Analysis of U.S. Energy Consumption and Census Data

Aidan Alai, Caroline Bablin, John Habib, Oren Merberg

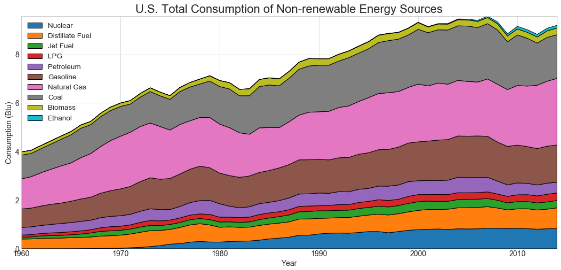
December 2019

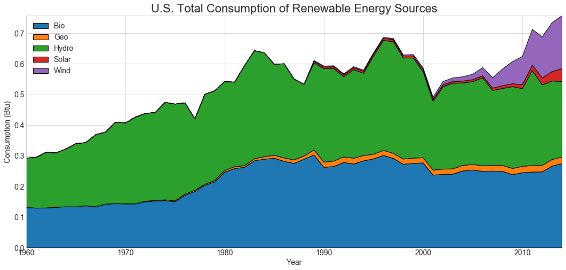
1 Introduction

In recent decades, the negative effects of climate change have become increasingly evident. Lessening harm to the environment is a prominent concern and necessitates special attention to improving energy efficiency. Energy efficiency is roughly described as consuming less energy while providing the same amount. This will be more formally defined in the following section. This is an especially important metric as the demand for energy is rapidly increasing. In this analysis, we examine energy efficiency in the United States with data on the consumption of various energy resources from 1960 to 2014. We first perform a dimension decomposition with principal and independent component analysis to give better insight into the data. We then aim to differentiate consumption patterns between states with regression analysis and we propose a classification of the states based on various factors using K-Means clustering.

2 Data

The data for this analysis is provided by the U.S. Energy Information Administration's State Energy Data System (SEDS) comprising of more than 200 energy features for each state for 55 years (1960 - 2014). There are over 600,000 values, so we select features that are significant for our analysis. The following energy consumption values are the columns in our data set: `Distillate Fuel`, `Jet Fuel`, `Liquid Propane Gas`, `Motor Gas`, `Nuclear`, `Coal`, `Natural Gas`, `Petroleum`, `Biomass`, `Ethanol`, `Geothermal`, `Hydropower`, `Solar`, `Wind`, `Electric`. To calculate a measurement for efficiency (described below), we aggregate values for gross domestic product (gdp) and population. This data is provided by the U.S. Bureau of Economic Analysis and the U.S. Census Bureau. An overview of our data is represented below, categorized by renewable and non-renewable sources.





3 Analysis

3.1 Normalization

Normalization is a preprocessing technique used on samples that may have a variety of parameter magnitudes. In the case of energy consumption, each resource type value was calculated as a part of the overall consumption. This brought the range of the data from 0 to 1xe6 into a space that exists in the range from 0 to 1. Now, each state carries the same weight and can be compared equally. The following equation was used in the normalization of all energy consumption resource types.

$$E_{norm} = \frac{E_{Btu}}{E_{total}}$$

3.2 Initial Dimension Reduction

After normalization is complete, further preprocessing is needed in order to take some assumptions into account. It was assumed that energy efficiency, the economy and whether the energy type was renewable or not had some correlation. Dimension reduction was a useful technique in refining the scope of the data and summarizing many parameters into a few. The following equations outline the process of dimension reduction to refine all parameters into four areas of interest.

$$E_{\text{Renewable}} = \frac{E_{\text{TotalRenewable}}}{E_{\text{MaxRenewable}}}$$

$$E_{\text{NonRenewable}} = -\frac{E_{\text{TotalNonRenewable}}}{E_{\text{MaxNonRenewable}}}$$

