K-means algorithm for mapping by utilizing google maps imagery

U Syaripudin1*, Y A Gerhana1, A P Fuzilesmana1, A R Atmadja1,2 and M A Ramdhani1

1Department of Informatics, UIN Sunan Gunung Djati, Bandung, Indonesia
2Department of Informatics, Sekolah Tinggi Tekologi Garut, Garut, Indonesia

*Corresponding author's email: undang.if@uinsgd.ac.id

Abstract. In several cases, data mapping that is available today has not much done digitally yet. Too much data that is involved in data filling process causing a long mapping process. The purpose of this research is to design the system by implementing K-Means Clustering Algorithm for data mapping using Google Maps imagery. The object of system design in this research is the plantation area in Sukabumi Regency, Indonesia. The system design methodology that used is the UML (Unified Modelling Language) approach. Based on the results of system testing, K-Means Clustering algorithm can determine and cluster the plantation area with potency of commodity with three level, among others high level, medium level, and low level, which is further indicated by spatial data.

1. Introduction

Remote sensing is the science of obtaining information about objects, regions, or symptoms by analysing data obtained using tools without direct contact with the object, blood or symptoms. In remote sensing using satellites will produce image data and one of the satellite image data that used to monitor the plantation area is image data can get from Google Maps. Remote sensing is part of digital imagery in the field of Geographic Information Systems (GIS) where there is a data that is processed to get information. Then the data will be processed to be an information that will be mapped by an image from Google Maps [1].

Therefore, it is necessary to do this mapping so that will be provided the plantation information more clearly. In this study, the plantation area was conducted in Sukabumi Regency with clustering techniques using the K-Means Clustering algorithm and utilizing the Google Maps image. K-Means Clustering is an algorithm that has advantages in grouping objects based on similarities between one object with other objects.

Several studies show that the K-Mean clustering algorithm is an effective technique for classifying objects including: Application of Clustering Methods with K-Means to Map the Potential of Rice Plants in Semarang City [2]; Implementation of K-Means Clustering Based on RFM Mofek as Mapping and Supporting Customer Management Strategies (Case Study: PT Herbal Penawar Alwahidah Indonesia Pekanbaru) [3]; Mapping of Teacher Spread in Banten Province Using Spatial Clustering K-Means Method (Case Study: Banten Province Region) [4]; Mapping of Street Crimes in Semarang City Using the K-Means Clustering Algorithm [5]; and Cluster Analysis Using Fuzzy C-means and K-means Algorithms for Clustering and Mapping of Agricultural Land in Southeast Minahasa [6]. Clustering in this study is used as a technique in data mining which is dividing objects...
into certain groups or clusters so that objects in a cluster have similarities with other objects, which in this study was conducted to classify the potential of plantation land.

2. Methods
The application system development methodology is carried out in stages below: (1) define design goals; (2) identify needs; (3) prototype design; (4) test and evaluate prototypes; and (5) additions and improvements to the prototype that has been made. In this study the application was built using a UML (Unified Modelling Language) approach that describes a system design that will be built into visual forms [7][8]. Reliability of system applications in this research, it is used: analytical, logical, conceptual, and operational verification by an expert[9].

3. Result and Discussion
Technology tools are made to improve the quality of human life [10], one of the technological tools that can be used to assist the effectiveness and efficiency of running a business process in the plantation sector is the information system. The information system is a combination of information technology and the activities of people using computerized technology [11], which is generally used to support operations and management [12]. Information systems are systems that process data in an organized manner [13], which are designed based on user requirements [14], information systems have a high degree of flexibility that allows them to be developed into better systems [15], and information systems are able to overcome complex data processing problems [16]. Based on many previous research prove that information systems have advantages in terms of: good data accessibility [8], efficient in time [17], accurate [18], supporting decision precisely [19], more economical [20], widely used [21], improve user understanding [22], improve the productivity [23], present data and information well [24], and data storage media [25].

