For office use only	Team Control Number	For office use only
T1 ________________	**86834**	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	**C**	F4 ________________

2018
MCM/ICM
Summary Sheet

Development Goals of The State Energy Agreement

Summary

First, we judge the integrity and redundancy of information, and filter the singularity according to the calculation of each feature. Then we extract 605 features into 10 features by data mining.

Next,we propose a Energy evaluation model.Then we put forward amulti index, called energy index, can describe the situation of energy consumption. We choose the principal component analysis to determine their weight. We model the evolution of energy profiles for each state based on multiple regression models.Finally, a new time series model prediction model (Genetic Algorithm Decision Tree Model) is constructed by combining data clustering algorithm, fuzzy decision tree and genetic algorithm to predict the energy profile of 2025-2050 without policy change.

According to the analysis of the fourth problems of Part I, we get: CA: energy rich, AZ: energy scarcity, TX: energy rich, NM: less energy. Comprehensive analysis of four problems of Part I,our criteria for best profile is: jointly develop energy in four states, and maximize the use of energy. Finally, using all the above analyzes, we propose three actions to achieve the energy compact goals.

We also conduct a sensitivity analysis based on the modeling. In turn, we change the parameters, introduce the error, draw the error prediction table and so on, to analyze the model. We find that these do not have a significant effect, indicating that our model is stable.

Finally, we analyze the problems in the model. Our evaluation model ignores the impact of different dimensions.If time permits, the evaluation index obtained can be inversely normalized to obtain the evaluation index that can reflect different dimension levels.

Keywords: energ; regression model; GADT model; evaluation model

Contents

1 Introduction

In the world, how to build a clean, renewable and economic energy and structure system is a problem to be solved at the moment. And in recent years, the development of renewable energy has been integrated into various industries. Many states have signed energy contracts with each other to promote the development of renewable energy.

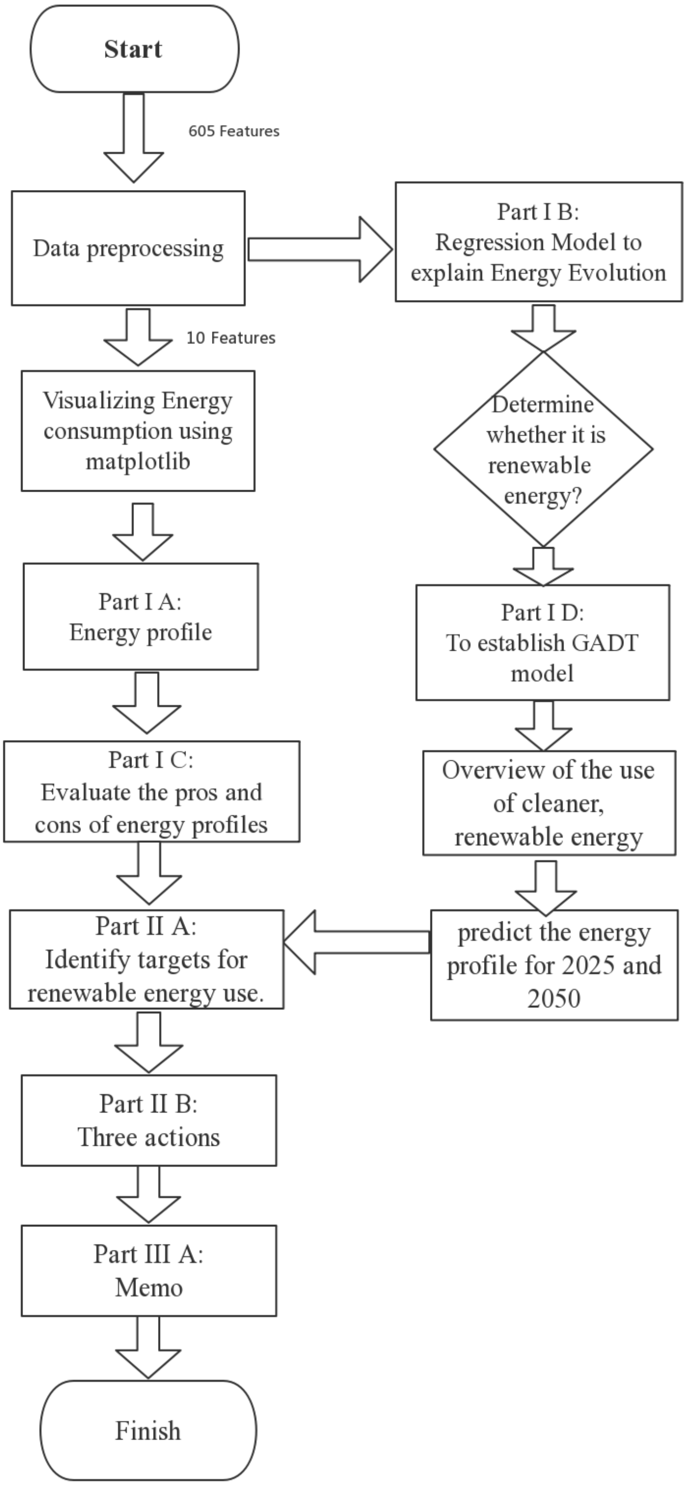
With the further consumption of energy, such as fossil fuels, such as resource depletion. Recyclable energy, with its inexhaustible, inexhaustible and clean features, has attracted more and more attention. How to make rational use of resources and how to make up the shortage of regional resources through cooperation has become the current trend.