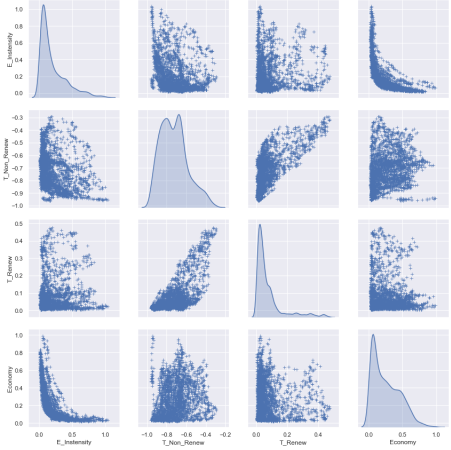
$$Intensity = -\frac{E_{\text{total}} - (\text{export} * E_{total})}{\text{gdp}/\text{Intensity}_{max}}$$

$$Economy = \frac{E_{\text{TotalRenewable}}}{E_{\text{MaxRenewable}}} / Economy_{max}$$

where,

$E_{Renewable} = E_{Hydro} + E_{Wind} + E_{Geo} + E_{Solar} + E_{Bio}$
$E_{NonRenewable} = E_{Coal} + E_{FuelOil} + E_{Ethanol} + E_{JetFuel} + E_{Gas} + E_{NatGas} + E_{Petrol} + E_{Gas}$

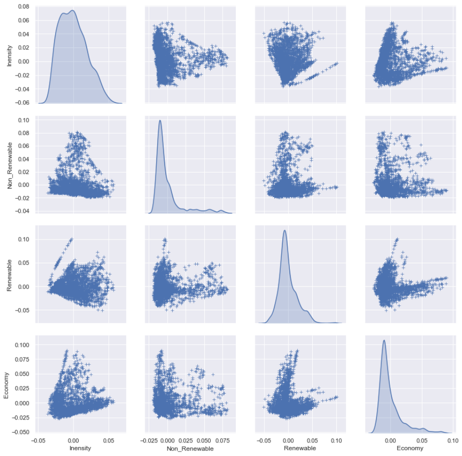
As seen in the Non-Renewable equation there was a sign change to the output values. The sign change was intended to differentiate values that were assumed to be beneficial to a state versus ones that could be assumed to have a negative impact. More non-renewables were determined to be a factor that acted in opposition to renewable energies and therefore the sign change was used to represent this. With the entirety of the data now reduced to four dimensions, it is possible to visualize the data in a meaningful way. The following plot shows how the value of each reduced dimension interact with one another. Each parameter was normalized one more time to bring the range of the vector space between 0 and 1. Additionally, the Intensity value denotes cost efficiency of the entire state. The Economy represents how the population may affect the GDP.



Note each plot's axial reflection is a transposed version of itself. From these plots it is observable the initial dimension reduction does provide some insight into how the two types of energy resources has an impact on the Intensity, yet clear correlation is difficult to define. The outlier values provide a lot of noise to the data which may be hiding grouping or trend. Removing this noise would assist in analysis. Further dimension reduction with ICA was decided to be the solution for optimizing refinement.

3.3 ICA- Independent Component Analysis

ICA was determined to be a good method for understanding the dataset further. It could provide a form of dimension reduction that would output matrix of weighted values for each parameter as they relate to one another per sample. The weighted matrix could then be used as a tool to determine the space where the input conditions are most likely to exist. Whitening is a first step of ICA that assisted in the reduction of noise of the data. It is a process that 'spheres' and centers the data. To 'sphere' the data is do alter the weight of an entire dimensions to make any Gaussian subspace less gaussian. Gaussian behavior can be observed when on dimension carries greater weight than the other, causing for points to cluster near the axis. Gaussian behavior can be observed in a couple of the previous section's plots. In particular, the E-Intensity/Economy plot displays a Gaussian relationship. Whitening will shift the weight of each dimension to spread or 'sphere' the data in attempt to pull values away from the axis. Centering is included within whitening to shift the origin to a true center of the data. The data must be non-Gaussian to properly run ICA, so whitening was a critical step for the dataset. By using the FastICA Python library, it was possible to implement ICA on the initial four reduced dimensions with whitening. The plot below represents weighted matrix that results from the ICA method. It is a matrix that displays the weighted relationship between each parameter.



