How to cite this paper:

Ahmad-Azani, N. I., Yusoff, N., & Ku-Mahamud, K. R. (2018). Fuzzy discretization technique for bayesian flood disaster model. *Journal of Information and Communication Technology*, *17* (2), 167–189.

FUZZY DISCRETIZATION TECHNIQUE FOR BAYESIAN FLOOD DISASTER MODEL

[1]Nor Idayu Ahmad-Azami, [2]Nooraini Yusoff & [2]Ku Ruhana Ku-Mahamud

[1]Faculty of Information & Communication Technology
Limkokwing University of Creative Technology, Malaysia
[2,]School of Computing, Data Science Research Lab
Universiti Utara Malaysia, Kedah, Malaysia

idayu.azami@limkokwing.edu.my; nooraini@uum.edu.my; ruhana@uum.edu.my

ABSTRACT

The use of Bayesian Networks in the domain of disaster management has proven its efficiency in developing the disaster model and has been widely used to represent the logical relationships between variables. Prior to modelling the correlation between the flood factors, it was necessary to discretize the continuous data due to the weakness of the Bayesian Network to handle such variables. Therefore, this paper aimed to propose a data discretization technique and compare the existing discretization techniques to produce a spatial correlation model. In particular, the main contribution of this paper was to propose a fuzzy discretization method for the Bayesian-based flood model. The performance of the model is based on precision, recall, F-measure, and the receiver operating characteristic area. The experimental results demonstrated that the fuzzy discretization method provided the best measurements for the correlation model. Consequently, the proposed fuzzy discretization technique facilitated the data input for the flood model and was able to help the researchers in developing effective early warning systems in

Received: 3 May 2017 **Accepted**: 17 December 2017

the future. In addition, the results of correlation were prominent in disaster management to provide reference that may help the government, planners, and decision-makers to perform actions and mitigate flood events.

Keywords: flood disaster, spatial data mining, Bayesian Network, fuzzy discretization.

INTRODUCTION

Floods are one of the natural hazards that commonly occur in many areas around the world. According to the Center for Research on the Epidemiology of Disasters (CRED) (2011) that reported on global natural disasters, flood events increased in 2010 compared to the previous year. Analyzing correlation, particularly in disaster research, provides a fascinating insight of understanding the disaster events. Furthermore, the effects of each factor in various areas are significantly different. According to Peerbolte and Collins (2013), correlations are used to represent a relationship between two or more variables.

In recent achievements, the use of Bayesian Network (BN) methods in the domain of disaster management has proven its efficiency in developing susceptibility models and risk models. Various researchers (Li, Wang, Leung, & Jiang, 2010; Liang, Zhuang, Jiang, Pann, & Ren, 2012; Peng & Zhang, 2012a; 2012b; Viglione, Merz, Salinas, & Blöschl, 2013; Vogel et al., 2013) present studies to develop flood models using BN. Although BN has to be highlighted as a powerful method to find dependencies, the challenge begins when dealing with the continuous variables (Nielsen & Jensen, 2009; Uusitalo, 2007; Zwirglmaier, Papakosta, & Straub, 2013). Dougherty, Kohavi, and Sahami (1995), Friedman and Goldsmith (1996), Aguilera, Fernández, Fernández, Rumí, and Salmerón (2011), and Vogel (2014) suggested the use of discretization to overcome this problem.

Therefore, this study proposed the fuzzy discretization method to handle continuous data. Data discretization is a process of converting continuous variables into partition boundaries with selected cut points. In spatial data mining, discretization has become one of the preprocessing techniques used to transform a continuous variable into a discrete one (Bakar, Othman, & Shuib, 2009; García, Luengo, & Herrera, 2015).

Some reviews of the discretization technique can be found in the literature (e.g. Liu, Hussain, Tan, & Dash, 2002; Yang, Webb, & Wu, 2010). The second section discusses previous research related to data discretization. Next, the paper describes the proposed data discretization. The performances of different discretization methods on correlation models are then discussed. Concluding remarks are provided in the last section.

DISCRETIZATION OF CONTINUOUS FLOOD INDUCING FACTORS

The main goal of discretization is to transform continuous attributes into discrete attributes. In this section, discretization will be discussed as a preliminary condition for data preprocessing in order to be fed into the Bayesian Network model. The presentations are focused on the supervised discretization methods. Supervised discretization methods utilize the class information in setting partition boundaries whereas unsupervised discretization methods do not utilize instance labels for the selection of cut points. These methods have been presented to work reasonably well when used in spatial data. Unsupervised methods such as equal interval (EI), natural breaks (NB), quantile (QU) and standard deviation (SD) are among the most common discretization methods implemented in the field of geovisualization and spatial data mapping (Fischer & Wang, 2011; Stewart & Kennelly, 2010).

Supervised methods have been presented widely in the research fields of spatial data mining, risk studies and prediction. Berger (2004) performed the Minimum Description Length Principle (MDLP) to discretize continuous environmental data using a rough set rule for agricultural soils and assess crop suitability. Bai, Ge, Wang and Lan Liao (2010) also used MDLP to discretize continuous risk factors and mined underlying rules between neural tube defects (NTD). Lustgarten, Visweswaran, Gopalakrishnan and Cooper (2011) provided an efficient supervised Bayesian discretization method to give better results of classification from a high-dimensional biomedical dataset. Ge, Cao and Duan (2011) compared the impacts of three supervised discretization methods which were used on remote sensing classification. The authors presented supervised methods for spatial data discretization.

Jenks and Caspall (1971) proposed the natural breaks method to determine the values of cut points. The author presented the choropleth map classes

using the unsupervised method that improved inputs of the choropleth map information system. Moreover, Dawod, Mirza and Al-Ghamdi (2012) also used the natural breaks method to identify the break points of total flood volume values. Although the natural breaks can handle volumes of spatial data, this method required predefined numbers of intervals before the discretization process.

