Clustering Module in OLAP for Horticultural Crops using SpagoBI

D Putri[1] **and I S Sitanggang**[1]

[1]Computer Science Department, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Jalan Meranti Wing 20 level 5 Kampus IPB, Bogor, West Java, Indonesia

E-mail: defrianiputri@gmail.com, imas.sitanggang@ipb.ac.id

Abstract. Horticultural crops data are organized by the Ministry of Agriculture, Republic of Indonesia. The data are presented annually in a tabular form and result a large data set. This situation makes users difficult to obtain summaries of horticultural crops data. This study aims to develop a clustering module in the SOLAP system for the distribution of horticultural crops in Indonesia and to visualize the results of clustering in a map using SpagoBI. The algorithm used for clustering is K-Means. Horticultural crops data include vegetables, ornamental plants, medicinal plants, and fruits from 2000 to 2013. The clustering module displays clustering results of horticultural crops in the form of text and table on SpagoBI. This module can also visualize the distribution of horticultural crops in the form of map on the HTML page. The application is expected to be useful for users in order to easily obtain summaries of the horticultural crops distribution data and its clusters. The summaries and clusters can be beneficial for the stakeholders to determine potential areas in Indonesia for horticultural crops.

1. Introduction

Ministry of Agriculture of Republic of Indonesia (Ministry of Agriculture RI) is the institution that provides data on agricultural sector including horticultural crops. The data are presented annually for all districts in Indonesia in a tabular form which results in a large data set. The situation makes data accumulation and further processing on the data is required if users need summaries of horticultural crops data. The horticultural crops data can be accessed on the website http://aplikasi.pertanian.go.id /bdsp/index.asp. The data include vegetables, ornamental plants, medicinal plants, and fruits from 2000 to 2013.

Spatial data are geographically oriented data and have a coordinate system as a reference [1]. Spatial data cube contains spatial dimensions and measures. Online Analytical Processing (OLAP) is one of application that integrates data warehouse in order to simplify a process of information extracting from large data sets [1]. A spatial data warehouse for horticultural crops has been constructed by integrating thematic and geographically referenced data from multiple sources and it provides a spatial data cube that is useful for multidimensional spatial data analysis [1].

Currently there are several open source business intelligence and reporting tools including Pentaho and SpagoBI. In this study, SpagoBI was used to display horticultural crops distribution as clustering results. Whereas, visualization of clustering results of horticultural crops is done in form of map on a HTML page. SpagoBI is an open source, complete, and flexible business intelligence tool with a wide

set of analytical engines and it can be used to find a best solution for end users, saving time, and economic resources [2]. SpagoBI has an advantage in integrating new modules and it gives the original developers the possibility to develop and to improve products [3]**.**

In this study, a clustering module in the SOLAP system was developed to provide distribution of horticultural crops in Indonesia and to visualize clustering results of horticultural crops in a map using SpagoBI. The SOLAP system is expected to be used by users to obtain summaries of the horticultural crops data and distribution of its clusters at district level in Indonesia. The summaries and clusters can be beneficial for the stakeholders to determine potential areas in Indonesia for horticultural crops.

2. Data and Method

2.1. Data

Horticultural crops data consist of four fields including commodities, status id, location, and time. This study uses horticultural crops data from 2000 to 2013. In addition, this study utilizes the administrative boundary map of districts in Indonesia in 2011. The map was collected from Geospatial Information Agency. There are several types of horticultural crops analysed in this study that are medicinal plants, fruits, vegetables, and ornamental plants. At the initial system development, we selected three types of data for each commodity category based on the completeness of data and data with less null. Medicinal plants commodity includes ginger, galangal, and turmeric. Fruit commodity includes avocados, star fruit, and durian. Vegetable crops include beans, carrots, and chili. Ornamental plants include orchids, roses, and tuberose flower.

2.2. Analysis of spatial data warehouse horticultural crops

Spatial data analysis of horticultural crops was done to identify characteristics of the data. Spatial data are geographically oriented data and have coordinate system as a reference [1]. A spatial data warehouse is a collection of spatial data that has several characteristics such as subject oriented, integrated, time variant, and non-volatile [1]. Data analysis is useful for determining the scheme of data warehouse. In this study, the data warehouse's scheme applied is the star schema.

