DISCUSSION PAPER SERIES

DP17429

Improved Causal Inference on Spatial Observations: A Smoothing Spline Approach

Morgan Kelly

ECONOMIC HISTORY

ISSN 0265-8003

Improved Causal Inference on Spatial Observations: A Smoothing Spline Approach

Morgan Kelly

Discussion Paper DP17429
Published 03 July 2022
Submitted 01 July 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Improved Causal Inference on Spatial Observations: A Smoothing Spline Approach

Abstract

With geographical observations, nearby places often have very similar treatments, controls, and outcomes. In such cases, even with perfect identification, difference in differences and synthetic controls return imprecise coefficients, while regression discontinuities and instrumental variables are prone to severe bias and spurious significance. This paper shows how this may be remedied by adding a spatial smoothing spline to the regression, something easily implemented in practice. The spline allows spatial structure to be separated out as a nuisance variable while simultaneously improving the bias-variance trade-off for the parameters of interest. For simulations and real examples, including a spline causes a marked shrinkage of coefficients, while standard errors change little for most types of cross-section but fall for panels.

JEL Classification: N/A

Keywords: N/A

Morgan Kelly - morgan.kelly@ucd.ie
University College Dublin and CEPR

Improved Causal Inference on Spatial Observations: A Smoothing Spline Approach

Morgan Kelly[*]

Abstract

With geographical observations, nearby places often have very similar treatments, controls, and outcomes. In such cases, even with perfect identification, difference in differences and synthetic controls return imprecise coefficients, while regression discontinuities and instrumental variables are prone to severe bias and spurious significance. This paper shows how this may be remedied by adding a spatial smoothing spline to the regression, something easily implemented in practice. The spline allows spatial structure to be separated out as a nuisance variable while simultaneously improving the bias-variance trade-off for the parameters of interest. For simulations and real examples, including a spline causes a marked shrinkage of coefficients, while standard errors change little for most types of cross-section but fall for panels.

1 Introduction

Natural experiments have become a cornerstone of empirical analysis in economics and are frequently applied to geographical observations such as cities, regions, and countries. However, when nearby places have similar treatments, controls and outcomes, as they frequently do, these techniques can return estimates of effect sizes that are imprecise at best, even when

[*] University College Dublin and CEPR. I would like to thank Timothy Conley and Benjamin Elsner for helpful conversations. Errors are mine.

their identification restrictions are fully satisfied. In particular, difference in differences and synthetic controls deliver reasonable confidence intervals but coefficient estimates that can be far from true values; while instrumental variables and regression discontinuities ones are prone to spurious significance and seriously biased coefficients. Because these are issues not of identification but of estimation they extend to regressions using spatial observations more generally

It is, of course, well known that the standard errors of spatial regressions need to be adjusted to reflect the fact that effective sample sizes can be smaller than they appear when many observations closely resemble their neighbours. Although a wide range of corrections has appeared in response, in practice their estimates vary too widely to be of practical utility.[1] However, even if the "true" adjustment were somehow known, the whole exercise of correcting standard errors ultimately comes down to wallpapering over the fact that the underlying spatial least squares estimates have a poor bias-variance trade-off to begin with, reflecting their tendency to overfit noise.

If you take some towns dotted across a landscape and represent their incomes by elevation on a map, you will generally find an undulating landscape where rich areas border on rich areas, and poor areas on poor ones.[2] Now take some unrelated variable where neighbour again resembles neighbour. If you regress one variable on the other, hills in one landscape will often match either hills in the other, giving large positive coefficients, or to hollows giving large negative ones. Even though unbiased, coefficients from these regressions will thus have a large variance around the true value of zero. What spatial standard error corrections aim to do is to inflate stand-

[1] These adjustments were pioneered by Conley (1999), and later contributions include the large cluster methods of Bester et al. (2016), Ibragimov and Müller (2010) and Canay, Romano and Shaikh (2017); and the principal components approach of Müller and Watson (2021). However, not only do different adjustments give widely varying estimates, but even minor changes to the tuning parameter (which must be set more or less arbitrarily by the user) of an individual correction can change significance levels substantially, sometimes by an order of magnitude. See Kelly, Mokyr and Ó Gráda (2023) for illustrations of this.

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

ard errors sufficiently to give reasonable confidence intervals around shaky coefficient estimates: to use two wrongs to make a (sort of) right.

Things are often a good deal worse. When two unrelated regressors share similar directional trends (for instance, most indices of development improve as you move away from the equator), coefficients can be severely biased leading to confidence intervals that are centred on wrong values regardless of standard error adjustment, and adding arbitrary polynomials in longitude and latitude generally does little to remedy this.

In brief, then, even when identification restrictions are fully satisfied, when applied to spatially correlated observations the effect sizes and confidence intervals returned by instrumental variables, regression discontinuities, difference in differences, and synthetic controls may be less trustworthy than might be hoped.

Given that these natural experimental techniques are so widely used on spatial data, is there any way to make their results more reliable? The simple solution proposed here is to add a spatial smoothing term to the regression in the form of a thin plate spline. This function has the remarkable property that it gives the optimal least squares fit to any smooth surface of unknown form subject to a second derivative penalty on overfitting whose weight can be determined by cross-validation.

Adding a spline term to a spatial regression offers two potential improvements. The first is that, by separating out the spatial correlation structure of the regression as a nuisance variable (although a very informative one as we will see below), it may be possible to carry out more reliable inference on the parameters of interest. The second is that because smoothing splines are penalized least squares estimators (of a ridge form) these parameters can be estimated with a better variance-bias trade-off than ordinary least squares ones. The procedure extends immediately to panels by adding time to the spline as a third dimension.

