For office use only	Team Control Number	For office use only
T1 _____	**82104**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2018
MCM/ICM
Summary Sheet

Energy production

Summary

Energy problem has always been a great concern for the U.S. Nowadays, it is urgent to strengthen cooperation between states, since a single state cannot achieve the optimal goals without others' assistance. The four states of America boarding with Mexico are going to reach an interstate compact on energy policy, and our work is to build mathematical models with historical data to set targets for the four states and to list some actions which can be taken by the governors.

For part I, firstly, in order to create an energy profile of each state, we pick some key variables to summarize the energy usage from the massive historical data. More specifically, we focus on the production and consumption of major conventional and renewable energy sources, and then picture them separately to give a vivid view. Secondly, we use XGBoost algorithm to create an energy index named W, which represents the comprehensive use of energy in each state. The evolution is shown clearly in the diagrams. Then we apply the same algorithm to renewable energy sources to describe their usage. Combined with other factors such as geography, population and climate, we give a further discussion about the energy profile of cleaner, renewable energy usage. Thirdly, we extract the data of major renewable energy sources of the four states in 2009, and applied Topsis method to determine the best energy profile for use of cleaner energy. Finally, we use the Prophet Forecasting Model to predict the energy profile of each state in 2025 and 2050, in which we introduce seasonality (financial crisis) to make the prediction more accurate and convincing.

For part II, we summarize the results of our models and set energy targets for the four states in the interstate compact. Furthermore, we give feasible advice for the governors on how to achieve these goals.

Of course, there are some defects in our model, but it will be more effective and practical with further studies.

Keywords: XGBoost, Topsis, Prophet, Time Series, Seasonality

Contents

1 Introduction

1.1 Background

While industrialization boosts productivity significantly, it also calls for a huge amount of resources, including energy. In fact, any industrialized body would be very fragile without a robust energy supplying system. Thus, among all the factors, energy production and usage has become one of the most vital aspects of an economy.

In the United States, without a centralized energy management department, energy policies are left to state governments to decide, which does make some sense considering the different features of every state. However, we cannot get the global optimum without many close interstate cooperations, but the decentralization of decision-making power makes this goal harder to achieve.

As an attempt to address this issue, the Western Interstate Energy Compact (WIEC) was formed in 1970 by twelve western states, aiming to foster cooperation between these states for the development and management of nuclear energy technologies. Thanks to WIEC, interstate cooperations are now easier with interstate compacts, which are contractual arrangements made between two or more states. These compacts are based on a specific policy issue and put into force either the adoption of a set of standards or the cooperation within these states on a particular regional or national matter.

2 Analysis of the Problem and Our Work

California, Arizona, New Mexico and Texas are four states standing in the transition stage of energy sources. Some of them have good conditions of exploiting cleaner, renewable energy sources, while others are in need of energy supplement or technological assistance. In the interstate compact, the utility of every resource ought to be maximized, which is our fundamental goal.

2.1 Part I

A Plentiful variables related to energy usage in four states are given in the data file, which calls for careful selection to reveal the specific energy profile of these states. Each type of energy is depicted in two aspects, production and consumption. The former might be detailed to the five sectors, that is transportation sector, commercial sector, industrial sector, residential sector and electric power sector, while the latter is mostly depicted in some macroscopic factors. Of course, the absence of certain data is a tough problem when choosing the factors we need.

After analysis, we decide to depict the energy profile in three different ways:

- Consumption in different sectors, varying from 1960 to 2009 in four states.
- Production of several major conventional energy source, varying from 1960 to 2009 in four states.
- Price of these energy source and the total average price, varying from 1960 to 2009 in four states.

B To characterize how energy profile of four states evolve from 1960 to 2009, we need to summarize the variables we choose and try to get an energy index W, representing integrated energy profile. To achieve this, we use XGBoost algorithm, and conduct more data analysis on the results, which turns out to be satisfying to some extent.

In order to inspect the usage of cleaner, renewable energy resources in the four states, we need to detail this part in our model, and analyze several important renewable energy sources further. The different orientation of utilizing renewable energy sources may be caused by the factors specific to a state, like geography, climate and population. Consequently, more information about the four states is required. With diagrams illustrating the similarities and differences between four states, the outcomes become more explicit.

C One of the four states has the best energy profile for using renewable energy resources. An evaluation method is required to decide which one is the best. Topsis algorithm is the easiest and most effective way to do that.

D Prediction is the vital part of the model, directly connecting to the goals of the intestate compact. Time series analysis can do that perfectly. Through the production, we also take some other potential factors into consideration, which will make our prediction more accurate.

2.2 Part II

A By summarizing our work in Part I, we need to further discuss the targets of renewable energy sources in 2025 and 2050, and formally state them in the interstate compact of the four states.

B The goals of the energy usage cannot be achieved without the four states' contribution and cooperation. Therefore, recommendation about several feasible actions is made.

2.3 Part III

We summarize our work in the former two parts and recommend the goals in the compact, and then write a memo to the governors of the four states.

3 Nomenclature

See table 1

4 Energy Profile

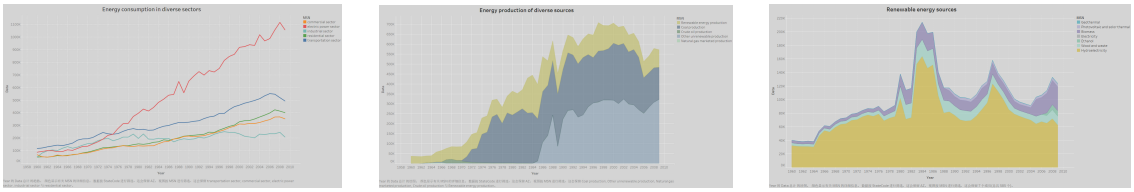
To illustrate the energy profile of each state, we analyze the data given in "ProblemC-Data.xlsx, which contains plentiful data of consumption and production of diverse energy sources. Some data lists are vital information of most frequently-used energy, such as natural gas, coal, motor gasoline, while others may be the renewable sources like

Symbol	Definition
$Obj^{(t)}$	the dependent variable of the objective function
W	the comprehensive variable of energy profile
$RenewP$	the proportion of renewable energy sources in production
$y(t)$	the dependent variable of the prediction equation
$g(t)$	the trend function representing the non-periodic changes
$s(t)$	the trend function representing the periodic changes
A	the data matrix of renewable energy usage
B	the matrix A normalized
w	the weight of each energy resource in Topsis
C	the matrix B with weight on the variables
S^*	the distance of each vector to the optimal solution
S^0	the distance of each vector to the negative ideal solution
f^*	the integrated index of renewable energy usage

Table 1: Nomenclature

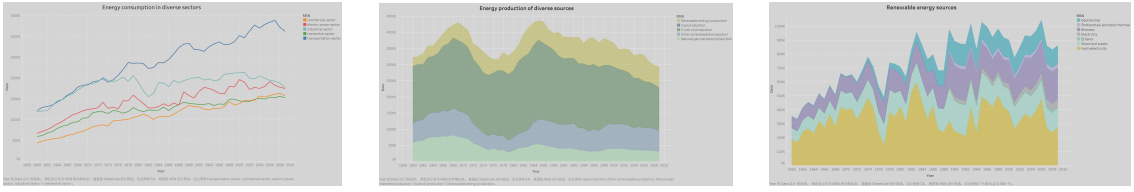
geothermal energy and hydroelectricity. Among these 605 variables collected by the official organization, we choose several of them that is significant to our study, and depict them in line charts or area plots based on the time series. In this case, we want to demonstrate the variation tendency of energy usage in the four states, and briefly compare them for the sake of further studies.

