| For office use only | Team Control Number | For office use only |
|---|---|---|
| T1 ________________ | **92564** | F1 ________________ |
| T2 ________________ | | F2 ________________ |
| T3 ________________ | Problem Chosen | F3 ________________ |
| T4 ________________ | **C** | F4 ________________ |

**2018**
**MCM/ICM**
**Summary Sheet**


**Summary**

We develop a model for analyzing big data on energy and succeed in proposing a long-term plan on renewable energy usage for the state governors and decision makers.

First, we operate on the data. We analyze the features of the original data and do data screening according to the integrity and usefulness of the information. Then we develop an index to quantify the energy profile of each state. The index is a linear composition of several significant indicators that reflect the energy profile best. Such denotion facilitates our analysis in the following sections.

Second, we put forward two kinds of model, SARMA for data fitting and ARIMA for data prediction. In the data fitting section, we apply the model into every dimension of the profile index. Combining the result of the fitting curve and other possible influential factors, we reach a conclusion on the difference of these four states in terms of renewable energy usage. Further, to find out the characteristics of energy usage in the future, we use ARIMA to make data prediction and get satisfactory results.

Third, we construct an evalution system to determine whose profile is better. Enlightened by the Fuzzy Comprehensive Evaluation Method, we define four factors that can affects whether the profile is good or not. The factors include the rates of clean energy, primary energy, energy loss and energy output, each describes the profile from a different angle. At last we find that Texas owns the best energy profile.

Finally, we propose a plan on future usage of renewable energy based on the previous analysis. Then we make sensitivity analysis for our model, changing the input parameters to analyze the different results of the output and to find the better parameters for ideal results.

To sum up, our model is a feasible and reasonable model with technical and data support. Because of the subjectivity, this model can be used flexibly after data training.

# Contents

# 1 Introduction

## 1.1 Problem Statement

In the United States, an interstate compact is an agreement between two or more states in which a set of standards or plans on a particular regional matter are adopted. The formulation of the policy usually requires analysis of big data and many factors. Western Interstate Energy Compact (WIEC), for example, is a compact that deal with cooperation between 12 western states for the development and management of nuclear energy technologies.

In respond to the need of optimizing the energy production and usage, four states in the U.S. − California (CA), Arizona (AZ), New Mexico (NM), and Texas (TX) − intend to form a realistic new energy compact focused on increased usage of cleaner, renewable energy sources. We are tasked with performing data analysis and creating models that can be applied to determine a set of energy usage targets for their interstate energy compact. The solution proposed in this paper will extract important information from the big data and objectively give a policy on energy use based on the historical evolution in these states.

## 1.2 Planned Approach

Because the data provided is quite huge, with 605 variables and about 100,000 samples available, we must select important indicators that can effectively reflect the energy profile of a state from the original data (since no one would like to deal with 605 variables in an equation). Then, to give a recommendation, we need to analyze the energy profile of each state, which includes three components: comparing their similarities and differences, finding criteria on deciding the best profile and predicting the possible change in the future. Based on the analysis above, we can finally propose a reasonable and feasible set of goals and actions for the energy compact to adopt.

Therefore, the problem can be divided into three parts with detailed steps as follows:

1. Data Overview and Focus Decision
    - minimizes the discomfort to the hands, or
    - maximizes the outgoing velocity of the ball.
2. Analysis and Prediction
    - minimizes the discomfort to the hands, or
    - maximizes the outgoing velocity of the ball.
3. **Strategy Making**
    - minimizes the discomfort to the hands, or
    - maximizes the outgoing velocity of the ball.

### 1.3 Assumptions

We make the following assumptions to simplify our model in this paper.

- Energy profile means a set of important factors that well depict the energy consumption and production of a state.
- Clean energy is viewed as an equivalence of renewable energy.
- In the evolution model, we focus on improving the degree of fitting rather than its ability to predict, though both of them are timing model.
- In the prediction model, we assume no policy changes happens during the 25-year period.

## 2 Definition of the Energy Profile

### 2.1 Data Overview

The attached data file *ProblemCData.xlsx* contains two worksheets, ***seseds*** describes 50 years of data in 605 variables on each state's energy production and consumption while ***msncodes*** defines the meaning of these variable names.

In the first worksheet, the variable values are shown as numbers (0 means the energy was not consumed or produced this year). We notice that some data is not recorded for some specific years, but for each state the data is intact. We will take this into consideration in the following analysis.

According to the second worksheet, the 605 variables can be categorized into three groups (along with some demographic and economic information): 1) Consumption 2) Prices and expenditures 3) Production. The consumption and expenditures of the energy are recorded in five different sectors − transportation, commercial, industrial, residential(end-use sector) and electric power(non end-use sector). Again, some data in certain sectors are not available.

### 2.2 Data Screening

Faced with the raw data with 605 variables, we should first do data screening according to the integrity and usefulness of the information, so that the energy profile proposed in the following section would be easy to understand for the governors and decision makers. The rules of data selection are as follows:

- Some variables do not denote a certain type of energy, such as TPOPP (resident population in the state). We pick these special variables from the raw data and put them together in a new table.
- We delete the energy that contains a lot of missing data. If most data are not accessible for certain kind of energy, analyzing them is meaningless.

- We combine some variables together. For example, we pick several kinds of renewable energy that are most widely used all over the world and combine the data of other energy together, named as "Other Renewables", using them as the input data in the following analysis.
- Some variables are related to each other. We choose one variable to represent all the related variables and delete all the other variables. For example, CLACB and CLACP both reflects the amount of coal consumption in the transportation sector, while CLACK is the converting factor between these two variables. We let CLACB be the representative and filter out the other two data. In this way, we can avoid duplication of data.

### 2.3 Definition of the Energy Profile

After data screening, there are about 15 types of energy that can be used to describe the energy profile of a state.

To help the governors better understand the energy profile in four states, we use data in 4 different years(1980, 1990, 2000, 2009) to draw several charts according to three dimensions: consumption, expenditures and production. The following are part of the charts in California, 2009.

