The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers

www.cesifo.org/wp

Spurious Weather Effects

Jo Thori Lind

CESIFO WORKING PAPER NO. 5365
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
MAY 2015

ISSN 2364-1428

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 5365*

Spurious Weather Effects

Abstract

Rainfall is a truly exogeneous variable and hence popular as an instrument for many outcomes. But by its very nature, rainfall in nearby areas tends to be correlated. I show theoretically that if there are also spatial trends in outcomes of interest, this may create spurious correlation. In panel data models where fixed features can be dummied out, the same problem can occur if time trends are spatially dependent. Using Monte Carlo analysis, I show that standard tests can reject true null hypotheses in up to 99% of cases. I also show that this feature is present in a study of the effect of precipitation on electoral turnout in Norway. Using precipitation on non-election days, I show that the distribution of parameter estimates is far away from the theoretical distribution. To solve the problem, I suggest controlling for spatial and spatio-temporal trends using multi-dimensional polynomial approximations.

JEL-Code: C130, C140, C210, D720.

Keywords: rainfall, spurious correlation, spatial correlation, Legendre polynomial.

Jo Thori Lind
Department of Economics
University of Oslo
PB 1095 Blindern
Norway – 0317 Oslo
j.t.lind@econ.uio.no

Monday 18th May, 2015
I got invaluable help with the meteorological data from Ole Einar Tveito. While carrying out this research I have been associated with the center Equality, Social Organization, and Performance (ESOP) at the Department of Economics at the University of Oslo. ESOP is supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 179552.

1 Introduction

In empirical economic research, truly exogenous variables are sought after, as they are a potential sources of exogenous variation which may provide causal inference. One such variable that has captures ample attention is the weather: Few suspect that human actions affect the weather in the short run, and the weather has a potential impact on a number of outcomes.

But by its very nature, rainfall is spatially correlated: If it's raining in one location, the likelihood of rain in nearby areas is high. In this paper I show that this induces a danger of spurious correlations if there are also trends in the outcomes of interest.

In cross sectional data, it is common to observe spatial patterns in many outcomes. When these are regressed on rainfall, the spatial patterns in the two variables is almost always going to coincide in one way or another. Even if there are no real relationship between the two, conventional tests will indicate a relationship. In panel data, where spatial trends can be controlled by fixed effects, the same problem may arise if there are spatially dependent trends in the outcomes of interest.

As an example, consider the relationship between electoral turnout and rainfall. There may be good reasons to expect a relationship between turnout and rainfall *on election day.* But rainfall on other days, with a possible exception of a few days prior to the election, should not have any impact. Using data from Norwegian municipal elections, I study the effect of rainfall on any day in window between 600 days before and 600 days after the election.[1]

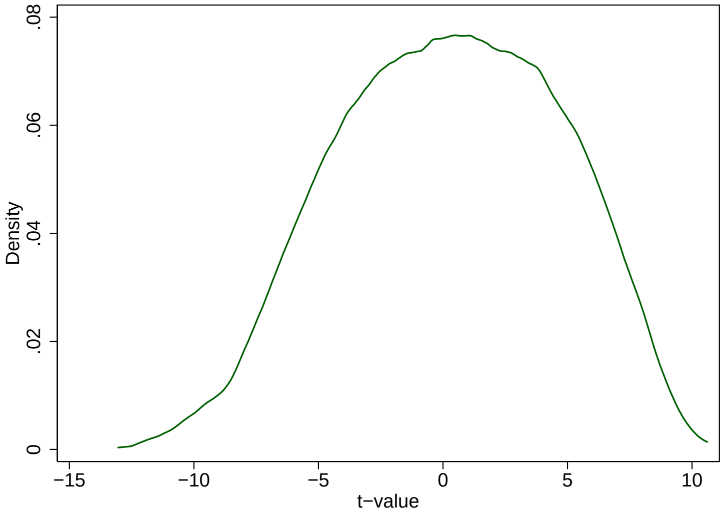
In these analyses we should only expect to find significant results due to the expected Type I errors determined by the level of significance. That is not the case. Rather, a 5 % significance test reject the hypothesis of no effect of precipitation 70.8 % of the cases.[2] The estimated t-values are shown in Figure 1. Although the distribution is symmetric around zero, the variance is much higher than the expected level of unity. Moreover, the distribution is not normal as the tails are lighter than the normal kurtosis.

In this paper, I provide an explanation for such spurious findings. In the case of Norwegian municipal elections, there is a spatio-temporal trend in turnout: in the eastern part turnout has decreased faster than national averages whereas the decline has been less fast in the western part. These trends are not controlled away by two way fixed effects. Moreover, as there is spatial dependency in rainfall data, the probability of generating either positive or negative correlation between the two is high. As trends are common in many types of

[1] The estimation uses data from ten elections between 1971 and 2007 using a two way fixed effects specification. See Section 5.3 for further details.

[2] One explanation for this feature could be that the distribution of precipitation has heavy tails or other irregularities in the data. However, results remain the same if one looks at dummies for precipitation above different thresholds, normalize by municipality means or variances, normalize the turnout variable and so on.

Figure 1: Regression coefficients



Notes: The graph shows the coefficient from two way fixed effects regression of electoral turnout on daily precipitation. Precipitation for 600 days before to 600 days after election day employed, but data from +/- 10 days are excluded.

outcomes, a proper understanding of these complications is key to a proper use of weather data in any analysis.

My suggested solution to the problem is to add a parametric trend. In the cross sectional spatial case, such a trend would be a low dimensional polynomial in geographical coordinates. In the case of panel data, we want a time trend whose slope varies geographically, so the slope of the trend is modeled by a similar polynomial in geographical coordinates. Although any polynomial can in theory be used, sequences of orthogonal polynomials have good numerical stability. In the current study, I focus on tensor products of Legendre polynomials, which seem to perform well.

The use of meteorological data in empirical analyses has skyrocketed in recent years. Some of these take worries of spurious correlations into account by running placebo studies, but far from all. Among the first applications where studies using annual and seasonal weather conditions to study agricultural output and hence serve as an instrument for income – see Dell et al. (2014) for a survey of this literature. More recently, short term weather conditions have also caught researchers' attention.One strand of literature is based on the relationship between weather conditions and people's mood.[3] An early contribution along these lines is Saunders' (1993) finding that US daily stock prices are affected by weather conditions in New York City, where they were traded.[4]

[3] See Cunningham (1979) for a seminal contribution and e.g. Denissen et al. (2008) and Keller et al. (2005) for more recent contributions.

[4] See e.g. Frühwirth and Sögner (2015) for an updated overview of the relationship between financial

Starting with Gomez et al. (2007) and Hansford and Gomez (2010), there is also by now a fairly large literature on the relationship between election day weather and turnout. Beyond the US, the question has been studied in Japan, Holland, Spain, Italy, Sweden, and Norway (Horiuchi and Saito, 2009; Eisinga et al., 2012b,a; Artés, 2014; Sforza, 2013; Lo Prete and Revelli, 2014; Persson et al., 2014; Lind, 2014). In many studies, it is found that rain on election day reduces turnout, but in Sweden there seems to be essentially no relationship between the two and in Norway the relationship is positive. Daily weather conditions have also been found to have an impact on participation in civil rights riots in the 1960s (Collins and Margo, 2007), Tea Party rallies (Madestam et al., 2013), and May day demonstrations (Kurrild-Klitgaard, 2013).

There is also clear evidence that the weather on a specific day affects the labor market: Male works have been found to work more on rainy days (Connolly, 2008) and labor productivity seems to be higher (Lee et al., 2014). Graff Zivin et al. (2015) find that NLSY survey respondents' math performance depends on the temperature on the day of observations. Connolly (2013) shows that answers to well-being surveys are affected by the weather on the interview day. Guven and Hoxha (2014) build on this research and use sunshine as an instrument for happiness to find the effect of happiness on willingness to take risk.

In other studies of the effect of daily weather conditions, Simonsohn (2010) find that the probability of enrollment into colleges is related to cloud cover of the day of visiting the college. Busse et al. (2014) find that the car purchase decisions are affected by daily weather conditions: it is more common to buy convertibles on warm days without rain and 4x4s on cold days with rain or snow. Carr and Doleac (2014) use variation in rainfall in the afternoon on incapacitating potential offenders to derive the causal effect of potential offenders on gun violence. Sen and Yildirim (2015) use rain on a given day as an instrument for the number of readers an online newspaper article gets, based on the idea that potential readers spend more time indoors on rainy days and hence have more time to read online newspapers.

The paper is also related to literature on spurious regressions in time series. In some ways it relates to the presence of spurious regression in regressions with non-stationary variables (Granger and Newbold, 1974; Phillips, 1986). Also, my suggested solution by estimation spatial or spatio-temporal trends relates to the literature on time trends (Sims et al., 1990). As this concerns units in space, it also relates to the massive literature on spatial statistics[5] and the more modest literature on spatial econometrics.[6]

The literature on spatio-temporal statistics has a strong focus on space-time autoregressive moving average (STARMA) type models (Cliff et al., 1975; Pfeifer and Deutsch, 1980), characterized by linear dependence lagged in both space and time. Such models can also

markets and the weather.

[5] Cressie (1993) and Ripley (2004) provide introductions to parts of the literature.

[6] See e.g LeSage and Pace (2009) for an introduction.

be extended to regression frameworks with spatial autoregressive distributed lags models (Elhorst, 2001). Although these models may be suited to handle the problem at hand, their main problem is that they are difficult to identify and estimate by themselves. When we also want to add panel data features, clustered standard errors, instrumental variables or discontinuity designs, they become intractable and not useful for practical applications. Hence I have chosen to rely on a simpler approach.

My suggested solution is to allow for a spatially varying time trend. This relates to the literature on varying coefficients (Hastie and Tibshirani, 1993) and particularly spatially varying coefficients (Gelfand et al., 2003). Specifically, Hoover et al. (1998) and Huang et al. (2002) estimate varying coefficients models where they model the coefficients by regularized basis functions as I suggest (albeit using B-splines rather than polynomial bases).[7] However, they consider coefficients varying in time, not in space. To the best of my knowledge, the only spatial application of the methodology is Zhu et al. (2014) who study MRI images.