The ICA plot provides much more clarity in comparison to the initial dimension reduction alone. It is not observable to see a much more refined space in which the parameters operate within. For example, the Intensity dimension has some distinct relationships with the energy type parameters. Renewable carry more weight when the Intensity has more impact. The two parameters can be corelated. Non-renewables exist in all ranges of intensity importance. This displays the dominance of non-renewable energies. The other plots are good representations of how the parameters might behave. The Non-Renewable parameter is likely to have very little impact on the way renewables and the economy behave. This can be seen from the high concentration of samples that landed close to the origin. This again shows the dominance of non-renewable energies and how small the impact of other facts play into their importance.

3.3.1 Conclusions from ICA

Dimension reduction with the assistance of Independent Component Analysis have shed away layers of noise and confusion. It has brought the dimension space from 16 dimensions to 4. It has also divided all energy types into categories of interest: renewable, non-renewable, intensity and economy. The categories of interest led to the creation of reduced dimensions representative of each parameter within those categories. It was then discovered manual dimension reduction is not enough and further efforts were needed to reform the data into a way that offers insight.

ICA became that solution. It whitened the data to reshape it and recenter for optimal ICA performance. After the analysis was through, the results proved to offer insight into how energy types could have an effect on each other and how that played a roll in the economy or cost effectiveness of state. The introduction of more types of data would be useful in completing this analysis. Understanding more about each state, their geography, politics and history could give more distinctive shape to the data. More resolution into how each state operates would be necessary to draw further conclusions on how a state consume energy most efficiently.

4 Clustering by Region

In this section we examine energy intensity by U.S. regions. The U.S. Census Bureau divides the states into four regions:

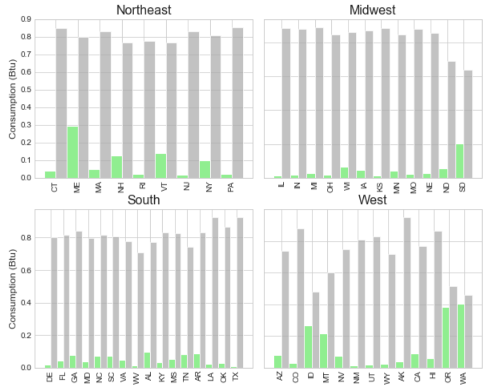
1. **Northeast**: Maine, Vermont, Massachusetts, New Hampshire, Rhode Island, New York, Pennsylvania, Connecticut, New Jersey, Delaware, Maryland, West Virginia

2. **South**: Virginia, North Carolina, South Carolina, Georgia, Florida, Tennessee, Alabama, Mississippi, Louisiana, Oklahoma, Texas

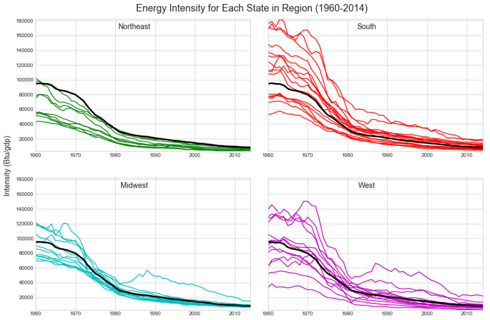
3. **Midwest**: Ohio, Michigan, Indiana, Kentucky, Wisconsin, Illinois, Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, Kansas

4. **West**: Washington Oregon, Nevada, California, Alaska, Hawaii, Montana, Wyoming, Idaho, Colorado, Utah, New Mexico, Arizona

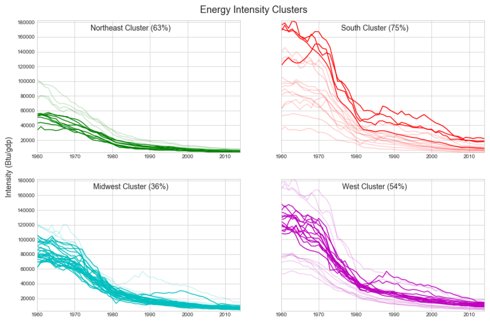
It is likely that the South and West have similar energy profiles and the Northeast and Midwest have similar profiles due to their environments. It could be insightful to examine this further to see if particular states stray from the norm of their region. To gain a simple understanding of energy consumption per region, a bar plot is created.