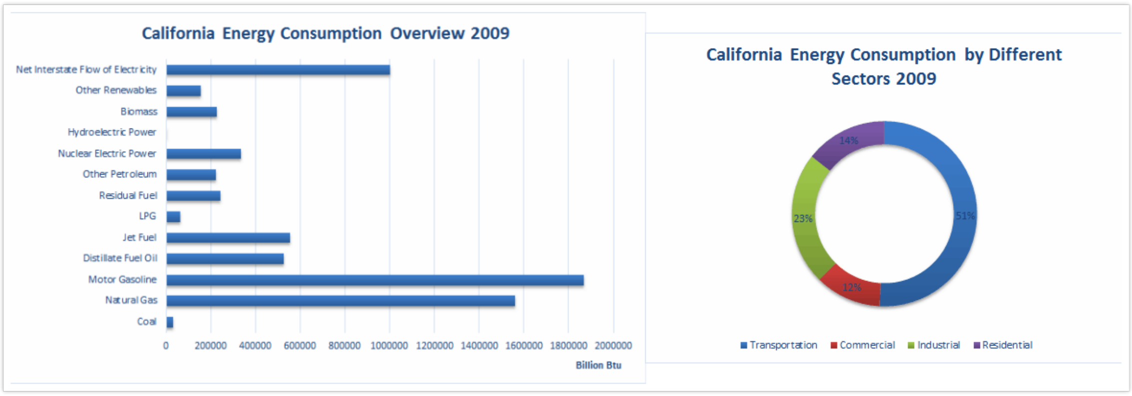


Figure 1: the AIC Values

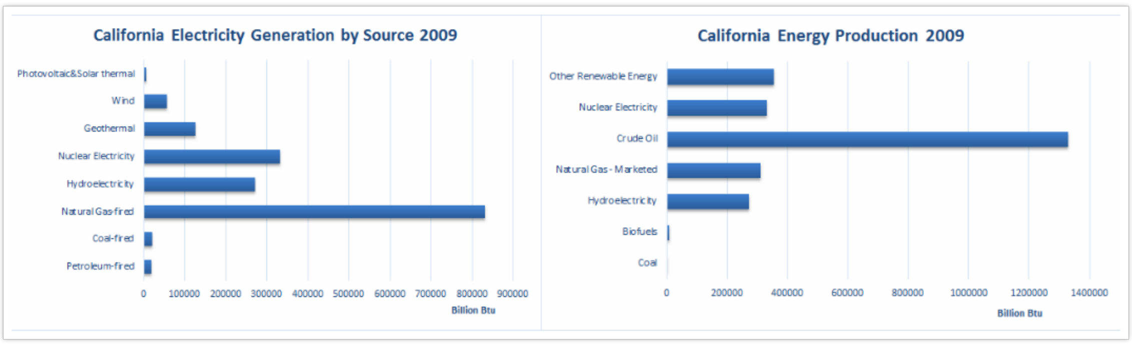


Figure 2: the AIC Values

For the sake of the following analysis and prediction, for each type of energy we denote ***energy profile*** as an $n \times 1$ vector whose structure is as follows:

$$energy\_profile = (type, year, state, indicators) \tag{1}$$

Since the effective indicators must be able to reflect the characteristics of each state, we can select the indicators that show great difference in those charts. After comparing the charts, we get the indicators shown in table 1.

We will use this structure to do the analysis in the following section.

| Number | Indicator | Description |
|---|---|---|
| 1 | total_expenditure | expenditures in total |
| 2 | total_consumption | consumption in total |
| 3 | primary_expenditure | expenditure of primary energy |
| 4 | clean_consumption | consumption of clean energy |
| 5 | tran_expenditure | expenditure in transportation sector |
| 6 | com_expenditure | expenditure in commercial sector |
| 7 | ind_expenditure | expenditure in industrial sector |
| 8 | res_expenditure | expenditure in residential sector |
| 9 | tran_consumption | consumption in transportation sector |
| 10 | com_consumption | consumption in commercial sector |
| 11 | ind_consumption | consumption in industrial sector |
| 12 | res_consumption | consumption in residential sector |

Figure 3: Indicators of the Energy Profile

## 3 Analysis of the Energy Profile

In this section, we aim to develop a model to characterize the evolution of the energy profile of each state from 1960 2009 and compare similarities and differences in the usage of clean and renewable energy sources.

From the previous discussion, we have determined the structure of energy profile. In order to fit the development of the four states' profiles, we model each dimension of the profile respectively. Since the data are non-linear, we choose the **Autoregressive Moving Average (ARMA)** model to fit the data.

The model is usually referred to as the ARMA(p,q) model where p is the order of the autoregressive part and q is the order of the moving average part (as defined below).

Because we don't need to make predictions here, we omit the procedure to test if the sequence is stationery and denote our model as **SARMA(Simplified Autoregressive Moving Average)**. In the following section we will improve it for the accuracy of prediction.

1. Choosing p and q

   The first step we do is to find an appropriate value of p and q. We use **Akaike Information Criterion (AIC)** to evaluate the quality of our choice.AIC is an estimator of the relative quality of statistical models for a given set of data.Let $k$ be the number of estimated parameters in the model. Let $L$ be the maximum value of the likelihood function for the model. Then the AIC value of the model can be calculated with the formula below.

$$AIC = 2k - 2ln(L) \tag{2}$$

   We calculate the AIC value for various combinations of p and q and determine the optimal pair when the corresponding AIC value is minimal.

2. Estimating coefficients

   The ARMA model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. Therefore, the equation of ARMA contains p autoregressive terms and q moving-average terms.We use this equation to fit the practical data and obtain the values of $\alpha$ and $\beta$.

$$Y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k x_{t-k} + \beta_0 + \sum_{k=1}^{q} \beta_k \varepsilon_{t-k} \tag{3}$$

We perform the steps above for every variable defined in the energy profile vector. Here we take Arizona as an example and characterize its evolution of energy expenditures.

| AIC | | q | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| p | 1 | 14.2241 | 14.2467 | 14.0845 | 14.1351 | 14.1628 |
| | 2 | 14.2739 | 14.2491 | 14.1053 | 14.1692 | 14.1087 |
| | 3 | 14.3090 | 14.1952 | 14.1444 | 14.1895 | 14.1646 |
| | 4 | 14.2029 | 14.0677 | 14.2005 | 14.2323 | 14.1040 |
| | 5 | 14.1942 | 14.2465 | 14.2195 | 14.0860 | 14.1615 |

Figure 4: The AIC Values

It can be seen from the table that the AIC reaches the minimum when p = 4, q = 2, so we choose them as the order of our model. Then we use the model to calculate the parameters and get the fitted curve below.

