Conley, Timothy G.; Kelly, Morgan

Working Paper

The standard errors of persistence

UCD Centre for Economic Research Working Paper Series, No. WP24/17

Provided in Cooperation with:
UCD School of Economics, University College Dublin (UCD)

Suggested Citation: Conley, Timothy G.; Kelly, Morgan (2024) : The standard errors of persistence, UCD Centre for Economic Research Working Paper Series, No. WP24/17, University College Dublin, UCD School of Economics, Dublin

This Version is available at:
https://hdl.handle.net/10419/306700

UCD CENTRE FOR ECONOMIC RESEARCH

WORKING PAPER SERIES

2024

The Standard Errors of Persistence

Timothy G. Conley
University of Western Ontario

Morgan Kelly
University College Dublin School of Economics

WP24/17

October 2024

UCD SCHOOL OF ECONOMICS
UNIVERSITY COLLEGE DUBLIN
BELFIELD
DUBLIN 4

The Standard Errors of Persistence

Timothy G. Conley and Morgan Kelly*

Abstract

Many studies of historical persistence find that modern outcomes strongly reflect characteristics of the same places in the distant past. However they rely on data that often exhibit extreme spatial trends and autocorrelation, suggesting that their unusually large t-statistics may be due to inadequately controlling for spurious correlation. To analyze this we introduce a new regression procedure and two diagnostic tests of no treatment effect: (a) a placebo test where the treatment is replaced with spatial noise and (b) a synthetic outcomes test of the hypothesis that the outcome is generated by a trend plus a spatial noise process independent of the treatment. We then show how reliable regression results can be obtained by adding a low dimensional spatial basis to the regression of interest, and applying a large cluster standard error correction. Examining 30 persistence studies in leading journals we find that few approach significance at conventional levels. Our procedure applies to regressions with spatial observations more generally and is implemented in an open source package.

1 Introduction

A substantial literature on deep origins or historical persistence finds that many modern outcomes such as income or social attitudes strongly reflect the characteristics of the same places in the more or less distant past, often centuries or millennia previously. Notable examples include showing how European mortality determines the quality of modern in-

*University of Western Ontario; and University College Dublin and CEPR. Conley gratefully acknowledges support from the Social Science and Humanities Research Council of Canada. The authors thank the editor, Alan Taylor, and two referees for constructive comments that greatly improved the paper. Hans Martinez Torres provided excellent research assistance. The R package for spatial inference is available at `https://github.com/morganwkelly/spatInfer`.

stitutions; how medieval pogroms prefigured Nazi zealotry; how the slave trade inhibits modern African development; and how colonial boundaries still drive poverty in Peru.[1]

Naturally, such findings are open to various charges of p hacking, of publication bias, of answers in search of questions, of ignoring mundane alternative explanations, of low quality data, of monocausal and largely atheoretical explanations of complex phenomena, of failures to explain what drives persistence, and so on. However, all of these objections crumble into irrelevance in the face of one blunt fact: the unusual explanatory power of these persistence variables. While a judicious choice of variables or time periods might coax a t statistic past 1.96, there would appear to be no way that the t statistics of four, five, or even larger that appear routinely in this literature could be the result of massaging regressions, no matter how assiduously. Such persistence results must instead reflect the workings of deep world historical processes: the enduring legacies of the past.

However, persistence regressions are spatial regressions—the values today of some variable in a given set of places are regressed on another variable for the same places in the past—and spatial data have two characteristics that make them prone to generate large t statistics. The first is the ubiquity of spatial trends and other forms of large scale spatial structure: Italian variables have strong north south trends, Europe enjoys better outcomes than Sub-Saharan Africa, and so on. As with time series, regressing two trending spatial series on one other will very often result in finding strong correlation, even when the variables are unrelated. Simply adding basic spatial trend controls in a systematic way—a quadratic in longitude and latitude for regional data, and indicators for World Bank regions in global ones—causes many of the most celebrated results in historical persistence to disappear immediately.

The second characteristic of spatial data is that they tend to show strong autocorrelation.[2] Neighbours resemble each other closely so that the number of useful observations is lower than it appears, sometimes far lower. As a result coefficient estimates can be highly imprecise which requires standard errors to be adjusted to get reliable inference. If no such adjustment is made and heteroskedasticity consistent or small cluster standard errors are used, large t-statistics are common even when the true effect size is negligible. In the next Section we illustrate this with synthetic regressions of two independent series on each

[1] These are, in turn, Acemoglu, Johnson and Robinson (2001), Voigtländer and Voth (2012), Nunn (2008), and Dell (2010). Review articles covering all aspects of historical persistence can be found in the volume edited by Bisin and Federico (2021).

[2] This is Tobler's First Law of Geography: "Everything is related to everything else, but near things are more related than distant things."

other, with each having empirically realistic spatial correlations. Using heteroskedasticity robust standard errors with no correction for spatial correlation, 38 percent of simulated treatment t-statistics are above 2. Twenty one percent are above 3, and eight percent above 4. The extreme t-statistics that are the hallmark of the persistence literature may be less a sign of incontrovertible historical truths than a warning of spurious results driven by spatial correlation.

In this paper we use simple models of spatial correlation to introduce powerful new simulation-based placebo and synthetic outcome diagnostic tests of no treatment effect. Standard placebo tests randomly shuffle treatments across subjects and cannot therefore be applied to spatial observations: assigning Sweden's treatment to Argentina, and Mali's to Brazil erases the spatial structure that defines the data. Instead, we generate synthetic noise placebo datasets that match the spatial structure of the treatment and are, by construction, independent of outcomes. The collection of t-statistics from these placebo regressions forms the reference distribution for a "randomization inference" style test of no treatment effect. Our simulations provide an accurate reference distribution, predicated on our assumption regarding the distribution functional form for the treatment variable.[3]

The placebo simulations provide, in addition, a simple Monte Carlo check on how well the regression's inference method deals with spatial correlation in the data. If the chosen standard error correction is appropriate then about five per cent of placebo regressions should be significant at five per cent. (For the original persistence regressions analyzed here, the median placebo rejection frequency is twenty seven percent.) These placebo Monte Carlos are especially useful, we will see below, for choosing the number of clusters for spatial standard error corrections.

Our second diagnostic is a synthetic outcome test of the hypothesis that there is no treatment effect and the outcome is generated by a trend plus spatially correlated noise. We replace the actual outcomes with spatial noise that has approximately the same distribution and run repeated simulations. The resulting collection of t-statistics serves as the reference distribution for a test of the hypothesis that the outcome is generated by a trend plus noise that is spatially correlated but independent of the treatment: there is no treatment effect.[4] Failure to reject the null suggests that the original finding of a signifi-

[3]Our maintained hypothesis is that we have a correctly specified model for the treatment variable so placebo test rejections could be driven by a mistake in this model. We use this placebo test as a diagnostic because of this potential for rejections due to its having more restrictive (maintained) assumptions relative to the original regression test of zero treatment slope.

[4]Note that this hypothesis test concerns a distribution rather than a parameter as is usually the case.

cant treatment effect may have resulted from mechanical prediction of outcome trends by treatment trends, or inadequate inference. A different measure of treatment or change in conditioning information may be required to sharpen the estimate of the treatment effect sufficiently to make it distinguishable from spatial noise.

To motivate our regression procedure we start by looking at what happens when we add simple spatial trend controls to 30 persistence regressions, while leaving their original standard error adjustment methods unchanged. For global studies the controls are dummies for World Bank regions, with quadratics in longitude and latitude for smaller scale ones. In 11 studies treatment slopes are no longer statistically different from zero at 5%, even with standard errors that may be too small. Placebo simulations indicate that in 14 of the remaining studies where slopes appear (at 5%) to be statistically significant, standard errors are too small: nominal 5% t-tests reject the correct null hypothesis more than 5% of the time, with 9 studies rejecting more than 10%. In other words, a failure to apply proper standard error corrections and controls for spatial trends can lead to seriously distorted regression results.

This leads us to introduce a new regression procedure that can reliably handle spatial trends in the data and autocorrelation in the residuals. A common practice in dealing with trends is to add polynomials of some arbitrary degree in longitude and latitude to the regression as a form of geographically correlated fixed effect. We systematize this procedure by including terms from a minimal tensor spline in coordinates that can flexibly capture a wide range of long range spatial structure in the data. Intuitively these splines are a mesh made up of products of, for example, four undulating threads that run north-south, and another four that run east-west.[5] In principle, we could add all the extra tensor variables directly to our regression but this would rapidly exhaust degrees of freedom so instead we use the first L principal components of these spline terms, where L is selected by a Bayes Information Criterion (BIC) penalty. We refer to these principal component regressors as spatial basis variables.[6]

These spatial basis variables serve to remove spatial trends and to reduce (but usually not to eliminate) spatial correlation in the residuals which improves the performance of standard error corrections. We then apply two complementary methods of large cluster spatial inference. Our first approach applied to all studies is that of Bester, Conley,

[5] A third time dimension could be added for panels; and locations could reflect the economic rather than geographical distances between observations.

[6] A formal analysis of the properties of spatial basis regressions is given by Conley, Kelly and Kozbur (2024).

and Hansen (2011) (BCH) which uses a standard cluster covariance estimator with a small number of large clusters chosen using the k-medoids classification procedure recommended by Cao et al. (2023). Placebo simulations indicate that BCH, usually with four clusters, works reliably.[7]

The second large cluster inference technique is that of Ibragimov and Müller (2010) (IM). This involves estimating a study's model within each of a set of clusters, again chosen via k-medoids. Estimated cluster-specific treatment slopes, $\hat{\beta}_c$, for each cluster c are then used as though they were observations in a classical Gaussian regression with just an intercept of β (a location model), with β being the treatment parameter of interest. This approach, while conservative, has very good robustness to cross-cluster heterogeneity but is limited to studies with enough data for the whole regression model to be well estimated within cluster. Placebo simulations again indicate that IM works well.

We apply the noise diagnostics and spatial basis regressions to 30 persistence studies that have appeared in leading journals. Each of these papers is a careful and lengthy statistical exercise that we do not attempt to replicate in full. In each case we reproduce one of the "leading" regressions that uses the full set of controls, usually located in one of the right hand columns of Table 2 or 3. Given the preoccupation of the persistence literature with treatment t-statistics, they will be our focus here.

This paper is concerned with detecting potential problems of inference caused by spatial trends and correlation, and introducing inference methods to handle these reliably. It is not concerned with issues of data construction. It is not concerned with the plausibility of the mechanism that is said to drive the claimed persistence, or alternative explanations of regression results, or with the quality of the underlying historical scholarship (although this can be extremely high, especially in regional studies).

Above all, and this cannot be emphasized too strongly, we are not concerned with somehow "validating" or "disproving" the findings of individual studies. In fact, the paper is not interested in any individual result except insofar as it illustrates the broader contours of the literature. Moreover, an individual regression's results being non-robust to spatial dependence issues does not necessarily imply that other regressions in the paper

[7]The common practice of clustering into many small clusters (like clustering household data by census tract) generally leads to unreliable results with spatial data because within-cluster averages are correlated between neighbouring small clusters. This applies especially to the clustering by observational units in longitudinal studies advocated by Bertrand, Duflo and Mullainathan (2004).

are equally problematic. In particular, changes in conditioning information can drastically change the impact of spatial dependence upon inference.[8]

Using our spatial basis regressions with BCH standard errors, inference about persistence effects changes drastically from the original studies (where most reported p-values on the order of 0.001 or lower) with only 2 studies of the 30 having a treatment effect significant at 5% and 2 more at 10%. Only 1 of our 30 studies can reject the synthetic outcome test no treatment effect null at 5% ($p = 0.047$).

The IM results largely agree with those using BCH. We find, moreover, that cluster-specific treatment estimates are often unstable. In two thirds of studies at least one cluster-specific estimate has a sign that is opposite to that of the full sample estimate. One study even exhibits Simpson's Paradox: every cluster's estimate has the opposite slope of the full sample one.

Of course, the fact that a treatment is insignificant at conventional levels does not necessarily mean that an effect is absent: it may simply not be possible to tie its magnitude down with any precision given the heavily trending and autocorrelated observations that are available. The legacies of history may be as overwhelming as their proponents insist, or as trivial as most historians suspect: we just cannot tell with the data at hand.

Our paper makes several contributions to the applied econometrics literature using spatial data. Besides enabling a test of no treatment effect, our placebo simulations facilitate the choice of the tuning parameter commonly required when implementing spatial standard error adjustments.[9] Although problems posed by trends in spatial and time series data have long been studied in statistics and econometrics going back to "Student" (W. S. Gosset) in 1914, applied work with spatial data has often not paid much attention to the issue as evidenced by many of the persistence studies that we examine. Our spatial basis method (which is similar in spirit to "pre-whitening" approaches that have long been popular in time series econometrics) offers a convenient way to de-trend in a general manner and improves the reliability of spatial standard error adjustments in general by reducing relevant spatial correlations.

The rest of the paper is as follows. The next Section gives an example of spurious regression results with spatial observations while the placebo and synthetic outcome tests

[8] See for example Conley and Topa (2002) and Conley and Ligon (2002).

[9] Such tuning "parameters" can be choice of large clusters, terms in an approximation, or a reference correlation function, e.g. Conley 1999, Ibragimov and Müller (2010), Kim and Sun (2011), Canay, Romano and Shaikh (2017), and Müller and Watson (2022, 2023). Our use of simulation is closely related to bootstrapping spatial data: see Conley et al. (2023).

are introduced in Section 3. These diagnostics are applied to persistence regressions with simple trend controls and no spatial standard error corrections in Section 4. The spatial basis procedure is introduced in Section 5 and applied to persistence studies in Section 6. Ibragimov and Müller (2010) inference and coefficient stability are considered in Section 7.

2 Spurious Results Due to Spatial Correlation

As we noted above, regressions based on spatial observations can be hard to estimate reliably because the data often show marked directional trends or other large scale structure, and are strongly autocorrelated. Here we illustrate the problems arising from spatial correlation, and show that even in the absence of trends, t-statistics are frequently inflated because standard errors without adequate adjustments for spatial correlation tend to be too small. This is a spatial version of spurious inference caused by serial correlation that has been a focus of the time series econometrics literature since Granger and Newbold (1974), and dealing with it has been a central part of the spatial econometrics literature since at least Conley (1999).