Our model can be divided into three parts. First, we build the Energy evaluation model. At the same time, we increase the energy index value, and normalized the data. We reduce the 10 parameters to 4 principal components by principal component analysis.According to the load matrix of the principal component and the eigenvalues of the 4 principal components, we determine the weight of each parameter on the 4 principal components respectively. Moreover, because the four principal components are of different importance, we weight the four principal components according to the variance of the four principal components, and get the two level weight model to build our best profile for use of cleaner, renewable energy. Then, we use multiple regression models to describe the evolution of energy in four states in 1960-2009 years, and make a correlation analysis of the similarities and differences. Finally, a new time series model prediction model (Genetic Algorithm Decision Tree Model) is constructed by combining data clustering algorithm, fuzzy decision tree and genetic algorithm to predict the energy profile of 2025-2050 without policy change. We expanded the traditional decision tree model to solve the problem of regression and classification.

Then we use the established model, and the data provided. Combined with the four parts of Part I,we can find that the best profile is jointly develop energy in four states, coordinate energy differences between States and states, and maximize the use of energy. And we find that in order to meet their energy compact goals,we should take three measures.

- To take the cooperation model, the proper use of related technologies, and jointly develop and use energy.
- In accordance with the provisions of national laws, the four states signed non-aggression agreements. Energy distribution according to the principle of energy distribution.The cost of mining energy should be, the four states jointly pay according to the principle of distribution.
- To put an end to extravagance and waste of energy resolutely between the four states.

Figure 1: Flow chart


Start
605 Features
Data preprocessing
Part I B:
Regression Model to explain Energy Evolution
10 Features
Visualizing Energy consumption using matplotlib
Determine whether it is renewable energy?
Part I A:
Energy profile
Part I D:
To establish GADT model
Part I C:
Evaluate the pros and cons of energy profiles
Overview of the use of cleaner, renewable energy
Part II A:
Identify targets for renewable energy use.
predict the energy profile for 2025 and 2050
Part II B:
Three actions
Part III A:
Memo
Finish

1.1 Literature Review

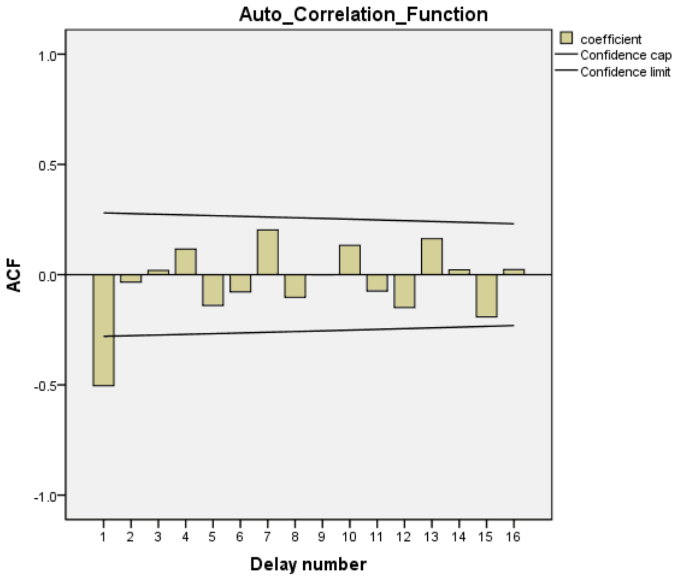
In the article, we set up our own energy evaluation model to evaluate the general situation of each state. In the model, we introduced the energy score to describe the use of energy in every state.

In the use of multiple regression models, we have made a correlation analysis of the factors of difference and similarity. We add the effects of population and GDP on energy consumption in the variables. flow is still not answered.

While analyzing the model, we considered the Auto-Regressive Moving Average Model, but we observed an overall upward trend in the energy profile, so we eliminated the effect of the overall trend by generating a differential sequence of energy profiles. Through the autocorrelation analysis for the differential sequence, we find that the differential sequence neither censoring nor trailing shows that the differential sequence does not satisfy the condition of wide stationary.

We have changed the traditional decision tree model for classification, combined it with genetic algorithms, using the clustering algorithm, Genetic Algorithm Decision Tree Model. In this method, a subset of the attributes is used as a training set to construct a decision tree, and the size of the subset is not affected by the number of attributes. Genetic algorithms are used to control the size of the decision tree. Thereby avoiding the situation of large-scale growth due to the increase of the number of attributes and further avoiding the occurrence of over-fitting. Applying our model to the prediction of the problem, combined with the actual results, the results show that our model is feasible and can make a reasonable forecast for the future energy situation.

Figure 2: Wide and steady



1.2 Assumptions

- The total consumption of each energy source represents an overview of that energy source.
- Use Electricity production to represen General situation of Nuclear power-Tidal energyWind energy.
- 3.All the units of energy used are Billion Btu.
- The first year of non-zero value represents the year for the use of energy.
- The factors that affect energy development in the United States are relatively stable in the future (such as population, economic growth).

