Data Mining and Regular Analysis on Traceability of Vital Agricultural Products

Jiahao Su, Xueting Zhang, Ming Fang, Jiang Jiang*, Yingwu Chen
(College of Information System and Management,
National University of Defense Technology, Changsha 410073, China)

Abstract: As to meeting the demand of food regulatory and actively promoting the "Internet + Agriculture" action, Commerce Department has established a meat traceability system. This paper selected the 2015 annual hog slaughter data from 10 pilot cities, By the method of data mining such as Anomaly Detection and Cluster Analysis to study agricultural traceability data, the changing pattern in hog slaughter amount is deeply analyzed. We find that the changing pattern is associated with the demographic composition of the city and the immigration during the holiday. Thus we may help production enterprises to develop a reasonable production plan based on changes in consumption characteristics of the city, and could also support the government's macro-control on hog supply of various cities.

Keywords: System of Traceability on Agricultural Products, Data Mining, Anomaly Detection, Cluster Analysis

1. Introduction

In order to meet the demand of food regulatory and macro-control of the food supply, and actively promote the "Internet + Agriculture" action, Commerce Department has established a meat traceability system in 58 pilot cities in batches with the use of Internet information technology in 2013, which has effectively and timely collected more than 2 billion of commercial data in the past three years day by day. It is well known that pork is the most daily consumption among the meat, its supply and safety are closely bound with the public. Thus this paper selects the 2015 annual hog slaughter data (500,901) from 10 pilot cities and uses data mining algorithms (e.g. statistical analysis, machine learning) to discover the value from spatial and temporal view. We aim to utilize some tools, such as MATLAB and SPSS, to analyze how different holidays and different types of cities affect the amount of hog slaughter. So that we could provide the real-time information for the consumers' safe and rational consumption, and guide the enterprises to analyze the market and make a scientific production plan in response to an emergent food shortage. Meanwhile we could also assist the government in understanding the distribution overall the market. Then the agricultural information network including production, currency, and consumption will be improved to release the supply and demand information more timely and comprehensively. In addition, the government will receive our support to make a macro-control on hog supply of various cities. The rest of the paper is organized as follows. Section 2 summaries the previous research results in agricultural products (e.g. price forecasting) and some application of clustering. In Section 3, we describe the data we get from Commerce Department and the pretreatment we make with some basic analysis. Next we introduce the method we choose and how it works in Section 4. Then we describe our experiment results in Section 5. The last section is conclusion.

2. Literature Review

The previous study towards agricultural products concentrated on market forecasting based on regression. According to the database of pork price, VAR model was established for forecasting (Xiaobin Ma[1] 2007). By using the Multiple Linear Regression Mode of SPSS, the price correlation of agricultural products between Nanning and other cities in Guangxi Province was found, which could be used to forecast the hog price of Nanning within a period of time in the future (Xie et al.[2] 2010). Liu and Li[3] demonstrated the price fluctuation of 2010 by studying the case of pork price between 2002 and 2009 in China, finding the features of periodicity and heteroscedasticity via descriptive analysis, establishing Time Sequence Model of periodicity and heteroscedasticity. Taking the fresh milk retail price as example and comparing different Time Sequence Models, they found that ARCH Model was with the highest forecast accuracy while the Holt-Winters possesses the highest stability of Non-season Smoothing Model (Dong et al.[4] 2010). Wang[5] (2008) used grey systems to forecast the price of eggs, and Ping[6] (2010) used artificial neural networks based on grey systems to improve the forecast accuracy of pork price in Jilin Province. A linear quantity approach was employed to forecast price movement intervals of pork, chicken and egg, it was found that the confidence intervals from 0.05th and 0.95th are good methods to solve this problem (Li[7] 2012). In addition, Cheng Peng et al.[8] (2007) have mined the spatial and temporal patterns of pork price in 23 provinces from 2010 to 2012. What they found is that the pork price presents a geometric pattern while it's stable. Since the Food Traceability System (FTS) was proposed, many scholars such as Moise[9] (2007) and Pan[10] (2014) shared their perspectives about the system's advantage, strengthening the supervision of food safety and improving the Punitive measures of food safety problem. In the meanwhile, they also pointed out that the FTS is not perfect, and the system deficiency is the starting and focal point. It makes the traceability data more available by clustering the complex and discrete data into continuous classification information, which was found out by clustering complex and strict time-sequence regularity base on the method of functional data analysis (Gao[11] 2012). Recently, some other scholars have studied consumer modes by using data from the FTS. In Stranieri's [12] (2009) research, it was raised that the people who pay more attention to the meat label cares a lot about the meat quality, and these people are always young females.