There are numerous studies including works by various researchers (Chang & Tsai, 2016; Güçlü & Şen, 2016; Lohani, Kumar, & Singh, 2012; Pulvirenti, Pierdicca, Chini, & Guerriero, 2011; Tsyganskaya et al., 2016) that exploit the concept of fuzzy logic in model development for disaster management analysis. It has been shown that fuzzy logic is extensively applied to analyze complex patterns with high accuracy.

In this study, the supervised method, which is the membership function (MF) graph in fuzzy logic, was used to discretize the continuous variables. Zadeh (2008) presented the fuzzy logic concept as a data preprocessing technique that provided more logical and scientific explanation to describe the attributes of the object. The fuzzy set intervals for each flood factor are represented as linguistic variables to a maximum of five intervals, which are very low, low, moderate, high and very high. Fuzzy logic is based on the theory of fuzzy sets that measure the ambiguity and believe that all things admit degrees (Kanagavalli & Raja, 2013; Negnevitsky, 2011). Hiwarkar and Iyer (2013) claimed that fuzzy logic presents the easier technique to clearly define the conclusion when it comes upon imprecise, vague, ambiguous, noisy or missing input information. In addition, Chin and Lim (2007) and Ku-Mahamud and Othman (2010) underlined the interpretations of linguistic variables that were claimed as a very natural and plausible way to obtain a better understanding to solve problems.

The major data acquisition for this study was focused on the environmental elements that can be classified into three categories: (a) time series data, which is the mean annual rainfall in 2010; (b) raster data, i.e. Interferometric Synthetic Aperture Radar (IfSAR); and (c) vector data, i.e. the data on the historical flooded area in 2010, the topographic map from the Department of Survey and Mapping Malaysia (JUPEM) and the soil map from the Minerals and Geoscience Department. Among the nine selected flood inducing factors, the attribute values of DEM, slope, SPI, TWI, river and rainfall need to be discretized and consequently fed into the BN model. Figure 1 shows the flowchart for the proposed data discretization technique based on fuzzy logic.

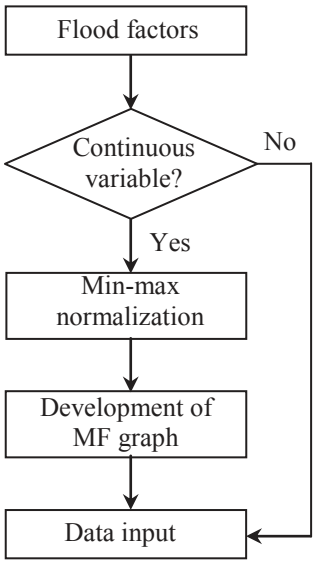

Figure 1. Flowchart for the proposed data discretization technique.

The proposed data discretization technique consists of two activities, which are the conversion of the actual data to Min-Max normalization, and the development of membership function (MF) to obtain fuzzy discretization. For the normalization process, the actual data is rescaled in the range of 0 to 1 using the max-min normalization procedure according to Al Shalabi, Shaaban, and Kasasbeh (2006) based on Eq. (1):

$$v' = ((v\text{-}mina) / (maxa - mina)) * (new\_maxa - new\_mina) + new\_mina \quad (1)$$

This operation was carried out in order to standardize the different intervals of each continuous variable for the attribute values of rainfall, DEM, slope, SPI, TWI and river. The construction of the MF graph was undertaken in the second step. The transformed data was then used to calculate the Entropy to develop the MF graph. The following equations were used in the Entropy estimation that was expressed by Christensen (1981) as follows:

$$S(x) = p(x) * S_p(x) + q(x) * S_q(x) \quad (2)$$

where from Eq. (2),

$$S_p(x) = -[p_1(x) * ln\ p_1(x) + p_2(x) * ln\ p_2(x)] \quad (3)$$

$$S_q(x) = -[q_1(x) * ln\ q_1(x) + q_2(x) * ln\ q_2(x)] \quad (4)$$

where, $p_k(x)$ and $q_k(x)$ conditional probabilities that class k samples are in the regions $[x_1, x_1 + x]$ and $[x_1 + x, x_2]$, respectively,

$p(x)$ and $q(x)$ probabilities that all samples are in the regions $[x_1, x_1 + x]$ and $[x_1 + x, x_2]$, respectively, and

$$p(x) + q(x) = 1 \quad (5)$$

Based on Eqs. (3) and (4), this is the formula to estimate the Entropy of $p_k(x)$, $q_k(x)$, $p(x)$ and $q(x)$:

$$p_k(x) = \frac{n_k(x)+1}{n(x)+1} \quad (6)$$

$$q_k(x) = \frac{N_k(x)+1}{N(x)+1} \quad (7)$$

$$p(x) = \frac{n(x)}{n} \quad (8)$$

$$q(x) = 1 - p(x) \quad (9)$$

where, $n_k(x)$ = the number of class k samples located in $[x_1, x_1 + x]$,
$n(x)$ = the total number of samples located in $[x_1, x_1 + x]$,
$N_k(x)$ = the number of class k samples located in $[x_1 + x, x_2]$, and
n = the total number of samples in $[x_1, x_2]$.

Table 1 shows the segmentation of x into two arbitrary classes. x is the transformed data that has been chosen randomly and has been classified into two classes. The value was divided into fuzzy partitions.

Table 1

The Segmentation of x into Two Arbitrary Classes

x	0	0.14	0.16	0.24	0.28	0.32	0.38	0.45	0.52	0.60	0.73	1.00
Class	1	1	1	1	1	1	1	1	2	2	2	2

Eqs. (2), (3), (4), (6), (7), (8), and (9) are used to compute p_1, p_2, q_1, q_2, $p(x)$, $q(x)$, $sp(x)$, $sq(x)$, and S, and the results are shown in Table 2. The value of x that gives the minimum value of the Entropy (S) is selected as the first threshold value partition point called the primary threshold (PRI) value.

