Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 14(2) 2021: 137-145
http://ejournal.undip.ac.id/index.php/media_statistika

COMPARATIVE STUDY OF DISTANCE MEASURES ON FUZZY SUBTRACTIVE CLUSTERING

Annisa Eka Haryati[1], Sugiyarto Surono[2]
[1] Magister Pendidikan Matematika, Universitas Ahmad Dahlan
[2] Matematika, Universitas Ahmad Dahlan

e-mail: sugiyarto@math.uad.ac.id

DOI: 10.14710/medstat.14.2.137-145

Article Info:
Received: 18 January 2021
Accepted: 12 December 2021
Available Online: 11 Januari 2022

Keywords:
Fuzzy Subtractive Clustering, Hamming, Combination of Minkowski Chebysev

Abstract: Clustering is a data analysis process which applied to classify the unlabeled data. Fuzzy clustering is a clustering method based on membership value which enclosing set of fuzzy as a measurement base for classification process. Fuzzy Subtractive Clustering (FSC) is included in one of fuzzy clustering method. This research applies Hamming distance and combined Minkowski Chebysev distance as a distance parameter in Fuzzy Subtractive Clustering. The objective of this research is to compare the output quality of the cluster from Fuzzy Subtractive Clustering by using Hamming distance and combine Minkowski Chebysev distance. The comparison of the two distances aims to see how well the clusters are produced from two different distances. The data used is data on hypertension. The variables used are age, gender, systolic pressure, diastolic pressure, and body weight. This research shows that the Partition Coefficient value resulted on Fuzzy Subtractive Clustering by applying combined Minkowski Chebysev distance is higher than the application of Hamming distance. Based on this, it can be concluded that in this study the quality of the cluster output using the combined Minkowski Chebysev distance is better.

1. INTRODUCTION

Clustering is a data analysis process which applied to classify the unlabeled data (Gan et al., 2007). In the optimum classification, each set of data will have a high percentage of similarity on certain cluster. Principally, cluster analysis will classify the given set of data which have high similarity into the same cluster and the dissimilar set of data into different cluster (Rencher, 2016).

Fuzzy clustering is a clustering method which applied to determine the set of data into certain cluster based on their membership value (Sharma & Verma, 2019). There are no certainty on each data to be included in one certain cluster. Hence, each data will have the possibility to be included as a member in different cluster (Jang et al., 2005). There are few known method for fuzzy clustering process, which are Fuzzy C-Means method and Fuzzy Subtractive Clustering.

Fuzzy Subtractive Clustering is one of the clustering method where the number of cluster classification is still undetermined yet. The basic concept of this method is to establish the coordinate of each data which possess the highest number of density value. Set of coordinate which have the highest number of neighboring coordinates will be chosen as the cluster centroid and its density values will get reduced (Sangadji,*et.al*,2018). In this method, the obtained number of cluster are affected by certain parameter, which is the radius (Dyvak *et al*., 2018).

Several studies have been carried out by applying clustering, such as the research conducted by (Gubu et al., 2021) for selecting a clustering portfolio and (Hidayatur Rifa & Pratiwi, 2020) using K-medoids and K-means for earthquake risk management in Indonesia. In addition, research using Fuzzy Subtractive Clustering was carried out by (Ghane'i Ostad *et al*., 2018) to determine the overlapped community on LBSN. In this research, every cluster centroid will be defined by the application of Fuzzy Subtracting Clustering based on each set of data potentials. This research concludes that the given method have higher accuracy than the previous used method.

The research which analyzed by (Salah *et al*., 2019) is applying subtractive classification which combined with swam particle optimization based on fuzzy classifier. Furthermore, (Abdolkarimi & Mosavi, 2020) also do a research related with fuzzy inference system for subtractive clustering which combined with wavelet for improving the navigation system. Other than that, (Banteng *et al*., 2019) and (Zeng *et al*., 2019) also applying subtractive clustering to analyze ad hoc cellular network based on information criteria for forecast process based on linear combination with independent variables.

Another research also worked by (Benmouiza & Cheknane, 2019) with the objective to forecast the solar radiation value by applying subtractive clustering, fuzzy c-means, and network partition. Three clustering methods used in this research is done for classifying every given set of data into certain classification so that each data included on certain classification have the high percentage of similarity to improve the understanding of the correlation between each data and to simplify the forecasting process.