1.3 Pretreatment

1. According to the formula of 605 variables [1], we get the meaning of the negative value and convert it to the corresponding positive value.
2. We delete the features of Thousand cords, Thousand barrels, Thousand short tons.
3. We extract all the representative features, a total of 34 features.
4. According to the formula [1], we get TETCB including FFTCB, NUETB, RETCB, ESTCB, FFTCB including: CLTCB, PMTCB , NGTCB(NNTCB); PMTCB including: ARTCB, AVTCB, DFTCB, JKTCB, JNTCB(JFTCB), KSTCB, LGTCB, LUTCB, MGTCB(MMTCB), POTCB, RFTCB and RETCB including: WWTCB(WDTCB, WSTCB), EMTCB, GETCB, HYTCB, SOTCB, WYTCB(WXTCB). Through the above screening, we finally get 11 features of CLTCB, NGTCB, PMTCB, NUETB, ESTCB, WWTCB, EMTCB, GETCB, HYTCB, SOTCB, WYTCB.
5. According to StateCode , we divide the features into four tables

2 The model

2.1 Evaluation model

In order to get a overview of "the best method" to use clean, renewable energy of four states, we set up a model for evaluating energy. After standardizing the data, we reduced the 10 parameters to four principal components by principal component analysis. According to the load matrix of the principal components and the eigenvalues of the four principal components, we determine the weight of each of the each parameter for four principal components. Since the four principal components have different degrees of importance, we weight the four principal components according to the square root of the eigenvalues of the four principal components to obtain a two-level weighted model. Finally, we have both the clean energy score and the non-clean energy score weighted by the clean energy parameter and the non-clean energy parameter. The two indices reflect the state's clean energy and non-clean energy development, respectively. We compare the clean energy score and the non-clean energy score to the conclusion that Texas uses clean energy " best".

We need a model to evaluate the energy structure and energy profiles of state.In order to eliminate the dimension, we limit the range of energy (related parameters) to [0, 1] by a standardized method of linear function transformation so that all values of the correlation coefficient matrix of normalized data are greater than zero.

$$x_L = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Through principal component analysis, we can conclude that the cumulative contributions of the first four principal components exceed 95%, and main differences of data can be retaine.

Sheet 1: Principal component analysis results.

Composition	Initial eigenvalue		
	Total	Variance %	Cumulative %
1	7.657	69.613	69.613
2	1.557	14.153	83.767
3	0.790	7.185	90.952
4	0.486	4.417	95.369
5	0.276	2.506	97.875
6	0.132	1.200	99.074
7	0.042	.377	99.452
8	.024	.220	99.671
9	.020	.183	99.854
10	.013	.123	99.977
11	.003	.023	100.000

Therefore, we choose the analysis of the first four principal components to get the factor loading matrix of the first four principal components.

Primary weight model:

$$\begin{cases} f_1 = U_{11} * x_1 + U_{12} * x_2 + \cdots + U_{L1} * x_L \\ f_2 = U_{12} * x_1 + U_{22} * x_2 + \cdots + U_{L2} * x_L \\ \vdots \\ f_m = U_{1m} * x_1 + U_{2m} * x_m + \cdots + U_{Lm} * x_L \end{cases} \tag{2}$$

The weight of each parameter is equal to the square root of each principal component load divided by the eigenvalue corresponding to the jth principal:

$$U_{ij} = \frac{u_{ij}}{\sqrt{\lambda_j}}, j = 1, 2, \ldots, m \tag{3}$$

According to formula (2) and formula (3), we calculate the value of each principal component and then substitute the following second-level weight model Second-level weight model:

$$F_j = f_j * \frac{w_{ij}}{\sum_{n=1}^{4} w_k} \tag{4}$$

which can derive to the principal component :

$$F = \sum_{n=1}^{4} f_j * \frac{w_{ij}}{\sum_{n=1}^{4} w_k} \tag{5}$$

Get the main composition score. The calculation formula of the composite score of principal component is:

$$\sum_{j=1}^{4}[\sum_{x \in x_{re}} \sum_{j=1}^{4}[\frac{u_{ij}}{\sqrt{\lambda_j}}]] \tag{6}$$

2.2 Multiple regression model

The relationship between related variables can be linear. Let $x_1, x_2, ..., x_p$ be p variables that can be measured accurately or controllable. If there is a linear relationship between variable y and $x_1, x_2, ..., x_p$, as the following formula [2]

$$\begin{cases} y_1 = b_0 + b_1 * x_{11} + b_2 * x_{12} + \cdots + U_p * x_{1p} \\ y_2 = b_0 + b_1 * x_{21} + b_2 * x_{22} + \cdots + U_p * x_{2p} \\ \vdots \\ y_m = b_0 + b_1 * x_{n1} + b_2 * x_{n2} + \cdots + U_{np} * x_n \end{cases} \tag{7}$$

Among them, $b_0, b_1, b_2, ..., b_p$ is p + 1 parameters to be estimated. We make multiple regression on the basis of year, population and GDP as dependent variables, total energy of all states, renewable energy and non-renewable energy as independent variables. A multiple regression equation for all parameters is created:

$$E = a * X_{Year} + b * X_{peo} + c * X_{GDP} + \xi \tag{8}$$

2.3 The GADT model based on genetic algorithm and decision tree

The traditional decision tree performs a recursive binary segmentation both in feature space [3]. We use genetic algorithms to blur the model based on the traditional model. The establishment of decision tree is divided into three parts:

