ICIC Express Letters
Part B: Applications
Volume **9**, Number **11**, November **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **1139–1146**

CLUSTER ANALYSIS FOR STUDENT PERFORMANCE IN PISA2015 AMONG OECD ECONOMIES

Dian-Fu Chang[1] and Chia-Chi Chen[2,*]

[1]Graduate Institute of Educational Policy and Leadership
[2]Doctoral Program of Educational Leadership and Technology Management
Tamkang University
No. 151, Yingzhuan Rd., Tamsui Dist., New Taipei City 25137, Taiwan
140626@mail.tku.edu.tw; *Corresponding author: sophiabv03@gmail.com

Received May 2018; accepted August 2018

Abstract. *This study selected OECD 35 economy members' science, math, and reading scores and related impact factors as targets to mine the patterns and explore the main factors impact on the PISA2015 performance. The data selection was the first step; then this study applied observation clustering function with Minitab to determining the optimal clusters. The 3D scatterplot and 3D surface plot have been used to display the data structure. The dendrogram with three clusters drawn by Ward linkage and Euclidean distance has a relatively high similarity level and a relatively low distance level in this study. The result reveals OECD economies in the cluster1 and cluster2 are needed to improve their students' performance. The teaching hours per year in OECD economies have negative relationship with PISA2015 performance. While the teaching hours per year in economies can explain only 12.50% of the OECD/PISA2015 performance in the regression model. The OECD/PISA data provides an excellent databank for mining practices.*

Keywords: Cluster analysis, Data mining, Regression analysis, OECD, PISA2005, OECD/PISA2015

1. **Introduction.** Data mining has been employed in different fields like medicine, marketing, production, banking, hospital, telecommunication, supermarket, bioinformatics and education. In this entire field, lots of data are generated day by day, and if that data are not processed properly then that data are useless. However, if that data are processed properly then they will be helpful in making some decision for any relative organization [1]. Data mining is one of the best computer based intelligent tools used to check the performance of the students [1-4]. For making the analysis on the student data, most of studies selected algorithms like decision tree, Naive Bayes, random forest, PART and Bayes network. This study conducted the cluster analysis to explore the dissimilarity of students' OECD/PISA2015 (Organization for Economic Co-operation and Development/Program for International Students Assessment 2015) performance issue in the economy member countries. Previous studies indicated, various techniques of data mining like classification and clustering can be applied to uncovering hidden knowledge from educational data [1,5]. Furthermore, the related studies are predominant by the point of view from the institutional administration, management, different stakeholder, faculty, students as well as parents [5-7]. This study selected the OECD economies as the research target. There are 35 member countries in OECD from North and South America to Europe and Asia-Pacific. They include many of the world's most advanced countries but also emerging countries like Mexico, Chile and Turkey [8].

OECD launched the triennial survey of 15-year-old students around the world known as PISA in 2000. The PISA2015 survey focused on science, with reading, mathematics and

DOI: 10.24507/icicelb.09.11.1139

collaborative problem solving as minor areas of assessment. Approximately 540,000 students completed the assessment in 2015, representing about 29 million 15-year-old in the schools of the 72 participating countries and economies. In this survey, computer-based tests were used, with assessments lasting a total of two hours for each student. Test items were a mixture of multiple-choice questions and questions requiring students to construct their own responses. The items were organized in groups based on a passage setting out a real-life situation. About 810 minutes of test items for science, reading, mathematics and collaborative problem solving were covered, with different students taking different combinations of test items [9]. PISA is an ongoing program that offers insights for education policy and practice, and that helps monitor trends in students' acquisition of knowledge and skills across countries and in different demographic subgroups within each country. PISA results reveal what is possible in education by showing what students in the highest-performing and most rapidly improving education systems can do. Current results reveal some 8% of students across OECD countries (and 24% of students in Singapore) are top performers in science, meaning that they are proficient at level 5 or 6. Students at these levels are sufficiently skilled in and knowledgeable about science to creatively and autonomously apply their knowledge and skills to a wide variety of situations, including unfamiliar ones. About 20% of students across OECD countries perform below level 2, considering the baseline level of proficiency in science. At level 2, students can draw on their knowledge of basic science content and procedures to identify an appropriate explanation, interpret data, and identify the question being addressed in a simple experiment. All students should be expected to attain level 2 by the time they leave compulsory education [9].

