Alina Ioana HĂLĂUCĂ, PhD Candidate
Financial-Banking University, Bucharest
Professor Crisan ALBU, PhD
The Bucharest Academy of Economic Studies

SEGMENTATION PROCESS OF CLIENTS' DATABASE

Abstract. *In order to identify and forecast the consumer's behavior, it was applied the principal components and cluster analysis in the segmentation process, identifying groups of clients with similar needs. Starting from a bank's database of reliable clients using consumer loans, I emphasized the main characteristics of the loan applicant, reducing also the number of targeted clients, using the two methods of analysis. .*

Key words: *consumer, lending, data mining.*

JEL Classification: C38, C81

I. Introduction

The high number of clients from a general market, one as financial-banking services, as well as the multitude of requests led to poor service coverage of a company regarding all possible consumers. Due to this the company needs to identify the market segments that could be served as efficient as possible. These details impose distinct buyer group identification and definition of characteristics, establishing main services/products advantages and bringing this information forward.

Starting point in any marketing analysis represents the variable identification that defines the problem studied and responds best to customer needs.

While marketing literature offers a wide range of variable groups that can be used in different circumstances, the multidimensional analysis offer also mathematical and technical ground needed for the development and solution of these studies.

The current paper will show how can we gain advantage from principal components and cluster analysis in order to resume the available data and extract correct information that characterize best the potential bank client groups.

II. Typology of principal components analysis

Developing studies regarding consumer behavior characterization and forecast requires analyze of existing connections between variables, identified and selected, based on data analyze methods.

Introduced by Pearson K. [15] (1901) and developed by Hotteling [8] (1931), the principal components analysis is one of the oldest and most used multivariate technique that has as purpose the variation description of a database set through uncorrelated variables. The analyze follows the non-redundant split of the total variance of the initial space with respect to the already existing information, in a small number of components in order to simplify the problem in study and retain the needed information.

For example if we want to measure the need of a bank client loan contract we could develop an appliance form where we could ask questions regarding activity sector and income level. If a high correlation between the two above questions will result we can consider the information as redundant. The relationship between the two variables, used in the form, can be transposed in a scatter plot, where a regression line could describe the linear connections between variables. If we define a variable that can approximate the regression line then this will contain the largest information gained from the original variables. So, the new factor defined as a particular linear combination of the original variables, could be used in later analyses. Summarizing two correlated variables in a single factor reinforces the main idea of the principal components analysis. If we extend this example to a larger number of variables, the above stated principle remains the same.

In a first dimension consisting of a n counted for variables notate x_1, x_2,...,x_n, where each measurement has a vector x shape of dimension m, m representing the characteristics number and considering the generic ad notation w for the principal component, each w_j coordinate is defined as:

$$w_j = \sum_{i=1}^{n} \alpha_{ji} x_i$$

Based on a matrix, identifying the linear combination coefficients requires choosing the highest value out of eigenvalues of the covariance matrix and calculating the vector components of percents, α[1],based on the eigenvector of matrix linked to its eigenvalue, Σ.

Identifying the vector components lead to the Lagrange multipliers method and following margin issue [18]:

$$\begin{cases} \max\limits_{\alpha} \quad \alpha^t . \Sigma \alpha \\ \text{SR: } \alpha^t \cdot \alpha = 1 \end{cases}$$

The solution is one of the eigenvectors of the covariance matrix Σ, respectively the one associated with the eigenvalue $\tilde{\lambda}$ of the same matrix. Assuming that all n eigenvalues of the covariance matrix are arranged as follows $\square_1$>= $\square_2$>=...>=$\square_n$, the first principle component that will have the maximum variance $\square_{1,}$ results from

[1] α is the ad notation for coefficient vector that define linear combination for a component

the following linear combination $w_1 = (\alpha^1)^t x$. α vector is that eigenvector of the covariance matrix sigma that corresponds to the highest eigenvalue and respects the boundaries:

$$(\square - \square_1 I)\alpha^1 = 0$$
$$(\alpha^1)^t\alpha^1 = 1$$

The eigenvalue is the square root of the characteristically equation $| \square - \square I| = 0$ where I is the Unit matrix.

