DOI: 10.12928/TELKOMNIKA.v17i3.11774

Cluster-based water level patterns detection

Friska Natalia Ferdinand*[1], Yustinus Soelistio[2], Ferry Vincenttius Ferdinand[3], I Made Murwantara[4]
[1,2]Universitas Multimedia Nusantara, Tangerang, Banten, Indonesia
[3,4]Universitas Pelita Harapan, Tangerang, Banten, Indonesia
*Corresponding author, e-mail: friska.natalia@umn.ac.id[1], yustinus.eko@umn.ac.id[2],
ferry.vincenttius@uph.edu[3], made.murwantara@uph.edu[4]

Abstract

Indonesian Disaster Data and Information in 2016 showed that flood has reached a soaring 32.2% overall. In one of the common flood region (2016), Tangerang, the flood had impacted 30,949, and destroys more than 400 residentials. In spite of this dreadful fact, Tangerang has no systematically ways of detecting the flood patterns. Therefore, there is urgency for a system that is able to detect potential flood risks in Tangerang. This study explores a mean to systematically find flood patterns in Tangerang and attempt to visualize the risks based on 11 years of data on four major river stations within Tangerang vicinity. All the data obtained from Ciliwung Cisadane River Basin Center (BBWS) between 2009 until 2017 with total data of 368,184 rows. This study proposes an interactive dashboard based on the water level data covering rivers of Angke, Pesanggrahan, and Cisadane. Three clustering methods are implemented, the K-Medoids, DBScan, and x-means, to segregate the water level data, taken from four stations obtained from Ciliwung Cisadane River Basin Center (BBWS), into meaningfull periodic flood patterns. The output of this research is an interactive dashboard created based on the newly found patterns. The dashboard is designed to be simple and easy to use for non-technical persons. We believe that the output of this research could be implemented into the decision-making process taken by the Ciliwung Cisadane River Basin Center (BBWS) in order to improve countermeasure attempts on the potentially flooded areas.

Keywords: *dashboard, DBscan, K-medoids, knowledge discovery in databases, X-means*

1. Introduction

Indonesia is prone to flood disaster. According to the National Disaster Management Agency in 2016 [1], flood is one of the most often occurring disaster in Indonesia byas much as 32.2% with 713 incidents Figure 1.

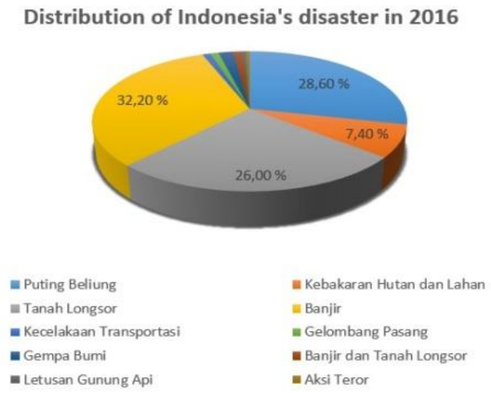

Figure 1. Distribution of Indonesia's disaster in 2016

According to the National Disaster Management Agency, the losses and damages caused by floods in Tangerang in 2016 alone reaches four mortality, 30,949 people suffering from floods, 5,313 displaced people, and immaterial losses of 403 residences, 11 education

Received June 5, 2018; Revised November 13, 2018; Accepted January 2, 2019

facilities, and 624 Ha of land damaged. The impact of this flood disaster can be reduced if the communities are informed by some prediction of potential flood risks that occur ahead.

Flood do not occur only in watersheds but also in urban areas or areas far from streams, for example in densely populated areas and roads that have no drainage or good uptake which makes it less obvious though still predictable. This flood behavior creates a need for an early warning system in such areas. One common approach to the flood early warning system is a visualization tool which has been used in many of similar systems.

According to the National Disaster Management Agency, the losses and damages caused by floods in Tangerang in 2016 alone reaches four mortality, 30,949 people suffering from floods, 5,313 displaced people, and immaterial losses of 403 residences, 11 education facilities, and 624 Ha of land damaged. The impact of this flood disaster can be reduced if the communities are informed by some prediction of potential flood risks that occur ahead. Flood do not occur only in watersheds but also in urban areas or areas far from streams, for example in densely populated areas and roads that have no drainage or good uptake which makes it less obvious though still predictable. This flood behavior creates a need for an early warning system in such areas. One common approach to the flood early warning system is a visualization tool which has been used in many of similar systems.

