For office use only	Team Control Number	For office use only
T1 ________	**90904**	F1 ________
T2 ________		F2 ________
T3 ________	Problem Chosen	F3 ________
T4 ________	**C**	F4 ________

2018

Mathematical Contest in Modeling (MCM/ICM) Summary Sheet

Energy Profile Model to Evaluate and Predict Based on Dynamic Factors

Summary

We establish an evaluation-prediction system of energy profile based on dynamic factor model and linear regression model. Our model has three parts:

In the first part, we operate on the data. We firstly do data preprocessing to get a cleaner and more complete data set by filtering features and addressing missing data. The method of filtering features combine with clustering and artificial screening. Secondly, we make statistic on the selected data. What we get are the proportion of clean energy usage, the consumption, production and price of different energy and so on, from where we can analyze the energy profile in four states.

In the second part, we develop a composite index which describes the energy profile of a state. Using dynamic factor analysis technique, we can calculate the composite index based on the extracted principal component factors and their variance contribution rates. Therefore, we can firstly apply the processed data from the previous step on our evaluation model to obtain composite indexes of four states from 1960 to 2009. Then we can easily figure out how the energy profile of the four states has evolved during these years as well as determine whether it is a greenest energy development mode or not.

In the third part, we build up our prediction model by combing linear regression model with our evaluation model from the previous step. Therefore, we can forecast the composite index and its error bounds according to this model. Taking the upper bounds of predicted composite indices in 2025 and 2050 as goals, we consider the factors with greater impact in our evaluation model, as well as location, population and some other factors. We also propose that government should make great efforts to develop clean energy, develop complementary advantages and develop scientific measures.

Finally, we change the number of features to examine the sensitivity of our evaluation-prediction system. The result shows that our system is robust.

Keywords: energy profile; dynamic factor analysis; linear regression

Memo

To: The group of Governors

From: Team#90904

Subject: Further Views on Energy Profile

After calculating the composite index on the basis of Energy Profile Model, it’ s observed that, in 2009, CA and TX got better scores. One of the reasons is that they are both good at using renewable energy from their natural resources. That makes the ratios of clean energy consumption to total energy consumption is much higher, which would make a difference in the score. However, AZ and NM have more similar lower negative scores, probably due to the lower utilization rate of clean energy.

We predicted that in the future, if there are not any policy changes, CA and TX will get a higher score but AZ and NM will get lower score. This situation is due to CA and TX have a better energy usage ratio of clean renewable energy to not-clean energy, and those two states are easier to develop renewable energy due to their economy, resources and so on than AZ and NM.

The energy compact we recommended could be summarized in two parts—sharing and changing. Sharing means all four states have different resources of energy, one state may be the biggest oil producer and the another one has plenty of hydroenergy. So the energy sharing could make a difference. Also, some of your states depend on too much fossil fuel which means possible heavier pollution and unsustainable urban development. It is necessary to make a change step by step. Developing renewable and clean energy is our goal forever.

Sincerely,

Team #90904

Contents

1 Introduction

1.1 Background

Energy is the most basic guarantee of all modern human activities. Without energy, almost all of the machinery and equipment in the world will be out of service, causing huge losses.

Energy can be divided into two categories:

- Renewable energy: solar energy, water energy...
- Non-renewable energy: coal, oil...

At present, the energy used by mankind is mainly non-renewable energy, specifically it is fossil fuel. The production and consumption of fossil fuels has affected global climate change, resulting in environmental pollution and ecological destruction.

As the quantity of proven and easily exploitable energy resources is declining and the increasing demand for energy consumption force people to look for new resources.

In USA, it's observed that in each state, the energy production and consumption are different due to the geographical, industrial and demographic factors. So it would be advantageous to establish an effective interstate energy compact.

This article is about energy profiles for the 4 states in USA. We aim to create a clear and useful energy profile by using data and find out the trend and usage of different kinds of energy.

1.2 Restatement of the Problem

The problems that need to solve in this article are:

- **Problem 1** Create an energy profile for each of the four states.
- **Problem 2** Develop a model to describe the energy status of each state in four states from 1960 to 2009.
- **Problem 3** Analyze status and determine which particular state that may have the best profile for use of cleaner, renewable energy.
- **Problem 4** Predict the energy profile each state for 2025 and 2050.
- **Problem 5** Determine the renewable energy usage targets for 2025 and 2050.
- **Problem 6** Make a suggestion to the four states to meet their energy compact goal.

1.3 Our work

To solve the above 6 problems, we establish 2 main models in 3 steps as listed:

1. **Data Analysis** In this part, we do data preprocessing to get a complete and smaller dataset, which is the used data in Model 1.0 as well as Model 2.0. We also make statistics about these selected data to analyze the energy profile in four states.

2. **Evaluation Model** Through dynamic factor analysis, we build up a model to get a composite index, which can evaluate the energy profile. Also, we apply the processed data to our evaluation model to obtain a composite index of four states from 1960 to 2009. We analyze the results as well as find the state which had the best profile in 2009.

3. **Prediction Model** We build up a linear regression model on the basis of Evaluation Model. We not only predict the energy profile of four states in 2025 and 2050, but also set some targets for them. What's more, we offer some suggestions to achieve these targets.

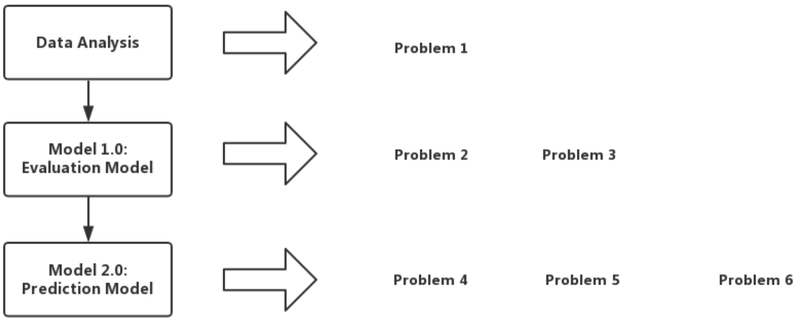

Figure 1: Overall structure diagram of our model constructions

2 Data Analysis

2.1 Data Preprocessing

The amount of raw data is large, so we should first do data preprocessing according to the usefulness and cleanness of the information.