Finally, there is a quite substantial literature on spatio-temporal modeling of weather phenomena(Stern and Coe, 1984; Brown et al., 2001; Velarde et al., 2004), but this literature generally has completely different objectives than the current paper.

2 The problem

A simple way to illustrate how the problem of spurious correlation may arise is the following specification: Consider N observations on a line. The explanatory variable is generated by K shocks $\nu_k \sim iid(0, \sigma_\nu^2)$ for $k = 1, \ldots, K$. The location of shock k is $p_k \in [1, N]$. In the case of precipitation, we may think of each shock as a weather system with intensity ν_k and center at p_k. At position i, the total effect of shocks is $r_i = \sum_k \frac{\nu_k}{1+d(i,p_k)}$ where d is a distance function which satisfies $d(i,i) = 0$ and $d(i,j) > 0$ when $i \neq j$. This is essentially a radial basis function network, which is commonly used to approximate functions (Buhmann, 2003). Hence this model should approximate a wide varieties of spatial patterns found in real life. To simplify the analysis, I here focus on $d(i,j) = |i-j|$; for a plane we can use the Euclidean distance.

The outcome variable is

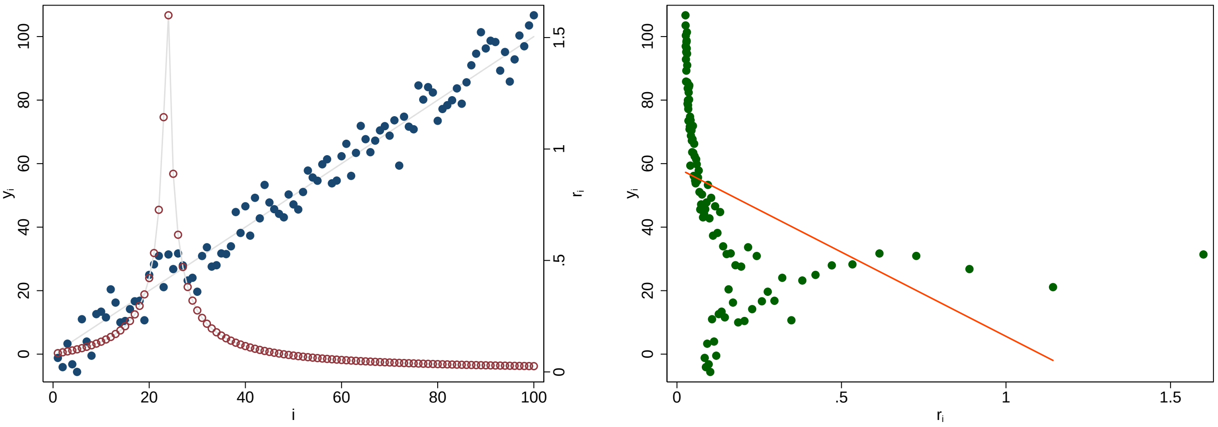
$$y_i = \alpha + \beta r_i + \tau_i + \epsilon_i \tag{1}$$

where τ_i is a trend that we simply define as $\tau_i = \tau i$ for some number τ. We want to test the hypothesis that $\beta = 0$, and the issue is the effect of neglecting the trend τ_i. The core of the problem is that the regression analysis may mistake the trend τ for the signal r_i.

Assume first that $K = 1$, i.e there is only one shock. The situation is illustrated in Figure 2. As is apparent from the figure, whenever the "position" of the shock is $p \neq \frac{N}{2}$, there is

[7] See also Matsui et al. (2011, 2014) for some recent development.

Figure 2: The stylized econometric model



(a) The set up

(b) The relationship

Notes: Panel (a) shows the simulated y_i and $r_i$against the observation number i. Panel (b) shows a scatter plot of y_i versus r_i as well as a linear fit of the data. Data are simulated for $\beta = 0$.

scope for the shock to pick up parts of the trend. I show formally that this is indeed so below. Moreover, as N grows, the problem does not diminish but rather get more acute.

To see the problem formally, consider the situation where the data are generated by (1), but where we fail to control for the trend in the analysis. The OLS estimator then yields

$$\hat{\beta} = \beta + \frac{\sum(r_i - \bar{r})\epsilon_i}{\sum(r_i - \bar{r})^2} + \tau\frac{\sum(r_i - \bar{r})i}{\sum(r_i - \bar{r})^2} \tag{2}$$

In a finite sample with $\epsilon_i \sim NID\,(0, \sigma^2)$, the first fraction has a normal distribution and is handled by ordinary hypothesis testing. [8] The second term, which stems from the omitted variable, is more problematic. In applied research much emphasis is on statistical significance, hence the t-values. Without loss of insight, I assume that σ is known so we can concentrate on z-values. We can split the z-value into to components

$$z = \underbrace{\frac{\sum\left(\frac{1}{1+|P-i|} - \bar{w}\right)\frac{\epsilon_i}{\sigma}}{\sqrt{\sum\left(\frac{1}{1+|P-i|} - \bar{w}\right)^2}}}_{A} + \underbrace{\frac{\tau}{\sigma}\frac{\sum\left(\frac{1}{1+|P-i|} - \bar{w}\right)i}{\sqrt{\sum\left(\frac{1}{1+|P-i|} - \bar{w}\right)^2}}}_{B} \tag{3}$$

The first term, A is a weighted sum of standard normally distributed variables so the first term, $A \sim N(0, 1)$. This is not the case for the second term, B. Here the numerator grows

[8] However, as $N \to +\infty$, we get $\frac{1}{N}\sum(r_i - \bar{r})^2 \to 0$, so even with $\tau = 0$, $\hat{\beta}$ would not be consistent with a single shock. When we let the number of shocks K grow as the sample size grows, this also assures convergence of estimators.

infinitely whereas the denominator goes to zero.

First, I show in Appendix (A.1) that the expression $\frac{1}{N}\sum\left(\frac{1}{|P-i|}-\bar{w}\right)i$ converges to a logarithmic function and hence diverges as $N\to\infty$. The proof is based on showing that the expression can be sandwiched between two harmonic sequences which both have logarithmic growth.

In Appendix A.2, I show that as $N\to\infty$, $\frac{1}{N}\sqrt{\sum\left(\frac{1}{1+|P-i|}-\bar{w}\right)^2}\to 0$. This is based on showing that $\sum\left(\frac{1}{1+|P-i|}-\bar{w}\right)^2$, which is closely related to a sum of the reciprocals of the squares of natural numbers. As $N\to\infty$, this is known to approach to the constant $\frac{\pi^2}{6}$, and similarly the sum at hand also converges to a constant. Consequently, the expression goes to 0 at rate $O\left(\frac{1}{N}\right)$. This leads to the following result:

Proposition 1. *For $K=1$, $B=\frac{\tau}{\sigma}\frac{\sum\left(\frac{1}{1+|P-i|}-\bar{w}\right)i}{\sqrt{\sum\left(\frac{1}{1+|P-i|}-\bar{w}\right)^2}}\to\infty$ as $N\to\infty$.*

It follows that the z-value of the test for $\beta=0$ diverges as $N\to\infty$, even when the true $\beta=0$. Consider next the case of multiple shocks. If there is a fixed number of shocks K, then the conclusions from the analysis above remains essentially unchanged when N becomes large: The numerator of (3) still has logarithmic growth[9] and the denominator goes towards zero. If, however K keeps growing linearly with N, then the situation improves. In this case the denominator converge to a non-zero constant. However, the numerator still diverges.

Table 1 shows a Monte Carlo analysis of the above model for sample sizes between 10 and 10000 and number of shocks varying from 1 to 20000. The simulations are based on a model where the true $\beta=0$ so t-tests should reject at the rate of the test. First, we recognize the diverging t-values: The larger the sample gets, the more likely the t-test is to reject. A test at the 5 % level rejects in about half of the cases for small samples and in more than 80 % of cases in larger samples.[10] Rejections rates and values of $|t|$ are slightly smaller for larger numbers of shocks, but this is not enought to take levels down to reasonable magnitudes.

3 More realistic models

3.1 Spatial models

In many real world applications, the assumption of a linear world is too restrictive.[11] A more realistic assumption is a spatial data structure where it is meaningful to talk about the

[9] There may be shock on either side of $\frac{N}{2}$, but with probability 1 the shows on one side or the other dominate the other.

[10] These numbers could of course be reduced by increasing the noise, i.e. increasing the variance of ϵ_i, but this does not reduce the importance of the problem.

[11] An exception is time series data, but the current modeling of shocks does not seem particularly relevant to that case.

Table 1: A Monte Carlo analysis of the simple model

	K						
N	1	2	5	10	100	N	2N
10	0.59 2.6	0.57 2.6	0.51 2.4	0.49 2.4	0.49 2.4	0.49 2.3	0.51 2.3
50	0.77 3.9	0.67 3.5	0.64 3.4	0.66 3.4	0.62 3.3	0.62 3.3	0.63 3.5
100	0.79 4.3	0.74 4	0.69 3.8	0.68 3.8	0.67 3.8	0.67 3.8	0.66 3.7
1000	0.86 6.2	0.82 6.1	0.82 6	0.80 5.7	0.78 5.6	0.75 5.4	0.82 5.9
10000	0.90 8.8	0.88 8.5	0.86 8.2	0.87 8.3	0.83 7.9	0.86 7.7	0.83 7.9

Notes:The table shows the fraction of cases where a t-test of $\beta = 0$ is rejected at the 5 % level (first line) and the average of the absolute value of the associated t-value (second line). The true model is $\beta = 0$, $\tau = 1$, $\epsilon_i \sim N(0,1)$, and for each k, $\nu_k \sim N(0,1)$ and the position $p_k \sim U(0,N)$. Each model is replicated 1000 times.

distance between two observations, and where units tend to be correlated with nearby units. Denoting observation i's geographical position (x_i, y_i), we can redefine the distance function as $d(i,p) = \sqrt{(x_i - x_p)^2 + (y_i - yx_p)^2}$ and the trend as a spatial trend $\tau_i = \tau_x x_i + \tau_y y_i$ for constants τ_x and τ_y. Such trends, sometimes with more sophisticated specifications, are widespread in geographical data and their study goes at least back to Krumbein (1959; 1963) and Tobler (1969). Without going into the formalism, it is easily seen that this model is essentially equivalent to the model studied in Section 2, and hence that the same problems arise. A Monte Carlo shown in Appendix Table A-1 shows that the problem is indeed still present and if anything stronger than in the basic model.

