GEOSPATIAL AND CLUSTERING ANALYSIS OF DENGUE CASES USING SELF-ORGANIZING MAPS: CASE OF QUEZON CITY, 2010 - 2015

J. J. Valles [1*], C. Perez [1], A. C. Blanco [1,2]

[1] Dept. of Geodetic Engineering, University of the Philippines – Diliman, Quezon City, Philippines
(jyvalles1@up.edu.ph, crperez@up.edu.ph, acblanco@up.edu.ph)
[2]Training Center for Applied Geodesy and Photogrammetry, University of the Philippines – Diliman, Quezon City, Philippines

Commission IV

KEY WORDS: Dengue, Self-Organizing Map, Clustering, OLS Regression

ABSTRACT:

Dengue is the most rapidly spreading disease in the world with more than 30% of the world's population at risk of contracting dengue. In 2016, more than 375,000 suspected cases of dengue were reported from the Western Pacific Region, and more than half of these were reported by the Philippines. Dengue virus inflicts significant health and economic burden to the Philippines. Thus, it is important to improve the country's current schemes for dengue surveillance and response thru better understanding and knowledge on the development of dengue. In this research, geospatial and clustering analyses of dengue cases in Quezon City through GIS and self-organizing maps (SOM) were performed. Two clusters were generated for each clustering method. After clustering the barangays, the coefficient of determination increased for most scenarios compared to the OLS regression of the ungrouped data. The R^2 values for the regression of whole Quezon City dataset ranged from 0.364 to 0.671, while it ranged from 0.468 to 0.839 for the SOM-clustered dataset. On the other hand, for the k-means-clustered dataset, R^2 values ranged from 0.395 to 0.945. Moreover, GWR models' adjusted R^2 values ranged from 0.675 to 0.876. Common predictors among the different regression models are the informal settlements and very low residential areas. Based on the significant predictors identified and the trend of the dengue cases, SOM produced more logical classification than the GIS Grouping Analysis. Although SOM takes a longer time compared to the GIS Grouping Analysis, SOM is easier and simpler to implement.

1. INTRODUCTION

1.1 Dengue Disease

Dengue is one of the major problems of tropical and sub-tropical regions of the world, including the Philippines. It is rapidly spreading with a dramatic increase in its incidence of almost 30 times. The US Centers for Disease Control reported that the estimated dengue cases worldwide each year is 50 to 100 million with more than 2.5 billion people in 100 countries living under the threat of dengue infection (Centers for Disease Control, 2019). Despite the alarming statistics and trends of the disease, it has been considered as one of the neglected tropical diseases with few joint and coordinated efforts from the national and international scene (World Health Organization, 2012). Dengue not only inflicts a significant health burden to the Philippines but it is also affecting and burdening the country's economy (Edillo, 2015).

Dengue, an infection caused by a virus (DENV), is the most common arthropod-borne viral (arboviral) illness in humans (Smith, 2019). DENV is carried by infected mosquitoes, specifically the *Aedes aegypti* and the female *Aedes albopictus*. The feeding time of these mosquitoes is usually during the daytime. Mosquitoes breed in stagnant, standing fresh water like puddles, oil tires, and water containers, thus, a neighborhood without consistent garbage collection has a greater chance of having more mosquitoes. Dengue has no specific antiviral treatment; however, it can be managed early and be prevented by eliminating places where mosquitoes can breed (Unilab, 2018).

Based on the combined reports available from the Philippines' Department of Health website, the National Capital Region (NCR) has one of the highest reported dengue cases (28,040) in 2018. Within NCR, Quezon City reported 9,114 dengue cases according to Metro Manila Center for Health Development. The existing integrated vector management initiatives of the city were implemented rigorously over the past years helped in mitigating the cases of dengue. Despite the local government's increasing efforts, unpredictable trends of dengue cases are happening. The success of dengue prevention and mitigation programs is determined by the proper understanding of the evolution and trend of dengue.

1.2 Spatial and Clustering Analysis in Epidemiology

Clustering analyses are performed to analyze a phenomenon at a more precise level. Spatial-temporal clustering is a method of grouping objects based on their spatial and temporal likeness. It is widely used in identifying disease distribution patterns, locating areas with active disease transmission, and evaluating the relationship between disease incidence and different factors (Xu et al., 2012). Spatial-temporal analysis of infectious diseases is widely used in understanding their development, transmission, spread, and dynamics for disease control and prevention strategies. One of the spatial analysis tools commonly used is hot spot analysis that is used to identify geographic clusters of disease and predict areas with a high risk of disease transmission (Sun et al., 2017). Therefore, spatial hotspot analysis and spatial-temporal clustering analysis are important tools in disease surveillance and spatial-temporal epidemiology.

This research discusses the use of geospatial and clustering analysis in understanding the incidence of dengue in Quezon City by determining significant predictors of dengue incidences in Quezon City among the candidate explanatory variables: demographic, land use and environmental. Moreover, the self-organizing map is introduced as a tool in the clustering analysis of dengue incidences in Quezon City.

2. SELF-ORGANIZING MAP

The Self-Organizing Map (SOM) was introduced by Kohonen (1982) in a theoretical study of self-organization of a low-dimensional output space induced by high-dimensional input space. The SOM is an effective tool for dimensionality reduction while preserving the important topological characteristics of the input space. Similar vectors in the input space appear to be neighbors when mapped into the output feature map, using certain distance metrics such as the Euclidean distance or the dot product. It is a neural network mimicking the brain, where a stimulus is assigned to a specific region for processing. The neural network has fully connected neurons that are not connected by weight vectors, but by adjacency.