2.3. OLAP operations and horticultural crops clustering

OLAP is an application that integrates data warehouse in order to simplify information extracting from large data sets [1]. In this study, OLAP operations are performed based on the type of each commodity, location, status id, year, harvest area, production, and productivity. Furthermore, clustering of horticultural crops is done on OLAP results. Clustering is a process of partitioning a set of data objects into several groups [1]. Data clustering is based on the similarity between objects. Objects that have high similarity are placed on one cluster. The similarity between two objects can calculated based on distance between those objects. This study uses Euclidean distance to measure the similarity between objects. The Euclidean distance (*d*) between objects *i* and *j* is define as follows in which x_i and x_j are features respectively in object i and object j [1].

$$d(i,j) = \sqrt{(x_{i1} - x_{i1})^2 + (x_{i2} - x_{i2})^2 + \cdots + (x_{ip} - x_{ip})^2} \quad (1)$$

The quality of clustering results is measured by Sum of Square Error (SSE). SSE is calculated to obtain number of clusters that minimizes the square error. For example $p \in C_i$ is each data point in cluster i, c_i is the centroid of the cluster i, k is total number of clusters, *dist* is the distance to each cluster i [1]. The Sum of Square Error (SSE) is defined as follows [2]

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} \text{dist}(p, c_i)^2 \quad (2)$$

This study applies the well-known K-Means clustering algorithm. The steps in the K-Means algorithm are as follows [1]:

1. Select k initial objects as the centroids of the clusters,

2. Insert the objects into a cluster whose objects that are most similar based on the average value of objects in a cluster,
3. Renew the centroid of clusters by calculating the average value of objects for each cluster,
4. Repeat step two and three until the objects in the clusters have not changed.

2.4. Integrations of clustering modules using R and SpagoBI

This study utilizes the tools R and SpagoBI. SpagoBI is utilized to analyze horticultural crops using OLAP operations whereas R is used to cluster the results of OLAP operations. The integration between R and the clustering module on SpagoBI aims to cluster the output of OLAP operations using the data mining module that is available on SpagoBI and to present clustering results of horticultural crops in the tabular format.

2.5. Visualization of clustering results

The visualization of the clustering results is provided in form of map on the HTML page. In the development of the visualization module, this study utilizes ggplot2, maptools, rgdal, and plyr packages that are available in R. The visualization module provides a map that plots distribution of horticultural crops in district level in Indonesia.

3. Result and Discussion

3.1. Analysis of spatial data warehouse horticultural crops

The data warehouse of horticultural crops was designed based on the star schema with a single fact table and several dimension tables. The fact table contains unique attributes from each dimension and measures. Measures contained in the fact table are production (in tons), productivity (in tons/tree), and harvested area (in ha). There are 4 dimensions that are created in the data warehouse including commodity, location, time, and status id. Data analysis was done based on horticultural crops category, commodity, islands, province, district, year, and status of measure values. Figure 1 represents the star scheme for data warehouse of horticulture. The red sign on figure 1 indicates the measures to be calculated when the OLAP operations are performed.

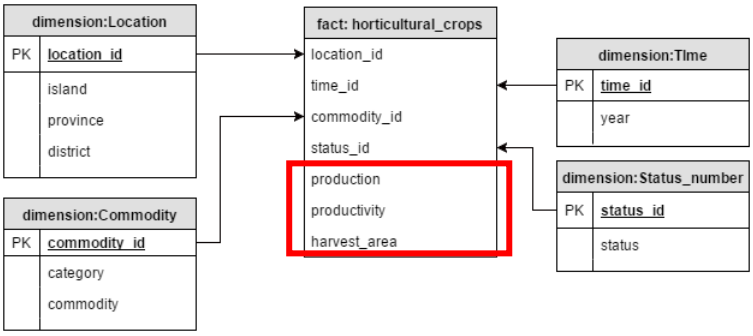

Figure 1. Star scheme for data warehouse of horticulture.

3.2. OLAP operations and horticultural crops clustering

OLAP operations are performed by selecting dimensions and measures on the data cube. OLAP operations on horticultural crops data include roll up, drill down, slice, and dice. The dimensions are commodity, location, status of measure values whereas the measures include harvested area (in hectares), production (in ton), and productivity (in ton/tree). The aggregation function sum is calculated for each measure in OLAP operations. Summary of horticulture data as results of OLAP operations are exported into the Microsoft Excel file on SpagoBI.

K-Means algorithm is applied to cluster datasets as the results of OLAP operations. Number of clusters (k) was chosen based on the lowest sum square error (SSE) to get the best clustering results.

Experimental results shows that the best clustering is obtained at the value k of 5. Therefore this work groups the horticulture data into 5 clusters with the label very low, low, medium, high, and very high. Table 1 shows the quality of clustering results at k of 5. Percentage of clustering quality in R is calculated by dividing *betweenss* with *totss*. The functions betweenss and totss are derived from the output component provided by the function of K-Means in R. For example, clustering results of harvested area, production, and productivity of durian fruit are presented respectively in table 2, table 3, and table 4.