$$\alpha = (1.0000, -2.9236, 3.0324, -0.8667, -0.2667) \tag{4}$$

$$\beta = (1.0000, -1.8011, 1.0000) \tag{5}$$

$$Y_t = 1.0000 - 2.9236x_{t-1} + 3.0324x_{t-2} - 0.8667x_{t-3} - 0.2667x_{t-4} + 1.0000 - 1.8011\varepsilon_{t-2} + 1.0000\varepsilon_{t-2} \tag{6}$$

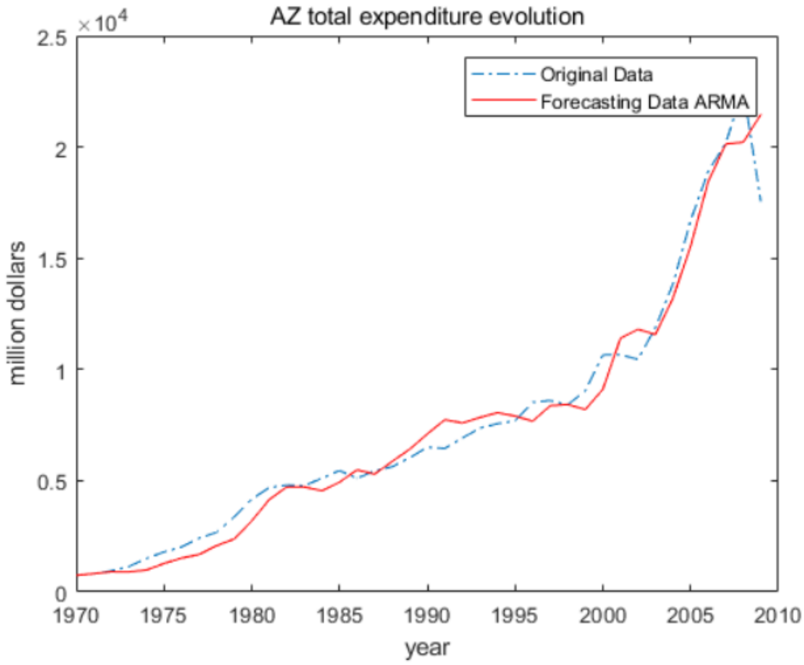


Figure 5: Evolution of Energy Expenditures in Arizona

# 4 Determine the Best Profile

## 4.1 Choose the Criteria

### 4.1.1 Clean Energy Rate

Firstly, we choose the ratio of clean energy and total energy consumption as a primary criteria, denoting it $Index1$.

$i.e.$

$$Index1 = \frac{the\ consumption\ of\ clean\ energy}{the\ consumption\ of\ total\ energy} \tag{7}$$

The reason for choosing this criteria is quite obvious. The larger the proportion of clean energy in total energy consumption, the better the overall energy structure is. In the ensuing discussion, we will add additional parameters that make this indicator a greater weight.

### 4.1.2 Primary Energy Rate

Then we choose the ratio of primary energy expenditure and total energy expenditure as the next criteria, denoting it $Index2$.

*i.e.*

$$Index2 = \frac{the\ expenditure\ of\ primary\ energy}{the\ expenditure\ of\ total\ energy} \tag{8}$$

Here the explanation becomes a little less intuitive. First things first, we should make it clear that apart from the primary energy we have the second energy which means that the direct energy is converted into electricity. And here we justify that the primary energy usage is of less benefit than the energy used to generate electricity. Therefore, the smaller the proportion of primary energy expenditure in total energy expenditure, the better the overall energy structure is.

#### 4.1.3 Loss Energy Rate

At last we choose the ratio of total energy loss and total energy consumption as the last criteria, denoting it $Index3$.

*i.e.*

$$Index3 = \frac{total\ energy\ loss}{the\ consumption\ of\ total\ energy} \tag{9}$$

We have seen some difficult to understand standards, yet this criteria is still satisfactory for us to explain. It is conspicuous that the smaller the ratio is, the better the overall energy structure is.

#### 4.1.4 Energy Amount Rate

Initially we did not consider the scale effect, leading to the result of choosing Arizona to be the best-structured energy profile state. After a further consideration, we add the fourth criterion to evaluate the proportion of the amount of energy in each state, denoting it $Index4$.

*i.e.*

$$Index4 = \frac{the\ consumption\ of\ total\ energy\ in\ each\ state}{the\ consumption\ of\ total\ energy\ of\ all\ states} \tag{10}$$

This useful benchmark helps us to take the real amount of energy consumption into consideration. If we do not use it, some state may score too high due to its high clean energy rate, whereas the total amount of energy the state consumed is insignificant compared to other high energy consumption states, which leads to a certain degree of unbalance.

### 4.2 Overview of the Profile under the Criteria

Now that we have defined all the criteria we need, here we present the scatter plot of the energy structures of the four states. One thing which needs to address is that we have found some data that is undefined in the total energy expenditure column and primary energy expenditure column between 1960 and 1969. We assume that these data are somehow lost or was not collected for some reason. At first we handle this situation by setting all these values to be zero. However, the result seems unsatisfactory. Thus we filling the missing data by setting them to be the average value of the rest of the data for each state. And here we give an overview of the four state energy structure according to all historical data.

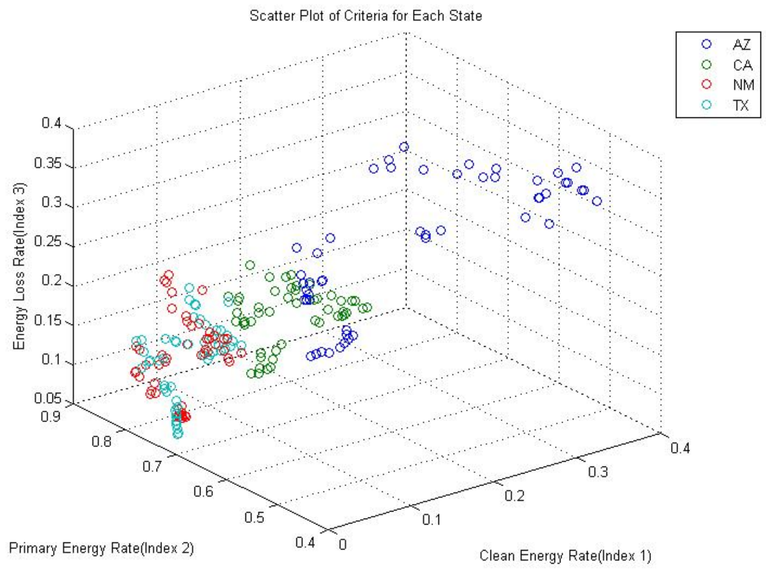


Figure 6: Scatter Plot of Criteria

Specifically we plot the scatter of the criteria in 2009 as required.

