Applied Mathematics, 2014, 5, 2408-2415
Published Online August 2014 in SciRes. http://www.scirp.org/journal/am
http://dx.doi.org/10.4236/am.2014.515232

Scientific Research

Cluster Analysis for Political Scientists

Dalson Britto Figueiredo Filho[1], Enivaldo Carvalho da Rocha[1], José Alexandre da Silva Júnior[2], Ranulfo Paranhos[2], Mariana Batista da Silva[1], Bárbara Sofia Félix Duarte[1]

[1]Department of Political Science, Federal University of Pernambuco, Recife, Brazil
[2]Institute of Social Science, Federal University of Alagoas, Maceió, Brazil
Email: dalsonbritto@yahoo.com.br, enivaldocrocha@gmail.com, mariana.bsilva@gmail.com, ranulfoparanhos@me.com, jasjunior2007@yahoo.com.br, barbarasfduarte@gmail.com

Received 5 June 2014; revised 8 July 2014; accepted 20 July 2014

Abstract

This paper provides an intuitive introduction to cluster analysis. Our targeting audience are both scholars and students in Political Science. Methodologically, we use basic simulation to illustrate the underlying logic of cluster analysis and we replicate data from Coppedge, Alvarez and Maldonado (2008) [1] to classify political regimes according to Dahl's (1971) [2] polyarchy dimensions: contestation and inclusiveness. With this paper, we hope to help novice scholars to understand and employ cluster analysis in Political Science empirical research.

Keywords

Cluster Analysis, Q Analysis, Political Regimes

1. Introduction

Classification of objects into meaningful groups is a central task in Science [3]. Cluster analysis is a statistical technique specialized to classify units into groups. Although cluster analysis is widely employed in other disciplines, its use in Political Science empirical research is limited when compared to linear regression [4], factor analysis and other multivariate statistical techniques [5].

The principal aim of this paper is to present an intuitive introduction to cluster analysis for political scientists. Our targeting audience are both undergraduate and graduate students in their initial training stage. Methodologically, we use basic simulation to illustrate the underlying logic of cluster analysis. In addition, we replicate data from Coppedge, Alvarez and Maldonado (2008) [1] to classify political regimes according Dahl's polyarchy dimensions: contestation and inclusiveness. On substantive grounds, we hope to facilitate the understanding and application of cluster analysis technique in Political Science.

The remainder of the paper is divided as follows. The next section briefly reviews the literature on cluster analysis. The third section presents the steps that should be followed to properly apply cluster analysis. The fourth section provides an example of research design using cluster analysis and presents the main statistics of interest. Finally, we present the conclusions of the article.

2. What Is Cluster Analysis?

During a long time cluster analysis was restricted to a limited group of researchers due to its mathematical complexity [6]. Technically, computational development facilitated the dissemination cluster analysis among different areas. Today, statistical packages can quickly perform mathematical distances calculations and therefore facilitate the use of cluster analysis by non-specialists.

But what is cluster analysis after all? A cluster can be defined as group of homogenous observations [7]. According to Aldenderfer and Blashfield, "cluster analysis is a generic designation for a large group of techniques that can be used to create a classification. Such procedures results in empirically clusters or groups of strongly similar objects" [4]. The main purpose of the technic is to group cases according to their degree of similarity. For Hair *et al*. (2009), "the cluster analysis gathers individuals or objects into clusters such that objects in the same cluster are more alike to each other than to other clusters" [8]. *i.e.* Observations within a specific cluster are more homogeneous than observations between clusters [9].

The underlying logic of cluster analysis is similar to factor analysis. The basic difference is that in factor analysis the researcher is concerned with representing a set of observed variables in a reduced number of factors, while in cluster analysis she seeks to represent a set of cases from a smaller number of groups (clusters). Factor analysis is concerned with variables while cluster analysis classifies cases. Considering another grouping technic, discriminant analysis, cluster analysis is different because in cluster analysis there is no prior knowledge about which elements belong to which clusters. The clusters are empirically defined using available data [5]. Cases are grouped according to the degree of mutual proximity, what the literature calls the distance/similarity. There are different ways of estimating how far/close observations are. In general, it is sought to ensure maximum homogeneity within the cluster, while it maximizes heterogeneity between groups. **Figure 1** illustrates an ideal type of cluster analysis[1].

