
Stefanie Intan Christienova, Yusep Suparman, Zulhanif and Choerul Afifanto

Department of Statistics, Padjadjaran University, Bandung, Indonesia
Computational Statistics, Institute of Statistics, Jakarta, Indonesia

Abstract: The geographical and socio-political conditions in remote areas, such as in Papua-Indonesia, may lead to the delay in reporting the results of data collection, changing sample and the increase of non-response rate. In many surveys, Statistics Indonesia (BPS) often excludes remote areas from sampling frame due to high transportation cost and difficult access. Nevertheless, it results in missing data problem, particularly of characteristics these areas. Therefore, it is essential to find substitutes which have similar characteristics (sister village) to the remote areas. Accordingly, in this paper we intend to identify sister villages for the remote villages by means of K-Harmonic Means (KHM) clustering method. The optimal solution from KHM is obtained by means of the Particle Swarm Optimization (PSO). The integration of PSO into KHM allows us to obtain a better clustering result, the best solution in a few number of iteration and to avoid trapping in local optima. This algorithm is applied to 143 villages data set in Lanny Jaya District, Papua. The validity indexes showed that PSOKHM performance is better than KHM algorithm. The analysis result indicates that the optimal number of cluster is ten clusters. The presence of non-remote and remote areas in a cluster enable us the substitution of remote areas by non-remote areas (sister village).

Key words: Remote Villages · Cluster Analysis · Substitution Sample

INTRODUCTION

As the national statistical agency, Statistics Indonesia (BPS) has an important role in providing trusted official statistics on a wide range of economic, social, population and environmental matters of importance to Indonesia. BPS’s main activities are undertaken in a regular cycle. There are three kinds of censuses which are conducted every ten years and a number of surveys. A survey is used when it is not possible nor practical to conduct a census [1]. In addition, a survey could be a solution because of its advantages, such as costs would generally belower than for a census, results may be available in less time, the scope of variables collected can be broader, activities are more easily monitored and non-sampling error tends to be smaller than a census.

On the other hand, survey is subject to various types of errors. First, error from the sampling design is called non-sampling error, such as coverage error, non-response error and measurement error. Non-response error is often encountered in remote areas, mostly in Papua. This problem may lead to a request of sample replacement. In order to reduce amount of those requests, BPS often excludes remote areas from sampling frame due to the inavailability of transportation means, high transportation cost and more travel time.

The exclusion of remote areas from sampling frame may have an impact on the sample representation. It is necessary to solve the problem so that survey results can make inference about the population as a whole. Therefore, the existence of other areas which have similar characteristics (sister village) to the remote area, is essential. The idea of sister village itself is inspired by the basic concept of sister city, as in Sinaga [2] that sister city is formed by two regions with similar characteristics. In this research, sister villages are built by using cluster analysis based on the socio-economic characteristics.
The most popular and well-known clustering method is K-Means (KM) algorithm [3]. According to Cheng et al. [4], KM was used to solve classification problems by involving training and teacher processes. Unfortunately, the cluster result of KM is sensitive to the selection of the initial cluster centers and may converge to the local optima [5]. Zhang et al. [6] proposed a method for solving problems in cluster center initialization on KM methods with K-Harmonic Means (KHM) methods, later modified by Hammerly et al. [7]. By using KHM algorithm, the problem of initialization of KM can be solved, but it also easily runs into local optima. Eberhart et al. [8] developed Particle Swarm Optimization (PSO) as a stochastic optimization technique to help the KHM algorithm escape from local optima. Yang et al. [9] and Saikhu et al. [10] proposed the hybrid based on KHM and PSO. The PSOKHM algorithm is superior to both the KHM and the PSO algorithms. The purpose of this research is to compare the clustering methods of KHM and PSOKHM and apply it to the data of village potential to build sister villages.

MATERIALS AND METHODS

Study Area: The proposed algorithm is applied on BPS data, Village Potential (PODES) 2014 data with 10 main variables related to socio-economics of 143 villages in Lanny Jaya District, Papua, Indonesia.

