Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 158 (2019) 145–152

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd World Conference on Technology, Innovation and Entrepreneurship (WOCTINE)

Clustering analysis application on Industry 4.0-driven global indexes

Merve Doğruel Anuşlu[a], Seniye Ümit Fırat[b]

[a]*İstanbul Gedik University,Industrial Engineering, Istanbul and 34876, Turkey*
[b]*Marmara University, Industrial Engineering, Istanbul and 34722, Turkey*

Abstract

Industry 4.0 is one of the most important topics in the academia and business world in recent years as a result of digital milestones in innovation area. Industry 4.0 is considered a great revolution in both manufacturing and services sectors. One of the reasons why Industry 4.0 is described as a revolution is the search for new solutions that unappeared before, to the challenges associated with energy, resources, environment, social and economic impacts by using modern technologies to ensure sustainable prosperity. Another reason is the use of modern technologies such as digital chains, smart systems and industrial internet to accelerate innovation as a result of faster implementation of new business models. The other one, is Industry 4.0 lighten the load of current challenges such as shorter product lifecycles, higher product complexity, and global supply chains for manufacturers in order to make the companies more flexible and responsive to business trends. According to global index scores in several areas as economic, environmental, sustainability, innovation etc. that published by international organizations, the positions of countries relative to other countries can be analyzed. Countries can evaluate their current status according to their index scores and have the opportunity to develop strategies for the performance level that they targeted. The aim of this study is to group countries within the scope of significant impact areas of Industry 4.0 by using Global Innovation Index, Sustainable Development Goals Index, Logistics Performance Index and Environmental Performance Index. By using the global indices mentioned above, countries are grouped and evaluated by using Clustering Analysis from data mining methods.

Peer-review under responsibility of the scientific committee of the 3rd World Conference on Technology, Innovation and Entrepreneurship

Keywords: Clustering Analysis; Environmental Performance Index; Global Innovation Index; Industry 4.0; Sustainable Development Goals Index

1. Introduction

Industry 4.0 Revolution, taking the whole world under the influence; it leads to rapid, deep and widespread changes. The most important supporting elements of this revolution are creativity and innovation. The other important issue on the world agenda is global warming and climate change. These issues are discussed under the "sustainability" framework in industrial sectors. Sustainability, which has three basic components, economic, environmental and social, is monitored by global indices on the world basis and progresses in SDGs (17 Goals) are

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd World Conference on Technology, Innovation and Entrepreneurship
10.1016/j.procs.2019.09.037

measured.

In the literature, it is argued that Industry 4.0 and sustainability issues affect each other in the same direction (positive effect), but it is suggested that there may be adverse (negative) effects in time dimension in terms of some indicators [1].The emerging new technologies of the Fourth Industrial Revolution (4IR) will inevitably change and transform the entire world in many ways [2]. There are many results that are expected to benefit greatly from the actualization of this transformation. There are also many undesirable risks and threats. The extent to which the advantages of the Industry 4.0 Revolution, which cannot be stopped and will not be left behind, are maximized and risks reduced; is extremely important in terms of social, economic and environmental sustainability [3]. In this respect, the sustainability principles and objectives should be taken into account and examined together in the Industry 4.0 Revolution researchs [1].

The purpose of this study is to define the picture of the interaction between the supportive elements and technologies of the Industry 4.0 Revolution and sustainability principles and objectives. With a clearer expression, aim of the research is to group countries based on the performance scores in the fields of innovation, sustainability and SDGs by applying clustering analysis. By analyzing the indicators of these different dimensions together, it is to determine the levels of concentration in the country groups. Thus, a combined performance profile can be obtained for country groups that are found as a result of clustering analysis. In other words, the level of environmental, social and economic sustainability dimensions and the indicators supporting and developing the industry 4.0 revolution and the level of innovation, innovation inputs and innovation outcomes, etc., are tried to be determined.

