A regional energy consumption analysis model using a novel outlier removal algorithm and k-means clustering method

Yuchen Wang[†], Shuxiang Xu[††] and Wei Liu[†]

[†]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
[††] School of Engineering and ICT, University of Tasmania, Launceston, Australia

Summary
The regional energy-saving work is an important part of China’s energy conservation projects. In this paper, we developed a regional energy consumption analysis model using the energy consumption data from a typical Chinese industrial city, Shaoxing. We incorporated a well-known data mining tool, the k-means clustering method, into our model to automatically classify our energy consumption data into low, medium and high clusters representing different energy consumption levels. This classification provides a basis for further analysis to help governments and enterprises to use energy more efficiently. However, there are a few of potential outliers in our data set, and the result of k-means might be strongly influenced by these outliers. To reduce the impact of these extremely large data points, we proposed a distance-based outliers removal algorithm as well as a corresponding parameters choosing algorithm which provides tuning parameters to make a balance between keeping and removing far away points. The experimental results show that our algorithms can effectively reduce the influence of outliers and make the k-means results more meaningful. The relationship between levels of consumption and industrial output was also examined as one possible way of further analysis based on our model.

Key words:
machine learning; data mining; k-means; outlier removal; regional energy consumption analysis model

1. Introduction

With the development of economy, China's energy consumption grew rapidly during these years. In 2014, the total consumption of China's primary energy (fossil fuels, nuclear fuel, biomass energy, wind energy, solar energy, etc.) reached 2.972 billion tons of oil equivalent, overtaking the figure in United States for five consecutive years and ranking at the first place around the world [1]. The environmental pollution is also becoming more serious, making China’s economic growth unsustainable [2]. To use energy efficiently has become one of China's national strategies because saving energy effectively can not only reduce the production cost of enterprises and increase their profits, but also dramatically reduce the pollution levels in China [3].

In order to improve the process of energy-saving work, the Chinese government has launched a series of energy conservation projects, such as the Top-1000 Enterprises Energy-Saving Program [4] and Low-Carbon Policy and Action [5]. These projects require local governments to collect and submit the data of high energy-consuming enterprises in different industries to ensure the high consumption companies are being specifically monitored and analyzed. This means the local governments are the ones who have the first hand data, and are able to communicate with companies in their region directly and efficiently. Therefore, developing a regional energy consumption analysis model for local governments is a practical way to help them guide the energy-saving work.

To analyze the energy consumption data and give effective guidance to the governments’ energy policy goals, a bunch of energy analysis models have been developed. In 1979, Abilock et al. [6] introduced MARKAL, which is a multi-period, linear-programming model for energy systems analysis. Jiang et al. [7] utilized MARKAL to estimate the future natural gas and coal consumption in Beijing, Guangdong and Shanghai provinces in China. Can et al. [8] applied LEAP model developed by Stockholm environment institute [9] to analyze the energy consumption in industries, transportation and construction on a global scale. However, these models are mainly focused on the energy supply and demand forecasting from a global or national perspective and cannot provide simple and direct guidance for the local governments to carry out energy-saving work.

In terms of the regional energy-saving work, the local governments want to use an effective way to explore the energy use condition in regional industries, especially in energy-intensive enterprises, so that they are able to analyze the reasons and guide the energy-saving work. For example, when manufacturing one same product, different enterprises will have different energy use rate because the techniques they use, such as equipment, processes, and integrated information systems, are more or less different from each other. In other words, in one type of industry, enterprises with high energy use efficiency may give valuable suggestions to those with low energy use rate. In addition, the local governments also want to figure out the

Manuscript received March 5, 2016
Manuscript revised March 20, 2016

energy use rate patterns in their regions. These patterns are very good references for newly established companies and companies in other regions. For instance, enterprises with a particular amount of industrial output in one region may be more energy-efficient than their counterparts in other regions. This will lead to more interaction from engineers, business executives to government officials between enterprises in order to do more effort to use energy more efficiently.

To meet the requirement stated above, we developed a regional energy consumption analysis model by utilizing k-means clustering algorithm, which classifies the energy efficiency data into low, medium and high consumption clusters representing different energy consumption levels. These classifications provide basic preparation for further analysis and we examined the relationship between levels of consumption and industrial output as one possible way of further analysis. However, before utilizing k-means, there are a few of potential outliers in data, and the result of k-means might be strongly influenced by these outliers. To reduce the impact of outliers in the consumption data, we proposed a distance-based outliers removal algorithm and a corresponding parameters choosing algorithm which can make a balance between keeping and removing far away points. The experimental results show that our algorithms can effectively reduce the influence of outliers and make the k-means results more meaningful.