We are interested in least squares estimation of the familiar linear regression model:

$$y_i = x_i'\beta + u_i. \tag{1}$$

The variance of the least squares estimator of β depends on the variance of the average of $x_i u_i$ across observations.[10] With spatial data many of the covariances between $x_i u_i$ and $x_j u_j$ may be positive, leading to a large increase in the variance of the estimator of β relative to its variance under spatial independence.

This is illustrated in Figure 1. There are 250 towns represented by white dots scattered at random over a square with sides of length one, and two independently constructed spatial noise variables called "Historical Variable" and "Modern Outcome." We use a Data Generating Process (DGP) with highly spatially correlated data of the form that we will see in real series below.[11] The values of the noise variables in each town are regressed on

[10] The least squares estimator of β, $\hat{\beta}$ can be written as: $\hat{\beta} = \beta + [(\frac{1}{N}\sum x_i x_i')^{-1}\frac{1}{N}\sum x_i u_i']$. The variance of $\hat{\beta}$ is determined by the term in brackets.

[11] In the notation of the next Section, $\mu = 0, \theta = \sqrt{2}/10, \tau^2 = 0.9, \sigma^2 = .1$. The variance of the sample mean for this process is about 20.5 times what it would be without spatial dependence, approximately the same amount of dependence an an AR1 time series with a slope of about 0.91.

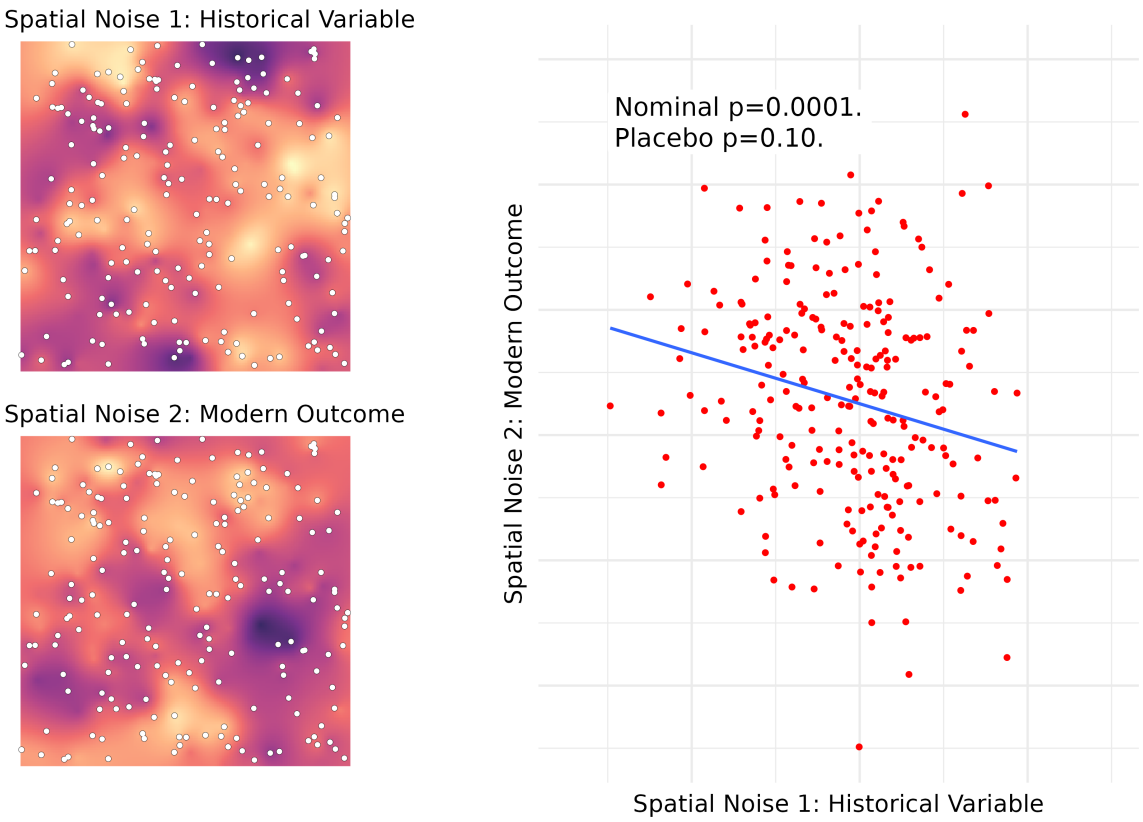

Figure 1: Regressions of one empirically realistic spatial noise series on another commonly return the sort of t statistics reported in historical persistence studies. If we regress noise values at the white dots (towns) on each other and use HC standard errors we get $t = -3.8$ with nominal significance $p = 0.0001$. The Monte Carlo significance level is 0.1: ten per cent of simulations had t statistics with $|t| > 3.8$.

each other and using heteroskedasticity consistent standard errors (without any adjustment for spatial correlation) we get a t statistic of -3.8 with nominal significance of 0.01%. However, the Monte Carlo significance level of the regression is actually 10%: one tenth of simulations had t statistics with absolute value above 3.8. Thirty-eight per cent of simulations have t statistics above 2, 21% above 3, and 8% above 4. In fact, a t-statistic of 4.4 is required for Monte Carlo significance of 5%, and of 5.9 for 1%. The simulated data do not include a spatial trend: were this done the inflation of t-statistics would be yet more severe.

1.1 Spatial Standard Error Corrections

It is, of course, well known that for spatially correlated data heteroskedasticity consistent standard errors can be incorrect (and perhaps too small), requiring some sort of correction.

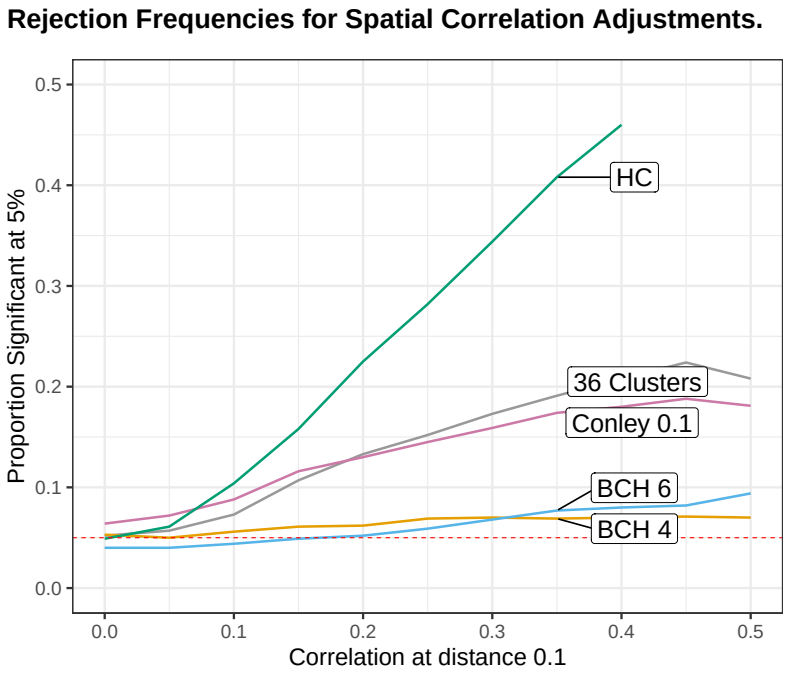

Figure 2: Spatial standard error adjustments. The diagram gives simulation rejection frequencies of nominal 5% t-tests for a correct null hypothesis of zero slope from an OLS regression of two independent but spatially correlated series on each other. The spatial correlation of the series increases from left to right, described by the correlation between points at distance .10.

Figure 2 illustrates how the performance of different standard error corrections can vary as the level of spatial correlation between observations increases.

The simulations have 250 points distributed at random across a unit square and involve regressing two sets of independent spatial noise on each other. The correlation between points at distance 0.1 apart ranges from 0 to 0.5.[12] A correlation of 0.5 might not appear large by the standards of time series but it is important to remember that there are many pairs of observations within that distance so seemingly modest pairwise correlations add up to very large amounts: the variance of the sample mean for the spatially correlated component is about 23 times greater than it would be if the data were uncorrelated. An analogous AR1 time series with this level of dependence would have a slope of approx-

[12]The DGP from the equation (3) is used with $\mu = 0, \theta = \sqrt{2}/10$, with τ^2 varying from zero to one, and $\tau^2 = 1 - \sigma^2$.

imately 0.92. Therefore, from left to right the Figure illustrates a wide range of spatial correlation levels.

Figure 2 illustrates the performance of three commonly used spatial dependence corrections along with two of the sort used in the remainder of this paper. The line labeled "HC" illustrates that heteroskedasticity consistent inference with no correction for spatial dependence performs very poorly as dependence increases. The line labeled "Conley 0.1" uses a Conley (1999) correction with a rectangular kernel allowing for correlations up to a distance of 0.1. It provides a considerable improvement over HC but still over-rejects substantially, in particular for larger levels of correlation.[13]

The line labeled "36 Clusters" reports results using cluster standard errors with 36 medium-sized groups based on a square 6×6 grid, and it performs similarly to the "Conley 0.1" adjustment. As spatial correlation rises, the level of across-cluster correlations that are unaccounted for increases, and this undermines its performance. Increasing the number of clusters would aggravate this problem further.

The corrections labeled "BCH 4" and "BCH 6" illustrate the performance of the primary inference method used in the remainder of this paper. They use regressions augmented with spatial basis functions and four or six large clusters chosen by k-medoids classification. As the level of dependence increases, the lower number of clusters performs better in terms of rejection frequencies at the cost of having fewer degrees of freedom for t-tests and hence a tendency to have larger confidence intervals.

3 Spatial Diagnostic Tests

Given the strong risk of spurious results arising from trends and spatial correlations in spatial data, we begin with two new diagnostic tests that can be applied once a researcher has decided on a regression specification and inference method. The first diagnostic is a placebo test where the real treatment is replaced with spatial noise. The second tests the null hypothesis that the outcome is generated by a "trend plus spatial noise" process, with no connection to the treatment. Both are joint tests of no treatment effect and the functional form assumed for the distribution of either treatments or outcomes. We refer to them as diagnostics relative to a benchmark test of zero treatment slope because they impose these additional restrictions.

[13] This type of DGP is challenging for Conley (1999) estimators because important correlations extend for large distances. See Conley, Kelly and Kozbur (2024) for details.

3.1 Placebo Treatment Test

Our first diagnostic is a placebo test of no treatment effect. As we noted above, to shuffle treatments randomly across subjects, as standard placebo tests do, is to ignore the spatial correlation structure that defines the data. Instead we first de-trend the treatment variable by regressing it upon controls for spatial trends, and then estimate a model of detrended treatments' spatial correlation structure. We take draws from this estimated distribution as synthetic treatments. These synthetic treatment variables are independent of the outcome by construction, but they have approximately the same spatial correlation structure as the true (detrended) treatment. Repeatedly running the regression of interest using synthetic treatments gives a collection of placebo treatment t-test statistics that forms the reference distribution for a "randomization inference" type test of the hypothesis that there is no treatment effect.

To use the placebo diagnostic we compare the study's treatment slope p-value with the p-value from the placebo reference distribution. Suppose that the study's t-test has a low p-value, rejecting a treatment effect slope of zero. If the placebo test has a similarly low p-value rejecting "no treatment effect" this lends support to the paper's t-test being reliable. If however the placebo test clearly does not reject then there is a contradiction. In such a case, because the placebo test's reference distribution is well approximated by the simulation, there is reason to suspect that the paper's inference method is not properly accounting for spatial correlations.

The placebo simulations offer another substantial benefit: they provide a simple Monte Carlo check on the performance of the regression's inference method. If the inference method accounts well for the spatial correlation in simulated data, rejection frequencies in the simulations should match the nominal level of the test. For example, a 5% level t-test should be significant in about 5% of placebo simulations. These simulations are not likely to be a perfect match for the true data generating process, but they serve as an empirically relevant Monte Carlo evaluation.[14] We use our placebo Monte Carlo to evaluate tuning parameter choice for spatial inference procedures, in particular how many clusters to use with our BCH and IM large cluster adjustments.

[14] It is possible for these placebo simulations to have more spatial dependence in residuals than under the true data generating process if the true coefficient is non-zero.

3.2 Gaussian Spatial Processes

To estimate spatial structure for simulating our diagnostics we use an additive Gaussian model. For a vector of observations of a variable V at sites s we assume it has two components, one with spatial dependence and the other idiosyncratic noise:

$$V(\mathbf{s}) = \mu(\mathbf{s}) + \psi(\mathbf{s}) + \eta(\mathbf{s}). \tag{2}$$

$\mu(\mathbf{s})$ is a mean which we will parameterize as a function of location coordinates. $\psi(.) \sim N\left(0, \tau^2 K\right)$ is a stationary Gaussian process that has spatial correlation matrix K with ij-th element $K(s_i, s_j)$ the correlation between sites s_i, s_j; and $\eta(.) \sim N\left(0, \sigma^2 I\right)$ is Gaussian idiosyncratic noise that is independent of ψ and location. So the distribution of $V(\mathbf{s})$ is:

$$V(\mathbf{s}) \sim N\left(\mu(s), \tau^2 K(\mathbf{s}) + \sigma^2 I\right). \tag{3}$$

A standard functional form for K in applied geostatistics is the Matérn function (see, for example, Gneiting and Gutthorp 2010) which takes on different shapes ranging from exponential to Gaussian depending on the value of a smoothness parameter. For the data here, the Matérn function with an exponential decay for K is a parsimonious choice across studies. For locations s_i, s_j at distance h apart their spatial covariance component equals

$$K(s_i, s_j) = \exp\left(-\frac{h}{\theta}\right) \tag{4}$$

where θ is a range parameter. All Matérn functions regardless of shape have the same correlation, approximately 14%, at distance 2θ, so we report estimates of 2θ as a convenient descriptor of spatial covariances. To describe the relative importance of the spatial signal (τ^2) and idiosyncratic noise (σ^2) components of V we will use ρ defined as:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{5}$$

3.3 Synthetic Outcomes Test

Our second diagnostic test uses simulated data to directly test the null hypothesis that outcomes are generated by a combination of spatial trend plus spatially correlated noise that match the marginal distribution of the real outcome variable. We estimate a parametric "trend plus spatially correlated noise" model for the outcome and run regressions with

the simulated outcomes replacing the true one. The trend in these simulated outcomes may be mechanically related to trends in the treatment, but the remaining variation is by construction independent of the treatment and other predictors. The resulting collection of simulation t-statistics forms the reference distribution under the null hypothesis that outcomes are generated by a "trend plus spatially correlated noise" process with no treatment effect, that the real t-statistic can be compared against. The treatment effect slope is not necessarily an optimal test statistic but the simulations provide a very accurate approximation of its reference distribution under the null hypothesis.