For this purpose, the GII and its sub-indexes are identified as Industry 4.0 indicators. SDGI and EPI coverage are included for the sustainability performance evaluation. In addition, since the logistics sector is an area that provides infrastructure for all sectors, has the best practices of new technologies and is one of the emission sources with the largest share in CO2 emissions released to our planet, LPI has also been included in the study.

In the first stage of the applied study, SDGs is analyzed by Principle Compenent Analysis for dimension reduction. Then obtained factors and other determined indicators of the indices are used as input variables in cluster analysis in order to obtain country groups.

2. Indices

Four global indices related to Industry 4.0 were selected and examined. GII was chosen to assess the potential of countries to transition to Industry 4.0. SDGI was chosen to address the impact of Industry 4.0 on sustainability. LPI was chosen to assess the logistics sector, one of the sectors most affected by Industry 4.0. And EPI was chosen to assess the environmental policies that are at the focus of Industry 4.0 and sustainability.

2.1. Global Innovation Index (GII)

The Global Innovation Index (GII), published in collaboration with Cornell University, INSEAD and the World Intellectual Property Organization (WIPO) in 2018, was first published in 2007. The Global Innovation Index (GII) aims to provide tools that can help in adopting multi-dimensional innovation aspects and adapting policies to promote long-term output growth, improved productivity and business growth [4]. The theme of the 2018 edition of the Global Innovation Index (GII) is "Energizing the World with Innovation". The 2018 GII covers 126 countries. As seen Figure 1, in the framework of GII, four measures are calculated: the overall GII, innovation efficiency ratio, the input and output sub-indices. The input sub-index contains five dimensions, the output sub-index contains two dimensions, and in 2018 there are a total of 82 indicators [5]. Countries are scored between 0 and 100 according to the overall GII.

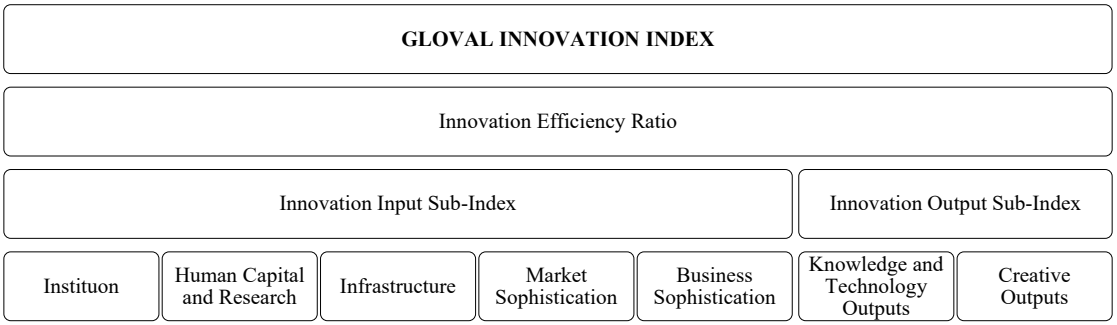

Figure 1: The 2018 GII Framework

2.2. Sustainable Development Goals Index (SDGI)

The SDG Index and Dashboard Report is the first worldwide study to assess their position to reach countries' Sustainable Development Goals (SDGs). The report has been prepared annually by Bertelsmann Stiftung and the Sustainable Development Solutions Network (SDSN) since 2016 [6]. The SDG Index and Dashboard Report identifies countries' current positions in terms of the sustainability target item and provides important clues about what issues should be prioritized in the SDGs targets expected to be realized by countries up to 2030 [6]. In order to group countries on a "traffic light" table (green, yellow, orange and red) during the evaluation and ranking phase, numerical threshold assignments are made to each indicator. The traffic light color scheme shows how far a country is from reaching a particular goal [7; 8].