3.2 Panel data models

In many applications including most of those mentioned in the introduction, we have access to a panel of observations. This allows for controlling for unit fixed effects, which would rule out the problem of the spatial trend τ_i. Time trends are also unproblematic as they are routinely handled by year dummies. But if time trends depend on geography, that is we have spatio-temporal trends, the problem studies above reappears. Consider the case where

$$z_{it} = \alpha_i + \beta \sum r_i + \tau_i t + \epsilon_{it} \tag{4}$$

with the trend $\tau_i = \tau_x x_i + \tau_y y_i$ for constants τ_x and τ_y. If we assume a balanced panel so we can differentiate expression (4), we get

$$\Delta z_{it} = \beta \sum \Delta r_i + \tau_i + \Delta \epsilon_{it}$$

which essentially is specification (1). De-meaning of course yields similar results. The only major difference is that we look at differenced shocks (or deviations from means). However, these have the exact same properties of spatial correlation as the undifferenced shock, so the issues studied in Section 2 still remain.

To see the effect of omitted spatio-temporal trends, Table 2 shows the results from some Monte Carlo simulations of model (4) for different panel lengths, sample sizes, and number of shocks. The conclusions are generally as above – the null hypothesis of no relationship which should have been rejected in 5% of cases is rejected far too often and t-values are typically high. Moreover, the problem is exacerbated by increasing sample sizes. There are some indications that increased panel lengths reduces the problem. As time periods are independent of each other, increasing T increases the (random) variation in Δr_i which helps uncover its independence to Δz_{it}.

4 Detecting and solving the problem

The problem can usually be detected by examining the weather at counterfactual dates as in Figure 1. If rejection rates differ markedly from the expected rates, some spatial or spatio-temporal dependency may be the explanation although of course other explanations obviously also exist. The next step should be to try to get some impression of the spatial dependency. One way to do this is to simply plot maps of spatial values or estimated spatial trends. In some cases it may also be useful to use testing procedures such as Moran's I statistic.

If a spatial pattern is found, two possible solutions can be pursued. The ideal solution is to find the source of the dependency and expand the econometric specification to take this into account. If, for instance, geographically different trends are due to geographical differences in demographic patterns (say young people moving toward large cities), one could potentially solve the problem by adding demographic controls. However, it may not always be easy to find a simple explanation and there may not be a single explanation for the geographical trend. In such cases, it may be a better option to attempt to control for the geo-spatial trend. In the time series literature, this is usually done by simply including the date as a variable, sometimes with a few polynomial terms. In the case of geographical data, this may be too limiting.

What we want is to estimate a function $T(x, y)$. Usually, the shape of T is unknown, so

Table 2: A Monte Carlo analysis of the panel data model

T		2							5						
K		1	2	5	10	100	N	2N	1	2	5	10	100	N	2N
N	9	0.90 4.51	0.83 4.18	0.82 4.23	0.81 4.16	0.82 4.22	0.81 4.11	0.84 4.18	0.58 2.84	0.57 2.87	0.55 2.73	0.51 2.61	0.54 2.67	0.54 2.73	0.56 2.74
	25	0.93 7.23	0.92 6.90	0.88 6.71	0.89 6.54	0.89 6.44	0.88 6.57	0.87 6.52	0.71 4.32	0.73 4.39	0.72 4.52	0.71 4.36	0.71 4.25	0.71 4.23	0.71 4.28
	49	0.94 9.75	0.91 9.16	0.89 8.78	0.90 8.80	0.91 8.94	0.90 8.87	0.89 8.66	0.79 5.98	0.77 5.66	0.75 5.68	0.78 5.73	0.78 5.88	0.76 5.65	0.75 5.57
	100	0.94 13.29	0.94 11.87	0.92 11.84	0.93 11.50	0.91 11.28	0.92 11.45	0.91 11.86	0.84 8.15	0.85 8.32	0.84 7.92	0.85 8.06	0.85 7.90	0.86 7.90	0.84 7.72
	400	0.96 22.06	0.96 20.94	0.94 19.66	0.95 19.16	0.95 19.73	0.94 19.39	0.95 19.69	0.92 14.43	0.91 14.25	0.90 14.21	0.91 14.17	0.91 14.30	0.90 13.55	0.91 14.02
	1024	0.97 31.18	0.96 28.84	0.96 28.25	0.97 28.79	0.97 28.70	0.95 27.38	0.96 27.07	0.93 22.64	0.94 20.83	0.94 21.14	0.94 21.14	0.94 21.49	0.94 20.80	0.94 21.08
	10000	0.99 76.86	0.99 70.88	0.98 65.82	0.98 68.79	0.98 68.76	0.99 67.55	0.98 66.95	0.97 59.49	0.98 57.21	0.98 58.50	0.97 55.25	0.98 56.76	0.98 54.96	0.98 57.06

T		10							20						
K		1	2	5	10	100	N	2N	1	2	5	10	100	N	2N
N	9	0.49 2.40	0.48 2.28	0.51 2.42	0.52 2.43	0.51 2.44	0.49 2.37	0.50 2.42	0.49 2.21	0.48 2.30	0.49 2.30	0.50 2.23	0.48 2.24	0.47 2.21	0.50 2.31
	25	0.69 3.87	0.67 3.89	0.67 3.91	0.66 3.66	0.69 3.77	0.67 3.75	0.66 3.67	0.66 3.55	0.66 3.63	0.68 3.68	0.66 3.66	0.65 3.61	0.69 3.72	0.64 3.57
	49	0.76 5.09	0.73 5.01	0.76 5.13	0.74 5.09	0.73 5.04	0.78 5.41	0.75 5.02	0.73 4.74	0.73 4.72	0.74 5.03	0.75 4.92	0.73 4.75	0.76 4.96	0.74 5.05
	100	0.82 7.09	0.82 6.78	0.85 7.70	0.82 7.30	0.82 7.16	0.82 7.43	0.81 7.01	0.82 7.07	0.83 6.76	0.81 6.72	0.83 6.97	0.81 6.61	0.83 6.77	0.82 7.03
	400	0.92 14.32	0.90 12.66	0.90 13.23	0.92 13.55	0.91 13.30	0.91 13.23	0.92 12.84	0.92 13.27	0.91 12.24	0.90 12.47	0.90 12.80	0.89 12.16	0.90 13.02	0.89 12.07
	1024	0.94 20.93	0.94 20.36	0.94 20.47	0.93 19.70	0.93 19.85	0.94 18.82	0.95 20.23	0.95 19.54	0.94 19.17	0.93 19.44	0.92 19.41	0.94 18.31	0.94 18.95	0.92 18.06
	10000	0.97 53.43	0.98 54.44	0.97 55.01	0.98 52.13	0.97 53.25	0.97 52.16	0.98 53.80	0.97 53.64	0.99 53.30	0.98 52.11	0.98 51.07	0.98 51.52	0.98 51.83	0.98 51.21

Notes: The Table shows Monte Carlo simulations of model (4) for different panel lengths T, sample sizes N, and number of shocks K. The table shows the fraction of cases where a t-test of $\beta = 0$ is rejected at the 5 % level (first line) and the average of the absolute value of the associated t-value (second line). The true model is $\beta = 0$, $\tau = 1$, $\epsilon_i \sim N(0,1)$, and for each k, $\nu_k \sim N(0,1)$ and the position $p_k \sim U([0,N]^2)$. Each model is replicated 1000 times.

a flexible estimator in two-dimensional space is called for. Kernel based and other standard non-parametric estimators are computationally intensive, and as their rate of convergence is typically below $\sqrt{n}$, inference of the other variables in the regression can't always be made using standard techniques. Consequently, a simpler form may be advisable in many cases.

In the case of a panel, we to estimate a function $T(x, y, t)$. As this is a function of three variables, a fully flexible non-parametric approach gets even more demanding. At least for short panels, it seems reasonable that the trend may be kept linear, so we can rewrite $T(x, y, t) = P(x, y)t$ for some function P. One solution that seems to work well for the electoral turnout data considered below is one where P is specified as a tensor product of Legendre polynomials.[12] The choice of orthogonal polynomials is to reduce problems of multicollinearity and improve numerical stability. One justification for choosing Legendre polynomials is their orthogonality property with regard to an L^2 inner product given a uniform spatial distribution of units. Although the distribution is not exactly uniform, this approach is likely to give better behavior than most other orthogonal polynomial bases that provide orthogonality given various bell shaped distributions.

Given dimensionalities K and L, we then specify

$$T(x, y, t) = t\sum_{k=0}^{K}\sum_{\ell=0}^{L}\theta_{k\ell}P_k(x)\,P_\ell(y) \tag{5}$$

where $P_i(\cdot)$ is the i'th order Legendre polynomial. [13] The $(K+1)(L+1)$ parameters $\theta_{k\ell}$ can be estimated together with the other parameters in an ordinary regression model.

The choice of the dimensions K and L has to be chosen to make the polynomial (5) provide a reasonable fit of the data. If K and L are chosen too high, there is both a danger of over fitting Hastie et al. (2008, Ch. 7) and loosing so much variation that it becomes impossible to identify the effect of the variable of interest. Hence we want would like a good fit with a low dimensional polynomial. To make a good trade off, I recommend to consider choosing K and L by maximizing a linear penalty function

$$R^2 - \xi(K+1)(L+1) \tag{6}$$

where ξ is a penalty for more parameters to estimate. This is closely related to maximizing the AIC and BIC criteria, but for varying penalties for the degrees of freedom. Varying the parameter ξ, we can trace out the class of potentially good polynomial compositions. It is also important to undertake counter factual estimations as in Figure 1 to check that

[12] See e.g. Judd (1998, Ch. 6) for an overview of Legendre polynomials and other polynomial basis with applications in economics and Totik (2005) for the mathematical background.

