Energy production

Team 93490

February 11, 2018

1 Part I

1.1 A. Energy profile of the four states

1.1.1 Classifying, integrating raw data and selecting main date

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 the most eager work.

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:

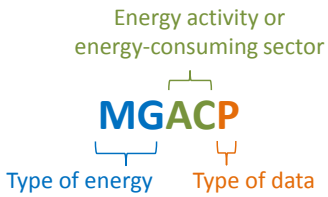

Figure 1: Components of a MSNCODE

As we can see, the first two letters of MSN code are used to represented 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)
- 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, MSNCODE=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.

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

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

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

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 3 and figure 2. 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, but 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) is 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 has 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.

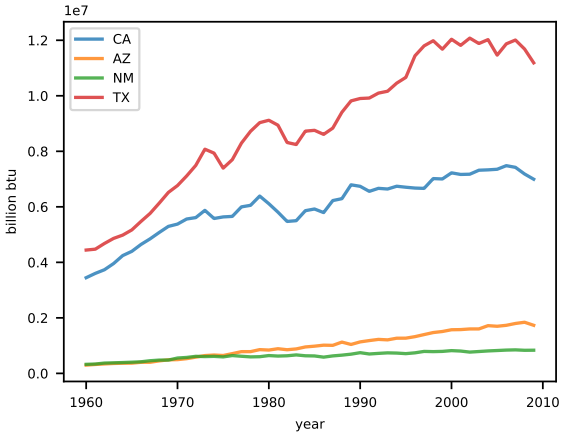

Figure 2: Total energy consumption of the four states from 1960 to 2009

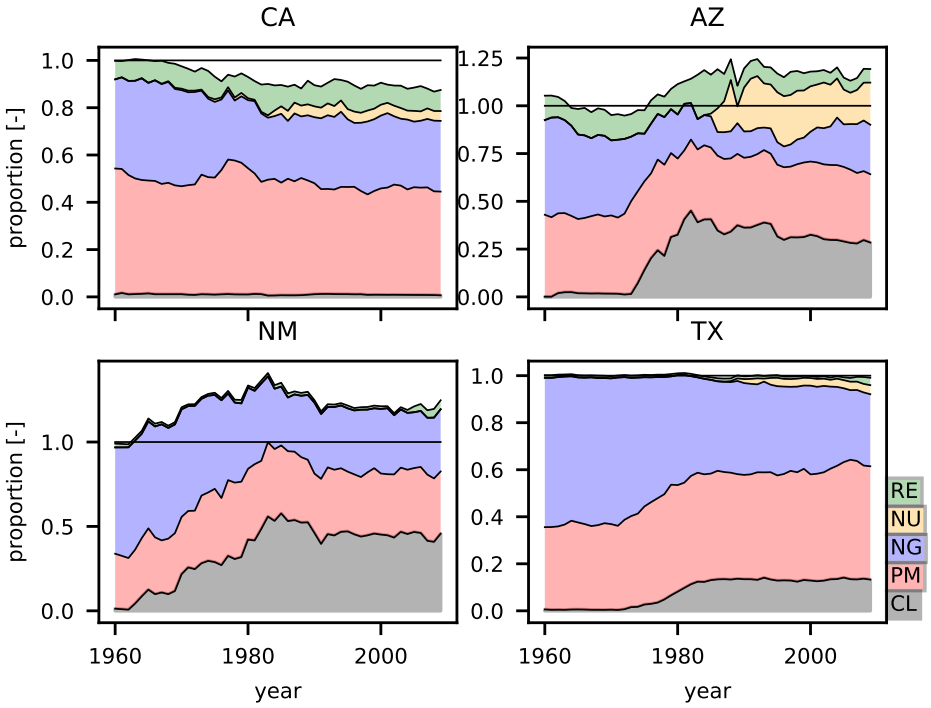

Figure 3: Proportion of five main energy types of the four states from 1960 to 2009

1.1.2 Summary of A

Energy production and usage are a major portion of any economy.

- 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 consumption structure evolution

1.2.1 Introduction of information entropy

Energy consumption structure 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 structure 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 consumption structure 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 [3]. 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 [1]. The equation above is called Shannon Equation, it can describe the degree of randomness and disorder of any system or material [2].

Follow what Shannon has done, we introduce the information entropy to the energy consumption structure 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, this satisfies the consumption of dissipative structure system. For example, consider an energy consumption structure 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 consumption structure 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, $S_{\min} = 0$; on the contrary, when all the energy types are equal, then $H_1 = H_2 = \cdots = H_m = H/m$, then $S_{\max} = \ln m$. 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 can be seen from 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 decrease by time, which can also suggest that the energy consumption structure is more balanced.

Combine figure 3 and 4, we can see in the four states, the Arizona state is most diversified, 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.

