ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 11 (2021) No. 4, pp. 87-112
DOI:10.33168/JSMS.2021.0405

Spatial-temporal Analysis using Two-stage Clustering and GIS-based MCDM to Identify Potential Market Regions

Ernawati[1,*], Safiza Suhana Kamal Baharin[2] and Fauziah Kasmin[2]

[1]Universitas Atma Jaya Yogyakarta, Indonesia

[2]Universiti Teknikal Malaysia Melaka, Malaysia

ernawati@uajy.ac.id (corresponding author); safiza@utem.edu.my; fauziah@utem.edu.my

Abstract. Promotion is essential in a competitive environment. Promotion to the right areas increases success and saves resources. However, due to Indonesia's vast territory and numerous regions of origin school, universities with student markets from all over the country must select target areas for promotion to meet their objectives and save resources. Unlike for-profit businesses, besides quantity factors, educational institutions need to consider student quality factors in selecting promotional locations. This study aims to conduct a data-driven spatio-temporal analysis to identify potential regions for university promotions targets. This study uses enrollment and academic data from one private university in Indonesia for the empirical study. In Geographic Information System (GIS) environment, the origin schools' locations were geocoded, and various thematic maps were analyzed. This study integrates two-stage clustering and GIS-based multi-criteria decision-making (MCDM) to identify potential market regions. A potential region is one that consistently sends many qualified students. First, time-series clustering is conducted to groups regencies/cities based on the enrolled students' patterns over time in the university. Subsequently, the origin schools' regencies/cities were clustered using the k-prototypes algorithm based on their time-series pattern category, the consistency in sending students, average cumulative grade point average (CGPA), and dropout (DO) rate. The clusters are scored using the sum weighting method. The highest valued cluster that consists of eight regencies and 18 cities that consistently contributed high quantity and quality students were selected as the priority regions. The proposed approach's results were compared to the Simple Additive Weighting (SAW) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) methods for evaluation. The proposed method can assist the university management in determining potential regions for promotion purposes.

Keywords: GIS, promotion, spatial temporal analysis, clustering, MCDM.

1. Introduction

Recognizing a potential market is critical to the success of businesses. Adopting information technology into companies' business operations enables them to explore potential markets using data-driven analysis. Unified information is a crucial component that can be used to put optimal decision-making knowledge into practice (Mekvabidze, 2020). As the market becomes more competitive, businesses seek more valuable insights into their customers, and data mining tools can assist in better understanding their customers (Dahiya et al., 2021). By empowering data mining techniques, businesses can extract hidden knowledge from their customers' data and use it to develop more efficient marketing strategies and personalize promotional offers (Roshan & Afsharinezhad, 2017). Likewise, in Higher Education Institutions (HEIs), with large volumes of student data such as registration and academic data, HEIs can utilize data mining to explore students' historical data to find targeted customers (Abaya & Gerardo, 2013).

High schools are students' source for universities because most prospective students are senior high school fresh graduates. The number of senior high schools and vocational high schools in Indonesia, the world's fourth most populous country, reaches 28,240 (Badan Pusat Statistik, 2020). They represent a potential market for universities, particularly for most universities in Daerah Istimewa Yogyakarta Province, a special region where many out-of-town students study. The schools are spread over a very wide area (1,196,862.20 km^2) in 514 regencies/cities (BPS-Statistics Indonesia, 2019), so school-to-school promotion and recruitment require huge resources. However, student populations tend to congregate in specific geographic locations (Tang & Mcdonald, 2002), so selecting target areas for promotion and recruitment needs to be done to achieve university goals and save resources.

Previous studies identified target school districts using a model based on spatial analysis. Ayad (2007) recommended potential school districts for recruitment using the gravity index. The model considers parameters such as SAT takers in a specific year, median household income, and driving distance to the university. Martin (2001) developed a model to measure the school district's potential to increase the university enrollment of minority populations. It depends on the district's number of university-bound students and is inversely proportional to the distance between the university and the school district. The target school district is determined based on the potential of the school district, the ethnic composition of university-bound students, and the "quality" of the district as measured by SAT score. These models used single-year data. On the other hand, enrollment is a trend (Morris & Thrall, 2010), so detecting changes in school enrolment patterns should be done by including university-bound features for school districts over a longer period (Martin, 2001). Therefore, in addition to spatial analysis, the temporal dimension of spatial data should be examined; as suggested by Kelly (2019), geospatial education researchers should pay

more attention to the temporal dimensions of spatial data.

The purpose of this research is to identify potential regions for university promotion and recruitment targets. The spatial and temporal analysis is conducted using historical data from one university in Daerah Istimewa Yogyakarta Province to discover prospective areas that meet both quality and quantity objectives. The cumulative grade point average (CGPA) and students' dropout (DO) rate at the end of the fourth semester of the university studies are used to assess the quality. Besides the fact that many students drop out during their first and second years of study (Ousley, 2010), this data also assesses students' readiness to study at a university during their early years of study. The quantity aspect is represented by time-series data on enrolled students and the region's consistency in sending students to the university. The motivation of this study is to utilize internal institutional data to support higher education institution management in developing marketing strategies.

Location is critical in marketing (Libório et al., 2020). Many location analyses have been used to reveal particular spatial patterns or identify suitable locations for resource management, service provision, and business decisions (Tong & Murray, 2017). Moreover, researchers also can utilize big data to review location theories, develop location applications and frameworks, and integrate Geographic Information System (GIS) more deeply into location analysis (Tong & Murray, 2017). GIS is a useful and effective tool for assisting decision-making in spatial analysis (Chacón-garcía, 2017; Oliveira et al., 2020). GIS could provide a more comprehensive and holistic approach, where decisions are supported by visual solutions (Jurisic et al., 2016). In the education domain, GIS is applied to analyze educational facilities distribution (Lagrab & Aknin, 2015; Murad et al., 2020), accessibility of education service (Bulti et al., 2018), provision of educational services (Constantinidis, 2019), which is beneficial for facility management and site planning.

The spatio-temporal analysis involves temporal aspects of spatial data. The spatio-temporal analysis is a technique for identifying and visualizing trends in spatial patterns over time (Rosyidah & Surjandari, 2019). Spatio-temporal analysis has been applied in business practices for supporting decision-making processes in marketing and customer relationship development, investment, resource, and asset management (Surjandari & Rosyidah, 2017). In the education domain, spatio-temporal analysis has been used to investigate the spatial distribution of schools over time (Ghodousi, et al., 2020; Zhang et al., 2020).

Because many factors influence location decisions, some studies combined GIS and multi-criteria decision-making (MCDM) to solve the site selection problem. Herlawati et al. (2020) proposed GIS-based MCDM to analyze the suitable district for central business. The weighted sum method was applied to six criteria of the overlaid layers. The weighted sum is popular in the GIS-MCDM approach because it is straightforward to implement in the GIS environment utilizing algebra operations and simple to grasp (Malczewski, 2006). Another popular MCDM technique that has

been applied with GIS is Analytic Hierarchy Process (AHP) (Chacón-garcía, 2017; Saha et al., 2020; Yıldırım, 2021). GIS-based MCDM is also employed in the education domain to identify the most proper school location and area (Baser, 2020; Lagrab & Aknin, 2017; Prasetyo et al., 2018). GIS-based MCDM has demonstrated promising results because it considers geographical data in decision-making (Herlawati et al., 2020; Saha et al., 2020).

Besides many criteria in decision making, location selection frequently includes many alternatives, so data mining techniques are involved to group data. Clustering is one of these techniques used to group similar data into the same cluster while placing different data in different clusters. Aryaee (2019) used GIS, clustering, and MCDM to choose the locations of bank branches. The best worst method was used to determine the weight of criteria and sub-criteria. After importing the criteria, sub-criteria, and weights into GIS layers and analyzing them on a district map grid, the 34 areas with the highest priority were chosen as candidate points. The Fuzzy C-means method was used to cluster the candidate points, and then the most important cluster was determined using the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method. The Complex Proportional Assessment (COPRAS) method prioritizes the points in the most important cluster, and the proper locations are determined. However, this study has not yet included the temporal aspect.