Spline regressions are not new to economics, but have been largely neglected for a generation. Shiller (1984) considered splines in the context of smoothing priors; and Engle, Granger, Rice and Weiss (1986) added a smoothing spline to linear regressions of electricity demand to capture seasonal trends of varying forms. Despite their elegance, simplicity and power, splines never took off in economics, perhaps because of their excessive computational burdens by the standards of the time.[3] However, they have continued to be actively developed in statistics under the name Generalized Additive Models (Hastie and Tibshirani 1990, Wood 2017) along with software that makes their estimation extremely straightforward.[4] Because their results are fully interpretable, spline regressions more recently have become a popular tool in machine learning where their predictive power often matches black box methods (James et al., 2021, 289–310).

To assess whether splines really do lead to more reliable estimates for the sort of regressions that interest us, we need to run Monte Carlo simulations. When it comes to spatial observations, generating empirically realistic data is not trivial. Previous simulations in econometrics, such as those used in the standard error correction literature mentioned above, are largely confined to highly stylized AR1 processes on regular lattices. Here, by contrast, we introduce empirically realistic correlation patterns based on Matérn processes that are the basis of empirical geostatistics. To mimic the spatial clumping of actual data, the simulated stochastic processes are sampled at point patterns taken from real locations.

The simulations turn up consistent patterns that also appear in the example regressions reported in the final Section. For cross-sectional regressions, the coverage of least squares deteriorates as the spatial correlation of observations rises, leading to spurious significance. When spatial trends are added to the variables, coefficient estimates become severely biased and coverage declines further. The performance of regression discontinuities and instrumental variables (when there are superficially strong but spuri-

[3] Other nonparametric smoothers have attracted somewhat more interest, kernels especially (Härdle, 1990); besides a literature on nonlinear instrumental variables (Blundell and Powell, 2001).

[4] All estimation here is done with the *R* package `mgcv` of Wood (2017). To add a spline in longitude and latitude to a linear regression of `y` on `x`, the command is $\texttt{gam}(\texttt{y} \sim \texttt{x} + \texttt{s}(\texttt{lon}, \texttt{lat}))$. The package can fit most distributional families in common usage including logistic, zero inflated Poisson, and ordered categorical; has extensive diagnostic and visualization tools; and extends to quantile regressions through the related `qgam` package.

ous instruments that proxy for spatial trends) is especially poor. In all cases, adding a spline leads to accurate coverage and tight, unbiased coefficient estimates.

As we would expect, conventional panel regressions perform better than cross-sectional ones: locational dummies absorb a good deal of spatial structure, leading to unbiased coefficient estimates and decent confidence intervals even when there are long distance trends. However, where least squares estimates again fall down is in their bias-variance trade-off. Coefficients vary widely across simulations leading to effect sizes that can be substantially too big or too small with high probability. The shrinkage induced by a spatio-temporal spline gives considerably tighter coefficient estimates and accurate coverage, along with smaller standard errors.

When added to some well known empirical studies, splines cause the same changes that we observe in the simulations. In cross sections, coefficients shrink markedly. Standard errors stay the same in standard regressions, but rise somewhat for instrumental variables and a lot for regression discontinuities. For panels, both coefficients and standard errors fall notably. Moreover, for most examples analyzed, the splines add considerable explanatory power suggesting that they may be picking up the effects of important explanatory variables with a strong spatial (or spatio-temporal) structure that have been omitted.

In brief, including a smoothing spline would appear to offer a simple means to improve the reliability of cross-sectional and longitudinal regressions on spatial data. Naturally, although splines improve estimation, they still leave the identification necessary to move from regression coefficients to causal effect sizes to be demonstrated by the user.[5]

The rest of the paper is as follows. Splines are introduced in Section 2. Monte Carlo simulations to assess the reliability of spline regressions for cross-sectional and longitudinal observations are presented in Sections 3

[5] One possible contribution of smoothing splines to improved identification may be in absorbing some of the spillovers between locations that can cause spatial data to violate SUTVA, but that is not explored here.

and 4. Finally Section 5 takes some empirical studies to examine how their results change when splines are added.

2 Smoothing Splines

This Section gives a brief overview of smoothing splines: more detailed treatments may be found, for instance, in Hastie, Tibshirani and Friedman (2008, Ch. 5) and Wood (2017, Ch. 5). Beginning in one dimension, there are n observations $\left(y_i, x_i\right), i = 1, \ldots, n$ generated by

$$y_i = f\left(x_i\right) + e_i \tag{1}$$

where f is an unknown smooth function and $E\left(e_i\right) = 0$.

A starting point is to represent f as a sum of basis functions $\{h_j\}_{j=1}^{m+k+1}$ of a pre-specified form

$$f\left(x_i\right) = \sum_{j=1}^{m+k+1} \beta_j h_j\left(x_i\right). \tag{2}$$

allowing (1) to be fitted by least squares. An obvious choice of basis function might appear to the simple polynomial $h_j\left(x_i\right) = x_i^{j-1}$, but the value of every observation then influences the entire estimated curve, even at points distant from it. Better results can be obtained from a spline basis.

The range of x is partitioned into $m + 1$ intervals by choosing m points $l_1, \ldots, l_m$ called knots. A kth order spline is a continuous piecewise polynomial of order k that has continuous derivatives of order $1, \ldots, k-1$ at these knots. The truncated power basis parameterizes the spline at its knots $l_1, \ldots, l_m$ as follows

$$g_1\left(x\right) = 1, g_2(x) = x, \ldots, g_{k+1}\left(x\right) = x^k \tag{3}$$

$$g_{k+1+j}\left(x\right) = \left(x - l_j\right)_+^k, j = 1, \ldots, m.$$

where x_+ is $\max\{x, 0\}$. To improve behaviour around end points, the natural power basis sets the spline to be a polynomial of order $(k-1)/2$ outside the knots. The coefficients $\beta_1, \ldots \beta_k$ can be solved through least squares by minimizing

$$\left(\sum_{i=1}^{n} y_i - \sum_{j=1}^{m+k+1} \beta_j g_j(x_i)\right)^2. \tag{4}$$

Typically $k = 3$, a cubic spline.