First, we do data screening on the 605 variables.

In this section, we are surprised to find that many of the features have the same or similar meanings. Therefore, we extract the features in the following two ways:

- Cluster the features according to the units.
- Calculate the correlation coefficient and covariance matrix between features in each cluster.
- Choose and classify some features on the basis of statistic data.
- Further reduce the amount of features according to the meaning of the features in an artificial manner.

In this way, we gain 37 features including the consumption and production of different energy, price, population, GDP and so on, decreasing the number of the features from hundreds to dozens.

Secondly, we do data preprocessing on the timeline.

We find out that some Data from 1960 to 1970 and other years are missing. What's more, there are many zeros in the dataset, some of which are considered as assigned value by our observation. While some zeros are correct value in the provided data.Therefore, we take the following measures to solve these two problems:

- For the missing part, we calculate the average value among the latest 3 years and assign to it.
- For the zeros, we adopt no measures because they are mixed which we can not distinguish.

Through the above two ways, we get a complete data which may be not completely clean. But it is good enough for our following model construction.

After data screening, complete data include 4 states with dozens attributes such as different fuel's consumption, production and price and population and so on.

2.2 Data Statistics and Analysis

From the selected data, firstly we use SPSS to analyze the average consumption and production of different fuels in the latest 3 years(2007-2009) in each of the given 4 states. We also analyse the average consumption and average price of different sectors in the same conditions.

What we get are the following numbers and pictures:

- Arizona

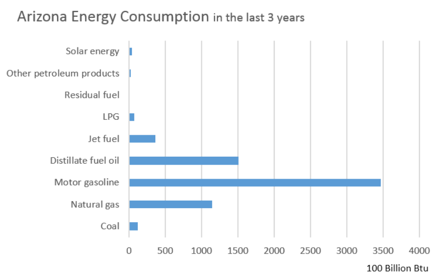

Figure 2: Consumption of different Fuels

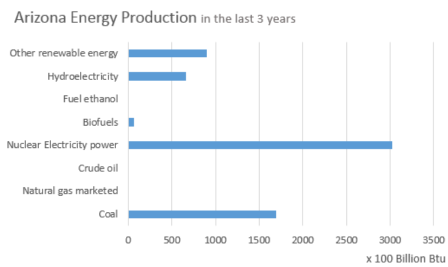

Figure 3: Production of different Fuels

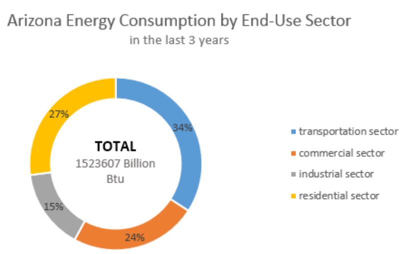

Figure 4: Consumption from different Sectors

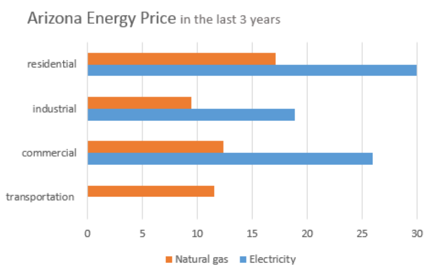

Figure 5: Price of different Sectors

From the above calculation, we can easily see the consumption of Motor gasoline and Distillate fuel oil were top two in Arizona, while clean energy such as Nuclear Electricity power and Hydroelectricity contributed quite a lot to the production of energy.

As for different sectors, the transportation sector ranked first in energy consumption. What's more, the price of Electricity in residential was quite high.

- California

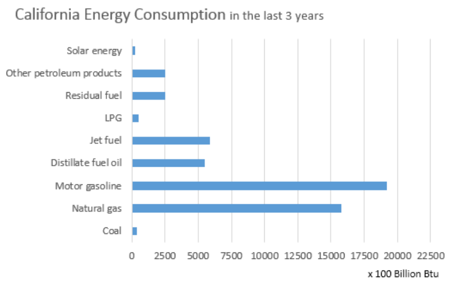

Figure 6: Consumption of different Fuels

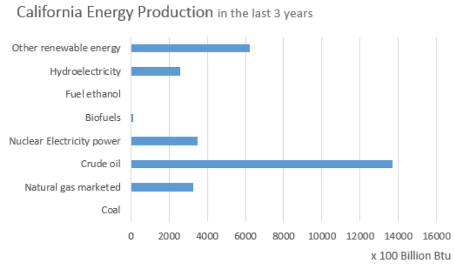

Figure 7: Production of different Fuels

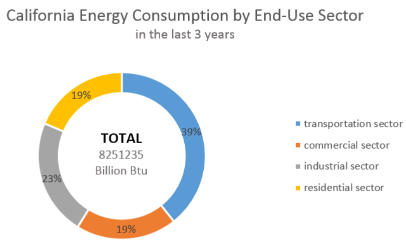

Figure 8: Consumption from different Sectors

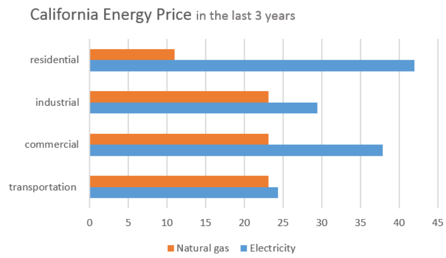

Figure 9: Price of different Sectors

From the statistic numbers and pictures, we can easily see the consumption of Motor gasoline and Natural gas were top two in California. Though renewable energy such as Nuclear Electricity power and others contributed quite a few to the production of energy, Crude oil still accounted for the most.

As for different sectors, the transportation sector ranked first in energy consumption. What's more, the price of Electricity in residential was a lot higher than the price of Natural gas in residential.

