Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 13(3) (2021) 3248–3272

IJCI
International Journal of Curriculum and Instruction

# Educational data mining: Examination of science instruction methods and science literacy within the scope of self-organizing maps

Mehmet Taha ESER [a] *, Derya ÇOBANOĞLU AKTAN [b]

[a] *Aydın Adnan Menderes University, Faculty of Education, Aydın, 09010, Turkey*

[b] *Hacettepe University, Faculty of Education, Ankara, 06800, Turkey*

**Abstract**

By applying educational data mining methods to big data related to large-scale exams, functional relationships are discovered in a basic sense and hidden pattern(s) can be revealed. Within the scope of the research, to show how the self-organizing map (SOM) method can be used in terms of educational data mining, how SOM differs from other clustering methods in terms of visual outputs (map) and how to interpret the outputs, and it is aimed to give information about how effective the variables are in grouping individuals into groups according to the answers given to the items. In this study, students of OECD countries participating in the 2015 PISA were modeled using SOM and the outputs of the created model were examined. In this respect, the study can be accepted as a descriptive survey model. According to the results of the analysis, outputs were obtained for the educational process of the data set, the state of neurons, neighborhood distance, code vectors, heat maps, the number of clusters and the distribution of the number of students to countries and clusters. At the same time, it was determined that 4 clusters were formed according to the analysis results, and the most effective variables in clustering by examining the heat maps were perceived feedback from science teachers, teacher-directed science instruction, average of plausibe values in science, enquiry based science instruction and adaptive instruction in science lessons. Researchers who want to clearly determine the effectiveness of the input variables in cluster analysis can be advised to use SOM.



***Keywords:*** Cluster analysis; large-scale exams; PISA

* Corresponding author: Mehmet Taha ESER
*E-mail address*: m.taha.eser@adu.edu.tr

## 1. Introduction

### *1.1. Introduce the problem*

Achievement tests that include knowledge and skills from different grade levels and/or courses and consist of more than one subtest or dimension are called "Large-scale tests" (Roeben, 1997; Ehmke, Van Den Ham, Sälzer, Heine & Prenzel, 2020). Various large-scale tests are administered at national and international level. The data sets obtained from these exams are classified as "big data". As in other big data sets, revealing the patterns in these data and interpreting these patterns require the use of appropriate analysis methods (Peña-Ayala, 2014; Baker, Mardin & Rossi, 2017; Aldowah, Al-Samarraie & Fauzy, 2019). Analyzes and researches conducted to determine the patterns belonging to these big data in the field of education are evaluated within "Educational data mining" (Romerao & Ventura, 2013; Baker & Yacef, 2009). In this context, it will be useful to use educational data mining methods to reveal the patterns of big data sets obtained from PISA (Program for International Student Assessment) and similar international large-scale tests.

By applying educational data mining methods to the big data of large-scale tests, basic functional relationships can be discovered, and hidden pattern(s) can be revealed (Jain & Dubes, 1998; Wu, 2012). Principal component analysis and multidimensional scaling methods are generally used in educational research for dividing the complex data set into more meaningful and smaller structures, determining the functional relations between structures and the profiles of individuals (Kruskal & Wish, 1978; Barlow, 1989; Becker & Plumbley, 1996). Regarding principal component analysis and multidimensional scaling, although they produce solution up to a point in terms of dimension reduction and the discovery of functional relations related to the dimensions, the problem of visualization continues and therefore the analysis results need to be supported with more descriptive visuals. Moreover, there are problems in the interpretation of the findings (Gurney, 1997; Haykin, 2009; Kohonen, 2014). At this point, “Clustering analysis”, one of the data mining analysis, is of great importance in overcoming these problems, revealing hidden patterns related to the data set and determining the profile of the individuals.

Regarding educational data mining, cluster analysis can provide information about the distribution of the data, structural relations between the clusters and the properties of the clusters formed by individuals (Baker & Yacef, 2009; ALMazroui, 2013). Clustering methods used in cluster analysis allows to specify meaningful clusters and discover useful patterns in the data set by using more or less a priori knowledge (Anderberg, 1973; Garson, 2014). Cluster analysis includes k-means, hierarchical and classical clustering methods. However, the data sets cannot be characterized in detail with the mentioned clustering methods, thus student trends in groups smaller than the clusters cannot be analyzed in detail (Kaufman & Rousseeuw, 1990; Jain & Dubes, 1998).

Moreover, these methods cannot provide outputs that clearly visualize the relationships between the clusters and variables. Visuals are of great importance for the researcher to deepen the knowledge about the structure of the data set as a whole (Thuneberg & Hotulainen, 2006; Lee, 2019).

In this context, Kohonen's Self-Organizing Map, which is a relatively new clustering method, provides solution to this visualization problem (Kohonen, 2001;Schreck, Bernard, von Landesberger & Kohlhammer, 2009 ). As an alternative clustering method Kohonen Self-Organizing Map (SOM) provides additional information that clarify the nature of the clusters for identifying the pattern, determining student trends, revealing important properties related to topological structures and input data, and observing these properties on the maps (Kuo, Ho & Hu, 2002; Lupaşcu & Tegolo, 2011). The maps obtained as a result of this method are grouped and displayed in a way that allows to observe student groups having similar response patterns. SOM allows the visualization of the properties of the groups and thus big groups are analyzed in more detail (Kohonen, 2001; Nielsen & Yezierski, 2016).

Another advantage of SOM is that statistical tests are not based on any assumptions that must be meet (Kiang, 2001; Kiang & Kumar, 2001; Wu & Chow, 2003). SOM does not require to meet any assumptions regarding the initial number of clusters, probability distributions of the variables, and independence of the variables; and it gives much better results than other methods, especially when working with multidimensional data sets (Multidimensionality makes statistical correlations insignificant and therefore statistical methods are inadequate and powerless in analyzing such data sets), which make SOM more useful than the other methods (Dunham, 2003; Dasu & Johnson, 2003; Penn, 2005). Moreover, the analyzes made by SOM (training process) are based on artificial neural network. SOM appears to be a method that can be used easily in terms of visualization of information, dimensionality reduction, data aggregation and data mining (Murtagh &Hernández-Pajares, 1995; Kohonen, 2001).

This research is important in terms of providing information to the researchers about the functioning of SOM; showing how SOM can be used in educational data mining; and presenting the differentiation of SOM from other clustering methods in terms of visual output (map) and the way of interpreting these outputs. In addition, the literature review revealed that there is just a few study in educational sciences in which SOM is used (Nielsen & Yezierski, 2016; Qiao & Jiao, 2018; Lee, 2019), which can be considered as another factor that increases the importance of the study. The aim of the study is to show the unique aspects of SOM in visualizing, understanding and interpreting educational data, which are different than other clustering methods. For this purpose, the data of PISA 2015 exam, one of the large-scale tests, were used in the study. The cluster analysis conducted in the research was carried out using the R program. Literature review showed that in educational data mining, there is no study in which SOM is applied by

using the R program, whereas there are studies in which SOM is applied using other statistical programs. As a result of the self-organizing map analysis performed with the R program, richer outputs can be obtained compared to the outputs obtained with other statistical programs. These rich outputs provide great convenience for the researcher in interpreting the results.

PISA is an exam administered every three years to measure literacy levels of 15-year-old students in science, mathematics and reading. Science was specified as the priority area in 2015. Most of the scientific principles and theories that 15-year-old students possess are taught at school. As in other fields, the way science is taught in schools does not only affect the achievements of the students in science, but also affects those who want to be involved in science in their higher education and career planning. Considering the expected growth in science-related employment worldwide and the decrease in students' interest in science at school, it has become more important to examine why some students are more interested in science-related careers. This has created the need to analyze the resources offered for science in detail, such as science learning opportunities at school, laboratory applications, science teachers and science activities, and the ways science is taught at school. For this reason, the studies involving the factors affecting science instruction are considered important. Various studies have been conducted on the factors affecting science instruction in the last 30 years (Taber, 2009; Lin, Yen, Liang, & Guo, 2016; Langdon, McKittrick, Beede, Khan and Doms, 2011; Vedder-Weiss, & Fortus, 2011). The input variables used in the model for cluster analysis are students answers about teacher-directed instruction, perceived feedback, adaptive instruction and inquiry-based instruction from in PISA 2015 data, which have been determined to be effective in science achievement, and students' average possible science achievement scores. The second objective of the study is determining the effectiveness of the variables in grouping individuals into clusters according to the answers given to the items on science instruction in the PISA student questionnaire.

