A VIEW OF ICT IN EUROPE

Ionela-Cătălina ZAMFIR [1]
Ana-Maria Mihaela IORDACHE [2*]

ABSTRACT

ICT (Information and Communication Technology) represent the one of the most important features of a developed country. Through the characteristics that can describe the development level for a country are: the access to technology, the use of internet for shopping, the individual's skills regarding ICT and the reasons for not using technology. This research take into account the most 22 relevant indicators for 2017 regarding ICT and use data mining techniques like: principal components analysis and Gaussian mixture models for classification along with statistics techniques for identifying the normal distribution and the number of optimal clusters, in order to group European countries in several classes of ICT development. The results show that Romania is still behind the rest of European countries in embracing the modern technology and its utilities.

KEYWORDS: *clustering, ICT in Europe, GMM, PCA*

JEL CLASSIFICATION: C38, L86

1. INTRODUCTION AND LITERATURE REVIEW

The ICT (Information and Communication Technology) incidence for a country is one of the most relevant indicators for a general development level of that country. The higher the incidence of modern technology is the higher is the level of development and the quality of life. Internet and computers are made for ease the life in general and taking decisions in particular.

In 2015, Savulescu obtained similar results as in this research regarding the development of north countries for ICT usage and the low development level for south-eastern countries (for example Romania, Bulgaria, Greece). The author underlines "the efforts of the European Union in view to reduce the digital divide in Europe and create a genuine internal digital market" (Savulescu, 2015). On the other side, Kleibrink et.all (2015) analyzed 97 regions from 9 Western Europe countries, 29 of the regions "having a dedicated ICT strategy". The authors findings suggest that "having a dedicated ICT strategy has not had a clear effect on performance in terms of Internet and broadband

[1] Assistant teacher, Phd, The Bucharest University of Economic Studies, Bucharest
[2*] corresponding author, Lecturer, Phd, School of Computer Science for Business Management, Romanian-American University, Bucharest iordache_ana_mari_mihaela@yahoo.com

access" (Kleibrink et.all, 2015). In 2018, Becker et.all used k-means to study Central European countries from ICT usage point of view. Their findings were that "Slovenia and Austria were leaders in Central Europe in 2017, and the worst-performing country was Poland, preceded by Hungary" (Becker, et.all, 2018).

The second section presents the methodologies used for having an overview of European countries from ICT point of view, the third section shows the dataset used and results, while the last section is about conclusions, discussions and further research.

2. THE METHODOLOGIES

The principal components analysis (PCA) is one of the most used techniques for reducing the dimension of a dataset. The big amount of data is stored in many variables that are essential for describing and modeling an issue, like the development of ICT in Europe. Also, the correlations between variables that represent the same fact have a big influence for the outcome of a model. PCA reduces the number of variables by creating new variables using a maximization problem. These new variables (PC) are not correlated, so the redundant information is reduced significantly, these take about 80%-90% of information from initial variables and their number is significantly reduced compared with the number of variables.

The new variables (PC) were tested for normality. Shapiro-Wilk test of normality have the null hypothesis that the selected sample come from a population that have a normal distribution. A value for p-value lower than 0.05 (the chosen alpha value), the null hypothesis is rejected and the distribution is considered not-normal. Another way to test from the visual point of view, that a variable have normal distribution is to generate a normal distributed sample with mean and standard deviation equal with the tested sample and represent graphically the probability density function.

Unless the number of classes is not imposed by the analyzed problem, there are several methods to decide what is the optimal and the right number of the classes. That assures the high variability between classes and low variability within each class:

- The elbow method suppose calculating and plotting the within total sum of squares for a number of clusters from 1 to 10 (for example), using the k-means algorithm. The optimal number of clusters is represented in the graph by an "elbow".

- The average silhouette method is similar to elbow method, only that the average silhouette is computing and plotted. The k number of clusters is selected to the highest value of the average silhouette width. This technique also "measures the quality of a clustering"[1].

- The function NbClust in R, that "provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods"[2].

[1] https://uc-r.github.io/kmeans_clustering#elbow
[2] https://www.r-bloggers.com/finding-optimal-number-of-clusters/

The GMM (Gaussian Mixture Models) technique for clustering is a probabilistic model that supposes that a population distribution is composed by K normally distributed subpopulations. For each subpopulation, a mean and a variance are computed. For multivariate case the covariance is calculated instead the variance. Each subpopulation has a weight in total population (sum of weights is one) that is learned with other parameters, using expectation maximization technique.