Cases are grouped according to the degree of mutual proximity, what the literature calls the distance/similarity. There are different ways of estimating how far/close observations are. In general, it is sought to ensure maximum homogeneity within the cluster, while it maximizes heterogeneity between groups.

Our simulated data can be clustered in three different groups: A, B and C. Left figure illustrates the distribution of two simulated variables (X_1 and X_2). The Pearson correlation between them is r = 0.980 (p-value < 0.001;

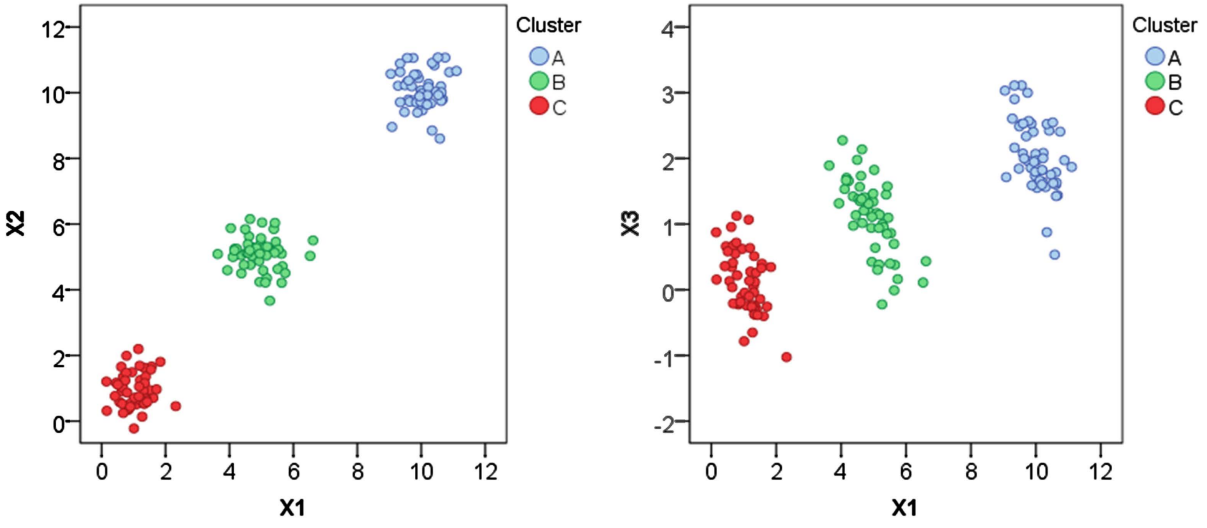

Figure 1. Cluster analysis examples.

[1]See the Appendix for simulation sintax.

n =150) considering all cases together as a unique group. When we compare the clusters, we observe that the correlation is not statistically significant for any group: (A; r = 0.019; p-value = 0.897; n = 50), (B; r = −0.096; p-value = 0.509; n = 50) and (C; r = 0.052; p-value = 0.719; n = 50). Similarly, right figure shows the relationship between X_1 and X_3. Taking all the observations at once, the Pearson correlation between them is 0.776 (p-value < 0.001; n = 150). However, disaggregating by cluster, the correlation between X_1 and X_3 is negative for each group (A; r = −0.528; p-value < 0.001; n = 50), (B; r = −0.701; p-value < 0.001; n = 50) and (C; r = −0.501; p-value < 0.001; n = 50). Substantively, these simulations show that ignore the clustered nature of data can lead to wrong conclusions.

3. Planning a Cluster Analysis

This section summarizes the requirements that must be met to proper use cluster analysis, focusing on the properties common to most methods. Political Scientists should follow six steps:

1) Data selection and treatment
2) Variables selection
3) Similarity measure
4) Cluster method determination
5) Number of clusters definition
6) Results validation

Different from other statistical techniques, sample size in cluster analysis is not related to statistical inference since the aim is not to estimate to what extent the results found in the sample can be extended to the population [10]. In fact, sample size should ensure the representation of small groups. Moreover, unlike other multivariate techniques, there is no general rule to specify a minimum sample size [11]. Our recommendation is that the greater the number of variables, more cases should be collected due the statistical instability of estimates from small samples. As long cluster analysis is sensitive to outliers it is also important to check for atypical observations. Hair *et al*. (2009) suggest profile diagram graphical inspection. The researcher can also use the blox-plot and scatter plots to identify outliers, in addition to the standard tests available in different statistical packages. Depending on variables measurement level, the use of raw data can complicate the interpretation of the concept of homogeneity. In this situation, it is recommended to standardize the variables before applying cluster analysis [12].