The 2018 report includes many improvements and additions compared to previous versions. The most important of these is that the 2018 report includes trend data to evaluate the progress of countries in meeting the 2030 SDF history, which was not previously. Due to several changes in the indicators and some adjustments the methodology, it is not correct to compare and interpret the results of the 2018 SDG Index and Dashboards with the results of 2017. The 2018 SDG Index covers 156 countries. The SDG Index score consisting of 17 targets as shown in Table 1 results a country's position between the worst (0) and the best or target (100) results.

Table 1: The Sustainable Development Goals

No	SDG	No	SDG
SDG 1	No poverty	SDG 10	Reduced inequalities
SDG 2	Zero hunger	SDG 11	Sustainable cities and communities
SDG 3	Good health and well-being	SDG 12	Responsible consumption and production
SDG 4	Quality education	SDG 13	Climate action
SDG 5	Gender requality	SDG 14	Life below water
SDG 6	Clean water and santation	SDG 15	Life on land
SDG 7	Affordable and clean energy	SDG 16	Peace, justice and strong institutions
SDG 8	Decent work and economic growrh	SDG 17	Partnership for the goals
SDG 9	Industry, innovation and infrastructure		

Data in the 2018 Global Index and Dashboards are taken from official and non-official data sources. Most of the data is provided by International Organizations (World Bank, OECD, WHO, FAO, ILO, UNICEF, other). Other data sources are household surveys (Gallup World Poll), non-governmental organizations and networks (Oxfam, Tax Justice Network, other) and peer-reviewed journals [8].

2.3. Logistics Performance Index (LPI)

The sixth edition of the Logistics Performance Index (LPI) report, the first edition of which was published in 2007, was published by the World Bank in 2018. The Logistics Performance Index is an interactive benchmarking tool that allows countries to see the difficulties and opportunities they face in their commercial logistics and determine what they can do [9]

There are six indicators in the calculation of the Logistics Performance Index (LPI) [10]:

1. The efficiency of customs and border management clearance
2. The quality of trade- and transport-related infrastructure
3. The ease of arranging competitively priced international shipments
4. The competence and quality of logistics services
5. The ability to track and trace consignments
6. The frequency with which shipments reach consignees within the scheduled or expected delivery time

LPI data are survey data (between 1 and 5), which include evaluations of logistics experts worldwide for these six general dimensions. The 2018 LP Index covers 160 countries [7].

2.4. Environmental Performance Index (EPI)

World countries should now create data-driven environmental policies. The Environmental Performance Index (EPI) provides a global view on the environmental performance of countries. The EPI is a major contributor to countries in achieving the goals of the United Nations 2015 Sustainable Development Goals and the Paris Climate Agreement. The EPI ranks the performance of countries in the main categories of environmental health and ecosystem vitality [7]. The 2018 Environmental Performance Index (EPI) includes 180 countries on 24 performance indicators, as shown in Table 2, and consists of ten issues. Countries are scored on a scale of 0 to 100. EPI is produced jointly by the World Economic Forum, Yale University and Columbia University. In 2018, the McCall MacBain Foundation and Mark T. DeAngelis also contributed. 2018 EPI data sources are international organizations, research institutions, academia and government institutions [11].

Table 2: The EPI 2018

Policy Objective	**Issue Category**
Environmental Health (%40) (6 indicators)	Air Quality
	Water & Sanitation
	Heavy Metals
Ecosystem Vitality (%60) (18 indicators)	Biodiversity & Habitat
	Forests
	Fisheries
	Climate & Energy
	Air Pollution
	Water Resources
	Agriculture

3. Cluster Analysis

Cluster Analysis or clustering also known as segmentation, is the most widely used multivariate descriptive method of data analysis and data mining [12; 13].Cluster Analysis is used in many different areas such as psychology, biology, medicine, sociology, economics and marketing. The main purpose of the Clustering Analysis is to maximize in terms of specified features both the homogeneity within a cluster and the heterogeneity among the different clusters

[12]. Creation of clusters is performed using dissimilarity measures such as edit distance, density in a euclidean or non euclidean data space, distance calculated using Minkowski measures, proximity measures or probability distributions [14].