This research uses GIS-based spatio-temporal analysis, two-stage clustering, and the MCDM method. The origin schools' spatial distribution is examined using Global Moran's I test. Time-series clustering identifies the temporal pattern of the school region based on its enrolled students, while the mixed-data type clustering is employed to cluster schools' regions based on the regions' temporal pattern, average CGPA, DO rate, and consistency in sending students. The sum weighting method is applied to determine the potential cluster as a promotion target. The potential regions identified by the proposed approach are compared to the Simple Additive Weighting (SAW) (Hwang & Yoon, 1981; Sembiring et al., 2019), and TOPSIS (Hwang & Yoon, 1981) results for evaluation. The contributions of this paper are summarized as follows. This study proposes a novel approach integrating two-stage clustering: time-series clustering and mixed-data type clustering with GIS-based MCDM to support higher education institutions' decision-making. This study considers not only spatial but also temporal aspects. The empirical findings of this research show that the combination of spatial and temporal analysis, two-stage clustering, and GIS-based MCDM successfully identified potential regions and may be recommended for university promotion targets.

2. Theoretical background

2.1. Global Moran's I

Global Moran's I statistics, first proposed by Moran in 1948 (Anselin, 2020), is a commonly used method to investigate global spatial autocorrelation. It is useful to

reveal whether the data have clustered or random patterns. Moran's I inference is based on the null hypothesis of spatial randomness on data. In Geoda software (Anselin, 2020), there are six outputs of Global Moran's test, namely the Moran's index (I), the expected values (E(I)), the mean, the standard deviation, the z-score, and the pseudo-p-value. If I > E(I) or the z-score is high with a p-value less than the significance value, the null hypothesis is rejected, indicating that the data is clustered. The Moran's I index is calculated using Equation (1) (Ghodousi et al., 2020):

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{S_0 \sum_{i=1}^{n} z_i^2} \quad (1)$$

where n denotes the number of observations, z_i denotes the deviation of each observation from the mean, w_{ij} denotes the spatial weight between observations i and j, and S_0 is the sum of all spatial weights.

2.2. Time-series clustering

Clustering is the process of partitioning observations into clusters of similar objects without having prior knowledge of the clusters. Objects in the same cluster are like one another, but not to objects in other clusters. The process of grouping time-series data based on their similarity is known as time-series clustering (Guijo-Rubio et al., 2020). Time-series clustering is classified into three types: whole time-series clustering, subsequence clustering, and time-points clustering (Aghabozorgi et al., 2015). A set of individual time-series data is clustered based on their similarity in whole time-series clustering. Subsequence clustering objects are obtained by dividing a single long time-series data based on similarity. In time-points clustering, the clustering objects are single points from a single time-series. In this study, the whole time-series clustering is performed to identify the regions' pattern shapes based on the number of enrolled students in the university.

For matches time-series based on shapes, two time-series data are aligned by a non-linear stretching and contracting of the time axes. This method mainly uses traditional clustering methods and works with the raw time-series data directly. Shape-based methods are proper for short time-series data (Aghabozorgi et al., 2015). The most popular shape-based similarity measures are the Euclidean distance and Dynamic Time Warping distance (DTW) (Aghabozorgi et al., 2015; Guijo-Rubio et al., 2020). Although Euclidean distance is the most efficient (Guijo-Rubio et al., 2020), the DTW method is the most accurate (Aghabozorgi et al., 2015; Javed et al., 2020), so DTW is used in this study along with Agglomerative Hierarchical Clustering (AHC) for time-series clustering (Abbasimehr & Shabani, 2020).

A hierarchical clustering method groups objects in a dataset into a dendrogram, a hierarchy of cluster structures (Han et al., 2012). It is classified as either an agglomerative or a divisive approach. AHC starts with each object in a separate cluster, then gradually merges the clusters based on similarity until all objects are grouped into a single cluster. The basic AHC algorithm is comprised of the following steps (Bramer, 2016):

1. For each object, make a single-object cluster and determine the distance between each pair of clusters.
2. Select the two closest clusters and merge them into a single cluster.
3. Calculate the distance between each of the old and new clusters.
4. Repeat steps 2 and 3 until all objects are grouped in a single cluster.

The DTW's goal is to discover correspondences between two time-series sequences. In the case of two time-series, A = (a_1, a_2, …, a_n) and B = (b_1, b_2, …, b_m), the DTW algorithm would first create an n x m matrix to minimize discrepancies between A and B. Calculate the element of the matrix (M_{ij}) starting from the left bottom corner using equation (2):

$$M_{ij} = \left|a_i - b_j\right| + \min\left(M_{(i-1)(j-1)}, M_{i(j-1),} M_{(i-1)j}\right) \quad (2)$$

Subsequently, starting from the top right corner element of the matrix, find its neighbor with the minimum value as the traversal path until it reaches the bottom left to find the warping path P = (p_1, p_2, ..., p_K). Finally, the DTW distance between A and B can be calculated by Equation (3) (Sammour et al., 2019):

$$d_{AB} = min \frac{\sum_{k=1}^{K} p_k}{K} \quad (3)$$

2.3. K-prototypes mixed-data types clustering

Besides hierarchical clustering, another type of clustering algorithm is partitioning clustering. K-means is the most used partitioning algorithm. However, because this algorithm is only suitable for numerical data, several developments, such as k-modes used to group categorical data and k-prototypes used to group mixed data, have been made to overcome its shortcomings (Huang, 1998). Sulastri et al. (2021) clustered schools using k-prototypes based on student admission data. This current study uses the k-prototypes algorithm to cluster origin schools' regions based on numerical data (average CGPA and DO rate) and categorical data (origin schools' regions temporal pattern and consistency in sending students). The k-prototypes algorithm divides a dataset with n objects, p numerical attributes, and m-p categorical data into k different clusters by minimizing the cost function value in equation (4) (Huang, 1998; Jia & Song, 2020):

$$F(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{il} d(x_i, q_l) \quad (4)$$

subject to equation (5):

$$\sum_{l=1}^{k} w_{il} = 1, \;\; w_{il} \in \{0,1\}, \;\; 1 \le i \le n, \;\; 1 \le l \le k \quad (5)$$

where W is an n×k partition matrix, Q = {q_1, q_2, ..., q_k} is the cluster center set, q_l is the l-th cluster's center, and d(x_i,q_l) is the dissimilarity measure between data object and cluster center. Huang (1998) integrated Euclidean distances (used in k-means) to calculate numerical data dissimilarity and the simple matching method (used in k-modes) for categorical data dissimilarity in the k-prototypes algorithm. The

dissimilarity between two mixed-type data X and Y, which has p numerical attributes and m-p categorical attributes, can be calculated by equation (6):

$$d(X,Y) = \sum_{i=1}^{p}(x_i - y_i)^2 + \gamma \sum_{i=p+1}^{m} \delta(x_i, y_i) \quad (6)$$

The first term in equation (6) is the Euclidean distance between X and Y, while the second term is the simple matching of X and Y and its weight (γ). The dissimilarity measure of X and Y using the simple matching method is defined in equation (7) (Huang, 1998):

$$\delta(x_i, y_i) = \begin{cases} o \; if \; x_i = y_i \\ 1 \; if \; x_i \neq y_i \end{cases} \quad (7)$$

The basic steps of the k-prototypes algorithm are as follow (Jia & Song, 2020):

1. Select k data objects randomly from the dataset as the initial cluster centers.
2. Compute the dissimilarity between object x_i and the cluster center q_l using equation (6), then assign x_i to the nearest prototype cluster. Apply this step to all x_i objects in the dataset and the k cluster centers.
3. Determine the new cluster centers. Use the average value for numerical attributes and the highest frequency (mode) for categorical attributes as the center.
4. Repeat steps 2 and 3 until the cost function in equation (4) is no longer changes.

Like the k-means algorithm, the number of clusters must be specified before using the k-prototypes algorithm, so this study employs k-prototypes and the well-known Elbow method. The Elbow method uses a sum of squared errors (SSE), which is the variation of each cluster member concerning its centroid.