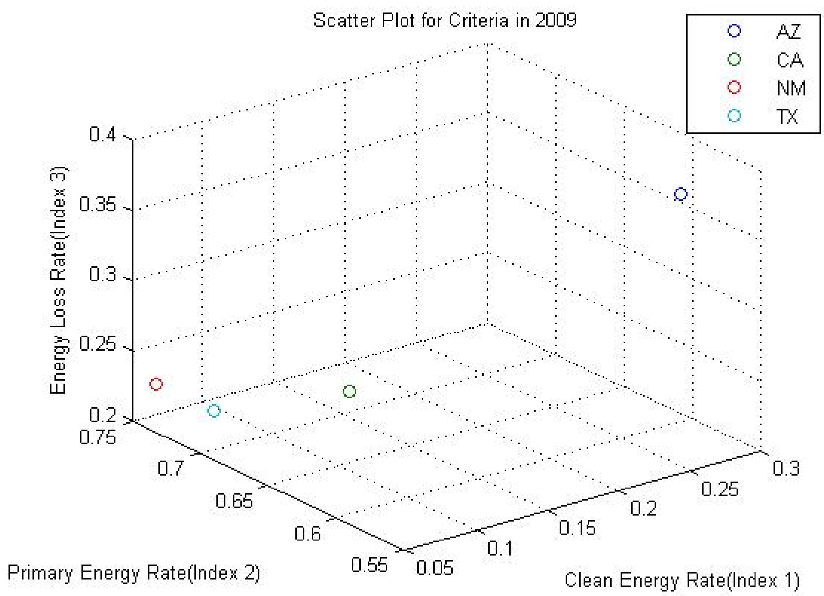


Figure 7: Scatter Plot of Criteria

### 4.3 Analysis of the Results

Reminding that the larger $CleanEnergyRate$ and $EnergyAmountRate$ and the smaller $PrimaryEnergyRate$ and $EnergyLossRate$ is, the better the overall energy system is.

Thus, intuitively we can derive that Arizona has the largest $CleanEnergyRate$ followed by California. Whereas Arizona also has the largest $EnergyLossRate$ followed by New Mexico. According to $PrimaryEnergyUse$, New Mexico has the largest rate, indicating that a substantially large proportion of the energy is directly put into use rather that for electricity generation.

Now enough for the analysis of graph. Hereafter we introduce a method to measure how well the energy structure is for a certain state.

First, we define some parameters to change the share of these criteria in the final comprehensive evaluation model.

$$\lambda = [\lambda_1 \ \lambda_2 \ \lambda_3 \ \lambda_4] \tag{11}$$

where it satisfies the constraint:

$$\sum_{i=1}^{4} \lambda_i = 1 \tag{12}$$

To simplify denotation, use $X$ to aggregate the indexes:

$$X = [X_1 \ X_2 \ X_3 \ X_4] = [Index1 \ Index2 \ Index3 \ Index4] \tag{13}$$

Note that the optimal structure is $X$ = [1 0 0 1]. Accordingly we can use the following function to represent the distance between the current state energy structure and the optimal energy structure.

$$D = -log_2(\lambda_1 \cdot X_1 + \lambda_2 \cdot (1 - X_2) + \lambda_3 \cdot (1 - X_3) + \lambda_4 \cdot X_4) \tag{14}$$

To put an emphasis on the proportation of renewable and clean energy use, we set the $\lambda$ as following.

$$\lambda = [0.30 \ 0.15 \ 0.25 \ 0.30] \tag{15}$$

After processing on the state data, we get the distance to the optimal energy structure of four states in 2009 and the distance on average.

From the figure and the table we can easily observe that the energy structure of Texas is the nearest to the optimal energy structure. Thus we come to a conclusion that Texas has the "best" profile for use of cleaner, renewable energy in 2009.

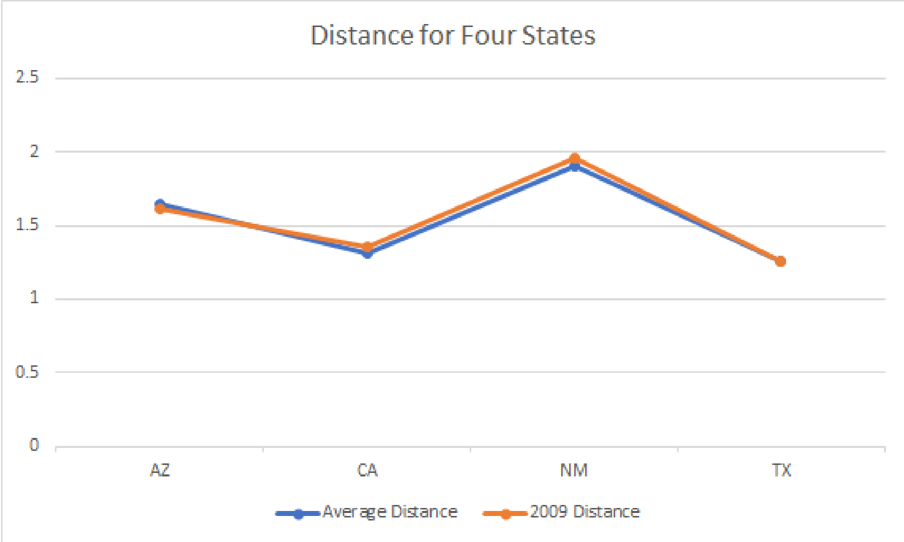


Figure 8: Distance to the Optimal Energy Structure in 2009

| State | Distance |
|---|---|
| AZ | 1.611534395 |
| CA | 1.351416151 |
| NM | 1.959632054 |
| TX | 1.259546381 |

Table 1: Distance to the Optimal Energy Structure in 2009

# 5 Prediction of the Energy Profile

## 5.1 Analyse the model in theory

In this section, our task is to predict the energy profile of the states in 2025 and 2050 based on the previous models and analysis. Here we assume there is no influence of policy factors. For ARMA, if the time series is not stationary, the forecasting result would be greatly deviated from the actual result. In order to improve the accuracy of the prediction, we apply the Autoregressive Integrated Moving Average model (ARIMA) into our prediction. The basic idea of ARIMA is similar to that of ARMA. Here we only focus on the improvements of ARIMA: a. Verify that it is a stationary time series. b. Differentiate several times until a stationary time series is obtained.