As expected, non-renewable sources are the overwelming majority for each region, however we expected the South to be the most prominent. There is also some interesting variability with renewables and non-renewables in the West, which can likely be attributed to variance in population. We turn our focus to energy intensity. To understand how each region's total intensity varies over time, a line plot is created (below) that shows each state's intensity for the four regions over 55 years. Each plot also has the United State's average intensity for reference.



Clearly, the U.S. has come a long way since 1960 as most regions seem to approach zero. Notably, the Northeast outperforms other regions with every state at or below the national average. Moreover, the South and West seem to have quite similar changes in intensity and are both much more varied than the Northeast and Midwest, so clustering with KMeans may tend to group in this fashion. To begin, we try running KMeans with four cluster centers. Below is a similar graph with the KMeans clusters highlighted.



Because KMeans is an unsupervised learning algorithm, accuracy does not make much sense. To get around this, a custom accuracy function is created which computes (for each region) how many states were in each cluster. The results are described as follows:

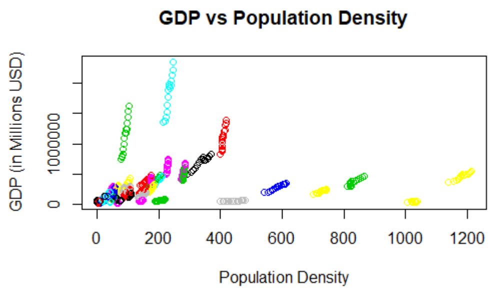
Northeast: [0. 4. 0. 5.]
Midwest: [5. 7. 0. 0.]
South: [6. 5. 4. 1.]
West: [5. 4. 2. 2.]

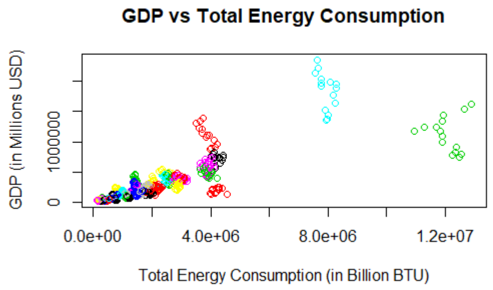
So the first cluster had 0 states in the Northeast, the third cluster had 0 states in the Northeast and Midwest, etc. We can see that the South and West indeed were poorly clustered with at least one state in each cluster. The second cluster has at least four states from each region.
Overall, KMeans did not do an excellent job at splitting up the data into regions. Intensity is probably not the best metric to differentiate between regions as there are no major differences. Chooseing different features that a vary per state would likely have more success

5 Regression

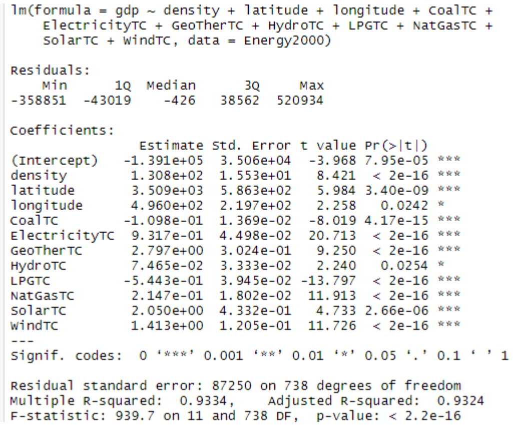
The goal in applying linear regression to the data was to see if there was a linear relationship between GDP, the response variable, and consumption of various energy sources, population density, and geographic location, the explanatory variables. After inspecting the data in its entirety, that is from the year 1960 to 2014, it was decided for this application to focus merely on the data since the year 2000. This would help simplify the model, as over the course of 55 years it was clear there was much variation between the response variables and the explanatory variables.

Before creating any models, it was necessary to inspect the variables and how they interact with the response variable. Scatter plots were created with GDP versus Population Density and GDP versus Total Energy Consumption, with both graphs colored by state.