This paper is organized as follows. Section 2 introduces the regional energy consumption data, the k-means method and the building process of our analysis model. Section 3 presents the proposed distance-based outliers removal algorithm and the parameters choosing algorithm. Section 4 shows the experimental results and section 5 makes a conclusion.

2. Data and methodology

2.1 Data source

In many countries, certain industrial types are often found in a certain region and this could be called industrial clustering [10]. This is partly because the set up of enterprises often needs strong social networks and involves tremendous interaction and trust [11]. China's industrial enterprises also have this kind of regional concentration. As an important industrial city, Shaoxing is a typical example for this. The information of energy-intensive enterprises in Shaoxing is extracted from the Regional Energy Monitoring and Warning System deployed there. Textile industry in Shaoxing is very famous in China [12]. From our data we can also see that, among the 252 energy-intensive enterprises, 155 of them are textile enterprises, accounting for 62% of total energy-intensive enterprises. The proportion of energy-intensive industries and the detailed number are illustrated below. The classification is according to the Chinese National Industries Classification GB/T4754-2011 [13].

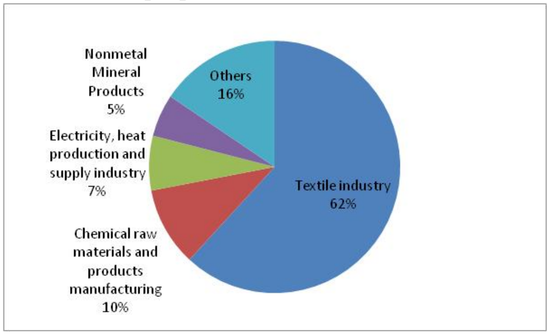

Figure 1. Proportion of energy-intensive industries in Shaoxing

Table 1.Industry types and corresponding number of companies

Industry types	Number
Textile industry	155
Chemical raw materials and products manufacturing	26
Electricity, heat production and supply	18
Nonmetal Mineral Products	14
Special Equipment	7
Paper products	6
Metal Products	5
Garment and dress industry	5
Leather, fur, feather and their products and shoe manufacture	4
Medical and Pharmaceutical Products	4
Rubber and plastic products	3
Electric Equipment and Machinery	2
Wine, beverages and refined tea manufacturing	1
Automobile	1
Food manufacturing	1

In this data, we are mostly interested in the indicators which represent the performance (industrial output, industrial added value, etc.) and the energy use condition, like coal, electricity consumption, of a company. The abbreviation and explanation for the indicators we use are listed below:

Table 2. Nomenclature

Abbreviation	Explanation	Unit
CNY	China Yuan	
TSC	Tons of standard coal	
TTC	Ten thousand CNY	
Total consumption indicators		
IO	Industrial output	TTC
IAV	Industrial added value	TTC
EEVC	Energy equivalent value consumption	TSC
TCC	Total coal consumption	TSC

EC	Electricity consumption	TSC
TC	Total cost	TTC
TEC	Total energy cost	TTC
Relative consumption indicators		
EEVC-IO	Energy equivalent value consumption per 10,000 CNY output value	TSC / Per TTC
EEVC-IAV	Energy equivalent value consumption per 10,000 CNY industrial added value	TSC / Per TTC
EC-IO	Electric consumption per 10,000 CNY output value	TSC / Per TTC
EC-IAV	Electric consumption per 10,000 CNY industrial added value	TSC / Per TTC
TCC-IAV	Coal consumption per 10,000 CNY industrial added value	TSC / Per TTC
P-TEC-TC	Proportion of total energy cost accounting for total cost	Proportion

2.2 K-means clustering algorithm

K-means [14] is a famous clustering method which can automatically classify data into different categories. The meaning of these categories and the number of categories, k, could be defined manually based on specific applications. After performing k-means, every classified sample belongs to the category which is the nearest to the sample based on the selected distance function. These clusters then provide further information for more detailed analysis. Due to its simplicity and ease of computing even on a large dataset, the k-means clustering algorithm has been applied to various fields such as computer vision [15], market segmentation [16] and web mining [17]. In the energy analysis field, the k-means method has also been applied. Xu and Liu [18] used k-means to determine the energy losses level. Lou and Zou [19] extracted the basic energy consumption situation in 1 ton of aluminum production. Hernández et al. [20] applied k-means to analyze energy consumption patterns in industrial parks. Figueiredo et al. [21] utilized k-means and other data mining techniques to build an electricity consumer characterization framework.

