IS415 Geospatial Analytics and Application
AY 2020/2021 Term 1
Group: T14
Project title: Spatially Constrained Clustering analysis of East Kalimantan province for Indonesia's new capital city
Prepared for: Professor Kam Tin Seong

Erika Aldisa Gunawan
Singapore Management University

Lee Jun Hui Sean
National University of Singapore

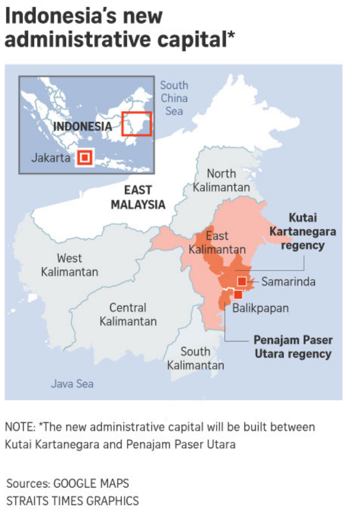
Ng Xun Jie
Singapore Management University

ABSTRACT

In light of the news about Indonesia's decision to move their capital from Jakarta to East Kalimantan, there has been numerous debates on where the new mega-city should be constructed. A project of this scale has never been seen before in Indonesia, where multiple institutions and millions of people will be relocated across the Indonesian Archipelago. This results in the need to properly spatially profile the region to help urban planners strategise and ease the planning process of the new Capital City. This paper aims to achieve the above by using agglomerative hierarchical clustering and as well as spatially constrained methods such as SKATER and ClustGeo. The findings indicate that the regions surrounding current population city centres in East Kalimantan, namely Balikpapan and Samarinda, are potentially good areas to look into. Another significant finding is that the areas in the North-Western region of East Kalimantan are relatively underdeveloped as compared to the coastal regions.

1. INTRODUCTION

On 26th August 2019, Indonesian President Joko Widodo announced that the capital city of Indonesia would be moved to East Kalimantan, between Penajam Paser Utara district & Kutai Kartanegara district. The proposed location is also in between two bigger cities of Balikpapan and Samarinda (Figure 1).


Indonesia's new administrative capital*

NOTE: *The new administrative capital will be built between Kutai Kartanegara and Penajam Paser Utara

Sources: GOOGLE MAPS
STRAITS TIMES GRAPHICS

Figure 1: Proposed location of new capital

The intent was to move away from the heavily populated Java island to slow down the environmental degradation of Jakarta and to unify the archipelago by developing Kalimantan, which is the geographical center of Indonesia (Septiana & Sumarlam, 2018). The move garnered criticisms from urban planning experts and environmentalists because they felt that the move was not strategic. With Kalimantan being a relatively underdeveloped region of Indonesia that is highly forested (Figure 2), and the main income of most of its people coming from agriculture, the aforementioned critics have lambasted this decision as they believe it can potentially cause significant environmental damage via deforestation in the name of urban development. Furthermore, they posit that it won't necessarily mean that Jakarta's problems will be significantly resolved, and could potentially occur to the new Capital city without strong urban planning and mindset changes.

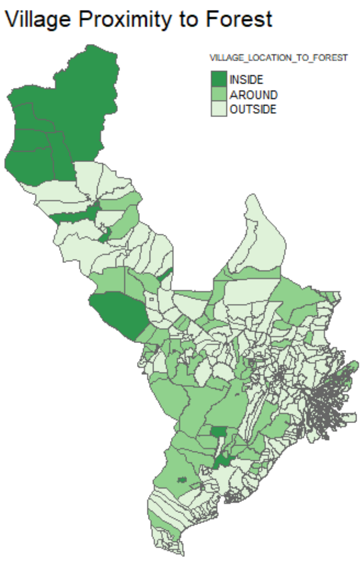


Figure 2: Proximities of Villages to Forests in East Kalimantan

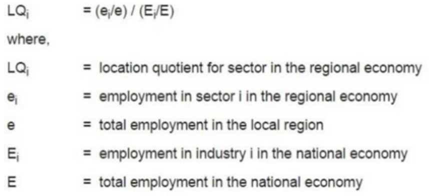
Thus, our project aims to conduct geospatial analysis on the proposed area of the new capital city in East Kalimantan to learn more about the area's socio-economical potential to be considered as the new capital city of Indonesia.

