Fuzziness Evaluation on Hybrid Context Based Clustering Methods with Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization Algorithm

Siti Nurmardia Abdussamad¹, Suci Astutik², Achmad Efendi²

¹,²Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya

*Corresponding author: sitinurmardiaabd@gmail.com

Abstract. This research contains to hybrid Context Based Clustering Method integrated with Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization (CFGWC-PSO). The computation time using CFGWC-PSO Algorithm is faster than other algorithms. CFGWC-PSO algorithm was applied on 11 variables from data factors causing the spread of dengue in East Java. One of the parameters used in this analysis is fuzziness (m), which is the parameter used to measure the level of obscurity from the clustering results. In this paper will use different fuzziness (m) values to evaluating best fuzziness value (m) which are appropriate used to clustering with CFGWC-PSO algorithm. CFGWC-PSO algorithm using fuzziness (m) = 1.5 and fuzziness (m) = 2, number of clusters = 2 then CFGWC-PSO will evaluated using IFV index. Based on IFV index found that the best clustering in this case with CFGWC-PSO algorithm is with using fuzziness value (m) = 2.

1. Introduction
Context based clustering is a clustering process that centralizes clustering according to the specified context variables without reducing the quality of the clustering. Integration of context based clustering with Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization (CFGWC-PSO) is to solve the problem of slow computing time. Fuzzy Geographically Weighted Clustering (FGWC) has a weakness, namely that it lies in the initialization process where the limitations in choosing the initial value of the cluster center because it is done randomly, and can cause the iteration process to fail to reach the optimum global solution. The method that is often used to solve the problem of finding an optimum global solution is metaheuristics. Metaheuristics are likely to achieve better solutions with less computation processing or shorter time than simple heuristic methods or other optimization algorithms [1]. The metaheuristic algorithm used in this study to solve the problem of finding an optimum global solution is the Particle Swarm Optimization (PSO) algorithm.

Generally, fuzzy clustering is minimizing the objective function where the main parameter is the membership function in fuzzy (membership function) or also known as fuzzier [2]. One of the stages of the analysis process of the hybrid Context Based Clustering method with Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization (CFGWC-PSO) is to determine the fuzziness (m) value that will be used in the study. Fuzziness (m) is a parameter used to measure the level of obscurity of the clustering results. Therefore, the focus in this paper is to evaluate the fuzziness or
fuzzier values used in the CFGWC-PSO hybrid method. Fuzziness (m) used in this study is fuzziness (m) = 1.5 and fuzziness (m) = 2. Determination of the appropriate fuzziness value is used by using the evaluation of the IFV index value. This method was applied on factors related to dengue fever in East Java. Vector control still used for cut off the chain of transmission in preventing dengue fever [3]. Therefore, a method is needed that can provide regional cluster analysis results based on the factors causing the spread of dengue fever. So, the government can be more focused in handling dengue fever cases.

2. Method

The fundamental difference between Fuzzy Geographically Weighted Clustering and Classical Cluster Analysis is the member clustering method. In classical cluster analysis, a member can only occupy one place in several clusters, and is not a member of other clusters. Meanwhile, in Fuzzy Geographically Weighted Clustering, a member can be a member of all existing clusters with a degree of membership at the center of the cluster. Fuzzy geographically weighted clustering is a geographically aware alternative by supporting the ability to apply population and distance effects to analyze geodemographic clusters.

The FGWC algorithm has several limitations in the initialization stage. First, the number of geodemographic clusters (clusters) must be defined manually by the user. Second, the center of the cluster (centroid) is determined randomly so that the iteration process fails to reach the optimum global solution. To overcome this limitation, the PSO algorithm is used to select a cluster center or membership matrix in the FGWC initialization phase. The objective functions of the FGWC that will be minimized are [4]:

\[ J_{FGWC}(U, V; X) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m |v_i - x_k|^2 \rightarrow \text{min} \]

where \( m \) is the exponential weight that determines the fuzziness of the clusters, \( u_{ik} \) is an element of the partition matrix, \( v_i \) is the cluster center, and \( x_k \) is the data point.

