The Study of Smokers' Attitude and Behavior in China Based on Data Mining Methods

Fei Yu
School of Statistics and Mathematics
Yunnan University of Finance and Economics
Kunming, China
1350691353@qq.com

Fu Xiaojing
Automobile Insurance Department
China Pacific Property Insurance Co.
Nanchang, China
765588126@qq.com

***Abstract*—Using data mining methods, this paper studies the smokers' attitude and behavior in China. The sample data are from 12 provinces including Yunnan, Shandong, Guangxi, Guangdong, Jiangsu, Shanghai, Beijing, Zhejiang, Jiangxi, Hubei, Hunan and Sichuan. Based on the research and *Frame Work Convention on Tobacco Control* of World Health Organization, the paper puts forward corresponding strategies of smoking control in our country.**

Keywords—Attitude and behavior; Correspondence analysis; Association rules; Tobacco control strategy.

I. INTRODUCTION

With the development of the tobacco industry, china has been the largest tobacco producing and consuming country in the word. The development of tobacco is related to the economy of nation, but tobacco do a great harm to the public's health. WHO(2010) has pointed out that tobacco dependence is a chronic addiction disease, that is listed in the international diseases classification, and Xiao *et al*(2008) confirm that tobacco is one of the critical threats to human health. Therefore, it is essential to put forward an effective tobacco control strategy which will be very propitious on the public health.

In recent years, studies on tobacco control strategy, which were done at home and abroad, while were mostly focused on teenagers, few were focused on smoking attitude or behavior and the reasons why they smoke. Diverse population characteristics of smokers are considered in this paper, it is more comprehensive. In addition, the previous studies on smokers' attitude and behavior were few related with tobacco control strategy, the existing research did not make a profound analysis. Therefore, it is a key point in this paper of combining results of analyzing the smokers' attitude and behavior with the perspective of tobacco control strategy in all countries, especially in developed countries, and we propose some suggestions on tobacco control strategy in China.

II. THE DESIGN OF QUESTIONNAIRE AND DATA PREPROCESSING

A. The Design of Questionnaire

This paper adopts the hierarchical sampling method, selecting smokers from 12 provinces, such as Yunnan, Shandong, Guangxi, Guangdong, Jiangsu, Shanghai, Beijing, Zhejiang, Jiangxi, Hubei, Hunan and Sichuan. The following three parts are the main survey contents:

- Background information: including gender, age, smoking year, city, educational attainments, profession and other basic information of the smokers.
- The importance level of the demand for cigarettes. There are 11 indexes related to the cigarette demand: cigarette packaging, flavor, tar, popularity, quality, price, brand, convenience of purchasing, made in China, advertising and friend recommendation. The level of the option is judged and measured by the degree of evaluation, which means 1 represents the least important, 2 represents not important, 3 represents general, 4 represents important, and 5 represents the most important.
- Smoking causes (Smokers social psychology): this part has 29 variables, also using the degree of importance to judge and determine the causes.

In the questionnaire, using the digital rules, the background of smokers is ranked by 1, 2, 3, 4, 5... respectively. Such as the occupation, 1, 2, 3…and 11 means civil servants, farmer, teacher, worker, professional, freedom occupation, manager, clerk, services staff, personnel, student and others.

B. Data Preprocessing

In this paper, due to the rarely missing values and sufficient data samples, simple deletion method is used for the exits missing values in the processing data.

Due to the presence of outliers, Wang (2012) has touched that we must carry out the detection and inspection of abnormal value. In this paper, we use *Grubbs* to test the abnormal value and the software package is *outliers*.

In *R* programming software, outlier test usually uses the *Outliers* in *Grubbs* test. Generally this test could only check one or two outliers of a data set .The specific code is as follows:

Grubbs. Test (*x*, *type*=10, *opposite*=FALSE, *two. sided* = FALSE)

In the command above, *x* means the detection data vector, *type* means the detection type. If *type*=10 means that one

This work is supported by Yunnan Province and Shanghai University of Finance and Economics Education Cooperation Consulting Project （42111217003）.Corresponding author: Fu Xiaojing.

outlier is been detected. Furthermore, two outliers at the same side means *type*=20.