In the light of the above information and considering the students in the OECD sample and who answered the PISA 2015 student questionnaire, the research questions are as follows:

1) Considering science teaching methods and the average of plausible values, how many clusters are students divided into?

2) Considering science teaching methods and the average of plausible values, what can be said about the importance level of input variables in the formation of the clusters?

To better understand the study, further information is given about the structural properties of SOM and the processes related to SOM are explained before proceeding to the methodology part.

### *1.1. Structural properties of SOM*

SOM is a type of neural network that was defined by Tuevo Kohonen in 1982 and explained with different application areas in the articles and books that he presented later, that does not need control through output or a different type of feedback during training (Penn, 2005). SOM also appears as an analysis that allows the visualization of multidimensional data in an easily understandable way. As a neural network, it usually consists of an input and an outlet layer; usually there are single-dimensional processing elements in the input layer, whereas output layer consists of two-dimensional processing elements deployed in various geometries (Kohonen, 1984; 2001; 2014). There is a link between each processing element of the input layer and the processing elements of the output layer, and this link is kept in the reference vector belonging to the output processing element (Kohonen, 2001; Thuneberg & Hotulainen, 2006).

In Figure 1, the relationship between input vectors and SOM neurons is illustrated. Each colored circle is called neuron (also called a node or reference vector), and the structure formed by neurons is called grid. The network needs to be trained to reveal the relationships between the input layer and the output layer.

When data is entered for network training, the output layer (competitive layer) automatically learns the inner topology structure based on the learning algorithm. For this purpose, the learning algorithm uses an iterative process. Competition between neurons begins at the output layer when the data is entered for training. In the training process, the weight vectors of the winning neurons and neighboring neurons are updated to approximate the input data. Update processes are repeated until a predetermined stopping criterion is met, and network training ends when the stopping criterion is met (Bagan et al., 2005).

The training process of SOM, visualization and determining the ideal number of clusters are discussed below.

### *1.2. Training process of SOM*

SOM emerged from neural network models, especially from associative memory and adaptive learning models (Kohonen, 1984). The method is based on obtaining results based on the observations of the cerebral cortex in explaining the spatial organization of brain functions. The spatial sequential line detectors of Malsburg (1973) and the neural field model of Amari (1980), which have been developed before Kohonen, formed the basis of SOM, which is based on self-organizing neural networks. The ability of self-organizing sets the stage for new possibilities. In addition, this feature is the most natural form of learning, shaped in the human brain. New possibilities take shape in the learning process. SOM provides network groups that use self-organizing, competitive type learning method.

In SOM, a signal is generated on the network input and then the neuron that best corresponds to the input vector, in other words the winning neuron, is determined. The scheme of the competition of neurons and modifications of synaptic cells can take various forms. There are many subtypes of the method that can be differentiated by a competitive and self-organizing algorithm. Perhaps the most important of these subsystems is the competitive neural network approach that has adopted the win-win function. In addition, there is another subsystem controlled by the neural network and alters the local synaptic flexibility of learning neurons. Learning is limited to the neighborhood of the most active neurons. The flexibility of the subsystem involved in control may be based on nonspecific neural interactions, but this is mostly a chemical control effect. The formation of the self-organizing system is possible by separating neural signal transfer and flexibility control (Kaski, Kangas, & Kohonen, 1998; 2014). However, SOM can also be expressed in a pure and abstract mathematical form without reference to any underlying neural or other component (Murtagh, F, & Hernández-Pajares, 1995; Kaski, Kangas & Kohonen, 1998).

In SOM, networks are formed as a result of data competition starting with a training algorithm. For this purpose, a variable that will give a relational dimension to the data set should be chosen. The learning phase consists of four steps. In the first step, a reference grid is created according to the size of the input; weighted vectors are randomly placed in each cell of the grid with the corresponding colors.

In the second step, input data vectors are assigned to the colored nodes of the grid that share the closest possible weight vectors, and this is the cornerstone of the training process. When a match occurs, the unit that best matches is called as "winning unit".

In the third step, after the "winning unit" (also called the best matching unit) has been determined, the allocation of neighborhood relations begin to appear on the map.

In the fourth step, an update process is carried out. The "winning unit", which is the most similar unit to the existing learning object, is updated to become even more similar to the learning object. The weighted average is used during the update process and the weight of the new object is one of the training parameters of the analysis. For the training process, the learning rate is called alpha and is usually set as the default value of 0.05 (Kohonen, 2001; Kohonen, 2014).

### *1.3. Visualization of SOM*

The visualization stage, which is important in terms of data analysis and obtaining results, begins after the completion of training process. In SOM, a single graph is used to show the cluster density of the different areas of the data. The increase in the density of reference vectors in an self-organized map means that more units come together in that area and thus clustering occurs (Kohonen, 2014; Ritter, 1991). Reference vectors will be close to each other in clustered areas and sparser in empty spaces. In this way, the

cluster structure of the data set gets visible by showing the distances between the reference vectors of neighboring units.

When creating a cluster map, the distance between each reference vector pair is calculated and scaled, in this way the distances fall between a certain minimum and maximum value. On the map screen, each scaled distance determines the grayness level or color of the corresponding map unit. The grayness values of the points corresponding to the map units are adjusted according to the average of some of the closest distance values (on a hexagonal grid, for example, the average of three of the six distances towards the lower right corner). After these values are specified, the map can be created, or the values can be flattened spatially. The shapes of the clusters in the resulting map may not provide realistic information about the actual shapes of the clusters. Most of the clustering algorithms prefer particular cluster shapes (Jain & Dubes, 1988; Anderson, 1999).

### *1.4. Determining the ideal number of clusters in SOM*

Elbow method and silhouette coefficient are used to determine the ideal number of clusters in SOM. The elbow method is used in the interpretation of the visual that illustrates the change of the within-cluster sum of squares according to the number of clusters. At this point, the elbow method considers the amount of explained variance as an indicator of the number of clusters. The clusters should be modeled in such a way that they do not overlap with each other, considering the amount of explained variance. While interpreting the visual about the change of the within-cluster sum of squares according to the number of clusters, the projection of the point where the marginal gain falls, that is, the graph begins to take the form of a flat plateau, indicates the ideal number of clusters. The silhouette coefficient, which is considered a combined measure of both cohesion and separation, is also used to determine the ideal number of clusters. The silhouette coefficient gives information about the distance of cluster elements to neighboring clusters. The silhouette coefficient takes a value in the range of [-1, +1]. The scores around +1 indicate that it is too far from neighboring clusters, whereas the values around 0 indicate proximity to the boundaries between clusters. Negative values indicate the possibility of misclassified samples (Jain & Dubes, 1988; Kaufman & Rousseeuw; 1990).

## 2. Method

In this section, information about the research type, study group and analysis of the data is given.

### *2.1. Research type*

In this study, students of OECD countries participating in the 2015 PISA were modeled using SOM and the outputs of the created model were examined. In this respect, the study can be accepted as a descriptive survey model.

