Grouping the districts in Indonesia based on value of science subjects of National Exam using K-Means clustering method

R Ferdhiana¹*, T FAbidin², A Mardhiah³
1 Department of Statistics, Syiah Kuala University, Banda Aceh
2 Department of Informatics, Syiah Kuala University, Banda Aceh
3 Department of Statistics, Syiah Kuala University, Banda Aceh

E-mail: ridha.ferdhiana@unsyiah.ac.id

Abstract. Implementation of the National Examination (NE) for senior high schools for each district in Indonesia is done as one way to mapping the success of education in each the district. Mapping can be done by grouping districts based on attribute values obtained on the implementation of the NE, which in this study using the K-means clustering method. The attribute of NE values used are the values for the science majors, which consist of 6 subjects, where from 514 districts only 510 districts own it. The purpose of the study in addition to grouping districts, will also predict the group of 4 districts that have no NE values. Prediction is done by entering the value of the district Human Development Index (HDI) that has no UN value into 95% confidence interval of HDI from groups formed from k-means clustering. Once known where the district is in a group based on the HDI group then the value of its UN will follow the confidence interval of the UN value of the group. The results of this study indicate that the number of clusters (k) selected to classify districts in Indonesia as many as 5 clusters. Group prediction results show that one district belongs to the V cluster, while the other three districts are predicted to be members of the I cluster.

1. Introduction
The National Exam (NR) is the final evaluation to measure the student learning outcomes and the quality of education in Indonesia and been implemented since 1965 [1]. Regulation of the Ministry of Education and Culture of the Republic of Indonesia No. 5 of 2015, article 1, paragraph 5, stated that the National Examination, hereinafter referred to as the NE, is an activity of measuring and assessing the achievement of national graduate competencies on certain subjects [2]. The NE has undergone several changes and the development of government policies from the beginning set. By 2015, the government sets the NE as one of the minimum passing standards that students must achieve and become a reference point for continuing education to the next level.

High school education in Indonesia has not yet become a compulsory education. The 2015/2016 educational dataset [3] shows that the number of new high school students, both those who attend public high school and vocational schools only amounts to 3,175,407 people, while the new junior high school (SMP) students are 3,435,984 people, difference of about 260,000 people. In fact, the high school level is the level of education that must be passed in order to continue studies to the higher school level, i.e., college or university.

*Corresponding author: ridha.ferdhiana@unsyiah.ac.id
The high school level is divided into 4 majors: Natural Science (IPA), Social Sciences (IPS), Language, and Religion, but the dominant majors are the science and social science. Data from the Education Assessment Centre (Puspendik) of the Ministry of Education and Culture (Kemendikbud) shows that the science and social sciences majors are majors with the interest of more than two other majors. Department of science is also more desirable than IPS because the selection of study programs in universities more freely done by students majoring in science.

Data from Puspendik stated that the average value of UN in Indonesia in the academic year 2014/2015 for high school level is 61.29 with an average score of UN for science majors of 61.84 out of 100.00, thus the average UN for IPS slightly below IPA. The subjects included in the science are Bahasa Indonesia, English, Mathematics, Physics, Chemistry, and Biology. Furthermore, the average value of the UN in 2015 is grouped in four categories, which is very good if the average of UN ≥ 85, good if ≥ 70, enough if ≥ 55, and less if <55.

UN value, in addition to being used as a reference for continuing study to the next level, is also used as a mapping of the success of teaching and learning activities [2]. This mapping can then be used to improve learning and decision-making activities. This mapping, if done using only the average reference value of the NE which are very good, good, enough, and less like the guidance of puspendik, will give a misleading mapping result. The misleading is due to the NE score of 6 subjects, one school could be very strong in mathematics-based sciences but lacking in language and biology, or strong on biology and chemistry but lacking in mathematics and physics.

Economics development in one region certainly has a significant effect on the success of the education in that region. Indonesia is an unequal country for economic development [4], so there is a lack of success in education in all its regions. There are 514 districts in Indonesia spread over 34 provinces and in the 5 largest islands. The islands of Java and Bali are the ones with higher economic levels than the other islands, thus the districts within them have more success in their educational system. If the mapping only sees the average NE score, then only districts in Java and Bali and some cities outside them are included in good category. For that, another way of mapping is needed by looking at the all value of NE subjects. One way to do that is by using the clustering method.