C. Correspondence Analysis

Wu (2010) referred that, in *R* software, we use *Corresp* function to realize correspondence analysis, *biplot* function to draw a graph of correspondence analysis, and the function is as follows:

Corresp(*x*, *nf* =1, …),

Where *x* means carrying on the corresponding analysis of the data matrix, *nf* means the number calculation factor, the function *biplot* can draw intuitive correspondence analysis figure.

D. Association Rules Analysis

The algorithm used common in association rules analysis is *Apriori* algorithm. In *R* language, we use *arules* package to realize the algorithm.

III. The Empirical Analysis

A. Analysis of Smokers Descriptive Background

This part mainly collects the data and does a basic description. We can see from sample data that the number of valid questionnaires is 4960, in which men accounted for 97.9%, women accounted for only 2.1%. In term of the sample size for each city, Yunnan province has the largest amount of samples, accounted for 14.2% of the total, Shanghai and Hunan has less sample volume accounted for only 5.9% and 6%.From the point of educational attainments, the number of undergraduate accounted for 28.0%, followed by high school education accounted for 27.2%, graduate or above only accounted for 8.8%. From the age, 20-30 years old accounted for 36.5%, people over the age of 60 accounted for only 4.8%. From the monthly income, the number of consumers that monthly income is 1001-2000 Yuan accounted for 37.5%, above 5000 Yuan only accounted for 8.0%. From the point occupation, the number of students accounted for 13.26%, the number of managers accounted for only 5.48%.

B. Analysis of Smoking Attitudes and Behaviors of Different Population Groups

According to previous studies, we know that smokers' attitude and behavior is different. For example, men are more likely to smoke than women and smokers of different age, occupation and others have different demands for cigarettes. On the one hand, in order to dig out the related information in the context of existing smokers, correspondence analysis and association rules on the relationship between the different characteristics of population background have been studied. On the other hand, in order to dig out other useful information, different analyses of cigarette demand have been studied.

1) Background analysis of the different characteristics of population

From Fig. 1, we can see that the people who smoke 5-10 Yuan cigarette usually have smoked for 1-3 years. The people who smoke 10-20 Yuan cigarette usually have smoked for 5-10 years.

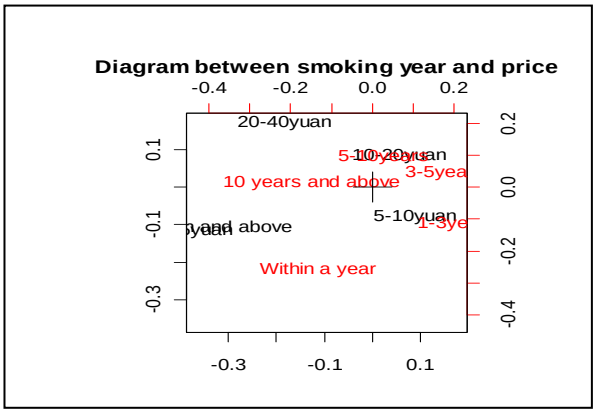

Fig. 1. Correspondence analysis diagram of the relationship between the smoking year and the price of cigarettes

As the same as the above analysis, we can also draw these following conclusions (diagrams are omitted):

By correspondence analysis of the relationship between the age of smokers and the price of cigarettes, we can see that 20-30 years old smokers, known as teenagers, usually buy 5-10 Yuan or 10-20Yuan cigarettes. This also shows that the price of cigarettes is relatively lower for teenagers.