In Fuzzy Subtractive Clustering method, the similarity measurement to determine the number of points which have the highest number of neighboring points is needed. The most frequently used distance parameter to determine the similarity measurement is Euclidean distance. In (Rezaei & Rezaei, 2020), the most known distance measurement to measure the distance between two set of fuzzy clusters is Hamming distance. This statement became the basis on the application of Hamming distance for this research. Other than the Hamming distance, the combined Minkowski Chebysev distance application also had been analyzed by (Rodrigues, 2018), for classifying process by using K-Nearest Neighbors (KNN) and (Surono & Putri, 2020) for clustering with Fuzzy C-Means and Principal Component Analysis application. The comparison of the two distances aims to see how well the clusters are produced from two different distances. Based on the statement below, this research will compare the achieved result through the application of Hamming distance and Minkowski Chebysev combination distance for clustering process using Fuzzy Subtractive Clustering method.

2. LITERATURE REVIEW

2.1. Fuzzy Subtractive Clustering

Fuzzy clustering is a clustering method based on the membership values which enclose fuzzy cluster as a clustering measurement basis. Every set of data are provided with probability value for their classification into existing group, this means that this each of data is not absolutely included in only one cluster and they will have a probability value to be classified in different cluster which have the highest percentage of membership level.

Fuzzy Subtractive Clustering (FSC) is one of the fuzzy clustering method where the number of existing cluster is unestablished yet. The basic concept of this method is to determine every coordinates of each data which have the highest density with their neighboring points. The coordinate with the highest number of neighboring points will be used as the cluster centroid. Then, the density level of the coordinate which used as a cluster centroid will be reduced and algorithm will determine the different coordinates which have the highest number of neighboring points to become another cluster centroid. This process will be executed until every set of coordinates have been tested (Kusumadewi & Purnomo, 2010).

2.2. Hamming Distance

Suppose the fuzzy sets A and B are subsets of $U = \{u_1, u_2, \cdots, u_n\}$. Hamming distance is defined as follows (Chen & Deng, 2020):

$$d_{hamming} = |\mu_A(u_i) - \mu_B(u_i)| \tag{1}$$

2.3. Combination of Minkowski Chebysev Distance

Rodrigues brings up a new distance that is a combination of the Minkowski and Chebysev distances. If w_1 is greater than w_2 then the distance is more like Minkowski, whereas if it's the other way around then the distance is more like Chebysev. The combination of the Minkowski Chebysev distance is as follows (Rodrigues, 2018):

$$d_{minkowski\ chebysev} = w_1 \sqrt[p]{\sum_{m=1}^{k} |x_m - y_m|^p} + w_2 max_{m=1}^{k} |x_m - y_m| \tag{2}$$

3. RESEARCH METHODS

3.1. Fuzzy Subtractive Clustering Steps

The method used in this research is Fuzzy Subtractive Clustering (FSC) with distance parameters, namely Hamming distance and Minkowski Chebysev combination distance. The obtained clusters will be evaluated using the Partition Coefficient to see the quality of the clusters. The data used in this study is hypertension patient data. The variables used were age (X_1), gender (X_2), systolic pressure (X_3), diastolic pressure (X_4), and body weight (X_5). The data used was obtained from one of the public health centers in Yogyakarta. Data processing is done with the help of the Python programming language. Information about the methods used in this study will be described as follows:

a. Determine the parameter values, which are, r (radius), q (squash factor), accept ratio, reject ratio. Squash factor is a constant to determine the radius of the data points around the center of the cluster whose data potential decline will be measured. The radius is a vector that will determine how much influence the cluster center has on each data. Accept

ratio is the lower limit where a data point that is a candidate for the center of the cluster is allowed to become the center of the cluster. While the reject ratio is the upper limit where a data point that is a candidate for the center of the cluster is not allowed to become the center of the cluster.

