e-Navi

Available online at http://www.e-navi.kr/
e-Navigation Journal

Original article

Fuzzy Monte Carlo clustering for analysing the distribution of fishing vessels in the coastal waters of South Korea

Gyei-Kark Park[a], Behrouz Fathi-Vajargah[b,*], Taeho Hong[c], GiJong Jo[d], YoungKi Kim[e], Kiyeol Seo[e]

[a]Division of Maritime Transportation, Mokpo National Maritime University, Korea, gkpark@mmu.ac.kr

[b,*]Deptartment of Statistics, University of Guilan, Iran, fathi@guilan.ac.ir, behrouz.fathi@gmail.com

[c] Division of Navigation & Information Systems, Mokpo National Maritime University, Korea, ds1pnp@mmu.ac.kr

[d] GMT Co. Ltd., Korea, jgj@gmtc.kr

[e]Korea Research Institute of Ships and Ocean Engineering, Korea, ykkim@kriso.re.k, kyseo@kriso.re.kr

Abstract

The issue of marine accidents can be based on the traffic/distribution of vessels in the waterways. These accidents are often associated with human and financial losses and require special attention. Usually, these accidents include collision of two fishing vessels with each other, collision of a fishing vessel with other types of vessels in the course and collision of a fishing vessel with an obstacle in the course (Yancai, et al, 2020).

In this article, we first want to deal with analysing the recorded statistical samples in 7 fishing areas in coastal waters of South Korea in 2023, while fuzzy clustering them. Then, according to analysing the sample data and finding the probabilistic structure and the membership of data sets the determined clusters, through Monte Carlo simulation, we will generate similar data in each of the 7 studied regions and model them in unsupervised mode. The generated data by Monte Carlo simulation based on the statistical distribution will able us to study the reality of distribution and possible accident in our target areas and find the model for future demands. We show that how the simulated data reduce the cost of data analysis and deliver us the facts of clusters for fishing vessels collisions. Finally, we reach to the most notified area for preventing the fishing vessels accidents and to make more preparations for reducing the human and costly damages in future activities.

Keywords: Fuzzy Clustering, Monte Carlo simulation, Unsupervised Data, Fishing Vessels, Collisions.

https://doi.org/10.52820/j.enavi.2024.22.026

1. Introduction

It is well-known that clustering is an unsupervised machine learning technique that divides the given data into different clusters based on their distances from each other. This difference also presents similarity/dissimilarity of data from a desired cluster. The unsupervised clustering can be considered by hard algorithms such as K-means that give the values of any point lying in some particular cluster to be either as 0 or 1 i.e., which assign each data point to a single cluster, only. In contrast, soft clustering algorithm gives the values of any point lying in different clusters in interval (0,1). It means that a particular data of a given data set may belong to two or more clusters of models, at the same time. Then, fuzzy clustering assigns a membership degree between 0 and 1 for each data point for each cluster.

It also provides us a different possibility to have uncertain clustering and can be also analysis it via probability and uncertain theories.

Large amounts of data are collected every day from many sources and cluster analysis is for them, such as maritime, satellite images, bio-medical, marketing, security, web searching, geo-spatial, cancer research, traffic flow, risk assessment and city planning. For example, in cancer research for classifying patients into subgroups according their gene expression profile. This can be useful for identifying the molecular profile of patients with good or bad prognostic, as well as for understanding the disease. In marketing for market segmentation by identifying subgroups of customers with similar profiles and who might be receptive to a particular form of advertising. In city-planning for identifying groups of houses according to their type, value and location (Kassambara, 2017).

In this paper, we are going to do the same job in fishing vessels distribution in coastal areas around South Korea. We focus on 7 busy area of fishing vessels and based on the all-recorded data of their distributions, we analysis each individual area. Then, we compare their results to reach more sensitive and notable area for preventing accidents and any waterway controls.

2. Monte Carlo Simulation

Monte Carlo methods estimate the solutions of a variety of mathematical, physical and engineering problems by performing a sample of experiences of a given population. In fact, using the Monte Carlo method with regard to realisation of statistical sampling, an approximation of a parameter will be obtained.

With regards to the consistency axiom for the Monte Carlo estimator, when we increase the number of sample size N, the absolute value of the error of the solution (estimation of parameter) decreases. Therefore, to obtain a good approximation of a parameter by the Monte Carlo method we should consider the size of the sample data to be big enough to realise an accurate estimation (Rubinstein,1981).