### 5.1.1 Unit root test

Unit root test is to verity whether there is a time series is stationary. We use Augmented DickeyFuller test to help us to do this.The testing procedure for the ADF test is applied to the model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{16}$$

where is a constant, is the coefficient on a time trend and p is the lag order of the autoregressive process. Imposing the constraints =0 and =0 corresponds to modelling a random walk and using the constraint =0 corresponds to modelling a random walk with a drift. The unit root test is then carried out under the null hypothesis =0 against the alternative hypothesis of <0 .Once a value for the test statistic

$$DE_{\tau} = \frac{\gamma}{SE(\gamma)} \tag{17}$$

is computed it can be compared to the relevant critical value for the DickeyFuller Test. If the test statistic is less than the critical value, then the null hypothesis of =0 is rejected and no unit root is present.

### 5.1.2 Differential processing

Differencing in statistics is a transformation applied to time-series data in order to make it stationary. A stationary time series' properties do not depend on the time at which the series is observed. In order to difference the data, the difference between consecutive observations is computed. Mathematic ally, this is shown as

$$y_t^{'} = y_t - y_{t-1} \tag{18}$$

Differencing removes the changes in the level of a time series, eliminating trend and seasonality and consequently stabilizing the mean of the time series. Sometimes it may be necessary to difference the data a second time to obtain a stationary time series, which is referred to as second order differencing

$$y_t^* = (y_t^{'} - y_{t-1}^{'}) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2} \tag{19}$$

## 5.2 Results of prediction

We perform the steps above for every variable defined in the energy profile vector. Here we take Texas as an example and make a prediction on its total expenditure in energy for the next 40 years.

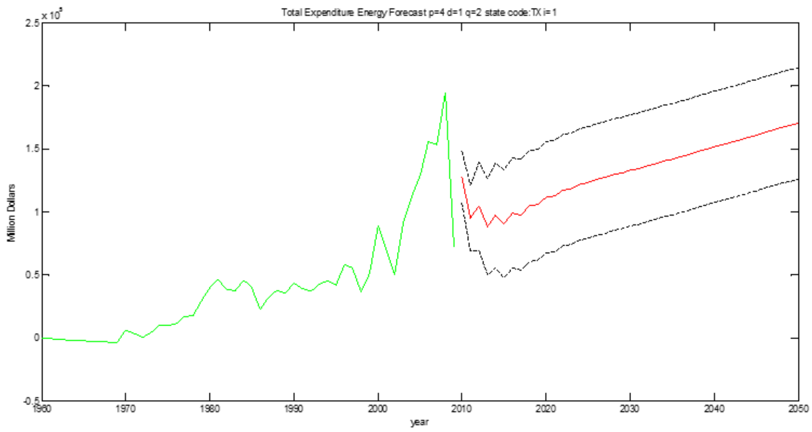


Figure 9: Prediction for Texas's total expenditure in the next 40 years

In this figure, d=1 means after one difference a non-stationary time series is gained.

The results of prediction ( unit: millon dollars, billion btu ):

| State | Year | Total_e | Pri_e | Tran_e | Com_e | Ind_e | Res_e |
|---|---|---|---|---|---|---|---|
| AZ | 2025 | 21006.87 | 14251.61 | 10482.30 | 3760.81 | 2244.39 | 9665.54 |
| | 2050 | 28574.90 | 20962.24 | 10676.24 | 5312.61 | 3188.34 | 22223.65 |
| CA | 2025 | 125992.80 | 68388.60 | 85197.36 | 25686.44 | 10135.80 | 26198.63 |
| | 2050 | 170189.30 | 77456.77 | 122185.7 | 37387.19 | 9894.23 | 37956.90 |
| NM | 2025 | 7737.28 | 5826.13 | 4158.86 | 1490.87 | 1049.24 | 1772.00 |
| | 2050 | 10963.24 | 8314.34 | 5699.15 | 2154.13 | 1473.41 | 2612.13 |
| TX | 2025 | 148819.20 | 113354.30 | 66785.96 | 18221.39 | 54116.71 | 26747.38 |
| | 2050 | 215708.5 | 166407.20 | 97600.04 | 26999.95 | 80928.18 | 40114.82 |
| | | | | | | | |
| State | Year | Total_c | Clean_c | Tran_c | Com_c | Ind_c | Res_c |
| AZ | 2025 | 1875358 | 547483 | 589738 | 446525 | 236434 | 523224 |
| | 2050 | 2496683 | 688726 | 746526 | 610730 | 266908 | 720405 |
| CA | 2025 | 9014439 | 1277306 | 3591687 | 1936050 | 1897985 | 1730776 |
| | 2050 | 11088716 | 1571148 | 4362116 | 2446112 | 1898919 | 2011701 |
| NM | 2025 | 823607 | 47112 | 263375 | 136019 | 245217 | 145319 |
| | 2050 | 1020719 | 61957 | 318757 | 157198 | 291836 | 180599 |
| TX | 2025 | 13174244 | 1169064 | 3408834 | 1900912 | 5977306 | 2061235 |
| | 2050 | 16376489 | 1628761 | 4394668 | 2545992 | 6359091 | 2679410 |

Figure 10: Prediction for the energy profile in 2025 and 2050

Also, we use the predicted data to caculate the criteria defined in section 4 and gain their values for 2025 and 2050.

| State | Year | Index 1 | Index 2 | Index 3 | Index 4 |
|---|---|---|---|---|---|
| AZ | 2025 | 0.246745 | 0.604370 | 0.421072 | 0.078419 |
| | 2050 | 0.170107 | 0.633281 | 0.512517 | 0.096750 |
| CA | 2025 | 0.152441 | 0.666364 | 0.258603 | 0.373353 |
| | 2050 | 0.179998 | 0.640072 | 0.304670 | 0.373465 |
| NM | 2025 | 0.072258 | 0.702513 | 0.299447 | 0.029456 |
| | 2050 | 0.096216 | 0.681310 | 0.385583 | 0.027123 |
| TX | 2025 | 0.126733 | 0.724737 | 0.268853 | 0.534854 |
| | 2050 | 0.169627 | 0.723198 | 0.346115 | 0.545043 |

Figure 11: Prediction for the criterias in 2025 and 2050

# 6 Renewable Energy Usage Target

We have already defined our own criteria for determining the best energy profile and made predictions on the energy profile of each state in the year 2025 and 2050. In this section we intend to determine the renewable energy usage targets for 2025 and 2050 of each state. Targets are discussed separately for each state. In the last subsection we will state them as goals for this new four-state energy compact.