Table 1. Quality of clustering results.

Number of cluster (k)	Clustering quality (%)		
	Harvested area (ha)	Production (ton)	Productivity (ton/tree)
3	91.07	81.63	92.09
4	96.72	86.68	95.39
5	98.16	87.81	96.12

Table 2. Clustering results of durian's harvested area.

Cluster	Cluster size	Centre	Minimum	Maximum.	Average	Percentage
1	12	32,321.500	23,335	44,972	32,322	2.73%
2	2	181,125.500	167,559	194,692	181,126	0.45%
3	399	208.251	0	6303	208.3	90.68%
4	19	13,199.368	7500	20,778	13,199	4.32%
5	8	63,805.625	52,895	75,813	63,806	1.82%

Table 3. Clustering results of durian's production (ton).

Cluster	Cluster size	Centre	Minimum	Maximum.	Average	Percentage
1	16	335,996.438	246,868	703,107	335,996	3.64%
2	327	3986.538	0	24,241	3987	74.32%
3	3	1,158,382.000	824,391	1,752,916	1,158,382	0.68%
4	35	130,450.314	88,980	224,386	130,450	7.95%
5	59	46,455.424	26,334	79,257	46,455	13.41%

Table 4. Clustering results of durian's productivity (ton/tree).

Cluster	Cluster size	Centre	Minimum	Maximum.	Average	Percentage
1	366	297.79	0	5725	297.79	83.18%
2	4	286,599.75	240,189	373,656	266,277	0.91%
3	7	106,352.14	79,605	130,677	106,352	1.59%
4	51	12,646.63	6993	21,131	12,647	11.59%
5	12	31,593.92	22,421	53,383	31,594	2.73%

3.3. Integration between R and data mining module on SpagoBI

This study uses the tools R version 3.2.4 and SpagoBI version 5.1.0. Clustering datasets is performed by integrating R with SpagoBI server. The integration between R and SpagoBI requires the rJava package. In rJava package, there are JRI files which are needed to execute scipts from R to Java. Furthermore, this study did the configuration on catalina.bat file that is contained in the folder bin in SpagoBI server. The configuration includes adding a path of JRI file that is available in rJava folder. The R code of Catalina.bat file configuration is given in figure 2.

```
set JAVA_OPTS="-Djava.library.path=C:\Program Files\R\R-3.2.4\library\rJava\jri\x64"
set R_HOME ="C:\Program Files\R\R-3.2.4"
set JAVA_OPTS= %JAVA_OPTS% -Xms1024m -Xmx1024m -XX:MaxPermSize=512m
```

Figure 2. Configuration on Catalina.bat file.

Integration between R and SpagoBI was implemented using the XML code template. On that template, R scripts are inserted in the SCRIPT tag. The program template is used when a new document is created on the data mining module in SpagoBI. Figure 3 shows the XML template to run R scripts on SpagoBI.

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<DATA_MINING>
  <DATASETS>
    <DATASET name="data" type="file" readType="table" canUpload="true" label="Upload">
      <![CDATA[header=TRUE,sep =",",dec="."]]>
    </DATASET>
  </DATASETS>
  <SCRIPTS>
    <SCRIPT name="k-means" mode="manual" datasets="data" label="k-means clustering" libraries="fpc,RCurl,R2HTML,plyr">
      <![CDATA[

