How to cite this paper:

Nor Idayu Ahmad Azami, Nooraini Yusoff, & Ku Ruhana Ku-Mahamud. (2017). Effect of fuzzy discretization in the association performance with continuous attributes in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 29-36). Sintok: School of Computing.

EFFECT OF FUZZY DISCRETIZATION IN THE ASSOCIATION PERFORMANCE WITH CONTINUOUS ATTRIBUTES

Nor Idayu Ahmad Azami[1], Nooraini Yusoff[2], and Ku Ruhana Ku-Mahamud[3]

[1] *Universiti Utara Malaysia, Malaysia, ayuazami.uum@gmail.com*
[2]*Universiti Utara Malaysia, Malaysia, nooraini@uum.edu.my*
[3] *Universiti Utara Malaysia, Malaysia, ruhana@uum.edu.my*

ABSTRACT. Flood is one of the natural disasters caused by complex factors such as natural, breeding and environmental. The variability of such factors on multiple heterogeneous spatial scales may cause difficulties in finding correlation or association between regions. The interaction between these factors has resulted in provision of either diverse or repeated information which can be detrimental to prediction accuracy. The complex and diverse available database has triggered this study to incorporate multi-source heterogeneous data source in finding association between regions. Bayesian Network based method has been used to quantify dependency patterns in spatial data. However, a group of variables may be relevant for a particular region but may not be relevant to other region. To overcome the weakness of Bayesian network in handling continuous variable, this study has proposed data discretization technique to produce spatial correlation model. The effect of the proposed fuzzy discretization on the association performance is investigated. The comparison between different data discretization techniques proved that the proposed fuzzy discretization method gives better result with high precision, good F-measure, and a better receiver operating characteristic area compared with other methods. The results of correlation between the spatial patterns gives detailed information that may help the government, planners, decision makers, and researchers to perform actions that help to prevent and mitigate flood events in the future.

Keywords: spatial data mining, Bayesian network, fuzzy discretization

INTRODUCTION

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. Several studies can be seen, including works by various researchers (Li et al., 2010; Liang et al., 2012; Peng & Zhang, 2012a; 2012b; Viglione et al., 2013; Vogel et al., 2013) that 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, 2009; Uusitalo, 2007; Zwirglmaier et al., 2013). Dougherty, Kohavi, and Sahami (1995), Friedman and Goldsmith (1996), Aguilera, Fernández, Fernández, Rumí, and Salmerón (2011), and Vogel (2014) suggested to use discretization to overcome this problem.

Therefore, this study proposes 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 uncertainty in the association, discretization has become one of the preprocessing techniques that used to transform a continuous variable into a discrete one (Bakar, Othman, & Shuib, 2009; García, Luengo, & Herrera, 2015).

The advantage of discretization on continuous data can lead to data reduction and the simplification of data. Subsequently, this process will make the learning faster and produce shorter and compact results. Some reviews of the discretization technique can be found in the literature (e.g., Liu, Hussain, Tan, & Dash, 2002; Yang, Webb, & Wu, 2010). Second section discusses the previous studies of 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, the 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 to the supervised discretization methods. Supervised discretization methods utilize the class information in setting partition boundaries. Unsupervised discretization methods do not utilize instance labels for the selection of cut points. These methods 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) performs the minimum description length principle (MDLP) to discretize continuous environmental data using 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) provide an efficient supervised Bayesian discretization method to give better results for classification from a high-dimensional biomedical dataset. Ge, Cao, and Duan (2011) compared the impacts of three supervised discretization methods which are 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 unsupervised method that improved inputs of choropleth map information system. Moreover, Dawod, Mirza, and Al-Ghamdi (2012) also used natural breaks method to identify the break points of total flood volume values. Although the natural breaks method can handle volumes of spatial data, this method required predefined numbers of intervals before the discretization process. As explained by Marcot, Steventon, Sutherland, and McCann (2006), the maximum number of intervals or the discretization should be limited in five states to improve the precision and the network structure.

In this study, the membership function (MF) graph in fuzzy logic has been 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 all things admit of 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.

The major data acquisition for this study is focused on the environmental elements that can be classified into three categories: (1) time series data, which is the mean annual rainfall in 2010; (2) raster data, i.e. Interferometric Synthetic Aperture Radar (IfSAR); and (3) 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.