Furthermore, data mining techniques has been used in various areas contributed by its rapid development. Anomaly Detection is often applied to economic, meteorological and medical fields, such Zhou[13] (2008), who integrated the KNN with time-sequence partition and then proposed an efficient time-sequence anomaly detection algorithm, which could be proved effective and reasonable after implementing the experiment with ECG data. Clustering is adopted more often in biology, marketing and management decisions, such as Ketchen[14] (1996) introduced Clustering into Policy Management and Chen et al. [15](2009) obtained time regularity of varied crimes with the method of Hierarchical Cluster Analysis. Apparently, Clustering is available to be applied to agriculture, such as Peng's[8] (2007) Spatial Clustering of pork price mentioned above and Yan's (2010) [16]finding on drought resistance by Clustering of different paddies. In addition, Ming[17] (2010) described the principle of Temporal Data Mining to try to increase applicable value of agriculture data by implementing data mining. In order to analyze the amount pattern of slaughter, we also chose Clustering as the principal method in this paper.

3. Data Pretreatment and Analysis of Changing Pattern

Commerce Department established an agricultural products traceability system for the supervision of food currency and safety. It has collected hog data from cultivation to sale. This paper selects the 2015 annual hog slaughter data from farms to slaughters of 10 cities (Chengdu, Chongqing, Dalian, Hangzhou, Ningbo, Nanjing, Qingdao, Shanghai, Wuxi and Kunming) as the Table1.

At first, some pretreatments are made on these data. We consider every pig weights 90 kilogram, and plus each trade which occurs in the same city and on the same day to get the slaughter data of each city and each day. If there is some default value, we think the slaughter may not work on this day, so it is considered as 0. Particularly, the data of City 10 is serious deficiency, we only select the data end with May 18th.

Table 1 The pilot cities

City Code	City Name	Data Times（T）
C_1	Chengdu	2015.01.01——2015.12.31
C_2	Chongqing	2015.01.01——2015.12.31
C_3	Dalian	2015.01.01——2015.12.31
C_4	Hangzhou	2015.01.01——2015.12.31
C_5	Ningbo	2015.01.01——2015.12.31
C_6	Nanjing	2015.01.01——2015.12.31
C_7	Qingdao	2015.01.01——2015.12.31
C_8	Shanghai	2015.01.01——2015.12.31
C_9	Wuxi	2015.01.01——2015.12.31
C_{10}	Kunming	2015.01.01——2015.05.18

Before analysis the amount of hog slaughter's changing pattern, every city's weight should be normalized per day. Because different cities bring about different population, different income and different dietary habits, we focus on how the important holidays affect the changing pattern of hog slaughter's amount. These factor which may cause differences in total amount should be excluded. Thus we assume $C_i(i=1,2,...,10)$ as 10 pilot cities, $y_{C_id}(d=1,2,...,T)$ as the weight of City C_i per day, then the standard consumption of City C_i per day is

$$X_{C_i d} = \frac{T * y_{C_i d}}{\sum_{d=1}^{T} y_{C_i d}} \tag{1}$$

Figure 1 presents the changing of the 2015 annual hog slaughter from 10 pilot cities. The vertical axis (10 ribbons) represents the 10 pilot cities, and the horizontal axis represents the date. There are great changes in various cities from the seasonal point of view, some cities have a large amount of slaughter in spring, such as C_8 and C_9, and some in summer and autumn, such as C_1 and C_3.Such data dispersion characteristics reflect that the frequent pattern is not easy to be excacated. Therefore, the particle size of the study is set to “day”, whether there are some special regulations in various cities on various days.

