Predicting Clean Energy Market for Next 6 Years in the U.S.

Hanyang Xu; Kejia Wu; Rahul Avadhoot; Tong Zhang
Direct Program, University of Washington, Clean Energy Institute

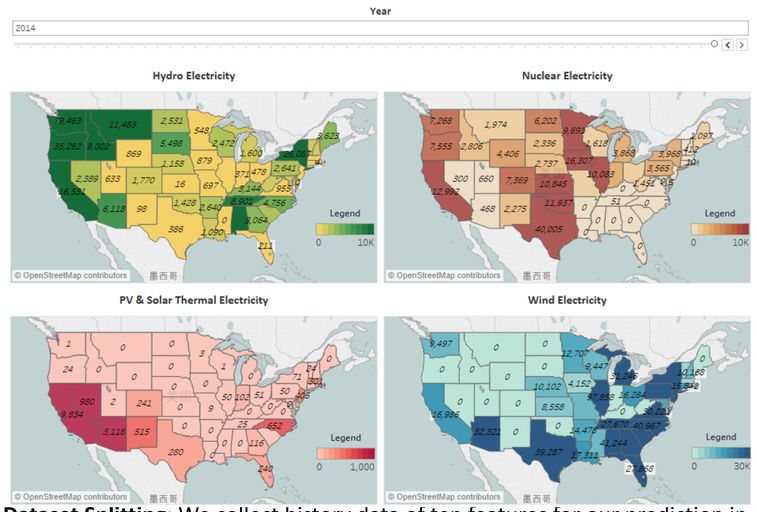
Motivation

Nowadays the clean energy market attracts great attention, especially within states that heavily rely on some renewable energy. With instructive productivity prediction, the companies and policy makers will be able to make more proactive decisions for the local energy markets respectively. The market outlook of clean energy highly depends on the joint effects of a variety of factors, such as crude oil price, GDP, climate condition, and the production of traditional energy as well as other clean energies. Therefore, we are going to analyze these factors and correlate them to clean energy and predict future clean energy output for different states with powerful machine learning methods. These results will be of great value when leaders of energy sector need to make decisions or policy on clean energy development.

Data and Processing

Data Source: State Energy Data System (SEDS) provides comprehensive state-level estimates of energy production, consumption, prices, and expenditures by source and sector from 1960 to 2014. From the dataset, we chose four types of clean energy production (Unit = Million kWh) as our prediction objective: HYTCP(hydro), WYTCP(wind), SOEGP(solar), NUETP(nuclear).

Visualization of Clean Energy Status from History Data



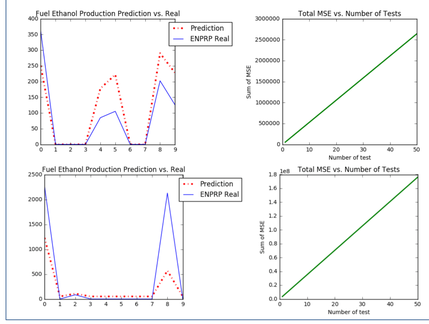
Dataset Splitting: We collect history data of ten features for our prediction in total, which includes other traditional energies, climate factors, GDP values and crude oil price with similar time ranges. As for the data splitting, we take 80% of the data to train our models and use the rest 20% with 5-fold cross-validation method.

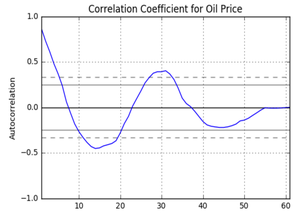
Missing Data: To deal with some features with missing data or varied shapes, we use linear regression for missing data while auto-regression for the prediction of time series variables.

Features Prediction

To predict future clean energy, we need to predict all the time series features first. We manage to solve this problem by building a time series analysis model using auto-regression (AR) model.

After conducting statistical tests, we pre-process the data and then compare order k (e.g. k=1, 2, 3, 4, 5) to see which order number fit best for each feature. For each order, we split data to training and test set at ratio 0.8:0.2.





1. Correlation coef (closer to 1 better) change of oil price with different k
2. Time series prediction of ethanol fuel and MSE(smaller better) for k=3 and k=1 (upper left & lower left)

Clean Energy Prediction I

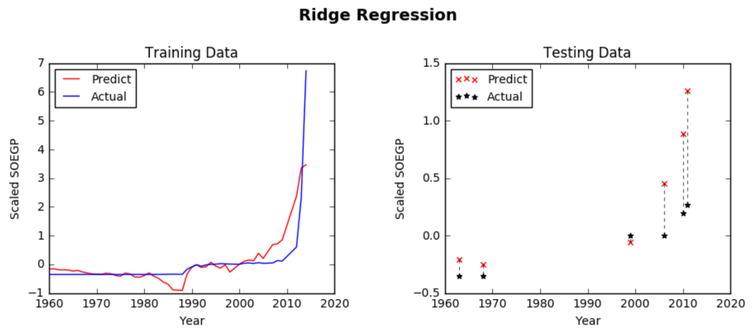
Lasso Regression

Considering multiple features, we try Lasso Regression first inherently performing feature selection. Since each non-zero coefficient adds to the penalty, it forces weak features to have zero as coefficients. The results are good except that it is unstable even on small data changes with correlated features in the data, which brings us to ridge regression.