[13] These polynomials are usually defined recursively with $P_0(x) = 1$, $P_1(x) = x$, and for $i \geq 2$, $P_i(x) = [(2i-1)xP_{i-1}(x) - (i-1)P_{i-1}(x)]/i$ where the variable x is normalized to be in the interval $[-1, 1]$.

the polynomial at hand actually solves the problem. If the fit is good enough, most of the placebo variables should have little effect on the outcome. Another approach could also be to choose K and L high, but constrain the $\theta_{k\ell}$ by employing ridge regression, LASSO, or other versions of constrained estimation (Belloni et al., 2014; Hastie et al., 2008, Ch. 3).

5 Turnout in Norwegian elections

5.1 Data

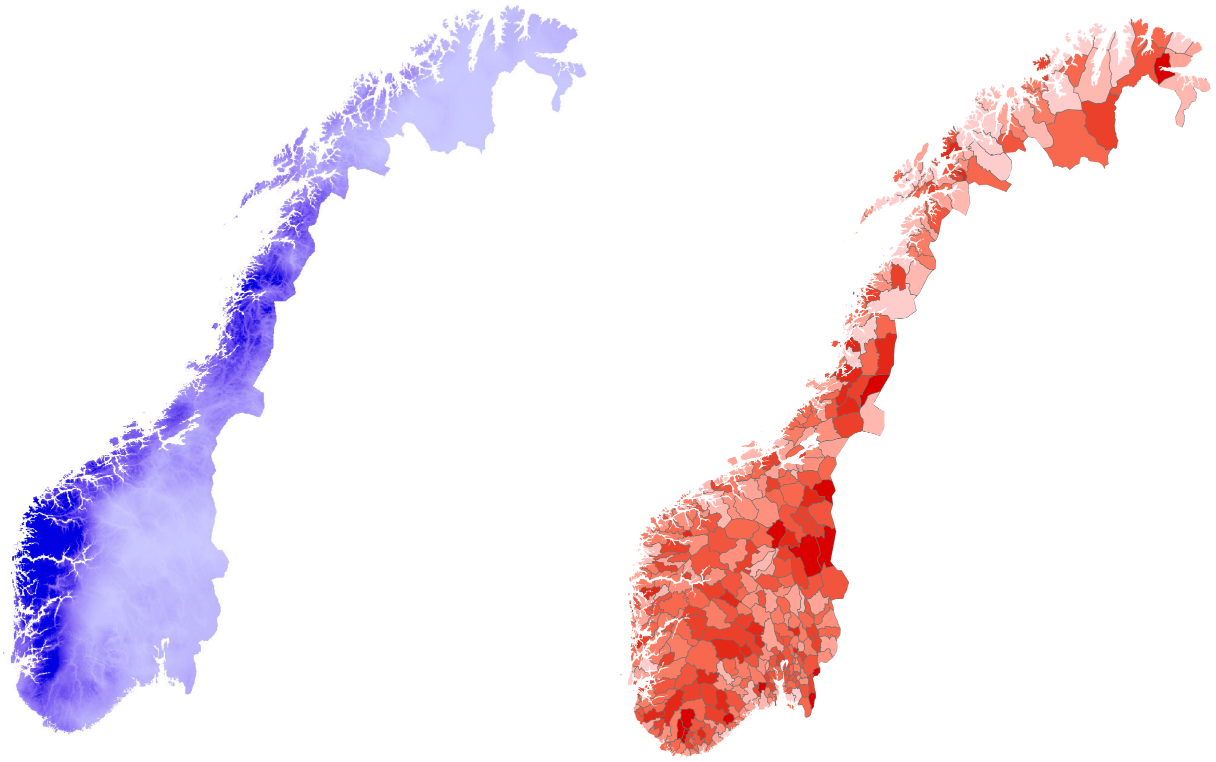
As an application, consider the relationship between electoral turnout and rain considered, that was also discussed in the Introduction. The meteorological data for this application are created by the Norwegian Meteorological Institute (met.no). The data are based on daily observations of precipitation at all 421 measurement stations in Norway, and based on spatial interpolation using a residual kriging approach is applied Tveito and Førland (1999). First, each observation is regressed on a number of geographic properties to separate between a deterministic and a stochastic part. The residuals are then interpolated using kriging and combined with deterministic parts to obtain a grid of 1 × 1 km cells for Norway. Average precitipation values on election days are shown in Panel (a) of Figure 3. As one would expect, average rainfall is larger along the west coast and in parts of the north.

I combine these data with GIS data on municipal boundaries to construct data on average precipitation by municipality for each election year. Municipal boundaries have changed over time, and GIS data on past municipal borders are essentially non-existent. To solve this I map municipalities that no longer exist into their current municipality and use weather data from the present day municipality. Data on electoral turnout taken from the recent collection of Norwegian municipal data made available by Fiva et al. (2012), originating from Statistics Norway and the Norwegian Social Science Data Services. See Lind (2014) for full details of the data used. Panel (b) of Figure 3 show the average election day precipitation and turnout for the period 1971-2007. There are no clear geographical trends in average turnout.

5.2 The spatio-temporal trend in turnout

As was already mentioned in the Introduction, when we do not control for spatio-temporal trends we typically get large t-values when regressing turnout on precipitation both on election day and almost any other day. One explanation for this finding could be outliers in precipitation, which is well know to have a heavy right tail, and turnout. To show that this cannot be the sole explanation, Figure 4 shows the distribution of the t-values in a number of specifications that reduces the leverage of outliers. Panel (a) is the specification shown in the introduction, where the level of turnout is regressed on the level of rain in millimeters.

Figure 3: Spatial averages

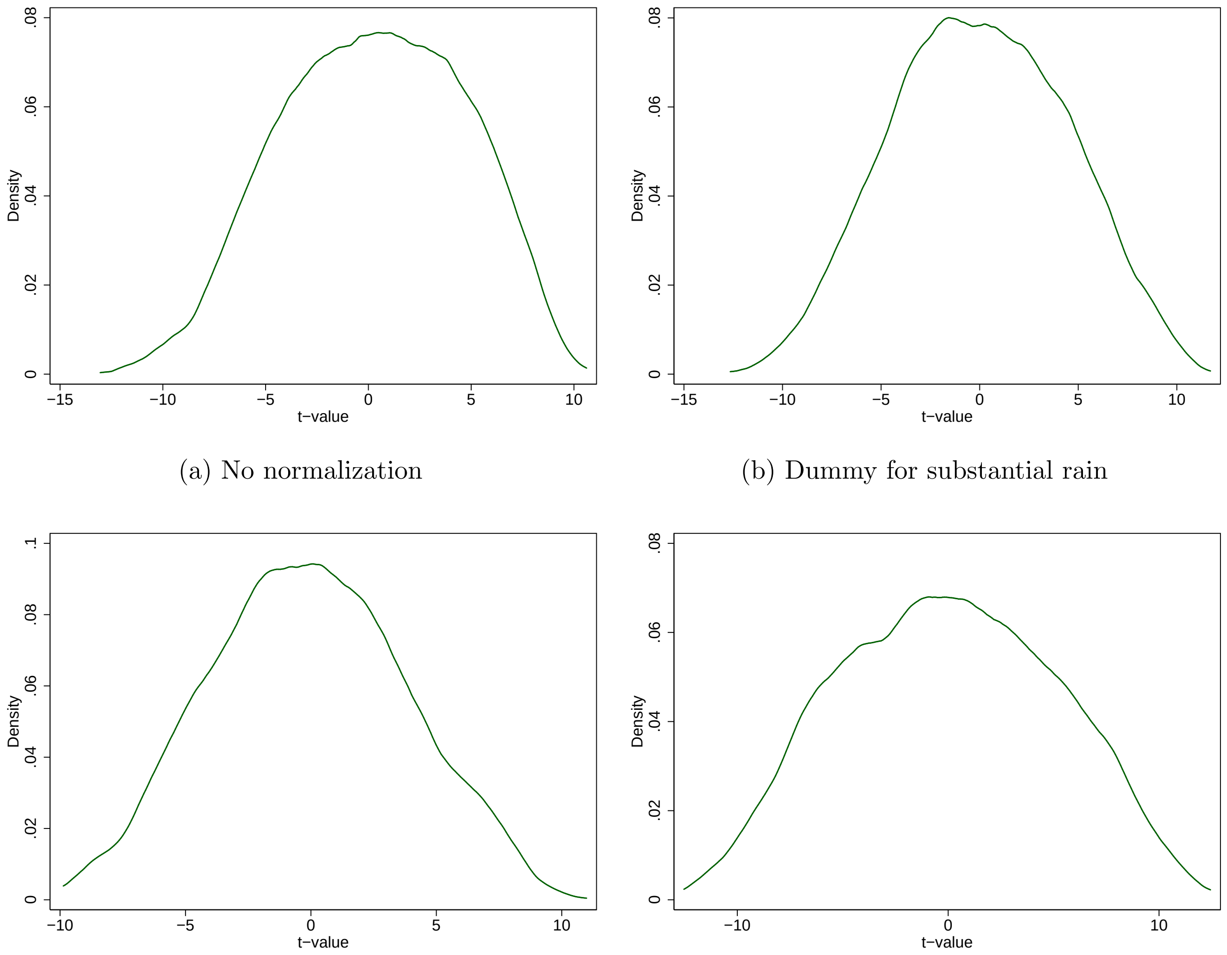


(a) Precipitation

(b) Turnout

Notes: Panel (a) shows average precipitation on election day, averaged over the elections 1971-2007. Dark colors indicate high levels of precipitation. Panel (b) shows municipal average turnout in the same elections. Dark colors indicate high turnout.