Based on the understanding, this study selected OECD 35 economies' science, math, and reading scores and related impact factors as targets to mine the patterns and explore the main factor impact on the OECD/PISA2015 performance. Use cluster analysis result to observe 35 OECD economies to gain deeper insight into students' performance. Specifically, the purposes of this study are as follows: a) to realize the patterns of PISA2015 performance among the OECD economies; b) to determine the main factor impact on the OECD/PISA2015 results; c) to provide implication of policy applications. Given these purposes, the structure of this paper is as follows. First, the method section provides a brief description of the research methods. Second, display the result of cluster analysis and regression analysis. Finally, the conclusions are displayed.

2. **Method.**

2.1. **PISA data set.**

The PISA2015 database contains the full set of responses from individual students, school principals and parents. These files include countries/economies/sub-regions that fully met adjudication criteria. This study is designed by using PISA data to transform and interpret the meanings for OECD 35 economies. On OECD web, the files available on the page include background questionnaires, data files in ASCII format (from 2000 to 2015), codebooks, compendia and SAS™ and SPSS™ data files in order to process the data [10]. This study selected the science, math and reading scores in PISA2015 results report by using Minitab statistical package to transform the data. The selected data set has been presented in Table 1. The impact factors include central government's spending on education, teaching hours per year, and teaching staff in primary and upper secondary education which have been selected in the regression model. In this study, the technical terms include:

PISA2015 refers to the database of OECD's PISA survey in 2015;

OECD/PISA2015 refers to the 35 OECD economies' PISA2015 data.

Table 1. OECD/PISA2015 for 35 economies

Country (code)		PISA scores			
		Reading	Science	Math	Average
1	Chile (CHL)	459	447	423	443
2	Mexico (MEX)	423	416	408	416
3	Turkey (TUR)	428	425	420	424
4	Austria (AUT)	485	495	497	492
5	Czech Republic (CZE)	487	493	492	491
6	France (FRA)	499	495	493	496
7	Greece (GRC)	467	455	454	459
8	Hungary (HUN)	470	477	477	475
9	Iceland (ISL)	482	473	488	481
10	Israel (ISR)	479	467	470	472
11	Italy (ITA)	485	481	490	485
12	Latvia (LVA)	488	490	482	487
13	Luxembourg (LUX)	481	483	486	483
14	Portugal (PRT)	498	501	492	497
15	Slovak Republic (SVK)	453	461	475	463
16	Spain (ESP)	496	493	486	492
17	Sweden (SWE)	500	493	494	496
18	United States (USA)	497	496	470	488
19	Australia (AUS)	503	510	494	502
20	Belgium (BEL)	499	502	507	503
21	Canada (CAN)	527	528	516	524
22	Denmark (DNK)	500	502	511	504
23	Estonia (EST)	519	534	520	524
24	Finland (FIN)	526	531	511	523
25	Germany (DEU)	509	509	506	508
26	Ireland (IRL)	521	503	504	509
27	Japan (JPN)	516	538	532	529
28	Korea (KOR)	517	516	524	519
29	Netherlands (NLD)	503	509	512	508
30	New Zealand (NZL)	509	513	495	506
31	Norway (NOR)	513	498	502	504
32	Poland (POL)	506	501	504	504
33	Slovenia (SVN)	505	513	510	509
34	Switzerland (CHE)	492	506	521	506
35	United Kingdom (GBR)	498	509	492	500

2.2. **The key result of 3D plot.** The scatter and surface plot contain the following elements:

- Predictors on the x- (reading scores) and y-axes (math scores).
- A continuous scatter and surface that represent the response values on the z-axis (science scores).

2.3. **Cluster analysis.** Cluster analysis is a popular statistical method. It is also called segmentation analysis or taxonomy analysis, partitions sample data into groups or *clusters*. Clusters are formed such that objects in the same cluster are very similar, and objects in different clusters are very distinct. Basically, cluster evaluation determines the

optimal number of clusters for the data using different evaluation criteria in diverse settings. This study found previous studies have provided various examples for conducting cluster analysis [11-14]. In this study, the data selection was the first step; then hierarchical clustering with Minitab was applied to determining the clusters. Basic cluster algorithms are as follows:

- Select k point as initial centroids,
- Repeat,
- From k clusters by assigning each point to its closest centroids,
- Re-compute the centroids of each cluster,
- Until centroids do not change.