3. DATASET AND RESULTS

The dataset has 22 variables that mainly refer to population that use or have computer or internet. Eurostat is the source of data, and 2017 represent the year of interest. There are 35 countries, including Romania.

Table 1. The dataset description

Indicator code	Description: % of population who
X1	daily use a computer
X2	last use of computer <3 months
X3	use of computer >1 year
X4	never used a computer
X5	ever used a computer
X6	have access to a computer
X7	use mobile (smart) phone for internet
X8	have access to internet
X9	use internet via broadband
X10	use internet via fixed broadband
X11	bought online <3 months
X12	bought online <1 year
X13	bought goods or services online >1 year or never
X14	bought online shares, bonds or funds
X15	took a credit online
X16	don't buy online (<1 year) because prefer to go shopping
X17	don't have the skills to buy online (<1 year)
X18	don't have a payment card to shop online (<1 year)
X19	don't have internet because of costs
X20	don't have internet because of lack of knowledge
X21	are not interested (no need) of internet use
X22	don't have internet because of equipment costs

The table from above represents the variables considered. There are 22 quantitative variables that are measured in percent of individuals and describe the incidence of ITC in European countries from the point of view of equipment with technology, use of internet for different purpose or reasons for not using internet.

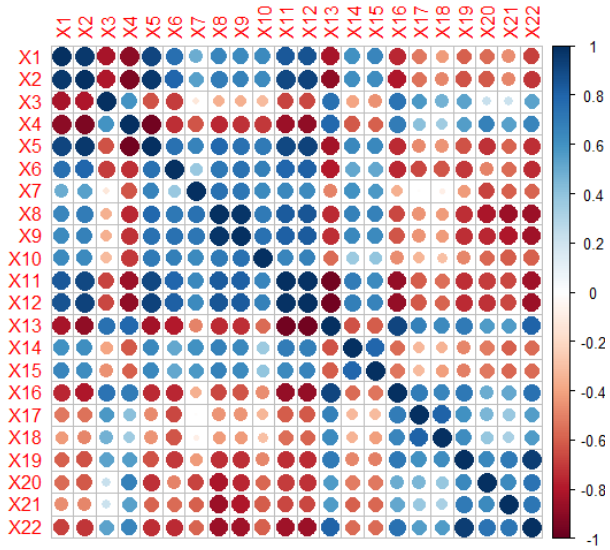

Figure 1. The correlation matrix of variables

Some strong connections between variables are presented in the correlation matrix from above. The bigger and intense the dot is, the stronger is the correlation between the variables, while the color represent the signification of connection. For example the color red is for negative correlation, while blue is for positive correlation. The strong connections between variables represent some redundant information contained by variables, parts of information that is common to two or more variables. This is the reason for applying variables reduction techniques, like factor analysis (FA) or principal components analysis (PCA).

	l	p	pc
Comp.1	14.73518	66.97811	66.97811
Comp.2	2.17290	9.87683	76.85494
Comp.3	1.40334	6.37881	83.23375
Comp.4	0.93068	4.23037	87.46413
Comp.5	0.54657	2.48439	89.94851
Comp.6	0.51179	2.32630	92.27481
Comp.7	0.38356	1.74346	94.01827
Comp.8	0.32200	1.46365	95.48192
Comp.9	0.21049	0.95676	96.43867
Comp.10	0.19472	0.88508	97.32375
Comp.11	0.17899	0.81359	98.13734
Comp.12	0.11277	0.51258	98.64992
Comp.13	0.09677	0.43987	99.08979
Comp.14	0.06368	0.28945	99.37925
Comp.15	0.05905	0.26840	99.64764
Comp.16	0.03093	0.14059	99.78824
Comp.17	0.02048	0.09309	99.88133
Comp.18	0.01096	0.04982	99.93115
Comp.19	0.00915	0.04159	99.97274
Comp.20	0.00404	0.01835	99.99109
Comp.21	0.00143	0.00651	99.99760
Comp.22	0.00053	0.00240	100.00000