Figure 4: Distribution of the t-values



(a) No normalization

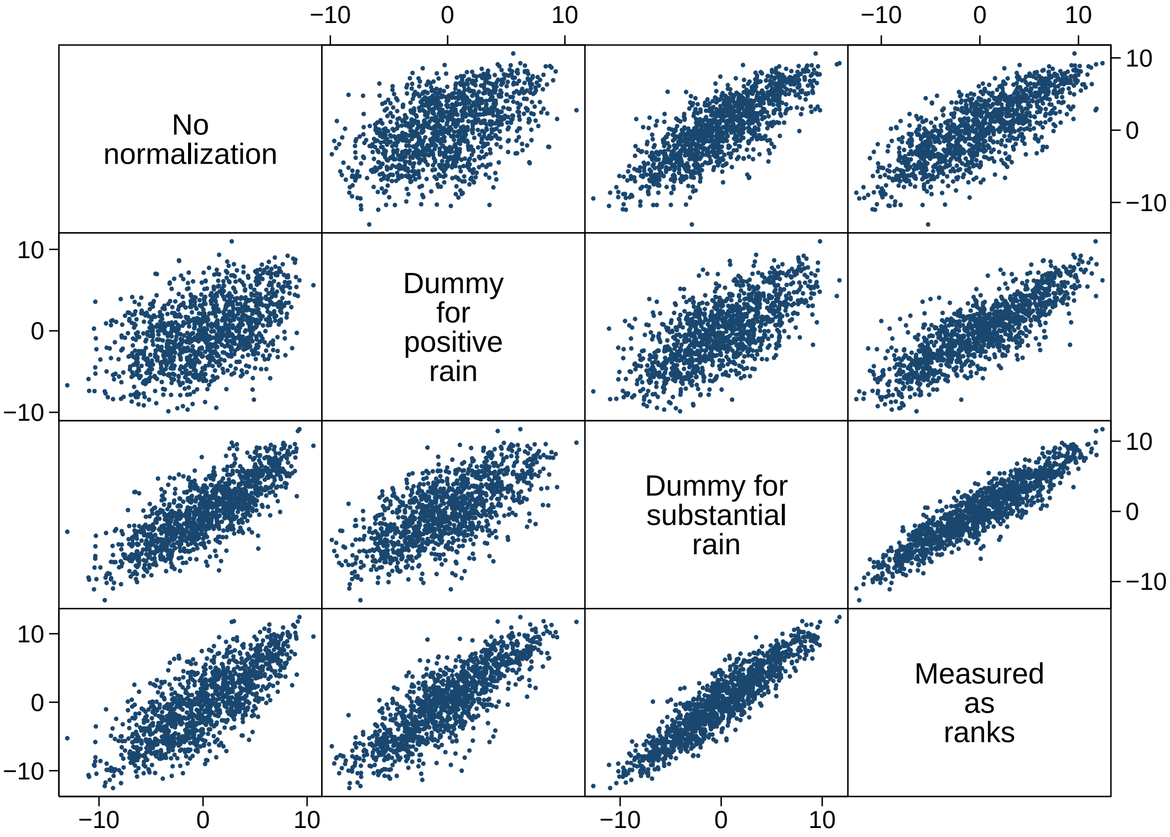
(b) Dummy for substantial rain

(c) Dummy for positive rain

(d) Measured as ranks

Notes: The graph shows the distribution of the t-values when regressing municipal turnout on daily precipitation for 600 days before and after election day. The 10 days before and after the actual election day are omitted. Panel (a) shows results from regressing levels on levels. Panel (b) shows the regression of turnout on a dummy for more than 25 mm rain while Panel (c) employs a dummy for any rain. Panel (d) shows results from a regression where the rank of turnout is measured on the rank of rain, i.e. both variables are uniform on the unit interval.

Figure 5: Association between the t-values in the different specifications



Notes: The graph shows the association between the t-values when regressing municipal turnout on daily precipitation for 600 days before and after election day using four different specifications. The 10 days before and after the actual election day are omitted.

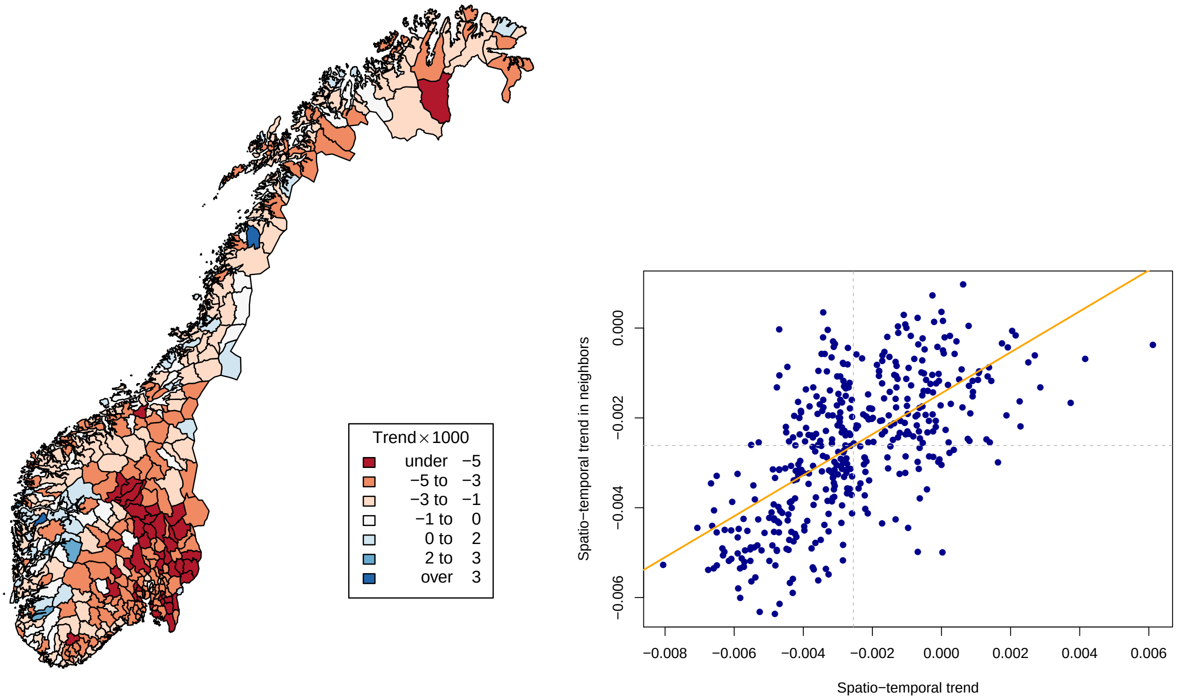
Panels (b) and (c) replace the measure of precipitation with dummies for substantial rain, defined as above 2.5 mm, and any rain at all. Finally, in Panel (d) both rainfall and turnout are measured using their ranks so they both have a uniform distribution on the unit interval. In all four cases, the distribution is far from the standard normal or t-distributions we would expect. The four measured t-values are indeed heavily correlated as seen from the matrix plot in Figure 5. This should indicate that mere outliers cannot explain the findings.

There is indeed strong spatio-temporal trends in the turnout data. Figure 6 shows the municipality specific coefficients δ_i from a regression of the type

$$Turnout_{it} = \alpha_i + \tau_t + \delta_i t + \epsilon_{it} \tag{7}$$

Panel (a) shows the geographical distribution of temporal trends. It is clear that there is a strong negative trend in the eastern part of the country and a positive trend in parts of the

Figure 6: Spatio-temporal patterns in turnout



(a) Trends by municipality

(b) Trends in neighboring municipalities

Notes: The figure shows municipality specific coefficients δ_i *from the regression* (7). *Red areas are strong negative, blue areas strong positive.*

west and the center. Panel (b) shows a Moran plot where the municipality specific coefficient δ_i is plotted against the average δ_i in the adjacent municipalities. Again it is clear that there is a spatial pattern. Formally, Moran's I statistic is $I = 0.456$ and the Moran test for no spatial dependency rejects with a p-value of 2.2×10^{-16}. We conclude that when controlling for two way fixed effects, turnout has been declining in the eastern part of the country and increasing in the western part. As shown in Sections 2 and 3, this can explain the t-values shown in Figure 4.

5.3 Controlling for the spatio-temporal trend

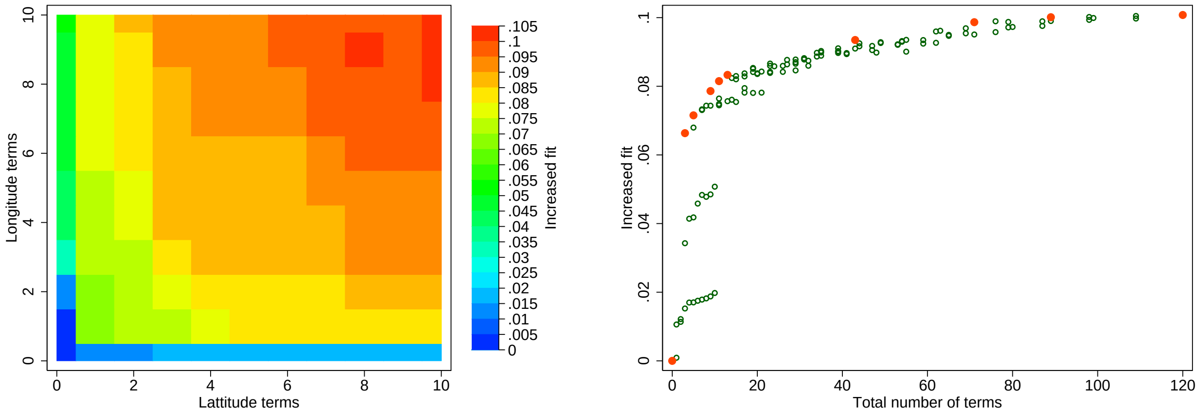
As argued in Section 4, one way to handle the problem of spation-temporal trends is to control them out in the estimation. I approximate the trend with the tensor product of Legendre polynomials. The first step needed is to make a choice of how many polynomial terms to include in each of the two dimensions. Figure 7 shows the the model's fit (net of a baseline model without spatio-temporal controls) for each combination of between 0 and 10 terms in each dimensions. Combinations of polynomial orders K and L that are maxima of the penalized model (6) for some value of ξ, i.e. those which are elements of the convex hull of the points, are shown in red. There is a strong increase in fit going up to about 15 terms, then the effect of additional terms seems to flatten out. To avoid over fitting the data an preserve some degrees of freedom, my preferred model specifies spatio-temporal trends using a first order polynomial in the longitude and a sixth order polynomial in the latitude, using 13 terms and increasing the fit as measured by R^2 by 0.083.