The methods used in clustering analysis are divided into two main groups as hierarchical and non-hierarchical as in Figure 2 [15].

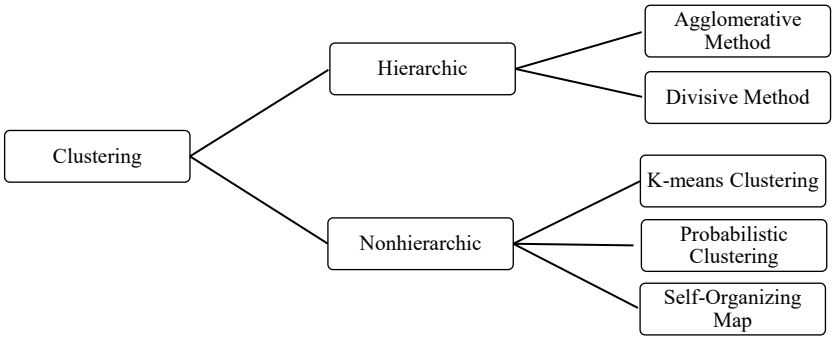

Figure 2: Types of clustering methods

Hierarchical clustering techniques are a series of sequential mergers or a series of splits processes. In the agglomerative method, the clusters are formed by merging smaller clusters into the larger ones bottom-up. In the divisive method, the clusters are formed by splitting the larger one into smaller cluster top-down [16]. The results of both agglomerative and divisive methods can be shown in a two-dimensional diagram, known as the dendrogram, which shows mergers or splits at consecutive levels [17].

Nonhierarchical clustering techniques find clusters that are appropriate for the number of sets that are predetermined by the user [15].

The K-means algorithm is the most preferred nonhierarchical clustering method. The steps of the traditional k-means algorithm can be summarized as follows [18; 19; 20].

Input: dataset $S = \{X'_1, X'_2, \dots, X'_N\}$ and desired number K of clusters

Output: K disjointed clusters

1- K units are randomly selected as initial cluster centers.
2- Cluster decentralized units are assigned to the closest initial cluster centers.
3- The cluster center is recalculated.
4- The new cluster centers are compared with the cluster center in the previous iteration. If there is no difference or too small, the algorithm is terminated; otherwise, steps 2 and 3 are repeated.

4. Application

The data set consists of the GII, SDGI, LPI and EPI indices of 116 countries which are common in the reports of 2018. In the data pre-processing stage, the missing data has been completed. For this, 12 missing observations in the SDG10 and 25 missing observations in the SDG14 were completed with the MissForest [21]which is nonparametric missing value imputation using Random Forest method in R programming.

The correlation coefficients between the indexes were calculated and found significant at 5% level. This also indicates that the indexes chosen are related to Industry 4.0.

Principal Component Analysis (PCA) was applied to investigate whether the 17 targets included in the Sustainable Development Goals Index can be represented by fewer variables. According to the results of the Kaiser-Meyer-Olkin (KMO) test, the sampling was found to be adequate (0,867>0,5). Bartlett test result was also found significant. According to these two metrics, PCA results were accepted as usable and the data are sufficient for analysis. All of the anti-image correlations were found significant in the anti-image matrix . When component matrices were examined, it was seen that the factor structure obtained from Varimax rotation application was better interpreted. In this structure, 6 factors were determined by scree plot was examined and Kaiser's criteria (eigenvalue >1) was applied. Total variance explained of 6 factors is %83,425. As a result of examining the coefficients in the rotated

component matrix, the variables in the extracted 6 factors and defined labels of factors are as in Table 3.