*2.2. Data*

2015 PISA data collected from the students of OECD member countries, except Slovenia, were used in the study (The link related to data: https://figshare.com/articles/dataset/Data_and_Syntax/15138660). While performing the cluster analysis in the R program, the computation time of the computer was too long due to the size of the data set at hand. To prevent this delay, systematic sampling was applied to the data set and accordingly the analyzes were carried out with the data of 9,870 students (see Apendix 2 for country codes). The data obtained from the students of 34 countries were included in the cluster analysis. Regarding Figure 1, which shows the distribution of the number of students by country, the highest number of participants are from Canada and the lowest number from Iceland. The number of participants shows a significant decrease after Canada, which is followed by Australia and the UK. Poland and Lithuania follow Iceland as the countries with the lowest number of participants.

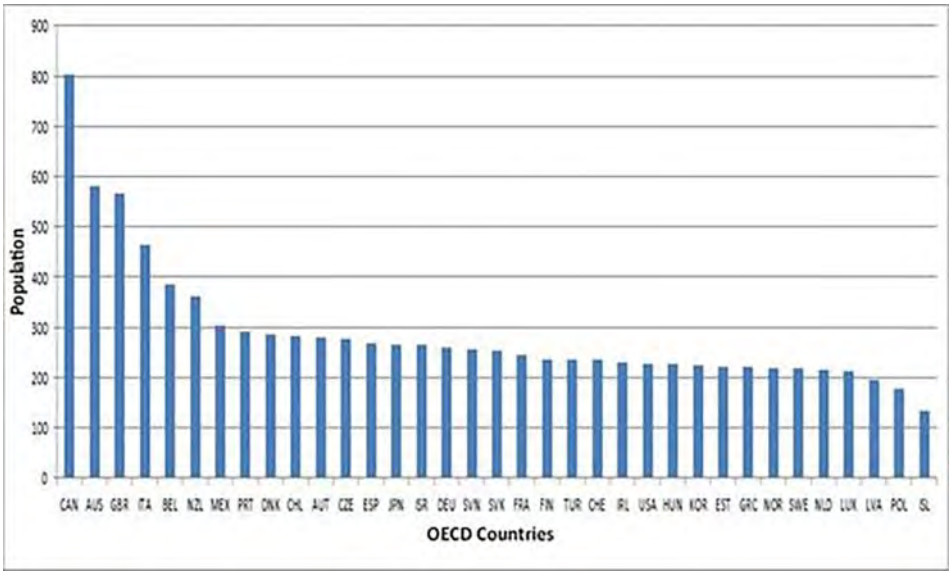


Figure 1. Short term retention of new vocabulary

Sub-dimensions of science instruction in PISA 2015 student questionnaire and the average of possible science achievement scores were used as input data (see Apendix 1 for information about items). The total number of items in science instruction's sub-dimensions of PISA 2015 was 21, but if the scores of all these 21 items were used as input, the comments to be made on these scores will suppress the comments of the model, thus factor scores were used for the sake of ease of interpretation. In addition, another reason for preferring to use factor scores, they are weighted combinations of associated variables, which makes this score type more reliable and higher quality compared to actual scores (Fiedler & Mcdonald, 1993; Milligan, 1980). Input variables used in the study were coded as follows: Factor 1 - Teacher-directed science instruction, Factor 2 -

Perceived feedback from science teachers, Factor 3 - Adaptive instruction in science lessons, Factor 4 - Enquiry-based science instruction, Factor 5/PVSCIENCE - Average of plausible values in science.

"Data Pre-Processing" was performed before analysis, to make the data ready for analysis. In this context, lost data analysis was performed first.

Multiple value assignment method was used in lost data analysis. The assignment of multiple values was based on logistic regression due to the structure of the data set. Multiple value assignment is a data assignment method that can be applied to data sets of different structures. The working principle of the method is to make a predetermined number of iterations for the incomplete data set and assign possible values to the missing values in each iteration. Multiple value assignment method is a highly valid method resistant to bias and extreme values. The method is applied to data sets that have random missing data (Little & Rubin, 1987; Graham, 2009; Bodner, 2008;). Due to the random structure of the missing data in the data set of PISA surveys, multiple value assignment method, one of the missing data assignment methods, was applied to the data set used in the study (Adams, Lietz, & Berezner, 2013; Kaplan & Su, 2016). As a result of missing data assignment, the data of Slovenia, which had a high amount of lost data, were removed from the data set considering the variables used in the research.

Systematic sampling was applied to the data set after the missing data assignment process. The reason for administering systematic sampling is that the computer's computation time was too long since PISA exam data set obtained from 253,140 students was too big to perform cluster analysis in the R program and high processor computers are needed to overcome this delay. A data set containing the responses of 10,000 individuals would be sufficient to perform the cluster analysis in the R program and the proportion constant was set as k = 25 (253.140/10.128). A macro was created in Excel and 1 out of every 25 students in the universe was taken into the sample. As a result of systematic sampling, the data set was observed to include the data of 9,870 students. Systematic sampling is a non-random sampling method that includes people selected from the universe at regular intervals (Monette, Sullivan, & Jong, 1990).

### *2.3. Data analysis*

The number of neurons in the analysis was chosen as 30 × 30 = 900, which was determined as a reasonable number by the program, in order to fit the 9,870 observations of the data set to be used as the initial grid. The shape specified for the grid was a hexagon. The number of iterations and learning rate were kept at the default values of 100 and 0.05 (learning rate decreases linearly by 0.01 for each iteration). Normally, the learning rate is set at the beginning of the analysis and does not change as the number of iterations changes. But, in clustering by SOM, it is necessary to reduce the learning rate to ensure convergence. In other words, if the learning rate does not change, the training process may not end.

Self-Organizing Map analysis performed in the research was carried out by the R program. Kohonen package was used to carry out the analysis in the R program (Wehrens & Buydens, 2007). As a result of the analysis: average distance - number of iterations graph was examined; counts plot was used to determine the nature of the two-dimensional map created for the model; neighborhood distance chart (U-Matrix) was used to obtain information about the distance between the vectors of neighboring neurons; code vectors distribution map was used to get information about the distribution pattern of the units and variables in the model; heatmap was used to obtain information about the significance of clustering variables; the change of within-cluster sum of squares was used to determine the ideal number of clusters; silhouette plots and calibration plots were used to determine the number of clusters.

## 3. Results

The first stage of SOM cluster analysis is the training process. As the number of iterations increase, the distance between the samples represented by the neuron decreases due to the weights of each neuron. Figure 2 shows the change of within-cluster distances according to the number of iterations.

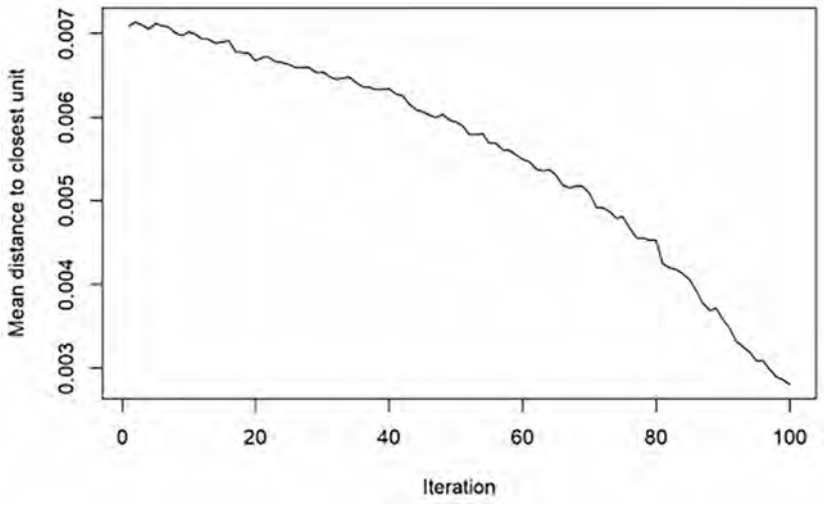


Figure 2. Training process of data set

According to Figure 2, as the number of iterations in the SOM training process increases, the distance between each neuron and the units represented by that neuron decreases because of the weight of the neuron. Moreover, regarding the graph of the training process, although the average distance between the observations and the unit closest to them is not stabilized, a faster downward trend is observed in the last iterations, therefore the training process seems to be effective. At this stage, what is

desired is that the line graph reaches a flat plateau, as in the scree plot. At the same time, it should be noted that increasing the number of iterations would not result with a better fit, as it tends to decrease rapidly; therefore due to the risk of overfitting that may occur as the number of iterations increases, it has been decided to keep the number of iterations at 100, which is the default number. Accordingly, the number of iterations was found to be sufficient for cluster analysis.