Table 2

Calculations for Selection of Partition Point Primary Threshold (PRI) Value

x	0.13	0.15	0.22	0.26	0.30	0.35	0.42	0.51	0.54	0.64	0.85
p_1	1.00	1.00	1.00	1.00	1.00	0.86	0.88	0.89	0.80	0.73	0.67
p_2	0.50	0.33	0.25	0.20	0.17	0.29	0.25	0.22	0.30	0.36	0.42
q_1	0.58	0.55	0.50	0.44	0.38	0.43	0.33	0.20	0.25	0.33	0.50
q_2	0.50	0.55	0.60	0.67	0.75	0.71	0.83	1.00	1.00	1.00	1.00
$p(x)$	0.08	0.17	0.25	0.33	0.42	0.50	0.58	0.67	0.75	0.83	0.92
$q(x)$	0.92	0.83	0.75	0.67	0.58	0.50	0.42	0.33	0.25	0.17	0.08
$S_p(x)$	0.35	0.37	0.35	0.32	0.30	0.49	0.46	0.44	0.54	0.60	0.64
$S_q(x)$	0.66	0.66	0.65	0.63	0.58	0.60	0.52	0.32	0.35	0.37	0.35
Sx	0.64	0.61	0.58	0.53	0.47	0.55	0.49	0.40	0.49	0.56	0.61

The same process as displayed in Table 2 is repeated for the negative and positive partitions for different values of x. Table 3 shows the calculations to determine the secondary threshold value known as TER1 for the negative side. The value of x that has been selected is lower than the PRI value.

Table 3

Calculations to Determine the Secondary Threshold Value (TER1): NG Side

x	0.14	0.20	0.24	0.28	0.32	0.38	0.45
p_1	1.00	1.00	1.00	1.00	1.00	0.86	0.88

(continued)

p_2	0.50	0.33	0.25	0.20	0.17	0.29	0.25
q_1	0.88	0.86	0.83	0.80	0.75	1.00	1.00
q_2	0.25	0.27	0.33	0.40	0.50	0.33	0.50
$p(x)$	0.13	0.25	0.38	0.50	0.63	0.75	0.88
$q(x)$	0.88	0.75	0.63	0.50	0.38	0.25	0.13
$S_p(x)$	0.38	0.37	0.35	0.32	0.30	0.50	0.46
$S_q(x)$	0.46	0.49	0.52	0.55	0.56	0.37	0.35
Sx	0.45	0.46	0.46	0.44	0.40	0.46	0.45

Table 4 shows the calculations to determine the secondary threshold value known as TER2 for the positive side. The value of x that has been selected is higher than the PRI value.

Table 4

Calculations to Determine the Secondary Threshold Value (TER2): PO Side

x	0.61	0.80	0.95
p_1	0.50	0.33	0.25
p_2	1.00	1.00	1.00
q_1	0.25	0.33	0.50
q_2	1.00	1.00	1.00
$p(x)$	0.25	0.50	0.75
$q(x)$	0.75	0.50	0.25
$S_p(x)$	0.35	0.37	0.35
$S_q(x)$	0.347	0.366	0.347
Sx	0.347	0.366	0.347

Table 5 shows the calculations to determine the tertiary threshold value known as TER3 for the negative side. The value of x that has been selected is lower than the PRI value.

Table 5

Calculations to Determine the Tertiary Threshold Value (TER3): NG Side

x	0.18	0.22	0.26	0.30
p_1	1.00	1.00	1.00	1.00
p_2	0.50	0.33	0.25	0.20
q_1	1.00	1.00	1.00	1.00
q_2	0.20	0.25	0.33	0.50
$p(x)$	0.20	0.40	0.60	0.80
$q(x)$	0.80	0.60	0.40	0.20
$S_p(x)$	0.35	0.37	0.35	0.32
$S_q(x)$	0.32	0.35	0.37	0.35
Sx	0.33	0.36	0.35	0.33

Table 6 shows the calculations to determine the tertiary threshold value known as TER4 for the positive side. The value of x is higher than the PRI value.

Table 6

Calculations to Determine the Tertiary Threshold Value (TER4): PO Side

x	0.69	0.85
p_1	0.50	0.33
p_2	1.00	1.00
q_1	0.33	0.50
q_2	1.00	1.00
$p(x)$	0.33	0.67
$q(x)$	0.67	0.33
$S_p(x)$	0.35	0.37
$S_q(x)$	0.37	0.35
Sx	0.36	0.36

As explained by Marcot, Steventon, Sutherland and McCann (2006), the maximum number of intervals or discretization should be limited in five states

to improve the precision and the network structure. In the present study, the fuzzy set intervals for each flood factor were represented as linguistic variables as follows:

Table 7

Linguistic Variable

Linguistic Variable	Interval
Very low	1
Low	2
Moderate	3
High	4
Very high	5

DIGITAL ELEVATION MODEL

Digital elevation models (DEM) are the major source to derive topographic factors that have a direct effect on runoff velocity and flow size. DEM was created using the IfSAR data with a resolution of 10m. x 10m. IfSAR is an active remote sensing technology that is able to easily collect data from huge areas. The resultant dataset is the base of the digital surface and elevation models. Since the surface conditions are the leading factors that determine the formation of flood events, the use of high-resolution synthetic data is the perfect source to derive the topographic factors of elevation, which are DEM, slope angle, curvature, SPI, TWI and distance from the river. Figure 2 shows the original data and reclassified data using fuzzy discretization.

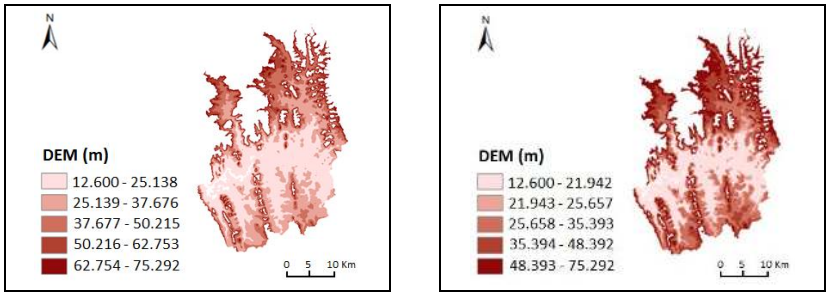

Figure 2. DEM maps: (a) the original DEM and (b) reclassified DEM.

