For office use only
T1 ________________
T2 ________________
T3 ________________
T4 ________________

Team Control Number
93490

Problem Chosen
C

For office use only
F1 ________________
F2 ________________
F3 ________________
F4 ________________

2018
MCM/ICM

Summary

In any times and any country, the energy production and usage are a major portion of any economy. In 1970, 12 western states in the U.S. formed the Western Interstate Energy Compact. Following it, California, Arizona, New Mexico and Texas wish to form a realistic new energy compact focused on increased usage of cleaner, renewable energy sources. In order to accomplish it, we must comprehend the historical trend of energy profile evolution, evaluate the current situation of cleaner, renewable energy usage and forecast the future. Our solution consists of five main sections:

1. Classification, extraction of the raw data provided and creation of energy profiles. The raw data contains 605 variables, and we must simplify the analysis objects. By analyzing the energy sources and features and checking the meaning of the MSN code, we choose five main energy types including coal (CL), natural gas (NG), petroleum products (PM), nuclear electric power (NU) and renewable energy (RE). We extract the consumption of these types of energy for the four states from the raw data, and create energy profiles (the proportion and amount of different types of energy from 1960 to 2009) for them.

2. Establishment of models to help understand the evolution of energy in the four states easily, especially the cleaner, renewable energy. Firstly, to capture the feature of energy profile evolution, we introduce information entropy in information theory to the energy profile to describe the diversity and complexity of it. We find the energy profiles are becoming more and more diversified, and the structures are becoming more and more balanced. Next we utilize the multiple linear regression model to simulate the evolution of information entropy and consumption of renewable energy. We manage to seek possible relevant factors to explain the differences of the four states. These factors belong to different classes, including geographic, industrial, economic, demographic and climatic ones. The R^2 of regression of information entropy and RE on these factors all reach 0.96, shows a highly significant relativity of regression equation. By comparing the P value of different factors or the coefficients of the normalized factors, we detect the factors that are strongly relevant to the energy profile evolution and renewable energy usage, and explain the influence mode of the significant factors and the differences of these factors of the four states.

3. Building a criterion to describe the best profile for use of cleaner, renewable energy in 2009.

4. Forecast of the energy profile in 2025 and 2050.

5. Setting goals of renewable energy usage in 2025 and 2050 and suggesting actions to reach the goals.

Energy production

Team 93490

February 12, 2018

Contents

1 Part I **1**

1.1 A. Energy profile of the four states 1

1.1.1 Classification, extraction of raw data and creation of energy profile 1

1.1.2 Summary of A 3

1.2 B. Analysis of the energy profile evolution 3

1.2.1 Introduction of information entropy 3

1.2.2 Summary of information entropy 4

1.2.3 Regression analysis of energy consumption evolution 4

1.3 C. Criteria of profile for use of cleaner, renewable energy 9

1.3.1 Introduction of Analytic hierarchy process 9

1.3.2 Evaluation of energy profile 11

1.4 D. Energy profile prediction 13

1.4.1 Introduction of ARIMA model 13

1.4.2 Prediction of energy profile 14

2 Part II **16**

1 Part I

1.1 A. Energy profile of the four states

1.1.1 Classification, extraction of raw data and creation of energy profile

To help my team analyze date and build the model as well as make four governors understand easily the next analysis we will address, simplifying these date becomes quite eager.

State Energy Data System (SEDS) describes the data identification codes exhaustively. The MSN (Mnemonic Series Names) of 605 variable names are defined in *State Energy Data System 2015 Consumption Technical Notes* with five characters:

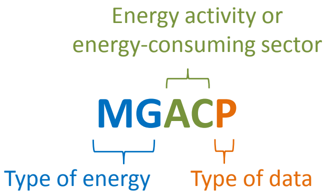


Figure 1: Components of a MSNCODE [1]

As we can see, the first two letters of MSN code are used to represent the energy source and products, the three and four letters of MSN are used to represented the energy-consuming sectors and the last letter identifies the type of data.

Energy sources can be categorized as renewable, nuclear energy, non-renewable sources and net interstate sales of electricity:

- Non-Renewable Sources
  - Fossil fuels
    - coal (CL)
    - net imports of coal coke (U.S. only)
    - natural gas excluding supplemental gaseous fuels (NN)
    - petroleum products excluding fuel ethanol blended into motor gasoline (PM)
- Non-renewable but clear: nuclear electric power (NU)
- Renewable Sources
  - fuel ethanol minus denaturant (EM)
  - geothermal direct use energy and geothermal heat pumps (GE)
  - conventional hydroelectric power (HY)
  - solar thermal direct use energy and photovoltaic electricity net generation (SO)
  - electricity produced by wind (WY)
  - wood and wood-derived fuels and biomass waste (WW)
- Net interstates sales of electricity and associated losses

  This value can both be negative and positive, MSN code=ELISB.

It should be noted that due to the amount of the renewable sources are always small, we sum all the renewable sources as an energy type RE.

Based on our demand and the actual situation of date provided, we select some main MSN Codes to describe the energy profiles of the four states. To check the correctness of our data, we add the total energy consumption in the data-set. The details of MSN Codes we used are as follows:

Table 1: Meaning of the MSNCODE used for analyzing energy profile

MSN	Description	Unit
CLTCB	Coal total consumption.	Billion Btu
NGTCB	Natural gas total consumption (including supplemental gaseous fuels)	Billion Btu
PATCB	All petroleum products total consumption	Billion Btu
NUETB	Electricity produced from nuclear power	Billion Btu
EMTCB	Fuel ethanol, excluding denaturant, total consumption	Billion Btu
RETCB	Geothermal energy total consumption	Billion Btu
ELISB	Net interstate sales of electricity and associated losses (negative and positive values)	Billion Btu
TETCB	Total energy consumption	Billion Btu

Additional Explanation: the code of energy-consuming sectors we choose TC, but as for NU, we choose ET.

Through these steps, we finally get the energy consumption proportion and amount of the four states from 1960 to 2009, as separately shown in figure 2 and figure 3. In all the four states, the equation below continuously holds:

$$\mathrm{TE} = \mathrm{CL} + \mathrm{NG} + \mathrm{PM} + \mathrm{NU} + \mathrm{RE} + \mathrm{EL} \tag{1}$$

Note: when EL is negative, which means the state exports electricity to other states, the proportion of $\mathrm{CL}+\mathrm{NG}+\mathrm{PM}+\mathrm{NU}+\mathrm{RE}$ to the total energy will exceeds 1, and the area above 1 is the proportion of EL. When EL is positive, the blank area below 1 is the proportion of EL. In general, the total consumption of all energy-consuming sectors increase with fluctuations, and Arizona increase fastest while New Mexico slowest. The amount of Texas and California, however, is several times that of Arizona and New Mexico. Natural gas (NG) and petroleum products (PM) are the main energy, but coal (CL) is important energy in Arizona and New Mexico still. Most of cleaner, renewable energy sources (RE) like electricity produced by wind and conventional hydroelectric power is in its infancy, but California and Arizona have a great beginning already. Specially, the Nuclear electric power (NU) has played an important role in Arizona. Besides, in the view of interstate electricity exchange, California imports more and more electricity, while Arizona and New Mexico export a relatively large proportion of electricity. Texas is nearly self-sufficient.