By correspondence analysis of the relationship between the occupation and the price of cigarettes we can see that smokers who buy 3-5 Yuan cigarettes are mainly farmers; Smokers who buy 5-10 Yuan cigarettes are mainly services staff and students; smokers who buy 10-20 Yuan cigarettes are mainly teachers, professionals, technical personnel and company staff; Smokers who buy above 40 Yuan cigarettes are mostly managers.

2) Analysis of attitude and behavior on demand for cigarettes

a) Correspondence analysis

In order to clearly show correspondence analysis diagram of the relationship between attitude and behavior of smokers and demand for cigarettes, the demand for cigarettes with corresponding A, B, C…and K mean packaging, flavor, tar, popular, price, brand, convenient, made in China, advertising and recommendation respectively.

From Fig. 2, we can see 30-40 years old smokers are more focused on A, which means cigarette packaging.

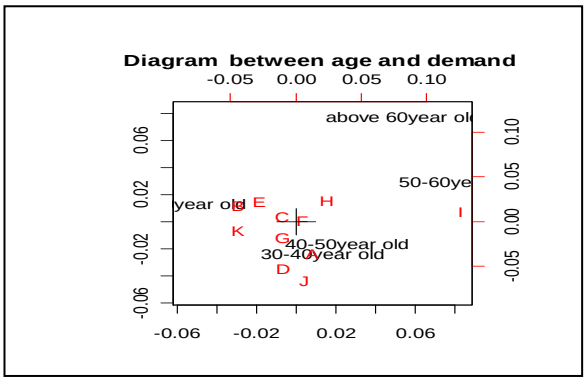

Fig. 2. Correspondence analysis diagram of the relationship between age and cigarette demand

From Fig. 3, we can see that the distance between1001-2000 and the word I is close, according to the mean of demand for cigarettes, the word I means the demand that cigarettes made in China, it also said that the smokers whose monthly income is 1001-2000 Yuan like to buy cigarettes made in China; As the same as the above description, 2001-3000, A and D are scarcely overlapped, A and D mean packaging and popularity respectively. So we can reach a conclusion that those whose monthly income is 2001-3000 Yuan pay more attention to the cigarette packaging and popularity.

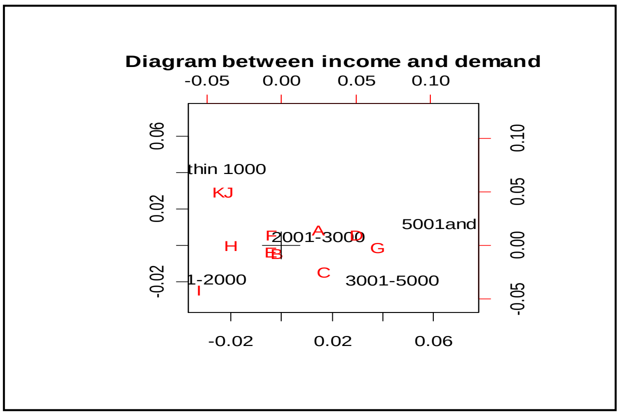

Fig. 3. Correspondence analysis diagram of the relationship between monthly income and cigarette demand

From Fig. 4, we can find that the distance between 1, 4, 7 and A, D, H is very close, and combined the rank of occupation and the mean of demand for cigarettes, we know 1, 4 and 7 rank civil servants, clerk and worker, A, D and H mean packaging, convenient and popular, so we can reach another conclusion that civil servants, clerks and workers will choose convenient place to buy cigarettes and are more concerned about popular and packaging of cigarette.