Slope

Another important aspect to consider is the slope in the study area. Slope is the basic index extracted from DEM to describe the terrain. Heavy rainfall will cause slope failure during flood events. This situation might give a great impact on the breeding of disasters as the sliding surface for the runoff process. The slope gradient in degrees are shown in Figure 3.

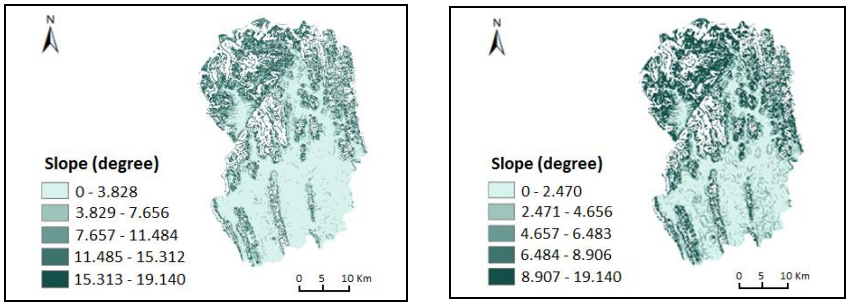

Figure 3. Slope maps: (a) the original slope and (b) reclassified slope.

Stream Power Index

Stream power index is the rate that the energy of flowing water expands on the bed and banks of a channel. High stream power values generally correspond with steep, straight, scoured reaches and bedrock gorges. Low stream power values occur in broad alluvial flats, floodplains and slowly subsiding areas, where the valley fill is usually intact and deep. The given equations are calculated and the generated SPI is as shown in Figure 4.

SPI=Ln((“facc_dem” + 0.001) * ((“slope_dem” / 100) + 0.001))

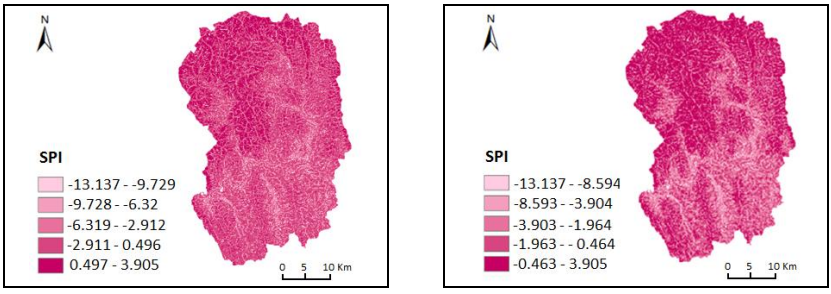

Figure 4. SPI maps: (a) the original SPI and (b) reclassified SPI.

Topographic Wetness Index

Topographic Wetness Index (TWI) is a steady-state wetness index. The value for each cell in the output raster (the TWI raster) is the value in a flow accumulation raster for the corresponding DEM. Higher TWI values represent drainage depressions; lower values represent crests and ridges. Figure 5 shows the original and reclassified TWI. In creating the TWI, the following equation is calculated to produce the TWI:

TWI =Ln((“facc_dem” + 0.001) / ((“slope_dem” / 100) + 0.001))

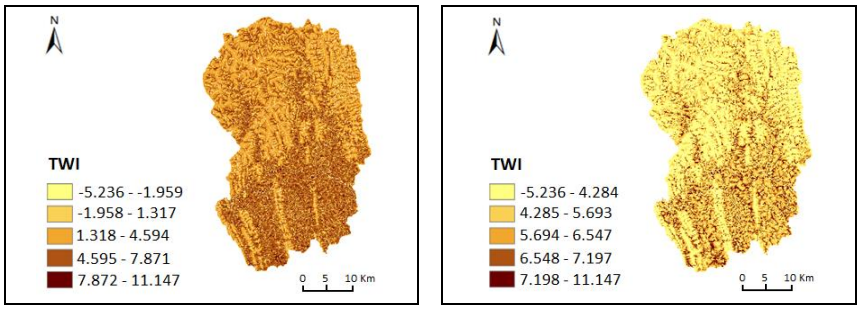

Figure 5. TWI maps: (a) the original TWI and (b) reclassified TWI.

River

Distance from river is a factor that calculates the approximate point between the consecutive points along rivers (polygon). At first, the main river in the study area was extracted using the IfSAR data. Next, the Euclidean Distance tool was used to create a raster of the distance from river. Figure 6 shows the original and reclassified distance from river.

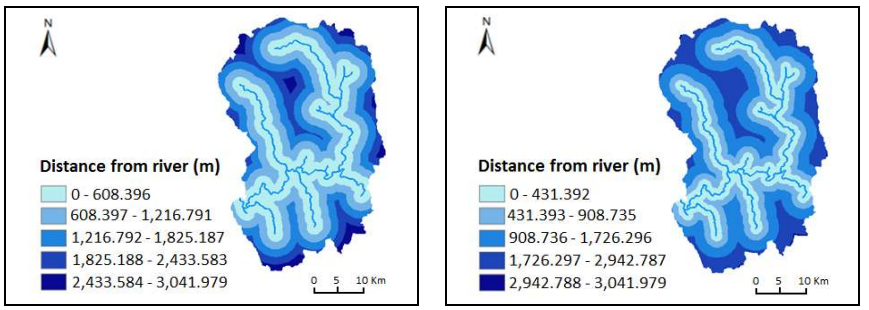

Figure 6. River maps: (a) original river and (b) reclassified river.

