ICIC Express Letters
Part B: Applications
Volume **10**, Number **5**, May **2019**

ICIC International ©2019 ISSN 2185-2766
pp. 413–418

INTERACTIVE DASHBOARD OF FLOOD PATTERNS USING CLUSTERING ALGORITHMS

Friska Natalia[1], Yustinus Eko[1], Ferry Vincenttius Ferdinand[2]
I Made Murwantara[3] and Chang Seong Ko[4]

[1]Department of Information Systems
Universitas Multimedia Nusantara
Jl. Scientia Boulevard, Gading Serpong, Tangerang 15811, Indonesia
{ friska.natalia; yustinus.eko }@umn.ac.id

[2]Department of Mathematics
[3]Department of Computer Science
Universitas Pelita Harapan
Jl. M. H. Thamrin Boulevard, 1100 Lippo Vilage, Tangerang 15811, Indonesia
{ ferry.vincenttius; made.murwantara }@uph.edu

[4]Department of Industrial and Management Engineering
Kyungsung University
309, Suyeong-ro, Nam-gu, Busan 48434, Korea
csko@ks.ac.kr

Received October 2018; accepted January 2019

Abstract. *Floods are the most common natural disaster and the leading cause of natural disaster mortality worldwide. Flooding has many impacts. It damages property and endangers the lives of humans and other species. Indonesian Disaster Data and Information in 2018 showed that the disaster that often occurs in Indonesia is flooding disaster. Flood control will become an ever-increasing issue in many countries. This research explored the flood patterns and found the potential area of the flood that revealed in the interactive dashboard in the rivers of Tangerang, Indonesia. Tangerang has three rivers which are Angke, Pesanggrahan, and Cisadane. This research used three variables which are water level, river, and stations. This paper used three clustering methods: K-medoids, DBSCAN, and X-Means to analyze possible water level rises patterns in Tangerang and the results are visualized in the form of the interactive dashboard that is simple and easy to use for non-technical users.*
Keywords: K-medoids, DBSCAN, X-Means, Knowledge discovery in databases, Interactive dashboards

1. **Introduction.** Indonesian Disaster Data and Information in 2018 showed that flood has reached the highest position of overall disasters in Indonesia [1] which is described in Figure 1. There are some previous researches about disaster management. First, river water levels can be predicted using several methods, namely single exponential smoothing, double exponential smoothing, and holt [2]. Second, a flood detection system connecting to hardware is developed to perform an action when a flood is detected, such as turn on flood alarm by using fuzzy logic with Mamdani model (max-min) [3]. Third, real-time floods can be predicted using rainfall and run-off models [4].

The difference compared with the previous researches is that this paper proposes an interactive dashboard on the river stations within Tangerang area which is built based on the result of these three clustering methods and implemented by using Power BI software. On the dashboard, users can see the average of water level based on the date (per year, month, and day), station, and river. Also, if users choose the day, they can find the water

