Short Paper: Understanding Residential Energy Usage Patterns and Future Solar PV

Rick Grahn
Carnegie Mellon University
rgrahn@andrew.cmu.edu

ABSTRACT

Demand charges have long been implemented in the utility industry for large scale commercial and industrial users as a cost for the utility to reserve and maintain the infrastructure to supply peak power to users during the billing period. This charge has not yet been implemented at the residential level due to the similarity and predictability or residential users. However, with all the energy saving appliances and technologies and the widespread adoption of solar photovoltaic systems, residential loading has become less predictable and less equitable. With demand charges for residential users becoming a hot topic of discussion, an understanding of residential energy demand (total and separated) and solar pv potential will be integral for the customer and utility. Customers will be informed about ways to shift loads to avoid peak loading charges. The utilities will benefit by understanding solar pv potential and appropriately plan for these grid variations. This will also enable utilities to reduce grid uncertainty and appropriately prepare for a less stable grid by charging residential end users.

CCS Concepts

• **Information systems**→Data Analytics, Clustering, Regression Trees.

Keywords

Electricity; Solar PV; Cluster Analysis; Regression

1. INTRODUCTION

Since the industrial revolution, the United States electricity grid has operated using steady, reliable sources such as fossil fuels and hydroelectric. Due to the negative health and climate effects produced by the burning of fossil fuels, the United States and the world have begun a transition to more renewable energy sources. The widespread adoption of these new sources, namely wind and solar, have posed a problem for utilities because of their intermittency and lack of reliability. Current policy allows for the private owners of renewable generation to sell excess power to the grid which creates problems for the utility due to their unpredictability. As the grid moves to a more dynamic state with many individual "generators" with residential sized solar pv systems, utilities will need to better understand and predict grid conditions to ensure system reliability.

In addition to understanding the impacts of mass residential solar adoption, utilities much also understand and update their dated residential pricing scheme to ensure equitability across many users. As new energy saving technologies emerge, such as smart thermostats and energy efficient appliances, residential customers can no longer be placed in one category and assumed to all behave in the same way. A prime example being a residential customer with a large solar pv system that uses a fair amount of energy. During the sunlight hours, the customer is selling all excess solar energy to the grid offsetting the customer's high peak loads during the evening. Current pricing would just charge per kW*hr, which in this case might be zero, even though the customer is contributing to a highly dynamic, unstable grid.

This paper explores solar potential across regions of the United States and well as energy demand to better understand and predict future conditions of the grid integrated with renewable energy. A study of energy usage by component across regions will be presented for customers to understand components contributing to peak loading to appropriately plan and reduce peak loading charges. A clustering approach to detect houses using electricity for water heating is presented as well as a regression model to predict solar potential for different regions based on the input variables "hour of day", "month of year", "latitude", and "longitude".

2. DATA

The purpose of this paper is to look at residential energy demand datasets as well as solar irradiance data to better understand residential energy usage and influence of solar power for the user and utility. This data will help inform utilities about regions of high solar potential and their constituent's energy demand trends. The solar irradiation data used for the project was obtained from the National Solar Radiation Database maintained by the National Renewable Energy Laboratory [1]. The residential energy usage data was obtained from the US Department of Energy open data catalog [2]. Several locations were selected with a large regional spread to explore regional variability. The locations were Twenty Nine Palms, CA, Yuma, AZ, Abilene, TX, Key West, FL, New York, NY, and Pittsburgh, PA. For each of these locations, total meteorological year (TMY) solar irradiation data was used to determine average irradiation throughout the year. Since TMY data isn't for a specific year, a direct comparison to the energy usage data cannot be made. However, this data allows us to predict annual solar potential for different regions seasonally and by hour of day. The residential house datasets used were all considered baseload cases with square footage being approximately 2000ft^2. All houses were one story in an urban setting with similar building characteristics. With many factors being held approximately constant, a regional comparisons can be explored.