b. Transform the given set of data into fuzzy number by applying membership function (Debnath & Gupta, 2019; Mahajan & Gupta, 2019):

$$\mu(x) = \begin{cases} 1 & x \leq a \\ \dfrac{e^{-\left(\frac{x-a}{b-a}\right)} - e^{-s}}{1 - e^{-s}} & a \leq x \leq b \\ 0 & x \geq b \end{cases} \quad (3)$$

c. Determine the potential of every given set of data D_i ; i = 1,2,3, … , n with steps as follows:
First, measure the distance for every given set of data by applying:
Hamming Distance

$$Dist_{ij} = \left(\frac{|\mu_A(u_i) - \mu_B(u_i)|}{r} \right) \quad (4)$$

Combination of Minkowsi Chebysev distance

$$Dist_{ij} = \left(\frac{w_1 \sqrt[p]{\sum_{m=1}^{k} |x_m - y_m|^p} + w_2 max_{m=1}^{k} |x_m - y_m|}{r} \right) \quad (5)$$

Second, determine the initial potential of each data by applying:

$$D_i = \sum_{k=1}^{n} e^{-4\left(\sum_{j=1}^{m} Dist_{ij}^2(x_i)\right)} \quad (6)$$

d. Find the data coordinates with the highest level of potential:

$M = max[D_i | i = 1,2, \dots,]$; for the first iteration.

$Z = max[D_i | i = 1,2, \dots,]$; for the second, third, and the following iteration.

e. Measure the ratio (R) of the cluster centroid candidates by applying the following equation:

$$R = \frac{Z}{M} \quad (7)$$

In initial iteration, value Z=M.

f. Check the suitability of the cluster centroid candidates to become the cluster centroid by using 3 condition as follows:

Condition 1: If ratio > accept ratio, the cluster centroid candidates are suitable to become the cluster centroid.

Condition 2: If reject ratio < ratio ≤ accept ratio, further suitability test will be executed to determine whether the cluster centroid candidates are suitable to become the new cluster centroid. If the result of suitability test still below the minimum requirement for cluster centroid suitability value, the iteration process will be terminated because there

are no further data available for further consideration. The steps for executing the second condition process are:

$$Md = -1$$

for $k = 1,2, \dots, p$, where p= the number of cluster.

$$Sd_k = \sum_{j=1}^{m} \left(\frac{V_j - C_{kj}}{r} \right)^2 \quad (8)$$

If $(Md < 0)$ or $(Sd_k < Md)$, then $Md = Sd_k$,

$$Mds = \sqrt{Md};$$

Where M*ds* is the nearest distance between the coordinates of the cluster centroid candidates to the cluster centroid. If $(ratio + Mds) \geq 1$; the cluster centroid candidates are suitable to become the new cluster centroid. Meanwhile, if $(ratio + Mds) < 1$; then, the cluster centroid candidates are not suitable and won't be reconsidered as the new cluster centroid (potential of the data is set to 0)

Condition 3: if ratio ≤ reject ratio, there will be no further data consideration to become the cluster centroid candidates and the iteration process will be terminated.

g. If the new cluster centroid has already acquired, the data potential around the existing cluster centroid will be reduced by using equation:

$$D_i^t = D_i^{t-1} - D_{c_{ki}} \quad (9)$$

where

$$D_{c_{ki}} = Z * e^{-4\left[\sum_{j=1}^{m}\left(\frac{C_{kj} - x_{ij}}{r*q}\right)^2\right]} \quad (10)$$

h. Retransform the cluster centroid into the its data original form:

$$x = (a - b) \ln(\mu - \mu e^{-s} + e^{-s}) + a \quad (11)$$

i. Measure the membership degree by using equation:

$$\mu_{k_i} = e^{-\sum_{j=1}^{m}\left(\frac{x_{ij} - C_{kj}}{\sqrt{2}\sigma_j}\right)} \quad (12)$$

where

$$\sigma_j = \frac{r * \left(X_{max_j} - X_{min_j}\right)}{\sqrt{8}} \quad (13)$$

j. Determine the validity index by applying Partition Coefficient (Utomo & Marutho, 2018)

$$PC = \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} \mu_{ij}^2\right) \quad (14)$$

with N is the number of research object, K is the number of cluster, and μ_{ij} is the value of the i-th membership at the center of cluster j.

