For office use only	Team Control Number	For office use only
T1 _____	**80325**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2018
MCM/ICM
Summary Sheet

Summary

In this paper, we construct two evaluation systems: Renewable Energy Indicator (REI) and Network Instability Indicator (NII). By using data mining methods and developing two models, we succeed in visualizing, characterizing, evaluating and predicting the energy flow network and the use of renewable energy.

First, we process the data. We do data screening to select two groups of variables to characterize the energy profile. As some data is missing, we do imputation based on existing data and data from other reliable sources. The data constructing key indicators are normalized with all fifty states. To illustrate energy profile, we also do data visualization with vivid diagrams like Sankey diagram and radar diagram.

Second, we construct two evaluation systems REI and NII to trace and compare the energy profiles of four states and their evolution. NII is derived from a Constrained-Ridge model to characterize the structural stability of energy flow. To specify on renewable energy, we define a REI with five key indicators. We use combined Analytic Hierarchy Process(AHP) and Entropy Weight Method(EWM) to construct REI, which characterizes the profile for use of cleaner, renewable energy. We compare REI and NII of four states and Arizona performs best.

Third, we use several time series models to predict the energy profile of each state in the future. Structural profile of energy flow is predicted with Constrained-Ridge model, and renewable energy profile is predicted with Autoregressive Integrated Moving Average Model(ARIMA) and Long Short Term Memory(LSTM). Based on the comparison and prediction, we determine the renewable energy usage target for four states and propose several feasible actions.

Finally, we conduct sensitivity analysis of our model. We try several machine learning models to predict 5 key indicators, and use independent validation datasets to evaluate their performances. We also do a perturb-and-profile test to evaluate Constrained-Ridge model's power to characterize network profile.

Our model is reasonable and legible with theoretical and data support. The model can be easily applied to characterize the energy profile of renewable energy and energy flow network after data training.

Memo for energy compact

The energy profile of each state is summarized with relate to their energy flow and renewable energy usage.

California is one of the biggest states and has the largest population among all states in the US. Among primary energy, petroleum and natural gas account for most of the energy production and consumption in California, while coal only accounts for a little. Among 4 states, California has the largest proportion of renewable energy consumption, leading the electricity generation from solar, geothermal, hydro energy and so on.

Texas is also one of the biggest states in the US. Compared with California, its energy production and consumption are even higher. Petroleum and natural gas also account for most of energy production and consumption. Renewable energy only accounts for a very small percentage of energy consumption, with wind energy playing a relatively more important role in electricity generation.

Compared with two states above, the total energy consumption of **Arizona** is significantly low. Apart from the fossil fuel which still play a vital role, nuclear energy also accounts for a large percentage of energy consumption, which may due to the low fossil fuel production in Arizona. Among renewable energy, hydro energy counts most because of the high elevations.

New Mexico is also low at total energy consumption. It heavily relies on fossil fuel, and coal accounts most in all energy sources, even more than petroleum. New Mexico is the only one who does not utilize nuclear energy among 4 states, and this can be compensated by its high production of fossil fuel. Wind energy is the prime renewable energy but still accounts for little in energy consumption.

According to our prediction based on historical energy profile evolution, without policy change and severe energy crisis, the energy profile will not change much in the next tens of years. Fossil fuel will keep dominating in energy consumption for each state, and renewable energy will remain a slight proportion in energy consumption. Definitely this will go against sustainable development and environment protection.

Therefore we suggest that some policies should be enacted in support of the use of renewable energy sources. Here we propose some goals about renewable energy:

- For Arizona and California, renewable energy consumption should reach 15% of total energy consumption in 2025, and 25% in 2050.
- For New Mexico and Texas, renewable energy consumption should reach 10% of total energy consumption in 2025, and 20% in 2050.

We hope that these goals can be adopted in the energy compact and practical actions can be taken to achieve the goals.

A model for characterizing, evaluating and predicting energy profile in US states

Contents

1 Introduction

1.1 Background

US has seen a fast increase in energy consumption in recent years. By 2010, US ranks second in total energy consumption, the majority of which comes from fossil fuels, while only a fraction is from renewable energy. Some states, such as California, have to rely on import to make up for the difference between production and consumption. Therefore a energy compact will help improve the energy situation of 4 states.

Several questions should be solved for this compact:

- How to characterize the energy profile of each state? How to evaluate it?
- How does this profile evolve over time? What will it become in future?
- How do other factors (population, GDP, technology, geography etc.) influence the profile?
- What are the goals of the compact? How to take action to reach the goals?

Solution to some questions are quantitative while others are qualitative. We try our best to characterize, evaluate and predict the energy profile in our model and propose reasonable goals and actions.

1.2 Overview of Our Work

We notice the problem has some key points:

- The volume of data is large, highly correlated and redundant. It is hard to have a general and visualized understanding of provided 605 variables across 50 years.
- The energy profile we need to define and illustrate should have very strong and clear interpretability. Common feature selection and dimensionality reduction is not suitable to find crucial variables we need. We need to look into provided data carefully to pick and combine useful variables.
- There are some missing and zero values, which makes profile construction and prediction difficult.
- Since some variables have strong correlation, it is better to consider them in a network to get a better understanding of the energy profile.
- Due to the data limitation, the prediction of energy profile in a far future may have serious deviation from truth. Also it is hard to validate our model.

To determine energy profile and characterize its changes in the past and future, we boil down our tasks in the following steps:

- We do data screening and visualization. We use a energy flow Sankey diagram to vividly illustrate 29 key variables in three categories: energy sources, sectors and flow between them. The energy flow diagram helps us get a very clear and general view of one state's energy profile in a year.[3]
- Second, we further use two variables to characterize energy profile: Renewable Energy Indicator **REI** and Network Instability indicator **NII**.
- To construct REI, we calculate five crucial indicators concentrating on general profile and renewable energy. We do data imputation to fill up missing values. Then we combine Analytical Hierarchy Process(AHP) and Entropy Weight Method(EWM) together to determine weights of the five indicators and construct REI.
- To construct NII, we develop a Constrained-Ridge model to capture the interactions of 29 key variables in energy flow network. By applying constraints to Ridge, the model captures the network character better. We further use perturb-and-profile method to predict the network change after an energy sources suddenly decrease in previous year. We finally construct an indicator NII to characterize the network's instability after a sudden perturbation.
- To predict REI in 2025 and 2050. We use several time series model to predict five crucial indicators needed. We validate our results in a independent datasets to pick up two models: ARIMA and LSTM. To predict NII, we use Constrained-Ridge to predict 29 key variables and then construct it.
- We do sensitive analysis on some parameters and constraints. We also use an independent dataset to validate our prediction. The perturb-and-profile method also helps to validate Constrained-Ridge model.

1.3 Assumptions

- To construct good indicators of energy profile, we should focus on renewable energy usage as well as energy flow network.
- There are a few keys factors to characterize energy profile. Like energy sources, sectors and electricity generation in energy flow network.
- A well structured and stable energy flow network should be resistant to sudden perturbation, so the elements in the network shouldn't change a lot. That's to say, the changing ratio of each 29 variables in the network shouldn't be too big.