This study focus on exploring ways to visualize the flood periodic occurrence in Tangerang area due to its urgencies to reduces the impact of flood disaster. Currently Tangerang has no application that can detect and tell potential flood and geographical location of Tangerang (and its nearby areas i.e. Jakarta and Bogor) which makes it more vulnerable to upcoming flood event. This study proposes an interactive dashboard on the river stations within Tangerang area. Tangerang has three rivers which are Angke, Pesanggrahan and Cisadane (illustrated in Figure 2). The information in the dashboard are collected from clustering method as is common in solving high dimensionality logistic problem (c.f. [2-5]).

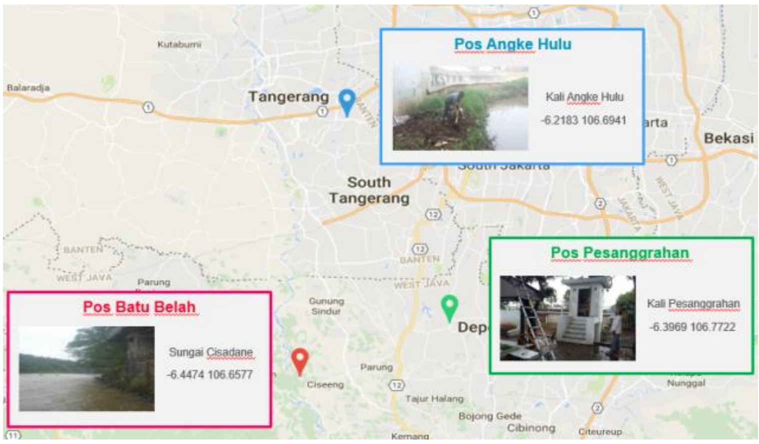

Figure 2. Geographical location of Tangerang

We implement three clustering methods: K-Medoids, DBScan and x-means to analyze possible water level rises patterns (i.e. high risk flood event) in Tangerang based on these four stations. The visualization in the dashboard is built based on the result of these three clustering methods and implemented by using Power BI software. We use the hourly water level data on the four stations obtained from Ciliwung Cisadane River Basin Center (BBWS) between 2009 until 2017 with total data of 368,184 rows.

2. Problem Statement and Research Method

The goal of this study is to explore possible Tangerang's periodic flood patterns and visualize the patterns in form of dashboard visualization. The visualization could be used as a benchmark to prioritize flood prevention attempts such as preparing water pumps at high risk points and rivers bed maintenance schedules in areas with great potential for flooding.

The visualization could also be used to help the Ciliwung Cisadane River Basin Center (BBWS) conducts maintenance services. Given this visualization data as a Knowledge Discovery in Databases (KDD), Ciliwung Cisadane River Basin Center (BBWS) can see the water level patterns that occurred during the period that has been predicted.

As research method, we follow the KDD process [6, 7] of discovering useful information from a collection of data. In the [6] by Julian & Natalia, they conducted a research to build an application with a purpose to recommend to its users in assembling computer that suit their needs so they can get a better price from build a computer that they need by using web scrapping. Similarly, Monica et.al. in [7], this paper presents the finding from analysing the large amount of data that the Indonesian Government Tourism Office, specifically regarding tourism in Bali. They use K-Means and X-Means algorithms to cluster the various type of tourist attractions in Bali according to their popularity and Power BI to develop the interactive dashboard. The difference between the previous is in this paper we use KDD and uses three methods in the clustering to analyze possible water level rises patterns. Figure 3 ilustrate the four steps of the KDD based on [6, 7]. The details of these steps are explain in section 2.1 through 2.4 and the results are presented in section 3.

