Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 12, No. 1, Juli 2023

ISSN 0216 – 0544
e-ISSN 2301– 6914

APPLICATION OF MULTISTAGE CLUSTERING FOR MAPPING ECONOMIC POTENTIAL IN EAST JAVA PROVINCE

[a]Ronny Susetyoko, [b]Edi Satriyanto*, [c]Alfi Fadliana, [d]Muhammad Syahfitra

[a,b,c]Politeknik Elektronika Negeri Surabaya
[d]Politeknik Pertanian Negeri Payakumbuh
E-mail: rony@pens.ac.id, edi@pens.ac.id, alfi@pens.ac.id, syahfitra@politanipyk.ac.id

Abstract

This study aims to map the economic potential in East Java Province based on GRDP according to business field category. Multistage clustering is a method developed for outlier data and datasets with large variance. Multistage clustering is a combination of Ordering Points to Identify the Clustering Structure (OPTICS) and K-Means. The first stage was grouped using OPTICS. The outlier data resulting from the clustering stage is used as a dataset in the second stage using K-Means. The performance of this method is compared with several other methods, namely: K-Means, DBSCAN – K-Means, Agglomerative, Fuzzy C-Means (FCM), Possibilistic C-Means (PCM), and Fuzzy Possibilistic C-Means (FPCM) based on the characteristics of the Silhouette score and Davies-Bouldin score. Multistage clustering was chosen as the best method with a Silhouette score of 0.442 and Davies-Bouldin score of 0.388. With the Elbow method and the two metrics, the optimum number of clusters is 8 clusters. The results of this mapping method, the City of Surabaya forms a separate cluster which has the highest economic potential in 15 categories of business fields. Next Gresik, Pasuruan, Sidoarjo, and Probolinggo have the second highest economic potential with 10 categories of business fields ranking in the top 3.

Key words: Davies-Bouldin score, mapping, multistage clustering, OPTICS, Silhouette score.

INTRODUCTION

There are 17 targets and policies in the Sustainability Development Goals (SDGs) whose achievement requires cooperation from all stakeholders, both at the central and regional levels. Indonesia's demographic composition is dominated by the productive age population (15-64) as much as 68.7% of the total population in 2019 (Bappenas, BPS, and UNFPA, 2018), which has the potential for a large workforce that can accelerate economic growth [1].

Gross Regional Domestic Product (GRDP) is an important indicator to determine the economic condition of a country in a certain period of time. GRDP calculation is a very important part in macroeconomics, especially regarding the economic analysis of a region [2][3]. GRDP can measure the rate of economic growth [4]. Based on data from the Central Statistics Agency (BPS), Indonesia's economic growth in the second quarter of 2022 reached 5.44% (y-o-y). The acceleration of economic performance was supported by domestic demand which continued to increase, especially household consumption and export performance which remained high. Economic improvement was driven by several business fields such as processing, transportation and warehousing, as well as wholesale and retail trade [5].

In 2021, East Java Province will contribute 14.57% to Indonesia's Gross Domestic Product

(GDP). East Java has a variety of potentials in its sectors ranging from agriculture, processing industry, trade to services. Diversity in regions geographically and socio-culturally is the driving force for the various potentials that exist in the East Java region [6]. East Java's economic growth continues to show good progress. In the second quarter of 2022, the business sector that experienced significant growth was transportation and warehousing by 22.21%. Next, other service business fields by 13.07%, and electricity and gas procurement by 9.58% [7].

The 2015 World Bank report shows that Indonesia's economic growth is only enjoyed by 20% of the population in the highest income group identified as the consumer class. And in 2021 Indonesia's economic growth which reached 7.07% will only be enjoyed by the upper middle class [8].

The purpose of this research is to map the economic potential in the East Java Province. Mapping of economic potential based on clustering results of gross regional income (GDP) based on current prices according to business field. Some researchers use various clustering methods for mapping. The Fuzzy C-Means Clustering method is used for a mapping system in East Java province for classification of cities/regencies with potential for transmigration [9]. Fuzzy C-Mean and K-Means algorithms are also used to manage agricultural data from data mining results, which are used to find and form clusters of agricultural land areas according to the type of commodity based on the supporting attributes used. The results of this analysis are implemented to provide land information such as the number of clusters, land area, area, location and level of productivity, which can be used as input in the process of conversion and arrangement of agricultural land [10].

The Location Quotient (LQ) is used to determine the leading/base and non-competitive/non-base sectors to support economic growth. The data analyzed is the Gross Regional Domestic Product (GDP) at constant prices in 2010, Arfak Mountains District and West Papua Province in 2014-2019. The result is that there are 6 leading sectors that contribute significantly to the economy of Arfak Mountains Regency. As a recommendation, the Regional Government of Arfak Mountains Regency is expected to be able to manage and improve the quality of sectors that are not competitive/non-base [11]. Analysis of Location Quotient, Shift Share, and Klassen Typology is used to classify the growth of the economic sector and find out the basic sectors and leading sectors in Pamekasan Regency [12]. Location Qoutient (LQ) method, Shift-Share analysis, Growth Ratio Model (MRP), Overlay Analysis are also used to map the potential of Ngawi Regency based on GRDP in 2015-2019. The results of the analysis show that the base sectors in Ngawi Regency are agriculture, transportation and communication, and services [13].

