Replication

Tahmid Ahmed

4/12/2021

Introduction

For my replication project, I will be replicating the results from "A global analysis of progress in household electrification" by Alkin et al. (2018). This paper focuses on the United Nations Sustainable Development Goal (SDG) #7, which strives for access to affordable, sustainable, and modern energy. The authors emphasize that access to energy is crucial to human well-being as literature highlights that low-cost electricity and clean energy can lead to healthier and more productive lives (Dinkelman, 2011; Barnes, 2014; Greenstone, 2014; Aklin et al., 2016). The authors note that the World Bank currently has a data set that measures progress in energy over time called the Global Tracking Framework (GTF). However, even though the data set is ground breaking in terms of measuring progress towards universal access to energy, the data lacks a robust baseline for national electrification rates across the globe since the data on rural electrification rates are very poor. The data is also not only incomplete and only available for the post-1990 period but also consists of multiple values that are based on simulation or interpolation. The authors further argue that in order to understand how to achieve universal access to electricity, scientists and policymakers must compare changes in total and urban electrification rates to that of rural electrification rates based on high-quality data and learn from past successes and failures. Thus, the main research question the authors are answering is: How can we better understand progress toward universal electricity access with high-quality data? To address this question, the authors introduced a new database of total, urban, and rural electrification rates around the world. This database uses nationally representative surveys and official reports from 124 countries. It is also the most detailed data set on electrification to date as it does not have data based on simulations, has data for the pre-1990 period, and includes clear coding criteria.

The authors went through a rigorous process to create the data. The authors use a multitude of sources to create the database, including the national census where available, nationally represented surveys that asked questions about electricity access, published journal articles, and government agency statistics. The database consists of 1065 observations for 124 countries across 15 different regions in Asia, Africa, Latin America, Middle East, Eastern Europe and the Caribbean between years 1949–2015. The database would also have other variables, such as polity scores, urban population share, and hydro potential.

Additionally, the authors ran a coherent analysis on the data to explore global electrification trends. In their first model, the authors made a linear model with least squares (OLS) and standard errors clustered by country. The dependent variable was electrification rate and the predictor was time. They found that their results are robust. Moreover, to examine the variation in pace of electrification, they first plotted the relationship between logarithmized GDP per capita (USD, 2010 constant prices) and rural or total electrification rates. They then used a generalized linear model to explore the nonlinear relationship. Thus, in their second model, they also made another linear model. The model used the following covariates- Population Density (k/sqkm); Urban Pop. Share; Hydro Potential per Capita (log) (kWh/k); Oil, Gas, Coal Rents (% of GDP); Nat. Resource Rents (% of GDP); and Democracy. The dependent variable was then deviations from the electrification rates predicted by logarithmized GDP per capita alone. The results showed that high population densities are a robust predictor of electrification rates totally and rurally.

Moreover, the paper concluded that previous estimates of progress toward universal electrification over time are very underestimated due to data caused from stimulations. They have also concluded that total, rural, and urban electrification rates have increased dramatically in the world. Last but not least, they also concluded that while GDP per capita is a strong predictor of electrification rates, urbanization and high population densities have also explained that some poor countries are able to improve their electrification rates at a quick rate.

In my replication, I have replicated all models in the paper using R. I was able to have the same conclusions and used different libraries, specifically plm, to make models with fixed effects and clustered standard errors. I also made similar maps and plots using ggplot. From my replication, I was able to conclude very similar results as the authors. An additional conclusion I made is that in addition to population density, natural resource rents (% of GDP) is a significant predictor in the second model with a p-value less than 0.05. This might be because regions with better economies are able to better afford technological advancements for access to electricity. Moreover, another conclusion I made was that even though many countries had higher electrification rates over time, access to electricity is very segregated in many continents and it is important that policymakers and scientists address this segregation for better wellbeing among all people. I also expanded .... The implications are that the United Nations and countries can learn from the results of this project, and inform better policies in terms of helping regions, especially those from rural areas, get better access to electricity.