- Choose features
- Generate a decision tree
- Prune decision tree

In the feature selection, we use K-means clustering algorithm, let K = 4, using historical data as a training set for the model When building a decision tree, the entropy of the fuzzy information is used as a heuristic for the decision tree. After constructing the decision tree, the extended attributes are selected by using the maximum information gain of the leaf nodes. The value of this attribute is an important basis for selecting the decision tree. In information theory and probability theory, entropy is a measure of the uncertainty of a random variable, X is a discrete random variable with a finite number of values, and its distribution is [4]

$$P(X = x_i) = p_i \tag{9}$$

So that the entropy of random variable X is defined as

$$H(X) = -\sum i = 1^n * p[i] * \log p_i \tag{10}$$

The greater the entropy is, the greater the uncertainty of the variable can be obtained. We can see that when P = 0 or 1, H (P) = 0, there is no uncertainty in the random variables. When P = 0.5, H (P) = 1 and the entropy is the largest. Uncertainty of random variables is the largest. By comparing the definition of entropy, we get the formula of conditional entropy

$$H(Y|X) = \sum i = 1^n * p[i] * H(Y|X = x_i) \tag{11}$$

The information gain we have given is characterized by the degree of reduction of the uncertainty of the partitioning of the dataset due to the characteristic A[4].

$$g(D, A) = H(D) - H(D|A) \tag{12}$$

According to the formula, we can make use of the definition of information gain, compare the result of the above clustering to the gain of information, and select the largest to divide the data set. Then use ID3 algorithm to generate decision tree.

Finally, prune the minimize fuzzy decision tree. We do this by minimizing the cost function in machine learning.

For genetic algorithms, all that is required is to evaluate each chromosome generated by the algorithm, express the solution to the problem as chromosomes, and select the chromosomes based on fitness to make more adaptive chromosomes have more breeding opportunities. Each individual breeds to produce a more environment-adaptable new generation by crossing and mutation. Simulate the next round of evolution of the new generation until a value that fits the environment. [5]

Genetic algorithm is evolved from the natural selection, and generally consists of the following five basic steps [6] the optimization of the object coding, and generate chromosomes, generating chromosomes containing the initial matrix; with the established evaluation function of the matrix Chromosomes are evaluated for fitness; at each generation, genetic manipulations of the chromosomes are performed, selecting, crossing, and mutating; performing the above steps iteratively until the evaluation function meets the requirements. Genetic algorithm is indeed successful, he avoids the traditional local optimal value problem.
par We will use four important operators: population numbers, generations of evolutionary, crossover rates, mutation rates:

Continuing the above five steps, sorting and forming a new parent can be used to optimize the structure of the decision tree and take advantage of the resulting new parent cycle until the optimal value is generated.

Sheet 2:Optimum genetic factor

Factor	EPISTAR	SIS	UMC
Chromosome	20	20	20
Evolutionary algebra	100	10	100
Crossing rate	0.9	0.9	0.9
Variation rate	0.1	0.1	0.3

Generate a decision tree, use the genetic algorithm to obtain the weight and then get the result.

2.4 Notations

Sysbol	Meaning
X_{year}	Year independent variable
X_{GDP}	GDP independent variable
X_{peo}	People independent variable
E	Energy
U_{ij}	The weight of the jth eigenvalue of the ith principal component
u_{ij}	The factor loading of the jth eigenvalue of the ith principal component
λ_j	The eigenvalue of the jth principal component
w_j	The contribution of variance
f_i	The first i principal component
ξ	Constant term
N	The total number of population
$\overrightarrow{x_i}$	The input instance
X	Discrete random variable.
D	The data set
A	Feature
$H(X)$	The entropy of random variable X
D_{ik}	The set of samples belonging to class C_k in subset D_i is D_{ik}
KMO	Correlation test variable

3 Disposal of problem

3.1 Part I A

Ten features were extracted according to the preprocessing part and then visualized according to Matplotlib in Python.

As shown in figure 1, it shows separately the total consumption of 10 characteristic energy sources in four states.

As shown in figure 2,the renewable energy profile of four states.

3.2 Part I B

We use python to reorganize the data, the use of groupby, unstcak function seseds data table Year as the index of the line, the MSN as the index of the column, to achieve the purpose of re-shaping a DataFrame.

Then use the Matplotlib module to visualize the 10 energy sources in each state.

As shown in figure 4,we take the multiple linear regression equation in Texas.

