ISSN 2614-0047

Job applicants clustering using self-organizing map

Arum Handini Primandari [1,*], Nur Aini Ikasakti [2]

Department of Statistics, Islamic University of Indonesia, Indonesia
[1] primandari.arum@uii.ac.id *; [2] 213611135@student.uii.ac.id
* corresponding author

ARTICLE INFO

Article history
Received October 2, 2017
Revised October 29, 2017
Accepted November 10, 2017

Keywords
Clustering
Mapping
Self-organizing maps

ABSTRACT

Yogyakarta Government through Directorate of Manpower and Transmigration (Disnakertrans) have been canvassing people looking for job. An employment program was provided by Disnakertrans to allow job applicants meet companies. This research was carried out to identify educational background of applicants, in order to obtain the suitable worker. One of the ways to identify educational background is by district clustering in Yogyakarta. Clustering method is employed to reveal the characteristic of educational quality in every district in Yogyakarta. Clustering is a grouping method which is done by minimalize the characteristic among class members and minimalize the characteristic among clusters. This research used Self Organizing Maps to grouping districts in Yogyakarta according to educational background of its job seekers. The clustering results 3 clusters: 6 districts belong to cluster 1, 4 districts belong to cluster 2, and 4 districts belong to cluster 3. Then, Yogyakarta map is used to visualize the result of district clustering.

1. Introduction

Result of Sakernas (Survei Angkatan Kerja Nasional) conducted by Badan Pusat Statistik (BPS) recorded unemployment rate or TPT (Tingkat Pengangguran Terbuka) in Indonesia in August 2014 reached 3.33% of the total workforce in Yogyakarta. This figure rose significantly in August 2015 reached 4.07% [1]. After the economic crisis in mid-1997, unemployment rate in Indonesia reached over 5% from 1998 to 2004. Even in the period from 1999 to 2002, unemployment rate has reached over 8% every year.

The unemployment problem in Indonesia as a developing country into a particular concern because it resulted in the emergence of problems in other areas such as economic, social, and political [2]. In the economic field, unemployment means unproductiveness a resource that will result in the increase of the number of people as the divisor in each capita income. The larger the denominator, then each capita income, as one indicator of economic development, is getting smaller. In the social sector, the increase in unemployment causes insecurity increase in social ills such as crime. As the impact of the imbalance in economic and social fields, the political situation becomes conducive.

Unemployment caused by the inability of jobs to absorb the labor force [3]. In other words, employment opportunities are much smaller than the labor force. According to a survey from the Institute of Management Development IMD World Competitiveness 2015, Indonesia was ranked 41, down 16 ranking of 2014. The position of Indonesia is far from neighboring countries such as Malaysia and Singapore, and even Thailand [4]. Factors affecting the development and assessment that is factor of investment, attractiveness factors of a country, and the readiness factors of human resources. One of the factors that led to Indonesia down the ranks is the third factor is the readiness of human resources is the most dominant figure accounts for skilled labor downgrade Indonesia in 2015.

Based on the exposure to the experts above, it can be concluded that the factors causing the increase in the number of unemployment in Indonesia covered lack of employment opportunities and quality of labor.

To overcome these problems, Yogyakarta Government through Directorate of Manpower and Transmigration (Disnakertrans) seeks employment opportunities for job seekers through a program of labor market information with the identification of employment (Job Canvassing). These programs bring together between job seekers with labor users who are in need of manpower making it easier for job seekers to obtain complete information on job vacancies. In addition, job seekers can choose or define their own desired job in accordance with the education and skills they have.

However, the efforts could not overcome the number of job seekers continues to increase every year. Because each time there will be students who graduated from the school and then become job seekers. Based on these problems, the author will discuss about whether the lack of success of efforts to reduce job seekers registered with the Disnakertrans Yogyakarta city caused by a lack of job opportunities or education quality job seekers. To find the right solution to the problem, the author uses the analysis of Clustering SOM to be visualized with a map.

2. Problem Formulation

Based on the background as to whether the lack of success of efforts to reduce the number of job seekers registered with the Disnakertrans Yogyakarta city caused by a lack of job opportunities or education quality search work.

To find the right solution to the problem, the authors use the SOM Clustering analysis to be visualized with a map, then the problem in this research is how the quality of education on unemployment in the city of Yogyakarta in 2015 and how to perform clustering (grouping) on unemployment in the city of Yogyakarta in 2015 by education level.