## 6.1 Target for Arizona

Energy profile prediction of Arizona has been shown in the figures. For the year of 2025 and 2050 we present the energy profile in detail.

According to our prediction, the Clean Energy Rate for 2025 and 2050 is 0.291935311 and 0.275856303 respectively. After a simple calculation, we estimate the appropriate renewable energy usage target for Arizona is 547483 and 688725 Billion Btu respectively.

According to our comparison between the four states, we find that although Arizona has a smaller proportion of total energy use, it has a comparatively high rate of renewable energy use. Thus the target for the Arizona is to simply hold its status.

## 6.2 Target for California

As for California, its remarkable characteristic is the both high Energy Amount Rate and high Clean Energy Rate. While certain caution should be taken to our policy maker. The main target is still maintain the Clean Energy Rate and seek ways to increase the proportion of clean energy. Calculation could be done similarly as that of Arizona and we could draw the conclusion that the appropriate renewable energy usage target for California is approximately 1277305 and 1571147 Billion Btu for the year 2025 and 2050 respectively.

## 6.3 Target for New Mexico

Contrary to the characteristics of California, the state of New Mexico is characterized by the both low Energy Amount Rate and Clean Energy Rate. The most significant problem for New Mexico is the lowest Clean Energy Rate compared with other three states. Thus raising clean energy must be placed at the highest priority. One appropriate way to do this it to import renewable energy from other clean energy surplus state to keep a balance. After the calculation we estimate the renewable energy usage target to be 47113 and 61957 Billion Btu for the year 2025 and 2050 respectively.

## 6.4 Target for Texas

Remember that we think the energy structure in Texas is the best according to our own set of criteria. However, according to the scatter plot of criteria we can see that the Clean Energy Rate of Texas is just a little higher than the New Mexico. Texas is also characterized by its highest Energy Amount Rate. Thus we should take the problem of rising the Clean Energy Rate of Texas into thorough consideration. It is not difficult to imagine what if a state with the highest Energy Amount Rate has a comparatively low Clean Energy Rate. We could tackle this problem by increasing the production of clean energy and promoting the usage of renewable energy. After the calculation we estimate the renewable energy usage target to be 1169063 and 1628760 Billion Btu for the year of 2025 and 2050 respectively.

### 6.5 Energy Compact Goals

All in all, here we state the energy compact goals for the four states as listed in the table and shown in the graph.

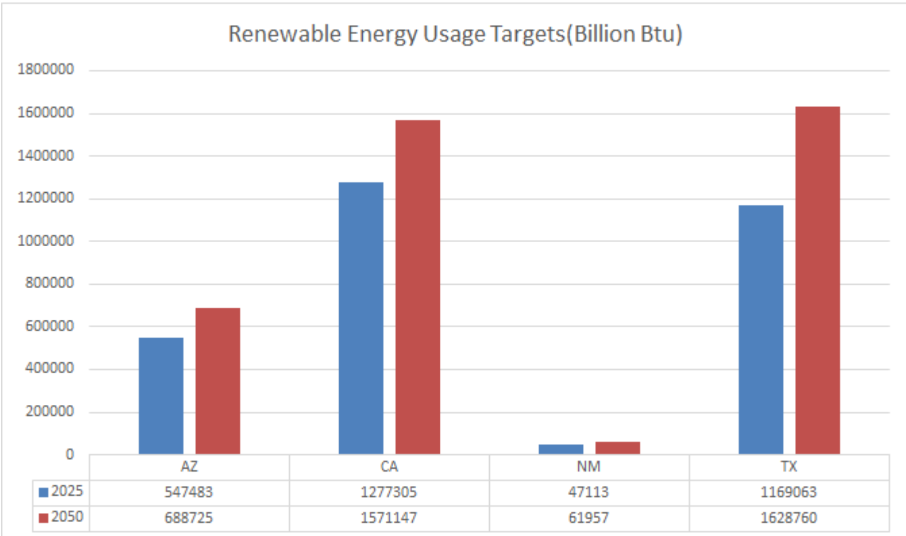


Figure 12: Renewable Energy Usage Targets

| State | Renewable Energy Usage Targets(Billion Btu) | |
|---|---|---|
| AZ | 547483 | 688725 |
| CA | 1277305 | 1571147 |
| NM | 47113 | 61957 |
| TX | 1169063 | 1628760 |

Table 2: Renewable Energy Usage Target for 2025 and 2050

## 7 Actions to be Taken

Here we briefly introduce some actions could be taken to meet the energy compact goals.

- Each state should try to develop their own strength and help each other. For example, Arizona has a high Clean Energy Rate while its total energy production is rather small. Hence the renewable energy could be transferred to the state that suffers from less usage of clean energy such as New Mexico.

- Increase energy use while ensuring increased clean energy use. A sound way to do this is that develop renewable energy first so that the Clean Energy Rate is increased along the increment of total energy usage.
- Try to reduce the loss energy rate. Apparently reduce energy loss could promote efficiency which leads to a better energy structure.
- Try to reduce the direct use of energy. This action can reduce the Primary Energy Rate. The heavy use of secondary energy can reduce the environmental pollution and help to achieve the goal of clean energy usage.