2.3 Model building process

In this paper, we developed a regional energy consumption analysis model using k-means clustering method and our proposed outlier removal algorithm. The k-means clustering algorithm is utilized to automatically classify our energy efficiency data into different levels. A practical classification strategy is to divide the data into three categories which represent low, medium and high energy consumption levels. The k-means clustering result is essentially a basis for further analysis. Once getting this classified data, the government officials could do more detailed analysis on the consumption data and give more guidance to help companies save energy. We incorporated the relationship between levels of consumption and industrial output into our analysis model as one possible way of further statistical analysis. However, before utilizing k-means, there are a few of potential outliers in some of the indicator data, and the result of k-means might be strongly influenced by these outliers. Based to the distribution of consumption data, we proposed a distance-based outliers removal algorithm and a corresponding parameters choosing algorithm to reduce the impact of outliers. These algorithms will be explained in next section. The whole model building process is illustrated in below figure:

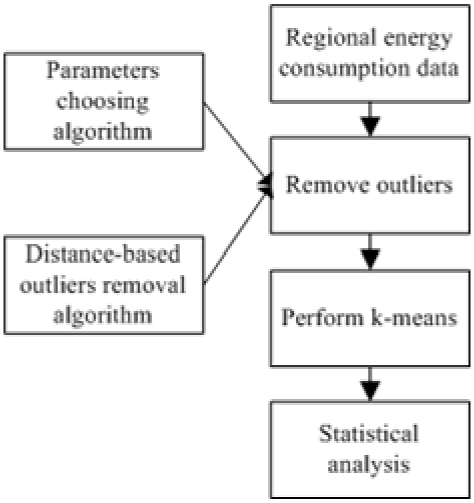
Regional energy consumption data
Parameters choosing algorithm
Remove outliers
Distance-based outliers removal algorithm
Perform k-means
Statistical analysis

Figure 2.Model building process

3. Proposed outliers removal algorithm

3.1 Data distribution and problems

Figure 3 and figure 4 are two cases of the data distributions of the indicators we use. In fact, all these indicators follow very similar patterns: they are all very right-skewed, with very few data points in each indicator standing far away from others. These data points that extremely deviate from other points can be named as outliers. K-means is an algorithm which will be strongly influenced by outliers [22]. Using k-means directly to cluster data may fail to get a reasonable result because these outliers may significantly influence the cluster centers. Furthermore, companies with these extremely high consumption indicators may have very different ways of energy use so that these companies need to be analyzed independently rather than being considered into a group. Therefore, it is necessary to develop a method which can automatically detect these outliers before performing the k-means algorithm.

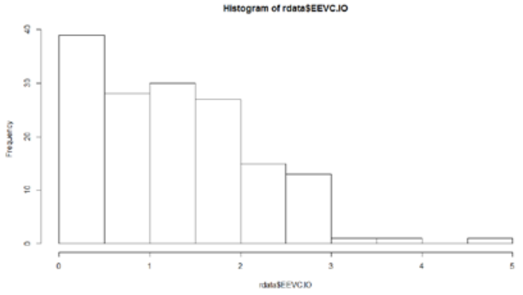

Figure 3. Distribution of EEVC-IO

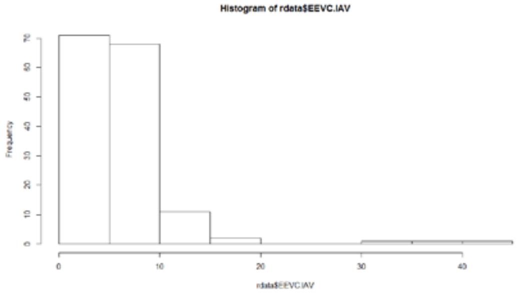

Figure 4. Distribution of EEVC-IAV

3.2 Criteria for defining outliers

Although the data points standing far away from other points should be considered as outliers, in practice of regional energy consumption analysis, if there are a certain number of points close with each other and are much greater than other data, they should not be regarded as outliers because there must be some potential reasons that make these points much larger than normal value. Simply removing all these bigger points may destroy the availability and integrity of data. Hence, developing a proper way to determine whether a set of extremely large points should be deleted is necessary in our algorithm. In addition, the exact distance which makes a point as outlier also needs to be well determined.

