Spatial Analysis of Bureaucrats' Attempts to Resist Political Capture in a Developing Democracy: The Distribution of Solar Panels in Ghana

Appendix (Online): Technical Discussion of Data and Methods

Part A: Dataset Creation

District Shapefiles

For the analysis in "Spatial Analysis of Bureaucrats' Attempts to Resist Political Capture in a Developing Democracy: The Distribution of Solar Panels in Ghana," we created a new spatial dataset by merging several types of data. We use a shapefile for Ghana's districts that was created by the RS/GIS Lab at the University of Ghana-Legon. Existing shapefiles were of 2012 jurisdictions, requiring us to create new shapefiles that represent district boundaries in 2008, before the solar panel commenced. This was necessary because some districts were created between 2008 and 2012 by splitting jurisdictions. After identifying the new districts created in 2012, we were able to merge them back together to form their 2008 counterparts.

Fitting Variables to Shapefiles

Most variables in the analysis required some form of transformation in order to fit into these shapefiles. We took three approaches to variable transformation, depending on data type. First, for variables that originated at the constituency level (NDC vote share 2008; voter turnout 2008; and ethnic fractionalization), we created district-level data by computing average values among the constituencies in the district. These averages are weighted by population (although population-weighted averages and averages that are not weighted by population are almost identical). Out of 230 total constituencies, we implement this transformation for 97 constituencies that fall within 37 districts. Second, for geo-located data (solar panels and World Bank projects) we spatially joined the existing data to the district shapefile. Third, for line shapefiles (electric grid and roads), we used the Intersect feature in ArcMap. We then took the resultant file and used the Dissolve feature in ArcMap to reduce the road and electric grid files down to one unit per district. The Calculate Geometry tool then allowed us to calculate line lengths. For electric grid per capita, we divided the length of grid by population. For road density, we divided the length of roads by area of the district. We display descriptive statistics in Table A1 below.

Table A1: Descriptive Statistics

	Mean	Min	Max	SD	N
Dependent Variable					
Solar projects	7.28	0	202	27.38	170
Political Variables					
Reward NDC loyalists (2008 NDC vote share)	0.49	0.11	0.96	0.18	170
Induce swing voters (NDC vote share volatility)	0.05	0	0.21	0.04	170
Mobilize voters where turnout level varies (turnout volatility)	0.15	0	0.27	0.04	170
Need Variables					
Electric grid per capita (*1000)	0.28	0	4.35	0.44	170
Road network density	0.12	0.03	0.42	0.05	170
Control Variables					
Ethnic fractionalization	0.41	0.06	0.82	0.2	170
Population density	391.59	7.97	10124.2	1311.76	170
World Bank projects	5.36	0	52	5.14	170
Percent nonvoters 2008	0.29	0.16	0.46	0.05	170
Poverty rate	30.52	1.95	92.4	19.96	170
Inequality level (Gini)	38.69	27.2	64	6.02	170
Number of health facilities	21.71	3	276	26.32	170
Literacy	0.68	0.2	0.93	0.18	170
Female ratio	0.51	0.47	0.55	0.02	170

Part B: Additional Methodological Discussion and Analysis

Accounting for Spatial Autocorrelation: Selecting Spatially Lagged Dependent Variable (SLDV) Models for Use in the Analysis

In the body of our manuscript, we present SLDV models that account for spatial clustering. It has been suggested that we might instead aggregate our district data into higher-level units (regions) in order to employ multi-level models, because fixed effects for given spatial units offer leverage to examine within-unit effects. Such an approach is not appropriate for several reasons. First, the approach does not account for clustering. Second, region fixed effects are not theoretically relevant for the questions asked. Third, given the way that bureaucrats selected locations, a "nested" approach is not as well-matched, theoretically, as the SLDV approach. We do present OLS models in the paper and additional models in this appendix, and those models cluster standard errors by region.

We thus highlight modelling strategies that account for spatial clustering. To confirm that this was appropriate, first, we use an LM test to examine whether an OLS model sufficiently fits our data, or if the model needs a spatially lagged variable or spatial error term. Spatially lagged dependent variable models account for the possibility that the values of the dependent variable in one location are influenced by the values in nearby locations. Spatial error models, on the other hand, are intended to account for omitted variables that are spatially autocorrelated. Such

variables are often omitted because they are either very difficult or impossible to effectively measure.