- There are some conservations in our energy flow model: electricity generation equals electricity consumption, total amount of energy sources equals total activity consumption.
- In the time series model, we assume that the influence of exogenous factors don't influence factors we need to predict.

2 Data Processing

We use data provided as well as two other datasets: all 50 states data in 1960-20009 and data we collect in 2010-2015. We do several data processing in the following ways.

Data Screening

We do data screening on 605 variables across 50 years in 50 states. All states are used for a better data normalization results.

To construct REI indicator and energy flow chart as well as NII indicator. We mainly collect and combine variables into two categories. By using MSN code abbrievation, we find 35 variables in total to use.

REI system:

5 key indicators are chosen to create a overall profile and calculate REI. We choose six variables from the dataset and combine them. These indicators are shown in the (Table1).

NII system:

We also need 29 variables in the dataset to construct the energy flow network. The variables can characterize the following factors' interaction: energy sources, sectors, electricity generation. We must collect explicit data to characterize energy flow. For example, we use SOEGB to quantify the solar power for electricity generation and ESCCB to quantify electricity generation for commercial sector.

- For energy sources, we have Solar, Nuclear, Hydro, Wind, Geothermal, Natural, Coal and Petroleum. We don't consider Biomass for missing of values.
- For sectors, we have Residential, Commercial, Industrial and Transportation.
- Electricity generation is a crucial part of the network, it connect almost all energy sources and sectors.

Using collected 29 variables, we can quantify the interactions between energy sources, sectors and electricity generation. The variables are like the connect weights(edge) between the network. We can also do the simple math to have nodes values.

Table 1: 5 key indicators

Indicator	Abbreviation	Definition	Choosing reason
TETPB	TETPB	Total energy consumption per capita	It implies the general consumption level.
TETCR	TETCR	Total energy consumed per dollar of real gross domestic product	It implies the energy efficiency.
ESTCB/TETCB	E/T	ratio of total electricity consumption to total energy consumption	It is a sign of energy efficiency level and is related to several electricity-generating renewable energy.
FFTCB/TETCB	F/T	ratio of total fossil fuel consumption to total energy consumption	It implies the dependency on fossil fuel.
RETCB/TETCB	R/T	ratio of total renewable energy consumption to total energy consumption	It implies the development level of renewable energy.

The first use of 29 variables is to plot Sankey diagram of energy flow. The plot we give us a vivid illustration of the energy flow profile. The second is to help construct a network model. We further use these variables to train a constrained machine learning to consider time and network. By doing some test method (perturb-and-profile) on the model, we can construct NII.

Data Imputation

Luckily all the 29 variables in energy flow network have past 50 years data. And there is only one key indicator TETGR for construction of REI misses 17 years data. So we do data imputation on it. Since all 50 states miss TETGR data in 1960-1976, we cannot use unsupervised learning methods like clustering to fill up the missing value.

We notice that TETGR can be calculated by dividing GDPRX with TETCB. Although some GDPRX data is also missing, We can use some methods to estimate one state's GDP growth.

We calculate the average GDP growth rate for each state in last 33 years, which is denoted as GR_{ZZ}. Then we search for the GDP growth rate of US in previous 17 years, which is denoted as $GR_{US,i}$ for i-th year. The GDP growth rate

of each state is estimated by:

$$GR_{ZZ,i} = 0.3 * GR_{US,i} + 0.7 * GR_{ZZ}$$

With the estimated GDP growth rates, we estimate the TETGR in previous 17 years by:

$$GDPRX_{ZZ,i} = \frac{GDPRX_{ZZ,i+1}}{GR_{ZZ,i+1}}$$

,

$$TETGR_{ZZ,i} = \frac{GDPRX_{ZZ,i}}{TETCB_{ZZ,i}}$$

Data Normalization

The 5 key indicators we need to construct REI are of different kinds and have different units. Data normalization can provide an approach to compare different kinds of data and help to easily combine different indicators.

So before applying AHP and EWM to construct REI, we normalize the data of 5 key indicators.

Here we use feature scaling, which can convert values into $[0, 1]$. The problem here is the datasets only have four states, which is not suitable for normalization. We then use the full datasets of 50 states and normalize each indicators in 50 states background. So for value of each indicator x, we collect the values of all 50 states: $(x_1, x_2, ..., x_{50})$. The normalized value is:

$$x_i' = \frac{x_i - \min x_i}{\max x_i - \min x_i}$$

.

The final $x_{min}^{'} = 0$ and $x_{max}^{'} = 1$

Data Visualization

We use many diagrams to help understand interactions and correlation of variables and compare four states' similarity and differences.

- We use basic line graph to illustrate changes of variables and indicator.
- We use Sankey diagram to vividly illustrate basic energy profile and energy flow. We can have a good view of variables in the network and consider their interactions.
- We also use Radar diagram to compare REI system's key indicators of different states.

3 Basic Profile: an energy flow network

To characterize basic profile of energy, we consider to illustrate it from the view of network. Several energy sources(petroleum, natural gas, coal, nuclear

energy, solar energy, geothermal energy, hydro energy, wind energy and electricity), and sectors(electric power sector, industrial sector, transportation sector, commercial sector and residential sector) are chosen as nodes in the diagram. The size of nodes indicates the amount of the total energy consumption of a source or sector, and the width of bands linking two nodes implies the amount of energy consumption.

From the Sankey diagram of four states energy profile in 2009, we can get a general idea of these states' basic profiles, which is shown in memo.

4 Profile Evaluation System: REI and NII

In order to characterize energy profile changes, study its similarity and differences and find the best profile. We should develop some evaluation system to quantify energy profile.

We consider the energy profile for two parts: The first part is the overall profile related with cleaner, renewable energy. The second part is the structural profile characterized by energy flow and their interactions. We use data explained in Data Screening part to construct the two parts' evaluation systems: **REI system** and **NII system**

4.1 REI system

4.1.1 Construct REI

We construct an indicator called REI(Renewable Energy Indicator) to evaluate 4 states' profile of renewable energy usage.

As we have explained in Data screening part, 5 key indicators are chosen to create a overall profile and calculate REI. We choose six variables from the dataset and combine some of them. These indicators are shown in the (Table1).

We apply data imputation to fill up missing values of TETCR. And we apply data normalization to this five indicators we calculated. To have a better normalization effect, we use all 50 states five indicators to normalize. The five indicators of Arizona, California, New Mexico and Texas are picked out for further analysis and are shown in (figure2). The measurement of the use of cleaner, renewable energy sources(REI) is defined as the linear combination of 5 key indicators:

$$\textbf{REI} = \sum_{j=1}^{n} a_j x_j = \vec{a}^T \vec{x}$$

where $\vec{x} = (\text{TETPB}, \text{TETCR}, \text{E/T}, \text{F/T}, \text{R/T})^T$.

We use combined weighting method of Analytic hierarchy process(**AHP**) and Entropy weight method(**EWM**) to evaluate five key indicators' weights.