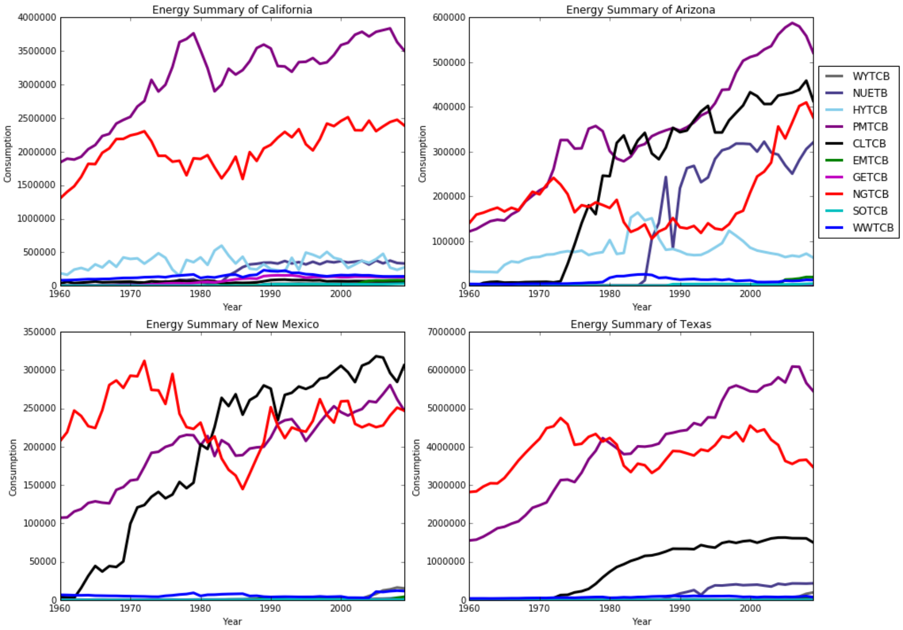


Figure 3: Energy profile

3.3 Part I C

As shown in figure 5,according to the division of pretreatment, we divide the energy into renewable and non-renewable energy, and use div and sum to adjust the sum of renewable energy and non-renewable energy corresponding to each index year to 1.Then use Matplotlib to draw a columnar stack plot that shows the ratio of renewable energy to non-renewable energy.

3.4 Part I D

We establish the GADT model to get the forecast result, as shown in figure 5:

Sheet 3 Energy Score in 2009

non-renewable energy score	2009			
	CA	TX	AZ	NM
non-renewable energy score	0.7427	0.7534	0.8950	0.5810
renewable energy score	1.2196	1.2844	1.2402	1.1194
total energy score	1.9622	2.0379	2.1352	1.7004

3.5 Part II A

We enter the predictive value of the four states in 2025 and 2050 into the evaluation model, then get predictive values of the energy score of the four states.

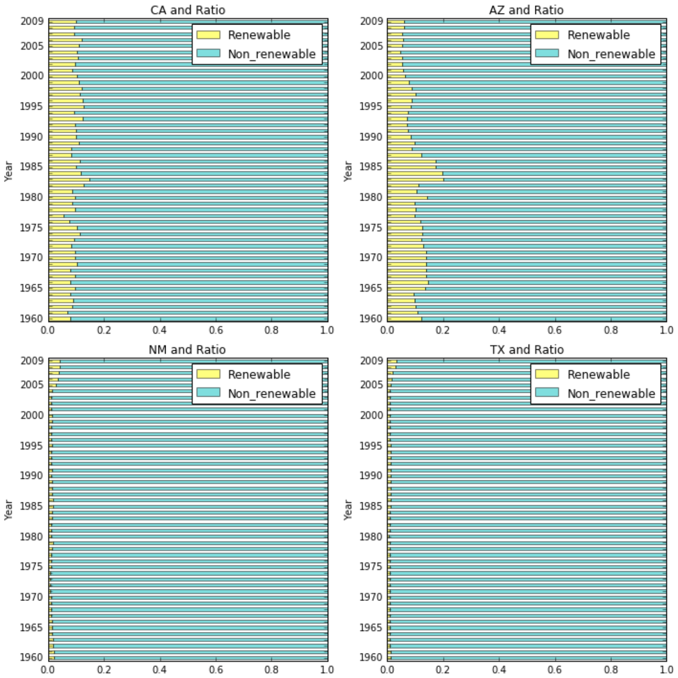


Figure 4: Renewable energy profile

Sheet 4 Energy Score in 2025 and 2050

Energy type	2025			
	CA	TX	AZ	NM
non-renewable energy score	0.111	1.461	1.723	0.137
renewable energy score	3.716	2.778	2.437	2.401
total energy score	3.827	4.239	4.160	2.538
	2050			
	CA	TX	AZ	NM
non-renewable energy score	0.638	1.368	1.277	0.585
renewable energy score	2.098	1.785	1.692	1.639
total energy score	2.737	3.153	2.969	2.225

4 Sensitivity analysis

- Part I B: In regression analysis, we analyze the sensitivity of some parameters. According to the coefficients of the regression equation, we can see that the error between the population and the GDP is not significant to the model disturbance in comparison with the model.

- Part I C: In our model, our evaluation model of the integrated use of various energy obtained by, we will increase or reduce the energy feature of standardized results by 0.05 to simulate error, single error of energy use has