Typically, the clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The dendrogram function plots the cluster tree. Based on the dendrogram, this study decides the level or scale of clustering that is most appropriate for the data application. The Ward method was used to identify the minimum variability as the criterion for merging the within-cluster sum of squares; it can indicate that the similarity within the group is high. The Ward method transformed the data according to the following format:

$$d_{A,B} = n_A \|\bar{x}_A - \bar{\bar{x}}\|^2 + n_B \|\bar{x}_B - \bar{\bar{x}}\|^2$$

$d_{A,B}$ refers to the calculated distance between A and B. n_A and n_B refer to the number of variables in clusters A and B. $\bar{x}_A$ and $\bar{x}_B$ represent the statistical language signature (SLS) vector for language, $\bar{x}_A$ and $\bar{x}_B$ in clusters A and B, and $\bar{\bar{x}}$ is the centroid of cluster A or B, in other words to calculate the minimum distance squared of $\|\bar{x}_A - \bar{\bar{x}}\|^2$ and $\|\bar{x}_B - \bar{\bar{x}}\|^2$.

2.4. **Regression analysis.** In this study, the PISA_t in terms of the total scores of science, math, and reading for OECD economies will be assigned as the dependent variable in the regression model. The impact factors, which will be verified in regression model, include the central government's spending on education, teaching hours per year, and teaching staff in primary and upper secondary education for specific OECD economies. Stepwise method was applied in regression analysis to determining which independent variables can be used to interpret the PISA_t in the model. This study also considered the residual testing by using normal probability plot, versus fits, histogram, and versus order to check whether the model building has violated the statistical assumption.

3. Result.

3.1. **Data structure with 3D display.** The 3D scatterplot shows the relationship with science (z), math (y), and reading scores (x) among OECD economies presented in Figure 1. Data points that tend to rise together suggest a positive correlation. Outliers of OECD/PISA2015 science scores are increasing from the main group of data points. It means the science scores with related to math and reading among these economies. 3D surface plot is a three-dimensional graph that is useful for investigating desirable response values and operating conditions. The peaks and valleys correspond with combinations of x (reading) and y (math) that produce local maxima or minima. Minitab uses interpolation to create the surface area between the data points, see Figure 1.

3.2. **Cluster analysis.** The key outputs of cluster observations analysis include the similarity and distance values, the dendrogram, and the final partition. The higher the similarity level is, the more similar the observations are in each cluster. The lower the distance level is, the closer the observations are in each cluster. Ideally, the clusters should have a relatively high similarity level and a relatively low distance level. The dendrogram