2. MOTIVATION

With the above being said, the specific objectives of this paper are as follows.

1. Analyse the necessary reasons that led to the change of Indonesia's capital city from Jakarta to East Kalimantan.
   - The analysis variables will be derived from research, dataset and extra variables needed for development of capital city
2. Perform an Exploratory Data Analysis on the Region to confirm/alleviate the concerns of naysayers
   - Look at the Spatial Distribution of Natural Disasters (E.g. Earthquakes, Tsunamis, etc) to identify suitable regions for the construction of the new capital
   - Look at the Spatial Distribution of Natural Vegetation (E.g. Forests) to identify suitable regions where the construction of the new capital will minimise damage to existing forests
3. Conduct Spatially constrained clustering analysis & segmentation
   - Profiling of East Kalimantan province to derive the most suitable area for the new capital city in East Kalimantan.
   - Perform analysis on the derived clusters to understand the outputs of the clustering algorithms. This will be useful for planners to understand the different clusters and their characteristics.

3. RELATED WORK



$LQ_i = (e_i/e) / (E_i/E)$

where,

LQ_i = location quotient for sector in the regional economy

e_i = employment in sector i in the regional economy

e = total employment in the local region

E_i = employment in industry i in the national economy

E = total employment in the national economy

Figure 3: Location Quotient Formulae

In search of related research papers on spatially constrained clustering techniques, our team has come across Location Quotient (LQ) as the measurement of a region's industry concentration. The purpose of using the location quotient technique is to generate a coefficient to represent an industry in a given study area compared with the same industry in a larger reference area (Miller & Gibson, 1991). Location quotient analysis commonly assumes that the whole state, as denoted as E to be the benchmark, and is then used to derive the characteristics of the industry in a particular study area within the state.

The quotient value of less than one indicates that the industry in the study region has less of a share of the total than is more generally found in the reference region. A quotient value of more than one indicates that the industry in the study region has a high numerical concentration compared to the average share of the same industry in the reference region. A quotient value of one indicates that the share of industry in the study region is identical with the share of industry of the reference region. Benefits of using location quotient are that it requires minimal amount of data and analytical skills and it can be implemented speedily (Isserman, 1977).

On the other hand, according to Guillain & Le Gallo (2010), spatial agglomeration occurs proportionally with regional development. In their paper on identifying the manufacturing and service sector around Paris in 1999, they have employed Spatial Autocorrelation analysis using location quotient (LQ) to compute the spatial delimiting agglomeration. Spatial Autocorrelation techniques like global Gini Index coupled with Moran's I statistics are argued to be able to present insights as to the relative distribution patterns and spatial patterns of observations in a dataset, but the inclusion of LQ amplify the analysis method and allow one to better understand the mechanism behind events of clustering. The researchers also included Moran scatter plots and local indicator of spatial association (LISA) statistics, to combat LQ's weakness of (1) arbitrary cut-off for agglomeration identification and (2) failure to identify the delimitations of clusters that are spatially autocorrelated.

4. DATASET

For this project, it is important to understand about the government administrative region of Indonesia. In Indonesia, nationally one country, is divided into 34 provinces which includes East Kalimantan. A province (provinsi in Indonesian language) is divided into cities or regencies. Regencies (kabupaten) are usually larger than cities (kota) in size and generally have agricultural activities. In the East Kalimantan province, there are 3 major cities, consisting of Balikpapan, Samarinda, Bontang, and 7 regencies. The administrative region of regencies or cities are then further divided into districts (kecamatan) and the smallest government administrative region is denoted as village (desa) or sub-district (kelurahan). A common observation is that villages, which are often located in rural areas, will combine to form a regency while a sub-district, which are often located nearer to urban areas, will combine to form a city. According to the 2019 report by the Ministry of Home Affairs, there are 8,488 urban villages and 74,953 rural villages in Indonesia.

The dataset used in this analysis, also known as the PODES (Potensi Desa) data, has been obtained through Professor Kam Tin Seong. The original dataset has been produced through extensive surveying of the region by Badan Pusat Statistik, which is the Central Statistics Body Indonesia.

