Evaluating predictive models for solar energy growth in the US states and identifying the key drivers

IOP Conf. Series: Earth and Environmental Science **127** (2018) 012002 doi:10.1088/1755-1315/127/1/012002

Evaluating predictive models for solar energy growth in the US states and identifying the key drivers

Joheen Chakraborty[1,2] and Sugata Banerji[1]

[1] Department of Mathematics and Computer Science, Lake Forest College, 555 North Sheridan Road, Lake Forest, IL 60045, USA
[2] Worked as a summer research student

Correspondence should be directed to banerji@lakeforest.edu

Abstract. Driven by a desire to control climate change and reduce the dependence on fossil fuels, governments around the world are increasing the adoption of renewable energy sources. However, among the US states, we observe a wide disparity in renewable penetration. In this study, we have identified and cleaned over a dozen datasets representing solar energy penetration in each US state, and the potentially relevant socioeconomic and other factors that may be driving the growth in solar. We have applied a number of predictive modeling approaches - including machine learning and regression - on these datasets over a 17-year period and evaluated the relative performance of the models. Our goals were: (1) identify the most important factors that are driving the growth in solar, (2) choose the most effective predictive modeling technique for solar growth, and (3) develop a model for predicting next year's solar growth using this year's data. We obtained very promising results with random forests (about 90% efficacy) and varying degrees of success with support vector machines and regression techniques (linear, polynomial, ridge). We also identified states with solar growth slower than expected and representing a potential for stronger growth in future.

1. Introduction and previous work

The total installed solar capacity in the world grew by about 200-fold between 2000 (1.4 GW) and 2015 (257 GW). About 13.3% of global solar capacity was in the US as of 2016, but the growth of solar has been uneven. We have collected and cleaned extensive data on solar installations and their potential drivers over the past 17 years across all 50 US states, and applied a variety of machine learning and other modeling techniques in an attempt to construct the most effective model to predict solar growth as well as identify the relative importance of different factors that are driving it.

Most of the work to date that deals with predicting the growth in solar energy have been conducted by government and industry organizations. For example, [1] uses a model named National Energy Modeling System that aims to capture interactions of economic factors with energy supply, demand and prices. The work [2] provides an analysis of the solar growth in the US and [3] analyzes trends in global renewable energy installations. However, none of these works addresses state-level solar penetration or tries to identify the relative importance of contributing factors.

2. Our approach

For this research, we first created a list of candidate factors / independent variables (Table 1) that have the potential to drive the growth in solar energy in US states. There are two dependent variables that we used in our study to represent the solar growth: (1) % growth in solar penetration this year

compared to the last year and (2) % growth in solar penetration next year compared to this year. We used the datasets listed in Section 3 to obtain data for the years 1997-2014.

After cleaning and consolidating the raw data, we systematically applied different modeling techniques including machine learning and regression to predict the two dependent variables using the independent variables. We also analyzed the model residuals to characterize state and year-specific bias in solar growth with respect to the model-predicted values.

3. Datasets

State Energy Data System (SEDS), U.S. Energy Information Administration [4]: SEDS provides comprehensive energy statistics at the state level including energy production and consumption by energy source. We extracted total energy production data from all sources per state for the years 1997-2014 as well as total solar energy (solar photovoltaic and solar thermal) consumed for the same period. We computed solar penetration per state from these data.

US Census Bureau, Median Household Income by State [5]: From this data source we obtained median household income per state for every year between 1997 and 2014 in terms of 2015 dollars.

Kaiser Family Foundation, Population distribution by race/ethnicity [6]: We obtained the demographic makeup of each state's population in 2015 broken down by White, Black, Hispanic and Asian. Given that these distributions generally change slowly over time and lack of easily available and reliable historical data, we assumed the breakdown to remain unchanged over the study period.

National Renewable Energy Laboratory (NREL) Sun Index [7, 8]: This per state index captures the solar energy potential of the state based on the total sunlight received after factoring in daylight duration, cloud cover, etc. Since solar potential of a region changes extremely slowly, we used the same values of the index for the entire study period although the sun index available was for 2006.