Once identified, the matter becomes how many principal components should be extracted, taking into account that the principal components analysis is a method of variables number reduction. In theory the number of principal components is equal to the number of original values. The decision of stopping is related to the moment when only a poor variability is left, of interest being now only these linear combinations information significant, characterized by high variability. So the nature of the decision is rather arbitrarily although in theory we have some guidelines like Kaiser[2] criteria, variance percentage criteria[3], a priori principle[4], screen diagram (Cattell)[5]. We can analyze a number of solutions through selecting a different number of factors and choosing the one that provides the best interpretation.

Once identified, a low number of variables can be used as starting point for the cluster analysis.

III. Typology of cluster analysis

Cluster analysis follows the classification of multivariate data organized in groups, so that we can observe the resemblance and difference between studied cases. For example, in marketing research, we can use cluster analysis in order to group a high number of consumers regarding their needs on a market where a certain product exists. So, as uncontrolled recognition technique of shapes, the cluster analysis classifies objects while assuring a minimum variation inside classes and maximum one outside. Any ranking is a pure split of the studied objects based on a set of rules that can be studied depending of their usefulness.

Clusters identification can be done beginning with charts analyze. As well the results of principal components analysis may be used.

Regardless if we speak about agglomerative or crumbling hierarchical clustering techniques or about partitioned algorithms, summing up the literature regarding

[2] Choosing pincipal components with eigenvalues higher than 1

[3] Selecting a number of principal components in order to reach a certatin procent from the total variance

[4] The researcher orders a computer to stop after a pre-established number of factors

[5] Diagram impling the identification of point where eigenvalues slowly decerase and lead to overlapping of their line with horizontal plot axe

cluster analysis allows three principles on which most users develop most applications: aggregation, optimization, models based on density functions. We need to underline that the number and group consistence are unknown [7, 16, 18, 20].

Hierarchical clustering techniques, resulting from more than one merger, allow data grouping in a certain number of classes or clusters.
The algorithm consists of: clusters C_1, C_2,…,C_n that contain only one subject; identifying pairs of clusters, for example C_i and C_j, and merging them; deleting C_j and decreasing the number of clusters until 1. If the number of clusters is equal to 1 the algorithm stops, otherwise it returns to phase 1. Hierarchical clustering techniques are different with respect to the cluster distance identifying methods.

The single linkage method (nearest-neighbor method) defines the similarity between clusters as the shortest distance from any object in one cluster to any object in the other.
The complete linkage method (farthest- neighbor method) determines the cluster similarity based on maximum distance between observations in each cluster. Similarity between clusters is the smallest sphere that can enclose all observations in both clusters.
Similarity of any two clusters is defined by the average linkage as the average similarity of all individuals in one cluster with all individuals in another. This method does not depend on extreme values as the above two and tends to generate clusters with small within-cluster variation.
In the centroid method the similarity between two clusters is the distance between clusters centroids, respectively between the mean values of the observations in one cluster. If a new individual is added to the group, the cluster centroid changed.
Ward's method defines the similarity between two clusters as the sum of squares within the clusters summed over all variables. The purpose is to select the combination of clusters that minimizes the within-clusters sum of squares.

All agglomerative hierarchical methods lead to only one final cluster that contains all subjects; even if there are more sophisticated methods that can select the number of clusters, the researcher is the one that has to decide the solution that describes the problem best.
Different studies developed in order to point out which grouping method is more efficient recommend using complete linkage or group average [5]. Problems in agglomerative hierarchical clustering techniques are pointed out: simple linkage can lead to problems regarding unwanted results if between groups separate individuals interpose (usually these individuals are integrated in an already available cluster instead of creating a new one); complete linkage and group average tend to impose a new structure instead of using the one available; the use of a sum of squares measure makes this ward's method easily distorted by outliers.

In nonhierarchical procedures the assignment of objects into clusters is done once the number of clusters is specified, without involving the treelike construction process. Several clustering algorithms have been proposed, as: sequential, parallel and optimization.

The starting point of the sequential method implies including all objects within a prespecified distance in one selected cluster seed. Then a second cluster seed is selected and all objects within the prespecified distance of that seed are included and so on. The main disadvantage of the method is that when an observation is assigned to a cluster it cannot be reassigned to another cluster even if that cluster is more similar.

The parallel method considers all clusters seeds simultaneously and assigns observations to the nearest seed.

