interest affects the null-rejection rates (in the presence of spatial correlation in the distribution of the outcome variable), we replace the idiosyncratic placebo policy shock in equation 1, $Policy_c$, with the spatially correlated policy shock, $PolicySC_c$. We estimate the following equation using OLS both correcting and not correcting for the presence of spatial correlation across counties within arbitrary spatial clusters:

$$Y_c = \alpha_{2c} + \beta_2 PolicySC_c + \delta_2 X_c' + \varepsilon_c \tag{2}$$

We expect the null hypothesis that $\beta_2 = 0$ will be rejected more than 5% of the time at the 95% confidence interval if spatial correlation in the model is not accounted for and the null-rejection rate to approach 5% when spatial correlation is accounted for as the number of observations approaches infinity for a sufficiently large number of Monte Carlo draws.

ENDOGENEITY. We introduce endogeneity in the model by generating a random policy shock that is correlated with the outcome variable. We do so by forcing the counties that receive a placebo policy shock to be among a sample of counties that are above the median in terms of log median earnings in 2000. To select the counties that receive an endogenous policy shock, we rely on the same random variable used to select counties that receive an exogenous placebo policy shock. We select, among counties that are above the median in terms of log median earnings, those that are in the top half of the distribution of this random variable as counties that receive an endogenous policy shock.[6] Panel (a) of Figure 3 visualizes an example of the distribution of the endogenous policy shock variable we draw at random. This endogenous random policy shock, by construction, is correlated with the county-level distribution of log median earnings in 2000 depicted in Figure 2.

To introduce endogeneity to the model, we replace the exogenous placebo random shock in equation 1, $Policy_c$, with the endogenous random policy shock, $PolicyEnd_c$, and estimate the

[6] Both our exogenous and endogenous random policy shocks take the value of 1 for 25% of the counties and 0 for the rest.

following equation:

$$Y_c = \alpha_{3c} + \beta_3 PolicyEnd_c + \delta_3 X_c' + \mu_c \qquad (second-stage) \quad (3)$$

where Y_c and X_c are defined as in equation 1. $PolicyEnd_c$ is a dummy variable indicating whether county c receives an endogenous placebo policy shock and $Corr(PolicyEnd_c, \mu_c) \neq 0$. We instrument the endogenous random policy variable, $PolicyEnd_c$, with the exogenous random variable $Policy_c$. Given that $Cor(Policy_c, \mu_c) = 0$, the instrumental variable $Policy$ has an impact on Y only through its impact on $PolicyEnd$. We estimate the following first-stage equation:

$$PolicyEnd_c = \alpha_{4c} + \beta_4 Policy_c + \delta_4 X_c' + \omega_c \qquad (first-stage) \quad (4)$$

We expect the null hypothesis that $\beta_3 = 0$ will be rejected 5% of the time at the 95% confidence interval if it is estimated with 2SLS as the number of Monte Carlo draws approaches infinity.

We introduce spatial correlation to the 2SLS model by generating a spatially correlated endogenous policy shock, $PolicyEndS$, in the same way we generate spatially correlated exogenous policy shocks. Then, we estimate the following sets of equations with 2SLS both correcting and not correcting for the fact that the dependent variable and regressor are spatially correlated across observations within arbitrary spatial clusters:

$$Y_c = \alpha_{5c} + \beta_5 PolicyEndSC_c + \delta_5 X_c' + \mu_c \qquad (second-stage) \quad (5)$$

$$PolicyEndSC_c = \alpha_{6c} + \beta_6 PolicySC_c + \delta_6 X_c' + \omega_c \qquad (first-stage) \quad (6)$$

We expect the null hypothesis that $\beta_5 = 0$ will be rejected more than 5% of the time at the 95% confidence interval if spatial correlation in the model is not accounted for even if it is estimated with 2SLS. Moreover, we expect the null-rejection rate to approach 5% if spatial correlation is accounted for while estimated with 2SLS, as number of observations approaches infinity for a sufficiently large number of Monte Carlo draws.

3.1.2 Results

In this subsection, we describe the Monte Carlo simulations we performed and report the results we obtained. We consider four scenarios:

1) a spatial setting without endogeneity or spatial correlation (equation 1) ;

2) a spatial setting with endogeneity and without spatial correlation (equation 3);

3) a spatial setting without endogeneity but with spatial correlation (equation 2);

4) a spatial setting with endogeneity and spatial correlation (equation 5).

We perform Monte Carlo simulations with 10,000 iterations. In each Monte Carlo draw, we generate random policy variables for these four scenarios as described in subsection 3.1.1. As a benchmark, we estimate equation 1 (scenario 1) using OLS and equation 3 (scenario 3) using 2SLS with heteroskedasticity-robust standard errors. Then, we estimate equation 2 (scenario 3) and equation 5 (scenario 4) without taking into account spatial correlation using OLS and 2SLS, respectively. Finally, we estimate equation 2 (scenario 3) and equation 5 (scenario 4) using the estimator we propose that corrects for arbitrary within-cluster correlation in both OLS and 2SLS settings.

Table 1 displays the simulation results. Each cell reports, for a different scenario-estimation pair, the average null-rejection rate for the randomly generated policy variables over 10,000 draws. We start, in column 1, with the simulation results obtained using the full sample of counties (N=3,141). We expect to reject the null hypothesis 5% of the time at the 95% confidence interval in the absence of spatial correlation. Consistent with the theoretical prediction, the null-rejection rate for the policy variables without spatial correlation is 5.2% for OLS estimates without endogeneity and 5.1% for 2SLS estimates with endogeneity at the 95% confidence interval (lines (1) and (2)). Then, we impose spatial correlation within spatial clusters. If we do not correct for it, the null-rejection rate increases to 9.1% and 9.0% in the case of OLS estimates without endogeneity and that of 2SLS estimates with endogeneity, respectively (lines (3) and (4)). Clustering the standard errors at the state level performs better than heteroskedasticity-robust standard errors, reducing the null-rejection rates to 6.8% in OLS and 6.6% in 2SLS with endogeneity (lines (5) and (6)). Many of the counties that are in the same spatial cluster are also in the same state; therefore, clustering at the state level approximates the existing spatial correlation structure to a certain extent. Finally, we correct for the presence of spatial correlation across counties using our `acreg` estimator.[7] We obtain a null-rejection rate of 5.5% in OLS and 5.3% in 2SLS with endogeneity, very close to the theoretical prediction of 5% (lines (7) and (8)).

Next, we replicate the analysis by splitting the sample into two in terms of whether counties in a given spatial cluster are all in the same state (within-state clusters in column 2S) or whether they cross state boundaries (cross-state clusters in column 3). Our proposed estimator

[7] Our estimator requires as input either a distance cutoff value or an adjacency matrix showing which observations are within the same spatial clusters. While correcting for spatial correlation across counties within arbitrary clusters, we use as input the distance cutoff that we use while generating spatial correlation across counties, which ensures that there are five observations in each spatial cluster on average.

performs equally well in both samples, producing null-rejection rates of approximately 5.5% to 5.8% for the spatially correlated random shocks. By contrast, there is a substantial difference in the null-rejection rates we obtain from clustering the standard errors at the state level across two samples. Null-rejection rates are lower if all the counties in a given spatial cluster are in the same state (within-state clusters) compared to the case where they cross a state boundary (cross-state clusters). This is expected as clustering at the state level in the former case approximates the existing spatial correlation structure much better than it does in the latter case. However, even in a sample of counties that are a part of within-state clusters, our proposed estimator performs better than clustering at the state level. As seen in Figure 2, the outcome variable is not uniformly correlated across all counties within the same state. The correlation is greater across counties that are closer to one another within the same state. Therefore, taking into account the physical distance between counties (spatial units) performs much better than treating all units within the same state (the greater administrative unit).