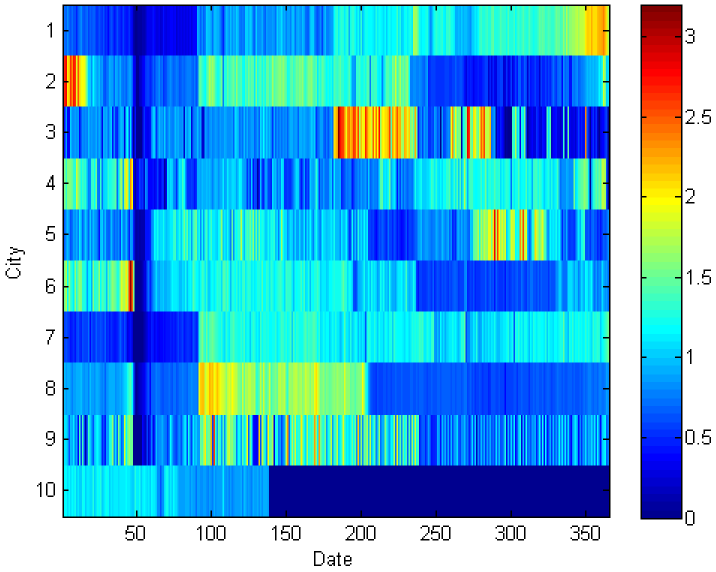

Figure 1 The changes in amount of the 2015 annual hog slaughter in 10 pilot cities

4. Anomaly Detection Based on Statistics and Hierarchical Clustering

Anomaly Detection is used to explore the special regulation may exist among different periods for the amount of hog slaughter in various cities. Then similar cities and similar holidays will be clustered into the same group by Clustering in order to discover changing regulation in various cities.

4.1 Anomaly Detection

Anomaly Detection, also known as Outlier Analysis, which is an important aspect of data mining is used to find out the data significantly different from other data. The reason behind it is always not stochastic but inevitable. Table 2 presents some usual methods of Anomaly Detection, and gives their fundamental with their merits and demerits.

Since the data of this study is one-dimensional data, and follows normally distribution, we choose the method based on statistics. Because a distribution model is assumed and the confidence interval is set, when some data exceeds the interval, it is considered as an outlier, because the further the data is away from its centre, the smaller the probability is. Therefore, it is judged as follow:

$$|x-\mu| > k * \sigma \tag{2}$$

where x is the value of the observation, μ is the average of the samples, σ is the standard deviation of the samples, and k is a constant given by the confidence interval we need. The advantage of this method is that the method is built on sophisticated statistics, the outliers are defined unambiguous, and the computational time is shorter.

4.2 Clustering Analysis

Clustering Analysis is based on the truth 'like attracts like', and is a classification of the samples. It could make a reasonable classification according to their characters without any prior knowledge. Since the units of data selected are different, the magnitudes are different, a normalization is absolutely necessary. While z-score is commonly used as a standardized method, which can score a real reaction from the average relative standard distance. It is calculated as follow:

Table 2 Some usual methods of Anomaly Detection

The Methods of Anomaly Detection	Fundamental	Merits and Demerits
Based on Statistics	The normal data is assumed as following a distribution with Θ as parameter. The object x's probability is given by probability density function, the value is smaller, the more probable x is an outlier,	Merits: there is a statistical basis； Demerits: data detection results for high-dimensional or multiple data are not satisfied.
Based on KNN	If an object x is far away from the most points, it is likely to be an outlier.	Merits: it has broad applicability and simple. Demerits: it needs a long computational time ($O(n^2)$)， is highly sensitive to the value of k, and cannot deal with the data with different densities.
Based on Density	If an object x is in a low density area, it is likely to be an outlier.	Merits: Outlier is described quantitatively, even the data with different densities could be handled. Demerits: it needs a long computational time ($O(n^2)$), and is hard to choose parameters.

$$z = \frac{x-\mu}{\sigma - \sqrt{\mu}} \tag{3}$$

where x is the value of the observation, μ is the average of the samples, σ is the standard deviation of the samples, and n is the number of the samples.