The optimizing procedure is similar to the sequential one, except the fact that allows the reassignment of the observation based of the goal of creating the most distinct clusters. Optimization methods may imply individuals grouping through numerical criteria optimization (maximization or minimization) of numeric principles: total dispersion, between-group or within-group dispersion. Being unable to examine each partition separately has led to the development of algorithms oriented in identification of optimal value of grouping criteria through rearranging existing partitions [7].

Grouping observations with respect to cluster belonging (Scott and Symons [19]), re-parametrization of observations through eigenvalues decomposition (Banfield and Raftary [2]) are ways to express optimization criteria's.

In order to check the validity of a cluster solution, we may test the hypothesis that there are no groups in the population analyzed [16]. For example, the hypothesis could be that the population represents a single multivariate normal distribution, or that the observations arose from a uniform distribution. Also, we may randomly divide data into two subsets, say *A* and *B*, and carry out the cluster analysis on each *A* and *B*. The results should be similar if the clusters are valid.

A particular context and the examination of strengths and weakness of each method may suggest which is most appropriate.

IV. Development and study results

In order to point out the need of usage of analysis models for principal components and cluster analysis in the segmentation process of consumers, we have analyzed the behavior, need of loans, of the bank's customers. Starting from a database of clients that have requested consumer loans through the two methods of analyze we have followed the main characteristics of the loan consumer; the whole process is directed on simplifying variables that describe the loan client's profile and identification of segments, considering that a narrower client identification can have its benefits as well as regarding business opportunity, that would lead in selling a credit card, and as for reducing costs.

For a number of 408 clients (good payers) that already have a personal needs loan, with or without warranties, we have selected a number of nine variables in order to describe their behavior: sex, age (in years), children number (chl_no), income as appliance from (income, in RON), education, loans from other banks (loan_oth), loan balance at 01 January 2011(loan, in RON), Region, Marital status. We underline that the objective of this study is to identify the profile of a customer interested in buying a credit card.

Each client is identified through an unique ID. The information regarding the client is extracted for the form filled in when requesting a loan.

In order to have a comprehensive approach I made the following adjustments:

✓ noted the variable “sex” as “1” for male and “2” for female

✓ split the levels of education in two: “1” represents university and post university studies, “2” represents high school and other forms of study[6]

✓ the variable “loan_oth” can take values “1” for “yes” and “2” for “no”

✓ in order to have numbers also for countries I have made following adjustments (I have used the 8 development regions of Romania): 1 – North-East, 2 –West, 3 –North-West, 4 – Middle, 5 – South-East, 6 – South-Muntenia, 7 – Bucharest-Ilfov, 8 – South-West Oltenia[7]

✓ the variable “marital status”: can take “1” for “Married”, “2” for “not married” and “3” for “widow”.

A first analyze of the existing information points out that the Region Bucharest-Ilfov has the best percent between the number of credits and their value - in the graph beneath a relative low loan volume is related with a high balance, therefore we may find a better economical situation in this part of Romania:

[6] Bank practices say that level 1 includes upper studies and level 2 middle studies

[7] North-East: Ia□i, Boto□ani, Neamt, Suceava, Bacău, Vaslui; Vest: Arad, Car□-Severin, Hunedoara, Timi□; Nord-West: Bihor, Bistri□a-Năsăud, Cluj, Maramure□, Satu-Mare, Sălaj; Middle: Alba, Sibiu, Mure□, Harghita, Covasna Bra□ov, South-East: Vrancea, Gala□i, Brăila,Tulcea, Buzău, Constan□a, South-Muntenia: Prahova, Dâmbovi□a, Arge□, Ialomi□a, Călăra□i, Giurgiu, Teleorman; Bicharest-Ilfov: Bucure□ti, Ilfov; South-West Oltenia: Mehedin□i, Gorj, Vâlcea, Olt, Dolj

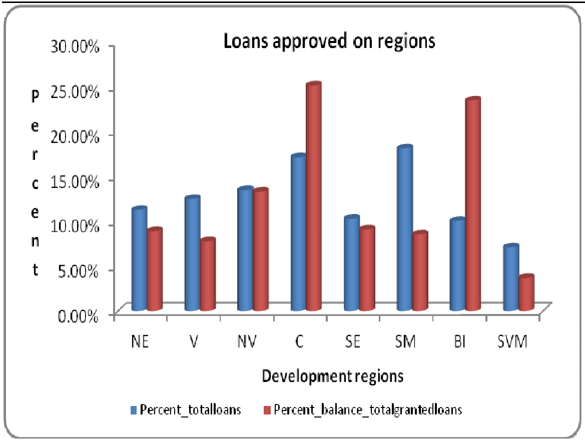

Figure 1 – Loans approved on regions – January 2011
Source: information based upon data from the bank