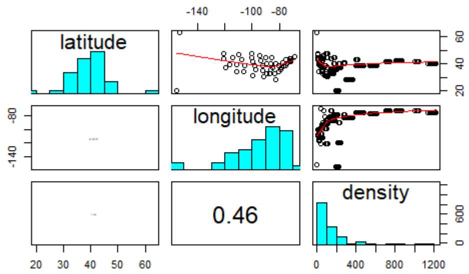
Analyzing these graphs, it is clear that there is a linear relationship between GDP and Population Density. However, there appears to be a significant difference between states, as seen by the clustering of points. The same appears to be true for the relationship between GDP and Total Energy Consumed. Larger, more densely populated states, such as Texas, California, and Florida, tend to follow a different trend and be further from the general trend for other states. From this, it was deduced that perhaps the large variation between states and their population densities and energy consumption, that a linear model may not be a good fit on a nationwide scale, but could perform better on a statewide scale. Nevertheless, this was disregarded to see if a model using all states would be a good fit. A linear model was created (shown below) using the several energy sources, population density, latitude and longitude.



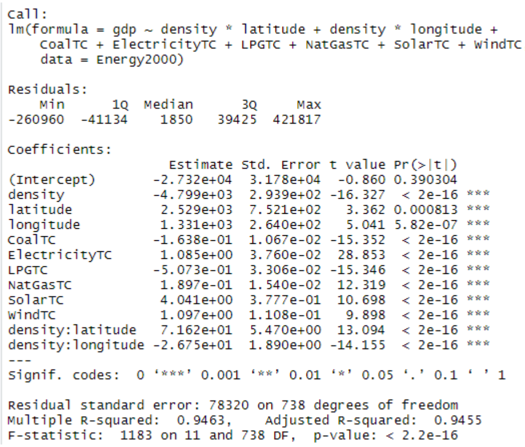
```
lm(formula = gdp ~ density + latitude + longitude + CoalTC +
    ElectricityTC + GeoTherTC + HydroTC + LPGTC + NatGasTC +
    SolarTC + WindTC, data = Energy2000)

Residuals:
    Min      1Q  Median      3Q     Max
-358851  -43019    -426   38562  520934

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.391e+05  3.506e+04  -3.968 7.95e-05 ***
density        1.308e+02  1.553e+01   8.421  < 2e-16 ***
latitude       3.509e+03  5.863e+02   5.984 3.40e-09 ***
longitude      4.960e+02  2.197e+02   2.258   0.0242 *
CoalTC        -1.098e-01  1.369e-02  -8.019 4.17e-15 ***
ElectricityTC  9.317e-01  4.498e-02  20.713  < 2e-16 ***
GeoTherTC      2.797e+00  3.024e-01   9.250  < 2e-16 ***
HydroTC        7.465e-02  3.333e-02   2.240   0.0254 *
LPGTC         -5.443e-01  3.945e-02 -13.797  < 2e-16 ***
NatGasTC       2.147e-01  1.802e-02  11.913  < 2e-16 ***
SolarTC        2.050e+00  4.332e-01   4.733 2.66e-06 ***
WindTC         1.413e+00  1.205e-01  11.726  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87250 on 738 degrees of freedom
Multiple R-squared:  0.9334,    Adjusted R-squared:  0.9324
F-statistic: 939.7 on 11 and 738 DF,  p-value: < 2.2e-16
```

Inspecting this model, it is evident that all variables are significant in the model due to their small p-values. The adjusted R^2 value is 0.9324, which is also a sign of a good model. However, this model has 11 variables. It is preferred for models to have fewer variables and be less complex. Analyzing the estimates for the coefficients of the variables, it is seen that the coefficient for some of the variables are one or two orders of magnitude of ten lower. To create a new, perhaps more appropriate model, variables whose coefficients were much smaller were removed. In essence, the energy sources that were most widely consumed were used, while other the others were dropped from the model.

In the first model, the variables of population density, latitude, and longitude were all included separately. However, it was hypothesized that these variables may interact with each other. It was proposed that population density would increase as latitude decreased and population density would increase in longitudes closer to either coast of the country. To test this idea before putting it into the model, a correlation plot was made, as seen below.



The histograms along the diagonal of this graph show the distribution of that variable. The scatter plots in the upper right scatter the two respective variables. The lower left corner shows the absolute value of the correlation coefficient between the two respective variables. The size of the number corresponds to how close the correlation is to 1. Clearly, the correlation between the variables latitude and longitude and latitude and population density are so small they are close to zero. This signifies that the variables of population density and latitude are not well correlated. The same goes for latitude and longitude, but the combination between those variables was not considered. Despite having such a low correlation coefficient, the interaction term between population density and latitude were included in the new model along with population density and longitude. The new proposed model is shown in Figure.