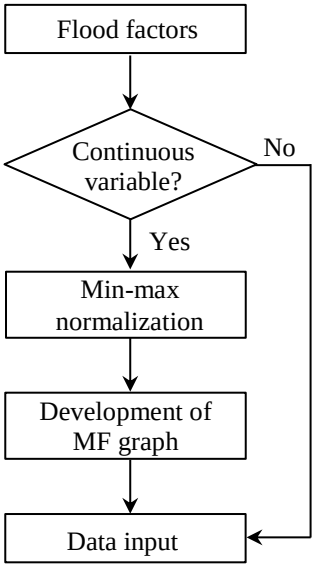

Figure 1. Flowchart for the proposed data discretization technique.

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 development of membership function graph, the entropy method is used to find the threshold value in order to develop the graph.

Digital Elevation Model

Digital elevation model (DEM) is 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 resulted dataset is the base of elevation models and digital surface. Since the surface conditions are the leading factors that determine the formation of flood events, therefore, the use of high-resolution synthetic data was the perfect source to derive the topographic factors of elevation, which are DEM, slope angle, curvature, SPI, TWI, and distance from river. Figure 2 shows the original data and reclassified data using fuzzy discretization.

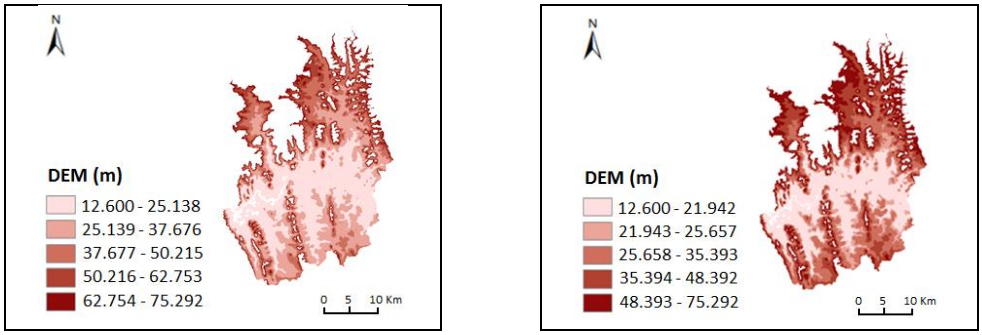

Figure 2. DEM maps: (a) the original DEM and (b) reclassified DEM.

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 great impact for the breeding of disasters as the sliding surface for the runoff process. The slope gradient is in degrees are shown in Figure 3.

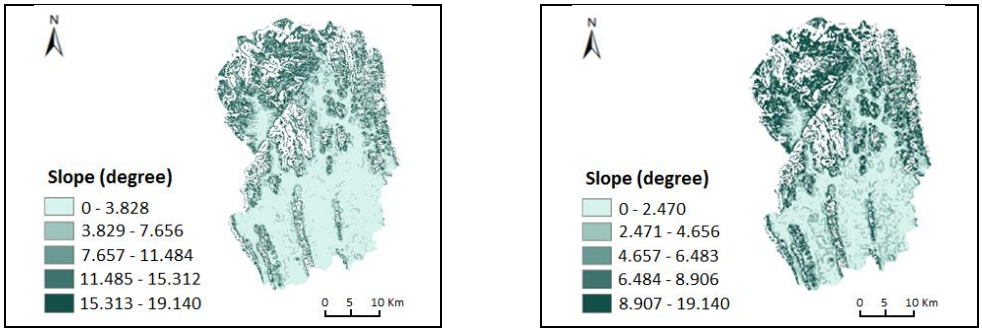

Figure 3. Slope maps: (a) the original Slope and (b) reclassified Slope.

Stream Power Index

Stream power index is the rate that the energy of flowing water is expended 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 flood plains, broad alluvial flats, and slowly subsiding areas, where valley fill is usually deepening and intact. The given equations have calculated and generated SPIs as shown in Figure 4.

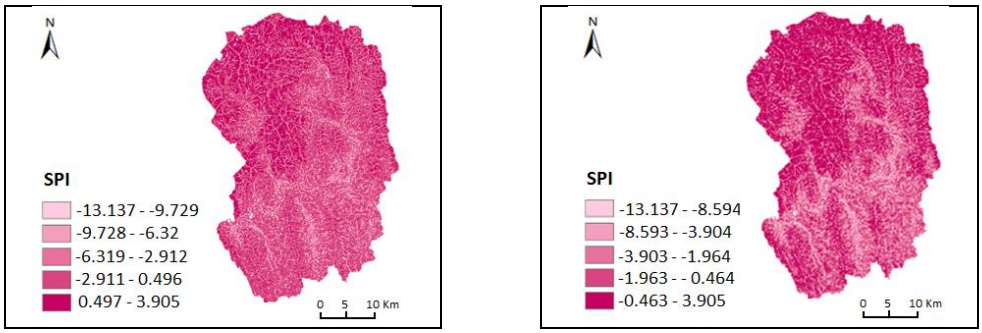

Figure 4. SPI maps: (a) the original SPI and (b) reclassified SPI.