            ]]>
    </SCRIPT>
  </SCRIPTS>
  <COMMANDS>
    <COMMAND name="k-means" scriptName="k-means" action="kmeans" label="k-means cluster" mode="manual">
      <OUTPUTS>
        <OUTPUT type="text" name="LabelCluster" value="Cluster" mode="manual" label="Label Cluster"></OUTPUT>
      </OUTPUTS>
    </COMMAND>
  </COMMANDS>
</DATA_MINING>
```

Figure 3. XML template to run R scripts on SpagoBI.

A new document in the data mining module on SpagoBI is created by selecting the analytical documents feature in the data mining folder. This work selects the generic document as one of the analytical documents types. Figure 4 shows the document fields in the data mining module of SpagoBI.

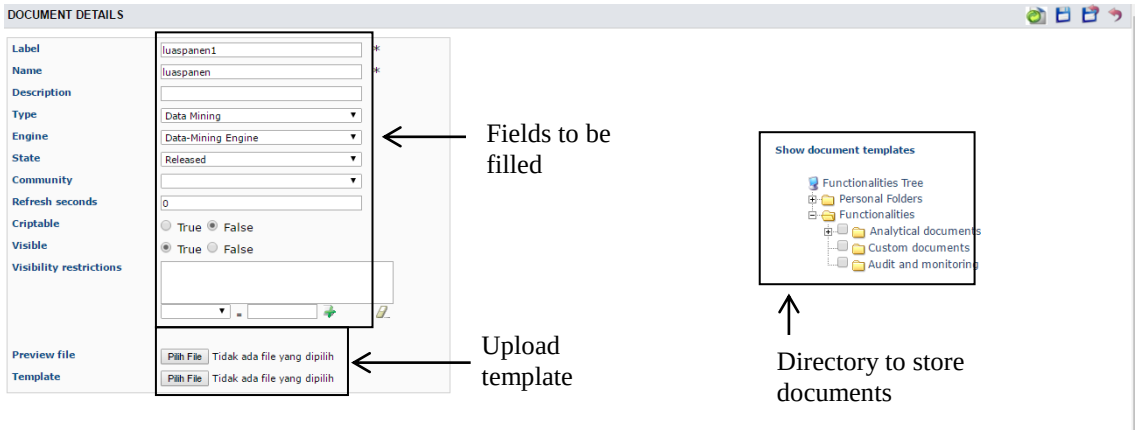

Figure 4. Document fields on the data mining module of SpagoBI.

Clustering results in SpagoBI are displayed in three types of output that are text, table, and image. The output type can be selected by adding the argument of text, html, or image on the output type in the XML template. Figure 5 shows the example of clustering results of durian commodity with the measure production that are represented in the text format.

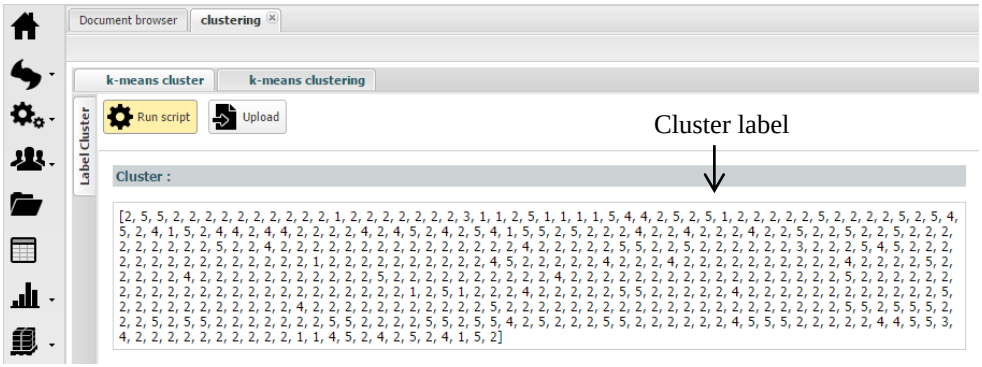

Figure 5. Example of clustering results represented in the text format.

In addition, clustering results can also be displayed in a tabular form using the data frame function. This feature requires the package R2HTML in R to display the output in tabular format on GUI of SpagoBI. Figure 6 represents the example of clustering results of durian commodity with the measure production represented in the tabular form.

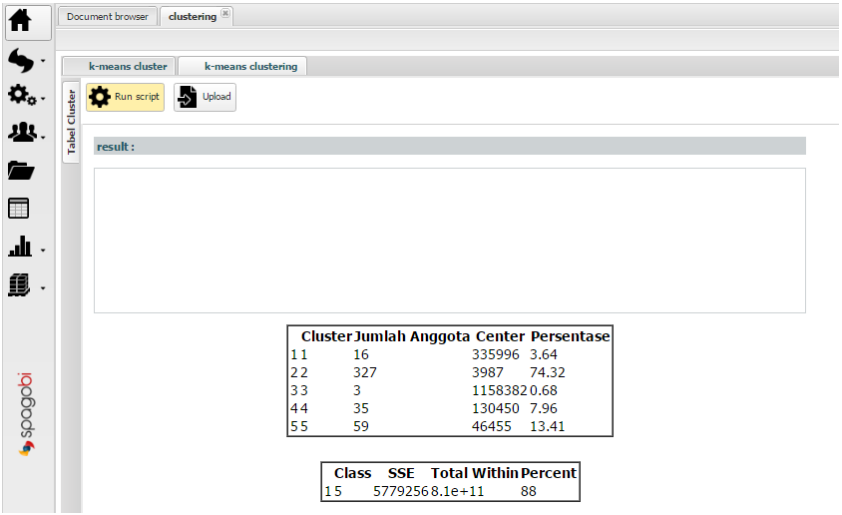

Figure 6. Example of clustering results represented in the tabular form.

3.4. Visualization of clustering results in a map

The visualization module of horticultural crops clusters was implemented using ggplot2, maptools, rgdal, and plyr packages that are available in R. The visualization module describes the distribution of horticultural crops in district level in Indonesia which is displayed on the HTML page. For example, figure 7 represents the distribution map of durian commodity based on harvested area.