Adding more terms not only have a minor impact on the model's fit, it turns out that the exact specification of the spatio-temporal has little importance once we reach a minimum level of complexity. Figure 8 shows the distribution of t-values for six specification with increasing complexity of the tensor product of Legendre polynomials and with linear and quadratic time trends. The distributions are almost perfectly overlapping for each of the four models. Indeed, the correlation between the most and the least complex models are between .85 and .9.

Moreover, we notice that the distribution of t-values is much more well behaved than the extreme values found in Figure 4. The distribution is somewhat fatter than the theoretical Student's t distribution. Still, the distribution is much more sensible t work with.

Table 3 shows estimation results from the preferred specification. The general pattern is that rain seems to increase turnout in Norway – see Lind (2014) for a discussion of the rationale behind this. Column (1) shows the plain regression of turnout on precipitation in cm. The effect of 1 cm increase in precipitation is about .3 percentage point increase in turnout. Columns (2) and (3) turns the attention to dummies for positive rain and substantial rain, defined as above 2.5 mm. Comparing elections with and without rain,

Figure 7: The number of terms in the nonparametric trend model



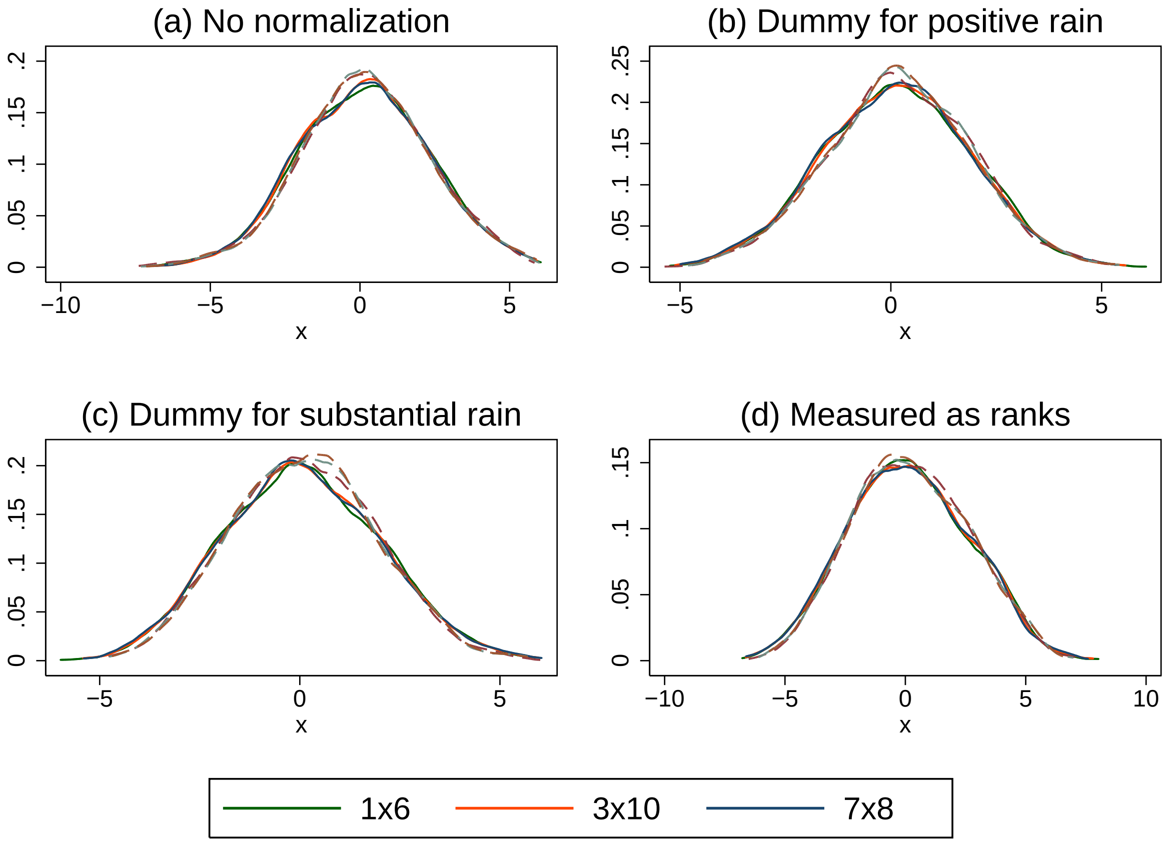
(a) Model fit and number of longiude and latitude terms

(b) Model fit and total number of terms

Notes: Panel (a) shows model fit as a function of the number of terms in the longitudinal and latitudinal polynomials, whereas Panel (b) shows fit as a function of the total number of terms included in the tensor product. Approximation is with tensor products of Legendre polynomials of varying degrees. In Panel (b), combinations that belong to the convex hull are shown with solid orange dots and other combinations with hollow green dots.

turnout is about .5 to .7 percentage points higher in the former. Columns (5) and (6) tests for the presence of a change in the parameters estimates over time. The effects seem to be fairly stable. Finally, Columns (7) and (8) test for non-linearities in the relationship. There is a weak tendency for extreme amounts of precipitation to reduce turnout, but the overall pattern is still close to linearity.

Figure 8: Distribution of t-values controlling for spatio-temporal trends



Notes:The graph shows the distribution of the t-values when regressing municipal turnout on daily precipitation for 600 days before and after election day. The 10 days before and after the actual election day are omitted. Panel (a) shows results from regressing levels on levels. Panel (b) shows the regression of turnout on a dummy for more than 25 mm rain while Panel (c) employs a dummy for any rain. Panel (d) shows results from a regression where the rank of turnout is measured on the rank of rain, i.e. both variables are uniform on the unit interval.

Spatio-temporal trends are controlled for using tensor products of Legendre polynomials with 1×6*,* 3×10*, and* 7×8 *terms. Linear temporal trends are shown in solid lines and quadratic linear trends in dashed lines.*

Table 3: The effect of precipitation on turnout

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Rain (in cm)	0.00299***			0.00244***	0.00283***	0.00284***	0.00387***	0.00943***
	(5.34)			(3.58)	(5.29)	(3.95)	(3.86)	(4.01)
Rain positive		0.00742***						
		(5.31)						
Rain above 2.5 mm			0.00511***	0.00215*				
			(4.94)	(1.69)				
Rain×Year					-0.000356			
					(-1.56)			
Rain×After 1990						0.000416		
						(0.44)		
Rain^2							-0.000170	-0.00429***
							(-0.96)	(-2.85)
Rain^3								0.000843***
								(2.79)
Rain^4								-0.0000497***
								(-2.70)
Mean dep. var	0.681	0.681	0.681	0.681	0.681	0.681	0.681	0.681
Obs	4417	4417	4417	4417	4417	4417	4417	4417
R^2	0.698	0.697	0.697	0.698	0.698	0.698	0.698	0.699

Notes: Outcome variable is municipal electoral turnout. All specifications include municipal and year fixed effects. All specifications include the tensor product of Legendre polynomials with 1×6 *terms to control for spatio-temporal trends. Standard errors are clustered at the municipality level (using the 2010 municipal structure).*
*t-values in parentheses,and *, **, and *** denotes significant at the 10%, 5%, and 1% levels.*

6 Conclusion

In this paper, I have shown that when outcomes of interest are regressed on weather data, there is a danger of spurious correlations. The reason is that spatial patterns in weather conditions are likely to align up with spatial or spatio-temporal patterns in the outcomes of interest. This can be shown theoretically in simple models, and occur in Monte Carlo analyses in a wider range of models. I also illustrate the problem using real data on Norwegian electoral participation, where turnout is correlated with rainfall on irrelevant days in the majority of cases.

To solve the problem, I suggest introducing controls for spatial or spatio-temporal trends in regressions. This is a simple remedy that can easily be combined with other techniques, such as instrumental variables of regression discontinuity designs. In the sample of Norwegian elections, this is shown to substantially improve the behavior of estimators.

The question of more sophisticated approaches to controlling for spatial and spatio-temporal trends, possibly borrowing from the literature on spatial statistics and econometrics is left for future research. There are probably possibilities to do better, but it is unclear that such approaches are sufficiently simple to implement that they actually matter for the applied researcher.

As weather data are typically available for a large number of periods, of which only a few matter, there is ample supply of placebo data. One question is whether these placebos could be used to construct a more correct null distribution of the parameter of interest, somewhat along the lines of bootstrapping techniques. Saunders (1993) implements a version of this estimator, but does not go into its statistical properties and potential advantages compared to ordinary inference.

References

Artés, J. (2014): "The rain in Spain: Turnout and partisan voting in Spanish elections," *European Journal of Political Economy*, 34, 126 – 141.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650.

Brown, P. E., P. J. Diggle, M. E. Lord, and P. C. Young (2001): "Space-time calibration of radar rainfall data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50, 221–241.

Buhmann, M. D. (2003): *Radial basis functions: theory and implementations*, Cambridge University Press.

Busse, M. R., D. G. Pope, J. C. Pope, and J. Silva-Risso (2014): “The Psychological Effect of Weather on Car Purchases,” *The Quarterly Journal of Economics*, Forthcoming.

Carr, J. and J. L. Doleac (2014): “Keep the Kids Inside: Juvenile Curfews, Bad Weather, and Urban Gun Violence,” Batten Working Paper 2014-003.

Cliff, A. D., P. Haggett, J. K. Ord, K. A. Bassett, and R. B. Davies (1975): *Elements of Spatial Structure: A Quantitative Approach*, Cambridge University Press.

Collins, W. J. and R. A. Margo (2007): “The Economic Aftermath of the 1960s Riots in American Cities: Evidence from Property Values,” *The Journal of Economic History*, 67, 849–883.

Connolly, M. (2008): “Here Comes the Rain Again: Weather and the Intertemporal Substitution of Leisure,” *Journal of Labor Economics*, 26, pp. 73–100.

——— (2013): “Some Like It Mild and Not Too Wet: The Influence of Weather on Subjective Well-Being,” *Journal of Happiness Studies*, 14, 457–473.

Cressie, N. A. C. (1993): *Statistics for Spatial Data*, John Wiley & Sons.