4. RESULT AND DISCUSSION

FSC measurement process will resulting in number of acquired cluster and Partition Coefficient value. The parameter value which applied in this research are q = 1.25, accept ratio = 0.8, reject ratio = 0.2, r = 0.72, 0.79, 0.97, 1.12, 1.31, and 1.6. The radius (r) is obtained by simulating the data and the sqash factor is obtained from (Azizah *et al*., 2018). Data which used for FSC measurement will be transformed into fuzzy number by applying equation (1) with the result as follows

Table 1. Fuzzy Numbers

X_1	X_2	X_3	X_4	X_5
0	0	0.4609	0.1289	0.4018
0.3913	1	0.5516	0.4785	0.1468
⋮	⋮	⋮	⋮	⋮
0.5734	0	0.3775	0.2862	0.6813

Furthermore, the results of clustering for Hamming distance using r = 0.72 obtained 4 clusters; by using r =0.79 obtained 3 clusters; and by using r = 0.97 obtained 2 clusters. The achieved clusters centroid are described as follows:

$$C_{0.72} = \begin{bmatrix} 0.4340 & 0 & 0.4609 & 0.2862 & 0.4018 \\ 0.4634 & 1 & 0.4609 & 0.2862 & 0.6501 \\ 0.4196 & 1 & 0.4609 & 0.2862 & 0.1468 \\ 0.3913 & 0 & 0.5144 & 0.1579 & 1 \end{bmatrix}$$

$$C_{0.79} = \begin{bmatrix} 0.4340 & 0 & 0.4609 & 0.2862 & 0.4018 \\ 0.4634 & 1 & 0.4609 & 0.2862 & 0.6501 \\ 0.3114 & 1 & 0.3008 & 0.1289 & 0.0775 \end{bmatrix}$$

$$C_{0.97} = \begin{bmatrix} 0.3373 & 0 & 0.4101 & 0.3775 & 0.4785 \\ 0.4634 & 1 & 0.4609 & 0.2862 & 0.6501 \end{bmatrix}$$

The C matrixes that mentioned below are the result of cluster centroid for each applied radius. The $C_{0.72}$ matrix points the matrix of the acquired cluster centroid with radius 0.72. The first row on $C_{0.72}$ matrix shows the first cluster centroid, the second row shows the second cluster centroid, the third row shows the third cluster centroid, and the fourth row shows the fourth cluster centroid.

Furthermore, the application of the Minkowski Chebyshev combination distance produces 4 clusters with r = 1.12, produces 3 clusters with r = 1.31, and produces 2 clusters with r = 1.6. The value of q is taken based on (Azizah et al., 2018) and the value of r is obtained from the simulation results. The achieved cluster centroid are described as follows:

$$C_{1.12} = \begin{bmatrix} 0 & 0 & 0.4609 & 0.1289 & 0.4018 \\ 0.5900 & 1 & 0.4609 & 0.1289 & 0.5900 \\ 0.5092 & 0 & 1 & 1 & 0.6196 \\ 0.7894 & 0 & 0.7571 & 0.2862 & 0.2439 \end{bmatrix}$$

$$C_{1.31} = \begin{bmatrix} 0 & 0 & 0.4609 & 0.1289 & 0.4018 \\ 0.5900 & 1 & 0.4609 & 0.1289 & 0.5900 \\ 0.5092 & 0 & 1 & 1 & 0.6196 \end{bmatrix}$$

$$C_{1.6} = \begin{bmatrix} 0 & 0 & 0.4609 & 0.1289 & 0.4018 \\ 0.7508 & 1 & 0.6501 & 0.4785 & 0.7132 \end{bmatrix}$$

The C matrixes that mentioned below are the output of cluster centroid for each applied radius. The $C_{1.12}$ matrix points the matrix of the achieved cluster centroid with radius 1.12. The first row, the second row, the third row, and the fourth row on $C_{1.12}$ matrix shows the first, second, third, and fourth cluster centroid respectively.

Afterwards, the achieved number of clusters are evaluated by applying Partition Coefficient value. Then, the result of the Partition value for FSC method with the application of Hamming distance and Minkowski Chebysev combination distance is compared. The result comparison is shown in figure 1.