The second step is to select which variables will be used to estimate the distance/similarity between cases. The choice of variables is one of the most important steps, but unfortunately one of the least understood [6]. This stage requires a theoretical stated research problem [8]. Hair *et al.* (2009) state that it should be included only variables that characterize the objects to be grouped and specifically related to the goals of cluster analysis. Ideally, the research design should include only theoretically relevant variables to classify cases. The authors warn that, otherwise, there is a serious risk of naive empiricism, producing results conceptually empty and that do not contribute knowledge accumulation.

After selecting the variables, the next step is to define the similarity measure. Substantively, the question is how to check if individual A is more similar to individual B than to individual C. Pohlmann (2007) [13] argues that the similarity or dissimilarity between objects is a measure of correspondence or distance between the objects to be grouped. There are different ways to calculate this measure and different measures tend to produce distinct solutions.

Three methods are the most usual in cluster analysis: correlational measures, distance measures and association measures [8]. The correlational measure is based on the correlation/similarity between the profiles of the two objects. In this case the columns represent the variables and the rows represent the objects, instead of the correlation between two sets of variables commonly used. Despite its widespread use in other multivariate technics, correlational measures are not the most used in cluster analysis. The most commonly used is a distance measure, specially the Euclidean distance and its variants, such as the squared distance, and the city-block distance, that uses the sum of the absolute differences of the variables. Association measures are used with non-quantitative variables such as ordinal and nominal. The result is a matching coefficient based on the presence or absence of determined attribute [5].

Having selected the distance measure, we must choose the clustering method or algorithm. That is, the researcher must define how clusters will be created and how many. There are three general approaches to

creating clusters: 1) hierarchical clustering; 2) nonhierarchical clustering and 3) two steps or combined clustering.

Hierarchical techniques can be divided in agglomerative, starts with one cluster and progressively agglomerates until only one cluster is formed, and divisive methods, that stars with only one cluster until each observation is a cluster [8] [10]. The differential aspect of hierarchical technics is that the final result is linked to earlier stages culminating in the first stage, similar to a tree. The hierarchical clustering approach can be used when there is no prior knowledge of how many clusters should be formed. However, the researcher is still responsible for selecting the final solution to represent the data structure, the "stopping-rule" [10]. The clustering algorithm defines how similarity is measured when the clusters have more than a single member.

The single linkage method or the nearest neighbor is based on the shortest distance between individuals in two different clusters. This method has broad use, however, its flexibility generates poorly defined clusters. The complete linkage or the farthest neighbor method is based on the opposite premise, meaning that considers the maximum distance, solving the indetermination problems of the single linkage. The average linkage (between groups) is based on the average similarities between all members of one cluster with all members of any other cluster. The benefit of the average linkage is that it reduces the effect of extreme values (close or distant). The centroid method is based on the distance between the centroids. The centroid is the mean value of the observation on the variables in the cluster [10]. The distance between clusters is defined as the distance between the centers [8]. Like the average linkage, this method also reduces the importance of extreme values. The Ward's method differs from other methods because is based on an analysis of variance approach. The total sum of squared deviations from the mean of a cluster is calculated and the criterion for joining a cluster is to produce the smallest increase in the error sum of squares [10].

Nonhierarchical clustering has as start point the specification of the number of clusters. Once defined the number, the objects are assigned into clusters. It is a two-stage process. First, the cluster seed is specified. This is a start point that can be defined by the researcher or at a systematic or random selection. Then, observations are assigned according to its similarities to the pre-defined seed. The nonhierarchical group of algorithms is the K-means. The K-means works by separating the data into a pre-specified number and systematically assigning observations to the clusters. The K-means method is more suitable for large samples ($n > 1000$) since it does not compute the proximity matrix between all cases.

The two-step or combined clustering tries to accommodate hierarchical and non-hierarchical technics. First, a hierarchical method is used to create a complete set of cluster solutions and establish the appropriate number of clusters. Then the outliers are eliminated and the remaining observations are assigned to cluster by a non-hierarchical method [10].