4 Motivation: Persistence Regressions with Trend Controls

To motivate our regression procedure, we begin by examining the robustness of each persistence study's results to adding simple spatial trend terms to its regression. We do not yet correct standard errors for spatial autocorrelation but instead apply whatever method was used in the original study and therefore anticipate that standard errors may be too small. We then apply the placebo and synthetic outcomes tests to each study. In Section 5 we repeat the exercise but with spatial basis regressions and large cluster standard errors.

4.1 Robustness to Simple Spatial Trends

Large scale spatial structure is common in spatial data and needs to be controlled for to obtain reliable confidence intervals or hypothesis tests. This is occasionally done in persistence studies, and here we simply add the most widely used control variables systematically to every regression. For global studies based on country data, many studies use distance from the equator and a dummy for continents as controls. We modify this slightly, using World Bank regions (Middle East and North Africa, Sub-Saharan Africa, and so on) as a more useful classification than the continent dummies. For studies on a smaller geographical scale we add a quadratic in longitude and latitude.

If the addition of trend variables renders the estimated treatment effects not statistically different from zero (even with standard errors that may be too small), this suggests that there is insufficient evidence for the researcher to distinguish a persistence treatment effect from a spatial trend. Such a treatment effect might indeed exist but it cannot be detected with the available data.

	Original Specification				Simple Trend Added			
	Nominal p	Placebo p	Synth p	Plac. 5%	Nominal p	Placebo p	Synth p	Plac. 5%
Acemoglu, Colonial.	0.000	0.003	0.346	0.202	0.250	0.286	0.263	0.057
Acemoglu, Reversal.	0.002	0.018	0.112	0.119	0.191	0.254	0.272	0.078
Acharya, Slavery.	0.013	0.044	0.031	0.119	0.013	0.044	0.031	0.119
Alesina, Plough.	0.000	0.033	0.022	0.298	0.045	0.085	0.069	0.096
Alsan, Tsetse.	0.002	0.020	0.111	0.141	0.006	0.055	0.052	0.175
Ambrus, Cholera.	0.000	.	0.128	.	0.065	.	0.398	.
Arbatli, Conflict.	0.019	0.074	0.072	0.160	0.103	0.189	0.180	0.100
Ashraf, Diversity.	0.125	0.519	0.105	0.405	0.002	0.276	0.007	0.558
Bazzi, Individualism.	0.000	0.000	0.002	0.302	0.000	0.001	0.002	0.330
Becker, Weber.	0.000	0.003	0.000	0.233	0.000	0.001	0.000	0.184
Buggle, Serfdom.	0.002	0.012	0.011	0.152	0.010	0.051	0.066	0.122
Caicedo, Mission.	0.024	0.516	0.461	0.585	0.193	0.678	0.583	0.545
Comin, 1000BC.	0.002	0.118	0.480	0.370	0.829	0.874	0.867	0.138
Dell, Mita.	0.001	.	0.126	.	0.321	.	0.465	.
Drelichman, Inquisition.	0.000	0.000	0.000	0.074	0.000	0.000	0.000	0.103
Enke, Kinship.	0.000	0.021	0.107	0.345	0.000	0.013	0.104	0.274
Galor, Impatience.	0.001	0.011	0.094	0.170	0.016	0.036	0.055	0.084
Giuliano, Climate.	0.011	0.128	0.187	0.258	0.280	0.292	0.349	0.042
Iyer, Raj.	0.000	.	0.008	.	0.001	.	0.030	.
Juhasz Blockade.	0.006	0.030	0.773	0.228	0.261	0.263	0.281	0.034
LaPorta, Legal Origins.	0.001	.	0.003	.	0.009	.	0.000	.
Michalopoulos, Folklore.	0.004	0.030	0.006	0.158	0.000	0.001	0.000	0.089
Michalopoulos, Precolonial.	0.003	0.004	0.005	0.047	0.000	0.001	0.002	0.068
Montero, Saints.	0.009	.	0.012	.	0.014	.	0.018	.
Nunn, Mistrust.	0.000	0.000	0.004	0.140	0.000	0.003	0.031	0.114
Nunn, Slavery.	0.003	0.006	0.006	0.070	0.022	0.031	0.026	0.059
Schulz, Kinship.	0.000	0.066	0.246	0.377	0.039	0.062	0.086	0.075
Spolaore, Diffusion.	0.000	0.002	0.036	0.386	0.038	0.157	0.201	0.184
Squicciarini, Devotion.	0.001	0.124	0.020	0.411	0.835	0.855	0.826	0.108
Voigtlaender, Persecution.	0.012	.	0.026	.	0.012	.	0.017	.

Nominal, placebo and synthetic outcome significance levels for persistence studies, using their original specification, and with simple trends added. Standard errors are not corrected for spatial autocorrelation of residuals: the regressions use whatever standard error procedure was applied in the original studies. Trends are a second level polynomial in longitude and latitude for regional studies, and World Bank region and absolute latitude for global ones. Plac. 5% denotes the proportion of simulations where the placebo treatment had a t-statistic above 2. Empty cells are for studies with binary treatments where placebos were not constructed.

Table 1: Diagnostic tests for persistence studies. Standard errors are computed using the same method as the original study.

The impact of these trend controls is shown by comparing the blocks of results in Table 1 under "Original Specification" and "Simple Trend Added". Comparing the first columns of these blocks shows that trends are clearly an issue for a substantial fraction of these persistence regressions. Although the median significance level originally reported was 0.002, after adding our simple trend controls it rises an order of magnitude to 0.02. Ten of 30 studies have t-statistics that become not significantly different from zero at 5%, even with standard errors that may be too small.[15] The p-values for the 24 studies with non-binary treatments (where placebo tests are constructed) are plotted in Figure 3. The grey dots are the p-values originally reported, and the blue dots give p-values after a spatial trend is included in the regression. For most studies in the top third of the Figure, it is clearly possible that the persistence variables are acting as a proxy for omitted spatial trends. The clear importance of spatial trends leads us in Section 5 to introduce a spline procedure that allows a very flexible specification for spatial trends.

4.2 Placebo Tests

We conduct placebo tests for the 24 cross sectional studies that have continuous treatment variables where the Gaussian model in Section 3 is a reasonable specification.[16] Maximum likelihood estimates of spatial correlation parameters for the detrended persistence treatment variables are given under the columns labelled x in Table 2 and these turn out with few exceptions to be extreme. Given the very different geographical scales of the different studies, we report estimates of 2θ—the distance when correlation K is about 14%—as a proportion of the 95-th percentile of distances between locations. These ranges tend to be substantial in most cases indicating that spatial correlations are "slow" to decline with distance.

The ratio $\rho = \tau^2/(\tau^2+\sigma^2)$ provides a useful summary of the size of spatial components relative to idiosyncratic noise when, as is the case here, the estimate of 2θ is large relative to observed distances. The large majority of studies report ρ values near one, showing that spatial correlation is large compared with idiosyncratic noise.

[15] One study, Ashraf and Galor (2013), requires a joint significance Wald test of a quadratic in treatment parameters and has a p-value of 0.12 originally and 0.02 after including a simple trend.

[16] Placebo tests are feasible with binary treatments but proper specification of a binary outcome model with trends is not straightforward and would possibly need to be tailored to the individual application. See Phillips, Jin and Hu (2007) for a related discussion.

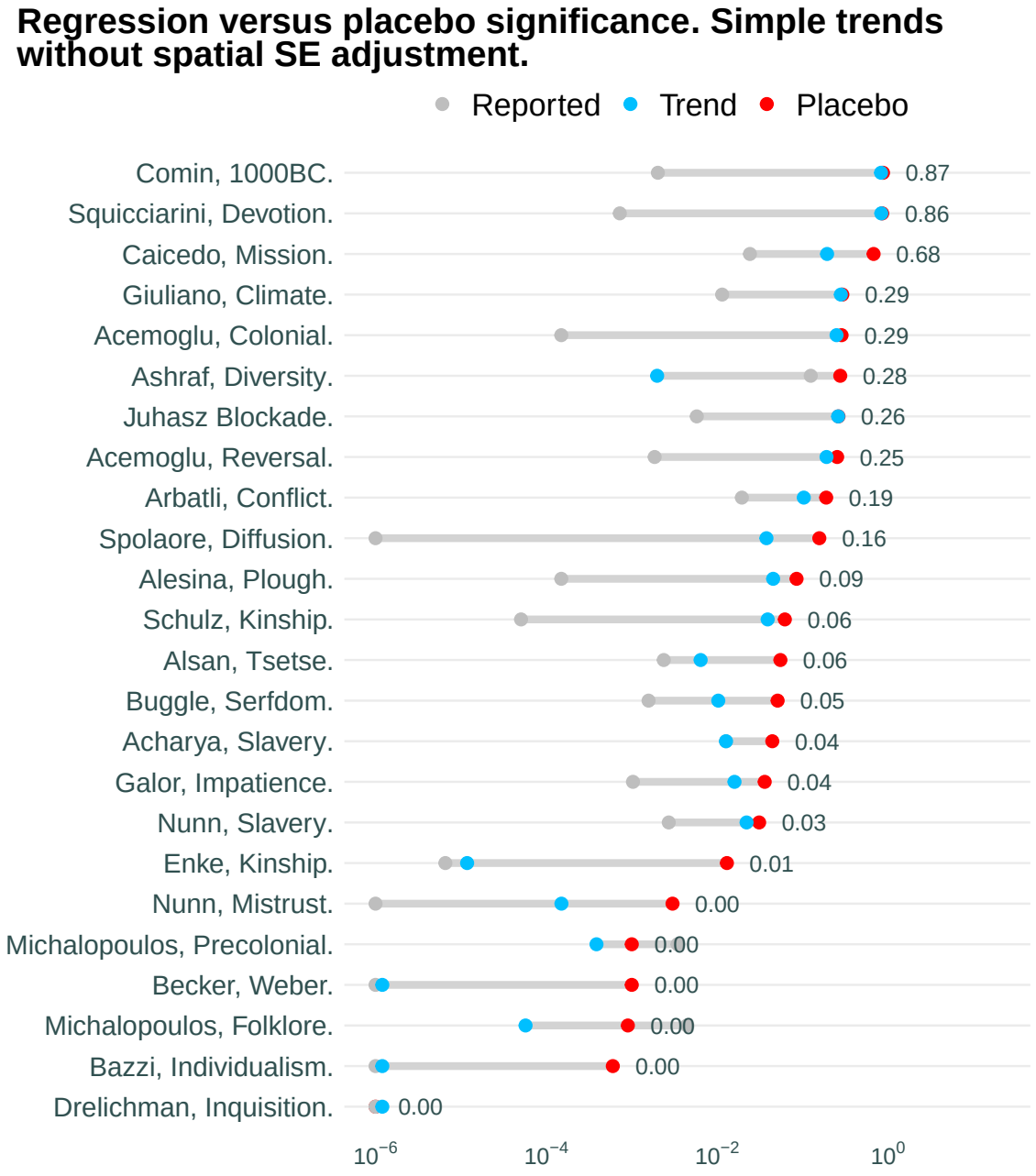

Figure 3: Trend and placebo significance levels for persistence studies (excluding those with binary treatments) using original standard error procedures. Grey dots are reported p-values from persistence papers' regressions, and blue dots give significance after a simple spatial trend is included in the regression. Red dots are placebo test p-values.

	Trend R^2		Range 2θ		Structure ρ	
	x	y	x	y	x	y
Acemoglu, Colonial.	0.55	0.45	0.15	0.05	0.41	0.23
Acemoglu, Reversal.	0.53	0.42	0.20	0.05	0.99	0.81
Acharya, Slavery.	0.29	0.05	0.24	0.12	0.99	0.10
Alesina, Plough.	0.63	0.40	0.55	0.05	0.64	0.92
Alsan, Tsetse.	0.37	0.04	0.29	0.62	0.99	0.91
Ambrus, Cholera.	.	0.13	.	0.15	.	0.95
Arbatli, Conflict.	0.69	0.12	0.66	0.05	1.00	0.72
Ashraf, Diversity.	0.95	0.37	1.99	0.35	1.00	0.99
Bazzi, Individualism.	0.10	0.35	0.25	0.30	0.99	0.82
Becker, Weber.	0.41	0.54	0.16	0.31	0.81	0.99
Buggle, Serfdom.	0.36	0.27	1.40	0.70	0.75	0.20
Caicedo, Mission.	0.96	0.03	1.98	0.15	1.00	0.88
Comin, 1000BC.	0.73	0.65	0.55	0.10	0.93	0.77
Dell, Mita.	.	0.04	.	0.05	.	0.07
Drelichman, Inquisition.	0.07	0.05	0.05	0.00	0.26	0.87
Enke, Kinship.	0.48	0.33	0.05	0.05	0.70	0.67
Galor, Impatience.	0.42	0.56	0.15	0.20	0.96	0.53
Giuliano, Climate.	0.57	0.47	0.05	0.10	0.96	0.44
Iyer, Raj.	.	0.39	.	0.25	.	0.80
Juhasz Blockade.	0.93	0.43	2.12	0.00	0.99	0.00
LaPorta, Legal Origins.	.	0.10	.	0.20	.	0.14
Michalopoulos, Folklore.	0.30	0.44	0.05	0.15	0.90	0.60
Michalopoulos, Precolonial.	0.08	0.14	0.04	0.24	0.75	0.70
Montero, Saints.	.	0.21	.	0.00	.	0.88
Nunn, Mistrust.	0.34	0.03	0.10	0.10	0.30	0.10
Nunn, Slavery.	0.31	0.30	0.06	0.19	0.96	0.30
Schulz, Kinship.	0.65	0.58	0.04	1.74	0.55	0.77
Spolaore, Diffusion.	0.63	0.47	0.25	0.21	0.88	0.86
Squicciarini, Devotion.	0.20	0.47	0.26	0.00	0.94	0.95
Voigtlaender, Persecution.	.	0.16	.	0.26	.	0.46

Maximum likelihood estimates of spatial parameters of the treatment (x) and outcome (y) variables. Trend R^2 is the fraction of the variance explained by trend variables. For global studies the trend variables are absolute latitude and World Bank region, and for smaller scale studies they are a quadratic in longitude and latitude. The remaining statistics are for the residuals from this regression. Range is the effective range 2θ, expressed as a fraction of the 95th percentile of the distance between sites. Structure is the ratio of signal to signal plus noise ρ.