Table 3: 6 factors and included variables

Factor	Variables	Name
F1	SDG 12, SDG 16, SDG 9, SDG 8, SDG 13, SDG 2	Innovative industry and sustainable society
F2	SDG 1, SDG 7, SDG 3, SDG 4	Goals to raise the standard of living of individuals
F3	SDG 6, SDG5, SDG 11	Sustainable and modern urbanism
F4	SDG 15, SDG 10	Life on land
F5	SDG 14	Life below water
F6	SDG 17	Partnerships fort he goals

After PCA, the variables that were determined for cluster analysis are: 6 factors score representing 17 SDG, GI input sub-index, GI output sub-index, overall LPI, environmental health objective and ecosystem vitality objective for EPI. Thus, the dataset consists of 11 variables and 116 observations.

In the Cluster Analysis stage, first the dendrogram was obtained by applying the hierarchical methods and then K-means was tried as a nonhierarchical method.

Between-groups linkage and Ward's methods have been tried in hierarchical clustering analysis application. Dendrogram obtained by between-groups algorithm was not suitable for interpretation, but according to the dendrogram result obtained by the Ward method, the country groups was interpreted as 3, 4 or 5 clusters.

Applications were carried out for 3, 4 and 5 cluster with K-means method. As the results were close to each other, 3 clustered result was preferred with the highest conceptual meaning. Factor 5 and factor 6 were not significant (>0.05) in the ANOVA table as a result of 3 cluster analysis with K-means method. Therefore, the analysis was repeated by removing these variables. In the second ANOVA table, factor 4 was found to be insignificant so the analysis was repeated by removing this variable also. As a result, the 8 variables that generate 3 clusters are significant differences between country gruops and this clustering with 3 clusters has been chosen as a final model. As seen in Table 4, there are 28 countries in cluster 1, 59 countries in cluster 2 and 29 countries in cluster 3.

Table 4: Countries in each cluster

Country	CLS	Country	CLS	Country	CLS	Country	CLS
Australia	1	Algeria	2	Lebanon	2	Bangladesh	3
Austria	1	Argentina	2	Lithuania	2	Benin	3
Belgium	1	Armenia	2	Malaysia	2	Burkina_Faso	3
Canada	1	Bahrain	2	Mauritius	2	Cambodia	3
Cyprus	1	Belarus	2	Mexico	2	Cameroon	3
Denmark	1	Bolivia	2	Moldova	2	China	3
Estonia	1	Bosnia_and_Herzegovina	2	Mongolia	2	Côte_d'Ivoire	3
Finland	1	Brazil	2	Montenegro	2	Ghana	3
France	1	Bulgaria	2	Morocco	2	Guinea	3
Germany	1	Chile	2	Oman	2	India	3
Greece	1	Colombia	2	Panama	2	Indonesia	3
Iceland	1	Costa_Rica	2	Paraguay	2	Kenya	3
Ireland	1	Croatia	2	Peru	2	Madagascar	3
Israel	1	Czech_Republic	2	Philippines	2	Malawi	3
Italy	1	Dominican_Republic	2	Poland	2	Mali	3
Japan	1	Ecuador	2	Qatar	2	Nepal	3
Luxembourg	1	Egypt	2	Romania	2	Niger	3
Malta	1	El_Salvador	2	Russian_Federation	2	Nigeria	3
Netherlands	1	Georgia	2	Saudi_Arabia	2	Pakistan	3
New_Zealand	1	Guatemala	2	Serbia	2	Rwanda	3
Norway	1	Honduras	2	Slovak_Republic	2	Senegal	3
Portugal	1	Hungary	2	Slovenia	2	South_Africa	3
Singapore	1	Iran_Islamic_Rep.	2	Sri_Lanka	2	Tajikistan	3
Spain	1	Jamaica	2	Trinidad_and_Tobago	2	Thailand	3

Sweden	1	Jordan	2	Tunisia	2	Togo	3
Switzerland	1	Kazakhstan	2	Turkey	2	Uganda	3
United_Kingdom	1	Kuwait	2	Ukraine	2	Vietnam	3
United_States_of_America	1	Kyrgyz_Republic	2	United_Arab_Emirates	2	Zambia	3
Albania	2	Latvia	2	Uruguay	2	Zimbabwe	3