3. APPROACH

First, the paper presents ways to visualize the data through exploratory data analysis to look at regional variations in solar potential, energy usage, component energy usage, and seasonal trends. A basic understanding of these trends and regional solar potential by season will enable customers and utilities to make more informed decisions regarding energy usage and generation. Once certain trends are observed, clustering techniques will be utilized to try and predict presence of components for specific customers. In this paper, the presence of electric water heaters is explored. If customers can be easily separated into groups that use electric water heating vs. gas water heating, utilities can target these customers as potential energy dumps or thermal storage when high renewable penetration is present. The KMeans clustering technique was used (using the Python package sklearn) to try and group the houses with electric vs. gas heating. The number of clusters was chosen to be two. The algorithm then adjusts these cluster centers until the intra-cluster variation is minimized using the sum of the squares.

Since solar potential varies seasonally and regionally due to different climates, weather systems, and earth tilt, it is important for utilities to understand these variations and try to predict cumulative solar penetration in their operation area for future years. A regression tree (using the Python package sklearn) was fit with the input variables of month, hour of day, latitude, and longitude for five of the six locations (Texas, Florida, California, Pennsylvania, and New York) and the sixth and final location (Arizona) was used to test the prediction capabilities of the model. When the regression tree is fit with training data, optimal splits (nodes) are determined by minimizing the sum of the squares (MSE) for both regions on each side of the node. Parameters can be set on the regression tree to avoid overfitting the tree to the training data set. Parameters such as the max depth of the tree, the minimum samples needed at a node to allow for another split, the maximum number of leaf nodes, etc. The python package GridSearchCV was used with a 10-fold cross validation to output optimal parameters for the regression tree to minimize the MSE. The cross validation technique randomly grabs 90% of the data from the data set to train the model and uses the remaining 10% of the dataset to test the model. A MSE value is output and this process is done 10 times (i.e. 10-fold). When the sum of the MSE values are minimized, then the given model parameters are optimized.

4. RESULTS

4.1 Exploratory Data Analysis

The regional variation of solar irradiation and energy demand is show in figures 1-2.

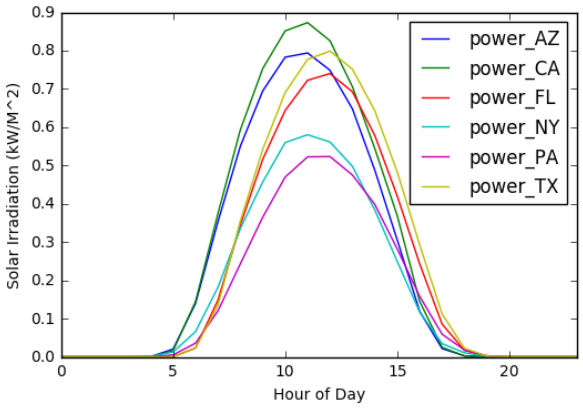


Figure 1 – Solar irradiation potential by region

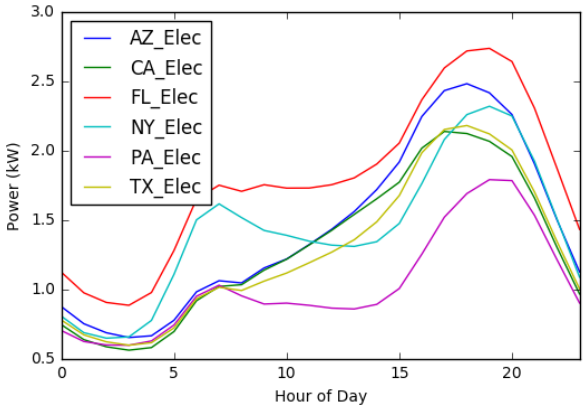


Figure 2 – Average daily demand by region

It can be observed that the solar potential is greatest in the warm, sunny climates in the southwest. Energy usage patterns are similar across regions except Florida and New York show a high morning peak. This is due to electric water heating present in these two locations as opposed to gas water heating.

In the next figure the net difference between cumulative annual energy demand and solar generation is presented assuming perfect energy storage. It can be observed that the high solar potentials of the southwest can offset the increased cooling loads in the region. However, the solar energy storage problem is a large obstacle to overcome.

