Supplemental Information for The Curse of Geography: How Governments Preempt Secession Attempts

Rob Williams*

March 5, 2020

Contents

A Descriptive Statistics

Figure A.1 presents descriptive statistics for all predictors included in the various models. Due to their skewed untransformed-distributions, *nightlights*, *population*, *capital distance*, *area*, and *GDP* are log-transformed. Figure A.1 depicts these transformed distributions. Continuous predictors are centered and scaled before analysis.

*Postdoctoral Research Associate, Department of Political Science, Washington University in St. Louis, rob.williams@wustl.edu, jayrobwilliams.com.

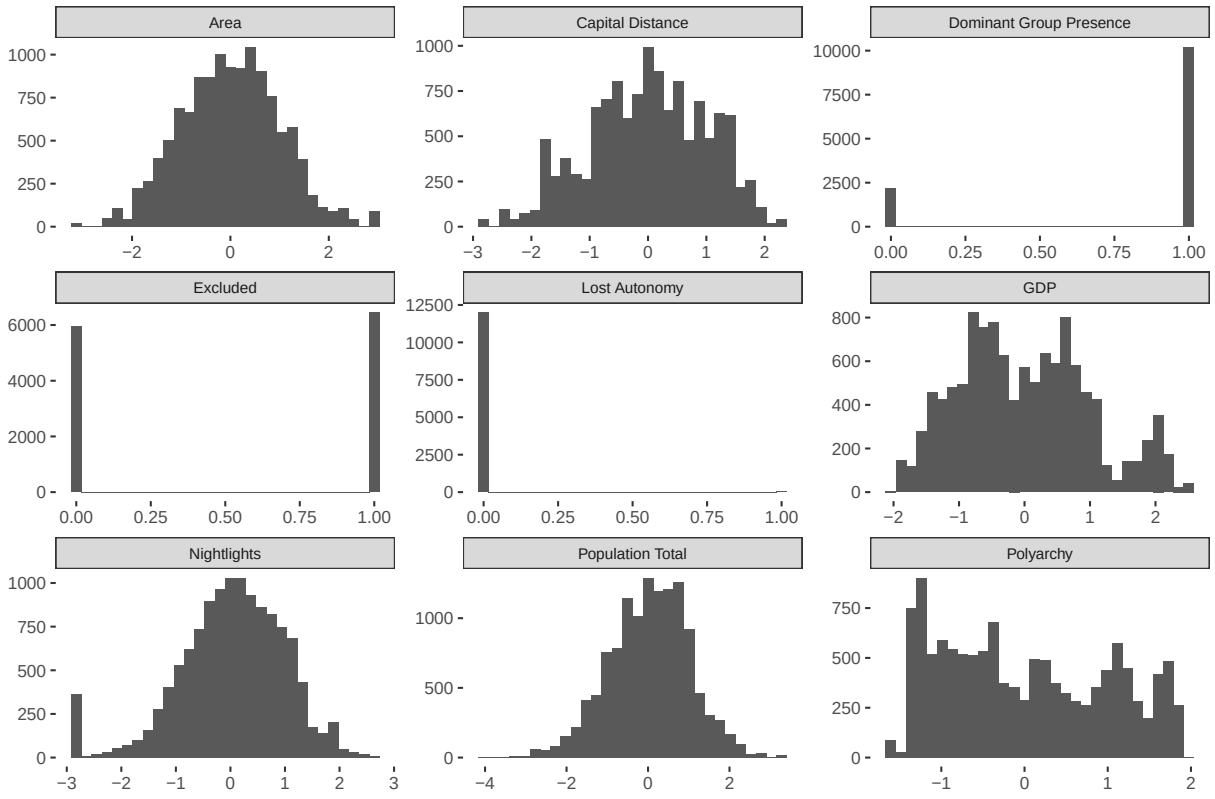
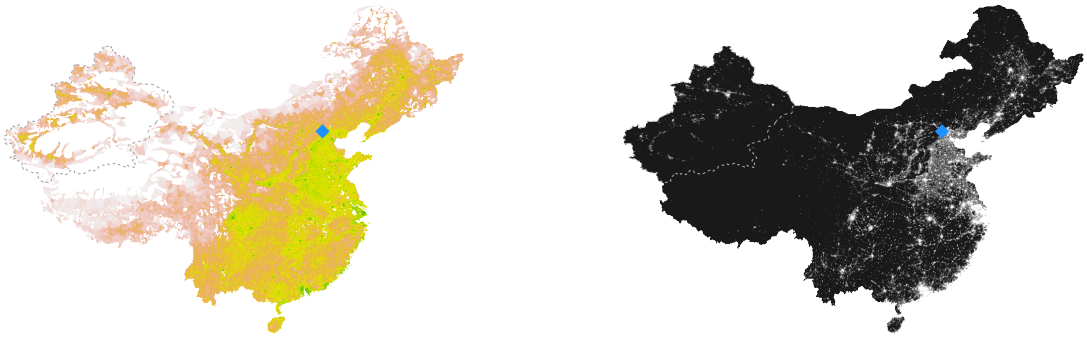