We next assess whether the arbitrary clustering estimator is affected by sample size. To do so, in each Monte Carlo simulation we keep the n largest counties in each state (excluding Washington D.C.) in the sample, where $n = \{3, 4, ..., 20\}$, and the sample size is approximately 50 states times n selected counties. We introduce spatial correlation to the model as before. Figure 4 presents the null-rejection rates in the presence of spatial correlation by sample size. Panel (a) and (b) present the results for the OLS and 2SLS settings, respectively. Each connected point in the figures represents the average null-rejection rates for a different scenario-estimation pair separately $\forall n = \{3, 4, ..., 20\}$.[8] The performance of our proposed estimator improves as the sample size increases and the null-rejection rate converges toward the theoretical benchmark 5%.

Kelly (2019) recently argued that spatial studies suffer from strong inference problems using artificial data. Figure A.1 in Appendix presents the null-rejection rates we obtain when we regress a randomly generated variable on another randomly generated variable as done in Kelly (2019).[9] Regressing a spatially correlated random outcome variable on a spatially correlated random variable of interest leads to average null-rejection rates of approximately 40% – much larger than what we observe when we regress an observed outcome variable on a randomly generated

[8] To ensure that the first stage has enough predictive power even in the case of small sample sizes, we run 10,000 Monte Carlo simulations in each iteration but report the average null-rejection rates for the top half of the Monte Carlo draws in terms of F-statistics of the first stage.

[9] We generate random variables, Y and X, that are independent and identically distributed (iid): $Y, X = \sim N(0, 1)$. To introduce spatial correlation to these variables, we impose a Bartlett kernel decay across observations within the same cluster. In other words, we spread the random variables across observations within the cluster as an inverse function of the distance between them. Then, we sum them up. Formally: $Y_{i,sc} = \sum_{j \neq i}^{N} [1 - (dist_{ij}/distcut)] \times Y_j$ and $X_{i,sc} = \sum_{j \neq i}^{N} [1 - (dist_{ij}/distcut)] \times X_j$, where N is the number of observations in the cluster of observation i, $dist_{ij}$ is the distance between observations i and j, and $distcut$ is the distance cuttoff. To introduce endogeneity to the model, we define an endogenous variable, End, as a function of Y, X, and IV. IV is a random variable and iid to Y and X: $IV = \sim N(0, 1)$. Then, we instrument End with IV.

policy variable. However, our proposed estimator reduces the null-rejection rates by taking into account the spatial correlation across observations within the same cluster. As is the case with observed data, the performance of our proposed estimator improves as the sample size increases and the null-rejection rate converges toward the theoretical benchmark 5%. For a sample size of 3,141 counties, our proposed estimator reduces the null-rejection rates in the presence of spatial correlation from 27.4% (heteroskedasticity-robust standard errors) to 6.5% in the OLS setting and from 27.1% to 5.4% in the 2SLS setting.

3.1.3 Understanding Spatial Correlation: A Practitioner's Guide

We conduct further Monte Carlo simulations to shed light on the way that spatial correlation affects the likelihood of Type 1 error. First, we focus on whether presence or lack of spatial correlation in the outcome variable affects the null-rejection rates. Then, we investigate how presence of controls affects them. Last, we document how to set the optimal correction threshold.

Spatial correlation in the outcome variable. The results presented previously show that in the absence of spatial correlation in the treatment variable, $Policy$, the null-rejection rates are close to the theoretical 5% despite the presence of spatial correlation in the outcome variable. This implies that the presence of spatial correlation leads to an increase in the likelihood of making a Type 1 error if unaccounted for, only if both the outcome variable and the variable of interest exhibit spatial autocorrelation. We test whether this is the case by randomizing the outcome variable by reshuffling observed log median incomes across counties without imposing any restriction.

Table 2 presents the average null-rejection rates obtained from 10,000 Monte Carlo simulations with the data generating process as described in section 3.1.1. Column 1 presents the baseline results obtained using observed log median income as the outcome variable, whereas column 2 presents those obtained using the randomized outcome variable, i.e., reshuffled across counties. In the absence of spatial correlation in the outcome variable, neither introduction of spatial correlation in the treatment variable nor correction for it considerably affects the null-rejection rates, which remain in the vicinity of the theoretical 5%. Next, we reintroduce spatial correlation to the outcome variable. We take the randomized outcome variable and impose a Bartlett kernel decay across observations within the same cluster. In other words, we spread the randomized outcome variable across observations within the cluster as an inverse function of the distance between them. Column 3 shows that reintroduction of spatial correlation to the outcome variable leads to an increase in the null-rejection rates when the treatment variable also exhibits spatial correlation.

Our findings indicate that spatial correlation has to be present in both the outcome variable

and variable of interest for an increase in the likelihood of Type 1 error if spatial correlation in the model is not accounted for. This implies that presence of spatial correlation in residuals of a model is not enough to identify whether the model would suffer from an inflation of t-statistics. This insight contradicts the procedure proposed by Kelly (2019).

Presence of controls. Next, in Table 3, we investigate how inclusion of covariates in the model affects null-rejection rates. In column 1, as before, we present the baseline null-rejection rates. Column 2 shows that the average null-rejection rate obtained when spatial correlation is not corrected for increases from 9.1% to 12.8% in the OLS setting (and from 9.0% to 12.7% in the 2SLS setting) when we do not include covariates in the vector X_c presented in equation 1. Column 3 shows that when we control for state fixed effects in addition to the covariates in baseline specification, it decreases to 8.4% and 8.3% in the OLS and 2SLS settings, respectively. Our proposed estimator performs as well in both of these specifications as it does in the baseline specification. These results suggest that the magnitude of the inflation in the likelihood of Type 1 error due to the presence of spatial correlation in the model depends on the degree of spatial autocorrelation in the residual variation left in the outcome variable and variable of interest, conditional on the set of covariates.

Optimal correction threshold. Finally, we investigate the presence of the optimal correction threshold. This time, we define a distance cutoff such that there are on average 50 counties in spatial clusters, corresponding to 168 kilometers. Then, using our proposed estimator, we correct for the spatial correlation in the model using different distance thresholds, namely: 56 kms (one-third of the true threshold, 5 counties on average), 82 kms (~half of the threshold, 12 counties on average), 117 kms (~two-thirds of the threshold, 25 counties on average), 168 kms (the true threshold), 242 kms (~1.5 times the threshold, 100 counties on average), 327 kms (~twice the threshold, 175 counties on average), and 478 kms (~three times the threshold, 350 counties on average).

Figure 5 presents the average null-rejection rates obtained from using each of these different thresholds for error correction over 10,000 draws. Panel (a) considers the case of a single policy treatment. When spatial correlation is not corrected for, we obtain a null-rejection rate of 11.9%. Even correcting for spatial correlation using very small correction thresholds such as 56 kms and 82 kms or very large thresholds such as 478 kms performs better than using heteroskedasticity-robust standard errors, producing average null-rejection rates of 10.5%, 9.1%, and 9.1%, respectively. However, they perform worse than clustering the standard errors at the state level, which reduces the average null-rejection rate from 11.9% to 7.5%. Correcting for spatial correlation using correction thresholds 117 kms, 168 kms, 242 kms, and 327 kms yields average null-rejection rates of 7.5%, 5.9%, 6.3%, and 7.3%, respectively–all of which are equal to or below the average

null-rejection rejection rate we obtain from clustering the standard errors at the state level. Importantly, correcting for spatial correlation using the (true) threshold that matches the data generating process leads to the lowest null-rejection rate, close to the theoretical benchmark of 5%. Correcting for spatial correlation using thresholds both larger and smaller than the one matching the data generating process yields greater null-rejection rates, which suggests that there is an optimal distance threshold to use while correcting for spatial correlation in a model.