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 corre-

sponding DEM. Higher TWI values represent drainage depressions; lower values represent crests and ridges. In creating the TWI, the following equation is calculated to produce the TWI. Figure 5 shows the original and reclassified TWI.

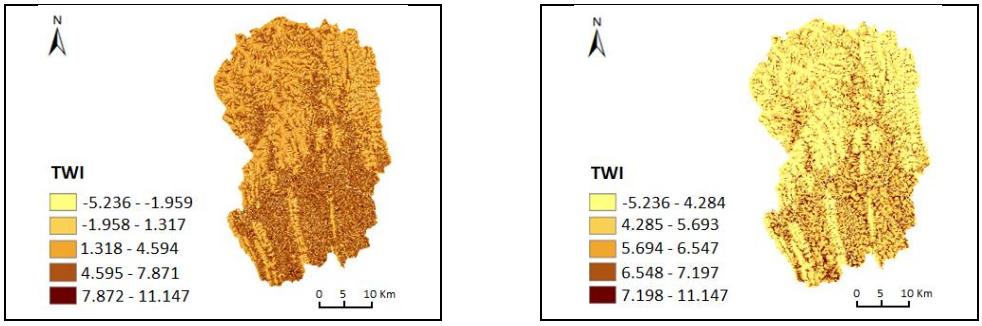

Figure 5. TWI maps: (a) the original TWI and (b) reclassified TWI.

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 is used to create a raster of the distance from river. Figure 6 shows the original and reclassified distance from river.

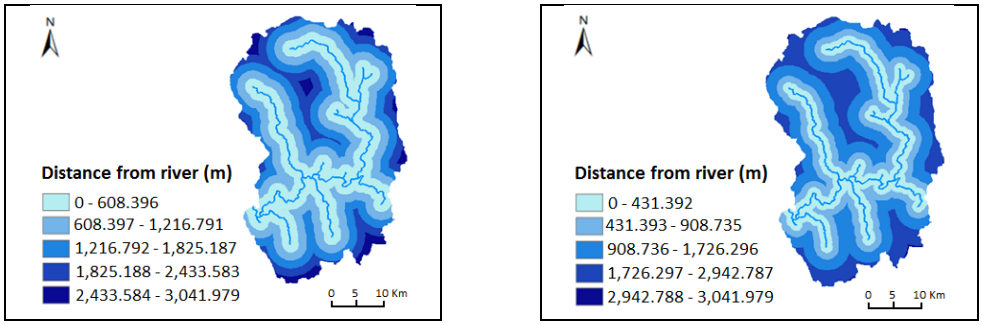

Figure 6. River maps: (a) the original River and (b) reclassified River.

Rainfall

The historical data that includes 18 rainfall stations with mean annual rainfall are obtained. In producing the mean annual rainfall intensity, the historical data are considered as the primary source of information. The available rainfall data is recorded at permanent but very disperse 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.

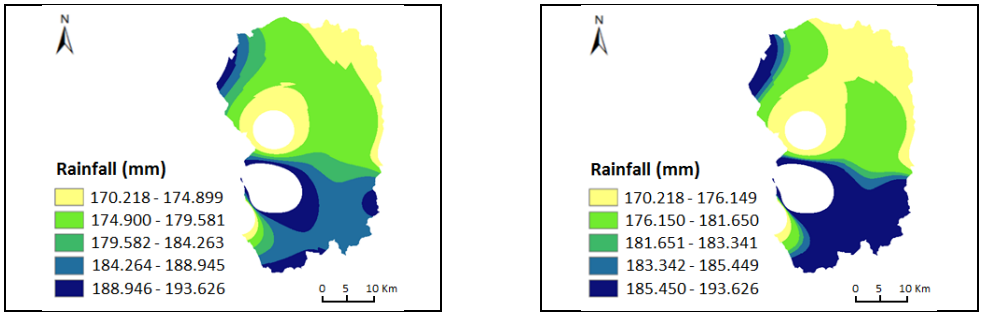

Figure 7. Rainfall maps: (a) the original Rainfall and (b) reclassified Rainfall.