2.4. Multi-criteria decision-making (MCDM)

MCDM is a problem-solving methodology that includes many criteria and alternatives that must be considered when deciding. MCDM methods such as SAW, TOPSIS, AHP, Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE), and COPRAS are widely used by decision-makers. Aside from CORPAS, SAW and TOPSIS are regarded as the most efficient and user-friendly MCDM methods (Sotoudeh-Anvari et al., 2018).

The basic idea behind SAW is to assess the weighted sum of each alternative's criteria score (Sembiring et al., 2019). For example, if the problem has m criteria (C_1, C_2, …, C_m) with corresponding weights (w_1, w_2, …,w_m) and n alternatives (A_1, A_2, …, A_n), the SAW solution steps are as follows. First, create a decision matrix (X_{nxm}) and normalize it to get a normalized decision matrix (R_{nxm}) using equation (8):

$$r_{ij} = \begin{cases} \dfrac{x_{ij}}{\max(x_{ij})}; \; if \; C_j \; is \; a \; benefit \; criterion \\ \dfrac{\min(x_{ij})}{x_{ij}}; \; if \; C_j \; is \; a \; cost \; criterion \end{cases} \quad (8)$$

Then, for each alternative, compute the weighted sum $S_i = \sum_{i=1}^{n} w_i r_{ij}$. The better the alternative, the higher the S_i value.

The fundamental idea behind TOPSIS (Hwang & Yoon, 1981) is that the best alternative should be the one that is closest to the best solution and farthest apart from the worst (Zemlickienė, 2019). The TOPSIS method, like the SAW method, solves the problem through the normalization steps of the decision matrix using equation (9):

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \tag{9}$$

The weighted normalized decision matrix $s_{ij} = w_i r_{ij}$ is then computed. Using equations (10) and (11), calculate the positive ideal solution $A^+ = (s_1^+, s_2^+, \dots, s_m^+)$ and the negative ideal solution $= (s_1^-, s_2^-, \dots, s_m^-)$.

$$s_j^+ = \begin{cases} \max(s_{ij}) ; \ if\ C_j\ is\ a\ benefit\ criterion \\ \min(s_{ij}); \quad if\ C_j\ is\ a\ cost\ criterion \end{cases} \tag{10}$$

$$s_j^- = \begin{cases} \min(s_{ij}) ; \ if\ C_j\ is\ a\ benefit\ criterion \\ \max(s_{ij}); \quad if\ C_j\ is\ a\ cost\ criterion \end{cases} \tag{11}$$

The next step is to determine the distances of each alternative from the positive ideal solution (d_i^+) and the negative ideal solution (d_i^-) using equations (12) and (13):

$$d_i^+ = \sqrt{\sum_{j=1}^{m} (s_{ij} - s_j^+)^2} \tag{12}$$

$$d_i^- = \sqrt{\sum_{j=1}^{m} (s_{ij} - s_j^-)^2} \tag{13}$$

Finally, compute the relative closeness of each alternative to the ideal solution (D_i) using equation (14). The higher the Di score, the better the alternative.

$$D_i = \frac{d_i^-}{d_i^- + d_i^+} \tag{14}$$

3. Study area, materials, and methods

3.1. The study area

The study area is the country of Indonesia, and the spatial analysis unit is the administrative regency/city boundary. Indonesia has 418 regencies and 96 cities spread across 34 provinces. These provinces are spread across five main islands and four archipelagos, as shown in Figure 1, namely: Sumatra Island, Kalimantan Island, Sulawesi Island, Java Island, Papua Island, Riau Archipelago, Bangka Belitung Archipelago, Nusa Tenggara Archipelago, and Maluku Archipelago (BPS-Statistics Indonesia, 2019).

3.2. Data and tools

This study uses enrollment and academic data from one Indonesian private university in Daerah Istimewa Yogyakarta. This province is located on Java Island. Indonesia

province and regency shapefile are obtained from the Indonesian geospatial portal (www.tanahair.indonesia.go.id), while the latitude and longitude of the high schools' locations are obtained from Google Map. Data preprocessing and analysis were carried out using R version 4.0.4. We used QGIS 3.18 and Geoda 1.18.0.0 for spatial analysis and visualization in a GIS environment.

Fig. 1: The study area

3.3. Methods

Figure 2 depicts the research framework. First, data for analysis are collected and preprocessed. Data preprocessing involves integrating, selecting, and cleaning data. This study uses 11,861 data of students enrolled in 2014 to 2018 that cleaned down to 11,199 records. This study aggregates students to their feeder school and includes 1,799 feeder schools from 338 regencies/cities for analysis. For spatial analysis, the feeder school locations are geocoded and overlaid on the regency/city shapefile. The Global Moran's I spatial autocorrelation test is conducted to see the spatial pattern of the school distribution. Various thematic maps are depicted and analyzed. Meanwhile, to determine potential areas, two-stage clustering and MCDM methods are performed. Finally, for assessing the outcome, the proposed method's results are compared to the results of two commonly used MCDM methods, SAW (Hwang & Yoon, 1981; Sembiring et al., 2019) and TOPSIS (Hwang & Yoon, 1981).

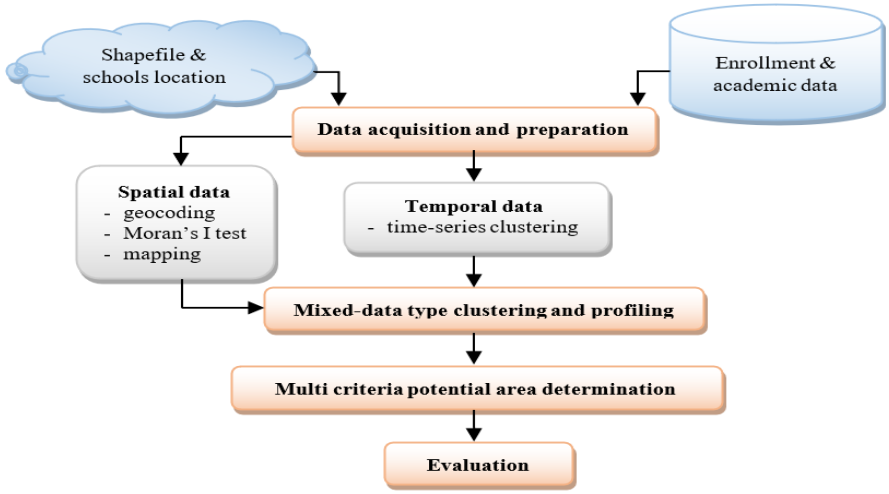

Fig. 2: The research framework

Figure 3 presents the flowchart of two-stage clustering and MCDM for identifying potential areas. Time-series clustering is performed first to identify relatively homogeneous regions' groups with similar student enrollment patterns over time. The enrolled students' trends are identified and analyzed. The time-series clustering is carried out using the AHC method and DTW similarity measure, whereas for determining the number of clusters, this study employs the Elbow method and Calinski Harabasz validity index. The cluster categories which state the same enrolled students' pattern over time produced by time-series clustering become one of the inputs for the second stage clustering. The second stage clustering employs the k-prototype since the data are mixed-data type. This step groups regions based on time-series cluster patterns, consistency in sending students, average CGPA, and average DO rates. Before clustering, the average CGPA and DO rates were normalized using Z-score normalization, and the consistency was discretized into five categories: inconsistent, less consistent, moderately consistent, consistent, and very consistent. Then, profiling is done based on the clustering results to determine the characteristics of each cluster. Subsequently, each feature in the cluster center is ranked and scored based on its importance to university promotion for identifying potential regions. The maximum score is equal to the number of clusters, and the lowest score is 1. If the ranks are equal, the score receives the average value of the existing scores. A cluster's score is the weighted sum of its criteria scores. In this case, all criteria are given equal weight. The cluster with the highest score is chosen as the potential cluster. The regencies/cities in the potential cluster are the potential regions for promotion targets.

