DISCOVERING CONSUMER PREFERENCES ON THE MARKET OF DRINKING WATER CONSUMPTION

Prof.dr.sc. Sanja Bijakšić

Faculty of Economics University of Mostar, B&H

E-mail: sanja.bijaksic@sve-mo.ba

Prof.dr.sc.Brano Markić

Faculty of Economics University of Mostar, B&H

E-mail: brano.markic@sve-mo.ba

Prof.dr.sc.Slavo Kukić

Faculty of Economics University of Mostar, B&H

E-mail: slavo.kukic@sve-mo.ba

-Abstract-

The goal of the research is to analyze the market supply of drinking water and consumers choice behavior when are buying the drinking water packed in the bottles. The paper is result of empirical research where the data are collected through the survey and the analysis was conducted using the method of principal component analysis. The respondents are evaluated on ordinal scale the importance of each criterion when deciding whether or not to by the product. Principal component analysis captures the essence of correlation among data and reduces the n-dimensional space to a few most important components that represent consumer preferences.

Key words: *consumer behavior, knowledge discovering, principal component analysis*.

JEL Classification: M31

1. INTRODUCTION

Subject to study consumer behavior is the behavior of individual and group customers in the process of buying, using and deprivation of products and services

on one hand, and the impact of this process on consumers and society on the other. Through the process of research is being done and collect data about the factors that influence consumers' decisions and overall behavior. The goal of the research is, ultimately, to provide an understanding of the ways in which consumers buy and use the purchased goods and services.

The creation and design of each marketing program requires insight into the preferences, expectations, beliefs, acquiring information and a variety of other processes going on in the minds of consumers..

The knowledge of the addressed theoretical premises we entered in the detection consumer preferences, behavior and demand in the market of bottled water. Large number of variables that describe the behavior of consumers of water from the bottle can be replaced by a smaller number of components without significant loss of information about the behavior and preferences of consumers. This is the main hypothesis of the study. The empirical study, which would be to test this hypothesis has been realized on the territory of Bosnia and Herzegovina, and the obtained results are the same or have a similar importance in general, but especially in countries that have similar overall economic and social development as well as Bosnia and Herzegovina.

True, this kind of assessment will not be able to substantiate and research in other countries because they do not have information th at a similar type of study in them, at least as far as the region of Southeastern Europe, until now, has been realized.

2. THE METHODOLOGY USED

To test the hypothesis, we entered an empirical research, which in the market of B&H was realized by surveys and questionnaires technique as an instrument composed of eighteen questions. The survey was carried out on stratified sample of 150 respondents, where the stratification was done according to the four elements - gender, age, education, and the average monthly household income.

3. THE RESEARCH RESULTS

The methodology in this paper is based on principal component analysis and R programming language. At the same time R denotes three things: data analysis software, calculator and programming language.[1]

3.1. Principal component analysis

Principal component analysis is a statistical method that belongs to the multivariate analysis. Formally, the main idea of principal component analysis can show the following: a set of n-dimensional vector samples $X = \{x_1, x_2, x_3, ..., x_m\}$ can be transformed into a new set $Y = \{y_1, y_2, y_3, .., y_m\}$ of the same size [1] but y has the property that its largest information content stored in the first few dimensions. Principal component analysis transforms a set of correlated variables (performance indicators) to a smaller set of mutually orthogonal (uncorrelated) variables called principal components. General data matrix with n rows of them correlated variables x_1, x_2, ... x_m transform the mind of new variables that are uncorrelated [5]. Suppose that is given data matrix X with **n** rows and **p** columns:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

The rows represent observations (in our example the answears of respondents in survey) and the columns variables (in our example, the values of answears to the questions). From the matrix X is calculated covariance matrix [1]:

$$A = \frac{1}{n-1} X^T X = \frac{1}{n-1} \begin{pmatrix} \sum_{i=1}^{n} x_{i1}x_{i1} & \dots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{ip}x_{i1} & \dots & \sum_{i=1}^{n} x_{ip}x_{ip} \end{pmatrix} = \begin{pmatrix} c_{11} & \dots & c_{1p} \\ \dots & \dots & \dots \\ c_{p1} & \dots & c_{pp} \end{pmatrix}$$

Covariance matrix S is symmetric. New variables y_i are linear combinations of the original variables x_i:

[1] R costs nothing and is completly free. To install R on your computer visit the site 1. http://cran.r-project.org/mirrors.html and choose the nearest mirror.

$y_i = a_{i1}x_1\ a_{i2}x_2 + + ...\ a_{ip}x_p$; i= 1,2,...p.