- New Mexico

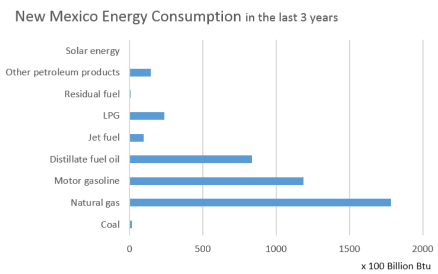

Figure 10: Consumption of different Fuels

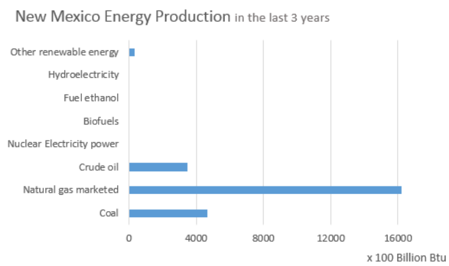

Figure 11: Production of different Fuels

From the statics, we can easily see the consumption of Natural gas, Motor gasoline and Distillate fuel oil were Big Three. However, Natural gas contributed quite a lot to the production of energy, which is consider more cleaner than Coal and Petroleum products.

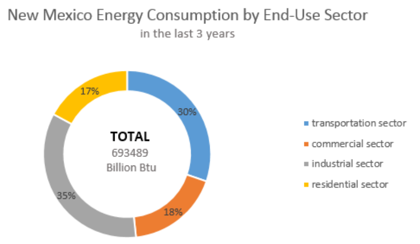

Figure 12: Consumption from different Sectors

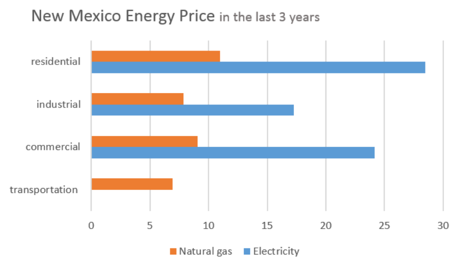

Figure 13: Price of different Sectors

As for different sectors, the industrial sector and transportation sector ranked first and second in energy consumption,respectively. What's more, using Electricity in transportation cost little in New Mexico.

- Texas

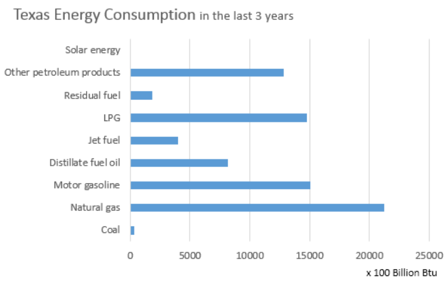

Figure 14: Consumption of different Fuels

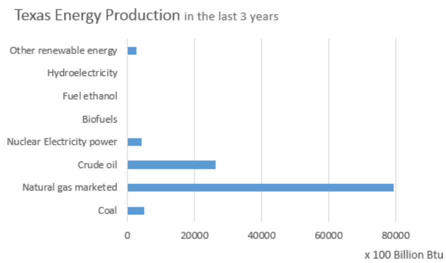

Figure 15: Production of different Fuels

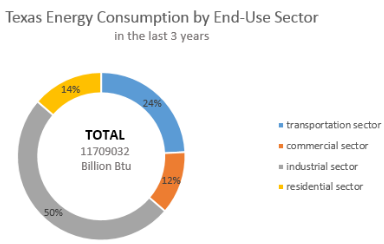

Figure 16: Consumption from different Sectors

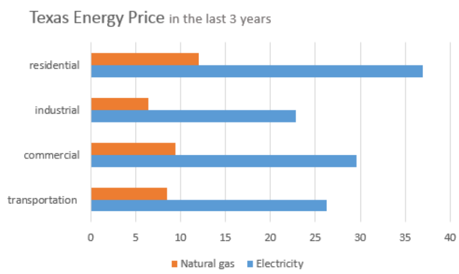

Figure 17: Price of different Sectors

From the numbers and pictures of Texas, we can easily see the consumption of many kinds of fuel were quite close. But as for the production of energy, Natural gas contributed quite a lot ,accounting for over a half.

As for different sectors, the industrial sector accounted for a half, ranking first in energy consumption. What's more, the price of Natural gas in every sector was much lower than the price of Electricity.

Secondly, we compare the total consumption, total production and the proportion of clean energy between the given 4 states.

Here are what we find:

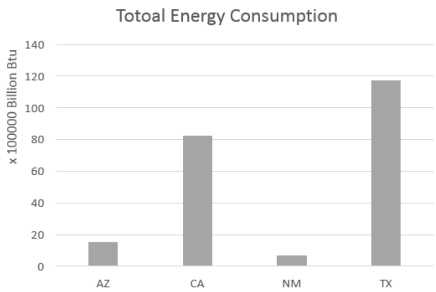

Figure 18: Total Consumption of different states

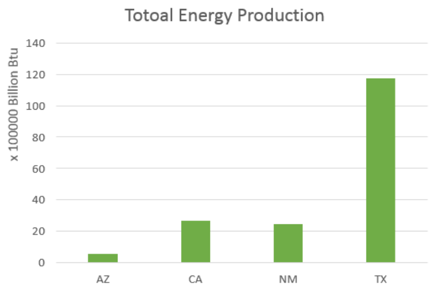

Figure 19: Total Production of different states

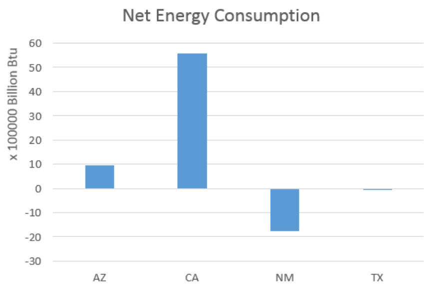

Figure 20: Net consumption of different states

- Consumption and Production

  From Figure 18 to 20, we can find that Texas ranked first between the 4 states in both energy consumption and production. As for California, the amount of consumption is large, while the amount of production is quite small.

  For the consideration of both consumption and production, we can calculate that New Mexico produced most energy in the given 4 states. We are not surprised by the large net energy consumption of California, but surprised by the small amount of production from Texas. The result is reasonable since the industrial sector in Texas consumed a lot energy according to the Figure 16.