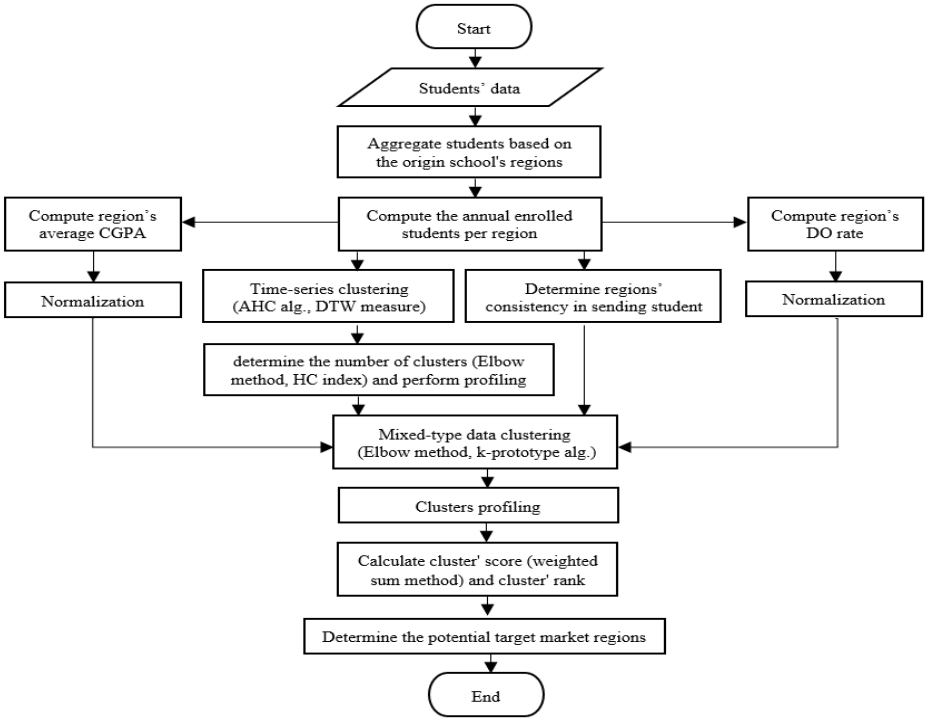

Fig. 3: The flowchart for determining potential regions

4. Results and discussion

4.1. Spatial analysis

For spatial analysis, the number of enrolled students is mapped at the origin school level. Figure 4 depicts the distribution of origin schools and their contribution to enrollment in the 2014-2018 period. Each bubble represents an origin school, and its size corresponds to the number of students enrolled. The map shows that the origin schools are spread throughout Indonesia but are denser around the university and several other regions. It is supported by the Global Moran's I test result, which can be seen in the right up corner of Figure 4. Based on Moran's index I = 0.5062, z-value = 9.1697, and pseudo-p-value = 0.001 (less than 0.05), the null hypothesis that origin schools are randomly distributed is rejected, concluding that origin schools are clustered. Hence, knowing where the origin schools are clustered is essential.

The origin schools are in 338 (66%) of Indonesia's 514 regencies/cities, spanning all provinces. Two hundred forty-eight regions (73%) are regencies administrative areas, while 90 regions (27%) are city administrative areas. Table 1 shows the distribution of the original schools' regencies/cities in each province. The table compares the number of regencies and cities where the origin school is located to the total number of regencies/cities in the province. For example, in Aceh Province,

schools of origin are found in one of the 18 regencies and four of the five cities. According to the data, the origin schools came from almost all existing cities (92%), except for a few provinces on the Sumatra Island and the Maluku Archipelago. On the other hand, the origin schools are only located in 59% of all existing regencies. Therefore, it is critical to pay attention to which regions the promotions will take place, implying that management needs to devise different marketing strategies to reach out to different regions.

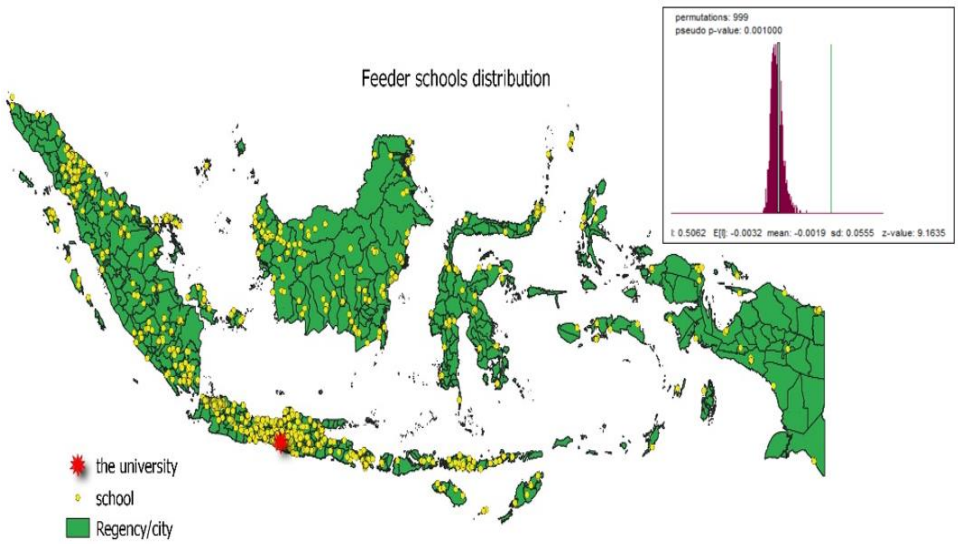

Fig. 4: The feeder schools distribution and the Moran's I test result

Among 1,799 origin schools, 922 (51%) were in regencies and 877 (49%) in cities. Figure 5 depicts a thematic map of the number of school origins in each regency/city. On average, each region had about five schools of origin, three to four origin schools in each regency, and nine to ten origin schools in each city. According to school type, 84% were senior high schools, 14% were vocational high schools, and 2% were others like Madrasa, Islamic boarding school, homeschooling. According to the school owner, 59% were public, while 41% were privately owned. Most of the public schools (59%) were in the regencies. In contrast to the private schools, most were in cities (60%). City schools had an average DO rate of 3.74%, while the regency schools had 4.08%. The average CGPA of origin schools was 2.97 (on the scale of 4), greater than in the regencies (2.90). Schools in cities were also more consistent in sending students, on average, four times in five years, while schools in regencies were three times in five years.

Table 1: Distribution of the origin schools' regencies/cities in each province

Province	Number of regencies	Number of cities	Province	Number of regencies	Number of cities
Sumatera Island:			Kalimantan Island:		
11. Aceh	1 (18)	4 (5)	61. Kalimantan Barat	12 (12)	2 (2)
12. Sumatera Utara	18 (25)	8 (8)	62. Kalimantan Tengah	11 (13)	1 (1)
13. Sumatera Barat	2 (12)	4 (7)	63. Kalimantan Selatan	5 (11)	2 (2)
14. Riau	8 (10)	2 (2)	64. Kalimantan Timur	6 (7)	3 (3)
15. Jambi	4 (9)	1 (2)	65. Kalimantan Utara	3 (4)	1 (1)
16. Sumatera Selatan	8 (13)	3 (4)			
17. Bengkulu	4 (9)	1 (1)	Sulawesi Island:		
18. Lampung	12 (13)	2 (2)	71. Sulawesi Utara	5 (11)	4 (4)
			72. Sulawesi Tengah	8 (12)	1 (1)
Bangka Belitung Archipelago:			73. Sulawesi Selatan	4 (21)	3 (3)
19. Kepulauan Bangka Belitung	5 (6)	1 (1)	74. Sulawesi Tenggara	1 (15)	2 (2)
			75. Gorontalo	0 (5)	1 (1)
Riau Archipelago:	3 (5)	2 (2)	76. Sulawesi Barat	1 (6)	0 (0)
21. Kepulauan Riau					
			Maluku Archipelago:		
Java Island:	0 (1)	5 (5)	81. Maluku	7 (9)	1 (2)
31. DKI Jakarta	15 (18)	9 (9)	82. Maluku Utara	3 (8)	1 (2)
32. Jawa Barat	29 (29)	6 (6)			
33. Jawa Tengah	4 (4)	1 (1)	Papua Island:		
34. DI Yogyakarta	22 (29)	9 (9)	91. Papua	10 (12)	1 (1)
35. Jawa Timur	3 (4)	4 (4)	92. Papua Barat	5 (28)	1 (1)
36. Banten					
Nusa Tenggara Archipelago:	8 (8)	1 (1)	TOTAL	248 (416)	90 (98)
51. Bali	4 (8)	2 (2)			
52. Nusa Tenggara Barat	17 (21)	1 (1)			
53. Nusa Tenggara Timur					