K-Harmonic Means (KHM): KHM is a center-based cluster. It provides weight to each data point dynamically. The purpose of this algorithm is to minimize the harmonic means of the distance from each data point to all cluster centers. On the contrary, KM searches the total distance if the membership value is the biggest one of all data point to the cluster center. Harmonic average of K numbers \( \{a_1, \ldots, a_k\} \) are defined by the following equation [9]:

\[
HA(\{a_1, \ldots, a_k\}) = \frac{K}{\sum_{k=1}^{K} a_k}
\]

The harmonic means gives a good (low) score for each data point is close to any one center. This is a property of the harmonic means and similar to the minimum function used by KM, but it is a smooth differentiable function. The KHM steps are as follows [11]:

- Initialize randomly chosen initial cluster centers.
- Calculate objective function value according to:

\[
f(X,M) = \sum_{i=1}^{n} \sum_{l=1}^{K} w(x_i) \frac{1}{\left\| x_i - m_l \right\|^{p-2}}
\]

- For each data point \( x_i \), compute its membership \( \text{mem}(m_l|x_i) \) in each center \( m_l \) according to:

\[
\text{mem}(m_l|x_i) = \frac{1}{\sum_{k=1}^{K} \left\| x_i - m_k \right\|^{p-2}}
\]

- For each data point \( x_i \), compute its weight \( w(x_i) \) according to:

\[
w(x_i) = \frac{\sum_{k=1}^{K} \left\| x_i - m_k \right\|^{p-2}}{\left(\sum_{k=1}^{K} \left\| x_i - m_k \right\|^{p-2}\right)^\frac{1}{2}}
\]

- For each center \( m_l \), recompute its location from all data points \( x_i \) according to its memberships and weights:

\[
m_l = \frac{\sum_{i=1}^{n} \text{mem}(m_l|x_i) w(x_i) x_i}{\sum_{i=1}^{n} \text{mem}(m_l|x_i) w(x_i)}
\]

- Repeat steps 2-5 until fitness function does not change significantly.
- Assign data point \( x_i \) to cluster \( l \) with the biggest \( \text{mem}(m_l|x_i) \).

\( x_i \) is a member of the cluster with the \( m_l \) cluster center if the \( \text{mem}(m_l|x_i) \) membership value is the biggest one compared to its membership value to the other cluster center point.

Particle Swarm Optimization (PSO): PSO method was introduced by James Kennedy and Russel Eberhart in 1995 and inspired by the social behavior of swarm. In PSO algorithm, birds in a swarm are represented as particles. Each particle is an individual and swarm is composed of particles. PSO uses an individual population to investigate promising areas of a solution space. Particles cooperate to find the best position (best solution) in the solution space. Each particle moves according to the velocity. At each iteration, the particle movement is computed as follows [11]:

\[
v_{ij}^{t+1} = v_{ij}^{t} + c_1 \cdot r_1 \cdot (p_{best} - x_{ij}^t) + c_2 \cdot r_2 \cdot (g_{best} - x_{ij}^t)
\]

\[
x_{ij}^{t+1} = x_{ij}^t + v_{ij}^{t+1}
\]
$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

$$v_i^{t+1} = k \left( w v_i^t + c_1 \text{rand} \left( pbest - x_i^t \right) + c_2 \text{rand} \left( gbest - x_i^t \right) \right)$$

where $x_i^t$: position of particle $i$ at time $t$

$v_i^{t+1}$: velocity of particle $i$ at time $t$

$pbest$: the best position found by particle $i$ itself

$gbest$: the best position found by the whole swarm

$\text{rand}$: random variable between 0 and 1

$c_1$ and $c_2$: two acceleration coefficients

$k$: denotes the constriction factor or the number of cluster

The value is determined by following equation:

$$k = \frac{2}{2 - c - \sqrt{c^2 - 4c}}$$

$\omega$: denotes the inertia weight factor is given by following equation:

$$\omega = \omega_{\text{max}} - \frac{\omega_{\text{max}} - \omega_{\text{min}}}{t_{\text{max}}} \times t$$

