Novelties in Intelligent Digital Systems
C. Frasson et al. (Eds.)

doi:10.3233/FAIA210091

A Fuzzy C-Means-Based Algorithm for the Surveillance of Dengue Cases Distribution in Local Communities

Jozelle C. ADDAWE[a,1], Jaime D.L. CARO[a] and Richelle Ann B. JUAYONG[a]
[a] *Service Science and Software Engineering Laboratory, University of the Philippines Diliman , Quezon City Philippines*

Abstract. The analysis of disease occurrence over the smallest unit of a region is critical in designing data-driven and targeted intervention plans to reduce health impacts in the population and prevent spread of disease. This study aims to characterize groups of local communities that exhibit the same temporal patterns in dengue occurrence using the Fuzzy C-means (FCM) algorithm for clustering spatiotemporal data and investigate its performance in clustering data on dengue cases aggregated yearly, monthly and weekly. In particular, this study investigates similar patterns of Dengue cases in 129 barangays of Baguio City, Philippines recorded over a period of 9 years. Results have shown that the FCM has promising results in grouping together time series data of *barangays* when using data that is aggregated weekly.

Keywords. data mining, dengue, fuzzy c-means algorithm, time series clustering, disease surveillance

1. Introduction

Health systems around the world are continuously challenged with the emergence and re-emergence of diseases that continue to threaten the lives of many people. In the Philippines, emerging and re-emerging infectious disease (EREIDs) such as leptospirosis, dengue, meningococcemia, tuberculosis among others, continues to threaten the health system.

While the emergence of any disease is unpredictable, its health impact can be managed through planning and execution of proper interventions and proper surveillance [1]. In 2019, the Philippines suffered public health crises due to measles, dengue and polio outbreaks. Dengue cases during this time was noted to be the worst in the decade after recording more than 450,000 cases with about 1,500 deaths [2].

In particular, it is important to consider strengthening health surveillance systems even at the most local communities, commonly called *barangays* in the Philippines, in order to prevent major health crises or disease outbreaks at the national level. This is because health intervention planning is unique to an area or population due to regional differences in terms of available resources, health personnel, environmental conditions, and accessibility [3].

[1] Corresponding Author: Jozelle C. Addawe, Service Science and Software Engineering Laboratory, University of the Philippines Diliman, Quezon City Philippines E-mail: jcaddawe@up.edu.ph

The unpredictability of disease outbreaks, the sporadic growth of dengue, the increase in the number of deaths and strengthening surveillance in localized areas are the primary motivations of this research.

This study explores disease pattern at the barangay level by grouping together barangays that exhibit similar time series patterns and identify shared characteristics of the barangays belonging to the same group using a data clustering based approach. Furthermore, this research investigates how data granularity may affect the findings. This approach in analyzing disease data discovers patterns in data while considering the geographic location of the barangays which might affect disease distribution.

Particularly, this paper investigates patterns of dengue cases recorded in 129 barangays of Baguio City collected over a 9-year period. Baguio City is a highly urbanized city and a tourist spot located in the northern part of the Philippines which was under close monitoring after exceeding the dengue alert threshold following the declaration of the dengue alert in the country in 2019. The city government intensified its dengue preventive drives after observing that more than half of the barangays have recorded cases of mosquito-borne diseases. According to the City Heath Office, as of the end of July 2019, there has been a 29.77% increase in the dengue cases in the city as compared to 2018 with one dengue related death. Despite not yet being declared for an outbreak, government officials are alarmed with the increase in the deaths due to dengue fever [4].

The next sections of the paper presents previous works, the fuzzy C-means (FCM) algorithm used in this paper, the methodology and the results and discussion of the simulations. The conclusions and recommendations of the research are presented thereafter.

2. Previous Work

Dengue is a fast spreading vector-borne disease transmitted by Aedes female mosquitoes. Dengue fever thrives and spreads faster in sub-tropical and tropical regions such as the Philippines due to the mosquitoes' dependence on the ecological conditions present in these areas that support their lifecycle. Survival of mosquitoes highly depend on water, temperature, precipitation, human habitation, vegetation cover. Moreover, dengue fever often occurs with rapid urbanization [5].