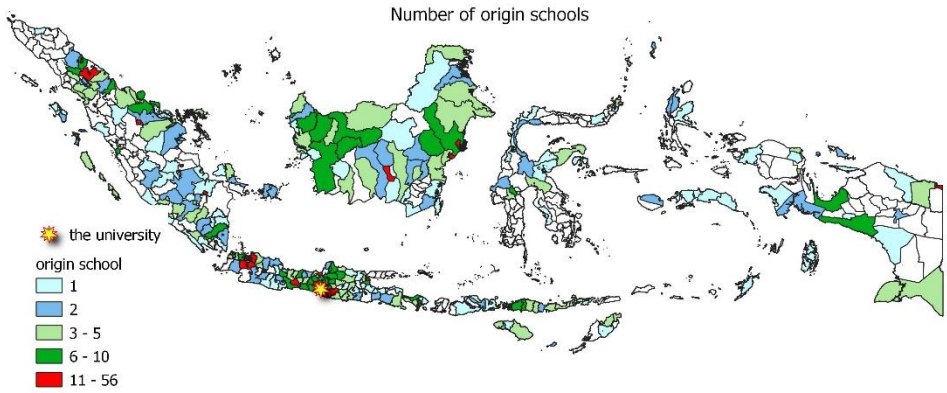

Fig. 5: Number of origin schools per regency/city

Figure 6 depicts the total number of enrolled students by regency/city. Each region sent an average of 33 students, with a minimum of 1 and a maximum of 2,113 students. The city of Yogyakarta, in the province where the university is located, supplied the most students. The thematic map can show areas with no enrollment (white color region) and other low-to-high student enrollment regions, which is useful for developing promotion planning. Most of the students came from schools located in cities (61%). Students from senior high schools accounted for 94% of the total, and most of them (64%) are graduated from private schools. Seventy percent of students came from city private schools, while for the regency, 55% came from public schools. The ratio of male to female students was 53:47, with city schools accounting for 58% of male students. Simultaneously, 64% of female students attended city schools.

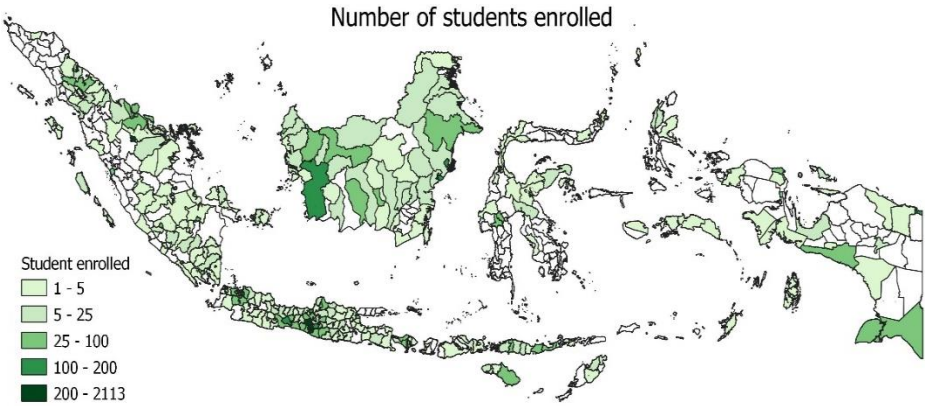

Fig. 6: The distribution of students enrolled in 2014-2018

4.2. Temporal analysis

After it is proven that the origin schools are clustered, two-stage clustering is used to determine the characteristics of the schools and determine potential market regions for promotion targets. This subsection explains the result of time-series clustering as

the first stage of the two-stage clustering. The AHC algorithm is used to group regions based on five-year time-series data of enrolled students. Based on the Elbow method and the Calinski Harabasz (CH) index, the optimum number of clusters (k) is five. The Elbow method graphic in Figure 7 shows that there is an elbow at k=5. Before k=5, there is a significant decrease in the SSE, but as k increases after k=5, the SSE does not decrease significantly. Likewise, the Calinski Harabasz graphic in Figure 7 shows that the maximum value of the CH index occurs at k=5.

The 338 regions were then clustered into five groups based on similar temporal patterns of enrolled university students. The time-series patterns of students enrolled in the 2014-2018 school year are presented in Figure 8. According to Figure 8, the profiles of the five clusters are described as follow:

- The first time-series cluster (TSC1) comprises 71 cities and 239 regencies that sent a very small number of students with minor fluctuations. Each region sent two students per year on average.
- TSC2 contains five cities and three regencies, sent a moderately large number of students, an average of 44 students per year, which had somehow increasing trend.

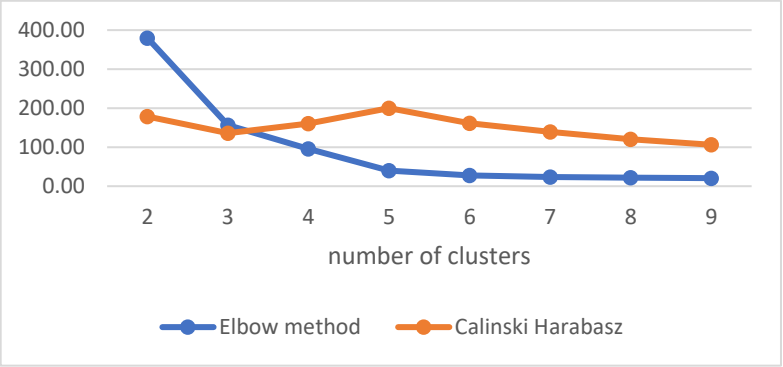

Fig. 7: Elbow method and Calinski Harabasz index for determining the number of clusters

- TSC3 includes 12 cities and six regencies, sent roughly 22 students annually with a horizontal trend and minor fluctuation.
- TSC4 contains one city and one regency, sent an average of 97 students per year. These regions experienced quite high fluctuations in the number of enrolled students. One of them is the regency where the university is located, whereas another is the city in the nearest province to the university.
- TSC5 includes the closest city to the university, sent the most students, approximately 422 students per year. The number of students fluctuated, with a downward trend in the last year.

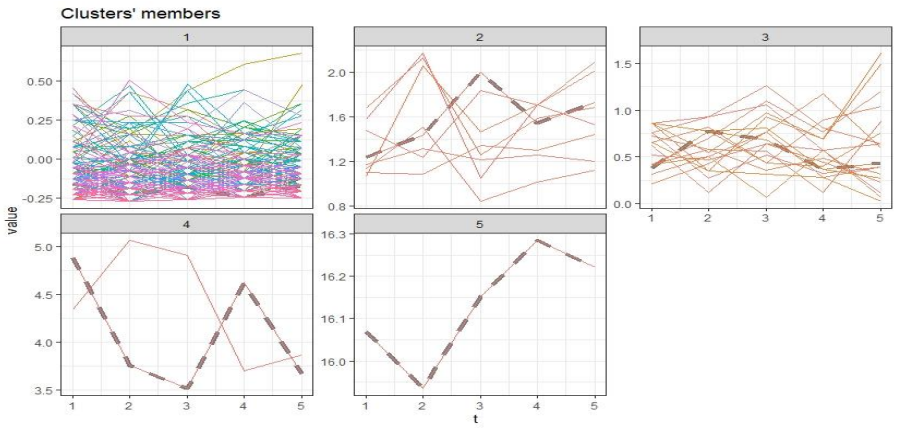

Fig. 8: Time-series patterns of the number of students enrolled in 2014-2018 for clusters: TSC1 (1), TSC2 (2), TSC3 (3), TSC4 (4), TSC5 (5)

4.3. Mixed-data type clustering and profiling

The k-prototype was applied on the second stage of the two-stage clustering to identify potential regions by grouping the average CGPA and DO rate (numerical data), the time-series clusters of the students enrolled, and consistency in sending students (categorical data). According to the Elbow method, as shown in Figure 9, we chose seven clusters rather than ten as the optimum number of clusters because it is easier to interpret for a small number of clusters (Ernawati et al., 2021) and does not produce clusters with a too-small number of members.