5. METHODS

The team employed 3 different clustering methods in the project, namely agglomerative hierarchical clustering, spatially constrained clustering (SKATER Method) and also the ClustGeo clustering method (D0,D1,Neighbours). The decision to use three different methods rather than one was so that we could provide the urban planners with a more comprehensive view of the study area. Below details the workings of each algorithm.

5.1 Agglomerative Hierarchical Clustering

The first method that the team employed would be Hierarchical Clustering. This is a method that involves the calculation of euclidean distances between the various Desas (villages). The euclidean distance between any two points refers to the length of the line segment between them in a particular dimension space (Tabak, 2014). Prior to this calculation (Figure 4), the various variables used for the clustering method were normalised using the Min-Max standardisation method in order to ensure that the variables used in clustering had the same range and were unitless.

$$d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_i-q_i)^2 + \cdots + (p_n-q_n)^2}.$$

Figure 4: Euclidean Distance Formula

After calculating the Euclidean Distances, the optimal clustering method was selected by looking at the agglomerative coefficient. The method with the highest agglomerative coefficient, the Ward's minimum variance method, was then subsequently used for the hierarchical clustering as it results in the strongest clustering structure.

Afterwards, the agglomerative hierarchical clustering was performed. The steps are as follows:

1. Each Desa in the proximity matrix represents one point in the dimension space.
2. While there are still more than one cluster, repeatedly merge the two closest clusters together and update the proximity matrix.
3. Stop once only one cluster remains.

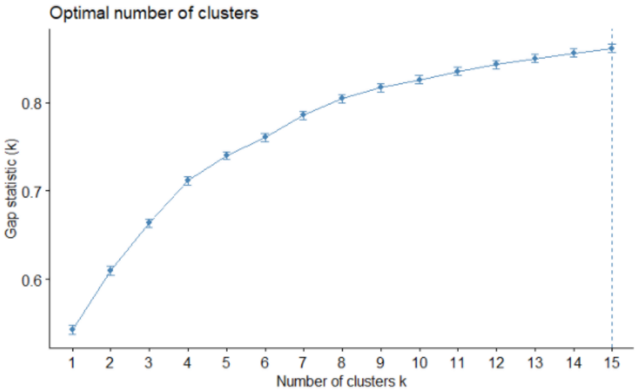


Figure 5: Typical Gap Statistic Plot in the Project

After creating the dendogram, the next step is to choose the optimum number of clusters for the analysis. This was done through the use of the Gap Statistic. The Gap Statistic compares the total intra-cluster variation for each cluster number with that of their expected values under the null reference data distribution (Tibshirani, Walther & Hastie, 2000). Ideally, the cluster number should be chosen such that it maximises the Gap Statistic. However, throughout the project, the results have consistently been such that the gap statistic increases as cluster number, K, increases (Figure 5).

According to Tibshirani, Walther and Hastie, when this situation occurs, it becomes important to pick the first cluster number i where the gradient from cluster i-1 to i is steeper than the gradient from i to i+1. This is because simply picking a large K would not result in meaningful clusters for the analysis. The created dendrogram is then cut into K groups, according to the optimum cluster number K. The cluster numbers were then mapped onto a choropleth map (Figure 6). Further analysis of the results can be found in the 'Results' section.

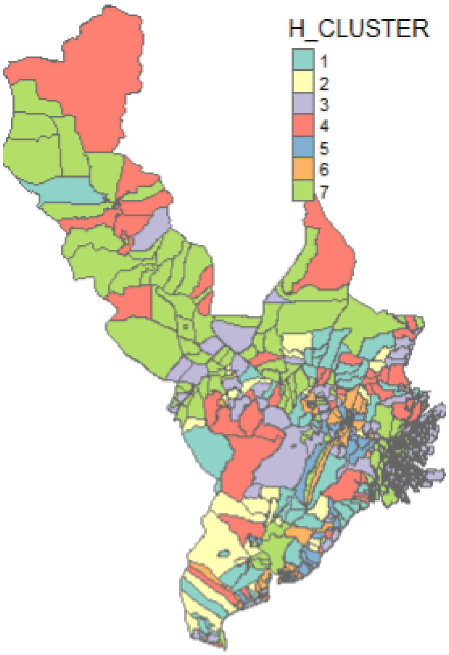


Figure 6: Example of a Choropleth map for Agglomerative Hierarchical Clustering Method

Overall, this method tends to produce spatially fragmented clusters as the geographical locations of the Desas are not taken into account, as seen in Figure 6. To counteract this, the Skater and ClustGeo methods were explored.