2.1 Random Variables Simulation

The random variable that its values make a (finite or infinite) countable set, is called discrete random variable. Some well-known discrete random variables are, discrete Uniform, Bernoulli, Binomial, Poisson, H

hyper-Geometric and Geometric random variables.

In contrast, the random variable that its values make an uncountable set is called continuous random variables. Some examples of famous continuous random variables are: continuous Uniform, Exponential, Chi-Square, Gamma, Beta, Normal, student's t, F fisher.

There are two common methods for simulating both discrete and continuous random variables:

- Inverse method
- Acceptance-Rejection method

Normal distribution is one of the most important continuous distributions, and we are able to simulate it by:

- Acceptance-Rejection method,
- Box-Muller method

They are also available in most statistical software such as SPSS, R, Minitab.

2.1.1 The Inverse Transform Method

Suppose we want to generate the value of a discrete random variable X having probability mass function

$$P\{X = x_i\} = p_i \quad ,i = 0, 1, \cdots n \quad \text{where} \quad \sum_{i=1}^{n} p_i = 1,$$

to accomplish this, we generate a random number U– that is, U is uniformly distributed over (0, 1)–and set

$$X = \begin{cases} x_0 & \text{If } U < p_0 \\ x_1 & \text{If } p_0 \le U < p_0 + p_1 \\ \vdots & \\ x_j & \text{If } \sum_{i=0}^{j-1} p_i \le U < \sum_{i=0}^{j} p_i \\ \vdots & \end{cases}$$

Since, for $0 < a < b < 1$, $p\{a \le U < b\} = b - a$, we have that

$$p\{X = x_j\} = p\left\{\sum_{i=0}^{j-1} p_i \le U < \sum_{i=0}^{j} p_i\right\} = p_j$$

so, X has our desired distribution.

Then, the preceding can be written algorithmically as

Generate a random number U

If U < p0 set X = x0 and stop,

If U < p0 + p1 set X = x1 and stop,

If U < p0 + p1 + p2 set X = x2 and stop,

As an example, if we wanted to simulate a random variable X such that p1 = 0.25, p2 = 0.10, p3 = 0.35, p4 = 0.30 where pi = P{X = i}, then we could generate U and do the following:

If U < 0.25 set X = 1 and stop.

If U < 0.35 set X = 2 and stop.

If U < 0.70 set X = 3 and stop.

Otherwise set X = 4.

One case where it is not necessary to search for the appropriate interval in which the random number lies is when the desired random variable is the discrete uniform random variable. That is, suppose we want to generate the value of X which is equally likely to take on any of the values 1,..., n. That is, P{X = j} = 1/n, j = 1,..., n. Using the preceding results, it follows that we can accomplish this by generating U and then setting

$$X = j \quad \text{if} \quad \frac{j-1}{n} \le U < \frac{j}{n}$$

Therefore, X will equal j if $j - 1 \le nU < j$; or, in other words,

X = int(nU) + 1,

where int (x) sometimes written as [x] is the integer part of x i.e., the largest integer less than or equal to x (Ross, 2006).

Suppose we have k clusters C1, …, Ck and n data set x1,…, xn $\in$ S. Using experimental data as a sample of

total data we clustered them and we obtain the degree of each data to our target clusters. Based on fuzzy clustering, we find xi $\in$ Cj where i=1,2,…,n and j=1,2,…,k.

In this step we want to simulate the relative data based on probabilistic structure, where we have now. The best method is here the discrete inverse method, as many as we would like to generate them (Kima et al, 2004).

4. Fuzzy Monte Carlo Algorithms

Algorithm1:

for *m* from *1* to *N*:

1. Set *k, n, sum*=0.
2. Generate *U* from uniform distribution on (0,1).
3. for *j* from *1* to *k*
   sum=sum + p_j
   for *i* from *1* to *n-1*
   if(*U*<sum) set x=i.
   print(" $i \in C_j$") and stop.
   Go to step 3.
4. Set *X=n*.
5. Print(" $i \in C_n$").
6. End

In the case that the given data set has the same degree function to each cluster, i.e. discrete uniform distribution we employ the algorithm2 which is simpler than the algorithm1.