DOI: 10.24507/icicelb.10.05.413

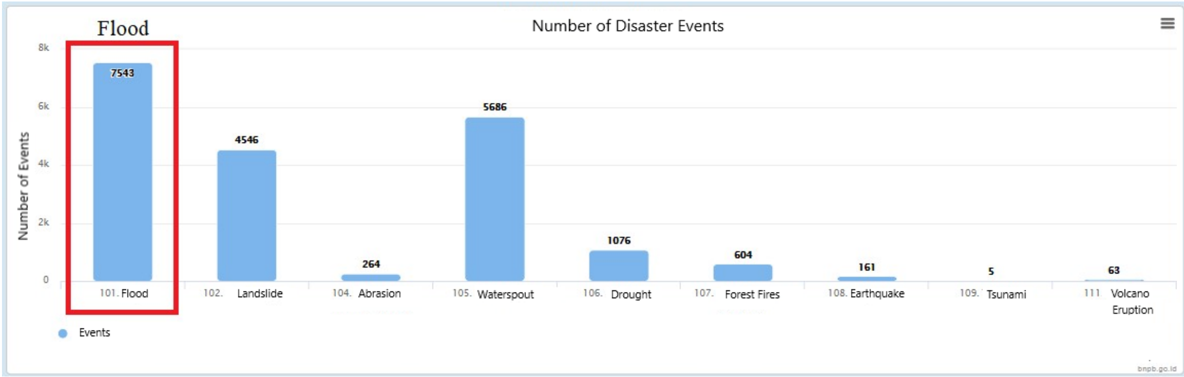

FIGURE 1. Indonesian Disaster Data and Information in 2018

level per hour. All the data are obtained from Ciliwung Cisadane River Basin Center (BBWS) from 2013 to 2017 with total data of 315,551 rows. In this paper three clustering methods: K-medoids, DBSCAN, and X-Means are used to analyze possible water level rises patterns in Tangerang based on these four stations: Serpong (Cisadane), Batubeulah (Cisadane), Pamulang (Angke), and Sawangan (Pesanggrahan).

2. **Problem Statement and Research Methodology.** The objective of the study is to explore the flood patterns in Tangerang and develop the interactive dashboard to visualize the patterns that are simple and easy to use for non-technical users. The interactive dashboard perhaps can be used to help the BBWS to conduct maintenance services and also can see the water level patterns that occurred during the period that has been predicted possible water level rises. Knowledge Discovery in Databases (KDD) is a process of discovering useful information from a collection of data [5,6]. This method is used and there are five stages to detect the potential flood.

a. Collection and Selection. Not all the data that the BBWS collects are useful to this research. The previous data in the excel are shown in Figure 2. To perform data selection, only the important and necessary information is required for this study, and the following target data will be appropriate to this research.
   1) The complete list of water level in Tangerang based on time.
   2) The complete list of water level in Tangerang based on day (per hour).
   3) The location of each river (Angke, Pesanggrahan, and Cisadane).

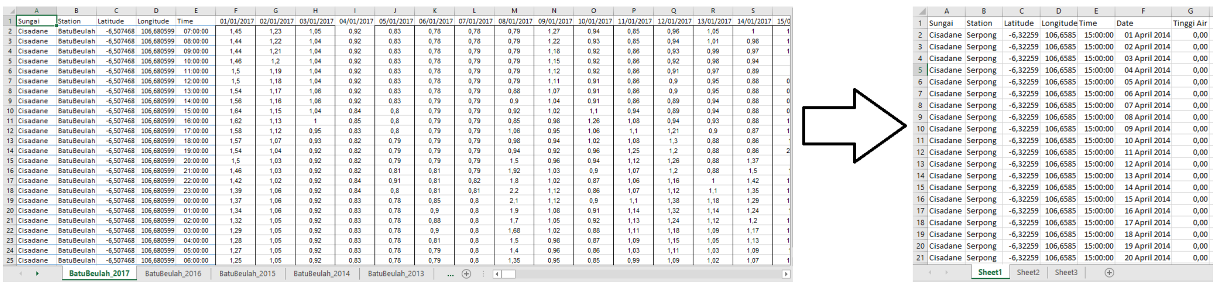

FIGURE 2. BBWS data from 2013 to 2017

b. Preprocessing/Cleaning. Data preprocessing is vital in the KDD steps. This stage is necessary because real data are likely to contain errors, inconsistent and sometimes incomplete. Therefore, in this study Rapid Miner and Power BI are used as a tool to complete this stage by deleting unused rows and columns and equalizing the name of seven columns; those are river, station, latitude, longitude, time, date, and water level.

c. Transformation. The transformed data will be made into several clusters based on the number of water level in each station using clustering method with K-medoids, DBSCAN, and X-Means clustering algorithms.
d. Data Mining. Rapid Miner Studio is used in this study as a visual workflow designer for data scientists. The Bayesian Information Criterion is implemented for the calculation of X-Means method in Rapid Miner using X-Means operator which calculates the optimal number of clusters automatically. The K-Means algorithm is implemented in Rapid Miner using Cluster Distance Performance operator with assuming $k = 3$ that will be used as Davies-Bouldin Index of the resulting clusters.
e. Interpretation/Evaluation. In this stage, the patterns which are mined are presented to the end user using visual aids. In this study, this step involves data visualization techniques which represented in the interactive dashboard to present the results to the users in an easily understandable way.

