Energy Profile Analysis Based on PLSR Algorithm

Ji Tao

College of Computer Science and Technology, Nanjing University of Technology, Nanjing, Jiangsu, 211816, China

Keywords: Energy Profile, PLSR Algorithm, Regression

Abstract: Issue with energy source always arouses people's interest. To deal with the problem in America, this paper summarizes the energy profile of the four states from year 1960 to year 2009 respectively employing PLSR Algorithm. To be specific, in terms of one certain state, this paper compares their energy profile year by year; when concerning the relationship between states, this paper compares their data of a certain year. Having a general idea of energy development, this paper comes up with a practical evaluation system. It vividly demonstrates the situation of each factor. With the help of it, this paper also succeed in offering guidance for those four states.

1. Introduction

As world develops, the level of economics and finances are making a rocketing progress. Meantime, resources are drawing our attention. In other words, power like electricity must be consumed during producing procedure. Simultaneously, neither all resources are clean, nor all the resources are inexhaustible. In America, it obviously has been a case. What's good, twelve states in west America compacted with each other on the part of energy, which was created to aid in the development and management of new energy technologies for the sake of all member states, which could then benefit the rest of the country in terms of economic growth and energy sustainability. However, in south America along the border with Mexico, another four states: California (CA), Arizona(AZ), New Mexico (NM), and Texas (TX) recently are aware of the significance of energy crisis, they tend to construct a new energy platform.

2. Energy Profile

With the data sorted, we can focus on observing the condition of energy in the four states. To displaying the results more visually, we will illustrate the data employing images.

2.1 Arizona State

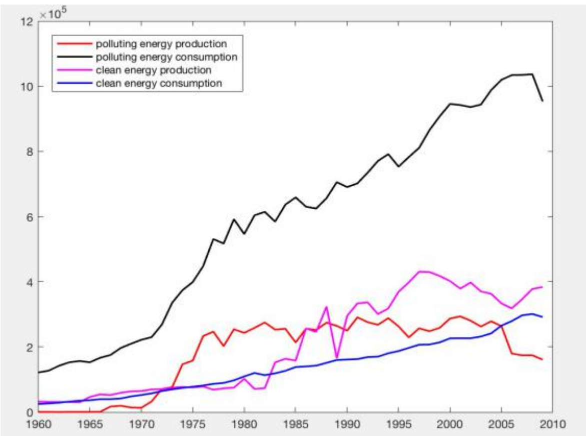

Figure 1. Arizona State

From the image, we can see the black line is far higher than others and it shows a rising trend with fluctuation. Thus, in Arizona, polluting energy requires a lot, which shows there are many industries there. The purple and the red lines are not high and are have a steady trend, showing that the raw materials were not rich and even rely on importing.

To study the data of our 14 indexes again. In AZ, the production of natural gas and petroleum maintained at a low level recently while coal production had been at a same level those years, which shows it was poor at fossil fuel. At the same time, geothermal, winds and solar energy was developing, distributing little for industry. Hydropower energy and nuclear energy was the main clean power generation mode, which produced electricity at thousand MK level and ten thousand MK level respectively.

2.2 California State

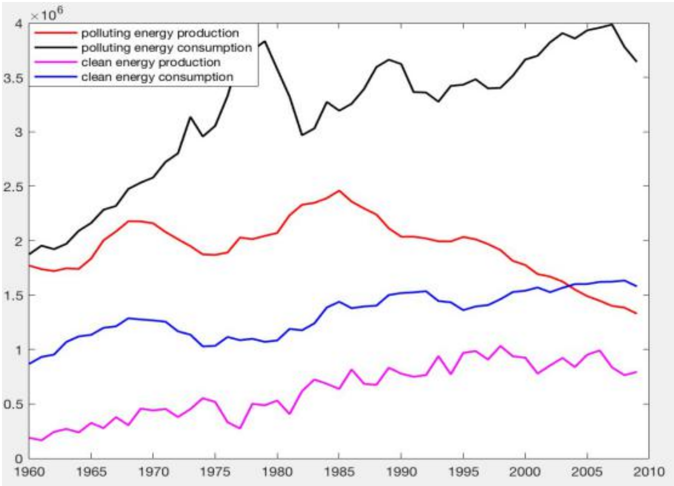

Figure 2. California State

Black line's high level and its rising trend shows it asked for big amount of polluting energy sources. Red line has a decline trend while purple line tends to rise, which shows more attention was drawn to clean energy. However, polluting energy was less and less in CA. With regard to fossil fuel, coal production as well as consumption was at a low level, petroleum showed a trend of first rise then drop. When it comes to natural gas, the consumption of it was nearly 5 times the production of it, stating that importing was a must.