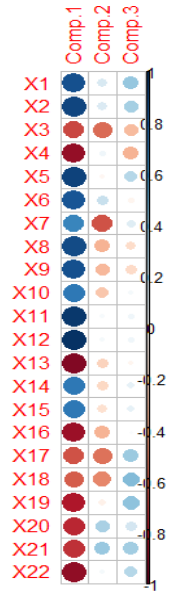

Figure 2. The Principal Components Analysis (PCA) results

Principal components analysis (PCA) results are presented in the figure 2. According to Kaiser criteria applied for standardized data 3 components should be considered in further analyses. These components have the variance higher than unit - 1 column from above. These synthesize 83% of total information, the pc column from above, from all 22 variables. The correlation matrix from the right side of the image from above is the factor matrix, which shows the connections between PC and initial variables. From this point of view, the first PC (W_1) is highly correlated with most of the variables and is positively correlated with variables that represent the use/access to computer/internet and negatively correlated with variables that show the "indifference" to technology and could be named the incidence and interest in ICT. The second PC (W_2) is highly negatively correlated with X_3 and X_7 and shows the use of computer more than a year ago and the use of mobile internet, while W_3 represent the lack of skills of individuals to buy online and the lack of a payment card for online shopping.

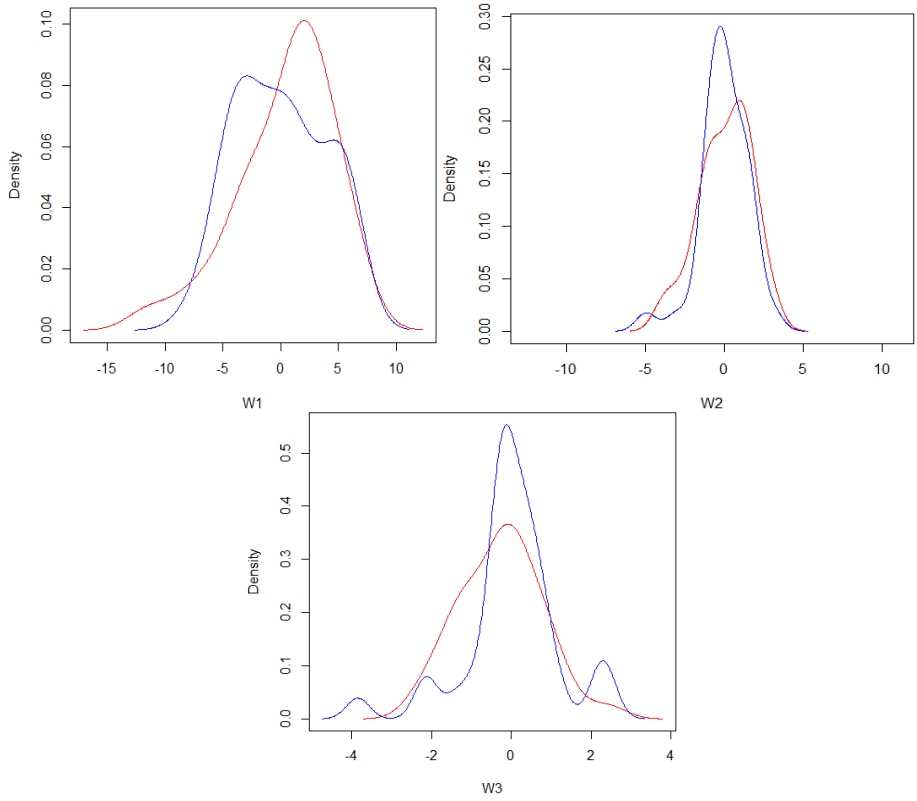

Figure 3. The PC's PDF

The figure from above show the comparison between normal distribution PDF and each of three PC retained for further analysis. For each component it was generated a normal distribution with the same mean and standard deviation (n=35) and PDF was plotted for both component (blue line) and generated data (red line) in order to compare the shape of PDF graph. First two PC have similar PDF graph with normal distribution while the third is not normal. Each component is a linear combination of initial variables, variables that are generally not normal.

```
> st1=shapiro.test(c[,1:1])
> st1

        Shapiro-Wilk normality test

data:  c[, 1:1]
W = 0.94142, p-value = 0.06184

> st2=shapiro.test(c[,2:2])
> st2

        Shapiro-Wilk normality test

data:  c[, 2:2]
W = 0.90666, p-value = 0.04431

> st3=shapiro.test(c[,3:3])
> st3

        Shapiro-Wilk normality test

data:  c[, 3:3]
W = 0.91413, p-value = 0.009697
```

Figure 4. The Shapiro-Wilk test for PCs

The Shapiro-Wilk test (figure 4) shows the results similar to PDF graphs from above. The first test corresponding to the first PC has a p-value higher than 0.05 and show the acceptance of null hypothesis that states that the variable is normally distributed. The second test's p-value is not far from 0.05, but show the rejection of null hypothesis, and the non-normality, while the last test's p-value is obviously lower than 0.05.