3. **Solution Procedure.** The clustering method is used to separate the data into meaningful groups [7]. In this study we used three different clustering methods on several parameters to obtain the flood patterns and found the potential area of the flood that revealed in interactive dashboard in the rivers of Tangerang, Indonesia. The analysis is accompanied based on the results obtained from implementing three clustering methods: K-medoids, DBSCAN, and K-Means. The results of all of the clustering methods are compared to make the final conclusion on the pattern and showed in the interactive dashboard by using Power BI.

3.1. **Clustering with K-medoids.** The implementation of K-medoids is done by using Rapid Miner. The data will be clustered based on the water level. The basic principle of K-medoids method is that the minimization of the total sum of the distance of dissimilar points from a reference point should be done for partitioning [8,9]. In this study, we chose to use that index as a metric to evaluate the performance of each cluster with using assumption number of $k = 3$ (meaning: high, medium, and low). The visualization of K-medoids can be seen in Figure 3. Based on Figure 3 that shows the dashboard, it can

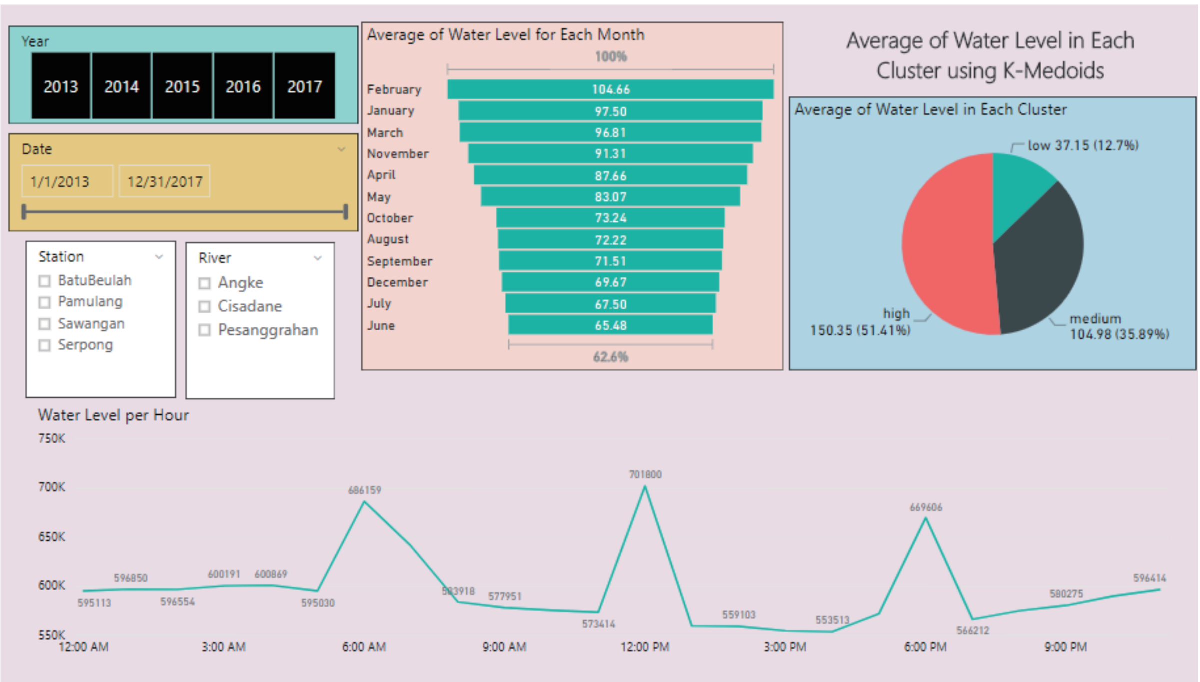

FIGURE 3. Visualization of K-medoids

be seen the characteristics of each cluster's member. For 2013 to 2017 of all rivers, the averages of water level on each cluster are:

- Cluster 0 (High): the group of a period with a relatively high number such as an average water level of 150.35 cm (51.41%);
- Cluster 1 (Medium): the group of a period with a moderate average water level such as an average water level of 104.98 cm (35.89%);
- Cluster 2 (Low): the group of a period with a relatively low number of average water level such as an average water level of 37.15 cm (12.7%).