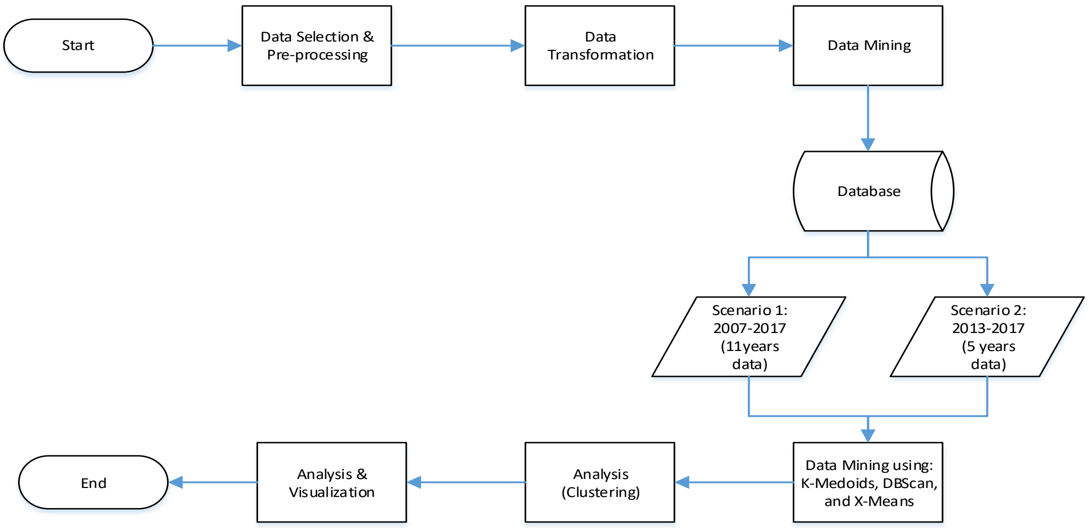
Start
Data Selection & Pre-processing
Data Transformation
Data Mining
Database
Scenario 1: 2007-2017 (11years data)
Scenario 2: 2013-2017 (5 years data)
Data Mining using: K-Medoids, DBScan, and X-Means
Analysis (Clustering)
Analysis & Visualization
End

Figure 3. The four steps of the KDD implemented in this research

2.1. Selection

The obtained data is water level per-hour from 2007 to 2017 on the four river stations: Serpong (Cisadane), Batubeulah (Cisadane), Pamulang (Angke), and Sawangan (Pesanggrahan) with total data of 368,184. The data is obtained from Ciliwung Cisadane River Basin Center (BBWS).

2.2. Selection

The steps undertaken in data preprocessing mainly fall into two categories, namely the removal of noise or outliers, and strategies for handling missing data fields. Specifically, in this study, we perform data preprocessing by following these steps using Power BI:

a. Delete unused rows and columns.
b. Equalizing the name of 7 columns; those are River, Station, Latitude, Longitude, Time, Date, and Water Level.

2.3. Transformation

Data that has been through the preprocessing phase will be transformed so that can be used for data mining process. In this case, we merge the data into one continuous vectors to be processed using clustering methods in R Version 3.4.4.

2.4. Data Mining

In the data mining process, clustering is done using K-Medoids, DBScan, and X-Means on two scenarios based on [8]:
Scenario 1: Using the complete data between 2007 until 2017.
Scenario 2: Based on the previous research, using the data from 2013 to 2017.

The clustering methods are used to segregate the data into meaningful groups. The patterns are strongly detected when there are some agreements between the results on both scenarios (i.e. since the clustering structures are still intact regardless of the number of data).

3. Analysis and Results

The analysis is conducted 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 final conclusion on the pattern.

3.1. Data Clustering on K-Medoids

The implementation of K-Medoids is done by using R. The data will be clustered based on hourly water level. 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 steps of K-Medoids process are [9-14]:

a. Arbitrarily choose k=3 data items as the initial medoids
b. Assign each remaining data item to a cluster with the nearest medoid
c. Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.
d. If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of K-Medoids.
e. Repeat steps 2, 3, and 4 till the medoids stabilize their location until 50 iterations.