These regression splines require the number and position of knots to be set by the user. A fully nonparametric solution is to find among all functions $f(x)$ with continuous second derivatives the one that minimizes the penalized sum of squares

$$\hat{f} = \arg\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(u)\, du. \tag{5}$$

The least squares fit of f is traded off against an over-fitting penalty that rises according to how wiggly (the technical term) the function is. In the case where there is no penalty so $\lambda = 0$, f linearly interpolates the observations, whereas when $\lambda = \infty$, f is the ordinary least squares line. Although the solution to (5) lives on an infinite dimensional functional space, there exists a unique solution which is the natural cubic spline basis with n knots, each at an observation x_i. The solution to (1) can therefore be written as

$$f(x_i) = \sum_{j=1}^{n} \beta_j g_j(x_i). \tag{6}$$

The estimation problem (5) then reduces to finding the parameters β that minimize the penalized objective function

$$(y - B\beta)'(y - B\beta) + \lambda\beta'\Omega\beta \tag{7}$$

where $B_{ij} = g_j(x_i)$ and $\Omega_{jk} = \int g_j''(u)\, g_k''(u)\, du$. The regularization term, which takes the ridge form $\beta'\Omega\beta$, acts to prevent over-fitting by imposing more shrinkage on the coefficients β_j of the wiggliest basis functions g_j according to the size of the penalty parameter λ.

The penalized least squares coefficients are estimated as

$$\hat{\beta} = \left(B^{'}B + \lambda\Omega\right)^{-1} B^{'}y. \tag{8}$$

The smoother matrix is defined as $S_\lambda = B\left(B^{'}B + \lambda\Omega\right)^{-1} B^{'}$ so $\hat{y} = S_\lambda y$. Analogous to ordinary least squares, the effective number of parameters of the smoothing spline is $\mathrm{tr}\left(S_\lambda\right)$. The penalty parameter λ can be estimated by cross-validation or maximum likelihood (Wood, 2017, 255–269).

For spatial observations x_i has two dimensions corresponding to the coordinates in longitude and latitude of each observation $x_i = (s_{i1}, s_{i2})$, and the smoothing spline generalizes to two or more dimensions in the form of the thin plate spline (Wood, 2017, 214–219). A practical limitation of smoothing splines is that they require n knots, so that estimation is slow and uses up all observations. However, a truncated eigen-decomposition can be used to obtain a lower dimensional approximation of B that speeds estimation and leaves degrees of freedom to estimate other model parameters (Wood, 2003).

Such parameters appear when more is known about each location than its spatial coordinates s_{i1}, s_{i2} and outcome y_i. These other covariates X can be included in the regression

$$y = f\left(s_1, s_2\right) + X\delta + e. \tag{9}$$

Following Engle et al. (1986), define a new matrix $W = (B, X)$ and coefficient vector $\gamma^{'} = (\beta, \delta)^{'}$. Letting Λ be the penalty matrix Ω bordered by an appropriate number of zeroes, the solution is

$$\hat{\gamma} = \left(W'W + \lambda\Lambda\right)^{-1} W'y. \tag{10}$$

It is straightforward to allow explanatory variables to have a non-linear impact on y by including them as additional spline terms, but loss of degrees of freedom may start to become an issue for the precision of estimates when datasets are small.

For longitudinal data with observations at times $t = 1, \ldots, T$, the temporal coordinates are measured in different units than the spatial ones so that a thin plate spline in three dimensions is not appropriate. Instead the surface can be fitted as the tensor product of two thin plate splines, in space and time respectively (Wood, 2017, 227-238).

3 Cross-Sectional Simulations

We want to assess how smoothing splines affect regression estimates, and this requires simulations of an empirically realistic structure. For spatially correlated data, the workhorse of geostatistics is the Matérn function, which has the appealing property that the rate of decay in correlation can be made to vary between Gaussian and exponential by varying a smoothing parameter (Gneiting and Gutthorp, 2010). In practice spatial correlation tends to fall off exponentially with distance, and in what follows we will generate data as standard normal variables with mean zero and covariance between sites s_i, s_j at distance h apart equal to

$$\Sigma_{ij} = \rho \exp\left(-h/\theta\right) + \tau^2 \mathbf{1}_{ij} \tag{11}$$

where $\mathbf{1}_{ij} = 1$ when $i = j$ and zero otherwise.[6] The parameter ρ gives systematic correlation while τ^2 represents idiosyncratic variability which equals $1 - \rho$ for the standard normal variables here. The range parameter θ controls how fast correlation decays with distance: correlation reaches about 0.14 at distance 2θ (Gneiting and Gutthorp, 2010).

We will analyse cases where there is strong or weak spatial structure in the variables, with and without systematic trends. The sites are located

[6] For instance, if we take Ttetse fly suitability from Alsan (2015), the likelihood function is maximized with exponential falloff of range $\theta = 1750$ km, and structure $\rho = 1$. For Nazi vote share from Voigtländer and Voth (2012) the respective parameters are 125 km and 0.55. If data are simulated with a slower than exponential decay, the increased spatial structure exacerbates the distortions associated with least squares estimates while leaving the performance of splines unchanged.

on a unit square. In the strong case $\theta = 0.1, \rho = 0.9$, whereas in the weak case $\theta = 0.05, \rho = 0.5$. In simulations where a spatial trend was added to variables, it takes the form of two peaks on a northwest-southeast diagonal from Wood (2003, 104).[7]

We will simulate three types of cross-sectional regression. The first is a standard least squares regression of one simulated noise series on another. Next there is a regression discontinuity where the explanatory variable is set to zero for points in the left half of the sample space, and to one on the right. Finally, there are simulations for instrumental variables regressions where x and y are driven by a common confounder, and there is a strong (but in some cases spurious) instrument for x.

The question arises of how to choose the observational points for sampling the spatial stochastic processes. Possible choices include a uniform grid, or a Poisson process. These tend to give fairly similar results but are empirically unrealistic. In practice sites tend to clump together: cities, for instance, are located disproportionately on coasts or large rivers. This clustering can be modelled, for example, by a Thomas process where imaginary central points are laid down as a Poisson process, and then clusters of observed points are set around each as a second Poisson process. However, we then face the task of choosing suitable generating parameters.