3.3 Algorithm description

In this paper, we proposed a novel distance-based outliers removal algorithm to effectively detect outliers from our data set. The main process of our algorithm is described below:

Given an energy consumption data set D = {d1,d2,…,dn}, where n is the number of data records and each record has m indicators. By sorting data for each indicator, we are able to calculate m number of average distance, avgDistance, between every pair of adjacent data points. These average values can be used as benchmarks for finding out outliers.

Then, we introduce two parameters, minPoint and multiplicator, and one concept connection link to help us find out outliers. These parameters and concept are explained as follows:

Initially, every point is the only point in its connection link. If two adjacent points in a sorted list are within distance multiplicator * avgDistance, they are considered as connected with each other and are in one connection link. If one point is connected with a point which has been connected with another point, all these points are then in the same connection link. If at least minPoint number of points are connected by a connection link, they are regarded as normal data and should not be removed. By contrast, if the number of points in a connection link is less than minPoint, they should be regarded as outliers and the consumption data records that these outliers belong to should be removed from the original data set. For example, considering a list with only one extremely large point, if minPoint is 3 and the largest point is more than multiplicator * avgDistance away from its neighbor, the number of points in its connection link is only 1, thus it should be regarded as an outlier and the corresponding data record should be removed.

As the distribution of our data are strongly right-skewed, we only consider the outliers in the large end of the coordinate axis instead of the small end. For each indicator, our algorithm starts from the largest data point, comparing the distance to its smaller neighbor with multiplicator * avgDistance. This comparison process repeats from large points to small and then a connection link can be found. If the number of points in this link is less than minPoint, these points are labeled as outliers and the algorithm continues its process from the next largest point of the list without the outliers. When at least minPoint number of points in a connection link are found, the detecting process for current indicator stops and starts to detect outliers in the next indicator.

The parameter minPoint is introduced because a few of outliers may influence the distribution and the k-means result significantly. Hence, the corresponding enterprises that these outliers represent should be investigated separately. However, it is also necessary to keep the far away points if there are too many of them, and these points should be included in the k-means clusters. This parameter is essentially a trade-off between keeping and removing far away points and could be determined by domain experts.

The pseudo-code of our algorithm is given below:

Table 3. Distance-based outliers removal algorithm

```
Algorithm 1: distance-based outliers removal algorithm
Input: D (original data)
       minPoint (number of points connected that makes them normal)
       multiplicator (scale factor that determines how easy one point can connect with another)
Output: new data ND without outliers

  potentialOutliers = Empty;
  deletedOutliers = Empty;

  for j = 1 to D.column do
     AscendingSort(D,j);
     for i = 2 to D.row do
        distance[i][j] = D[i][j] – D[i – 1][j];
     end
     avgDistance[j] = calculateAvg(distance[j]);
  end

  for j = 1 to D.column do
     AscendingSort(D,j);
     for i = D.row to 2 do
        if distance[i][j] >= multiplicator * avgDistance [j] do
           count = 0;
           for each po in potentialOutliers do
              deletedOutliers.add(po.key);
           end
           deletedOutliers.add(D[i].key);
           potentialOutliers.clear();
        else
           count++;
           potentialOutliers.add(D[i].key);
           if count == minPoint – 1 do
              potentialOutliers.clear();
              count = 0;
              break;
                 end
        end /* end checking distance */
     end /* end loop i */
  end /* end loop j */

  ND = D;
  for each o in deletedOutliers do
     deleteRow(ND,o.key);
  end

  return ND;
```

First, this algorithm sorts each indicator in ascending order and calculates all the distance between each pair of adjacent data points for every indicator. The average value avgDistance for each indicator can also be computed during this process. Ascending order for each column of data is just for easy understanding because it follows the same order as distribution figures. Next, from the largest pair of data points of each indicator, the algorithm compares the distance with multiplicator * avgDistance, if the distance is less than that threshold, the larger point will be put into potentialOutliers, which stores potential outliers. Then the algorithm compares the number of potential outliers with minPoint – 1. If this number reach minPoint – 1, the detecting process for current indicator breaks and starts to check the next indicator. In contrast, if the distance is greater than the threshold, this means some actual outliers have been found. The algorithm transfers all data in potentialOutliers to deletedOutliers which contains the actual outliers, and also put current checking data points into deletedOutliers. The key, which is an unique identification for a particular data record, is used for identifying and removing outliers. After getting all outliers, the algorithm deletes these outliers according to their identification key and returns the final list without outliers.