Rainfall

The historical data that included 18 rainfall stations with mean annual rainfall were obtained. In producing the mean annual rainfall intensity, the historical data were considered as the primary source of information. The available rainfall data was recorded at permanent but very dispersed rain gauges. Therefore, this study used the Inverse Distance Weighted (IDW) method to reproduce the spatial distribution of rainfall data for the entire study areas. The spatial distribution of rainfall data is illustrated in Figure 7.

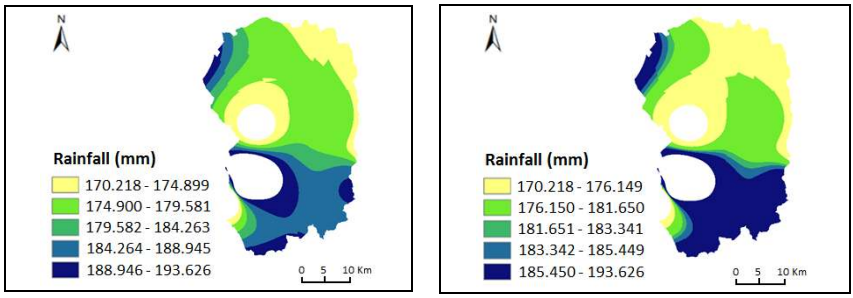

Figure 7. Rainfall maps: (a) the original rainfall and (b) reclassified rainfall.

RESULT AND DISCUSSION

Figure 8 illustrates the MF graph for rainfall data after the calculation of entropy was complete. The development of the MF graph was to standardize the differences in rainfall data for each *mukim* that can be measured by using one graph. x0, x1, x2, x3 and x4 were the threshold values estimated using the Entropy method, which were 0.180, 0.320, 0.510, 0.610 and 0.690, respectively. Very low, low, moderate, high and very high are the standard stages for all levels. The y axis is the value of MF in the range of zero to one, while the x axis is the transformed value from the range of 0 to 1.

By using this graph, the converted data was transformed into new representations for interval boundaries. The fuzzy set interval was then defined as shown in Table 8. The rainfall data was been normalized in the range of 0 to 1 and then transformed into new representations of fuzzy discretization by using the MF graph.

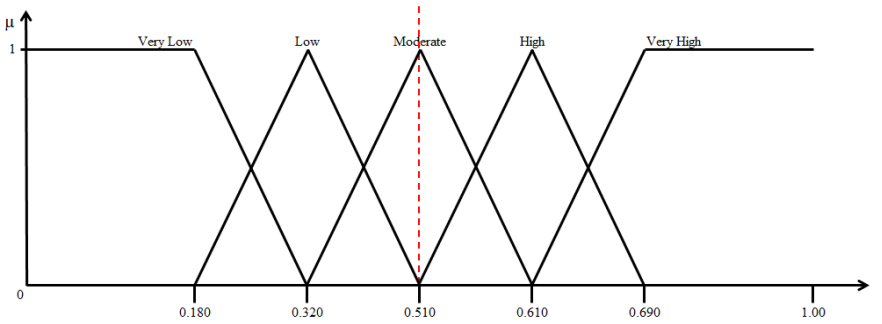

Figure 8. The membership function graph for rainfall data

Table 8

Sample of Transformed Rainfall Data

Rainfall data	Normalized classes	Linguistic variable	Fuzzy discretization
172.10	0.08	Very low	1
173.33	0.13	Very low	1
175.88	0.24	Very low	1
176.15	0.25	Low	2
178.54	0.35	Low	2
181.65	0.49	Moderate	3
183.19	0.55	Moderate	3
184.12	0.59	High	4
185.42	0.65	High	4
187.89	0.76	Very high	5

This new representation was used to enhance the correlation model of BN in the data discretization phase. This was applied to all data with continuous variables. The second MF graph was developed for DEM. The estimated threshold values for the DEM level were 0.120, 0.164, 0.240, 0.482 and 0.560, respectively. Figure 9 illustrates the MF graph for DEM.

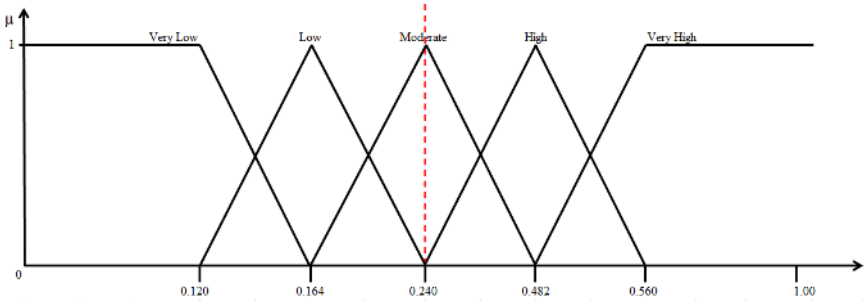

Figure 9. The membership function graph for DEM.

The third MF graph was developed for the slope data. The estimated threshold values for the data level were 0.040, 0.210, 0.272, 0.404 and 0.540, respectively. Figure 10 illustrates the MF graph for the slope data.

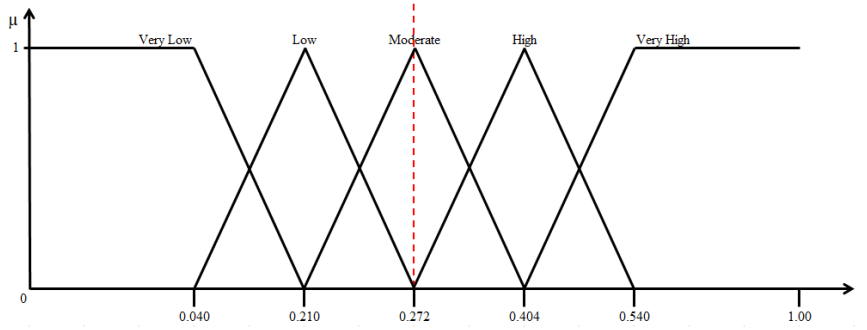

Figure 10. The membership function graph for slope data.