The best performance fitness value are updated at each generation based on the equation:

$$pbest_i^{t+1} = \begin{cases} pbest_i^t f(x_i(t+1)) \leq f(x_i(t)) \\ x_i(t+1) f(x_i(t+1)) > f(x_i(t)) \end{cases}$$

**Particle Swarm Optimization K-Harmonic Means (PSOKHM):** KHM algorithm tends to get faster convergence than the PSO algorithm because it requires fewer function evaluations but it usually gets stuck in local optima [10]. The integration of KHM with PSO forms a hybrid clustering algorithm, called PSOKHM, which maintains the merits of KHM and PSO. The fitness function of PSOKHM algorithm is the objective function of KHM algorithm. The algorithm for PSO based KHM is given below [11]:

- Set the initial parameters
- Set iterative count $t=0$
- Apply the PSO operator to update the swarm particles
- Calculate the position of particle’s $i^{th}$ cluster centroid vector by KHM algorithm
- If $t < t_{\text{max}}$, go to step 3, else
- Assign data point $x_i$ to cluster $l$ with the biggest $\text{mem}(m_{ix})$, as the initial cluster

**Validity Index:** Clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters) without any class information [12]. Based on the information intrinsic to the data alone, estimate the most optimal number of clusters is determined by the maximum value of silhouette index, dunn index and connectivity index.

- **Silhouette index**

  For a given cluster, $X_j = (j = 1,...,c)$, the silhouette technique assigns to the $i^{th}$ sample of $X_j$ a quality measure, $s(i) = (i=1,...,m)$, known as the silhouette width. This value is a confidence indicator on the membership of the $i^{th}$ sample in the cluster $X_j$ and it is defined as [12]:

  $$s(i) = \frac{(b(i) - a(i))}{\text{Max}[a(i),b(i)]}$$

  $a(i)$ is the average distance between the $i^{th}$ sample and all of samples included in $X_j$ and $b(i)$ is the minimum average distance between the $i^{th}$ and all of the samples clustered in $X_i = (k = 1,...,c; k \neq j)$.

- **Dunn index**

  $$Dunn = \min_{1 \leq i \leq c} \left\{ \min \left\{ \frac{d(c_j, c_j)}{\text{Max}_{x \in \mathcal{X}}(d(X_k))} \right\} \right\}$$

  Bigger value of $\text{SepG}(C)$ indicates that a valid optimal partition to the different given partitions is found.

- **Connectivity index**

  The connectivity index indicates the degree of connectedness of the clusters. It has a value between 0 and infinity and should be minimized [13].

  $$\text{Conn}(C) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,m_{ix}}$$

  $m_{ix}$ is the nearest neighbor of the $i^{th}$ data from the $i^{th}$ data. If it is in the same cluster, $y_{i,m_{ix}}$ is 0 and if it is in a different cluster then it is $1/j$. $N$ is the number of observation and $K$ is the number of cluster.
To objectively assess the validity of the cluster analysis results, the intra-cluster and inter-cluster standard deviations are computed and compared [14]. The intracluster standard deviation is calculated for a given method by using the equation:

\[ S_w = K^{-1} \sum_{k=1}^{K} S_k \]

is the standard deviation for a given variable in the \( k \)th of \( K \) clusters. Similarly, the intercluster standard deviation was calculated by using the following expression:

\[ S_B = \left[ (K-1)^{-1} \sum_{k=1}^{K} (\bar{X}_k - \bar{X})^2 \right]^{1/2} \]

\( \bar{X}_k \) is the cluster mean for a given variable and \( \bar{X} \) is the total mean for all the \( K \) clusters (\( \bar{X} = \) zero for each of the principal components).