Table 3 shows some usual methods of Clustering Analysis. Hierarchical Clustering is chosen, since the data set of our study is not large and the number of clusters is needed exactly. At the beginning we consider each sample as a cluster, then we calculate closeness between each cluster by a given algorithm. Particularly, the maximum Euclidean distance is applied in this paper (the furthest neighbor clustering algorithm). The distance between two clusters P_i and P_j is defined

$$dist_{\max}(P_i, P_j) = \max_{x \in P_i, x' \in P_j} \{|x - x'|\} (i \neq j) \tag{4}$$

where $|x - x'|$ is the Euclidean distance between x and x'. As we have get the distance, then the closest clusters are classified as a new cluster. Afterwards It is repeated until all samples clustered into one group.

Table 3 The usual methods of Clustering Analysis

The methods of Clustering Analysis	Fundamental	Merits and Demerits
Partitional	Data objects are divided into non-overlapping subsets (clusters) such that each data object is in exactly one subset	Merits: it takes heuristic algorithm which costs a short computational time, and is suitable to discover a small globular cluster. Demerits: the number of clusters are needed to be set, and this method is sensitive to the initial, and is hard to deal with outliers and noises.
Hierarchical	A set of nested clusters is organized as a hierarchical tree.	Merits: it is not sensitive to the initial, and could handle outliers and noises. Also there is no need to set the number of clusters. Demerits: the error of the division cannot be corrected, and it takes a long computational time ($O(n^2 logn)$).
Based on Density	A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.	Merits: the clusters of arbitrary shape cab be discovered, and the outliers can be filtered. Demerits: it takes a long computational time ($O(n^2)$).

5. Experiment Results and Analysis

In this paper, the possible existence time of special rules, which is in the changes of the amount of the annual hog slaughter, is selected by Anomaly Detection. The various cities and holidays are clustered and analyzed in order to explore the patterns and influencing factors behind the changes of the amount of the hog slaughter.

5.1 Anomaly Detection on the Annual Data

There are usually some fluctuations in the amount of hog slaughter, so it is lack of significance to discover the normal fluctuation. Thus we only study on the case of outliers. While Anomaly Detection can assist us to find where the outliers are in various cities. At first the amount of the hog slaughter is assumed as following the Gauss Distribution, and 95% is the confidence interval that we need. Then the outliers are sentenced by:

$$|x-\mu| > 1.96 * \sigma \quad (5)$$

where x is the amount of the slaughter of the observation, μ is the average of the annual slaughter, and σ is the standard deviation of the annual slaughter. When the observation point is far away from the 1.96 times standard deviation (beyond the confidence interval), this point is an outlier and should have some special laws. Figure 2 demonstrates the outliers in various cities during the whole year. We find that the dates there are outliers over five cities are 49 and 50, while the 50th day is the Spring Festival. Obviously, it is reasonable to infer there are some special changing patterns during the Spring Festival. The changing regulation of the amount of the hog slaughter during the Spring Festival need to be discovered, and cities can be clustered according to these regulations. Meanwhile whether there are some different changing patterns in different holidays can be analysed.

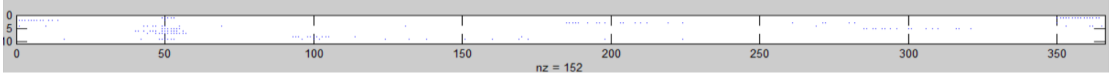

Figure 2 The outliers in the annual hog slaughter

5.2 Clustering Analysis of Cities and holidays

As described above, the changes in the amount of the hog slaughter during the Spring Festival have a special pattern, but there are some similarities and differences between the cities, for this the Clustering Analysis is utilized to classify the cities, and to explore their inherent variation patterns.