- Proportion of Clean Energy

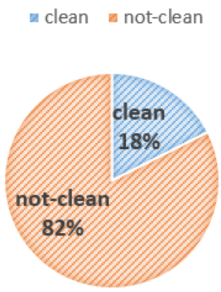

Figure 21: Arizona's Proportion on Consumption

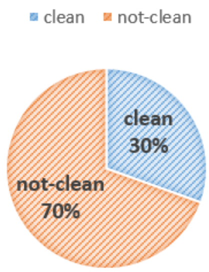

Figure 22: California's Proportion on Consumption

As for the amount of consumption, the main component of clean energy was Natural gas, which illustrates the reason why New Mexico had the biggest ratio.

As for the amount of production, clean energy shared a big part in Arizona because of its Nuclear Electricity power according to Figure 3

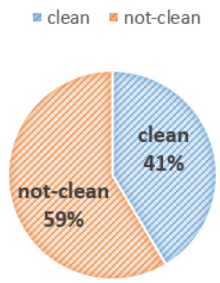

Figure 23: New Mexico's Proportion on Consumption

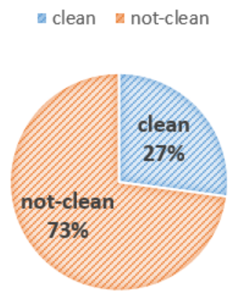

Figure 24: Texas's Proportion on Consumption

Now, we have a better command of the energy profile about the given 4 states, by calculating and comparing some important statics about the states. Therefore, we are going to construct a more comprehensive model to describe and illustrate the energy profile of a state.

3 Preparation of the Models

3.1 Assumptions

We make the following basic assumptions in order to simplify the problem. Each of our assumptions is justified and is consistent with the basic fact.

- Consider the zeros in data are all correct values since they are mixed with assigned value that we can not recognize. That's to say, we consider all data after our screening as clean data.
- According to the requirement, we pay great attention to the use of clean energy when we build up a composite evaluation index model. We care less about cost performance though we still consider the price of energy.
- Ignore inflation and deflation of money and other time value of money. The value of money remains unchanged.
- For evaluation, we assume that there are no policy changes made by each governor's office after 2009.
- Typically in ARIMA's timing model, we assume that the influence of other factors which are excluded outside the model can be ignored.

3.2 Notations

The primary notations used in this paper are listed in **Table 1**.

Table 1: Notations

Symbol	Definition
I	number of states
J	number of features
T	number of years
x_{ijt}	value of j feature from i state in t year
z_{ijt}	z-score value of x_{ijt}
σ	mean value
s	standard deviation value
S	covariance matrix
λ	eigenvalues
v	eigenvectors
c	score in the DFA model
E	composite index of energy profile
F	principal component
ϵ	error term
$\hat{E}$	estimate of composite index E

4 Model 1.0: Basic Model to Evaluate

In this section, we build a model and drive a composite index to evaluate the energy profile of a state.

4.1 Theory about Dynamic Factor Analysis

Dynamic factor analysis, is a methodology to combine, form a descriptive point of view.[1] It is a cross-section analysis through Principal Component Analysis and the time series dimension of data through linear regression model.

Comparing with common factor analysis model, dynamic factor analysis take the consideration of the influence from time series.Here are some mathematical introductions about it:

Assume that there are J indicators for a thing. Assume a vector $X(I, J, T) = x_{ijt}, i = 1...I, j = 1...J, t = 1...T$ that i means each subject, j means each indicator, t means each period.

Secondly, DFA use the principal component analysis and linear regression model to decompose the variance-covariance matrix into two parts:

$$S = S_t + *S_t$$

$$X_{\#jt} = a_j + b_{jt} + e_{jt}, j = 1...J, t = 1...T$$

which S_t is the variation produced by linear regression model in different periods and residual above should satisfy conditions below:

$$cov(e_{jt}, e_{j^{'}t^{'}}) = \begin{cases} w_j & j = j^{'}, t = t^{'} \\ 0 & otherwise \end{cases} \tag{1}$$

The model described above shows the relation effected by the principal component between indicators.

4.2 Construction of our Energy Profile Evaluation Model

From the result of our Data Preprocessing, we know that there are many mixed zeros and missing values in the data. However,for the reason that dynamic factor analysis can analyze short, non-stationary time series containing missing values[2], we decide to use this technique to form our model without worrying the above problems.

(1) Our selected features from Data Preprocessing are shown as following. Though some of them may not imply certain meanings, we need to drive a composite index to evaluate the energy profile of a state on the basis of them.

Table 2: Features after Screening

Features	Approximate meaning
X_1	Population in the area
X_2	Condition of Economy
X_3	Total consumption of not-clean Energy
X_4	Total consumption of clean Energy
X_5	Total production of not-clean Energy
X_6	Total production of clean Energy
X_7	Price of not-clean Energy
X_8	Price of clean Energy
X_9	Ratio in consumption of clean Energy
X_{10}	Ratio in production of clean Energy
X_{11}	Consumption of clean energy per capita
X_{12}	Production of clean energy per capita
X_{13}	GDP per capita
X_{14}	Sum of consumption
X_{15}	Sum of production

(2) The first step is the normalization of the considered variables. In this way, we eliminate the differences brought by different dimensions:

$$z_{ijt} = \frac{x_{ijt} - \sigma_j}{s_j} \tag{2}$$

where

$$\sigma_j = \frac{1}{I}\frac{1}{T}\sum_{i}^{I}\sum_{t}^{T} x_{ijt} \tag{3}$$

and

$$s_j = \sqrt{\frac{1}{I}\frac{1}{T}\left(x_{ijt} - \sigma_j\right)^2} \tag{4}$$

(3) Then compute $S(t)$, the covariance matrix of different years, on the basis of data in each year. Compute S_T as well, which indicates the average dispersion matrix within times(within variability):

$$S_T = \frac{1}{T}\sum_{t}^{T} S(t) \tag{5}$$

where $S(t)$ is the covariance matrix at year t. The dispersion within times jointly mirrors two effects: the effect due to the static structure of data and the one due to the differential dynamic of the single units[1].