Here we provide the energy profile for each state through data visualization. For each of the four states, figure (a) depicts the energy consumed by various sectors; figure (b) shows the energy produced through different means; figure (c) demonstrates the trend of renewable energy usage.



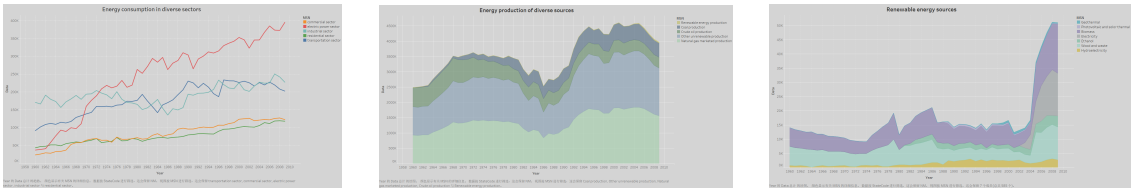
(a) **(b)** **(c)**

Figure 1: The energy profile for Arizona (a) Energy consumption in diverse sectors (b) Energy production in diverse sources (c) Renewable energy usage



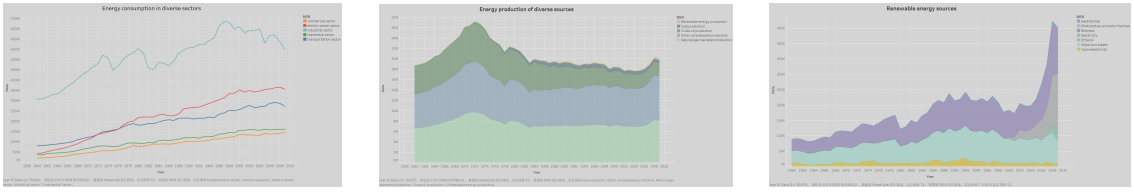
(a) **(b)** **(c)**

Figure 2: The energy profile for California (a) Energy consumption in diverse sectors (b) Energy production in diverse sources (c) Renewable energy usage



(a) (b) (c)

Figure 3: The energy profile for New Mexico (a) Energy consumption in diverse sectors (b) Energy production in diverse sources (c) Renewable energy usage



(a) (b) (c)

Figure 4: The energy profile for Texas (a) Energy consumption in diverse sectors (b) Energy production in diverse sources (c) Renewable energy usage

5 XGBoost Modeling

Our work is to find the relationship between types of main energy and y, which represents the energy consumption of the state. Here we define y as "total energy consumption per capita", and we hope to find the importance of each energy by fitting y through the model. Initially, we try to use liner-regression model such as general linear model and ridge regression to fit the y and get the weights of every feature, but linear model doesnt fit very well and the error is too large. Then it occurs to us that we can use XGBoost Algorithm, which can fit y well and output the importance of every feature.

5.1 Introduction to XGBoost Algorithm

XGBoost is an improvement of GBDT, which has been popular for many years. The biggest difference between XGBoost and GBDT is its definition of the objective function (1):

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant \tag{1}$$

where l is the loss function, Ω is a regularization term which consists of two regularization terms, L_1 and L_2, and $constant$ is a constant term.

Applying Taylor's theorem (2), we can get an approximation of the objective function (3), which in this case only depends on the first and second derivative.

$$f(x + \Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2} f''\Delta x^2 \tag{2}$$

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \right] + \Omega(f_t) + constant \tag{3}$$

where

$$g_i = \frac{\partial l(y_i, \hat{y}^{(t-1)})}{\partial \hat{y}^{(t-1)}}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y}^{(t-1)})}{\left[\partial \hat{y}^{(t-1)}\right]^2} \tag{4}$$

Let's do a little refinement to the definition of f. Like GBDT, we will have some decision trees., but here the tree is split into the structure part Q and the leaf weight part w. The structure function Q maps the input to the index number of the leaf, and W gives the leaf score corresponding to each of the cable quotes. Defining this complexity contains the number of nodes in a tree, and the L_2 modulus square of the output score on each tree leaf node.

Using this new definition, we can rewrite the objective function as follows, in which I_j is defined as the set of samples on each leaf.

$$I_j = \{i|q(x_i) = j\} \tag{5}$$

Here is the objective function, where g is the first derivative, and g is the second derivative.

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{6}$$

$$= \sum_{i=1}^{n} \left[g_i \omega_q(x_i) + \frac{1}{2} h_i \omega_q^2(x_i) \right] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} \omega_j^2 \tag{7}$$

$$= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j^2 \right] + \gamma T \tag{8}$$

This goal contains T independent single variable quadratic functions. We define

$$G_j = \sum_{i \in I_j} g_i \tag{9}$$

$$H_j = \sum_{i \in I_j} h_i \tag{10}$$

Finally we can get

$$Obj^{(t)} = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j^2 \right] + \gamma T \tag{11}$$

$$= \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2 \right] + \gamma T \tag{12}$$

Now, take the derivative of w_j equal to 0,

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{13}$$

and then add the optimal solution to the formula,

$$Obj = -\frac{1}{2}\sum_{j=1} T\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{14}$$

5.2 Advantages and Parameter Adjustment

Due to its boosting thought, XGBoost can fit the target accurately, and thanks to its bagging thought of feature selection, it can prevent overfitting to a certain extent. Also, it helps us to calculate the weights of features. In addition to the difference of objective function, XGBoost has a C++ implementation of Gradient Boosting Machine. The biggest feature of XGBoost is that it can automatically use CPU's multithreading to parallelism, which means performing many calculations at the same time. This improves the accuracy of the algorithm. More importantly, XGBoost is a boosting algorithm, but it select the features randomly. Consequently, it can avoid overfitting and tell us the importance of every features by multiple iterations. Due to its high accuracy, the weights it gives are more efficient than other algorithm such as linear-regression, simple neural network and so on.