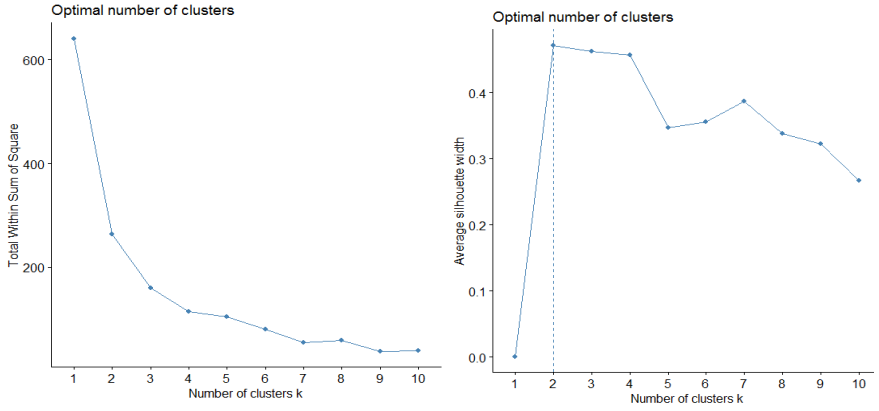

Figure 5. Elbow and Silhouette methods for number of clusters

For the optimal number of classes, the elbow method from above, represented in the first graph, show that three or four clusters should be taken in consideration. From a higher than 4 number of classes, the ratio between the between sum of squares and total sum of squares "tends to change slowly and remain less changing as compared to other k's"[1]. On the other side, the Silhouette method shows that for two classes, the average silhouette values is maxim.

[1] https://www.r-bloggers.com/finding-optimal-number-of-clusters/

```
*****************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 12 proposed 3 as the best number of clusters
* 2 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*****************************************************************
```

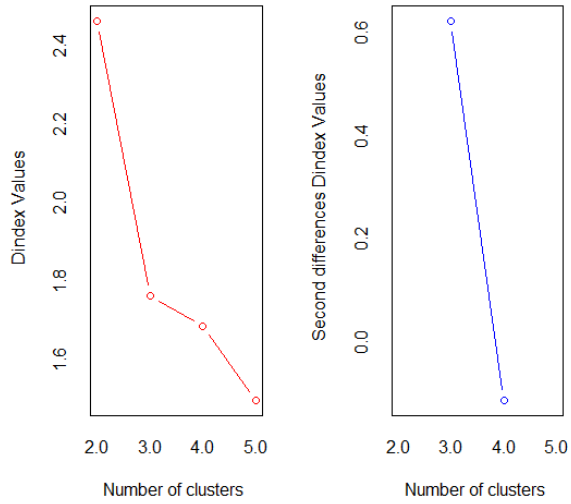

Figure 6. Criteria for optimal number of clusters

The results provided by NbClust package (figure 6) show that three is the optimal number of clusters. There were 12 proposals for 3 as the best number of clusters, so the conclusion according to the majority rule is 3. From this point of view, in clustering algorithm from below, 3 clusters will be considered.

```
$centroids
           [,1]       [,2]       [,3]
[1,]  5.2417644 -0.5838452  0.1996014
[2,]  0.7359372  0.6921286 -0.2897132
[3,] -3.7976108 -0.1834192  0.1042975

$covariance_matrices
          [,1]      [,2]      [,3]
[1,] 0.5067563 0.2061801 0.1618604
[2,] 1.2740371 0.8831907 0.2188617
[3,] 2.2933996 3.7734135 2.9930092

$weights
[1] 0.2562706 0.3266945 0.4170350
```

Figure 7. The GMM results

The figure from above show the centroids, the covariance matrices and the weights for each of the three classes selected. Taking into account the significance for each PC, given by the factor matrix and the fact that the centroids represent a virtual observation that could be described as the average value for components, each class has a different signification. The first class has the biggest average value for W_1 and W_3 and the lowest value for W_2. The countries that belong to this class have a high percent of individuals that have/use computer/internet, including for online purchasing. Also, in these countries, the individuals use the internet on mobile or smart phones, have the skills to buy online and also have an online payment card. The third class has countries where individuals are reserved in using technology and consider that the costs for equipment or for internet are too high. Also, the lack of skills prevents individuals to use internet and many individuals do not have a payment card for online transactions or do not have the skills to buy online. The class 2 represents the middle class and it contains countries that are developing from ICT point of view.