Algorithm2:

1. Generate *U* from uniform distribution on (0,1).
2. for *i* from *1* to *k*
   set *X=int(nU)+1*
   print ("$X \in C_i$")
3. End

5. Fuzzy Clustering

It is very clear that fuzzy clustering has more natural situation and also has more chance to fit on our target data set. Fuzzy clustering and its developed algorithms have been used in many branches of our life and play a

main roll in their analysis and help us to make more efficiency in our future decisions (Farnam, et al 2021).

Some useful applications of fuzzy clustering are listed here (Bezdek, 1981 and Effati, et al. 2013):

1. Image segmentation: fuzzy clustering can be used to segment images by grouping pixels with similar properties together, such as color or texture.
2. Pattern recognition: fuzzy clustering can be used to identify patterns in large datasets by grouping similar data points together.
3. Marketing: fuzzy clustering can be used to segment customers based on their preferences and purchasing behavior, allowing for more targeted marketing campaigns.
4. Medical diagnosis: fuzzy clustering can be used to diagnose diseases by grouping patients with similar symptoms together.
5. Environmental monitoring: fuzzy clustering can be used to identify areas of environmental concern by grouping together areas with similar pollution levels or other environmental indicators.
6. Traffic flow analysis: fuzzy clustering can be used to analyze traffic flow patterns by grouping similar traffic patterns together, allowing for better traffic management and planning.
7. Risk assessment: fuzzy clustering can be used to identify and quantify risks in various fields, such as finance, insurance, and engineering.

5.1 FCM Algorithm for Data Clustering:

1. Initialize β (, c number of clusters and m=2, and data set S.

2. Select centers of clusters from data point, arbitrary.

3. At step r, calculate the centers vector $C^{(r)} = [c_j]$ with $U^{(r)} = [u_{ij}]$ where

$$c_j = \frac{\sum_{k=1}^{n} {u_{ij}}^2 x_j}{\sum_{k=1}^{n} {u_{ij}}^2} = \frac{{u_{1j}}^2 x_1 + \cdots + {u_{nj}}^2 x_n}{{u_{1j}}^2 + \cdots + {u_{nj}}^2}$$

4. Update $U^{(r)}, U^{(r+1)}$:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_j - c_j||}{||x_j - c_k||} \right)^2}$$

where i ∈ S (data set)and $j \in C$ (cluster set).

5. If $||U^{(r)} - U^{(r+1)}|| \leq \beta$ or if $||C^{(r)} - C^{(r+1)}|| \leq \beta$, then stop the algorithm.
Otherwise, go to step 3.

5.2 Advantages of Fuzzy Clustering

1. Flexibility: fuzzy clustering allows for overlapping clusters, which can be useful when the data has a complex structure or when there are ambiguous or overlapping class boundaries.
2. Robustness: fuzzy clustering can be more robust to outliers and noise in the data, as it allows for a more gradual transition from one cluster to another.
3. Interpretability: fuzzy clustering provides a more nuanced understanding of the structure of the data, as it allows for a more detailed representation of the relationships between data points and clusters.

6. Numerical Results for Fishing Vessels Distribution

Here, we are going to analysis and cluster 7 sections of fishing vessel distribution around in the coastal waters of South Korea in year 2023.

Figure1: South Korea 7 sectors of fishing area

1. Gyeonggiman : 38.0N, 124.0E 38.0N, 126.9E 36.6N, 126.9E 36.6N, 124.0E.

2. Cheonsuman-Anmagundo :36.6N, 124.0E 36.6N, 126.9E 35.3N, 126.9E 35.3N, 124.0E

3.Adjacent Seas of Mokpo: 35.3N, 124.0E 35.3N, 126.6E 34.0N, 126.6E 34.0N, 124.0E

4. Wondo-Tongyeong : 35.3N, 126.6E 35.3N,

128.5E 34.0N, 128.5E 34.0N, 126.6E

5.Adjacent Seas of Jeju Island: 34.0N, 124.0E 34.0N, 128.2E 32.0N, 128.2E 32.0N, 124.0E

6.Pohang-Tongyeong: 34.0N, 128.5E 36.3N, 128.5E 36.3N, 131.0E 34.0N, 131.0E

7.Goseong-Pohang : 36.3N, 28.5E 38.5N, 128.5E 38.5N, 131.0E 36.3N, 131.0E

We consider total 3449 data of 7 sector areas as follow:

Table 1: mass probability or degree function for clusters

Area	1	2	3	4	5	6	7
No. of data set	134	116	341	1214	279	868	497
No. of clusters	2	2	4	4	2	4	3
Prob. mass function	*U{1,2}*	*U{1,2}*	*U{1,2,3,4}*	*U{1,2,3,4}*	*U{1,2}*	*U{1,2,3,4}*	*U{1,2,3}*

Based on the fuzzy algorithm we have made the following R program(Venables, et al 2024) for clustering the given data:

```
fathi=read.table("beherouz11.txt", sep="\t", header=TRUE)
fathi_new<-fathi[,unlist(lapply(fathi, is.numeric))]
library(cluster)
df <- scale(fathi_new)
res.fanny <- fanny(df, 4)
head(res.fanny$membership, 10)
res.fanny$coeff
head(res.fanny$clustering, 20)
library(factoextra)
fviz_cluster(res.fanny, ellipse.type = "norm", repel = FALSE,
palette = "jco", ggtheme = theme_minimal(), legend = "right")
```

Tabel 2: number of simulated data

Area	1	2	3	4	5	6	7
No. of data set	134	116	341	1214	279	868	497
Simulated data set	402	348	1023	3642	837	2604	1988

In fact, we simulate for the area 1-7 of data sets, with $N_i=3n_i$ as follow 402, 348, 1023, 3642, 837, 2604 and 1988. We can simulate these data as we want to do (Fathi-Vajargah et al, 2013, 2016).

Tabel 3: number of simulated clusters

Area	1	2	3	4	5	6	7
No. of clusters	2	2	4	4	2	4	3
No. of Optimal simulated clusters	2	2	4	4	2	4	3

We can refer to the clustering graphs obtained for the real data and also the clustering obtained by simulated data (Fig. 2-8). The results show that there is not significant different between both clustering cases. It shows the performance of Monte Carlo simulation in clustering data, and can be very useful that whenever we cannot completely reach to the real data.