It is often the case that researchers are interested in the inference of more than one parameter in their model. What is the optimal distance threshold to use in the presence of two treatment variables? We investigate this question by performing the same analysis as above but including in the model a second random policy variable that is spatially correlated across counties within arbitrary clusters of 242 kilometers in radius. This simulation is interesting because the spatial neighbors for the second policy are not the same as that for the first policy, as is reasonable in real-life applications. Panel (b) of Figure 5 presents the average null-rejection rates in the case of two policy treatments. If we do not take into account the presence of spatial correlation, we estimate a null-rejection rate of 11.4% and 10.5% for the first and second policy treatments, respectively. For both policy treatments, we obtain the lowest null-rejection rates when we correct for spatial correlation using the distance threshold that matches the data generating process of each treatment, namely, 6.2% and 6.5%. This implies that there is no *universal* distance threshold that minimizes the likelihood of Type 1 error for all treatments (or covariates) in a model. Moreover, the difference in null-rejection rates between the two treatments when the spatial correlation is not accounted for suggests that the degree of inflation in t-statistics and the likelihood of Type 1 error depends on the joint distribution of the outcome variable and the variable of interest in question.[10]

Implications. As our simulations have shown, only the presence of spatial correlation in both the outcome variable and variable of interest in a model leads to a greater likelihood of Type 1 errors if such spatial correlations are not accounted for. Controlling for covariates (that have spatial dimension) and clustering standard errors at a greater administrative unit can help with addressing the inflation in t-statistics due to spatial correlation in the model. However, a better approach is to explicitly model the spatial correlation structure in the model with our proposed estimator. When deciding on how to model the spatial correlation in your model, as Cameron and Miller (2015) put it: "You need to think carefully about the potential for correlations in your model errors, and how that interacts with correlations in your covariates." Our Monte Carlo simulations in a controlled environment suggest that an optimal correction threshold exists for each

[10] the t-statistic for the mean difference between the null-rejection rates with heteroskedasticity-robust standard errors (11.4% – 10.5% = 0) is 1.92 (p-value = 0.0547). The t-statistic for the mean difference between the null-rejection rates without robust standard errors (13.7% – 11.7% = 0) is 3.99 (p-value = 0.0001).

parameter.

In practice, however, it is possible that the correlation structure in the data cannot be approximated by spatial clusters defined as circles with a given radius. For example, topographic features such as mountain ranges could generate variations in the distribution of the outcome variable and covariates across spatial units that are in close proximity in terms of Euclidean distance. To help address this issue, our proposed estimator's companion statistical package (`acreg`) allows users to provide a bilateral-distance matrix of any metric between observations. Then, the distance threshold used for error correction can be defined as *effective distance* between observations in terms of time or cost of travel (flight, road, or walking) distance. Moreover, for a given model and variable of interest, the optimal distance threshold could vary depending on the outcome of interest and its spatial distribution.

The existing measures of spatial correlation, i.e., Moran's I and Geary's C, allow researchers to test the existence of spatial autocorrelation in a variable. Therefore, they can be used to identify whether an outcome variable and a variable of interest are spatially autocorrelated and thus whether inference for a given outcome of interest in a model is likely to suffer from inflation of t-statistics. However, they fall short on providing insights on the optimal threshold for error correction, as they do not provide any metrics on the joint spatial distribution of two variables (or of the residuals left in them conditional on controls) and on the degree of spatial correlation between them in a given sample. As a result, no clear-cut procedure currently exists to define the *potential* optimal threshold using observational data.

We suggest that researchers correct standard errors with varying distance thresholds (and potentially using different distance metrics) and select as the baseline the threshold that provides the largest standard errors for a given model. In the presence of multiple outcomes of interest, we advise selecting a correction threshold that provides the largest standard errors for most of the variables of interest as the baseline. Overall, we recommend that researchers, as a healthy practice, be transparent about their choice of baseline distance threshold and report the robustness of their findings to correcting the standard errors in their models using a wide range of distance thresholds.

3.2 NETWORK

To illustrate how correlation across units linked in networks leads to an overrejection of the null hypothesis if such correlation is not accounted for, we use data on coauthorship networks. We extract information on coauthorship links between researchers from IDEAS RePEc. We identify the researchers with the highest number of coauthors. Then, for each of these authors, we collect data on their research profile, i.e., the number of articles they have indexed on IDEAS RePEc

and the total number of citations these articles have received (from Google citations), and complement them with basic demographic information such as age and gender obtained from their CVs.

3.2.1 The Data Generating Process and Hypotheses

To quantify the problem induced by correlation within networks, we randomly generate productivity shocks that affect some authors and not others in each Monte Carlo draw. Similar to the approach we employ in the spatial setting, we first generate a random variable from a normal distribution. Then, we select the authors who are in the top quartile of the distribution of this random variable as authors that receive a "placebo" productivity shock. Panel (a) of Figure 6 visualizes our coauthorship network links and an example of the distribution of the productivity shock variable we draw at random.

We estimate the following equation using OLS:

$$Y_a = \alpha_{1a} + \beta_1 Productivity_a + \delta_1 X_a' + \epsilon_a 74 \tag{7}$$

where Y_a is the log number of citations author a receives. $Productivity_a$ is a dummy variable indicating whether author a receives a productivity shock. X_a is a vector of author-level controls, which comprises the log number of articles they have authored, their gender, their age and its squared value. Given that Y and $Productivity$ are independent and $Cor(Productivity_a, \epsilon_a) = 0$, statistical theory predicts that the null hypothesis that $\beta_1 = 0$ will be rejected 5% of the time at the 95% confidence interval as the number of Monte Carlo draws approaches infinity.

Next, we impose within-network correlation to the productivity shocks we draw at random. For each author, we compute the share of their first degree coauthors who are affected by the productivity shocks.[11] We define productivity shocks that are correlated within coauthorship networks as the sum of this share and a dummy variable indicating whether the author herself receives an idiosyncratic placebo productivity shock. Therefore, the productivity shocks that are correlated within coauthorship networks are the sum of the idiosyncratic productivity shocks and the productivity shocks that are shared by all authors who are part of a coauthorship network. Panel (b) of Figure 6 visualizes the distribution of productivity shocks that are correlated within coauthorship networks.

To quantify how introduction of within-network correlation to the main variable of interest affects the null-rejection rates, we replace the idiosyncratic placebo productivity shock in equation 7, $Productivity_a$, with the productivity shock that is correlated within coauthorship

[11] We adopt a setting where shocks are correlated in coauthor neighborhoods of degree 1. Larger neighborhoods and decay in shocks can be accommodated in our estimator as well.

networks, $ProductivityNC_a$. We estimate the following equation using OLS both correcting and not correcting for the presence of within-network correlation across coauthors:

$$Y_a = \alpha_{2a} + \beta_2 ProductivityNC_a + \delta_2 X_a' + \varepsilon_a \tag{8}$$

We expect the null hypothesis that $\beta_2 = 0$ will be rejected more than 5% of the time at the 95% confidence interval if within-network correlation in the model is not accounted for and the null-rejection rate to approach 5% when within-network correlation is accounted for, as number of observations approaches infinity for a sufficiently large number of Monte Carlo draws.