U.S. Census Bureau, American Community Survey [9]: This data source provides the percentage of a state's population that have a high school degree or higher, or a bachelor's degree or higher.

U.S. Census Bureau, Population and Housing Unit Estimates [10]: State population numbers are available for each year starting 2000. For years prior to that included in the study (1997-1999), we used raw data from the Intercensal State and County Characteristics Population Estimates and aggregated them up to the state levels.

Federal Election Commission, U.S. Presidential Election Results [11, 12]: We obtained percentage of votes received by Democratic and Republican candidates per state in each presidential election between 2000 and 2016 from this source. For the 1996 presidential election, we processed the raw data available in the ASCII files to obtain the state level vote breakdowns. To estimate the political leaning in the intervening years between presidential elections, we used linear interpolation of the data from two consecutive presidential elections.

Bureau of Economic Analysis (BEA), U.S. Department of Commerce [13]: This data source provided the GDP per state for each year during the study period.

Lawrence Berkeley National Laboratory, Price of PV systems in the US [14]: We obtained installed price of grid-connected, residential and non-residential solar photovoltaic (PV) systems in the US from this report for each year in the study. We assumed the prices to be the same across US due to unavailability of reliable and comprehensive solar cost time series data at state level for the study period.

4. Random Forests

Random forests [15] are an ensemble machine learning technique that constructs many decision trees from the training data and computes the final result by taking a mean of the results predicted by each individual tree. This is an effective way to reduce the problem of overfitting by averaging away any biases in the individual trees. The algorithm used for this research is implemented in Python and is freely available as part of the SciPy module.

4.1. Predicting next year's solar growth

We applied a random forest regressor [15] on our dataset by randomly selecting 75% of the (year, state) data for training and the rest 25% for testing. Using all of the 22 features, the model's efficacy

(measured by the R^2 value) is around 90% (0.901). The importance score of the different independent variables is given in Figure 1. The feature serial numbers on the x-axis are taken from Table 1. From these results, this year's solar growth seems to be the strongest predictor of next year's growth.

4.2. Modeling current year solar growth using independent variables

We also constructed a random forest model to predict this year's solar growth based on other features to identify the strongest factors that explain the growth in solar energy within a state. Using 19 features, the model's efficacy is around 86% (0.862). The relative importance of features is shown in Figure 1.

5. Regression techniques

We utilized three different regression techniques to predict values for 1-year-forward solar growth: linear, polynomial, and ridge. Linear regression fits the data points in the model to a line and returns a linear equation that can be used to predict unknown values. Polynomial regression does this for degrees higher than one. Ridge regression is slightly more involved: first, L2 regularization (penalty equivalent to the square of the magnitude of coefficients) is performed. Next, a λ (tuning parameter) is specified. The value of λ determines the amount of emphasis given to minimizing the sum of the square of coefficients. For models with $\lambda = 0$, ridge regression outputs the same coefficients as simple linear regression. And for models with $\lambda = \infty$, the coefficients will all be zero due to the infinite weightage on the square of coefficients [16]. We used Python implementations from the scikit-learn library.

Table 1: Factors used in developing our models for explaining current year's solar growth and predicting next year's solar growth

Serial No.	Feature Name	Serial No.	Feature Name
1	Solar potential of the state	12	Cost of solar installation
2	Democratic vote / Republican vote	13	Growth in cost of solar installation
3	Total population	14	This year's growth in Democratic vote / Republican vote
4	Median household income	15	This year's population growth
5	GDP	16	Growth in median household income
6	% of population with a bachelor's degree or higher	17	GDP growth %
7	% of population with a high school degree or higher	18	Total energy produced in the state
8	% of population that is white	19	This year's growth in total energy produced
9	% of population that is black	20	Total solar energy consumed in the state
10	% of population that is Hispanic	21	% of energy coming from solar
11	% of population that is Asian	22	This year's solar growth %