The choice of the number of clusters is fundamental in cluster analysis and should be theoretically guided. For example, if previous studies suggested the existence of three groups, an analytical possibility is to replicate the number of groups in order to evaluate solution stability. In absence of theoretical guidance, the researcher can adopt an exploratory approach and run different analysis varying the number of groups. Different solutions should be compared with the literature searching for substantive explanation.

Finally, results should be validated in order to ensure practical significance [10]. To do so, the researcher can divide the original sample and compare the solutions obtained in both cases. Another way is to test the predictive capacity of the solution comparing it with a random variable that has not been used in the initial solution. For example, when clustering groups according to smoking habits, it is expected that, on average, the physical strength of non-smokers is higher than smokers. Thus, after separating the groups, the researcher can conduct a battery of physical tests and see if the group of nonsmokers in fact presents a superior performance. Or, to classify political regimes according to their level of democratization, the researcher can estimate the extent to which income inequality varies among different groups of countries, assuming that democracies tend to promote more equal income distribution than non-democracies.

4. Example of Research Design in Political Science: Classifying Political Regimes

Following Dahl's (1971) original typology base on the contestation and inclusiveness dimensions, we observe the ideal types of political regimes: polyarchies (high scores in both dimensions), competitive oligarchies (high contestation, low inclusiveness), inclusive hegemonies (low contestation, high inclusiveness) and closed hegemonies (low scores in both dimensions). **Figure 2** illustrates these dimensions.

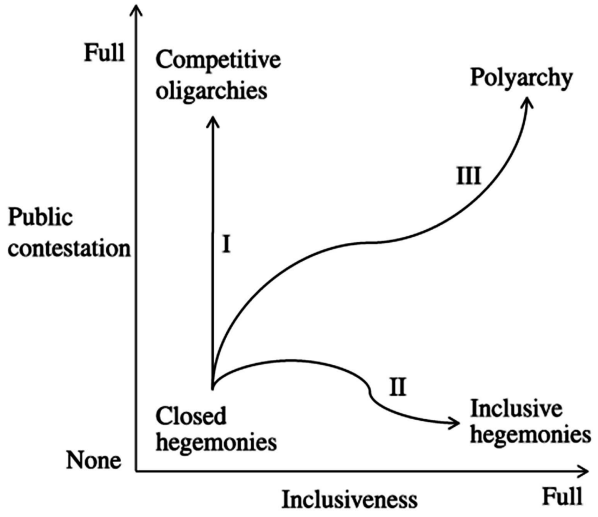

Figure 2. Dah'ls polyarchy dimensions.

The goal here is to group real political regimes according to these ideal types. This procedure improves the understanding of the usage of cluster analysis to construct typologies and its relevance to translate concepts into observable categories. Methodologically, we replicate data from Coppedge, Alvarez and Maldonado (2008) [1] to classify political regimes according to the two dimensions of polyarchy proposed by Dahl (1971): contestation and inclusiveness. The data can be found in Coppedge's website[2].

First, we analyzed data for 192 countries in the year 2000. The variable selection was theoretically driven, according to Dahl's dimensions of contestation and inclusiveness. So two variables were used to group the countries. Contestation and inclusiveness are indexes constructed by Coppedge, Alvarez and Maldonado (2008). The variables' values were standardized to be comparable across countries.

Before deciding the similarity measure and cluster/agglomeration methods we should graphically analyze observed countries distribution regarding Dahl's two dimensions. **Figure 3** displays this information.

The scores of both inclusiveness and contestation are standardized (mean = 0; std = 1) and we use the average as benchmark comparison. Denmark (DNMK) represents an example of a polyarchy (upper right). Vietnam (VNM) represents the political regimes that Dahl (1971) called inclusive hegemonies (lower left). Afghanistan (AFGN) can be classified as closed hegemony (lower right). Finally, countries in the upper left are named as competitive oligarchies, Ghana (GHNA) is an example of this institutional design. To use the average of the distribution to differentiate groups is a raw procedure to construct typologies. The mean approach is highly arbitrary, frequently classifying regimes extremely close in different groups just because they have scores a little higher or a little lower than the mean. Cluster analysis is a more systematic approach since it is based on the proximity between observations.