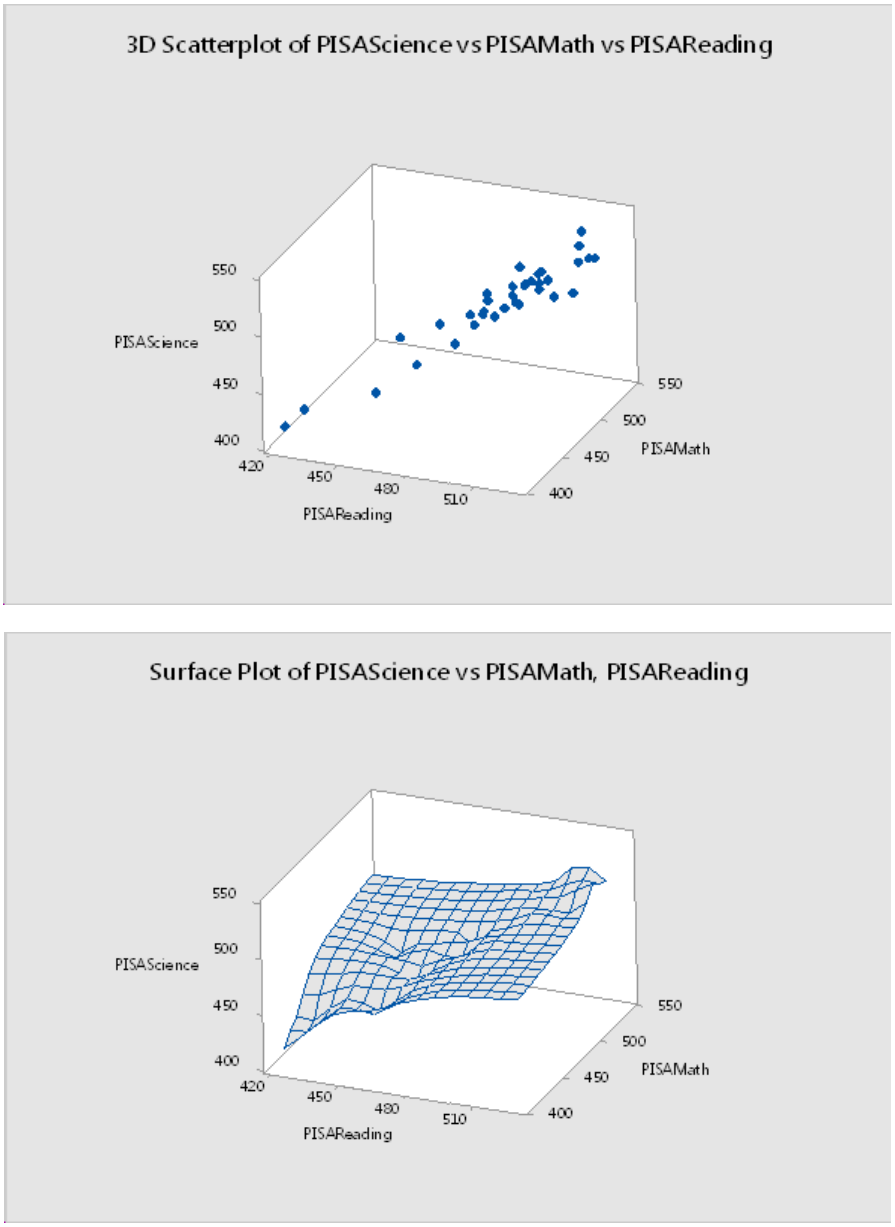

FIGURE 1. 3D scatterplot and surface plot for OECD/PISA2015

with three clusters drew by Ward linkage and Euclidean distance has a relatively high similarity level and a relatively low distance level, see Figure 2. Minitab also provides different colors with the groups to identify.

This dendrogram was created using a final partition of 3 clusters, which occurs at a similarity level of approximately 196. The cluster1 (far left) is composed of five observations (Chile, Greece, Slovak Republic, Mexico, and Turkey). The cluster2, directly in the middle, is composed of 13 observations (Austria, Czech Republic, France, Sweden, Portugal, Latvia, Spain, United States, Hungary, Israel, Iceland, Italy, and Luxembourg). The cluster3 is composed of 17 observations (Australia, United Kingdom, New Zealand, Belgium, Denmark, Switzerland, Germany, Netherlands, Slovenia, Ireland, Norway, Poland, Canada, Finland, Estonia, Japan, and Korea). After determining the final groupings, this study displays the final partition in Table 2. Table 3 shows the characteristics of each cluster with their centroids and distances.