Hierarchical cluster analysis is used to classify the level of welfare of districts/cities needed as input for policy formulation and as a tool to view conditions, monitor and evaluate the success of development in East Java in accordance with the SDGs [14]. Grouping areas based on economic potential is also done by clustering data with mixed attributes (numerical and categorical data). Clustering was carried out using the Fuzzy k-prototypes algorithm and modified Eskin distances to measure categorical attribute distances. The results of this clustering can be used as a guide in determining village development targets in increasing the Village Building Index in Demak Regency [15].

Clustering is the most important unsupervised learning technique as it deals with finding structure in unlabeled data collections. Several clustering algorithms such as K-means clustering, Fuzzy C-means hierarchical clustering, DBSCAN, OPTICS, STING, ROCK and CACTUS are used to select an appropriate clustering approach [16]. The DBSCAN method is also used for grouping earthquake data [17].

OPTICS is an unsupervised learning algorithm based on hierarchical density, which is not sensitive to parameters. This method can handle the problem of clusters that have non-uniform densities in the dataset and can visualize the cluster structure of the data set [18]. OPTICS utilizes the distance between neighboring points to construct reachability plots, which are used to distinguish groups of varying densities of noise [19].

A two-stage clustering method was used to capture a more representative temporal pattern of the loading shape and peak demand through a cluster merging approach [20]. A two-stage

clustering technique is also used for image based hyperspectral sensing Neighboring Union Histogram (NUH). The first stage of relatively coarse clustering using K-means for classify each group's NUH and the second stage uses K-means to perfect the results of the first grouping [21]. Multistage clustering is also used for summarizing unstructured text documents [22].

This research collaborates two clustering methods that are often used by other researchers, namely Ordering Points to Identification of the Clustering Structure (OPTICS) and K-Means. Multistage clustering was developed to handle outlier data and dataset cases that have large variances. This method is a combination of OPTICS and K-Means. In the first stage, the OPTICAL method is used to separate cluster results with noise or outliers. In the second stage, the remaining datasets that have been grouped as outliers are regrouped using the K-Means method. The multilevel clustering performance will also be compared with other clustering methods.

MATERIAL AND METHODS

The stages of this research are shown in Fig. 1 starting from dataset collection, performing extraction and transformation (ETL), preliminary analysis, OPTICS clustering (Stage-1), taking outliers as new datasets, K-Means clustering (Stage-2), combining labeling results (cluster), modeling with some other method, model evaluation to implementation and analysis.

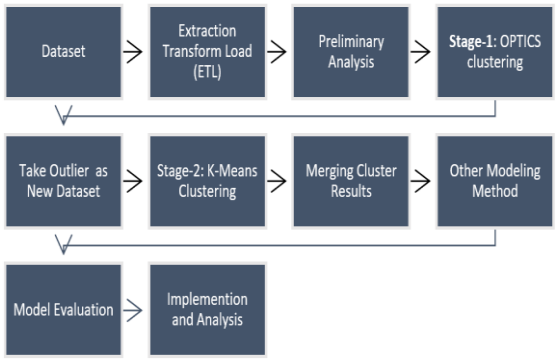

Fig 1. Methodology

In the early stages, descriptive analysis was carried out on the original data and some of the results of the dataset transformation using natural logarithms, square root, and a combination of both. In addition, data adequacy testing was also carried out using The Kaiser-Meyer-Olkin (KMO) test and normal multivariate testing using the Shapiro-Wilk normality test.

In stage-1, grouping was carried out using OPTICS. Data that are classified as outliers at this stage are grouped using K-Means, then the labeling results at stage-1 and stage-2 are combined. Furthermore, grouping was carried out using several other methods, namely: K-Means, Multistage clustering (DBSCAN – K-Means), Agglomerative, FCM, PCM, and FPCM. At the model evaluation stage, the best model was selected based on the characteristics of the Silhouette score and Davies-Bouldin score. The final stage is to apply the best method for mapping economic potential based on GRDP for each category of business field in regencies/cities in East Java Province.

Datasets

The source of the dataset in this study is the Publication of Regency/City Gross Regional Domestic Product in East Java Province According to Business Sector Central Bureau of Statistics or Badan Pusat Statistik (BPS) East Java for 2014 – 2021. Clustering attributes can be seen in Table 1.

Table 1. Attributes for clustering

Category	Business field
A	Agriculture, Forestry and Fisheries
B	Mining and excavation
C	Industry and Processing
D	Procurement of Electricity and Gas
E	Water Procurement, Waste Management, Waste and Recycling
F	Construction
G	Wholesale and Retail Trade, Car and Motorcycle Repair
H	Transportation and Warehousing
I	Provision of Accommodation and Food and Drink
J	Information dan communication
K	Financial Services and Insurance
L	Real Estate
MN	Company Services
O	Government Administration, Defense and Compulsory Social Security
P	Education Services
Q	Health Services and Social Activities
RSTU	Other Services

Multistage Clustering Algorithm

The inability to find clusters with different densities is the main drawback of DBSCAN [23]. For this reason, several DBSCAN authors developed the OPTICS method. Multistage clustering is two stages of clustering, the first stage uses OPTICS. Observations identified as outliers are new datasets for the second stage of clustering using K-Means. The OPTICS algorithm steps are shown in the pseudo code in [24]. Then in the second stage using the K-Means algorithm [25].

Model Evaluation

The performance of multistage clustering is compared to other methods, namely K-Means, DBSCAN – K-Means, Agglomerative, FCM, PCM, and FPCM based on the Silhouette score and Davies-Bouldin score.

Implementation and Analysis

In the final stage, the best method or model will be applied for classifying economic potential in the East Java Province area based on GRDP for all categories of business fields. The grouping will be carried out from 2014 – 2021. The shift of a district/city from one cluster to another will be analyzed at this stage.