New variables y_i are derived in decreasing order of their importance. These are called principal components (principal components). The next step is to calculate their eigenvalues **λ** and eigenvectors **e** of the matrix C. Then solve the equation

A * e = λ * e . Eigenvalues λ odf the matrix are sorted in descending order: $\lambda_1 \geq \lambda_2 \geq ... \lambda_n$. The aim is to present the data using a new set of mutually orthogonal coordinate (the new base). Eigenvalues are variances, one for each own vector.

3.2. The principal components analysis and R language

For principal component analysis we are using R language. "R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environment for data mining, statistical analyses, and data visualization"[7]. Unlike other programming languages in the R language is not necessary to declare variable because the type is ***determined automatically*** when variable is created with "<-" operator. This programming language commands are intuitive and easy to use. Therefore, we will immediately applied the statements of R language in the analysis of data collected through the questionnaire and stored in a data set called wa1An.csv.

3.3. Discovering customer behavior and preferences

The first thing that we will want to do to analyse our multivariate data will be to read it into R, and to plot the data. We can read data into R using the read.table() function. Our data set contains twelve answears of respondents to the survey. The data set looks like this:

```
> waPCA<-read.table("C://wa1An.csv",header=T,sep=";")
> WAPCA
   QB QT LB T PC BS PP PR LC BR PE MP[2]
```

[2] The abbreviations in the columns mean:
QB = Quality of bottled water; QT= Quality of tap water; LB= Label on bottle about the quality and health benefits; T= Taste of bottled water; PC= Color packaging of bottled water; BS= Shape of packaging of bottled water; PP= Convenience packaging of bottled water; PR=Price: LC= Impression of fewer chemicals in the water from a bottle: BR= Recyclable bottles filled with water; PE= The possibility of increasing employment for recycling and filling water bottles; MP= Media promotion.

1 1 5 2 3 2 3 3 3 4 1 1 3

2 4 3 2 4 2 2 2 2 3 1 2 4

.....

149 4 2 5 4 5 5 5 4 5 4 5 5

150 4 3 3 4 4 3 3 4 4 4 2 4

These twelve variables define the key reasons for the decision to purchase water from a bottle. This raises the hypothesis that these twelve variables can be reduced to a smaller number of variables and to retain key information about customer behavior in the buying process.

To standardise[3] the variables of consumer behavior in purchasing bottled water and carry out principal component analysis on the standardise values, we type the statement:

> stanVal<-round(as.data.frame(scale(waPCA[1:12])),1)

> stanVal

	QB	QT	LB	T	PC	BS	PP	PR	LC	BR	PE	MP
1	-2.1	1.6	-0.6	-0.2	-0.8	-0.3	-0.5	-0.3	0.6	-1.3	-1.4	-0.3
2	0.9	-0.1	-0.6	0.6	-0.8	-1.2	-1.4	-1.1	-0.2	-1.3	-0.6	0.5
.....												
150	0.9	-0.1	0.4	0.6	0.7	-0.3	-0.5	0.5	0.6	1.1	-0.6	0.5

The stanradised values of variables have the standard deviation 1 and the mean 0. The tortal variance is therefore equal to the number of variables.

The next step in principal component analysis is calcualtion of covariance using the next statement in R language (all the values are rounded to two decimal points):

> covaW<-round(head(cov(stanVal)),2)

> covaW

[3] The principal components were computed after replacing each original variable by a standardized version of the variable that has unit variance.

	QB	QT	LB	T	PC	BS	PP	PR	LC	BR	PE	MP
QB	0.97	-0.14	0.28	0.31	0.16	0.17	0.09	-0.09	0.19	0.12	0.07	0.24
QT	-0.14	0.97	-0.15	-0.07	0.06	0.10	0.01	-0.01	-0.14	0.20	0.19	0.08
LB	0.28	-0.15	1.00	0.38	0.32	0.24	0.31	0.18	0.16	0.25	0.21	0.08
T	0.31	-0.07	0.38	0.99	0.44	0.26	0.27	-0.02	0.26	0.16	0.08	0.17
PC	0.16	0.06	0.32	0.44	1.05	0.56	0.44	0.19	0.25	0.20	-0.03	0.34
BS	0.17	0.10	0.24	0.26	0.56	0.97	0.56	0.20	0.17	0.20	0.16	0.29