What we do instead is to choose sites based on real world coordinates. For the cross-sectional simulations here we take the location of 150 African tribes (normalized to lie on the unit square) used by Alsan (2015) and others. If we base the simulations on some other set of points, such as the 41 counties of England analyzed by Kelly, Mokyr and Ó Gráda (2023), or the 48 capitals of the contiguous US states, the results are similar but not identical, reflecting the different clustering patterns of the sampling points. For simulations based on uniform grids or homogeneous Poisson processes, standard least squares estimates perform less poorly than with these more realistic patterns.

[7] Using the other trend surface given by Wood did not change the pattern of OLS results materially, and left the spline ones effectively unchanged.

	95% Coverage			Coef Estimate			Coef RMSE		Standard Error		
Degree	Spl	OLS	Cnly	Spl	OLS	Ratio	Spl	OLS	Spl	OLS	Cnly
High Spatial Correlation. No Trend.											
0	0.92	0.64	0.82	-0.00	0.00	0.50	0.09	0.19	0.08	0.08	0.13
1	0.92	0.68	0.81	-0.00	0.00	0.56	0.09	0.17	0.08	0.08	0.11
2	0.92	0.71	0.79	-0.00	0.00	0.60	0.09	0.16	0.08	0.08	0.10
High Spatial Correlation. Trend.											
0	0.90	0.34	0.58	0.03	0.22	0.29	0.09	0.17	0.08	0.08	0.14
1	0.90	0.35	0.56	0.03	0.21	0.32	0.09	0.16	0.08	0.08	0.12
2	0.90	0.38	0.55	0.03	0.20	0.34	0.09	0.15	0.08	0.08	0.12
Low Spatial Correlation. No Trend.											
0	0.93	0.90	0.86	0.00	0.00	0.92	0.09	0.10	0.08	0.08	0.08
1	0.93	0.90	0.86	0.00	0.00	0.93	0.09	0.10	0.08	0.08	0.08
2	0.93	0.91	0.86	0.00	0.00	0.95	0.09	0.10	0.08	0.08	0.08
Low Spatial Correlation. Trend.											
0	0.87	0.34	0.51	0.05	0.20	0.36	0.09	0.10	0.08	0.08	0.10
1	0.87	0.41	0.52	0.05	0.18	0.40	0.09	0.10	0.08	0.08	0.09
2	0.87	0.49	0.58	0.05	0.16	0.45	0.09	0.10	0.08	0.08	0.09

The dependent and explanatory variables are 150 observations of exponentially decaying noise on a unit square. Strong spatial correlations have range θ of 0.1 and structure ρ of 0.9, while weak ones have parameters 0.05 and 0.5. Degree is the degree of polynomials in longitude and latitude added to each OLS regression. The true regression coefficient is zero. Cnly denotes Conley standard errors with a rectangular kernel that has a cutoff distance of 0.1. Ratio denotes the median absolute ratio of the spline to OLS coefficient estimates in each regression. Trend denotes a spatial trend added to both dependent and explantory variables.

Table 1: Monte Carlo simulations of cross-sectional regressions, using OLS and splines.

In the simulations that follow, the dependent variable has mean zero and standard deviation of one, as does the explanatory variable except in cases where it is a binary treatment. The regression coefficients are therefore

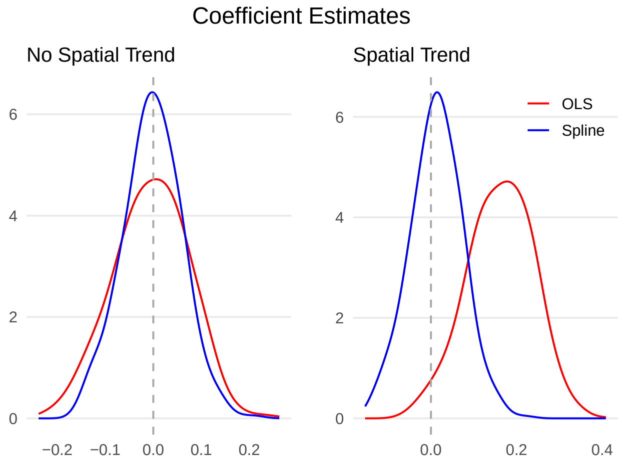


Figure 1: Coefficient estimates from Monte Carlo simulations for least squares and smoothing splines, with and without spatial trends.

the number of standard deviations that y changes by on average when x increases by one standard deviation, or one unit depending on the context.[8]

Each entry of Table 1 compares the performance of least squares, with or without a Conley (1999) standard error adjustment, against a regression that includes a thin plate spline in longitude and latitude. Successive rows give results when no longitude and latitude controls were added to the least squares regression, and then when they were included linearly and quadratically. When trends are added to the simulation these are the diagonal hills from Wood (2003, 104) mentioned earlier.

It is immediately evident that, except when there is a weak spatial structure and no systematic trend, least squares has poor coverage and imprecise coefficient estimates. In the first row, where there is high spatial structure and no trend, with no longitude and latitude terms in the least squares regression, the 95% confidence interval contains the true coefficient of zero only 65 per cent of the time, which improves to 80 per cent with a Conley adjustment, whereas the spline estimate has 92 per cent coverage.

[8] All tables report results for 1000 iterations: their statistics had effectively stabilized within 200.