The equation for spatial error models is as follows:

$$y=X(\beta) + \varepsilon, \varepsilon =\lambda(W) \varepsilon + u$$

In this equation, y is the dependent variable, X is a matrix of observations on the explanatory variables, λ and β are parameters, ε is a vector of spatially autocorrelated error terms, W is the spatial weights matrix, and u is a vector of independent identically distributed (i.i.d.) errors.

To estimate spatially lagged dependent variable models, the following equation is used:

$$y=(\rho)Wy + X(\beta) + \varepsilon$$

In this equation, Wy is the spatially lagged dependent variable for weights matrix W, X is a matrix of observations on the explanatory variables, ε is a vector of error terms, and ρ and β are parameters.

While the technical components of SLDV and spatial error models are distinct, they address closely related issues. Many models may exhibit spatially clustered error terms and spatial autocorrelation in the dependent variable. This produces a need to statistically diagnose which type of model is more appropriate. Since our data include both violations of the assumptions of OLS, we employ another LM diagnostic test, presented in Table A2.

Use of the LM diagnostic was originally suggested by Anselin and Rey.[1] This was refined by Anselin et al. to incorporate the robust forms of the statistics.[2] In all of the simple LM tests for error dependence and for a missing spatially lagged dependent variable, our test statistics were statistically significant. This led us to look to robust versions of these tests. The LM test for error dependence in the possible presence of a missing spatially lagged dependent variable is not significant at any of the levels of analysis. Meanwhile, the LM test for a missing spatially lagged dependent variable in the possible presence of error dependence is significant at the 0.1 level or lower throughout all models. This indicates that the spatially lagged dependent variable model is more appropriate than the spatial error model. The variables used in these analyses are not highly collinear (defined as Pearson's $r > 0.7$), and the variance inflation factor (VIF) clearly indicates that global multicollinearity is not present. We estimate the spatial models using the "spdep" package in R.[3]

Researchers often interpret the coefficients immediately estimated by SLDV models (Whitten, Guy D., Laron K. Williams & Cameron Wimpy 2019). Yet this approach conflates spatial dependencies and the relationship between the independent and dependent variable. Instead, we use the impacts function as a post-estimation step to isolate the direct effects of our independent variables. Direct effects are the effects of independent variables at a location i on the dependent variable at location i. Indirect effects are the effects of independent variables at a location i on the dependent variable on i's neighboring locations j. Direct effect coefficients can be interpreted like OLS coefficients. We obtain distributions of our direct effects by using a Markov Chain Monte Carlo simulation process. We ran 1000 simulations and calculated z-statistics from these simulations.

[1] Anselin and Rey 1991.
[2] Anselin et al. 1996.
[3] See Bivand et al. 2005. Since the R spatial analysis community is transitioning from spdep to spatialreg, future researchers may need to update their R script accordingly.

TABLE A2. Diagnostics for Spatially Lagged versus Spatial Error Model Fit

District Data		
	Statistic	p-value
Simple LM error test	5.4313	0.019779*
Simple LM spatial lag test	8.7718	0.003059*
Robust LM error test	1.0663	0.301779
Robust LM spatial lag test	4.4068	0.035795*
Portmanteau test: LMerr+RLMlag	9.8381	0.007306*
* p<0.05		

Second, we use an additional LM test to examine whether a model with only a spatially lagged dependent variable (SLDV) can effectively fit our data. If this LM test calculates a statistically significant parameter, then there is justification to use a spatial Durbin model (Elhorst 2010). In our case, this test fails to support the need for inclusion of spatially lagged independent variables, so we have confidence that either an SLDV or spatial error model is most appropriate.[4]

Robustness and Sensitivity Checks on Spatially Lagged Dependent Variable Models

Presentation of Additional SLDV Models

To confirm the validity of our results, we report in Table A3 a variety of model specifications. Models 2 and 3 include a variable for whether a jurisdiction borders Lake Volta, which is a dichotomous variable valued at 1 if the district touches the lake. (This variable is not a variable for the district being part of Volta Region, which is a formal administrative unit; see Figure 3 in the main text.) We include the variable for bordering the lake both because we were told that the program initially targeting island communities, and because there is clearly clustering in that area. Additionally, Model 4 includes a dummy variable for whether turnout in a district increased between 2004 and 2008. We include this variable to test whether the direction of turnout matters – whether more or fewer people voted in 2008 than in 2004 in a jurisdiction. In some models, we dropped population density, since it is clear from the maps that solar PV panels were not being targeted to major urban areas. In Model 5, we estimated an SLDV model with population instead of population density. Due to relatively high collinearity between population and road density (0.65), we had to manually increase the tolerance for this model to converge.