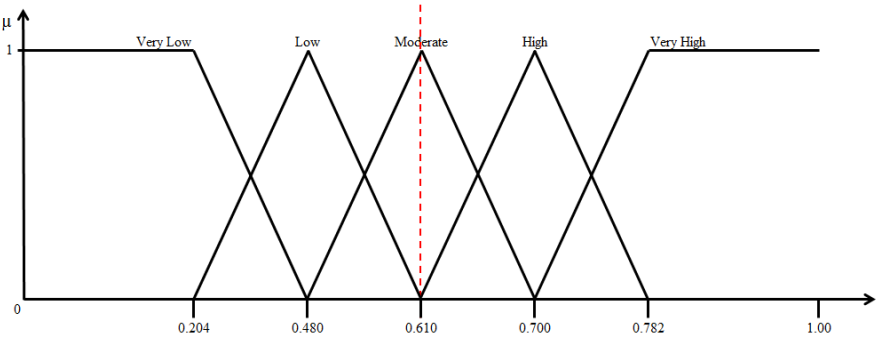

Figure 11. The membership function graph for SPI.

The fourth MF graph was developed for the SPI data. The estimated threshold values for the data level were 0.204, 0.480, 0.610, 0.700 and 0.782, respectively.

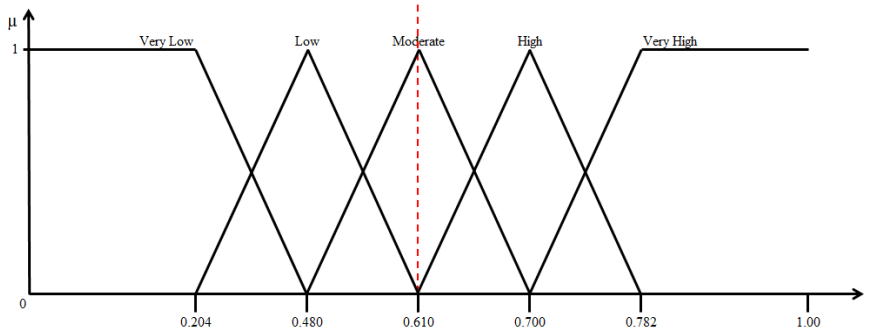

Figure 11. Illustrates the MF graph for SPI.

The fifth MF graph was developed for the TWI data. The estimated threshold values for the data level were 0.531, 0.633, 0.703, 0.737 and 0.779, respectively. Figure 12 illustrates the MF graph for TWI.

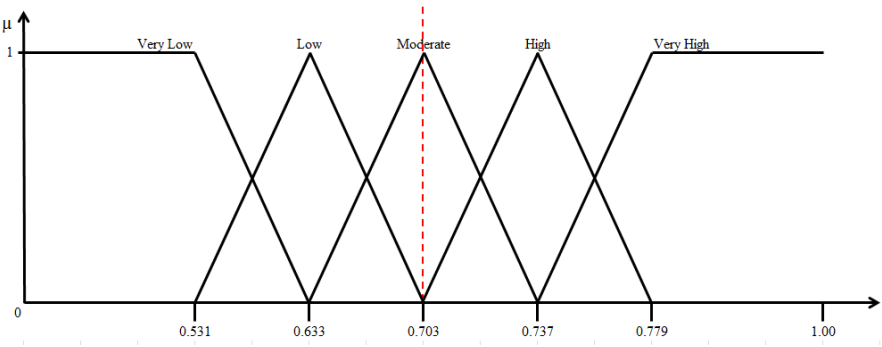

Figure 12. The membership function graph for TWI.

The last MF graph was developed for the river data. The estimated threshold values for the data level were 0.114, 0.170, 0.430, 0.630 and 0.970, respectively. Figure 13 illustrates the MF graph for the river data.

The structure of the Bayesian Network correlation model for the area of interest is displayed in Figure 14. This structure presents the complexity of the relationship between the occurrences of flood events that link with the flood inducing factors. The flood is the parent node of the BN structure, and each flood inducing factor is directly linked to it.

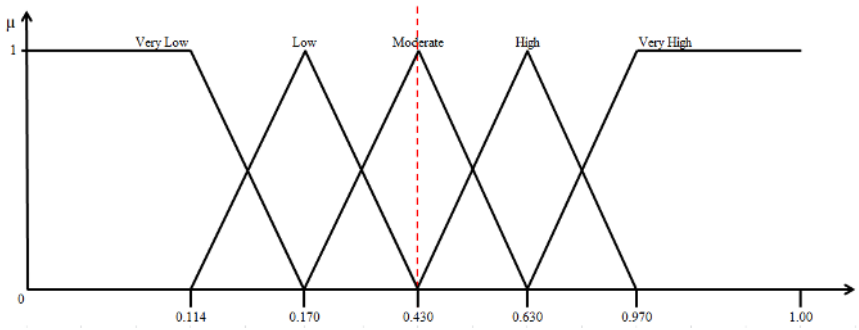

Figure 13. The membership function graph for river data.