Cunningham, M. R. (1979): “Weather, Mood, And Helping-behavior - Quasi Experiments with whe Sunshine Samaritan,” *Journal of Personality and Social Psychology*, 37, 1947–1956.

Dell, M., B. F. Jones, and B. A. Olken (2014): “What Do We Learn from the Weather? The New Climate-Economy Literature,” *Journal of Economic Literature*, Forthcoming.

Denissen, J. J. A., L. Butalid, L. Penke, and M. A. G. van Aken (2008): “The Effects of Weather on Daily Mood: A Multilevel Approach,” *Emotion*, 8, 662–7.

Eisinga, R., M. Grotenhuis, and B. Pelzer (2012a): “Weather conditions and political party vote share in Dutch national parliament elections, 1971-2010,” *International Journal of Biometeorology*, 56, 1161–1165.

——— (2012b): “Weather conditions and voter turnout in Dutch national parliament elections, 1971-2010,” *International Journal of Biometeorology*, 56, 783–786.

Elhorst, J. P. (2001): “Dynamic Models in Space and Time,” *Geographical Analysis*, 33, 119–140.

Fiva, J. H., A. Halse, and G. J. Natvik (2012): “Local Government Dataset,” Dataset available from esop.uio.no.

Frühwirth, M. and L. Sögner (2015): "Weather and SAD related mood effects on the financial market," *The Quarterly Review of Economics and Finance*, Forthcoming.

Gelfand, A. E., H.-J. Kim, C. F. Sirmans, and S. Banerjee (2003): "Spatial Modeling with Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396.

Gomez, B. T., T. G. Hansford, and G. A. Krause (2007): "The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections," *The Journal of Politics*, 69, 649–663.

Graff Zivin, J. S., S. M. Hsiang, and M. J. Neidell (2015): "Temperature and Human Capital in the Short- and Long-Run," NBER Working Paper 21157, National Bureau of Economic Research.

Granger, C. and P. Newbold (1974): "Spurious regressions in econometrics," *Journal of Econometrics*, 2, 111 – 120.

Guven, C. and I. Hoxha (2014): "Rain or shine: Happiness and risk-taking," *Quarterly Review of Economics and Finance*, Forthcoming.

Hansford, T. G. and B. T. Gomez (2010): "Estimating the Electoral Effects of Voter Turnout," *American Political Science Review*, 104, 268–288.

Hastie, T. and R. Tibshirani (1993): "Varying-Coefficient Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 757–796.

Hastie, T., R. Tibshirani, and J. Friedman (2008): *The elements of statistical learning: data mining, inference and prediction*, Springer, 2 ed.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998): "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika*, 85, 809–822.

Horiuchi, Y. and J. Saito (2009): "Rain, Election, and Money: The Impact of Voter Turnout on Distributive Policy Outcomes," Mimeo, Yale.

Huang, J. Z., C. O. Wu, and L. Zhou (2002): "Varying-coefficient models and basis function approximations for the analysis of repeated measurements," *Biometrika*, 89, 111–128.

Judd, K. L. (1998): *Numerical methods in economics*, MIT Press.

Keller, M., B. Fredrickson, O. Ybarra, S. Cote, K. Johnson, J. Mikels, A. Conway, and T. Wager (2005): "A warm heart and a clear head - The contingent effects of weather on mood and cognition," *Psychological Science*, 16, 724–731.

Krumbein, W. C. (1959): "Trend surface analysis of contour-type maps with irregular control-point spacing," *Journal of Geophysical Research*, 64, 823–834.

——— (1963): "Confidence intervals on low-order polynomial trend surfaces," *Journal of Geophysical Research*, 68, 5869–5878.

Kurrild-Klitgaard, P. (2013): "It's the weather, stupid! Individual participation in collective May Day demonstrations," *Public Choice*, 155, 251–271.

Lee, J. J., F. Gino, and B. R. Staats (2014): "Rainmakers: Why bad weather means good productivity," *Journal of Applied Psychology*, 99, 504–513.

LeSage, J. and R. K. Pace (2009): *Introduction to Spatial Econometrics*, CRC Press.

Lind, J. T. (2014): "Rainy Day Politics - An Instrumental Variables Approach to the Effect of Parties on Political Outcomes," CESifo Working Paper No. 4911.

Lo Prete, A. and F. Revelli (2014): "Voter Turnout and City Performance," SIEP Working Paper 10.

Madestam, A., D. Shoag, S. Veuger, and D. Yanagizawa-Drott (2013): "Do Political Protests Matter? Evidence from the Tea Party Movement," *Quarterly Journal of Economics*, 128, 1633–85.

Matsui, H., T. Misumi, and S. Kawano (2011): "Varying-coefficient modeling via regularized basis functions," *ArXiv e-prints.*

——— (2014): "Model selection criteria for the varying-coefficient modelling via regularized basis expansions," *Journal of Statistical Computation and Simulation*, 84, 2156–2165.

Persson, M., A. Sundell, and R. Öhrvall (2014): "Does Election Day weather affect voter turnout? Evidence from Swedish elections," *Electoral Studies*, 33, 335–342.

Pfeifer, P. E. and S. J. Deutsch (1980): "A three-stage iterative procedure for space-time modeling phillip," *Technometrics*, 22, 35–47.

Phillips, P. C. B. (1986): "Understanding spurious regressions in econometrics," *Journal of Econometrics*, 33, 311 – 340.

Ripley, B. D. (2004): *Spatial Statistics*, John Wiley & Sons, 2nd ed.

SAUNDERS, EDWARD M., J. (1993): "Stock Prices and Wall Street Weather," *The American Economic Review*, 83, 1337–1345.

SEN, A. AND P. YILDIRIM (2015): "Clicks and editorial decisions: Does popularity shape coverage?" Mimeo, Toulouse School of Economics.

SFORZA, A. (2013): "The Weather Effect: Estimating the effect of voter turnout on electoral outcomes in Italy," Mimeo, London School of Economics and Political Sciences.

SIMONSOHN, U. (2010): "Weather To Go To College," *The Economic Journal*, 120, 270–280.

SIMS, C. A., J. H. STOCK, AND M. W. WATSON (1990): "Inference in linear time series models with some unit roots," *Econometrica*, 58, 113–144.

STERN, R. AND R. COE (1984): "A model fitting analysis of daily rainfall data," *Journal of the Royal Statistical Society. Series A (General)*, 147, 1–34.

TOBLER, W. R. (1969): "Geographical Filters and their Inverses," *Geographical Analysis*, 1, 234–253.

TOTIK, V. (2005): "Orthogonal Polynomials," *Surveys in Approximation Theory*, 1, 70–125.

TVEITO, O. E. AND E. J. FØRLAND (1999): "Mapping temperatures in Norway applying terrain information, geostatistics and GIS," *Norwegian Journal of Geography*, 53, 202–212.

VELARDE, L. G. C., H. S. MIGON, AND B. D. B. PEREIRA (2004): "Space-time modeling of rainfall data," *Environmetrics*, 15, 561–576.

ZHU, H., J. FAN, AND L. KONG (2014): "Spatially Varying Coefficient Model for Neuroimaging Data With Jump Discontinuities," *Journal of the American Statistical Association*, 109, 1084–1098.

A Proofs

A.1 Proof of divergence of the numerator in (3)

Proof. Let $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the ceil and floor operators. [14] If we have $\lambda = \lceil pn\rceil - pn$ (so $1-\lambda = pn - \lfloor pn\rfloor$) we have

$$\begin{aligned}\sum \frac{1}{|pn-1|+1} &= \frac{1}{\lceil pn\rceil - pn + 1} + \frac{1}{\lceil pn\rceil - pn + 2} + \ldots + \frac{1}{\lceil pn\rceil - pn + (n - \lceil pn\rceil)} \\ &+ \frac{1}{pn - \lfloor pn\rfloor + 1} + \frac{1}{pn - \lfloor pn\rfloor + 2} + \ldots + \frac{1}{pn - \lfloor pn\rfloor + \lfloor pn\rfloor} \\ &= \frac{1}{1+\lambda} + \frac{1}{2+\lambda} + \ldots + \frac{1}{(n-\lceil pn\rceil)+\lambda} \\ &+ \frac{1}{1+(1-\lambda)} + \frac{1}{2+(1-\lambda)} + \ldots + \frac{1}{\lfloor pn\rfloor + (1-\lambda)}\end{aligned}$$

Define

$$S^1_{n-\lceil pn\rceil} = \frac{1}{1+\lambda} + \frac{1}{2+\lambda} + \ldots + \frac{1}{(n-\lceil pn\rceil)+\lambda}$$

and

$$S^2_{\lfloor pn\rfloor} = +\frac{1}{1+(1-\lambda)} + \frac{1}{2+(1-\lambda)} + \ldots + \frac{1}{\lfloor pn\rfloor + (1-\lambda)}$$

The series S^1_n is a generalized harmonic series of length $(n - \lceil pn\rceil)$. If we define $S^0_n = 1 + \frac{1}{2} + \ldots \frac{1}{n}$ as the standard harmonic series of length n, we see that $S^0_{n-\lceil pn\rceil+1} - 1 < S^1_{n-\lceil pn\rceil} < S^0_{n-\lceil pn\rceil}$. For large n we know that $S^0_n \to \gamma + \ln n$ where γ is the Euler–Mascheroni constant ($\gamma \approx .577$). Hence $\gamma + \ln \frac{n-\lceil pn\rceil+1}{e} < S^1_{n-\lceil pn\rceil} < \gamma + \ln(n - \lceil pn\rceil)$. From a similar reasoning, $\gamma + \ln \frac{\lfloor pn\rfloor+1}{e} < S^2_{\lfloor pn\rfloor} < \gamma + \ln(\lfloor pn\rfloor)$. It follows that

$$2\gamma + \ln \frac{n - \lceil pn\rceil + 1}{e} + \ln \frac{\lfloor pn\rfloor + 1}{e} < \sum \frac{1}{|pn-1|+1} < 2\gamma + \ln(n - \lceil pn\rceil) + \ln(\lfloor pn\rfloor)$$

Next, the term

$$\begin{aligned}\sum \frac{i}{|P-i|+1} &= \frac{\lceil pn\rceil}{\lceil pn\rceil - pn + 1} + \frac{\lceil pn\rceil + 1}{\lceil pn\rceil - pn + 2} + \ldots + \frac{n}{\lceil pn\rceil - pn + (n - \lceil pn\rceil)} \\ &+ \frac{\lfloor pn\rfloor}{pn - \lfloor pn\rfloor + 1} + \frac{\lfloor pn\rfloor - 1}{pn - \lfloor pn\rfloor + 2} + \ldots + \frac{1}{pn - \lfloor pn\rfloor + \lfloor pn\rfloor}\end{aligned}$$

[14] That is, for any $x \in \mathbb{R}_+$, $\lceil x\rceil = \min\{y \in \mathbb{N} : x \leq y\}$ and $\lfloor x\rfloor = \max\{y \in \mathbb{N} : x \geq y\}$.