[4] Another option has been suggested, but it is not currently possible to do. Specifically, spatial econometrics has long been interested in the possibility of including both a spatially lagged dependent variable and a spatial error term (Fischer and Getis 2009, p. 381-382). However, this is not possible if the spatial weights matrix is the same for the spatially lagged dependent variable and the spatial error term (Anselin and Bera 1998).

Table A3: Direct Effects of Spatially Lagged Total Solar Projects

	Model 1	Model 2	Model 3	Model 4	Model 5
Political Variables					
H1A: Reward NDC loyalists					
(2008 NDC vote share)	13.00	6.58	7.52	12.11	15.11
	(1.04)	(0.55)	(0.66)	(0.93)	(1.30)
H1B: Induce swing voters					
(NDC vote share volatility)	47.35	48.59	51.89	40.68	45.38
	(0.98)	(1.05)	(1.18)	(0.85)	(1.05)
H1C: Mobilize nonvoting incumbent supporters					
(nonvoters * NDC vote share)	123.30*	109.20*	102.80*	103.90	109.96*
	(2.12)	(2.06)	(1.95)	(1.80)	(1.92)
H1C: Mobilize voters where turnout level varies					
(turnout volatility)				-18.23	
				(-1.32)	
Need Variables					
H2A: Electric grid per capita	-4952	-4315	-3926	-4906	-4224
	(-1.07)	(-1.03)	(-0.95)	(-1.17)	(-1.02)
H2B: Road network density	-100.20*	-74.32	-111.20*	-153.10*	-114.09*
	(-2.14)	(-1.69)	(-2.23)	(-2.83)	(-2.58)
Control Variables					
Ethnic fractionalization	5.97	-3.11	-5.82	2.36	4.69
	(0.56)	(-0.31)	(-0.54)	(0.20)	(0.42)
Population density			0.003	0.004	
			(1.30)	(1.50)	
Population size					0.00004
					(1.78)
World Bank projects	-0.27	-0.54	-0.54	-0.14	-0.45
	(-0.63)	(-1.37)	(-1.39)	(-0.32)	(-1.09)
Percent voters 2008	-41.08	-33.05	-35.19	-63.23	-45.28
	(-0.81)	(-0.72)	(-0.75)	(-1.26)	(-0.90)
Poverty rate	-0.1	-0.1	-0.09	-0.08	-0.09
	(-0.71)	(-0.75)	(-0.69)	(-0.56)	(-0.61)
Inequality level (Gini)	0.04	0.08	0.13	0.03	0.11
	(0.14)	(0.27)	(0.41)	(0.07)	(0.33)
Number of health facilities	0.05	0.06	-0.01	-0.05	-0.17
	(0.59)	(0.73)	(-0.10)	(-0.50)	(-1.14)
Literacy	-16.93	-10.66	-6.99	-13.14	-13.20
	(-1.05)	(-0.69)	(-0.45)	(-0.79)	(-0.73)
Female ratio	-347.90*	-303.00*	-316.70*	-356.40*	-359.24*
	(-2.38)	(-2.26)	(-2.36)	(-2.48)	(-2.60)
Borders Lake Volta		33.29*	32.91*		
		(4.65)	(4.35)		
N	170	170	170	170	170

* p<0.05

Estimates are mean direct effects; simulated z-statistics are in parentheses

In all models, turnout volatility remains statistically significant and positive, and road density is statistically significant and negative, even when the Borders Lake Volta dummy is included. This Lake Volta variable is significant when included in models, but the dummy variable marking when voter turnout increased in a district is not.

OLS Models

Table A4 presents OLS results. Models 3 and 4 are the same as Models 1 and 2 in Table 2 of the main text. Models 1 and 2 in Table A4 exclude control variables. We cluster standard errors by region. Our OLS results are consistent with our SLDV results, except more variables are statistically significant. This is to be expected when spatial clustering is not fully taken into account.