The center of the cluster itself can be determined using the following formula:

\[ v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} \]

On the other hand, the fuzzy cluster membership matrix before geographic modification can be calculated as follows:

\[ u_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{|v_i - x_j|}{|v_i - x_k|} \right)^{m-1}} \]

More specifically, the objective function \( J_{FGWC}(U, V; X) \) will be minimized by performing optimization through parameters U and V. The lagrange multiplier \( \lambda_k \) with constraint \( \sum_{i=1}^{c} u_{ik} = 1 \) is used to find the optimum value of \( u_{ik} \) and \( v_i \). The lagrange function for FGWC is then derived for each parameter and equalized to zero to obtain the optimum value of \( u_{ik} \) and \( v_i \). The results of the two objective function formulations are as follows:

\[ J_{FGWC}(V; X) = \sum_{i=1}^{c} \sum_{k=1}^{n} |v_i - x_k|^2 \left( \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} - x_k \right)^m \rightarrow \text{min} \]

\[ J_{FGWC}(U; X) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \left( \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} - x_k \right)^2 \rightarrow \text{min} \]
There are 3 methods in determining context variable such as [5]
a. random matrix with size $n \times 1$,
b. using FCM with determine one of the dimensions of the dataset to be used as a context variable, separate it as context $Y$ which is matrix size $n \times 1$, input the number of clusters and value of the clustering parameters used, apply FCM to context $Y$, then get a membership matrix and select one of the columns the membership matrix which has the most representative of the $k$-th data point to the context $Y$, the column chosen is the context variable value which has the form of matrix $n \times 1$ consisting of membership degrees from 0 to 1.
c. calculate mean and standard deviation using the formula below:

$$
\mu_Y = \frac{\sum_{k=1}^{n} x_{ky}}{n} \quad (6)
$$

$$
\sigma_Y = \frac{\sqrt{\sum_{k=1}^{n} (x_{ky} - \mu_Y)^2}}{n} \quad (7)
$$

Where $\mu_Y$ is mean of context $Y$, $\sigma_Y$ is standard deviation, $x_{ky}$ is data point $k$-th context $Y$, and $n$ is number of record.

IFV index is used to determine the best algorithm in this study. The equation of the IFV index is described as follows [6]:

$$
IFV = \frac{1}{c} \sum_{j=1}^{c} \left\{ \frac{1}{N} \sum_{k=1}^{N} \mu_{kj}^2 \left[ \log_2 c - \frac{1}{N} \sum_{k=1}^{N} \log_2 \mu_{kj} \right]^2 \right\} \frac{SD_{\max}}{\sigma_D} \quad (8)
$$

The maximum distance between cluster centers is described as follows:

$$
SD_{\max} = \max_{k \neq j} \| v_k - v_j \|^2 \quad (9)
$$

The divisor between each object from the cluster center is described as follows:

$$
\sigma_D = \frac{1}{c} \sum_{j=1}^{c} \left( \frac{1}{N} \sum_{k=1}^{N} \| x_k - v_j \|^2 \right) \quad (10)
$$

Where,

$u_{kj}$ = Membership value $j$-th data point in $i$-th cluster center

$N$ = Number of data points (records)

$c$ = Number of cluster

$v_k$ = $k$-th cluster center

The following are the steps for the CFGWC-PSO Algorithm:

1. Enter the data then the data reading process.
2. Then, enter the clustering parameters is number of clusters $(c) = 2$, threshold $\epsilon = 10^{-4}$ and different fuzziness which first step using $(m) = 1,5$ and second step using $(m) = 2$
3. Determine parameter particle swarm optimization and parameter of geographic modification such as $\alpha = 0.5$, $\beta = 0.5$
4. Apply context variable by defining the dimension that will be used as context variable and separating the variable used as the context variable from the dataset. Determination of context
variable with random matrix size $n \times 1$ which consists of random values between 0 until 1 as a context variable.
5. If we get conditions $n>d$ then the application will use the cluster center as a swarm particle (equation 4) as an objective function, whereas if we get conditions $n<d$ the application will use the membership matrix as a swarm particle (equation 5) as an objective function.
6. Initialize the center of the cluster or membership matrix as a swarm particle based on the conditions in step 5. Then, initialize the number of swarm particles. Application will choose the best solution in the initial population. If it has reached the maximum iteration or find the best solution then this condition is same as global minimum.
7. Application will update weighting value of inertia. The new swarm and the swarm is used as input in fitness function.
8. Fitness function will calculate the membership matrix and the center cluster (centroid).
9. Update value of inertia weighting, number of swarm, and swarm matrix. If the fitness function is minimized, then update the best solution found during iteration. If there is no increase in the value of fitness function, then use the last best solution obtained from the previous iteration.

3. Results
Clustering with fuzziness $(m) = 1.5$ and $m = 2$ with number of cluster $= 2$ processed produces the validation index as in figure 1. Previously, calculations had been made to determine the best fuzziness value. Based on the value of the index evaluation determination of the fuzziness value, shows that the fuzziness value appropriate for the case in this study is clustering with the fuzziness value $(m) = 2$. Determination of the best algorithm from 2 variants of fuzziness value can be seen in the IFV Validity Index value. Figure 1 shows the comparison of IFV index for dataset East Java.