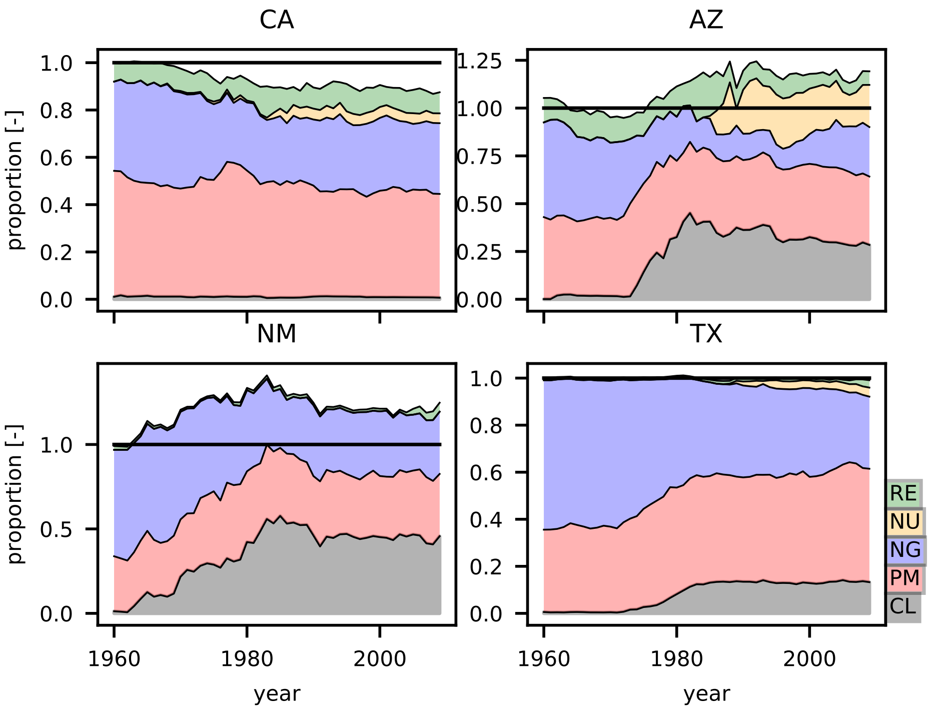


Figure 2: Proportion of five main energy types of the four states from 1960 to 2009

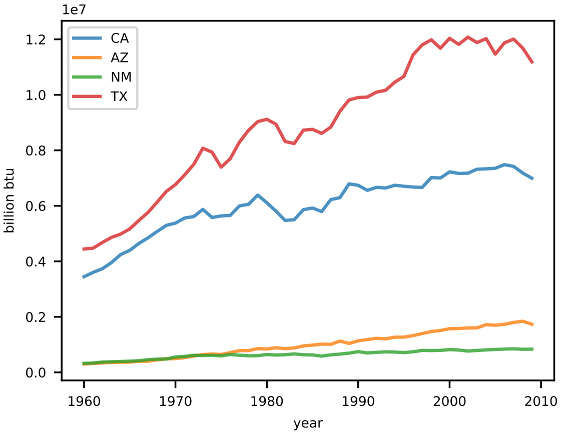


Figure 3: Total energy consumption of the four states from 1960 to 2009

1.1.2 Summary of A

- California: the energy is tending to be diverse, but the proportion of Nuclear electric power (NU) could be bigger.
- Arizona: the energy is tending to be diverse, and the amount increases fastest in all, but the amount is too small.
- New Mexico: the main energy is fossil fuels but there is little cleaner, renewable energy sources (RE).
- Texas: the main energy is natural gas (NG) and petroleum products (PM) and there is a little cleaner, renewable energy sources (RE).

1.2 B. Analysis of the energy profile evolution

1.2.1 Introduction of information entropy

Energy profile of different areas are always quite different because of different energy storage situation and economic level. Therefore, it's hard to adopt a specific index the quantify the profile evolution. As Hollis B. Chenery said, economic development is accompanied by the structural transformation. This structural transformation is mainly embodied in the next two aspects: the industrial structure upgradation and energy profile evolution. As the economic development, cleaner and higher-efficiency energy will replace the traditional and lower-efficiency energy. Commonly, also as figures of the four states' energy profile shows, the natural gas, nuclear energy and renewable energy take higher proportion.

Here we utilize the concept of *information entropy* to quantificationally describe the energy consumption evolution. Entropy is a state function to describe the irreversibility of spontaneous process, defined based on the second law of thermodynamics [9]. Boltzmann gave the following entropy function:

$$S = K_B \ln P \tag{2}$$

where K_B is Boltzmann constant, P is the probability that the system is in a certain state. In 1908, Shannon introduced the concept of entropy to the information theory and defined it as information entropy H:

$$H(X) = E[I(X)] = E[-\ln(P(X))] \tag{3}$$

where E is the expected value operator, and I is the information content of X. $I(X)$ is itself a random variable [3]. The equation above is called Shannon Equation, it can describe the degree of randomness and disorder of any system or material [4].

Following what Shannon has done, we introduce the information entropy to the energy profile analysis. The energy consumption system is a non-linear open system that has wide material, energy and information exchange with the environment. As time passes, it's structure shows a spontaneous and irreversible evolution, which satisfies the consumption of dissipative structure system. For example, consider an energy profile with m types of energy (the unit must be uniform), and the quantity of ith energy is H_i, thus is proportion is $P_i = H_i / \sum H_i = H_i/H$. Therefore, the information entropy of energy profile is:

$$S = -\sum_{i=1}^{m} P_i \ln P_i \tag{4}$$

Particularly, when there is only one type of energy in the system, $S_{\text{min}} = 0$; on the contrary, when all the energy types are equal, $H_1 = H_2 = \cdots = H_m = H/m$, then $S_{\text{max}} = \ln m$. In the real world, such two extreme situation will never appear, thus information entropy is between S_{min} and S_{max}, and it can reflects the complexity of the energy concumption structure.

Further more, we can define the equilibrium and dominance degree. Based on the information entropy equation, the equilibrium degree can be given by:

$$E = -\sum_{i=1}^{m} P_i \ln P_i / \ln m = S/\ln m \tag{5}$$

i.e., the equilibrium degree is the ratio of information entropy to the maximum entropy. Larger E, smaller difference of the proportion of different types of energy. The dominance degree is $D = 1 - E$, and it reflects the support ability of one or several types of energy to the energy consumption, which is the contrast to the equilibrium degree.