RESULTS

During the period of 2010-2016, BPS records the increase of the number of remote area in Indonesia. In 2016, the number of remote area in Indonesia is 1,842 villages. The province with the largest number of remote area is Papua, it is 283 villages in 2010-2014 or 24.38% of the total difficult villages in Indonesia. In 2016 the number increased to 456 remote areas. Among twenty nine districts in Papua, Lanny Jaya has the biggest number of remote area. In 2014 there were 100 villages categorized as remote areas out of 143 villages in Lanny Jaya.

Firstly, by using cluster analysis, this research performs a comparison between KHM and PSOKHM methods. In order to see the performance of the two clustering methods, experiments were proceeded by forming 5 different clusters and varying parameter \( p \) values. The number of clusters formed from 10 to 15 clusters. In addition, both method was measured using parameter values \( p = 2, 2.5, 3, 3.5 \). Furthermore, validity indexes were calculated to see and compare the performance of the clustering resulted from both methods.

Performance of Clustering Results: Validity index measurements can be used to measure the performance of the grouping resulted by looking at the maximum value from both of silhouette index and dunn index.

The silhouette index showed the differences between the accuracy of KHM algorithm and PSOKHM at particulars value of parameter \( p \). For all \( p \) value, PSOKHM has better performances than KHM at \( p = 2 \). Meanwhile, dunn index showed that PSOKHM has better performances at \( p = 3 \). However, compared to all parameter, the results of silhouette index showed that PSOKHM has better performance at \( p = 3 \). The same condition is also indicated by the results of dunn index. Higher values of silhouette index and dunn index indicate better performance.

In Figure (2), the connectivity index which also showed that PSOKHM clustering with \( p = 3 \) is better than the other \( p \) value of KHM. It is because the value of connectivity index generated by PSOKHM is smaller than KHM. Smaller connectivity index values showed better performance.

At PSOKHM with \( p = 4 \) and \( p = 5 \) all the silhouette index, dunn index and connectivity index values equal to 0. This condition occured because the grouping results are not convergent. Hence the silhouette index, dunn index and connectivity index values are not calculated.

Number of Optimum Groups: The purpose of cluster analysis is to obtain similarity within the group (the standard deviation value within the group, denoted by \( S_w \), the minimum) and different from the other group (the standard deviation value between the groups, expressed by \( S_B \), the maximum). However, in real situation, it is difficult to fulfill. Therefore, the approach that can be done is to use the ratio of \( S_w / S_B \) where the smaller ratio indicates better group precision quality.

Table 1. showed the result of ratios based on the PSOKHM clustering with \( p = 3 \) and tested for the number of clusters of 10 to 15. The ratio generated on each cluster shows the highest value at 13 clusters and the smallest value at 10 clusters. Therefore, the optimum number of clusters used in this analysis is 10.

From Table (2), it can be seen that the result of village clustering that built the sister village. In a cluster, there are a number of non-remote areas which are expected to become substitute samples for remote areas due to the characteristic sister city.

2190
Fig. 1: Comparison of Clustering Results Based on Silhouette Index and Dunn Index

![Silhouette Index](image1)

![Dunn Index](image2)

Fig. 2: Comparison of Clustering Results Based on Connectivity Index

![Connectivity Index](image3)

Table 1: Ratio Comparison of PSOKHM Clustering with $p=3$

<table>
<thead>
<tr>
<th>Number of Cluster</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.4373</td>
</tr>
<tr>
<td>11</td>
<td>1.4932</td>
</tr>
<tr>
<td>12</td>
<td>1.7248</td>
</tr>
<tr>
<td>13</td>
<td>1.8077</td>
</tr>
<tr>
<td>14</td>
<td>1.6015</td>
</tr>
<tr>
<td>15</td>
<td>1.6099</td>
</tr>
</tbody>
</table>
Table 2: Sister Village Clustering Result by Area Category