Data on number of disease cases collected over time and space is a spatiotemporal type of data wherein each sample in the data is composed of a spatial and the temporal component. In epidemiology, analysis of such data is useful in the exploration of the distribution of disease over a particular region over time. Various researches in diseases use spatial analysis to study disease distribution by means of disease mapping or hotspot analysis [6, 7, 8], or disease modelling to forecast disease cases. However, most studies recommend that other factors be considered to further explain disease distribution, such as socioeconomic factors or environmental factors including cleanliness or terrain [9].

Izakian [10, 11] presented a fuzzy clustering method to group together subsequences of time series data within time windows to reveal patterns in historical data. The Fuzzy C-means is used to cluster spatiotemporal data based on the following objective function:

$$O = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d^2(v_i, x_k) \tag{1}$$

where $U = U_{ik} \in [0,1]$ is a fuzzy partition matrix containing the degree of membership of the k^{th} data point to an i^{th} cluster and $\mathrm{m} > 1$ is the fuzzification coefficient. The distance between the i^{th} cluster, v_i and the k^{th} data point, $\mathrm{x_k}$, d is computed by the Euclidean distance between these points [10].

$$d_\lambda^2 = (v_i, x_k) = ||v_i(s) - x_k(s)||^2 + \lambda||v_i(t) - x_k(t)||^2 \quad (2)$$

In this equation, the first term and second term is the distance between the spatial components and the temporal components, respectively and $\lambda \geq 0$ is a factor that controls the effect of the temporal component to the overall distance, *d*. In this paper, the spatial component of the data is the location of the barangay, the temporal component is the dengue cases over time and $\lambda = 0.1$ is used in all experiments. Data is assigned to the cluster with the highest degree of membership. The FCM uses the same randomly generated partition matrix for partitioning n barangays into c clusters, regardless of the data granularity. Reconstruction error is used to evaluate the results. This error is the difference between the original data x_k and the reconstructed data, $\hat{\mathrm{x}}_\mathrm{k}$ obtained from the FCM clustering result computed as [10]:

$$\hat{x}_k = \frac{\sum_{i=1}^{c} u_{ik}^m v_i}{\sum_{i=1}^{c} u_{ik}^m} \quad (3)$$

3. Results and Discussion

The fuzzy c-means clustering was implemented on time series data of dengue case records from 129 barangays of Baguio City obtained from the City Health Office. Results from initial implementations of the FCM using all barangays had low clustering performance which could have been affected by outliers in data, such as those barangays that have very low or zero dengue cases. It was used to cluster all 129 barangays (Case A), top 14 barangays having high dengue cases yearly (Case B) and top 19 barangays with high dengue incidence rates yearly (Case C) in order to extract relevant and information specific to a set of barangays. Dengue incidence rate is the number of dengue cases by the average population of a barangay.

The FCM was performed on time series data of dengue cases aggregated yearly, monthly or weekly for each barangay for arbitrarily chosen number of clusters *c = 3, 4, 5* and *6*.

The reconstruction errors of the FCM on yearly data aggregated weekly for Case B and Case C are summarized in Table 1 where shaded cells correspond to the least reconstruction errors. Observe that the optimal number of clusters of barangays is different each year with very small differences. For simplicity, *c=5* clusters is chosen for Case B while *c=3* clusters is used for Case C. The reconstruction errors for Case A is not shown but the optimal number of clusters is *c = 5* for all years.

From these reconstruction results, the FCM was used to group the barangays in Case A, Case B and Case C using *c=5*, *c=5* and *c=3* clusters, respectively. Table 2 summarizes the reconstruction errors of the FCM in clustering the 9-year time series data of dengue cases aggregated weekly, monthly and weekly. Results show that using weekly aggregated data result to a lower reconstruction error although this is not the case for Case B.

Table 1. Reconstruction Errors of the FCM using weekly aggregated data

	Barangays with Top Cases				Barangays with Top Incidence			
YEAR	3 Clusters	4 Clusters	5 Clusters	6 Clusters	3 Clusters	4 Clusters	5 Clusters	6 Clusters
2010	10.030	9.818	9.727	9.936	10.118	10.118	10.118	10.118
2011	11.056	11.044	10.972	10.922	10.381	10.382	10.381	10.381
2012	9.968	9.968	9.926	9.876	12.086	12.086	12.086	12.086
2013	12.850	12.741	12.381	12.261	14.504	14.555	14.594	14.543
2014	11.278	11.192	11.050	11.182	10.527	10.542	10.542	10.542
2015	12.595	12.467	12.054	12.482	13.332	13.329	13.367	13.316
2016	12.670	12.510	12.434	12.516	16.833	16.923	16.989	16.917
2017	12.306	12.251	12.191	12.186	9.254	9.254	9.254	9.254
2018	11.091	11.086	10.976	11.037	9.704	9.705	9.706	9.706