![Figure 1. Comparison of IFV Index](image)

The evaluation results using IFV index are indicated by the greater IFV index value, which is the better clustering quality. Based on IFV index calculation results in Figure 1 shows that the index value for the fuzziness value $(m) = 2$ algorithm has a greater value than the index with fuzziness value $(m) = 1.5$. So, the fuzziness value $(m) = 2$ is a better for used in this algorithm for the case study in this research. The results of this study are confirmed by research by Klawonn and Hopner (2001) which informs that the value of fuzziness value $(m)$ is often used and is considered the most refined is fuzziness value $(m) = 2$. Next step, clustering with fuzziness value $(m) = 2$ and number of clusters $= 2$.

| ID | Districts/Cities | Membership Value |
|----|------------------|------------------|
| 1  | Kab. Pacitan      | 0.258            |
|    |                   | 0.741            |
From table 1, cluster determination can be determined from the value of the largest degree of membership. Then, the membership value is created in a Table 2.

|   | Kab. Ponorogo | 0.572 | 0.428 |
|---|---------------|-------|-------|
| 2 | Kab. Trenggalek | 0.281 | 0.719 |
| 3 | Kab. Tulungagung | 0.231 | 0.769 |
| 4 | Kab. Blitar | 0.168 | 0.832 |
| 5 | Kab. Kediri | 0.193 | 0.807 |
| 6 | Kab. Malang | 0.184 | 0.815 |
| 7 | Kab. Lumajang | 0.571 | 0.429 |
| 8 | Kab. Jember | 0.554 | 0.446 |
| 9 | Kab. Banyuwangi | 0.506 | 0.494 |
| 10 | Kab. Bondowoso | 0.575 | 0.424 |
| 11 | Kab. Situbondo | 0.671 | 0.328 |
| 12 | Kab. Probolinggo | 0.643 | 0.356 |
| 13 | Kab. Pasuruan | 0.272 | 0.727 |
| 14 | Kab. Sidoarjo | 0.341 | 0.658 |
| 15 | Kab. Mojokerto | 0.555 | 0.444 |
| 16 | Kab. Jombang | 0.533 | 0.466 |
| 17 | Kab. Nganjuk | 0.184 | 0.816 |
| 18 | Kab. Madiun | 0.175 | 0.825 |
| 19 | Kab. Magetan | 0.464 | 0.536 |
| 20 | Kab. Ngawi | 0.224 | 0.775 |
| 21 | Kab. Bojonegoro | 0.248 | 0.752 |
| 22 | Kab. Tuban | 0.485 | 0.515 |
| 23 | Kab. Lamongan | 0.284 | 0.716 |
| 24 | Kab. Gresik | 0.339 | 0.661 |
| 25 | Kab. Bangkalan | 0.542 | 0.458 |
| 26 | Kab. Sampang | 0.227 | 0.773 |
| 27 | Kab. Pamekasan | 0.238 | 0.761 |
| 28 | Kab. Sumenep | 0.315 | 0.684 |
| 29 | Kota Kediri | 0.244 | 0.756 |
| 30 | Kota Blitar | 0.306 | 0.693 |
| 31 | Kota Malang | 0.222 | 0.777 |
| 32 | Kota Probolinggo | 0.324 | 0.676 |
| 33 | Kota Pasuruan | 0.281 | 0.719 |
| 34 | Kota Mojokerto | 0.366 | 0.633 |
| 35 | Kota Madiun | 0.281 | 0.718 |
| 36 | Kota Surabaya | 0.509 | 0.491 |
| 37 | Kota Batu | 0.311 | 0.688 |
| 38 | Kota Surabaya | 0.509 | 0.491 |
| 39 | Kota Madiun | 0.281 | 0.718 |
| 40 | Kota Surabaya | 0.509 | 0.491 |
| 41 | Kota Madiun | 0.281 | 0.718 |
| 42 | Kota Surabaya | 0.509 | 0.491 |
| 43 | Kota Madiun | 0.281 | 0.718 |
| 44 | Kota Surabaya | 0.509 | 0.491 |
| 45 | Kota Madiun | 0.281 | 0.718 |
| 46 | Kota Surabaya | 0.509 | 0.491 |
| 47 | Kota Madiun | 0.281 | 0.718 |
| 48 | Kota Surabaya | 0.509 | 0.491 |
| 49 | Kota Madiun | 0.281 | 0.718 |
| 50 | Kota Surabaya | 0.509 | 0.491 |
| 51 | Kota Madiun | 0.281 | 0.718 |
| 52 | Kota Surabaya | 0.509 | 0.491 |
| 53 | Kota Madiun | 0.281 | 0.718 |
| 54 | Kota Surabaya | 0.509 | 0.491 |
| 55 | Kota Madiun | 0.281 | 0.718 |
| 56 | Kota Surabaya | 0.509 | 0.491 |
| 57 | Kota Madiun | 0.281 | 0.718 |
| 58 | Kota Surabaya | 0.509 | 0.491 |
| 59 | Kota Madiun | 0.281 | 0.718 |
| 60 | Kota Surabaya | 0.509 | 0.491 |
| 61 | Kota Madiun | 0.281 | 0.718 |
| 62 | Kota Surabaya | 0.509 | 0.491 |
| 63 | Kota Madiun | 0.281 | 0.718 |
| 64 | Kota Surabaya | 0.509 | 0.491 |
| 65 | Kota Madiun | 0.281 | 0.718 |
| 66 | Kota Surabaya | 0.509 | 0.491 |
| 67 | Kota Madiun | 0.281 | 0.718 |
| 68 | Kota Surabaya | 0.509 | 0.491 |
| 69 | Kota Madiun | 0.281 | 0.718 |
| 70 | Kota Surabaya | 0.509 | 0.491 |
| 71 | Kota Madiun | 0.281 | 0.718 |
| 72 | Kota Surabaya | 0.509 | 0.491 |
| 73 | Kota Madiun | 0.281 | 0.718 |
| 74 | Kota Surabaya | 0.509 | 0.491 |
| 75 | Kota Madiun | 0.281 | 0.718 |
| 76 | Kota Surabaya | 0.509 | 0.491 |
| 77 | Kota Madiun | 0.281 | 0.718 |
| 78 | Kota Surabaya | 0.509 | 0.491 |
| 79 | Kota Madiun | 0.281 | 0.718 |
| 80 | Kota Surabaya | 0.509 | 0.491 |
| 81 | Kota Madiun | 0.281 | 0.718 |
| 82 | Kota Surabaya | 0.509 | 0.491 |
| 83 | Kota Madiun | 0.281 | 0.718 |
| 84 | Kota Surabaya | 0.509 | 0.491 |
| 85 | Kota Madiun | 0.281 | 0.718 |
| 86 | Kota Surabaya | 0.509 | 0.491 |
| 87 | Kota Madiun | 0.281 | 0.718 |
| 88 | Kota Surabaya | 0.509 | 0.491 |
| 89 | Kota Madiun | 0.281 | 0.718 |
| 90 | Kota Surabaya | 0.509 | 0.491 |
| 91 | Kota Madiun | 0.281 | 0.718 |
| 92 | Kota Surabaya | 0.509 | 0.491 |
| 93 | Kota Madiun | 0.281 | 0.718 |
| 94 | Kota Surabaya | 0.509 | 0.491 |
| 95 | Kota Madiun | 0.281 | 0.718 |
| 96 | Kota Surabaya | 0.509 | 0.491 |
| 97 | Kota Madiun | 0.281 | 0.718 |
| 98 | Kota Surabaya | 0.509 | 0.491 |
| 99 | Kota Madiun | 0.281 | 0.718 |
| 100 | Kota Surabaya | 0.509 | 0.491 |
Table 2. Determination of Cluster Areas