Figure A.1: Descriptive statistics for predictors included in analysis. Continuous predictors are shown centered and scaled. Demographic balance, horizontal inequality, GDP, population density, nightlights, accessibility, and area are log transformed.

B Nightlights Considerations

One of the main downsides of the DMSP OLS data is that they are unable to distinguish variation within urban areas where light levels are high due to saturation from neighboring pixels (Hsu, Baugh, Ghosh, Zhizhin & Elvidge 2015). In these cases, all pixels in a saturated area receive the maximum value. This phenomenon can be clearly seen in the area around Beijing in Figure B.1b. Luckily, I am interested in variation between entire ethnic group territories, not within individual cities, so this is less problematic for my analyses.

Using the 'cookie cutter' approach (Cederman, Buhaug & Rød 2009, Cederman, Weidmann & Gleditsch 2011, Cederman, Weidmann & Bormann 2015) requires correcting for cells where multiple group territories overlap. I do this by dividing the cell value by the number of group polygons that cover it for each cell in the raster data. For example, a substantial portion of the Syrian Kurds' settlement area overlaps with areas inhabited by Sunni Arabs. Each raster cell in these areas has its nightlights value divided by 2 before aggregation to the group level, so the Kurds and the Sunnis each receive half of the cell's nightlights. While equal distribution of nightlights, and thus state capacity, between overlapping territories is a strong assumption, it introduces less bias than ignoring the problem. Doing nothing double counts the nightlights of overlapping cells, resulting in



(a) Population (b) Nightlights

Figure B.1: China in 2013. Panel (a) displays (log) population and Panel (b) displays nightlights. The gray dashed line denotes the Xinjiang Uyghur Autonomous Region, while Beijing is represented by the blue diamond.

the state devoting 'extra' attention relative to the total investment in a given region.

Another shortcoming of these data is that the units of brightness are not inherently meaningful and are not stable over time. In addition to sensor drift within a satellite over time, values are not comparable across satellites. The maximum value in the data is 63, but that does that not mean that 63 in two years of the same satellite is equivalent, or that 63 between two satellites is equivalent. Users of the data have developed an intercalibration method to deal with these issues (Wu, He, Peng, Li & Zhong 2013). Essentially, geographic regions that do not vary over time are identified, one year of data is chosen as a reference raster, and then a model is fit using all other years to explain the invariant region in reference year. The coefficients of this model represent the difference between a given satellite-year and the reference raster. Once this model is trained, it is applied to the rest of the world, adjusting estimates for all other years so that they can be compared to the reference year. Following Wu et al. (2013), I select the Japanese prefecture of Okinawa, the American territory of Puerto Rico, and the nation of Mauritius as invariant regions to calibrate the DMSP OLS data.

C Population Considerations

As the population data (Center for International Earth Science Information Network - CIESIN - Columbia University; United Nations Food and Agriculture Programme - FAO; Centro Internacional de Agricultura Tropical - CIAT 2005, Center for International Earth Science Information Network - CIESIN - Columbia University 2015) are only available in

five year intervals, I linearly interpolate the data for the intervening years. While a rather blunt method of imputation, there are two main reasons that this approach is appropriate. First, measuring population on a yearly time scale already involves significantly loss of information. Second, a parametric imputation approach that uses variables observed in all years would either only be able to use country level variables, or would require the collection of significant amount of data at the subnational level, which is prohibitively time consuming. In either case, such an approach is unlikely to improve sufficiently over linear interpolation to justify the time and effort.

D Missing Data

Table D.1 presents the missingness of explanatory and control variables. Due to the fact that no variable has more than 10% of data missing I treat these observations as missing not at random and multiply impute them (Rubin 1987) using the `mice` package (van Buuren & Groothuis-Oudshoorn 2011), generating five imputed datasets. For all models with missing data, I estimate two chains on each imputed dataset and then pool all 10 chains together for inference.