(4) Next, we calculate the eigenvalues λ_1 > λ_2 >...> λ_{15} and eigenvectors v_1,v_2...v_{15} from S_T, respectively. At the same time, we can get the proportion of contribution from different Principal Component e.

(5) In the next step, we compute the static score c_{ih}, using the data z after normalizing and eigenvectors v_h we gain from the previous step:

$$\bar{z_i} = \frac{1}{T}\sum_{t}^{T} z_{it}, i = 1...I \tag{6}$$

$$\bar{z.} = \frac{1}{I}\sum_{i}^{I} \bar{z_i} \tag{7}$$

$$c_{ih} = (\bar{z_i} - \bar{z.})^{'} \cdot v_h \tag{8}$$

where

$$z_{it}^{'} = (z_{i1t}, ...z_{iJt})\,, i = 1...I, t = 1...T \tag{9}$$

(6) In this step, we compute the dynamic score c_{ih}, as well as the final score - composite score E, which is our evaluation of the energy profile in different states during different periods:

$$c_{iht} = (z_{it} - \bar{z_{.t}})^{'} \cdot v_h \tag{10}$$

where

$$\bar{z_{.t}} = \frac{1}{I}\sum_{i}^{I} z_{it} \tag{11}$$

we get the composite score of energy profile:

$$E = \frac{1}{T}\sum_{t}^{T} c_{iht} \tag{12}$$

4.3 Application of our Model on the Dataset

To describe how the energy profile of each of the four states has evolved from 1960 - 2009, we use dynamic factor analysis to build an evaluation model for energy use. With the help of Python 3.6.2 and MATLAB R2016a, we analyze the data processed after our Data Preprocessing part.

Table 3: Data of Each Component

Principal Component	Eigenvalue	Variance Contribution	Cumulative Variance Contribution
$e1$	10.68192187	0.534096094	0.534096094
$e2$	6.809895581	0.340494779	0.874590873
$e3$	1.470072179	0.073503609	0.948094482
$e4$	0.373378403	0.018668920	0.966763402
$e5$	0.288210616	0.014410531	0.981173932
$e6$	0.170357961	0.008517898	0.989691831
$e7$	0.159140861	0.007957043	0.997648874
$e8$	0.019711299	0.000985565	0.998634439
$e9$	0.013547190	0.000677360	0.999311798
$e10$	0.007544982	0.000377249	0.999689047
$e11$	0.004272618	0.000213631	0.999902678
$e12$	0.001276909	0.000063845	0.999966523
$e13$	0.000364488	0.000018224	0.999984747
$e14$	0.000258675	0.000012933	0.999997680
$e15$	0.000046368	0.000002318	0.999999998

Use the mean-variance-covariance matrix to solve the eigenvalue, the eigenvector, the variance contribution rate and the cumulative contribution rate of variance. The result is in Table 3. Based on the chart, the eigenvalue of principal component 1 and principal component 2 is 10.68192187 and 6.809895581 which is bigger than any other principal components. The variance contribution rate of these components is 0.534096094 and 0.340494779, and the cumulative variance contribution rate reaches 0.874590873. These two principal components can fully reflect the main information of the original 15 features and are reasonable to represent the comprehensive level of clean energy in four states. Thus, we use the first and second principal components as the calculation factor, The principal component score coefficient matrix, also called the eigenvector of S_t, can be used to describe the relation between the first, second, third principal components extracted from each state and the original features after standardization.

The principal component score coefficient matrix is in the Table 4.

The matrix reflects the level of the contribution of each feature to the two principal components we choose. It's observed that for the first principal component, the biggest effect on it is the consumption of clean energy and the variance contribution rate reached 0.5341. According to this, the consumption of clean energy is the most critical factor in measuring the state's energy consumption, especially for the use of clean energy.

For the second principal components, the most influential feature is three, which are the per capita clean energy production, the proportion of clean energy consumption per capita and the total consumption of clean energy for all energy consumption. This principal component reflects the problem of proportions well, because the four states have different sizes. The production and consumption of clean energy per capita can better reflect the development of clean energy in the state. And proportion of clean energy consumption to all energy consumption is a fair measure of four states from another point of view. The variance contribution rate of second principal component factors is 0.3405, which indicates that the component is the second important factor affecting the evaluation system.

Table 4: Principal Component Score Coefficient Matrix

Principal Component	$e1$	$e2$
X_1	0.236024057	-0.302895213
X_2	0.240123001	-0.298132204
X_3	0.331980511	-0.133795632
X_4	0.350747700	-0.029888885
X_5	0.345861838	0.071106210
X_6	0.323335521	0.147468992
X_7	-0.111597088	-0.359304473
X_8	-0.227050292	-0.303668886
X_9	0.058609205	0.352377371
X_{10}	-0.076117602	0.303153413
X_{11}	0.211085368	0.340034284
X_{12}	-0.003755406	0.391947608
X_{13}	0.284373289	-0.21355182
X_{14}	0.345346767	-0.086790181
X_{15}	0.333607472	0.118340221

Combining the principal component score coefficient matrix and the principal component expression, we can get the static score of the four states for the first principal component and the second principal component as Table 5.

Table 5: The Static Score of Two Principal Components

State	$e1$	$e2$
AZ	-3.00923431	-0.77282275
CA	0.868948992	-3.228305515
NM	-2.07280171	2.644078853
TX	4.213087028	1.357049412

Calculate the dynamic score of each sample every year based on the extracted principal component and its partial variance contribution rate as Table 6. In this step, we have already got the composite indexes of 4 states from 1960 to 2009.

4.4 Analysis of the results

From the above calculation, we finally get the average comprehensive score as an index for evaluating the utilization of clean energy in different states. The higher the score is, the better the use of clean energy is. The following chart can be shown according to the Figure 25.

- According to our model, the higher the overall score, the better the state's clean energy use. From Figure 25, we can clearly see that California has the best profile for use of cleaner,renewable energy in 2009.