Walk Through and Analysis of Replication

In our analysis, we will first explore trends in global electrification. To do so, I replicated the maps in the journal. These plots essentially depict progress in overall electrification globally. To create the plots, I had to use ggplot. I first read data for replication. The uscoord_df refers to the geographic coordinates of countries and comes from the TM World Borders Dataset 0.3 (cite), which is a dataset that contains geographic information about countries and meant to be used for the public. I also read usdb_df, which contains the electrification rates of countries and provided by the authors. I then joined both data sets (by country name), which made an object called electric to create ggplots. Moreover, once I had the object named electric, I created maps using geom_polygon. With geom_polygon I am essentially able to make shapes and since the joined data set has all the geographical coordinates for each country, I can make shapes of countries. Thus, I used ggplot with x is longitude and y is latitude and filled shapes based on electrification rates. The results were essentially the same as the paper.

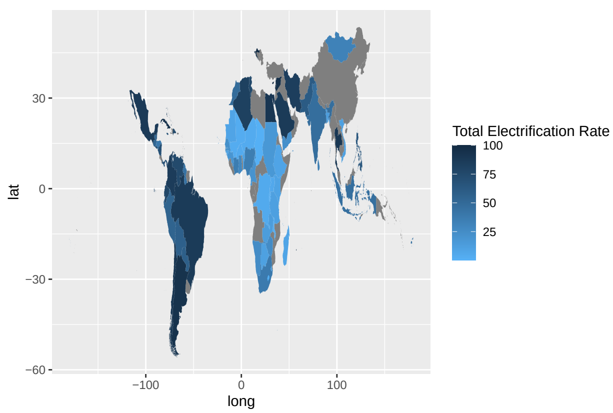

Figure 1a. 1986−1994

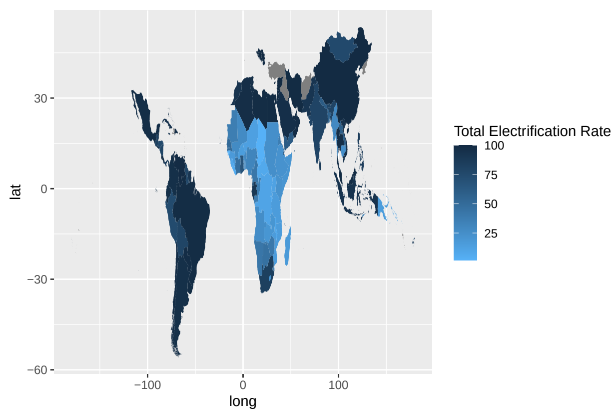

Figure 1b. 2006−2014

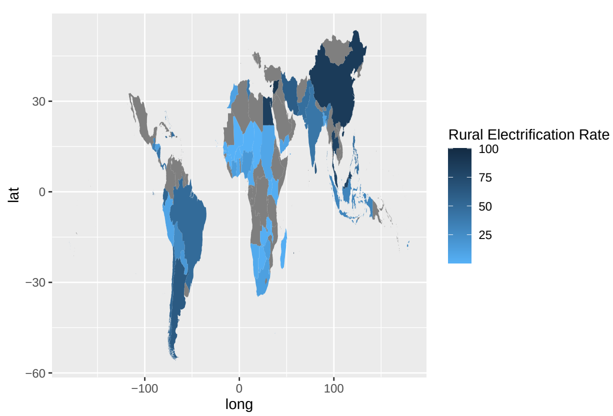

Figure 2a. 1986−1994

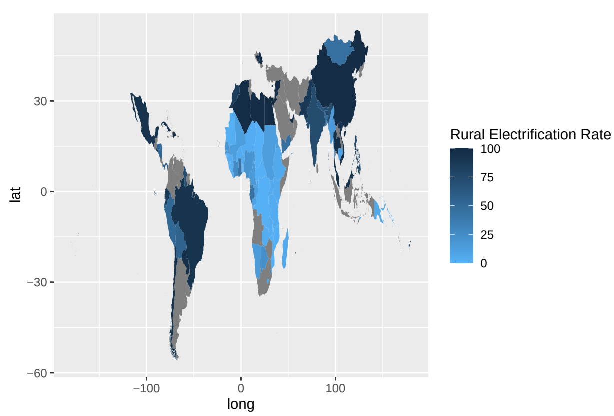

Figure 2b. 2006−2014

Figure 1a and Figure 1b show the electrification rate during around 1990 while Figure 2a and Figure 2b show the electrification rate at around 2010. Based on the first pair of maps, we can see that total electrification rates have increased across all regions over time. Additionally, based on these maps, Latin America, Asia, and the Middle East and North Africa seem to be reaching universal electrification while Sub-Saharan Africa is slowly behind. For the second pair of maps, we can see similar results for rural electrification rates. While there is less data in 1990 for rural electrification rates, especially in Africa, we can still see a dramatic increase in Latin America and some progress in other nations. Moreover, Sub-Saharan Africa still seems to be lacking in progress in rural electrification rates as not much has changed since 1990.

Now that we have our maps, we can first analyze our first model. In our first model, we are using a linear regression to predict progress in total, rural, and urban electrification rates over time. The electrification rate is regressed on country-specific intercepts, including country fixed effects, and on a linear time trend. Standard errors are clustered by country as well. Tables 1-6 pertain to different regions, including all regions, and based on data created by the authors. Table 7 pertains to the World Bank GTF data. E/SE Asia stands for East/Southeast Asia. LA stands for Latin America. MENA stands for Middle East and North Africa. SA stands for South Asia and SSA stands for Sub-Saharan Africa.