	% Missing
Polyarchy	0.90
Lost Autonomy	2.67
GDP per capita	6.00

Table D.1: Missingness of control variables.

E Estimation and MCMC Diagnostics

I estimate the models using the Stan probabilistic programming language (Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li & Riddell 2017) in R (**?**) via the RStan interface (Stan Development Team 2017). Due to missingness in the variables, I multiply impute the missing values using the `mice` package (van Buuren & Groothuis-Oudshoorn 2011). I generate 5 imputed datasets, run two chains on each, and then perform inference on all 10 chains pooled together, averaging over the uncertainty in different imputed values (Little & Rubin 2002, 217-218).[1] I run four chains for 2,000 warmup iterations followed by 2,000 sampling iterations. All inference is based on the sampling iterations. Standard diagnostics indicate good convergence of the chains.

This section presents diagnostics of MCMC samples for Model 6. Figure E.1 displays the traceplots for the regression coefficients β. Each shade of grey represents a different

[1]Although it is possible to employ a model that jointly specifies the probability of an observation's absence alongside the parameters of interest, doing so is unnecessary in this case. When the proportion of missing information in a dataset is low, this "uncongeniality" between separate imputation and analysis models does not affect inference of imputed data (Meng 1994). The percentage of missing data in the data is .24%, so this should not affect the validity of my inferences.

chain, and the overlap between them provides evidence that the chains have converged to the stationary distribution. Figure E.2 presents a plot of the Geweke diagnostic statistics for β. The diagnostic tests whether the chain has converged to the stationary distribution by comparing the means of the first 10% and final 50% of the samples in each chain. Almost all estimates are within $\pm$ 1.96 standard deviations of the mean, offering further evidence that the chains have converged to the stationary distribution.

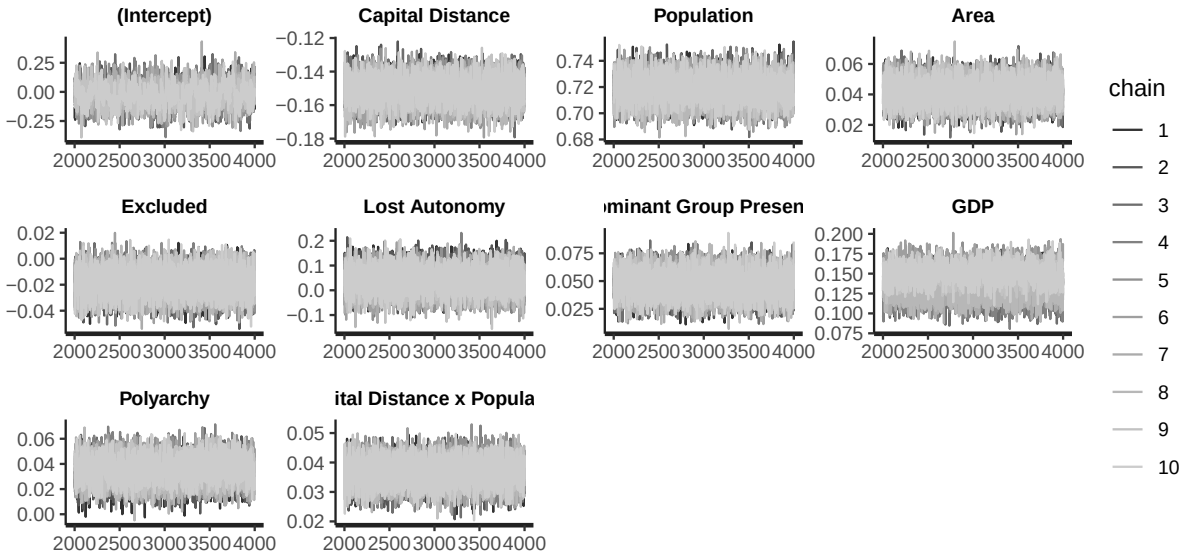


Figure E.1: Traceplot of samples for β in Model 4. Each shade of grey represents samples from one chain initialized at different starting values.

F Political Exclusion

I also estimate models explaining the level of nightlights in a group's territory using only the subsample of politically excluded groups. Table F.1 replicates Table 1, while Table F.2 replicates Table 2. The results are substantively similar, with the distribution of *lost autonomy* being significantly more uncertain.

Figure F.1 displays a similar pattern of an increasing marginal effect of population on nightlights as distance from the capital increases.