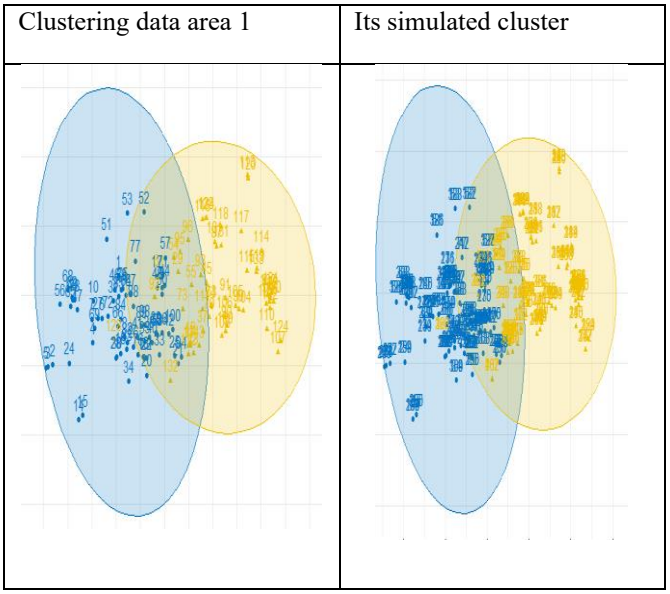

Figure 2: Cluster of data and simulated cluster in area 1

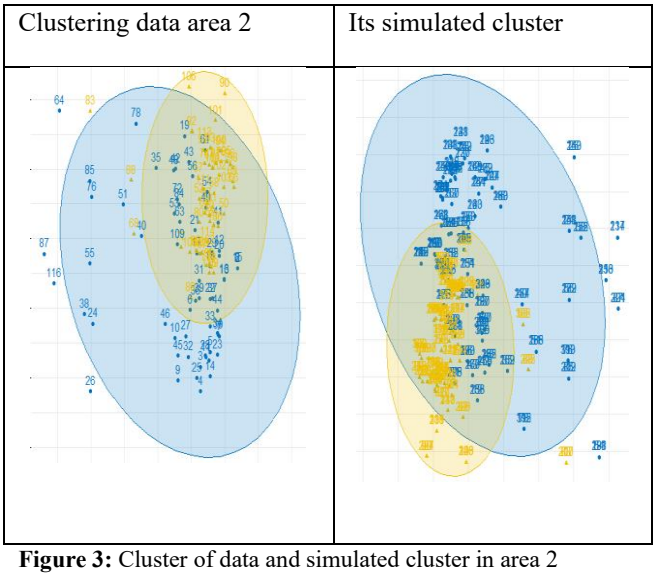

Figure 3: Cluster of data and simulated cluster in area 2

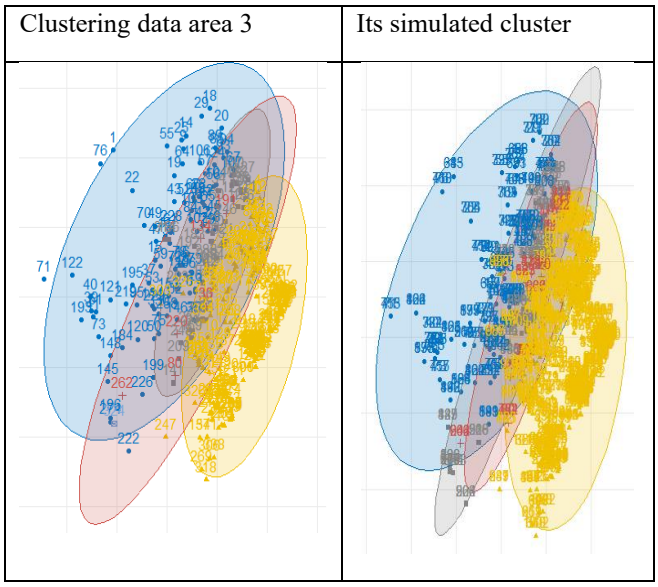

Figure 4: Cluster of data and simulated cluster in area 3

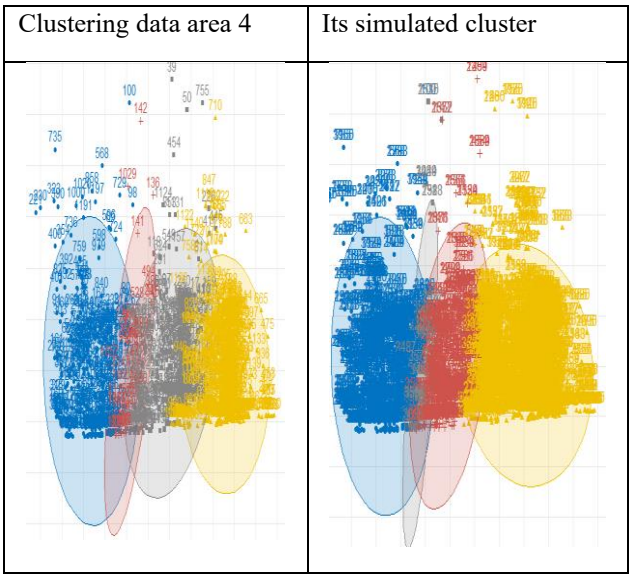

Figure 5: Cluster of data and simulated cluster in area 4

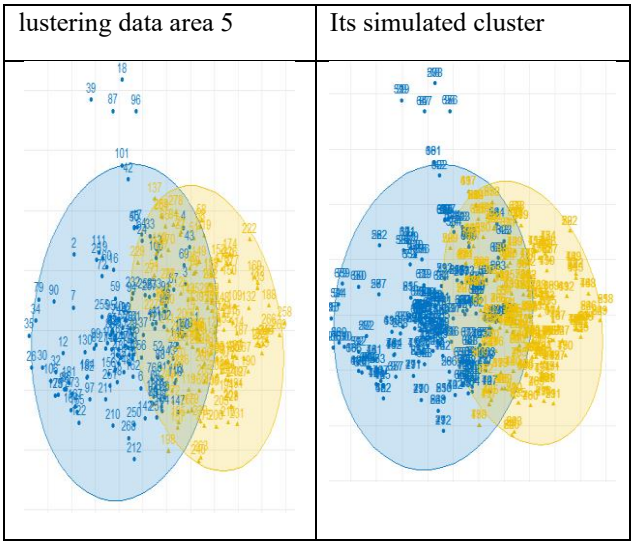

Figure 6: Cluster of data and simulated cluster in area 5

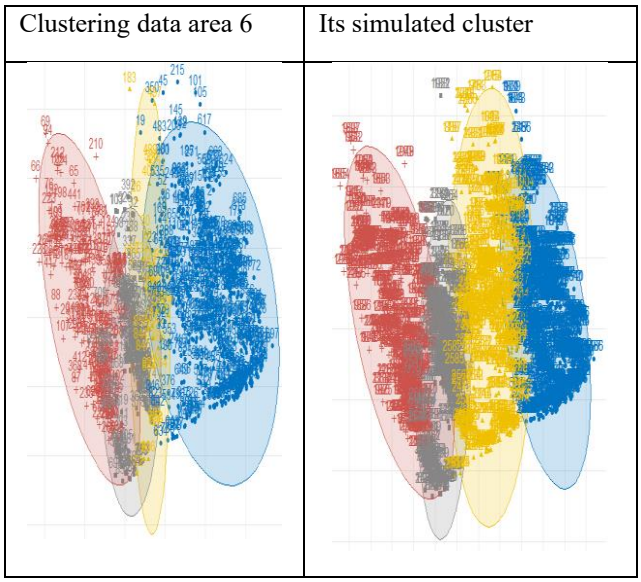

Figure 7: Cluster of data and simulated cluster in area 6

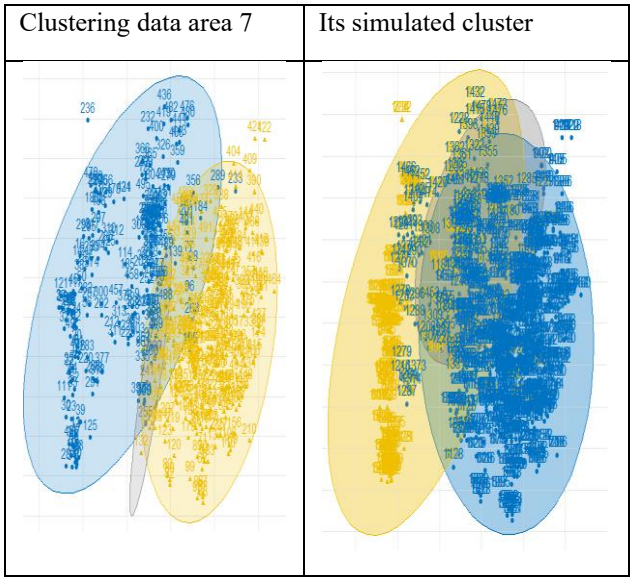

Figure 8: Cluster of data and simulated cluster in area 7

7. Converging Fuzzy Monte Carlo Clusters

We generally suppose that θ is a parameter of a h ypothesis population M. With the Monte Carlo meth od we consider independent sample paths from mea sured space (M,Ω, P) and from these samples we d etermine an approximate value of θ. The basis of t his discussion is consistent with the following defini tions:

Definition: If θ is a parameter of a hypothesis population and $T = T(X)$ is a function of random variable (r.v.) X such that, E[T]=θ with Var(T)=σ2 < ∞ then T is called Unbiased estimator for θ (with finite variance).

From this unbiased estimator T, we can make another

estimator as defined below:

If T is an unbiased estimator and $X_1, X_2, ..., X_n$ sampled from (M,Ω, p) independently and $\hat{\Theta}(X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$ such that, $E(\hat{\Theta}) = \theta$, $Var(\hat{\Theta}) = \frac{\sigma^2}{n} < \infty$, then $\hat{\Theta}$ is called secondary estimator (Monte Carlo estimator) for θ (Rohatgi, 1979).

The basic idea in this definition is the definition of T is such that it must be an unbiased estimator (or asymptotically convergent) with finite variance. Then the application of that as a secondary or Monte Carlo estimator is not difficult (Fathi-Vajargah et al, 2012).

The secondary estimator of parameter θ is convergent to θ with probability one as n tends to ∞ by Strong Law Large Numbers (SLLN). Then for a large sample size *n* we have $\hat{\Theta} \approx E(T) = \theta$.

Now, in clustering instead of general random variable X in the SLLN we use cluster C which is it has a mean μc i.e. E(C)=μc and Var(C)=σ2 <∞ (Variance here can be area of the cluster). When we simulate n cluster such as C1, C2, ..., Cn then the mean of this sample is converging with probability one to the μc.

