Impacts of Income on Wind Turbine Locations

Rei Bertoldi

5/10/2020

Link : Project Git Repo

Background

According to the Office of Energy Efficiency and Renewable Energy of the United States, wind energy is one of the fastest-growing energy sources in the world. According to data from the U.S. Energy Information Administration (EIA), in 2016, five Heartland states - Iowa, South Dakota, Kansas, Oklahoma and North Dakota - sourced over 20% of their electricity generation from wind power. The data also showed that wind energy supplied over 5.5% of electricity nationally. In the wake of climate change and the struggling petroleum industry, currently taking major hits from the covid-19 outbreak and price wars with Saudi Arabia and Russia, the greater use of wind energy is particularly salient. Given the economics of energy demand, cost-effective renewables are going to be key in substituting oil demand. Land-based utility-scale wind is one of the lowest-priced energy sources today, costing between two and six cents per kilowatt-hour, and it happens to be clean and sustainable. Wind development, however is often met with pushback against communities not wishing to live near a turbine farm. Sound and visual impact of wind turbines are associated with nuisance factors that discourage individuals from willingly living near turbine farms. I am interested in the potential relationship between income and wind turbine locations. The goal of my research project is to analyze the effects of median income on the number of wind turbines on the state level in the greater United States, or on the county level within a particular state.

Analysis

The analysis of this project is twofold and therefore, the paper will be split into two parts. The initial results of this paper come from aggregated analysis at the state level within the United States. In this aggregated model, controlling for median income, land area, average wind speed, mean elevation, max elevation and minimum elevation, I found no significant spatial effects nor significant income effects on the number of wind turbines in a state. The secondary results of this paper come from a disaggregated analysis of a binary outcome of either having wind turbines or having no wind turbines at the county level in the state of Texas. The control variables are identical to the controls in the intial analysis. The results of this model are significant spatial autocorrelation and non-significant income effects.

State Level Analysis

Data

The most important data underlying my model comes from three distinct sources: US Census Bureau data, wind speed metadata and wind turbine shape data. The main data source of this project has been the recently released United States Wind Turbine Database which contains a publicly available, comprehensive data set of U.S. wind turbine locations and characteristics. I am using this data in collaboration with NREL metadata on average wind speed and a median income regressor variable from the US Census Bureau. I am also using elevation and land area data from the U.S. Geologrical Survey's National Geospatial Program and the U.S Census Bureau, respectively.

Variables

Outcome Variable

My outcome variable comes from the US Wind Turbine Database. It is a count variable of the number of wind turbines in each state. I had concerns about the efficiency of running a simple OLS linear model with a count outcome variable. To explore the nature of my outcome variable, I preformed some basic plotting to visualize the nature of the distribution of the variable. Fig 1 is a plot of my outcome variable by my regression variable of interest: median income. The distribution

of the outcome variable in this plot looks continuous enough for me to feel comfortable proceeding with my analysis using a basic OLS linear regression model.

It is also worth noting, that Texas appears to be a large outlier in the data. To account for the exceptionally high number of turbines located within the Texas, I included a control variable for the land area of each variable. Texas may have a larger number of turbines, but is has very large land area so I am hoping that controlling for land area I am able to account for the nature of this outlier in the data. To futher explore the effects Texas has as an outlier in the data on my model, I ran a secondary OLS regression model at the state level which utilizes an outcome variable in which the count of wind turbines have been divided by the land area of the state.

Regressor variables

My regressor variable of interest is a variable acquired from the US Census Bureau, containing information on median income amounts for each state. I am interested in this variable because the ultimate goal of my project is to develop an understanding of the relationship between income and the number of wind turbines within a given state or county, as a proxy for living near a wind turbine farm. The coefficient on this regressor will help me in my analysis in understanding whether there is a positive or negative relationship between income and wind turbine counts. I wanted to add a quadratic income term as a regressor in my model to allow for the slope of the coefficient to vary across income values. Judging by Fig 1, a very naïve inference from the plot would tell us that turbine count increases as median income increases to a point, then begins to decrease. However, I was unable to run spatial models using a median income interaction term because of multicollinearity issues.

The rest of my variables are regressor variables to help control for observable characteristics that we might expect to influence the presence of a wind turbine within a state. For instance, we might consider the characteristic of a good wind power site. The most obvious control variable to include in the model would be average wind speed. The amount of power generated by a wind turbine is proportional to the cube of the wind speed. This means that increasing the average wind speed from 6 m/s to 7 m/s yields a 60/ increase in power from the same turbine. This being the case, wind speed is going to be an important indicator of wind turbine locations. My wind speed variable

comes from NREL metadata on average wind speeds across the United States. It is important to note that this dataset was missing Hawaii and Alaska, so these states were dropped from my dataset and were therefore not included in my model.

Additionally, when thinking about the characteristics of a wind power site, we might consider the accessibility of the land. Turbines are very large and heavy, so there needs to be access roads and tracks to the site with no weak bridges, excessively tight corners or steep gradients. A common land characteristic that would make it difficult for the development of a wind turbine farm is the presence of mountains. To control for the reasonability of wind turbine development, I include average elevation as well as maximum elevation and minimum elevation. I believe including mean, max and min gives us robust information about the characteristics of a state's landscape. Elevation data was collected by the U.S. Geological Survey.

The turbine data set also includes generation capacity variables, which I did not include in the model because they lead to R-squared values of 0.99, as they perfectly predict the presence of each turbine.

Weights Matrices

I created two weights matrices to quantify the spatial relationships that may or may not exist in the data and formally incorporate spatial dependence into my model. I created both to visualize and conceptualize which foundational spatial structure would be best based on theoretical assumptions of what I would reasonably expect the nature of the spatial structure of the data to be. I used kNN and queen criterion. Figure 2 illustrated the weights matrix as defined using kNN, specifying four neighbors. I am weary of this specification because many states in middle America are only connected to states on their right and not on their left, as a result of calculating distances from state centroids. For example, Kansas is a neighbor of Missouri but not of Colorado even though they are neighbors. However, I do still believe this neighbor specification might be reasonable since states to the left of these 'middle' states that are only connected to states on their right border, are much more mountainous. In this way, these states being neighbors of states only to their right may actually be a reasonable specification.