We have

$$\frac{\lceil pn\rceil+1}{\lceil pn\rceil-pn+1}+\frac{\lceil pn\rceil+2}{\lceil pn\rceil-pn+2}+\ldots+\frac{n}{\lceil pn\rceil-pn+(n-\lceil pn\rceil)}$$
$$=(\lceil pn\rceil-1)\left(\frac{1}{\lceil pn\rceil-pn+1}+\frac{1}{\lceil pn\rceil-pn+2}+\ldots+\frac{1}{\lceil pn\rceil-pn+(n-\lceil pn\rceil)}\right)$$
$$+\frac{1}{\lceil pn\rceil-pn+1}+\frac{2}{\lceil pn\rceil-pn+2}+\ldots+\frac{n-\lceil pn\rceil}{\lceil pn\rceil-pn+(n-\lceil pn\rceil)}$$
$$=(\lceil pn\rceil-1)\,S_n^1+T^1_{n-\lceil pn\rceil}$$

where the serie

$$T_n^1=\frac{1}{1+\lambda}+\frac{2}{2+\lambda}+\ldots+\frac{n}{n+\lambda}$$

We know that

$$\frac{n}{1+\lambda}<T_n^1<n$$

Similarly,

$$\frac{\lfloor pn\rfloor}{pn-\lfloor pn\rfloor+1}+\frac{\lfloor pn\rfloor-1}{pn-\lfloor pn\rfloor+2}+\ldots+\frac{1}{pn-\lfloor pn\rfloor+\lfloor pn\rfloor}$$
$$=(\lfloor pn\rfloor+1)\left(\frac{1}{pn-\lfloor pn\rfloor+1}+\frac{1}{pn-\lfloor pn\rfloor+2}+\ldots+\frac{1}{pn-\lfloor pn\rfloor+\lfloor pn\rfloor}\right)$$
$$-\left(\frac{1}{pn-\lfloor pn\rfloor+1}+\frac{2}{pn-\lfloor pn\rfloor+2}+\ldots+\frac{\lfloor pn\rfloor}{pn-\lfloor pn\rfloor+\lfloor pn\rfloor}\right)$$
$$=(\lfloor pn\rfloor+1)\,S_n^2-T^2_{\lfloor pn\rfloor}$$

where the serie

$$T_n^2=\frac{1}{1+(1-\lambda)}+\frac{2}{2+(1-\lambda)}+\ldots+\frac{n}{n+(1-\lambda)}$$

We know that

$$\frac{n}{2-\lambda}<T_n^1<n$$

It follows that

$$\frac{1}{n}\sum\frac{i}{|P-i|+1}=\frac{(\lceil pn\rceil-1)\,S^1_{n-\lceil pn\rceil}+(\lfloor pn\rfloor+1)\,S^2_{\lfloor pn\rfloor}+T^1_{n-\lceil pn\rceil}-T^2_{\lfloor pn\rfloor}}{n}$$

so the full numerator becomes

$$N=\frac{1}{n}\sum\left(\frac{1}{|P-i|+1}-\bar{w}\right)i=\left(\frac{\lceil pn\rceil-1}{n}-\frac{n+1}{2n}\right)S^1_{n-\lceil pn\rceil}$$
$$+\left(\frac{(\lfloor pn\rfloor+1)}{n}-\frac{n+1}{2n}\right)S^2_{\lfloor pn\rfloor}+\frac{T^1_{n-\lceil pn\rceil}-T^2_{\lfloor pn\rfloor}}{n}$$

When $n \to +\infty$, we see that the two first parentheses converge to $p - \frac{1}{2}$ and the last fraction to ... Hence N converges to a log function, and hence diverges. □

A.2 Proof of convergence of the denominator in (3)

Proof. We want to study the behavior of $\sqrt{\frac{1}{N^2}\sum\left(\frac{1}{|P-i|+1} - \bar{w}\right)^2}$. We have $\frac{1}{N^2}\sum\left(\frac{1}{|P-i|+1} - \bar{w}\right)^2 = \frac{1}{N^2}\sum\left(\frac{1}{|P-i|+1}\right)^2 - \frac{1}{N^2}N\bar{w}^2$. We know from the proof in Appendix A.1 that $\bar{w}$ converges to a log function so $\frac{\bar{w}^2}{N} \to 0$ as $N \to \infty$. As for the proof in A.1, define $\lambda = \lceil pn \rceil - pn$ (so $1 - \lambda = pn - \lfloor pn \rfloor$). Then we have

$$\begin{aligned}\sum\left(\frac{1}{|pn-1|+1}\right)^2 &= \left(\frac{1}{\lceil pn \rceil - pn + 1}\right)^2 + \left(\frac{1}{\lceil pn \rceil - pn + 2}\right)^2 + \ldots + \left(\frac{1}{\lceil pn \rceil - pn + (n - \lceil pn \rceil)}\right)^2 \\ &+ \left(\frac{1}{pn - \lfloor pn \rfloor + 1}\right)^2 + \left(\frac{1}{pn - \lfloor pn \rfloor + 2}\right)^2 + \ldots + \left(\frac{1}{pn - \lfloor pn \rfloor + \lfloor pn \rfloor}\right)^2 \\ &= \left(\frac{1}{1+\lambda}\right)^2 + \left(\frac{1}{2+\lambda}\right)^2 + \ldots + \left(\frac{1}{(n - \lceil pn \rceil) + \lambda}\right)^2 \\ &+ \left(\frac{1}{1+(1-\lambda)}\right)^2 + \left(\frac{1}{2+(1-\lambda)}\right)^2 + \ldots + \left(\frac{1}{\lfloor pn \rfloor + (1-\lambda)}\right)^2\end{aligned}$$

Define the series

$$Q^1_{n-\lceil pn \rceil} = \left(\frac{1}{1+\lambda}\right)^2 + \left(\frac{1}{2+\lambda}\right)^2 + \ldots + \left(\frac{1}{(n - \lceil pn \rceil) + \lambda}\right)^2$$

and

$$Q^2_{\lfloor pn \rfloor} = \left(\frac{1}{1+(1-\lambda)}\right)^2 + \left(\frac{1}{2+(1-\lambda)}\right)^2 + \ldots + \left(\frac{1}{\lfloor pn \rfloor + (1-\lambda)}\right)^2$$

and define the sum of the the reciprocals of the squares of natural numbers $Q^0_n = \sum_{i=1}^n \left(\frac{1}{i}\right)^2$. Then we see that $Q^0_{1+n-\lceil pn \rceil} - 1 \leq Q^1_{n-\lceil pn \rceil} \leq Q^0_{n-\lceil pn \rceil}$ and $Q^0_{1+\lfloor pn \rfloor} - 1 \leq Q^2_{\lfloor pn \rfloor} \leq Q^0_{\lfloor pn \rfloor}$. Hence for given p then $\lim_{N\to\infty}\sum\left(\frac{1}{|pn-1|+1}\right)^2 = Q_\infty$. And as $\lim_{N\to\infty} Q^0_n = \frac{\pi^2}{6}$, we have $\frac{\pi^2}{3} - 2 \leq Q_\infty \leq \frac{\pi^2}{3}$. As $\lim_{N\to\infty}\sqrt{\frac{1}{N^2}\sum\left(\frac{1}{|P-i|+1} - \bar{w}\right)^2} \to \sqrt{\frac{Q_\infty}{N^2}}$, it is clear that $\lim_{N\to\infty}\sqrt{\frac{1}{N^2}\sum\left(\frac{1}{|P-i|+1} - \bar{w}\right)^2} = 0$, and this happens at rate $O\left(\frac{1}{N}\right)$. □

B Additional Monte Carlo results

Table A-1: A Monte Carlo analysis of the spatial model

	K						
N	1	2	5	10	100	N	2N
9	0.37	0.36	0.31	0.32	0.30	0.29	0.25
	1.76	1.71	1.59	1.59	1.56	1.53	1.42
25	0.57	0.58	0.55	0.54	0.54	0.54	0.51
	2.84	2.78	2.57	2.52	2.44	2.49	2.43
49	0.65	0.66	0.66	0.65	0.63	0.66	0.64
	3.8	3.84	3.59	3.47	3.28	3.41	3.35
100	0.73	0.76	0.73	0.74	0.72	0.71	0.75
	5.21	5.12	4.79	4.87	4.61	4.6	4.64
400	0.85	0.85	0.84	0.85	0.85	0.85	0.83
	9.15	8.98	8.5	8.6	8.12	8.56	7.79
1024	0.88	0.90	0.90	0.91	0.89	0.89	0.90
	13.1	12.9	12.7	12.7	12.3	11.7	12.2
10000	0.95	0.96	0.96	0.94	0.96	0.95	0.96
	31.5	31.6	31.5	30.3	29.8	31	31.4

Notes:The table shows the fraction of cases where a t-test of $\beta = 0$ is rejected at the 5 % level (first line) and the average of the absolute value of the associated t-value (second line). The true model is $z_i = \alpha + \beta \sum r_i + \tau\,(x_i + y_i) + \epsilon_i$ with$\beta = 0$, $\tau = 1$, $\epsilon_i \sim N(0,1)$, and for each k, $\nu_k \sim N(0,1)$ and the position $p_k \sim U([0,N] \times [0,N])$. Each model is replicated 1000 times.