RESULT AND DISCUSSION

Statistic Descriptive

Based on Fig. 2, in 2021 the average GRDP for East Java Province is Rp. 64610.5 billion. There are 12 regencies/cities that have GRDP above the average, namely Surabaya City (Rp. 590228 billion), Sidoarjo (Rp. 210643.9 billion), Pasuruan (Rp. 157150.8 billion), Gresik (Rp. 144435.2 billion) , Kediri City (Rp. 141466.9 billion), Malang (Rp. 107036.4 billion), Mojokerto (Rp. 87261.7 billion), Banyuwangi (Rp. 85915.8 billion), Bojonegoro (Rp. 83439.2 billion), Jember (Rp. 81069 billion), Malang City (Rp. 76617.4 billion), and Tuban (Rp. 65901.6 billion).

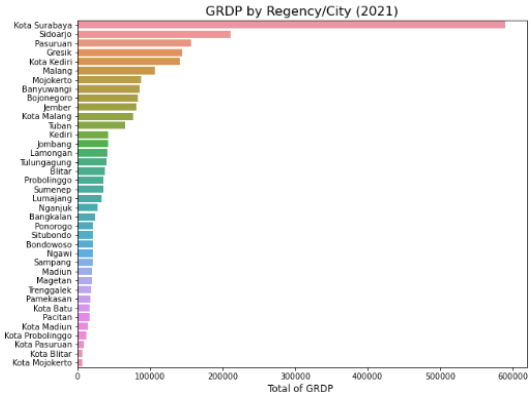

Fig 2. GRDP by regency/city 2021 (Rp. billion)

Based on Fig 3., GRDP based on category A (Agriculture, Forestry, and Fisheries) the first highest is Banyuwangi (Rp. 25028.7 billion), the second highest is Jember (Rp. 21089.5 billion), and the third highest is Malang (Rp. 15836 billion) .

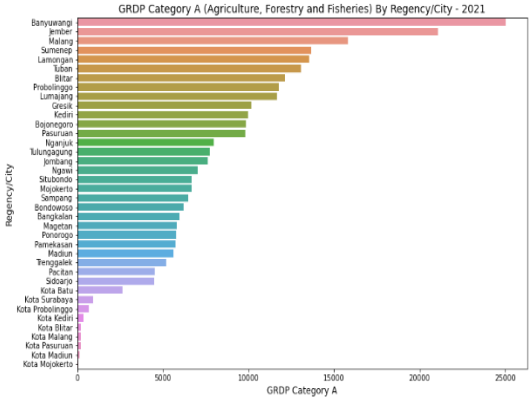

Fig 3. GRDP category A (agriculture, forestry and fisheries) 2021

Preliminary Analysis

The distribution of GRDP in each business field which is then used as the specified attributes in clustering is shown in the histogram matrix in Fig. 4. All attributes are not normally distributed and tend to have positive skewness.

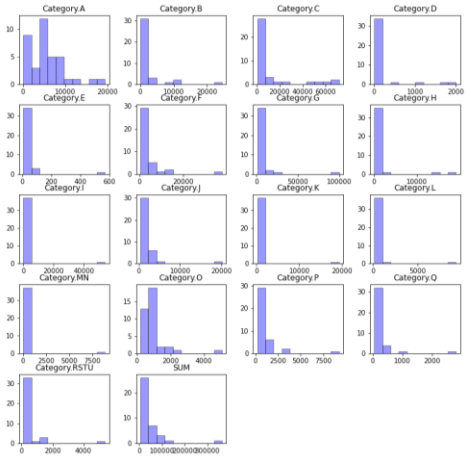

Fig 4. Histogram matrix of original data attributes

Furthermore, several transformations are carried out to see the distribution of the data, the number of outlier data, the data adequacy test (KMO test), and the Shapiro-Wilk normality test. Based on Table 2, it is found that the results of the transformation of natural logarithms, square roots, and the combination of the two data are still not normally distributed and tend to be exponentially distributed. This is evidenced by the p-value which is less than 5%. The number of outlier data from the original data and the data resulting from the transformation is the same, namely as many as 10 observations and the KMO test value is more than 0.5 (sufficient sample).

Table 2. Average skewness, outliers, KMO test and shapiro-wilk normality test

Dataset Transformation	Avg. Skewness	Outliers	KMO test	Shapiro-Wilk normality test (p-value)
Original (X)	4.554	10	0.86	1.79E-13
Ln(X + 1)	1.190	10	0.85	1.79E-13
Sqrt(X)	3.270	10	0.85	1.79E-13
Ln(Sqrt(X))	0.845	10	0.86	1.96E-09
Sqrt(Ln(X+1)	1.253	10	0.85	8.49E-11

The distribution of data resulting from the transformation of natural logarithms is shown in the histogram matrix in Fig. 5. Using Kolmogorov – Smirnov (a = 1%), there are 53% of the attributes that have a Normal distribution, namely category C, category E, category F, category G, category J, category K, category, L, category O, and category Q.

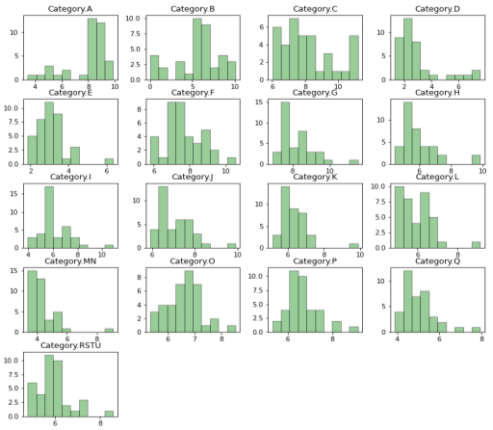

Fig 5. Histogram matrix of transformation data attributes (natural logarithm)

Model Evaluation

Based on the results of the initial analysis, at the clustering stage the data set resulting from the natural logarithm transformation was used.