However, the parameters of XGBoost are hard to adjust, because it have so many parameters, for example, eta, tree max depth, to mention just a few. We take 5000 iteration and the algorithm can finally converge, which also can produce less error. To avoid overfitting, we choose a small number for tree max depth. As for the percentage of subsample, we set 70 like other sets, and the results show that these parameters work.

5.3 Modeling Process

We select a bunch of features from the 605 MSNs, such as $CLTCB$ and $HYTCB$, from 1960 to 2009. They form this a 50×14 matrix, it contains the most important information of energy use about the state from 1960 to 2009, including both unrenewable and renewable energy. To get the weights, we define $TETPB$ as y. According to XGBoost Algorithm, we create a model between different energy use and total energy consumption per capita. The weights here can be explained as the influence of different energy types on energy consumption per capita. In view of our target, the weights of XGBoost can be used to measure the importance of the energy.

After using XGBoost, we can get the importance of our features(w_j), and we calculate the percentage of the importance, which we define as the features' weight.

Next, we define W as $\sum_{i=1}^{14} w_i x_i$, which represents the state's energy use in that year. After getting w_i, we just need to add x into the formula and get W of every year.

5.4 Analysis

From the result we find that, Arizona's total energy consumption per capita is mainly influenced by coal production, coal total consumption and residual fuel oil total consumption, whereas that of New Mexico is mainly influenced by coal production, natural gas total consumption and residual fuel oil consumption. It's easy to see that these two states

are similar, which both focus on coal and residual fuel oil. Nonetheless, Arizona's coal total consumption is more important than natural gas total consumption. Thus, compared to Arizona, New Mexico needs more natural gas but not coal. On the other hand, Arizona uses more coal but also produce more coal. When it comes to California, its total energy consumption per capita is mainly influenced by coal total consumption, fossil fuel total consumption and crude oil production. Finally, Texas is mainly influenced by coal production, natural gas total consumption and natural gas marketed production. Although they all influenced by coal, California is influenced by using coal but Texas is affected by producing lots of coal. This means Texas and New Mexico can output coal to California and Arizona. It's interesting to observe that Texas needs a lot of natural gas, but they also produce a large amount of natural gas, which can be seen in its natural gas marketed production. California uses coal and fossil fuel but produce more crude oil, yet it can use crude oil to get more energy.

Adding w to the formula $W = Xw$, we calculate the W of 4 states from 1960 to 2009 (5).

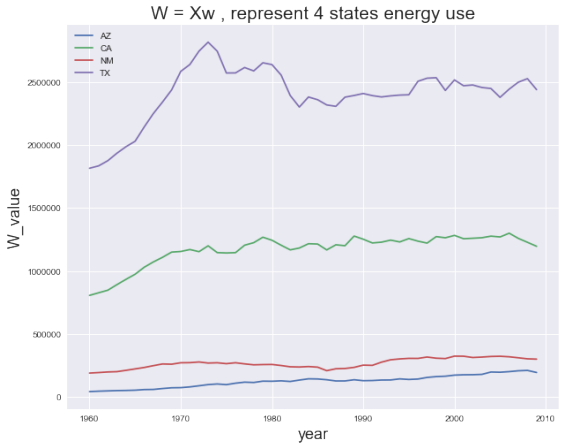


Figure 5: Energy use of four states

W is the comprehensive use of energy, which is the weighted sum of some important unrenewable energy and renewable energy total consumption, so it's effective to define W as the energy profile. The picture 5 depicts the change of the energy profile about 4 states from 1960 to 2009. From this picture, we can see that Texas uses the most energy and California ranks second. New Mexico and Arizona use less energy and have no obvious difference. Texas's energy use changes violently but others' changes more smoothly.

So we can see that, in terms of unrenewable energy, no matter production or consumption, coal influences the most on total energy consumption per capita, followed by fuel oil, fossil fuel and natural gas. These energy are also very important, but as for other energy like motor gasoline, these states can use more renewable energy instead of these insignificant energy. On the one hand renewable energy is cleaner, but on the other hand these energy influence less on total energy consumption per capita, so it's a good idea to increase the usage of renewable energy.

Closer observation reveals that some energy sources important to specific states, are not vital to the others however. For example, asphalt and road oil total consumption ranks low in Texas but rank high in Arizona. Fossil fuel total consumption ranks low in

New Mexico but ranks high in California. Consequently, the four states can exchange the energy which is unimportant to them but vital to others according to the ranking picture above. However, generally speaking, these states all need to increase renewable energy use to a certain extent.

6 Predictions

Based on our observation, the data appear to carry some seasonality. In order to model this characteristic properly, we use the Prophet forecasting model.

6.1 Methodology

The basic idea is shown is the following equation (15):

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{15}$$

where $g(t)$, $s(t)$ are trend functions representing the non-periodic and periodic changes respectively, $h(t)$ represents the effects of irregularly-scheduled events, and ϵ_t is the error term.

For $g(t)$, a piece-wise constant rate of growth provides a parsimonious and often useful model. Here the trend model is (16)

$$g(t) = \left(k + a(t)^T\delta\right)t + \left(m + a(t)^T\gamma\right) \tag{16}$$

where k is the growth rate, δ has the rate adjustments, m is the offset parameter, and γ_j is adjusted using automatic change-point selection to make the function continuous.

With the help of Fourier series (17), we can construct a flexible model of periodic effects, or $s(t)$.

$$s(t) = \sum_{n=1}^{N}\left(a_n\cos\left(\frac{2\pi nt}{P}\right) + b_n\sin\left(\frac{2\pi nt}{P}\right)\right) \tag{17}$$

When it comes to $h(t)$, we are aware that events without a regular pattern such as El Niño and La Niña may have an impact on the data. However, due to the inherent unpredictability of future anomalies, we decide not to include them to our model.

6.2 Application

We use a standard Prophet forecasting model to model the trend of W.

As mentioned before, we detect a wave with a period of 25 years, which is presumably a result of the Kuznets swing, a medium-range economic wave with a period of 1525 years connected with demographic processes. Taking this into consideration, we add a 25-year seasonality into the model.