Table A4: OLS Analysis of Spatially Lagged Total Solar Projects

	Model 1	Model 2	Model 3	Model 4
Political Variables				
H1A: Reward NDC loyalists				
(2008 NDC vote share)	19.55*	-51.67	19.69	-56.75
	(9.16)	(75.86)	(12.47)	(70.11)
H1B: Induce swing voters				
(NDC vote share volatility)	54.30	67.01	48.68	60.39
	(65.14)	(68.84)	(73.04)	(76.22)
H1C: Mobilize nonvoting incumbent supporters				
(nonvoters * NDC vote share)		247.28		262.97
		(270.65)		(253.41)
H1C: Mobilize voters where turnout level varies				
(turnout volatility)	133.47*	115.90*	145.22*	128.93*
	(57.87)	(52.26)	(59.38)	(55.95)
Need Variables				
H2A: Electric grid per capita	-5351.00*	-5297.54*	-5059.93+	-4939.84+
	(2703.17)	(2671.00)	(2669.37)	(2613.32)
H2B: Road network density	-198.85**	-193.22**	-171.33*	-162.41*
	(72.80)	(69.69)	(72.61)	(69.07)
Control Variables				
Ethnic fractionalization	21.92+	23.93+	7.00	9.27
	(13.20)	(14.20)	(12.24)	(12.25)
Population density	0.003+	0.003+	0.004+	0.004+
	(0.001)	(0.001)	(0.002)	(0.002)
World Bank projects	-0.32	-0.37	-0.22	-0.28
	(0.23)	(0.25)	(0.25)	(0.27)
Percent nonvoters 2008	29.04	-88.91	36.97	-88.89
	(43.02)	(114.51)	(50.49)	(106.52)
Poverty rate			-0.06	-0.07
			(0.12)	(0.12)
Inequality level (Gini)			0.14	0.17
			(0.30)	(0.29)
Number of health facilities			-0.04	-0.04
			(0.05)	(0.05)
Literacy			-8.70	-10.49
			(18.24)	(18.02)
Female ratio			-377.37	-372.23
			(237.45)	(231.38)
N	170	170	170	170

+ p<0.10 ; * p<0.05 ; ** p<0.01

Estimates are mean effects; standard errors clustered by region are in parentheses

Spatial Analysis Using Count Models

Ideal analyses for the data would use models that reflect the fact that our data more closely resembles a Poisson distribution than a normal distribution; a count model would be most appropriate. Given limitations of existing statistical packages, however, it is not possible to combine global analyses that account for spatial autocorrelation with local analysis using geographically weighted regression, if count models are used. Specifically, the only available statistical packages for count models with spatially autocorrelated data employ Bayesian analysis techniques, whereas GWR models employ frequentist maximum likelihood approaches. Because these rely on fundamentally different procedures for inference, we do not believe they should be used together. Moreover, there does not yet exist a reliable statistical package for GWR with count models (discussed more below).

As a result, the body of our paper presents models that assume normally distributed data in Table 2, but we present spatial count models here to show that results are not highly sensitive to model specification. To do so, we employ the relatively recent *CARBayes* package in R.[5] These count models address spatial autocorrelation through the incorporation of vectors of random effects. This approach is different than that of SLDV models, in which the dependent variable is lagged to incorporate values of neighboring units on the right-hand side of the equation. These count model approaches tend to rely upon simulation-based techniques for inference, as opposed to maximum likelihood.

Table A5 shows the results of estimating Poisson models that include a vector of random effects. Specifically, we fit a multivariate spatial generalized linear mixed model, where the random effects are modelled by a multivariate conditional autoregressive model.[6] The conditional autoregressive prior was proposed by Leroux et al., so we refer to these as Leroux models.[7] Between variable correlation is captured by a between variable covariance matrix with no fixed structure. Further details are provided in the package vignette.[8] For our models, we ran 150,000 Markov Chain Monte Carlo (MCMC) simulations with 50,000 burn-in. The high number of simulations was necessary for us to obtain a reasonable acceptance rate for the Markov chain, although it may also make these models more likely to detect statistical significance.

[5] Lee 2013.
[6] Within *CARBayes*, this is done by the MVS.CARleroux command.
[7] Leroux et al. 2000.
[8] Lee 2017.

TABLE A5. Leroux Model Estimated Effects on Total Solar Projects

	Districts
Political Variables	
H1A: Reward NDC loyalists (2008 NDC vote share)	0.5112
	(-3.6493 – 8.8148)
H1B: Induce swing voters (NDC vote share volatility)	8.3577
	(-11.2945 – 32.0298)
H1C: Mobilize voters where turnout level varies (turnout volatility)	24.2897
	(-16.7662 – 46.6628)
Need Variables	
H2A: Electric grid per capita	1013.2836
	(-1493.8302 – 2523.9297)
H2B: Road network density	-143.8257*
	(-162.7773 – -0.8783)
Control Variables	
Ethnic fractionalization	-9.6067
	(-14.6354 – 5.7254)
Population density	-0.0036
	(-0.0073 – 0.0018)
World Bank projects	-1.2172*
	(-1.4687 – -0.0487)
Percent voters 2008	-35.2699
	(-52.0530 – 8.3183)
Poverty rate	0.2187*
	(0.0599 – 0.2692)
Inequality level (Gini)	-0.4224*
	(-0.6377 – -0.0543)
Number of health facilities	0.2068
	(-0.0374 – 0.3287)
Literacy	12.3708
	(-2.3480 – 20.8424)
Female Ratio	86.3596
	(-26.4562 – 142.3604)
N	170

*p<0.05

Range between 2.5th percentile and 97.5th percentile in parentheses.