Out of 514 districts in Indonesia only 510 districts have the value of NE majoring in science in the academic year 2014/2015, thus the grouping will be done for the 510 districts. Four districts that do not have the value of NE majoring in Science are Administrasi Kepulauan Seribu, Nduga, Puncak, and Intan Jaya. The absence of NE value data might be caused by the absence of science majors in the area, or there is data input error, or indeed no data. Because education closely related to the Human Development Index (HDI), those 4 districts NE will be predicted based on the value of HDI.

2. Proposed Method

2.1 Data and Variable

The data used in this study was taken from the website of the national exam results report by the Center of Education Assessment (Puspendik) and consist of the average NE values from each district in Indonesia. The HDI variable, which taken from Indonesia Bureau of Statistics (BPS), will be used to predict the value of NE for the four districts that have missing NE value. The variables used are the average values of Bahasa Indonesia (X1), English (X2), Mathematics (X3), Physics (X4), Chemistry (X5), Biology (X6), and the average value of standard deviation of Bahasa Indonesia (X7), English (X8), Mathematics (X9), Physics (X10), Chemistry (X11), Biology (X12), and HDI (X13).

2.2 Data Analysis Procedure

The procedure of data analysis in this study has the following stages:

1. Grouping of districts/cities based on the value of NE majoring in Science using K-means method.
   The stages in performing clustering analysis using K-means method are as follows:
   a) Determine the number of clusters (k) using the Elbow method.
b) Measure the quality of cluster number (k) using clustering quality measurement method: R-square ($R^2$), Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), and Figure of Merit (ADM).

c) Collect the districts groups according to the best number of clusters (k) selected.

2. Grouping the value of HDI for each districts based on the result of cluster in step 1a.

3. Predict the clusters of four districts that do not have the NE value based on the group of their HDI.

3. Result and Discussion

3.1 Determining the Best Number of Cluster (k)  
The number of clusters (k) will be determined based on the Sum of Squares Error (SSE) value or Total Within Sum of Squares (TWSS). The TWSS value shows the cluster variance, the smaller the result of TWSS, the better-formed cluster [5]. The best candidates of the number of cluster are chosen when the graph of TWSS shown a sharp turn, which is similar to elbow, thus called elbow method [6]. The graph of elbow method is shown in Figure 1 and it is clear that the number of clusters (k) to be selected for clustering lies between 3 up to 5 clusters.

![Figure 1. Graph of Elbow method](image)

3.2. Measuring the quality of cluster number (k)  
To determine the best number of cluster, the R-square ($R^2$) and the qualities of measurements are used. $R^2$ is a measurement to check the height of the quality of clustering results by looking at its ability to distinguish existing data on a cluster with existing data on other clusters. The value of $R^2$ ranges from 0 to 1, and the best $R^2$ value is when it is close to 1, which shows the maximum differences between the clusters [7].

Quality of measurements consists of three values. First is the Average Proportion of Non-overlap (APN), which performs calculations on the average proportion of observations that are not placed within the same cluster, with the grouping based on complete data and grouping by data with a single column deleted. The value of APN ranges from 0 to 1 [0.1], the smaller the resulted APN value can show highly consistent grouping results [8]. Second, Average Distance (AD) is a cluster validation calculation based on the average distance between observations placed on the same cluster, with the groupings based on the complete data and grouping by data with one column deleted [8]. The value of AD ranges from 0 to infinity (∞) [0, ∞], the smaller the resulted AD value can show very consistent grouping results. The third is Average Distance Between Means (ADM) which its calculation based on the average distance between cluster centres with the observations on the same cluster, with groupings based on complete data and groupings based on data with one column deleted. The ADM value ranges
from 0 to infinity (∞) \([0, \infty]\), the smaller the resulted ADM value can show very consistent grouping results.

The quality measurements result is shown in Table 1. And the bolded number shows the optimal number of the measurements, the 3 clusters has two optimal measurements while 5 clusters has three, thus it is decided to use 5 clusters.