G Alternative Measures

The population Gini measure is calculated by treating each grid cell in the population data as an individual in the standard Gini index formula in Equation G.1:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}} \tag{G.1}$$

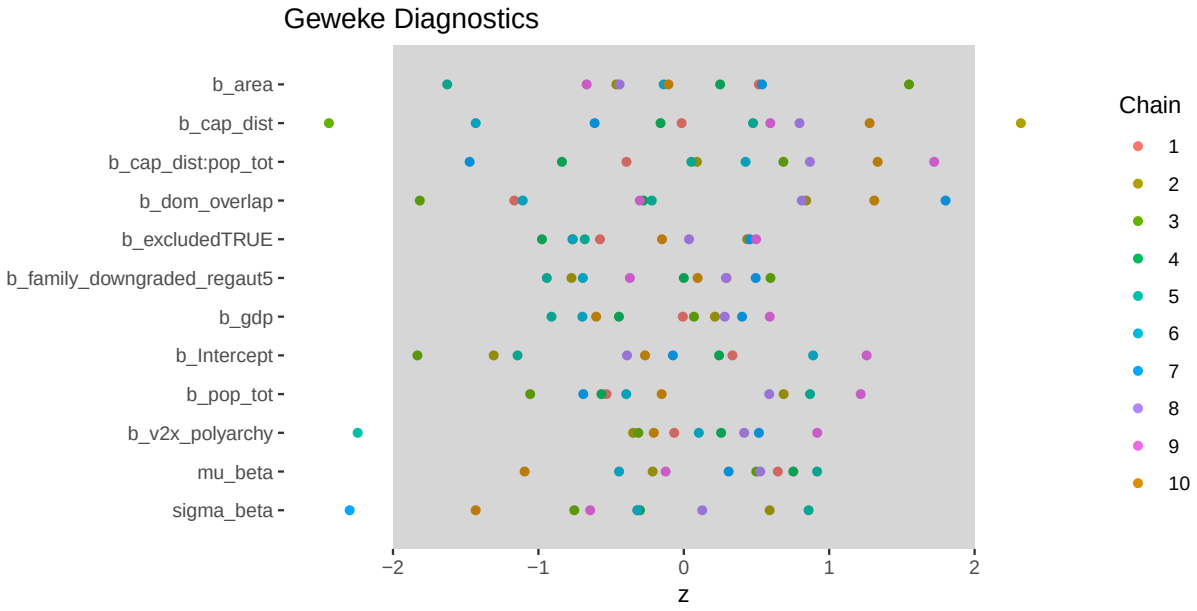


Figure E.2: Geweke diagnostic plot for β in Model 4. Dots are z-scores of the difference in means of the first 10% and final 50% of the samples in each chain.

This excellently captures the theoretical concept of population concentration. While Weidmann (2009) uses the Herfindahl-Hirschman index to measure population concentration, his unit of analysis is ethnic group territory polygons, not grid cells within a polygon. Thus, his data will have no instances of a unit with zero population. As the Herfindahl-Hirschman index is a diversity measure, it ignores observations with a zero value. This property is inappropriate when many observations have zero population and these unpopulated grid cells indicate a more concentrated population. Each grid cell with no population contributes to a higher Gini coefficient because between two territories with equal population, the one with more unoccupied areas will have a more concentrated population overall.

Figure G.1 presents results for the reestimated Models 3 & 4.

The relationship between population Gini and nightlights is similar to that of total population. Effect sizes are smaller, and model fit is worse when comparing WAIC and RMSE. However, the relationship remains positive.

H Out of Sample Accuracy

Due to the stratified nature of the data, I conduct grouped k-fold cross-validation. In this modification of k-fold cross-validation, entire states of ethnic groups are included or excluded from the folds at a time. The reported RMSE of each model thus captures its ability to predict nightlight levels in countries it has not seen before. In doing so, it provides a more honest estimate of out of sample accuracy than the random split into training and test sets provided by traditional k-fold cross-validation.

	Model 1	Model 2
Population	0.82*	
	[0.81; 0.83]	
Capital Distance		−0.39*
		[−0.42; −0.37]
(Constant)	0.08	−0.07
	[−0.16; 0.33]	[−0.24; 0.10]
σ_α	0.66*	0.94*
	[0.58; 0.76]	[0.83; 1.07]
σ_γ	0.50*	0.12*
	[0.37; 0.70]	[0.09; 0.18]
WAIC	9509.90	19196.96
5-fold RMSE	0.37	0.58
Observations	11908	11908

* 0 outside 95% credible interval

Table F.1: Linear models explaining nightlights as a function of excluded ethnic group population and capital distance. The standard deviation of the country and year random intercepts are represented by σ_α and σ_γ, respectively. Continuous variables logged and standarized.