In R, I used the plm package to create my models. Since we are looking at data across different cross sections of time or essentially panel data, I used the plm function to make my models. The plm package allows you to make linear models that have panel data. Moreover, I also used plm because this is a fixed effects regression model. In the plm function, the fixed effects estimator is also called the within estimator, so I set model = 'within'. Now I can make the fixed effects regression. Moreover, since standard errors are clustered by country, we can use vcovHC on the model from the sandwich package to extract the covariance of the coefficient estimates in the model and set cluster = "group" to cluster by country. I then found the square root of the covariance in order to get the standard error. I repeated this process for each type of electrification rate and for different regions, including all regions. I also repeated this process based on data from the World Bank. Stargazer was used to make the respective tables. The results were nearly identical to that of the paper.

Table 1: Regression Results for Total Electrification Rate

	Dependent variable:			
	Total Electrification Rate			
	All	E/SE Asia	LA	MENA
	(1)	(2)	(3)	(4)
time	1.295^{***}	2.185^{***}	1.080^{***}	1.446^{***}
	(0.084)	(0.395)	(0.083)	(0.189)
Observations	927	127	222	116
R^2	0.695	0.755	0.843	0.679
Adjusted R^2	0.648	0.719	0.821	0.620
F Statistic	$1{,}825.652^{***}$ (df = 1; 802)	338.589^{***} (df = 1; 110)	$1{,}039.223^{***}$ (df = 1; 194)	205.384^{***} (df = 1; 97)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: Regression Results for Total Electrification Rate Cont.

	Dependent variable:	
	Total Electrification Rate	
	SA	SSA
	(1)	(2)
time	1.849^{***}	0.992^{***}
	(0.072)	(0.131)
Observations	65	381
R^2	0.887	0.556
Adjusted R^2	0.874	0.495
F Statistic	449.266^{***} (df = 1; 57)	418.299^{***} (df = 1; 334)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: Regression Results for Rural Electrification Rate

	Dependent variable:			
	Rural Electrification Rate			
	All	E/SE Asia	LA	MENA
	(1)	(2)	(3)	(4)
time	1.513^{***}	1.941^{***}	1.569^{***}	2.287^{***}
	(0.100)	(0.293)	(0.141)	(0.202)
Observations	643	85	158	62
R^2	0.701	0.793	0.819	0.856
Adjusted R^2	0.638	0.748	0.784	0.817
F Statistic	$1,244.239^{***}$ (df = 1; 531)	264.411^{***} (df = 1; 69)	595.520^{***} (df = 1; 132)	285.089^{***} (df = 1; 48)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression Results for Rural Electrification Rate Cont.