3. Method

3.1. Self-Organizing Maps (SOM)

Kohonen Self Organizing Maps is a network that was discovered by Teuvo Kohonen network is one of the most widely used. Named "self-organizing" because this method does not need a special surveillance and SOM competitive approach followed by unsupervised probation [5]. The word "maps" itself because this method uses the map in the weighting of input data. Each node in this network works presented each input data, therefore the network can also be called "Self Organizing Feature Maps", the concept of "features" into something important and valuable, in specific topological relations between inputted data will remain intact and original when mapped in a SOM network [6].

Kohonen self-contained within the SOM two most important characteristics of this network which explains that the SOM is a device for data visualization and analysis of high-dimensional [7]. However, the network is able to be used also for clustering, dimensionality reduction, classification, vector quantization and data mining [8]. In perspective, SOM can be seen not just as a tool but as a toolbox containing features numbers and make it more attractive in different situations.

The work can be done by SOM among other groupings, in the context of Clustering, SOM can be used as a grouping alternative to K Means. Knowledgeable amount SOM Cluster will divide the available data into different groups. The main advantage of the SOM are less likely to get results than using a branching K Means algorithm, and can be used as a good initialization algorithm for K Means method. In fact, the SOM can be substituted with the same K Means and the SOM algorithm produces the same algorithm with K Means. Other advantages of the SOM algorithm are to obtain a sequence which typically Cluster topologically similar spliced together.

Kohonen network is used to divide the data into a high-dimensional pattern with dimensions lower. Data shown to have a relationship with the topology of the original data, thus, a pattern that is composed can visualize the results of the training to see the data, for example, the structure of the Cluster. Suppose the input of vector of n components to be grouped in a maximum of m pieces of the group. Exodus networks are among the most close / similar to a given input.

Weight vectors example serves as a determinant of the sample vector proximity to a given input. During the setup process, the vector example at the time closest to the input will emerge as the winner. Vector winner (and vicinity vectors) will be modified weights.

The Clustering algorithm is the Kohonen network patterns with initializing the form of weights (Wij) obtained randomly for each node. After weight (Wij) is given then select an input sample (xi). Once the input received by the network, and then calculating the Euclidean distance vector Dj(x) is obtained by summing the difference between the weight vector(Wij) with the input vector (xi).

$$D_j(x) = \Sigma(w_{ij} - x_{ij}) \quad (1)$$

In addition to the distance between nodes are known then the specified minimum value of the calculation of distance vector Dj(x),then the next step to change the weights.

$$W_{ij}(new) = w_{ij}(old) + a[x_i - w_{ij}(old)] \quad (2)$$

In the process of getting new weight requires a value of learning rate (α) is $0 \leq \alpha \leq 1$. The value of learning rate for each epoch would be reduced to.

$$a(i+1) - 0.5a \quad (3)$$

The termination condition testing is done by calculating the difference between the weights Wij (new) with Wij (old), if the value Wij only changed slightly, testing means have reached convergence so that it can be stopped [9].

Table 1. Unemployment in the city of Yogyakarta

District	No school	SD	SMP	SMA	SMK	DI /DII /DIII	S1	S2/ S3
MANTRIJERON	8	57	112	148	155	18	36	0
KRATON	4	43	108	392	73	10	44	0
MERGANGSAN	15	117	203	369	159	17	52	1
UMBULHARJO	27	147	303	483	262	35	79	0
KOTAGEDE	19	100	212	418	260	39	49	2
GONDOKUSUMAN	14	114	205	347	312	42	43	1
DANUREJAN	7	54	140	269	119	12	23	0
PAKUALAMAN	3	32	62	194	74	5	17	0
GONDOMANAN	17	96	121	149	68	8	3	0
NGAMPILAN	5	50	92	134	56	4	16	1
WIROBRAJAN	35	114	201	295	170	23	20	1
GEDONGTE-NGEN	20	98	170	198	205	17	20	0
JETIS	22	129	184	228	202	19	21	1
TEGALREJO	15	130	226	308	260	26	35	1
Total	211	1281	2339	3932	2375	275	458	8

SOM itself can be considered as a spatial form of the K Means Cluster analysis. The analogy, each unit in accordance with a Cluster and the Cluster number is determined by the size of the grid which is usually arranged in a square or hexagonal shape. SOM grid use in the mapping process. So when the two-dimensional objects are very similar, then the position in the mapping will be very close together. This algorithm is more concentrated on the biggest similarity [5].

The results of Clustering using the SOM class in addition to producing the appropriate criteria also has an output in the form of fan charts. When interpreting output fan diagram including subjective because it depends associate researcher in color. For example if a variable is used more dominant, it

can be associated to a group 1. Next group / class is also divided into several corresponding circle determined matrix multiplication. The variables that exist within one characteristic will have a circle of the same color. Then for the outcome of the class can be directly mapped with the help of other software. Later the difference between these classes can be distinguished by their color.