Also, from the dashboard it can be seen that the three highest average water levels are in February (104.66 cm), January (97.50 cm), and March (96.81 cm). Based on the time, the three highest average water levels are on 12:00 PM, 6:00 AM, and 6:00 PM.

3.2. **Clustering with DBSCAN.** The implementation of DBSCAN is done by using Rapid Miner. The data will be clustered based on the water level. The algorithm gradually converts the unclassified objects into classified or noise objects. Density-Based Spatial Clustering and Application with Noise (DBSCAN) was a clustering algorithm based on density [10]. The parameters are: Eps: 1.0 and MinPts: 5. The number of clusters will be divided based on those two parameters (Eps and MinPts). The optimum number of clusters based on the parameters (Eps: 1.0 and MinPts: 5) is 2, that is high cluster and low cluster. The visualization for DBSCAN algorithm is displayed in Figure 4.

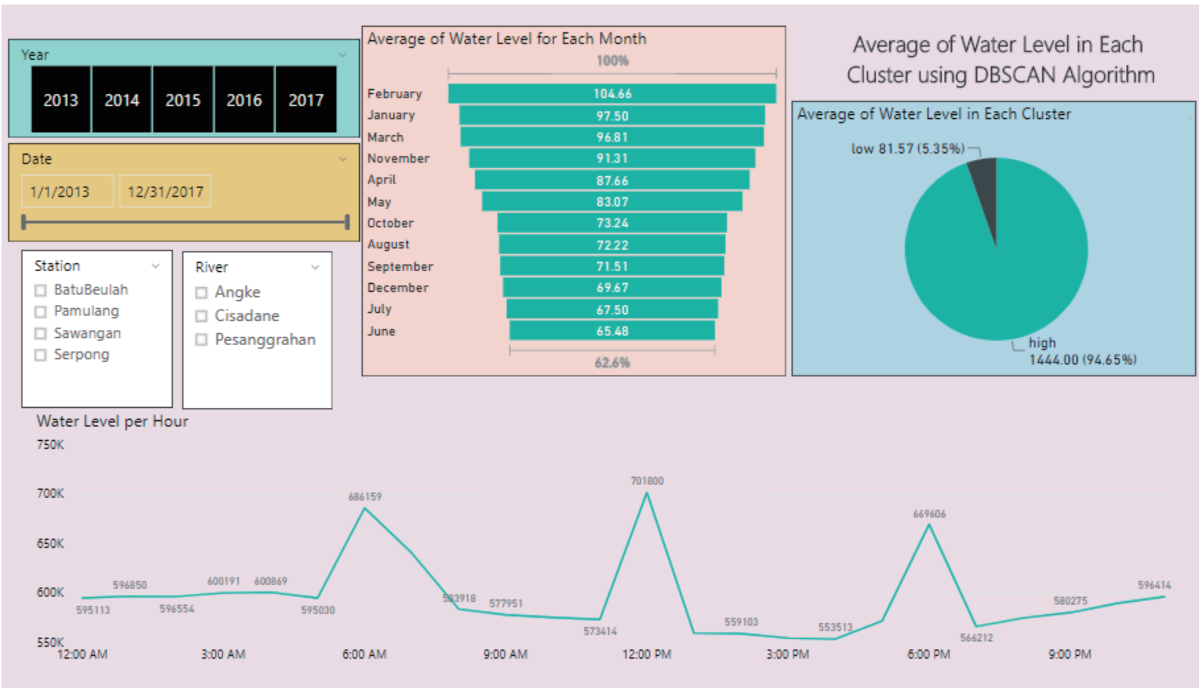

FIGURE 4. Visualization of DBSCAN

Based on Figure 4 that shows the dashboard, it can be seen the characteristics of each cluster's member. For 2013 to 2017 of all rivers, the averages of water level on each cluster are:

- Cluster 0 (High): the group of a period with a relatively high number such as an average water level of 1444.00 cm (94.65%);
- Cluster 1 (Low): the group of a period with a relatively low number such as an average water level of 81.57 cm (5.35%).