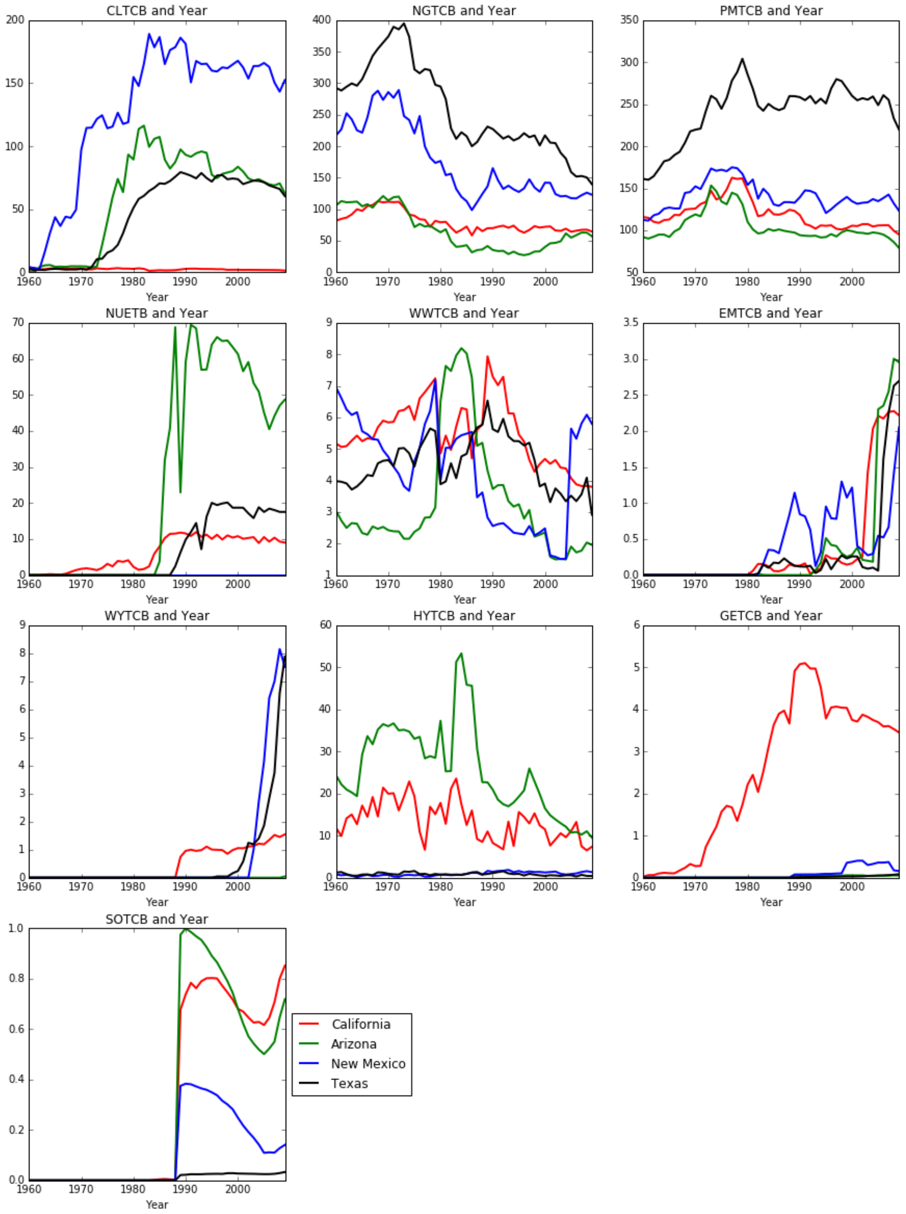


Figure 5: energy difference

little influence on our evaluation model, prove that our energy evaluation model is robust.

- Part I D: In our GADT model, we draw the results of the error prediction table as follows: The error of the actual value is controlled in a small range, so it can be shown that our GADT model is more effective. Through the sensitivity test above, we can prove that our model has good stability.

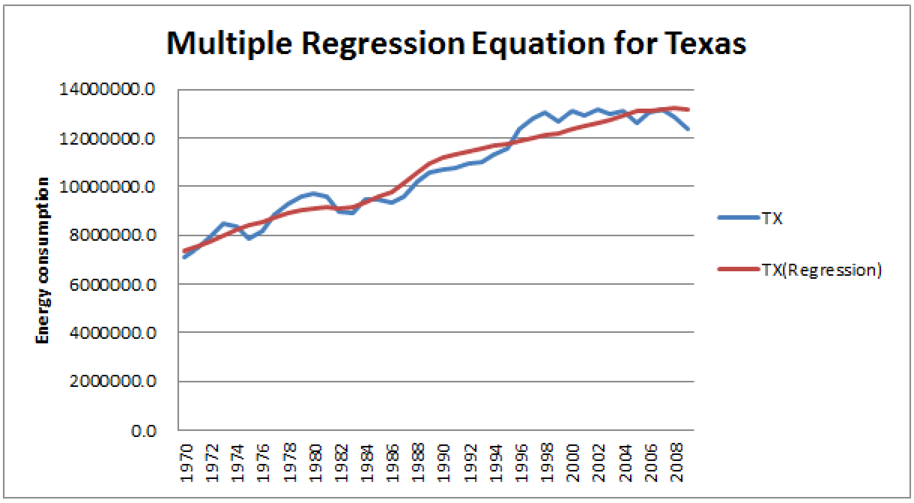


Figure 6: Multiple Regression Equation for Texas

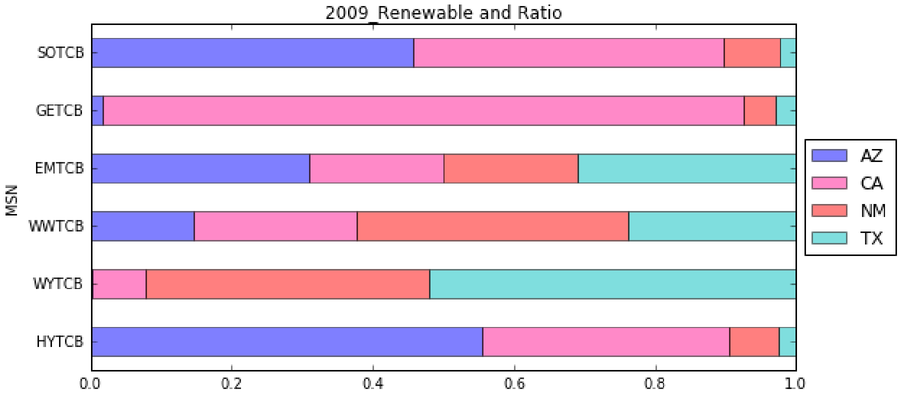


Figure 7: The percentage of states with different clean energy sources.