### Table 1. Results of clustering quality measurement based on number of clusters (k)

| Clustering Quality Measurement | The Number of Cluster (k) |
|-------------------------------|--------------------------|
|                               | 3 | 4 | 5 |
| R-square(R²) (%)              | 67,640 | 72,264 | 75,600 |
| Average Proportion of Non-overlap (APN) | 0.0527 | 0.1014 | 0.2120 |
| Average Distance (AD)         | 26,521 | 24,987 | 24,533 |
| Average Distance between Means (ADM) | 2,5643 | 4,0911 | 7,0979 |
| Figure of Merit (ADM)         | 5,7873 | 5,4181 | 5,3170 |

#### 3.3. Cluster of Districts Based on NE Value

Based on the clustering quality measurement, the K-Means algorithm is run to get five clusters of district based on NE value with the results of the number of districts in each cluster are 141, 169, 77, 86, and 177 respectively. The summary for the number of each cluster of districts per province is shown in Table 2.

Figure 2 shows the boxplots of each subject for each cluster. The purpose of the boxplot is to make comparison of the subject average value by cluster. Based on the Figure, it seen that the IV cluster is the cluster with the best ranking, followed by the II cluster, then the V cluster, the III cluster, while the I cluster is cluster with the lowest rank for every subjects. This finding is match the preliminary epiphany that the district with higher economic status will have a better education system, as it is clearly shown that most districts in Java and Bali Island are laid in the IV, II, and V clusters. While, most districts in Papua, Nusa Tenggara Barat, and Nusa Tenggara Timur mostly laid in the I and III cluster.

### Table 2. Cluster of districts/cities per province based on the NE value (1/2)

| No. | Province     | Cluster | I | II | III | IV | V | Total |
|-----|--------------|---------|---|----|-----|----|---|-------|
| 1   | DKI Jakarta  | n       | % | n  | %  | n  | % | n   | %  | N   | %  |
| 2   | West Java    | -       | - | 3  | 60.00 | - | - | 2   | 40.00 | - | 5   | -  |
| 3   | Central Java | -       | - | 11 | 40.74 | 2 | 7.41 | 5   | 18.52 | 9 | 33.33 | 27 |
| 4   | DI Yogyakarta| -       | - | 8  | 22.86 | - | - | 27  | 77.14 | - | 35  | -  |
| 5   | East Java    | -       | - | 20 | 52.63 | - | - | 14  | 36.84 | 4 | 10.53 | 38 |
| 6   | Aceh         | -       | - | 13 | 56.52 | 1 | 4.35 | 8   | 34.78 | 1 | 4.35  | 23 |
| 7   | North Sumatera| 1      | 3.03 | 10 | 30.30 | 3 | 9.09 | 17  | 51.52 | 2 | 6.06  | 33 |
| 8   | West Sumatera| 1      | 5.26 | 11 | 57.89 | - | - | 2   | 10.53 | 5 | 26.32 | 19 |
| 9   | Riau         | -       | - | 9  | 75.00 | - | - | 3   | 25.00 | - | 12   | -  |
| 10  | Jambi        | 1      | 9.09 | 3 | 27.27 | 3 | 27.27 | -   | -   | 4 | 36.36 | 11 |
| 11  | South Sumatera| -     | - | 7  | 41.18 | 1 | 5.88 | 5   | 29.41 | 4 | 23.53 | 17 |
| 12  | Lampung      | 1      | 6.67 | 2 | 13.33 | 4 | 26.67 | -   | -   | 8 | 53.33 | 15 |
| 13  | West Kalimantan| -     | - | 4  | 28.57 | 3 | 21.43 | -   | -   | 7 | 50.00 | 14 |
| 14  | Central Kalimantan| 1   | 7.14 | - | -  | 5 | 35.71 | -   | -   | 8 | 57.14 | 14 |
| 15  | South Kalimantan| -    | - | 7  | 53.85 | 1 | 7.69 | 1   | 7.69 | 4 | 30.77 | 13 |
| 16  | East Kalimantan| -     | - | 1  | 10.00 | 4 | 40.00 | -   | -   | 5 | 50.00 | 10 |
| 17  | North Sulawesi| -      | - | 7  | 46.67 | 1 | 6.67 | 2   | 13.33 | 5 | 33.33 | 15 |
| 18  | Central Sulawesi| 2    | 15.38 | 1 | 7.69 | 8 | 61.54 | -   | -   | 2 | 15.38 | 13 |
| 19  | South Sulawesi| -     | - | 14 | 58.33 | - | - | 9   | 37.50 | 1 | 4.17  | 24 |
| No. | Province            | Cluster          | Total |
|-----|---------------------|------------------|-------|
|     |                     | I    | %   | II  | %   | III | %   | IV  | %   | V   | %   | Total |
| 20  | Southeast Sulawesi  | -    | -   | 11  | 64.71 | -    | -   | 3   | 17.65 | 3   | 17.65 | 17    |
| 21  | Maluku              | 5    | 45.45 | 3   | 27.27 | -    | -   | -   | -    | 3   | 27.27 | 11    |
| 22  | Bali                | -    | -   | -   | -    | -    | -   | 9   | 100   | -   | -     | 9     |
| 23  | West Nusa Tenggara | -    | -   | 3   | 30.00 | 2    | 20.00 | 1   | 10.00 | 4   | 40.00 | 10    |
| 24  | East Nusa Tenggara | 12   | 54.55 | 7   | 31.82 | -    | -   | 2   | 9.09  | -   | -     | 22    |
| 25  | Papua               | 10   | 38.46 | 7   | 26.92 | 4    | 15.38 | 1   | 3.85  | 4   | 15.38 | 26    |
| 26  | Bengkulu            | -    | -   | 1   | 10.00 | 5    | 50.00 | -   | -     | 4   | 40.00 | 10    |
| 27  | North Maluku        | 2    | 20.00 | 2   | 20.00 | 3    | 30.00 | 1   | 10.00 | 2   | 20.00 | 10    |
| 28  | Bangka Belitung     | -    | -   | 7   | 100   | -    | -   | -   | -     | -   | -     | 7     |
| 29  | Gorontalo           | -    | -   | 2   | 33.33 | 2    | 33.33 | 1   | 16.67 | 1   | 16.67 | 6     |
| 30  | Banten              | -    | -   | 2   | 25.00 | 2    | 25.00 | -   | -     | 4   | 50.00 | 8     |
| 31  | Riau Islands        | 1    | 14.29 | -    | -   | 2    | 28.57 | -   | -     | 4   | 57.14 | 7     |
| 32  | West Sulawesi       | -    | -   | 1   | 16.67 | 3    | 50.00 | -   | -     | 2   | 33.33 | 6     |
| 33  | West Papua          | 3    | 23.08 | 5   | 38.46 | -    | -   | 1   | 7.69  | 4   | 30.77 | 13    |
| 34  | North Kalimantan    | 1    | 20.00 | -    | -   | 4    | 80.00 | -   | -     | -   | -     | 5     |