To classify countries into clusters, we chose Euclidian distance as similarity measure (the most commonly used) and run three different agglomeration methods—two-step, Nonhierarchical K-means and Hierarchical average linkage method. The next step is to choose the number of clusters. In this specific case the choice was theoretically driven and four clusters were selected based on Dahl's typology. **Figure 4** shows the observed results in comparative perspective.

For the two-step solution, we used the Euclidean distance and for the others the square of the Euclidean distance. For the k-means, we defined a maximum of 10 iterations. In all three solutions, cluster analysis identified four clusters. However, we can see different classifications for the three different methods used. This means that depending on the method used, we achieve very different results. It's possible to identify clearly two of Dahl's four regimes—closed hegemonies (low contestation and low inclusiveness) and poliarchies (high contestation and high inclusiveness)—the other two are somewhere in between the two opposites.

In the two-step solution, cluster analysis identified two groups located in the upper right quadrant, one group-located at the left bottom that could be classified as closed hegemony and one group ant the right bottom, that could be classified as inclusive hegemony. In the k-means solution we highlight an intermediate group with av-

[2]See http://www3.nd.edu/~mcoppedg/crd/datacrd.htm

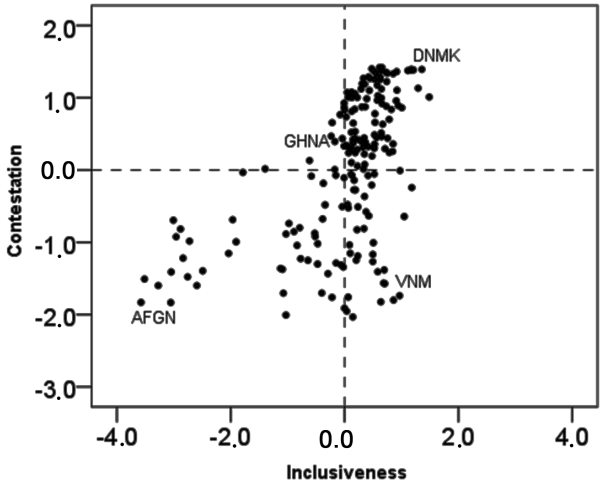

Figure 3. Observed countries according to polyarchy dimensions.

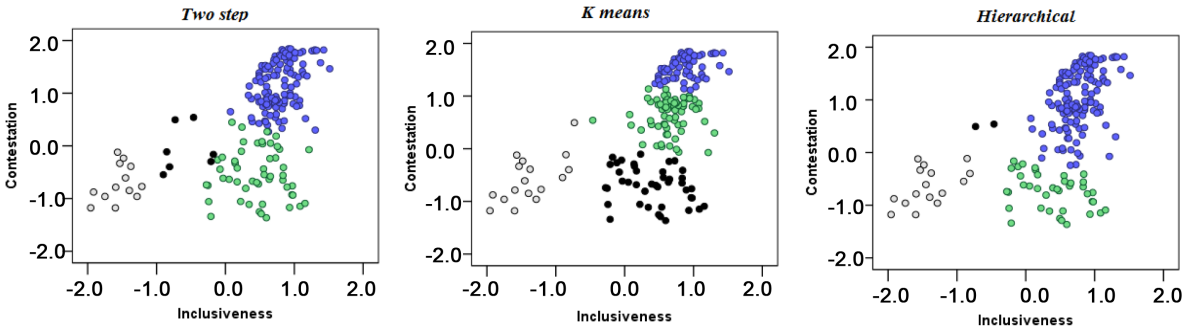

Figure 4. Agglomeration methods comparison.

erage contestation and average inclusiveness. Lastly, the hierarchical solution shows three defined groups and an unspecified political system that varies across two different ideal types (closed hegemonies x inclusive hegemonies).

Figure 5 compares different linkage methods in hierarchical clustering. The goal is to show the different results achieved when an alternative linkage method is applied.

Comparing the solutions we can see slightly different results, especially concerning the classification of the observations located around the mean value. The between groups and the furthest neighbor (maximum distance) solutions are relatively similar. The nearest neighbor (minimum distance) is somewhat different, presenting one large group that concentrates the majority of the observations. The centroid and the Ward's methods are more complex, focusing on the centroid and an analysis of variance respectively. The solution using the centroid is very similar to the average linkage. However, the Ward's method presents groups more clearly delimited. The comparisons show that that depending on the method used, the classification will be different. The choice rests on the researchers' ability to connect theoretical expectations and empirical classification.