Taking into account that the purpose of this analyze is to select these clients that could be possible credit card buyers, we are interested mainly in the regions that would require such a card. So with respect to a balance between the number of approved loans and their value, five Regions, North-East, West, North-West, South-East and South-West Muntenia, could fit our needs.
Another aspect that could characterize the possible credit card request form is the number of bank's clients that have loans granted by other banks. So, almost 39% of the customers have loans at other financial institutions. With respect to development regions its structure is like followed represented:

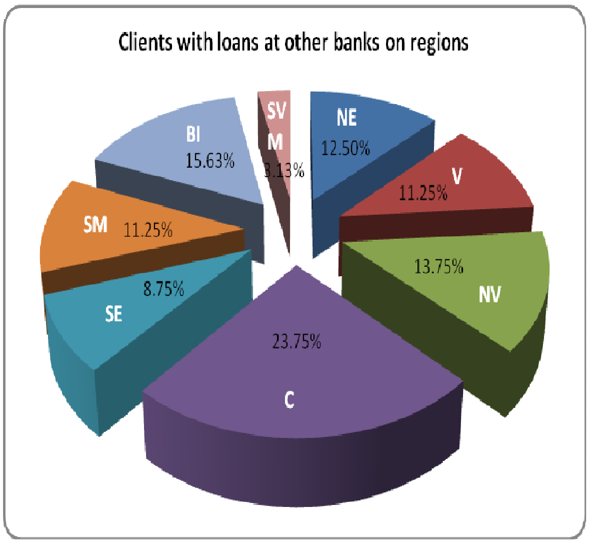

Figure 2 – Clients with loans at other banks on regions – January 2011
Source: information based upon data from the bank
For the second time the Region Middle and Bucharest-Ilfov have the upper hand when speaking about customers that already have a loan contracted from another

financial institution, another thing that underlines a more dynamic attitude of this Region's inhabitants regarding credits.

Out of the bank total clients more than 61% have medium studies, fact that can be linked with the need to increase the revenue for current needs, but not with the knowledge about the banks products. The Regions Middle and Bcharest-Ilfov have more clients with upper studies, a fact that can represent an advantage in analyze of education level that lead to loan decision:

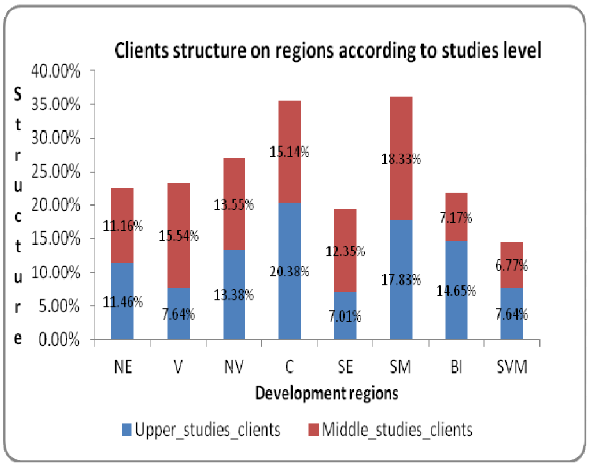

Figure 3 – Clients structure on regions according to studies level – January 2011

Source: information based upon data from the bank

The high percentage of clients with loans granted by other banks, living in regions like North-East, North-West, South-East and South-East Muntenia may point out an inhabitant's need of finding different sources of financing their daily spendings.

So, as a first conclusion we may split our selected regions in two categories:

✓ In North-East, West, North-East, South-East and South-West Muntenia customers may prefer to take a new loan, under the restriction of being able to comply with bank's requires regarding levels of incomes

✓ In the Middle and in Bucharest – Ilfov region, clients may be more selective in taking a decision linked to evolving in a new credit contract.

Further on in the study, over the clients' date base I applied principal components analysis typology.

The idea of using quantitative, numerical data in principal components and cluster analysis may be pointed out. Everitt and Dunn show that analyzing qualitative data through these techniques may reveal some differences in the correlation matrix and in the estimated factor loadings in comparison with the results obtained based on the methods proper for continuous variable; although the interpretation of factors in each case would be similar[8].