Table 2. Cluster of districts/cities per province based on the NE value (2/2).

Figure 2. Boxplot cluster rating by subjects
3.4. HDI Score of Cluster of Districts Based on NE Value
Based on the cluster obtained above, then the HDI value of each district in the cluster is grouped. This grouping aims to see whether educational success is comparable to overall human development. Statistical measurements of the grouped HDI values are then calculated as shown in Table 3. Based on the mean HDI value for each cluster, it can be seen that the HDI rating is the same as the result rating of the NE value. Thus, we can conclude that education is an important factor in Human Development Index.

### Table 3. Summary statistics of HDI Score for each cluster

| Summary Statistics | I  | II | III | IV  | V  |
|--------------------|----|----|-----|-----|----|
| Minimum            | 52,78 | 43,55 | 44,32 | 53,73 | 40,91 |
| Maximum            | 70,89 | 80,96 | 79,63 | 83,37 | 81,20 |
| Mean               | 61,82 | 68,13 | 64,97 | 70,57 | 66,90 |
| Standard Deviation| 4,53 | 6,72 | 5,41 | 6,14 | 5,68 |
| Median             | 60,96 | 67,45 | 64,36 | 70,36 | 67,12 |
| 95% Confidence Interval |    |      |      |      |      |
| Lower Limit        | 60,39 | 69,64 | 64,69 | 69,25 | 67,43 |
| Upper Limit        | 63,25 | 72,18 | 67,56 | 73,02 | 70,15 |

3.5. The Prediction of Cluster of Districts that has no NE Value
The HDI grouping aims to predict the NE value for districts that does not have it, which are district of Administrasi Kepulauan Seribu, Nduga, Puncak and Intan Jaya. The 95% Confidence Interval (CI) shown in Table 3 is then be used to predict the district cluster according to their HDI values. The cluster where those districts belong is given in Table 5. Thus, the Administrasi Kepulauan Seribu is belong to the V cluster, while those other 3 belong to the I cluster though their HDI values are far below the 95% CI of HDI in the I cluster. Table 4 shows the prediction values of each subjects for those 4 districts. Figure 3 give overall mapping of cluster rating of NE values in Indonesia. The mapping shows that most Java and Bali Islands are in highest rating while Papua and Sumbawa and Kalimantan islands are in the lowest rating.