5.2 Spatially Constrained Clustering (SKATER Method)

This method takes into account the proximity of the Desas when clustering them. The first step taken requires the computation of the neighbours list. This refers to a matrix where each row has two columns; one that refers to a desa and another that contains a list of the neighbouring desas of that particular desa.

For this particular dataset, there is one Desa that is not physically connected to the Kalimantan mainland (Figure 7). Upon further inspection, this is the desa corresponding to *"Kutai Kartanegara Muara Jawa Muara Kembang"*. Using the knearneigh() function, its nearest neighbour in terms of coordinate distance is then added to its neighbour list. To ensure that subsequent algorithms can work as intended, this particular desa was also added to its nearest neighbour's neighbours list.

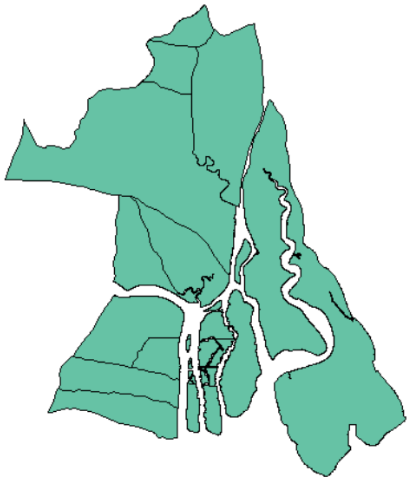


Figure 7: Unconnected Desa

The distances between the desas, which involves the pairwise dissimilarity between the desa and its neighbours, are then computed using nbcosts() of the spdep package. With this, the minimum spanning tree is calculated. The minimum spanning tree of a graph refers to a subset of the edges of a connected, undirected graph, where the edges in this subset connect all the desas together with the lowest possible total distance (refers to the computed distance in the previous part).

Subsequently, the SKATER method can then be implemented. This uses the skater() present in the spdep package. This method uses the created minimum spanning tree, and groups the desas similarly to the agglomerative hierarchical clustering method. The cluster numbers used in this section followed the number used in the previous agglomerative hierarchical clustering method. The cluster numbers were then mapped onto a choropleth map (Figure 8).

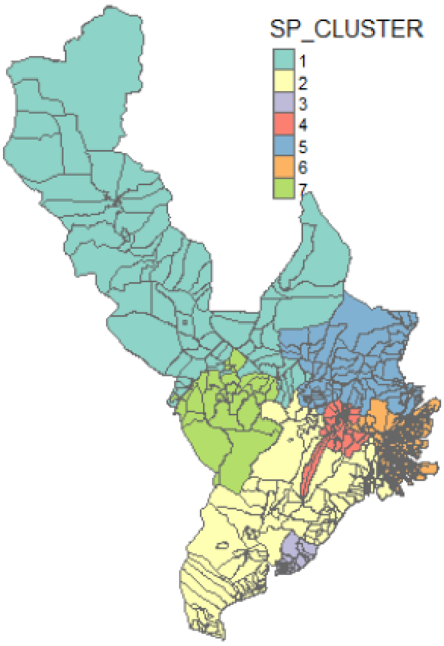


Figure 8: Example results from the SKATER Clustering method

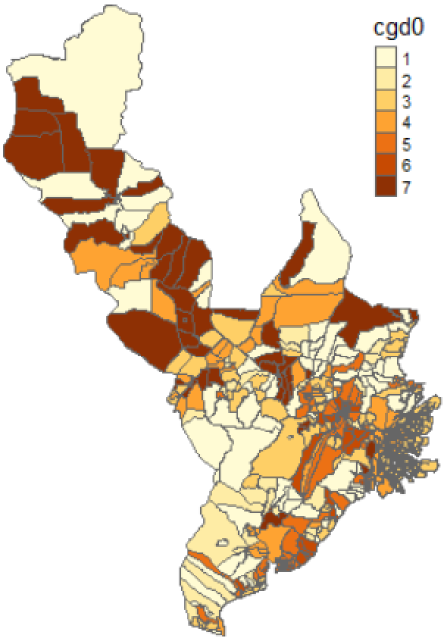


Figure 9: Example results of the D0 Method

This method produces spatially continuous clusters and can be more relevant to urban planners, as more often than not, they want to find an entire continuous plot of land to develop it.