ENDOGENEITY. We introduce endogeneity to the model by generating a random productivity shock that is correlated with the outcome variable. We do so by forcing the authors who receive a placebo productivity shock to be among a sample of authors who are above median in terms of log number of citations. To select the authors that receive an endogenous productivity shock, we use the same random variable from a normal distribution that we use to select authors who receive an exogenous placebo productivity shock. We select among authors who are above the median in terms of the log number of citations earnings, those who are in the top half of the distribution of this random variable as authors who receive an endogenous productivity shock.[12] By construction, this endogenous random productivity shock is correlated with the distribution of the log number of citations.

To introduce endogeneity to the model, we replace the exogenous random productivity shock in equation 7, $Policy_c$, with the endogenous random productivity shock, $PolicyEnd_c$, and estimate the following equation:

$$Y_a = \alpha_{3a} + \beta_3 ProductivityEnd_a + \delta_3 X_a' + \mu_a \qquad (second-stage) \tag{9}$$

where Y_a and X_a are defined as in equation 7. $ProductivityEnd_a$ is a dummy variable indicating whether author a receives an endogenous placebo productivity shock and $Corr(ProductivityEnd_a, \mu_a) \neq 0$. We instrument the endogenous random productivity variable, $ProductivityEnd_a$, with the exogenous random variable $Productivity_a$. Given that $Cor(Productivity_a, \mu_a) = 0$, the instrumental variable $Productivity$ has an impact on Y only through its impact on $ProductivityEnd$. We estimate the following first-stage equation:

$$ProductivityEnd_a = \alpha_{4a} + \beta_4 Productivity_a + \delta_4 X_a' + \omega_a \quad (first-stage) \tag{10}$$

[12] Both our exogenous and endogenous random productivity shocks take the value of 1 for 25% of the counties and 0 for the rest.

We expect the null hypothesis that $\beta_3 = 0$ will be rejected 5% of the time at the 95% confidence interval if it is estimated with 2SLS as the number of Monte Carlo draws approaches infinity.

We introduce spatial correlation to the 2SLS model by generating an endogenous productivity shock that is correlated within coauthorship networks, $ProductivityEnd$, in the same way we generate an exogenous productivity that is correlated within coauthorship networks. Then, we estimate the following sets of equations with 2SLS both correcting and not correcting for the fact that the dependent variable and regressor are correlated across authors that are in a coauthorship relationship:

$$Y_a = \alpha_{5a} + \beta_5 ProductivityEndNC_a + \delta_5 X_a' + \mu_a \qquad (second-stage) \qquad (11)$$

$$ProductivityEndNC_c = \alpha_{6a} + \beta_6 ProductivityNC_a + \delta_6 X_a' + \omega_a \qquad (first-stage) \qquad (12)$$

We expect the null hypothesis that $\beta_5 = 0$ will be rejected more than 5% of the time at the 95% confidence interval if within-network correlation in the model is not accounted for even if it is estimated with 2SLS. Moreover, we expect the null-rejection rate to approach 5% if within-network correlation is accounted for while estimated with 2SLS, as number of observations approaches infinity for a sufficiently large number of Monte Carlo draws.

3.2.2 Results

In this subsection, we briefly describe the Monte Carlo simulations and then discuss the results. As was the case with the spatial setting, we consider four scenarios:

1) a network setting without endogeneity or within-network correlation (equation 7) ;
2) a network setting with endogeneity and without within-network correlation (equation 9);
3) a network setting without endogeneity but with within-network correlation (equation 8);
4) a network setting with endogeneity and within-network correlation (equation 11).

We perform Monte Carlo simulations with 10,000 iterations. In each Monte Carlo draw, we generate random productivity shock variables for these four scenarios as described in subsection 3.2.1. As the benchmark, we estimate equation 7 (scenario 1) using OLS and equation 9 (scenario 3) using 2SLS with heteroskedasticity-robust standard errors. Then, we estimate equation 8 (scenario 3) and equation 11 (scenario 4) without taking into account spatial correlation using OLS and 2SLS, respectively. Finally, we estimate equation 8 (scenario 3) and equation 11 (scenario 4) using the estimator we propose that corrects for arbitrary within-cluster correlation in both OLS and 2SLS settings.

Figure 7 presents the null-rejection rates in the presence of within-network correlation by sample size.[13] Panel (a) and (b) present the results for the OLS and 2SLS settings, respectively. Each connected point in the figures represents the average null-rejection rates for a different scenario-estimation pair and a sample size separately.[14] In the absence of within-network correlations, we can reject the null hypothesis that random productivity shocks have a statistically significant effect on the log number of citations at the 5% level approximately 5% of the time, consistent with the predictions of the statistical theory.

However, when we impose the random productivity shocks to be correlated within coauthorship networks, we obtain average null-rejection rates of approximately 8% if we do not take into account the existence of within-network correlations. Our proposed estimator that takes into account the network links to correct for standard errors significantly reduces the null-rejection rates. Its performance improves as the sample size increases and the null-rejection rate converges toward the theoretical benchmark 5%. For a sample size of 1,000 observations, we obtain average null-rejection rates of 5.5% in the OLS setting and 6.2% in the 2SLS setting.

4 CONCLUSION

We implement a procedural approach to obtain an asymptotically valid inference in settings with spatial or network topology, allowing for any type of dependence between observation units. Our proposed variance-covariance matrix (VCV) estimator, accompanied by a companion statistical package `acreg` for Stata, allows researchers to obtain cluster-robust inference in OLS and 2SLS settings with arbitrary dependence across observations and over time. Arbitrary here refers to the way units could be correlated with each other in space and time. Our approach allows units to be correlated with each other in any possible way: the estimator can account for indirect links in the cross-sectional dependence, time dependence and alteration in the correlation structure over time. This allows our estimator to be suitable for many applications. Choosing the right spatial bandwidth or distance in the network is a central practical challenge. We provide simulation results that suggest that inference is reasonably precise when standard errors are largest. While not conclusive, these simulations can offer some practical guidance for implementation. We also discuss other implementation issues such as including control variables or multiple spatially correlated regressors.

[13] To ensure that the first stage has enough predictive power even in the case of small sample sizes, we run 10,000 Monte Carlo simulations in each iteration but report the average null-rejection rates for the top half of the Monte Carlo draws in terms of F-statistics of the first stage.

[14] To ensure that the number of coauthors per author (links per node) and the network metrics are constant across different sample sizes, we start with a sample of 50 authors with the highest number of coauthors and increase the sample size by generating duplicates of observations in this sample.

REFERENCES

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?*. *The Quarterly Journal of Economics*, **119**(1), 249–275.

Cameron, A., Gelbach, J., and Miller, D. (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, **29**(2), 238–249.

Cameron, C. A. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50**(2), 317–372.

Conley, T. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, **92**(1), 1–45.

Kelejian, H. H. and Prucha, I. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, **17**(1), 99–121.

Kelejian, H. H. and Prucha, I. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, **40**(2), 509–33.

Kelly, M. (2019). The standard errors of persistence. *CEPR Discussion Paper Series 13783*.