Table 6: The Dynamic Score of each state shown in every three years

Year	AZ	CA	NM	TX	Year	AZ	CA	NM	TX
1960	-0.443	1.21	-0.487	0.721	1985	-0.424	1.214	-0.501	0.712
1963	-0.438	1.217	-0.486	0.708	1988	-0.41	1.238	-0.494	0.666
1966	-0.436	1.221	-0.485	0.701	1991	-0.412	1.239	-0.493	0.666
1969	-0.435	1.22	-0.487	0.702	1994	-0.413	1.219	-0.507	0.701
1972	-0.433	1.218	-0.49	0.706	1997	-0.411	1.208	-0.517	0.721
1975	-0.432	1.215	-0.493	0.71	2000	-0.402	1.217	-0.519	0.704
1978	-0.432	1.214	-0.495	0.712	2003	-0.398	1.215	-0.523	0.706
1981	-0.431	1.206	-0.501	0.726	2006	-0.391	1.213	-0.53	0.709
1984	-0.427	1.212	-0.501	0.715	2009	-0.396	1.198	-0.535	0.733

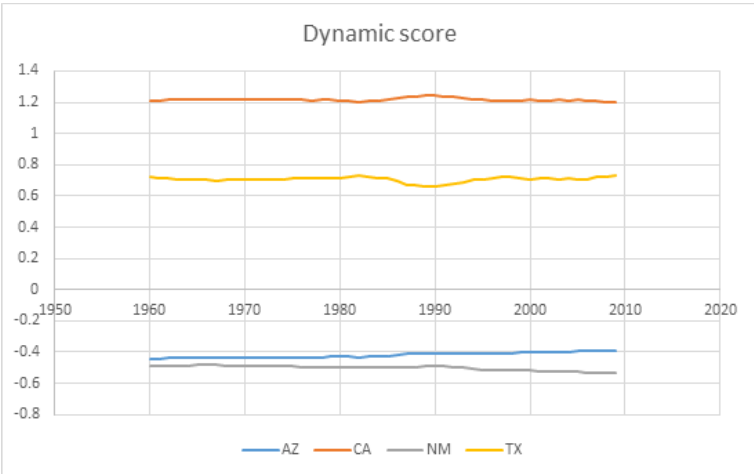

Figure 25: The Dynamic Score of each state

In general, according to the score, the difference of energy situation in these four states is distinct. From 1960 to 2009, CA always got the highest score. It shows that the use of clean energy is more common , the amount and proportion of CA are the highest, and has the best clean energy usage. The second place is TX. The scores of CA and TX are both positive, indicating that the use of clean energy in these two states is more optimistic in the four states, with higher amounts and proportions. AZ and NM have more similar scores and are negative scores, suggesting that these two states are less likely to use clean energy in the four states.

Focusing on the score of each state, we can see that for each state, there is a certain fluctuation from 1960 to 2009, and they have different trends.

From Figure 26, the average score of AZ between 1960 and 2009 is -0.421 and the overall trend of the score was slowly increased. The proportion of clean energy consumption to total energy consumption in AZ state has always been at a low level and it has been decreasing year by year. It is only 17.78% in 2009. AZ is more dependent on the use of non clean energy, which results in its score lower than CA and TX.

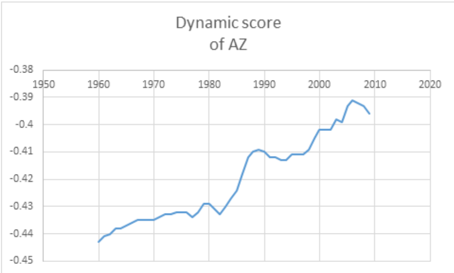

Figure 26: The Dynamic Score of AZ

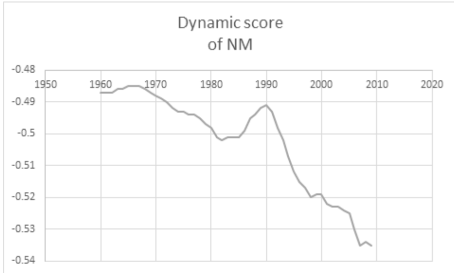

Figure 27: The Dynamic Score of NM

From Figure 27, NM is the lowest scoring state in four states with an average score of only -0.0502, and the score is still declining from the trend of scoring. NM's clean energy consumption has basically not changed, but the total energy consumption rose from 270776.7102 to 415739.1045, increased 53.54%. So it's a large increase in the consumption of not-clean energy, leading the ratio of clean energy consumption to total energy consumption has dropped one third. Obviously, NM tend to use non clean energy while the other three states improve their clean energy use ratio. When NM adjusting its energy structure, the use of not-clean energy ratio increases, which makes it has the lowest energy, and continuously reduce.

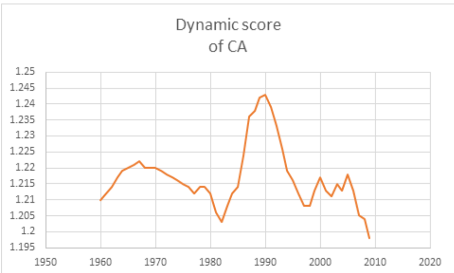

Figure 28: The Dynamic Score of CA

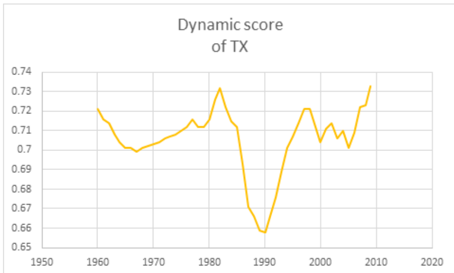

Figure 29: The Dynamic Score of TX

CA is the highest scoring state in four states with an average score of 0.0717, far ahead of the other states. One reason is that CA has rich solar energy resources, there are three large wind farms and plenty of natural gas power plants. Moreover, CA has a long coastline with a very livable Mediterranean climate. It has large population but less energy consumption per capita. It can attract more talents to develop clean energy industry based on its livable conditions.