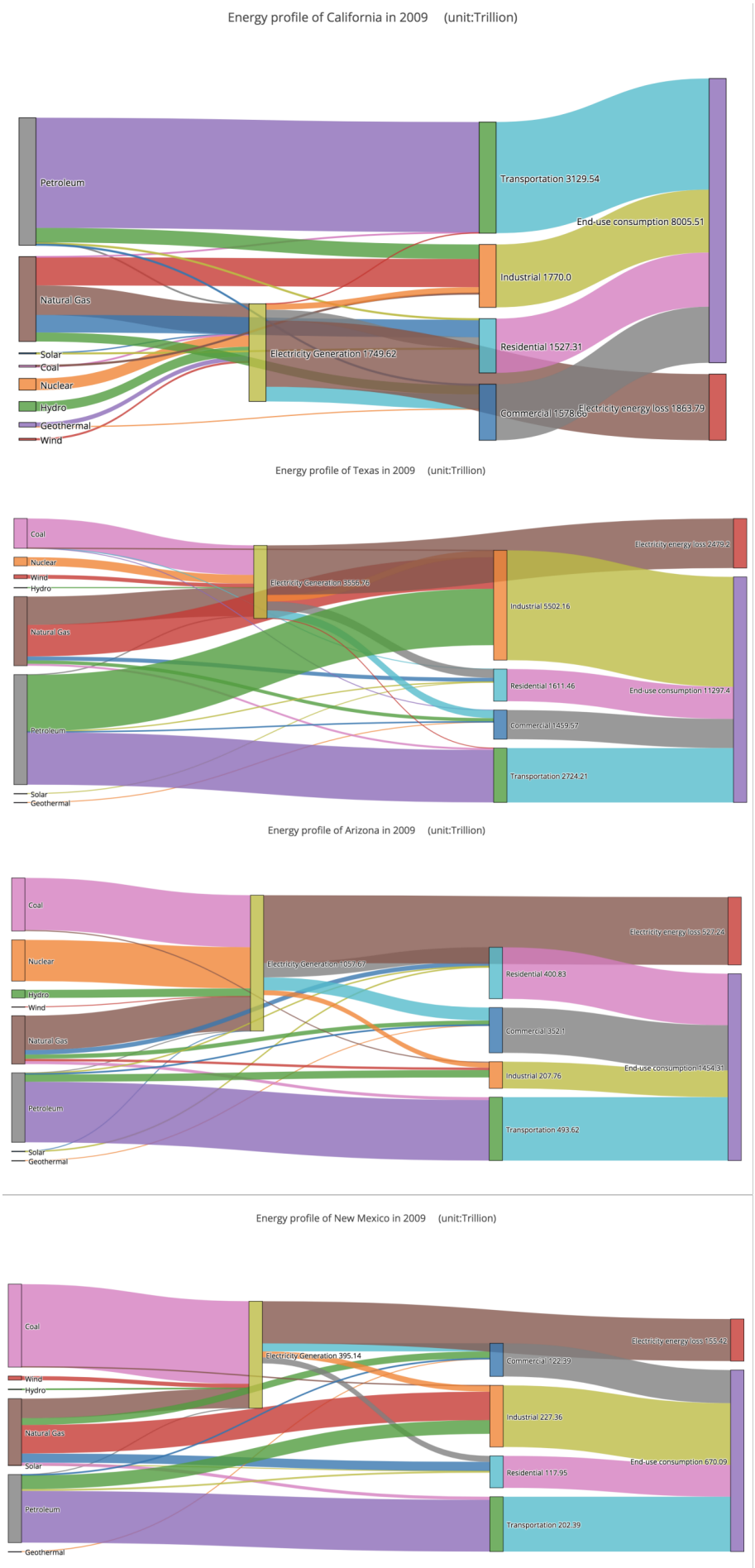

Figure 1: energy flow network of each state in 2009

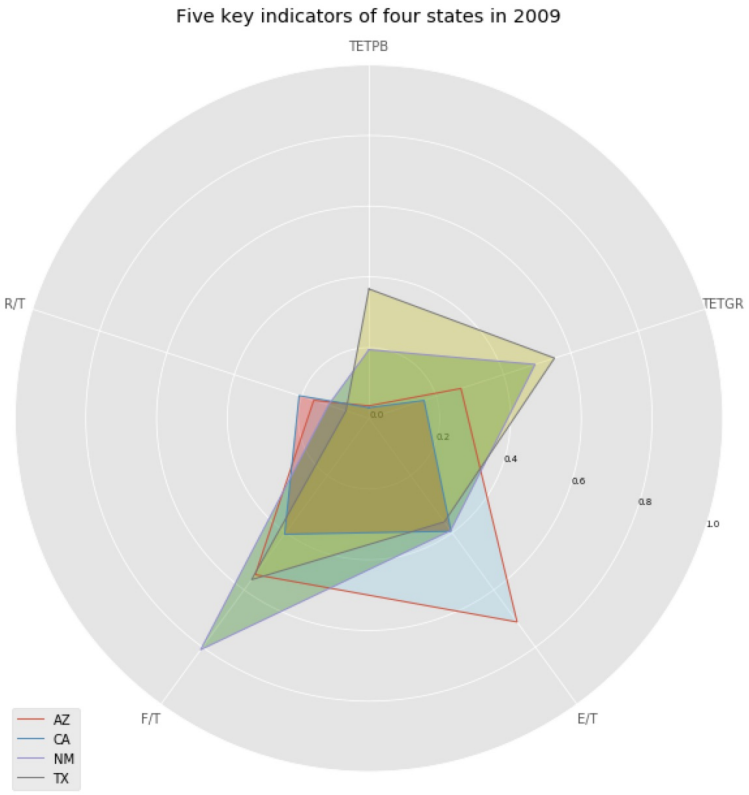

Figure 2: 5 key indicators of each state in 2009

Since the TETGR and F/T are usually negatively correlated with renewable energy, we set $\vec{c}_{AHP} = (a_1, -a_2, a_3, -a_4, a_5)^T$. Then we employ AHP to determine the coefficients $\vec{c}_{AHP}$. By their impact on the use of cleaner, renewable resources, we derive the following comparison matrix:

	TETPB	TETGR	E/T	F/T	R/T
TETPB	1	1	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{1}{7}$
TETGP	1	1	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{1}{7}$
E/T	5	5	1	$\frac{1}{2}$	$\frac{1}{2}$
F/T	3	3	2	1	$\frac{1}{3}$
R/T	7	7	2	3	1

The weight vector is the normalized eigenvector corresponding to the largest eigenvalue:

$$\vec{c}_{EWM} = [0.1028, 0.1031, 0.3908, 0.3991, 0.8166]^T$$

In the consistency check, we get **CI**=0.047<0.1, suggesting that these weights determined by AHP are reasonable.

In order to be more objective when determining the coefficients, EWM is also employed to determine the coefficients without prior knowledge about the effect of each indicator. The coefficients c_{EWM} are determined by the Shannon Entropy of each group of indicators: Determine ratio of x_{ij} in j-th indicator, i denotes year.