Here is the result. The blue line is the prediction of W until the year 2050, and the light blue part marks its 95% confidence interval. The energy profile of each state for 2025 and 2050 is listed below:

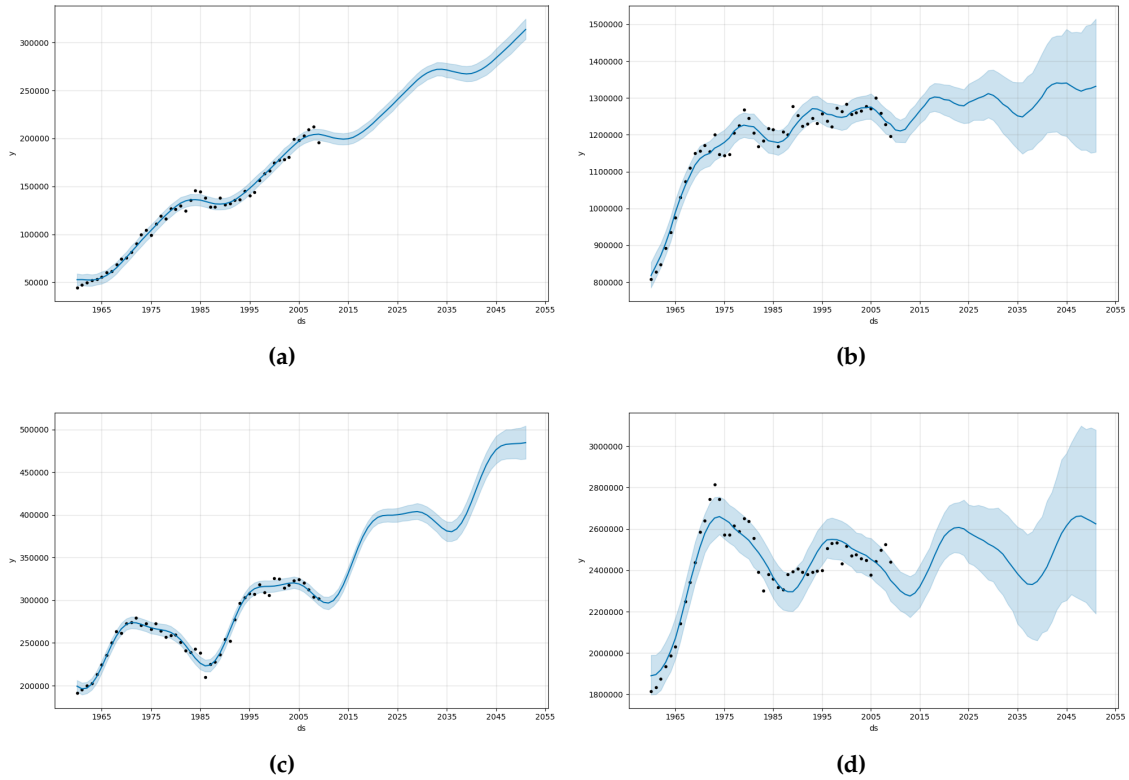


Figure 6: The prediction of W in (a) AZ (b) CA (c) NM (d) TX

	Arizona	California	New Mexico	Texas
2025	245620.8996	1293635.354	400980.219	2568748.067
2050	313629.0864	1331776.615	484511.6994	2624274.833

Table 2: The prediction of W for 2025 and 2050

6.3 Validating the Model

Time series cross validation is used to measure forecast error using historical data. We carried out a simulated historical forecast on W, where the model was fit to an initial history of thirty years, and a forecast was made on a one year horizon.

Here is the result.

Arizona	California	New Mexico	Texas
48.65%	54.05%	62.16%	32.43%

Table 3: Rate of correct predictions

While it doesn't seem very impressive, note that we've merely used the data of the first 30 years, which only contains a single 25-year cycle. However, when making the prediction, all 50 data points are used, and the two complete 25-year cycle contained would hopefully make a big difference.

7 Usage of Renewable Energy Sources

We choose main renewable energy sources of four states and calculate what percentage each source takes up in the total. It is worth mentioning that we get the production of fuel ethanol through a simple subtraction (18), since there is no direct data concerning that.

$$\text{Fuel Ethanol} = REPRB - ROPRB \tag{18}$$

where $REPRB$ and $ROPRB$ stands for "Renewable energy production" and "Renewable energy production, other than fuel ethanol", respectively.

To compare the usage of renewable energy sources, we introduce several more variables. $RenewP$ is a variable representing the proportion of renewable energy sources in the total energy production.

$$RenewP = \frac{REPRB}{TEPRB} \tag{19}$$

The results are shown in the following tables.

	Arizona	California	New Mexico	Texas
Hydroelectricity	62730.88 (50.59%)	272187.23 (31.65%)	2644.59 (5.17%)	10039.69 (2.22%)
Geothermal energy	329.09 (0.26%)	127461.12 (14.82%)	317.06 (0.62%)	2057.02 (0.45%)
Photovoltaic and solar energy	4732.12 (3.81%)	31397.01 (3.65%)	282.52 (0.55%)	819.76 (0.18%)
Electricity produced from wind	288.35 (0.23%)	56996.57 (6.62%)	15095.96 (29.55%)	195454.76 (43.24%)
Biomass	35412.81 (28.56%)	224662.52 (26.13%)	17295.22 (33.85%)	148263.57 (32.80%)
Wood and waste	12867.29 (10.37%)	140159.17 (16.30%)	11633.20 (22.77%)	72108.45 (15.95%)
Fuel ethanol	7623.61 (6.14%)	6861.25 (0.79%)	3811.80 (7.462%)	23217.36 (5.13%)

Table 4: Major renewable energy resources in four states

	Arizona	California	New Mexico	Texas
Proportion of renewable energy resources	14.17%	24.11%	1.24%	2.35%

As can be seen from the tables, each state has its own preference of using renewable energy resources, which may be affected by many factors like climate, geography, etc.

Here we give further discussion.

- **Arizona**
  Arizona has relatively low demand for energy use, however, in which renewable energy resources takes a large proportion (shown in the energy profile for AZ), about 14.18% in production, rankly only second to California. Thus, we come to

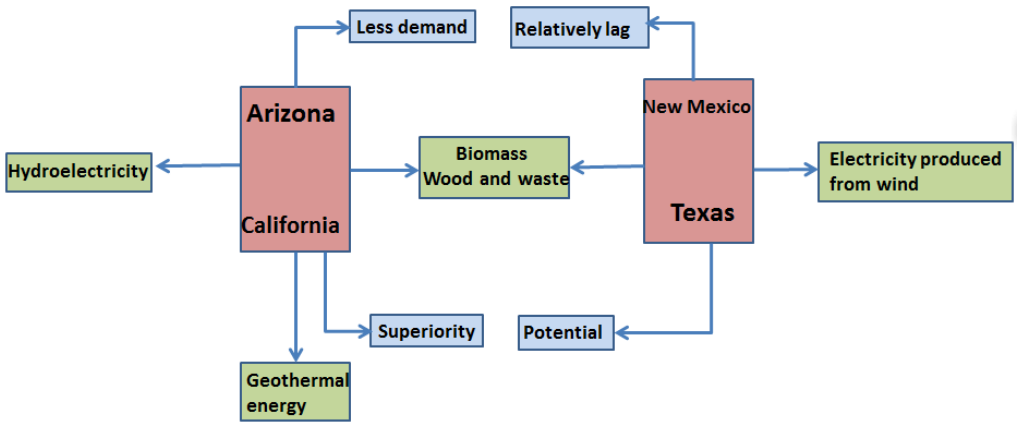


Figure 7: Energy features of the four states

conclusion that Arizona has a good basis of exploiting and utilizing renewable energy resources.