	Dependent variable:	
	Rural Electrification Rate	
	SA	SSA
	(1)	(2)
time	2.110^{***}	0.663^{***}
	(0.160)	(0.131)
Observations	37	289
R^2	0.859	0.369
Adjusted R^2	0.831	0.249
F Statistic	182.641^{***} (df = 1; 30)	141.429^{***} (df = 1; 242)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5: Regression Results for Urban Electrification Rate

	Dependent variable:			
	Urban Electrification Rate			
	All	E/SE Asia	LA	MENA
	(1)	(2)	(3)	(4)
time	0.961***	1.270**	0.732***	0.949***
	(0.073)	(0.510)	(0.099)	(0.173)
Observations	567	67	135	49
R^2	0.565	0.484	0.729	0.432
Adjusted R^2	0.460	0.332	0.663	0.243
F Statistic	591.322*** (df = 1; 456)	47.753*** (df = 1; 51)	289.861*** (df = 1; 108)	27.381*** (df = 1; 36)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Regression Results for Urban Electrification Rate Cont.

	Dependent variable:	
	Urban Electrification Rate	
	SA	SSA
	(1)	(2)
time	1.359***	1.217***
	(0.152)	(0.113)
Observations	36	269
R^2	0.856	0.572
Adjusted R^2	0.826	0.486
F Statistic	171.787*** (df = 1; 29)	298.382*** (df = 1; 223)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 7: Regression Results for Total Electrification Rate (WB)

	Dependent variable:		
	Total Electrification Rate	Rural Electrification Rate	Urban Electrification Rate
	All	Rural	Urban
	(1)	(2)	(3)
time	1.017***	0.978***	0.999***
	(0.086)	(0.109)	(0.133)
Observations	2,478	2,337	1,857
R^2	0.578	0.434	0.336
Adjusted R^2	0.556	0.406	0.295
F Statistic	3,222.949*** (df = 1; 2353)	1,704.139*** (df = 1; 2226)	884.312*** (df = 1; 1747)
Note:			*p<0.1; **p<0.05; ***p<0.01

The results of the models were very telling. The coefficients represent "the average annual change in the percentage of electrified households/populations" (Aklin et al., 2018). Thus, we can see that the World Bank greatly underestimates progress in electrification in terms of total, rural, and urban electrification. For example, the rate of progress for total electrification (coefficient is 1.30) is about 28% higher than the coefficient (coefficient is .02) in the model using World Bank data. Likewise, the rate of progress for rural electrification (coefficient is 1.51) is about 53% higher than the coefficient (coefficient is .98) in the model using World Bank data. However, the coefficient for urban electrification rate is about the same for both models using the author's data and the model using the World Bank simulations. Thus, based on the World Bank model, data with simulations seem to underestimate the rate of improvement in electrification whereas data without simulations have a much more positive effect. Additionally, the model shows that Sub-Saharan Africa made a lot less improvement in electrification compared to other regions. In fact, Latin America, which also has impoverished areas, has a higher coefficient for total electrification (1.08) compared to Sub-Saharan Africa (.99) but only slightly. However, Sub-Saharan Africa has shown to be the only region where urban electrification has made faster progress than that of rural electrification whereas all other regions had the opposite. This could be because regions outside of Sub-Saharan already have high urban electrification rates so they won't see much improvement. Another key finding is that all the models are significant where the time predictor has a p-value less than 0.05, indicating that electrification does seem to be changing over time. Also, the fact that the Middle East and North Africa has a lot higher coefficients for both total electrification rates and rural electrification rates than Sub-Saharan Africa shows that Africa is very divided in terms of progress in electrification. It would be interesting to examine why this difference is apparent between the northern and southern parts of Africa. One reason is that Sub-Saharan Africa is isolated from the world and may not be exposed to technological measures for access to more electricity.