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \quad i = 1, ..., 50, j = 1, ..., 5$$

Calculate j-th indicator's entropy:

$$e_j = -k \sum_{i=1}^{n} p_{ij} ln(p_{ij}) \quad j = 1, ..., 5, k = 1/ln(n)$$

Calculate each indicator's weight:

$$w_j = \frac{d_j}{\sum_{j=1}^{5} d_j} \quad j = 1, ..., 5$$

Combining these 2 methods, we set combined weights of each indicator:

$$c_{AVG} = \frac{c_{AHP} + c_{EWM}}{2}$$

4.1.2 Results of REI

With these coefficients, REI is calculated for each state in 50 years, shown in (figure3). The bigger REI is, the better use of cleaner, renewable energy.

4.2 NII system

To further analyze the energy profile, it is better to quantify the variables in a network to consider their interactions. We have found 29 variables characterize energy flow and plot the Sankey diagram. In this part we try to develop a model with constraints to characterize energy profile from a network view and construct another indicator called NII(Network Instability Indicator).

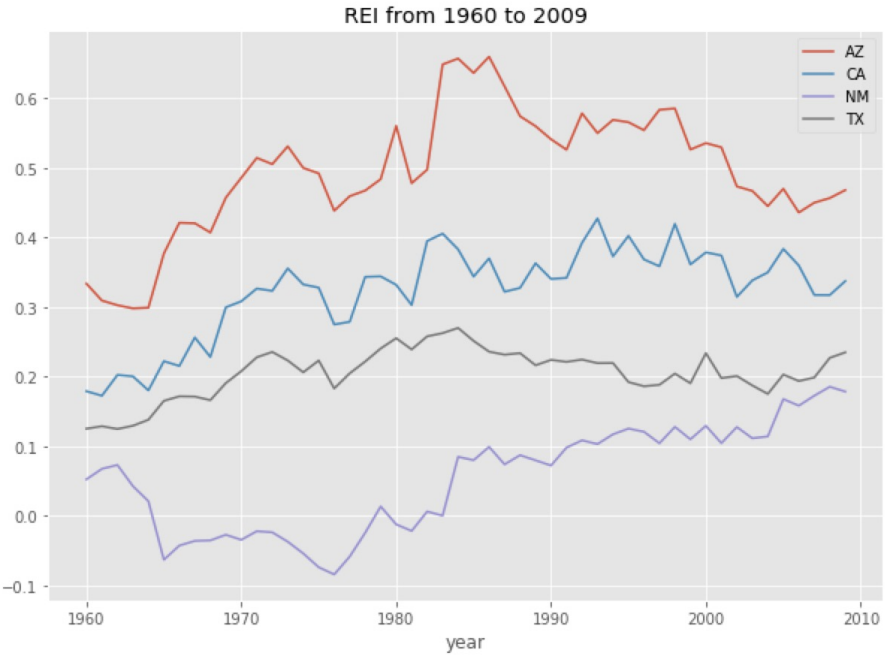

Figure 3: REI of each state in 50 years

4.2.1 Network Model: Constrained-Ridge Model

We use a Linear model with L_2 regularization called Ridge to characterize the network interaction:

$$y = X\beta + \epsilon$$

By minimizing the following function we can optimize parameter β:

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2 + \alpha||\beta||_2^2$$

We train Ridge to learn this year's 29 variables to predict next year's. We believe by training the model, it can capture the interaction of 29 variables. Parameters in machine learning model can construct a state-transition matrix, it is also a network to capture and predict the change of variables. (figure 8 in appendix) is Ridge's parameter β. It illustrates the interaction of 29 variables.

4.2.2 Constrained-Ridge

To make Ridge learns correctly from the data, we add some constraints to Ridge to help it learn better. From assumptions, we have

electricity generation $\equiv$ electricity consumption

total amount of energy sources $\equiv$ total activity consumptions

So it changes to a Constrained optimization problem:

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & ||f(\mathbf{X}) - \mathbf{y}||_2^2 + \alpha||\beta||_2^2 \\ \text{subject to} \quad & \vec{b_1^T} f(\mathbf{X}) = 0, \\ & \vec{b_2^T} f(\mathbf{X}) = 0. \end{aligned}$$

The constraints can be rewritten as:

$$\begin{bmatrix} \vec{b_1^T} \\ \vec{b_2^T} \end{bmatrix} f(\mathbf{X}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In order to optimize the Constrained-ridge function, we first determine the derivative of it. We notice that the data is in matrix format. So we determine the ridge function derivative in matrix form:
At first we use:

$$f(\mathbf{X}) = \mathbf{X}\beta$$

to rewrite the ridge function:

$$\begin{aligned} ||f(\mathbf{X}) - \mathbf{y}||_2^2 + \alpha||\beta||_2^2 &= \mathbf{tr}[(f(\mathbf{X}) - \mathbf{y})^T (f(\mathbf{X}) - \mathbf{y})] + \alpha \cdot \mathbf{tr}(\beta^T \beta) \\ &= \beta^T \mathbf{X}^T \mathbf{X}\beta - 2\mathbf{X}^T 2\mathbf{y}^T \mathbf{X}\beta + \mathbf{y}^T \mathbf{y} + \alpha\beta^T \beta \end{aligned}$$

Then we derive the derivatives of Ridge:

$$\begin{aligned}\frac{\partial L(\beta)}{\partial \beta} &= 2\mathbf{X}^T\mathbf{X}\beta - 2\mathbf{X}^T\mathbf{y} \\ &= 2(2\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}) - 2\mathbf{X}^T\mathbf{y}\end{aligned}$$

With Ridge and it's derivatives formula as well as the constraints in calculable form, we apply a optimization method called Sequential Least Squares Programming(SLSQP) to optimize parameter β.

4.2.3 Generate NII: Perturb-and-Profile

In order to construct an indicator of the network, we propose a method called perturb and profile. It is a method aiming to test network's stability. We assume this year's variables concerning petroleum suddenly decrease 20 percent, then we apply Constrained-Ridge to predict next year's 29 variables in the network. We compare the predicted values with real values and calculate its rate of change(we ignore zero values).

Then we calculate a weighted changing rate according to the variables value, which means the variables with big values(Coal for electricity generation or natural gas for industrial) can lead to bigger changes and damage in same change of rate. The definition of NII is:

$$\textbf{NII} = \sum_{i}^{29} x_{ij} \| \frac{x_{ij}^{'} - x_{ij}}{x_{ij}} \| \quad i \text{ denotes variables and } j \text{ denotes year}$$

The smaller NII is, the more stable a state's energy network are.

We perform the perturb method and observe the Sankey plot, we find that the Constrained-Ridge model tend to predict the variables average, that is predict their amounts similar. In order to fulfil the real condition, we add another inequality constraints to Ridge: All variables changing rate should be less than 0.2. For the constraints become very severe, we increase the optimization iteration and try different initialization arguments to optimize.

4.2.4 Results of NII

It is shown in (figure4) that all four states' NII slowly decrease, indicating that the energy network's structure develops towards a more stable and healthier situation.