SOM provides Counts Plot, which helps visualize the number of samples associated with each neuron. This visual is a measure of the quality of the clustering of the original map. In the graphic in Figure 3, the number of units associated with each neuron are shown by colors. A metric grading is made from blue to red in the color palette shown vertically on the left side of the chart. Blue tones indicate that the number of units associated with neurons is low; whereas red tones show high number of units associated with the neurons. High number of blue neurons indicates that the size of the map, i.e. the number of neurons, is too much for the data set and therefore should be reduced; whereas having many red neurons means that the number of neurons used for mapping should be increased. In addition, it is thought that having 5-10 neurons for each color of the color palette will contribute to homogeneity (Kohonen, 2001; Wehrens & Buydens, 2007). Counts Plot showing the number of units in each neuron is displayed in Figure 3.

Regarding the graph shown in Figure 3, it can be said that the number of observations per neuron is relatively homogeneous. In addition, in general terms, there are at least 5-10 neurons in the map for each color of the color palette. As a result, considering the map consisting of 900 neurons shown in Figure 3 and the number of units in data set, it is concluded that the number of neurons is sufficient.

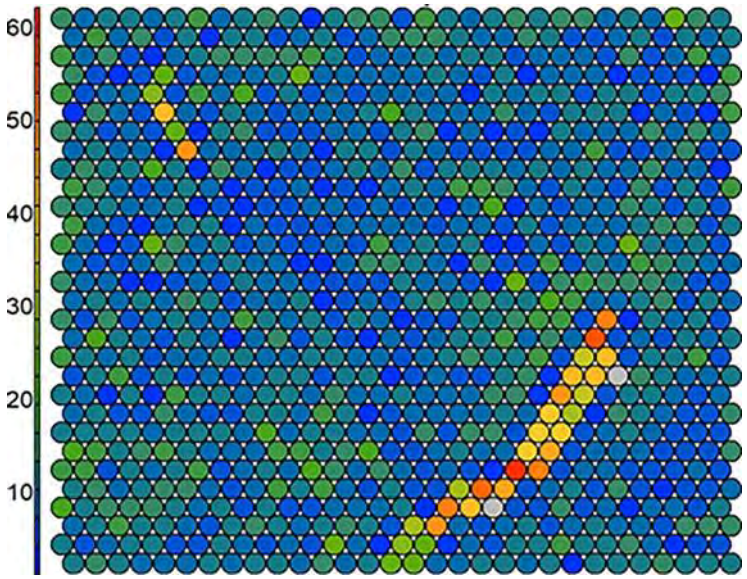

Figure 3. Counts plot for the units associated with neurons

Another graph obtained as a result of the analysis is "Neighborhood Distance Chart" which gives information about the distance between the vectors of neighboring neurons. It is also known as "U-Matrix". This graph gives information about the distance between each neuron in the map and suggests using the boundaries while bringing similar neurons together. The Neighborhood Distance Chart for clusters obtained from SOM and the distances between them are shown in Figure 5.

According to Figure 4, as the color palette progresses from blue to red, the neighborhood distances of neurons increase. The distance in question is Euclidean distance. In the graph showing the neighborhood distances, blue tones indicate low distance between neurons, meaning that the neuron groups are similar. The red tones in the graph show that the distance between neurons is high, so the neuron groups are different. Accordingly, Figure 4 can be used to describe the clusters created by SOM. Regarding the chart shown in Figure 4, neurons that are very close to their neighbors are dark blue, neurons at moderate distance to their neighbors are green, and those that are far from their neighbors are red. This is a clue to distinguish the clusters from each other visually.

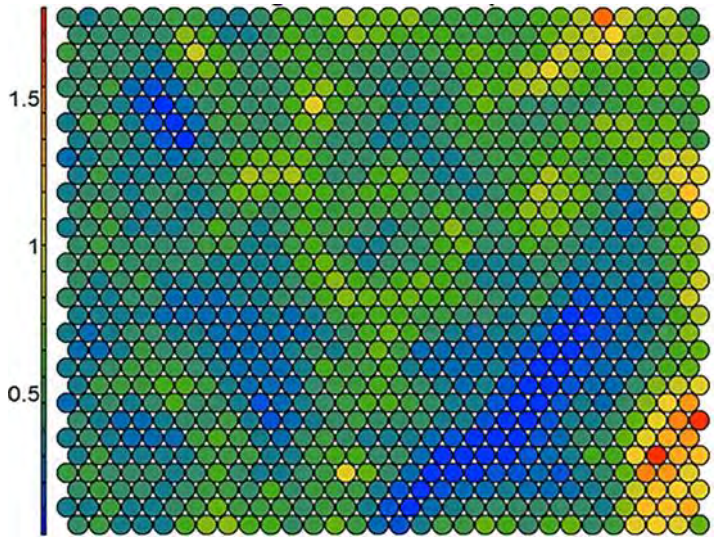


Figure 4. Neighborhood distance chart (u-matrix)

Weight vectors of the neurons, also known as "Codes," consist of normalized values of the variables used to create a self-organized map. The weight vector of each neuron is representative of the samples associated with that neuron. The visualization of the weight vectors on the map provides information about the distribution patterns of the units and variables. In short, a map showing the distribution of code vectors provides

information about the role of the variables taken into the analysis in defining different areas of the relevant map. The map of the Weight Vectors obtained as a result of the analysis is shown in Figure 5.

Figure 5 shows the weight of each neuron according to the variables covered in the analysis. The intensity of the colors defined for the variables in the map gives information about the relative effect of the relevant variable. Regarding code vectors distribution map, Factor 1 (fct.1) is present in the majority of the neurons that constitute the map and its weight is higher than other variables. The z variable, which is defined as the average possible achievement score for science literacy, is observed to be secondly present in most of the neurons and has a higher weigh than other variables.

As the number of neurons and the number of variables increases, reading code vectors distribution map becomes difficult. Code vectors distribution map created in the analysis consists of 900 neurons, which makes it difficult to interpret. At this point, instead of trying to determine the weights of all variables in a single map, a graph highlighting the contrast between the areas with high and low value of each variable can be created. These univariate graphics are easier to interpret than the code vectors map.

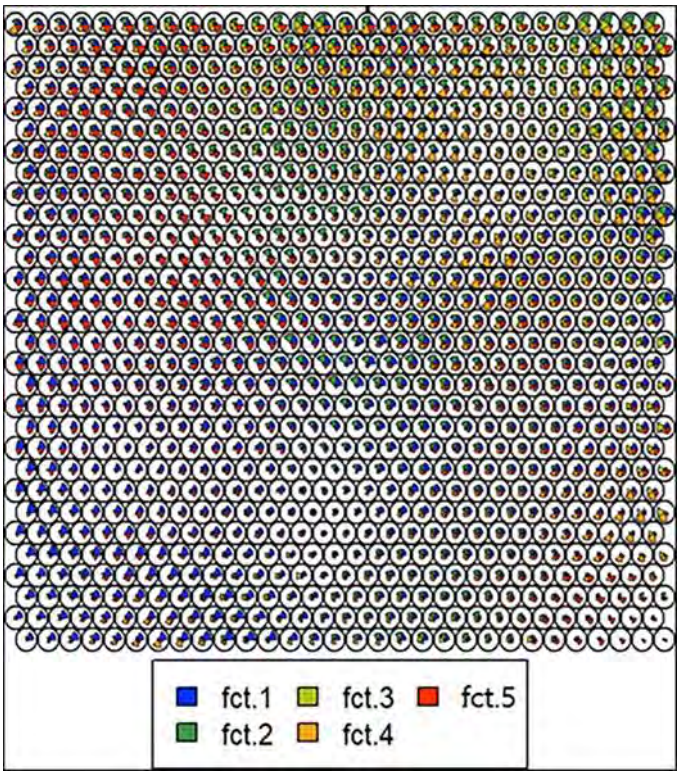


Figure 5. Code vectors distribution map

In Figure 6, there are heatmaps showing the importance of the five variables in the formation of the clusters. In the heatmaps, the unit numbers of the vertical axis are