| ID | Districts/Cities      | Membership Value | Cluster |
|----|-----------------------|------------------|---------|
|    |                       | 1    | 2   |         |
| 1  | Kab. Pacitan          | 0.258 | 0.741 | 2       |
| 2  | Kab. Ponorogo         | 0.572 | 0.428 | 1       |
| 3  | Kab. Trenggalek      | 0.281 | 0.719 | 2       |
| 4  | Kab. Tulungagung      | 0.231 | 0.769 | 2       |
| 5  | Kab. Blitar           | 0.168 | 0.832 | 2       |
| 6  | Kab. Kediri           | 0.193 | 0.807 | 2       |
| 7  | Kab. Malang           | 0.184 | 0.815 | 2       |
| 8  | Kab. Lumajang         | 0.571 | 0.429 | 1       |
| 9  | Kab. Jember           | 0.554 | 0.446 | 1       |
| 10 | Kab. Banyuwangi       | 0.506 | 0.494 | 1       |
| 11 | Kab. Bondowoso        | 0.575 | 0.424 | 1       |
| 12 | Kab. Situbando        | 0.671 | 0.328 | 1       |
| 13 | Kab. Probolinggo      | 0.643 | 0.356 | 1       |
| 14 | Kab. Pasuruan         | 0.272 | 0.727 | 2       |
| 15 | Kab. Sidoarjo         | 0.341 | 0.658 | 2       |
| 16 | Kab. Mojokerto        | 0.555 | 0.444 | 1       |
| 17 | Kab. Jombang          | 0.533 | 0.466 | 1       |
| 18 | Kab. Nganjuk          | 0.184 | 0.816 | 2       |
| 19 | Kab. Madiun           | 0.175 | 0.825 | 2       |
| 20 | Kab. Magetan          | 0.464 | 0.536 | 2       |
| 21 | Kab. Ngawi            | 0.224 | 0.775 | 2       |
| 22 | Kab. Bojonegoro       | 0.248 | 0.752 | 2       |
| 23 | Kab. Tuban            | 0.485 | 0.515 | 2       |
| 24 | Kab. Lamongan         | 0.284 | 0.716 | 2       |
| 25 | Kab. Gresik           | 0.339 | 0.661 | 2       |
| 26 | Kab. Bangkalan        | 0.542 | 0.458 | 1       |
| 27 | Kab. Sampang          | 0.227 | 0.773 | 2       |
| 28 | Kab. Pamekasan        | 0.238 | 0.761 | 2       |
| 29 | Kab. Sumenep          | 0.315 | 0.684 | 2       |
| 30 | Kota Kediri           | 0.244 | 0.756 | 2       |
| 31 | Kota Blitar           | 0.306 | 0.693 | 2       |
| 32 | Kota Malang           | 0.222 | 0.777 | 2       |
Based on table 2, it can be seen that the number of districts/ cities, which cluster 1 having 11 districts/ cities and cluster 2 having 27 districts/ cities. We get that number of districts/city in cluster 2 more than cluster 1. Data processing also get the average clustering variable results which shows in cluster 1 there are 4 high value variables that affect the factors causing the spread of dengue fever. Whereas in cluster 2 there are 7 high value variables that influence the causes of the spread of dengue fever. So, handling of dengue fever can be focused on areas in cluster 2 especially on the factors that most influence the spread of dengue fever.