Figure 3 illustrates the result of running a queen contiguity matrix. This specification seems fair

since it is connecting states by their borders. The issue when running a queen contiguity matrix on the United states stems from the variation in state sizes across the United States. States on the east coast tend to be much smaller than the rest of the states as you head west. Figure 4 is a connectivity histogram illustrating the variation in the number of neighbors. We see that there are more states with one neighbor than there are states with up to 8 neighbors. I do not think the variation here is unreasonable. States with one neighbor, such as Maine reasonably only have one neighbor. States with high numbers of neighbors, like Tennessee, which has 8 neighbors likely because it is a long and narrow state, are being connected by their borders. Given that a state is sharing a border with another state, it would be reasonable to believe that they share land characteristics. Since a large part of my model is including land characteristics, the queen contiguity matrix may also be reasonable. I tested for spatial dependence using both weight matrices.

Basic Linear Model

The basic linear regression model without accounting for any spatial dependence is as follows-

$$Y\text{turbine} = \beta_0 + \beta_1\text{inc} + \beta_2\text{land} + \beta_3\text{wind} + \beta_4\text{elev} + \beta_5\text{maxelev} + \beta_6\text{minelev} + \epsilon$$

Here, the outcome variable of interest is number of turbines (*turbine*). The independent control variables are land area in square miles (*land*), average state wind speed (*wind*), average state elevation (*elev*), maximum state elevation (*maxelev*), minimum state elevation (*minelev*). The regressor variables I am particularly interested in are (*inc*) and (inc^2), which are the median state income and quadratic median state income variables. The coefficient on the income variable will inform the direction of the relationship between income and number of turbines and the coefficient on the quadratic term will inform us of whether this relationship increases or decreases across income values.

Depending on the results of spatial dependence testing, this OLS model will be sufficient or we will need to run a spatial lag or spatial error model. Using LaGrange multiplier diagnostic testing for spatial dependence, if the p-value of `LMlag` is significant, I would want to run a spatial lag model. The reduced form equation of the spatial lag model would be - $(1 - \rho W)y = x\beta + e$. This would mean that the independent variables are explaining the variation in the dependent variable that is not explained by the neighbors' values. If, on the other hand, the p-value of `LMerror` is significant, I

would want to run a spatial error model with the reduced form equation - $y = x\beta + \epsilon$ with spatially correlated errors - $\epsilon = \lambda W\epsilon + u$. In this case, we would be correcting for the spatial autocorrelation in the error term, i.e. there is a missing variable that is spatially correlated. If both are significant, I would need to use robust testing.

Results

Turbine Count Outcome

Spatial Dependence Testing

The results of spatial dependence testing using the turbine count variable are illustrated Table 1.

Table 1: Spatial Dependence Testing Results

Weight	Test	p-value
Queen	Moran's I	0.66100
Queen	LMerr	0.26280
Queen	LMlag	0.72610
Queen	RLMerr	0.22800
Queen	RLMlag	0.57020
kNN	Moran's I	0.76820
kNN	LMerr	0.13560
kNN	LMlag	0.92500
kNN	RLMerr	0.06586
kNN	RLMlag	0.28050

We can see that we fail to reject the null hypothesis of spatial randomness for all the LaGrange multiplier tests using both the queen matrix and the kNN matrix. This means that there is no correlation of turbine counts conditional on my regressors. In other words, there is no spatial autocorrelation of the residuals. I also ran a Moran I test using both the queen matrix and the kNN matrix, which were also insignificant. This is telling us that there is an absence of any spatial

pattern in the data, so that there is spatial randomness in our data. If there was spatial dependence in our data, I would imagine it would likely be positive correlation associated with clustering of states with high counts of wind turbines. However, the results of our spatial dependence testing are telling us that there is in fact no significant spatial effects in our model at the state level or that the value at one state does not depend on the value of a neighboring state. Figure 5 is a choropleth map of turbine counts in the United States.

I believe that the resolution of our spatial units is driving the insignificance in our spatial effects. I would have expected to see positive correlation, or clustering of wind turbine farms around particular locations with choice environmental conditions. Aggregating up to the state level seems to be too coarse of a resolution to function as an effective spatial unit. I believe I am losing spatial dependence after aggregating up to such a large spatial unit. Regardless of the presence of significance, we are still learning that there is spatial randomness at the state level. This being the case, the basic OLS model without accounting for spatial effects is going to be sufficient. In the next section, I will interpret the results of this model.

Basic Linear Regression Results

Table 2 includes the regression results for the linear regression model I ran without any spatial autocorrelation modeling. The coefficients on land area and average wind speed are significant. This is not surprising because intuitively we would expect that the more land area a state has, the more space for wind turbines. However, it might also be import to consider the influence of outliers here. Given Texas' relatively high number of turbines and very large state land area, this coefficient might be mechanically significant. The significance on average wind speed is also significant, but this is to be expected since as discussed previously, wind speed is a strong determining characteristic of a good power site.

The coefficient on the median income variables is significant and positive. This is telling us that an increase in income is associated with an increase in turbines. However, the coefficient on the median income variable is quite small, `.05950` so the effect of income on the presence of wind turbines is significant but not large. I want to note that this model is being run without the use of a quadratic income variable so we are not seeing if the income trend on the number of turbines

changes as income increases. When running a OLS model using both a median income term as well as a quadratic interaction income term, the significance on median income goes away. In other words, the median income variable is no longer significant conditional on the inclusion of the squared variable. This means that when we are not controlling for the squared effect, we are over-estimating the linear effect of median income on the number of wind turbines in a state.