The basic algorithm is initialization, competition, and adjustment. In the initialization step, a lattice of certain size *m x n* is created with each node containing a weight vector equal in length as to that of the input vector. Initially, the weight vectors are just random values. The lattice can be arranged in different ways, with the rectangular and hexagonal ones being very common. The difference with these configurations is the number of neighbors of each node. In a rectangular arrangement, there are four (4) neighbors; while there are six (6) in the hexagonal pattern. The second step randomly selects an input vector and compares it to every node of the map using a certain distance metric. The distance metric can be a dot product or more commonly, the Euclidean distance. The algorithm selects the nearest node or most similar node to the input vector and assigns that node as the best matching unit (BMU), i.e., the winning node. In the process of adjustment, that node and its neighbors within a pre-set radius will be adjusted using an equation that makes their weights even closer to the input vector.

Different clustering techniques are available for different kinds of data. Common research on spatial epidemiology uses the self-organizing maps in conjunction with GIS for their analysis (Mutheneni et al., 2018; Zhang, J., Shi, H. & Zhang, Y., 2009; Basara & Yuan, 2008). Basara & Yuan's (2008) research results indicated that the variability between community clusters was significant with respect to the spatial distribution of disease occurrence. Moreover, clustering the SOM performed better than direct clustering of input data using k-means and partitive clustering (Alhoniemi, E. & Vesanto, J., 2000).

3. MATERIALS AND METHODS

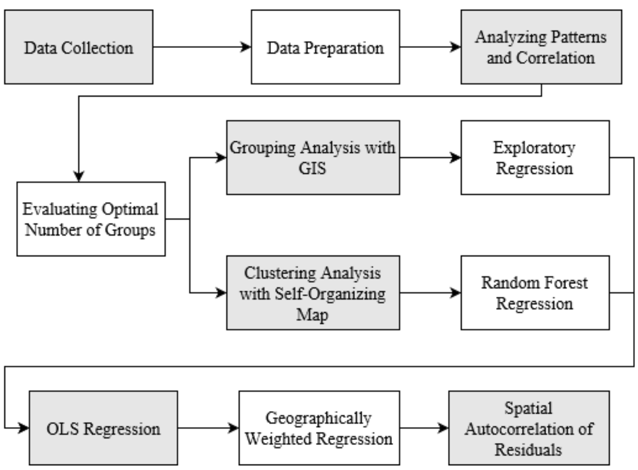


Figure 1. General workflow of methodology

The general workflow of this research is shown in Figure 1. Data Collection is the first procedure that involves gathering all the necessary data for the study. Table 1 shows all the data used in this study. The next procedure, Data Preparation, is done to increase the accuracy and consistency of data, which consists of data cleansing, feature selection, and data representation. Data cleansing consists of finding and removing incomplete, incorrect, and inconsistent data. Feature selection is done to improve the accuracy of model creation by removing factors that are not correlated with the dependent variable. Data representation is transforming data into a different form to enable applications to access and analyze data more accurately and effectively. Analyzing Patterns and Correlations consists of determining the relationship of each factor to the reported dengue cases and identifying which of the factors demonstrate a spatial pattern with respect to the spatial distribution of the reported dengue cases. The optimal number of groups is then evaluated using the Grouping Analysis tool of ArcMap 10.3.

The next phase of methodology will identify clusters of barangays. The Grouping Analysis tool and SOM are utilized to generate clusters of barangays based on the reported dengue cases per month. SOM was implemented on the Python programming language using the Numpy, Matplotlib, and Pandas as the main libraries. Significant variables that affect dengue incidence were identified using the Exploratory Regression tool and Random Forest Regression. The Random Forest algorithm was executed using Scikit-Learn's implementation on Python. Ordinary Least Squares (OLS) regression was then used to generate a model and determine its statistical strength. If non-stationarity exists, Geographically Weighted Regression (GWR) is carried out. The Spatial Autocorrelation tool was used to determine if the residuals' pattern is random.

Data	Source
Quezon City Land use Map	Quezon City Planning and Development Office
Quezon City Demographics	Philippine Statistics Authority
Rainfall	DOST - ASTI DOST – PAGASA
Land surface temperature	MODIS - Moderate Resolution Imaging Spectroradiometer
Monthly dengue incidence report	Quezon City Health Department (QCHD)

Table 1. List of all data gathered and their sources.

4. RESULTS

4.1 Pattern and Correlation Analysis

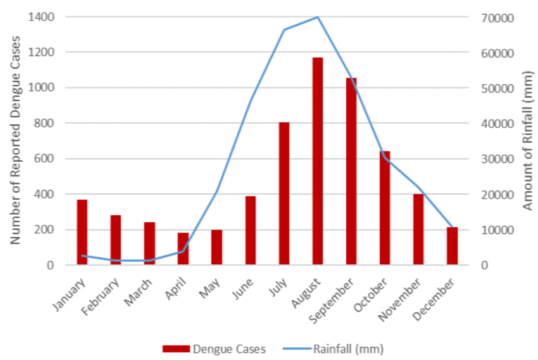


Figure 2. Quezon City's average monthly reported dengue cases and amount of rainfall from 2010 to 2015

https://doi.org/10.5194/isprs-archives-XLII-4-W19-455-2019 |

The total number of reported dengue cases per year were mapped and their spatial distribution were compared with that of dengue hotspots per year. Figure 3 shows the dengue hotspots per year for the period 2010 - 2015. It can be observed that the northern part of Quezon City is the usual location of the dengue hotspots while cold spots are located in the southern part of the city.