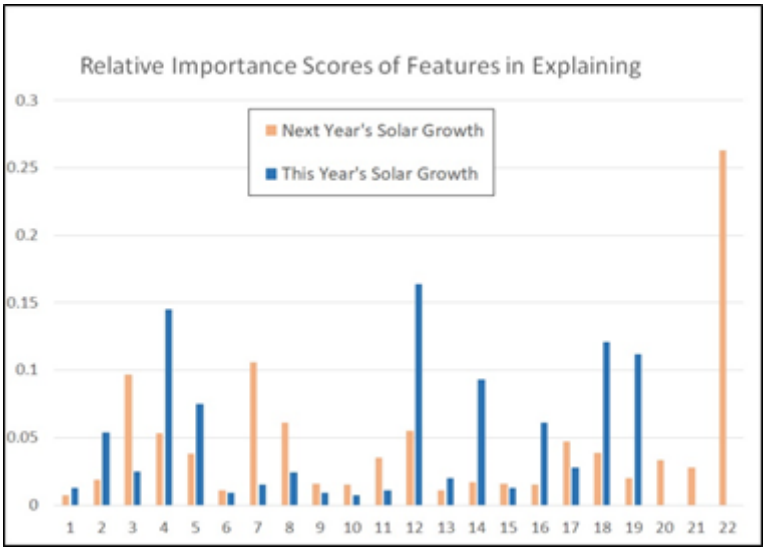
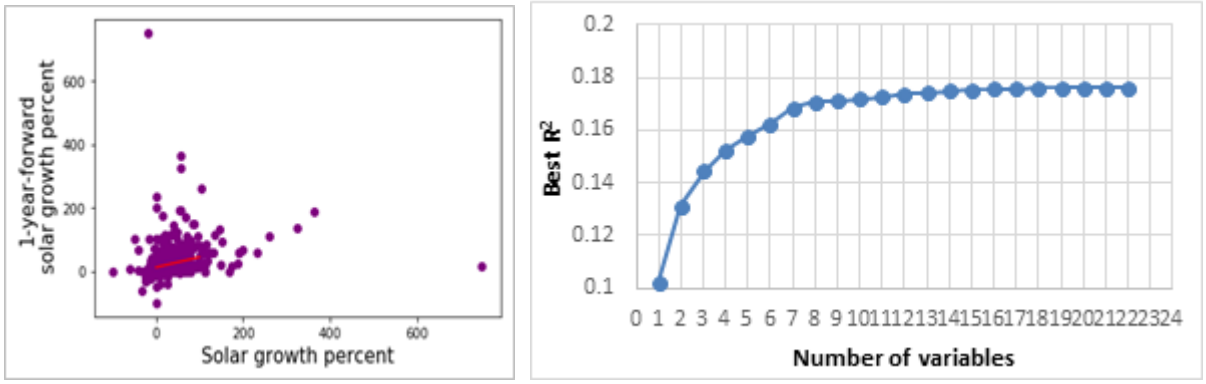


Figure 1: Relative importance of factors in determining next year's solar growth vs. Relative importance of factors in determining this year's solar growth

Using 1-year-forward solar growth % as the dependent variable, we first found the correlation of each independent variable. As anticipated, the most highly-correlated independent variable was this year's solar growth. Using it as the sole independent variable resulted in the model expressed in equation (1) and Figure 2(a) with an R^2 value of 0.1024.

$$y = 0.31959935\,x + 12.18611002 \quad (1)$$

We then used multivariate linear regression, using from 2 to 22 independent variables to predict 1-year-forward solar growth. To do this, we iteratively created every possible combination of parameters in the linear model. Each additional variable increased the R^2 by diminishing margins, shown in Figure 2(b). The R^2 plateaus off at around 10 variables. The highest correlation achieved with linear regression, using all 22 independent variables, was 0.1762. Both polynomial and ridge regression techniques produced poor results with very high error rates. Those results are not included here.