Turbine Count Over Land Area Outcome Variable

Spatial Dependence Testing

The results of the OLS model using the turbine count over land area variable are shown in Table 3 and the results of the spatial dependence testing are illustrated in Table 2.

Table 2: Spatial Dependence Testing Results

Weight	Test	p-value
Queen	Moran's I	0.020580
Queen	LMerr	0.252800
Queen	LMlag	0.338800
Queen	RLMerr	0.370000
Queen	RLMlag	0.521500
kNN	Moran's I	0.004068
kNN	LMerr	0.122300
kNN	LMlag	0.174000
kNN	RLMerr	0.380900
kNN	RLMlag	0.632200

We can see that, again, we fail to reject the null hypothesis of spatial randomness for all the LaGrange multiplier tests using both the queen matrix and the kNN matrix. However, now the Moran's I tests are significant, even more so when using the kNN weights. I think this is very interesting, and demands revisitation of my previous hypothesis that we are losing spatial dependence when aggregating up to the state level because the units are too large. Rather, when we are using this

ratio outcome variable, it seems as though we are better capturing smaller states with relatively high numbers of turbines. This is because our outcome variable is containing more information about the number of turbines in relation to amount of space that they could occupy within a given state. When we were looking simply at the number of turbines in a state, states with small numbers of turbines relative to larger states seemingly lose their informational power. Now, when we look at the choropleth map of this outcome variable rather than just turbine counts in the United States, we could be looking at a stronger positively correlated spatial process. See Fig 6.

However, I am reserved in making any claims of statistical significance in this model due to the fact that the regressors are doing an incredibly poor job of fitting the data. The results of the model can be see in Table 4. The R-squared for this model went down to `0.14` from `0.77`. For this reason, I will not be fully interpreting the results of this OLS model but I will note that the average wind speed variable is also significant in this model. The poor fit of this model, also lends itself to the notion that the land area of a state is a good indicator of the number of wind turbines in a state, since removing it as an indicator resulted in such a low R-squared.

County Level Analysis

Motivation

There are a couple of reasons why I wanted to extend my analysis beyond the state level. The two primary motivations for analyzing the data at a county level are the inefficiency of the model when treating a count variable outcome as a continuous variable outcome and the loss of spatial dependence at the aggregate level. When analyzing on the county level, there are a number of missing observations, i.e. missing states, as some states within the United State do not have any wind energy capacity. This motivates the use of a probit model with a binary outcome at the county level. Since there are many counties within states that do not have wind turbines, even more so than states without wind turbines, using a binary outcome here is going to be very helpful. Secondly, as I discussed in the previous section, at the aggregate level we do not observe spatial autocorrelation. I had hypothesized that this loss of spatial dependence was due to the large scale of our spatial units. Furthermore, in disaggregating down to the county level and utilizing a binary outcome, a probit model allows me to circumvent the potential bias of Texas as an outlier in the data in additional to

these primary limitations.

Scope

For the county level analysis, I chose to use Texas as my state of interest. I chose Texas because it has the largest number of wind turbines in the data set, which will give me the largest sample size and variance that I can achieve in the data.

Data

The data in my state level analysis primarily come from the same sources as the data used in the state level analysis. Similarly, important underlying data comes from the US Census Bureau, the United States Wind Turbine Database and the U.S. Geological Survey. Wind speed data used in the state level analysis was aggregated state level average wind speed, so complete county level wind speed data was not available. I am using rankings data from USA.com as my wind speed indicator variable.

Variables

Spatial Weights

Similar to the prior analysis, I generated two weights matrices for the county level spatial units. I created a queens matrix, which can be seen in Figure 8. This matrix is pretty interesting because of the structure of counties in Texas. If you look at the northern counties in the state, county lines are very foursquared. This generates an interesting queen matrix because the queen matrix will create neighbors of units that are diagonal and touching corners of adjacent counties. Because of this, we see that counties generally have a high number of neighbors, which we can see in the left skewed connectivity histogram for this weight matrix (Figure 9).

This is unlike the kNN weights matrix, specifying four neighbors, illustrated in Figure 10. In this case, specifying four neighbors ends up effectively generating a rook matrix. However, it seems that the location of a wind turbine would be indiscriminate about sharing a common boundary point rather than a segment, particularly since we have seen in our previous analysis that wind speed was a significant indicator of the number of wind turbines in a state. To argue in favor of using the kNN

matrix, we may consider the county sizes - if the counties are exceptionally large then we may not want to consider diagonal neighbors since wind speeds may be different if we are at great distances. However, this is not generally the case. Furthermore, we might expect using a rook matrix would be reasonable if the topography of the landscape was particularly mountainous since wind may be blocked on one side of the mountain. However, northern Texas is not particularly mountainous, so we would not expect wind to be getting caught between counties. This being the case, theoretically the queen matrix would be a better model of any potential spatial dependence in our data.

Basic Probit Model

The probit regression model without accounting for any spatial dependence is as follows-

$P(Y\text{turbine}|X_1, X_2, ..., X_6) = \phi(\beta_0 + \beta_1\text{inc} + \beta_2\text{land} + \beta_3\text{wind} + \beta_4\text{elev} + \beta_5\text{maxelev} + \beta_6\text{minelev})$ with $\phi(z) = P(Z \leq z), Z \sim N(0, 1)$.

Here, the right-hand side regressor variables are the same as the previous state level OLS regression model. The difference here is I am using a dichotomous outcome variable coded 1 if the county has at least one turbine and 0 if the county does not a turbine. Furthermore, I am using a probit link function, so I am assuming the error term is distributed normally. Using the binary outcome here is necessary and helpful because nearly half of the counties in Texas do not have wind turbines. So, in this sense, although I may be losing information by dichotomizing my outcome variable, the gain is more than sufficient to motivate the model.