	95% Coverage			Coef Estimate			Coef RMSE		Standard Error		
Degree	Spl	OLS	Cnly	Spl	OLS	Ratio	Spl	OLS	Spl	OLS	Cnly
High Spatial Correlation. No Trend.											
0	0.93	0.27	0.71	0.01	0.02	0.68	0.34	0.54	0.32	0.10	0.30
1	0.93	0.42	0.71	0.01	0.01	0.51	0.34	0.71	0.32	0.20	0.40
2	0.93	0.44	0.66	0.01	0.01	0.56	0.34	0.64	0.32	0.19	0.34
High Spatial Correlation. Trend.											
0	0.93	0.32	0.74	0.02	0.11	0.69	0.34	0.54	0.32	0.12	0.33
1	0.93	0.41	0.71	0.02	0.30	0.44	0.34	0.71	0.32	0.22	0.46
2	0.93	0.40	0.64	0.02	0.43	0.43	0.34	0.64	0.32	0.22	0.38
Low Spatial Correlation. No Trend.											
0	0.94	0.57	0.86	0.07	0.11	1.49	0.44	0.30	0.43	0.13	0.25
1	0.94	0.57	0.72	0.07	0.30	0.81	0.44	0.50	0.43	0.24	0.36
2	0.94	0.50	0.62	0.07	0.43	0.70	0.44	0.46	0.43	0.24	0.31
Low Spatial Correlation. Trend.											
0	0.94	0.56	0.82	0.00	0.02	1.50	0.44	0.30	0.42	0.11	0.20
1	0.94	0.62	0.71	0.00	0.00	0.88	0.44	0.50	0.42	0.22	0.28
2	0.94	0.65	0.69	0.00	0.00	0.94	0.44	0.46	0.42	0.22	0.26

The simulations are done with the same locations and parameters as Table 1. The explanatory variable is zero at sites to the left of a vertical line halfway across the space, and one to the right.

Table 2: Regression discontinuities.

Most notably, if we compare the ratio of the two coefficient estimates, the spline estimate is half as large on average (reflecting shrinkage due to penalization), while the standard errors returned by both estimators are similar, and about two thirds the size of Conley ones. When a quadratic in longitude and latitude is added, coverage improves slightly to 70 per cent but the precision of coefficient estimates does not improve.

Turning to a trend in both variables, the coverage of OLS is now 0.35, or 0.6 with a Conley adjustment; while the average bias in coefficient estimates is 0.2 and does not improve as directional polynomials are added. In other words, half the time a one standard deviation change in x will appear to cause a change of more than 0.2 standard deviations in y, something that does not occur once in the spline simulations.

The coefficients of the spline regressions are now a third of the least squares ones on average. In the final two panels where there is weak structure, when there is no systematic trend the spline makes little difference as we would expect, but a trend again causes OLS to perform badly. In all cases, the spline estimates are reliable, giving correct coverage and tight coefficient estimates. Figure 1 gives the distribution of the simulated coefficient estimates for the least squares and spline regressions for data with a high spatial structure. It is apparent that the spline estimates are considerably tighter and remain unbiased even when a trend is added, in contrast to the OLS ones.

3.1 Regression Discontinuities

Table 2 reports results for regression discontinuities. where the frontier is a vertical line at 0.5 that splits the study area in half. Using a horizontal line gave similar (but not identical, because the pattern of points on each side of the boundary is different) results. With strong spatial structure and no trend, OLS coverage is only 0.3 but increases to 0.45 as a quadratic polynomial is added, while Conley stays around 0.7. The spline coefficient is between 0.5 and 0.7 of the OLS one. Interestingly, when there is low spatial structure in the observations, OLS continues to perform poorly with coverage of 0.6 while the Conley coverage falls from 0.8 to 0.7 as a quadratic is included. In contrast to Table 1, the spline standard errors are now substantially larger that the OLS ones (something we will also see for real examples in Section 5), and resemble Conley ones. For simulations where a trend is added, the bias of OLS is 0.1 without a polynomial, but rises to 0.4 when a quadratic is included.

Structure		95% Cover.			Coef Estimate			Coef IQR		Std Error		
Space	Trend	Boot	Unadj	LS	Spl	LS	Ratio	Spl	LS	Unadj	LS	F
						Instrumental Variables.						
H	N	0.83	0.86	0.75	0.00	0.01	0.64	0.49	0.77	0.25	0.26	24
H	Y	0.80	0.82	0.33	0.06	0.40	0.39	0.27	0.37	0.14	0.13	93
L	N	0.88	0.85	0.91	0.00	0.01	1.20	0.47	0.38	0.25	0.25	20
L	Y	0.81	0.82	0.31	0.09	0.36	0.44	0.26	0.22	0.14	0.14	81

The simulations are carried out with the same locations and parameters as Table 1. To deal with some extreme coefficients from the IV simulations, reported coefficient estimates and standard errors are the median of simulated values, and the RMSE of coefficient estimates is replaced by inter-quartile range IQR. LS denotes standard IV estimates without a spline term in the second stage. Boot denotes spline coverage estimated from a semiparametric bootstrap, while Unadj uses unadjusted standard errors from the second stage spline regression. F is the median value of the F statistic from the first stage regression.

Table 3: Instrumental variables.

3.2 Instrumental Variables

The impact of including a spline term in IV regressions in shown in Table 3. The starting point is spatial noise simulations η_y, η_x, η_z with the same parameters and locations as in Table 1. There is a confounding variable c that can take two forms. Either it is also spatial noise, or else it is a trend surface of the same form as previously. The regressions involve a dependent variable $y = \eta_y + c$, an explanatory variable $x = \eta_x + c + 0.5z$, and an instrument z which equals η_z when the confounder is spatial noise, and equals $\eta_z + c$ when the confounder is a trend. In other words, x and y are connected only through the confounder. The instrument is valid when the confounder is spatial noise, but spurious when c is a trend, even though it will appear strong in first stage regressions that omit a spline, as we saw in Table 1.

There are possible ways to estimate confidence intervals for the spline regressions: either using the unadjusted values from the second stage regression and ignoring the fact that these are based on estimated values; or by a bootstrap. Because the dependent variable is spatially correlated, the bootstrap cannot simply be cased on random samples. Instead a semiparametric procedure is used where $\hat{y}$ is the predicted value of y from the IV regression, and ϵ is the vector of residuals which should be spatially uncorrelated, something that can be tested by comparing how the difference between each residual and the average of its neighbours (which will be low when there is autocorrelation) changes as the residuals are randomly permuted Wood (2017, 243). A new dependent variable for 500 bootstrap iterations is then computed as the sum of $\hat{y}$ and a sample with replacement from ϵ.