After establishing the solution, it is also important to validate the results. This is the final step of a proper cluster analysis and can be done by partitioning the dataset and applying the solution to a sample to verify if the solution holds. Another way is to correlate the clusters with an exogenous variable to identify if the expected relationship is empirically observed.

5. Conclusion

Considering that political scientists usually work with typologies, we believe that cluster analysis is an important tool to classify units into groups. Its main advantage is to produce objective and replicable classification that can develop our knowledge regarding political phenomena. This paper provided an intuitive introduction to cluster

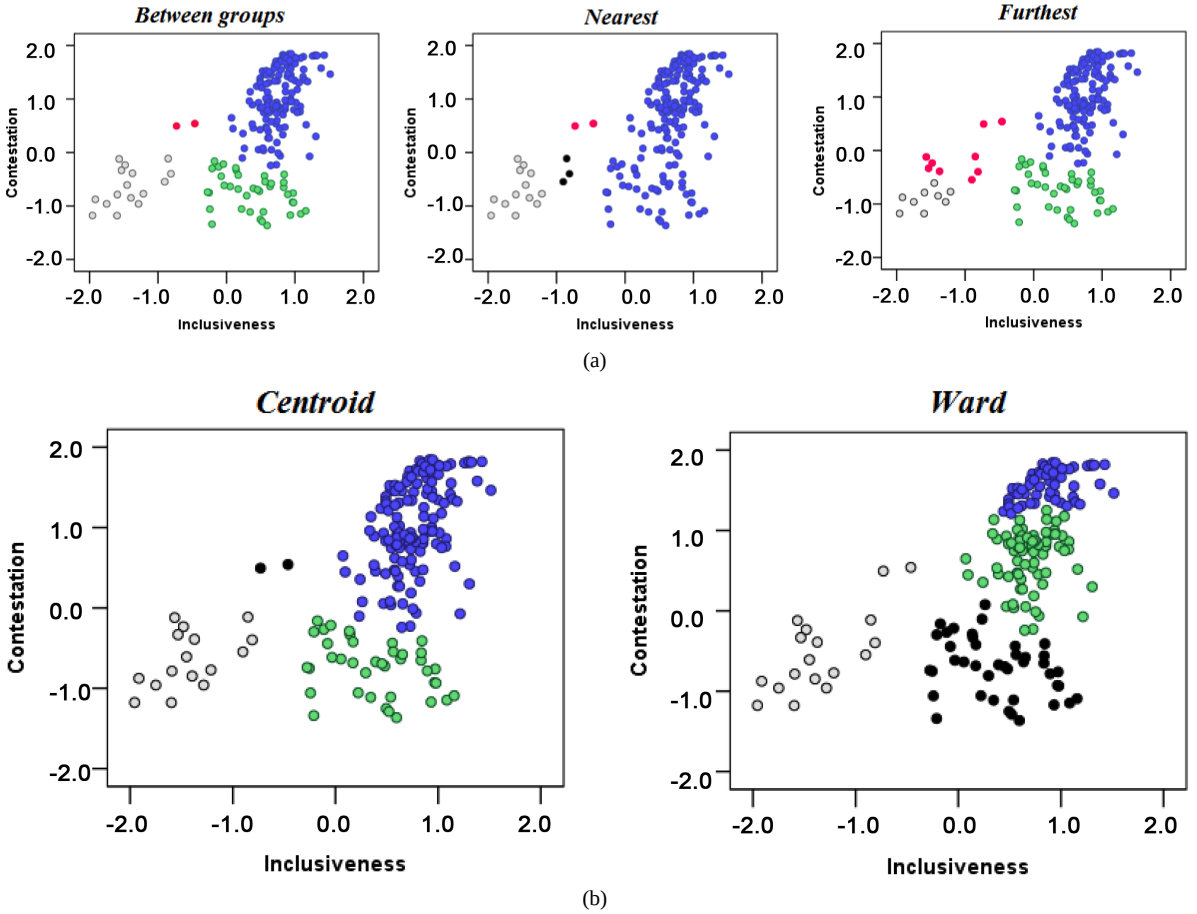

Figure 5. (a) Comparing linkages methods in hierarchical models; (b) Comparing linkages methods in hierarchical models.