Spatial Lag Probit Model

The practical method of testing for spatial dependence when using a probit regression model is different from the spatial autocorrelation testing I performed in the previous section of my paper when using a basic OLS model. Rather than running all of the LaGrange multiplier tests to determine which spatial model to run, I immediately ran a spatial lag model. When running the spatial lag model, the spatial lag, ρ, in the model is significant. This being the case, there is no need to run a spatial error model. I will continue in the following section and further discuss the results of the spatial lag probit model.

Results

The main results of the spatial lag probit model are twofold. The first major result of the model is the high significance of the spatial lag (See Table 7). The significance of the spatial lag model is telling us that the outcome variable in one unit is affected by the independent variables in the neighboring units. This is, perhaps, more in line with the results that I was expecting to see when I initially began the project. For instance, the regressor coefficient on elevation is significant in our model. If we think about the spatial distributions of elevation and its significant effects on the presence of a wind turbines in a county, the significance of the lag model makes sense. Elevation in one county is highly likely to be similar to elevation in its neighboring counties. This is Tobler's law - "everything is related to everything else, but near things are more related than distant things." And, again, we know that elevation significantly effects the probability of the presence of a wind turbine in any given county.

The second result of the model is the coefficient on median income. Unfortunately, we see here, again, a non-significant coefficient on the median income variable. This is telling us that there is no significant effect of median income on the probability of the presence of at least one wind turbine in a county in Texas. But, as I mentioned in the previous paragraph, although it does not necessarily support my underlying research question, we do see significant effects of elevation. We can derive the direct and total effects of this variable. The effect of this variable is actually quite small, with the direct effect of elevation being `4.700e-05` and a total effect of `1.384e-04` (See Table 6). I will say that the upper 95% percentile of the distribution does contain 0, so I am not fully convinced that this variable is necessarily significant enough to put too much weight on.

Conclusion

In conclusion, I found insignificant spatial and median income results at the state level of my analysis. This is pretty uninformative and inspired my decision to continue further analysis at the county level. At the county level, I found significance on the spatial lag which tells me that there is spatial autocorrelation at the levels of the dependent variable. I also found, again, insignificant median income effects at the county level. So, ultimately, I cannot make any claims of effects of median income on the number of wind turbines on the state level in the greater United States, or

on the county level within Texas.

Table 3:

	Dependent variable:
	paste0("t_count ~", reg_vars)
median_income	0.059^{***}
	(0.020)
square_miles_land_area	0.052^{***}
	(0.005)
ave_wind_speed	661.921^{**}
	(294.719)
mean_elev	-0.196
	(0.295)
max_elev	-0.106
	(0.087)
min_elev	-0.561
	(0.446)
Constant	$-9{,}168.038^{***}$
	(2,182.259)
Observations	48
R^2	0.773
Adjusted R^2	0.740
Residual Std. Error	1,174.368 (df = 41)
F Statistic	23.277^{***} (df = 6; 41)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Spatial Probit Lag Regression Summary Results

	Estimate	Std. Dev	p-level	t-value	Pr(>|z|)
(Intercept)	-1.6353724	0.9439813	0.035	-1.7324204	0.0844126
median_income	0.0000010	0.0000072	0.432	0.1395169	0.8891522
elevation	0.0001813	0.0000685	0.004	2.6452880	0.0086711
land_area	0.0000533	0.0001262	0.333	0.4227580	0.6728295

	Estimate	Std. Dev	p-level	t-value	Pr(>|z|)
ave_wind_speed	0.0669652	0.0460891	0.061	1.4529500	0.1474719
rho	0.6924870	0.0795186	0.000	8.7084872	0.0000000

Table 4:

	Dependent variable:
	paste0("count_land ~", reg_vars_norm)
median_income	0.00000
	(0.00000)
ave_wind_speed	0.011**
	(0.005)
mean_elev	0.00000
	(0.00000)
max_elev	−0.00000
	(0.00000)
min_elev	−0.00001
	(0.00001)
Constant	−0.070**
	(0.034)
Observations	48
R^2	0.139
Adjusted R^2	0.037
Residual Std. Error	0.018 (df = 42)
F Statistic	1.361 (df = 5; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5:

	Dependent variable:
	paste0("count_dummy ~", reg_vars)
median_income	−0.00000
	(0.00001)
elevation	0.0005***
	(0.0001)
land_area	−0.0001
	(0.0001)
ave_wind_speed	0.121**
	(0.052)
Constant	−2.779***
	(1.026)
Observations	254
Log Likelihood	−136.756
Akaike Inf. Crit.	283.511
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Spatial Probit Lag Regression Impact Results

	lower_005	posterior_mean	upper_095
median_income	-0.0000029	0.0000003	0.0000035
elevation	0.0000211	0.0000492	0.0000762
land_area	-0.0000430	0.0000150	0.0000717
ave_wind_speed	-0.0009733	0.0188501	0.0422278
	lower_005	posterior_mean	upper_095
median_income	-0.0000059	0.0000005	0.0000071
elevation	0.0000406	0.0000904	0.0001363
land_area	-0.0000808	0.0000300	0.0001477
ave_wind_speed	-0.0019742	0.0348882	0.0806480
	lower_005	posterior_mean	upper_095
median_income	-0.0000090	0.0000008	0.0000104
elevation	0.0000657	0.0001396	0.0001953
land_area	-0.0001249	0.0000450	0.0002132
ave_wind_speed	-0.0032225	0.0537383	0.1203979