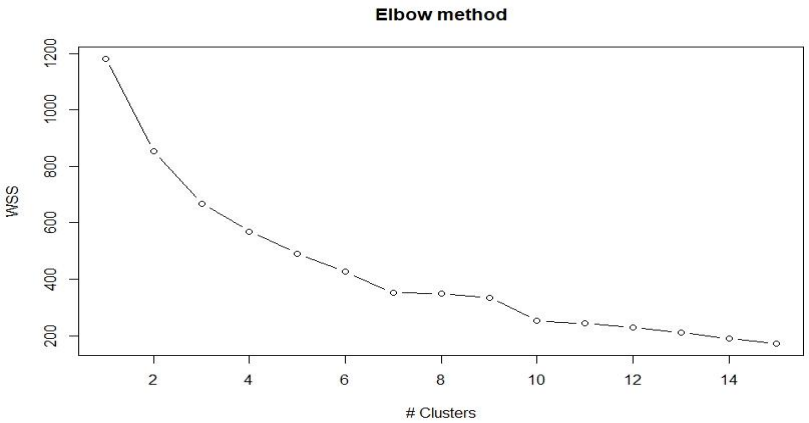

Fig. 9: The Elbow method for determining the optimum number of cluster

Each cluster region is categorized in a thematic map after clustering the input data into seven clusters using the k-prototypes algorithm. Figure 10 depicts the thematic maps of the input and output variables of the mixed-data type clustering.

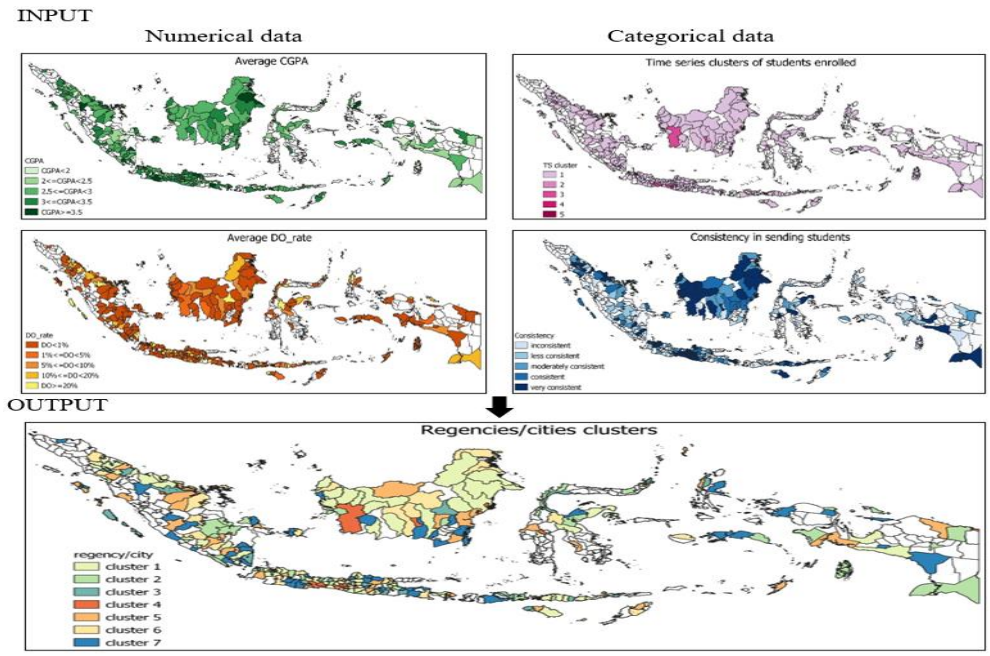

Fig. 10: The thematic map of the input and output variable of k-prototypes

Based on the results of the k-prototypes algorithm as shown in Figure 11, the profiles of each cluster are summarized as follows:

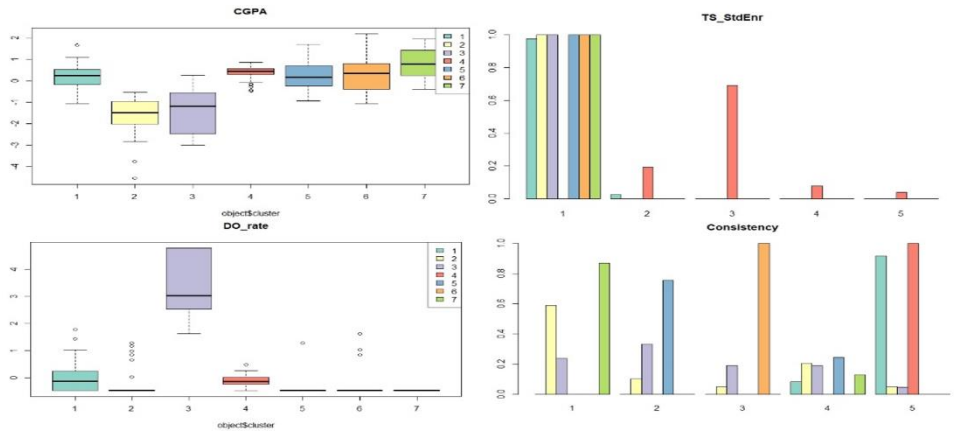

Fig. 11: Visualization of the k-prototypes clusters' characteristics

- Cluster1 consists of 35% regions (43 cities and 77 regencies). These regions contain the biggest number of origin schools (52.47%) that supplied 36.12% of students. The majority of them are very consistent in sending students, and the majority have a TSC1 pattern, with only a few having a TSC2 pattern. They have high-quality students with an average CGPA of 2.91 and a DO rate of 4.16%.

- Cluster2 contains 12% of the regions (5 cities and 34 regencies), with all regions exhibiting a TSC1 pattern. In terms of sending students, this cluster contains regions that are inconsistent to very consistent, but the majority are inconsistent. Despite the low DO rate (2.22%), academic performance is poor (an average CGPA of 2.12). This cluster contains 4.22% origin schools, which sent 1.95% of students.
- Cluster3 has 6% regions (2 cities and 19 regencies), all of which have a TSC1 pattern. The majority are areas that consistently send students, but have the lowest quality (DO rate 33.21% and CGPA 2.20). This cluster sent the fewest students (0.82%) from 2.83% origin schools.
- Cluster4 includes 8% regions (18 cities and eight regencies) with TSC2, TSC3, TSC4, and TSC5 patterns and is very consistent in sending students, so supplied the most students (57.04%). This cluster have 26.57% origin schools with very good quality students (CGPA 2.98 and DO rate 3.44%).
- Cluster5 comprises 13% of the regions (eight cities and 37 regencies), with all regions following a TSC1 pattern and the majority of them being less consistent in sending students. It has a very low DO rate (0.37%) and a CGPA of 2.93 on average. These regions contain 5.39% origin schools and 1.69% students.
- Cluster6 comprises 10% regions (7 cities and 26 regencies), all of which have a TSC1 pattern and are moderately consistent in sending students. It has an average DO rate of 2.46% and an average CGPA of 2.97. This cluster contains 4.50% origin schools which sent 1.43% of students.
- Cluster7 has 16% of the regions (7 cities and 47 regencies) that sent the best students (no dropout and CGPA 3.16). It has a TSC1 pattern, and most of them are inconsistent in sending students, so supplied only 0.96% of students. There are 4.00% of origin schools in these areas.

4.4. Multi-criteria potential area determination

As this study employs a weighting sum of the variable scores for deciding the best potential cluster, each cluster center criterion from the seven clusters is ordered based on its value in decision making and given a score with a maximum score of 7 and a minimum of 1. The cluster with the highest CGPA received a score of 7, while the lowest received a score of 1, whereas the DO rate was the inverse. The cluster that is more consistent in sending students is given a higher score. As this study employs a weighting sum of the variable scores for deciding the best potential cluster, each cluster center criterion from the seven clusters is ordered based on its value in decision making and given a score with a maximum score of 7 and a minimum of 1. The cluster with the highest CGPA received a score of 7, while the lowest received a score of 1, whereas the DO rate was the inverse. The cluster that is more consistent in sending students is given a higher score. A higher score is given to a cluster with

an enrolled student pattern that contributes more students. Since the enrolled student pattern category for Clusters 2,3,4,6,7 is the same, their score is the average value of their rank (5,4,3,2 and 1), which equals 3. The cluster score is a weighted sum of its variable scores, and we assigned the same weight to each criterion (0.25). For example, Cluster1's score is (3*0.25+2*0.25+6*0.25+6*0.25)=4.25. Table 2 displays the results of the given scoring and the ranking of each cluster.