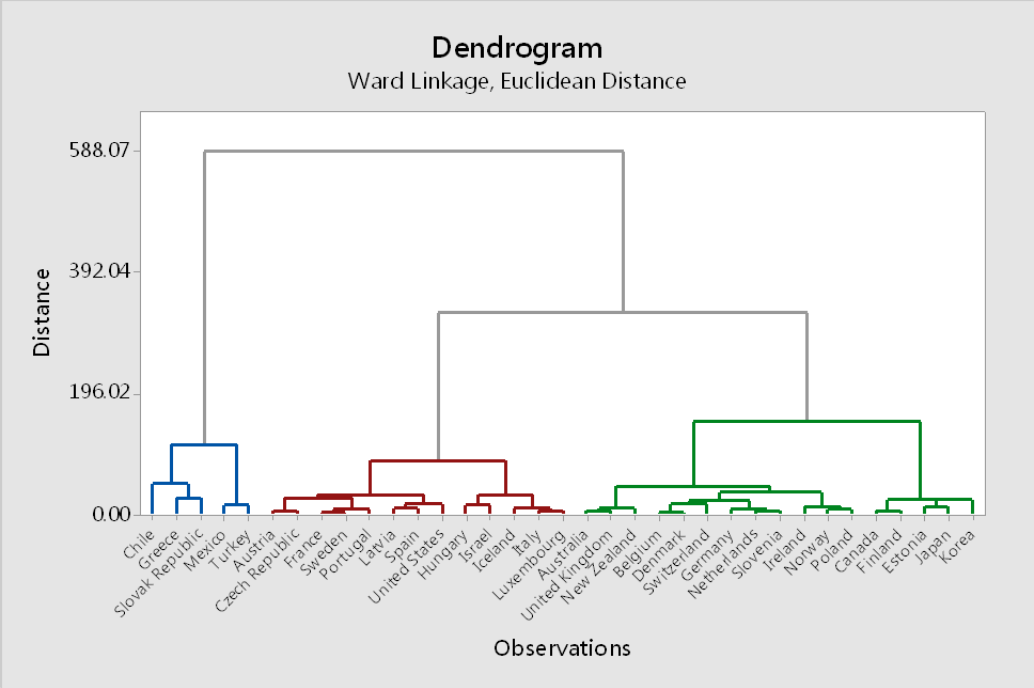

FIGURE 2. Three clusters among 35 OECD economies' PISA2015

TABLE 2. Final partition of cluster analysis OECD/PISA2015

Clusters	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	5	6078.80	33.5362	44.4752
Cluster2	13	3220.46	14.4183	27.5214
Cluster3	17	6013.29	17.3395	34.2162

TABLE 3. Cluster centroids and distances for OECD/PISA2015

Variables	Cluster1	Cluster2	Cluster3	Grand centroid
PISA Reading	446.0	488.231	509.588	492.571
PISA Math	436.0	485.923	509.471	490.229
PISA Science	440.8	487.462	513.059	493.229
Clusters distances	Cluster1	Cluster2	Cluster3	
Cluster1	0.000	80.331	121.090	
Cluster2	80.331	0.000	40.815	
Cluster3	121.090	40.815	0.000	

3.3. **Factors impact on PISA performance.** The proposed main impact factors in regression model include the central government's spending on education, teaching hours per year, and teaching staff in primary and upper secondary education for the individual OECD economies. Stepwise method was used to verify the regression model. The residual plots for PISA_t have been displayed in Figure 3. The result reveals only the teaching hours per year can explain 12.50% of the OECD/PISA2015 performance in the regression model ($R^2 = .125$, $p = 000$). Table 4 reveals the relationship between OECD/PISA2015 and the teaching hours per year is negative.

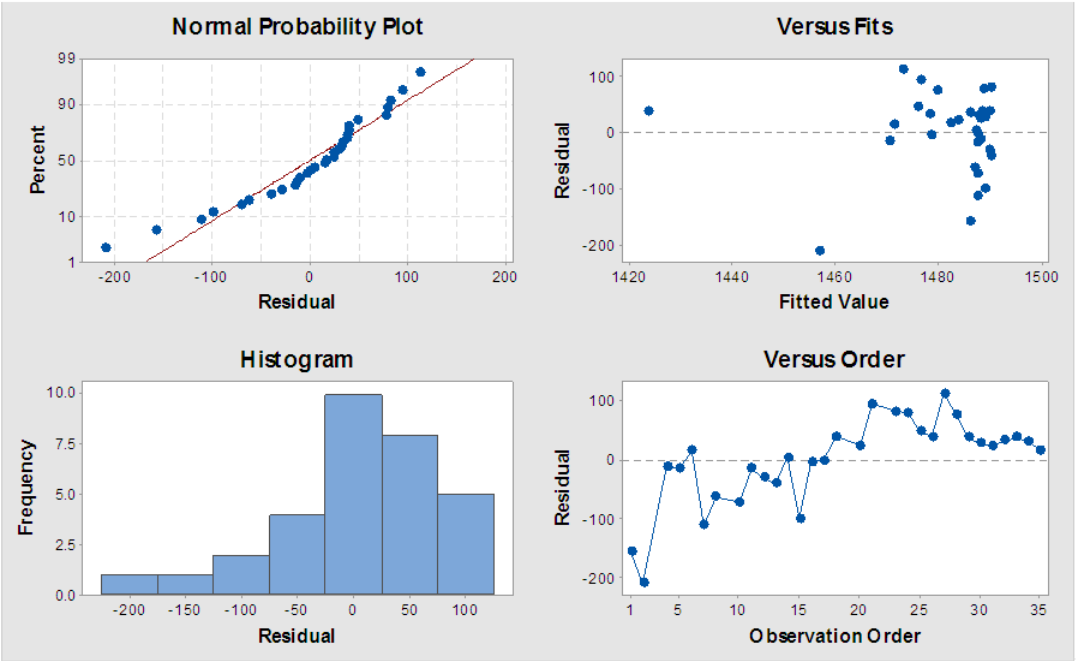

FIGURE 3. Residual plots for $PISA_t$ to build the regression model

TABLE 4. Analysis of variance and coefficients in the regression model

Source	**DF**	**Adj SS**	**Adj MS**	***F*-Value**	***P*-Value**
Regression	1	19980	19980	3.14	0.090
Teaching Hours Per Year	1	19980	19980	3.14	0.090
Error	22	139820	6355		
Total	23	159800			
Coefficients	**Coef**	**SE Coef**	***T*-Value**	***P*-Value**	**VIF**
Constant	1600.0	68.5	23.34	0.000	
Teaching Hours Per Year	−0.1770	0.0998	−1.77	0.090	1.00