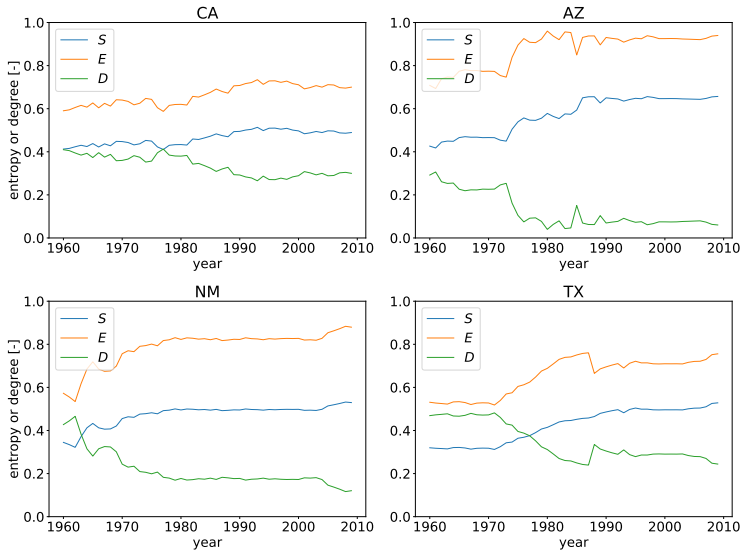

Figure 4: Information entropy S, equilibrium degree E and dominance degree D of four sates from 1960 to 2009

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 MSNCODE, we select the total energy (TE) of:

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

   (c) Economic factors

   As discussed above, the economic development means the structural transformation, which includes the energy consumption structure evolution. Economic factors include GDP (MSNCODE=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 (MSNCODE=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 adtion, 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 MSNCODE 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. MSNCODE here is TPOPP (resident population including armed forces).

(e) Climate factors

Climate influence 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=23, i.e., number of the indices). The linear regression assume 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 normalize. 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:

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \beta + \varepsilon_i \tag{8}$$

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 a $\boldsymbol{\beta}$'s estimation $\hat{\boldsymbol{\beta}}$ 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, suggest that the information entropy is strongly linear correlated to the factors. Take Arizona for example, the regression overview in Stata is shown in figure 6.

   - P value of the F test is 0.0000, 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), that's say, 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.

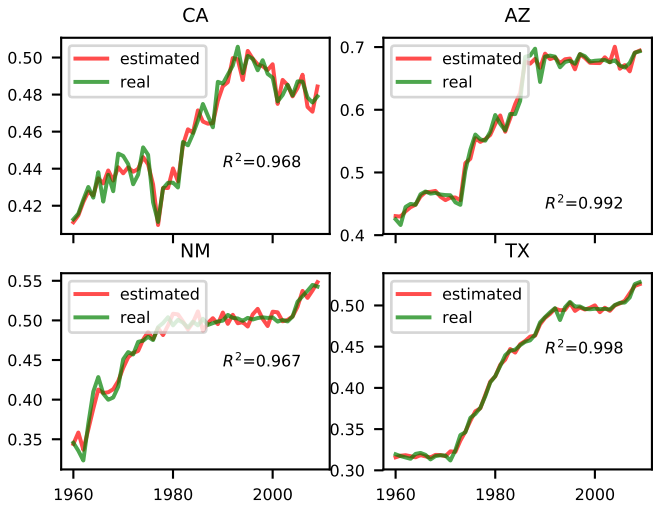

Figure 5: The real and estimated information entropy of the four states from 1960 to 2009

Source	SS	df	MS	Number of obs =	50
Model	.471416397	21	.0224484	F(21, 28) =	165.78
Residual	.003791579	28	.000135414	Prob > F =	0.0000
				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. Take Arizona for example, the regression overview in Stata is shown in figure 8.

   - P value of the F test is 0.0000, 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. A possible explanation is that several variable in the list is negative relevant to TEEIB and positive relevant to RE.

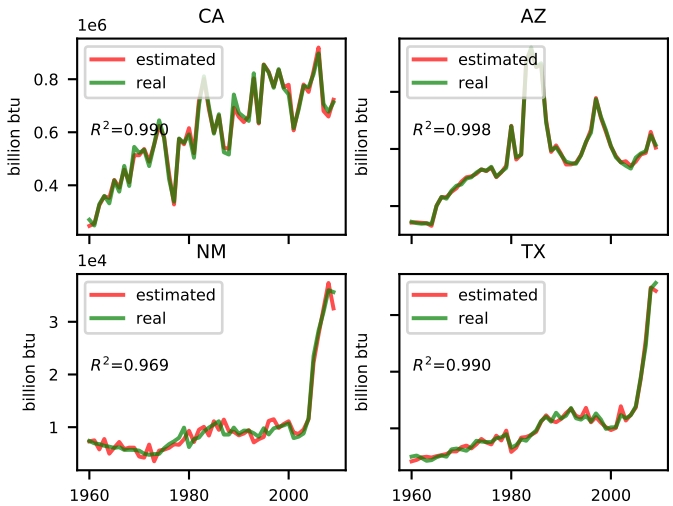

Figure 7: The real and estimated renewable energy of the four states from 1960 to 2009

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

We summerize all the variables and coefficients with $P < 0.05$ of the four states in table **??**

References

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

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

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