normalized values. In the color palette, the transition from blue to red was numerically corresponding to a scale from 0 to 60; whereas normalized values of the heatmaps vary between -2 and +2. According to the heatmaps, in the area where Factor 1 (Teacher-directed science instruction) is most effective in clustering, Factor 3 (Adaptive instruction in science lessons) has the least effect; in the area where Factor 3 has the least effect, Factor 4 has the relatively less effect. In addition, in the area where Factor 2 (Perceived feedback from science teachers) is most effective in clustering, Factor 4 is also effective but not as much as Factor 2. According to these results, it can be said that there is a negative relationship between Factor 1 and Factor 3, while there is a positive relationship between Factor 2 and Factor 4. Visually, the graphs of Factor 1 and Factor 3 are in different colors; whereas the graphs of Factor 4 and Factor 2 uses the same color.

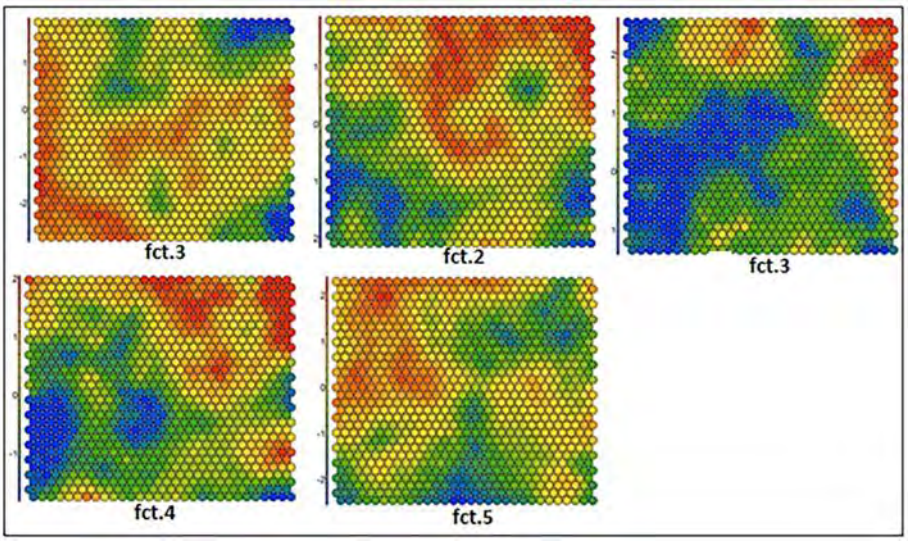


Figure 6. Heatmaps of the factors

After getting information from heatmaps about the effect of each variable on the cluster analysis, the stage of determining the number of clusters, under which the data is grouped, has started. For this purpose, first it is required to review the graph in Figure 7 showing how within-cluster sum of squares vary according to the number of clusters.

To decide the ideal number of clusters, the point where the graphic starts to form a flat plateau should be identified. Since it is quite difficult to determine the ideal number of clusters visually, from the graph, the silhouette plots should be reviewed for the number of clusters k = 2, 3, 4 and 5.

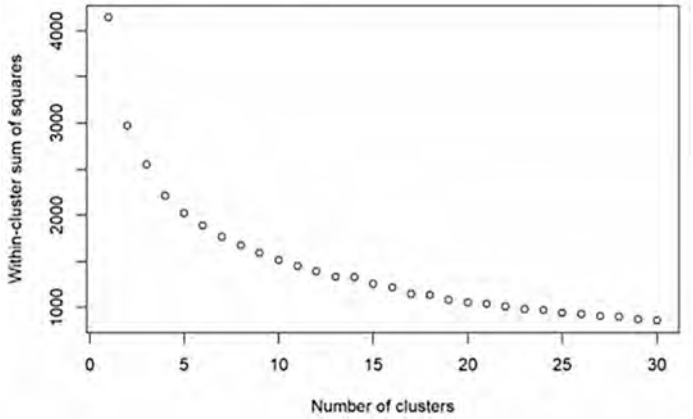

Figure 7. The change of within-cluster sum of squares according to the number of clusters

In Figure 8, silhouette plots obtained for different number of clusters are displayed. Regarding cluster analysis, silhouette plots can be used to calculate the distance between clusters, that is, to decide on the number of clusters. Silhouette plot is the visual evaluation of the parameters such as the distance of each point in a cluster to the points in neighboring clusters and the number of clusters. Silhouette coefficient takes a value between -1 and +1.

Silhouette coefficients are evaluated at four different levels: being equal to or less than 0.25 means that no significant cluster is found; between 0.26-0.50 means that the obtained structure is weak and different algorithms should be tested; a value between 0.51-0.70 indicates a reasonable structure; and a silhouette coefficient between 0.71-1.00 means a strong structure (Kaufman and Rousseeuw, 1990).

Regarding the plots obtained for different number of clusters; for 2 clusters, there are 337 neurons in the first cluster, 563 neurons in the second cluster and the average silhouette coefficient is 0.25; for 3 clusters, there are 238 neurons in the first cluster, 329 neurons in the second cluster, 333 neurons in the third cluster and the average silhouette coefficient is 0.19; for 4 clusters, there are 188 neurons in the first cluster, 270 neurons in the second cluster, 191 neurons in the third cluster, 251 neurons in the fourth cluster and the average silhouette coefficient is 0.53; for 5 clusters, there are 140 neurons in the first cluster, 259 neurons in the second cluster, 205 neurons in the third cluster, 124 neurons in the fourth cluster, 172 neurons in the fifth cluster and the average silhouette coefficient is 0.18. When the number of clusters is specified as four, the silhouette coefficients of each cluster are found to be above 0.51, in other words all four clusters have an acceptable structure, and the overall silhouette coefficient is 0.53. Regarding Figure 8, there may be neurons assigned to the wrong clusters (color extensions in the negative direction) when the number of clusters is set as two, three and five. As a result, considering that the average silhouette coefficient and within-clusters silhouette coefficients are above 0.50 and there are no neurons assigned to the wrong clusters, it can be said that the ideal number of clusters is 4 (four) and the chosen clustering method is appropriate.