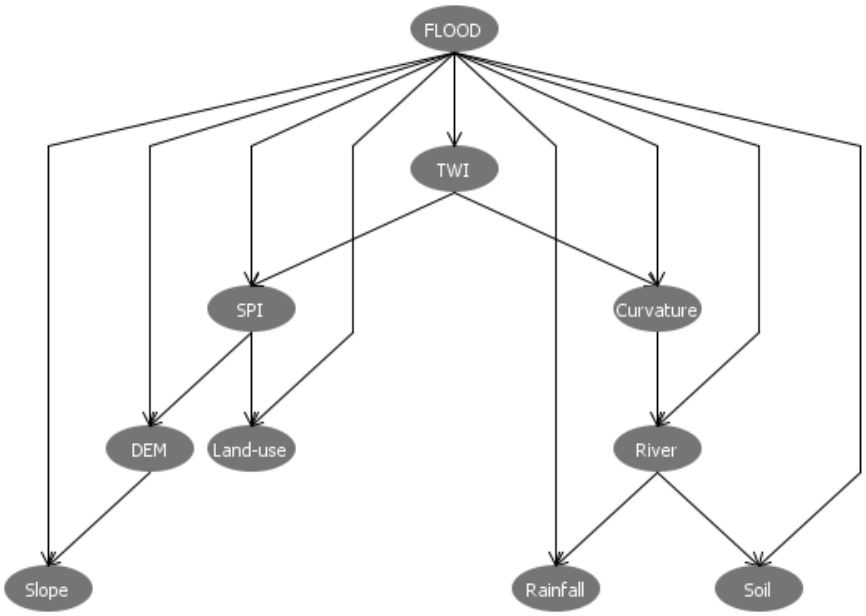

Figure 14. Structure of Bayesian Network correlation model.

This result reveals that among the nine flood inducing factors, TWI is the most significant factor that contributes to flood. TWI is one of the most influential topography factors that is used to depict the effect of flow accumulation in an area (Dehotin et al., 2015; Kafira, Albanakis, & Oikonomidis, 2015). Information of the correlation model listed in Table 9, shows high probability

(29%) of flood occuring when TWI is low (discretize value is 2). Lower TWI values represent ridges and crests. This means the velocity of runoff water will increase and cause a greater capacity of water to be distributed to all areas with lower elevation. The relationship between TWI and other flood influencing factors indicate that TWI is strongly associated followed by curvature, SPI, river, land use, DEM, soil types, rainfall and slope.

Table 9

Conditional Probability of the Node TWI

The value of the parent node	TWI				
(Flood)	Very low	Low	Moderate	High	Very high
Absence of flood	0.93	0.07	0.01	0.00	0.00
Presence of flood	0.18	0.29	0.25	0.14	0.14

Based on the experiments, it was found that the proposed fuzzy discretization method shows better performance. This indicates that incorporating the proposed fuzzy discretization with the BN model gives better results. The results from of performance metrics show that this method performed well compared to other discretization methods. Five data discretization techniques for modelling the BN were compared, namely Fuzzy Discretization, Equal Width, Natural Breaks, Quantile, and Geometrical Interval. The results are summarized in Table 10. The performance of the models is based on precision, recall, F-measure, and receiver operating characteristic (ROC).

Table 10

Comparison of Average Performance Assessment of BN Models

Technique	Precision	Recall	F-Measure	ROC Area	Class
Fuzzy discretization	0.99	0.97	0.98	0.99	Flood
	0.89	0.98	0.93	0.99	No flood
Equal width	0.82	0.61	0.68	0.81	Flood
	0.81	0.63	0.58	0.81	No flood
Natural breaks	0.83	0.60	0.67	0.80	Flood
	0.53	0.64	0.58	0.80	No flood
Quantile	0.84	0.58	0.66	0.81	Flood
	0.53	0.64	0.58	0.81	No flood
Geometrical interval	0.82	0.61	0.68	0.81	Flood
	0.54	0.63	0.58	0.81	No flood

The performance assessment of BN strongly depends on the choice of the different interval between the compared methods. Good results were obtained from the fuzzy discretization with the precision of 0.99, recall of 0.97, F-measure of 0.98, and receiver operating characteristic of 0.99 for the correlation model.

CONCLUSION

Bayesian Network has been widely used to represent the logical relationships between variables. However, many of the flood factors consist of continuous variables that introduce challenges for the data-mining task. Hence, the proposed data discretization method contributes in the process to re-encode the continuous variables into discrete variables. Nevertheless, if too many intervals are unsuited to the learning process, this will lead to a loss of information; and if there are too few intervals, this can lead to the risk of losing some interesting information. In brief, incorporating the proposed fuzzy discretization with the BN model for the flood event provides better results. Thus, the effects of the proposed fuzzy discretization method with continuous data lead to data reduction and the simplification of data. Subsequently, this process will make learning faster and produce shorter and compact results.

ACKNOWLEDGMENT

The authors wish to thank the Malaysia Ministry of Higher Education Malaysia for funding this study under the Long Term Research Grant Scheme (LRGS/ b-u/2012/UUM/ Teknologi Komunikasi dan Informasi).

REFERENCES

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software*, *26*(12), 1376–1388.

Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*(9), 735–739.

Bai, H., Ge, Y., Wang, J. F., & Lan Liao, Y. (2010). Using rough set theory to identify villages affected by birth defects: The example of Heshun, Shanxi, China. *International Journal of Geographical Information Science*, *24*(4), 559–576.

Bakar, A. A., Othman, Z. A., & Shuib, N. L. M. (2009, October). Building a new taxonomy for data discretization techniques. In 2009 2nd *Conference on Data Mining and Optimization* (pp. 132–140). *IEEE*.

Berger, P. A. (2004). Rough set rule induction for suitability assessment. *Environmental Management*, *34*(4), 546–558.

Chang, F. J., & Tsai, M. J. (2016). A nonlinear spatio-temporal lumping of radar rainfall for modeling multi-step-ahead inflow forecasts by data-driven techniques. *Journal of Hydrology*, *535*, 256–269.

Chin, W. C., & Lim, A. H. L. (2007). Web cluster load balancing via genetic-fuzzy based algorithm. *Journal of Information and Communication Technology*, *6*, 73-86.

Christensen, R. (1981). *Entropy minimax sourcebook: General description* (Vol. 1). Entropy Limited, Lincoln, MA.

CRED (Centre for Research on the Epidemiology of Disasters). (2011). *Natural disaster trends*. Retrieved from http://www.emdat.be.

Dawod, G. M., Mirza, M. N., & Al-Ghamdi, K. A. (2012). GIS-based estimation of flood hazard impacts on road network in Makkah city, Saudi Arabia. *Environmental Earth Sciences*, *67*(8), 2205–2215.