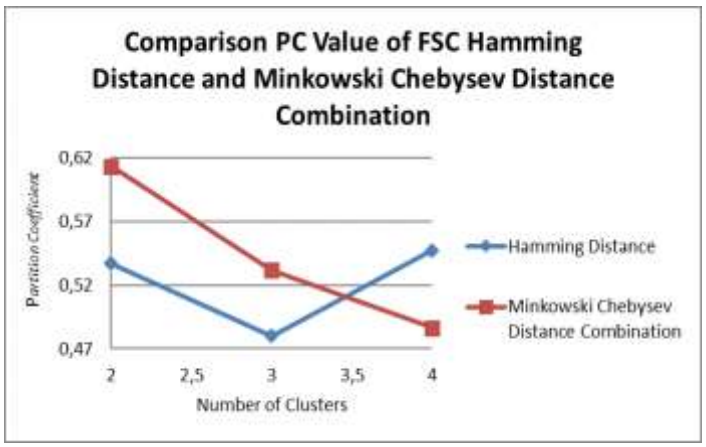

Figure 1 Partition Coefficient Value Comparson of Fuzzy Subtractive Clustering

Based on figure 1, the achieved Partition Coefficient value for Hamming distance application are lower than the Partition Coefficient value for Minkowski Chebysev combination distance application. The Partition Coefficient value from the Hamming distance for 2 achieved clusters is 0.5369, for 3 achieved clusters is 0.4801, and for 4 achieved clusters is 0.5473. Meanwhile, the Partition Coefficient value from Minkowski Chebysev combination distance for 2 achieved clusters, 3 achieved clusters, and 4 achieved clusters are 0.6135, 0.5319, and 0.4867 respectively. The greater the Partition Coefficient value (closer to 1), it means that the quality of the cluster obtained is getting better (Bezdek, 1981).

5. CONCLUSION

In this study, Fuzzy Subtractive Clustering using the Minkowski Chebysev combination distance produces a Partition Coefficient value that is better than the Partition Coefficient value obtained from the application of the Hamming distance. This conclusion is concluded because the Partition Coefficient value resulting from the application of the Minkowski Chebysev combination distance is generally higher than the Partition Coefficient value obtained from the application of the Hamming distance. This study simulates the value of the radius (r) to form the number of clusters. For further research, simulations can be carried out on other parameters such as q (squash factor).

REFERENCES

Abdolkarimi, E. S., & Mosavi, M. R. (2020). Wavelet-adaptive neural subtractive clustering fuzzy inference system to enhance low-cost and high-speed INS/GPS navigation system. *GPS Solutions*, *24*(2), 1–17. https://doi.org/10.1007/s10291-020-0951-y

Azizah, N., Yuniarti, D., & Goejantoro, R. (2018). Penerapan Metode Fuzzy Subtractive Clustering (Studi Kasus: Pengelompokkan Kecamatan di Provinsi Kalimantan Timur Berdasarkan Luas Daerah dan Jumlah Penduduk Tahun 2015). *Jurnal Eksponensial*, *9*(2), 197–206.

Banteng, L., Yang, H., Chen, Q., & Wang, Z. (2019). Research on the subtractive clustering algorithm for mobile ad hoc network based on the akaike information criterion. *International Journal of Distributed Sensor Networks*, *15*(9), 1–8. https://doi.org/10.1177/1550147719877612

Benmouiza, K., & Cheknane, A. (2019). Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting. *Theoretical and Applied Climatology*, *137*(1–2), 31–43. https://doi.org/10.1007/s00704-018-2576-4

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. In M. Nadler (Ed.), *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press. https://doi.org/10.1007/978-1-4757-0450-1

Chen, C., & Deng, X. (2020). Several new results based on the study of distance measures of intuitionistic fuzzy sets. *Iranian Journal of Fuzzy Systems*, *17*(2), 147–163. https://doi.org/10.22111/ijfs.2020.5225

Debnath, I. P., & Gupta, S. K. (2019). Exponential membership function and duality gaps for i-fuzzy linear programming problems. *Iranian Journal of Fuzzy Systems*, *16*(2), 147–163. https://doi.org/10.22111/ijfs.2019.4549

Dyvak, M., Maslyiak, Y., Voytyuk, I., & Maslyiak, B. (2018). Modified method of subtractive clustering for modeling of distribution of harmful vehicles emission concentrations. *CEUR Workshop Proceedings*, *2300*, 58–62.