These results are largely consistent with those in Table A3 in showing the interaction between when need-based and political factors influence the distribution of solar projects. Namely, need-based factors consistently are correlated with solar panel distribution. Meanwhile, these models suggest different stories about political factors. In the SLDV models (Table A3), our measures for need and politics are significant. In the Leroux model (Table A5), however, need variables are significant, but political ones are not. Before discussing these results in greater detail, we emphasize that the Leroux and SLDV models employ different procedures for inference (Bayesian MCMC simulations vs. Frequentist maximum likelihood). The substantial

differences between these models make it extremely difficult to obtain perfectly consistent results.

The Leroux model presented in Table A5 finds statistical significance for need variables. This result corresponds with the stated goals of the program implementers, and provides suggestive evidence, unlike the SLDV models, that they were able to avoid political capture.

Additional Information on and Results from Geographically Weighted Regression (GWR)

While spatially lagged dependent variable models offer many valuable insights for our analysis, they average effects across all units into a single coefficient estimate for each variable. To test whether differences exist in variable relationships over space, we employ geographically weighted regression (GWR) as a sensitivity analysis. GWR has been used for a wide variety of topics to explore the spatial heterogeneity in relationships, such as voter behavior during the US Democratic realignment of the 1920s and 1930s,[9] campaign contributing and volunteerism,[10] regional effects of terrorism on economic growth,[11] and even afforestation in Vietnam.[12] Here, we provide additional explanation of GWR analyses, as well as model results.

GWR estimates distinct coefficients for each unit in an analysis, allowing the researcher to examine the logic of distribution employed in individual jurisdictions. In our case, it allows us to check whether there appears to be cross-district variation in the logic of public goods distribution. GWR is critical for our analysis because distributed goods are visibly clustered in a particular area – the islands and remote villages in and around Lake Volta. GWR allows us to determine whether the logics at play in this region are consistent in the rest of the country. In doing so, we also illustrate the utility of using advanced spatial methods to explore and compare average national phenomena with local-level variation.

Using GWR, locations closer to a given location *i* receive greater weight than locations that are further away from location *i*. Such weight is assigned through a spatial weights matrix. The resulting equation, and its spatial weights matrix, can be displayed in two equations as follows:

$$\text{Equation 1: } y_i = \beta_0(u_i, v_i) + \Sigma_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

$$\text{Equation 2: } w_{ij} = \exp[-1/2(d_{ij}/b)^2]$$

where "u_i, v_i" denotes the coordinates of the ith point in space and $\beta_k(u_i, v_i)$ is a realization of the continuous function $\beta_k(u, v)$ at point i.[13] In the spatial weights matrix, d_{ij} is the distance between locations i and j and b is the bandwidth. Bandwidth reflects the distance-decay of the weighting function and affects the spatial smoothing of the estimates. For our analysis, we use the corrected AIC statistic to calculate the optimal distance for our spatial weights.

Estimates from GWR could reveal unit-level estimates that vary substantially. This could suggest that *different logics* exist across space. Alternatively, GWR results could reveal that the *same logic* holds across space, but that the magnitude of the relationship varies across units. Because GWR results do not report significance, interpreting them requires knowledge and analysis of local context.[14]

---

[9] Darmofal 2008.
[10] Cho and Gimpel 2010.
[11] Ocal & Yildirim 2010.
[12] Clement et al. 2009.
[13] Fotheringham, Brunsdon, and Charlton 2002.
[14] Clement et al. 2009.