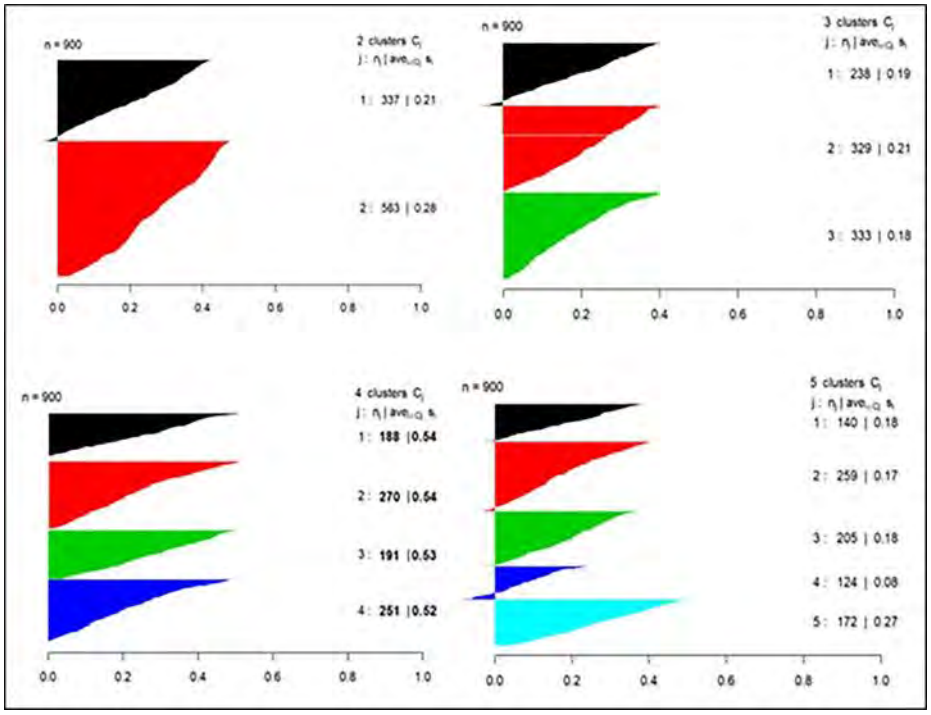


Figure 8. Silhouette plots for the validity of the clusters

The calibration plot, which is another criterion in determining the ideal number of clusters, is formed by merging similar and neighboring neurons and it is shown in Figure 10. At the beginning of the calibration process, each student (observation/unit) is considered as a separate cluster and therefore the number of clusters is equal to the number of observations. In the next steps of the algorithm, similar clusters are merged until they become a single cluster or provide the desired properties. In the merging process, both the distances and the positions of the clusters on the map are taken into account. In the calibration plot formed by merging the neurons, clusters located on the map are adjacent. However, depending on the distributions of input variables used in the analysis, there may be differences in the adjacencies of the clusters. In such a case, negativities may arise regarding the homogeneity of the clusters. Hierarchical clustering method is used to create adjacent clusters, and to merge similar and close neurons on the map (Kohonen, 2001). In this context, Hierarchical Agglomerative Clustering Method is the most commonly used method in the literature. This step, in which SOM and bulk hierarchical clustering method are used together, is called "Calibration" stage. In the calibration stage, a visual of the compressed representation of the distribution of neuron classes generated by SOM, and the information on the reliability of the neuron classes in homogeneous areas are obtained. Clustering results are visualized using the drawing function of hierarchical agglomerative clustering method, and the boundaries (shown by bold and black lines) are determined in a statistically accurate way. Regarding Figure 9,

there are clusters in four different colors: light blue, light green, dark blue and dark green, therefore the ideal number of clusters was determined as 4. Of these, the number of units in the blue cluster is the highest. Besides, the number of elements in the light green cluster is the lowest.

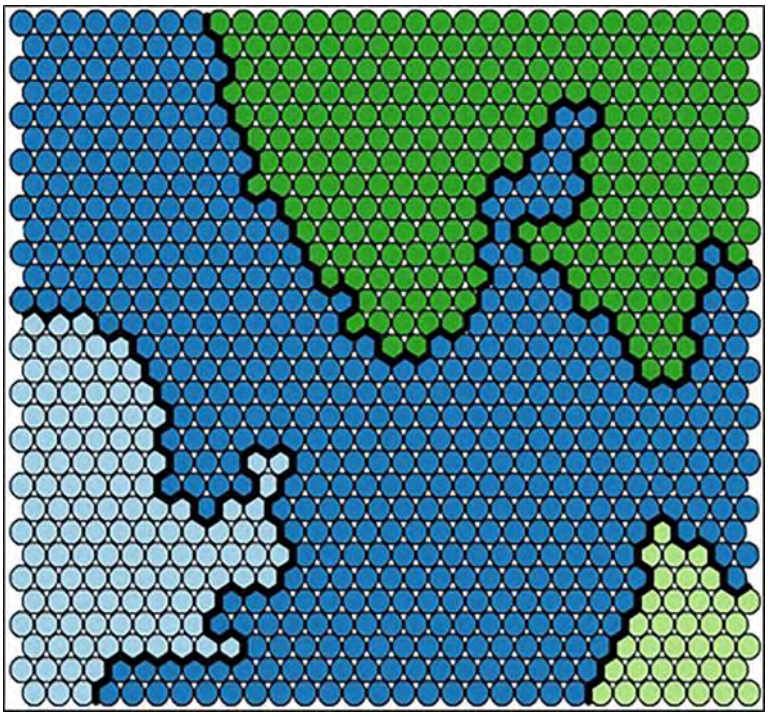

Figure 9. Calibration plot for clusters

In addition, the R program provides column charts showing the distribution of the students in these 4 clusters for each country. Student distribution of the countries is shown in Figure 10. Regarding Figure 10, the 3rd cluster, shown in dark blue, has the highest number of students in all countries, followed by 4th cluster shown in dark green, 1st cluster shown in light blue is in the third position and the 2nd cluster shown in light green in the fourth position. In other words, the ratio of the students falling in the clusters is Cluster 3, Cluster 4, Cluster 1 and Cluster 2 in descending order. Regarding the clusters, Canada, England, Italy, Belgium and New Zealand are observed to be the countries contributing to the third cluster with the highest number of students, which is the largest cluster in terms of students and shown in dark blue. Mexico, Poland, Ireland, Sweden, Greece and America are observed to be the countries contributing to the fourth cluster with the highest number of students, which is the second largest cluster in terms of students and shown in dark green.

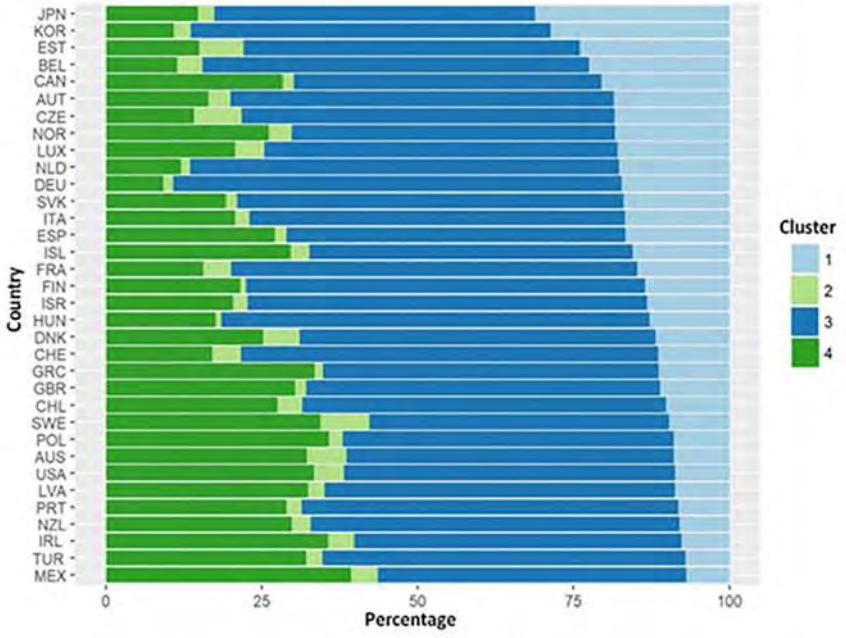


Figure 10. Distribution of the number of students by countries and clusters

Provide dates defining the periods of recruitment and follow-up and the primary sources of the potential subjects, where appropriate. If these dates differ by group, provide the values for each group.

## 4. Discussion

In this study, clustering results obtained from science instruction factor scores and PISA science literacy average scores of 34 OECD member countries, which participated in the PISA 2015 exam and determined by systematic sampling method, were analyzed using SOM. SOM allows to determine the correlations and trends between input variables in educational data mining. Using this method, the comparisons between input variables can be performed faster than other clustering methods; in addition, the resulting outputs provide more detailed information in terms of comparison of the variables.