Manson, S., Schroeder, J., Van Riper, D., and Ruggles, S. (2017). Ipums national historical geographic information system: Version 12.0 [database]. Minneapolis: University of Minnesota. http://doi.org/10.18128/D050.V12.0.

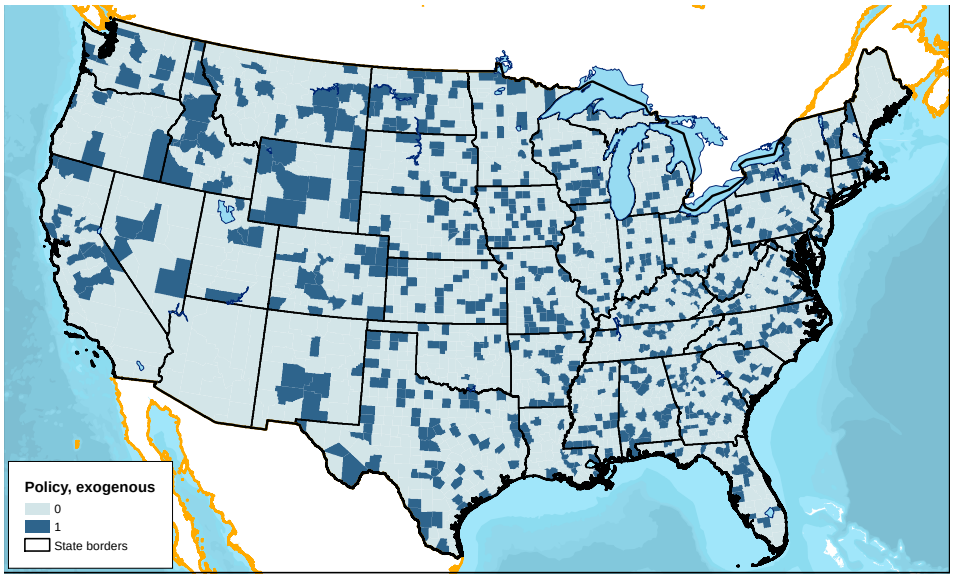
Michalopoulos, S. and Papaioannou, E. (2017). Spatial patterns of development: A meso approach. NBER Working Papers 24088, National Bureau of Economic Research, Inc.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**(4), 817–38.

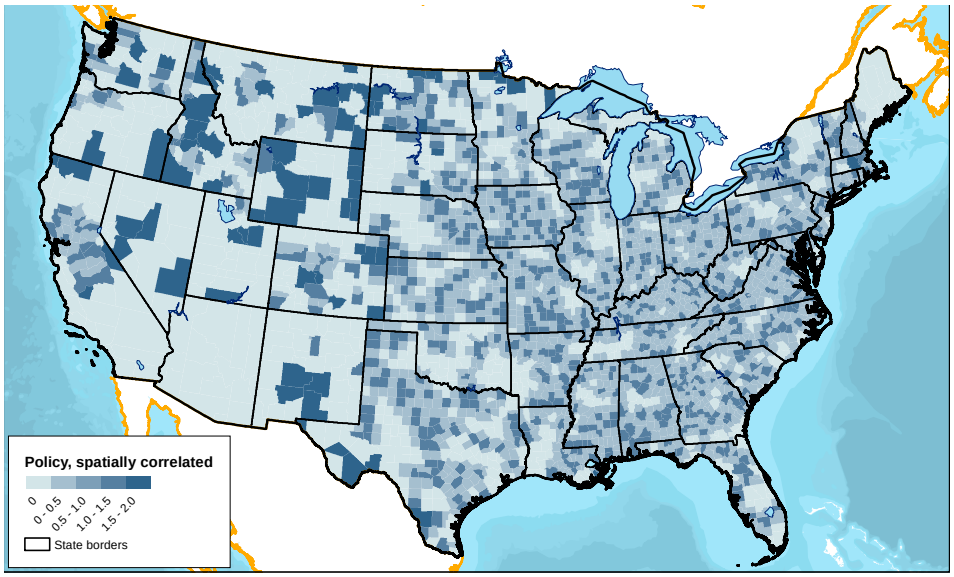
White, H. (1984). *Asymptotic Theory for Econometricians*.

Figure 1: Illustration of data generation process: exogenous shocks in U.S. counties

(a) Random policy shock

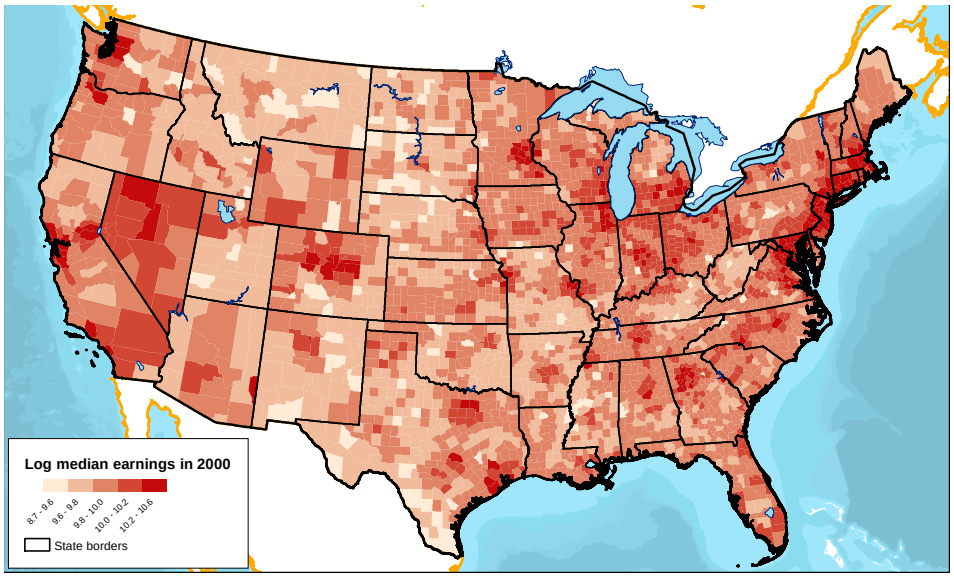


(b) Spatially correlated policy shock



Notes: Data source for the county boundaries: NHGIS (Manson *et al.*, 2017). The values of the exogenous policy shocks represented are randomly generated with the algorithm described in section 3.1.1.

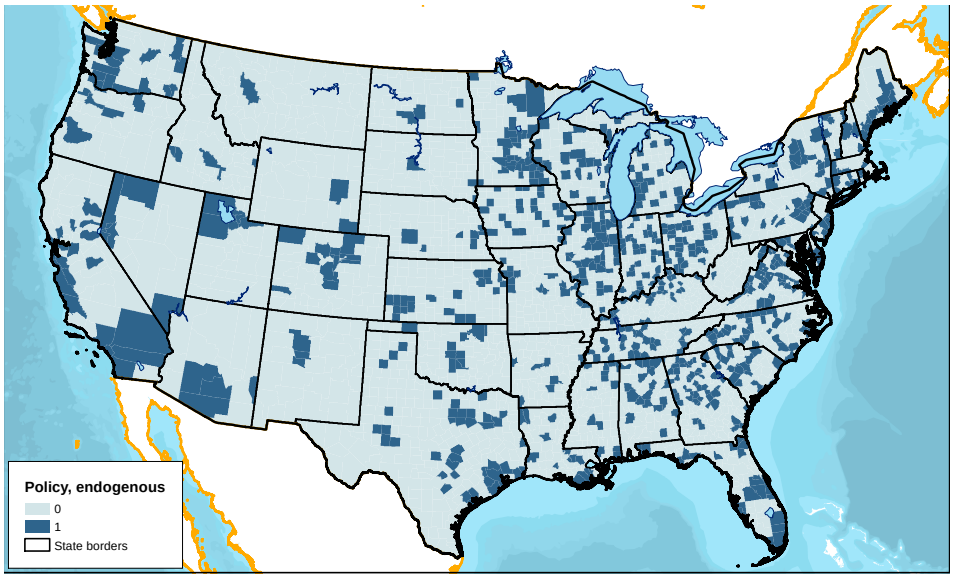
Figure 2: Log median earnings across US counties in 2000