One implementation of the information system that related for mapping is a GIS, which uses spatial data as a tool in conveying information. GIS can be implemented by utilizing the Google Maps feature in mapping and monitoring an area will be able to produce information that is useful by remote sensing [26]. An area mapping is closely related to a web-based GIS, because it can get information on maps of an area with internet network access. Plantation mapping is useful for identifying land potential. Information on land potential can be identified by looking at the area and the layout pattern of the plantation. The mapping process is important as one method of monitoring an area, which then provides information about the state of the area. Digital mapping will have a positive impact considering that the area's information can be accessed anywhere and anytime [27].
3.1. Weighted Product Method
Analysis of the problem in this study is necessary to digitize spatial data on plantations (Figure 1). Digital mapping can be analysed what is the lack and advantages are there in the area. Then, digitizing mapping is very useful because it can be accessed anywhere and anytime. The current mapping process is still constrained in plantation data as map input, because it requires a lot of data to be entered and after the data is entered it also needs to be manually classified, so that the mapping process will take a long time because there will be a signification process of a commodity potential.

In the current mapping, it has not been described in detail yet about the potential commodity whether it is high or low because it will determine the current plant growth. So that the current problems in the previous system include:

a. Too much data when entered
b. The classification process is not carried out by the system
c. The mapping process takes a long time

The area in Sukabumi Regency is a fairly high producer in the plantation sector because there are many types of plants that can be maximized and not a few species of plants exist, there are at least up to 20 types of plants that grow on the Sukabumi district plantations. So, the use of plantations would be good to have a good maintenance system function, one of which is the plantation mapping system

3.2. K-Means Clustering Algorithm
The K-Means Clustering method is used in data mining for data grouping into clusters or based on a similarity of variables or data attributes. There are several stages in the calculation of K-Means Clustering as presented in Figure 2. The K-Means algorithm will then repeat the following steps until there is stability, among others:

a. Determine the number of clusters. To do clustering with the K-Means algorithm the first step is to determine how many clusters that will be formed. In this study, there were three clusters, namely the Most Commodity Production (C1), Moderate Production (C2), and Slight Production (C3).

b. Determine the center point. After determining number of clusters that want to be formed, the next step is to determine the center point (centroid) of each cluster. Taking the center point at random (random). The center point that taken in this study is the central point taken in this study, among others: C1 is the center point for high production areas with a central point (1000; 10000); C2 is the center point for moderate production areas with a central point (500; 5000); and C3 is the center point for small rice production areas with a central point (50; 100).

c. Calculate the distance of each object to the center point. The next after determining the center point is to calculate the distance of each data with a predetermined center point (Equation (1)).

\[ d_{(x,y)} = \sqrt{(x_i - y_i)^2 + (x_l - y_l)^2} \]  

(1)

Where \( x \) is the production dataset, the dataset that used for mining calculations is the area and production. Whereas, \( y \) is the centre point that has been determined randomly.

| Table 1. Plantation data |
|--------------------------|
| **Commodities** | **Area (Ha)** | **Production/Year (Kg)** | **Commodities** | **Area (Ha)** | **Production/Year (Kg)** |
| Aren | 982.54 | 3.594.032.60 | Kemiri | 64,30 | 33.232,50 |
| Cengkeh | 7.661.48 | 6.620.985.00 | Kina | 00,00 | 00,00 |
| Geta Perca | 00,00 | 00,00 | Kopi Arabika | 643,00 | 96.220,00 |
| Jambu Mete | 14,05 | 2.523,75 | Kumis Kucing | 217,00 | 59.830,00 |
| Jarak Pagar | 00,00 | 00,00 | Lada | 89,50 | 22.095,00 |
| Kakao | 405,30 | 30.020,00 | Mindi | 00,00 | 00,00 |
| Kopak | 35,64 | 7.128,00 | Pala | 1.475,40 | 311.153,31 |
| Karet | 3.477,89 | 1.135.167,75 | Pandan | 40,00 | 00,00 |
| Kelapa Dalam | 9.263,71 | 4.300.866,00 | Panili | 297,00 | 47.657,50 |
| Kelapa Hibrida | 2.855,00 | 2.109.250,00 | Pinang | 51,00 | 10.427,50 |
| Kelapa Sawit | 00,00 | 00,00 | Teh | 8.316,28 | 9.121.415,00 |
Table 1 provides the plantation data as an input of implementation system. The data in Table 1 are the plantation data which will be clustered, because the purpose of this clustering process is to find the potential commodity, so that the data that required is the area and the production. The initial calculation process is conducted by setting the centre point randomly, among others: the highest, lowest and median data (Table 2). So, based on Table 1, the lowest commodity is “jambu mete” with 14.05 Ha of area and 2.532,75 kg of production per year. The median data is “kakao” with 405.30 Ha of area and 30.020,00 kg of production per year. And the last, for highest commodity is “kelapa dalam” with 9.263,71 Ha of area and 4.300.866,00 kg of production per year.