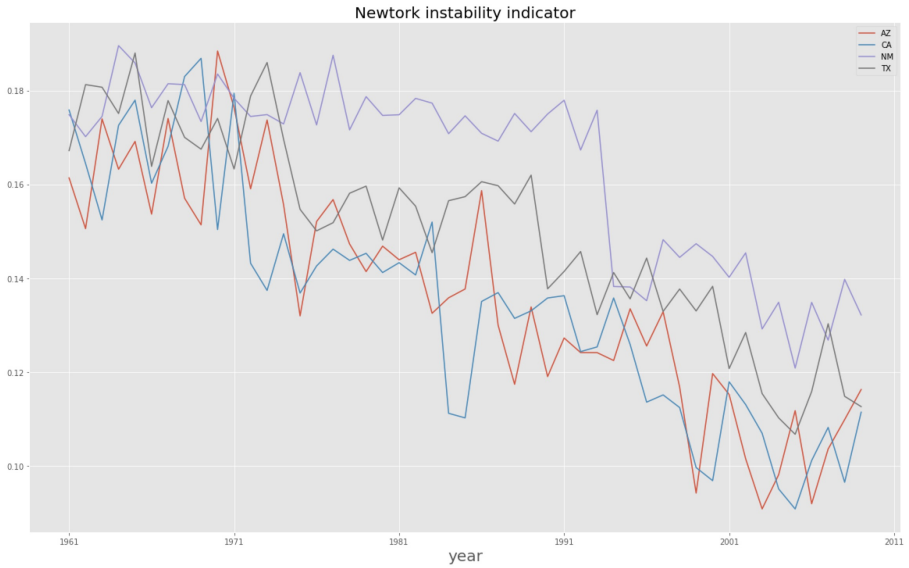

Figure 4: NII of each state in 50 years

4.3 Evaluation of profile

Arizona and California have better profile for usage of cleaner, renewable energy sources than New Mexico and Texas. These two states are similar in several aspects, including low energy consumption per capita, high energy intensity, low dependency on fossil fuel, high dependency on electricity and renewable energy sources in the US. On the contrary, Texas and New Mexico both have high energy consumption per capita, low energy intensity, high dependency on fossil fuel, low dependency on electricity and renewable energy sources. Besides, energy consumption per capita and energy intensity both tend to decrease in 4 states(figure10 in appendix).

Arizona has plentiful renewable energy sources like wind, solar and hydro energy due to its climatic and geographical factors. For example, low precipitation contributes to abundant solar energy and high elevation contributes to abundant hydro energy. Besides, it is lack in fossil fuel resources, probably promoting the use of renewable energy and nuclear energy.

California has a large population and many industries, which require much energy. However, California puts great effort into increase energy efficiency and promote the development of new energy. Besides, its mild climate contributes to some renewable energy like solar energy and biomass. It is worth noting that the environmental movement greatly promotes the increasing usage of renewable energy and reducing usage of coal, which causes serious environmental pollution.

Texas has a large population and many energy intensive industries, so it is heavily dependent on fossil fuel. Renewable energy consumption maintains a low level for a long time. Although wind and solar energy resources are available on high plains, they are still underutilized compared with other states.

New Mexico contains plentiful fossil fuel and renewable energy resources. It is a large supplier of fossil fuel including petroleum, coal and natural gas. Although there are substantial wind, solar and other renewable energy resources in New Mexico, the usage of these energy cannot match with those in California and Arizona. One possible reason may be relatively high cost of renewable energy compared with low cost of fossil fuel.

Because Arizona has biggest REI among 4 states and a small NII in 2009, it is considered to have the 'best' profile for usage of cleaner, renewable energy sources. The most important contributing factors are its relatively big R/T, big E/T and small F/T. I.e., among all states, Arizona relies much on renewable energy and electricity and little on fossil fuels. California follows Arizona on REI and has even smaller NII, so the profile of California is also great.

5 Predict REI and NII

To predict future's energy profile(in 2025 and 2050), we develop Time Series model to predict our two indicator systems: REI and NII.

There are several ways to predict time series data.

For NII, we can only use Constrained-Ridge to predict future NII, for the construction of NII depends on Constrained-Ridge model. As a machine learning model, Constrained-Ridge also has the ability to predict the future(in fact it is how perturb-and-profile works).

For REI, we can try different kinds of time series model to predict 5 indicators and then construct REI. We try a common model called Autoregressive Integrated Moving Average Model (**ARIMA**) and several kinds of machine learning models including Kernel Ridge, Random Forest, Support Vector Regression, K-Nearest Neighbor Regression, Gaussian Process Regression and a Deep Learning model called Long Short Term Memory model (**LSTM**).

5.1 Time Series Model

We try many kinds of time series model to predict 5 indicators in REI. The following two models have the seemingly best results.

ARIMA

To predict data in short term based on history, we consider using traditional ARIMA model. Take Texas's energy consumption per capita (TETPB) as an example. We use ADF test to determine a first-order difference of data. Then use BIC criterion to determine that both p (ACF order) and q (PACF order) equal 5.

Case by case, we can predict five indicators of four states in REI in 2025 and 2050.

LSTM

We try several kinds of machine learning models and choose the most suitable one based on two criteria.

- The pearson correlation coefficients(PCC) of prediction and truth value of 2010-2015.

$$PCC = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

  Higher PCC indicates better model performance.

- Long time range capture capability. Given that our dataset is small and the prediction is in far future, the prediction deviation may increase year by year. So we should find a model capturing long time range well.

  We have tested several machine learning models on 2010-2015 datasets and compare their PCC (figure5). Although the model performs well on 2010-2015 dataset, it may experience severe overfitting problem later.

  Apart from changing training batches of datasets and set back steps more than one, LSTM has the mechanism which naturally capture the long time range correlation. It may partially solve the over fitting problem, but the little amount of data will cause over fitting problems too. As we don't have validation data to test our future prediction, it is hard to check the model's performance. Here we choose LSTM because it is the best to capture the long time range correlation.

5.2 Prediction Results

We try both Autoregressive Integrated Moving Average Model(ARIMA) and Long Short Term Memory(LSTM) model to predict 5 key indicators and REI of each state. The results of REI are shown in (figure6) and (Table2).