The result for K-Medoids algorithm is displayed in a dashboard:

Figure 4 shows the characteristics of each cluster's member. For scenario I and scenario II of all rivers, the averages of water level on each cluster are represented in Table 1:

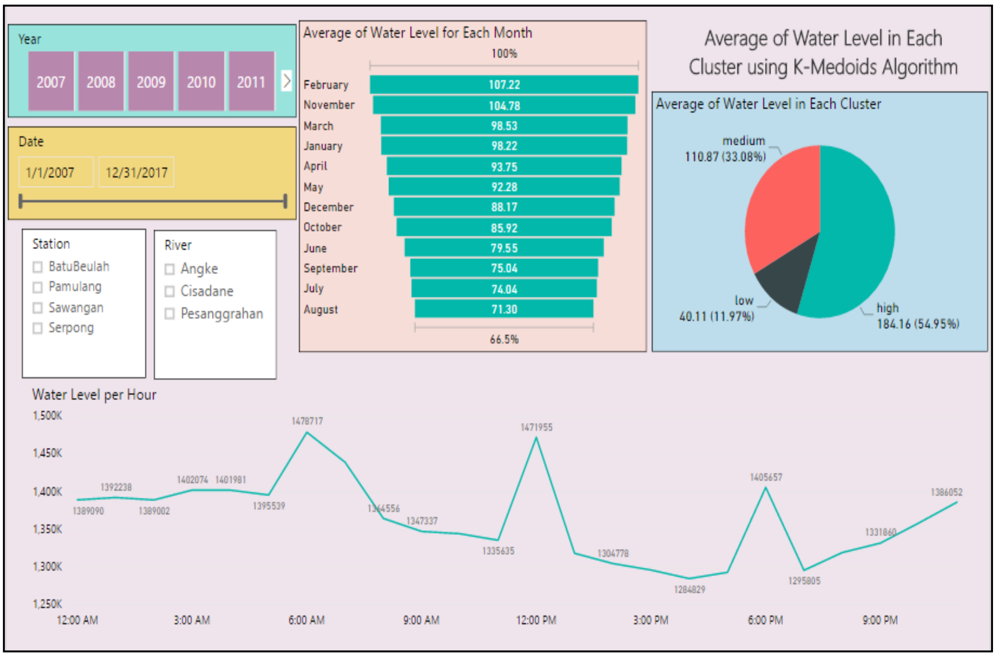

Figure 4. Dashboard example of K-Medoids as in scenario 1

Table 1. The Comparison of Scenario I and II using K-Medoids

Category	Scenario I	Scenario II
Clustering Results	Cluster 0 (High): group of period with average water level relatively high number, with an average water level of 184.16 cm (54.95%). Cluster 1 Medium: group of period with a moderate average water level, with an average water level of 110.87 cm (33.08%). Cluster 2 Low: group of period with relatively low number of average water level, with an average water level of 40.11 cm (11.97%).	Cluster 0 (High): group of period with average water level relatively high number, with an average water level of 150.35 cm (51.41%). Cluster 1 Medium: group of period with a moderate average water level, with an average water level of 104.98 cm (35.89%). Cluster 2 Low: group of period with relatively low number of average water level, with an average water level of 37.15 cm (12.7%).
Highest Average Water Level Based on Month	The three highest average water levels are on February (107.22 cm), November (104.78 cm), and March (98.53 cm).	The three highest average water levels are on February (104.44 cm), January (97.50 cm), and March (96.81 cm).
Highest Average Water Level Based on Time	The three highest average water levels is on 12:00 PM, 6:00 AM, and 6:00 PM.	The three highest average water levels is on 6:00 AM, 12:00 PM, and 6:00 PM

3.2. Data Clustering on DBScan

The implementation of DBScan is done by using R. The data will be clustered based on the water level. We implement different parameters from the previous study [8]:

a. Eps: Previous is 1.0, currently is 0.01.

b. MinPts: Previous is 5, currently is 1000 to 6000.

We use different values of the parameters because as the previous result has a very low clustering resolution (unbalance groups where one of the group has only one member). The number of cluster will be divided based on those two parameters (Eps and MinPts). These are the steps of DB Process [15-20]:

a. Arbitrary selection of a point p.

b. Retrieve all points density-reachable from p w.r.t Eps and MinPts (Eps: Previous is 1.0, currently is 0.01; MinPts: Previous is 5, currently is 5000).

c. If p is a core point, then a cluster is formed.

d. If p is a border point, no points are density-reachable from p and DBScan visits the next point of the database.

e. Continue the process until all of the points have been processed.