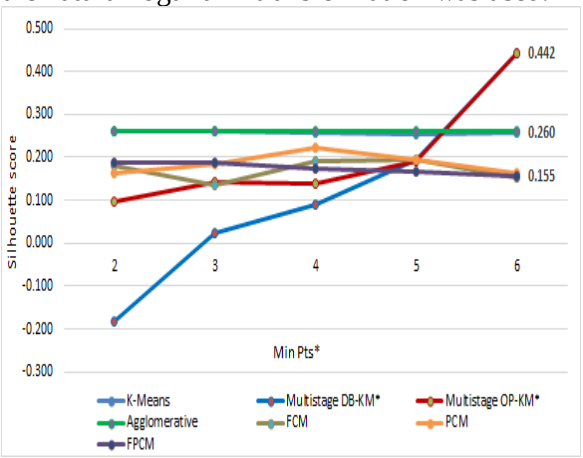

Fig 6. Comparison of silhouette score multistage clustering (OPTICS - K-Means) with other methods.

Before implementing Multistage clustering (OPTICS – K-Means), this method will be compared with several other methods, namely: K-Means, Multistage clustering (DBSCAN – K-Means), Agglomerative, FCM, PCM, and FPCM. The parameter used to compare is the Silhouette score. For multistage clustering, the Min Pts value varies from 2 to 6. While other methods are replicated 5 times. The silhouette score can be seen in Fig. 6. The silhouette score on K-Means and Agglomerative is relatively the same, which is around 0.26 (blue and green lines that almost coincide). As for the FCM, PCM and FPCM methods (each with a gray,

orange and purple line) the Silhouette score ranges from 0.137 – 0.222.

In multistage clustering (DBSCAN – K-Means) which is depicted by a bright blue line, the Silhouette score increases significantly as Min Pts increases from -0.183 to 0.442. Likewise multistage clustering (OPTICS – K-Means) which is illustrated with a red line, Silhouette score 0.098 at Min Pts = 2 and tends to increase up to 0.442 at Min Pts = 6. At this stage, multistage clustering (OPTICS – K-Means) can be considered to be applied because the Silhouetter score at Min Pts = 6 is the highest. For various values of Min Pts, OPTICS – K-Means is better than DBSCAN – K-Means.

The next step is to determine the optimal number of classes based on the Elbow method, Silhouette score and DBS score. At this stage the FCM, PCM, and FPCM methods are ignored. The elbow curve can be seen in Fig. 7. From this curve, the number of classes can be determined between 4 to 9 because there is no clear elbow.

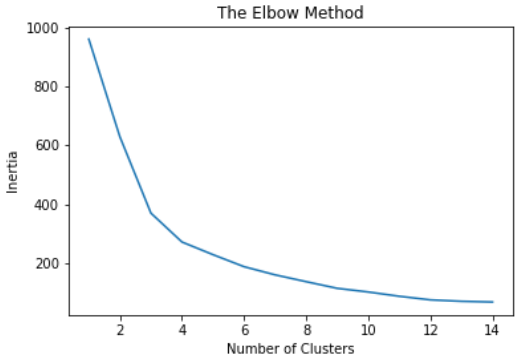

Fig 7. The elbow method

Fig. 8. is the Silhouette score on the number of classes 3 to 9 of the 4 methods. In the K-Means method, as the number of clusters increases, the monotone Silhouette score decreases. Agglomerative performance is not better than the K-Means method. For multistage clustering (OPTICS – K-Means), the Silhouette score for the number of clusters = 3, 5, 8, and 9 has a score of more than 0.44. Whereas in multistage clustering (DBSCAN – K-Means), the Silhouette score for the number of clusters = 4,5,7, and 9. When the number of clusters = 8, the Silhouette score is only 0.238. The next analysis focuses on comparing the performance of OPTICS – K-Means with DBSCAN – K-Means.