CA had a sharp smell of new energy. Geothermal energy, wind energy, solar energy and nuclear energy all had developed quickly here since 1980. Hydropower energy here started from 1960 and electricity production based on it increases quickly during early years. Then its production tended to be stay at an average level, however, its data fluctuated greatly.

2.3 New Mexico State

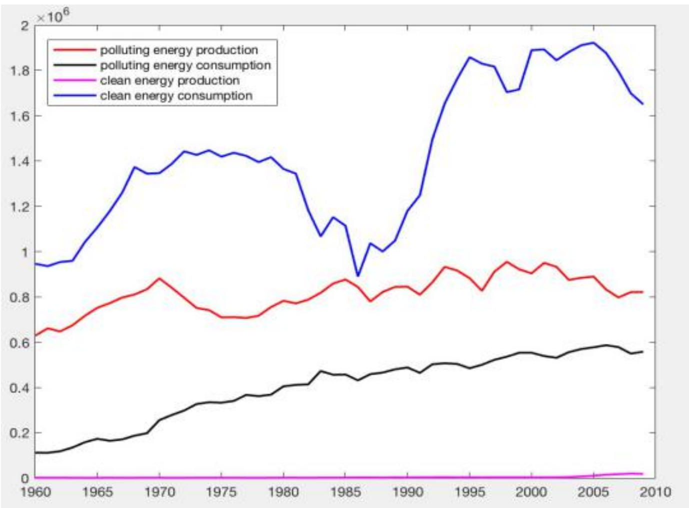

Figure 3. New Mexico State

The purple line remains nearly horizon and it is close to X axis, which shows NM had a low ability to produce clean energy. We assume that NM might make a breakthrough in clean energy production in few years. The blue line shows NM requires clean energy a lot, for instance, electricity. When it comes to polluting energy, the red and the black lines show the production and the consumption both tended to be stable in next few years.

It is actually an opposite situation when attention is focused on clean energy. Geothermal energy, solar energy and nuclear energy developed almost none during those years. Wind and hydropower energy using for generating electricity was limited at relative low level. All the above shows that it required New Mexico concentrate on clean energy sources as soon as possible.

2.4 Texas State

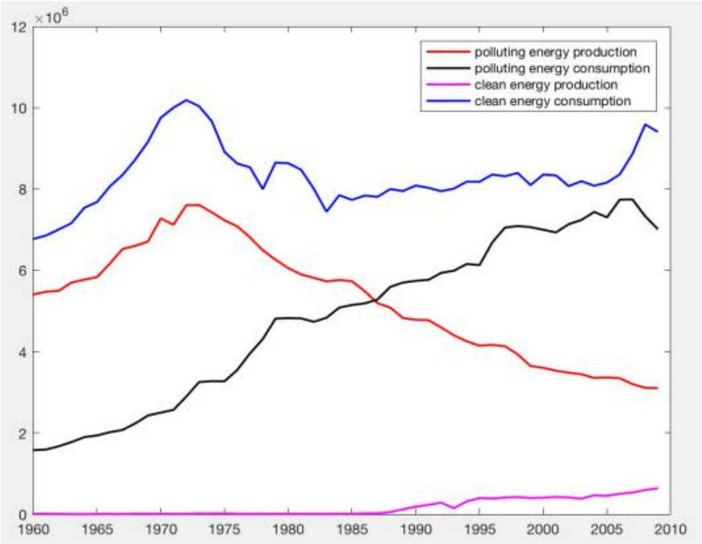

Figure 4. Texas State

The purple line demonstrates that TX was not sensitive to clean energy in the early years, but it gradually focused more on it. Red line declined obviously from about year 1973, showing the state produced polluting energy less. However, its requirement for the energy like coal was rising. Texas may need to import more polluting energy.

Texas State had an abundant storage of natural gas, although it consumed a lot, it had a large cardinal number, which supported TX to export. Differently, TX required coal for large amount, however its production was not able to supply for itself. It needed extra import then. The petroleum shows a tendency of first rise then drop and it may go on decline on.

3. Evolvement of Energy Profile

3.1 Data Collection

3.1.1 Variable Selecting

We select geography, climate, population and industry these four points as our adding research points, and divide the extra four points into two groups.