1.2.2 Summary of information entropy

Time series of the S, E and D of four states are calculated and is shown in figure 4. It should be emphasized that due to the existence of negative values in ELISB (net interstate sales of electricity and associated losses), it is ignored when calculate the information entropy. In this figure, the information entropy of the four states has a trend of continuously increasing by time. This means the degree diversification and complexity of them are increasing, which may be a result of economic development. Besides, the equilibrium degree always increases and the dominance equilibrium degree always decreases by time, which can also suggest that the energy profile is more balanced.

Combining figure 2 and 4, we can see Arizona state is the most diversified in the four states, with largest proportion of nuclear energy. New Mexico state also has larger information entropy as it has a balanced structure of coal, petroleum and natural gas. California and Texas shows similar diversity, their the proportion of natural gas and petroleum is similar but California consume more renewable and nuclear energy and Texas consume more coal.

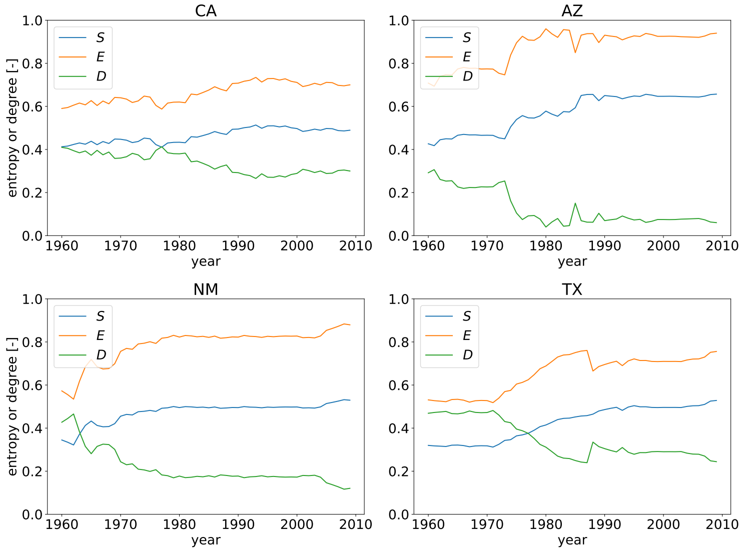


Figure 4: Information entropy S, equilibrium degree E and dominance degree D of four sates from 1960 to 2009

1.2.3 Regression analysis of energy consumption evolution

Here we need to find the factors in the data-set that are relevant to the energy consumption.

1. Relevant factors

   The factors we apply are divided into the next classes:

(a) Geographic factors

These factors are mainly the energy possessed by the region, such as coal, hydroenergy, natural gas, crude oil, related factors in the data-set are:

- CLPRB (coal production)
- HYTCB (hydroelectricity total production)
- NGMPB (natural gas marketed production),
- PAPRB (crude oil production (including lease condensate))

(b) Industrial factors

The energy consumed by different sectors can reflect the their volume. According to the third and forth alphabet of the MSN code, we select the total energy (TE) of:

- AC transportation sector consumption, MSN code=TEACB
- CC commercial sector consumption, MSN code=TECCB
- EG electric power sector generation (also generation), MSN code=TEEGB
- EI electric power sector consumption, MSN code=TEEIB
- IC industrial sector consumption, MSN code=TEICB
- RC residential sector consumption, MSN code=TERCB

(c) Economic factors

As discussed above, the economic development means the structural transformation, which includes the energy profile evolution. Economic factors include GDP (MSN code=GDPRX), GDP growth rate (calculated from GDP: $\mathrm{GR}_i = \frac{\mathrm{GDP}_{i+1}-\mathrm{GDP}_i}{\mathrm{GDP}_i}$) and total energy consumption per real dollar of GDP (MSN code=TETGR). It might also be noted that GDP data starts from 1977, and years before that GDP and it's growth rate is treated as zero.

In addtion, the average prices of different types of energy are also considered. These include:

- CLTCD (coal average price, all sectors)
- NGTCD (natural gas average price, all sectors (including supplemental gaseous fuels)
- PATCD (all petroleum products average price, all sectors)
- NUETD (nuclear fuel average price, all sectors)
- ESTCD (electricity average price, all sectors).
- Renewable energy average price, all sectors (with similar MSN code RETCD). This is not supplied in the data-set, and is calculated by the following equation:

$$\mathrm{RETCD} = \frac{\mathrm{TETCV} - \sum \mathrm{XTCB} \cdot \mathrm{XTCD}}{\mathrm{RE}} \tag{6}$$

where X can be CL, NG, PA or NU, and TETCV is total energy expenditures.

(d) Demographic factors

Population of the a state may affect the amount of the energy consumption if the difference of that of the individuals is not too large. MSN code here is TPOPP (resident population including armed forces).

(e) Climate factors

Climate influences the energy consumption by influencing people's behaviour. For example, the temperature. When it gets hotter, the air conditioning may run; when it becomes colder, the heating may work. As another example, the precipitation. When the state suffer drought, pumps may launch; when the state meet flood, life and production activities may be impeded and energy consumption may be reduced. NCEI (National Centers For Environmental Information) has collected climate and weather datasets (https://www.ncdc.noaa.gov). We adopt annually averaged temperature and precipitation from four stations in four states from U.S. Local Climatological Data, including station Tucson for Arizona, station Bakersfield for California, station Albuquerque for New Mexico, station Abilene for Texas.

2. Multiple linear regression model

The data set here is $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n years (n=50 and p=21, i.e., number of the factors). The linear regression assumes the relationship between the development variable y_i and the p-vector of regressors x_i is linear. Before establishing the regression model, all the data should be normalized. The normalized variable x_i' is given by:

$$x_i' = \frac{x_i - \bar{x}_i}{std(x_i)} \tag{7}$$

The normalization treatment can prevent the larger variable from hiding the influence of smaller variable.

The multiple linear regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \beta + \varepsilon_i \tag{8}$$

where β_i is the coefficient, ε_i is random variable. The vector form of the n equations is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{9}$$

where:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Select an estimation of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}$ by using the Ordinary Least Squares (OLS) estimator (minimize the sum of the squares of $\boldsymbol{\varepsilon}$), namely:

$$\min \sum \boldsymbol{\varepsilon}^2 = \min (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{10}$$

The solution of the equation above is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \left(\sum \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum \mathbf{x}_i y_i\right) \tag{11}$$

Finally, we can get the estimation of the dependent variable:

$$\hat{\mathbf{Y}} = \mathbf{X}^\top \hat{\boldsymbol{\beta}} \tag{12}$$

3. Regression results of energy profile evolution

To characterize the evolution of the energy profile of each of the four states, we first regress information entropy on all the possible factors listed above. Figure 5 shows the real and the estimated data. All the R^2 reach 0.96, suggesting that the information entropy is strongly linear correlated to the factors. Illustrated by the example of Arizona, the regression overview in Stata is shown in figure 6.