### Table 4. The Prediction of NE values per subjects of districts that has no NE values

| Districts/cities              | Cluster | Subjects | Confidence Interval 95% |
|------------------------------|---------|----------|-------------------------|
| Administrasi Kepulauan Seribu| V       | Indonesian | 73,39 | 75,15 |
|                              |         | English   | 60,68 | 62,44 |
|                              |         | Mathematics | 50,57 | 52,67 |
|                              |         | Physics   | 57,57 | 59,90 |
|                              |         | Chemistry | 53,48 | 55,63 |
|                              |         | Biology   | 57,44 | 59,16 |
| Nduga                        | I       | Indonesian | 59,07 | 64,54 |
|                              |         | English   | 40,19 | 44,01 |
| Puncak                       |         | Mathematics | 25,86 | 28,32 |
|                              |         | Physics   | 28,49 | 31,19 |
| Intan Jaya                   |         | Chemistry | 29,43 | 32,74 |
|                              |         | Biology   | 34,24 | 37,22 |
Table 5. The clusters of districts that has no NE value

| Districts                | HDI  | Cluster | Confidence Interval 95% | Lower Limit | Upper Limit |
|--------------------------|------|---------|-------------------------|-------------|-------------|
| Administrasi Kepulauan Seribu | 68.84 | V       | 67.43                   | 70.15       |
| Nduga District           | 25.47 | I       | 60.39                   | 63.25       |
| Puncak District          | 39.41 | I       | 60.40                   | 63.26       |
| Intan Jaya District      | 44.35 | I       | 60.41                   | 63.27       |

Figure 3. Map of cluster districts ranking by island in Indonesia

4. Conclusion
The conclusions of cluster districts analysis based on NE value and HDI score using K-means clustering method can be revised as follows:
1. The best number of cluster (k) to cluster districts based on UN value is 5
2. The order of the rating of clusters is the IV, II, V, III and last the I cluster.
3. The cluster of those districts that has no NE values are Administrasi Kepulauan Seribu City belong to the V cluster, while Nduga, Puncak and Intan Jaya belong to the I cluster.

5. Acknowledgments
This research was supported by Lembaga Penelitian dan Pengabdian kepada Masyarakat Syiah Kuala University through Penelitian Lektor Scheme. We thank our colleagues from Pusat Penilaian Pendidikan, Kementerian Pendidikan Dan Kebudayaan who provided data. We also thank our colleagues in Statistics Department of Syiah Kuala University for the support and discussions.
6. References

[1] Setneg RI (Sekretariat Negara Republik Indonesia) 1966 *Penetapan Presiden Republik Indonesia Nomor 19 Tahun 1965 tentang Pokok-Pokok Sistem Pendidikan Nasional Pancasila* (Jakarta: Setneg RI)

[2] Mendikbud RI (Menteri Pendidikan dan Kebudayaan Republik Indonesia) 2015 *Peraturan Menteri Pendidikan dan Kebudayaan Nomor 5 Tahun 2015 tentang Kriteria Kelulusan Peserta Didik, Penyelenggaraan Ujian Nasional, dan Penyelenggaraan Ujian Sekolah/Madrasah/ Pendidikan Kesetaraan pada SMP/MTs atau yang Sederajat dan SMA/MA/SMK atau yang Sederajat* (Jakarta: Mendikbud RI).

[3] Kementerian Pendidikan Dan Kebudayaan, Sekretariat Jenderal Pusat Data Dan Statistik Pendidikan Dan Kebudayaan. 2016 *Statistik Sekolah Menengah Atas (SMA) 2015/2016*.

[4] Andi Irawan 2015 *Bulletin of Indonesian Economic Studies*, 51:1, 148-149.

[5] Calinski, Tadeusz, and Harabasz, Joachim. 1974 *Communications in Statistics*. 3(1): 1-27.

[6] Kodinariya, Truptil M. and Maknawa, Prashant R. 2013 *International Journal of Advance Research in Computer Science and Management Studies*. 1(6): 90-95.

[7] Gudono 2011 *Analisis Data Multivaria* (Yogyakarta: BPFE)

[8] Datta, Susmita, and Datta, Somnath. 2003 *Bioinformatics* 19(4): 459–466.