A large area of Quezon City is medium density residential and low-density residential areas. Since it can be found in almost all parts of Quezon City, both land use classes could have low significance as to the characteristic spatial distribution of dengue incidences. Many informal settlements, on the other hand, were found to be concentrated at the northeast part of Quezon City where dengue hotspots are located. However, small areas of informal settlements can also be found in neutral and cold spot areas.

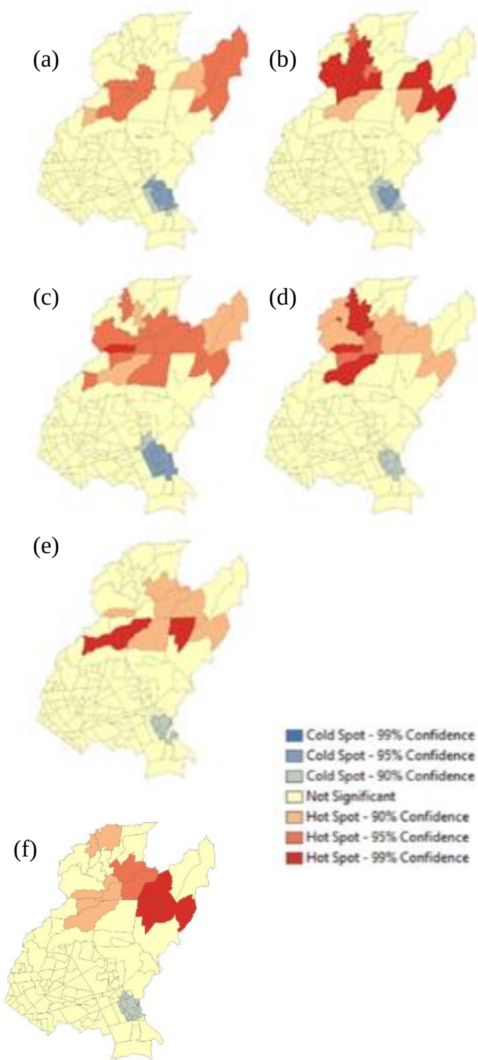


Figure 3. Dengue hotspots in Quezon City: (a) 2010, (b) 2011, (c) 2012, (d) 2013, (e) 2014, (f) 2015

Figure 4.a. shows the spatial trend of the average annual rainfall for the period 2010 - 2015. The rainfall and dengue hotspots shown in Figure 3 have a similar trend wherein the areas that experience heavier rainfall have the highest incidence of dengue compared to the areas that experience a smaller amount of rainfall. This conforms to the prior findings that rainfall amount is directly related to the incidence of dengue in Quezon City.

On the other hand, it can be observed in Figure 4.b that the land surface temperature (LST) exhibits a negative correlation with the incidence of dengue. Areas with a high incidence of dengue has lower LST while those with low incidence of dengue has a relatively higher LST. The hotspots and cold spots identified by the Optimized Hotspot Analysis tool also supported this observation.

As shown in Figure 4c, areas with higher elevations are found on the northeast side of Quezon City. An increase in altitude will result in a decrease in air temperature, thus, these areas were found to be not just experiencing a high amount of rainfall but also lower temperature. As shown in Figure 3, there are dengue hotspots in these areas.

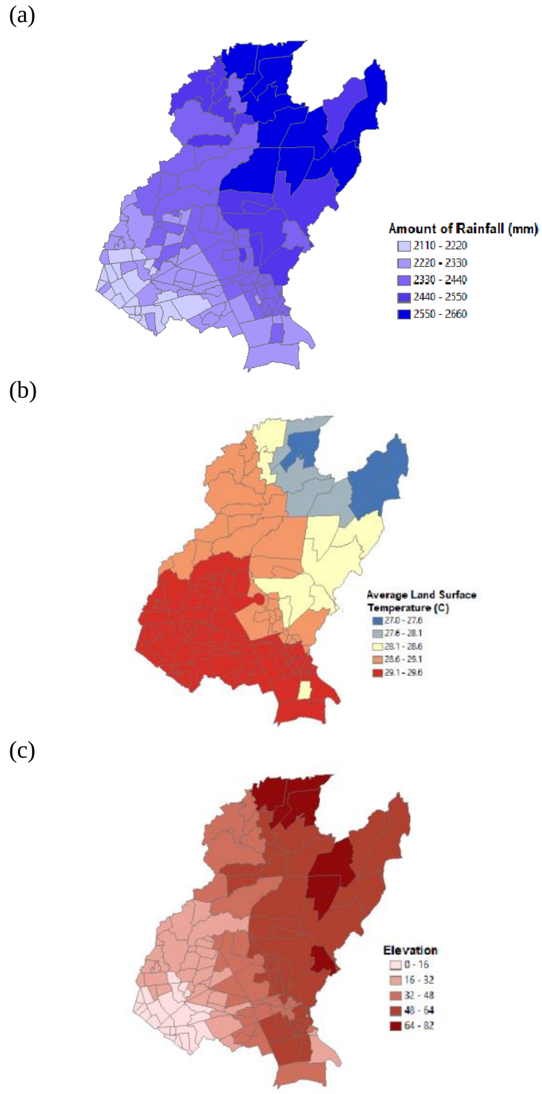


Figure 4. a) Six-year average rainfall and (b) six-year average land surface temperature (c) elevation of Quezon City from 2010 to 2015.