- P value of the F test is 0.0000, which shows a highly significant relativity for regression equation.
- Generally, if the P value of the t test of a variable is smaller than 0.05 (with a confidence level of 95%), we can refuse the hypothesis that the variable is not related to the dependent variable. These include: CLPRB, HYTCB, NGMPB, NUETD, RETCD, TPOPP, TEEIB, TERCB.
- Since all the factors has been normalized, their coefficients can reflect their contribution rate to the dependent variable. Top 3 variables of that are TERCB (0.206), TEEIB (0.194), TPOPP (0.166), which means the information entropy (the diversity or complexity of the energy profile) has a strong positive correlation with total energy consumed by the residential, total energy consumed by the electric power sector and residential population. This is an evidence of the economic development influences the energy profile.

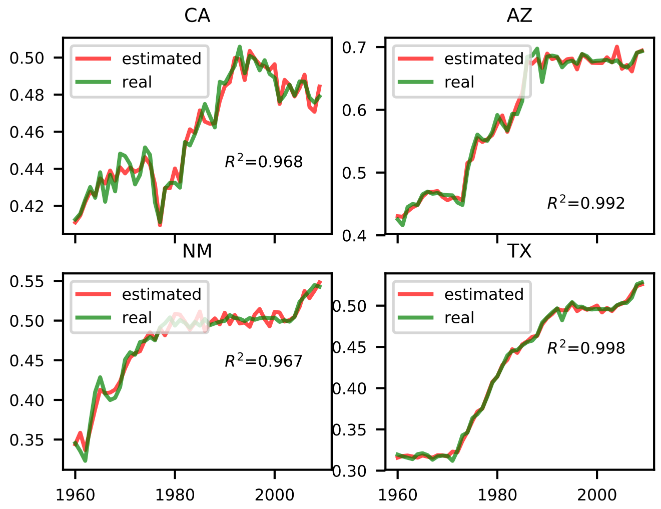
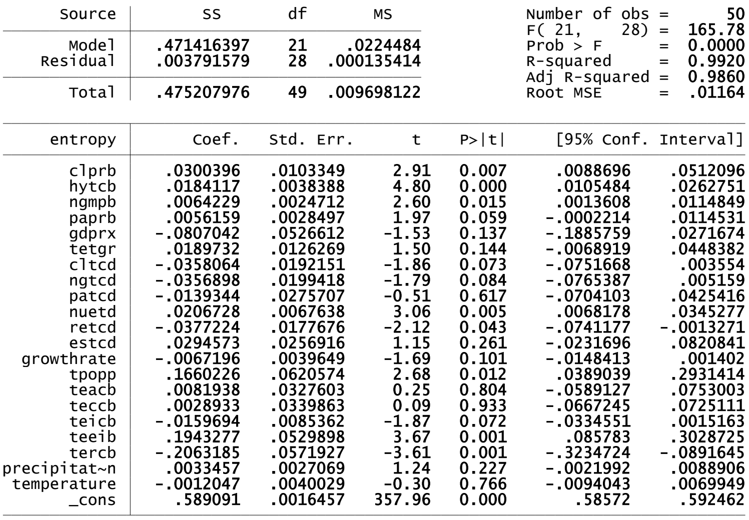


Figure 5: The real and estimated information entropy of the four states from 1960 to 2009



```
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F( 21,    28) =  165.78
       Model |  .471416397    21   .0224484            Prob > F      =  0.0000
    Residual |  .003791579    28  .000135414           R-squared     =  0.9920
-------------+------------------------------           Adj R-squared =  0.9860
       Total |  .475207976    49  .009698122           Root MSE      =  .01164

------------------------------------------------------------------------------
     entropy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       clprb |   .0300396   .0103349     2.91   0.007     .0088696    .0512096
       hytcb |   .0184117   .0038388     4.80   0.000     .0105484    .0262751
       ngmpb |   .0064229   .0024712     2.60   0.015     .0013608    .0114849
       paprb |   .0056159   .0028497     1.97   0.059    -.0002214    .0114531
       gdprx |  -.0807042   .0526612    -1.53   0.137    -.1885759    .0271674
       tetgr |   .0189732   .0126269     1.50   0.144    -.0068919    .0448382
       cltcd |  -.0358064   .0192151    -1.86   0.073    -.0751668     .003554
       ngtcd |  -.0356898   .0199418    -1.79   0.084    -.0765387     .005159
       patcd |  -.0139344   .0275707    -0.51   0.617    -.0704103    .0425416
       nuetd |   .0206728   .0067638     3.06   0.005     .0068178    .0345277
       retcd |  -.0377224   .0177676    -2.12   0.043    -.0741177   -.0013271
       estcd |   .0294573   .0256916     1.15   0.261    -.0231696    .0820841
  growthrate |  -.0067196   .0039649    -1.69   0.101    -.0148413     .001402
       tpopp |   .1660226   .0620574     2.68   0.012     .0389039    .2931414
       teacb |   .0081938   .0327603     0.25   0.804    -.0589127    .0753003
       teccb |   .0028933   .0339863     0.09   0.933    -.0667245    .0725111
       teicb |  -.0159694   .0085362    -1.87   0.072    -.0334551    .0015163
       teeib |   .1943277   .0529898     3.67   0.001      .085783    .3028725
       tercb |  -.2063185   .0571927    -3.61   0.001    -.3234724   -.0891645
precipitat~n |   .0033457   .0027069     1.24   0.227    -.0021992    .0088906
 temperature |  -.0012047   .0040029    -0.30   0.766    -.0094043    .0069949
       _cons |    .589091   .0016457   357.96   0.000       .58572     .592462
------------------------------------------------------------------------------
```

Figure 6: Overview of the regression of information entropy on all the listed factors of Arizona in Stata

4. Regression results of consumption of renewable energy

   Figure 7 shows the real and the estimated renewable energy. All the R^2 reach 0.96, suggest that the information entropy is strongly linear correlated to the factors. Illustrated by the example of Arizona, the regression overview in Stata is shown in figure 8.

   - P value of the F test is 0.0000, which shows a highly significant relativity for regression equation.
   - Significant variables include: HYTCB, PATCD, TEACB, TEEIB, and they all have relatively larger coefficient. Hydroelectricity belongs to renewable energy and is also connected to geographic features. If it plays an important role in renewable energy, it's not surprise that it's coefficient is large. Price of petroleum also positively influence RE, as petroleum is one of the substitute of RE. Total energy consumed by transportation sector and electric power sector both negatively influence RE. However, when we regress RE on these four variables, coefficient of TEEIB turns to be positive.