Table 2: Spatial parameters of treatment and outcome variables.

Returning to Table 1, the first column of each block labeled "Nominal p" reports significance levels of a t-test of zero treatment slope. In the "Original Specification" block, all studies have treatment slope p-values well below 5% with the exception of the Ashraf and Galor (2013) Wald statistic p-value of 0.125. The second column of the "Original Specification" block reports placebo test p-values, and these contradict an original test rejection of zero treatment slope at conventional levels in 7 of 24 cases. The fourth column of each block reports the fraction of placebo simulation test statistics that are deemed statistically significant at 5% using the study's t-test procedure. If the study's inference method is working well, this column should be about 5%. Only 3 studies have rejection frequencies below 10%. The remaining 21 studies markedly over-reject the true null hypothesis, indicating that their standard errors are too small.

In the trend-augmented regressions, reported in the "Simple Trend Added" block, only 20 of 30 studies have estimated treatment effects that are significantly different from zero at 5%. However the placebo test indicates a failure to reject no treatment effect (at 5%) in 6 of the 20, suggesting that these papers' standard errors are not adequately accounting for spatial dependence, even when augmented with simple trends. These discrepancies are illustrated in Figure 3 using blue and red dots respectively. Low values of both blue and red dots indicate a satisfactory methodology; but low blue values and high red ones are contradictory and suggest the paper's inference methods are inadequate. The fourth column of the "Simple Trend Added" block of Table 1 presents the rejections for the placebo test Monte Carlo exercise with 10 of 24 studies having rejection frequencies between 4% and 9% but the remaining 15 rejecting 10% or more: better than for the "Original Specification" block, but there is clearly room for improvement. This leads us in the next Section to introduce a new regression procedure to improve inference in the presence of spatial dependence.

4.3 Synthetic Outcome Tests

Table 1 also reports results from our synthetic outcome tests in the columns labeled "Synth-p". To construct synthetic outcomes, we use the same trend variables as in Table 1: absolute latitude and regional indicators for global studies, and a quadratic polynomial in latitude and longitude for smaller scale ones. We generate noise as we did for treatments in the placebo test, using maximum likelihood estimates for detrended outcomes. These parameter estimates are reported in the columns labelled y in Table 2. The outcomes (which

usually are fairly mundane economic variables such as consumption or GDP per capita) tend to have considerably lower spatial dependence than treatments but are nevertheless substantially spatially correlated in most cases.

We apply the synthetic outcome diagnostic to the 30 original regressions in the "Original Specification panel of Table 1. Strikingly, for 14 of 30 original specifications we are unable to reject (at 5%) the hypothesis that outcomes are driven by trend plus noise; that there is no true treatment effect. This suggests significance results in nearly half of the original regressions could be due to treatment trends predicting outcome trends or to inadequate handling of spatial correlation. The corresponding results in the "Simple Trend Added" panel for trend-augmented regressions are similar, with only 13 of 30 studies rejecting the trend plus noise no treatment effect hypothesis at 5%, and 18 rejecting at 10%. The specifications in our trend augmented regressions are sufficient to capture the trend component of our synthetic outcomes so results for this panel suggest many studies' original inference methods do not properly account for spatial dependence.

5 Regression with a Spatial Basis and Large Cluster Inference

5.1 Spatial Basis Regressions

Our approach is to augment regressions with a flexible function of locations based on spatial splines to make inference with spatial observations more reliable. We add a set of spatial regressors g_i to our regressions

$$y_i = x_i'\beta + g_i'\delta + \epsilon_i. \tag{6}$$

These additional regressors g_i serve two purposes. First, they provide a flexible way to detrend the data. Second, they reduce (but do not always eliminate) spatial correlation in residuals, making them closer to white noise so that it is easier to estimate reliable standard errors.[17]

The terms in g_i are motivated by considering a tensor spline approximation to the dependent variable surface in coordinate space. Intuitively we can think of this spline as a mesh made up of a small number r of undulating threads that run north-south, and an-

[17]The detrending role of our spatial basis is more important than its reduction of residual correlations because we will still use spatial dependence robust inference methods. The properties of this estimator are analyzed by Conley, Kelly and Kozbur (2024).

other r that run east-west. Each thread can capture a substantial amount of the variation along its path.

Consider the following parameterization of a surface as a function of two spatial coordinates m and n, such as latitude and longitude. In each dimension we can flexibly approximate a function of its coordinate as a linear combinations of r basis functions. This yields functions $h_1(m)$ and $h_2(n)$:

$$h_1(m) = \sum_{k=1}^{r} \alpha_k a_k(m), \; h_2(n) = \sum_{j=1}^{r} \beta_j b_j(n) \tag{7}$$

where the α_k and β_j are parameters and $a_k(m)$ and $b_j(n)$ are known basis functions. For basis functions we use piecewise linear B-splines, each of which is a triangle over an interval and then zero otherwise. To generalize this to flexible functions of (m, n) in two dimensions we allow for interactions between all of the components of our single coordinate basis functions, yielding

$$f(m,n) = \sum_{k=1}^{r} \sum_{j=1}^{r} \gamma_{kj} a_k(m) b_j(n). \tag{8}$$

We therefore have r^2 basis functions $a_k(m) b_j(n)$, each pair of which makes a three dimensional surface, and the $\gamma_{k,j}$ are an array of parameters.[18] We evaluate the set of interacted basis functions $\{a_k(m) b_j(n)\}$ at the coordinates of all observations and collect them in an N by r^2 matrix M.

We begin with low dimension tensor products of piecewise linear (triangular) B-splines. An example of one is given in Figure 4 where the simple surface explains 80 per cent of the variance of the outcome variable. To further conserve degrees of freedom, we do not use the matrix of basis functions M directly but take instead its first L principal components as our spatial regressors g_i and choose the value of L by a Bayesian Information Criterion penalty.

5.2 Spatial Correlation Inference Corrections: BCH and IM

We rely on including the g_i terms in the regression to de-trend our data and also anticipate they will capture some of the "lower frequency" spatial noise fluctuation in the data to

[18] A discussion of multidimensional splines can be found in Hastie, Tibshirani and Friedman (2008, Section 5.7) and includes a diagram (Figure 5.10) illustrating the components of a tensor B-spline.

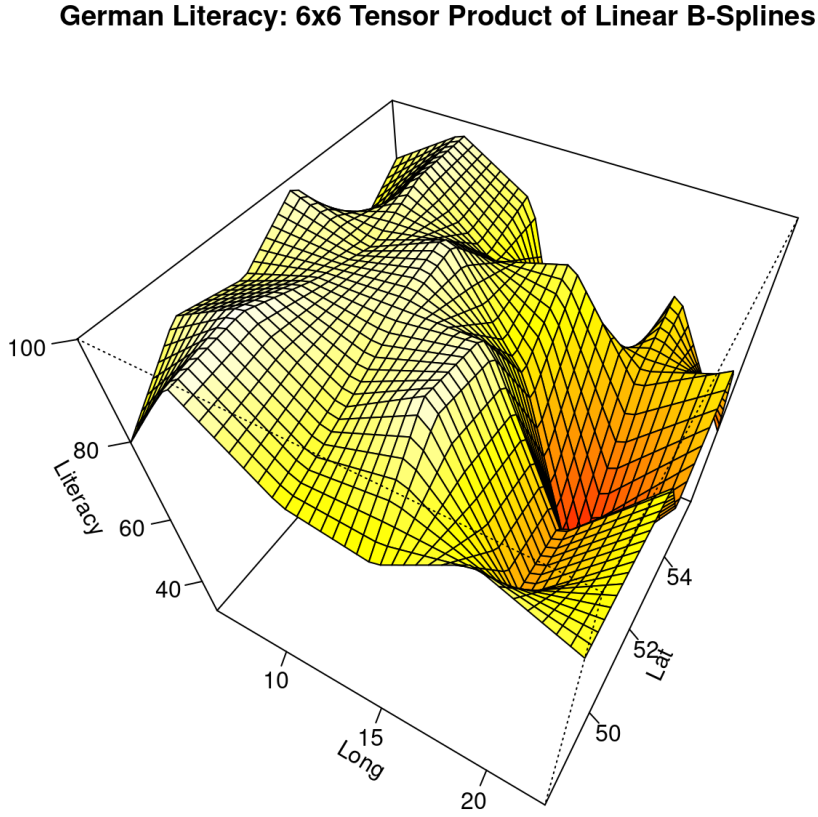

Figure 4: An example of a 6×6 tensor product of piecewise linear B-Splines in longitude and latitude, applied to literacy in nineteenth century German from Becker and Woessmann (2009). As Figure 6 shows, this minimal spatial surface captures 80 per cent of the variation in literacy.

make inference easier. In general, however, not all autocorrelation in residuals will be removed so we still need to use spatial dependence robust inference procedures.

We first compute spatial dependence robust standard errors using BCH. This procedure employs standard cluster robust covariance estimators but with a small set of very large, well-shaped clusters, and the resulting t-statistic has critical values based on a Student t distribution with degrees of freedom equal to the number of clusters minus one. In most of the studies we examine the selected number of clusters is four in which case a t-statistic of 1.96 has a p-value of 0.14 and the critical value for a 5% test is 3.2. The key

requirements for the approximations in BCH to work well are that (a) cluster averages are approximately independent due to their large interior relative to boundary, (b) cluster averages are approximately Gaussian due to large group size, and (c) regressors' variances are similar across clusters.

In Section 7 we check the sensitivity of our results to the BCH homogeneity assumptions by comparison to the alternate large cluster method of Ibragimov and Müller (2010) that has excellent robustness properties under general cross-cluster heterogeneity. IM inference can be used with the same k-medoids clusters as BCH as long as there are enough within-cluster observations to get reasonable cluster-specific regression estimates. We also use our placebo-based Monte Carlo to evaluate IM inference.

Large cluster methods like BCH or IM are known to work better than Conley (1999) standard errors when spatial dependence levels are very high, which is the case for most data here. With higher spatial dependence, the number of clusters used should decrease so that most interior observations become farther from and less correlated with interior observations of other groups, resulting in group averages remaining approximately independent. We allocate points to clusters using the k-medoids clustering algorithm analyzed by Cao et al. (2023) who demonstrate that a k-medoids partitioning algorithm produces well-shaped' clusters that have the requisite small boundary relative to interior, outside of pathological cases.[19]

The placebo simulations serve as a diagnostic to evaluate the performance of both BCH and IM with a chosen set of clusters. If around 5% of placebo simulations are significant for a nominal 5% test, this suggests that the number of clusters is appropriate. We start with a modest number of clusters, six, then if more than 8% are rejected we successively reduce the number of clusters. In most cases the chosen number is four. As noted above, the residuals in placebo simulations can exhibit higher spatial correlation than those in the real data so this diagnostic may lead us to be conservative in a sense, to err on the side of keeping the number of clusters relatively small.[20] This will still result in reliable inference, but confidence intervals may be wider than necessary.

[19] A k-medoids clustering algorithm is analogous to a k-means classification algorithm, but the cluster "centers' are locations of within sample data points. This allows cluster calculation using just inter-point (potentially non-Euclidean) distances between observations. See the Appendix for further discussion about implementing cluster choice.

[20] IM inference is also conservative in the usual sense, nominal 95% confidence intervals have 95% or greater coverage. See Section 7.

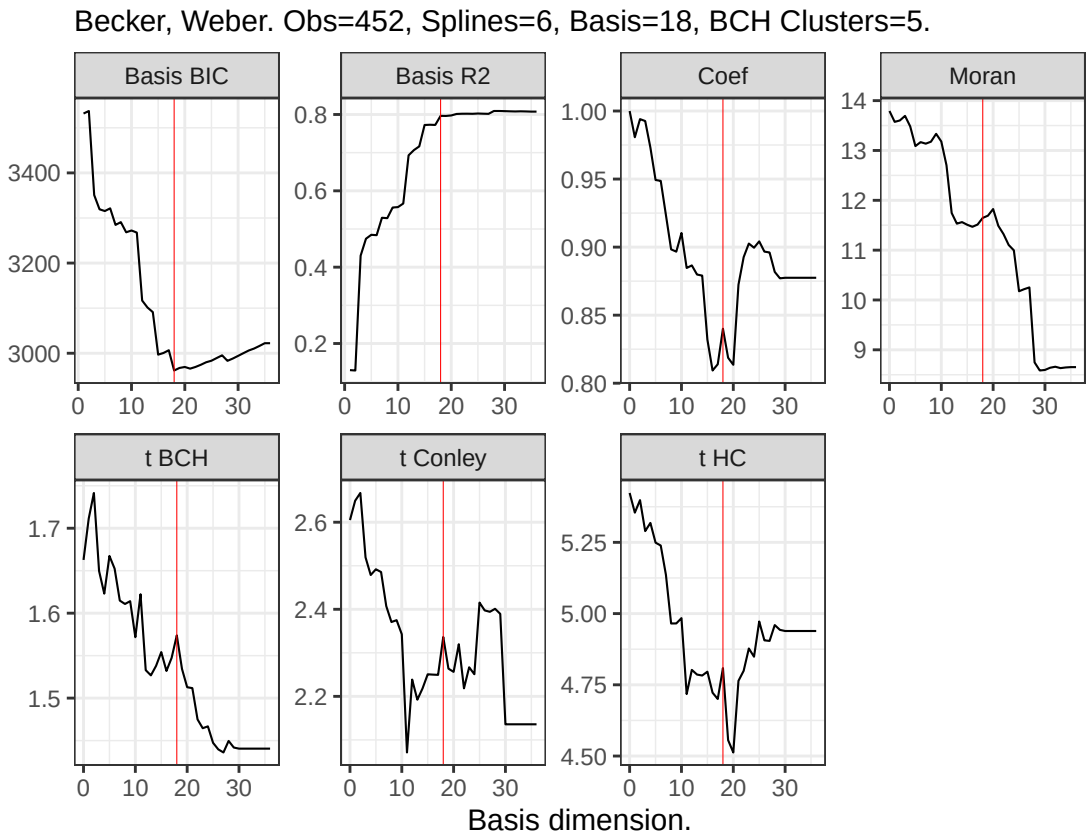

Figure 5: Example of the effect of adding spatial basis terms from a 6×6 spline to a regression. The optimal basis dimension based on BIC is 18 and explains 80% of variance in the outcome. With the basis, the treatment coefficient falls by about 16%. However, spatial correlation of residuals remains with a high five-nearest-neighbors Moran statistic of 12 (scaled to be standard normal under the null of zero spatial correlation). The t-statistic computed with heteroskedasticity robust standard errors remains over 4.5 and that computed with Conley standard errors (using a uniform kernel with cutoff of 100km) falls to 2.3. Given that the BCH t-statistic has a reference Student t distribution with four degrees of freedom making interpretation difficult, the BCH panel reports a pseudo t-statistic given by the value of a standard normal statistic that has the same p-value.