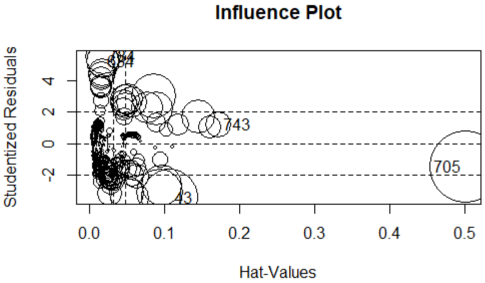


```
Call:
lm(formula = gdp ~ density * latitude + density * longitude +
    CoalTC + ElectricityTC + LPGTC + NatGasTC + SolarTC + WindTC,
    data = Energy2000)

Residuals:
    Min      1Q  Median      3Q     Max
-260960  -41134    1850   39425  421817

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.732e+04  3.178e+04  -0.860 0.390304
density           -4.799e+03  2.939e+02 -16.327  < 2e-16 ***
latitude           2.529e+03  7.521e+02   3.362 0.000813 ***
longitude          1.331e+03  2.640e+02   5.041 5.82e-07 ***
CoalTC            -1.638e-01  1.067e-02 -15.352  < 2e-16 ***
ElectricityTC      1.085e+00  3.760e-02  28.853  < 2e-16 ***
LPGTC             -5.073e-01  3.306e-02 -15.346  < 2e-16 ***
NatGasTC           1.897e-01  1.540e-02  12.319  < 2e-16 ***
SolarTC            4.041e+00  3.777e-01  10.698  < 2e-16 ***
WindTC             1.097e+00  1.108e-01   9.898  < 2e-16 ***
density:latitude   7.162e+01  5.470e+00  13.094  < 2e-16 ***
density:longitude -2.675e+01  1.890e+00 -14.155  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78320 on 738 degrees of freedom
Multiple R-squared:  0.9463,    Adjusted R-squared:  0.9455
F-statistic:  1183 on 11 and 738 DF,  p-value: < 2.2e-16
```

Comparing this model to the previous model, the variables still all have small p-values, so they are significant in the model. It is clear that the adjusted R^2 has increased to 0.9455 and this increase tells us this model is a better fit. Comparing the estimates for the standard deviation of both models, the new estimate has decreased from 87250 to 78320. A smaller standard deviation again is a sign of a better fit as it signifies that the difference between predicted values of GDP and the actual values are smaller. Despite these benefits, it is important to note that this model still has many variables and can be quite complex.

To see if there was a better model, backwards selection was performed on the most recent model. This removes one variable at a time and computes the resulting sum of square errors, residual sum of squares, and AIC, a penalty function for the residuals. Ideally, each of these values should be small and close to zero, and the smaller they are the better the performance of the model. The results of backwards selection show that removing any one of the variables increases each of these three values, signifying that no variables should be removed from the model at all.

Next, the data from this model was inspected to look for any outliers. An influence plot was created to help identify these points, as seen below.



This plot helps to find three types of outliers; a traditional outlier; one that does not follow the normal trend and has a large residual, a high leverage point; one that has extreme value for one or more of the explanatory variables, and a influential point; one that is both an outlier and a high leverage point. In the plot, the size of each circle corresponds to Cook's distance where the larger the circle the larger the value of Cook's distance for that observation. This distance is a measure of the influence of a point, or how much that point will change the estimates for the coefficients for each variable. It is clear there are outliers, such as observation 43, which may also be influential point due to its high value for Cook's distance. Perhaps the most notable point is observation 705, which has a large Cook's distance. These observations were found to belong to large states, such as Texas or California, which was not surprising due to the first inspection of the data where these states visibly followed different trends.

From this analysis, it was hypothesized that perhaps better models would be created from considering this data at a state wide level as opposed to nationwide. States like Texas often produced different results and could be affecting the quality of the model. In addition to this, future models could insect the variables further to add or take away variables. Looking at the correlation between more variables could be insightful, so as to predict which variables could be included as another interaction term. Overall, however, these models were a good starting point for examining the relationship between GDP and sources of energy, population density, and location.