For CA itself, it's complicated scoring trend in this period. It has reached its peak in 1990 or so, and then dropped back to the previous level. This is determined by multiple factors. We will analyze it in different periods.

The score between 1960 and 1970 increased slowly, between 1970 and 1982 is decreased. This is because during this period, the output and consumption of clean energy both increased and then decreased. And the score began to increase in 1983, and peaked in 1990, which is the result of clean energy increase and the continuous decrease of not-clean energy. From 1991 to 1998, it fell back to the initial value, because the trend was reversed. The amount of non clean energy is increasing, and the amount of clean energy is decreasing. When the score dropped to the initial value, it fell into a vibrate change.

TX scores second in four states, the average score is 0.705. TX is the nation's largest producer of lignite coal, the leading oil-producing state and the leading natural gas producer with superior location. TX has a wealth of wind energy resources, and wind energy is relatively stable and sustainable, very suitable for the development of wind power projects, so it can get a higher score. But compared to CA, TX has a bigger not-clean energy consumption and per capita energy use.

The score of TX first decreased and increased back to initial value between 1980 and 2000. As shown below, TX's non clean energy consumption has always kept increasing at a constant rate. The consumption of clean energy is also increasing at the beginning, but it began to decline near 1980, resulting in a continuous decline in the proportion of clean energy consumption, resulting in a reduction in the score.

5 Model 2.0: Enhanced Model to Predict

5.1 ARIMA Model to predict

Approach Autoregressive Integrated Moving Average Model(ARIMA) is one of the most famous model in time series analysis.

The ARIMA model is quite simple, only endogenetic variables are needed without the use of other exogenous variables. But notices that the time series data or its differencing should be stationary.

The ARIMA model has three parameter – p, q, d, p represents the lag number (lags) of the time series data used in the prediction model, also called the AR/Auto-Regressive, d represents how many times in differencing to become stationary, also called Integrated, q represents the lag number of prediction errors used in the prediction model , also called MA. With this model, it could be easy to predict the future data.

Result First, we used the score from 1960 to 1999 year as time series data to predict the next ten years'score of each state. Secondly, we accounted the relative error and worked out coefficient of determination.

Coefficient of determination provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.[3]

The Goodness of fitness associates with lower value of average relative error and higher value of coefficient of determination. The result is as following:

State	Average Relative Error	Coefficient of Determination
AZ	4.27%	20.9%
CA	4.67%	15.8%
NM	2.67%	57.9%
TX	3.31%	49.1%

Table 7: The Result From ARIMA Model

5.2 Our Energy Profile Model to predict

Approach A common framework of standard linear regression model is as listed[4]:

$$y_{t+1} = \beta x_t + \epsilon_{t+1} \tag{13}$$

In this equation, y is regressand and x is regressors, β is the regression coefficients and ϵ is error term which is satisfy the classical conditions. In our cases about energy profile, the regression model should be like that:

$$E_{t+1} = \beta x_t + \epsilon_{t+1} \tag{14}$$

On the basis of data $\{E_t, x_t\}_{t=1}^{T}$ got from our Model 1.0, we can predict the energy profile in next time stamp.

It is a common occurrence that x_t the feature itself is time-related variable when we forecast overall performance. For the consideration of time-related features, we improve the prediction model from the strict condition in ARIMA model.

In our data, a large number J of features $x_{jt}, i = 1...I$, which are collected, are driven by few unobserved common factors, summarized in F. Accordingly, dynamic factor models express the variables x_{jt} as the sum of a common component and an idiosyncratic component. Here is a static representation[5]:

$$x_{jt} = \Lambda_j F_j + \epsilon_{jt} \tag{15}$$

From these facts, the regression equation can be updated as listed, which is not only capable to deal with time-related variables, but also proper for our model[6]:

$$E_{t+1} = \alpha F_t + \epsilon_{t+1} \tag{16}$$

which means

$$\hat{E_{t+1}} = \alpha F_t \tag{17}$$

where F is the Principal Component in our Energy Profile Model 1.0 and coefficient α can be calculated as well. Therefore, there is a two-step forecasting approach:

(1) Compute principal component F and its coefficient as well from our Model 1.0

(2) Compute $\hat{E}$ iteratively from its previous states.

Result First, we used data from 1960 to 1999 year as input into our model to predict the next ten years'score of each state. Secondly, we accounted the relative error and worked out coefficient of determination. The Goodness of fitness associates with lower value of average relative error and higher value of coefficient of determination. The result calculated by SPSS is as following:

State	Average Relative Error	Coefficient of Determination
AZ	2.52%	63.9%
CA	2.78%	59.6%
NM	1.02%	87.8%
TX	1.56%	76.3%

Table 8: The Result From Our Energy Profile Model

Comparing the results from our model and ARIMA model (Table 9,) we are very sure that our model is better at predicting. Because according to average relative error, the average value of four states in our model is 1.97% lower than 3.73% in ARIMA model. What's more, the average value about coefficient of determination in our model is 71.9% much higher than 35.925% in ARIMA model.

Model	Mean	Maximum	Minimum	Range
ARIMA	35.925%	57.9%	15.8%	42.1%
Ours	71.9%	87.8%	59.6%	28.2%

Table 9: Coefficient of Determination between different Models

It is not surprising that our model work better. Since many input features are closely related with time variable, ARIMA model require stationary time series and is not particularly suitable for dealing with such problems. But our Energy Profile Model 2.0 is based on Dynamic Factor Model, which is capable to work well in such conditions.

Therefore, we predict the composite score(E value) of energy profile in 4 states on the basis of our Model 2.0 and the result is listed as following:

State	E-value in 2025	Upper Bound in 2025	E-value in 2050	Upper Bound in 2050
AZ	-0.40799	-0.38106	-0.42111	-0.37165
CA	1.1855	1.1962	1.14925	1.1593
NM	-0.54099	-0.52354	-0.60159	-0.56099
TX	0.82537	0.91148	1.03711	1.18035

Table 10: Predicted E-value in Four States

From the above table, we can see that E-value in Arizona and New Mexico is decreasing from 2009, which means the gap between AZ,NM and CA,TX is larger. What's more, the E-value of California is always the biggest. The energy profile of California is stable and excellent.