5.3 Spatially Constrained Clustering (ClustGeo Method)

This is the last method that the team has attempted for this study. This method was introduced as it allows one to specify the relative weighting of the input clustering variables against that of the geographical distances between the desas. This is in contrast to the SKATER method, where one is unable to specify this parameter. The number of clusters used in this section will also be the same as the number of clusters used in the agglomerative hierarchical clustering method, which was derived by looking at the Gap Statistic.

5.3.1 Input Clustering Variables only - D0 Method

This method is similar to the agglomerative hierarchical clustering method in the sense that it involves the calculation of the euclidean distances between the Desa to create the matrix D0. Therefore, the outputs of this method (Figure 9) are largely similar to the agglomerative hierarchical clustering method that was previously discussed, although there are some differences as well. This is due to the difference in the methods used.

5.3.2 Hybrid between Clustering Variables & Geographical Distances - D1 Method

This method involves the incorporation of geographical distances into the clustering algorithm.

Firstly, the geographical distances between the Desas had to be calculated. To do so, the projection of the Data had to be changed to “longlat” as it was previously in a Universal Transverse Mercator (UTM) format. Afterwards, the geodist() function of the geodist package was used to calculate the geographical distance between the Desas using the extracted latitude and longitude coordinates.

With this done, the optimal mixing parameter, alpha is then chosen. This value determines the relative importance we assign to the clustering variables versus that of the geographical distances. The higher the value, the greater the geographical differences are factored into the clustering algorithm. This is done using the choicealpha() function of the ClustGeo package, which results in a plot as seen in Figure 10.

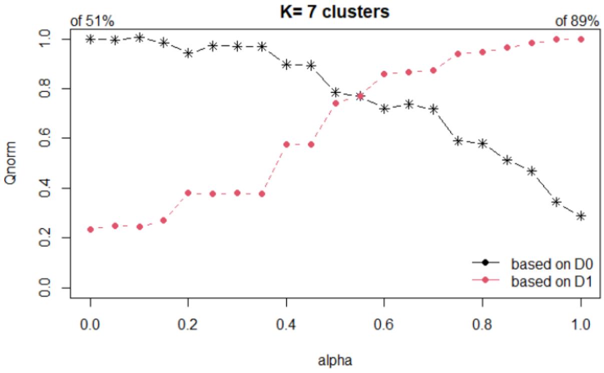


Figure 10: Typical choice alpha plot

From here, an alpha value is chosen by picking a value which results in a small loss of variable clustering homogeneity (represented by black line, D0), but large significant gains in terms of geographic homogeneity (represented by red line, D1). Afterwards, the hclustgeo() function is run using D0,D1, and the selected alpha variable to generate the required output (Figure 11).

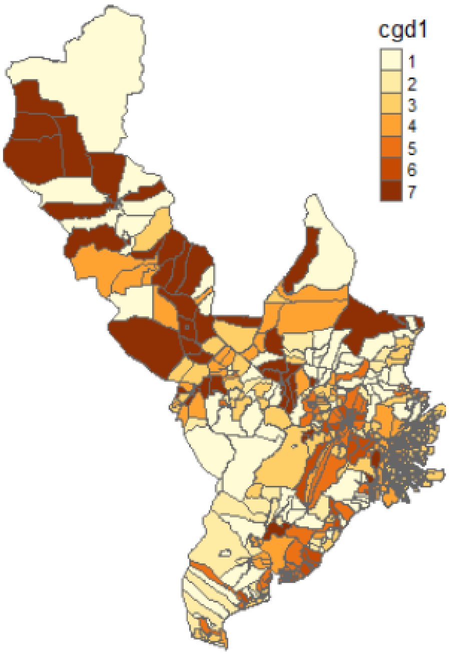


Figure 11: Example results of the D1 Method

5.3.3 *Hybrid between Clustering Variables & Neighbouring Relationship - Neighbours Method*

This method draws similarity to the SKATER method, except now it is possible to specify the relative weighting of the input clustering variables against that of the neighbouring relationship between the desas.