### 4. Conclusion

In this paper, we propose the hybrid methods is Context Based Clustering with Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization to contribute development of fuzzy clustering analysis. One of the parameters used in this analysis is fuzziness (m), this paper try to get best fuzziness (m) which used in this case of research and from data processing we get that best fuzziness (m) for this paper is fuzziness (m) = 2. Clustering with best fuzziness (m) = 2 and number of clusters = 2 gives result number of area in cluster 2 is more than in cluster 1.

### 5. References

[1] Hadi, BS. (2017). “Pendekatan Modified Particle Swarm Optimization dan Artificial BeeColony pada Fuzzy Geographically Weighted Clustering”. Tesis. Institut Teknologi Sepuluh Nopember.
[2] Klawonn dand Hoppner. (2001). “What is Fuzzy about Fuzzy Clustering? Understanding and Improving the Concept of the Fuzzier”. Science Journal, 2810, pp. 254-264.

[3] Abdussamad, S., Suci, A. dan Achmad, E. (2019). “Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization using Context Based Clustering (CFGWC-PSO)”. Journal of Experimental and Life Science, Vol. 9, No. 3, pp. 201-204.

[4] Wijayanto, A.W. dan Purwariantini, A. (2014). "Improvement of Fuzzy Geographically Weighted Clustering Using Particle Swarm Optimization", In Information Technology Systems and Innovation (ICITSI), IEEE.

[5] Nurmala, N. dan Ayu, P. (2015). “Improvement of fuzzy geographically weighted clustering-ant colony optimization using Context Based Clustering”. International Conference on Information Technology Systems and Innovation (ICITSI), IEEE.

[6] Chunchun, H., Lingkui, M. dan Wenzhong, S. (2008). "Fuzzy Clustering Validity For Spatial Data". Geo-spatial Information Science, Vol. 11, No. 3, pp. 191196.

[7] Son, L.H. (2014). Enhancing clustering quality of geodemographic analysis using Context Fuzzy Clustering Type-2 and Particle Swarm Optimization. J. Appl. Soft Comput. 22. 566- 584.

[8] Central Bureau of Statistics. (2018). East Java in numbers 2018. Central Bureau of Statistics. East Java Province.

[9] East Java Province Public Health Office. (2018). Profile of Health in East Java 2018. East Java Province Public Health Office. East Java Province.

Acknowledgement
Author thanks to thesis supervisor who has provided guidance and advice to this research. Author also thanks to Ministry of Research, Technology and Higher Education of Republic Indonesia for granting master thesis grants 2020 so it can help to completion this research.