Taking minPoint equaling 4 and 3 as examples, We illustrate the detecting process in below figures:

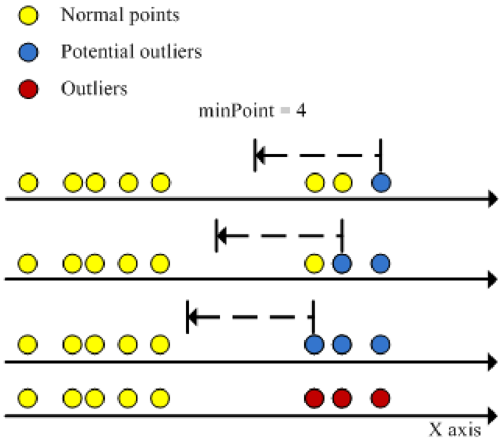

Figure 5. Detecting process when minPoint = 4

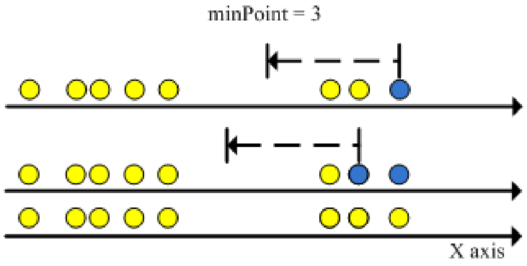

Figure 6.Detecting process when minPoint = 3

We assume that there are 3 points standing extremely far away from other points and minPoint is 4. At first, the algorithm starts from the largest point and finds that its nearest neighbor is within the distance threshold. As a result of this, the largest point is temporarily labeled as a potential outlier and then the algorithm does the same check on the next point. When the third point finds that it is not within the distance threshold to its neighbor and the number of points in its connection link is less than minPoint, these three points are detected as outliers. In another case, when the minPoint is 3 and the second largest point finds that its neighbor is within the distance threshold. The number of points in this connection link reaches minPoint so that all these far points are regarded as normal.

3.4 Choosing parameters

As minPoint defines the minimum number of points close to each other that should be regarded as normal points, the bigger the minPoint is, the easier a set of far points would be considered as outliers. The multiplicator controls how easy one point can reach its nearest neighbor, the bigger the easier. In practice, these two parameters should be set by experts based on well understanding on the distribution of data as well as the results under different combinations of parameter values. In this part, we give the general algorithm for choosing practical parameters.

Table 4. Parameters choosing algorithm

```
Algorithm 2: parameters choosing algorithm
Input: D (original data)
        minMinPoint (start value for minPoint)
        maxMinPoint (end value for minPoint)
        minMult (start value for multiplicator)
        maxMult (end value for multiplicator)
        meetCriteria (decision-function that defines acceptable
parameters)

Output: resultMapper (selected multiplicator for minPoint)

  for mp = minMinPoint to maxMinPoint do
    for mu = maxMult to minMult do
      ND = Distance-basedOutliersRemoval(D, mp, mu);
      if meetCriteria(ND) do
        resultMapper[mp] = mu;
        break;
      end
    end
  end

  return resultMapper;
```

Ultimately, the algorithm returns the multiplicator for every minPoint. At the beginning of this algorithm, the lower and upper bounds of parameters can be initialized according to the distribution of data. The parameter multiplicator ranges from maxMult to minMult, which means the number of expected outliers goes from small to large. After removing outliers by using our distance-based outliers removal algorithm with particular minPoint and multiplicator, if the newly generated data set meets the user-defined criteria, the algorithm puts that parameter pairs into its result. The criteria could be defined by domain experts as long as these criteria have meaningful explanation or the experimental results can give useful guidance. For example, we could perform k-means on the data without outliers and, if the proportion of clusters meets a defined standard, the algorithm puts the multiplicator, minPoint pair into its result. Alternatively, the algorithm finds a result pair when a specific number of points, or a particular percentage of points have been removed.