Population distribution and population density distribution of Quezon City from 2010 to 2015 was also observed. Comparing the spatial distribution of dengue, the population density shows a random distribution. On the other hand, barangays that are identified as dengue hotspots have high population and cold spots have low population. However, the barangays on the northwest part of Quezon City do not follow this trend. This could be because of the higher temperature experiencing by the areas in this location, thus, mosquitoes' growth is much slower and lower than areas located at the northeast.

4.2 Grouping Analysis with ArcMap

Using the monthly reported dengue cases, Quezon City was divided into two groups using the Grouping Analysis tool of ArcMap 10.3. As shown in Figure 5, Cluster 1 is displayed as the blue clusters while Cluster 2 as the red clusters. Cluster 2 has 16 to 23 barangays while Cluster 1 has 118 to 126 barangays. Cluster 2 can be observed to include barangays identified as dengue hotspots as shown in Figure 3. There are several barangays in Cluster 2 that can be found in the southwest portion of the Quezon City.

The number of monthly reported dengue cases of each barangay in each cluster were then plotted separately as shown in Figure 6. A low number of reported dengue cases in Cluster 1 can be observed ranging 0 – 20 reported dengue cases per month. On the other hand, Cluster 2 cases ranged from 2 to 56 reported dengue cases per month. Cluster 2 shows that the number of reported dengue cases mostly peaked in August. The incidence of dengue also peaked in August in some of the barangays in Cluster 1.

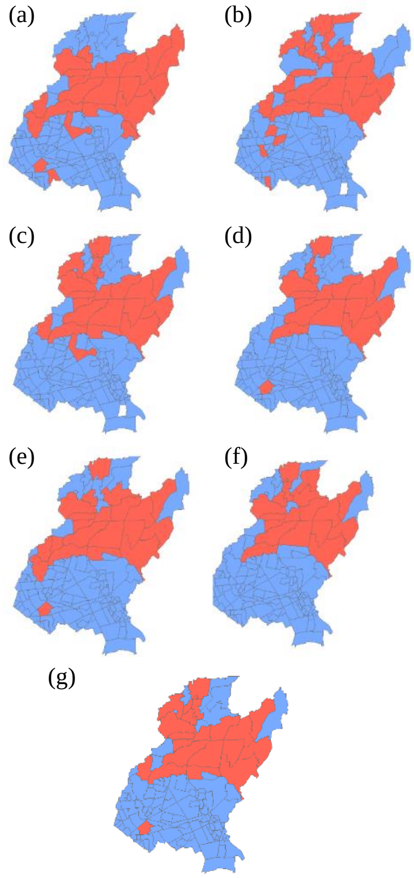


Figure 5. Clusters identified using ArcMap Grouping Analysis. (a) 2010 (b) 2011 (c) 2012 (d) 2013 (e) 2014 (f) 2015 (g) 2010 - 2015

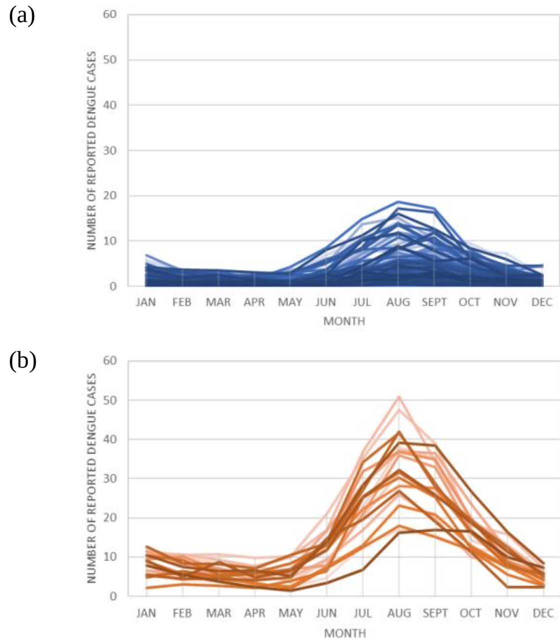


Figure 6. Monthly average reported dengue cases of each barangays in a) Cluster 1 b) Cluster 2

Exploratory data analysis was performed using the ArcMap Exploratory Regression tool to identify the variables with high significance to the incidence of dengue. The identified variables were then used in the OLS regression model. Table 2 shows the significant variables identified.

The amount of rainfall, land surface temperature, informal settlements, open spaces, and very low-density residential are the variables that appear the most in the analysis conducted considering entire city and Cluster A only. One possible cause may be due to the poor management of these spaces. Open spaces and very low-density residential may have open areas that are not managed properly; water can remain stagnant for days, making it more vulnerable to a rapid increase in the mosquito population. Due to their mobility, these mosquitoes may be able to transmit the virus to neighbouring high-density residential areas. The high correlation of informal settlement with dengue incidence is related to poor living conditions, improper waste disposal, inadequate drainage system, and poor water storage management which creates more suitable breeding places for mosquitoes.

Cluster 2 has varying significant predictors and are observed to be far different from the significant predictors identified on Cluster 1 and the whole Quezon City. The most common predictors identified on these clusters are Transport and Service Facilities and Informal Settlements. Transport and Service Facilities is an area designed for transport and service facilities where bus terminals.