3.2. Research Method

Analysis Cluster method SOMS (Self-Organizing Maps) that will be used to see a breakdown of the number of unemployed based on the characteristics of each cluster, as well as the mapping for the cluster is formed. Software used in this analysis is the R version 3.1.2 and Q GIS.

Fig. 1. Flowchart of research methodology

4. Analysis

4.1. Descriptive Statistics

Fig. 2, shows the percentage of unemployed from each level of education for all districts.

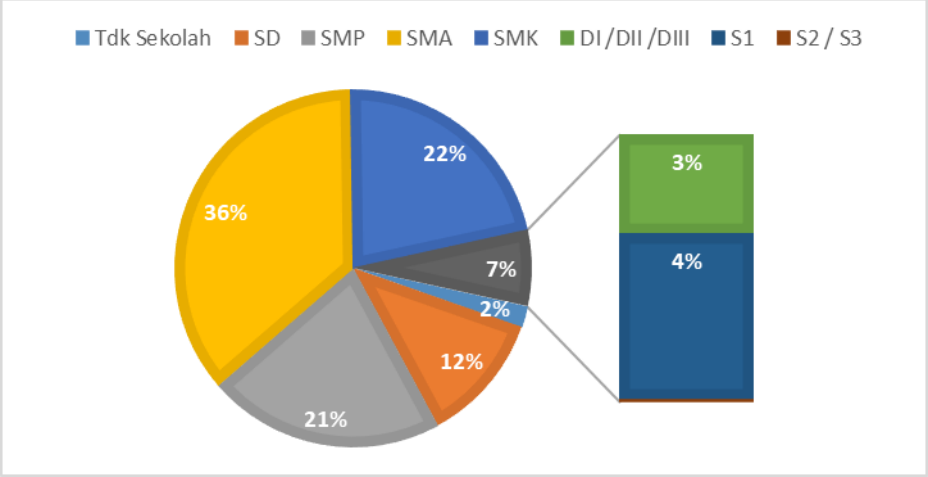

Fig. 2. Percentage unemployment judging from education level

Workforcein the category of the unemployed can be differentiated according to the level of education. The percentage of unemployed people have a high school education down, if specified at the most are those who have graduated from high school is as much as 36% (3,932 people). Next up was SMK graduates by 22% (2,375 people), SMP by 21% (2,339 people), and primary school by 12% (1,281 people). Based on Fig. 2, it was concluded that unemployment in the city of Yogyakarta majority of low-educated are generally only had high school down.

Unemployment is higher education that have graduated DI / DII / DIII and S1 and S2 / S3 reaches 741 people. That number suggests that people who have higher education is not necessarily accepted by the labor market. It is highly related to limited employment opportunities to absorb them. In addition, the number of job seekers is also abundant so the level of competition to be able to get a job to be very tight.