It first requires an adjacency matrix to be constructed from the underlying graph structure of East Kalimantan. As this was already previously constructed in the SKATER portion of this study, the constructed neighbours list then was simply converted to an adjacency matrix.

Afterwards, the dissimilarity matrix, denoted as D1.n, was calculated by taking as.dist(1-A), where A refers to the adjacency matrix representation of the graph. The next step would be to choose the optimal mixing parameter once again, which was done in a similar fashion to the D1 method above. The clusters are then formed by specifying D0,D1.n, and alpha as the input arguments for the hclustgeo() function, giving the outputs in Figure 12.

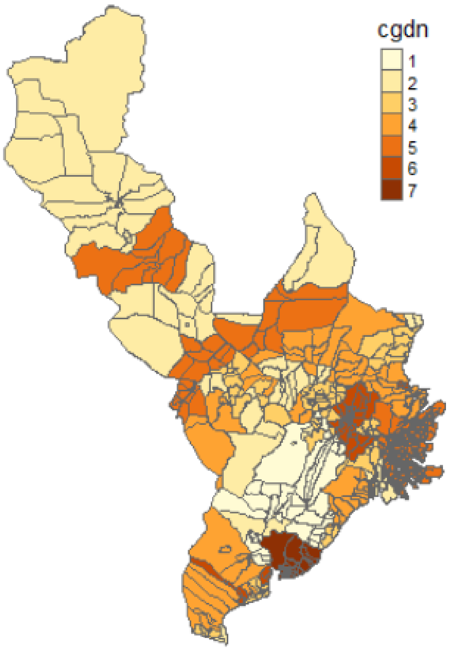


Figure 12: Example results of the Neighbours Method

Overall, there were not much differences observed between the D0 and D1 method, while incorporating the neighbors into the clustering analysis allowed more spatially homogeneous clusters to be formed. This is consistent with the results from the agglomerative hierarchical clustering and spatially constrained SKATER method used, where the former resulted in dispersed clusters while the latter allowed for more spatial homogeneity.

6. RESULTS & DISCUSSION

Upon obtaining the results of SKATER clustering method, our team decided to examine how each selected variable contributed to the formation of clusters seen in our SKATER map.The results from the SKATER method was analysed because the results offered were the most spatially homogeneous and we felt that would offer the greatest insights to urban planners.

To do so, the team decided to firstly perform a descriptive analysis on the created spatially constrained clusters created from the SKATER method.The cluster exploration was done in the form of bar charts (Figure 13), where the mean of selected variables are taken for the clusters.

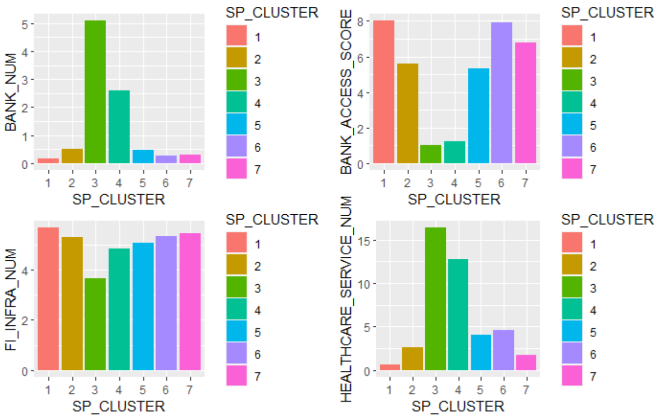


Figure 13: Bar charts of aggregated means for Finance and Healthcare industries

Furthermore, the team also reran the clustering algorithm for some selected variables only to better profile the different sectors in the region. For example, when running the algorithm for the Finance industry, the team selected a few finance related variables (e.g. ATMs, Number of Credit Loans, etc) to understand the Spatial Distribution for this particular sector within East Kalimantan.