A feature of the IV simulations is that coefficients and standard errors with extreme values were returned by several iterations of the least squares simulations, and a handful of the spline ones. To prevent these outliers from distorting the summary statistics, Table 3 reports the median of coefficient and standard error estimates, along with the interquartile range of the coefficient estimates. The Table reports four estimates, with high and low spatial structure, and with a noise or trend confounder.

It can be seen that the performance of regressions without a spline is similar to the OLS ones in Table 1 without polynomial terms while the spline ones perform slightly worse with coverage in the range 0.82 to 0.88 with unadjusted standard errors, and slightly lower (except in the low structure case) for bootstrap ones. The spline coefficients, moreover, are somewhat biased upwards when there is a trend in the variables, but again far less than standard least squares ones. The median first stage F statistics from the spurious trend instruments are above 80, but including a spline as a first stage check causes the apparent significance of the instruments to vanish.

4 Panel Simulations

Turning to longitudinal simulations, the variable at each site now evolves through time as an AR1 process with parameter α. If W_t is the vector of values at the n sites in period t then

$$W_{t+1} = \alpha W_t + \sqrt{1-\alpha^2} Z_{t+1} \tag{12}$$

where the innovation vector $Z \sim N(0, \Sigma)$ has the same exponential covariance matrix (11) as before. The $\sqrt{1-\alpha^2}$ term serves to make the process stationary in time.

We consider simulations with high and low spatial structure using the same parameter values as in Tables 1 and 2, but now with high and low temporal autocorrelations of $\alpha = 0.9$ and $\alpha = 0.5$.

The longitudinal regressions include dummies for location and time period. Locational dummies turn out to absorb long run directional trends effectively so we will not report simulations with added trend surfaces of the type in the previous cross-sectional tables. For these panel simulations, the sites are now the capitals of the 48 contiguous US states, again normalized to lie on a unit square; and there are ten time periods.

4.1 Fixed Effects and Difference in Differences

Table 4 simulates four types of panel regression. In all cases the outcomes are autocorrelated spatial noise of the form just described. As a benchmark we start with a standard fixed effects regression where the explanatory variable is also spatio-temporal noise with the same generating parameters. Next we consider zero-one treatments where half of the observations are treated simultaneously, in period 6. In each iteration of the simulation, a drawing of cross-sectional noise is first made and the 24 sites with the highest values are treated. When there is high spatial correlation, treated sites will tend to cluster geographically, whereas they are more randomly dispersed when spatial structure is low.

Structure		95% Cover		Coef Estimate			Coef RMSE		Std Error	
Space	Time	Spl	OLS	Spl	OLS	Ratio	Spl	OLS	Spl	OLS
					Fixed Effects.					
H	H	0.91	0.89	0.00	0.00	0.61	0.06	0.10	0.06	0.08
H	L	0.92	0.89	0.00	0.00	0.78	0.06	0.07	0.05	0.06
L	H	0.90	0.95	0.01	0.00	0.78	0.07	0.08	0.06	0.08
L	L	0.92	0.93	0.00	0.00	0.90	0.06	0.06	0.05	0.06
					Simultaneous zero-one treatment					
H	H	0.88	0.91	0.00	-0.00	0.69	0.16	0.22	0.13	0.19
H	L	0.90	0.89	0.00	-0.01	0.91	0.26	0.28	0.21	0.24
L	H	0.87	0.95	-0.00	-0.01	0.96	0.19	0.20	0.14	0.20
L	L	0.86	0.94	0.00	-0.01	1.19	0.30	0.25	0.23	0.24
					Staggered zero-one treatment.					
H	H	0.92	0.93	-0.00	-0.00	0.58	0.09	0.16	0.08	0.15
H	L	0.93	0.93	0.00	-0.00	0.83	0.17	0.21	0.16	0.20
L	H	0.89	0.94	0.00	-0.00	0.77	0.12	0.15	0.10	0.15
L	L	0.88	0.94	0.00	-0.01	1.03	0.22	0.21	0.18	0.20
					Simultaneous variable treatment.					
H	H	0.90	0.87	-0.00	-0.00	0.75	0.09	0.12	0.07	0.10
H	L	0.90	0.86	-0.01	-0.00	0.93	0.15	0.16	0.12	0.12
L	H	0.89	0.94	-0.00	-0.00	0.91	0.10	0.11	0.08	0.10
L	L	0.88	0.92	-0.01	-0.00	1.11	0.16	0.13	0.12	0.12

Here the sites are the coordinates of 48 US state capitals and there are ten time periods. The first column gives the strength of the spatial correlation, high or low, using the same parameters as earlier. The second column gives the strength of temporal autocorrelation, 0.9 or 0.5. The OLS standard error and coverage are based on clustering by location.

Table 4: Panel regressions.

In the third set of simulations, the treatment is introduced sequentially. There is a drawing of spatial noise at the start, and each period three more sites are added to the treatment group in the order of their noise drawing. For the final set of simulations, treatment varies across sites. A drawing of spatial noise is taken, and from periods six to ten this is the treatment administered at each site. Again, the higher the spatial correlation, the more that the treatments of nearby sites will resemble each other.

For each of the treatments, four sets of results are reported according to whether the observations have high or low spatial structure, and high or low temporal autocorrelation. For the least squares results, standard errors clustered by location are reported.

A spline to absorb both spatial and temporal correlation is now required. Following Section 2, because space and time are measured in different units, this is a tensor spline which is the outer product of a thin plate spline in longitude and latitude, and a thin plate spline in time.

As Table 4 shows, the widely used Bertrand, Duflo and Mullainathan (2004) clustering procedure works well everywhere in terms of coverage. The difficulty with least squares comes instead from the imprecision of coefficient estimates, and the inflated standard errors needed to compensate for this. If we look at the second panel, where half the observations are treated at time six, when spatial and temporal correlation are both high, the spline coefficients and standard errors are sixty per cent as large on average as the least squares ones. For cases where the spatial correlation is low, the difference is not as marked regardless of the strength of the temporal correlation. However, in simulations where a staggered treatment is applied, the spline estimates are noticeably more precise than the least squares ones, except when both spatial and temporal structure are weak. In all cases, the spline coefficients estimates come with considerably lower standard errors.