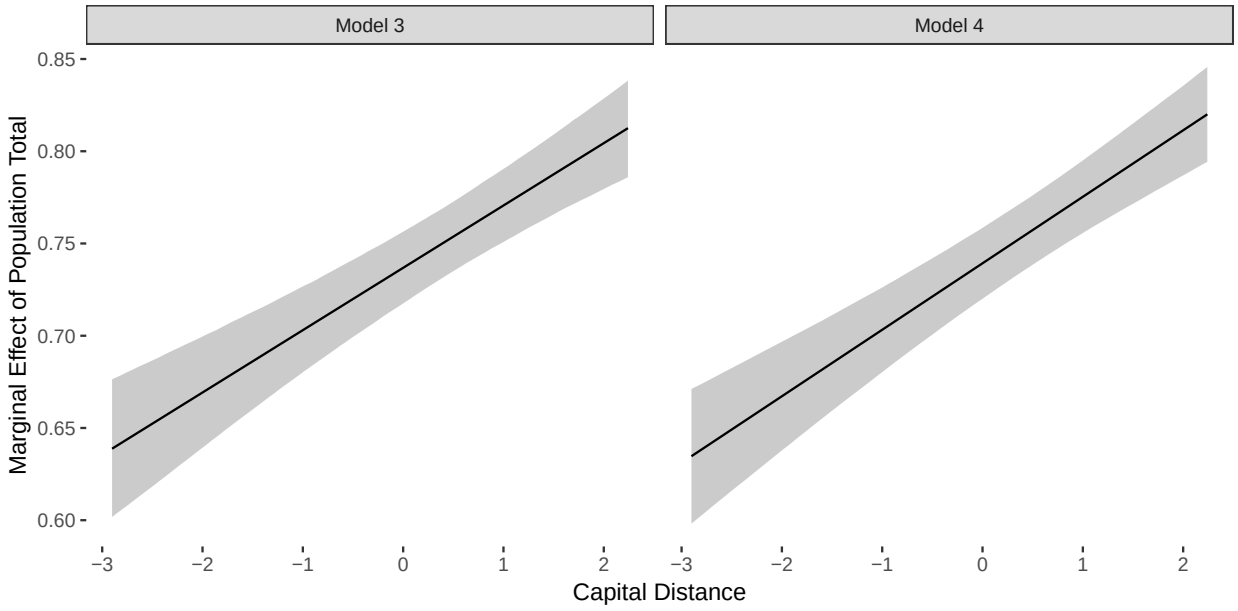


Figure F.1: Marginal effects of politically excluded ethnic group population on nighttime light levels, conditional on distance to the capital.

	Model 3	Model 4
Population	0.74*	0.74*
	[0.72; 0.76]	[0.72; 0.76]
Capital Distance	−0.16*	−0.16*
	[−0.18; −0.15]	[−0.17; −0.14]
Population Total × Capital Distance	0.03*	0.04*
	[0.02; 0.04]	[0.03; 0.05]
Area	0.05*	0.04*
	[0.03; 0.06]	[0.03; 0.06]
Dominant Group Presence		0.03*
		[0.00; 0.05]
Lost Autonomy		−0.01
		[−0.11; 0.09]
GDP_{PC}		0.25*
		[0.21; 0.28]
Polyarchy		0.05*
		[0.03; 0.07]
(Constant)	0.02	−0.05
	[−0.22; 0.26]	[−0.25; 0.15]
σ_α	0.66*	0.46*
	[0.59; 0.75]	[0.41; 0.53]
σ_γ	0.46*	0.42*
	[0.34; 0.63]	[0.31; 0.59]
WAIC	9113.98	8974.24
5-fold RMSE	0.37	0.37
Observations	11908	11908

* 0 outside 95% credible interval

Table F.2: Linear models explaining nightlights as a function of excluded ethnic group population and capital distance. The standard deviation of the country and year random intercepts are represented by σ_α and σ_γ, respectively. Continuous variables logged and standarized.

I Robustness to Nonlinearities

While marginal effects plots can improve our understanding of interactive regression models (Brambor, Clark & Golder 2006), they only provide part of the picture. Another way to improve interpretability is to estimate $\hat{Y}$ for a wide range of values and then observe the relationship between the components of the interaction term and the outcome. Figure I.1 presents the predicted value of nightlights as a function of capital distance and population, which allows us to get a more complete sense of the relationship between them. Predicted nightlights values are highest when capital distances are lowest and population is highest, which makes sense as territory close to the capital is often inhabited by ethnic groups in power and the state if frequently capable there.