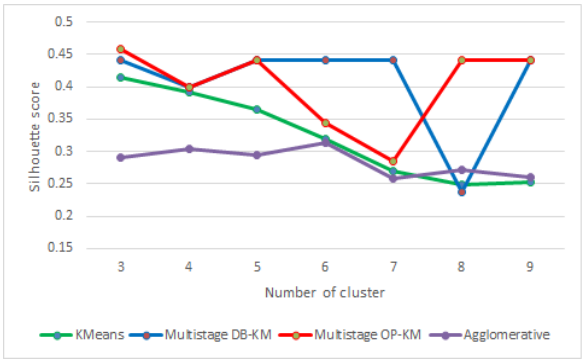

Fig 8. Silhouette score on the number of clusters 3 to 9

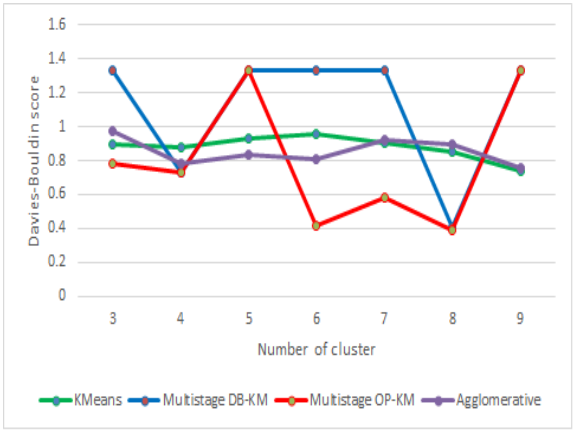

Fig 9. Silhouette score on the number of clusters 3 to 9

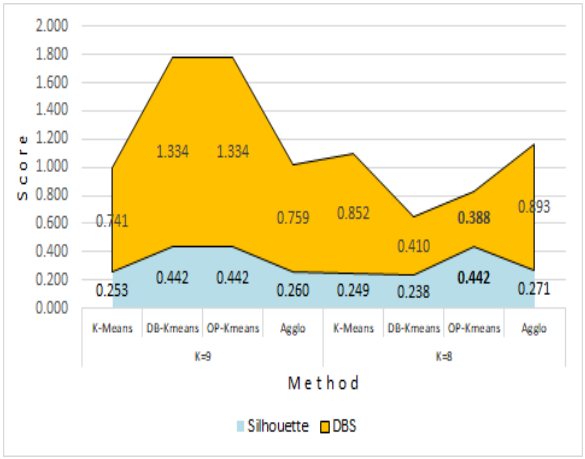

Fig 10. Silhouette score and DBS score maps

Fig. 9. is the DBS score for the number of classes 3 to 9 of the 4 methods. The lowest DBS score is achieved by multistage clustering (OPTICS – K-Means) when the number of clusters = 8, which is 0.388. This value is lower than multistage (DBSCAN – K-Means) with Davies-Bouldin score = 0.409. The map in Fig. 10 is clearer to compare the Silhouette score with the DBS score with the number of clusters 8 and 9 in the 4 methods.

Multistage clustering (OPTICS – K-Means) was selected with the number of clusters = 8 because it has the highest Silhouette score (0.442) and the lowest DBS score (0.388).

Implementation and Analysis

In the following analysis there is still a mention of the word "city" but no mention of the word "regency" for the regency area. The multistage clustering method (OPTICS – K-Means) was implemented for mapping the economic potential of East Java Province based on the GRDP of each business field category. Fig. 11. is a scatter plot between potential category A (Agriculture, Forestry and Fisheries) and potential category B (Mining and excavation). Cluster-4, especially Bojonegoro has the highest GRDP in this sector. It can be seen that the red dot has a high value on the category A axis close to Rp. 10,000 billion and category B more than Rp. 40,000 billion.

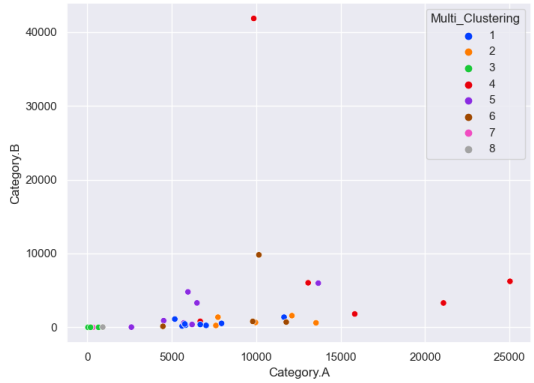

Fig 11. Scatter plot Category A vs. Category B

Fig. 12. is a scatter plot between potential category A (Agriculture, Forestry and Fisheries) and potential category D (Procurement of Electricity and Gas). Clusters with relatively high GRDP compared to other clusters in these two potentials are Cluster-6, namely Gresik, Pasuruan, Sidoarjo, and Probolinggo (dots colored brown). Surabaya City has the highest potential for the procurement of electricity and gas, but the potential for agriculture, forestry and fisheries is very low. Likewise, Cluster-4 tends to have high potential in the agriculture, forestry, and fisheries business fields.

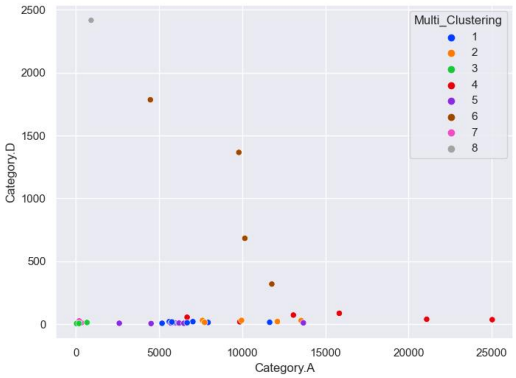

Fig 12. Scatter plot Category A vs. Category D

Fig. 13. is a scatter plot between the potential category C (Industry and Processing) and the potential category F (Construction). Cluster-8 (Surabaya City) has a very high GRDP in both categories with a very large gap compared to Cluster-6, Cluster-4, Cluster-7 and other clusters.

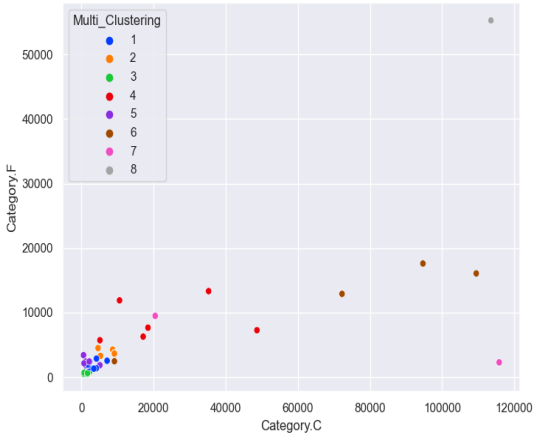

Fig 13. Scatter plot Category C vs. Category F

In this study, Multistage clustering (OPTICS – K-Means) was implemented for grouping economic potential from 2014 – 2021 to see cluster shifts from year to year.