Having as support Statistical Analysis System (SAS) /www.sas.com/, in the first step of analysis, a series of statistical information about mean and standard deviation of variable were obtained. An average age of 36 years old of personal need loan owners and an average value of income, as it was recorded in the client's credit request, of 1,151 lei, lower than the nominal average salary for January, of 1,424 lei[9] were observed. So, it is shown that, in most cases, the bank's client, loan owner, is a person with incomes lower than the economy average, fact that consists in a first aspect about consumer profile.

In the next output given by the program, through the uncorrected with the mean correlation matrix, some high correlation coefficients could be pointed out: between Education and Age (0.9521) or between Age and Loans granted by other banks (0.9484) or between Education and Loans granted by other banks (0.9209). Not analyzing all the variables, consequence due to the correlation between them, may imply some false assumptions accordingly to factor interpretation.
Using the percentage of total variation criteria as way of factors extraction, the study shown that the first two principal components sum up to 86% of information contained by the identified variables, while the first three components contain up to 93% from the initial value. Taking into account the purpose of the principle components analyze, we may say that the initial nine variables could be represented as two with a loss of information of 14%. Having so the principal components determined one may calculate the nine eigenvectors of correlation matrix.

According the mathematical model mentioned above, the principal components can be written as a linear combination of original variables:

Prin1 = 0.3657*Sex + 0.3811*Age + 0.2534*Chl_no + 0.2982*Income + 0.3653*Education + 0.3686*Loan_oth + 0.2132*Balance + 0.3550*Region + 0.3552*Marital_status
Prin2 = -0.1635*Sex – 0.1305*Age + 0.1609*Chl_no + 0.5244*Income – 0.2033*Education – 0.1759*Loan_oth + 0.7337*Balance - 0.1237*Region - 0.1711*Marital_status

One may notice that the second principal component is linked with loan balance at the time in question, fact that may point out that this can be interpreted as reflecting actual debt of clients. Also, we may define the first variable as an indicator of client's social status.

[8] Everitt Brian S., Dunn Graham, "*Applied Multivariate Data Analysis*", second edition, John Wiley & Sons ltd, 2001, page 286
[9] Data supplied by the National Institute of statistical Analyze

Reducing the frame dimensions allows the subjects representation (bank customers) based upon two axles. Therefore the institution may use the new variables in order to analyze the customer behavior regarding the availability to contract a credit card.

Once identified a lower number of synthetically factors that represent, on a base of accepted loss of information, the characteristics of the initial dimension, we may focus upon grouping clients through cluster analysis. This action is interpreted towards possible group selection in order to gain smaller groups that would allow a certain marketing strategy.

Using the results from the previous analysis, on each development region, I applied cluster analysis, using Ward method and SAS Software. Having as variables the number of children1[10] and loan balance, following cluster grouping, we may reduce the number of clients to be taken into account for a selling campaign of a credit card to 39 in the South-East Region from 42, for example. Using same way of thinking for all other regions we may reduce the database as follows: for North-East to 40 clients, West to 47, for North-West to 38, Middle 55, South-Muntenia 70 and for South-Oltenia 24. We may observe that for Bucharest-Ilfov due to high homogeneity such a process is not allowed.
Therefore we may use a marketing strategy that will have as target 314 clients, with 23% less than initial, identified through cluster analysis with respect to two variables.

V. Summary and conclusions

The use of information about customers, only when a certain product is requested, without being included in a centralized system, system able to offer the opportunity for each division to access and capitalize the data, makes difficult to develop an integrated image of clients and of their profile. A proper knowledge of client's needs and expectations is the base for a successful communication.

Principal components analysis, through data summary under a minimum lost of information; helps reduce data dimensions, base for simplifying further analysis. Summarizing the information obtained through the study developed, we can observe that the first two principal components assure a total of about 80% of the initial information. Also, starting from retaining the first two principal components in managing the client's data base and analyzing their coefficients, we may select as first variable number of children and loan balance in order to create the typology of a potential client interested in a credit card.

[10] Different solutions shown that variable Chl_does not provide better interpretation compare to the variable Age, both reflecting client's social status.