5.3 Two Example Regressions

An example of the process at work for one persistence study is given in Figure 6.[21] The first two panels give the BIC and adjusted R^2 of a regression of the dependent variable on

[21] The same diagram for all studies can be found in the Online Appendix.

the first L principal components of a 6×6 tensor, as basis dimension L rises. It can be seen that 18 principal components minimize BIC and explain about 80% of the variance in the dependent variable. Next is the regression coefficient with the outcome is rescaled so that the original zero basis coefficient is one: including 18 basis vectors causes this to fall by 17%.

The box labeled Moran reports Moran tests for spatial correlation. The Moran test that we use can be interpreted as the average correlation among five nearest neighbors.[22] The reported Moran statistics are scaled so they have approximately a standard normal distribution under the null hypothesis of zero spatial correlation. The Moran statistics generally decline with the number of basis terms, as is the case in all studies analyzed here, but remain larger than 8, clearly rejecting a null hypothesis of no spatial correlation even after including the spatial basis terms.

The panel labeled HC reports heteroskedasticity consistent t statistics that remain above 4.5 as the basis terms increase. However, as reported below in Table 3, 16% of placebo simulations are significant at 5% suggesting that HC standard errors are too small. For BCH with five clusters chosen by k-medoids, nominal 5% tests have 4% rejections in placebo simulations indicating it works satisfactorily. The panel labeled t Conley reports t-statistics created using Conley standard errors with a uniform kernel that is one for pairs of observations within 100km and zero otherwise, with results in between BCH and HC.

BCH critical values are based on a Student-t with four degrees of freedom rather than the standard normal critical values used for HC and Conley. Therefore, to make the results easily interpretable the panel reports a pseudo-t statistic which is the value of a standard normal statistic with the same p-value: for instance if the BCH p-value was 0.05 then the BCH pseudo-t would be 1.96. It can be seen that these pseudo t-statistics are near one: there is no evidence of a statistically significant treatment effect slope using basis regressions with BCH inference.

Figure 6 presents an illustration of our approach for a small sample size study with 64 observations, Acemoglu, Johnson and Robinson (2001) where BIC is minimized by a 3×3 tensor with 3 principal components that account for 35% of the variation in the outcome. In this case, as in other global studies, we no longer use regional indicators when we add

[22] Specifically, if we define e_i as the residual at site i, a Moran statistic I can be constructed using weights w_{ij} for sites i and j (with $w_{ii} = 0$) as $I = \frac{1}{\sum_{ij} w_{ij}} \sum_{i,j} w_{ij} e_i e_j / (\frac{1}{N} \sum_i e_i^2)$. Different weights will emphasize different covariances. We use nearest neighbor weights where $w_{i,j}$ is an indicator equal to one if j is among $i's$ five nearest neighbors and zero otherwise.

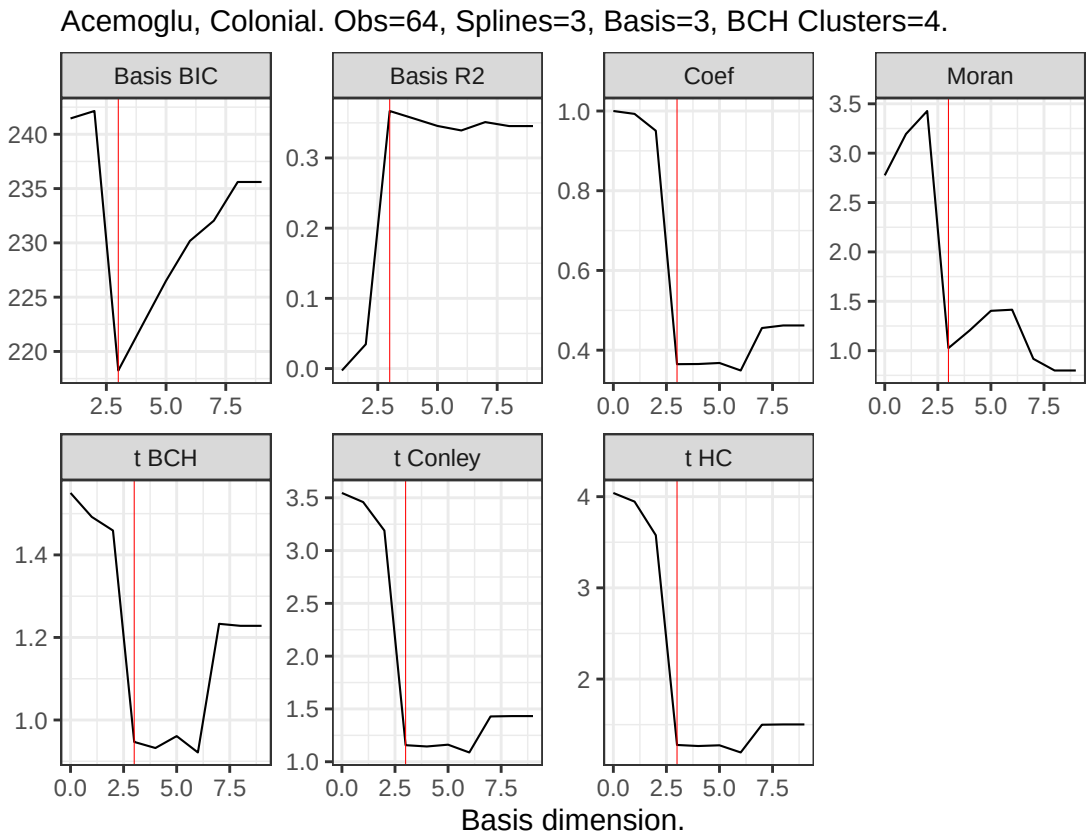

Figure 6: Acemoglu *et al*, "Colonial Origins". Adding 3 principal components from a minimal 3×3 spatial basis causes the HC t-statistic (which the 5% placebo level in Table 3 and low Moran statistic suggest is reliable here) to fall from 4 to 1.5.

a spatial basis. This study shows how adding even a small number of spatial basis terms can substantially alter empirical results, both point estimates and inference. The third subgraph indicates that with 3 basis terms the estimated coefficient of interest falls by over half and the fourth panel indicates that the Moran statistic falls below 1.5: residuals lose their spatial correlation. Our placebo simulations indicate that using BCH with four clusters chosen by k-medoids nominal 5% tests reject at 6%, indicating it works satisfactorily. A BCH t-test fails to reject a zero treatment slope with any basis and at 3 or more basis terms all three inference methods agree and fail to reject a zero slope.

6 Historical Persistence: Spatial Basis Regressions

We will now examine the impact of applying our spatial basis regression method and diagnostics to 30 studies of historical persistence that have appeared in leading journals. Most have continuous treatments but we do include five studies that use binary ones. The procedure extends immediately to panel regressions by using a three dimensional, spatiotemporal basis but we do not consider this here. Details of each regression are given in Appendix B. The simple trend adjustments used in Section 4—regional dummies and absolute latitutude for global studies and a quadratic in longitude and latitude for global ones—are now omitted because of the spatial basis.

Table 3 presents results from applying the spatial basis procedure and diagnostics to these persistence regressions. This choice of spatial basis terms and number of BCH clusters used in the Table are determined by a straightforward two step procedure.

Firstly, to minimize the risk of overfitting, we choose among spatial bases constructed from the principal components of relatively small tensors based on piece-wise linear B-splines with dimensions ranging from 3×3 to 6×6 for larger studies and 3×3 to 5×5 for studies with fewer than 100 observations. We select the the combination of tensor dimension and principal components that minimize BIC in a regression predicting the outcome with our spatial basis.[23] Secondly, we use the placebo simulations to evaluate rejection frequencies for a 5% nominal test for 3 to 6 k-medoids clusters and select the highest number of clusters where rejection frequencies are 8% or lower. In 19 of 24 studies, we found a cluster choice with 8% or fewer rejections with the number of PCs that minimized BIC. In 6 studies where rejection frequencies for all options were above 8% we systematically varied the number of PCs (up or down 1 or 2) and then repeated the Monte Carlo evaluation to choose clusters based on rejection frequency. [24] This slight adjustment of PCs resulted in specifications with 8% or lower rejections in 5 of 6 cases with the exception of one study (with 40 observations) where best-case Monte Carlo rejections remained at 9%. For binary treatments where placebo simulations are not conducted, results for four clusters are re-

[23] We select the model via BIC in this regression without other conditioning information to make sure we have a good basis for de-trending the outcome. It may contain a few basis terms more versus what would be selected with conditioning information. Having a few extra basis terms will tend to help when implementing spatial dependence robust inference as discussed in Conley, Kelly and Kozbur (2024).

[24] We first added one PC to the minimal BIC number, redid the Monte Carlo, stopping if we found a cluster size with 8% or fewer rejections. If necessary we repeated this procedure by subtracting one PC relative to BIC-optimal, adding two PCs, and finally subtracting two PCs.

	HC p				BCH p				Moran
	Estim-ated	Plac-ebo	Plac 5%	Synth Out	Estim-ated	Plac-ebo	Plac 5%	Synth Out	
Acemoglu, Colonial.	0.21	0.24	0.06	0.24	0.34	0.38	0.06	0.41	1.03
Acemoglu, Reversal.	0.80	0.82	0.06	0.86	0.80	0.86	0.09	0.82	0.20
Acharya, Slavery.	0.05	0.12	0.12	0.53	0.20	0.21	0.06	0.66	2.15
Alesina, Plough.	0.28	0.35	0.08	0.33	0.18	0.18	0.05	0.20	1.90
Alsan, Tsetse.	0.01	0.07	0.18	0.16	0.21	0.28	0.08	0.33	6.13
Ambrus, Cholera.	0.58	.	.	0.83	0.52	.	.	0.68	6.32
Arbatli, Conflict.	0.03	0.08	0.11	0.07	0.24	0.33	0.07	0.36	1.39
Ashraf, Diversity.	0.00	0.01	0.16	0.01	0.15	0.17	0.06	0.16	12.82
Bazzi, Individualism.	0.00	0.00	0.34	0.00	0.16	0.17	0.05	0.21	25.22
Becker, Weber.	0.00	0.00	0.10	0.00	0.12	0.11	0.04	0.16	11.65
Buggle, Serfdom.	0.00	0.02	0.37	0.08	0.21	0.27	0.06	0.30	.
Caicedo, Mission.	0.35	0.71	0.41	0.64	0.51	0.63	0.08	0.60	19.99
Comin, 1000BC.*	0.01	0.05	0.13	0.12	0.23	0.27	0.07	0.38	3.27
Dell, Mita.	0.00	.	.	0.06	0.19	.	.	0.33	5.85
Drelichman, Inquisition.	0.00	0.00	0.09	0.00	0.03	0.04	0.06	0.05	14.84
Enke, Kinship.	0.26	0.47	0.17	0.70	0.38	0.45	0.06	0.61	15.60
Galor, Impatience.*	0.00	0.03	0.14	0.02	0.15	0.19	0.07	0.12	1.45
Giuliano, Climate.*	0.01	0.04	0.15	0.07	0.07	0.11	0.07	0.07	3.83
Iyer, Raj.	0.01	.	.	0.05	0.33	.	.	0.41	8.75
Juhasz Blockade.	0.81	0.82	0.04	0.89	0.65	0.70	0.07	0.83	-1.08
LaPorta, Legal Origins.	0.00	.	.	0.00	0.11	.	.	0.23	2.34
Michalopoulos, Folklore.*	0.00	0.01	0.10	0.02	0.03	0.03	0.05	0.10	3.01
Michalopoulos, Precolonial.	0.00	0.01	0.14	0.01	0.16	0.17	0.05	0.17	7.22
Montero, Saints.	0.01	0.01	0.06	0.02	0.12	0.15	0.06	0.15	18.93
Nunn, Mistrust.	0.74	0.80	0.11	0.93	0.58	0.61	0.06	0.80	.
Nunn, Slavery.	0.09	0.10	0.06	0.10	0.24	0.27	0.08	0.27	-0.28
Schulz, Kinship.	0.19	0.25	0.08	0.35	0.28	0.34	0.07	0.39	0.40
Spolaore, Diffusion.	0.06	0.19	0.15	0.20	0.26	0.32	0.06	0.33	6.76
Squicciarini, Devotion.*	0.47	0.55	0.12	0.52	0.25	0.32	0.07	0.30	2.46
Voigtlaender, Persecution.	0.00	.	.	0.00	0.07	.	.	0.07	7.36

Significance levels for BCH and Heteroskedasticity Consistent standard errors. Estimated gives the significance level from the spatial basis regression, followed by the placebo and synthetic outcome ones. Placebo 5% gives the fraction of placebo simulations that are significant at 5%. Empty placebo cells correspond to studies with binary treatments. Moran gives the z test value for residual autocorrelation and is not reported for studies with multiple observations at the same site. Stars denote studies where the number of principal components was changed from the minimum BIC number in order to obtain 5% placebo values below 0.08. The change made is described in the details of each study given in Appendix B.

Table 3: Regression, placebo, and synthetic outcome significance levels for spatial basis regressions, using BCH and heteroskedasticity consistent standard errors.

ported in the Table, and Appendix Table 1 reports results for three to five clusters that are very similar across cluster choices.

The same spatial parameters reported for the simple trend regressions in Table 2 are repeated in Table 4, with three columns added. These are the tensor dimension and number of principal components that minimized BIC, and the number of BCH clusters chosen by the 5% placebo Monte Carlos. The 5 studies where adjustments to the BIC optimal number of splines were used are indicated with stars: details of each are given in Appendix B.

Placebos are constructed by regressing the treatment on the principal component basis and using the estimated spatial parameters of the residuals to generate spatial noise. Synthetic outcomes are derived, as previously, by generating noise that matches the residuals of a regression of the outcome on a quadratic in longitude and latitude.