Ridge Regression

Ridge Regression is used for cases where there are many predictors, but not so many number of observations (in our case, 50 year annual data, 10 attributes). Changing the alpha term prevents overfitting. Ridge regression works well with highly-correlated features (different energy sources).

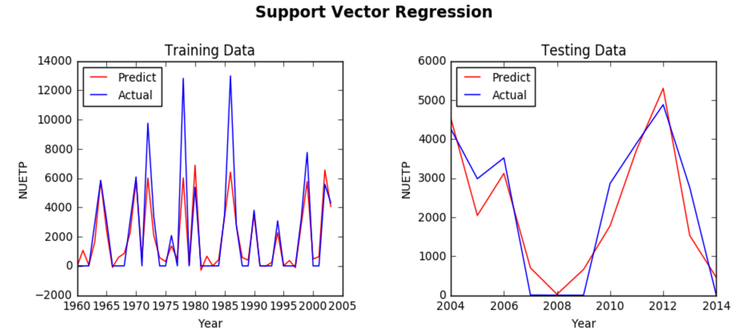
Example: ridge regression on CA data



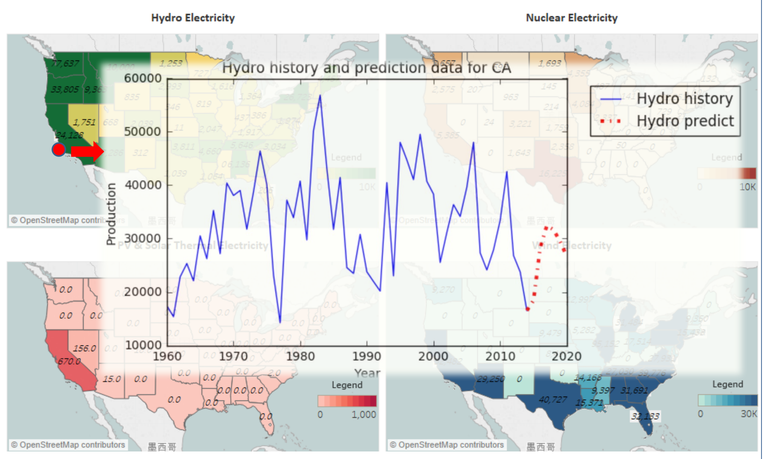
Clean Energy Prediction II

Support Vector Regression(SVR)

We use SVR to regress on the production of four clean energies, which is to find a regression curve with minimum margins to the feature points. After preprocessing all the features, we choose linear, rbf and sigmoid kernel for different energies; the test errors are calculated from cross-validation.



Visualization



Future Clean Energy Production in the contiguous US (in million kWh)

Conclusions

There are certainly rooms for improvements. First, we can also try further predictions on clean energies using autoregression for the energy itself with all other features (all are time series data). It would also be better to perform feature selection first to avoid overfitting. In addition, no matter in Lasso regression or SVR, we select satisfactory parameters for one state and then iterate the model 50 times due to limited time, but it is more accurate to fit parameters for 50 states respectively.

Contact

<Hanyang Xu; Kejia Wu; Rahul Avadhoot; Tong Zhang>
<University of Washington>
Email: hy118@uw.edu; kejiawu@uw.edu; rahulavd@uw.edu; zhangt04@uw.edu
Website: https://github.com/uwkejia/Clean-Energy-Outlook
Advisor: Jim Pfaendtner: jpfaendt@uw.edu ;
David Beck: dacb@uw.edu

Sources

1. State Energy Data System (https://www.eia.gov/state/seds/seds-data-complete.php)
2. Domestic Crude Oil Prices (in $/Barrel) (http://inflationdata.com/Inflation/Inflation_Rate/Historical_Oil_Prices_Table.asp)
3. Gross Domestic Product by State (https://www.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=2%23reqid=70&step=1&isuri=1#reqid=70&step=1&isuri=1)
4. Climate: NOAA Satellite and Information Service (https://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp)
5. Introduction to Statistical Learning by James et al.