Table 3 shows there is a shift in the economic potential clusters of Batu City from 2019 to 2020, namely from cluster-2 to cluster-5. From 2020 to 2021 there will be a shift in the Trenggalek economic potential cluster from cluster-5 to cluster-1. This also happened in Probolinggo , which shifted from cluster-2 to cluster-6 so that it became one cluster with Gresik, Pasuruan , and Sidoarjo.

Table 4 and Table 5 show the cluster centers (centroids) from the 2021 dataset clustering. In general, Cluster-8, namely the City of Surabaya has the highest economic potential in 15 categories of business fields, namely category C (Industry and Processing), category D (Procurement of Electricity and Gas), category E (Water Procurement, Waste Management, Waste and Recycling), category F (Construction), category G (Wholesale and Retail Trade, Car and Motorcycle Repair), category H (Transportation and Warehousing), category I (Provision of Accommodation and Food and Drink), category J (Information and communication), category K (Financial Services and Insurance), category L (Real Estate), category MN (Company Services), category O (Government Administration, Defense and Compulsory Social Security), category P (Education Services), category Q (Health Services and Social Activities), and the RSTU (Other Services) category.

Table 3. Result of Clustering 2019 - 2021

Cluster	Regency/City
1	Lumajang, Madiun, Magetan, Nganjuk, Ngawi, Ponorogo, Situbondo, Trenggalek**
2	Blitar, Jombang, Kediri, Lamongan, Probolinggo, Tulungagung, Batu City
3	Blitar City, Madiun City, Mojokerto, Pasuruan City, Probolinggo City
4	Banyuwangi, Jember, Malang, Mojokerto City, Tuban, Bojonegoro
5	Bondowoso, Batu City*, Pacitan, Pamekasan, Trenggalek, Bangkalan, Sampang, Sumenep
6	Gresik, Pasuruan, Sidoarjo, Probolinggo**
7	Kediri City, Malang City
8	Surabaya City

** cluster shift 2019 to 2020*
***cluster shift 2020 to 2021*

Cluster-6 namely Gresik, Pasuruan, Sidoarjo, and Probolinggo in general have the second highest economic potential with 10 categories of business fields each ranking in the top 3, namely: the third highest in category A (Agriculture, Forestry and Fisheries), second highest in category B (Mining and excavation), second highest in category C (Industry and Processing). The second highest in category D (Procurement of Electricity and Gas). Third highest in category E (Water Procurement, Waste Management, Waste and Recycling), the second highest in category F (Construction), the second highest in category G (Wholesale and Retail Trade, Car and Motorcycle Repair), the second highest in category H (Transportation and Warehousing), the third highest in category MN (Company Services), and third highest in category O (Government Administration.

Cluster-7, namely Kediri City, Malang City in general has the third highest economic potential with 10 categories of business fields each ranking in the top 3, namely: the third highest in category C (Industry and Processing).

The second highest in category E (Water Procurement, Waste Management, Waste and Recycling), the third highest in category G (Wholesale and Retail Trade, Car and Motorcycle Repair), the third highest in category I (Provision of Accommodation and Food and Drink), the second highest in category K (Financial Services and Insurance), the second highest in the MN (Company Services) category, the second highest in the P (Education Services) category, the second highest in the Q (Health Services and Social Activities) category, and the second highest in the RSTU (Other Services) category.

Table 4. Cluster Center of Category A - I

Cluster	Category								
	A	B	C	D	E	F	G	H	I
Cluster_1	6965	573	3387	15	25	1977	3808	415	534
Cluster_2	*10195*	*893*	7390	25	31	3890	8224	633	693
Cluster_3	243	1	1427	8	19	626	2606	635	495
Cluster_4	*15262*	*10001*	22545	*51*	57	*8692*	11862	*1257*	1808
Cluster_5	6456	2285	1722	8	18	2296	3604	293	427
Cluster_6	*9054*	*2865*	*71395*	*1038*	*83*	*12259*	*18332*	*4959*	*3848*
Cluster_7	269	25	*68152*	19	*86*	5890	*17585*	1205	*2788*
Cluster_8	918	33	*113545*	*2417*	*872*	*55274*	*163510*	*30520*	*91418*
Average	*6310*	*2135*	34378	419	139	10752	26828	4646	11733

Table 5. Cluster Center of Category J – RSTU

Cluster	Category							
	J	K	L	MN	O	P	Q	RSTU
Cluster_1	1350	579	411	79	1117	1028	206	486
Cluster_2	2655	890	878	125	1432	1776	394	569
Cluster_3	1078	834	289	66	438	569	135	319
Cluster_4	*4771*	1516	*1233*	212	*2221*	*2191*	*483*	*941*
Cluster_5	1294	477	328	65	972	813	146	564
Cluster_6	*4990*	*1548*	*1487*	*265*	*2171*	1452	471	740
Cluster_7	3164	*1766*	846	*448*	797	*3649*	*1245*	*1107*
Cluster_8	*35273*	*30561*	*15790*	*14543*	*8311*	*14682*	*5221*	*7340*
Average	6440	4434	2486	1814	2107	3101	974	1430

Cluster-4 namely Banyuwangi, Jember, Malang, Mojokerto, Tuban, and Bojonegoro in general have the fourth highest economic potential with 11 categories of business fields each ranking in the top 3, namely: the first highest in category A (Agriculture, Forestry and Fisheries), the first highest in category B (Mining and excavation), the third highest in category D (Procurement of Electricity and Gas), the third highest in category F (Construction), the third highest in category H (Transportation and Warehousing), third highest in category J (Information and communication), third highest in category L (Real Estate), second highest in category O (Government Administration, Defense and Compulsory Social Security), third highest in category P (Education Services), third highest in category Q (Health Services and Social Activities), and the third highest in the RSTU (Other Services) category.