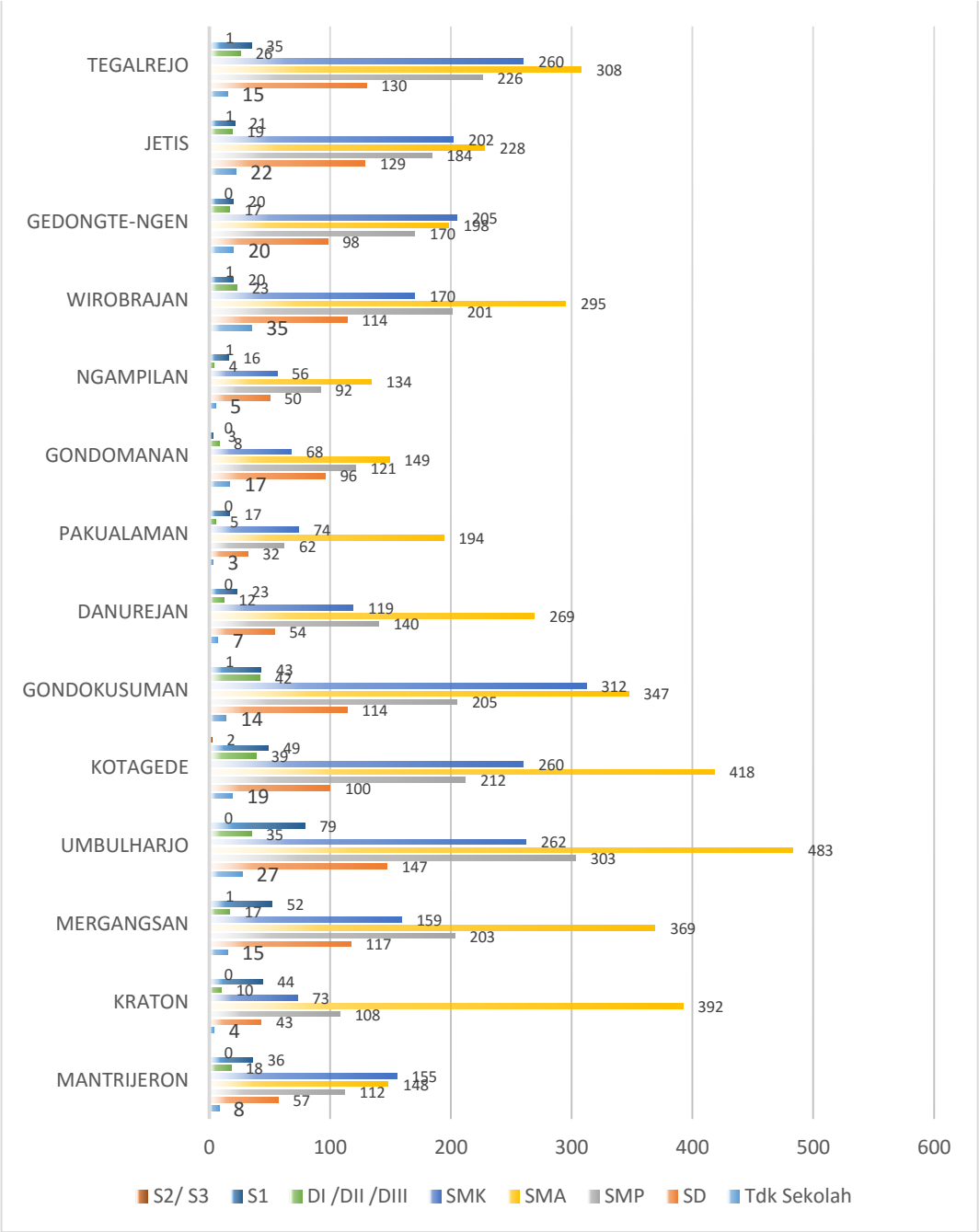

Fig. 3. Graph unemployment based on the number of districts in Yogyakarta

Based on the number of districts located in the city of Yogyakarta, on the level of education of the school not the lowest unemployment rate in the Pakualaman district which amounted to 3 people and the highest in Wirobrajan district which amounted to 35 people. For this level of education above the elementary school graduate unemployment is the lowest in Pakualaman district which amounted to 32 people and the highest in Umbulharjo district which amounted to 147 people. For this level of education graduated from junior high/equal lowest unemployment rate in the Pakualaman district is 62 people and the highest in Umbulharjo district which amounted to 303 people. For graduating high school education level of the lowest unemployment rate in the Ngampilan district which amounted to 134 inhabitants and the highest in Umbulharjo district which amounted to 483 people. Tertiary education for vocational school graduate lowest unemployment rate in the Ngampilan district amounted to 56 people and the highest in Gondokusuman district which amounted to 312 people. Tertiary education for graduate D1 and D2 / D3 lowest unemployment rate in the Ngampilan district which amounted to 4 people and the highest in Gondokusuman district which amounted to 42 people. For the highest level of education completed S1 lowest unemployment rate in Gondomanan district

which amounted to 3 people and the highest in Umbulharjo district which amounted to 79 people and graduated S2 / S3 highest unemployment rate in the Kotagede district which amounted to 2 people. Meanwhile, measures of statistical descriptive of unemployment in the city of Yogyakarta is shown in Table 2.

Table 2. Unemployment DIY central tendency

Measure	No School	SD	SMP	SMA	SMK	DI /DII /DIII	S1	S2/ S3
mean	15.07143	91.5	167.0714	280.8571	169.6429	19.64286	32.71429	0.571429
median	15	99	177	282	164.5	17.5	29	0.5

If payed of the measure of the centralization of data, between the mean values and the media is almost the same. It thus becomes an indicator that the unemployment data has no outliers.

4.2. Clustering Process

Determination of the number of clusters that classifying the city of Yogyakarta has been done by the government of Yogyakarta using geographical factors and divide Indonesia. In addition to referring to the number of groups that have been established, the researchers also used the approach Within Cluster Sum of Squares (WCSS) in determining number of the clusters [10]. WCSS is the distance between elements within the Cluster.