According to Table 2, Cluster4 has the highest ranking, thus chosen as the target market. This cluster consists of 26 regions (18 cities and eight regencies) set as potential regions for promotion, as displayed in Figure 12. Even though it only contains 8% of the regions, these areas contribute 57.04% of students. These regions also sent high-quality students. The highly potential regions are spread across 15 provinces in five main islands (Sumatra, Kalimantan, Sulawesi, Java, Papua) and two archipelagos (Riau and Nusa Tenggara Archipelago). Most of them are on the Island of Java, in cities/regencies close to the university.

Table 2: The clusters score and rank

Cluster	CGPA score	DO rate score	Std-Enrolled score	Consistency score	Cluster score	Cluster rank
Cluster1	3	2	6	6	4.25	4
Cluster2	1	5	3	2	2.75	6
Cluster3	2	1	3	3	2.25	7
Cluster4	6	3	7	7	5.75	1
Cluster5	4	6	3	4	4.25	4
Cluster6	5	4	3	5	4.25	4
Cluster7	7	7	3	1	4.5	2

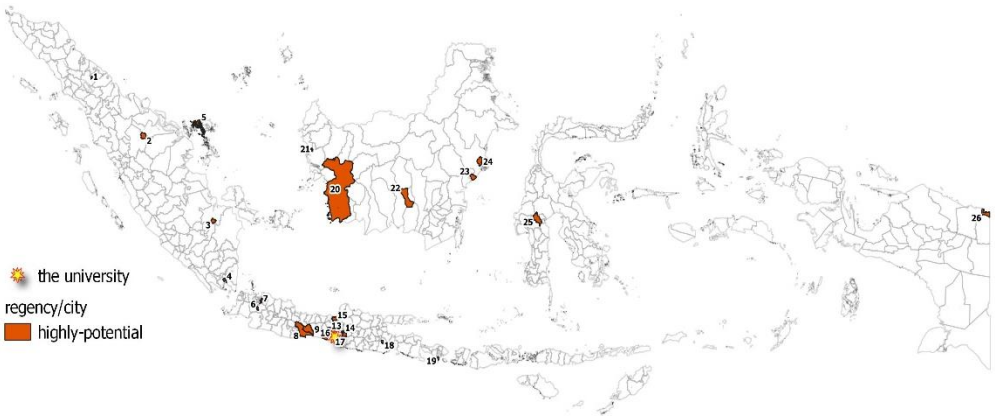

Fig. 12: The potential area for promotion

5. Evaluation

The proposed method's result is compared to the SAW and TOPSIS first 26 rankings for evaluation. According to co-location analysis, the three methods recommended ten identical regions. As seen in Figure 13, the proposed approach has 11 identical regions to the SAW and 15 to the TOPSIS. The recommended regions by the three methods are very consistent in sending students. The ten co-locations regions have an average DO rate of 2.81%, CGPA of 3.08, contain all temporal patterns except TSC1 with a very high number of students sent (462 students on average). Five overlap regions between the proposed method and TOPSIS and one overlap region between the proposed method and SAW have a TSC3 temporal pattern, so the average of students sent higher than the five overlap regions between the SAW and TOPSIS but lost in terms of the students quality (DO rates higher and CGPA lower). For regions that are only recommended by each method, SAW recommends the highest quality students regions but has a TSC1 temporal pattern with few students sent. TOPSIS suggests regions with a higher DO rate than SAW but nearly the same CGPA as the proposed method. The TOPSIS recommends regions with TSC1 and TSC2 temporal patterns with moderately enrolled students, whereas the proposed method recommends regions with higher DO rates with TSC3 temporal patterns that sent the highest number of enrolled students.

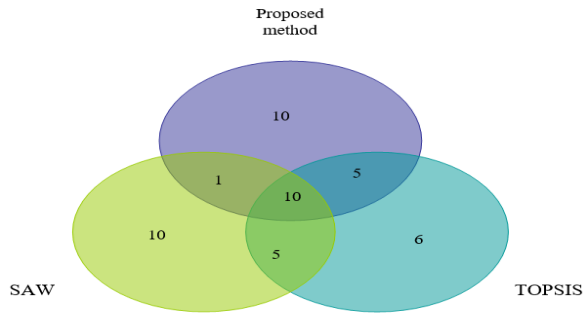

Fig. 13: Overlaps of the potential regions using the proposed method, SAW, and TOPSIS.

Table 3 summarizes all the regions' characteristics suggested by each method. The SAW method suggests areas with a lower DO rate and a higher average CGPA but significantly smaller enrolled students. Even though the average DO rate is slightly higher and the average CGPA is slightly lower, the proposed method can identify locations that supply more enrolled students. The TOPSIS method yields results that fall somewhere between SAW and the proposed method. Based on the analysis, it appears that the proposed method suggests regions with larger enrolled students than the other two methods, despite having a slightly higher DO rate and lower CGPA because areas with no student dropouts and high CGPA but low enrolled students are grouped in a separate cluster (Cluster7).

Table 3: Comparison of the highly potential regions' characteristics

Method	DO rate	CGPA	Consistency	Students	
				Count	%
SAW	1.49	3.17	5	5,294	47.27%
TOPSIS	2.12	3.04	5	6,285	56.12%
The proposed method	3.44	2.98	5	6,388	57.04%

6. Conclusion

This study applies spatio-temporal analysis, two-stage clustering, and the MCDM method to identify potential regions for university promotion targets. The origin schools' regency/city spatial unit is analyzed in the GIS environment using enrollment and academic data from one Indonesian private university. The temporal analysis is conducted using time-series clustering to reveal the enrollment trends, and then for determining the potential regions, mixed data type clustering and MCDM are used.

According to Global Moran's I test, the university students' origin schools are clustered. The origin schools are located in more regencies than cities, but most students come from schools located in cities. Students from city schools are relatively better in quality. Furthermore, schools in the city are more consistent in sending larger numbers of students. Five distinct patterns emerged from the temporal analysis. Horizontal trends and minor fluctuations were discovered in two clusters of regions with the fewest students enrolled. The trend tends to rise slightly in the cluster with moderately enrolled students, while it fluctuates with a trend to fall in the last analysis year for two clusters' areas with the highest student enrollments.

Potential areas for promotion targets were successfully determined using the k-prototype algorithm and the weighted sum of the clusters' variables score, based on the time-series pattern category, the consistency in sending students, average CGPA, and DO rate. Eight districts and 18 cities in Cluster4 are designated as target regions. The proposed approach's findings can help university management develop promotional plans and allocate resources for promotional activities. Compared to the SAW and TOPSIS methods, these 26 regencies/cities supply the most students. According to the analysis, by applying time-series clustering first to the number of enrolled students data, the proposed method generates regions with more enrolled students than the other two methods, although with slightly lower quality because areas with no dropouts and very high CGPA but relatively low enrolled students are grouped in a separate cluster (Cluster7). Thus decision-makers can select regions in Cluster7 and apply different marketing strategies. The proposed approach has not yet been applied to the target schools level, can be considered for future work.

Acknowledgements

The first author would like to respectfully appreciate the support from Universitas

Atma Jaya Yogyakarta, Indonesia, and Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.

References

Abaya, S. A., and Gerardo, B. D. (2013). An education data mining tool for marketing based on C4.5 classification technique. *2013 2nd International Conference on E-Learning and E-Technologies in Education*, 289–293.

Abbasimehr, H., and Shabani, M. (2020). A new framework for predicting customer behavior in terms of RFM by considering the temporal aspect based on time series techniques. *Journal of Ambient Intelligence and Humanized Computing*.

Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T. (2015). Time-series clustering - a decade review. *Information Systems*, 53, 16–38.