4. Experimental results

In this section, we present the whole process of building our energy consumption analysis model. We use six relative energy efficiency indicators for every enterprise. These indicators are described in below table:

Table 5. Relative energy efficiency indicators

Indicator abbreviation	Description
EEVC-IO	Energy equivalent value consumption per 10,000 CNY output value
EEVC-IAV	Energy equivalent value consumption per 10,000 CNY industrial added value
EC-IO	Electric consumption per 10,000 CNY output value
EC-IAV	Electric consumption per 10,000 CNY industrial added value
TCC-IAV	Coal consumption per 10,000 CNY industrial added value
P-TEC-TC	Proportion of total energy cost accounting for total cost

Based on these indicators, we apply k-means clustering method to automatically classify our energy efficiency data into three categories which represent low, medium and high energy consumption levels respectively. The detailed setup for k-means is showed in following table:

Table 6. Setup for k-means

Number of clusters	3
Distance function	Euclidean distance
Normalization in distance calculation	True
Maximum iterations	500
Seed	10

Before applying k-means, we need to remove potential outliers which may influence the result of k-means, so we first use the parameters choosing algorithm to find out ideal parameters, minPoint and multiplicator, for our distance-based outliers removal algorithm. The data and parameters for the choosing algorithm are described in following table:

Table 7. Data and parameters for the parameters choosing algorithm

Industrial type	Textile industry in Shaoxing
Data collected time	2009
Enterprise number	155
Decision-function	Performing k-means where k equals 3 representing low, medium and high consumption categories, and get more than 30 enterprises for each category for the first time (i.e. with the least outliers removed).
minMinPoint	2
maxMinPoint	6
minMult	10
maxMult	200

In our experiment, we have 155 textile industry companies' energy consumption information in Shaoxing city, Zhejiang province, China. The decision-function for selecting appropriate parameters can be defined by experts according to their understanding on the data. In this experiment, we also utilize k-means method as a decision-function. The k equals 3 representing low, medium and high consumption companies. If all categories have more than 30 enterprises, a pair of multiplicator, minPoint is found. Usually, 30 is a number that might show meaningful results when performing statistical tools. The minPoint and multiplicator range from 2 to 6 and 200 to 10 respectively. The minPoint is set experientially. The multiplicator is set by looking into the distributions of all indicators, which makes sure the number of outliers detected can range from 0 to a particular amount.

To have a clear view about the selecting process, we illustrate how the numbers of points in low, medium and high categories change when different number of outliers are deleted, and this is done by tuning parameter multiplicator from large to small. Here we take minPoint equaling 3 as an example.

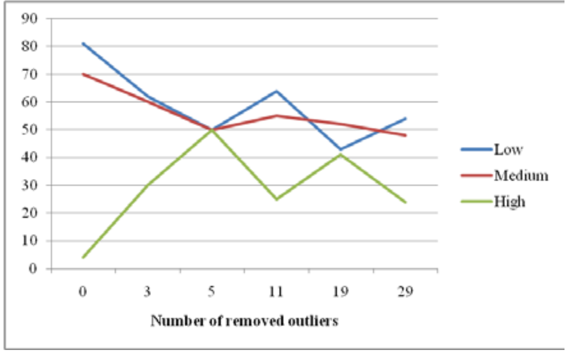

Figure 7. Number of points in k-means clusters with various number of removed outliers

We can see from above figure that, when no outliers are detected, the number of points in high consumption category is very low, containing only 4 points. This is not an ideal result for further analysis because the samples in high energy consumption level is too small to provide statistically reasonable results. Then, as more outliers are deleted, the size of high consumption category grows and the other two categories drop. After removing 5 outliers, the numbers of points in three categories all start to fluctuate. When 3 outliers are removed, the numbers of points in low, medium and high categories reach 62, 60 and 30 respectively for the first time. Based on our decision-function, these numbers meet our criteria and thus the multiplicator for minPoint 3 is obtained. As we use k-means both as decision-function and for automatic classification, the generated clusters are exactly our expected result. In summary, the influence of outliers on the result of k-means has been significantly reduced by our proposed outliers removal algorithm.

The k-means results for different minPoint are given below:

Table 8. K-means results for different minPoint

minPoint	Removed outliers	High	Medium	Low
2	2	60	60	33
3	3	62	60	30
4	3	62	60	30
5	3	62	60	30
6	3	62	60	30

In our data set, as the extremely large data points are rare, the minPoint has relatively low influence on the result. Bigger minPoint can capture more outliers, so it should be adjusted according to specific data sets. According to the k-means results for different minPoint, the result with 62, 60 and 30 number of points in low, medium and high categories would be reasonable and could be selected for further analysis.