	Model 6	Model 7	Model 8
Population Gini	0.41*	0.16*	0.15*
	[0.40; 0.43]	[0.14; 0.17]	[0.14; 0.16]
Capital Distance		−0.39*	−0.34*
		[−0.40; −0.37]	[−0.35; −0.32]
Population Gini × Capital Distance		0.00	0.02*
		[−0.01; 0.01]	[0.01; 0.03]
Area		0.51*	0.43*
		[0.50; 0.52]	[0.42; 0.44]
Excluded			−0.28*
			[−0.30; −0.26]
Dominant Group Presence			0.07*
			[0.05; 0.10]
Lost Autonomy			0.20*
			[0.09; 0.32]
GDP_{PC}			0.16*
			[0.12; 0.19]
Polyarchy			0.02
			[−0.00; 0.04]
(Constant)	0.13*	−0.01	0.05
	[0.00; 0.27]	[−0.13; 0.11]	[−0.05; 0.16]
σ_α	0.68*	0.68*	0.57*
	[0.60; 0.78]	[0.60; 0.78]	[0.51; 0.65]
σ_γ	0.09*	0.10*	0.07*
	[0.06; 0.12]	[0.07; 0.14]	[0.05; 0.10]
WAIC	21992.23	14428.88	13590.21
5-fold RMSE	0.57	0.43	0.41
Observations	13854	13854	13854

* 0 outside 95% credible interval

Table G.1: Linear models explaining nightlights as a function of ethnic group population Gini and capital distance. The standard deviation of the country and year random intercepts are represented by σ_α and σ_γ, respectively. Continuous variables logged and standarized.

At first brush, we would expect the level of state involvement to decline with distance from the capital as it becomes more difficult for the agents of state to travel to various locations. While distance still has a negative effect on state presence within a group's territory, highly populated territories have higher levels of state attention than similarly populous territories located closer to the centers of state power. Given the increasing cost of government activity in these more remote locations, this relationship suggests that there must be a particularly compelling reason for governments to make these investments. Fear of secession and loss of territory is a valid concern that justifies such costly behavior.

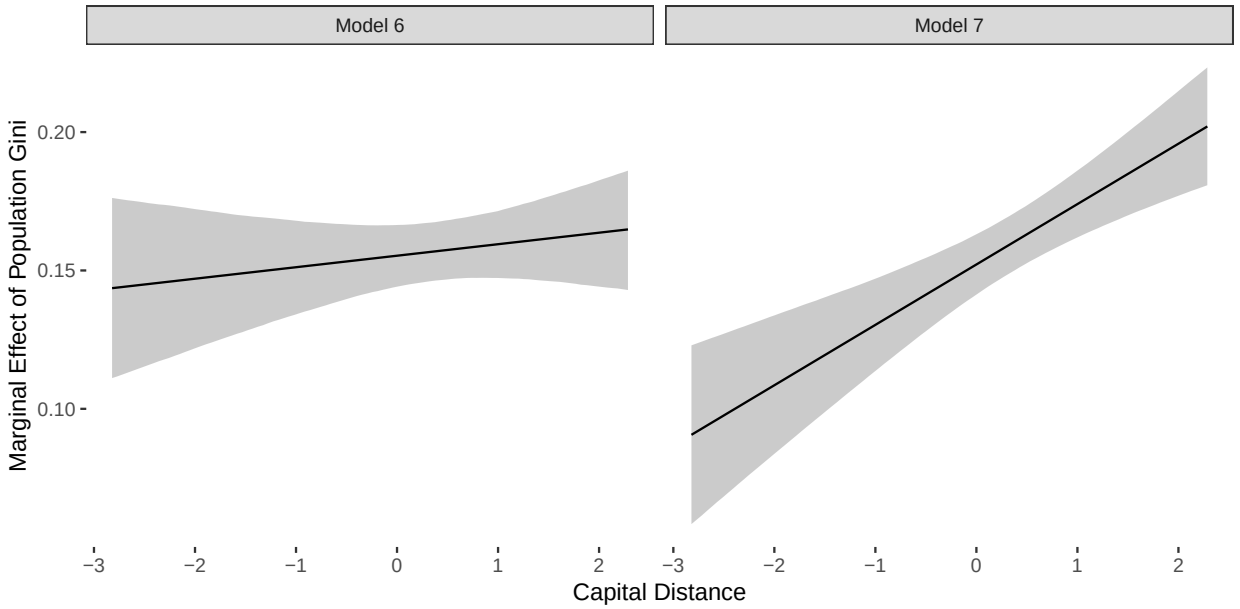


Figure G.1: Marginal effects of ethnic group population concentration on nighttime light levels, conditional on distance to the capital.