It is easy to check the existence of theoretical models and observe linear correlations between variables through the colors on the self-organizing maps. At the same time, the maps help in finding nonlinear correlations between different factors, which are very difficult to predict but affect the data set. In this sense, SOM can be used to test theoretical assumptions (Thuneberg & Hotulainen, 2006). In addition, the maps provide the researcher the opportunity to manipulate the data set so that the dimensions of the study can be seen more clearly. By using SOM in educational data mining, hidden characteristics of the sub-populations can be revealed, and specific groups can be identified. The identification of the hidden characteristics allows to carry out more

detailed analysis in the following stages. With this method, it is possible to have an idea about the common characteristics of the groups in the sample that constitute the data set.

SOM is a method that can provide practical benefits for the administrators and teachers in the education sector. Within the scope of the first research question, Using SOM, the ideal number of clusters is found to be four. Approximately 60% of the students are in the third cluster, 21% are in the fourth cluster, 15% are in the first cluster and 3% are in the second cluster. According to this result, the majority of students are in the third cluster, and the number of students in the second cluster is very low.

Within the scope of the second research question, according to the areas covered by the colors close to red and red in the heatmaps, one of the outputs of SOM, the most effective factors in clustering (variables) were found to be as follows: perceived feedback from science teachers, teacher-directed science instruction, average probable science achievement score, inquiry-based science instruction and adaptive instruction of science lessons. In other words, the most effective factor in clustering students is perceived feedback, and the least effective factor is adaptive instruction. Considering the vast majority of countries included in PISA 2015, it is striking that science teachers working in these countries are more tolerant of students' individual differences and tend to pay more attention to individual needs. At this point, feedback to science plays a huge role in teaching (Lipko-Speed, Dunlosky & Rawson, 2014). Forbes, Neumann & Schiepe-Tiska (2020) examined the relationship between science achievement and science instruction in 13 countries that participated in PISA 2015, and in terms of the effectiveness of teaching practices for the most successful student group in a sample of 13 countries, perceived feedback have reached the conclusion that it can be effective at the level of based education.

By using SOM in educational data mining, the variables related to school, classroom and student concepts can be analyzed easily. SOM can provide a clearer understanding of the factors affecting student achievement and the maps obtained by this method can contribute to the development of communication between groups such as school administrators, teachers and policy makers (Thuneberg & Hotulainen, 2006; Taniguchi et al., 2018). By using SOM as an educational tool, the relationships between the factors that are typical for a particular subgroup or for a particular setting, can be determined. SOM may allow to determine the students who are at risk of maladaptation according to any variable and even environmental factors related to the behavior of these students can be specified. Such an application provides an opportunity for both the teacher and the administrator to shape the educational process and to regulate the environmental factors related to the learning environment. In addition of being unique in many ways compared to other methods, SOM can also be used for validating the results obtained from different statistical methods. SOM can be used especially for the validity of the

analyzes involving dimension reduction including other clustering methods, principal component analysis, and factor analysis.

There is a need for self-organized maps in cluster analysis. Because these maps offer unique features in concretizing abstract features (Bagan et al., 2005; Thuneberg & Hotulainen, 2006). Regarding the secondary objective of the study, SOM provided information about the importance of the input variables in the formation of the clusters and student clustering according to their science literacy levels.

## 5. Conclusions

As a result of SOM, four cluster profiles were determined; the students with the highest possible science achievement were found to adopt inquiry-based science instruction. For the future, it is suggested to conduct studies to reveal the reason of these contradictory findings. Following the third cluster of SOM, which includes the students with highest science achievement, the cluster that includes students with the second highest science achievement is the first cluster. The students in the first cluster were found to adopt teacher-directed science instruction. Based on this result, it can be suggested to focus on teacher-directed science instruction in schools that want to increase science achievement.

Researchers who want to clearly determine the effectiveness of the input variables in cluster analysis can be advised to use SOM. Regarding the last results obtained from SOM, silhouette coefficients and calibration plots were found to be quite comprehensible in determining the ideal number of clusters. Considering this, researchers are advised to use SOM as a clustering method.

**Acknowledgements**

This article was produced from the doctoral dissertation entitled "Examination of the Program for International Student Assessment 2015 Data by Clustering Methods in Data Mining" written by Mehmet Taha ESER.

## References


Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Sacele Assessments in Education*, *1*, 5. **https://doi.org/10.1186/2196-0739-1-5.**

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. telematics and ınformatics, *37*, 13-49. **https://doi.org/10.1016/j.tele.2019.01.007.**

ALMazroui, Y. A. (2013). A survey of data mining in the context of e-learning. *International Journal of Information Technology & Computer Science*, *7*(3), 8-10.

Amari, S. (1980). Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, *42*, 339-364. **https://doi.org/10.1007/BF02460791.**

Anderson, B. (1999). Kohonen neural networks and language. Brain and Language, *70*(1):86–94. **https://doi.org/10.1006/brln.1999.2145.**

Anderberg, M. (1973). *Cluster analysis for applications*. New York: Academic Press.

Bagan, H., Wang, Q., Watanabe, M., Yang, Y., & Ma, J. (2005). Land cover classification from MODIS EVI times-series data using SOM neural network. *International Journal of Remote Sensing*, *26*(22), 4999-5012. **https://doi.org/10.1080/01431160500206650.**

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17. **https://doi.org/10.5281/zenodo.3554657.**

Baker, R. S., Martin, T., & Rossi, L. M. (2017). Educational data mining and learning analytics. In A. A. Rupp & J. P. Leighton (Eds.), The handbook of cognition and assessment: Frameworks, methodologies, and applications (pp. 379-396). Oxford, UK: John Wiley & Sons, Inc.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, *1*(3), 295-311. **https://doi.org/10.1162/neco.1989.1.3.295.**

Becker, S., & Plumbley, M. (1996). Unsupervised neural network learning procedures for feature extraction and classification. *International Journal of Applied Intelligence*, *6*, 185-203. **https://doi.org/10.1007/BF00126625.**

Bodner, T. E. (2008). What improves with increased missing data ımputations? *Structural Equation Modeling, 15*(4), 651–75. **http://dx.doi.org/10.1080/10705510802339072.**

Forbes, C. T., Neumann, K., & Schiepe-Tiska, A. (2020). Patterns of inquiry-based science instruction and student science achievement in PISA 2015. *International Journal of Science Education*. *42*(5), 783-806. **https://doi.org/10.1080/09500693.2020.1730017.**

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. USA: John Wiley&Sons.

Dunham, M. H. (2003). *Data mining ıntroductory and advanced topics*. USA: Prentice Hall.

Ehmke, T., Van Den Ham, A-K., Sälzer, C., Heine, J., & Prenzel, M. (2020). Measuring mathematics competence in international and national large scale assessments: Linking PISA and the national educational panel study in Germany. *Studies in Educational Evaluation*, *65*, [100847]. **https://doi.org/10.1016/j.stueduc.2020.100847**

Garson, D. G. (2014). *Cluser analysis (Statistical Associates Blue Book Series)*. Ashebero: Statistical Associates Publishing. Kindle edition.

Fiedler, J. A., Mcdonald J. J. (1993). Market figmentation: Clustering on factor scores versus individual variables. *AMA Advanced Research Techniques Forum*.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576. **https://doi.org/10.1146/annurev.psych.58.110405.085530.**

Gurney, K. (1997). *An ıntroduction to neural network*. London: UCL Press Limited.

Haykin, S. (2009). *Neural networks and learning machines*. New Jersey: Prentice Hall.

Jain, A. K. ve Dubes, R. C. (1998). *Algorithms for clustering data*. New Jersey: Prentice Hall.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, *41*, 51–80. **https://doi.org/10.3102/1076998615622221.**

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. *Neural Computing Surveys*, *1*, 102-350. **https://doi.org/10.1007/978-3-540-68860-0_15.**

Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kiang, M. Y. (2001). Extending the kohonen self-organizing map networks for clustering analysis. *Computational Statistics&Data Analysis*, *38*, 161-180. **https://doi.org/10.1016/S0167-9473(01)00040-8.**

Kiang, M., & Kumar, A. (2001). An evaluation of self-organizing map networks as a robust alternative to factor analysis in data mining applications. *Information Systems Research, 12*(2), 177-194. **https://doi.org/10.1287/isre.12.2.177.9696.**

Kohonen, T. (2014). *MATLAB Implementations and applications of the self-organizing map*. Helsinki: Unigrafia Oy.