- **California**
  California has both the largest gross and the biggest proportion (24.11%) of renewable energy resources usage, which indicates that California has already came to a relatively mature stage.

- **New Mexico**
  Undoubtedly, the usage of renewable energy sources in New Mexico is not satisfying. With great energy demand as California, New Mexico rarely make use of renewable ones. Thus a "cleaner" energy profile is the most essential target for New Mexico.

- **Texas**
  Compared with other three states, Texas has enormous energy production and consumption. With increasing proportion of renewable energy, Texas has the greatest potential of utilizing cleaner energy sources.

7.1 Other Factors

Now, we take geography, population, and climate into consideration.

	Arizona	California	New Mexico	Texas
Population(million)	6.93	39.25	2.08	27.86

Table 5: Population of the four states

California has the largest population and the most advanced technology to utilize renewable energy resources. Due to its location on the Pacific's "ring of fire" and because of tectonic plate conjunctions, California contains the largest amount of geothermal electric generation capacity in the United States. As a result, California is the only state that make good use of geothermal energy.

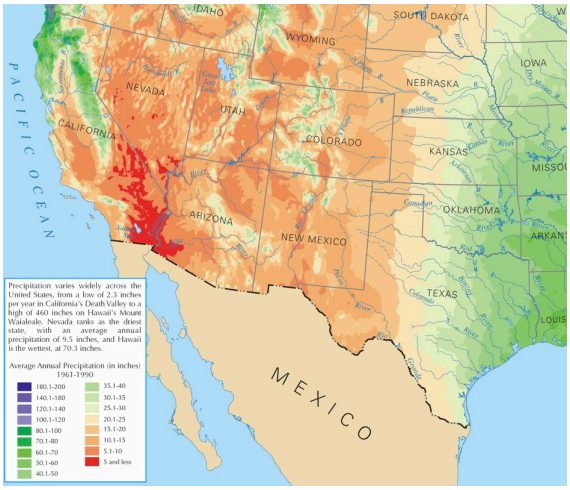


Figure 8: A geographic map of the U.S.

According to the geography, complex river networks are located in California and Arizona, making hydroelectricity a main source of renewable energy. On the contrary, without many rivers in New Mexico, electricity produced from wind becomes the top priority due to its frequent gale. However, as mentioned before, there is still a long way for New Mexico, since its production is still too low.

Last, with largest energy production and consumption, Texas also has every superiority to utilize varieties of renewable energy resources. Up to now, wind power in Texas has achieved good results, standing at the top of four states in amount. However, from the topographic map, we can see the river networks in Texas form favorable conditions for hydroelectricity. Therefore, we want to hightlight that Texas has the greatest potential in usage of renewable energy sources.

7.2 Which One is the Best? California

Considering the utility of main renewable energy sources and its proportion it the gross, we use TPOSIS method to determine the best energy profile.

Firstly, we normalize the data matrix A to get the decision matrix B. The variables chosen are all benefit factors, so we adopt the same method below.

$$b_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{4} a_{ij}^2}}, i = 1, 2, \ldots, 4; j = 1, 2, \ldots, 8 \tag{20}$$

Next, we form the normalized matrix C with certain weight for each variable. The weight is shown in the bubble diagram 9.

$$w = [0.14 \quad 0.06 \quad 0.03 \quad 0.16 \quad 0.18 \quad 0.10 \quad 0.03 \quad 0.30] \tag{21}$$

$$C_{ij} = w_i \times b_{ij}, i = 1, 2, \ldots, 4 \tag{22}$$

Then we get the optimal solution and the negative ideal solution

$$C_i^* = \max_i C_{ij}, j = 1, 2, \ldots, 8 \tag{23}$$

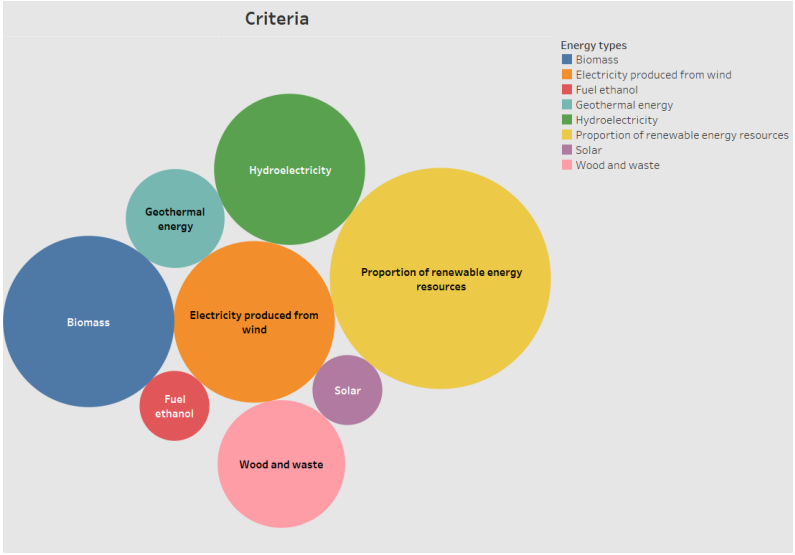


Figure 9: Criteria

$$C_i^0 = \min_i C_{ij}, j = 1, 2, \ldots, 8 \tag{24}$$

Finally, we calculate the distance of each data list to the optimal solution and the negative ideal solution.

Distance to the optimal solution:

$$S_i^* = \sqrt{\sum_{j=1}^{n} (C_{ij} - C_j^*)^2}, i = 1, 2, \ldots, 4 \tag{25}$$

Distance to the negative ideal solution:

$$S_i^0 = \sqrt{\sum_{j=1}^{n} (C_{ij} - C_j^0)^2}, i = 1, 2, \ldots, 4 \tag{26}$$

Integrated index:

$$f_i^* = \frac{S_i^0}{S_i^0 + S_i^*}, i = 1, 2, \ldots, 4 \tag{27}$$

	Arizona	California	New Mexico	Texas
s^*	0.2635	0.1102	0.3534	0.2771
s^0	0.1495	0.3262	0.0116	0.1817
f	0.3619	0.7475	0.0318	0.3961
rank	3	1	4	2

Table 6: The results of Topsis

The result of our model clearly shows that **California** has the best energy profile.

8 Prediction of Renewable Energy Usage Targets

Our prediction of the $RETCB$ of 2025 and 2050 is shown in figure 10 . The same technique is used as in the prediction of W.