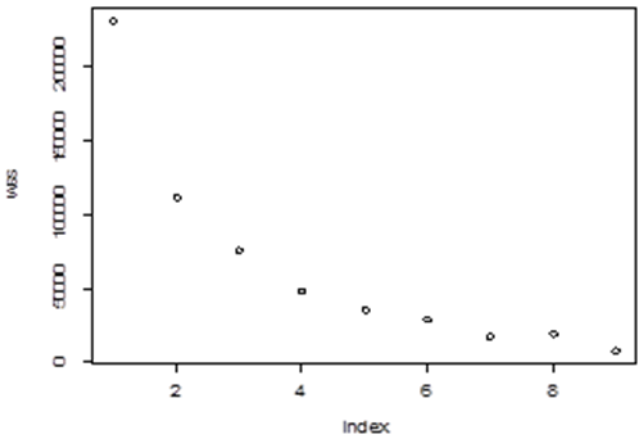

Fig. 4. Within cluster sum of squares (WCSS)

Based on the picture, when the point cluster index number 3 represents the movement that began ramps do not like change point cluster in previous index number that illustrate the change is quite steep. If you use multiple clusters 3, then the distance between elements in the cluster will not vary much if you use multiple cluster 4. Meanwhile, if you use multiple cluster goes higher, then clustering will be ineffective. This is because the number of districts will be grouped only 13 districts. After approaching the WCSS then, found the number of clusters as many as three classes, further implemented to methods Self Organizing Maps.

At algorithm Self Organizing Maps takes itersi to get the best grouping. Fig. 4. explaining the many training progress that shows the number of iterations and the impact on the average distance to the closest unit is getting smaller. It can be seen that the indicate iteration convergence began when iterating to 400.

Based on the graph in Fig. 5, it can be seen that in training progress has been made as much as 1000 iterations and produced a mean of distance to closest unit (average distance each unit Cluster) fewer than 4. It can be concluded that when researchers conducted iterations more and more so, the mean of distance cluster units are getting smaller and results the clustering will be better. After passing iterations to 400 shows that training progress is beginning to stabilize with a mean of distance cluster units fewer than 4.

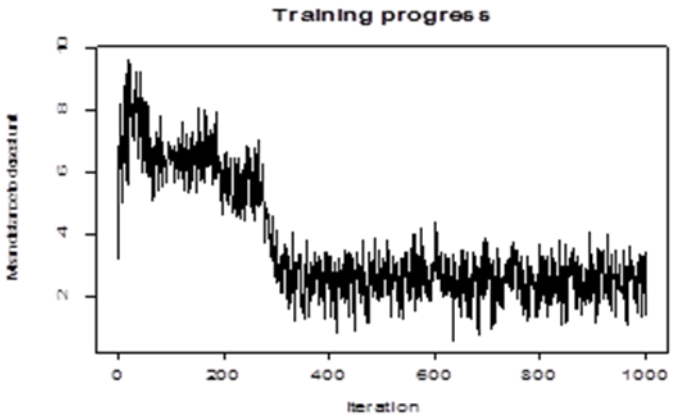

Fig. 5. Graph training progress

Process SOM algorithm produces a model of SOM and in the process using R will produce a diagram that contains several circles (circle) which topology will be adjacent if their characteristics same.

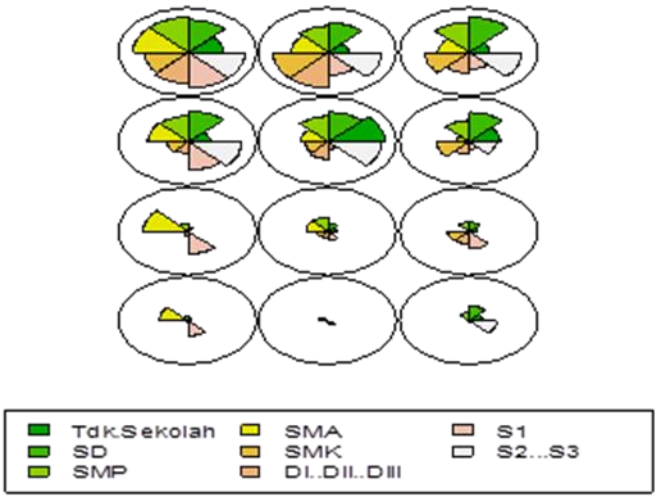

Fig. 6. Venn diagram kohonen

Based on the Fig. 6, algorithm can be seen authors make of fan use charts display rectangular with of grid 3 x 4. Diagram is formed based on the results of data if the Kohonen algorithm using eight variables. Once formed fan diagram can be known depiction and staining for each of variables: No School by dark green, graduate from elementary school is green, junior high school graduate was given a green, senior high school graduate is yellow, SMK graduate is orange, D1, D2 / D3 graduate is orange, S1 graduate is pink, and S2 / S3 graduate is white. Fan diagram shows the distribution of the variables on the map. Patterns can be seen by examining the dominant color.