The first Xi is taken from the initial data table and the first “aren” area is taken from the center of the first cluster. Whereas, the second Xi is taken from the initial “aren” production data and the second y is taken from the first cluster center in the second part. Example of calculating “aren” data with an area of 982.54 Ha and production of 3,594,032.60 kg below:

a. Calculation of the first centroid distance is:
\[ d(x,y) = \sqrt{(982.54 - 9263.71)^2 + (3594032.60 - 4300866.00)^2} = 706881.90 \]
b. Calculation of “aren” distance for the second centroid is:
\[ d(x,y) = \sqrt{(982.54 - 405.30)^2 + (3594032.60 - 30020.00)^2} = 3564012.64 \]
c. Calculation of palm distance for the third centroid is:
\[ d(x,y) = \sqrt{(982.54 - 14.05)^2 + (3594032.60 - 2523.75)^2} = 3591508.98 \]

After six times iteration, then in the Table 2 shows a comparison of the two closest cluster between iteration 5 and iteration 6 which no longer has a difference, which indicates the iteration process is complete, and the result is the final result of the clustering process.

| No | Commodity | 5th Iteration | 6th Iteration |
|----|-----------|---------------|---------------|
| 1  | Aren      | C2            | C2            |
| 2  | Cengkeh   | C1            | C1            |
| 3  | Geta Perca| C3            | C3            |
| 4  | Jambu Mete| C3            | C3            |
| 5  | Jarak Pagar| C3          | C3            |
| 6  | Kakao     | C3            | C3            |
| 7  | Kapok     | C3            | C3            |
| 8  | Karet     | C3            | C3            |
| 9  | Kelapa Dalam| C2        | C2            |
| 10 | Kelapa Hibrida| C2         | C2            |
| 11 | Kelapa Sawit| C3          | C3            |

| No | Commodity | 5th Iteration | 6th Iteration |
|----|-----------|---------------|---------------|
| 12 | Kemiri    | C3            | C3            |
| 13 | Kina      | C3            | C3            |
| 14 | Kopi Arabika| C3         | C3            |
| 15 | Kumis Kucing| C3         | C3            |
| 16 | Lada      | C3            | C3            |
| 17 | Mindi     | C3            | C3            |
| 18 | Pala      | C3            | C3            |
| 19 | Pandan    | C3            | C3            |
| 20 | Pamili    | C3            | C3            |
| 21 | Pinang    | C3            | C3            |
| 22 | Teh       | C1            | C1            |

Then, from Table 2 can be determined which commodities that have the highest potential (C1), medium potential (C2), and the lowest potential (C3). Based on the process above, the results of calculations using the K-Means Clustering algorithm produce commodity data where the data are high, medium, and low potential. From 22 data, C1 contain “cengkeh” and “teh”, C2 contain “aren”, “kelapa dalam”, and “kelapa hibrida”, and the others is in C3. Those data is the final result of determining the minimum distance for each commodity where C1 is the highest potential, C2 for medium potential and C3 is low potential.