5.3 Energy Profile in future

Goal According to the above,we use our energy profile model to predict the cleaner energy scores of four states in 2025 and 2050,choose the upper bound of the four states as their targets of 2025 and 2050the result is as following:

State	Score in 2009	Target in 2025	Target in 2050
AZ	-0.396	-0.38106	-0.37165
CA	1.198	1.1962	1.1593
NM	-0.535	-0.52354	-0.56099
TX	0.733	0.91148	1.18035

Table 11: Targets of four states in future

The dynamic factor analysis integrated scoring system we established combines a variety of features to make a comprehensive and credible assessment of the state's use of clean energy .The higher the score, the better the use of clean energy in the state in the four states.. Because state governments can adjust the use of clean energy according to more effective policies, The upper bound scores can be defined as the targets of the four states in 2025 and 2050, making the targets concisely and comprehensively reflect the state's clean energy use within a reasonable range and facilitates the four states to reach new clean energy treaties.

In our scoring system, the final score is the result of a combination of features.Among them, the consumption of clean energy, the ratio of clean energy consumption to total energy consumption, per capita clean energy consumption and per capita clean energy production have a significant impact on the scores. In order to achieve the set goals in these four states, They need to improve their use of clean energy and maximize the proportion of clean energy consumption and the proportion of clean energy production, making the use of energy focus shifted from non-clean energy to clean energy sources.

Measures Specific measures are as follows:

- **Develop clean energy** To develop the state's clean energy industry according to local conditions, Texas and California, which have large-scale wind power, solar energy and natural gas industries, should continue to strengthen and utilize the advantages of location and geography on the existing basis. Arizona's nuclear energy industry developed, has a large proportion of clean energy production, but the industry is single, should have multi-direction for the development of clean energy.
- **Develop complementary advantages** At the same time, two states with higher scores and more clean energy use can deliver clean energy to neighboring states with lower use of clean energy and guide them in developing clean energy, so as to jointly clean the energy industry. New Mexico's clean energy industry is weak with abundant non-clean energy resources, which can increase the proportion of its own clean energy use by transporting the non-clean energy it produces to needed state
- **Develop scientific measures** Industries that can use clean energy for energy supply can reduce the use of non-clean energy until they can not be replaced with clean energy, and can use scientific methods to improve the efficiency of energy use, reduce energy consumption and improve energy efficiency

6 Strengths and Weaknesses

6.1 Strengths

In our model, there are many advantages as listed:

- **Objective** For the evaluation, the weights of performance index are usually assigned by the experts in a subjective way. While in our model, our composite index of energy profile are calculated in an objective way.
- **Tolerance of missing value** To evaluate or predict the energy profile of a state, a large variety of data containing many aspects are usually needed. And it is a common situation that some values of the data will be missing. Therefore, our model on the base of DFA model is robust due to its tolerance of missing part.
- **Consideration of time series** Comparing with common Factor Analysis model, our model consider the influence of time series. Our 3-dimension model can lead us to observe the score of different items as well as different time stamps.

- **Ability to analyse non-stationary time series** For the prediction, it is a fact that some features, which can illustrate the energy profile of a state, are closed related with time. Because of this strength, our model can still work well under above conditions.

6.2 Weaknesses

- The complexity of our model is one of the disadvantages. If we want to evaluate the energy profile of a certain state, at least five steps with calculating many indexes and values of variables are needed after gaining the proper data. And if we want to do further to predict, steps, indexes and calculation are even more.
- The principal component factors are not easy to explain under all situations. Therefore, some results of the model sometimes may be understood in addition of other background knowledge.

7 Sensitivity Analysis

As can be shown from the figure, even if one of the previous features is deleted, we can still see that scoring trends and relative levels between four states are consistent with previous one, indicating that the sensitivity of the model is good.

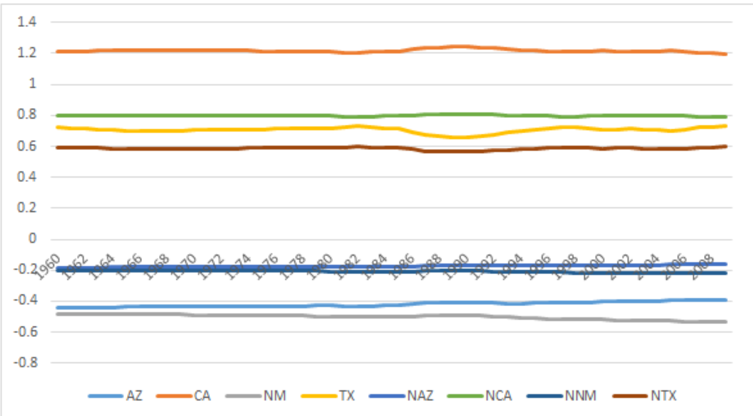

Figure 30: The Result of Sensitivity Analysis

8 Conclusions

This paper manages to develop two models - an evaluation model to find out the energy profiles of each states by using the huge database and an enhanced prediction model to predict the future. With the help of the DFA technique, we can easily figure out how the energy profile of the four states has evolved from 1960 to 2009. We can also distinguish whether it is a green and clean energy development mode or not. As a reference for signing the new energy compact, which focused on increased usage of cleaner and renewable energy sources, we predict the energy profile for 2025 and 2050 and identify the actions that the four states might take to meet their energy compact goals.

References

[1] Alessandro Federici. Dynamic Factor Analysis STATA, 2005.

[2] A.F.Zuur, R.J.Fryer. Estimating common trends in multivariate time series using dynamic factor analysis, 2003.

[3] https://en.wikipedia.org/wiki/Coefficient_of_determination

[4] https://en.wikipedia.org/wiki/Linear_regression

[5] S Eickmeier, C Ziegler. How successful are dynamic factor models at forecasting output and inflation. A meta analytic approach,2008

[6] JH Stock. Diffusion indexes, 1998