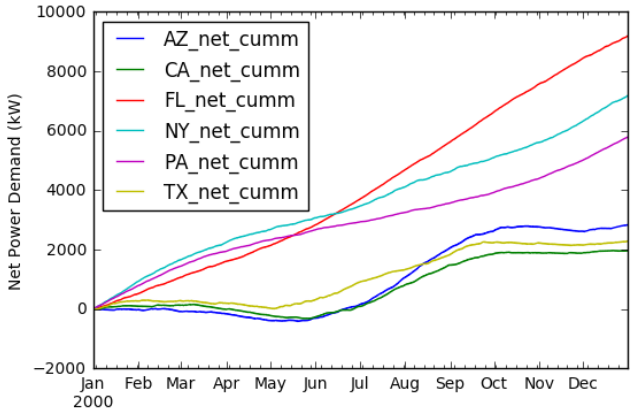


Figure 3 –Cumulative Demand – Solar Production

Another important aspect in trying to understand future impacts of solar penetration is the seasonal variability of solar power. In the following figure, it can be observed that the seasonal variability of available solar energy will need to be considered for reliable grid management.

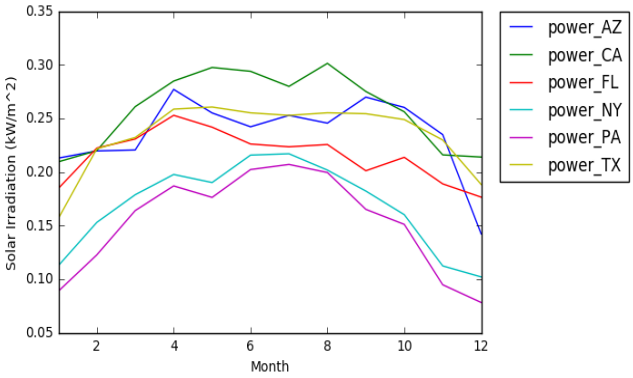


Figure 4 – Seasonal solar productivity by region

Regional solar studies will ensure solar energy reliability. The next scatter plot shows how solar irradiation varies from day to day across regions. Climate in the southwest provides very reliable solar energy while New York is a little harder to predict.

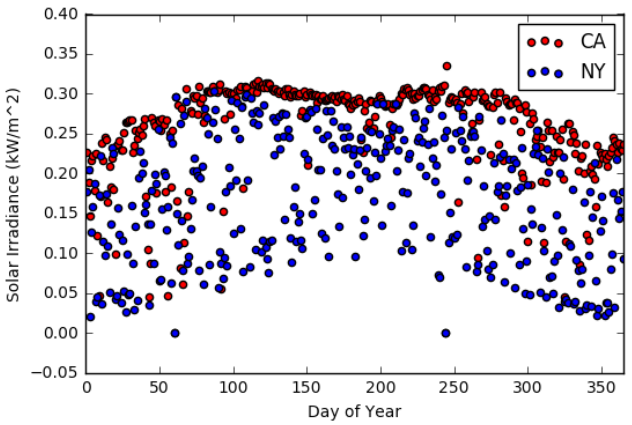
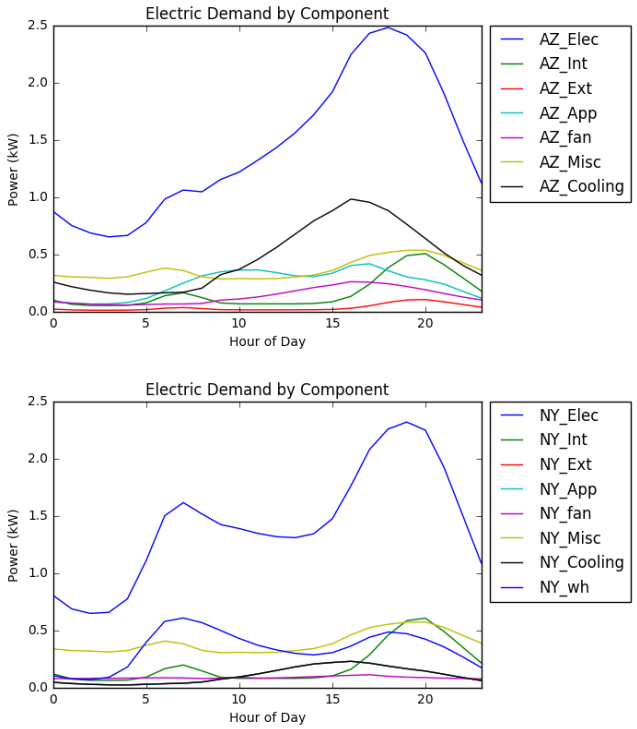