3.3. System Design

Figure 3 describes about the flow process of implementation system. There are admin as an actor who manage initial data and input plantation data into system, user as an actor who use the system and receive the final result, data and topography that data input for system, database as a plantation data storage, k-means clustering as an algorithm that used for determining commodity potential, internet as access media for retrieve data from database, PC an and device to show mapping information, and Google Maps and GPS as an support tools for showing the map of plantation information in Sukabumi regency.
3.4. User Interface
User interface for this research system is designed in accordance with easy use aspect. Figure 4 shows plantation data in Sukabumi Regency where contain the commodities, area of commodity and production per year of commodity. Figure 5 shows the final result of K-Means Clustering algorithm that have three cluster, among others highest commodity, medium commodity, and lowest commodity. And Figure 6 shows the plantation information of Sukabumi regency.

4. Conclusion
The K-Means Clustering algorithm accelerates the process of grouping the potential of plantation commodities. K-Means Clustering algorithm is also can determine the potential of the highest, medium, and lowest commodities. Mapping of plantations is conducted digitally by utilizing Google Maps features that have shown the coordinates of the area, which effectively displays information on plantation data in each area of Sukabumi Regency. Suggestions for the next development, it can use another feature besides Google Maps so that the mapping becomes more accurate, displays the location and distribution of the plantations in more detail and accurate (larger scale).

5. References
[1] Eddy P 2013 Sistem Informasi Geografis : Konsep-Konsep Dasar Perspektif Geodesi & Geomatika (Bandung: Informatika)
[2] Lianna F 2014 Penerapan Metode Clustering dengan K-Means Untuk Memetakan Potensi Tanaman Padi di Kota Semarang Semarang: Universitas Dian Nuswantoro
[3] Hadi F, Mustakim M, Rahmadia D O, Nugraha F H, Bulan N P and Monalisa S 2017 Penerapan K-Means Clustering Berdasarkan RFM Mofek Sebagai Pemetaan dan Pendukung Strategi Pengelolaan Pelanggan (Studi Kasus: PT. Herbal Penawar Alwahidah Indonesia Pekanbaru)
[4] Priambodo Y A and Prasetyo S Y J 2018 Pemetaan Penyebaran Guru di Provinsi Banten dengan Menggunakan Metode Spatial Clustering K-Means (Studi kasus: Wilayah Provinsi Banten) Indones. J. Comput. Model. 1 1 18–27

[5] Astuti W and Widodo D A 2017 Pemetaan Tindak Kejahatan Jalanan di Kota Semarang Menggunakan Algoritma K-Means Clustering J. Tek. Elektro 8 1 5–7

[6] Tamaela J, Sedyono E and Setiawan A 2017 Cluster Analysis Menggunakan Algoritma Fuzzy C-means dan K-means untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara J. Buana Inform. 8 3 151–160

[7] Bkhati D D 2016 Pemodelan Sistem Informasi Keuangan Daerah Pada Direktorat Evaluasi Pendanaan Dan Informasi Keuangan Daerah JOIN (Jurnal Online Inform.) 1 2 98–106

[8] Maylawati D S, Ramdhani M A and Amin A S 2018 Tracing the Linkage of Several Unified Modelling Language Diagrams in Software Modelling Based on Best Practices Int. J. Eng. Technol. 7 2.9 776–780

[9] Suhendi H Y, Ramdhani M A and Irwansyah F S 2018 Verification Concept of Assessment for Physics Education Student Learning Outcome Int. J. Eng. Technol. 7 3.21 321–325

[10] Ramdhani M A, Aulawi H, Ilkhana A and Mauluddin Y 2017 Model of green technology adaptation in small and medium-sized tannery industry J. Eng. Appl. Sci. 12 4 954–962

[11] Pamaragung A, Suryadi K and Ramdhani M A 2006 Enhancing the implementation of e-Government in Indonesia through the high-quality of virtual community and knowledge portal Proceedings of the European Conference on e-Government, ECEG 341–348