However, the smooth prediction surface highlights the simplification entailed in the model and emphasizes that it may not reflect more complicated relationships between capital distance, population, and nightlights. To address these concerns, I fit a random forest model to the data. A random forest is an ensemble of regression trees (Breiman 1984), each trained on a subset of the data (Breiman 2001). While random forests are designed to maximize predictive accuracy, they can also be used to detect nonlinearities in the relationship between variables and outcomes (Breiman 1984).

Figure I.2 presents a partial dependence plot (Friedman 2001, Greenwell 2017) of the relationship between population, capital distance, and nightlights.[2] A slight nonlinearity is observable in the lower 2/3 of the plot, where areas with lower population have higher nightlights close to the capital and very far away. This pattern supports my argument that states are increasing their capacity in areas most prone to secession because similarly populated areas at a middling distance from the capital have lower nightlights values. State capacity is naturally high in areas close the the capital, and strategically high in areas far from the capital and more governable.

The random forest model is fit using the `randomForest` package (Cutler & Wiener 2018) in R. The model is fit using the default parameters of 500 trees, $\frac{p}{3} = 1$ variable randomly chosen to make each split, $\frac{2}{3}$ of the data randomly sampled for each tree, minimum terminal node size of 5, and no cap on the number of terminal nodes in a tree. Partial dependence is assessed using the `pdp` package (Greenwell 2017) in R as a function of capital distance and population, marginalizing over the effect of area.

[2]This model includes population, capital distance, and the size of a group's territory as predictors. For full details, see the Supplemental Information.

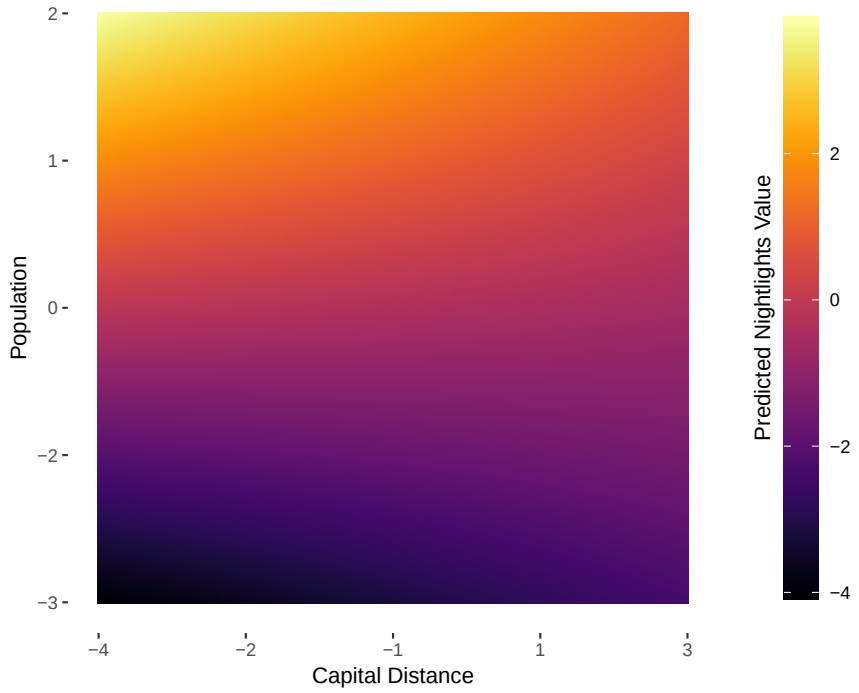


Figure I.1: Predicted nightlights as a function of capital distance and population.