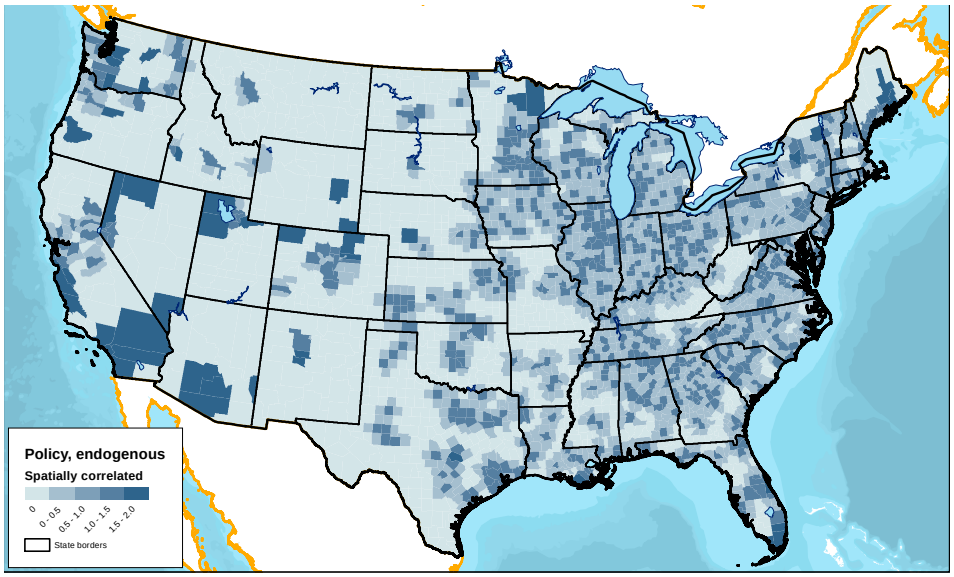
Notes: Data source for the county boundaries and log median earnings in 2000: NHGIS (Manson *et al.*, 2017).

Figure 3: Illustration of data generation process: endogenous shocks in U.S. counties

(a) Random endogenous policy shock



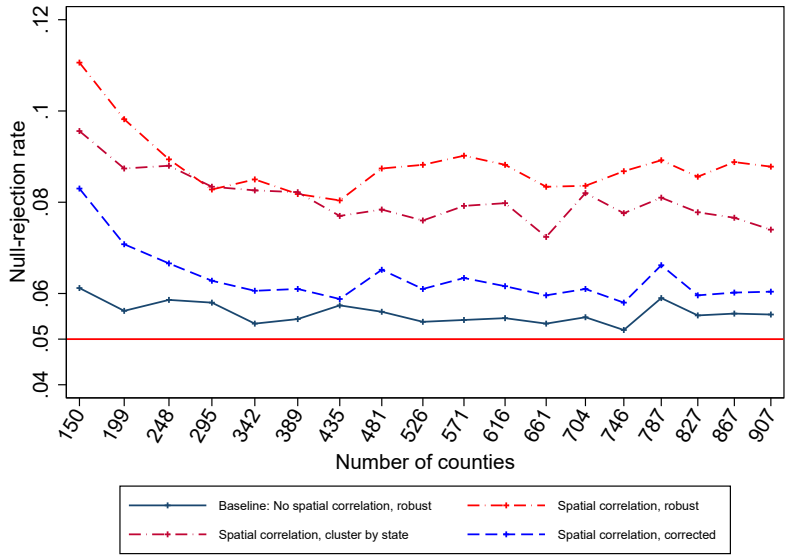
(b) Spatially correlated endogenous policy shock



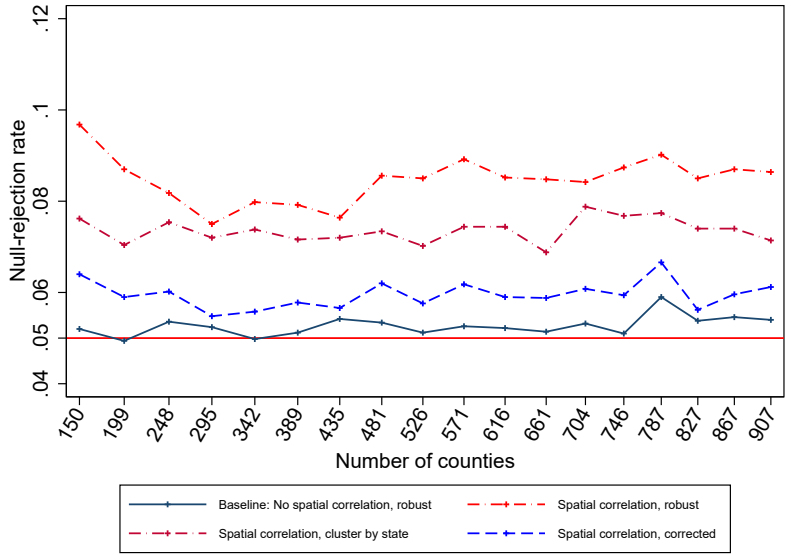
Notes: Data source for the county boundaries: NHGIS (Manson *et al.*, 2017). The values of the endogenous policy shocks represented are randomly generated with the algorithm described in section 3.1.1.

Figure 4: Null-rejection rate in the presence of spatial correlation: U.S. counties