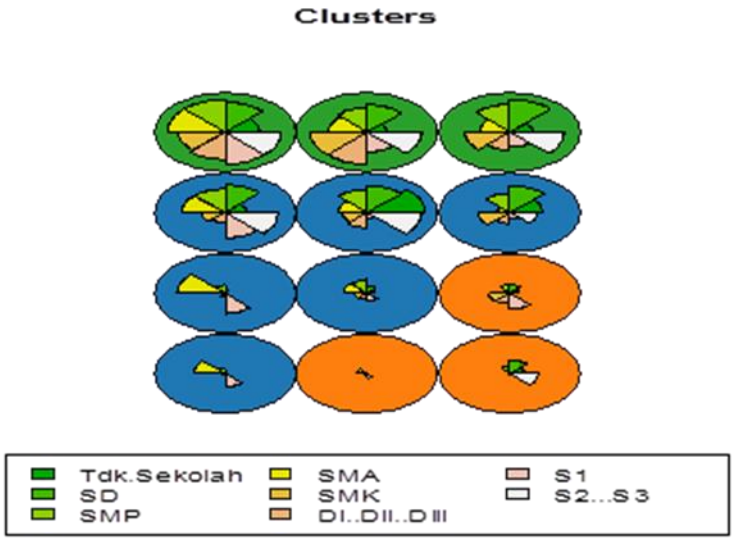

Fig. 7. Venn diagram of results clustering

Based on Fig. 7, viewable models created with Kohonen algorithm is then shaped into 3 clusters with hierarchical cluster method. Each cluster is formed has its own characteristics. Cluster 1 is marked in green, cluster 2 is marked in blue, cluster 3 is marked by the orange color. Fig. 8 shows the characteristics of each cluster:

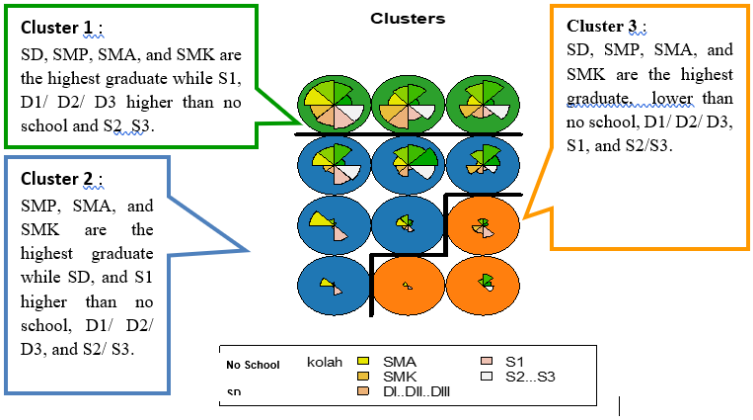

Fig. 8. Characteristics of each cluster

Table 3. Results grouping using self-organizing maps

District	Group
Mantrijeron	2
Kraton	1
Mergangsan	1
Umbulharjo	3
Kotagede	3
Gondokusuman	3
Danurejan	1
Pakualaman	2
Gondomanan	2
Ngampilan	2
Wirobrajan	1
Gedongtengen	1
Jetis	1
Tegalrejo	3

Processes of understanding the diagram of the SOM algorithm is when the diagram has colored and defined by the vectors are visualized in a plot mapping. Based on Figure 8., Obtained information that the circle of orange on the lower right is associated in a group that has a level of education completed primary school, junior high, high school and vocational high, but the level of education is not school, graduated D1 / D2 / D3, S1, and graduated S2 / S3 low. Graduate junior high school, vocational and higher and have completed primary school, and S1 were but the other levels of education no school, graduated D1 / D2 / D3 and S2 / S3 low associated in the blue circle in the top middle. A green circle is associated with a group that has the level of education completed primary school, junior high.

Table 4. Number and class members using self-organizing maps

Group	Number of	Members Group
1	6	Kecamatan Kraton, Mergangsan, Danurejan, Wirobrajan, Gedongtengen, and Jetis
2	4	Kecamatan Mantrijeron, Pakualaman, Gondomanan, and Ngampilan
3	4	Kecamatan Umbulharjo, Kotagede, Gondokusuman, and Tegalrejo

As for the results of the analysis to the classification of mappings using the SOM which members of each group have been known, would have seen his profile by using the average in each group.