4. **Conclusions.** Previous studies in educational data mining have created lots of interesting topics in the research community. PISA2015 data mining provides a unique example by using the OECD economies as a research target. Tryon, as we know the first time initiated the notion, indicated there are various algorithms and methods in cluster analysis [15]. Cluster analysis has become a popular statistical method in the research field. Taken advantaged from cluster analysis, it can group data over a variety of scales by creating a cluster tree or dendrogram. Based on the dendrogram, the study can easy to decide the level or scale of clustering that is most appropriate for the data application. While the group selection is not objective absolutely, the agglometrative methods in a study might include single linkage, average, centroid, Ward, etc. Cluster evaluation can be used to determine the optimal number of clusters for the data using different evaluation criteria in diverse settings.

The result of cluster analysis can explore deeply to insight the information of students' performance among 35 OECD economies. The result of cluster analysis for student performance in PISA can be used to scrutinize a specific educational system among the economies. Especially, the improvement suggestions for OECD economies in the cluster1 and cluster2 are needed. Moreover, this study found the teaching hours per year in OECD economies has negatively related with PISA2015 performance. The result of regression

analysis can be used to reboot the reasonable teaching system in individual economy. Mining the PISA results reveals what is possible in education by showing what students in the highest-performing and most rapidly improving education systems can do. For further studies, the PISA data could be an excellent databank as mining practices for related educational policy makers.

REFERENCES

[1] M. Kumar and A. J. Singh, Evaluation of data mining techniques for predicting student's performance, *International Journal of Modern Education and Computer Science*, vol.9, no.8, pp.25-31, 2017.

[2] M. Kumar, A. J. Singh and D. Handa, Literature survey on student's performance prediction in education using data mining techniques, *International Journal of Education and Management Engineering*, vol.6, no.6, pp.40-49, 2017.

[3] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, Analyzing undergraduate students' performance using educational data mining, *Computers & Education*, vol.113, no.177-194, 2017.

[4] J. Peral, A. Maté and M. Marco, Application of data mining techniques to identify relevant key performance indicators, *Computer Standards & Interfaces*, vol.54, no.2, pp.76-85, 2017.

[5] P. Kaur, M. Singh and G. S. Josan, Classification and prediction based data mining algorithms to predict slow learners in education sector, *Procedia Computer Science*, vol.57, pp.500-508, 2015.

[6] K. B. Bhegade and S. V. Shinde, Student performance prediction system with educational data mining, *International Journal of Computer Applications*, vol.146, no.5, pp.32-35, 2016.

[7] M. Durairaj and C. Vijitha, Educational data mining for prediction of student performance using clustering algorithms, *International Journal of Computer Science and Information Technologies*, vol.5, no.4, pp.5987-5991, 2014.

[8] OECD, *Members and Partners*, http://www.oecd.org/about/membersandpartners/, 2018.

[9] OECD, *PISA 2015 Results in Focus*, http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf, 2018.

[10] OECD, *PISA Database*, http://www.oecd.org/pisa/data/, 2018.

[11] O. R. Battaglia, B. D. Paola and C. Fazio, A new approach to investigate students' behavior by using cluster analysis as an unsupervised methodology in the field of education, *Applied Mathematics*, vol.7, no.15, pp.1649-1673, 2016.

[12] M. J. Brusco, R. Singh, J. D. Cradit and D. Steinley, Cluster analysis in empirical OM research: Survey and recommendations, *International Journal of Operations & Production Management*, vol.37, no.3, pp.300-320, 2017.

[13] M. L. Crowe, A. C. LoPilato, W. K. Campbell and J. D. Miller, Identifying two groups of entitled individuals: Cluster analysis reveals emotional stability and self-esteem distinction, *Journal of Personality Disorders*, vol.30, no.6, pp.1-14, 2015.

[14] N. K. Lankton, D. H. McKnight and J. F. Tripp, Facebook privacy management strategies: A cluster analysis of user privacy behaviors, *Computers in Human Behavior*, vol.76, pp.149-163, 2017.

[15] R. C. Tryon, *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*, Edwards Brothers, Ann Arbor, MC, 1939.