We further analyze variation in the pace of electrification in the world. Based on the paper, one reason that there is variation in the pace of electrification is that there is a strong relationship between economic growth and electrification (Barnes, 2014; Greenstone, 2014). Moreover, studies have also shown that electrification may be higher for higher income households since they can afford electricity (Foley, 1992). Thus, the authors looked to create a baseline by plotting logarithmized GDP per capita (USD, 2010 constant prices and based on general additive model) on the x axis versus electrification rate (total, rural, urban) on the y axis. In R, I used ggplot to create these plots with the same axes shown in Figure 3. Moreover, since the authors added a loess smoothing function to see the nonlinear effect of GDP per capita, I added method = loess to geom_smooth. I colored the points by region like the paper and got essentially very similar results.

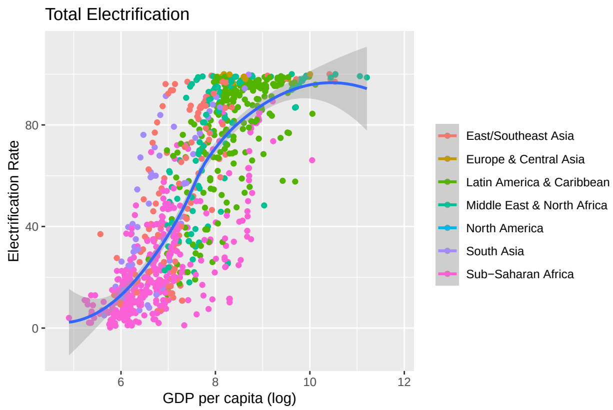

Figure 3a

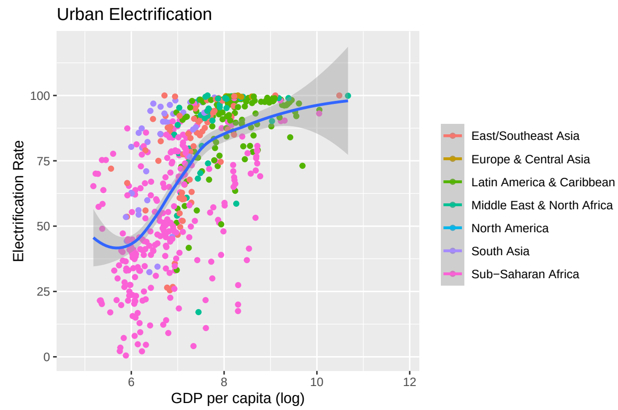

Figure 3b

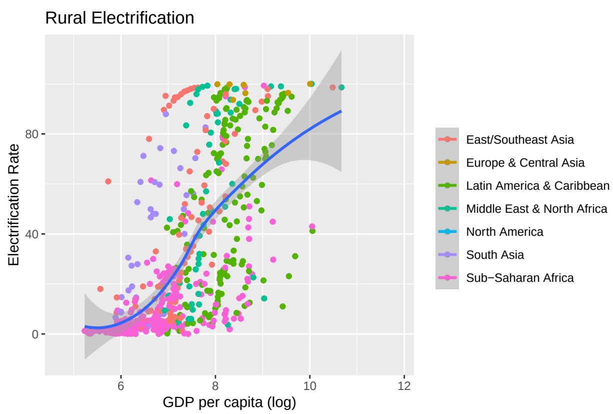

Figure 3c

Based on Figure 3, the relationship between logarithmized GDP per capita and electrification is robust. In both rural total electrification plots, electrification rates seem to be increasing dramatically with income. However, as the authors noted, there is also variation in the predicted relationship based on the general additive model in that countries tend to underperform or over perform when GDP per capita is neither very high nor very low.