Table 2. Reconstruction Errors using data aggregated yearly, monthly and weekly

	Yearly Aggregated	Monthly Aggregated	Weekly Aggregated
All Barangays (Case A)	136.160	132.72	43.868
Top Barangays (Case B)	11.580	11.612	12.270
Top Incidence Barangays (Case C)	19.959	20.013	16.791

Table 3. Hamming Distance between cluster assignments of barangays

	Case A	Case B	Case C
Yearly vs. Monthly	$D = 20$	(33333232221122) vs. (33333232221122) $D = 0$	(0011021010100011110) vs. (0011001010100011110) $D = 1$
Yearly vs. Weekly	$D = 25$	(33333232221122)vs. (33333232221122) $D = 0$	(0011021010100011110) vs. (0212002010100011010) $D = 5$
Weekly vs. Monthly	$D = 25$	$D = 0$	$D = 5$

Apart from the reconstruction error, a comparison of the cluster assignments of barangays obtained from the FCM using different data aggregations to tell more about the performance of the clustering approach. This was done by comparing the Hamming distances (D) between the cluster assignment of the barangays using different data aggregations. Results are summarized in Table 3, where the sequence of numbers enclosed in parentheses are the cluster number assignments of each barangay considered. Results show that the FCM results using yearly and monthly aggregated data generally partitions the barangays in the same clusters with very few differences. Furthermore, the FCM assigns Case B barangays to the same clusters when using different data aggregations but generally, a lower reconstruction error is obtained when using weekly aggregated data.

The FCM was tested in clustering barangays in Case B and Case C for different number of clusters and lambda values. Lower reconstruction errors were obtained with input c = 6 clusters and $\lambda = 0.04$ for Case B and c = 3 and $\lambda = 0.04$ for Case C.

The visualizations of the time series plots show similarities in the pattern of dengue cases. Figure 1 shows that barangays in Cluster 3 have relatively higher cases. Figure 2 shows that high incidence barangays belong to Cluster 0 while those with relatively the lowest dengue incidence belong to Cluster 1. Both results however show that there are a few barangays that are outliers of a cluster.

The barangays per cluster are mapped in Figure 3 to show the locations of these barangays within the city. Observe that barangays belonging to the same cluster

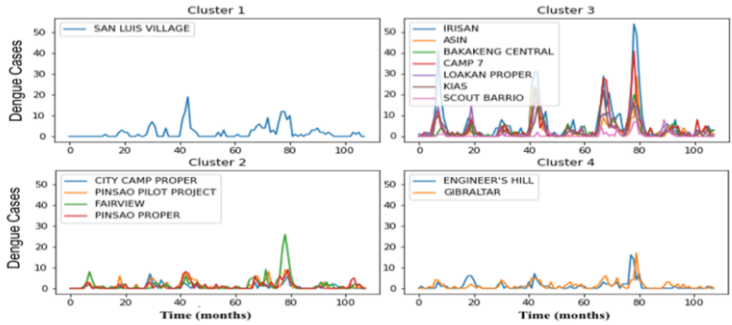
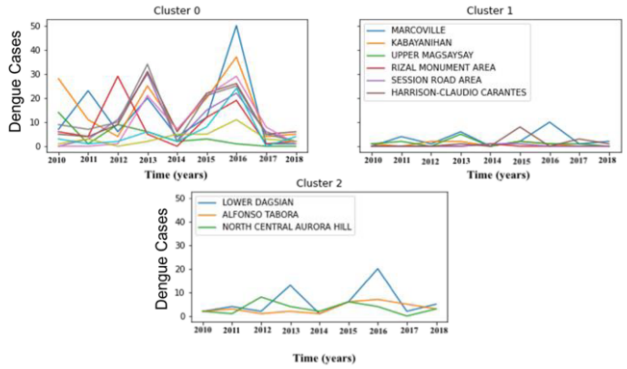


Figure 1. Cluster Assignments of Barangays with Top Dengue Cases



F**igure 2.** Cluster Assignments of Barangays with Top Dengue Incidence Rates

generally share the same boundaries. The FCM suggests that the barangays with top cases belonging to Cluster 3 are found in the western and southeastern part of the region. Barangays with top cases yearly relatively have high land area while those with high incidence areas are clustered closer to the city center with relatively smaller land areas. Moreover, top incidence barangays belonging to Cluster 1 are nearer to the central district while those in cluster 0 seem to be just around the center as shown in the map.