The characteristics of the countries were interpreted according to the table of final cluster centers and assigned profiles to the clusters. According to the final center table; the distances from the zero point of the variables included in the analysis were sorted from the largest to the smallest. This ranking is common to GII input sub-index, GII output sub-index, EPI environmental health, EPI ecosystem vitality, LPI, factor 1 and factor 3. The 1st rank is in cluster 1, 2nd rank is in cluster 2 and 3rd rank is in cluster 3. Only for factor 3, there is a difference in the ranking of the 1st and 2nd cluster, but the difference indistance are not very large. Therefore, the first cluster is named "high performer", the second cluster is a named "medium performer" and the third cluster is named a "low performer".

5. Results

It has been seen that countries can be clustered by chosen indexes and determined indicators. When the distance between final cluster centers are examined, clusters at the furthest distance from each other are seen as cluster 1 (high performer) and cluster 3 (low performer) (70,894). Clusters at the nearest distance from each other are determined as cluster 2 (medium performer) and cluster 3 (low performer) (34,648). The distance between cluster 1 (high performer) and cluster 2 (medium performer) is calculated as 39,030.

Considering the high performer class, Germany, which is the pioneer of Industry 4.0, and generally developed countries, is remarkable. Except for factor 2, the final cluster centers of all variables entering the analysis are high in this class. Cluster 2 consists of more countries than cluster 1 and 2. Cluster 2 is the cluster where factor 2 has the highest center with a very small margin. Except for factor 2, the final cluster centers of all variables entering the analysis are medium level in this class. The final cluster centers of all variables entering the analysis are the lowest in Cluster 3. Countries in Cluster 3 are generally underdeveloped countries.

The countries in cluster 1 are generally positioning among the top 30 according to the GII overall, EPI overall, LPI overall and SDGI overall. Especially in terms of global innovation index, the countries in this cluster are among the top 30 except for Greece. Australia and USA are the top 30 countries in the indexes, except for the SDGI overall. Malta and Iceland are among the first 30 countries in the indexes excluding LPI. Cyprus and Israel are among the first 30 countries in terms of GII overall and EPI overall. Estonia is among the first 30 countries in terms of GII overall and SDGI overall. Estonia is among the first 30 countries in the indices excluding EPI overall and LPI overall, while Singapore is among the first 30 countries in the indices excluding EPI overall and SDGI overall.

The countries in cluster 3 are generally made up of countries in the last 30 according to the GII overall, EPI overall, LPI overall and SDGI overall. China, Thailand and Vietnam are among the last 30 countries in only terms of EPI. India, South Africa, Kenya and Indonesia are among the last 30 countries in terms of GII overall and EPI overall.

Although China is a big economy, it is in the 3rd cluster. When the ranking of the countries in the indexes are examined; it is seen that China is 50th ranking in terms of SDGI, 15th ranking in terms of GII, 89th ranking in terms of EPI and 24th ranking in terms of LPI.

We conclude that the components and technologies of the Industry 4.0 Revolution act together with the principles and goals of sustainability. It was determined that there was a positive link between the indicators that explain the Industry 4.0 level and the overall sustainability performances of the countries. Intra group homogeneity was found for profiles in the group of three countries obtained in the clustering study.

It is thought that clustering analysis obtained with the indexes examined contributes to the Industry 4.0 and Sustainability relations area literature. In further studies, comparisons and interpretations with clusters obtained using different indexes and indicators can be made.

Acknowledgements

This work was supported by Scientific Research Projects Coordination Unit of Istanbul Gedik University.

References

[1] Doğruel Anuşlu, M. and Fırat, S. Ü. (2019) Endüstri 4.0 ve sürdürülebilirlik etkileşimi: Küresel endekslerle değerlendirmeler. In E. S. Bayrak Meydanoğlu, M. Klein and D. Kurt (Eds) *Dijital dönüşüm trendleri* (pp 56-100). İstanbul: Filiz Kitapevi.