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Area Category</th>
<th>Name of Village</th>
<th>Number of Village</th>
<th>Total Village</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Remote Area</td>
<td>Megagobak, Balimneri</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Kemiri, Ekanom, Longgalo, Balingga, Indawa</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Remote Area</td>
<td>Kulia, Yirene, Kwenukwe, Guninggame andugume, Bagi, Oka, Wahiragi, Argeneri,</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kuabaga, Gubo, Tiwi, Mabume, Kamuluk, Noho, Bonanip, Kukepake,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Golime</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Remote Area</td>
<td>Indugu, Guburni, Eyuni, Lelam, Yaneko, Kemiulume, Teyiko, Penggima, Juta,</td>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tigimia, Tinggipura, Luaren, Malagai, Dimba, Yugwena, Ekapame, Lualo, Yugumia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Mamiri, Kimbo, Bonom, Keloyak, Wanuga, Labora, Kondena, Gondura, Pirambor,</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Milinggame, Megalunik, Gipura</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Remote Area</td>
<td>Libome, Gume, Balime, Popome, Yeyugu, Bruyugu, Ekaba, Ogodome, Tinime, Yigemi</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mili, Tologi, Wiyagi, Lugobak, Bigipaga, Muara, Tikome, Anitila, Warngangome,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Waluwa, Langgime</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Remote Area</td>
<td>Ilunggume, Melendik, Karunggame, Tekul, Wamindak, Kuwanom, Kugame, Yamiga, Wi</td>
<td>19</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rin, Nileme, Yugumabur, Name, Dugume, Nenggeya, Munen, Wamiru, Lubutini, Nanim,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Giari</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Remote Area</td>
<td>Wenam, Umbanume, Yalipak, Pirime, Milimbo, Takobak, Golomi, Golo, Yiwili,</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Golopura, Nambume</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Remote Area</td>
<td>Gukop</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Oyi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Remote Area</td>
<td>Tina, Wame, Kelulome, Piwugun, Wupi, Salemo, Yudani, Yugumabur, Odika, Pirawun,</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Pindalo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Remote Area</td>
<td>Gamelia, Ayafofa</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Bokon</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Remote Area</td>
<td>Mokoni, Wulundia, Logom, Wupaga, Kuyawage, Dugu-Dugu, Ponalo, Tabakeker, Giwan,</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Non-Remote Area</td>
<td>Warituu</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gurika, Dura, Olume, Yakobak, Ninabua, Konikme</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Cluster 1 consists of 5 non-remote areas that can replace 2 remote areas. In the cluster 2, there are 1 non-remote areas that can be as replacement sample for 18 remote areas. In cluster 3, there are 12 non-remote areas and 18 remote areas. In cluster 4, there are 4 non-remote areas and 20 remote areas. In cluster 5, there are 11 non-remote areas and 19 remote areas. In cluster 6, there are 1 non-remote area and 1 remote area. In cluster 7, there are 2 non-remote areas and 11 remote areas and In cluster 10, there are 6 non-remote areas and 10 remote areas.

While some other clusters consist of remote areas that built an individual cluster, such as cluster 7 consists of 1 remote area only. Meanwhile, cluster 9 consists of just 1 non-remote area.

The existence of clusters covering areas of the same category can be an obstacle because these areas can not build sister village. The formation of such clusters can be caused by the existence of big differences in characteristics so that certain areas can not make a cluster with different area category.

CONCLUSION

Among 143 observation units of villages in Lanny Jaya, Papua-Indonesia, it obtained the result of the clustering of remote areas and non-remote areas. Non-remote areas within the same cluster with remote areas can be used as a substitute for remote areas samples in a survey. This is due to the similarity of characteristics so it is expected to overcome the problem of missing data on the characteristics of remote areas that
are not covered in the sample. The performance tests of clustering results show that PSOKHM is better than KHM and subsequently obtained the optimal cluster percentage formed is 10 clusters.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of Statistics Indonesia and Master Program in Applied Statistics of Padjadjaran University which help in terms of data and writing of this paper.

REFERENCES