Using above k-means clusters, relevant government departments, such as the energy sector, could do more analysis to guide the energy use of regional industries. In our experiment, we conducted several statistical methods to show some basic facts in companies with different consumption levels.

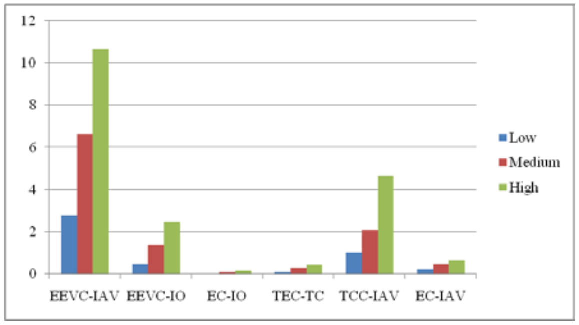

Figure 8. Average value of six indicators in three consumption levels

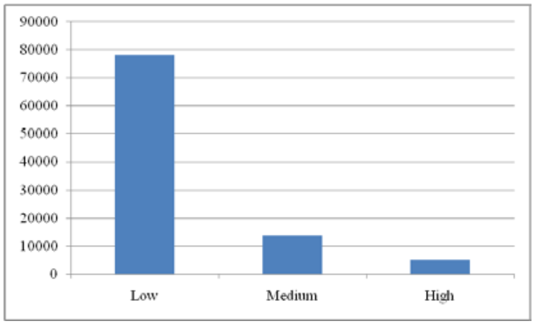

Figure 9. Average industrial output in three consumption levels

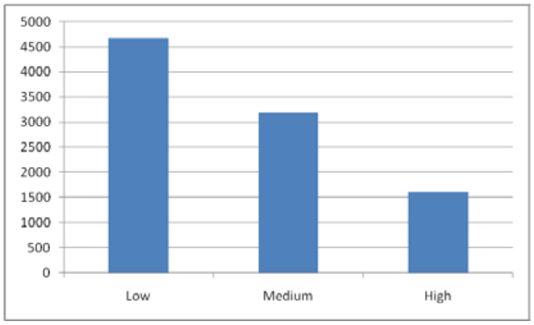

Figure 10. Average total energy cost in three consumption levels

Figure 8 shows the average value of six indicators in three consumption levels. It is clear that the k-means method has automatically classify the data according to the consumption indicators. In these three levels, the average industrial output and average total energy cost are calculated. We can see that companies with high consumption rate tend to have less industrial output as well as total energy cost.

To measure the trends statistically, we examined the correlation between industrial output of a company and its energy consumption situation. As the distributions of all these indicators are strongly right-skewed, we selected Spearman's coefficient, which does not have strong restriction on the distribution of data and does not need linear relationship between variables, to measure the statistical dependence between industrial output and some other indicators. The result correlation coefficient table is showed below:

Table 9. Correlation coefficient between industrial output and other indicators

	IO
IAV	0.943218
TC	0.976889
TEC	0.645084
EEVC	0.611584
EEVC-IAV	-0.65903
EEVC-IO	-0.7491
EC-IO	-0.63598
TEC-TC	-0.62593
TCC-IAV	-0.43276
EC-IAV	-0.51088

The result shows that the industrial output has very strong positive Spearman correlation with industrial added value, energy equivalent value consumption and other overall consumption indicators. By contrast, there is a very significant negative Spearman correlation between industrial output and all relative consumption indicators. This means when industrial output of a company grows up, its relative consumption indicators, such as energy equivalent value consumption per 10,000 CNY output value, tends to decrease. This is partly because enterprises with high industrial output may have more capital to utilize the state-of-the-art technology and do scientific research, which give them more advantages in terms of using energy efficiently.