(a) OLS



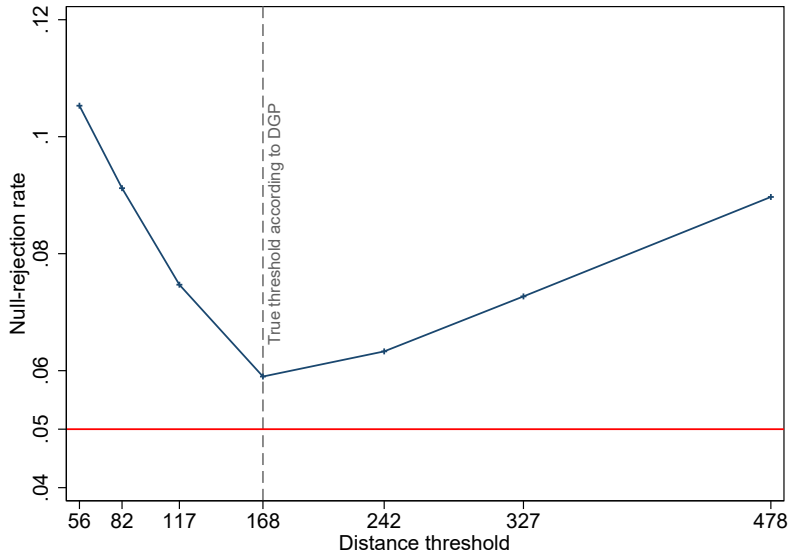
(b) 2SLS



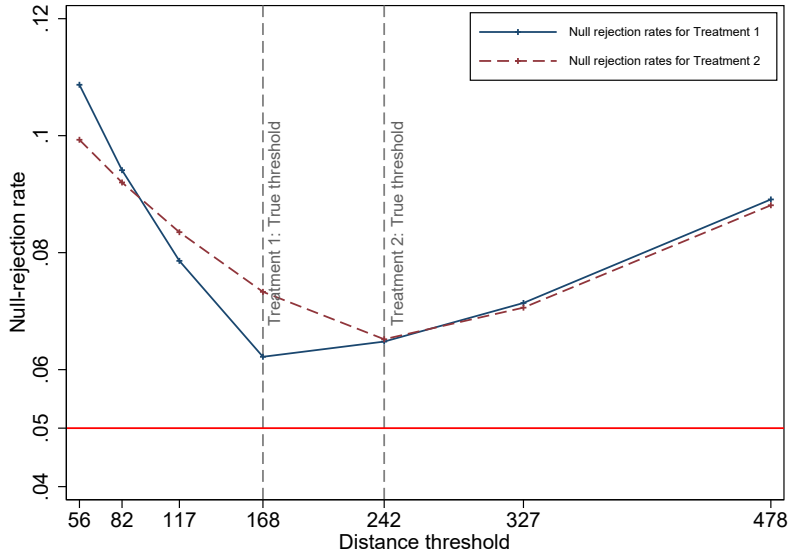
Notes: The red vertical line represents the benchmark null-rejection rate of 5%. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations. Each point in the figure represents a different Monte Carlo simulation × estimation pair. The horizontal axis represents the sample size.

Figure 5: Spatial setting: Optimal distance threshold and null-rejection rates

(a) Single policy treatment



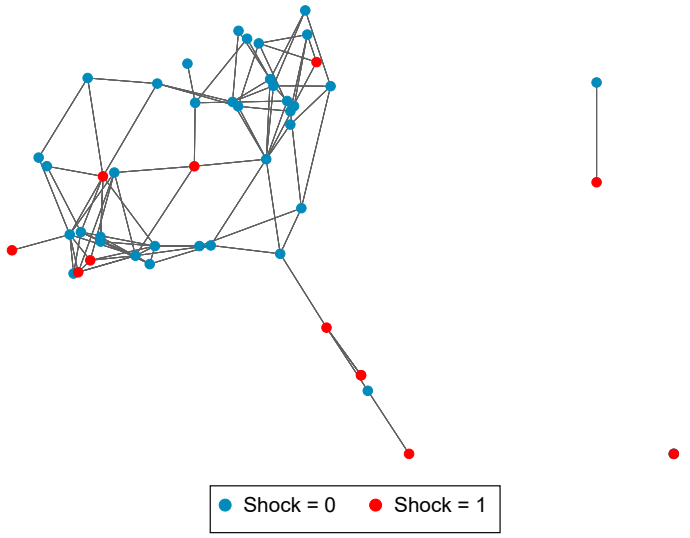
(b) Multiple policy treatments



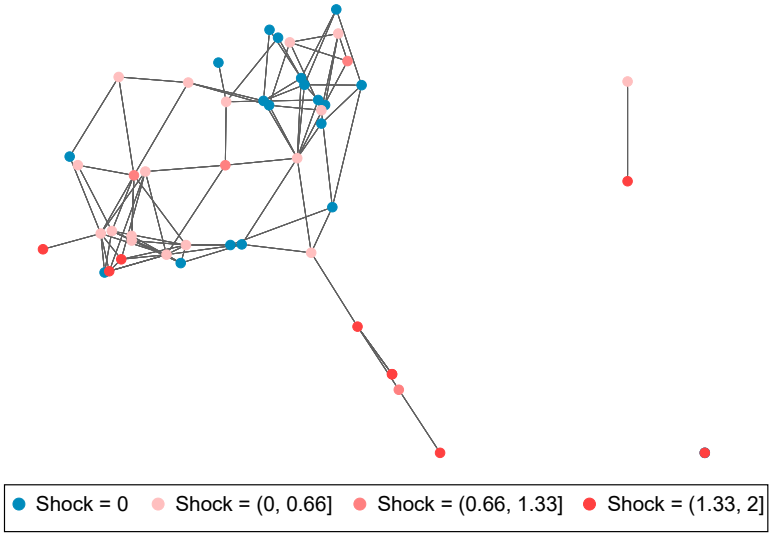
Notes: The red vertical line represents the benchmark null-rejection rate of %5. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations on a sample of 3,141 counties. Each point in the figure represents a different Monte Carlo simulation uses a different distance threshold used for error correction.

Figure 6: Illustration of data generation process: exogenous shocks in coauthorship networks

(a) Productivity shocks



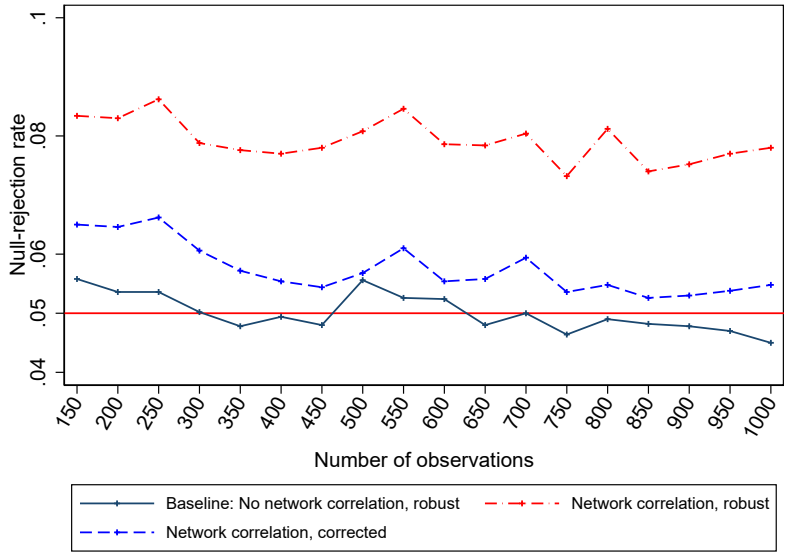
(b) Productivity shocks with within-network correlation



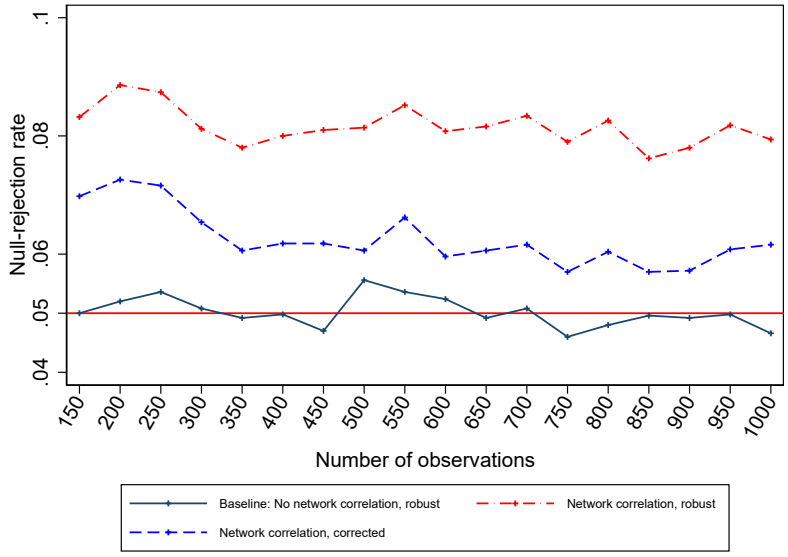
Notes: The figure maps the coauthorship links between authors. The sample consists of the top 50 researchers listed in the IDEAS RePEc in terms of the number of coauthors. The values of the exogenous productivity shocks represented are randomly generated with the algorithm described in section 3.2.1.