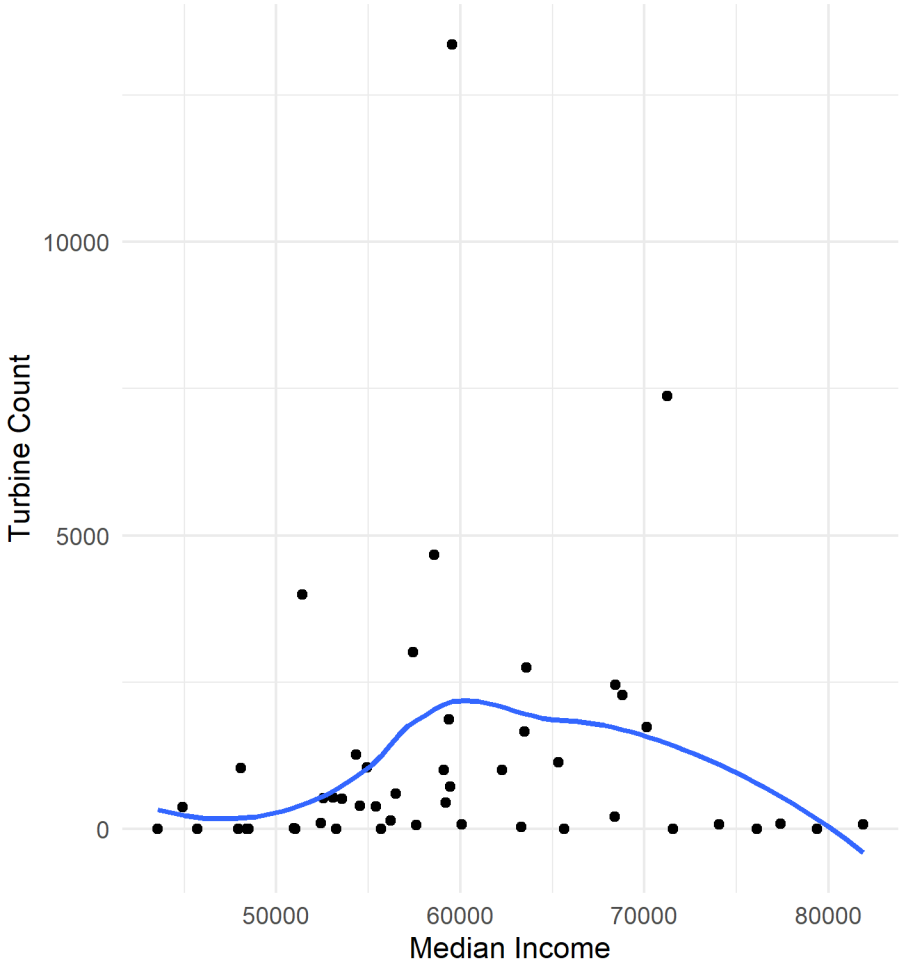

Figure 1: Medium income by turbine count

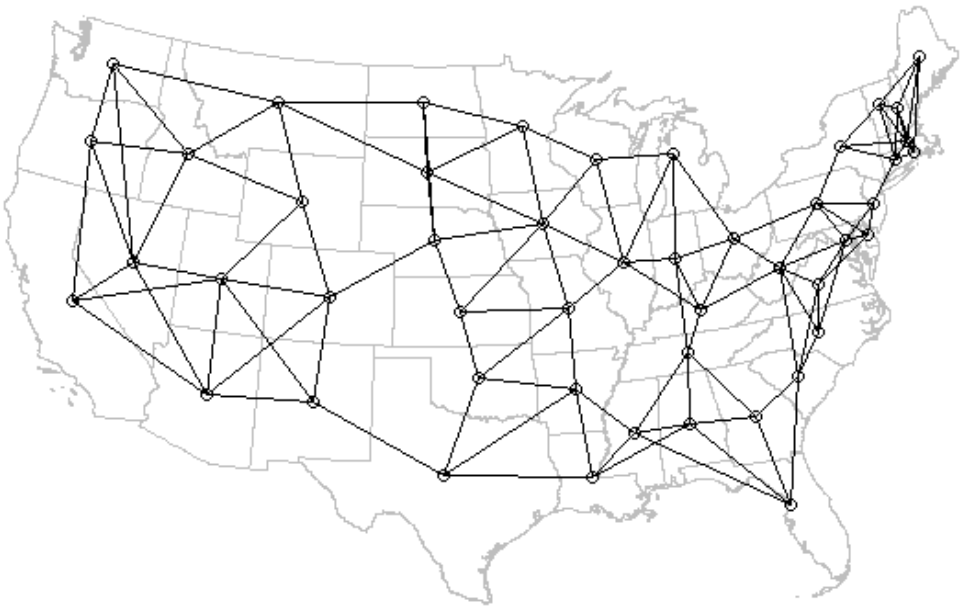

Figure 2: kNN 4 neighbors weight matrix

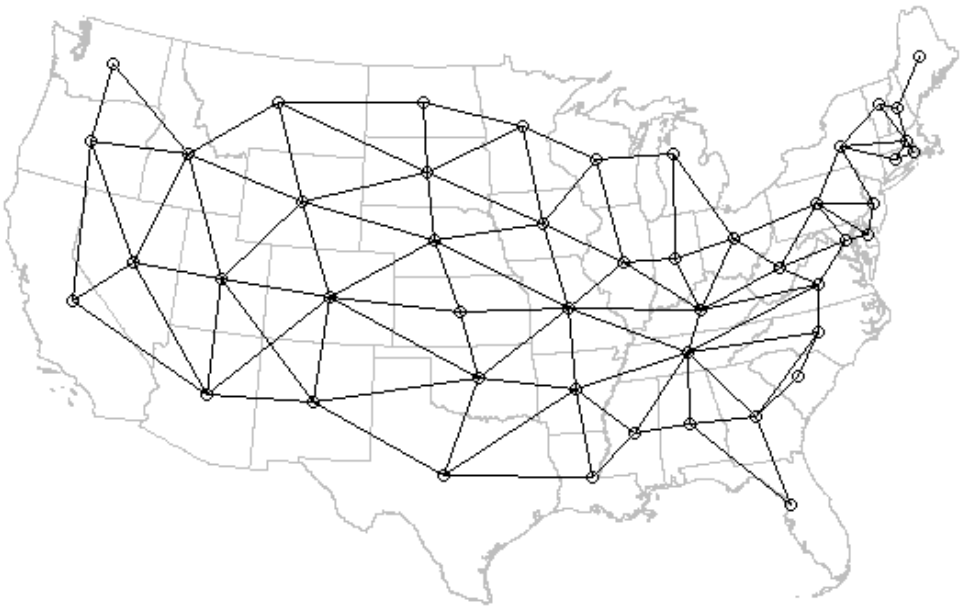

Figure 3: Queen neighbors weight matrix

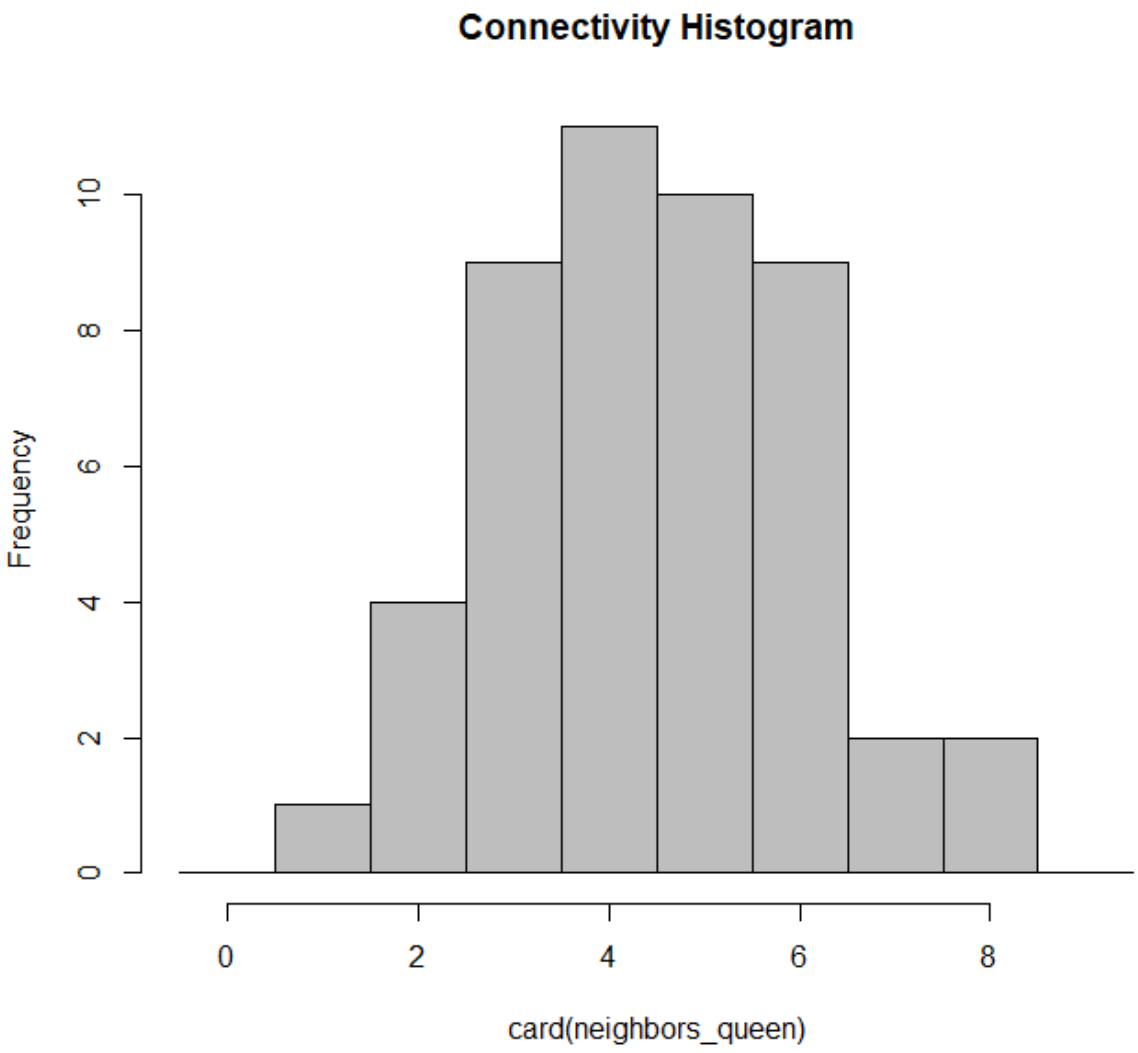