Table 5. Calculation of average results clustering SOM

Group	Group 1	Group 2	Group 3
Tdk Sekolah	17.17	8.25	18.75
SD	92.50	58.75	122.75
SMP	167.67	96.75	236.50
SMA	291.83	156.25	389.00
SMK	154.67	88.25	273.50
DI /DII /DIII	16.33	8.75	35.50
S1	30.00	18.00	51.50
S2 / S3	0.50	0.25	1.00

When depicted in the bar chart, then visualization average each variable for each Cluster are shown at Fig. 9.

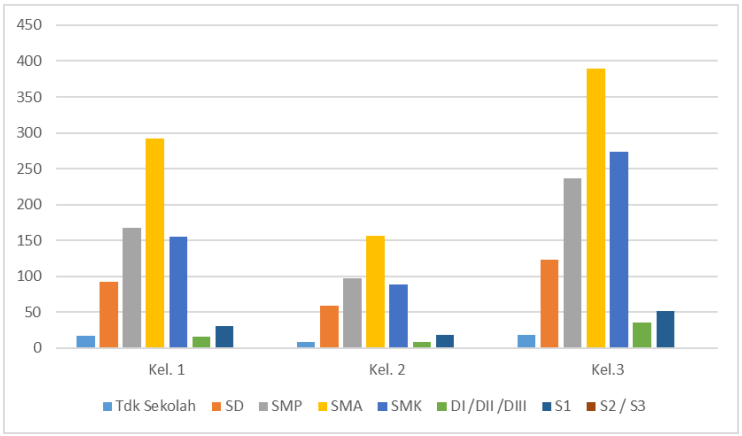

Fig. 9. Average Number of Unemployment According to Education for Every Cluster

When the note, then the group 3 has the number of unemployed from various levels of the education higher than group 1 and group 2. This indicates that the third group consists of the districts that have high unemployment rates. Group 2 is a group with an unemployment rate at various levels of education are the lowest among the three groups. This indicates that the second group consists of the districts which have a relatively low unemployment rate.

4.3. Validation

Selection of the best techniques and results grouping also able to use a validation in this case the researchers are using Cluster Variance. Validation of the SOM is to calculate the variance between members of a group (Sw) and the variance between groups(Sb), for the next available cluster variance.

The variance between members of a group will show better results when the value gets smaller. Meanwhile, the value of variance between groups will show good results when a large value. Values cluster variance which is a division of the variance between members in the group and the variance between groups, where the value of cluster variance would be better if the value is getting smaller. Based on the results obtained validations conducted cluster varianceto equal 1.48 for the method of Self Organizing Maps.

Table 6. Table validation SOM

No School	SD Sekolah	SD	SMP	SMA	SMK	DI/DII/DIII	S1	S2/S3
sw	44,17	274.00	500.92	837.08	516.42	60.58	99.50	1.75
	14.72	91.33	166.97	279.03	172.14	20.19	33.17	0.58
	4.39	1.00	0.35	120.48	224.29	10.95	7.37	0.01
	46.53	8372.25	27912.86	78880.73	28778.70	385.84	1070.22	0.33
	13.53	1190.25	3032.86	17651.02	214.41	2.70	10.80	0.33
	64.45	9563.50	30946.08	96652.23	29217.40	399.49	1088.39	0.66
	32.23	4781.75	15473.04	48326.12	14608.70	199.75	544.19	0.33
Sb	5.68	69.15	124.39	219.83	120.87	14.13	23.33	0.57
Sw/sb	2.59	1.32	1.34	1.27	1.42	1.43	1.42	1.02
Cluster Varianc	1.48							

[a.] Amount of the average of each variable by district.

[b.] On average variable by district reduced the overall average variable squared.

[c.] The sum of average variable in number 2.

[d.] The average of the variables have been added shared with many classes.

4.4. Mapping

The mapping results of clustering analysis using Self Organizing Maps in the Fig. 10. If you see the results table grouping and mapping SOM visually, and see the table group average then the group 1 consisting district of Kraton, Mergangsan, Danurejan, Wirobrajan, Gedongtengen, and Jetis education level SMP, SMA, and SMK graduate are highest, SD and S1 higher than no school, D1 / D2 / D3 and S2 / S3. The group is in accordance with which is associated in the blue circle in the Self Organizing Maps. Group 2 consists district of Mantrijeron, Pakualaman, Gondomanan, and Ngampilan in the blue circle the level of education SMP, SMA, and SMK are the highest graduate while SD, and S1 higher than no school, D1/ D2/ D3, and S2/ S3. The results of profiling in accordance with a circle of orange in the mapping of Self Organizing Maps. Group 3 consists district of Umbulharjo, Kotagede, Gondokusuman, and Tegal level of education SD, SMP, SMA, and SMK are the highest graduate, lower than no school, D1/ D2/ D3, S1, and S2/S3. The group is associated in a green circle in the mapping of Self Organizing Maps.