# 8 Sensitivity Analysis and Validation

For SARMA model, we use two parameters, that is, p and q. As we have mentioned in that part, we choose p=4 and q=2. In order to measure the effect of the selection, we introduce the goodness of fitting$--R^2$, which is defined as follows,

$$R^2 = (\sum_{i=1}^{L}(\frac{1-(predict(i,1)-practical(i,1))^2}{practical(i,1)^2}))/L \tag{20}$$

where i represents a year. As the figure shows, the fit is good. However, considering the data is a series of discrete points,$R^2$ can effectively help us quantify the degree of fitting. The closer $R^2$ is to 1, the higher the degree of fitting. For various p and q, we calculate the $R^2$ of AZ's total expenditure.

| | p | q | p | q | p | q | p | q |
|---|---|---|---|---|---|---|---|---|
| | 4 | 2 | 4 | 1 | 2 | 2 | 2 | 1 |
| **$R^2$** | 0.9891 | | 0.9875 | | 0.9879 | | 0.9889 | |

Figure 13: The value of $R^2$ for various p and q

All the results are close to each other as well as to 1. Thus, small changes of p and q will not affect the goodness of fitting and we have a very high degree of fitting by applying SARMA model.

For ARIMA model, we add a parameter d. We first have a look at the effect of prediction in accordance with the method introduced in part D to find proper d. To test our model, we collect data of four states' total consumption in 2015.

From the Figure 14, we could find that the predictive value and the actual value are fairly close. It proves that our model could efficiently predict the data in the future. While some external factors can also cause considerable impact like some policy changes, we could confirm that the model can be used for accurate prediction in the absence of great policy changes. We analyze the differences of the predictive values caused by a change of d and get this figure(when d=1, we get our practical prediction).

From the Figure 15, we can see that the influence of parameter d is very large, so we should strictly choose the proper value for this parameter.

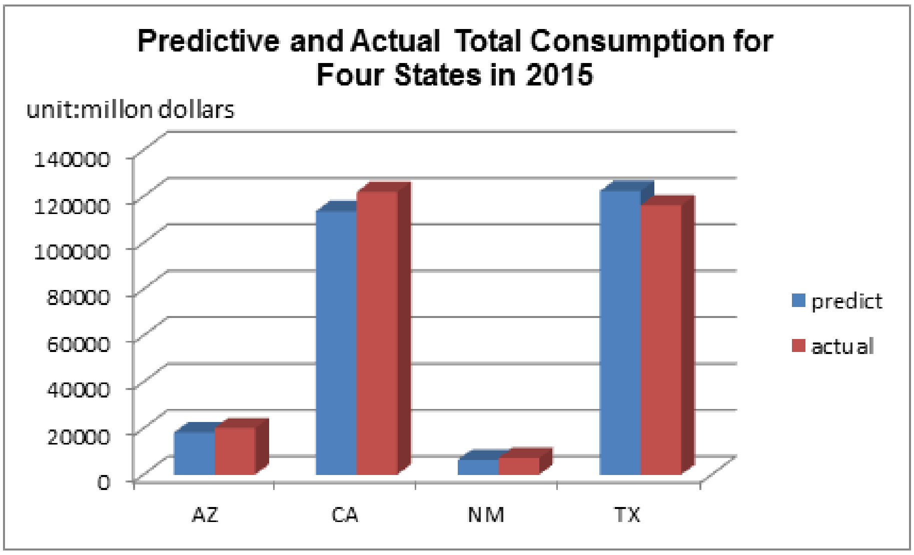


Figure 14: The comparison of the predictive value and the actual value

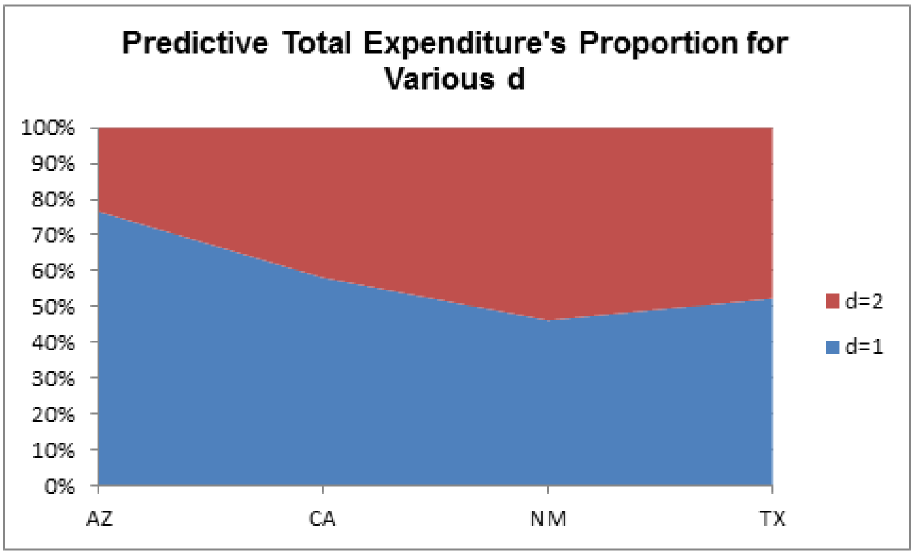


Figure 15: The comparison of predictive value for various d

# 9 Strengths and weaknesses

## 9.1 Strengths

- **Data processing**
  When working with the data, we extract valid data from it. The data is organized into different categories, according to the subject requirements and filtering out the reasonable data. By this means, we make the most use of the data attached.

- **Reasonable multi criterias**
  We creatively put forward a few criterias to measure the energy profile. By using a series of reasonable methods to handle this criterias, we could get a convincing assessment for four states especially for clearer and renewable energy.

- **Accurate fitting and prediction**
We use SARMA and ARIMA model to respectively chacaterize the evolution of data and predict the data in the future. By data analysis, both of them achieve good results.

- **Clear modeling process**
We give a clear process when introducing a model so that we could properly apply the models to practical data. Take the specific requirements into consideration, we choose the most effective method to model based on some changes to original model.

### 9.2 Weaknesses

- **Data vacancy**
There are too many vacancies in the data attached, so we gave up some useful data and fill in the blanks with the mean value for the data with less vacancy, which may be viewed as a main source of errors.

- **Empirically considering criterias**
In our assessment of energy profiles, we empirically set every criteria a weight, which means its degree of importance. We look forward to take a better way to quantify the importance of different criterias.

- **Prediction's limitations**
Although our model is suitable to predict, it has several limitations. On one hand, it doesn't consider some special factors like energy policy changes, economic crisis and so on, which may cause a huge impact in a short period. On the other, while it accurately predict the data in 2015, it may have more errors when predicting longer-term data, which cannot be tested now, in 2018.