analysis. Our targeting audience is both political science undergraduate and graduate students in their initial training stage. Methodologically, we used basic simulation to illustrate the underlying logic of cluster analysis. In addition, we replicated data from Coppedge, Alvarez and Maldonado (2008) to classify political regimes according to Dahl's (1971) polyarchy dimensions: contestation and inclusiveness. With this paper we hope to diffuse cluster analysis technique in Political Science and help novice scholars not only to understand but also to employ cluster analysis in their own research designs[3].

References

[1] Coppedge, M., Alvarez, A. and Maldonado, C. (2008) Two Persistent Dimensions of Democracy: Contestation and Inclusiveness. *Journal of Politics*, **70**, 632-647. http://dx.doi.org/10.1017/S0022381608080663

[2] Dahl, R. (1971) Poliarquia: Participação e Oposição. Edusp, São Paulo.

[3] Alquist, J.S. and Breunig, C. (2011) Model-Based Clustering and Typologies in the Social Sciences. *Political Analysis*, **20**, 92-112.

[4] Krueger, J. and Lewis-Beck, M. (2008) Is OLS Dead? *The Political Methodologist*, **15**, 2-4.

[5] Tabachnick, B. and Fidell, L. (2007) Using Multivariate Analysis. Allyn & Bacon, Needham Heights.

[6] Aldenderfer, M.S. and Blashfield, R.K. (1984) Cluster Analysis. Quantitative Applications in the Social Science, Sage University Paper Series.

[7] Burns, R. and Burns, R. (2008) Cluster Analysis. In: *Business Research Methods and Statistics Using SPSS*. Sage Publications.

[8] Hair, J., *et al.* (2009) Multivariate Data Analysis. 17th Edition, Prentice Hall, Upper Saddle River.

[9] Tan, P., Steinbach, M. and Kumar, V. (2005) Cluster Analysis: Basic Concepts and Algorithms. In: *Introduction to*

To get more information on cluster analysis see https://onlinecourses.science.psu.edu/stat505/node/138, http://www.statistics.com/clustering and https://www.statsoft.com/Textbook/Cluster-Analysis

Data Mining, Addison-Wesley, Boston.

[10] Bartlett, M.S. (1947) Multivariate Analysis. *Journal of the Royal Statistics Society*, **9**, 176-197.

[11] Dolnicar, S. (2002) A Review of Unquestioned Standards in Used Cluster Analysis for Data Driven Market Segmentation. Faculty of Commerce, Papers.

[12] de O. Bussab, W., Miazaki, S.E. and Andrade, D.F. (1990) Introdução à análise de agrupamento. In: *IX Simpósio Brasileiro DE Probabilidade E Estatísitica*, IME-USP, São Paulo, 105 p.

[13] Pohlmann, M.C. (2007) Análise de Conglomerados. In: Corrar, L.J., Edílson, P. and Dias Filho, J.M., Eds., *Análise Multivariada*, Atlas, São Paulo.

APPENDIX

Sintax for X_1 e X_2 per Cluster

```
IF (cluster = 1) ×2 = RV.NORMAL(10, 0.5).
EXECUTE.
IF (cluster = 1) ×1 = RV.NORMAL(10, 0.5).
EXECUTE.
IF (cluster = 2) ×1 = RV.NORMAL(5, 0.5).
EXECUTE.
IF (cluster = 2) ×2 = RV.NORMAL(5, 0.5).
EXECUTE.
IF (cluster = 3) ×2 = RV.NORMAL(1, 0.5).
EXECUTE.
IF (cluster = 3) ×1 = RV.NORMAL(1, 0.5).
EXECUTE.
```

Sintax for X_3 per Cluster

```
IF (cluster = 1) ×3 = ×1 * −0.6 + ×2 * SQRT(1 − 0.6 ** 2).
EXECUTE.
IF (cluster = 2) ×3 = ×1 * −0.6 +×2 * SQRT(1 − 0.6 ** 2).
EXECUTE.
IF (cluster = 3) ×3 = ×1 * −0.6 + ×2 * SQRT(1 − 0.6 ** 2).
```

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.