To interpret GWR results, it is common practice to compare GWR results for each unit of analysis (districts in our case) to average marginal effects reported in OLS regressions of the same variables.[15] One can then evaluate the range of variation in a coefficient by simply comparing the minimum, maximum and various percentile coefficient values for each variable of interest, which are estimated in GWR. GWR allows us to inspect whether the average marginal effect from the OLS regression falls close to the marginal effect of the median data point in GWR results. OLS average marginal effect values that are far above or below the median effect indicate the possibility of a strong spatial component to the variation in average marginal effect. In simple terms, if the average marginal effect from OLS regression falls below the 25th percentile or above the 75th percentile calculated in GWR estimation, some areas of the country are likely driving the results, and we can examine which areas those are. Table A6 thus shows OLS results for models with the variables that are in the GWR analysis, alongside minimum, maximum and several percentile GWR results.

We start with the need-based variable, road density, which was significant in our global models. The sign of the coefficients is consistent across the OLS result *and* the interquartile range, suggesting that the basic logic of distributing projects where existing infrastructure is limited is consistent countrywide. At the same time, however, the OLS result for road density falls just below the interquartile range, suggesting that the statistical significance in the global models may be driven by particularly strong relationships in a subset of the units. We explore this subnational variation qualitatively in the body of the text.

TABLE A6. GWR Results: Estimated Effects on Number of Solar Panels

	District Data					
	OLS	**Min**	**25%**	**Median**	**75%**	**Max**
Electric grid per capita	-6477.44	-22042.67	-10673.30	-5111.39	-1129.57	921.36
Road density	-162.81	-782.99	-153.12	-74.22	-23.14	62.78
NDC vote share 2008	19.55	-7.72	13.33	26.55	29.98	52.27
NDC vote share volatility	36.85	-158.27	-36.40	-14.42	44.45	641.92
Turnout volatility	177.21	-7.34	38.98	77.65	150.45	463.97

Results are largely the same for the measure of political influence that is statistically significant in the SLDV results, turnout volatility. The interquartile range results in Table A6 show that turnout volatility is positively associated with panel distribution in most of Ghana. However, the mean coefficient on turnout volatility is larger than the coefficients in the interquartile range. This means that the coefficient estimates for more than 75 percent of districts falls below the average marginal effect. As with the effect of road density, we can infer that while there is generally a positive relationship between turnout volatility and panel distribution, particularly strong relationships in a subset of jurisdictions may be driving the statistical significance in the global models.

[15] Although it would be ideal to compare GWR results with the results from our global SLDV model, they are not strictly comparable, because SLDV models account for spatial clustering while GWR models do not. As a result, it is standard to compare GWR and OLS results. While OLS model results may produce biased coefficients due to violations of the independent observation assumption, here our OLS results are similar in coefficient sign and magnitude when compared to our SLDV results.

Maps of GWR results—which are presented in Figure 4 in the main text—help us to further interpret these results.[16] Specifically, we see that across nearly the entire country, the relationship between solar panel distribution and road density is negative, and the relationship between solar panel distribution and turnout volatility is positive. But we also see that there is considerable variation in the *magnitude* of these coefficients across the country for both variables.

Here, we also recognize the limitations of GWR. Local collinearities, the lack of a base model to explain coefficient variation, repeated use of data to estimate model parameters, and sensitivity to kernel bandwidth can lead to unreliable estimates of coefficient signs, coefficient magnitudes, and standard error sizes.[17] While these concerns make GWR unreliable for explanatory assessments of statistical significance, GWR is suitable as it is used in our paper, where the goal is exploratory analysis of variation in coefficient signs and magnitudes.

A Note on GWR Analyses Using Count Models

As explained earlier, local analysis with geographically weighted regression is most frequently performed using OLS as the base model. Shifting to Poisson models to reflect the non-normal distribution of our dependent variable, GWR becomes more complex. The *spgwr* package in R is capable of estimating GWR Poisson models, but even the authors of the package are uncertain about whether scholars should use it. They state, "The use of GWR on GLM is only at the initial proof of concept stage, nothing should be treated as an accepted method at this stage".[18] Given this uncertainty, we chose not to use GWR Poisson models.

Hot Spot Analysis

In the main text of the paper, we overlay solar power projects, represented as dots, on maps of the electric grid and NDC vote share in Figure 1. We also overlay the dots on choropleth maps of road density and voter turnout volatility in Figure 5. These maps yield valuable opportunities for descriptive analysis. However, misperceptions from visually examining these variables are possible. Because of this, we leverage spatial statistics as a robustness check, which can be done with hot spot analysis.

Whereas GWR explores spatial variation in the relationship between variables at a specific unit of analysis, hot spot analysis is a statistical technique that assesses where a variable of interest is particularly clustered or dispersed. This method goes beyond a simple "eyeball test" of where individual solar projects are located because it is not susceptible to visual misperceptions. For our discussion, hot spot analysis also cuts through the complexity of searching for patterns in the spatial distribution of 1242 solar power projects.