4.2 Synthetic Controls

Finally, Table 5 presents simulation results for synthetic control regressions. Again the outcome y is allowed to have high or low spatial and temporal correlation. The points are the capitals of the 48 contiguous states, and the intervention now occurs in period 11 of 20 on a state that is chosen at random each time. The Table reports effect sizes and standard errors using the synthetic difference in differences estimator of Arkhangelsky et al. (2021): the

Structure		95% Cover.		Coef Estimate			Coef RMSE		Std Error	
Space	Time	Spl	Syn	Spl	Syn	Ratio	Spl	Syn	Spl	Syn
Synthetic Controls.										
H	H	0.92	0.94	-0.03	-0.00	0.62	0.47	0.71	0.45	0.70
H	L	0.93	0.94	-0.03	-0.02	0.90	0.62	0.68	0.58	0.67
L	H	0.92	0.94	-0.04	-0.02	0.72	0.54	0.75	0.53	0.73
L	L	0.92	0.94	-0.04	-0.02	1.05	0.78	0.72	0.72	0.72

The simulations are again based on 48 US state capitals with the same spatio-temporal parameters as Table 4. The coefficient and standard errors for the synthetic control estimators are calculated using the synthetic difference in differences estimator of Arkhangelsky et al. (2021).

Table 5: Synthetic controls.

original Abadie, Diamond and Hainmueller (2010) estimator gave similar results. It can be seen that the coverage of the synthetic difference in differences is good but, again, the spline regressions return tighter estimates of effect sizes and smaller standard errors, except in the final line where there is little spatio-temporal structure when the results are almost indistinguishable as we would expect.

5 Regression Illustrations

We now present some examples of how smoothing splines can change the parameter of interest in a variety of cross-sectional and longitudinal regressions. Table 6 gives the coefficient and standard error for the main regression variable, along with the adjusted R^2 for the least squares regression and the smoothing spline. It is notable that for each sort of regression the impact of the spline is in line with what the simulations above predict suggesting that these are reliable.

	OLS			Spline		
	Coef.	SE	R^2	Coef.	SE	R^2
Cross-sectional						
Ambrus, Cholera.	-0.44	0.09	0.38	-0.15	0.20	0.72
Chetty, Opportunity.	-0.62	0.05	0.36	-0.25	0.03	0.79
Dell, Mita.	-0.26	0.08	0.04	-0.14	0.29	0.10
Instrumental Variables, Stages 1 and 2						
Acharya, Slavery.	0.40	0.04	0.49	0.13	0.05	0.88
	-0.27	0.13	.	0.76	0.28	.
Autor, China.	0.95	0.10	0.44	0.97	0.04	0.45
	-0.89	0.14	.	-0.66	0.08	.
Panel						
Abadie, Smoking.	-15.60	8.71	0.87	-6.10	5.80	0.95
Donaldson, Railroads.	0.16	0.05	0.84	0.07	0.03	0.95
Fetzer, Brexit.	1.45	0.26	0.79	0.44	0.15	0.85
Stevenson, Divorce.	-0.06	0.02	0.68	-0.01	0.02	0.78

OLS gives the original coefficient, standard errors (clustered for panels) and adjusted R^2 reported for the main variable of interest in each study. Spline reports the same statistics after a smoothing spline has been included. For the Abadie Smoking regressions, the first set of coefficients and standard errors were computed using the synthetic difference in differences of Arkhangelsky et al. (2021).

Table 6: Illustrative cross-sectional and panel regressions.

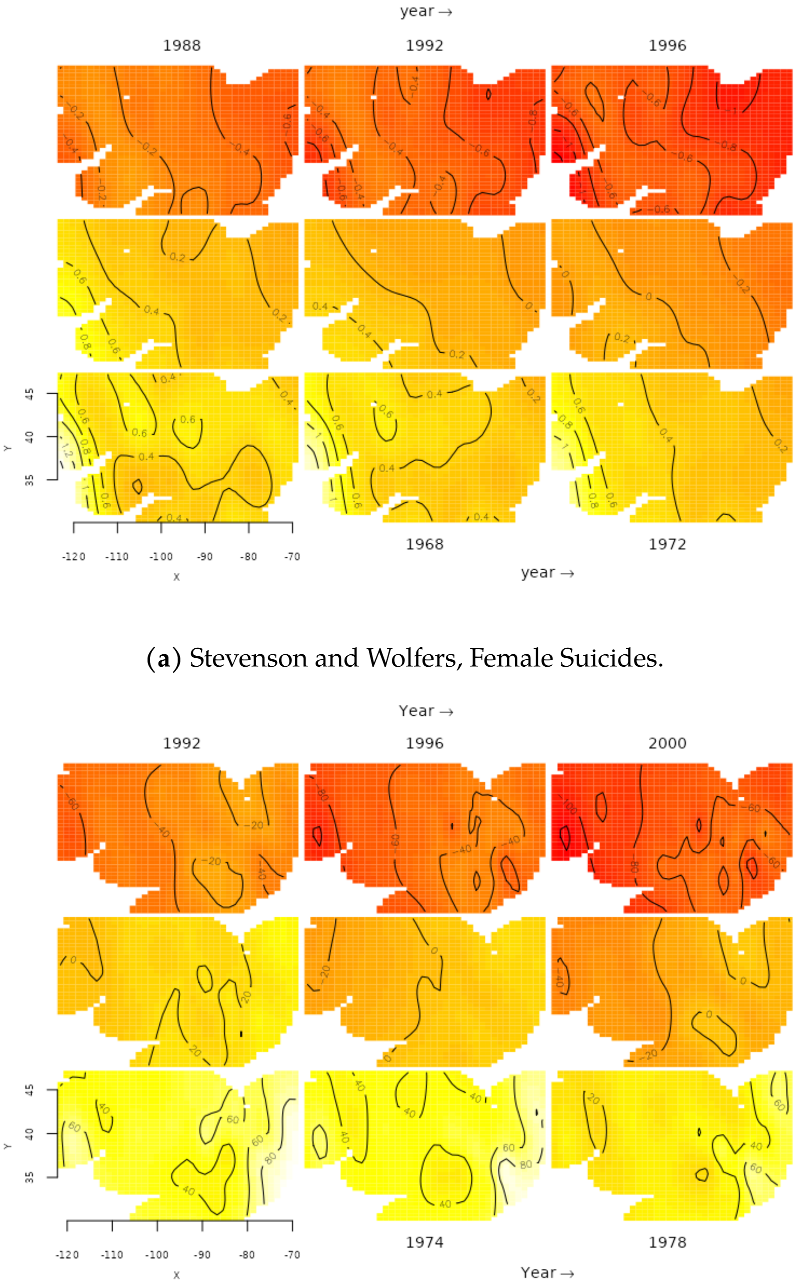
For cross-sectional examples, Table 6 uses the regression discontinuity studies of Ambrus, Field and Gonzalez (2020) and Dell (2010), and the Chetty et al. (2014) “Great Gatsby Curve” linking US income inequality and social mobility. The IV studies are Acharya, Blackwell and Sen (2016) on slavery and Democratic support, and Autor, Dorn and Hanson (2013) on the China Shock. The longitudinal examples are and Abadie, Diamond and Hainmueller (2010) on California’s tobacco control program, re-analyzed by Arkhangelsky et al. (2021); Donaldson (2018) on the impact of Indian railroads on income; Fetzer (2019) on austerity and Brexit; and Stevenson and Wolfers (2006) on unilateral divorce and female suicide.[9]