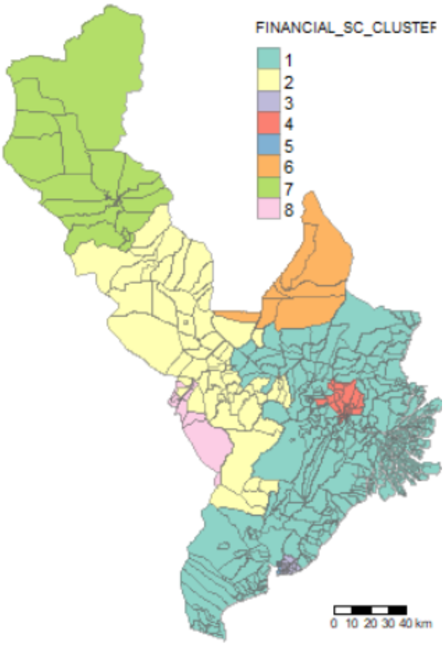


Figure 14: Skater clustering map of Finance industries

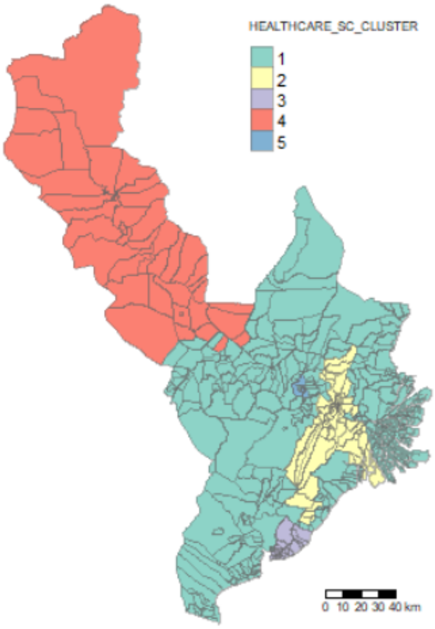


Figure 15: Skater clustering map of Healthcare industries

From this analysis, it is evident that cluster 3 is unique because it is presented in almost all bar charts as either the highest or lowest value. In the interest of time, the team then decided to look into two variables as cluster 3 stands out the most. As seen above, our team decided to look into the financial and healthcare sector (Figure 13), more specifically the number of service providers in these 2 sectors. This decision is made in accordance with how we perceive that the financial and healthcare sector forms part of the basic infrastructure of a city. We can clearly see that when compared with the cluster number formed with the SKATER analysis on all variables, cluster 3 in the financial sector is significantly lower than other clusters. Our team hypothesized that this could be due to low financial investment in the cluster and we hope that further research can be conducted to prove our hypothesis.

On the other hand, we can observe that cluster 3 is significantly higher than other clusters in terms of healthcare sector, signifying that there could be ample healthcare facilities in the cluster. However, our team proposes that an in-depth analysis must be conducted to identify whether the supply of healthcare services is meeting the demand of such services in the corresponding clusters.

In addition, for the industrial level SKATER analysis, the team looked at the same two industries, financial (Figure 14) and healthcare (Figure 15) in order to identify any geographical patterns in their distributions. In doing the analysis, we can quickly identify that the isolation of the variables associated with the two industries have yielded different numbers of clusters (8 for finance, 5 for healthcare), compared the overall SKATER analysis that contains 7 clusters. These numbers were derived through the Gap Statistic method as previously discussed.

Considering that the two bigger cities located in East Kalimantan province, namely Balikpapan and Samarinda, we can pinpoint that cluster 4 in the financial sector is relatively closer to the city of Samarinda, potentially indicating that the financial sector is stronger in that region. This could be done for future analysis, alongside more data to prove our hypothesis. The healthcare sector map on the other hand complements the cluster 3 formed from previous SKATER analysis, further strengthening our hypothesis that the region (cluster 3 as well) that is closer to the city of Balikpapan is indeed equipped with sufficient healthcare infrastructure.

7. FUTURE WORK

While the clustering analysis has elucidated out certain trends in the study area, there are some limitations that our team would like to work on in the future if given the chance.