Figure 7: Null-rejection rate in the presence of within-network correlation: Top-cited authors

(a) OLS



(b) 2SLS



Notes: The red vertical line represents the benchmark null-rejection rate of 5%. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations. Each point in the figure represents a different Monte Carlo simulation × estimation pair. The horizontal axis represents the sample size.

Table 1: Null-rejection rates in the spatial setting

Unit:					U.S. counties		
Sample:					All	Within-state	Cross-state
Sample size:					N=3,141	N=2,126	N=1,015
					(1)	(2)	(3)
	Spatial corr.	**Endogeneity**	**Estimator**	**Correction**	**Null-rejection rate**		
(1)			OLS	robust	5.2%	5.1%	5.1%
(2)		✓	2SLS	robust	5.1%	5.1%	5.0%
(3)	✓		OLS	robust	9.1%	8.2%	10.4%
(4)	✓	✓	2SLS	robust	9.0%	8.2%	10.1%
(5)	✓		OLS	cluster	6.8%	6.9%	9.2%
(6)	✓	✓	2SLS	cluster	6.6%	6.9%	8.8%
(7)	✓		OLS	acreg	5.5%	5.8%	5.6%
(8)	✓	✓	2SLS	acreg	5.3%	5.5%	5.6%

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Each column-row pair represents a different environment (data generating process and error correction) and sample pair. The data generating process simulates two different models: a baseline model without any spatial correlation in the policy treatment variable across units and another model imposing a spatial correlation in the policy treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is U.S. counties. The outcome variable is log median earnings.

Table 2: Null-rejection rates in the spatial setting: Presence of spatial correlation in the outcome variable

Unit:			U.S. counties, N=3,141		
Spatial correlation in the outcome:			Baseline observed	Random	Fake spatial correlation
			(1)	(2)	(3)
	Spatial corr.	**Correction**	**Null-rejection rate**		
(1)		robust	5.2%	5.5%	4.7%
(2)	✓	robust	9.1%	5.1%	8.8%
(3)	✓	cluster	6.8%	6.1%	6.2%
(4)	✓	acreg	5.5%	5.2%	5.1%

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Each column-row pair represents a different environment (data generating process and error correction) and different outcome. The outcome variable in column 1 is the observed log median earnings. In column 2, the outcome variable is the observed log median earnings randomly reshuffled across counties. In column 3, we impose spatial correlation to the randomly shuffled log median earnings used in column 2. The data generating process simulates two different models: a baseline model without any spatial correlation in the policy treatment variable across units and another model imposing a spatial correlation in the policy treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is U.S. counties.

Table 3: Null-rejection rates in the spatial setting: Controls

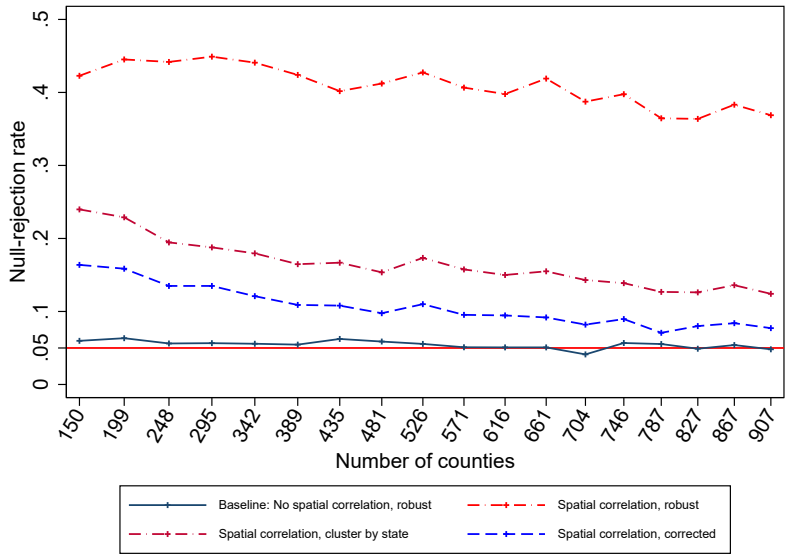
Unit:					U.S. counties, N=3,141		
Controls:					Baseline	No controls	State FEs
					(1)	(2)	(3)
	Spatial corr.	**Endogeneity**	**Estimator**	**Correction**	**Null-rejection rate**		
(1)			OLS	robust	5.2%	4.9%	5.1%
(2)		✓	2SLS	robust	5.1%	4.9%	5.0%
(3)	✓		OLS	robust	9.1%	12.8%	8.4%
(4)	✓	✓	2SLS	robust	9.0%	12.7%	8.3%
(5)	✓		OLS	cluster	6.8%	7.2%	6.2%
(6)	✓	✓	2SLS	cluster	6.6%	6.8%	5.8%
(7)	✓		OLS	acreg	5.5%	5.7%	5.6%
(8)	✓	✓	2SLS	acreg	5.3%	5.6%	5.5%

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Each column-row pair represents a different environment (data generating process and error correction) and sample pair. The data generating process simulates two different models: a baseline model without any spatial correlation in the policy treatment variable across units and another model imposing a spatial correlation in the policy treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is U.S. counties. The outcome variable is log median earnings.

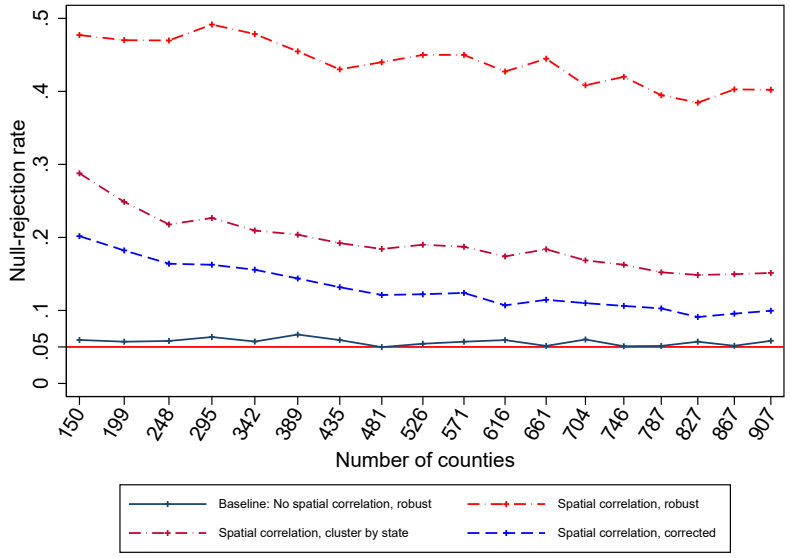
A APPENDIX

Figure A.1: Null-rejection rate in the presence of spatial correlation: U.S. counties with fake data

(a) OLS



(b) 2SLS



Notes: The red vertical line represents the benchmark null-rejection rate of 5%. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations. Each point in the figure represents a different Monte Carlo simulation × estimation pair. The horizontal axis represents the sample size.