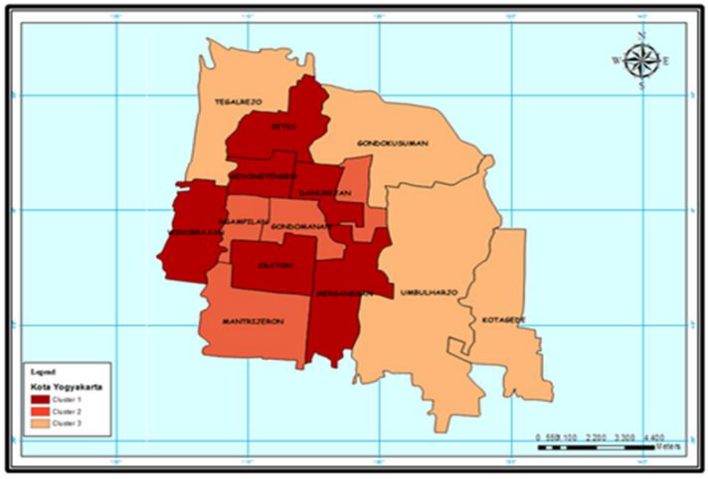

Fig. 10. Mapping algorithm results SOM

5. Conclusion

Based on the analysis and discussion that has been done in the previous chapter, it can be concluded as follows.

1. Unemployment in the city of Yogyakarta, the majority of low-educated are generally only had high school down.
2. The number of groups formed as many as 3 groups were determined by researchers with the approach of using within cluster Sum of Squares.
3. The use of algorithms Self Organizing Maps produces three groups with each group as follows.
   a. Six districts for cluster 1 to the characteristics of the level of education SMP, SMA, and SMK graduate are highest, SD and S1 higher than no school, D1 / D2 / D3 and S2 / S3.
   b. Four districts to cluster 2 has the characteristic graduated from SMP, SMA, and SMK are the highest graduate while SD, and S1 higher than no school, D1/ D2/ D3, and S2/ S3.
   c. Four districts to cluster 3 has the characteristic level of education SD, SMP, SMA, and SMK are the highest graduate, lower than no school, D1/ D2/ D3, S1, and S2/S3.
4. Group 1 consists district of Kraton, Mergangsan, Danurejan, Wirobrajan, Gedongtengen, and Jetis. Group 2 consists district of Mantrijeron, Pakualaman, Gondomanan, and Ngampilan, and group 3 comprised district of Umbulharjo, Kotagede, Gondokusuman, and Tegalrejo.

References

[1] Badan Pusat Statistik Provinsi D.I. Yogyakarta, "Keadaan Ketenagakerjaan Di Daerah Istimewa Yogyakarta Pada Agustus 2014 Tingkat Pengangguran Terbuka Sebesar 3,33 Persen," 2014. [Online]. Available: https://yogyakarta.bps.go.id/pressrelease/2014/11/05/321/keadaan-ketenagakerjaan-di-daerah-istimewa-yogyakarta-pada-agustus-2014-tingkat-pengangguran-terbuka-sebesar-3-33-persen.html. [Accessed: 14-Jul-2016].

[2] D. Suryadarma, A. Suryahadi, and S. Sumarto, "The Measurement and Trends of Unemployment in Indonesia: The Issue of Discouraged Workers," Jakarta, 2005.

[3] A. Kayahan Karakul, "Educating labour force for a green economy and renewable energy jobs in Turkey: A quantitave approach," *Renew. Sustain. Energy Rev.*, vol. 63, pp. 568–578, 2016.

[4] IMD World Competitiveness Center, "IMD world talent report 2015," Lausanne, Switzerland, 2015.

[5] R. Wehrens and L. M. Buydens, "Self-and super-organizing maps in R: the Kohonen package," *J. Stat. Softw.*, vol. 21, no. 5, pp. 1–19, 2007.

[6] S. M. Guthikonda, "Kohonen self-organizing maps." Wittenberg University, 2005.

[7] J. Fort, P. Letremy, M. Cottrell, I. Elie, and U. Nancy, "Advantages and drawbacks of the Batch Kohonen algorithm," *Proc Eur. Symp. Artif. Neural Networks ESANN*, pp. 223–230, 2002.

[8] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. neural networks*, vol. 11, no. 3, pp. 586–600, 2000.

[9] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of artificial neural networks*. Massachusetts: MIT Press, 1997.

[10] I. B. Mohamad and D. Usman, "Standardization and its effects on K-means clustering algorithm," *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 17, pp. 3299–3303, 2013.