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 result for DBScan method is displayed in a dashboard: Based on the Figure 5 that shows the dashboard, it can be seen the characteristics of each cluster's member. For scenario I and scenario II of all rivers, the averages of water level on each cluster is represented in Table 2:

Table 2. The Comparison of Scenario I and II using DBScan

Category	Scenario I	Scenario II
Clustering Results	Cluster 0 (High): group of period with average water level relatively high number, with an average water level of 100.21 cm (54.95%). Cluster 1 Medium: group of period with a moderate average water level, with an average water level of 7 cm (33.08%). Cluster 2 Low: group of period with relatively low number of average water level, with an average water level of 0 cm (11.97%).	Cluster 0 (High): group of period with average water level relatively high number, with an average water level of 1444.00 cm (94.65%). Cluster 1 Low: group of period with relatively low number of average water level, with an average water level of 81.57 cm (5.35%).
ighest Average Water Level Based on Month	The three highest average water levels are on February (107.22 cm), November (104.78 cm), and March (98.53 cm)	The three highest average water levels are on February (104.44 cm), January (97.50 cm), and March (96.81 cm).
Highest Average Water Level Based on Time	The three highest average water levels is on 12:00 PM, 6:00 AM, and 6:00 PM.	The three highest average water levels is on 6:00 AM, 12:00 PM, and 6:00 PM

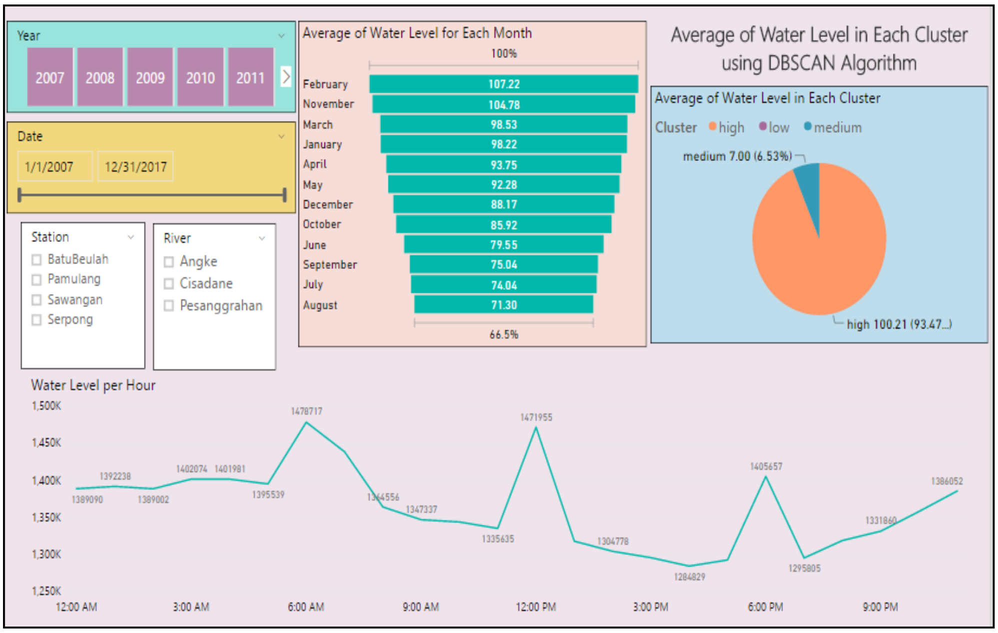

Figure 5. Dashboard example of DBScan implementation of scenario 1

3.3. Data Clustering on X-Means

The number of cluster will be divided based on Bayesian Information Criterion (BIC) of those two parameters (*max_k* and *min_k*) [21]. The optimum number of clusters based on the parameters (*min_k*: 2 and *max_k*: 10) is 2, that is high cluster and low cluster. The result for X-Means algorithm is displayed in a dashboard in Figure 6. Based on the Figure 6 that shows the dashboard, it can be seen the characteristics of each cluster's member. For scenario I and scenario II on all rivers, the averages of water level on each cluster are represented in Table 3.