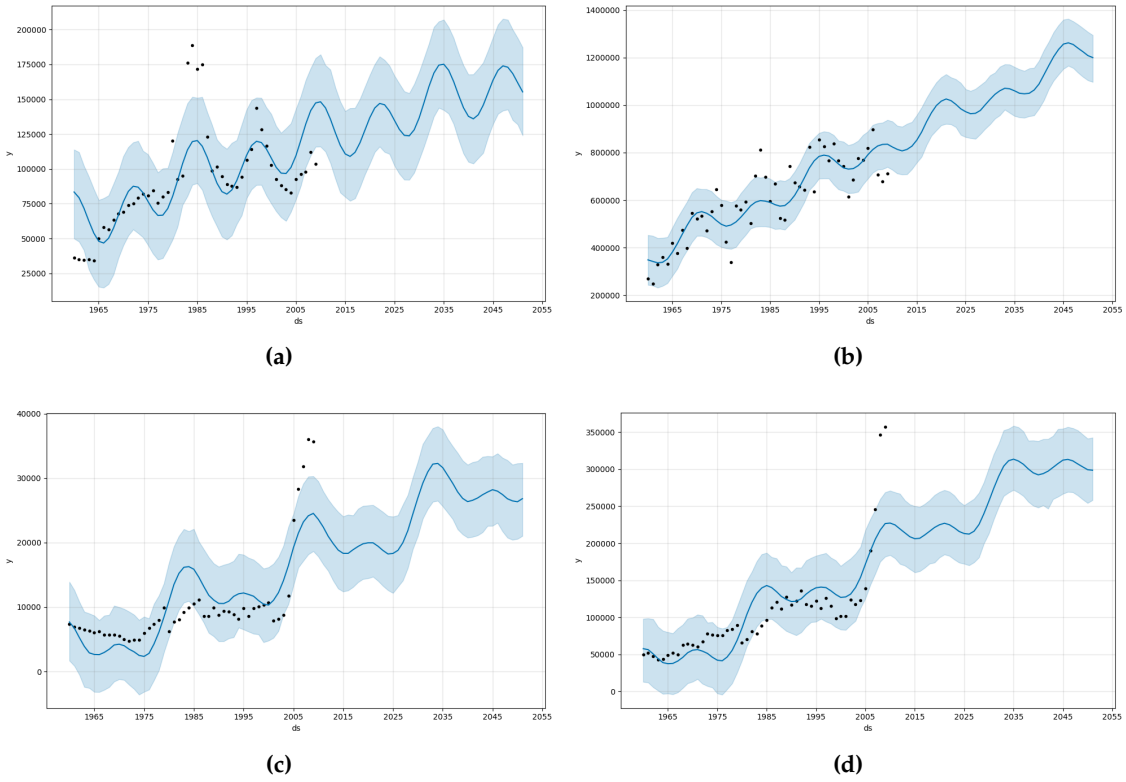


Figure 10: The prediction of $RETCB$ in (a) Arizona (b) California (c) New Mexico (d) Texas

As before, the blue line is the prediction of W until the year 2050, and the light blue part marks its 95% confidence interval. The prediction of renewable energy usage targets of each state for 2025 and 2050 is listed in table 7:

	Arizona	California	New Mexico	Texas
2025	128327.9859	963652.5339	18797.11625	212369.961
2050	155304.4328	1199923.714	26812.56441	298504.3516

Table 7: The prediction of $RETCB$ for 2025 and 2050

8.1 Validating the Model

Table 8 shows the result of cross validation, using the same configuration as before.

Arizona	California	New Mexico	Texas
29.73%	70.27%	29.73%	43.24%

Table 8: Rate of correct predictions

Again, for the same reason, we believe our prediction is better than what's shown in table 7.

9 Targets and Actions in the Compact

9.1 Goals

Based on our comparison between the energy profile of each state, our criteria for the best profile, and our predictions, we mainly discuss the goals of renewable energy usage in three aspects.

1. Make energy profile of every state cleaner and more renewable According to our results, the proportion of renewable energy usage is not ideal in all the four states, even California which has the best energy profile. Therefore, more utilization of cleaner energy should be included in the interstate compact. As each state has its own preponderance of using certain renewable energy, this goal coordinate with the following ones to some extent. The targets are listed below:
   (a) Increase the percentage that renewable energy takes in each state, California better more than 30%, while others have dramatic improvements, especially New Mexico.
   (b) Substitute the most polluted energy resources gradually.
   (c) Exploit the great potential cleaner energy resources in Texas.
2. Technological assistance from superior state to the ones lag behind Among the four states, California is the most developed one with the largest population, of which the technology for utilizing renewable energy resources stands at the top. Therefore, California ought to contribute more to improve the skills of exploiting cleaner energy of other three states, especially Texas and New Mexico.
3. Energy redistribution to maximize its utility To maximize the utility of energy usage, redistribution of energy sources is the best way, because some resources, not important to specific states, might be vital to other states. Therefore, the transportation of some certain energy from the states, that it ranks low, to those with high rank is a beneficial method. This goal can be achieved through the advice given below.

9.2 Recommended Actions

9.2.1 Arizona

1. Exploit geothermal energy like California.
2. Make good use of photovoltaic and solar energy.
3. Export motor gasoline to Texas and California.
4. Develop renewable energy to replace motor gasoline.

9.2.2 Texas

1. Exploit hydroelectricity and wind power.
2. Transmit electricity to other states in order to increase the demand.
3. Export asphalt and road oil to Arizona and New Mexico.
4. Decrease crude oil production and develop renewable energy.

9.2.3 California

1. Provide technical assistance to other states, especially New Mexico and Texas.
2. Take its advantage of rich geothermal energy by making more use of it.
3. Decrease the consumption of environmentally unfriendly energy, and import cleaner energy from Texas and Arizona instead.
4. Exploit photovoltaic and solar energy.
5. Export residual fuel oil to other states, especially Arizona.
6. Decrease coal production and develop renewable energy.

9.2.4 New Mexico

1. Change the energy structure gradually yet completely, by exporting more resources and importing technology of utilizing cleaner energy.
2. Exploit wind energy.
3. Export fossil fuel and motor gasoline to California
4. Export motor gasoline to Texas.
5. Develop renewable energy to replace motor gasoline and fuel fossil fuel.

10 Memorandum

To Governors

From Team # 82104

Date February 13, 2018

Subject Energy Production

As of 2009, Arizona and California is doing fairly well in terms of energy profile, whereas New Mexico and Texas need some work. However, all of the four states depends greatly on fossil energy, which is not very clean.

According to our model, if no policy changes are made, the energy efficiency of all states will keep growing, though in a very slow rate and no without some serious fluctuating. Coincidentally, the same result applies to the usage of renewable energy. This is not bad, but we can do better with an energy compact.

we recommend the following goals for the compact.