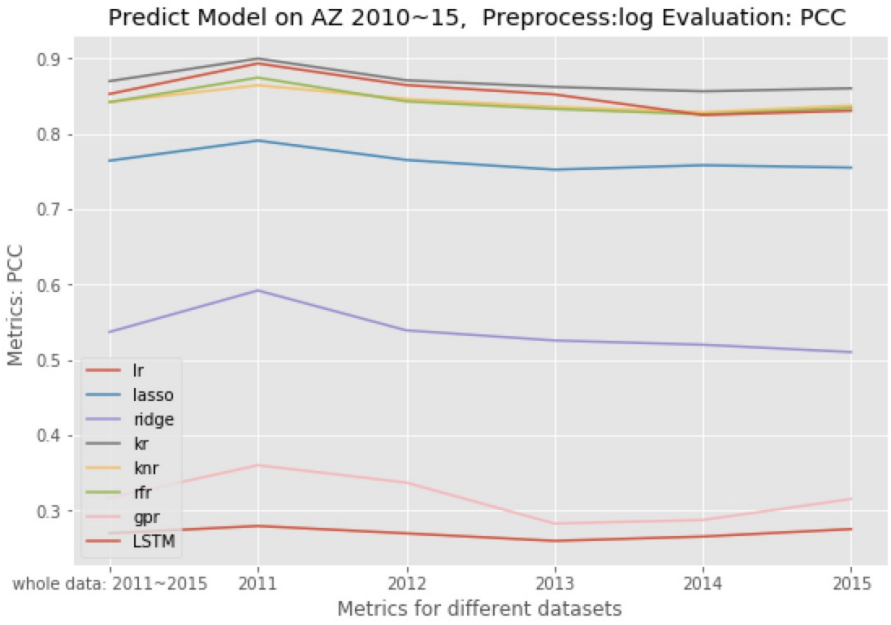

Figure 5: PCC of 8 models

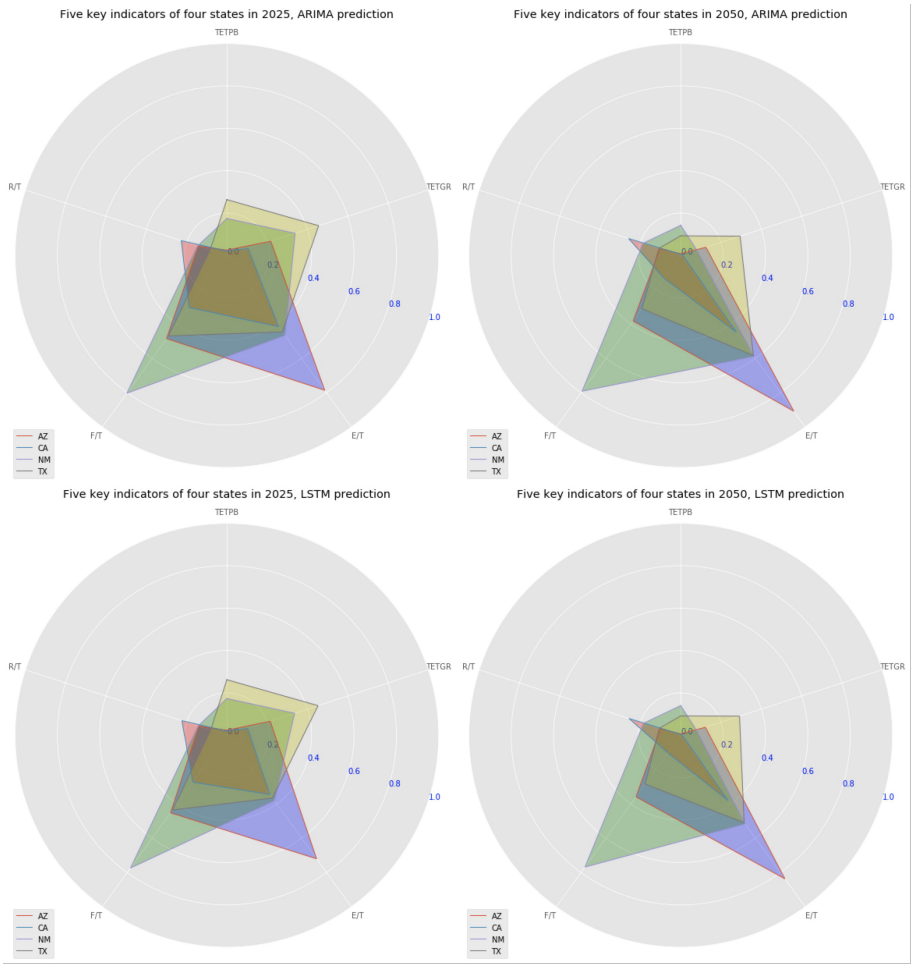

Figure 6: predicted 5 key indicators

Table 2: predicted REI

State / Model	AZ	CA	NM	TX
ARIMA,2025 year	0.718	0.516	0.451	0.480
ARIMA,2050 year	0.805	0.606	0.595	0.577
LSTM, 2025 year	0.665	0.459	0.380	0.415
LSTM, 2050 year	0.744	0.551	0.534	0.516

We use Constrained-Ridge to predict the energy flow and NII of each state. The results are shown in (figure7 in appendix,Table3).

5.3 The goal and actions

In order to simplify this goal, we consider the ratio of renewable energy consumption to total energy consumption(R/T) as the most important indicator. We notice that in the last 5 years(2005-2009), the values of R/T of Arizona and California are close to 0.1, and those of New Mexico and Texas increase quickly(although they are just around 0.05). In out prediction, the values of R/T do not change a lot without policy changes. Besides, in the years after 2009, 4 states adopted more aggressive renewable goals and improved a lot in use of renewable energy sources[9, 10]. As renewable energy will become more and more important in future because of fossil fuel resources reduction and environmental concerns, we suggest the following reasonable but challenging goal:

- For Arizona and California, renewable energy consumption should reach 15% of total energy consumption in 2025, and 25% in 2050.
- For New Mexico and Texas, renewable energy consumption should reach 10% of total energy consumption in 2025, and 20% in 2050.

Some feasible actions

Interstate communication and cooperation on use of renewable energy sources should be strengthened. For example, California leads the technology development of several renewable energy usage like solar energy. The experience and technology of California is valuable for other states. All 4 states have large renewable energy potential but underutilize it. By governmental and commercial communication and cooperation, common development of renewable energy usage can be attained in 4 states.

Governmental program and legislation should be in favor of renewable energy usage. For example, the Renewable Energy Program Emerging Renewables Program proposed by the California Energy Commission helped "provide market-based incentives for new and existing utility-scale facilities powered by renewable energy". They contributed to total renewable electricity production increase statewise. New programs should be proposed every several years in each state to fit in with future renewable energy usage tendency[8].

Table 3: predicted NII

State / year	AZ	CA	NM	TX
2025	0.0789	0.0976	0.1201	0.0918
2050	0.0343	0.0503	0.0721	0.0447

Increasing energy efficiency is also an important action, as proved by Arizona and California. For example, by reducing the energy loss in electricity transmission and use energy-efficient technologies, California manages to restrain energy demand(especially fossil fuels) even with many energy-intensive industries[4].

Distributed renewable generation is also promising, as shown in Arizona. Although customer-sited electricity generation only contributes to a small part in total generation now, the potential is far from being completely realized. Infrastructural and financial support should be provided to customers to encourage the development of renewable energy-dependent electric utilities.

6 Sensitivity Analysis and Validation

Independent Dataset Test

For a data science problem, the best way to test a model is using independent datasets, especially for complicated machine learning models with so many variables. So we find another datasets in 2010-2015, recording same variables as the 1960-2009 datasets.

To check if the two datasets have very high similarity, we do correlation test of the two datasets. 99% of variables have PCC larger than 0.99, so we can say the validate datasets in 2010-2015 is effective in evaluating our time series model's prediction. Then we test several model's prediction results with real value and calculate its PCC (figure9 in appendix). But as we have mentioned in previous time series model part, the model may perform poorly in far future and it is very hard to evaluate the model performances.