With these trends, the authors create their second type of model. The authors consider the covariates that may predict the deviations from the pattern predicted by GDP per capita alone. Thus, "the dependent variable in the new model is deviation from the value predicted by logarithmized GDP per capita," which is measured by "observed (total, rural, or urban) electrification rate less the expected electrification rate based on logarithmized GDP per capita." Additionally, the authors looked at the following possible predictors that may explain the over performance or underperformance - population density since it can reduce the cost of household electrification (Oparaku, 2003); natural resource rents since they can provide funds for national electrification programs (Squalli, 2007); and democracy since democratic countries have stronger incentives to provide electricity to households in comparison to authoritarian governments (Min, 2015). They examine the extent of these predictors on the dependent variable of this model to explain the "divergence in the degree of total, rural, and urban electrification across the world" (Alkin et al., 2018).

To replicate their models, I had to use plm for the same reasons as to why I used plm for the first type of model. However, I used method = "pooling" here as this is essentially a pooled regression now. Since this model also had standard errors clustered by country, I essentially did the same thing as before. The models differ by the inclusion and exclusion of the unique predictors and the last model includes all explanatory variables. Tables 8 through 16 show the results of the models where Tables 10, 13, and 16 include all predictors.

Table 8: Regression Results for Total Electrification Rate

	Dependent variable:		
	Total Electrification Rate		
	(1)	(2)	(3)
Population Density (k/sqkm)	21.599***		
	(7.595)		
Urban Pop. Share		0.169***	
		(0.059)	
Hydro Potential per Capita (log) (kWh/k)			−1.440
			(0.911)
Time	0.626***	0.588***	0.610***
	(0.073)	(0.074)	(0.073)
Constant	−32.507***	−35.664***	−26.701***
	(3.461)	(3.474)	(4.386)
Observations	894	899	898
R^2	0.189	0.178	0.157
Adjusted R^2	0.187	0.176	0.155
F Statistic	103.941*** (df = 2; 891)	97.131*** (df = 2; 896)	83.570*** (df = 2; 895)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 9: Regression Results for Total Electrification Rate Cont.

	Dependent variable:		
	Total Electrification Rate		
	(1)	(2)	(3)
Oil, Gas, Coal Rents	−0.263*		
	(0.141)		
Nat. Resource Rents		−0.435***	
		(0.111)	
Democracy			0.098
			(0.277)
Time	0.667***	0.720***	0.634***
	(0.086)	(0.085)	(0.090)
Constant	−31.212***	−31.088***	−30.833***
	(4.425)	(4.467)	(4.111)
Observations	849	838	863
R^2	0.126	0.164	0.150
Adjusted R^2	0.124	0.162	0.148
F Statistic	61.033*** (df = 2; 846)	81.607*** (df = 2; 835)	75.937*** (df = 2; 860)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 10: Regression Results for Total Electrification Rate Cont.

	Dependent variable:
	Total Electrification Rate
Population Density (k/sqkm)	26.626^{**}
	(12.587)
Urban Pop. Share	0.230^{***}
	(0.067)
Hydro Potential per Capita (log) (kWh/k)	-0.088
	(1.011)
Oil, Gas, Coal Rents	0.007
	(0.229)
Nat. Resource Rents	-0.394^{**}
	(0.185)
Democracy	-0.298
	(0.308)
Time	0.644^{***}
	(0.125)
Constant	-39.798^{***}
	(7.193)
Observations	806
R^2	0.245
Adjusted R^2	0.238
F Statistic	36.974^{***} (df = 7; 798)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 11: Regression Results for Rural Electrification Rate