## 10 Conclusions

In our paper, we have done lots of data processing work, defined the energy profile and its corresponding evaluation model, put forward SARMA and ARIMA model to respectively characteriz the evolution and predict the energy profile. We use many theory and methods such as AIC, ADF, difference, euclidean metric and so on to help us to complete our work. Meanwhile, we find all these models have great efficiency and accuracy.Based on their results, we analyse various energy circumstances, especially the clear energy, for four states and put forward some reasonable advice.

## References

[1] Ellabban, Omar; Abu-Rub, Haitham; Blaabjerg, Frede (2014). "Renewable energy resources: Current status, future prospects and their enabling technology". Renewable and Sustainable Energy Reviews. 39: 748−764 [749]. doi:10.1016/j.rser.2014.07.113

[2] Brockwell, P. J.; Davis, R. A. (2009). Time Series: Theory and Methods (2nd ed.). New York: Springer. p. 273. ISBN 9781441903198

[3] Zhou Yang, Wu Wenxiang, Hu Yin, Zhang Yushu. Forecast and Analysis of China 's Energy Supply and Consumption. Energy and Environment. ISSN1672-9064. CN35-1272/TK

[4] Ma Yue, Dong Zhou. Research and Configuration of Distributed Energy System. Energy Conservation,Vol.4(2011)

[5] Aho, K.; Derryberry, D.; Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. Ecology. 95: 631−636, doi:10.1890/13−1452.1

[6] Wikipedia: Augmented Dickey−Fuller test https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

[7] Sandra Lach Arlinghaus, PHB Practical Handbook of Curve Fitting. CRC Press, 1994

[8] U.S. Energy Information Administration https://www.eia.gov/state/seds/seds-data-complete.php?sid=US

# Appendices

## Appendix A First appendix

MEMORANDUM

To: the Governor From: Team 92564 Date: February, 2018 Subject: Suggestions for the energy compact to adopt

Dear Governor, We've attached our development plan to this email, but we also want to quickly discuss the trends we've noticed in the energy data over the past few years. The rate of clean energy utilization in Arizona is the highest among the four states, followed by CA, TX and NM respectively. Such difference in energy utilization can be attributed to the geographical and climatic characteristics in these states. Texas and New Mexico depends more on traditional energy over the past 50 years, while Arizona and California have made great efforts to encourage the use of clean energy. Based on the data over the 50-year period, we've made a prediction on the profiles in 2025 and 2050 respectively. The result has shown that the gap between these states on renewable energy use would be wider over time. However, the extreme disequilibrium of energy use among states can cause serious problems. To bridge the gap between states, we have put forward the following suggestions:

Each state should develop their own strength and help each other. For example, Arizona has a high Clean Energy Rate while its total energy production is rather small. Hence the renewable energy could be transferred to the state that suffers from less usage of clean energy such as New Mexico. Increase energy use while ensuring increased clean energy use. A sound way to do this is that develop renewable energy first so that the

Clean Energy Rate is increased along the increment of total energy usage. Reduce the loss energy rate. Apparently reduce energy loss could promote efficiency which leads to a better energy structure. Reduce the direct use of energy. This action can reduce the Primary Energy Rate. The heavy use of secondary energy can reduce the environmental pollution and help to achieve the goal of clean energy usage.

Please let me know if you have any questions.

Best, Team 92564

## Appendix B Second appendix

model for characterizing the evolution **Input MATLAB source:**

```
function [yp,phi,zeta]=mysimulate(y1)

z=iddata(y1);
test=[];
for p=1:5
     for q=1:5
         m=armax(z(1:length(y1)),'na',p,'nc', q);
         AIC=aic(m)
         test=[test;p q AIC];
     end
end
for k=1:size(test,1)
   if test(k,3)==min(test(:,3))
       p_test=test(k,1);
       q_test=test(k,2);
       break;
   end
end
m=armax(z(1:length(y1)),'na',5,'nc',2);
yp =predict(m,z,1);
cc=y1-yp.outputdata
RR=0
for i=1:length(cc)
    RR=RR+(1-cc(i,1)^2/y1(i,1)^2)
end
R=RR/length(cc)
phi=m.a'
zeta=m.c'

plot(y1,'-.');
hold on
plot(yp,'r')
po=yp.outputdata
print po
grid
legend('Original Data','Forecasting Data ARMA')
title(['elvolve p=' num2str(p_test)  ' q=' num2str(q_test)])
```

model for prediction **Input MATLAB source:**

```
function [yp]=mypredict(y1)

t=41
```

```
d=1
y=diff(y1)
x=adftest(y)
i=1
while(x==0)
    temp=y
    y=diff(temp)
    x=adftest(y)
    i=i+1
end
z=iddata(y);
z1=iddata(y1);

test=[];
for p=1:4
     for q=1:3
         m=armax(z(1:length(y)),'na',p,'nc', q);
         AIC=aic(m)
         test=[test;p q AIC];
     end
end
for k=1:size(test,1)
   if test(k,3)==min(test(:,3))
       p_test=test(k,1);
       q_test=test(k,2);
       break;
   end
end
model=arima(p_test,d,q_test)
fit=estimate(model,y1)
res=infer(fit,y1)
figure
subplot(2,2,1)
plot(res./sqrt(fit.Variance))
title('Standardized Residuals')
subplot(2,2,2)
qqplot(res)
subplot(2,2,3)
autocorr(res)
subplot(2,2,4);
parcorr(res)
[Yf,YMSE] = forecast(fit,t,'Y0',y1);
UB = Yf + 1.96*sqrt(YMSE);
LB = Yf - 1.96*sqrt(YMSE);
yp=Yf
figure
h1 = plot(y1,'Color',[.75,.75,.75]);
hold on;
endtime = length(y1)+t-1;
h2 = plot(length(y1):endtime, Yf, 'r', 'LineWidth', 2);
h3 = plot(length(y1):endtime, UB, 'k--', 'LineWidth', 1.5);
plot(length(y1):endtime, LB, 'k--', 'LineWidth', 1.5);
set(gca,'XTick',1:10:N);
set(gca,'XTickLabel',datestr(dates(1:10:N),17));
legend([h1,h2,h3],'Initial Data','Forecast',...
        'Forecast Interval');
title(['Energy Forecast p=' num2str(p_test) ' d=' num2str(d) ' q=' num2str(q_test)]);
```