Perturb and Profile Test

We do perturb and profile test to check Constrained-Ridge model's effectiveness as well as construct NII. It is shown in (figure4) that the Constrained-Ridge performs well at characterizing the network feature.

ARIMA Test

For ARIMA model, we do several kinds of test to check the model. First we do ADF test to check the stationery after first-order difference. We also do autocorrelation and partial autocorrelation test to check the difference effect. We then do Durbin-Watson test to check first-order autocorrelation. Next, we run Ljun-Box test to check the overall randomness based on a series of lags. This test is more comprehensive than deciding randomness at each lag. Finally, we run test of residuals.

We test all the ARIMA models on different variables according to the test pipeline until the result passes the test. As log value of the data is needed, we ignore the zero values.

7 Future Work

- **Adjust the subjective evaluation process.**

  When choosing variables and determining its weights, we do many subjective work. Although 29 variables that we choose to characterize energy flow profile seem reasonable, they still need more evaluation. The REI evaluation system is more subjective, due to determining five key indicators and using AHP to determine weights. We also apply some subjective but incomplete model constraints. The experts in energy field may help us evaluate and modify our model.

- **Use more data** As a data science problem, more data is always helpful. Although we have used as many data as we can (data from 2010-2015 datasets and all 50 states), the prediction still remains really hard. Both our Constrained-Ridge model and Time Series models like ARIMA and LSTM need more data. For example, we observe that LSTM model's loss function MSE decreases really slowly, indicating that it is very difficult to use so small datasets to train.

  But this problem isn't a typical machine learning problem and the data volume of same variables is restricted. We can hardly use more data than provided. So we should consider more prior knowledge of the energy field. For example, a better way to use more data is to establish a larger network to include more data, but it need much work of case by case data screening and a lot of experience. It should also be noticed that a bigger network means more parameters in the model, for example a non-linear model like Random Forest or Gaussian Process Regression. These models have more potential risks of overfitting.

8 Conclusion

We have done much data processing work including screening, imputation, normalization and visualization. Two evaluation systems are constructed to characterize energy profile: REI and NII. We use combined AHP and EWM to construct REI. We also use a Constrained-Ridge model to characterize the energy flow in a network and construct NII. For future indicator prediction, We use Constrained-Ridge to predict NII and try several kinds of machine learning models to predict REI. We finally choose one basic time series model ARIMA and one deep learning model LSTM to predict REI.

Our models have some strengths and some weaknesses.

8.1 Strengths

- **Integrity** We try to use as many data as we can. We use all 50 states for normalization and 10-15 datasets for validation.

- **Theory-Based** All the models and methods are theory-based. By doing every step we do checks to be reasonable. For example the whole process of ARIMA test and the PCC test before using 2010-2015 data.
- **Cross Field** We get many inspiration from biological field. For example the energy flow chart is similar to biological energy flow. Also the perturb-and-profile method gets some inspiration from a state-of-art testing RNA secondary structure method.
- **Flexible and Extendable** The time series models could include additional factors. The network model can also be applied to other variables satisfying the characterization of network.

8.2 Weaknesses

- **Data Limitation** As a machine learning model, not enough data makes it hard to train and easy to overfit. As we have discussed in future work part, it is hard to include large amount of data. So we should consider designing and optimizing our model with more prior knowledge.
- **Subjective and Simple Assumptions** We use some very subject methods and variables discussed in future work part too. Our own experience and intuition may be biased. Also some simple assumptions will omit some potential useful data and information.

References

[1] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning[M]. New York: Springer series in statistics, 2001.

[2] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. 1999.

[3] Estimated Energy Use in 2012. Lawrence Livermore National Laboratory: http://flowcharts.llnl.gov/

[4] Arizona State Energy Profile. US Energy Information Administration. https://www.eia.gov/state/print.php?sid=AZ

[5] California State Energy Profile. US Energy Information Administration. https://www.eia.gov/state/print.php?sid=CA

[6] Texas State Energy Profile. US Energy Information Administration. https://www.eia.gov/state/print.php?sid=TX

[7] New Mexico State Energy Profile. US Energy Information Administration. https://www.eia.gov/state/print.php?sid=NM

[8] California Renewable Energy Overview and Programs. http://www.energy.ca.gov/renewables/

[9] Arizona regulator proposes 80% clean energy mandate, 3 GW storage target. https://www.utilitydive.com/news/arizona-regulator-proposes-80-renewable-energy-mandate-3-gw-storage-targ/515872/

[10] New Mexico considers more aggressive renewable energy goals. https://americaninfrastructuremag.com/new-mexico-renews-energy-goal/

Appendices

Table 4: 29 variables used in network, unit: Btu

MSN	Description
SOEGB	Electricity produced from photovoltaic and solar thermal energy by the electric power sector.
NUETB	Electricity produced from nuclear power.
HYTCB	Hydroelectricity total production.
WYTCB	Electricity produced from wind energy.
GEEGB	Electricity produced from geothermal energy by the electric power sector.
NGEIB	Natural gas consumed by the electric power sector (including supplemental gaseous fuels).
CLEIB	Coal consumed by the electric power sector.
PAEIB	All petroleum products consumed by the electric power sector.
CLRCB	Coal consumed by the residential sector.
NGRCB	Natural gas consumed by (delivered to) the residential sector (including supplemental gaseous fuels).
GECCB	Direct use of geothermal energy and heat pumps in the commercial sector.
NGCCB	Natural gas consumed by (delivered to) the commercial sector (including supplemental gaseous fuels).
NGICB	Natural gas consumed by (delivered to) the industrial sector (including supplemental gaseous fuels).
CLCCB	Coal consumed by the commercial sector.
CLICB	Coal consumed by the industrial sector.
NGACB	Natural gas consumed by the transportation sector.
ESRCB	Electricity consumed by (i.e., sold to) the residential sector.
PARCB	All petroleum products consumed by the residential sector.
ESCCB	Electricity consumed by (i.e., sold to) the commercial sector.
PACCB	All petroleum products consumed by the commercial sector.
ESICB	Electricity consumed by (i.e., sold to) the industrial sector.
PAICB	All petroleum products consumed by the industrial sector.
ESACB	Electricity consumed by (i.e., sold to) the transportation sector.
PAACB	All petroleum products consumed by the transportation sector.
LOTCB	Total electrical system energy losses.
TERCB	Total energy consumed by the residential sector.
TECCB	Total energy consumed by the commercial sector.
TEICB	Total energy consumed by the industrial sector.
TECCB	Total energy consumed by the commercial sector.