(a) (b)

Figure 2: (a) This year's solar growth vs. Next year's solar growth (b) Increase in R2 with increasing number of explanatory variables

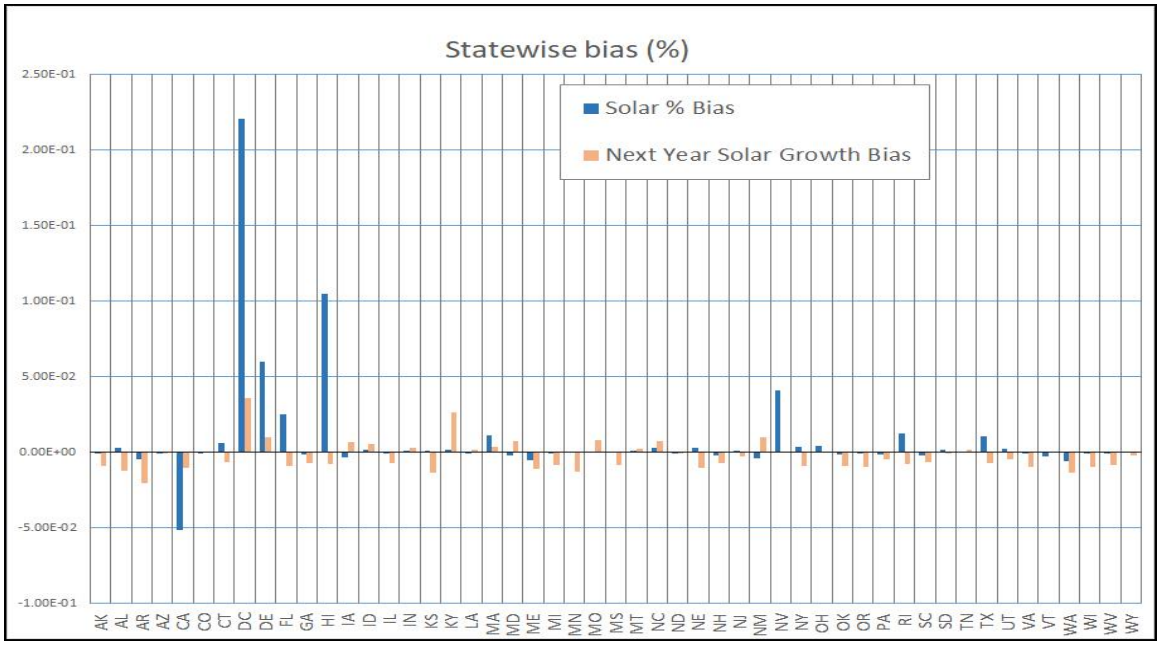


Figure 3. State-wise biases for current year solar % and next year solar growth as obtained from residual analysis of the SVM regressor

6. Support Vector Machines

Support vector machine (SVM) is a machine learning tool first introduced by Vapnik et al [17]. In SVM regression, the input **x** is first mapped onto an m-dimensional feature space using a nonlinear mapping, and then a linear model is constructed in this feature space. The SVM regressor used in this work is the one distributed with MATLAB. The linear model in the feature space *f(x,w)* is given by equation (2) where $g_j(x)$, $j=1,\ldots,m$ denotes a set of nonlinear transformations, and *b* is the 'bias' term [18].

$$f(x,w) = \sum_{j=1}^{m} w_j g_j(x) + b \tag{2}$$

We randomly selected about 75% of the years and used the data from these years for training an SVM and the remaining years for testing. We found the models generated by our program had a low R^2 value, calculated at less than 0.30 for predicting next year's solar growth from the current year's data, and less than 0.60 for predicting the current year's solar production from the other factors for the current year. To understand these results better, we ran two different sets of experiments as explained below.