Also, from the dashboard it can be seen that the three highest average water levels are in February (104.66 cm), January (97.50 cm), and March (96.81 cm). And based on the time, the three highest average water levels are on 12:00 PM, 6:00 AM, and 6:00 PM.

3.3. **Clustering with X-Means.** The implementation of X-Means is done by using Rapid Miner. The data will be clustered based on the water level. X-Means is an extended clustering algorithm of K-Means. X-Means can identify the best number of clusters k by itself based on the Bayesian Information Criterion (BIC) [11]. The parameters are: min_k: 2 and max_k: 10. The number of clusters will be divided based on BIC of those two parameters (max_k and min_k). The optimum number of clusters based on the parameters (min_k: 2 and max_k: 10) is 2, that is the high cluster and low cluster. The visualization for X-Means algorithm is displayed in Figure 5.

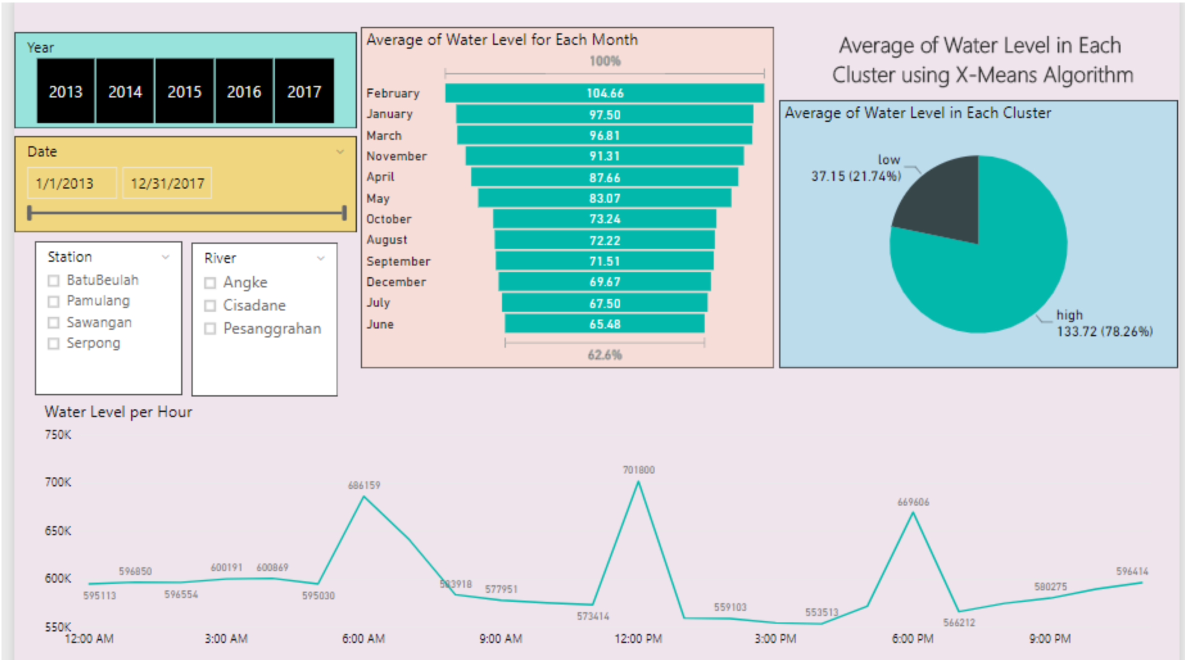

FIGURE 5. Visualization of X-Means

Based on Figure 5 that shows the dashboard, it can be seen the characteristics of each cluster's member. For all rivers from 2013 to 2017, the averages of water level on each cluster are:

- Cluster 0 (High): the group of a period with a relatively high number such as an average water level of 133.72 cm (78.26%);
- Cluster 1 (Low): the group of a period with a relatively low number such as an average water level of 37.15 cm (21.74%).