Figure 5 – Daily Solar – New York vs. California

While solar productivity is important to understand as the grid shifts to more renewable sources, an understanding of component electricity demand and energy usage trends will assist the customer when trying to reduce peak loading and the utility to target customers for efficiency upgrades or thermal storage. In the next two figures, a comparison is shown between Arizona and New York. New York uses electric water heating while Arizona doesn't which can be observed by the early morning peaks.

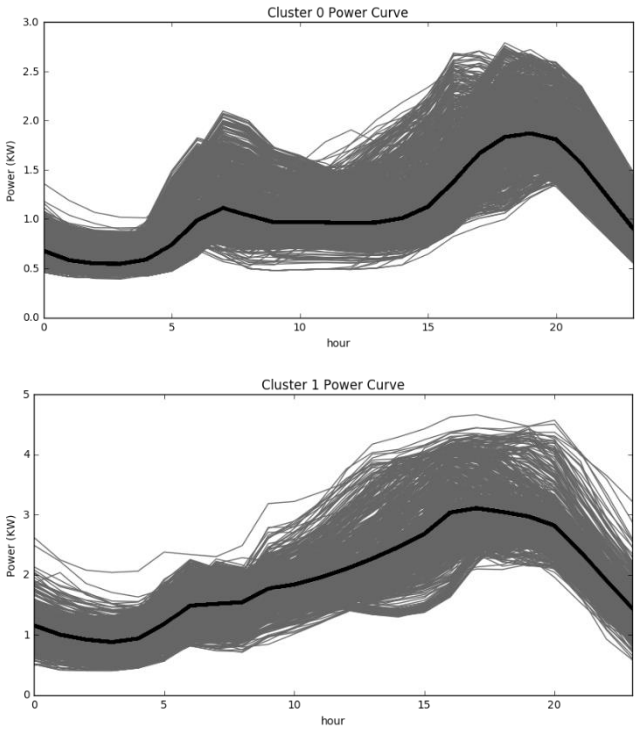


Figures 6a,b – Component electric demand – AZ vs. NY

The shape of the total demand curve for New York presents an opportunity to look at only total energy demand for a specific house and try to determine if they use electricity for water heating. This information might be beneficial to the utility because they can target users for efficiency upgrades if they know the users system as well as using these houses to dump excess energy to store thermally.

4.2 Clustering

Due to observed trends in load curve shapes, a KMeans clustering analysis is presented with the goal of separating total demand load curves by the presence of electric water heating. In this situation, the houses with electric water heating is known. However, reliability of the kmeans clustering using total energy demand is presented to use when future water heating methods are unknown.