(1) Clustering Analysis of Cities during the Spring Festival

This paper selects the data from 6th Feb. to 8th Mar. in 2015 (fifteen days around the Spring Festival) to mine the mode of the hog slaughter of different cities during the Spring Festival. In order to analyze the changing patterns, the normalization method described as above is applied to deal with the data, so that the whole year's factors will be removed. Then daily amount of the hog slaughter is taken as the features, a bottom-up aggregation strategy is chosen, the Euclidean distance is used to measure how close it is between clusters, and we use SPSS to assist us to cluster.

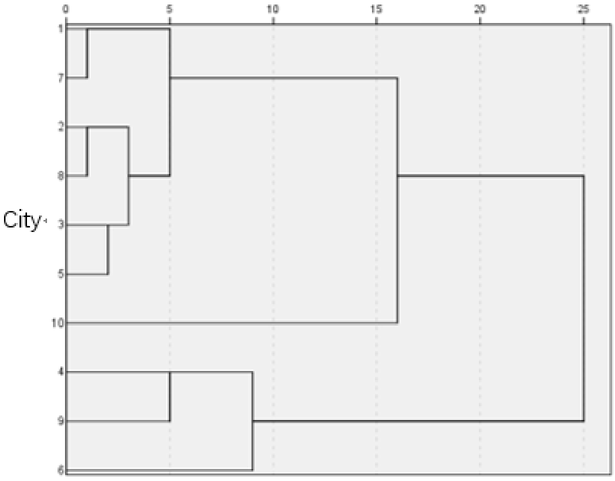

Figure 3 Clustering and Analysis's dendrogram of the changes in the amount of the hog slaughter in pilot cities during the Spring Festival

Figure 3 shows the Clustering and Analysis's dendrogram of the changes in the amount of the hog slaughter in pilot cities during the Spring Festival. The vertical axis represents 10 cities, and the horizontal axis represents the distance between the cities. It tells that three clusters is chosen most appropriately to guarantee that the cities among the same cluster are similar and there is big difference between any different groups. The result of clustering is that the first category consists of C_1, C_2, C_3, C_5, C_7 and C_8, the second category consists of C_{10}, and the third category consists of C_4, C_6 and C_9.

Figure 4 demonstrates that the changing pattern is associated with the demographic composition of the city. Specifically, the city in the first category (such as Shanghai) showed as solid lines contains many large enterprises with large external population. Then a massive outflow appears during the Spring Festival, resulting in a hog slaughter plunged, while large enterprises' Spring Festival vacation is short and people return soon, so the slaughter quickly rebounds after the vacation. The city in the second category (such as Kunming) showed as a short line is dominated by tourism, with a large influx of tourists during the Spring Festival, hog slaughter remains high. After the vacation, a large number of visitors leave, then it occurs a decline. The city in the third category (such as Nanjing) showed as dotted lines contains many small and medium enterprises(SMEs) with many migrant laborers, and plenty of laborers outflow during the Spring Festival, so hog slaughter plunged appears, and the laborers in SMEs return late, leading to slow recovery in hog slaughter. In order to cope with the amount of the Spring Festival, a large number of factories are slaughtered before the holiday, so there is a pre-holiday surge.