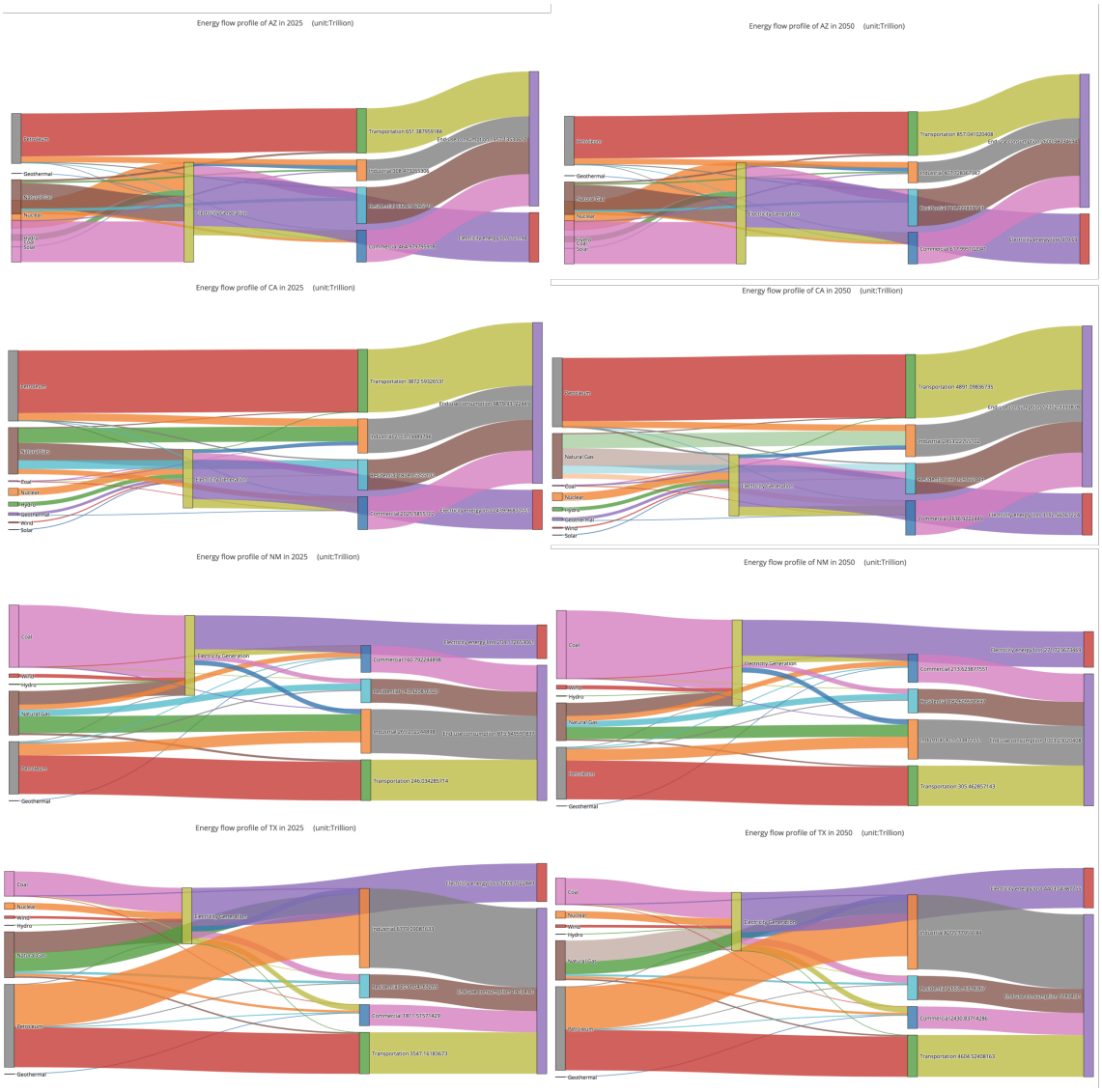

Figure 7: predicted energy flow network

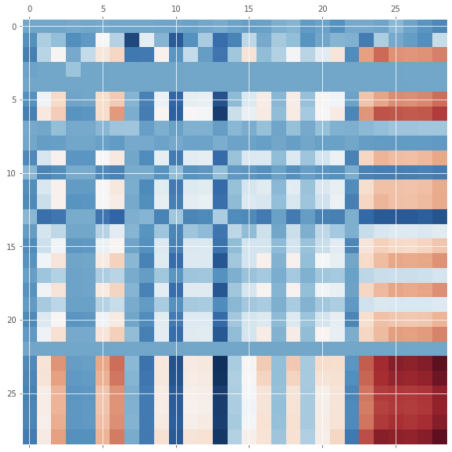

Figure 8: β in the Constrained-Ridge model

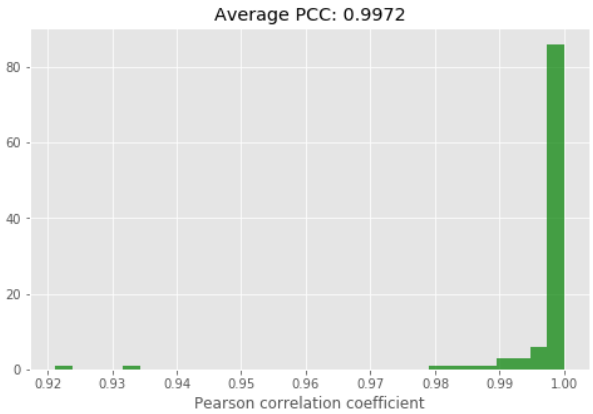

Figure 9: PCC of two datasets

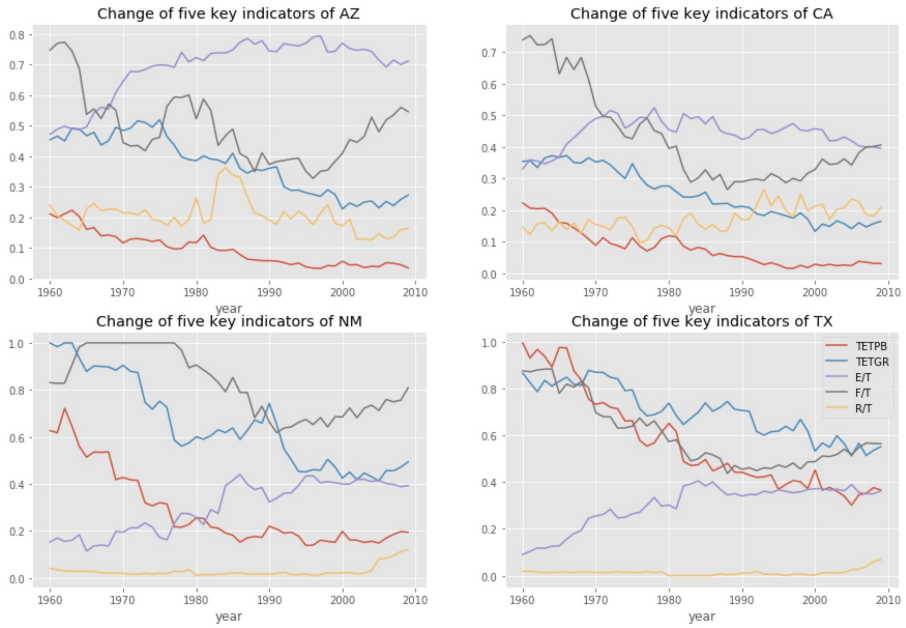

Figure 10: 5 key indicators