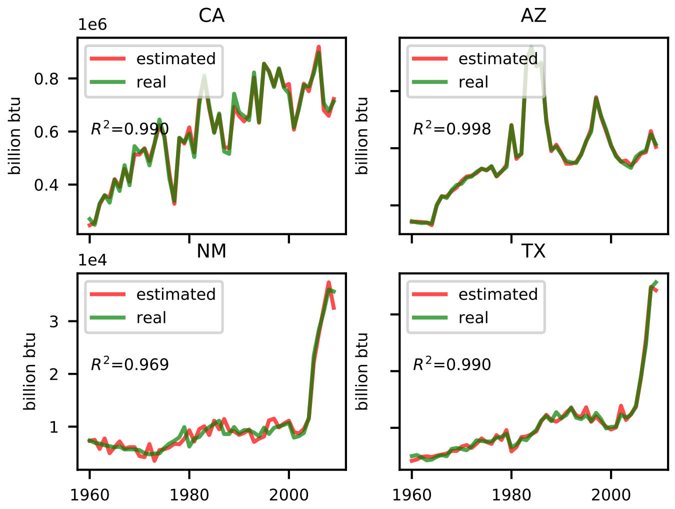
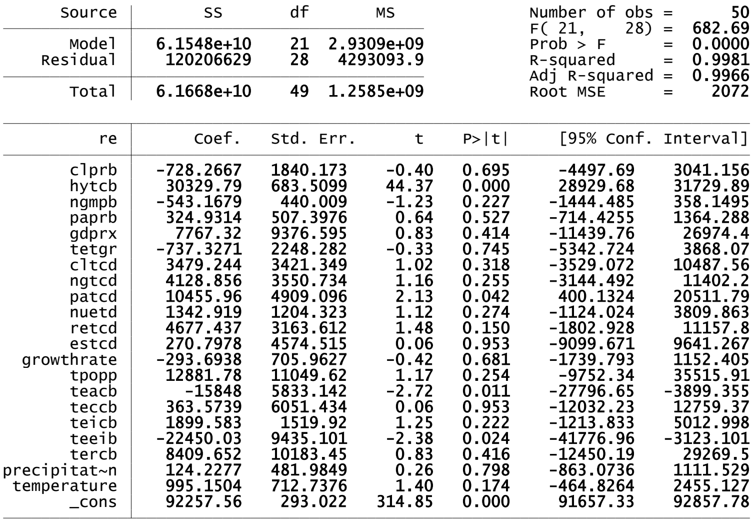


Figure 7: The real and estimated renewable energy of the four states from 1960 to 2009



```
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F( 21,     28)  =    682.69
       Model |  6.1548e+10        21  2.9309e+09   Prob > F        =    0.0000
    Residual |   120206629        28   4293093.9   R-squared       =    0.9981
-------------+----------------------------------   Adj R-squared   =    0.9966
       Total |  6.1668e+10        49  1.2585e+09   Root MSE        =      2072

------------------------------------------------------------------------------
          re |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       clprb |  -728.2667   1840.173    -0.40   0.695     -4497.69    3041.156
       hytcb |   30329.79   683.5099    44.37   0.000     28929.68    31729.89
       ngmpb |  -543.1679    440.009    -1.23   0.227    -1444.485    358.1495
       paprb |   324.9314   507.3976     0.64   0.527    -714.4255    1364.288
       gdprx |    7767.32   9376.595     0.83   0.414    -11439.76     26974.4
       tetgr |  -737.3271   2248.282    -0.33   0.745    -5342.724     3868.07
       cltcd |   3479.244   3421.349     1.02   0.318    -3529.072    10487.56
       ngtcd |   4128.856   3550.734     1.16   0.255    -3144.492     11402.2
       patcd |   10455.96   4909.096     2.13   0.042     400.1324    20511.79
       nuetd |   1342.919   1204.323     1.12   0.274    -1124.024    3809.863
       retcd |   4677.437   3163.612     1.48   0.150    -1802.928     11157.8
       estcd |   270.7978   4574.515     0.06   0.953    -9099.671    9641.267
  growthrate |  -293.6938   705.9627    -0.42   0.681    -1739.793    1152.405
       tpopp |   12881.78   11049.62     1.17   0.254     -9752.34    35515.91
       teacb |     -15848   5833.142    -2.72   0.011    -27796.65   -3899.355
       teccb |   363.5739   6051.434     0.06   0.953    -12032.23    12759.37
       teicb |   1899.583    1519.92     1.25   0.222    -1213.833    5012.998
       teeib |  -22450.03   9435.101    -2.38   0.024    -41776.96   -3123.101
       tercb |   8409.652   10183.45     0.83   0.416    -12450.19     29269.5
precipitat~n |   124.2277   481.9849     0.26   0.798    -863.0736    1111.529
 temperature |   995.1504   712.7376     1.40   0.174    -464.8264    2455.127
       _cons |   92257.56    293.022   314.85   0.000     91657.33    92857.78
------------------------------------------------------------------------------
```

Figure 8: Overview of the regression of renewable energy on all the listed factors of Arizona in Stata

5. Summary of B

We summerize all the variables and coefficients with $P < 0.05$ of the four states in table 2. Followings are our annalysis:

- Price of petroleum influences positively RE of California, Arizona, New Mexico and Texas at the same time. As the main energy of four states, growth of petroleum price makes the demand of petroleum decrease, so the amount of RE will grow as RE is the substitute goods of petroleum.
- In California and Arizona, HYTCB is also an important influencing factors. Hydroelectricity belongs to renewable energy and is also connected to geographic features. If it plays an important role in renewable energy, its not surprise that its coefficient is large.
- In California and Texas, population is a main influencing factors because the population of California and Texas is large. The more energy will be demanded with the larger of population, therefore the coefficient of TPOPP is large.

- In New Mexico and Texas, CLPRB is a main important influencing factors and the CLPRB influence negatively RE. When coal production becomes large, the price of coal will decline and people are more willing to use coal, which results in the decreasing of RE.
- Whats more, the price of renewable energy (RETCD) is significant in New Mexico and Texas. It is obvious that people wont use renewable energy with high price. As the same reason, the amount of RE will increase when the price of coal is going to be higher in Texas.
- Specially in Arizona, total energy consumed by the transportation sector (TEACB) and total energy consumed by the electric power sector (TEEIB) both negatively influence RE. However, when we regress RE on these four variables, coefficient of TEEIB turns to be positive. A possible explanation is that several variable in the list is negative relevant to TEEIB and positive relevant to RE.

Table 2: Variables and their coefficients whose $P < 0.05$ of the four states

CA		AZ		NM		TX	
variable	coef.	variable	coef.	variable	coef.	variable	coef
HYTCB	66649.39	HYTCB	30329.79	CLPRB	-6233.096	CLPRB	-30335.52
GDPRX	-113341.8	PATCD	10455.96	PATCD	7464.639	CLTCD	34015.41
TETGR	37955.93	TEACB	-15848	RETCD	3200.461	PATCD	65595.21
PATCD	104757.7	TEEIB	-22450.03	TERCB	8226.356	RECTD	43622.03
TPOPP	210763					ESTCD	-45093.49
						TPOPP	119151.5

1.3 C. Criteria of profile for use of cleaner, renewable energy

1.3.1 Introduction of Analytic hierarchy process

The analytic hierarchy process (AHP) is a structured technique for organizing and analyzing complex decisions, based on mathematics and psychology [5]. It has particular application in group decision making, especially when several factors make contributions to the result, and significance of each factor cannot be easily recognized and defined. The model is used around the world in a wide variety of decision situations, in fields such as government, business, industry and healthcare. The evaluation of energy profile is just the case.