Year	Quezon City	Cluster 1	Cluster 2
2010	Informal Settlements Open spaces Average LST	Open spaces Average LST Informal Settlements	Education Institution Total Rainfall
2011	Open spaces Informal Settlements Very Low DR Total Rainfall	Transport and Service Facilities Total Rainfall	Population Density Informal Settlements

https://doi.org/10.5194/isprs-archives-XLII-4-W19-455-2019 | © Authors 2019. CC BY 4.0 License.

2012	Informal Settlements Open spaces Elevation Very Low DR	Very Low DR Informal Settlements Open spaces	Religious and Cemetery Transport and Service Facilities
2013	Informal Settlements Total Rainfall Average LST Open spaces Elevation	Informal Settlements Open spaces Health & Welfare Total Rainfall Average LST Very Low DR	Medium DR Transport and Service Facilities
2014	Informal Settlements Total Rainfall Open spaces	Informal Settlements Open spaces Total Rainfall	Medium DR High DR Transport and Service Facilities
2015	Informal Settlements Open spaces Very Low DR Total Rainfall	Very Low DR Total Rainfall Open spaces Informal Settlements	Informal Settlements Elevation Total Rainfall
2010 - 2015	Total Rainfall Informal Settlements Open spaces Average LST	Open spaces Very Low DR Average LST Informal Settlements	Elevation Religious and Cemetery Informal Settlements

Table 2. Significant Variables of the Quezon City, Cluster 1 and Cluster 2 in each year and whole period identified by the Exploratory Regression tool, (*DR – Density Residential)

Each significant predictor identified was then used as explanatory variables for each OLS regression. Variables with high variance inflation factor (VIF > 7.5) were removed from the model to address the problem of multicollinearity. The total rainfall and average LST are the variables that exhibit multicollinearity and thus, OLS regression was performed multiple times to determine which variable is more fit to be used. Moreover, an increase in the coefficient of determination (R^2) can be observed in the OLS regression applied to clusters compared to the OLS regression of the whole dataset (see Table 3).

Year	Coefficient of Determination		
	All	Cluster 1	Cluster 2
2010	.670	.656	.941
2011	.573	.606	.854
2012	.685	.395	.337
2013	.612	.739	.817
2014	.606	.615	.779
2015	.744	.662	.942
2010-2015	.601	.701	.945

Table 3. Coefficient of determination for the whole Quezon City, Cluster 1 and Cluster 2 in each year and whole period

Year	Variables	R^2	
		Cluster 1	Cluster 2
2010	Informal Settlements Open spaces Average LST	.656	.935
2011	Open spaces Informal Settlements Very Low Density Residential Total Rainfall	.590	.899
2012	Informal Settlements Open spaces Elevation Very Low Density Residential	.565	.932
2013	Informal Settlements Average LST Open spaces Elevation	.701	.944
2014	Informal Settlements Total Rainfall Open spaces	.615	.899
2015	Informal Settlements Open spaces Very Low Density Residential Total Rainfall	.662	.947
2010 to 2015	Informal Settlements Open spaces Average LST	.627	.951

Table 4. Coefficient of Determination of Cluster 1 and Cluster 2 for each year and whole period using the significant predictors of the whole dataset of same year

Year	Coefficient of Determination		
	All	Cluster 1	Cluster 2
2010	.670	.656	.935
2011	.570	.569	.887
2012	.671	.574	.945
2013	.677	.677	.944
2014	.608	.608	.901
2015	.703	0.590	.945

Table 5. Coefficient of Determination of Cluster 1 and Cluster 2 for each year and whole period using the significant predictors of 2010 – 2015 whole dataset

Significant predictors of the whole dataset in each year were then used as the explanatory variables in each cluster of the same year. Table 4 shows that the R^2 values for Cluster 1 are not much different from the R^2 for the whole dataset. However, a significant increase in the R^2 in Cluster 2 could be seen. OLS regressions were then performed using the significant predictors identified as explanatory variables for the whole period of 2010-2015, namely, informal settlements, open space, and average land surface temperature. The R^2 for the whole Quezon City and Cluster 1 ranges from 0.569 to 0.703 while Cluster 2 showed higher R^2 ranging from 0.887 to 0.945 which indicates a strong correlation between the model and the data in Cluster 2, however, this does not mean that the model is already valid. The explanatory variables should be tested for significance.

GWR	Explanatory Variables	Adjusted R^2
2010	Informal Settlements, Open spaces	0.648
2011	Informal Settlements, Open spaces	0.623
2012	Informal Settlements, Open spaces	0.781
2013	Informal Settlements, Open spaces	0.659
2014	Informal Settlements, Open spaces	0.658
2015	Informal Settlements, Open spaces	0.868

Table 6. Explanatories used for the GWR model of each year and their respective overall adjusted R^2 values.

Geographically Weighted Regression (GWR) was performed for each year on the dengue cases as the dependent variable and the predictors determined by the Exploratory Regression tool and used in the OLS regression for the whole data as independent variables. However, when the GWR can't proceed due to local multicollinearity some variables were removed from the model. Table 6 shows the overall adjusted R^2 for each GWR model.