Figure 4: Connectivity Histogram of queen neighbors weight matrix

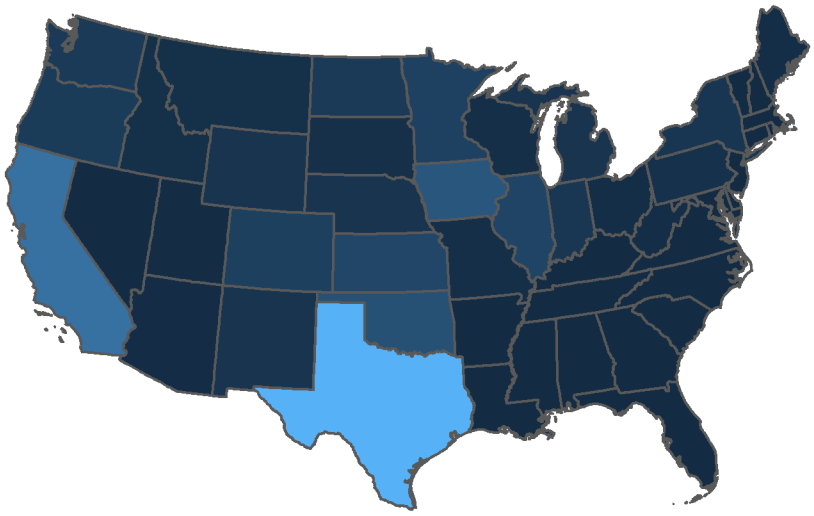

Figure 5: Choropleth map of turbine count in the United States

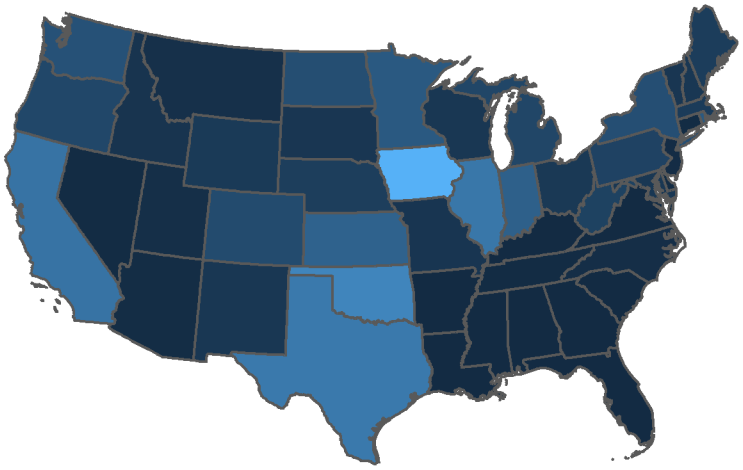

Figure 6: Choropleth map of turbine count divided by land area in the United States