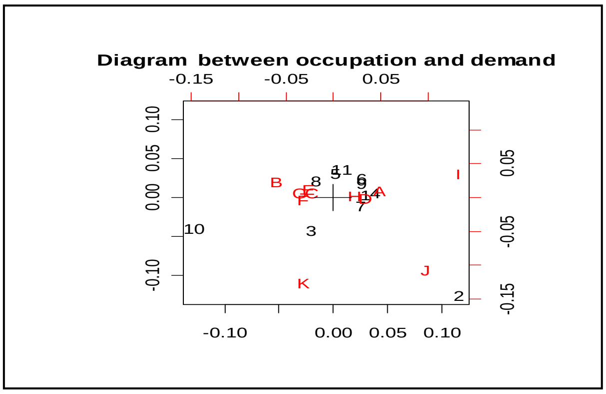

Fig. 4. Correspondence analysis diagram of the relationship between occupation and cigarettes demand

b) Association rules analysis

Analysis between monthly income, age, price of cigarettes, smoking age, educational attainments and cigarette demand, the importance of attribute of cigarettes is two division variables, The importance level 1-3 means less important, 4-5 means important.

Analysis of smokers' background and cigarette demand, the results are as follows:

```
> inspect(sort(x, y="support")[1:5])
  lhs               rhs                        support     confidence    lift
1 {} => {b1Quality = important}      0.8221762    0.8221762    1.0000000
2 {} => {b1Flavor= important}        0.8031088    0.8031088    1.0000000
3{}=> {b1Advertising= less important }0.7670466  0.7670466 1.0000000
4 {b1Flavor=important }
        => {b1Quality=important }  0.7187565    0.8949677     1.0885353
5 {b1Quality=important }
        => {b1Flavor=important }   0.7187565    0.8742123     1.0885353
```

The results show that, for 82.2% of consumers of the questionnaire, quality is important. For 80.3% of the consumers, flavor is also very important. However, for 76.7% of consumers, advertising is less important; the results of frequent two item sets show that the smokers who think that quality and flavor is both important accounted for 71.8% of the total number. Among those who like flavor, the percent people also like the quality is accounted for 89.5%.

C. Analysis of Smoking Behavior and Reason

According to previous studies, there are many reasons for the smoking behavior, so the variables in the questionnaire are very rich. Firstly we use factor analysis to reduce dimension to extract the main factors that influent. Secondly, dig out association rules relations from main factors.

1) Factor Analysis

Reduced dimension to extract the main five factors in 29 variables of smoking behavior by factor analysis, which means motivation of smoking, they are Charming image, communication, distinctive, psychological dependence and identity respectively, and are said by 1, 2, 3, 4 and 5 in the following TABLE1.

TABLE I. ROTATED COMPONENT MATRIX

Rotated component matrix					
	Component				
	1	**2**	**3**	**4**	**5**
q24Smoking great style	0.704				
q26Smoking can enhance manhood	0.697				
q15Smoking can increase personal charm	0.609				
q21Smoking is a symbol of mature man	0.568				
q22 I prefer imported brands of cigarettes.	0.511				
q23The cigarette is a good gift in China		0.649			
q25Embarrassment and tension can be hided in some occasions		0.644			
q17In communication, offering cigarettes is a good start.		0.643			
q27 Smoking can help your work and business.		0.602			
q11 Smoking is due to curiosity or imitate.			0.649		
q19 It is a kind of treason.			0.561		
q14 When I smoke, never consider the feelings of others.			0.504		
q3 The smoking is pleasant.				0.618	
q7 Smoking help thinking.				0.607	
q13 Get self satisfaction during				0.571	

the smoking process.					
q5 The brand of cigarettes can display personal identity.					0.648
q8 The consumption of a brand of cigarettes can make others understand what kind of person I am.					0.558
q6 The people around me buy the same brand of cigarettes and so do I.					0.524

2) Association rules analysis

By association analysis in main reasons for the smoking, the results are as follows:

items	support
1 { q16Smoking helps me relax=4}	0.4997909
2 { q25Embarrassment and tension can be hided in some occasions=4}	0.4623588
3 {q6sh The people around me buy the same brand of cigarettes and so do I,=3}	0.4316186
4 {q27ch Smoking can help your work and business =4}	0.4268089
5 {q17sh In communication, offering cigarettes is a good start =4}	0.4171895

The first record shows that about 50% of smokers think that smoking helps them to relax; the second record shows that 46.2% of smokers consider that smoking can hide embarrassments and tension. 43.1% of smokers do not agree that the people around them buy the same brand of cigarettes as they buy.