4. Conclusions and Future Works

This paper has shown the performance of an existing fuzzy c-means based approach in grouping together barangays that exhibit similar patterns of dengue cases. Results have shown that weekly aggregated data distinguishes similarities in time series data with lower reconstruction error. Further studies following the current findings may be to discover how cluster memberships of barangays change over the years or per quarter to observe significant disease pattern changes during a certain period. Necessary adjustments such as data transformation and proper scaling of the spatial and temporal components of the data is recommended.

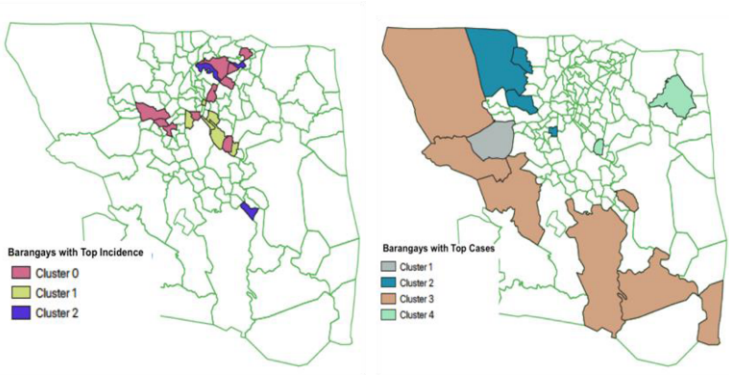


Figure 3. Cluster Assignments of Barangays

Acknowledgements

The authors would like to acknowledge the DOST - ERDT for funding the publication of this paper. The authors also thank the Baguio City Epidemiology and Surveillance Unit for the raw data on dengue cases.

References

[1] (WHO) WHO. Public Health Surveillance;. Available from: https://www.who.int/immunization/monitoring_surveillance/burden/vpd/en/.
[2] Santos AP. Philippines struggling to cope with back-to-back disease outbreaks; 2019. Available from: https://www.dw.com/en/philippines-struggling-to-cope-with-back-to-back-disease-outbreaks/a-51084152.
[3] Esparagoza C. Challenges in the Philippine Healthcare System: Social Determinants to health, Health system strengthening and Health engagement in development towards issues in equity. 2020 05.
[4] Rizavel C Addawe KSC Louie Ville A Balino. Spatio-temporal pattern distribution of dengue infections in baguio city. In: A Preprint; 2019. .
[5] Walsh M. Dengue Part 2: The Mosquito and its ecology'; 2011. Available from: http://www.infectionlandscapes.org/2011/01/dengue-part-2-mosquito-and-its-ecology.html.
[6] Pangilinan MAP, Gonzales DPG, Leong RNF, Co F. Spatial analysis of the distribution of reported dengue incidence in the National Capital Region, Philippines. Acta medica Philippina. 2017;51(2):126-32.
[7] Mutheneni SR, Mopuri R, Naish S, Gunti D, Upadhyayula SM. Spatial distribution and cluster analysis of dengue using self-organizing maps in Andhra Pradesh, India, 2011–2013. Parasite epidemiology and control. 2018;3(1):52-61.
[8] Ansari MY, Ahmad A, Khan SS, Bhushan G, et al. Spatiotemporal clustering: a review. Artificial Intelligence Review. 2019:1-43.
[9] Datoc HI, Caparas R, Caro J. Forecasting and data visualization of dengue spread in the Philippine Visayas island group. In: 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE; 2016. p. 1-4.
[10] Izakian H, Pedrycz W, Jamal I. Clustering spatiotemporal data: An augmented fuzzy c-means. IEEE transactions on fuzzy systems. 2012;21(5):855-68.
[11] Izakian H, Pedrycz W. Anomaly detection and characterization in spatial time series data: A cluster-centric approach. IEEE Transactions on Fuzzy Systems. 2014;22(6):1612-24.