A map of the coefficients of each model was produced for the variables informal settlements and open spaces. As seen in the Figure 7, presence of informal settlements shows a negative relationship with the incidence of dengue on some barangays in the northwest and south of Quezon City from 2010 to 2015.

https://doi.org/10.5194/isprs-archives-XLII-4-W19-455-2019

However, most barangays showed positive association of informal settlements with the number of dengue incidences. Open spaces, on the other hand, as shown in Figure 8, has a more varying relationship with dengue incidence. Models of years 2011, 2012, and 2013 show that many barangays, which are mostly located at the west side of the Quezon City, are showing negative relationship with the incidence of dengue. But still most of the barangays showed positive association between informal settlements and the number of dengue incidences.

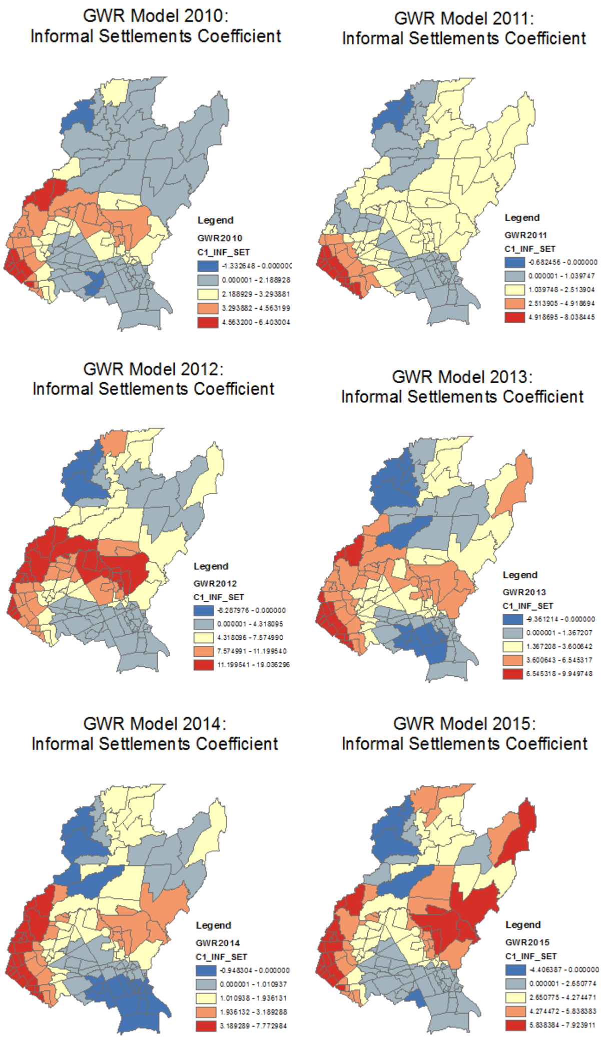
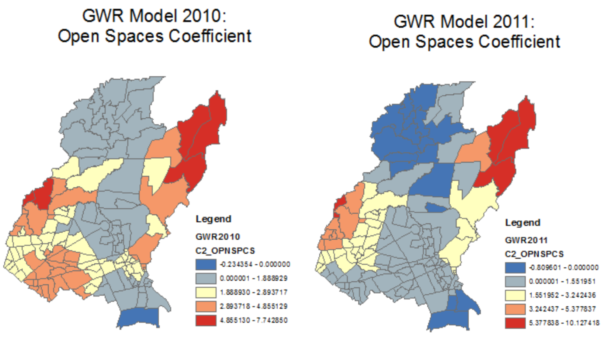


Figure 7. Map of the coefficients of informal settlements as predictor for each GWR model from 2010 to 2015.



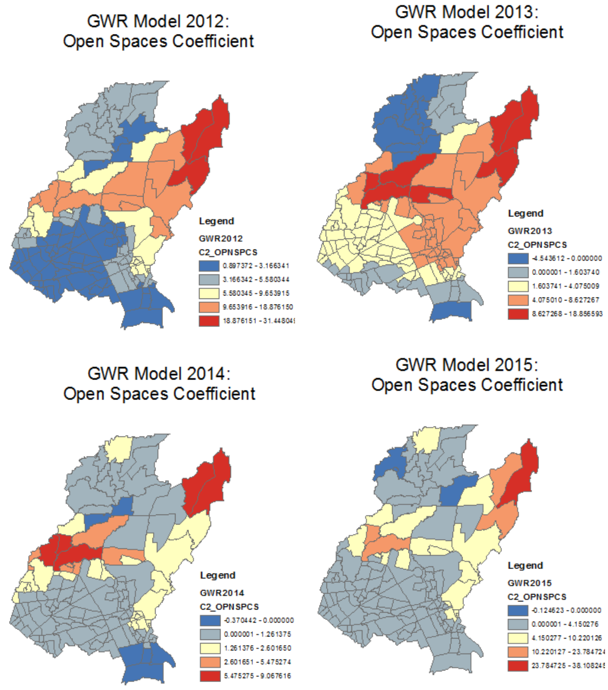


Figure 8. Map of the coefficients of open spaces as predictor for each GWR model from 2010 to 2015.

4.3 Clustering Analysis with Self-Organizing Map

The dot-product SOM was used to cluster the 142 barangays in Quezon City. The dengue incidence data from 2010 to 2015 were used to produce seven (7) SOMs, one for each year and another for the combined dataset 2010-2015. Figure 9 shows the u-matrix and cluster map for the combined dataset.