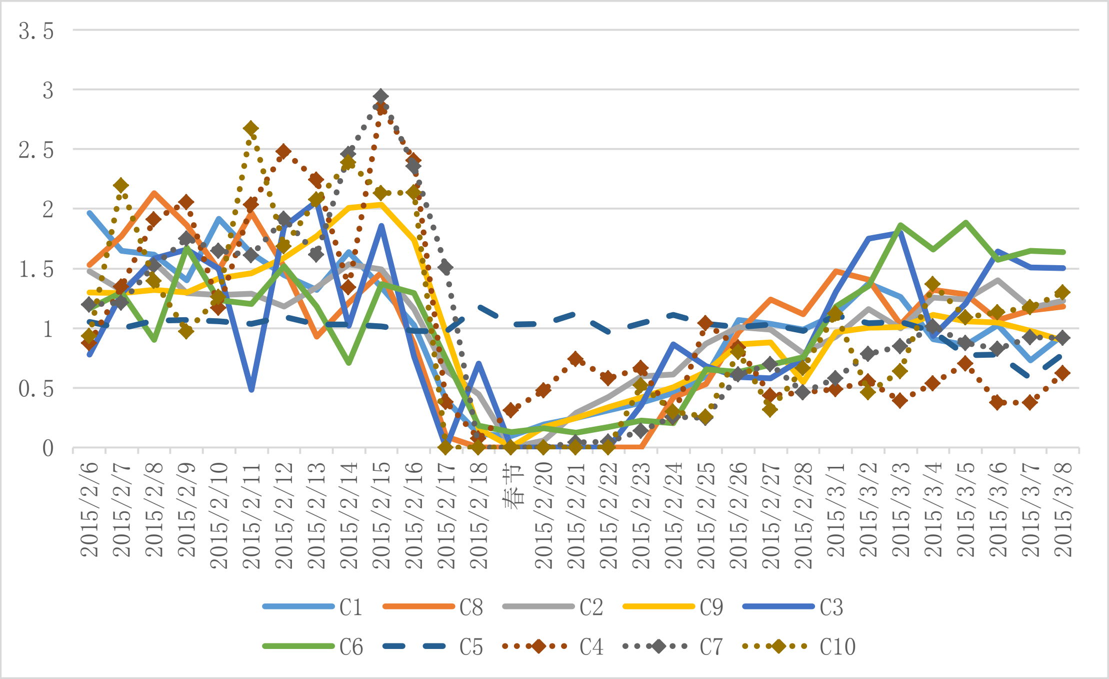

Figure 4 The trend of three categories during the Spring Festival

(2) Clustering Analysis of various holidays

Because of the data imperfection of Kunming (Kunming's hog slaughter amount just be collected up to 18th May), we choose the other 9 cities as the examples to analyze changing patterns of hog slaughter amount during festivals. The changing patterns of hog slaughter amount is mined in different pilot cities during several traditional festivals by Comparing the hog slaughter amount during the Spring Festival with four other traditional festivals, which are the Tomb-sweeping Day (from 28th Mar. to 13th Apr.), the International Labor Day (from 24th Apr. to 10th May), the Dragon Boat Festival (from 13th Jun. to 29th Jun.) and the Mid-Autumn Festival (from 19th Sep. to 14th Oct.).

The mean, variance, range, and the median of the hog slaughter amount in each pilot city are taken as the 4 features of every city's data, meaning that there are 36 data features of each festival. Likewise, the proper cluster number is unknown, therefore we integrate the hierarchical clustering method with SPSS. Particularly, the meanings of features are different, leading to the gap of data in the order of magnitude, and that is why we standardize the data before Clustering according to the Z-Score.

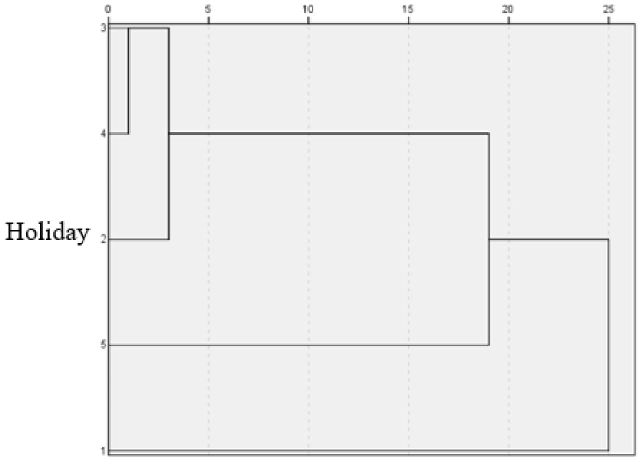

Figure 5 Clustering and Analysis's dendrogram of the changes in the amount of the hog slaughter during different holidays