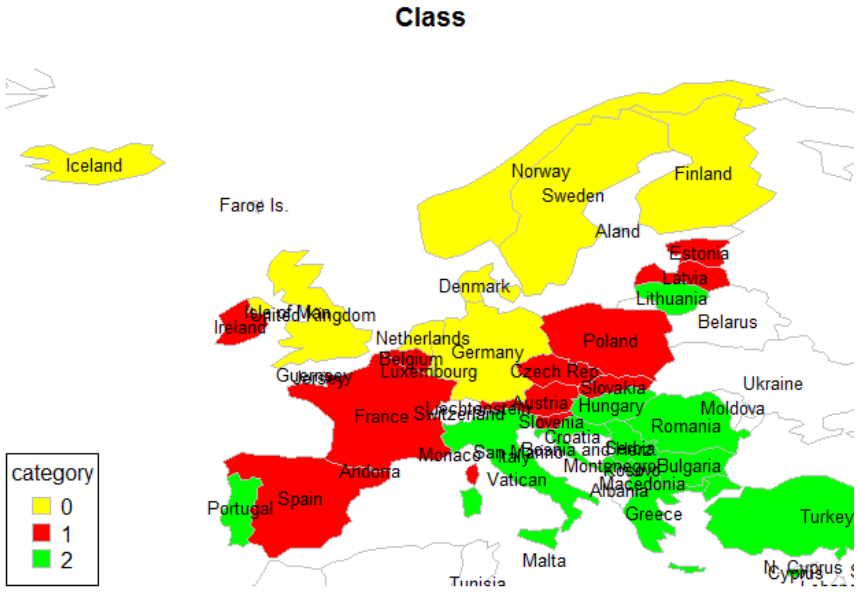

Figure 8. Europe's map with 3 classes

The most relevant representation for classes is the Europe's map from the above figure. Category 0 correspond to class 1 from above, 1 is for the middle class, and 2 is for the third class. The results show a stratification of European countries, so that countries like Romania, Bulgaria, Greece, Italy, Turkey or Portugal have the lowest development rate for ICT usage point of view. On the other side, the most developed countries are the countries from north Europe, like Norway, Sweden, Finland, Iceland, UK or Germany.

4. CONCLUSIONS AND DISCUSSIONS

The unsupervised learning technique used for clustering 35 European countries into three groups of ICT development show that Romania and other south countries are behind other countries regarding the modern technology usage, equipment, skills or even interest. Even

if the efforts for modern technology usage are continuously made, it could pass years and, why not, generations before the full acceptance of a new lifestyle that includes technology. Some similar studies or analyses that show the evolution in time of ICT usage could be made to show the progress of each country.

As further research, other techniques could be applied to predict the evolution of European countries regarding the ICT usage and development. Also, correlations between economic growth and ICT usage can be made, with the assumption that a big economic growth is based on high percentage of ICT usage in providing services.

REFERENCES

[1] Savulescu, C. (2015), Dynamics of ICT development in the EU, Procedia Economics and Finance, 23, pp. 513 – 520.

[2] Kleibrink, A., et.all. (2015), Regional ICT Innovation in the European Union: Prioritization and Performance (2008–2012), Journal of the Knowledge Economy, Vol. 6, Issue 2, pages 320-333

[3] Becker, J. et.all. (2018), ANP-based analysis of ICT usage in Central European enterprises, Procedia Computer Science, 126 (2018), pp. 2173–2183

[4] https://cran.r-project.org/web/packages/ClusterR/vignettes/the_clusterR_package.html

[5] https://brilliant.org/wiki/gaussian-mixture-model/

[6] https://www.r-bloggers.com/finding-optimal-number-of-clusters/

[7] https://uc-r.github.io/kmeans_clustering#elbow

[8] https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#elbow-method