Cluster-2, namely Blitar, Jombang, Kediri , Lamongan, Probolinggo, and Tulungagung have quite high economic potential, namely the second highest in category A (Agriculture, Forestry and Fisheries) and the third highest in category B (Mining and excavation). Cluster-1 namely Lumajang, Madiun, Magetan, Nganjuk, Ngawi, Ponorogo, and Situbondo in general have quite high economic potential in category A (Agriculture, Forestry and Fisheries).

Cluster-3 namely Blitar City, Madiun City, Mojokerto City, Pasuruan City, Probolinggo City. Some potentials that can be developed and optimized are business fields in category C (Industry and Processing), category G (Wholesale and Retail Trade, Car, and Motorcycle Repair), category J (Information and communication), category O (Government Administration, Defense and Compulsory Social Security), and category P (Education Services). Whereas in Cluster-5, namely Bondowoso , Batu City, Pacitan , Pamekasan , Trenggalek, Bangkalan , Sampang , and Sumenep , there are 6 economic potentials that can be developed to boost economic growth, namely business fields in category A (Agriculture , Forestry and Fisheries), category B (Mining and excavation), category C (Industry and Processing), category F (Construction), category G (Wholesale and Retail Trade, Car and Motorcycle Repair), and category J (Information and communication).

In [11][12][13], the mapping is based on the average location quotient which is classified as sector basis and non basis. On [14] using the Internal Cluster Dispersion Rate (ICDRate) to evaluate clustering. While this research is a case of grouping by looking for the optimal number of groups based on Silhouette score and DBS.

There are similarities between this study and research [17], which used a Silhouette score at the clustering evaluation stage. In [17] using 4 attributes for clustering with a fairly high Silhouette value between 0.979 – 0.811 at MinPts 3 – 5 with the number of clusters = 3. The DBSCAN method is better than OPTICS. While in this study using 17 attributes with a fairly high variance and outlier data. The highest Silhouette score is 0.442 in the multistage clustering method (OPTICS – K-Means) with the number of clusters = 8. This method is better than DBSCAN – K-Means, K-Means, and Agglomerative.

CONCLUSION

In the early stages several transformations have been carried out but the normal multivariate assumptions cannot be fulfilled because there are 10 outliers detected in the data. However, the KMO test dataset is sufficient for clustering analysis. Multistage clustering (OPTICS – K_Means) is one solution that can be used as a clustering method for datasets with numeric types that have high variability and outliers. Although this method has a relatively low Silhouette score, it is the best method compared to several other methods, namely K-Means, DBSCAN - K-Means, Agglomerative, and fuzzy clustering. Multistage clustering can be applied and further developed for data sets with a relatively large number of observations.

Acknowledgment

The authors would like to thank for Directorate General Vocational Education, Ministry of Education, Culture, Research and Technology Republic of Indonesia and Center for Research and Community Service of Politeknik Elektronika Negeri Surabaya which has provided funding and facilities for this research.

REFERENCES

[1] Bappenas, "Roadmap of SDGs Indonesia : A Hihglight," pp. 27–36, 2019, [Online]. Available: https://www.unicef.org/indonesia/sites/unicef.org.indonesia/files/2019-07/ROADMAP OF SDGs INDONESIA_final draft.pdf.

[2] N. Oktaviana and N. Amalia, "Gross Regional Domestic Product Forecasts Using Trend Analysis: Case Study of Bangka Belitung Province," *J. Ekon. Stud. Pembang.*, vol. 19, no. 2, 2018, doi: 10.18196/jesp.19.2.5005.

[3] D. Wasani and S. I. Purwanti, "The GRDP Per Capita Gap between Provinces in Indonesia and Modeling with Spatial Regression," *J. Mat. Stat. dan Komputasi*, vol. 19, no. 1, pp. 65–78, 2022, doi: 10.20956/j.v19i1.20997.

[4] N. Kholifia, S. Rahardjo, M. Muksar, N. Atikah, and D. L. Afifah, "Spatial analysis of factors influencing Gross Regional Domestic Product (GRDP) in East Java: A spatial durbin error model analysis," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, 2021, doi: 10.1088/1742-6596/1918/4/042044.

[5] BPS Indonesia, "Berita Resmi Statistik," *Bps.Go.Id*, vol. 19, no. 27, pp. 1–16, 2021.

[6] Dinas kominfo, "Jatim Jadi Provinsi dengan Nilai Tambah Barang dan Jasa Terbesar Kedua di Jawa," *Dinas Kominfo Provinsi Jawa Timur*. 2022, [Online]. Available: https://kominfo.jatimprov.go.id/berita/jatim-jadi-provinsi-dengan-nilai-tambah-barang-dan-jasa-terbesar-kedua-di-jawa.

[7] BPS, "Pertumbuhan Ekonomi Jawa Timur Triwulan II-2020," *Ber. Resmi Stat.*, no. 13, p. 12, 2020, [Online]. Available: https://jatim.bps.go.id/pressrelease/2020/08/05/1141/ekonomi-jawa-timur-triwulan-ii-2020-terkontraksi-5-90-persen.html.