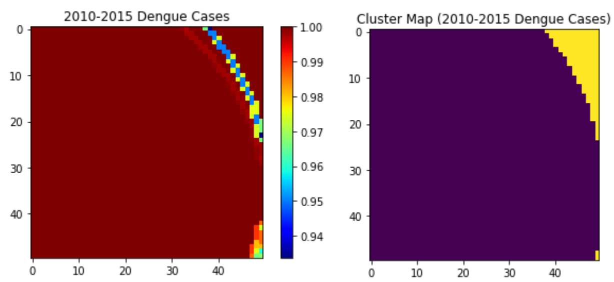


Figure 9. U-Matrix and cluster map for SOM 2010-2015.

The clusters were then mapped to the geographic space. Figure 10 shows the clusters in Quezon City. Using the combined data from 2010 to 2015 appeared to have divided Quezon City into two clusters, the north and south areas. The north area (red polygon) has barangays with a relatively higher number of dengue cases. On the other side, the south area (blue polygon) has barangays with a relatively lower number of dengue cases. The comparison between the dengue cases between the two clusters can be seen in the time-series plot in Figure 11.

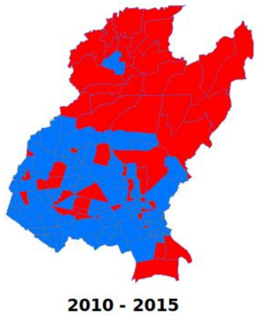


Figure 10. Geographic map of the clusters identified using SOM. Cluster 1 is red; Cluster 2 is blue.

It can be seen that the first cluster (red) has higher dengue cases compared to the second cluster (blue) (see Figure 11). Although there are more barangays in the second cluster, these barangays are commonly the small ones. This may contribute to the fact that these barangays have relatively lower dengue cases.

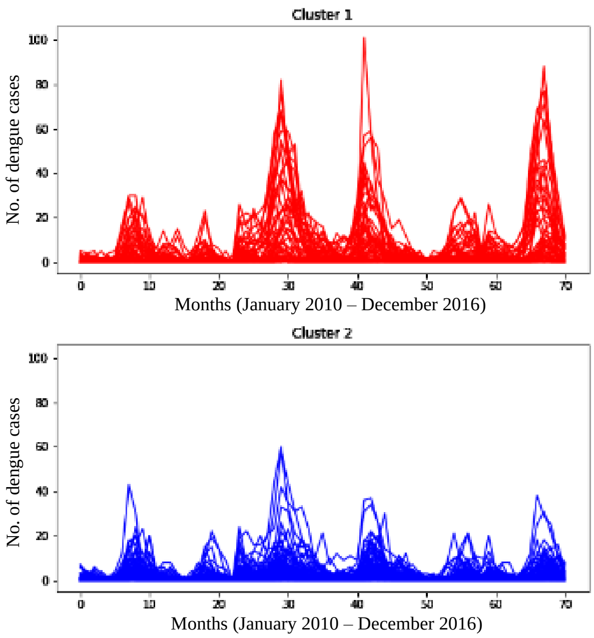


Figure 11. Time-series plot of dengue cases per cluster from the SOM 2010-2015.

Prior to ordinary least squares regression, random forest regression was performed in Python using the Scikit-Learn library primarily to get the variables with high importance in predicting dengue cases. These variables, shown in Table 7, were then used in the OLS regression model. The variance inflation factors were also calculated in Python using the Statsmodels library. Variables with VIF greater than 7.5 were removed further. Another set of OLS regression models were also generated for each cluster using the variables that are found to be significant for the whole dataset using random forest regression. Table 8 shows the results of the OLS regression for the whole dataset and each cluster. In most cases, it can be seen that the R^2 increased when the OLS regression was applied to clusters, compared to the OLS regression of the whole dataset.

Year	Quezon City	Cluster 1	Cluster 2
2010	Informal Settlements Very Low DR	Open spaces Informal Settlements	Open spaces Informal Settlements Very Low DR
2011	Informal Settlements Very Low DR	Informal Settlements Elevation Total Rainfall	Informal Settlements
2012	Informal Settlements Very Low DR	Very Low DR Informal Settlements Water Related	Informal Settlements Elevation Very Low DR
2013	Informal Settlements Average LST Commercial	Informal Settlements Average LST	Informal Settlements Average LST Very Low DR
2014	Informal Settlements Very Low DR	Informal Settlements Very Low DR	Informal Settlements Average LST
2015	Informal Settlements Very Low DR Elevation	Informal Settlements Elevation Commercial	Informal Settlements Very Low DR Total Rainfall
2010 - 2015	Informal Settlements Open spaces Very Low DR	Informal Settlements Open spaces Total Rainfall	Informal Settlements Commercial

Table 7. Significant Variables of the Quezon City, Cluster 1 and Cluster 2 in each year and whole period, (*DR – Density Residential)

Year	Coefficient of Determination		
	Quezon City	Cluster 1	Cluster 2
2010	0.369	0.468	0.588
2011	0.435	0.728	0.695
2012	0.432	0.290	0.839
2013	0.544	0.510	0.695
2014	0.364	0.501	0.560
2015	0.651	0.746	0.776
2010-2015	0.671	0.765	0.583

Table 8. Coefficient of determination from OLS regression for the whole Quezon City, Cluster 1 and Cluster 2 in each year and whole period

Among the significant predictors are informal settlements and very low residential areas, as can be observed in Table 9. One possible cause may be due to the poor management of these spaces. Last of all, OLS regressions were performed using the significant predictors identified as explanatory variables for the whole period of 2010-2015 which are informal settlements, very low density residential, and average land surface temperature. The R^2 for the whole Quezon ranges from 0.57 to 0.75 and Cluster 1 ranges from 0.417 to 0.73, while Cluster 2 showed a higher R^2, ranging from 0.58 to 0.90. Both methods yielded strong correlation between the models produced and the data in Cluster 2.