Table 3. The Comparison of Scenario I and II using X-Means

Category	Scenario I	Scenario II
Clustering Results	Cluster 0 (High): group of period with average water level relatively high number, with an average water level of 184.16 cm (54.95%). Cluster 1 Medium: group of period with a moderate average water level, with an average water level of 110.87 cm (33.08%). Cluster 2 Low: group of period with relatively low number of average water level, with an average water level of 40.11 cm (11.97%).	Cluster 0 (High): group of period with average water level relatively high number, with an average water level of 133.72 cm (78.26%). Cluster 1 Low: group of period with relatively low number of average water level, with an average water level of 37.15 cm (21.74%).
Highest Average Water Level Based on Month	The three highest average water levels are on February (107.22 cm), November (104.78 cm), and March (98.53 cm).	The three highest average water levels are on February (104.44 cm), January (97.50 cm), and March (96.81 cm)
Highest Average Water Level Based on Time	The three highest average water levels is on 12:00 PM, 6:00 AM, and 6:00 PM.	The three highest average water levels is on 6:00 AM, 12:00 PM, and 6:00 PM

Similarly, the implementation of X-Means [21-26] is done by using R. The data will be clustered based on the water level using parameters:

a. min_k: 2
b. max_k: 10

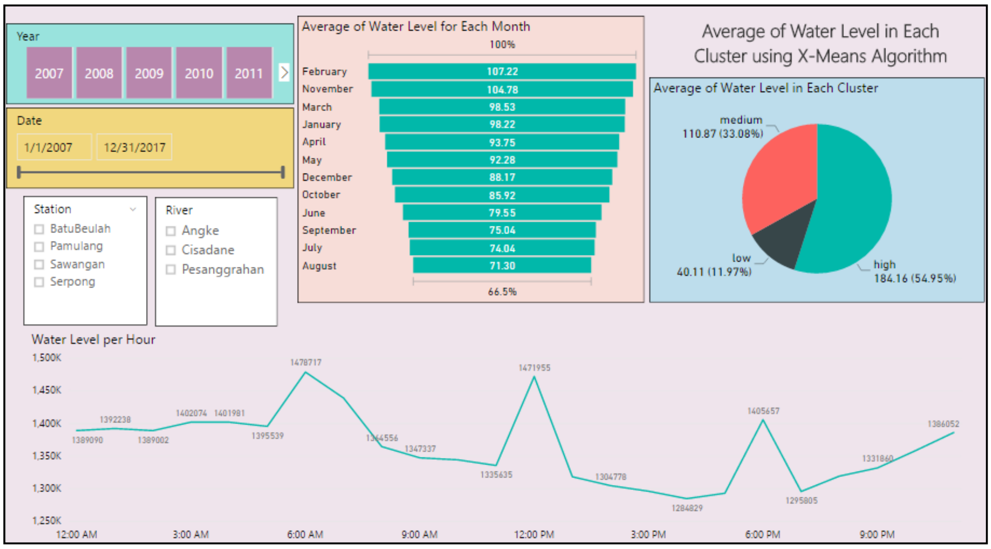

Figure 6. Dashboard example of the x-means method results on scenario 1

4. Discussion

Figure 4 through 6 show some example of the interactive dashboard. 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 design is simple and should be able to be used by common people without specific technical expertise. Comparing the results from the three clustering methods used, K-Medoids is the only method that give consistent results over both scenarios. DBScan and K-Means clustering produce the same clustering variation results on the scenarios. This differences could enlighten some properties of the data. First, the stability of K-Medoids can be interpreted as similar L1 (median) error on both scenarios which indicate similar range of water level between scenarios (i.e. the range of water level is similar between 2007-2017 and 2013-2017. Second, the disparity results from DBScan and K-Means indicate an L2 (mean) gaps between scenarios (i.e. the distribution of the data is different between 2007-2017 and 2013-2017).

The combination of these two properties signify a shift in water level distribution in the first seven years (2007-2013), yet still within persistent range. Furthermore, the third property, the distribution shift should happen in the month of November and January where the peak of water level in 2007-2013 is shifted from November to Januari in 2013-2017. As the comparison conclusion, we propose the use of DBScan and K-Means when the distribution's gap is negligible while the K-Medoids as a more general means to cluster data with unknown or equally sparse distribution.