Also, from the dashboard it can be seen that the three highest average water levels are in February (104.66 cm), January (97.50 cm), and March (96.81 cm). And based on the time, the three highest average water levels are on 12:00 PM, 6:00 AM, and 6:00 PM. Based on the clustering methods in Figure 3 through Figure 5 it shows the final result of the interactive dashboard to analyze possible water level rises patterns in Tangerang. Using the dashboard, users can see the average of water level based on the date (per year, month, and day), station, and river. It is also able to show daily status to see hourly water level. The interactive dashboard design is user-friendly, simple and should be able to be used by common people without specific technical expertise. Based on the interactive dashboard, the result can be described in the same pattern. Based on the months, the highest average water level is always in February, January, and March. Also, based on the months, the three highest average water levels are on 12:00 PM, 6:00 AM, and 6:00 PM which shows a high flood risk on these periods.

4. **Conclusion.** In this research, the data of water level in Tangerang from 2013 to 2017 is clustered by using K-medoids, DBSCAN, and X-Means clustering algorithms. The data

collected then cleaned in the preprocessing stage. To take this into account, we considered dividing the experiment into two same results, which calculated based on month and time. On data mining process, clustering is done with three methods. The clustering result for K-medoids is with three clusters, DBSCAN is with two clusters, and X-Means is with two clusters. Based on the results, the pattern on the flood for five years can be seen in February, January and March. These results are visualized in form of the interactive dashboard that is simple and easy to use for non-technical users. Furthermore, other problems in disaster management can be solved and finding a better solution procedure will be studied in future research.

Acknowledgements. The study is conducted by grant from Indonesian Ministry of Research, Technology and Higher Education 034/KM/PNT/2018. All computations are performed in Big Data Laboratory of Universitas Multimedia Nusantara. All data are provided by Ciliwung Cisadane River Basin Center (BBWS).

REFERENCES

[1] BNPB, *Indonesia Disaster Data and Information*, 2018, http://dibi.bnpb.go.id/, accessed in June 2018.

[2] N. S. Muhamad and A. M. Din, Exponential smoothing techniques on time series river water level data, *Proc. of the 5th International Conference on Computing and Informatics (ICOCI 2015)*, pp.644-649, 2015.

[3] P. C. Nayak, K. P. Sudheer and K. S. Ramasastri, Fuzzy computing based rainfall-runoff model for real time flood forecasting, *Hydrological Processes*, vol.19, pp.955-968, 2005.

[4] A. S. Baharom, Z. Idris, S. S. M. Isa, M. Nazir and A. Khan, Prediction of flood detection system: Fuzzy logic approach, *International Journal of Enhanced Research in Science Technology & Engineering*, vol.3, no.1, pp.335-339, 2014.

[5] L. R. Julian and F. Natalia, The use of web scraping in computer parts and assembly price comparison, *Proc. of the 3rd International Conference on New Media (CONMEDIA)*, pp.8-13, 2015.

[6] S. Monica, F. N. Ferdinand and S. Sudirman, Clustering tourism object in Bali province using K-Means and X-Means clustering algorithm, *The 16th IEEE International Conference on Smart City (SmartCity)*, 2018.

[7] K. Supangat and Y. E. Soelistio, Bus stops location and bus route planning using mean shift clustering and ant colony in west Jakarta, *IOP Conference Series: Materials Science and Engineering*, vol.185, no.1, 2017.

[8] K. G. Soni and D. A. Patel, Comparative analysis of K-means and K-medoids algorithm on IRIS data, *International Journal of Computational Intelligence Research*, vol.13, no.5, pp.899-906, 2017.

[9] P. Sengottuvelan and T. Gopalakrishnan, Efficient web usage mining based on K-medoids clustering technique, *International Journal of Computer and Information Engineering*, vol.9, no.4, pp.1044-1048, 2015.

[10] C. Raj, Comparison of K-Means, K-medoids, DBSCAN algortihms using DNA microarray dataset, *International Journal of Computational and Applied Mathematics*, vol.12, no.1, pp.344-355, 2017.

[11] N. T. Nguyen, R. Kowalczyk, C. Orłwski and A. Ziółowski, *Trans. Computational Collective Intelligence XXV*, Springer, Berlin, 2016.