Anselin, L (2020). Visualizing Spatial Autocorrelation, *Geodacenter*. *Available online: https: //geodacenter.github.io/workbook/5a_global_auto/lab5a.html (accessed on October 27th, 2021).*

Aryaee, S. (2019). Locating using clustering and capabilities of GIS (case study: Bank). *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 1443–1471.

Ayad, Y. M. (2007). Challenges in student recruitment for educational institutions: materials and methods. *ESRI UC 2007*.

Badan Pusat Statistik. (2020). Statistik Pendidikan 2020. *Badan Pusat Statistik*.

Baser, V. (2020). Effectiveness of school site decisions on land use policy in the planning process. *ISPRS International Journal of Geo-Information*, 9(11).

BPS-Statistics Indonesia. (2019). Statistical Yearbook of Indonesia 2019. *BPS-Statistics Indonesia.*

Bramer, M. (2016). Introduction to Data Mining (Third Edition). *Springer-Verlag London.*

Bulti, D. T., Bedada, T. B., and Diriba, L. G. (2018). Analyzing spatial distribution and accessibility of primary schools in Bishoftu Town. *Ethiopia. Spatial Information Research.*

Chacón-garcía, J. (2017). Geomarketing techniques to locate retail companies in regulated markets. *Australasian Marketing Journal (AMJ)*.

Constantinidis, B. R. (2019). Geomarketing as a location public - private decision support tool or schools in the City of Buenos Aires, Argentina. *Territorio Italia*,1,2, 31–54.

Dahiya, A., Gautam, N., and Gautam, P. K. (2021). Data mining methods and techniques for online customer review analysis: a literature review. *Journal of System and Management Sciences*, 11(3), 1–26.

Ernawati, E., Baharin, S. S. K., and Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, 1869 01208.

Ghodousi, M., Sadeghi-niaraki, A., Rabiee, F., and Choi, S. (2020). Spatial-temporal analysis of point distribution pattern of schools using spatial autocorrelation indices in Bojnourd City. *Sustainability*, 12, 7755.

Guijo-Rubio, D., Duran-Rosal, A. M., Gutierrez, P. A., Troncoso, A., and Hervas-Martinez, C. (2020). Time-series clustering based on the characterization of segment typologies. *IEEE Transactions on Cybernetics*, 1–14.

Han, J., Kamber, M., and Pei, J. (2012). Data Mining Concepts and Techniques (Third Edit). *Morgan Kaufmann Publishers*.

Herlawati, H., Abdurachman, E., Heryadi, Y., and Soeparno, H. (2020). GIS-based MCDM for central business suitability in a small city. *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 1–5.

Huang, Z. (1998). Extensions to the k-Means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2, 283–304.

Hwang, CL., Yoon K. (1981). Methods for Multiple Attribute Decision Making. In: Multiple Attribute Decision Making. *Lecture Notes in Economics and Mathematical Systems*, 186. Springer, Berlin, Heidelberg.

Javed, A., Lee, B. S., and Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications*, 1, 100001.

Jia, Z., and Song, L. (2020). Weighted k-Prototypes clustering algorithm based on the hybrid dissimilarity coefficient. *Mathematical Problems in Engineering*.

Jurisic, M., Ravlic, S., Loncaric, R., and Pugelnik, I. (2016). Implementation of geographic information technology in marketing - GIS marketing. *Interdisciplinary Management Research*.

Kelly, M. G. (2019). A map is more than just a graph: geospatial educational research and the mmportance of historical context. *AERA Open*, 5(1).

Lagrab, W., and Aknin, N. (2015). Analysis of educational services distribution-based geographic information system (GIS). *International Journal of Scientific & Technology Research*.

Lagrab, W., and Aknin, N. (2017). A suitability analysis of elementary schools-based geographic information system (GIS) a case study of Mukalla districts in Yemen. *Journal of Theoretical and Applied Information Technology*, 95(4), 731–742.

Libório, M. P., Bernardes, P., Ekel, P. I., Ramalho, F. D., and Santos, A. C. G. dos. (2020). Geomarketing and the locational problem question in the marketing studies. *Brazilian Journal of Marketing*, 19(2), 448–469.

Malczewski, J. (2006). GIS-based multicriteria decision analysis: a survey of the literature. *International Journal of Geographical Information Science*, 20(7), 703–726.

Martin, R. N. (2001). Identifying target enrollment areas to improve diversity. *Proceedings of the 21st Annual ESRI International Users Conference*.

Mekvabidze, R. (2020). From business modeling to business management: an exploratory study of the optimal decision making on the modern university level. *Journal of Logistics, Informatics and Service Science*, 7(1), 67–86.

Morris, P., and Thrall, G. (2010). Using geospatal technique to address institutional objective: St. Petersburg College Geo-demographic analysis. *IR Applications*, 27.

Murad, A. A., Dalhat, A. I., and Naji, A. A. (2020). Using geographical information system for mapping public schools distribution in Jeddah City. *International Journal of Advanced Computer Science and Applications*, 11(5), 82–90.

Oliveira, M. F. F., Albuquerque, P. H. M., Hao, P. Y., and Henrique, P. A. (2020). Mapping regional business opportunities using geomarketing and machine learning. *Gestão & Produção*, 27(3).

Ousley, C. (2010). A geographic-information-systems-based approach to analysis of characteristics predicting student persistence and graduation. *Ph.D dissertation, The University of Arizona*.

Prasetyo, D. H., Mohamad, J., and Fauzi, R. (2018). A GIS-based multi-criteria decision analysis approach for public school site selection in Surabaya, Indonesia. *Geomatica*, 72(3), 69–84.

Roshan, H., and Afsharinezhad, M. (2017). The new approach in market segmentation by using RFM model. *Journal of Applied Research on Industrial Engineering*, 4(4), 259–267.

Rosyidah, A., and Surjandari, I. (2019). Exploring customer data using spatio-temporal analysis: case study of fixed broadband provider. *International Journal of Applied Science and Engineering*, 16(2), 133–147.

Saha, S., Sarkar, D., Mondal, P., and Goswami, S. (2020). GIS and multi-criteria decision-making assessment of sites suitability for agriculture in an anabranching site of sooin river, India. *Modeling Earth Systems and Environment*.

Sammour, M., Othman, Z. A., Rus, A. M. M., and Mohamed, R. (2019). Modified dynamic time warping for hierarchical clustering. *International Journal on Advanced Science, Engineering and Information Technology*, 9(5), 1481–1487.

Sembiring, B. S. B., Zarlis, M., Sawaluddin, Agusnady, A., and Qowidho, T. (2019). Comparison of SMART and SAW methods in decision making. *Journal of Physics: Conference Series*, 1255 01209.

Sotoudeh-Anvari, A., Sadjadi, S. J., Molana, S. M. H., and Sadi-Nezhad, S. (2018). A new MCDM-based approach using BWM and SAW for optimal search model. *Decision Science Letters*, 7, 395–404.

Sulastri, S., Usman, L., and Syafitri, U. D. (2021). K-prototypes algorithm for clustering schools based on the student admission data in IPB University. *Indonesian Journal of Statistics and Its Applications*, 5(2), 228–242.

Surjandari, I., and Rosyidah, A. (2017). Fixed broadband customer area mapping using spatial analysis. *Proceeding 2017 IEEE 8th International Conference on Awareness Science and Technology*, 109–114.

Tang, H., and Mcdonald, S. (2002). Integrating GIS and spatial data mining technique for target marketing of university courses. *Symposium A Quarterly Journal In Modern Foreign Literatures*.

Tong, D., and Murray, A. T. (2017). Location analysis: developments on the Horizon. *Advances in Spatial Science*, 9783319505893, 193-208.

Yıldırım, Ü. (2021). Identification of groundwater potential zones using GIS and multi-criteria decision-making techniques: a case study Upper Coruh River Basin (NE Turkey). *ISPRS Int. J. Geo-Inf*, 10(396).

Zemlickienė, V. (2019). Using TOPSIS method for assessing the commercial potential of biotechnologies. *Journal of System and Management Sciences*, 9(1), 117–140.

Zhang, D., Zhou, C., and Xu, W. (2020). Spatial-temporal characteristics of primary and secondary educational resources for relocated children of migrant workers: the case of Liaoning Province. *Complexity*.