	Dependent variable:		
	Rural Electrification Rate		
	(1)	(2)	(3)
Population Density (k/sqkm)	56.496^{***}		
	(20.852)		
Urban Pop. Share		0.059	
		(0.112)	
Hydro Potential per Capita (log) (kWh/k)			-3.409^{**}
			(1.473)
Time	0.707^{***}	0.758^{***}	0.710^{***}
	(0.112)	(0.104)	(0.112)
Constant	-39.893^{***}	-39.865^{***}	-28.027^{***}
	(5.394)	(7.385)	(7.166)
Observations	616	621	621
R^2	0.223	0.140	0.168
Adjusted R^2	0.220	0.137	0.165
F Statistic	87.947^{***} (df = 2; 613)	50.167^{***} (df = 2; 618)	62.439^{***} (df = 2; 618)
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 12: Regression Results for Rural Electrification Rate Cont.

	Dependent variable:		
	Rural Electrification Rate		
	(1)	(2)	(3)
Oil, Gas, Coal Rents	-0.248		
	(0.225)		
Nat. Resource Rents		-0.504^{***}	
		(0.181)	
Democracy			0.038
			(0.448)
Time	0.717^{***}	0.810^{***}	0.762^{***}
	(0.149)	(0.149)	(0.122)
Constant	-34.191^{***}	-35.499^{***}	-37.655^{***}
	(7.915)	(7.843)	(5.606)
Observations	579	571	611
R^2	0.086	0.120	0.135
Adjusted R^2	0.083	0.116	0.132
F Statistic	27.250^{***} (df = 2; 576)	38.547^{***} (df = 2; 568)	47.356^{***} (df = 2; 608)
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 13: Regression Results for Rural Electrification Rate Cont.

	Dependent variable:
	Rural Electrification Rate
Population Density (k/sqkm)	52.300^{**}
	(23.593)
Urban Pop. Share	0.147
	(0.113)
Hydro Potential per Capita (log) (kWh/k)	-1.250
	(1.685)
Oil, Gas, Coal Rents	0.254
	(0.331)
Nat. Resource Rents	-0.495^{*}
	(0.278)
Democracy	-0.350
	(0.464)
Time	0.758^{***}
	(0.171)
Constant	-41.868^{***}
	(10.016)
Observations	566
R^2	0.222
Adjusted R^2	0.213
F Statistic	22.787^{***} (df = 7; 558)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 14: Regression Results for Urban Electrification Rate

	Dependent variable:		
	Urban Electrification Rate		
	(1)	(2)	(3)
Population Density (k/sqkm)	32.294^{***}		
	(8.210)		
Urban Pop. Share		0.120	
		(0.074)	
Hydro Potential per Capita (log) (kWh/k)			-0.576
			(1.159)
Time	0.293^{***}	0.308^{***}	0.312^{***}
	(0.085)	(0.078)	(0.087)
Constant	-16.252^{***}	-19.157^{***}	-13.172^{**}
	(4.204)	(4.439)	(5.686)
Observations	562	567	567
R^2	0.094	0.056	0.044
Adjusted R^2	0.091	0.052	0.040
F Statistic	29.001^{***} (df = 2; 559)	16.667^{***} (df = 2; 564)	12.846^{***} (df = 2; 564)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 15: Regression Results for Urban Electrification Rate Cont.

	Dependent variable:		
	Urban Electrification Rate		
	(1)	(2)	(3)
Oil, Gas, Coal Rents	-0.371^{**}		
	(0.182)		
Nat. Resource Rents		-0.427^{***}	
		(0.160)	
Democracy			-0.026
			(0.260)
Time	0.326^{***}	0.424^{***}	0.336^{***}
	(0.110)	(0.102)	(0.096)
Constant	-13.676^{**}	-15.907^{***}	-15.341^{***}
	(5.818)	(5.521)	(4.622)
Observations	526	518	557
R^2	0.052	0.081	0.043
Adjusted R^2	0.048	0.078	0.040
F Statistic	14.251^{***} (df = 2; 523)	22.760^{***} (df = 2; 515)	12.492^{***} (df = 2; 554)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 16: Regression Results for Urban Electrification Rate Cont.