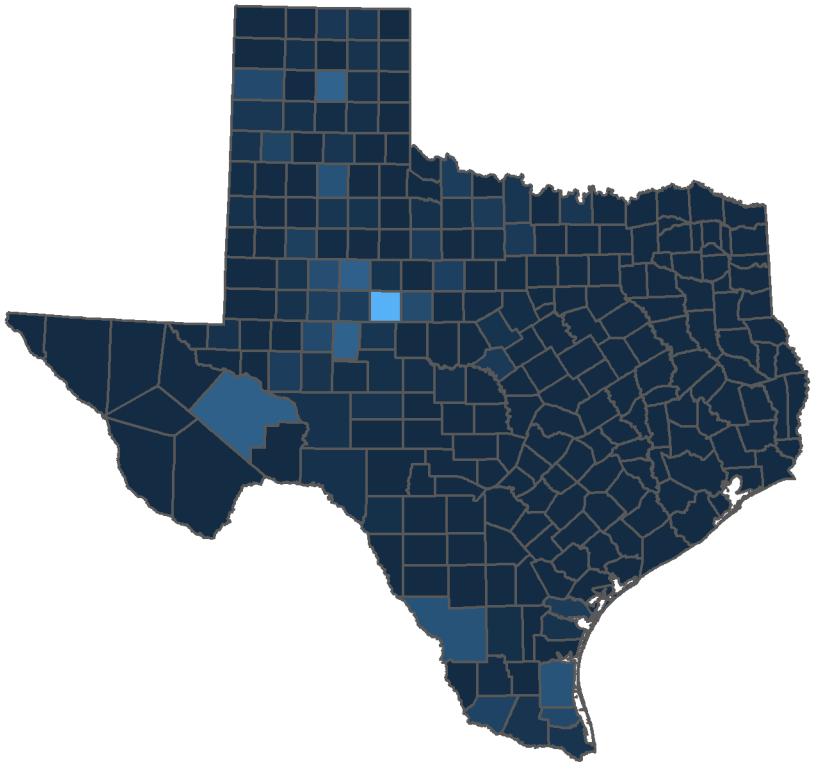

Figure 7: Choropleth map of turbine count in Texas

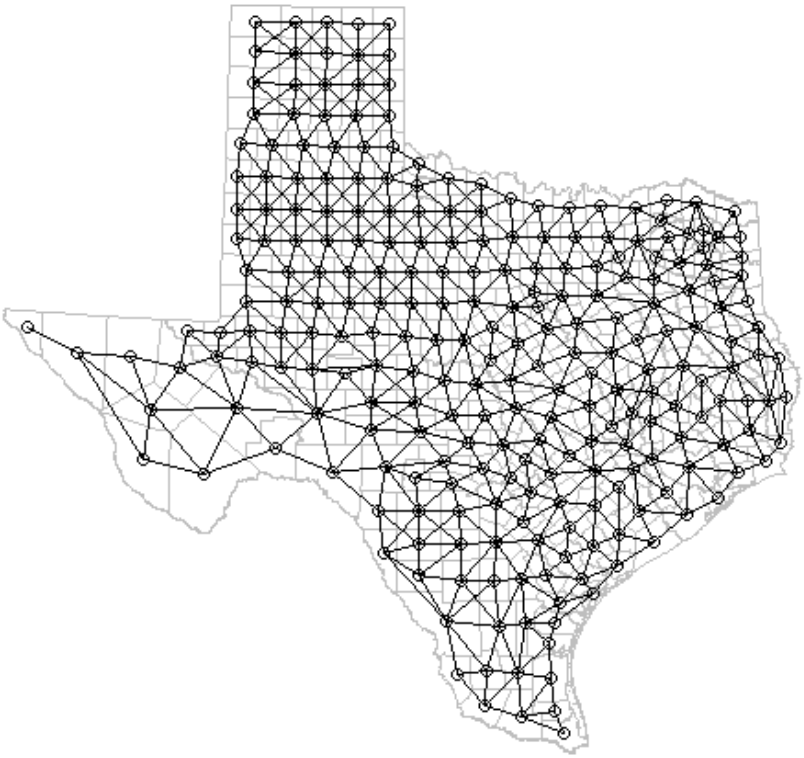

Figure 8: Texas county level queen neighbors weight matrix