Take Your Advantages

States are suppose to use good of every natural resource they own. For example, Texas may develop hydroelectricity and wind power more vigorously, since it's windy and full of river.

Exchange Material and/or Technology

Some states are rich in natural resources but leak the technology to exploit it. In this case, states with these technologies could consider providing technical assistance to them in exchange for their resources. This would be a win-win deal for both side.

Think Ahead

In the short run, conventional energy is more economical, since it's cheaper and yields more energy. However, renewable energy can provide huge strategic advantages in the long run, most notably a cleaner environment and an energy source which will never be exhausted.

References

[1] Ümran engül, Miraç Eren, Seyedhadi Eslamian Shiraz, Volkan Gezder, Ahmet Bilal engül, Fuzzy TOPSIS method for ranking renewable energy supply systems in Turkey, Renewable Energy, Volume 75, 2015, Pages 617-625

[2] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016

[3] Crude oil price forecasting using XGBoost,Published in: Computer Science and Engineering (UBMK), 2017 International Conference on, Date of Conference: 5-8 Oct. 2017

[4] Taylor SJ, Letham B. (2017) Forecasting at scale. PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2

[5] `http://pbpython.com/prophet-overview.html`

[6] `https://en.wikipedia.org/wiki/Kuznets_swing`

Appendices

Appendix A MATLAB Source Code

matlab.m

```
Matrix=[62730.8872,272187.2346,2644.59888,10039.69525;
    329.09862,127461.1205,317.06702,2057.02446;
    4732.12723,31397.00543,282.52701,819.76122;
    288.3592,56996.57722,15095.96768,195454.7653;
    35412.81309,224662.5224,17295.22312,148263.5744;
    12867.29943,140159.1761,11633.20739,72108.45032];
REPRB=[88571.38442,635062.3653,33785.17435,303697.0626];
ROPRB=[80947.77168,628201.1138,29973.36798,280479.6965];
TEPRB=[570994.0459,2605311.838,2412219.049,11914996.72];
Matrix=[Matrix;REPRB-ROPRB;REPRB./TEPRB];
Per=zeros(8,4);
for i=1:4
    Per(:,i)=Matrix(:,i)/sum(Matrix(:,i));
end
Matrix=Matrix';
[m,n]=size(Matrix);
for j=1:n
    Mend(:,j)=Matrix(:,j)/norm(Matrix(:,j));
end
w=[0.14,0.06,0.03,0.16,0.18,0.10,0.03,0.30];
c=Mend.*repmat(w,m,1);
cmax=max(c);
cmin=min(c);
for i=1:m
    smax(i)=norm(c(i,:)-cmax);
    smin(i)=norm(c(i,:)-cmin);
end
```

```
smax,smin
f=smin./(smax+smin)
[sf,ind]=sort(f,'descend')
```

Appendix B Python Source Code

model.py

```
# coding: utf-8

# In[1]:

import pandas as pd
import numpy as np

data_az = pd.read_csv('...\ProblemCData_AZ.csv')
data_ca = pd.read_csv('...\ProblemCData_CA.csv')
data_nm = pd.read_csv('...\ProblemCData_NM.csv')
data_tx = pd.read_csv('...\ProblemCData_TX.csv')

feature1 = ['CLPRB','CLTCB','FFTCB','MGTCB','NGMPB','NGTCB',
    'P1TCB','PAPRB','REPRB','RETCB','RFTCB']
feature2 = ['HYTCB','GETCB','SOTCB','WYTCB','BMTCB','WWTCB']
#Coal production.
#Coal total consumption.
#Fossil fuelS, total consumption.
#Geothermal energy total consumption.
#Motor gasoline total consumption.
#Natural gas marketed production.
#Natural gas total consumption
#Asphalt and road oil
#Crude oil production
#Renewable energy production.
#Renewable energy total consumption.
#Residual fuel oil total consumption.
def create_x(data,k):
    X = pd.DataFrame(index = range(1960,2010))
    if k==1:
        for i in feature1:
            X[i] = data[data['MSN']==i]['Data'].values
        #TETPB is total consumption per capita
        y = data[data['MSN']=='TETPB']['Data'].values
    elif k==2:
        for i in feature2:
            X[i] = data[data['MSN']==i]['Data'].values
        y = data[data['MSN']=='RETCB']['Data'].values

    return X,y



# In[2]:

x1_az,y1_az = create_x(data_az,1)
x1_ca,y1_ca = create_x(data_ca,1)
x1_nm,y1_nm = create_x(data_nm,1)
x1_tx,y1_tx = create_x(data_tx,1)
```

```
x2_az,y2_az = create_x(data_az,2)
x2_ca,y2_ca = create_x(data_ca,2)
x2_nm,y2_nm = create_x(data_nm,2)
x2_tx,y2_tx = create_x(data_tx,2)


# In[3]:

import xgboost as xgb
from xgboost.sklearn import XGBClassifier
import operator

def xgbTraining(X,y):
    X_ = (X-X.mean())/X.std()
    y_ = (y-y.mean())/y.std()

    dtrain=xgb.DMatrix(X_,y_)
    dtest=xgb.DMatrix(X_)
    param = {}
    param['eta'] = 0.05
    param['max_depth'] = 4
    param['mmin_child_weight'] = 3
    param['subsample'] = 0.7
    param['colsample_bytree'] = 0.7
    param['silent'] = 1

    alg = xgb.train(param,dtrain,5000)
    Y = alg.predict(dtest)
    Y = Y*y.std() + y.mean()

    importance = alg.get_fscore()
    importance = sorted(importance.items(), key=operator.itemgetter(1))

    df = pd.DataFrame(importance, columns=['feature', 'fscore'])
    df['fscore'] = df['fscore'] / df['fscore'].sum()
    return Y,df


# In[75]:

Y1_az,df1_az = xgbTraining(x1_az,y1_az)
Y1_ca,df1_ca = xgbTraining(x1_ca,y1_ca)
Y1_nm,df1_nm = xgbTraining(x1_nm,y1_nm)
Y1_tx,df1_tx = xgbTraining(x1_tx,y1_tx)

Y2_az,df2_az = xgbTraining(x2_az,y2_az)
Y2_ca,df2_ca = xgbTraining(x2_ca,y2_ca)
Y2_nm,df2_nm = xgbTraining(x2_nm,y2_nm)
Y2_tx,df2_tx = xgbTraining(x2_tx,y2_tx)


# In[76]:

import seaborn as sns
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')

def plot(df_az,df_ca,df_nm,df_tx):
    plt.figure(figsize=(20,20))
    plt.subplot(221)
```

```
    plt.barh(range(df_az.shape[0]),df_az['fscore'],facecolor='#B8860B')
    plt.yticks(range(df_az.shape[0]),df_az['feature'],fontsize=18)
    plt.title('AZ feature importance',fontsize=21)
    plt.subplot(222)
    plt.barh(range(df_ca.shape[0]),df_ca['fscore'],facecolor='#8B0000')
    plt.yticks(range(df_ca.shape[0]),df_ca['feature'],fontsize=18)
    plt.title('CA feature importance',fontsize=21)
    plt.subplot(223)
    plt.barh(range(df_nm.shape[0]),df_nm['fscore'],facecolor='#8FBC8F')
    plt.yticks(range(df_nm.shape[0]),df_nm['feature'],fontsize=18)
    plt.title('NM feature importance',fontsize=21)
    plt.subplot(224)
    plt.barh(range(df_tx.shape[0]),df_tx['fscore'])
    plt.yticks(range(df_tx.shape[0]),df_tx['feature'],fontsize=18)
    plt.title('TX feature importance',fontsize=21)
    plt.show()


# In[77]:

plot(df1_az,df1_ca,df1_nm,df1_tx)


# In[78]:

plot(df2_az,df2_ca,df2_nm,df2_tx)


# In[79]:

def cal_w(df,X):
    w = df['fscore'].values
    W = np.dot(X,w)
    return W

w1_az = cal_w(df1_az,x1_az)
w1_ca = cal_w(df1_ca,x1_ca)
w1_nm = cal_w(df1_nm,x1_nm)
w1_tx = cal_w(df1_tx,x1_tx)

plt.figure(figsize=(10,8))
plt.plot(x1_az.index,w1_az,label = 'AZ')
plt.plot(x1_ca.index,w1_ca,label = 'CA')
plt.plot(x1_nm.index,w1_nm,label = 'NM')
plt.plot(x1_tx.index,w1_tx,label = 'TX')

plt.title('W = Xw , represent 4 states energy use',fontsize=21 )
plt.legend(loc='best')
plt.xlabel('year',fontsize=18)
plt.ylabel('W_value',fontsize=18)
plt.show()


# In[80]:

w2_az = cal_w(df2_az,x2_az)
w2_ca = cal_w(df2_ca,x2_ca)
w2_nm = cal_w(df2_nm,x2_nm)
w2_tx = cal_w(df2_tx,x2_tx)

plt.figure(figsize=(10,8))
```

```
plt.plot(x2_az.index,w2_az,label = 'AZ')
plt.plot(x2_ca.index,w2_ca,label = 'CA')
plt.plot(x2_nm.index,w2_nm,label = 'NM')
plt.plot(x2_tx.index,w2_tx,label = 'TX')

plt.title('W = Xw , represent 4 states renewable energy use',fontsize=21 )
plt.legend(loc='best')
plt.xlabel('year',fontsize=18)
plt.ylabel('W_value',fontsize=18)
plt.show()


# In[73]:

print('In 2009 these four states renewable energy use:')
print('Arizona:%d \nCalifornia:%d \nNew Mexico:%d \nTexas:%d' % \
    (w2_az[-1],w2_ca[-1],w2_nm[-1],w2_tx[-1]))
```