The first step in the AHP is to model the problem as a hierarchy. An AHP hierarchy is a structured means of modeling the decision at hand. It consists of an overall goal, a group of options for reaching the goal, and a group of criteria that relate the alternatives to the goal. Figure 9 represents an example of a hierarchy with four criteria and three alternatives, each of which is connected with multiple lines. The weight of alternatives on the goal can be determined by the weight of criteria on the goal and the weight of alternatives on criteria.

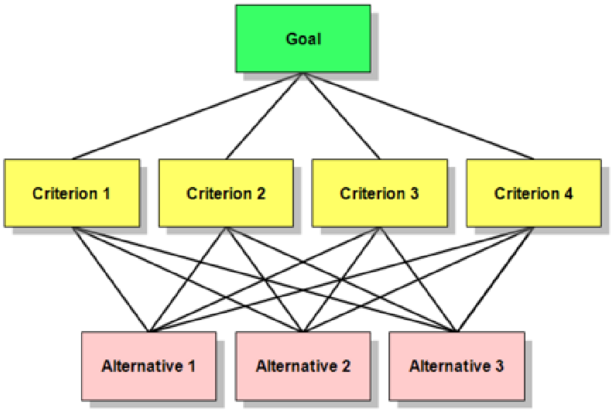


Figure 9: Structure of AHP model

The model can be processed with following steps:

1. Comparisons between criteria

   Let x_1,x_2...x_n be the criteria for decision, then the result of each alternative can be defined as linear combination of the criteria, i.e.,

   $$y = w_1x_1 + w_2x_2 + ... + w_nx_n \tag{13}$$

   where w_1,w_2,...,w_n are weight coefficients of criteria on the goal which satisfy $w_i \geq 0$. To determine the values of w_i, we compare the significance between each two of the x_i, and give numerical scales for the measurement. The number are given by the following method:

   Table 3: Methods for determining the weight

$\frac{x_i}{x_j}$	Meaning
1	x_i and x_j are of same significance
3	x_i is slightly more significant than x_j
5	x_i is more significant than x_j
7	x_i is far more significant than x_j
9	x_i is absolutely more significant than x_j
2,4,6,8	between the two states above
$\frac{1}{x_j/x_i}$	x_i is less significant than x_j

   Using the method above, we can get the $n \times n$ matrix $\mathbf{A}$ in which A_{ij} is the significance of x_i to x_j. Diagonal numbers in matrix $\mathbf{A}$ are all 1, and symmetrical elements are reciprocal. Example:

   $$\begin{bmatrix} 1 & 1/2 & 4 & 3 & 3 \\ 2 & 1 & 7 & 5 & 5 \\ 1/4 & 1/7 & 1 & 1/2 & 1/3 \\ 1/3 & 1/5 & 2 & 1 & 1 \\ 1/3 & 1/5 & 3 & 1 & 1 \end{bmatrix}$$

   In theory, for any $1 \leq i, j, k \leq n$, $A_{ik} = A_{ij}A_{jk}$. If all the elements in matrix $\mathbf{A}$ satisfy the equation, we call matrix $\mathbf{A}$ a consistent matrix. The consistent matrix has only one non-zero eigenvalue n, and all eigenvectors correspond to eigenvalue n. In practice, it is accepted that the equation may not be always true. The consistency test checks whether the matrix falls in acceptable inconsistency.

   Define consistency index: $\mathrm{CI} = \frac{\lambda - n}{n-1}$, where λ is the biggest eigenvalue of matrix $\mathbf{A}$. Define random consistency index RI. The value is defined in the following form.

   Table 4: Value of RI

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

   Define Consistency Ratio $\mathrm{CR} = \mathrm{CI}/\mathrm{RI}$. If $\mathrm{CR} < 0.1$, the matrix $\mathbf{A}$ passes the consistency test. In the example of the matrix given above, n = 5, λ = 5.073, CI = 0.018, RI = 1.12 ,CR = 0.016<0.1. The consistency test passes. The eigenvectors p corresponding to λ is the weight of criteria of the goal.

2. Comparisons between alternatives on the criteria

   For each criterion x_1, x_2, ..., x_n, compare the weight between each two of the m alternatives P_1, P_2, ..., P_n, and give measurement using the numerical scale mentioned above. Then, we get n matrices B_1, B_2, ..., B_n with size $m \times m$. Calculate the maximum eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ and corresponding eigenvectors w_1, w_2, ..., w_n of each matrix, and combine them into vector $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots \lambda_n]$ and matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2 \cdots \mathbf{w}_n]$.

   In the second consistency test, define vector $\mathbf{K} = (\boldsymbol{\lambda} - n)/2$, define $\mathrm{CI}_2 = \mathbf{K}p$, RI_2 is the value corresponding to m in the form. $\mathrm{CR}_2 = \mathrm{CR} + \mathrm{CI}_2/\mathrm{CR}_2$. If $\mathrm{CR}_2 < 0.1$, the second consistency test passes. Define $\mathbf{\Omega} = \mathbf{W}p$. $\mathbf{\Omega}$ is an $m \times 1$ vector, representing the weight of alternatives to the goal, i.e. the result of the evaluation. The alternative with the maximum result can be the best decision.

1.3.2 Evaluation of energy profile

5 criteria – information entropy, greenhouse gas emission, renewable energy percentage, financial cost and energy consumption per capita – are chosen to evaluate the priority of the energy profile. After discussion and comparison, we give the matrix shown as below:

$$\begin{bmatrix} 1 & 0.33 & 0.14 & 0.2 & 0.33 \\ 3 & 1 & 0.33 & 1 & 1 \\ 7 & 3 & 1 & 2 & 2 \\ 5 & 1 & 0.5 & 1 & 1 \\ 3 & 1 & 0.5 & 1 & 1 \end{bmatrix}$$

λ = 5.043, $\mathbf{p}$ = [0.1023, 0.3230, 0.7821, 0.3895, 0.3489]$^\top$, CI = (5.043 - 5) / (5 -1) = 0.011 RI = 1.12 CR = CI / RI = 0.0096 < 0.1, the first consistency test passes. (0.1023, 0.3230, 0.7821, 0.3895, 0.3489) corresponds to the weight of (information entropy, greenhouse gas emission, renewable energy percentage, financial cost and energy consumption per capita) to the goal.

To evaluate the significance of two states on each criteria, we download all the energy data of 50 states (including data of District of Columbia), and calculate the values of 5 criteria of 50 states. For each criteria, we sort the value of each state and give them rank in 50 states.If state A rank 0-5 higher than state B, we give 1 score, and 2 score for 6-10 higher, 3 score for 11-15 higher ... 8 score for 36-40 higher and 9 score for 41 higher or more.

1. Information Entropy

   Based on the information entropy calculation method mentioned above, we calculate information entropy of 50 states in 2009 and rank them in ascending order. Here is the result.