Additional controls were added in three cases. Malaria prevalence is included as a control in all studies of Africa and is added to Nunn (2008). A dummy is included in Michalopoulos and Papaioannou (2013) for areas within 200 kilometers of the coast, and for Nunn and Wantchekon (2011) indicators were added for the three main cities of western Nigeria. Obvious outliers were omitted in two studies: Iceland and Djibouti for Alesina, Giuliano and Nunn (2013), and Hong Kong for Acemoglu, Johnson and Robinson (2002).

Table 3 presents significance levels for tests of zero treatment slope for two inference methods: heteroskedasticity consistent standard errors (with no spatial correlation adjustment), labelled HC; and BCH large clusters. The columns labeled "Estimated" reports p-values from a t-test of zero treatment slope for the spatial basis regressions. The column labeled "Placebo" reports the p-value for the placebo test of no treatment effect, with binary treatments left blank. The columns labeled "Synth-Out' reports the p-value of a test that the outcome is generated by a spatial trend plus noise. The final column of each block labeled "Plac-5%' reports the fraction of placebo simulations that are significant at 5%, using that block's inference method. This diagnostic indicates a problem with the inference method if these rejection frequencies are substantially different from 5%. In particular, if the rejection frequency is well above 5% it raises the concern that the standard errors are too small and confidence intervals are too narrow. The last columns gives the z-score of a Moran test for zero spatial correlation in residuals (using five nearest neighbours). Table 4 lists the BIC choice of tensor dimension and number of principal components along with the number of clusters used.

In the 24 studies with available placebo Monte Carlo simulations, 21 studies had a nominal 5% test placebo rejection frequencies of 4% to 7%. Three studies had 8% rejec-

					Trend R^2		Range 2θ		Structure ρ	
	N	Spl.	PCs	Clust	x	y	x	y	x	y
Acemoglu, Colonial.	64	3	3	4	0.46	0.36	0.15	0.05	0.53	0.39
Acemoglu, Reversal.	40	4	5	4	0.44	0.37	0.25	0.05	0.99	0.96
Acharya, Slavery.	1152	3	2	4	0.54	0.05	0.20	0.10	0.99	0.10
Alesina, Plough.	139	4	7	3	0.74	0.07	0.10	0.10	0.61	0.92
Alsan, Tsetse.	379	6	15	4	0.65	0.04	0.05	0.60	0.42	0.91
Ambrus, Cholera.	357	6	7	4	.	0.13	.	0.15	.	0.95
Arbatli, Conflict.	147	3	3	4	0.80	0.11	0.15	0.05	0.91	0.77
Ashraf, Diversity.	143	4	6	5	0.52	0.62	1.45	0.10	1.00	0.36
Bazzi, Individualism.	2040	5	16	4	0.29	0.35	0.25	0.30	0.96	0.82
Becker, Weber.	452	6	18	5	0.66	0.54	0.05	0.45	0.77	0.91
Buggle, Serfdom.	17155	5	24	5	0.72	0.27	0.10	0.70	0.25	0.20
Caicedo, Mission.	548	6	5	4	0.76	0.03	2.00	0.15	1.00	0.88
Comin, 1000BC.	105	3	5	5	0.61	0.53	0.40	0.20	0.99	0.84
Dell, Mita.	1477	3	4	4	.	0.04	.	0.05	.	0.08
Drelichman, Inquisition.	546	6	6	4	0.08	0.05	0.05	0.10	0.24	0.87
Enke, Kinship.	1216	5	10	5	0.56	0.20	0.05	0.10	0.68	0.72
Galor, Impatience.	87	5	3	4	0.05	0.48	0.30	0.15	0.99	0.66
Giuliano, Climate.	72	3	2	4	0.48	0.34	0.10	0.15	0.78	0.54
Iyer, Raj.	271	4	8	4	.	0.39	.	0.30	.	0.80
Juhasz Blockade.	63	3	3	4	0.90	0.43	2.00	0.05	0.99	0.00
LaPorta, Legal Origins.	130	3	1	4	.	0.11	.	0.05	.	0.37
Michalopoulos, Folklore.	101	4	4	4	0.20	0.41	0.10	0.10	0.79	0.66
Michalopoulos, Precolonial.	683	5	14	5	0.15	0.14	0.05	0.35	1.00	0.60
Montero, Saints.	1625	3	5	4	.	0.21	.	0.20	.	0.55
Nunn, Mistrust.	17644	4	8	4	0.02	0.03	2.00	0.15	0.34	0.11
Nunn, Slavery.	52	4	6	4	0.33	0.30	0.20	0.20	0.99	0.30
Schulz, Kinship.	85	3	6	4	0.29	0.55	0.05	0.90	0.96	0.70
Spolaore, Diffusion.	141	4	6	4	0.59	0.40	0.05	0.20	0.96	0.90
Squicciarini, Devotion.	82	4	6	4	0.72	0.47	0.30	0.10	0.93	0.95
Voigtlaender, Persecution.	320	6	2	4	.	0.16	.	0.15	.	0.50

Spl. gives the number of B-splines in each direction used to construct the tensor and PCs the number of principal components used. Clust is the number of BCH clusters selected by 5% placebos. For the treatment (x) and outcome (y) variables trend R^2 is the fraction of the variance explained by the spatial basis in the case of treatments, and a quadratic in longitude and latitude for outcomes. The remaining statistics are for the residuals from this regression. Range is the effective range 2θ, expressed as a fraction of the 95th percentile of the distance between sites. Structure is the ratio of signal to signal plus noise ρ.

Table 4: Parameters of spatial basis placebos and synthetic outcomes.

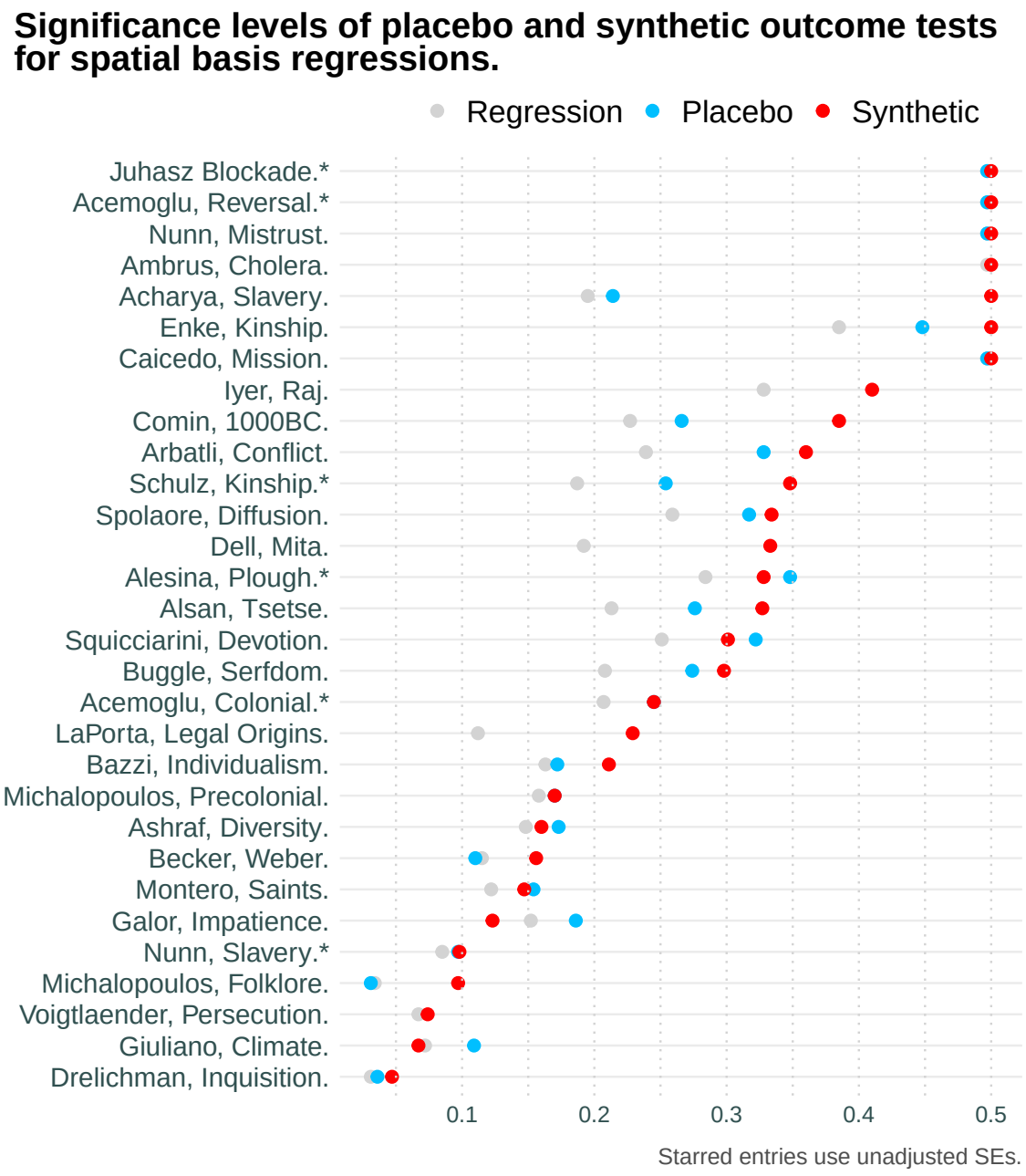

Figure 7: Regression, placebo, and synthetic outcome significance levels for persistence studies. HC inference is used for starred studies and BCH inference is used for the remaining, non-starred studies.

tions and one had 9% rejections. Overall, these placebo simulation rejection rates promote confidence in the BCH inference procedure.

For HC standard errors, 7 studies had placebo test rejection frequencies at 8% or below but of these one had Moran tests with z-scores markedly above two indicating spatial dependence in residuals. In the 6 remaining cases, HC inference appears adequate. This is a result of these studies' conditioning information and/or the presence of spatial basis regressors. When adequate, HC inference is arguably preferable to BCH inference as it may have shorter confidence intervals.

Figure 7 presents our results for treatment slope t-tests and our two diagnostic tests across 30 studies. P-values are from BCH inference in 24 studies, with HC standard errors for the other 6 (indicated by stars). We anticipate that some researchers will want to use our spatial basis regression and HC standard errors as a starting point, then check the placebo and Moran test diagnostics and if the diagnostics appear acceptable, to stick with HC standard errors. We note, however, that BCH returns qualitatively similar significance levels in 5 of 6 cases where we report HC results.[25] So this Figure would be similar if we used BCH throughout. Significance levels for spatial basis regressions are given by the grey dots. P-values for our placebo and synthetic outcome tests are give by blue and red dots, respectively. The diagram is truncated at 0.5 for ease of interpretation.

The results illustrated in Figure 7 are strikingly different from those reported in the original persistence studies. Only two studies have treatment effect slope estimates that are statistically different from zero at 5% and only three more have p-values between 7% and 9%. Our placebo diagnostic tests of no treatment effect return significance levels that qualitatively agree with those for t-tests. The synthetic outcome tests, which are applicable across all studies, echo these results with only one study rejecting the null at 5%. Even at a 10% significance level, only five studies reject the hypothesis that outcomes are spatial trend plus noise with no true treatment effect.

7 Ibragimov-Müller Inference and Coefficient Stability

This Section presents estimates using IM, an alternative large cluster estimator that is conservative but also robust to across-group heterogeneity. IM involves estimating spa-

[25] The one exception has a 9% p-value with HC inference versus a 26% p-value with BCH.

tial basis regressions to get $\hat{\beta}_c$ for each cluster in a set of C clusters. [26] The $\hat{\beta}_c$ are then used as though they were observations in a classical Gaussian linear model with just an intercept of β (a location model). A point estimate of β is the cross-cluster average of these $\hat{\beta}_c$, $\bar{\beta} = \frac{1}{C}\sum_c \hat{\beta}_c$. The usual formula for the sample variance of a mean is utilized, $S = [\frac{1}{C-1}\sum_c(\hat{\beta}_c - \bar{\beta})^2]/C$, to form a t-statistic: $t_{IM} = \frac{\bar{\beta}-\beta_0}{S^{1/2}}$. Ibragimov and Müller (2010) show that inference based on a Student-t distribution with $C-1$ degrees of freedom is valid and conservative even with quite general heterogeneity across clusters for confidence percentages and cluster intervals we use here. So, an IM 95% confidence interval will have 95% or greater coverage probability even with heterogeneity in the variance of $\hat{\beta}_c$ across clusters. This not only provides a complementary, heterogeneity-robust approach to inference but also provides a nice illustration of treatment effect stability (or lack thereof) by simply examining the group-specific estimates: $\hat{\beta}_c$.

Estimating cluster-specific versions of a given study's specification requires a large enough cluster size relative to the complexity of its regression model. Therefore, we only apply IM to the 15 studies with over 250 observations where the full study specification augmented with our spatial basis still has reasonable degrees of freedom within cluster. Table 7 reports results with cluster number again chosen by 5% placebo Monte Carlo rejection results as for BCH. For studies with a binary treatment, four clusters are again used in the Table with full results for 3, 4 and 5 clusters given in Appendix Table 1. We report 95% confidence intervals for each study that generally agree with those for BCH. Only one study has a statistically significant treatment effect significant at 5% with two more significant at 10%. Placebo tests find two studies reject the no treatment effect hypothesis at 5% and another at 10%. Placebo test p-values for the remaining 12 studies generally agree with confidence interval results. Our placebo simulation Monte Carlo indicates approximately correct rejection frequencies for 5% nominal tests across all the studies where this experiment is conducted. Finally, our synthetic outcome test rejects its no treatment effect hypothesis for only two studies at 5% and a total of three at 10%.

Although we cannot implement Ibragimov-Müller inference with the full regression specification for studies with smaller sample sizes, it is still informative to estimate specifications with less conditioning information to illustrate the prevalence of subsample slope estimate variability across the large majority of studies. To conserve degrees of freedom

[26] We use the spatial basis tensor dimension (3x3 to 6x6) and number of PCs chosen by BIC for the full sample, rather than attempt cluster-specific basis choice. We anticipate that our spatial basis will detrend better with a full sample basis choice.