Year	Coefficient of Determination		
	All	Cluster 1	Cluster 2
2010	0.634	0.659	0.741
2011	0.632	0.417	0.825
2012	0.635	0.550	0.891
2013	0.574	0.532	0.666
2014	0.593	0.611	0.579
2015	0.748	0.727	0.770

Table 9. Coefficient of Determination from OLS regression of Cluster 1 and Cluster 2 for each year and whole period using the significant predictors of 2010 – 2015 whole dataset

https://doi.org/10.5194/isprs-archives-XLII-4-W19-455-2019 |

5. CONCLUSIONS

The reported monthly dengue incidences data allowed us to divide Quezon City into two clusters. Thru GIS and SOM clustering analysis, the two clusters produced can be characterized by the difference in their number of reported dengue cases. SOM was able to take into consideration all the monthly reported dengue cases from 2010 to 2015, and on another hand, because of its limitations, GIS grouping analysis of ArcMap 10.3 was only able to use the average monthly dengue cases from 2010 to 2015. Thus, the barangays within clusters produced using SOM showed more similarity in their trend of dengue incidence than the barangays within clusters produced using GIS grouping analysis.

The common predictors of dengue cases for both methods are the presence of informal settlements and very low-density residential. The SOM clustering algorithm produced more logical classification than the GIS Grouping Analysis. The barangays clustered using SOM showed more reasonable significant predictors than the clusters generated using GIS.

The OLS regressions performed in this study show that clustering analysis is an important process in finding data patterns for the epidemiological data. The coefficient of determination in each cluster is higher than the results of the whole dataset which indicates a strong correlation between the model and the data. Thus, it could be said that clustering data would be a better process to see relationships between the attributes.

Lastly, in the clustering analyses done, SOM is found to be very simple to implement. The superiority of Kohonen's SOM algorithm in preserving the topology of data can be observed in the resulted clusters in this research. While SOM is easy to implement, training takes a longer time as compared to k-means clustering used in the grouping analysis performed in the ArcMap.

To further improve the results of this study, it is recommended to use other regression models such as the Poisson, Negative Binomial Poisson, and Zero-Inflated Poisson regression models which are more applicable to count data (e.g., number of votes, death incidence, disease incidence, etc.). Moreover, actual environmental data (rainfall and air temperature) of the barangays are also recommended to be used for a more precise and accurate correlation of the incidence of dengue to the environmental factors. It is also recommended to have a more accurate dengue case recording for the surveillance unit of the Department of Health - Epidemiology Bureau.

REFERENCES

Alhoniemi, E. & Vesanto, J., 2000. Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks. Vol. 11 No. 3. Pp. 586-600.

Basara, H. G., & Yuan, M., 2008. Community health assessment using self-organizing maps and geographic information systems. International Journal of Health Geographics, 7(1), 67. doi:10.1186/1476-072x-7-67

Centers for Disease Control and Prevention, 2019. Retrieved from Centers for Disease Control and Prevention: www.cdc.gov/dengue/epidemiology/index.html

Edillo, F. E., Halasa, Y. A., Largo, F. M., Erasmo, J., Amoin, N. B., Alera, M., Shepard, D. S., 2015. Economic cost and burden of dengue in the Philippines. The American journal of tropical medicine and hygiene, 92(2), 360–366. doi:10.4269/ajtmh.14-0139

Kohonen, T., 1982. Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43. Pp. 59-69.

Monthly Dengue Report, 2019. Quezon City, Metro Manila, Philippines.

Mutheneni, S. R. et al., 2016. Spatial distribution and cluster analysis of dengue using self-organizing maps in Andhra Pradesh, India, 2011–2013. Parasite Epidemiology and Control. Vol 3 Issue 1, Pages 52-61.

Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011). Pages 2825-2830.

Smith, D. S., 2019. Dengue. Retrieved from Medscape: https://emedicine.medscape.com/article/215840-overview

Sun, Wei & Xue, Ling & Xie, Xiaoxue, 2017. Spatial-temporal distribution of dengue and climate characteristics for two clusters in Sri Lanka from 2012 to 2016. Scientific Reports. 7. 10.1038/s41598-017-13163-z.

Unilab, 2018. What are the Basic Symptoms of Dengue? Retrieved from Unilab: https://www.unilab.com.ph/articles/what-are-the-basic-symptoms-of-dengue/

World Health Organization, 2012. Global Strategy for Dengue Prevention and Control, 2012 - 2020. France.

Xu, B., Madden, M., Stallknecht, D. E., Hodler, T. W. & Parker, K. C., 2012. Spatial and spatial-temporal clustering analysis of hemorrhagic disease in white-tailed deer in the southeastern USA: 1980–2003. Preventive Veterinary Medicine 106, 339–347.

Zhang, J., Shi, H. & Zhang Y., 2009. Self-Organizing Map Methodology and Google Maps Services for Geographical Epidemiology Mapping. 2009 Digital Image Computing: Techniques and Applications. Pp. 229-235. DOI 10.1109/DICTA.2009.46