Based on the clustering results on both scenarios and all methods, we find several interesting patterns. First, the highest average water level is always on February and diminished on March. This suggests an important period of diminishing water level pattern on the start of the year. Second, there are differences in the clustering results on the scenarios where the first scenario put November as the three largest clusters and the second scenario put January as the three largest clusters. These results suggest a shift in the water level pattern from starting to increase on November (and start to diminish on March) to January. This shift in pattern provides a clue that the important period of flood has shifted from November through March to January through March in the last five years.

Third, the average water level is constantly reduced in the past five years compares to the last 11 years as shown in Figure 7. Fourth, the function from the second scenario is better fit than the first (in spite of both have $R^2 = 1$). This function signifies a continuous slope which might be usable for predicting the water level between January to March period. And finally, fifth, based on the time, the three highest average water levels are on 12:00 PM, 6:00 AM, and 6:00 PM respectively which suggest a high flood risks on these time frame.

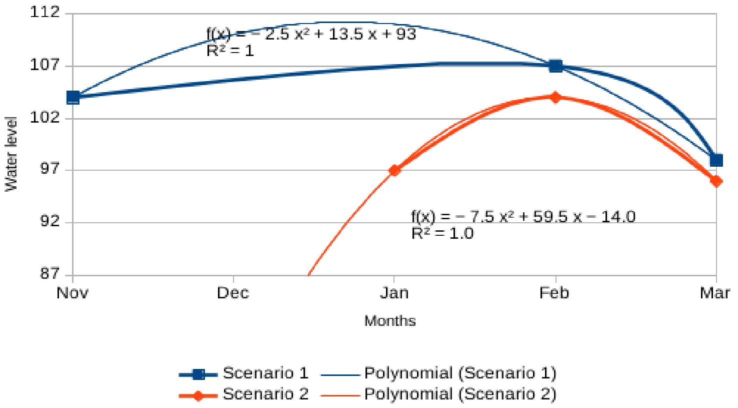

Figure 7. Water level functions on the three largest clusters

However, we found a warning that may invalidate these results that these patterns occur on limited clustering method parameters. Although we try to make as much generalization as possible by using three different clustering methods on several parameters, high value parameters (e.g. high value of k and low member value in DBScan method) might suggest different patterns. Therefore, further investigation should be done more thoroughly before implementing the findings into the real flood early warning system.

5. Conclusions

In this study, the data of water level in Tangerang for 2007 to 2017 is clustered by using K-Medoids, DBScan, and x-means clustering methods. The data are cleaned in the preprocessing stage. The data is then experimented on two scenarios based on hourly appearance. In the data mining process, clustering is done using three methods. The clustering result for K-Medoids is 3 clusters, DBScan is 2 clusters, and x-means is 3 clusters. Based on the results, the K-Medoids and x-means clustering appear better since the member of each cluster is more evenly distributed. The methods performed on scenario 1 appear better due to larger data availability; nevertheless the data from scenario 2 have a better predictive function. The cluster results show almost a complete convergence where they suggest a high potential flood risk on the first two months of the year. These results are visualized in form of interactive dashboard that is simple and easy to use for non-technical users. As final conclusion, we believe that the results show as an interesting potential for a flood early warning system in Tangerang.

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.

References

[1] BNPB. Indonesian Disaster Data and Information, 2017 (In Indonesia: Data dan Informasi Bencana Indonesia, 2017). Available: [URL] dibi.bnpb.go.id. [Accessed 13 Februari 2017].
[2] Supangat K, Soelistio YE. *Bus Stops Location and Bus Route Planning Using Mean Shift Clustering and Ant Colony in West Jakarta*. IOP Conference Series: Materials Science and Engineering. Bali. Indonesia. 2017; 185(1).
[3] Maaten LVD, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*. 2008: 2579-2605.
[4] LeCun Y, Bengio Y, Hinton, G. Deep learning. *Nature*. 2015:436.
[5] Gupta S, Kumar R, Lu K, Moseley B, Vassilvitskii, S. *Local search methods for K-Means with outliers*. Proceedings of the VLDB Endowment. 2017:757-768.