[12] Ramdhani M A 2013 Metodologi Penelitian untuk Riset Teknologi Informasi. (Bandung: UIN Sunan Gunung Djati)

[13] Maylawati D S, Darmalaksana W and Ramdhani M A 2018 Systematic Design of Expert System Using Unified Modelling Language IOP Conf. Ser. Mater. Sci. Eng. 288 1 012047

[14] Ramdhani M A, Maylawati D S, Amin A S and Aulawi H 2018 Requirements Elicitation in Software Engineering Int. J. Eng. Technol. 7 2.29 772–775

[15] Aulawi H, Ramdhani M A, Slamet C, Ainissyifa H and Darmalaksana W 2017 Functional Need Analysis of Knowledge Portal Design in Higher Education Institution Int. Soft Comput. 12 2 132–141

[16] Darmalaksana W, Ramdhani M A, Cahyana R and Amin A S 2018 Strategic Design of Information System Implementation at University Int. J. Eng. Technol. 7 2.29 787–791

[17] Slamet C, Rahman A, Sutedi A, Darmalaksana W, Ramdhani M A and Maylawati D S 2018 Social Media-Based Identifier for Natural Disaster IOP Conf. Ser. Mater. Sci. Eng. 288 1 012039

[18] Slamet C, Andrian R, Maylawati D S, Darmalaksana W and Ramdhani M A 2018 Web Scraping and Naïve Bayes Classification for Job Search Engine IOP Conf. Ser. Mater. Sci. Eng 288 1 1–7

[19] Gerhana Y A, Zulfikar W B, Ramdani A H and Ramdhani M A 2018 Implementation of Nearest Neighbor using HSV to Identify Skin Disease IOP Conf. Ser. Mater. Sci. Eng. 288 1 012153

[20] Rahman C, Slamet C, Darmalaksana W, Gerhana Y A and Ramdhani M A 2018 Expert System for Deciding a Solution of Mechanical Failure in a Car using Case-based Reasoning IOP Conf. Ser. Mater. Sci. Eng. 288 1 012101

[21] Slamet C, Rahman A, Ramdhani M A and Darmalaksana W 2016 Clustering the Verses of the Holy Qur’an Using K-Means Algorithm Asian J. Inf. Technol. 15 24 5159–5162

[22] Maylawati D S, Ramdhani M A, Zulfikar W B, Taufik I and Darmalaksana W 2017 Expert system for predicting the early pregnancy with disorders using artificial neural network 2017 5th Int. Conf. Cyber IT Serv. Manag. CITSIM 2017

[23] Zulfikar W B, Jumadi, Prasetyo P K and Ramdhani M A 2018 Implementation of Mamdani Fuzzy Method in Employee Promotion System IOP Conf. Ser. Mater. Sci. Eng. 288 1 012147
[24] Maylawati D S, Ramdhani M A, Rahman A and Darmalaksana W 2017 Incremental technique with set of frequent word item sets for mining large Indonesian text data 2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017 1–6

[25] Taofik A, Ismail N, Gerhana Y A, Komarujaman K and Ramdhani M A 2018 Design of Smart System to Detect Ripeness of Tomato and Chili with New Approach in Data Acquisition IOP Conference Series: Materials Science and Engineering 288 1 012018

[26] Tomasoa L E and Prasetyo S Y J 2018 Analisis Index Vegetasi Pesisir Pantai Aceh Pasca Tsunami Menggunakan Citra Satelit Landsat 7 dan Landsat 8 dengan Metode Clustering Algoritma K-Means Indones. J. Comput. Model. 1 1 28–35

[27] Rohmi G F, Zulfikar W B and Gerhana Y A 2018 Implementasi citra digital berdasarkan nilai HSV untuk mengidentifikasi jenis tanaman mangga menggunakan algoritma K-Nearest Neighbor INSIGHT 1 1 142–147

Acknowledgement
The authors expressed the appreciation and thank you to the Research and Publishing Center of UIN Sunan Gunung Djati Bandung, which has provided funding support for the publication of this article.