6.1. Residual analysis by state

We fitted an SVM regressor to all of the data, and calculated the model residuals for each state for each training year. Then we averaged these residuals for each state over all training years to find the mean residual for each state. These results are shown in Figure 3. We see particularly large variation in the residuals for the current year's solar % by state, which we call state-wise "bias". A negative bias for a state indicates that the state is performing poorer than expected on the solar energy production front, and so, that state has a larger potential for growth. A positive bias, on the other hand, means a state is already performing better than expected and there is a smaller potential for higher growth.

On inspecting the biases, we can find some intuitive understanding for some of the states with the highest bias absolute values. For instance, the highest positive bias is for the District of Columbia which may be due to its small size (smallest among all 'states'), or high degree of urbanization (100%), or some other factor not captured by our data. The second highest positive bias is for Hawaii which may be because it is an island without easy access to sources of conventional energy.

6.2. Residual analysis by year

We perform another set of experiments where we fit an SVM regressor to the data and find the mean residuals for each year averaged over all states. We found the year-wise solar % bias to follow a nearly monotonically increasing trend over the years. This is likely due to some factors not considered in our models, such as changing government policy, or increasing focus on environment, energy security, etc. We plan to analyze these biases further in the future to better understand their cause.

7. Conclusion and future work

In this paper we have tried to solve the hitherto unaddressed problem of predicting the solar energy production and next year's solar growth at a US state level. Our contribution with this research is threefold. First, we consolidate data from over a dozen different datasets and identify factors that have the potential to influence solar energy. Second, we perform a comparative evaluation of different machine learning and regression models for solar growth and identify a random forest regression model that can predict/explain the dependent variables with 90% efficacy. Other models, including linear, polynomial, and ridge regression, were also tested, yielding mixed results. Finally, we analyze the residuals obtained by fitting an SVM regressor on the data and identify state-wise and year-wise biases. The state-wise biases can help us identify states with high or low potentials for growth. In future, we wish to continue analyzing these biases to better understand their cause and improve our models further.

8. References

[1] U.S. Energy Information Administration 2017 Annual energy outlook 2017 with projections to 2050.
[2] Sherwood L, Interstate Renewable Energy Council 2013 U.S. solar market Trends.
[3] Office of Energy Efficiency and Renewable Energy, United States Department of Energy 2012 *Renewable Energy Data Book*.
[4] U.S. Energy Information Administration 2016 State Energy Data System *2016 updates by energy source.*
[5] US Census Bureau 2015 Median household income by state: historical income tables
[6] The Henry J. Kaiser Family Foundation 2015 State health facts- population distribution by race/ethnicity
[7] Renewable Resource Data Center, National Renewable Energy Laboratory 2006 Sun index
[8] Nebraska Government 2006 Comparison of solar power potential by state http://www.neo.ne.gov/statshtml/201.htm
[9] U.S. Census Bureau 2010 *American Community Survey 1-Year Estimates: Education Attainment for States, Percent with High School Diploma and with Bachelor's Degree*
[10] U.S. Census Bureau 2015 *Population and Housing Unit Estimates*
[11] Federal Election Commission 2012 *Election Results for the U.S. President, the U.S. Senate, and the U.S. House of Representatives*
[12] Wasserman D 2016 National popular vote tracker, *Cook Political Report*.
[13] Bureau of Economic Analysis, U.S. Department of Commerce 2015 *Regional Economic Accounts – Annual Gross Domestic Product by State*
[14] Barbose G L and Darghouth N R 2016 Tracking the sun IX: the installed price of residential and non-residential photovoltaic systems in the United States, *Lawrence Berkeley National Laboratory Report*
[15] Ho T K 1995 Random decision forests, *Proc. 3rd Int. Conf. on Document Analysis and Recognition (Montreal)*
[16] Tibshirani R 2013 *Modern regression 1: ridge regression* http://www.stat.cmu.edu/~ryantibs/datamining/lectures/16-modr1.pdf
[17] Vapnik V 1995 *The Nature of Statistical Learning Theory* (New York: Springer).
[18] Smola A and Schölkopf B 1998 A Tutorial on Support Vector Regression, *NeuroCOLT Technical Report NC-TR-98-030* (Royal Holloway College, University of London, UK).