5 Conclusion

1. For Part I (A):

   CA:Energy structure is relatively simple, oil: In the past 50 years, the main energy consumption dominated by fossil fuels, of which the largest share of oil consumption, but this year's oil consumption has a downward trend. Nuclear power, though being used in a lower percentage, has been on the rise. Hydropower: Relatively more energy is consumed than fossil fuels.

   AZ:The energy structure is relatively rich, fossil fuels still dominate, the use of nuclear power rather late, but has been on the rise. Other primary energy consumption can increase the proportion of primary key.

   NM:The main energy consumed is still fossil fuels, but there is a replacement of main energy, and the consumption of coal gradually displaces the dominant position of natural gas as energy consumption. However, nuclear power generation has never been used. According to the US River Distribution, [7]. It is concluded that there is a shortage of water resources

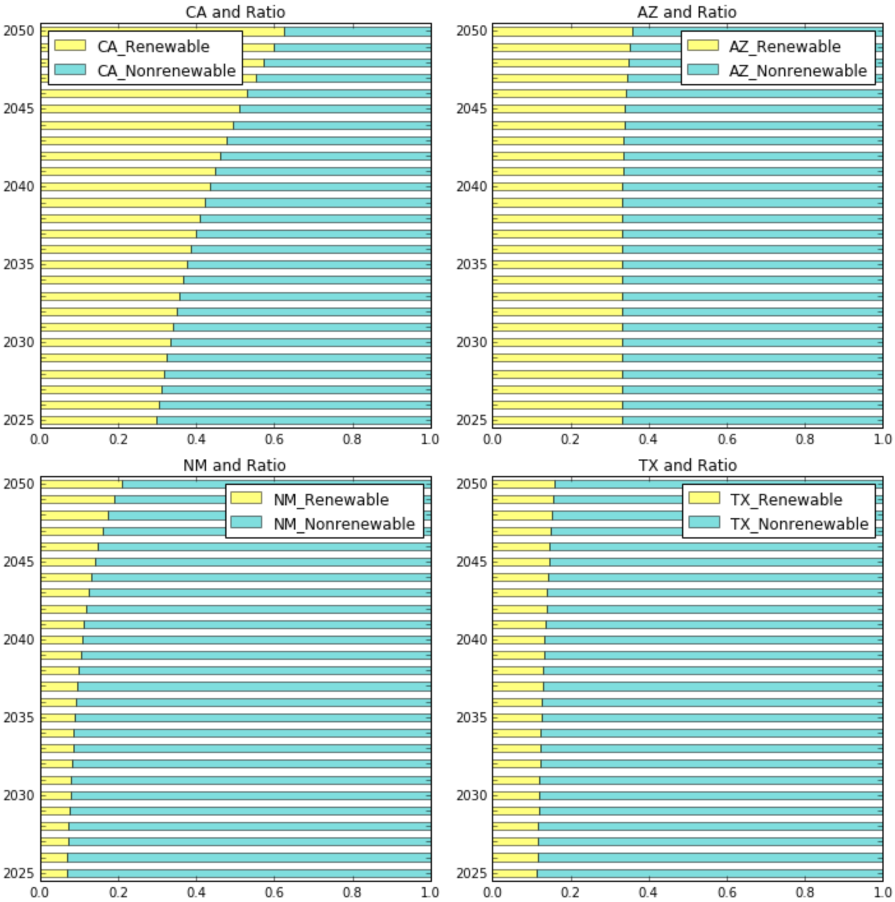
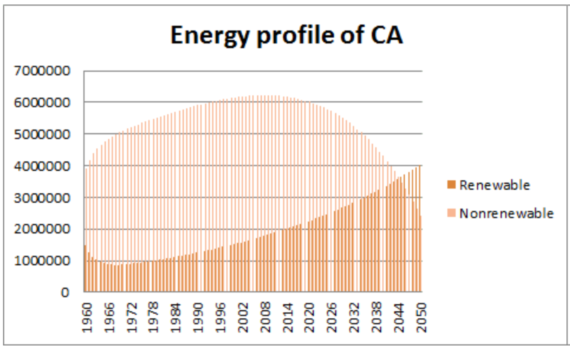


Figure 8: The percentage of states with different clean energy sources.



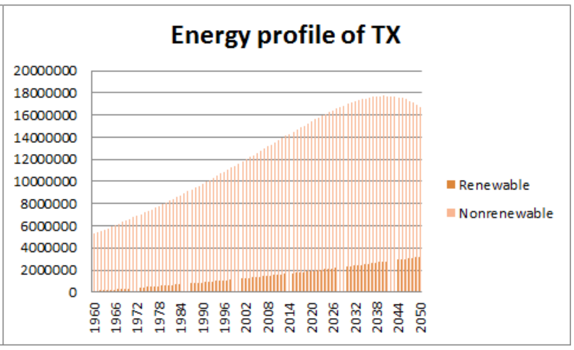


Figure 9: Enengy profile of CA

Figure 10: Energy profile of TX

in the NM state, and that large amounts of cooling water are required for nuclear power generation. The state uses very little hydroelectric power, further demonstrating our conclusion.

TX:As in most states, the main source of energy is fossil fuels, but the main source of natural gas is converted to coal. What's more, renewable energy is used later and less.