[6] Julian LR, Natalia F. *The Use of Web Scraping in Computer Parts and Assembly Price Comparison.* International Conference on New Media (CONMEDIA). Tangerang. Indonesia. 2015: 8-13.

[7] Monica S, Ferdinand FN, Sudirman S. *Clustering Tourim Object in Bali Province Using K-Means and X-Means Clustering Algorithm.* The 16th IEEE International Conference on Smart City (SmartCity). 2018.

[8] Ferdinand FN, Soelistio Y, Murwantara IM, Ko CS. *Clustering The Water Level Patterns in Ta'ngerang Using Three Clustering Algorithms Period 2013-2017*. The 13th International Conference on Innovative Computing, Information and Control (ICICIC2018). 2018 [in press]

[9] Bhat A. K-Medoids Clustering Using Partitioning Around Medoids for Performing Face Recognition. *International Journal of Soft Computing, Mathematics and Control (IJSCMC)*. 2014: 3(20: 1-12

[10] Sengottuvelan P, Gopalakrishnan T. Efficient Web Usage Mining Based on K-Medoids Clustering Technique. *World Academy of Acience, Engineering and Technology International Journal of Computer and Information Engineering*. 2015: 9(4): 1044-1048.

[11] Agarwal S, Mehta S. Approximate shortest distance computing using K-Medoids clustering. *Annals of Data Science*. 2017: 547-564.

[12] Yu D, Liu G, Guo M, Liu X. An improved K-medoids algorithm based on step increasing and optimizing medoids. *Expert Systems with Applications*. 2018: 92(C): 464-473.

[13] Newling J, Fleuret F. K-Medoids for K-Means seeding. *Advances in Neural Information Processing Systems*. 2017: 5195-5203.

[14] Kim TY, Kim, S Kim, JA, Choi JY, Lee JH, Cho Y, Nam YK. Automatic identification of Java Method Naming Patterns Using Cascade K-Medoids. *KSII Transactions on Internet & Information Systems*. 2018; 10(2): 873-891.

[15] Raj C. Comparison of K-Means, K-Medoids, DBScan Algortihms using DNA Microarray Dataset. *International Journal of Computational and Applied Mathematics*. 2017: 12(1): 1819-4966.

[16] Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBScan revisited, revisited: why and how you should (still) use DBScan. ACM Transactions on Database Systems (TODS). 2017; 42(3): 19.

[17] Chen Y, Tang S, Bouguila N, Wang C, Du J, Li H. A Fast Clustering Algorithm based on pruning unnecessary distance computations in DBScan for High-Dimensional Data. *Pattern Recognition*. 2018.

[18] Sharma S, Sharma AK, Soni D. Enhancing DBScan algorithm for data mining. In 2017 International Conference on Energy, Communication. *Data Analytics and Soft Computing (ICECDS)*. 2017: 1634-1638.

[19] Ozkok FO, Celik M. A New Approach to Determine Eps Parameter of DBScan Algorithm. *International Journal of Intelligent Systems and Applications in Engineering*, 2017: 247-251.

[20] Gan J, Tao Y. On the hardness and approximation of Euclidean DBScan. *ACM Transactions on Database Systems (TODS)*. 2016: 1(1): 1-44.

[21] Nguyen NT, Kowalczyk R, Orłowski C, Ziółkowski A. Transactions on Computational Collective Intelligence XXV. Berlin. 2016.

[22] Kettani O, Ramdani F. A Parameter Free Clustering Algorithm. *International Journal of Computer Applications*. 2017; 164(1): 34-39.

[23] Bogucharskiy S, Mashtalir V. *Image segmentation via X-means under overlapping classes*. Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). 2015: 45-47.

[24] Kapil S, Chawla M, Ansari MD. On K-Means data clustering algorithm with genetic algorithm. *Parallel, Distributed and Grid Computing (PDGC)*. 2016: 202-206.

[25] Pelleg D, Moore AW. X-means: Extending K-Means with efficient estimation of the number of clusters. 2000: 727-734.

[26] Hamerly G, Elkan C. Learning the k in K-Means. *Advances in neural information processing systems*. 2004: 281-288.