Even if the higher homogeneous character of the database may be considered as a flow of the developed study, the use of the cluster analysis determined the diminution of potential clients, based on the output of the principal component analysis. Also, based on the dendrograms of each region, the institution may develop the analysis to the grouping and selection of clients the best suitable to cost and time restrictions imposed to the company.

In conclusion, the bank may reduce the targeted client number with up to 23% accordingly the study results.

Also, a potential client interested in buying a credit card lives in North-East, West, North-West, South-East or South-West Muntenia Region, is around 36 years old, with an income lower than the economy average wage and has already other loans.

REFERENCES

[1] **Awasthi S**. (?????? anul) ***Principal Components and Factor Analysis*** http://www.uta.edu/faculty/sawasthi/Statistics/stfacan.htm

[2] **Banfield J. D., Raftery A. E. (1993),** ***Model-based Gaussian and Non-Gaussian clustering***, Biometrics, vol. 49, page 803 – 821;

[3] **Casin Ph. (2001),** ***A Generalization of Principal Component Analysis to K Sets of Variable*** .Computational Statistics & Data Analysis, vol 35, page 417- 428;

[5] **Everitt Brian S., Dunn Graham (2001),** ***Applied Multivariate Data Analysis***, second edition, John Wiley & Sons ltd.;

[6] **Ferré Loius (1995),** ***Selection of Components in Principal Component Analysis: A comparison of Methods*** ;Computational Statistics & Data Analysis, vol. 19, page 669 – 682;

[7] **Hair Joseph F. Jr., Black William C., Babin Barry J., Anderson Rolph E. (2009),** ***Multivariate Data Analysis***, seventh edition, Ed .Pearson College Div.;

[8] **Hotelling H. (1933),** ***Analysis of a Complex of Statistical Variables into Principal Components*** ;Journal of Educational Psychology, vol 24, page 417 – 441;

[9] **Jolliffe I.T. (2002),** ***Principal Component Analysis***, second edition, Springer;

[10] **Kaiser H.F. (1958),** ***The Varimax Criterion for Analytic Rotation in Factor Analysis***, Psychometrika, vol. 23, page 187 – 200;

[11**] Linoff Gordon S., Berry Michael J. A.** (**2011**), ***Data Mining Techniques for Marketing, Sales and Customer Relationship Management***, third edition, Wiley Publishing;

[12] **Lu Wei-Zhen, He Hong-Di, Dong Li-Yun (2011),** ***Performance Assessment of air Quality Monitoring Networks Using Principal Components Analysis and Cluster Analysis***, Building and Environment, vol 46, page 577 – 583;

[13] **Marriott F.H.C. (1982),** ***Optimization Methods of Cluster Analysis***, Biometrika, vol. 69, page 417 – 421;

[14] **Păun C., Hunya G., Munteanu C. (2010),** ***Factorial Analysis on Risk Management Practices. The Case of the Romanien Companies***. Economic Computation and Economic Cybernetics Studies and Research, vol. 45 pg 221-236;
[15] **Pearson K. (1901),** ***On Lines and Planes of Closest Fit to Systems of Points in Space***, Philosophical Magazine, vol. 2, page 559 – 572;
[16] **Rencher Alvin C. (2002),** ***Methods of Multivariate Analysis***, second edition, John Wiley & Sons Ltd.;
[17] **Romesburg H. Charles (2004),** ***Cluster Analysis for Researchers***, Ed. Lulu Press North California;
[18] **Ruxanda Gh. (2009**), ***Analiza multidimensionala a datelo****r*, course paper PhD School Economical Studies The Bucharest Academy of Economic Studies;
[19**] Scott A. J., Symons M. J. (1971),** ***Clustering Methods Based on Likelihood Ratio Criteria***, Biometrics, vol. 27, page 387 – 398;
[20] **Simar Léopold, Härdle Wolfgang (2003),** ***Applied Multivariate Statistical Analysis***, Springer, 2003;
[21] **Shlens J. (2009),** ***A Tutorial on Principal Components Analysis***, http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf ;
[22] **Spircu Liliana, Calciu Mihai, Spircu Tuberiu (1994),** ***Analiza datelor de marketing***, All Publishing House;
[23] **Thomaz Carlos Eduardo, Giraldi Gilson Antonio (2010),** ***A New Ranking Method for Principal Components Analysis and its Aplication to Face Image Analysis***, Image and Vision Computing, vol 28, page 902 – 913;
[24]***www.insse.ro .