Figures 7a,b – Cluster centroids with associated load curves

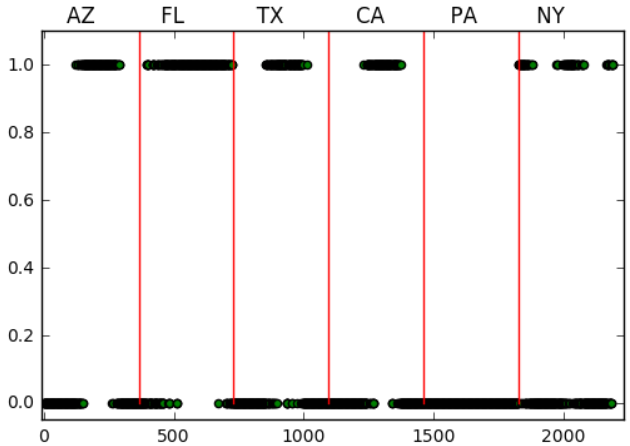


Figure 8 – Cluster assignment by region

When observing figures7a,b, it is hard to tell how the different load curves were clustered. Cluster 0 has more of a defined first peak while cluster 1 actually shows a higher power reading for the early "peak" around 7am. Since appropriation is unclear, a scatter plot was created by location to better understand how and where the different load curves clustered from location to location. It can be observed from the scatter plot that during the summer months,

Arizona, Texas, and California clustered with higher morning peaks. This might be explained by the higher summer cooling loads at these locations. New York is harder to explain. During the mild weather days in the spring and fall, New York clusters with the group identified with lower early morning peak loads because New York's afternoon peak is very small. Visual representation of this can be observed in Figure 9.

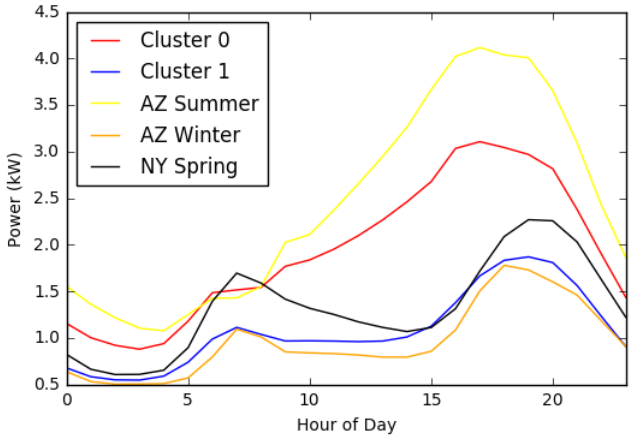


Figure 9 – Seasonal Comparison of Load Curves

The variation in load curve shape from season to season for the different regions makes it hard to recognize the presence of electric water heaters. Even though the NY Spring curve shows a defined early peak, the total curve is more similar to the cluster group without a defined early morning peak. Since load curve shape is most important during cluster appropriation, normalizing due to seasonal variation will not improve the model.

Table 1 offers a more clear representation of cluster appropriation. Cluster 1 represents the higher morning peak when looking at amplitudes.

Table 1 – Cluster Appropriation by Region

Location	Cluster0 - Counts	Cluster1- Counts
Arizona	224	141
Florida	97	268
Texas	247	118
California	249	116
Pennsylvania	365	0
New York	267	98