IV. THE MAIN CONCLUSIONS AND POLICY SUGGESTIONS

According to the basic requirements of the *Fame Work Convention on Tobacco Control* and the combination of empirical analysis, we put forward the following suggestions:

- By correspondence analysis of the relationship between the age of smokers and the price of cigarettes, this also shows that the price of cigarettes is relatively lower for teenagers, we can give the following suggestions on teenagers: 1) As we all know, the teenagers are more sensitive to the price than the adults, the most effective way to prevent them smoking is to increase the tobacco tax. The higher the price is the fewer teenagers to start smoking, and also higher price helps reduce the smokers' consumption. 2) Try to prohibit branch or small package of cigarette sales, due to this kind method of marketing will improve the teenagers' purchasing ability.
- From Fig. 2, it can be seen that, for the 30-40 year old smokers, they pay more attention on cigarette packaging, so we can take some warning notes like "smoking is harmful to health" on cigarette packs, then they can recognize the dangers of smoking from awareness. The specific measures are as follows: 1) In tobacco packaging promotion, deceiving the consumer by false means is forbidden. Such as printing words like "low tar" or "light" in the packaging to hide the harmful substances in tobacco, and making false report of release to give a wrong impression to the consumers and mislead them. 2) Mark some warnings or suitable information on tobacco product packaging and labels. These information and warnings must be clearly printed by pictures, and the visible area of the pictures should cover more than 50% of a packet. 3) Raw materials and releases of cigarettes should also be printed on the packet of cigarettes, and should also follow the provisions of the national authorities.
- From Fig. 4, it can be seen that civil servants, clerks and workers will choose convenient place to buy cigarettes and are more concerned about popular and packaging of cigarette. From the perspective of tobacco control strategy, the amount of tobacco retail shops should be as few as possible around companies; this can prevent civil servants and clerks buying cigarettes. The packaging is referred in the above conclusion, we can put into effect.
- We do a comprehensive association analysis on the background of smokers and their demand for cigarettes, on the result, we find that there are 82% of consumers care about the cigarette quality. So we may find a way to make those smokers be less eager cigarettes buyers. Just imagine if the manufacturers and importers of tobacco products are asked to disclose composition of tobacco products and their release to the government authorities. Once consumers know that the cigarettes they smoke contain toxic substances, it will prevent some consumers who care about the quality.
- Now let us consider the factor analysis in smoking behavior. It can be seen from TABLE I the motivation of smoking is divided into five aspects, and they are Charming image, communication, distinctive, psychological dependence and identity. Thus, we can take measures to control the phenomenon of smoking in the interpersonal communication, make smokers feel smoking is no longer a good shape and symbol of identity. Based on this, the measures we can make.1) Education and communication and training on tobacco should be expanded, it will raise public awareness, so as to achieve to prevent smoking.2) Let us do my best to minimize the exit time and frequency of smoking in television, media, advertising and other related smoking scenes.

REFERENCES

[1] Science and Technology Daily, WHO: "Smoking is a Chronic Addiction Disease" [EB/OL], April 2010.

[2] Xiao Dan, Wang Chen and Wen Xinzhi, "Tobacco dependence is a chronic disease", China Health Education, 24 (9)2008, pp.721-722.

[3] Wang Huailiang, "R language identification and statistical data of abnormal value", Research and development of electronic technology, May, 2012

[4] Wu Xizhi, " The complex data statistical method based on R application (Second Edition) ", Renmin University of China Press, September, 2010

[5] Qin Ning, "Study on the development of tobacco industry from the perspective of global tobacco control", Chinese Ocean University of Chinapress, June,2011