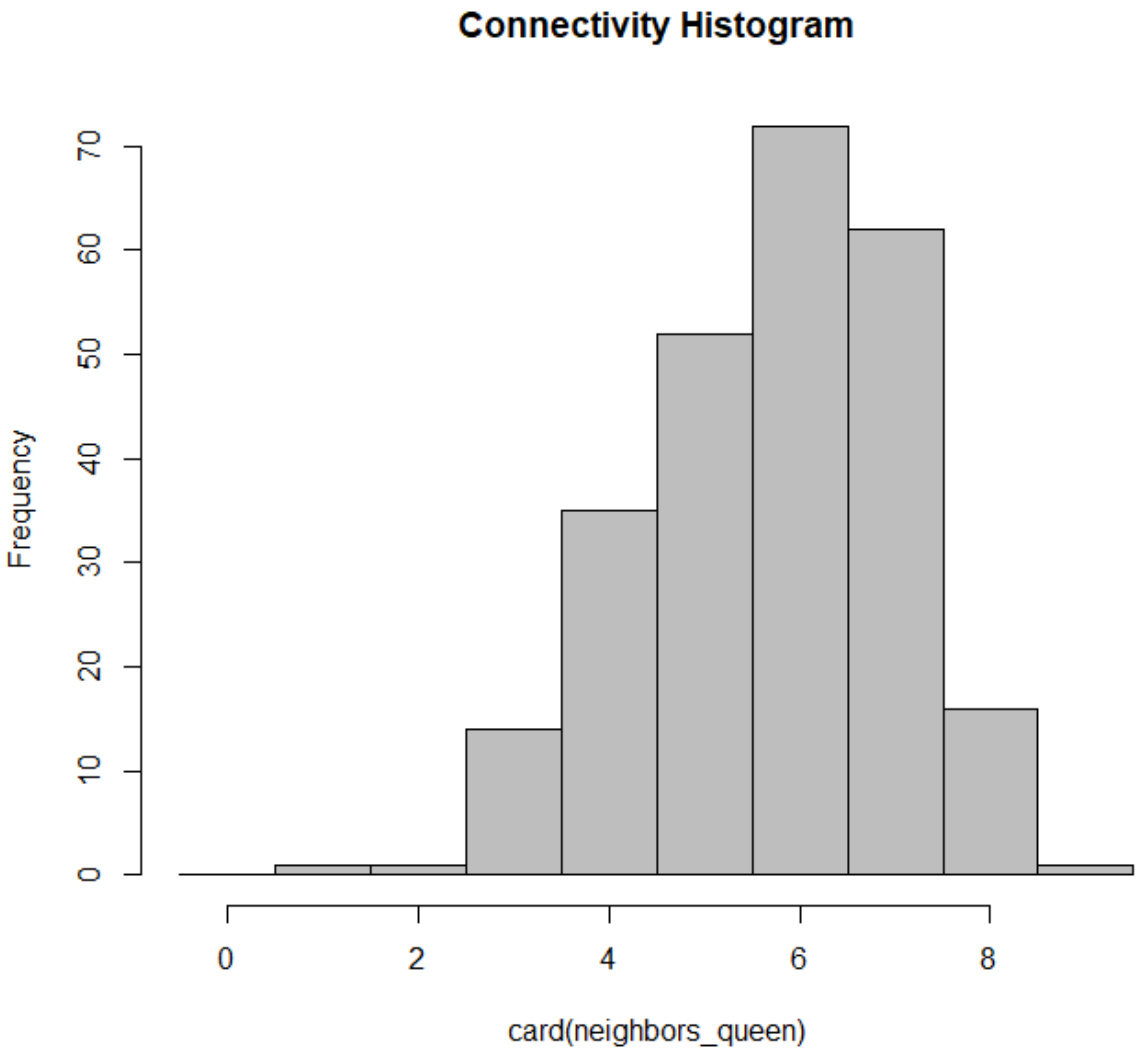

Figure 9: Texas county level connectivity histogram of queen neighbors weight matrix

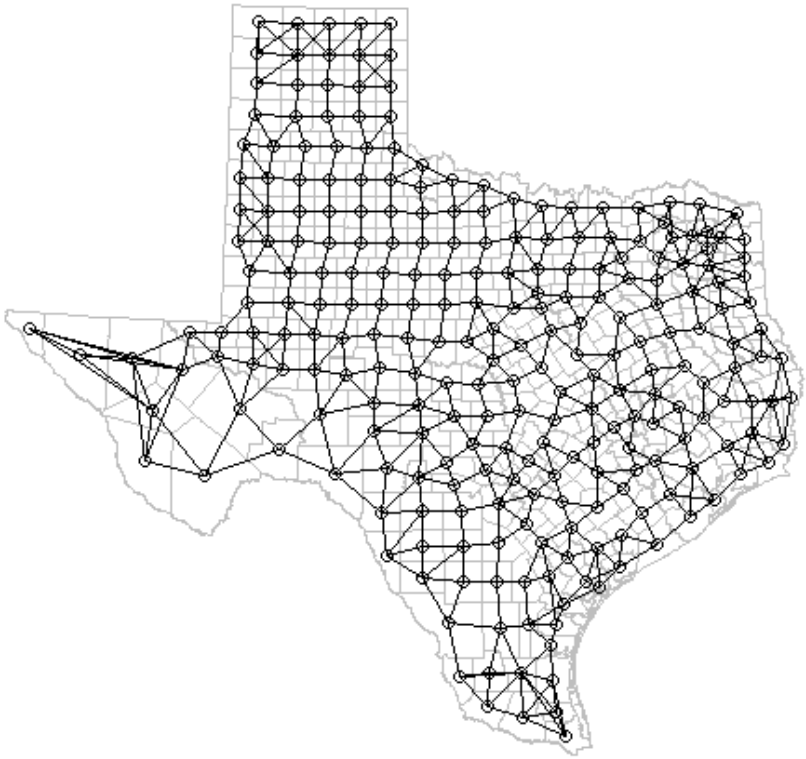

Figure 10: Texas county level kNN neighbors weight matrix