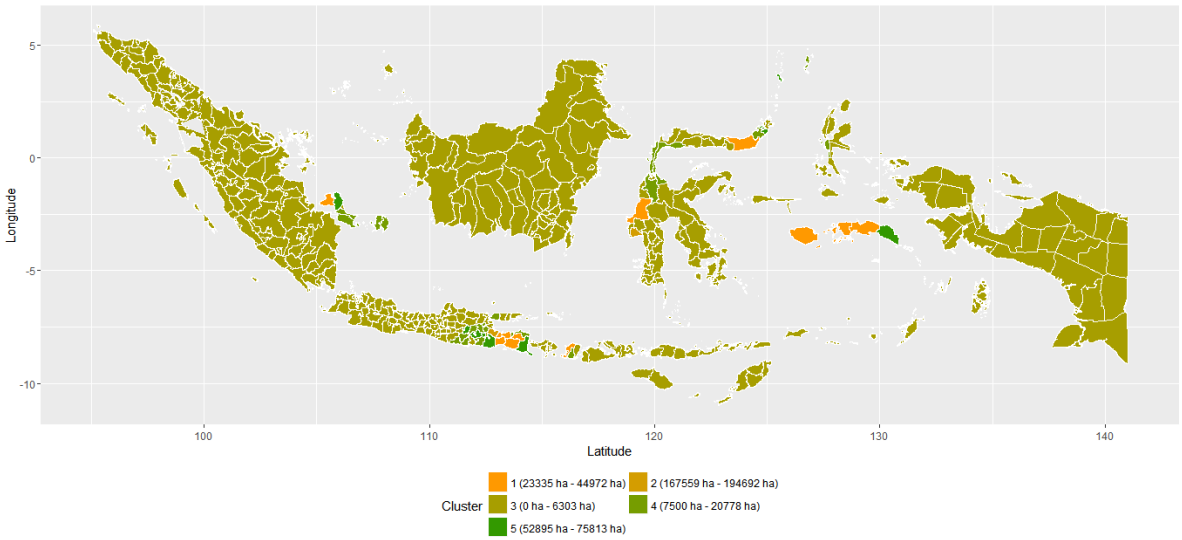

Figure 7. Distribution of durian commodity based on harvested area.

Different colors of each polygon on the distribution map in figure 7 represent different clusters. Each polygon in the map shows a district in Indonesia. Figure 7 shows that most of all districts in Indonesia are members of cluster 3 in which the harvested area in these districts is less than or equal to 6,303 ha. The main features of clustering module were tested based the black box approach. The main features include

- Load horticultural commodities dataset for clustering process
- Run the R scripts and displays the clustering results in the text format
- Run the R scripts and display the clustering results in the tabular format
- Plot clustering results on a map on the HTML page

The testing results show that all main features work properly.

4. Summary

A clustering module of horticultural crops data in the SOLAP system has successfully developed using SpagoBI integrated with R. Clustering results of horticultural crops are visualized in maps utilizing the ggplot2, maptools, rgdal, and plyr packages that are available in R. Maps are displayed on HTML pages that provide information about the distribution of horticultural crops in district level in Indonesia. Distribution of horticultural crops can be used to identify potential area for horticultural production based on harvested area (ha), production (tons), and productivity (tons/tree). System testing was carried out to evaluate the features in the visualization module. The testing results show that the clustering module in SpagoBI has successfully displayed horticulture crops clusters in the three types of output namely text, tabular and map.

References

[1] Han J, Kamber M and Pei J 2012 *Data Mining*. *Concept and Techniques* 3rd Edition (Amsterdam: Elsevier) pp 83-443
[2] Cazzin G 2012 *Business intelligence with SpagoBI* (Padua (IT) SpagoBI Competency Center) pp 2-59
[3] Golfarelli M 2009 Open source BI platforms: a functional and architectural comparison, *11th International Conference DaWak* (Austria. Bologna (IT): University of Bologna) pp 289-297