W.py

```
from datetime import date
import pandas as pd
import numpy as np
from fbprophet import Prophet
from fbprophet.diagnostics import cross_validation

results = {}
results_rate = {}

for state in ['az', 'ca', 'nm', 'tx']:
    df = pd.read_csv('W_data/w_{0}.csv'.format(state))
    df['ds'] = [str(y)+'-1-1' for y in df['ds']]

    m = Prophet()
    m.add_seasonality(
        name='financial_crisis',
        period=25*365.25,
        fourier_order=2)
    m.fit(df)
    future = m.make_future_dataframe(periods=42, freq='Y')
    forecast = m.predict(future)

    df_cv = cross_validation(m,
            horizon='365.25 days',
            initial='10958 days')
    print('#'*79)
    result = [row['yhat_lower'] <= row['y'] <= row['yhat_upper'] \
            for index, row in df_cv.iterrows()]
    results[state] = result
    results_rate[state] = result.count(True) / len(result)

    fig = m.plot(forecast)
    fig.savefig('W_out/fig/prediction_{0}.png'.format(state))
    forecast['ds'] = [str(y)[0:4] for y in forecast['ds']]
    forecast.to_csv('W_out/csv/prediction_{0}.csv'.format(state))


for state in ['az', 'ca', 'nm', 'tx']:
    print('{0}: {1}'.format(state, results_rate[state]))
```

MSN.py

```
import os
from datetime import date
import pandas as pd
import numpy as np
from fbprophet import Prophet
from fbprophet.diagnostics import cross_validation


for root, dirs, files in os.walk('MSN_data'):
    msns = files

for msn in msns:
    df = pd.read_csv('MSN_data/' + msn)
    ds = df['ds']
    df['ds'] = [str(y)+'-1-1' for y in df['ds']]

    m = Prophet()
    m.add_seasonality(
        name='financial_crisis',
        period=25*365.25,
        fourier_order=3)
    m.fit(df)
    future = m.make_future_dataframe(periods=42, freq='Y')
    forecast = m.predict(future)

    fig = m.plot(forecast)
    fig.savefig('MSN_out/fig/prediction_{0}.png'.format(msn.split('.')[0]))
    forecast['ds'] = [str(y)[0:4] for y in forecast['ds']]
    forecast.to_csv('MSN_out/csv/prediction_{0}'.format(msn))
```

RETCB.py

```
from datetime import date
import pandas as pd
import numpy as np
from fbprophet import Prophet
from fbprophet.diagnostics import cross_validation

results = {}
results_rate = {}

for state in ['az', 'ca', 'nm', 'tx']:
    df = pd.read_csv('RETCB_data/RETCB_{0}.csv'.format(state))
    df['ds'] = [str(y)+'-1-1' for y in df['ds']]

    m = Prophet()
    m.add_seasonality(
        name='financial_crisis',
        period=25*365.25,
        fourier_order=2)
    m.fit(df)
    future = m.make_future_dataframe(periods=42, freq='Y')
    forecast = m.predict(future)

    df_cv = cross_validation(m,
            horizon='365.25 days',
            initial='10958 days')
    print('#'*79)
    result = [row['yhat_lower'] <= row['y'] <= row['yhat_upper'] \
```

```
            for index, row in df_cv.iterrows()]
    results[state] = result
    results_rate[state] = result.count(True) / len(result)

    fig = m.plot(forecast)
    fig.savefig('RETCB_out/fig/prediction_{0}.png'.format(state))
    forecast['ds'] = [str(y)[0:4] for y in forecast['ds']]
    forecast.to_csv('RETCB_out/csv/prediction_{0}.csv'.format(state))


for state in ['az', 'ca', 'nm', 'tx']:
    print('{0}: {1}'.format(state, results_rate[state]))
```