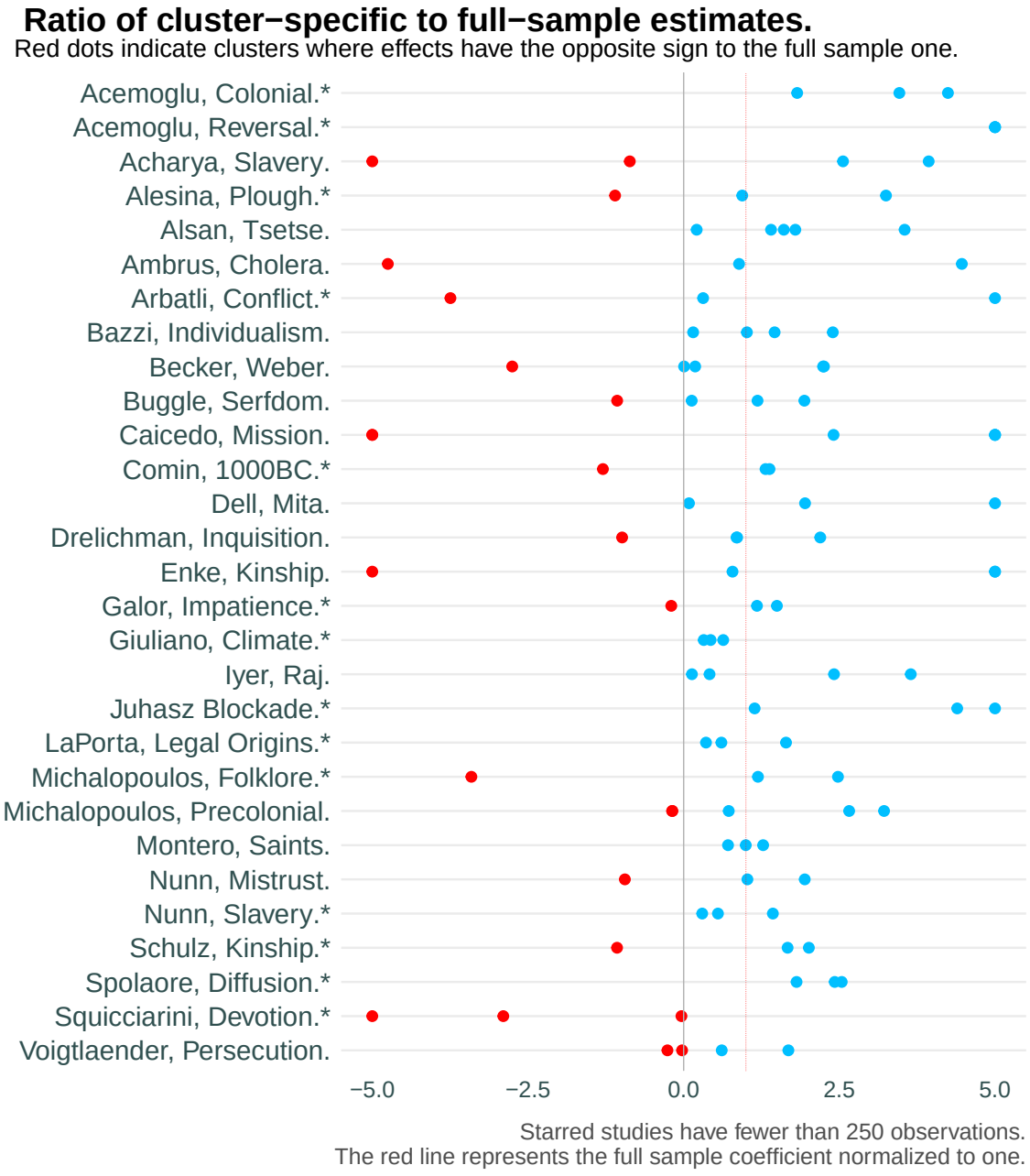

Figure 8: Estimated treatment effects by cluster, expressed as a proportion of the full sample estimate. Starred entries, having fewer than 250 observations, do not appear in Table 5: they are partitioned into three clusters. Red dots denote clusters with the opposite sign to the full sample one. Two thirds of studies at least one region with an opposite treatment effect to the full sample one, and one exhibits Simpson's Paradox: all clusters have the opposite sign to the full sample one. The axis is truncated at plus and minus five.

	CI	IM p			
		Placebo	Plac 5%	Synth Out	Clusters
Acharya, Slavery.	[-0.35, 0.34]	0.97	0.05	0.96	4
Alsan, Tsetse.	[-1.45, -0.1]	0.04	0.05	0.03	5
Ambrus, Cholera.	[-1.2, 0.62]	.	.	0.40	4
Bazzi, Individualism.	[-0.03, 0.31]	0.06	0.04	0.07	4
Becker, Weber.	[-0.12, 0.17]	0.77	0.03	0.92	5
Buggle, Serfdom.	[-0.69, 0.4]	0.47	0.04	0.63	4
Caicedo, Mission.	[-0.15, 0.32]	0.42	0.05	0.48	6
Dell, Mita.	[-22.37, 11.19]	.	.	0.20	4
Drelichman, Inquisition.	[-0.73, 0.35]	0.34	0.05	0.37	4
Enke, Kinship.	[-0.71, 0.56]	0.71	0.04	0.87	4
Iyer, Raj.	[-0.05, 0.22]	.	.	0.14	4
Michalopoulos, Precolonial.	[-0.15, 0.64]	0.15	0.05	0.15	5
Montero, Saints.	[-0.32, -0.06]	.	.	0.02	3
Nunn, Mistrust.	[-0.03, 0.02]	0.53	0.05	0.55	3
Voigtlaender, Persecution.	[-0.02, 0.04]	.	.	0.32	4

Confidence intervals and significance levels for the Ibragimov-Mueller procedure for studies with at least 200 observations. Empty placebo cells correspond to studies with binary treatments. Confidence intervals for coefficients are based on the ratio of the coefficient in each cluster to the full sample one.

Table 5: Ibragimov-Müller confidence intervals and significance levels for regression estimates, and placebo and synthetic outcome tests.

we use three k-medoids clusters and estimate each regression using only the full sample BIC optimal number of spatial basis variables and the treatment variable of interest: other control variables used above are omitted. While in many cases omitting controls does not appreciably impact slope estimates, it is important to note that it can do so and the IM treatment effect estimates may not be comparable to estimates with the full conditioning information. Nevertheless, they illustrate the extent of subsample variability in the study's data in the context of an interesting regression.

Estimated cluster-specific coefficients are shown in Figure 8, with stars denoting studies with fewer than 250 observations, except for two studies with fewer than 60 observations that are not reported and Ashraf and Galor (2013) that does not allow IM inference

because it uses a quadratic in the treatment variable.[27] Full-sample treatment effect estimates are far enough from zero that they can be used as a normalizing factor to make the scale of cluster-specific estimates roughly comparable across studies. Thus, the Figure plots cluster-specific estimates divided by the full-sample treatment effect estimate. It is evident that these ratios often vary markedly by cluster with few studies having all estimates close to unity. In most studies (18 out of 27) at least one regional group has a coefficients of the opposite in sign to the full-sample one. This substantial variation of treatment effects across clusters in Figure 8, suggests that these studies' original reported findings should be treated with caution. It can be seen that one study even displays Simpson's paradox: all regions have the opposite sign to the full sample.

8 Extensions and Conclusions

If your pickaxe unearths a large gold nugget every time it sinks into the ground, you may eventually start to wonder about what sort of gold it is that you are finding in such abundance. This paper originated in a concern that the unusually large t statistics that are the hallmark of persistence literature were less evidence of deep historical processes than statistical artefacts arising from a failure to control for spatial trends and autocorrelation.

Reliable estimation with spatial observations poses two challenges in the form of trends or other large scale structure in the data, and of spatial autocorrelation. Standard error corrections have received considerable attention in the econometrics literature (although recent theoretical advances have been largely ignored by practitioners) but often require that the data have already been detrended, an issue that has been largely neglected until now.

In this paper we introduced a new regression procedure and two diagnostic tests. The regression procedure was to add spatial basis variables to the regression, and then to estimate standard errors using a large cluster procedure. The first diagnostic was a placebo test of no treatment effect using simulated treatments that have approximately the same spatial correlation as the true treatment. This diagnostic is particularly useful in providing a reliable, data driven means to evaluate tuning parameter choice for spatial correlation corrections, in our case the number of chosen large clusters in BCH and IM. The second synthetic outcome diagnostic tested the null hypothesis that the outcome was generated

[27] The cross-cluster variation of coefficients is still pronounced: the linear and quadratic terms for each cluster, expressed as a multiple of the full sample ones, are $(32.9, 1.5, 2.5)$ and $(-3.0, 3.2, 1.7)$ respectively.

was generated by a trend plus spatial noise process, with no true treatment effect. Applying these procedures to 30 historical persistence regressions, few remained significant at even five per cent.

These procedures are applicable to regressions using spatial observations more generally. In particular, modern causal inference techniques are often applied to geographical data, possibly resulting in estimated treatment effects with overstated precision regardless of validity of their identification strategies. Investigating this issue is left for further research.

References

Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 95:1369–1401.

Acemoglu, Daron, Simon Johnson and James A. Robinson. 2002. "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution." *Quarterly Journal of Economics* 117:1231–1294.

Alesina, Alberto, Paola Giuliano and Nathan Nunn. 2013. "On the Origin of Gender Roles: Women and the Plough." *Quarterly Journal of Economics* 128:469–530.

Alsan, Marcella. 2015. "The Effect of the TseTse Fly on African Development." *American Economic Review* 105:382–410.

Ambrus, Attila, Erica Field and Robert Gonzalez. 2020. "Loss in the Time of Cholera: Long Run Impact of a Disease Epidemic on the Urban Landscape." *American Economic Review* 110:475–525.

Arbatli, Cemal Eren, Quamrul H. Ashraf, Oded Galor and Marc Klemp. 2020. "Diversity and Conflict." *Econometrica* 88:727–797.

Ashraf, Quamrul and Oded Galor. 2013. "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development." *American Economic Review* 103:1–46.

Bazzi, Samuel, Martin Fiszbein and Mesay Gebresilasse. 2020. "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States." *Econometrica* 88:2329–2368.

Becker, Sascha O. and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124:531–596.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119:249–275.

Bester, C. Alan, Timothy G. Conley and Christian B. Hansen. 2011. "Inference with Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165:137–151.

Bisin, Alberto and Giovanni Federico. 2021. *Handbook of Historical Economics*. New York: Academic Press.

Canay, Ivan M., Joseph P. Romano and Azeem M. Shaikh. 2017. "Randomization Inference under an Approximate Symmetry Assumption." *Econometrica* 85:1013–1030.

Cao, J., C. Hansen, D. Kozbur and L. Villacorta. 2023. "Inference for Dependent Data with Learned Clusters." *Review of Economics and Statistics* .

Comin, Diego, William Easterly and Erick Gong. 2010. "Was the Wealth of Nations Determined in 1000 BC?" *American Economic Journal: Macroeconomics* 2:65–97.

Conley, Timothy G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92:1–45.

Conley, Timothy G. and E. A. Ligon. 2002. "Economic Distance and Cross-Country Spillovers." *Journal of Economic Growth* 7:157–187.

Conley, Timothy G. and G Topa. 2002. "Socio-economic Distance and Patterns in Unemployment." *Journal of Applied Econometrics* 17:303–327.

Conley, Timothy G., Morgan Kelly and Damian Kozbur. 2024. Improved Spatial Dependence Robust Inference via 'Pre-Whitening'. Working paper ETH Zurich.
URL: *https://econ.uzh.ch/en/people/faculty/kozbur.html*

Conley, Timothy G, Sílvia Gonçalves, Min Seong Kim and Benoit Perron. 2023. "Bootstrap Inference under Cross-Sectional Dependence." *Quantitative Economics* 14:511–569.

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining *Mita*." *Econometrica* 78:1863–1903.

Drelichman, Mauricio, Jordi Videl-Robert and Hans-Joachim Voth. 2021. "The Long-run Effects of Religious Persecution: Evidence from the Spanish Inquisition." *Proceedings of the National Academy of Science* 118:e2022881118.

Enke, Benjamin. 2019. "Kinship, Cooperation, and the Evolution of Moral Systems." *Quarterly Journal of Economics* 129:953–1019.

Galor, Oded and Ömer Özak. 2016. "The Agricultural Origins of Time Preference." *American Economic Review* 106:3064–3103.

Giuliano, Paolo and Nathan Nunn. 2021. "Understanding Cultural Persistence and Change." *Review of Economic Studies* 88:1541–1581.

Gneiting, Tilmann and Peter Gutthorp. 2010. Continuous Parameter Stochastic Process Theory. In *Handbook of Spatial Statistics*, ed. Alan E. Gelfand, Peter Diggle, Peter Guttorp and Montserrat Fuentes. Boca Raton: CRC Press.

Granger, C. W. J. and P. Newbold. 1974. "Spurious Regressions in Econometrics." *Journal of Econometrics* 2:111–120.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second ed. New York: Springer.

Ibragimov, Rustam and Ulrich K. Müller. 2010. "t-Statistic Based Correlation and Heterogeneity Robust Inference." *Journal of Business and Economic Statistics* 28:453–468.

Iyer, Lakshmi. 2010. "Direct versus Indirect Colonial Rule in India: Long Term Consequences." *Review of Economic and Statistics* 92:693–710.

Kim, Min Seong and Yixiao Sun. 2011. "Spatial Heteroskedasticity and Autocorrelation Consistent Estimation of Covariance Matrix." *Journal of Econometrics* 160:349–371.

La Porta, Rafael, Florencio Lopez de Silanes and Andrei Shleifer. 2008. "The Economic Consequences of Legal Origins." *Journal of Economic Literature* 46:285–322.

Michalopoulos, Stelios and Elias Papaioannou. 2013. "Pre-colonial Ethnic Institutions and Contemporary African Development." *Econometrica* 81:113–152.

Michalopoulos, Stelios and Melanie Meng Xue. 2021. "Folklore." *Quarterly Journal of Economics* .

Montero, Eduardo and Dean Yang. 2022. "Religious Festivals and Economic Development: Evidence from the Timing of Mexican Saint Day Festivals." *American Economic Review* 112:3176–3214.

Müller, U. K. and M. W. Watson. 2023. Spatial Unit Roots. Working paper Princeton University.

Müller, Ulrich K. and Mark W. Watson. 2022. "Spatial Correlation Robust Inference." *Econometrica* 90(6):2901–2935.

Nunn, Nathan. 2008. "The Long-term Effects of Africa's Slave Trades." *Quarterly Journal of Economics* 123:139–176.

Nunn, Nathaniel and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101:3221–3252.