Dehotin, J., Breil, P., Braud, I., de Lavenne, A., Lagouy, M., & Sarrazin, B. (2015). Detecting surface runoff location in a small catchment using distributed and simple observation method. *Journal of Hydrology*, *525*, 113–129.

Dougherty, J., Kohavi, R., & Sahami, M. (1995, July). Supervised and unsupervised discretization of continuous features. In *Machine learning: Proceedings of the Twelfth International Conference* (Vol. 12, pp. 194–202).

Fischer, M. M., & Wang, J. (2011). *Spatial data analysis: Models, methods and techniques*. Springer Science and Business Media.

Friedman, N., & Goldszmidt, M. (1996, July). Discretizing continuous attributes while learning Bayesian networks. *ICML* (pp. 157–165).

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. New York: Springer.

Ge, Y., Cao, F., & Duan, R. F. (2011). Impact of discretization methods on the rough set-based classification of remotely sensed images. *International Journal of Digital Earth*, *4*(4), 330-346.

Güçlü, Y. S., & Şen, Z. (2016). Hydrograph estimation with fuzzy chain model. *Journal of Hydrology*, *538*, 587–597.

Hiwarkar, T. A. & Iyer, R. S. (2013). New applications of soft computing, artificial intelligence, fuzzy logic and genetic algorithm in bioinformatics. *International Journal of Computer Science and Mobile Computing*. Vol. 2, Issue 5, 202–207.

Jenks, G. F., & Caspall, F. C. (1971). Error on choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers*, *61*(2), 217–244.

Kafira, V., Albanakis, K., & Oikonomidis, D. (2015). *Flood risk assessment using remote sensing and geographical information systems* (GIS). An example from Kassandra Peninsula, Chalkidiki, Greece, 287–308.

Kanagavalli, V., & Raja, K. (2013). A fuzzy logic based method for efficient retrieval of vague and uncertain spatial expressions in text exploiting the granulation of the spatial event queries. *International Journal of Computer Applications* (0975-8887). National Conference on Future Computing CoRR.

Ku-Mahamud, K. R., & Othman, M. (2010). Fuzzy subjective evaluation of Asia Pacific airport services. *Journal of Information and Communication Technology*, *9*, 41-57.

Li, L., Wang, J., Leung, H., & Jiang, C. (2010). Assessment of catastrophic risk using Bayesian network constructed from domain knowledge and spatial data. *Risk Analysis*, *30*(7), 1157–1175.

Liang, W. J., Zhuang, D. F., Jiang, D., Pan, J. J., & Ren, H. Y. (2012). Assessment of debris flow hazards using a Bayesian network. *Geomorphology*, *171*, 94–100.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data mining and knowledge discovery*, *6*(4), 393–423.

Lohani, A. K., Kumar, R., & Singh, R. D. (2012). Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. *Journal of Hydrology*, *442*, 23–35.

Lustgarten, J. L., Visweswaran, S., Gopalakrishnan, V., & Cooper, G. F. (2011). Application of an efficient Bayesian discretization method to biomedical data. *BMC Bioinformatics*, *12*(1), 309.

Marcot, B. G., Steventon, J. D., Sutherland, G. D., & McCann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, *36*(12), 3063–3074.

Negnevitsky, M. (2011). *Artificial intelligence a guide to intelligent systems* (3rd ed.). England: Pearson Education.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. New York: Springer Science & Business Media.

Peerbolte, S. L., & Collins, M. L. (2013). Disaster management and the critical thinking skills of local emergency managers: Correlations with age, gender, education, and years in occupation. *Disasters*, *37*(1), 48–60.

Peng, M., & Zhang, L. M. (2012a). Analysis of human risks due to dam-break floods—part 1: A new model based on Bayesian networks. *Natural Hazards*, *64*(1), 903–933.

Peng, M., & Zhang, L. M. (2012b). Analysis of human risks due to dam-break floods—part 2: Application to Tangjiashan landslide dam failure. *Natural Hazards*, *64*(2), 1899–1923.

Pulvirenti, L., Pierdicca, N., Chini, M.,Guerriero, L. 2011. An algorithm for operational flood mapping from Synthetic Aperture Radar (SAR) data using fuzzy logic. *Nat. Hazards Earth Syst. Sci*., *11* (2), 529–540.

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3), 312–318.

Stewart, J., & Kennelly, P. J. (2010). Illuminated choropleth maps. *Annals of the Association of American Geographers*, *100*(3), 513–534.

Tsyganskaya, V., Martinis, S., Twele, A., Cao, W., Schmitt, A., Marzahn, P., & Ludwig, R. (2016). A fuzzy logic-based approach for the detection of flooded vegetation by means of Synthetic Aperture Radar data. ISPRS-*International Archives of the Photogrammetry*, *Remote Sensing and Spatial Information Sciences*, 371–378.

Viglione, A., Merz, R., Salinas, J. L., & Blöschl, G. (2013). Flood frequency hydrology 3: A Bayesian analysis. *Water Resources Research*, *49*(2), 675–692.

Vogel, K., Riggelsen, C., Scherbaum, F., Schröter, K., Kreibich, H., & Merz, B. (2013, June). Challenges for Bayesian network learning in a flood damage assessment application. In *11th International Conference on Structural Safety and Reliability* (pp. 16–20).

Vogel, K. (2014). Applications of Bayesian networks in natural hazard assessments.

Yang, Y., Webb, G. I., & Wu, X. (2010). Discretization methods. *Data Mining and Knowledge Discovery Handbook*, 101–116. Springer.

Zadeh, L. A. (2008). Is there a need for fuzzy logic? *Information Sciences*, *178*(13), 2751–2779.

Zwirglmaier, K., Papakosta, P., & Straub, D. (2013). Learning a Bayesian network model for predicting wildfire behavior. In *Proceedings of the 11th International Conference on Structural Safety and Reliability*.