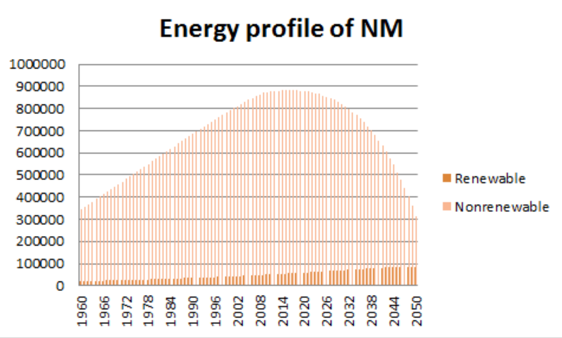


Figure 11: Enengy profile of NM

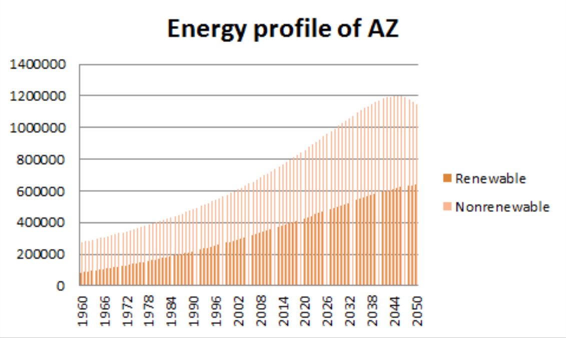


Figure 12: Enengy profile of AZ

2. For Part I (B):We can find that California is large in area and population, and requires large amounts of energy generation. We suspect that geographical location will also affect the energy structure of the four states to a certain extent.For example, California has a Mediterranean climate. Every year, the higher-temperature seawater forms a low pressure and attracts westerlies so that the westerly wind is greatly strengthened. The geography of the western coast and warm current from the North Pacific contributed to California's well-developed hydropower.Texas coastline in the Gulf of Mexico is not long, so hydroelectric power is weaker than California. Considering that the state of New Mexico is located in the southern Rocky Mountains, the annual increase of river water during snow melting in the Rocky Mountains also provides favorable conditions for hydroelectric power generation. Arizona, on the other hand, is inland and cannot use hydropower efficiently, but is located on a plateau that could be suitable for photovoltaic power generation.

3. For Part I (C):For energy profile in 2009, Texas is best, followed by Arizona, for renewable energy development. Arizona is the best in terms of overall energy development, followed by Texas. All kinds of energy development in New Mexico are not optimistic.

4. For Part I (D):For energy profile in 2025 and 2050, for renewable energy profile,Texas is best, followed by Arizona.For overall energy profile ,Arizona is best, followed by Texas.New Mexico energy profile is not optimistic.

5. For Part II (A),we get that, CA: Rich in energy. AZ: Energy scarce. NM: Rich in energy. TX: Less energy.

6. For Part II (B),We advocate three actions:

   Take the cooperation model, the proper use of related technologies, and jointly develop and use energy

   2.Set up a board of directors,and in accordance with the provisions of national laws, the four states signed non-aggression agreements. Energy distribution according to the principle of energy distribution.The cost of mining energy should be, the four states jointly pay according to the principle of distribution.

3.Put an end to extravagance and waste of energy resolutely between the four states.

References

[1] Appendix A. Mnemonic Series Names (MSN)

[2] Jaccard J, Turrisi R. Interaction effects in multiple regression[M]. Sage, 2003.

[3] H.Jiaozhuan,W.Zhenglin.Python vs. Machine Learning[M].1.Electronic Industry Press, 2017.3

[4] L.Hang.Statistical learning methods[M].1.Tsinghua University Press, 2012.3 :56-75.

[5] Janikow C Z. A genetic algorithm method for optimizing fuzzy decision trees[J]. Information Sciences, 1996, 89(3-4): 275-296.

[6] Z.Jinwu.Application of MATLAB in Mathematical Modeling[M].2.Beijing University of Aeronautics and Astronautics Press, 2011.

[7] James P E, Jones C F. American geography, inventory and prospect[M]. Syracuse University Press, 1954.

6 memo

Memorandum

To: Governors
From: Team 86834
Date: 2018.2.13
Subject: An energy contract based on historical data.

An analysis of four states along the United States border with California (CA), Arizona (AZ), New Mexico (NM), and Texas (TX), in terms of energy production, consumption, population and economic information over the 50 years from 1960 to 2009, Using regression, evaluation and GADT models, we can get the energy profile of the four states from 2025 to 2050, especially the development trend and general situation of clean and renewable energy. It is hoped that the governors of the states will adopt it.

CA's renewable energy is growing, the largest proportion of energy is about 60%, and it is expected to reach 30% in 2025.

AZ's renewable energy is growing, the largest proportion of energy is about 35.8%, and it is expected to reach 33% in 2025.

NM's renewable energy is growing, the largest proportion of energy is about 21.2%, and it is expected to reach 7% in 2025.

TX's renewable energy is growing, the largest proportion of energy is about 16.1%, and it is expected to reach 11% in 2025.

Suggest:

1. Renewable energy can promote the adjustment of energy structure, give priority to the development and utilization of renewable energy, and expand the utilization of renewable energy and increase the proportion of renewable energy in energy consumption as an important binding indicator of energy development in various regions.

2. Renewable energy is highly dependent on policies. Market-led efforts should be made to improve policies and mechanisms such as reducing the intensity of subsidies for new energy generation, implementing a full-scale guaranteed acquisition system for renewable energy generation and upgrading renewable energy sources Consumption level.

3. Renewable energy can not be effectively used, we must speed up technological progress and improve the ability of industrial innovation as the main direction to guide the development of renewable energy.