RESULT AND DISCUSSION

For brevity, we discuss an example of the MF graph for rainfall data in Figure 2 after the calculation of entropy is complete. The development of the MF graph is to standardize the differences of rainfall data for each *mukim* that can be measured by using one graph. x0, x1, x2, x3, and x4 are the threshold values estimated using the Entropy method, which are 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.

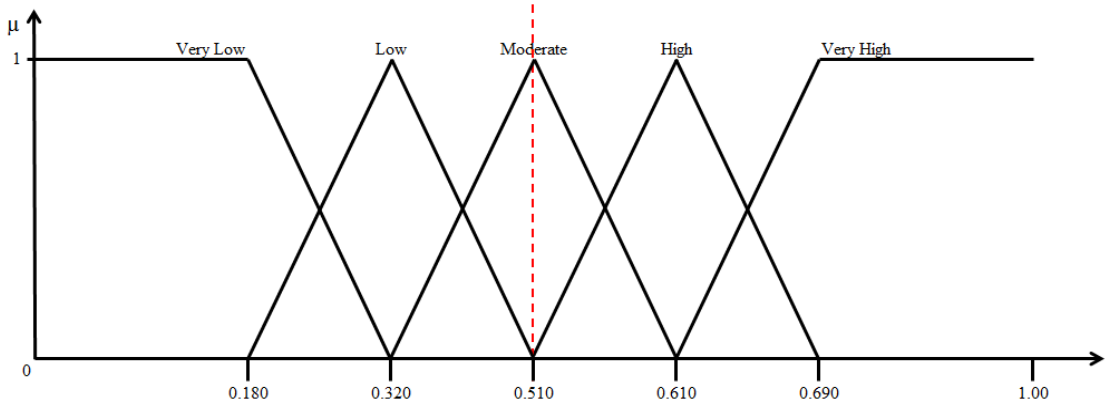

Figure 2. The membership function graph for rainfall data

By using this graph, the converted data is transformed into new representations for interval boundaries. The fuzzy set interval is then defined as shown in Table 1. The rainfall data has been normalized in the range of 0 to 1 and then transformed into new representations of fuzzy discretization by using the MF graph. This new representation has been used to enhance the correlation model of BN in the data discretization phase. This is applied to all data with continuous variables.

Table 1: Sample of transformed rainfall data

Rainfall data	Normalized classes	Linguistic variable	Fuzzy discretization
172.097	0.076	Very Low	1
173.332	0.130	Very Low	1
175.876	0.239	Very Low	1
176.150	0.251	Low	2
178.540	0.354	Low	2
181.651	0.488	Moderate	3
183.187	0.554	Moderate	3
184.117	0.594	High	4
185.417	0.650	High	4
187.886	0.756	Very High	5

Based from the experiments, it has been found that the proposed fuzzy discretization method shows better performance. This indicates that incorporating the proposed fuzzy discretization with the BN model give better results. The results from the performance metrics have shown that this method performed well as compared to other discretization methods.

In this study, five data discretization techniques for modelling the BN have been compared, namely Fuzzy Discretization, Equal Width, Natural Breaks, Quantile, and Geometrical Interval. The results are summarized in Table 2. The performance of the models is based on precision, F-measure, and receiver operating characteristic (ROC).

Table 2: Comparison of average performance assessment of BN models

Technique	Precision	F-Measure	ROC Area	Class
Fuzzy Discretization	0.992	0.980	0.984	Flood
	0.875	0.917	0.984	No Flood
Equal Width	0.820	0.680	0.805	Flood
	0.812	0.579	0.805	No Flood
Natural Breaks	0.831	0.669	0.803	Flood
	0.531	0.578	0.803	No Flood
Quantile	0.839	0.661	0.813	Flood
	0.529	0.578	0.813	No Flood
Geometrical Interval	0.819	0.682	0.814	Flood
	0.535	0.578	0.814	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 fuzzy discretization with the precision of 0.992, F-measure of 0.980, and receiver operating characteristic of 0.984 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.

ACKNOWLEDGMENTS

The authors wish to thank the 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.

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.

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.

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. In *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.

Hiwarkar, T. A. & Iyer, R. S. (2013). New applications of soft computing, artificial intelligence, fuzzy logic and genetic algorithm in bioinformatics.In *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.

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. In *International journal of computer applications* (0975-8887), national conference on future computing CoRR.

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.

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), 1.

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.* Springer Science & Business Media.

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.

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.

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*, pp. 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. *Proc. ICOSSAR* 2013