5. Conclusion

In this paper, we developed a regional energy consumption analysis model based on the energy consumption data from Shaoxing city, Zhejiang province, China. By using k-means clustering method, we classified the energy efficiency data into low, medium and high clusters representing different energy consumption levels. However, before utilizing k-means, there are a few of potential outliers in some of the indicator data, and the result of k-means might be strongly influenced by these outliers. To reduce the impact of these outliers, we first defined the criteria of outliers in our experiment. Then, by examining the distribution of consumption data, we proposed a distance-based outliers removal algorithm and a corresponding parameters choosing algorithm. The experimental results show that our algorithms can effectively reduce the influence of outliers. In particular, when removing 3 outliers, the numbers of points in low, medium and high categories change from 81, 70 and 4 to 62, 60 and 30 respectively, which makes further statistical analysis more feasible. As one possible way of further analysis, the relationship between levels of consumption and industrial output was examined. In our future plan, more statistical analyses could be conducted based on the results of our analysis model.

References

[1] Dudley, B. (2015). BP statistical review of world energy. June 2015. London, UK.

[2] Chen, S. (2014). Environmental pollution emissions, regional productivity growth and ecological economic development in china. China Economic Review, 35, 171-182.

[3] Yan, H. (2015). The integration of energy, environment and health policies in china: a review. Amse Working Papers.

[4] Ke, J., Price, L., Ohshita, S., Fridley, D., Khanna, N. Z., & Zhou, N., et al. (2012). China's industrial energy consumption trends and impacts of the top-1000 enterprises energy-saving program and the ten key energy-saving projects. Energy Policy, 50, 562-569As the access to this document is restricted, you may want to look for a different version under "Related research" (further below) orfor a different version of it.

[5] Wu, C. (2014). Low-carbon policy and action in the chinese mainland: an overview of current development. Chinese Studies, 03(4), 157-164.

[6] Abilock, H., Bergstrom, C., & Brady, J. (1979). Markal: a multiperiod, linear-programming model for energy systems

analysis (bnl version). Nasa Sti/recon Technical Report N, 80.

[7] Jiang, B. B., Chen, W., Yu, Y., Zeng, L., & Victor, D. (2008). The future of natural gas consumption in beijing, guangdong and shanghai: an assessment utilizing markal. Energy Policy, 36(9), 3286-3299.

[8] Can, S. D. I. R. D., & Price, L. (2008). Sectoral trends in global energy use and greenhouse gas emissions. Energy Policy, 36(4), 1386–1403.

[9] Hippel, D. V., Suzuki, T., Williams, J. H., Savage, T., & Hayes, P. (2011). Energy security and sustainability in northeast asia. Energy Policy, 39(11), 6719-6730.

[10] Storper, M. (2000). The regional world: territorial development in a global economy. Economic Geography, 76.

[11] Barnes, T. J., & Gertler, M. S. (1999). The new industrial geography : regions, regulation and institutions. Routledge.

[12] Jiang, A. Q., & Zhang, R. (2011). An analysis of shaoxing textile industry's international competitiveness: the perspective of the export evaluation indexes. Advanced Materials Research, 331, 722-725.

[13] Zheng, X., Zhang, Z., Yu, D., Chen, X., Cheng, R., Min, S., ... & Wang, J. (2015). Overview of membrane technology applications for industrial wastewater treatment in China to increase water supply. Resources, Conservation and Recycling, 105, 1-10.

[14] Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 40(1), 200-210.

[15] Martins, L. D. O., Junior, G. B., Silva, A. C., Paiva, A. C. D., & Gattass, M. (2009). Detection of masses in digital mammograms using k-means and support vector machine. Elcvia Electronic Letters on Computer Vision & Image Analysis, 8(2), 39-50.

[16] Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and k -means algorithm for market segmentation. Computers & Operations Research, 29(11), 1475-1493.

[17] Lingras, P., & West, C. (2004). Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems, 23(1), 5-16.

[18] Xu, G. A., & Liu, X. (2012). Research on financial industry classification and calculation methods of quarterly financial value-added. Statistical Research.

[19] Lou, X. F., & Zou, F. X. (2010). Energy Consumption Optimization of the Aluminum Industrial Production Based on K-means Algorithm. Proceedings of 2010 International Conference on Computer,Mechatronics, Control and Electronic Engineering (CMCE 2010) Volume 3 (Vol.3, pp.61-64).

[20] Hernández, L., Baladrón, C., Aguiar, J. M., Carro, B., & Sánchez-Esguevillas, A. (2012). Classification and clustering of electricity demand patterns in industrial parks. Energies, 5(12), 5215-5228.

[21] Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005). An electric energy consumer characterization framework based on data mining techniques. Power Systems, IEEE Transactions on, 20(2), 596-602.

[22] Patel, V. R., & Mehta, R. G. (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. International Journal of Computer Science Issues, 8(5).