Hot spot analysis has been used to help explain a variety of political phenomena, such as terrorism.[19] A "hot spot" is a geographic area in which the variable of interest – in our case, solar panel distribution – is clustered. A "cold spot" is therefore where there is a high degree of dispersion at the local level. We use the Getis-Ord Gi* statistic for our hot spot analysis, as

[16] For our maps, we display five gradients. The intervals for these gradients are calculated by ArcGIS using the Natural Breaks (Jenks) method. As Cho and Gimpel (2010) state, the natural breaks method maximizes within-class homogeneity and between-class heterogeneity.
[17] Wheeler and Tiefelsdorf 2005, Wheeler 2014.
[18] Bivand et al. 2017.
[19] Nemeth, Mauslein, and Stapley 2014.

calculated by ArcGIS.[20] This statistic is calculated using the equations in the figure below.[21] Since the Gi* statistic is a z-score, we can intuitively interpret the positive and significant values as hot spots and the negative and significant values as cold spots.

Figure A1: Getis-Ord Statistic Equations

The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (1)$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (2)$$

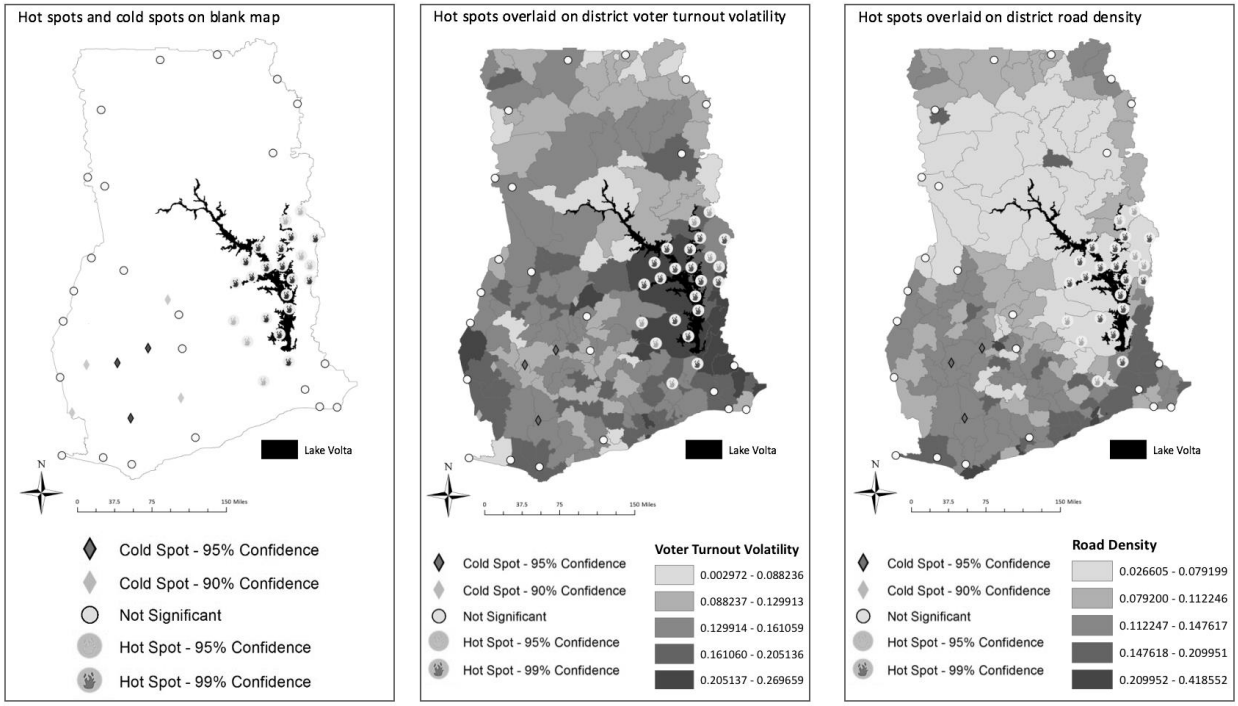
$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \quad (3)$$

The G_i^* statistic is a z-score so no further calculations are required.