State	AZ	CA	NM	TX
Value	1.531	1.123	1.213	1.217
Rank	4	42	32	31

2. Greenhouse Gas Emission According to the information on the Internet, greenhouse gas emission of different energy sources are shown in the form below.

   Table 5: Lifecycle greenhouse gas emission estimates (g CO_2/ kWh) [7]

Coal	Petroleum	NaturalGas	Nuclear	FuelEthanol
820	778	490	66	519
Geothermal	Hydroelectric	SolarEnergy	WindEnergy	Wood&Waste
38	24	45	12	230

   Define GGE = EE% W, where GGE is total greenhouse gas emission, EE% is the percentage of one energy source in total energy consumption. W is the dimensionless coefficient of emission, whose value is same as the one in the form. There are two reasons for the definition of the GGE

   (a) The amount of emission is directly associated with population and GDP. By using percentage, the effect caused by GDP and population can be reduced.

   (b) Comparison of two alternatives are only needed in the model, and the values in the form not only show the real amount of emission, but also show the ratio between each two energy sources.

   By multiply the percentage by the value, we get the quantity relationship between emissions of two alternatives.

   After calculation, we get the result:

State	AZ	CA	NM	TX
Value	593.7	578.7	755.6	646.0
Rank	25	19	48	38

$\mathbf{B}_2$ can also be calculated.

3. Renewable Energy Percentage

State	AZ	CA	NM	TX
Value	5.97%	10.19%	4.26%	3.19%
Rank	27	16	39	47

$\mathbf{B}_3$ can also be calculated.

4. Financial Cost

Table 6: Financial cost of energy sources ($/MWh) [6]

Coal	Petroleum	NaturalGas	Nuclear	FuelEthanol
107	59	128	117	68
Geothermal	Hydroelectric	SolarEnergy	WindEnergy	Wood&Waste
100	140	197	55	69

Define FC = EE%$\mathbf{M}$, where FC is total financial cost. M is the dimensionless coefficient of financial cost, whose value is same as the one in the form.

State	AZ	CA	NM	TX
FC	104.4	90.4	107.0	89.0
Rank	45	15	47	11

$\mathbf{B}_4$ can also be calculated.

5. Energy Consumption Per Capita

State	AZ	CA	NM	TX
Population [8]	6392307	37252895	2059192	25146105
Consumption(bBTU)	1734973	6997016	836524	11196446
ECPC	0.271	0.188	0.406	0.445
Rank	16	4	38	44

$\mathbf{B}_5$ can also be calculated.

6. Summary

λ = [4.0623, 4.0340, 4.0583, 4.0052, 4.1145]

$$\mathbf{W} = \begin{bmatrix} 0.9606 & 0.5173 & 0.3961 & 0.1041 & 0.4365 \\ 0.0892 & 0.8248 & 0.8969 & 0.6744 & 0.8884 \\ 0.1747 & 0.1122 & 0.1684 & 0.0968 & 0.1189 \\ 0.1967 & 0.1987 & 0.1018 & 0.7246 & 0.0781 \end{bmatrix}$$

CI_2 = 0.0525 , RI_2 = 1.12 CR_2 = 0.0690 $<$ 0.1 The second consistency test passes. $\mathbf{\Omega} = \mathbf{Wp}$= [0.7680, 1.5497, 0.2650, 0.4734], i.e. the result of Arizona, California, New Mexico and Texas is 0.7680, 1.5497, 0.2650 and 0.4734. California has the best energy profile in 2009.

1.4 D. Energy profile prediction

To simplify the prediction, we firstly select common factors of different sources of energy. Two principles are followed during the selection process: these factors should be strongly correlated to sources of energy and they should be easier to predict both in short term and in the long term. These factors are defined as the main factors. To predict the time series of the main factors, we adopt the ARIMA model. Finally with the multiple linear regression model of the known period, we can predict the energy profile in 'known' future.

1.4.1 Introduction of ARIMA model

1. Mathematical theory

   ARIMA (Autoregressive integrated moving average) model is proposed by DJ Bartholomew in 1997, and it's an generalization of an ARMA (Autoregressive Moving Average Model). [2] It's used for fitting the time series of past and for predicting time series of future. Given a time series of data X_t where t is an integer index and the X_t are real numbers, an ARMA(p, q) model is given by:

   $$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{14}$$

   The equal form is:

   $$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t \tag{15}$$

   where L is the lag operator, the $alpha_i$ are the parameterd of the autoregressive part of the model, the θ_i are the parameters of the moving average part and the ε_t are error terms. ε_t are generally assumed to be independt, and $\sim N(0, \sigma)$.

   Assume the polynomial $\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right)$ has a unit root (a factor $(1 - L)$) of multiplicity d. Then it can be rewritten as:

   $$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) = \left(1 - \sum_{i=1}^{p'-d} \phi_i L^i\right) (1 - L)^d \tag{16}$$

   An ARIMA(p,d,q) process expresses this polynomial factorisation property with $p = p' - d$, and is given by:

   $$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t\} \tag{17}$$

2. Recognition of stationary difference order

   To apply the ARIMA model, we firstly need to recognize the stationary difference order. A stationary time series is one whose properties do not depend on the time at which the series is observed [10]. ARIMA models can be applied to eliminate the non-stationarity. The first order difference of y_t is:

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

   Similarly, the second order difference is:

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

   Higher orders are defined in the same way. Generally by checking the time series figure or checking autocorrelation or partial correlation coefficient, we can recognize the stationary difference. The order is d in ARIMA's order.

3. ACF and PCF

   ACF is Autocorrelation Function, and PCF is Partial Autocorrelation Function. By checking the ACF and PCF figures and finding the lags that over stand error, we can get two another orders of ARIMA.

1.4.2 Prediction of energy profile

1. Summary of the main factors of the four states

   All the main factors are in table **??**. With the main factors, five sources of energy of each state reach 0.6 or more, these main factors are strongly correlated with the energy profile. In addition, these main factors are generally easily to predict with ARIMA model, such as GDP, population, different prices that are non-stationary.

Table 7: Main factors of the four states

CA		AZ		NM		TX	
main factors	R^2	main factors	R^2	main factors	R^2	main factors	R^2
GDPRX	CL (0.59)	GDPRX	CL (0.96)	CLPRB	CL (0.96)	GDPRX	CL (0.97)
TPOPP	NG (0.90)	TPOPP	NG (0.71)	NGTCD	NG (0.83)	TPOPP	NG (0.90)
TETGR	PM (0.95)	TETGR	PM (0.99)	PATCD	PM (0.90)	PATCD	PM(0.99
CLTCD	NU (0.93)	TE	NU (0.85)	TE	NU (none)	ESTCD	NU (0.94)
TE	RE (0.94)	HYTCB	RE (0.98)		RE (0.87)	CLTCD	RE (0.85
PATCD							
HYTCB							

2. Prediction of main factors with ARIMA model

   Take California for example, with Stata we find the these factors include: GDPRX, TETGR, TPOPP, GLTCD, TETCB, PATCD, HYTCB. Regress all sources of energy on these factors, and the R^2s are 0.59 (coal), 0.90 (natural gas), 0.95 (petrolium), 0.93 (nuclear energy), 0.94 (renewable energy). One reason for the 'bad performance' of coal may be the dramaticly fall of coal consumption in 1983, however, this occasional event may not happen in the future.