Phillips, Peter C.B., Sainan Jin and Ling Hu. 2007. "Nonstationary Discrete Choice: A Corrigendum and Addendum." *Journal of Econometrics* 141:1115–1130.

Schulz, Jonathan, Duman Bahrami-Rad, Jonathan Beauchamp and Joseph Henrich. 2019. "The Church, Intensive Kinship, and Global Psychological Variation." *Science* 366. Issue 6466.

Spolaore, Enrico and Romain Wacziarg. 2009. "The Diffusion of Development." *Quarterly Journal of Economics* 124:469–529.

Squicciarini, Mara. 2020. "Devotion and Development: Religiosity, Education, and Economic Progress in Nineteenth-Century France." *American Economic Review* 110:3454–3491.

Valencia Caicedo, Felipe. 2019. "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America." *Quarterly Journal of Economics* 134:507–556.

Voigtländer, Nico and Hans-Joachim Voth. 2012. "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany." *Quarterly Journal of Economics* 127:1339–1392.

"STUDENT". 1914. "The Elimination of Spurious Correlation Due to Position in Time or Space." *Biometrika* 10(1):179–180.

	BCH Estimate			BCH Synth			IM Confidence Interval			IM Synth		
Clusters	3	4	5	3	4	5	3	4	5	3	4	5
Ambrus	0.52	0.61	0.58	0.68	0.74	0.74	[-30.09, 28.09]	[-4.14, 2.14]	[-2.97, 0.97]	0.9	0.4	0.18
Dell	0.19	0.12	0.06	0.33	0.21	0.15	[-5.12, 3.12]	[-4, 2]	[-3.13, 1.13]	0.21	0.2	0.14
Iyer	0.34	0.33	0.2	0.43	0.41	0.26	[-1.9, 3.9]	[-0.61, 2.61]	[-0.35, 2.35]	0.27	0.14	0.11
LaPorta	0.14	0.07	0.06	0.21	0.12	0.11	.	.	.	.	.	.
Voigtlaender	0.12	0.1	0.09	0.12	0.11	0.08	[-4.61, 6.61]	[-1.75, 3.75]	[-1.77, 3.77]	0.57	0.32	0.4
Montero, Saints.	0.02	0.07	0.02	0.01	0.06	0.02	[-1.57, -0.43]	[-2.16, 0.16]	[-1.76, -0.24]	0.01	0.05	0.01

Regression and synthetic outcome p-values using BCH and IM for studies with binary treatments where placebo significance is not estimated. Results are given for 3, 4 and 5 k-medoids clusters.

Table A.1: Results for studies with binary treatments using 3 to 5 clusters.

Appendix A Basis Regression Results for Studies with Binary Treatments

For the BCH and IM results in Tables 3 and 5 we relied on 5% placebo Monte Carlos to select the optimal number of clusters. However this was not possible with studies that used binary treatments where placebos could not be calculated and in those cases we reported results using four clusters. To examine the robustness of this assumption, Table 1 reports BCH and IM values using from three to five clusters. It is immediately evident that changing the cluster size does not affect the estimates materially.

Appendix B Studies Examined.

Here we give details of the regressions we examined from the papers analysed above. In every case, we chose the column of the table with the maximum number of controls applied, and regional or country indicators if applicable.

Acemoglu, Johnson and Robinson (2001). "The Colonial Origins of Comparative Development: An Empirical Investigation." Table 3.1. Regress property rights on settler mortality.

Acemoglu, Johnson and Robinson (2002). "Reversal of Fortune." Table 3.1. Regress income on urbanization in 1500. Robustness check is to add a dummy for Hong Kong.

Alesina, Giuliano and Nunn (2013). "On the Origin of Gender Roles: Women and the Plough." Table 3.1. Regress women's labour force participation on plough adoption. Robustness check is to omit Iceland and Djibouti.

Alsan (2015). "The Effect of the TseTse Fly on African Development" Table 1.4. Regress historical population density on tsetse fly suitability.

Ambrus, Field and Gonzalez (2020). "Loss in the Time of Cholera: Long Run Impact of a Disease Epidemic on the Urban Landscape" Table 3.4. Regress rents on cholera pump dummy.

Arbatli et al. (2020) "Diversity and Conflict." Table 1.8. Regress civil conflict on diversity.

Ashraf and Galor (2013). "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development" Table 3.6. Regress population density on quadratic diversity testing the hypothesis that they have no joint effect.

Bazzi, Fiszbein and Gebresilasse (2020). "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States" Table 2.2, top panel. Regress uncommon names on total frontier experience.

Becker and Woessmann (2009). "Was Weber Wrong? A Human Capital Theory of Protestant Economic History" Table 3.4. Regress literacy on the percentage Protestant.

Comin, Easterly and Gong (2010). "Was the Wealth of Nations Determined in 1000 BC?" Table 8A.1. Regress income on technology level in 1000 BC. PC number changed from 4 to 5.

Dell (2010). "The Persistent Effects of Peru's Mining *Mita*" Table 2.1, panel three. Regress household consumption on Mita dummy.

Drelichman, Videl-Robert and Voth (2021) "The Long-run Effects of Religious Persecution: Evidence from the Spanish Inquisition." Table 1.1. Regress GDP on inquisitorial intensity.

Enke (2019) "Kinship, Cooperation, and the Evolution of Moral Systems" Table 3.3. Regress kinship tightness on malaria ecology.

Galor and Özak (2016). "The Agricultural Origins of Time Preference" Table 1.2. Regress long term orientation on crop yield. PC number changed from 5 to 3.

Giuliano and Nunn (2021) "Understanding Cultural Persistence and Change" Table 1.2. Regress importance of tradition on climatic instability. PC number changed from 3 to 2.

Iyer (2010) "Direct versus Indirect Colonial Rule in India: Long Term Consequences" Table 3.1. Regress fertiliser usage on British direct rule.

La Porta, de Silanes and Shleifer (2008). "Economic Consequences of Legal Origins" Table 1.A.3. Regress creditor rights on a common law dummy.

Michalopoulos and Papaioannou (2013). "Pre-Colonial Ethnic Institutions" Table 3.8. Regress nighttime illumination on political centralization. Robustness check is to replace log distance from sea with a dummy for sites within 200 kilometres of coast.

Michalopoulos and Xue (2021) "Folklore." Table 5.2. Regress trust on punishment of tricksters. PC number changed from 3 to 4.

Montero and Yang (2022). "Religious Festivals and Economic Development: Evidence from the Timing of Mexican Saint Day Festivals" Table 2A.5 Regress income on festival overlapping planting or harvesting.

Nunn (2008). "The Long Term Effect of the Slave Trade" Table 3.5. Regress GDP per capita on slave exports. The robustness check is to add the share of population at risk of malaria from Ashraf and Galor (2013).

Nunn and Wantchekon (2011). "The Slave Trade and the Origins of Mistrust in Africa" Table 2.3. Regress trust in neigbours on slave exports. Robustness check is to add dummies for three regions (Fon, Yoruba, and Ibo) corresponding roughly to the main cities of western Nigeria: Lagos, Ibadan, and Port Harcourt.

Valencia Caicedo (2019). "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America" Table 2.2. Regress modern literacy on distance from a Jesuit mission.

Voigtländer and Voth (2012). "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany." Table 4.2. Regress Nazi vote share on pogroms. The robustness check is to use a log dependent variable.

Schulz et al. (2019) "The Church, Intensive Kinship, and Global Psychological Variation" Table 2.3, first row. Regress individualism on kinship intensity.

Spolaore and Wacziarg (2009). "The Diffusion of Development" Table 1.3. Regress GDP per capita on genetic distance from the US.

Squicciarini (2020). "Devotion and Development: Religiosity, Education, and Economic Progress in Nineteenth-Century France" Table 3.1. Regress industrial employment on refractory clergy. PC number changed from 5 to 6.

Appendix C Online Appendix. Outline of Method.

Our approach involves four steps that are carried out as one line commands in the `spatInfer` package.

1. **Optimal spatial basis.** From equation (8), using piecewise linear (triangle) B-splines, construct the matrix of basis functions M_r for several tensors of differing dimensions $r = 3, 4, \ldots$. For each M_r, calculate its matrix of principal components P_r. Candidate spatial basis models for each r will consist of sets of the first L_r principal components of M_r for all values of $L_r = 1, ..., r^2$. So there will r^2 models for each value of r. The whole set of models to choose from is the union of all models for all r. Run a set of regressions of an outcome y upon each candidate spatial basis model. Select among these spatial basis models using a BIC criterion.

2. **Placebo Regressions.** Estimate the model in (3) with a mean given by the BIC-optimal combination of tensor and principal components from step 1 via maximum likelihood. Use MLE estimates to generate placebo treatment draws from the estimated distribution. For each simulation, run a regression using the simulated treatment in place of real treatment plus BIC-optimal spatial basis and other control variables. The resulting collection of simulated t-statistics forms the reference distribution for the placebo test of no treatment effect. The fraction of simulations with nominal p-values lower than 5% gives a Monte Carlo test of the inference procedure.

The number of clusters in BCH and IM are chosen as the largest number of clusters to have close to 5% Monte Carlo rejections, we use less than 8% as the operational definition of close. Points are allocated to clusters by k-medoids. Cluster allocations are manually checked to rule out pathological cases, e.g. with European countries, Iceland being a cluster of its own or clusters with large numbers of boundary observations (relative to their interior).[28] Clusters are also checked for having a large enough number of observations for within-cluster averages to be approximately normally distributed. Too small clusters can be aggregated with others or the k-medoids algorithm re-run choosing fewer clusters. If large enough clusters have quite different sample sizes, we recommend using IM.

3. **Synthetic Outcome Regressions.** Regress the true outcomes on a quadratic in longitude and latitude rather than the spatial basis. Obtain MLE estimates of the distribution of residuals just as was done for placebo treatments. Generate draws of synthetic outcomes by taking draws from the estimated residual/error distribution and adding them to the estimated quadratic trend. Then re-run regressions for each simulation and the resulting set of simulation t-statistics forms the reference distribution for the synthetic outcome test of no treatment effect.

4. **Spatial Basis Regression.** From Step 1 we know the BIC-optimal spatial basis P_{rl} and add this to the regression of interest. From Step 2 we know a well-performing choice of BCH or IM clusters so utilize this number of k-medoids clusters in performing large cluster inference.

[28] This could happen for example with sample locations being on three equal-length horizontal line segments with slight perturbations up and down. A k-medoids choice of two clusters could be (a) all points on the top line and randomly about half on the middle segment and (b) all points on the bottom line and the other half of the middle segment points. These two clusters would share a large boundary comprised of middle segment locations.

Appendix D Online Appendix. Diagrams of Changing Basis Dimension for All Studies.

UCD CENTRE FOR ECONOMIC RESEARCH – RECENT WORKING PAPERS

WP23/22 Ronald B. Davies, Dieter F. Kogler, Guohao Yang: 'Construction of a Global Knowledge Input-Output Table' October 2023
WP23/23 Taipeng Li, Lorenzo Trimachi, Rui Xie, Guohao Yang: 'The Unintended Consequences of Trade Protection on the Environment' October 2023
WP23/24 Constantin Bürgi, Mengdi Song: 'Do Professional Forecasters Believe in Uncovered Interest Rate Parity?' November 2023
WP23/25 Morgan Kelly, Kevin Hjortshøj O'Rourke: 'Industrial policy on the frontier: lessons from the first two industrial revolutions' November 2023
WP23/26 Morgan Kelly, Kevin Hjortshøj O'Rourke: 'Industrial policy on the frontier: lessons from the first two industrial revolutions' November 2023
WP23/27 Constantin Bürgi, Prachi Srivastava, Karl Whelan: 'Oil Prices and Inflation Forecasts' November 2023
WP23/28 Cormac Ó Gráda, Chihua Li, Ann Arbor, L. H. Lumey: 'How Much Schizophrenia Do Famines Cause?' December 2023
WP23/29 Judith M. Delaney, Paul J. Devereux:'Gender Differences in Teacher Judgement of Comparative Advantage' December 2023
WP23/30 Vincent Hogan, Patrick Massey: 'Different Strokes: Winning Strategies in Women's (and Men's) Big Bash Cricket.'
WP24/01 Ronald B. Davies, Lena S. Specht: 'Brexit and Foreign Students in Gravity' February 2024
WP24/02 Ronald B. Davies, Guohao Yang: 'A Comparison between Traditional and Knowledge Input Output Tables' February 2024
WP24/03 Tadgh Hegarty, Karl Whelan: 'Comparing Two Methods for Testing the Efficiency of Sports Betting Markets' February 2024
WP24/04 Matthew Amalitinga Abagna, Cecília Hornok, Alina Mulyukova: 'Place-based Policies and Household Wealth in Africa' February 2024
WP24/05 David Madden: 'The Trajectory of Obesity in a Cohort of Irish Children and their Mothers: An Application of Sequence Analysis' March 2024
WP24/06 Aline Bütikofer, Deidre Coy, Orla Doyle, Rita Ginja: 'The Consequences of Miscarriage on Parental Investments' March 2024
WP24/07 Håkan J. Holm, Margaret Samahita, Roel van Veldhuizen, Erik Wengström: 'Anchoring and Subjective Belief Distributions' April 2024
WP24/08 Judith M. Delaney, Paul J. Devereux: 'Gender Differences in Graduate Degree Choices' April 2024
WP24/09 Ciarán Mac Domhnaill: 'All hail? The impact of ride hailing platforms on the use of other transport modes' April 2024
WP24/10 Margaret Samahita: '"Luxury beliefs": Signaling through ideology?' June 2024
WP24/11 Alan de Bromhead,Seán Kenny: 'Irish Regional GDP since Independence' June 2024
WP24/12 Ronald B. Davies, James R. Markusen: 'Capital Ideas: Modelling and Measuring Factors in the Knowledge Capital Model' July 2024
WP24/13 Karl Whelan: 'Samuelson's Fallacy of Large Numbers With Decreasing Absolute Risk Aversion' July 2024
WP24/14 Cormac Ó Gráda: 'H1N1 and WW1: The Spanish Flu and the Great War' July 2024
WP24/15 Benjamin Elsner, Eoin T. Flaherty, Stefanie Haller: 'Brexit Had no Measurable Effect on Irish Exporters' August 2024
WP24/16 Eoin T. Flaherty: 'Are workers with multinational experience a determinant in startup success?' August 2024

UCD Centre for Economic Research
Email economics@ucd.ie