In principal component analysis we need to calculate the eigenvalues and corresponding eigenvectors:

```
> eigenvalues<-eigen(cov(stanVal))$values
> eigenvalues
 [1] 3.3237593 1.5806841 1.3092924 1.1384527 0.9787556 0.7466877 0.7016972
0.5562358  0.5014641 0.4349769 0.3696358 0.2827669
> eigenvectors<-eigen(cov(stanVal))$vectors
>round(eigenvectors,2)
      [,1]  [,2] [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10] [,11] [,12]
 [1,] 0.21 -0.21  0.49 -0.19 -0.04 -0.03  0.60 -0.46 -0.05  0.02  0.21 -0.12
 [2,] 0.02  0.44 -0.31 -0.43  0.12  0.50  0.18 -0.13  0.33  0.30  0.11  0.01
 [3,] 0.32 -0.10  0.25  0.35  0.39  0.11  0.23  0.50  0.18  0.44 -0.07  0.10
 [4,] 0.32 -0.23  0.25 -0.10  0.27  0.39 -0.40 -0.17  0.32 -0.46 -0.16  0.12
 [5,] 0.40 -0.22 -0.27 -0.19  0.17  0.20 -0.15  0.13 -0.55  0.06  0.17 -0.50
.....
[11,] 0.22  0.60  0.24  0.13  0.01 -0.16 -0.04 -0.07  0.06 -0.09 -0.40 -0.56
[12,] 0.26 -0.09 -0.05 -0.43 -0.46 -0.08  0.25  0.56  0.20 -0.28 -0.17  0.00
```

In order to decide how many components should be retained we can use three different methods:

a) screeplot function and mark the change in slope

b) Kaiser's criterion reatains components for which is the variance above 1 when principal component analysis was applied to standardised data

c) keep the number of components required to explain at least some minimum of toal variance.

By implementation plot() function is obviously that the change in slope occurs at component fourth.

> plot(waPCA.pca,type="l")

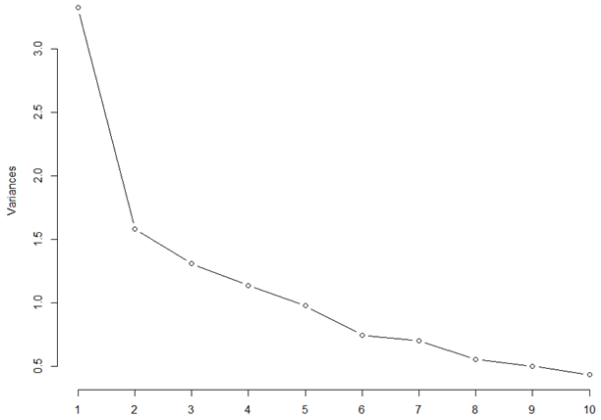

Figure-1: The screeplot function and determination the number of components

We will use the Kaiser's criterion. The values for the first four components is the variance above 1 (3.32, 1.58, 1.31 and 1.14, respectivaly). Therfore, using the Kaiser's criterion we will retain the first four principal components. Loadings for the principal component analysis are the values of eigenvectors.[4] The result is the matrix with the loadings of each principal components. This means that the first principal component (PC1) is the linear combination of the variables:

PC1 = 0.21 *QB +0.02 * QT + 0.32 *LB +0.32 * T+ 0.40 * PC +0.38 * BS +0.40 * PP + +0.15*PR +0.28 * LC + 0.29 * BR+0.22 *PE + 0.26 *MP

[4] The same result can be obtained by immplementation the function prcomp() on standardised data set named ***stanVal.***

PC2= 0.21 *QB -0.44 * QT + 0.10 *LB +0.23 * T+0.22 * PC +0.04 * BS -0. 05 * PP + 0.05*PR +0.08 * LC - 0.52 * BR-0.60 *PE + 0.09 *MP

PC3= - 0.49 *QB +0.31 * QT - 0.25 *LB -0.25 * T+0.27 * PC + 0.33 * BS +0. 24 * PP + 0.44 *PR - 0.13 * LC - 0.19 * BR-0.24 *PE + 0.05 *MP

PC4= 0.19 *QB +0.43 * QT - 0.35 *LB +0.10 * T+0.19 * PC + 0.15 * BS -0. 08* PP - 0.60 *PR - 0.17 * LC - 0.06 * BR-0.13 *PE + 0.43*MP

.....

PC12= 0.21 *QB -0.01* QT -0.10 *LB -0.12 * T+ 0.5 * PC +0.40 * BS -0. 06 * PP + +0.05*PR -0.03 * LC - 0.49 * BR + 0.56 *PE + 0.00 *MP

The first principal component has the highest loadings for PC (0.40), PP (0.40), BS (0.38), LB (0.32) and T (0.32 - taste of bottled water). Therefore the interpretation the first principal component could be that it represents the packaging (appearance) of bottled water.

The second principal component has the highest loadings (in absolute value) for QT (0.44), BR (0.52) and –PE (0.60). This component reflects the social and economic context regarding the use of water from the bottle. Specifically, higher spending means more jobs. The second component is the social importance of the consumption of water from the bottle.

The third component has highest loading for QB(-0.49), PR (0.44), BS (0.33),

QT(0.31). The loadings for PR, BS and QT are positive and loading for QB is negative. The third component reflects the economic criteria of price, quality of water and component packaging design (price performance component).