Gan, G., Ma, C., & Wu, J. (2007). Data clustering: theory, algorithms, and applications. In *Data Clustering: Theory, Algorithms, and Applications* (Issue January 2007). https://doi.org/10.1137/1.9780898718348

Ghane'i Ostad, M., Vahdat Nejad, H., & Abdolrazzagh Nezhad, M. (2018). Detecting overlapping communities in LBSNs by fuzzy subtractive clustering. *Social Network Analysis and Mining*, *8*(1), 1–11. https://doi.org/10.1007/s13278-018-0502-5

Gubu, L., Rosadi, D., & Abdurakhman, A. (2021). Robust Portfolio Selection With Clustering Based on Business Sector of Stocks. *Media Statistika*, *14*(1), 33–43. https://doi.org/10.14710/medstat.14.1.33-43

Haqiqi, B. N., & Kurniawan, R. (2015). Analisis Perbandingan Metode Fuzzy C-Means Dan Subtractive Fuzzy C-Means. *Media Statistika*, *8*(2), 59–67. https://doi.org/10.14710/medstat.8.2.59-67

Hidayatur Rifa, I., & Pratiwi, H. (2020). Clustering Of Earthquake Risk in Indonesia Using K-Medoids and K-Means Algorithms. *Media Statistika*, *13*(2), 194–205. https://doi.org/10.14710/medstat.13.1.194-205

Jang, J. S. R., Sun, C. T., & Mizutani, E. (2005). Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence. In *IEEE Transactions on Automatic Control* (Vol. 42, Issue 10). 1482-1484. https://doi.org/10.1109/tac.1997.633847

Kusumadewi, S., & Purnomo, H. (2010). *Aplikasi logika fuzzy untuk pendukung keputusan*. Graha Ilmu.

Mahajan, S., & Gupta, S. K. (2019). On fully intuitionistic fuzzy multiobjective transportation problems using different membership functions. *Annals of Operations Research*, 1–31. https://doi.org/10.1007/s10479-019-03318-8

Rencher, A. C. (2016). *Methods of multivariate analysis* (Second edi, Vol. 4, Issue 1). John wiley & sons, inc.

Rezaei, K., & Rezaei, H. (2020). New distance and similarity measures for hesitant fuzzy sets and their application in hierarchical clustering. *Journal of Intelligent and Fuzzy Systems*, *39*(3), 4349–4360. https://doi.org/10.3233/JIFS-200364

Rodrigues, O. (2018). Combining minkowski and cheyshev: new distance proposal and survey of distance metrics using k-nearest neighbours classifier. *Pattern Recognition Letters*, *110*, 66–71. https://doi.org/10.1016/j.patrec.2018.03.021

Salah, H., Nemissi, M., Seridi, H., & Akdag, H. (2019). Subtractive Clustering and Particle Swarm Optimization Based Fuzzy Classifier. *International Journal of Fuzzy System Applications*, *8*(3), 108–122. https://doi.org/10.4018/ijfsa.2019070105

Sangadji, I., Arvio, Y., & Indrianto. (2018). Dynamic segmentation of behavior patterns based on quantity value movement using fuzzy subtractive clustering method. *Journal of Physics: Conference Series*, *974*(1), 0–7. https://doi.org/10.1088/1742-6596/974/1/012009

Sharma, R., & Verma, K. (2019). Fuzzy shared nearest neighbor clustering. *International Journal of Fuzzy Systems*, *21*(8), 2667–2678. https://doi.org/10.1007/s40815-019-00699-7

Surono, S., & Putri, R. D. A. (2020). Optimization of fuzzy c-means clustering algorithm with combination of minkowski and chebyshev distance using principal component analysis. *International Journal of Fuzzy Systems*, *23*(1), 139–144. https://doi.org/10.1007/s40815-020-00997-5

Utomo, V., & Marutho, D. (2018). Measuring hybrid sc-fcm clustering with cluster validity index. *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*, *C*, 322–326. https://doi.org/10.1109/ISRITI.2018.8864459

Zeng, S., Chen, S. M., & Teng, M. O. (2019). Fuzzy forecasting based on linear combinations of independent variables, subtractive clustering algorithm and artificial bee colony algorithm. *Information Sciences*, *484*, 350–366. https://doi.org/10.1016/j.ins.2019.01.071