Kohonen T. (1984). *Self-organization and associative memory*. Berlin: Springer.

Kohonen, T. (2001). *Self-organizing maps*. Berlin: Springer-Verlagn.

Kruskal, J. B. & Wish, M. (1978). *Multidimensional scaling*. Newbury Park: Sage Publications.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and k-means algorithm for market segmentation. *Computers&Operations Research*, *29*(11), 1475-1493. **https://doi.org/10.1016/S0305-0548(01)00043-0.**

Langdon, D., Mckittrick, G., Beede, D., Khan, B., & Doms, M. (2011). STEM: Good jobs now and for the future, U.S. *Department of Commerce Economics and Statistics Administration*, 3(11), 2.

Lin, J.-W., Yen, M.-H., Liang, J., Chiu, M.-H., & Guo, C.-J. (2016). Examining the factors that ınfluence students' science learning processes and their learning outcomes: 30 years of conceptual change research. *Eurasia Journal of Mathematics, Science and Technology Education, 12*(9), 2617-2646. **https://doi.org/10.12973/eurasia.2016.000600a.**

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*(3), 171–176. doi:**10.1016/j.jarmac.2014.04.002.**

Little. R., & Rubin. D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lupaşcu C. A., & Tegolo D. (2011). *Automatic unsupervised segmentation of retinal vessels using self-organizing maps and k-means clustering*. Heidelberg: Springer.

Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14*, 85–100. **https://doi.org/10.1007/BF00288907.**

Milligan, G .W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45***,** 325–342 **https://doi.org/10.1007/BF02293907.**

Monette, D. R., Sullivan, T., & De Jong, C. R. (1990). *Applied Social Research*. New York: Harcourt Broce Jovanovich, Inc.

Murtagh, F., & Hernández-Pajares, M. (1995). The Kohonen Self-Organizing Map Method: An Assessment. *Journal of Classification*, *12*(2), 165-190. **https://doi.org/10.1007/BF03040854.**

Nielsen, S. E., & Yezierski, E. J. (2016). Beyond academic tracking: using cluster analysis and self-organizing maps to investigate secondary students' chemistry self-concept. *Chemistry Education Research and Practice* , *17* (4), 711-722. https://doi.org/10.1039/C6RP00058D.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, 9, 2231.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert System with Applications, *41*(4), 1432-1462. **http://dx.doi.org/10.1016/j.eswa.2013.08.042**

Penn, B. S. (2005). Using self-organizing maps to visualize highdimensional data. *Computers & Geosciences*, *31*(5), 531-544. **https://doi.org/10.1016/j.cageo.2004.10.009.**

Roeben, E. D. (1997). The lechnical and practical challcnges in developing innovalivc assessment approaches for use in statesvide assessment programs. Contemporary Education, *69*(1), 6- 10.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley ınter disciplinary reviews. *Data Mining and Knowledge Discovery*, *3*(1), 12–27. **https://doi.org/10.1002/widm.1075.**

Schreck, T., Bernard, J., von Landesberger, T., & Kohlhammer, J. (2009). Visual cluster analysis of trajectory data with ınteractive kohonen maps. *Information Visualization*, *8*, 14-29. **https://doi.org/10.1057/ivs.2008.29.**

Taber, K. S. (2009). Progressing science education: Constructing the scientific research programme into the contingent nature of learning science. The Netherlands: Springer.

Taniguchi, T., Maruyama, Y., Kurita, D., & Tanaka, M. (2018). Self-organizing map analysis of educational skills using questionnaire to university students in computing classes. In: Proceedings of 15th International Conference Cognition and Exploratory Learning in Digital Age, pp. 103–110.

Thuneberg, H., & Hotulainen, R. (2007). Contributions of data mining for psycho-educational research: what self-organizing maps tell us about the well-being of gifted learners, *High Ability Studies*, *17*(1), 87-100. **https://doi.org/10.1080/13598130600947150.**

Vedder-Weiss, D., & Fortus, D. (2012). Adolescents' declining motivation to learn science: A follow up study. *Journal of Research in Science Teaching*, *49*(9), 1057–1095. **https://doi.org/10.1002/tea.21049.**

Wehrens, R., & Buydens, L. M. C. (2007). "Self- and Super-Organizing Maps in R: The kohonen Package." _Journal of Statistical Software_, *21*(5), 1-19. doi: 10.18637/jss.v021.i05 (URL: **https://doi.org/10.18637/jss.v021.i05).**

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

Wu, J. (2012). *Advances in k-mean clustering: A data mining thinking*. Hedilberg: Springer Science & Business Media.

Wu, S., & Chow, T. W. S. (2003). Self-organizing-map based clustering using a local clustering validity index. *Neural Processing Letters*, *17***,** 253–271. **https://doi.org/10.1023/A:1026083612746.**

## Appendix A. Information about items

| Dimensions | Item Code | Items |
|---|---|---|
| Teacher-Directed Science Instruction | ST103Q01NA | The teacher explains scientific ideas |
| | ST103Q03NA | A whole class discussion takes place with the teacher |
| | ST103Q08NA | The teacher discusses our questions |
| | ST103Q11NA | The teacher demonstrates an idea |
| Perceived feedback from science teachers | ST104Q01NA | The teacher tells me how am I performing in this course |
| | ST104Q02NA | The teacher gives me feedback on my strengths in this class |
| | ST104Q03NA | The teacher tells me in which areas I can still improve |
| | ST104Q04NA | The teacher tells me how I can improve my performance |
| | ST104Q05NA | The teacher advises me on how to reach my learning goals |
| Adaptive instruction in science lessons | ST107Q01NA | The teacher adapts the lesson to my class's needs and knowledge |
| | ST107Q02NA | The teacher provides individual help when a student has difficulties understanding a topic or task |
| | ST107Q03NA | The teacher changes the structure of the lesson on a topic that most students find difficult to understand |
| Enquiry based science instruction | ST098Q01TA | Students are given opportunities to explain their ideas |
| | ST098Q02TA | Students spend time in the laboratory doing practical experiments |
| | ST098Q03NA | Students are required to argue about science questions |
| | ST098Q05TA | Students are asked to draw conclusions from an experiment they have conducted |
| | ST098Q06TA | The teacher explains how a science idea can be applied to a number of different phenomena |
| | ST098Q07TA | Students are allowed to design their own experiments |
| | ST098Q08NA | There is a class debate about investigations |
| | ST098Q09TA | The teacher clearly explains the relevance of science concepts to our lives |
| | ST098Q10NA | Students are asked to do an investigation to test ideas |

## Appendix B. Country codes

| Code | Country Name | Code | Country Name |
|---|---|---|---|
| JPN | Japan | ISR | Israel |
| KOR | Korea | HUN | Hungary |
| EST | Esthonia | DNK | Denmark |
| BEL | Belgium | CHE | Switzerland |
| CAN | Canada | GRC | Greece |
| AUT | Australia | GBR | Great Britain |
| CZE | Czech Republic | CHL | Chile |
| NOR | Norway | SWE | Sweden |
| LUX | Luxemburg | POL | Poland |
| NLD | Netherlands | AUS | Australia |
| DEU | Germany | USA | America |
| SVK | Slovakia | LVA | Lithuania |
| ITA | Italy | PRT | Portugal |
| ESP | Spain | NZL | New Zealand |
| ISL | Iceland | IRL | İreland |
| FRA | France | TUR | Turkey |
| FIN | Finland | MEX | Mexico |