The fourth component has highest loading for PR (-0.60), QT (0.43), MP (0.43) and LB (-0.35). These component reflects a contrast between water price, label of bottles on one side and media promotion and the quality of tap water on other side. This component can be described by saying that higher prices and better information on the bottle have the opposite effect in relation to the media promotion and the quality of tap water. These component describe the demand of bottled water determined by price, packaging, promotion and the quality of tap water.

The new standardise values (scores) can be calculated by implementation the next statement in R language (formed by multiplying the loadings with the original data): >PC<-as.matrix(stanVal)%*%eigenvectors

>round(PC,2)

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	-1.94	-0.10	-1.65	-0.52	-0.37	1.06	-1.18	0.25	1.18	1.09	-0.24	0.65
.....												
[150,]	1.04	-0.49	0.65	-0.01	-0.48	1.14	0.44	0.19	-0.73	-0.32	0.52	0.44

The values of the principal components are stored in a matrix with the principal components, where the first column in the matrix contains the first principal component, the second column the second component, and so on.

Thus, in our example, " PC[,1]" contains the first principal component, and " PC[,2] " contains the second principal component.

We can make a scatterplot of the first two principal components, and label the data points with the price or how the price of bottled water inluences to the consumption, by typing:

> plot(PC[,1],PC[,2])

> text(PC[,1],PC[,2], waPCA$PR, cex=0.5, pos=3, col="blue")

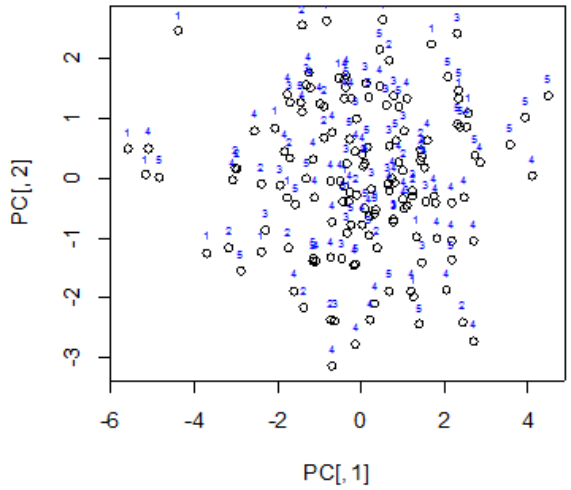

Figure-2: The scatterplot of first (PC[,1]) and second (PC[,2]) principal component

The scatterplot displayes the first principal component on the x-axis, and the second principal component on the y-axis. It is obviously from the scatterplot that does not exist any regularity regarding the values of the two principal components and the price of bottled water. The values of the first and second do not separate the answears of respondents regarding the price and are not are reasonably useful for distinguishing the answears to query about the price of bottled water.

4. CONCLUSION

This research confirmed the hypothesis that it is possible to detect the key components that describe the preferences and behavior of buyers in the market of bottled water. Methodology applied to the data collected about the market and consumer behavior showed applicative power to detect consumer preferences. In the data set built up on answears of respondents in survey, we have 12 variables describing the behavior of customers on the market of bottled water in B&H. By carrying out a principal component analysis, we found that most of the variation between answears of respondents can be captured using the first four principal components, where each of the principal components is a particular linear combination of the 12 variables. These four components are packaging (appearance), social importance of the consumption, price performance (or cost benefit) and demand. The fourth component demand is combination of the price, promotion, packaging of bottled water and the quality of tap water. A professional marketing and promotion strategy has to consider these four components to be succesful.

BIBLIOGRAPHY

1. A.L. Yuille (2004), Dimension Reduction & PCA, Department Statistics Ucla.
2. Husic, M., (2009), Lifestyle and Luxury Consumption, Doctoral thesis, School of Economics and Business in Sarajevo.
3. Husić-Mehmedović, M., Kukić, S., Čičić, M. (2012), Consumer Behaviour, School of Economics and Business University of Sarajevo, Sarajevo.
4. I.T, Jollife, (2008), Principal Component Analysis, Springer Verlag.

5. Kantaradžić, M., (2003), Data mining, Concepts, Models, Methods, and Algorithms, Wiley-Interscience.
6. Lawson, R. and S. Todd (2002), Consumer Lifestyles: A Social Stratification Perspective, Marketing theory, Volume 2(3), pp. 295-307.
7. Markić, B. (2011), Integrating Theory and Practice: Knowledge Discovery and Unsupervised Learning, Meeting of Management Departments the Faculties of Economics, Faculty of Economics University of Split, September 2011,pp. 23-31.
8. Schiffman, L.G. and L.L. Kanuk, (2004), Consumer Behavior, Pearson/Prentice Hall, 8. Edition.