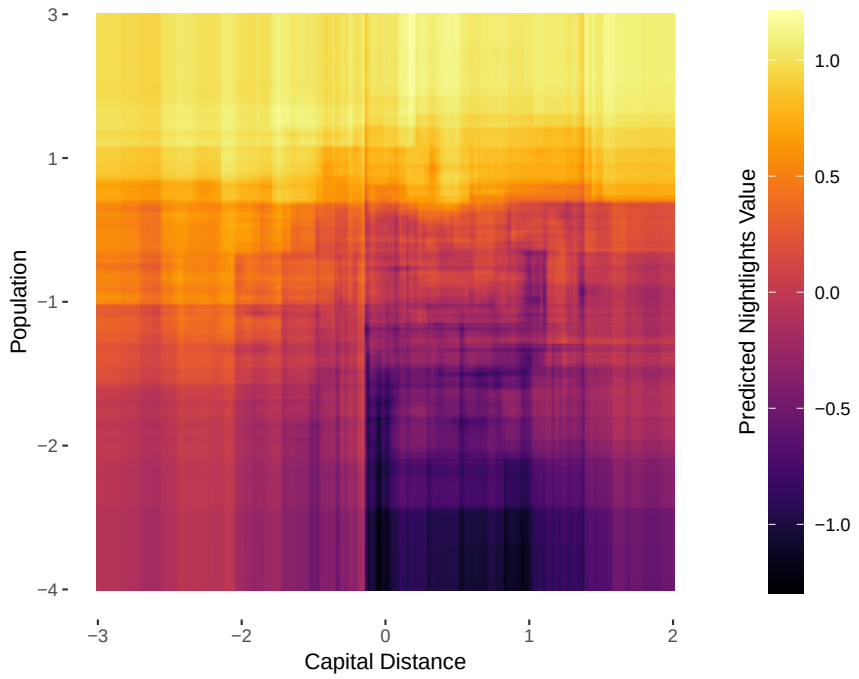


Figure I.2: Partial dependence of nightlights on capital distance and population.

References

Brambor, Thomas, William Roberts Clark & Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1):63–82.

Breiman, Leo. 1984. *Classification and Regression Trees*. New York, N.Y.: Chapman & Hall.

Breiman, Leo. 2001. "Random Forests." *Mach. Learn.* 45(1):5–32.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. "Stan A Probabilistic Programming Language." *Journal of Statistical Software* 76(1).

Cederman, Lars-Erik, Halvard Buhaug & Jan Ketil Rød. 2009. "Ethno-Nationalist Dyads and Civil War A GIS-Based Analysis." *Journal of Conflict Resolution* 53(4):496–525.

Cederman, Lars-Erik, Nils B. Weidmann & Kristian Skrede Gleditsch. 2011. "Horizontal Inequalities and Ethnonationalist Civil War: A Global Comparison." *The American Political Science Review* 105(3):478–495.

Cederman, Lars-Erik, Nils B. Weidmann & Nils-Christian Bormann. 2015. "Triangulating Horizontal Inequality: Toward Improved Conflict Analysis." *Journal of Peace Research* 52(6):806–821.

Center for International Earth Science Information Network - CIESIN - Columbia University. 2015. Gridded Population of the World, Version 4 (GPWv4): Population Density. Technical report NASA Socioeconomic Data and Applications Center (SEDAC) Palisades, NY: .

Center for International Earth Science Information Network - CIESIN - Columbia University; United Nations Food and Agriculture Programme - FAO; Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. Technical report NASA Socioeconomic Data and Applications Center (SEDAC) Palisades, NY: .

Cutler, Fortran original by Leo Breiman and Adele & R. port by Andy Liaw and Matthew Wiener. 2018. "randomForest: Breiman and Cutler's Random Forests for Classification and Regression.".

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29(5):1189–1232.

Greenwell, Brandon M. 2017. "Pdp: An R Package for Constructing Partial Dependence Plots." *The R Journal* 9(1):421–436.

Hsu, Feng-Chi, Kimberly E. Baugh, Tilottama Ghosh, Mikhail Zhizhin & Christopher D. Elvidge. 2015. "DMSP-OLS Radiance Calibrated Nighttime Lights Time Series with Intercalibration." *Remote Sensing* 7(2):1855–1876.

Little, Roderick & Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. ed. Hoboken, N.J.: Wiley.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9(4):538–558.

Rubin, Donald. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Stan Development Team. 2017. "Rstan: R Interface to Stan.".

van Buuren, Stef & Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(3).

Weidmann, Nils B. 2009. "Geography as Motivation and Opportunity Group Concentration and Ethnic Conflict." *Journal of Conflict Resolution* 53(4):526–543.

Wu, Jiansheng, Shengbin He, Jian Peng, Weifeng Li & Xiaohong Zhong. 2013. "Intercalibration of DMSP-OLS Night-Time Light Data by the Invariant Region Method." *International Journal of Remote Sensing* 34(20):7356–7368.