Figure 5 shows the Clustering and Analysis's dendrogram of the changes in the amount of the hog slaughter during different holidays. The vertical axis represents 5 holidays, and the horizontal axis represents the distance between the holidays. It tells that the Spring festival has obviously the biggest difference from the others, and it is clustered by its own as the first category, which is consistent with the results of Anomaly Detection. Then the second category contains the Tomb-sweeping Day, May Day and the Dragon Boat Festival. And the third category only contains the Mid-Autumn Festival. Figure 6 presents that the average amount of various cities of the first category is obviously smaller than others (it is seen by the mean and the median) and the volatility is larger than others (it is seen by the variance and the range). It is because that people are accustomed to return home during the Spring Festival holidays and vacation is long enough to travel, a massive outflow will appear before the Spring Festival and a mass reflux will occur after vacation. Also, the vacation of the second category is short and people will not travel frequently, as a result, the average amount is totally larger than others and the volatility is totally smaller than others. While not only the average amount but also the volatility of the third category is between the other two. As some citizens in cities have plenty of time to travel, which leads to a population flow. While the flow is smaller than the Spring Festival's travel, so that the average and volatility of the third category are between the others.

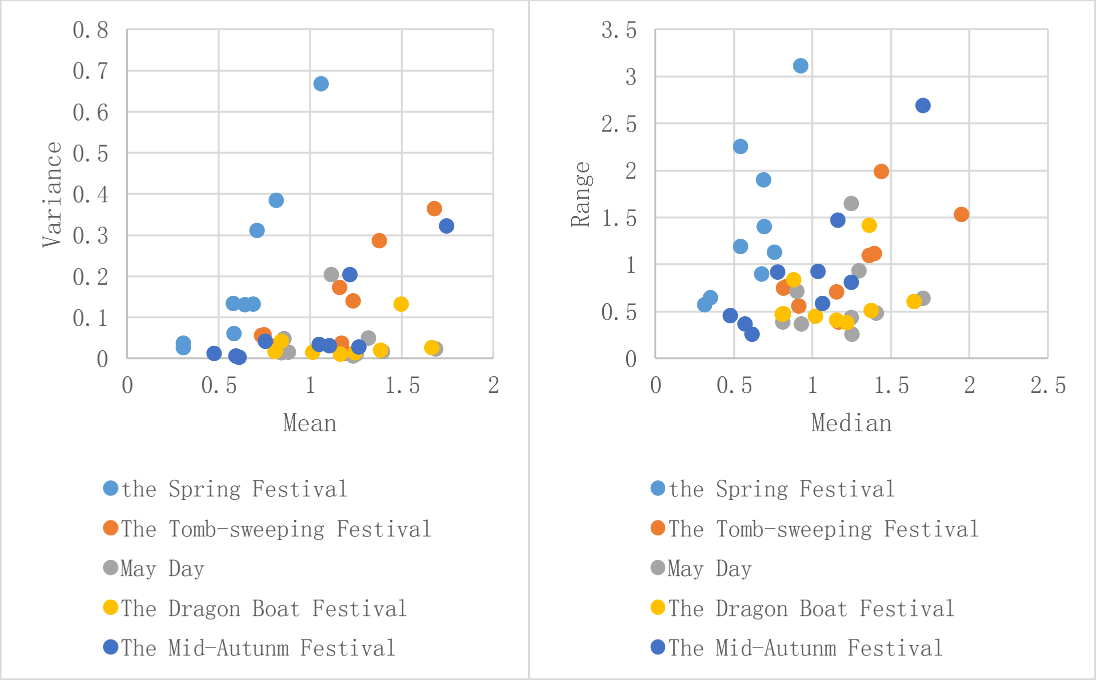

Figure 6 The Scatter Plot of Mean-Variance and the Scatter Plot of Median-Range of various cities