We extend this idea to the following m-tiple clusters. Suppose we have C1=(C11, C12,…, C1m), C2=(C21, C22,…, C2m), …, Cn=(Cn1, Cn2,…, Cnm) and C=(C1, C2, ..., Cn) with E[C]=(E[C1], E[C2], ..., E(Cn)) i.e. μc=(μC1, μC2, … , μCn) then $\overline{C} = \frac{1}{n}\sum_1^n C_i$ $E[C] = \mu_C$ by SLLN. This is the main reason for converging the simulated clusters to the real data clusters. The SLLN convergent and other types of convergence of simulated clusters can be investigated based on uncertainty theory (Liu, 2015, Ghaffari-Hadigheh et al, to be appeared).

8. Conclusions

In this article we successfully developed the clustering processes via fuzzy Monte Carlo simulation. Comparing the obtained clusters using real data and simulated clusters (using Monte Carlo simulated data in clustering) shows that when we find the degree function of sample data sets to each cluster, then we can easily employ the Monte Carlo algorithm to simulate the all required data for analysing the target data and cluster them by fuzzy C means method.

Based on presented results in the coastal areas 1-7, for the distribution of fishing vessels, we eventually conclude that the areas 3,4 and 6 are the most notified area since their clusters are 4 (Fig. 4, 5, 7). It means that the distribution of fishing vessels in these are categorized in 4 types. In contrast, the areas 1,2,5 categorize in 2 clusters and the area 7 also with 3 clusters, only (Fig. 2, 3, 6, 8).

If we compare the whole results together, we conclude that controlling the traffic/accidents of fishing vessels in the area 1,2,5 and also 7 are simpler than the same control for the area 3,4 and 6. Finally, we reaccommodate to prevent the accidents of fishing vessels in 7 studied area, more preparation should be considered for the area of 3,4 and 6 since these area can be more critical areas in happening fishing vessels accident.

9. Acknowledgements

This research was supported by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (RS-2023-00262855).

References

Bezdek, J. C., (1981), Pattern Recognition with Fuzzy Objective Function Algorithms.

Effati, S., et al., (2013), Fuzzy clustering algorithm for fuzzy data based on α-cuts, Journal of Intelligent & Fuzzy Systems, 24(3), pp. 511- 519.

Farnam, M. and Dareh-Mirki, M., (2021), Fuzzy data clustering using an algorithm (FCM) based on parametric distance measurement, Journal of fuzzy system and applications, pp. 93-119.

Fathi-Vajargah, B., and Eskandari Chechaglou, (2013), A., "Optimal Halton Sequence via Inversive Scrambling." Communications in Statistics: Simulation and Computation 42.2, pp. 476-484.

Fathi-Vajargah, B., Javadzadeh-Moghtader, R., (2012), P-quasi Random Number Generators for Obtaining Stochastic Integrals, Advances in Information Technology and Management (AITM), 207 Vol. 2, No. 1, pp. 207-215.

Fathi-Vajargah, B., and Ghasemalipour, S., (2016), Random fuzzy numbers generation with cubic Hermit membership function and its application in simulation, Int. J. Computing Science and Mathematics, Vol. 7, No. 4, pp. 301-311.

Fathi-Vajargah, B., Ghasemalipour, S., (2016), Simulation of a random fuzzy queuing system with multiple servers, Journal of Contemporary Mathematical Analysis, Vol. 51, pp. 103-110.

Ghaffari-Hadigheh, A., Fathi-Vajargah, B., On the convergence of uncertain set processes (submitted).

Kassambara, A., (2017), Practical guide to cluster analysis in R, Unsupervised machine learning, Published by Sthda (http://www.sthda.com).

Kima, D-W, et al, (2004), On cluster validity index for estimation of the optimal number of fuzzy clusters, Pattern Recognition, Vol. 37, pp. 2009–2025.

Liu, B., (2015), Uncertainty theory (4nd ed.). Berlin: Springer.

Rohatgi V. K., (1976), An introduction to probability theory, John Wily and Sons, New York.

Ross, S., (2006), Simulation, Academic Press.

Rubinstein R.Y., (1981), Simulation and the Monte Carlo method, John Wiley &Sons, New York.

Yancai Hu, Gyei-Kark Park, (2020), Collision risk assessment based on the vulnerability of marine accidents using fuzzy logic, International Journal of Naval Architecture and Ocean Engineering, Vo1. 12, pp. 541-551.

Venables, W. N., et al, (2024), An Introduction to R, R Core Team.

Received **06 June 2024**

1st Revised **19 June 2024**

Accepted **20 June 2024**