Firstly, there were many missing fields present in the dataset. The dataset was collected through surveying villages and geographical difficulties may have caused difficulties in surveying the regions accurately. When working with the data, many missing fields were present. Therefore, the study had to drop columns where there are too many nulls present, which could have resulted in some level of distortion present in the clustering analysis. To counteract this, perhaps secondary sources could have been found to further complement the Podes dataset used in this study.

Secondly, the team also believes that a Local Moran's I test, incorporating Location Quotient calculation could be performed in order to help Urban Planners better visualize where the "hot spots" and "cool spots" are in terms of a univariate analysis. The team hoped that by using the above mentioned techniques will bring better insights about the distribution of different socio-economic drivers or industries. This would have tied in and reinforced the box maps that we used as part of the exploratory data analysis portion of the study.

Thirdly, the team has processed the Podes data in such that the different variables are aggregated together based on their common functionalities. We propose that more exploratory data analysis should be done to ascertain the inclusion of socio-economic factors or variables that may have left out during the course of our research, perhaps through understanding the characteristics, proportion and impacts of a particular variable to a region.

Fourthly, the team also faced some computational problems when trying to cluster the desas. This is because there was a limit to the amount of computational power that our computers had. Therefore, this prevented us from using as many features as we wanted to do a complete profiling of the study area.

Lastly, the team was also unable to take into account the domain knowledge of urban planners into the project analysis. Therefore, we could only try our best to profile the region according to what we felt was appropriate. Being able to receive feedback from urban planners will definitely add insight to this report and further strengthen the credibility of the analysis presented. For instance, perhaps the minimum area needed for the new capital city state could have been provided by urban planners to guide the recommendation on the site suitability.

8. CONCLUSION

In conclusion, this paper explored the East Kalimantan area using various clustering methods in order to profile the area for urban planners. These methods include the agglomerative hierarchical clustering, SKATER method, and as well as the ClustGeo method. This was done to provide the intended users with a more comprehensive overview of the area. While the study has some limitations, it is believed to still be useful as it provides urban planners with a geographical overview and understanding of the East Kalimantan area. It is hoped that the findings from this paper will help guide the Urban Planners in their decision in building a new Indonesian capital city.

9. REFERENCES

Brown, N.S. & Watson, P. (2012) "What can a comprehensive plan really tell us about a region?: A cluster analysis of county comprehensive plans in Idaho", Western Economics Forum. Pp.22-37. (https://ageconsearch.umn.edu/record/176591/files/WEFFall2012v11n2_Brown.pdf)

Demeter, T. and Bratucu, G. (2013) "Statistical Analysis Of The EU Countries from A Touristic Point of View", Bulletin of the Transilvania University of Braşov, 6(55): 121-130. (https://search-proquest-com.libproxy.smu.edu.sg/docview/1510289237?rfr_id=info%3Axri%2Fsid%3Aprimo)

Guillain, R., & Le Gallo, J. (2010). Agglomeration and Dispersion of Economic Activities in and around Paris: An Exploratory Spatial Data Analysis. Environment and Planning B: Planning and Design, 37(6), 961–981. https://doi.org/10.1068/b35038

Isserman, A. M. (1977). The Location Quotient Approach to Estimating Regional Economic Impacts. Journal of the American Institute of Planners, 43(1), 33–41. https://doi.org/10.1080/01944367708977758

Miller, M. M., & Gibson, L. J. (1991). Location quotient: A basic tool for economic development... Economic Development Review, 9(2), 65.

Rovan, J. and Sambt, J. (2003) "Socio-economic Differences Among Slovenian Municipalities: A Cluster Analysis Approach", Developments in Applied Statistics, pp. 265-278. (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.4636&rep=rep1&type=pdf)

Septiana, D., & Sumarlam, S. (2018). Palangka Raya the Capital City of Indonesia: Critical Discourse Analysis on News about Moving the Capital City from Jakarta. Advances in Social Science, Education and Humanities Research, 280, 190-202.

Tabak, J. (2014), Geometry: The Language of Space and Form, Facts on File math library, Infobase Publishing, p. 150, ISBN 9780816068760