	Dependent variable:
	Urban Electrification Rate
Population Density (k/sqkm)	40.220^{***}
	(11.760)
Urban Pop. Share	0.165^{**}
	(0.080)
Hydro Potential per Capita (log) (kWh/k)	1.301
	(1.206)
Oil, Gas, Coal Rents	−0.044
	(0.340)
Nat. Resource Rents	−0.391
	(0.290)
Democracy	-0.567^{*}
	(0.301)
Time	0.451^{***}
	(0.136)
Constant	-29.685^{***}
	(8.374)
Observations	513
R^2	0.162
Adjusted R^2	0.150
F Statistic	13.907^{***} (df = 7; 505)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results from models were very insightful. As the authors also mentioned, population densities are a robust predictor of overperformance in both total and rural electrification. For example, the range for the population density coefficient ranges from 21.6 to 56.5 and the coefficient always had the highest value among other predictors. Additionally, the relationship between urban population share and rural electrification is fairly weak and didn't show statistical significance. The value of the coefficient of urban population share was small most of the time and only significant in three out of the six models. This might be because it is easier to electrify in urban regions than rural regions as households are close and clustered together. Moreover, another interesting finding is that hydro potential per capita seemed to only be significant in the model with rural electrification where all other predictors were excluded. This is interesting and it could be that rural areas are becoming more technologically advanced over time in terms of hydropower. The authors also pointed this out but another finding is that there is a negative correlation relationship between fossil fuel rents and electrification. However, I also noticed that natural resource rents is a significant predictor in most of the models with a p-value less than 0.05. A possible reason is that regions with better economies can afford the technological resources needed to access electricity. Democracy was also shown as a weak predictor in that it was very small and not significant.

Extensions

In the first extension, I will be

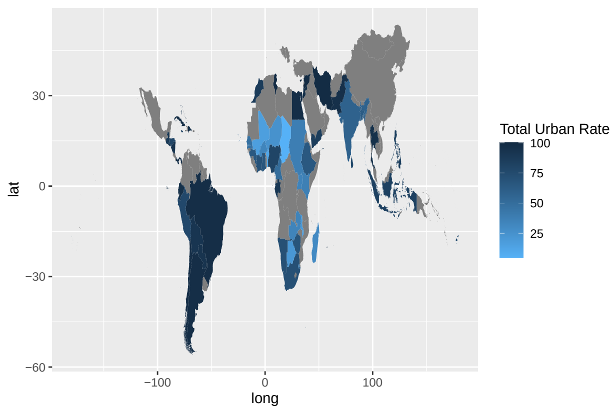

Figure 4a. 1986−1994

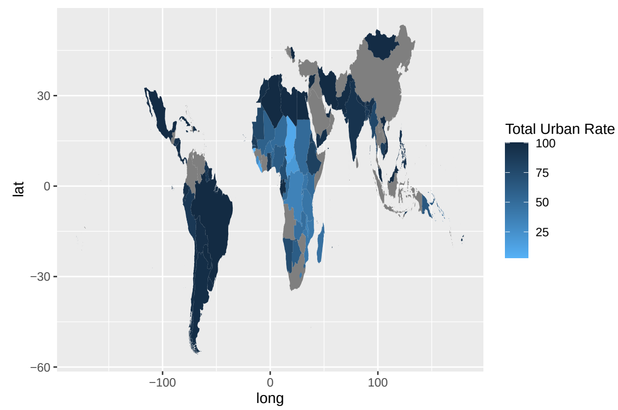

Figure 4b. 2006−2014

Table 17: Regression Results from Extension.

	Dependent variable:
	Total Electrification Rate
Log Population	2.596^{***}
	(0.964)
Time	0.626^{***}
	(0.078)
Constant	-72.275^{***}
	(16.500)
Observations	900
R^2	0.197
Adjusted R^2	0.195
F Statistic	109.886^{***} (df = 2; 897)
Note:	*p<0.1; **p<0.05; ***p<0.01