[2] Fırat, O. Z. and Fırat, S. Ü. (2017). Endüstri 4.0 yolculuğunda trendler ve robotlar. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*. 46(2). 211-223.

[3] Fırat, S. Ü. O. and Fırat, O. Z. (2017) Dördüncü Sanayi Devriminde riskler robotlar ve yapay zekanın yönetişim sorunları. *Global Sanayici Dergisi*. http://www.sanayicidergisi.com.tr/dorduncu-sanayi-devriminde-riskler-robotlar-ve-yapay-zekanin-yonetisim-sorunlari-makale,644.html

[4] Global Innovation Index. History. (2019, 16th April) https://www.globalinnovationindex.org/about-gii#history

[5] Cornell University, INSEAD, and WIPO. (2018). *The Global Innovation Index 2018: Energizing the World with Innovation*. Ithaca, Fontainebleau, and Geneva.

[6] SDG Index & Dashboards. Overview. (2019, 3rd April). http://sdgindex.org/overview/

[7] Fırat, S. Ü., Yurtsever, Ö., İleri, Ç. and Kıvılcım, İ. (2017). *Sürdürülebilir Bir Dünyaya Doğru: Küresel Gündem ve Türkiye*. İstanbul: İktisadi Kalkınma Vakfı.

[8] Sachs, J., Schmidt-Traub, G., Kroll, C., Lafortune, G., and Fuller, G. (2018). *SDG Index and Dashboards Report 2018*. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN).

[9] The World Bank. LPI About. (2019, 6th April). https://lpi.worldbank.org/about

[10] Arvis, J.-F., Ojala, L., Wiederer, C., Shepherd, B., Raj, A., Dairabayeva, K., and Kiiski, T. (2018). *Connecting to Compete 2018: Trade Logistics in the Global Economy: The Logistics Performance Index and Its Indicators*. Washington: The International Bank for Reconstruction and Development/The World Bank.

[11] Wendling, Z.A., Emerson, J. W., Esty, D. C., Levy, M. A., de Sherbinin, A. et al. (2018). *2018 Environmental Performance Index*. New Haven, CT: Yale Center for Environmental Law & Policy. https://epi.yale.edu/

[12] Tufféry, S. (2011). *Data mining and statistics for decision making*. United Kingdom: John Wiley & Sons, Ltd.

[13] Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Edinburg Gate, Harlow: Pearson Education Limited.

[14] Nerurkar, P., Shirke, A., Chandane, M., and Bhirud, S. (2018). Empirical Analysis of Data Clustering Algorithms. *Procedia Computer Science, 12*, 770-779.

[15] Blattberg, R. C., Kim, B-D., and Neslin, S. (2008). *Database marketing: Analyzing and Managing Customers*. New York, USA: Springer Science+Business Media, LLC.

[16] Mirkin, B. (1996). *Nonconvex optimization and its applications: Mathematical classification and clustering*. The Netherlands: Kluwer Academic Publishers.

[17] Johnson, R. A., Wichern, D. W. (2002). *Applied multivariate statistical analysis*. The USA: Pearson Education International.

[18] Yu, S-S., Chu, S-W., Wang, C-M., Chan, Y-K., and Chang, T-C. (2018). Two improved k-means algorithms. *Applied Soft Computing, 68,* 747-755.

[19] Steinbach, M., Ertöz, L., and Kumar, V. (2004). The Challanges of Clustering High Dimensional Data. In L. T. Wille (Ed), *New directions in statistical physis: Econophysis, bioinformatics, and pattern recognition* (pp 273-309). Germany: Springer-Verlag.

[20] Tim, N. H., (2002). *Applied multivariate analysis*. New York: Springer-Verlag.

[21] Package 'missForest'. (2019, 29th April). https://cran.r-project.org/web/packages/missForest/missForest.pdf