6. Conclusion

By the methods of data mining such as Anomaly Detection and Clustering to study agricultural traceability data, the whole year's changing pattern of daily hog slaughter amount in 2015 is deeply analyzed. It comes the conclusion that the changing pattern of hog slaughter amount during the Spring Festival is related to the characteristics of a city. To illustrate further, Shanghai mainly consists of large enterprise, leading to a hog slaughter plunged during the Spring Festival, while the slaughter quickly rebounds after the vacation. Tourism is the main economic support of Kunming, which contributes to the hog slaughter remains high during the Spring Festival, while it declines after the vacation. Small and medium-sized enterprises is the main component of the Nanjing corporate structure, corresponding to this situation, the hog slaughter plunged appears during the Spring Festival, while it recovers slowly and there is a pre-holiday surge. In the meanwhile, we do some research on changing patterns of hog slaughter amount in different cities, the result shows that there is a significant relationship between the duration and the habits of holiday and hog slaughter, the longer the holiday is and the more major the festival is, the lower the average slaughter during the holiday is and the greater the volatility is. Which may help production enterprises to develop a reasonable production plan based on changes in consumption characteristics of the city, and could also support the government's macro-control on hog supply of various cities. In subsequent studies, price changes will be put into consideration and the research methods will be spread, expanding data from the national level to the provincial level, and from the meat to aquatic products, medicines or any other product which can be traced back. So that we can master the market rules more adequately, assist the government in scientific decision-making, instruct enterprises in efficient production, and help consumers rational consumption.

References

1. Ma X B, Wang T, Dong X, Wang C D. Using VAR to forecast pig price. Chinese Journal of Animal Science,2007(23), 4-6.
2. Xie H W. Exploring prediction method for agricultural product prices in Guangxi - Case studies in Nanning City. Guangxi Agricultural Sciences, 2010(41), 862-865.
3. Liu X, Li J Z. Analysis and forecast on China's pork price based on periodicity and heteroscedasticity time series model. Journal of the Central University for Nationalities (Natural Science Edition), 2009(18), 106-109.
4. Dong X X, Li G Q, Liu Z J. Choice and application of short-term forecast method for agricultural products price-taking fresh milk retail price as example. Shandong Agricultural Sciences, 2010(42), 109-113.
5. Wang S H. Application of grey forecasting model in forecasting egg price. Guide to Chinese Poultry, 2008(25), 48-50.
6. Ping P, Liu D Y, Yang B, Jin D, Fang F, Ma S J, Tian Y, Wang Y. Research on the combinational model for predicting the pork price. Computer and Engineering and Science, 2010(32), 109-112.
7. Li G Q, Xu S W, Li Z M, et al. Using quantile regression approach to analyze price movements of agricultural products in China[J]. Journal of Integrative Agriculture, 2012, 11(4):674-683.
8. Cheng P, Wu H R, Huang F, et al. Data Mining on Agricultural Products' Price Based on Spatial Statistics[J]. Research of Agricultural Modernization, 2014, 35(1):29-32.
9. Resende-Filho M. A Principal-Agent Model for Investigating Traceability Systems Incentives for Food Safety[J]. Mois és Resende Filho, 2007.
10. Pan W J, Wang J. Study on Food Safety in the Perspective of Supply Chain Network [J]. Forum on Science and Technology in China 2014(9):155-160
11. Gao R, Wang Q, Luo D, Qin Z G, et al. Functional Cluster Analysis of Time Series Data in Food Traceability[J]. Journal of University of Electronic Science and Technology of China, 2012, 41(4):561-563.
12. Stranieri S, Banterle A, Fritz M, et al. Fresh Meat and Traceability Labelling: Who Cares?[J]. Alessandro Banterle, 2009:663-673.
13. Zhou D Z, Liu Y F, Ma W X. Effective time series outlier detection algorithm based on segmentation[J]. Computer Engineering and Applications, 2008, 44(35):145-147.
14. Ketchen D J, Shook C L. The Application of Cluster Analysis Strategic Management Research: an Analysis and Critique[J]. Strategic Management Journal, 1996, 17(6):441-458.
15. Chen P, Shu X M, Yan J, et al. Timing of criminal activities during the day[J]. Journal of Tsinghua University (Science and Technology), 2009(12):2032-2035.
16. Yan M J, Huang W Z, Hu J T, et al. Application of Cluster Analysis in Drought-resistant Material Categorised on Rice [J]. Journal of Anhui Agricultural Sciences, 2010, 38(19):9998-10000.
17. Ming D G, Xu Q, Zhou Z B, et al. The Application Research of Temporal Data Mining in Predicting the Price of Farm Products[J]. Software Guide, 2010, 9(3): 140-142.