We create the "weighted" points that are necessary for hot spot analysis with the Integrate command in ArcGIS. With this command, we grouped points that are within 10 kilometers of each other. In our analysis, we display hot spots (local clustering, shown as flames) and cold spots (local dispersion, shown as diamonds) at the 90%, 95%, and 99% confidence levels (Figure A2). In Figure A2, the highest levels of clustering are marked with dark flames, and the highest levels of dispersion are marked with dark diamonds. We can obtain additional insights by overlaying our hot spot analysis on maps of road density and voter turnout volatility.[22] These maps reveal that the main cluster around Lake Volta is the only area in Ghana with both low road density and high voter turnout volatility. This yields a combined motivation to place solar power projects in this area: high need and high voter mobilization potential.

[20] Nemeth, Mauslein, and Stapley 2014.
[21] The figure comes from ArcGIS documentation here: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm
[22] It is worth noting that hotspot analysis does not recognize sub-national boundaries; districts are shown in Figure A2 in order to display descriptive statistics of road density or voter turnout volatility only.

FIGURE A2. Hotspot Analyses



REFERENCES

Anselin, Luc, and Anil K Bera. 1998. "Spatial dependence in linear regression models with an introduction to spatial econometrics." *Statistics Textbooks and Monographs* 155:237-290.

Anselin, Luc, Anil K Bera, Raymond Florax, and Mann J Yoon. 1996. "Simple diagnostic tests for spatial dependence." *Regional science and urban economics* 26 (1):77-104.

Anselin, Luc, and Serge Rey. 1991. "Properties of tests for spatial dependence in linear regression models." *Geographical analysis* 23 (2):112-131.

Ayee, Joseph RA. 2012. "The political economy of the creation of districts in Ghana." *Journal of Asian and African Studies*:0021909612464334.

Bivand, Roger, Andrew Bernat, Marilia Carvalho, Yongwan Chun, Carsten Dormann, Stéphane Dray, Rein Halbersma, Nicholas Lewin-Koh, Jielai Ma, and Giovanni Millo. 2005. "The spdep package." *Comprehensive R Archive Network, Version*:05-83.

Bivand, Roger, Danlin Yu, Tomoki Nakaya, Miguel-Angel Garcia-Lopez, and Roger Bivand. 2017. "Package 'spgwr'." *R software package*.

Cho, Wendy K. Tam, and James G. Gimpel. 2010. "Rough Terrain: Spatial Variation in Campaign Contributing and Volunteerism." *American Journal of Political Science* 54 (1):74-89. doi: 10.1111/j.1540-5907.2009.00419.x.

Clement, Floriane, Didier Orange, Meredith Williams, Corinne Mulley, and Michael Epprecht. 2009. "Drivers of afforestation in Northern Vietnam: Assessing local variations using geographically weighted regression." *Applied Geography* 29 (4):561-576. doi: http://dx.doi.org/10.1016/j.apgeog.2009.01.003.

Darmofal, David. 2008. "The political geography of the new deal realignment." *American Politics Research*. 36 (6): 934-961.

Elhorst, J Paul (2010) Applied spatial econometrics: raising the bar. *Spatial economic analysis* 5(1): 9-28.

Fischer, Manfred M, and Arthur Getis. 2009. *Handbook of applied spatial analysis: software tools, methods and applications*: Springer Science & Business Media.

Fotheringham, AS, C Brunsdon, and Martin Charlton. 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. West Sussex, UK: John Wiley.

Lee, Duncan. 2013. "CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors." *Journal of Statistical Software* 55 (13):1-24.

Lee, Duncan. 2017. "CARBayes version 5.0: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors." https://cran.r-project.org/web/packages/CARBayes/vignettes/CARBayes.pdf.

Leroux, Brian G., Xingye Lei, Norman Breslow, M Halloran, and Berry Donald Elizabeth. 2000. "Estimation of disease rates in small areas: a new mixed model for spatial dependence." *Statistical models in epidemiology, the environment, and clinical trials*:179-191.

Nemeth, Stephen C, Jacob A Mauslein, and Craig Stapley. 2014. "The primacy of the local: Identifying terrorist hot spots using geographic information systems." *The Journal of Politics* 76 (02):304-317.

Öcal, Nadir, and Jülide Yildirim. 2010. "Regional effects of terrorism on economic growth in Turkey: A geographically weighted regression approach." *Journal of Peace Research* 47 (4): 477-489.

Wheeler, David C. 2014. "Geographically weighted regression." In *Handbook of Regional Science*, 1435-1459. Springer.

Wheeler, David, and Michael Tiefelsdorf. 2005. "Multicollinearity and correlation among local regression coefficients in geographically weighted regression." *Journal of Geographical Systems* 7 (2):161-187.