Looking at the cross-sectional results first, it can be seen that the spline regressions have markedly lower coefficient estimates, while their standard errors tend usually to be somewhat larger than the robust or clustered ones originally used. The addition of the spline term tends to cause a considerable rise in explanatory power suggesting that some important spatially correlated variables may have been omitted from the regressions. Besides the first stage of the Autor et al. regression in Table 6, an example of a spatial study where splines have no impact is Kelly, Mokyr and Ó Gráda (2023).

For the longitudinal regressions, the shrinkage in estimated effect sizes associated with the spline are substantial but, in contrast to the cross-sectional regressions, standard errors also tend to fall considerably. This again is in keeping with the simulations in Tables 4 and 5.

The estimated spline values from the Stevenson and Wolfers (2006) regression are shown in Figure 2. These give the systematic correlation structure of the component of suicides that is not explained by the other variables: divorce, and fixed effects for time and location. There is a marked downward trend, particularly after the mid-1980s and in the west, indicating the absence of important covariates that would explain these. The spline from Abadie et al (2010) is shown in the second panel and again indicates a downward trend everywhere that accelerates in the early 1990s, that is not explained by the existing variables in the regression.

[9] For the cross-sectional studies these are taken respectively from Table 3.4; Table 2.1; and Table 5.1. For instrumental variables these are Table 2.1 and 2.2; and Table 2.1. For longitudinal studies, these are Table 1.1; Table 4.1; Table 1, entry 1 (coefficients multiplied by 100 are reported here); and Table 1.1. For the last study, the public replication files used here gave somewhat different results than the published ones. Replications of Chetty et al. and Stevenson and Wolfers omit Alaska and Hawaii.



(**a**) Stevenson and Wolfers, Female Suicides.

(**b**) Abadie et al, Smoking.

Figure 2: The longitudinal spline components of female suicides from Stevenson and Wolfers (2006), and of smoking from Abadie et al (2010). The downward trend in suicides, particularly after the mid-1980s and in the west; and in smoking, particularly after the early 1990s, is evident and suggests that important explanatory variables are absent from each regression.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105:493–505.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "The Political Legacy of American Slavery." *Journal of Politics* 78:621–641.

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.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111:4088–4188.

Autor, David H., David Dorn and Gordon Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103:2121–2168.

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, Christian B. Hansen and Timothy J. Vogelsang. 2016. "Fixed-b Asymptotics for Spatially Dependent Robust Nonparametric Covariance Matrix Estimators." *Econometric Theory* 32:154–186.

Blundell, Richard and James L. Powell. 2001. Endogeneity in Nonparametric and Semiparametric Regression Models. In *Advances in Economics and*

Econometrics, ed. Mathias Dewatripont, Lars Peter Hansen and Stephen J. Turnovsky. Cambridge: Cambridge University Press.

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

Chetty, Raj, Nathaniel Hendren, Patrick Kline and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129:1553–1623.

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

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

Donaldson, Dave. 2018. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review* 108:899–934.

Engle, Robert F., C. W. J. Granger, John Rice and Andrew Weiss. 1986. "Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Journal of the American Statistical Association* 81:310–320.

Fetzer, Thiemo. 2019. "Did Austerity Cause Brexit." *American Economic Review* 109:3849–3886.

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.

Härdle, Wolfgang. 1990. *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hastie, Trevor and Robert Tibshirani. 1990. *Generalized Additive Models*. New York: Chapman and Hall.

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.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. New York: Springer. 289–310.

Kelly, Morgan, Joel Mokyr and Cormac Ó Gráda. 2023. "The Mechanics of the Industrial Revolution." *Journal of Political Economy* . Forthcoming.

Müller, Ulrich K. and Mark W. Watson. 2021. Spatial Correlation Robust Inference. Working paper. Department of Economics Princeton University.

Shiller, Robert J. 1984. "Smoothness Priors and Nonlinear Regression." *Journal of the American Statistical Association* 79:609–615.

Stevenson, Betsey and Justin Wolfers. 2006. "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress." *Quarterly Journal of Economics* 121:267–288.

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.

Wood, Simon N. 2003. "Thin Plate Regression Splines." *Journal of the Royal Statistical Society. Series B* 65:95–114.

Wood, Simon N. 2017. *Generalized Additive Models*. Boca Raton: CRC Press.