Table.1. 2 Groups

A	B
population	geography
industry	climate

Firstly, we have to admit these four factors affect the profile to some extent. Thus, we can't leave any of them out. Secondly, we can search the data of factors in group *A*, which means elements in group *A* could combine with the 4 achieved attributes. When we consider the six points together, we may have a chance to accurately describe the evolvement process. Thirdly, although factors in B can't participate in the quantitative analysis process, there is also information supporting us go on a qualitative analysis.

3.1.2 Data integrity

Before all the analysis, what we have to do is sorting all the data well. First and foremost, we look up in the excel and list all the MSNs we may need below.In terms of the first four attributes, data integrity has been ensured in Data Preprocessing part, it can be directly in use. And the MSN which describes the population information is TPPOP (*Resident population including Armed Forces*).The data in TPPOP line is valid without any loss. However, we have to deal with the TEICV (*Total energy expenditures*) for the data of first ten years is missing.

We decide to fill up the missing data with Polynomial curve fitting method. That means we fit a curve with several given data, then we conclude the missing data. For accuracy, we decide to employ cubic polynomial and we just conclude 3 missing numbers using 10 numbers following them at a time.

In the algorithm, we admit the corresponding relationship below:

$$\binom{x_1}{y_1}\binom{x_2}{y_2}\cdot\ldots\cdot\binom{x_k}{y_k}$$

It shows that a certain x corresponds to a certain y.

Our aim is to obtain a equation like $y = ax^3 + bx^2 + cx + d((a,b,c,d)^T \in R^4)$ But in order to calculate the equation, we tend to make the value of the following equation to be the minimum:

$$\sum_{i=1}^{k}[y(x_i) - y_i]^2$$

On account of the fact that we have 10 missing data, we have to repeat the algorithm for 4 times for each state.

3.2 Process Research

To depict the evolving process, we regard equation as an effective method. According to equations, we can safely conclude the evolvement of energy profile.

3.2.1 PLSR Algorithm

PLSR [1] (*Partial least-squares regression*) provides a method for modeling multi-linear regression, especially when the number of two sets of variables is very high, there are multiple correlations among them, and the number of observations is small. Adopting PLSR model has the advantages that other classical regression analysis methods don't have.

PLSR analysis combine the principal component analysis, canonical correlation analysis and linear regression analysis method of characteristics in the process of modeling, so in the results of the analysis, in addition to providing a reasonable regression model, it can also finish at the same time some related work similar to the principal component analysis and canonical correlation analysis. What's more, it can provide more rich, and some further information.

Considering the modeling problem of p dependent variables ($y_1,y_2,...,y_p$) and q independent variables ($x_1,x_2,...,x_q$), we will firstly extract the first component t_1 (the linear combination of $x_1,x_2,...,x_q$, and extract as much variation information as possible from the original set of variables). At the same time, extract the first component u_1 in the dependent variable concentration, and require the correlation degree between t_1 and u_1 to be maximized. Then, establish the regression between the dependent variables $y_1,y_2,...,y_p$ and t_1. If the regression equation has satisfied the accuracy, the algorithm can be aborted. Otherwise, extract the second component until satisfactory accuracy is achieved.

If we finally extract r components from the set of independent variables ($t_1,t_2,...,t_r$), PLSR will be employed to establish a regression equation between $y_1,y_2,...,y_p$ and $t_1,t_2,...,t_r$, and then transit the regression equation to a new one between $y_1,y_2,...,y_p$ and the original independent variables .The new equation is called the *least squares regression equation*.

To be more convenient, we might as well assume the p dependent variables ($y_1,y_2,...,y_p$) and the q independent variables ($x_1,x_2,...,x_q$) are all standardized variables. The n time standardized observation data matrix of the dependent variable group and the independent variable group is denoted as:

$$F_0 = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} \qquad E_0 = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

3.2.2 PLSR Steps

In order to calculate the results with high accuracy, we take these steps in sequence:

1. Find the eigenvectors w_1 corresponding to the maximum eigenvalues of the matrix $E_0{}^T F_0 F_0{}^T E_0$, and find the component $t_1 = w_1{}^T X$ as well. Then calculate the component score vector $\hat{t}_1 = E_0 w_1$ and residual matrix $E_1 = E_0 - \hat{t_1}\alpha_1{}^T$ (among it, α_1 equals to $E_0{}^T \hat{t}_1/\text{k}\hat{t}_1\text{k}^2$).