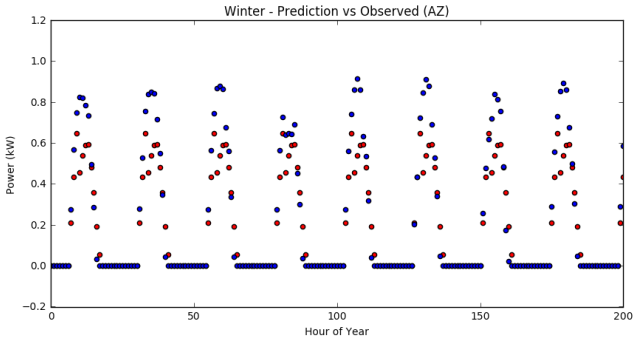
4.3 Regression Analysis

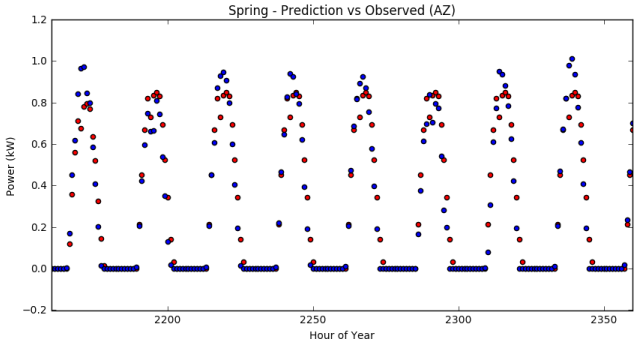
Due to solar pv systems decreasing in cost, increasing in efficiency, and net metering policy, residential rooftop solar systems have grown from almost zero production to well over 2,000MW in the last ten years [3]. Due to the quick rise, general understanding of the impacts of high solar penetration lags behind and creates a situation with many uncertainties when trying to plan for future grid reliability. This paper presents a simple regression model could be created with simple, easy to obtain input variables to enable utilities to assess regional solar potential. The model created uses four input variables (month, hour of day, latitude, and longitude) to predict solar irradiation for a specific location. Florida, California, New York, Pennsylvania, and Texas datasets were used to train the regression model. Once the optimized parameters were chosen, the model output an R^2 value equal to 0.87. The importance factors for each input can be seen in Table 2. As expected, the hour of day was the most important parameter by far. Many more datasets will need to be explored to see the importance of latitude and longitude on a specific locations solar irradiance.

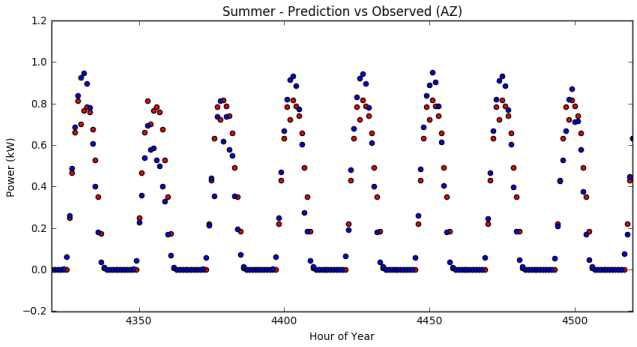
Once the model was fit, the Arizona dataset was used to test the prediction capabilities of the model. The subsequent figures show the seasonal variability of the prediction model compared to the actual data. The winter showed the most prediction error while the spring, summer, and fall showed little error. The predicted values are in red.

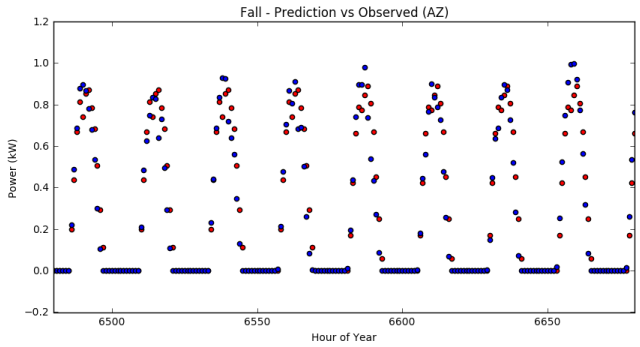
Table 2 – Input Value Importance Factors

	Regression Model
Month	0.036
Hour of Day	0.896
Latitude	0.052
Longitude	0.016









Figures 10a,b,c,d – Season Variation of Regression Model

The R^2 value for the prediction model was 0.85. This seems like a pretty good number, however, this might be skewed a little bit because the model can easily predict all the zero values based on hour of day, which are values we aren't interested in anyway.

5. CONCLUSIONS/FUTURE WORK

This paper explores general energy data across regions and takes a preliminary look at how viable solar pv systems vary from region to regions. Much of this data can be further explored to quantify energy usage trends and solar potential. One area of future work is to split up the clustering data to look at just the morning hours to try and appropriate load curves based on the presence of electric water heaters. By splitting the data, the influence of the second peak will not play a role in the clustering algorithm.

Second, by training the regression tree with many data sets from many locations will enable for better predictions of solar potential from region to region. With only five data points for latitude and longitude, it is hard to extract valuable results from this model.

6. REFERENCES

[1] National Solar Radiation Database http://rredc.nrel.gov/solar/old_data/nsrdb/1991-

[2] US Department of Energy Open Data Catalog http://en.openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states

[3] Solar Energy Industries Association. http://www.seia.org/research-resources/solar-industry-data