[8] World Bank, "Ketimpangan yang Semakin Lebar: Aku Akhir Untuk Indonesia," pp. 1–33, 2015, [Online]. Available: http://documents.worldbank.org/curated/en/870151468197336991/pdf/101668-BAHASA-WP-PUBLIC-Box394818B-Executive-Summary-Indonesias-Rising-Divide.pdf.

[9] S. N. Aisah, A. Nurcahyani, and D. C. Rini, "Implementasi Fuzzy C–Means Clustering (Fcm) Pada Pemetaan Daerah Potensi Transmigrasi Di Jawa Timur," *J. Tek. Inform. UNIKA St. Thomas*, vol. 07, pp. 33–40, 2022, doi:

10.54367/jtiust.v7i1.1841.

[10] J. Tamaela, E. Sediyono, and A. Setiawan, "Cluster Analysis Menggunakan Algoritma Fuzzy C-means dan K-means Untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara," *J. Buana Inform.*, vol. 8, no. 3, 2017, doi: 10.24002/jbi.v8i3.1317.

[11] M. G. G. N. A. Yuniarti, "Pemetaan Sektor Unggulan Untuk Mendukung Pertumbuhan Ekonomi di Kabupaten Pegunungan Arfak Provinsi Papua Barat Lapangan Usaha Industri Pengolahan Pengadaan Listrik dan Gas Pengadaaan Air , Sampah , dan Daur Ulang Kontruksi Perdagangan Besar dan Eceran ;," *J. Ilm. Muqoddimah*, vol. 5, no. 1, pp. 104–111, 2021, [Online]. Available: https://jurnal.ugm.ac.id/jgs/article/view/63890/30608;

[12] M. Wahed, "Pemetaan_Potensi_Ekonomi_Sektoral_dan_Estimasi_Per," *J. Ekon. dan Bosnis*, vol. Vol.5 No.1, no. 1, pp. 1–16, 2018.

[13] E. D. A. Widyaningrum and H. Cahyono, "Pemetaan Potensi Wilayah Guna Mendorong Pembangunan Ekonomi Kabupaten Ngawi," *J. Din. Ekon. Pembang.*, vol. 3, no. 2, pp. 117–139, 2020, doi: 10.14710/jdep.3.2.117-139.

[14] I. Wahyuni and S. P. Wulandari, "Pemetaan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesejahteraan Rakyat Menggunakan Analisis Cluster Hierarki," *J. Sains dan Seni ITS*, vol. 11, no. 1, 2022, doi: 10.12962/j23373520.v11i1.63092.

[15] H. Prasetyo, "Pengelompokan wilayah menurut potensi ekonomi menggunakan modifikasi algoritme fuzzy k-prototypes untuk penentuan target pembangunan desa Regional clustering based on economic potential with a modified fuzzy k-prototypes algorithm for village developing t," *J. Teknol. dan Sist. Komput.*, vol. 10, no. 1, pp. 46–52, 2022, doi: 10.14710/jtsiskom.2022.14247.

[16] A. Ali *et al.*, "Systematic Review: A State of Art ML Based Clustering Algorithms for Data Mining," *Proc. - 2020 23rd IEEE Int. Multi-Topic Conf. INMIC 2020*, 2020, doi: 10.1109/INMIC50486.2020.9318060.

[17] R. Rizky, A. Rahman, and A. W. Wijayanto, "Pengelompokan Data Gempa Bumi Menggunakan Algoritma Dbscan Grouping Earthquakes Data Using Dbscan Algorithm," *J. Meteorol. dan Geofis.*, vol. 22, no. 1, pp. 31–38, 2021.

[18] J. Liu, Z. Tian, R. Zheng, and L. Liu, "A Distance-Based Method for Building an Encrypted Malware Traffic Identification Framework," *IEEE Access*, vol. 7, pp. 100014–100028, 2019, doi: 10.1109/ACCESS.2019.2930717.

[19] B. Bataineh, "Fast Component Density Clustering in Spatial Databases: A Novel Algorithm," *Inf.*, vol. 13, no. 10, 2022, doi: 10.3390/info13100477.

[20] M. Afzalan, F. Jazizadeh, and H. Eldardiry, "Two-Stage Clustering of Household Electricity Load Shapes for Improved Temporal Pattern Representation," *IEEE Access*, vol. 9, pp. 151667–151680, 2021, doi: 10.1109/ACCESS.2021.3122082.

[21] W. Yang, K. Hou, B. Liu, F. Yu, and L. Lin, "Two-Stage Clustering Technique Based on the Neighboring Union Histogram for Hyperspectral Remote Sensing Images," *IEEE Access*, vol. 5, pp. 5640–5647, 2017, doi: 10.1109/ACCESS.2017.2695616.

[22] M. Y. Saeed, M. Awais, R. Talib, and M. Younas, "Unstructured Text Documents Summarization with Multi-Stage Clustering," *IEEE Access*, vol. 8,

pp. 212838–212854, 2020, doi: 10.1109/ACCESS.2020.3040506.

[23] M. Hahsler, M. Piekenbrock, and D. Doran, "Dbscan: Fast density-based clustering with R," *J. Stat. Softw.*, vol. 91, no. 1, 2019, doi: 10.18637/jss.v091.i01.

[24] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure," *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)*, vol. 28, no. 2, pp. 49–60, 1999, doi: 10.1145/304181.304187.

[25] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 248–254, 2020, doi: 10.14569/IJACSA.2020.0110832.