Repeat the procedure in Step 1 for $r-1$ times to find the eigenvalues $w_i(i = 2,3,...,r)$, the components $t_i(i = 2,3,...,r)$, the component score vector $\hat{t}_i(i = 2,3,...,r)$, the residual matrixes $E_i(i = 2,3,...,r-1)$ (among them, $\alpha_i(i = 2,3,...,r-1)$ equals to:

$$E_{i-1}{}^T \hat{t_i}/\|\hat{t_i}\|^2 (i = 2, 3, ..., r-1)\ ).$$

According to cross-validation, we determine to extract r components ($t_1,t_2,...,t_r$) in all, then a satisfactory prediction model can be obtained. Afterwards, find the ordinary least squares regression equation of F_0 in $\hat{t}_1,...,\hat{t}_r$:

$$F_0 = \hat{t_1}\beta_1{}^T + ... + \hat{t_r}\beta_r{}^T + F_r$$

Put $t_k = w^*_{k1}x_1 + ... + w^*_{km}x_m (k = 1, 2, ..., r)$ into $Y = t_1\beta_1 + ... + t_r\beta_r$, and we can achieve p partial least-squares regression equation for dependent variables:

$$yj = aj1x1 + ... + ajmxm$$

3.2.3 PLSR Outcome

We regard year (X_y), population (P_{tp}), industry (V_{te}) as the independent variables. Take clean energy production (C_p), clean energy consumption (C_c), polluting energy production (P_p), polluting energy consumption (P_c) as the dependent variables.We can finally achieve the equations in the following form:

$$y_i = k_1X_y + k_2P_{tp} + k_3V_{te} + c(i = 1,2,3,4)$$

Table.2. PLSR results

Arizona	$P_p(y_1)$	$P_c(y_2)$	$C_p(y_3)$	$C_c(y_4)$
C	-2.46E+07	-3.28E+07	-2.40E+07	-4.82E+06
k_1	1.25E+04	1.67E+04	1.22E+04	2456.8841
k_2	-5.4797	42.5306	9.9747	16.4268
k_3	-182.5826	-31.6958	-112.6655	38.6868
California	$P_p(y_1)$	$P_c(y_2)$	$C_p(y_3)$	$C_c(y_4)$
C	5.71E+06	-2.28E+07	-1.98E+07	-1.13E+07
k_1	-1821.6693	1.26E+04	9939.9485	6145.7871
k_2	-2.4632	28.7368	32.2434	17.3738
k_3	-14.5364	34.1438	-27.8541	-2.6193
New Mexico	$P_p(y_1)$	$P_c(y_2)$	$C_p(y_3)$	$C_c(y_4)$
C	-7.45E+06	-1.00E+07	1.32E+05	-2.57E+07
k_1	4098.9212	5106.6597	-67.1850	1.34E+04
k_2	144.5527	196.2155	-1.4561	484.9499
k_3	-201.4202	0.8467	17.5622	-480.9540
Texas	$P_p(y_1)$	$P_c(y_2)$	$C_p(y_3)$	$C_c(y_4)$
C	7.84E+07	-1.47E+08	-1.04E+07	1.29E+07
k_1	-3.58E+04	7.48E+04	5166.2187	-2293.7339
k_2	-121.1662	261.5108	17.2198	-10.4197
k_3	-7.7741	-21.4674	2.2297	11.1846

With the equations above, we can safely conclude the effects the three factors have on the energy profile. When positive, larger the coefficient of an independent is, larger the effect the factor has on the dependent in positive correlation. For example, in CA, the k_1 in the column P_c. The great number means with the population increased those years in CA, polluting energy consumption had a great enough increase with it. However, when in situation of negative numbers, the thing is just the opposite.

4. Conclusion

Energy is the essential part in industry while an energy profile can play the role of a guider. In our mind, a qualified energy profile should consist of the development level of energy source, the trend of its development and the momentum of energy sources' development. This paper summarizes the energy profile of the four states from year 1960 to year 2009 respectively employing PLSR Algorithm. Through it, we can do comparison among these states by adopting Correlation Coefficient Algorithm.

References

[1] Liu Wanying, Huang Guisheng, Wang Bin, etc.New talent [J]. Experimental science and technology, 2013, 11 (1): 140-142.

[2] Katherine Meacham-Hensold, Christopher M. Montes, Jin Wu, Kaiyu Guan, Peng Fu, Elizabeth A. Ainsworth, Taylor Pederson, Caitlin E. Moore, Kenny Lee Brown, Christine Raines, Carl J. Bernacchi. High-throughput field phenotyping using hyperspectral reflectance and partial least squares regression (PLSR) reveals genetic modifications to photosynthetic capacity [J]. Remote Sensing of Environment, 2011.