   Time series of GDP, first order difference of GDP, second order difference of GDP are shown in figure 10. The third order difference of GDP (GDPdiff3) is staionary. Check the ACF and PCF of GDPdiff3, as shown in figure 11 and figure 11. Therefore, we can apply ARIMA(2,3,1) model for GDP prediction (figiure 13).

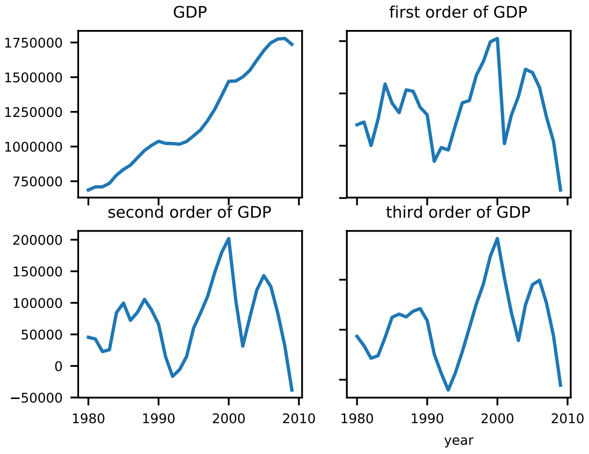


Figure 10: GDP, first, second and third order difference of GDP

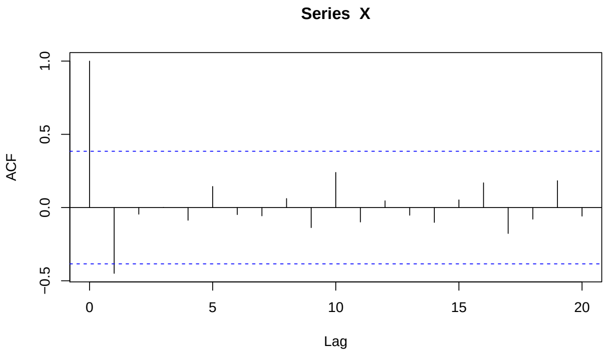


Figure 11: Autocorrelation correlation function of first order difference of GDP

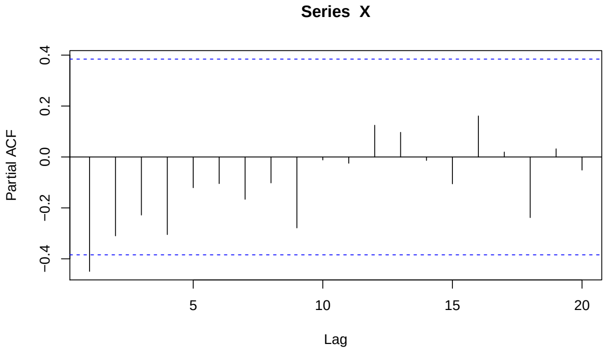


Figure 12: Partial autocorrelation function of first order difference of GDP

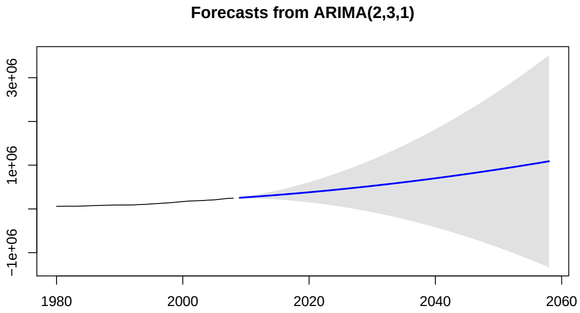


Figure 13: Forecast GDP with ARIMA (2, 3, 1)

3. Prediction of energy profile with predicted main factors

   The prediction of the energy profile is shown in figure 14. We can draw out the features in the figure during 2020∼2050:

   - California: proportion of petroleum will decrease slowly, while proportion of renewable energy will increase slowly.
   - Arizona: proportion of natural gas will decrease, and nearly goes to zero. While proportion of nuclear energy increase quickly and exceed coal at the end of 2050.

- New Mexico: proportion of coal will transfer to natural gas, it is the only state that the proportion of natural gas keeps increasing.
- Texas: proportion of natural gas will fade away, will that of coal, natural gas and renewable energy increase gradually.

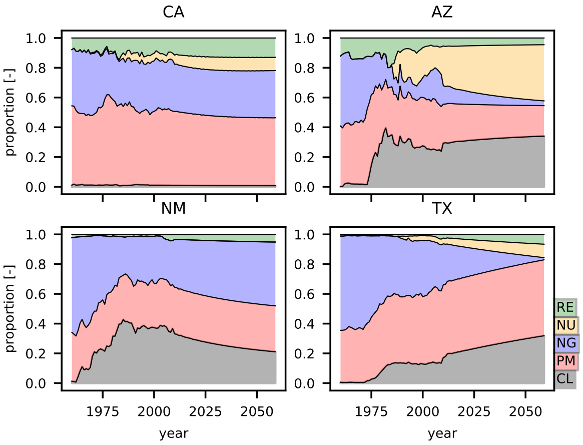


Figure 14: Prediction of energy profile (profile before 2009 is regressed with real main factors, after 2009 is predicted based on multiple linear regression model before 2009 with predicted main factors)

2 Part II

References

[1] U.S. Engergy Information Administration. State energy data system 2015 consumption technical notes, 2015.

[2] DJ Bartholomew. Time series analysis forecasting and control. *Journal of the Operational Research Society*, 22(2):199–201, 1971.

[3] Monica Borda. *Fundamentals in information theory and coding*. Springer Science & Business Media, 2011.

[4] Shaohong Cai and Shizheng Peng. *Dissipation Structure and the Change of Non-equilibrium State: Principle and Application*. Guizhou Science and Technology Press, 1998.

[5] Wikipedia contributors. Analytic hierarchy process — wikipedia, the free encyclopedia, 2018. [Online; accessed 12-February-2018].

[6] Wikipedia contributors. Cost of electricity by source — wikipedia, the free encyclopedia, 2018. [Online; accessed 12-February-2018].

[7] Wikipedia contributors. Life-cycle greenhouse-gas emissions of energy sources — wikipedia, the free encyclopedia, 2018. [Online; accessed 12-February-2018].

[8] Wikipedia contributors. List of u.s. states and territories by population — wikipedia, the free encyclopedia, 2018. [Online; accessed 12-February-2018].

[9] Josiah Willard Gibbs. *The scientific papers of J. Willard Gibbs*, volume 1. Longmans, Green and Company, 1906.

[10] Otext. Stationarity and differencing, 2017. [Online; accessed 12-February-2018].