Effect of Data from Neighbouring Regions to Forecast Dengue Incidences in Different Regions of Philippines Using Artificial Neural Networks

K. Darshana Abeyrathna[a], Ole-Christoffer Granmo[b], Morten Goodwin[c]
Department of Information and Communication Technology
University of Agder, Grimstad, Norway
darshana.abeyrathna@uia.no[a], ole.granmo@uia.no[b], morten.goodwin@uia.no[c]

Abstract

Disease outbreaks forecasting is a vital component in public-health resource planning and emergency preparedness. However, the existing procedures have limitations due to the lack of analytical knowledge about spatiotemporal data. In this paper, we investigate how spatiotemporal data can be leveraged to forecast future disease outbreaks, using dengue incidences in the Philippines for demonstration purposes. Our approach is based on identifying highly correlated regions and using inputs from these regions to train and forecast dengue incidences using Artificial Neural Networks. We then removed the spatial aspect, focusing separately on each region to measure the effect of introducing spatial data. In all the experiments, monthly dengue incidences in 2016 were used as the testing data. Our empirical results show that including spatial data reduces the Mean Absolute Error by approx. 54 % compared to only using data from the target region. We conclude that adding data from neighbouring regions for forecasting can enhance the traditional approaches for forecasting dengue outbreaks, and we recommend that a spatio-temporal analysis is introduced as a standard component of disease outbreak forecasting.

1. Introduction

Disease Outbreaks Forecasting

According to the World Health Organization, an early warning system should be able to predict an outbreak in terms of the time it occurs, the area it affects, as well as its magnitude (Drake, 2005). Forecasting disease outbreaks is a vital component in public-health resource planning, emergency preparedness and helps in reducing morbidity and mortality due to serious illnesses (Nsoesie et al., 2014). However, the task is challenging since it is controlled by a number of internal and external factors. A large number of researchers in the field have been trying to analyse the disease outbreak forecasting problem with different approaches. Many of them are focusing on predicting the magnitude of the disease outbreaks while far fewer are working on forecasting the spatiotemporal data related to disease outbreaks. As Drake mentioned (Drake, 2005), forecasting the magnitude of disease outbreaks is difficult due to the infectious nature of diseases. Also, the factors which fluctuate the disease outbreak time series, such as characteristics of the disease, the environment or climate fluctuations, human demography, international and interregional travels (specially of the pathogen), and the health facilities of a country or of a region of a country (Myers et al., 2000) reduce forecasting performance. If there is a way to include these factors into forecasting models as input variables, the accuracy of the outcomes could increase significantly: (Linthicum et al., 1999) and (Hii et al., 2012) use climate and satellite data to forecast fever epidemics and temperature and rainfall data to forecast dengue incidence, respectively.

Particularly, factors that decisively affect the dengue fever are identified as international and inter-regional transports, population growth, the rate of urbanization, health infrastructure of the country, climate changes, and availability of disease control systems (Descloux et al., 2012). According to Descloux et al., each year, approximately, 25,000 deaths have been being recorded out of 500,000 patients with dengue haemorrhagic fever or dengue shock syndrome. This is a portion of 50 million people

This paper was presented at the NIK 2018 conference. For more information see http://www.nik.no/

who have been affected by one of the four serotypes of dengue virus: DENV 1 – DENV 4. The virus is mainly transmitted by Aedes Aegypti (also known as yellow fever mosquito) (Barbazan et al., 2002). Although the casualty rate of dengue haemorrhagic fever incidences has been reduced by early diagnosis and treatments of patients, still it is a serious concern in the public health.

To control the spreading rate, an early warning system is a necessity. Usage of weather variables is ample in forecasting models to forecast climate-sensitive inflection diseases such as Dengue and Malaria. Since many research are being conducted to find the influence of temperature and rainfall data series on forecasting outcomes (Descloux et al., 2012, Hii et al., 2012, Mattar et al., 2013), objective of this research is to find the influence of data from neighbouring regions to forecast the dengue incidences in the Philippines. Nevertheless, the existing processes for forecasting disease outbreaks are not reliable due to several reasons. Inadequate forecasting models are one of the major reasons among them. Since in this research, our emphasis is on forecasting disease outbreaks in terms of time and location, a study of spatiotemporal data forecasting techniques is essential.

Disease Outbreaks Forecasting Techniques

In the recent years, problems of data scarcity and computational power have been replaced by new challenges. In all brevity, knowledge discovery, data mining, and forecasting techniques are now challenged by the abundance of data available from efficacious data harvesting tools that produce real-time spatiotemporal data massively, for different applications in environmental science, meteorology, precision agriculture, oceanography, and other domains (Haworth and Cheng, 2012). In parallel to the above trend, many papers and books are being published with the purpose of improving the people's understanding of spatiotemporal data and encouraging them to develop new models which can precisely analyse such data (Li et al., 2002).

In addition to the simple statistical strategies, different regression models have been used to forecast dengue incidences: Multiple Regression (Hii et al., 2012, Phung et al., 2015), Seasonal Autoregressive Integrated Moving Average (SARIMA) (Choudhury et al., 2008, Gharbi et al., 2011, Phung et al., 2015), Autoregressive Integrated Moving Average (ARIMA) (Luz et al., 2008, Promprou et al., 2006, Silawan et al., 2008), Poisson Distributed Lag Model (PDLM) (Phung et al., 2015). Althouse and Cummings (Althouse et al., 2011) compare three regression techniques to forecast dengue incidences in Singapore and Bangkok (Step-down Linear Regression, Generalized Boosted Regression, and Negative Binomial Regression) and show that Linear Regression outperforms other models.

Available auto-regression techniques to analyse spatiotemporal data are summarized in (Pokrajac and Obradovic, 2001) under spatial data forecasting, temporal data forecasting, and spatiotemporal data forecasting. Haworth and Cheng (Haworth and Cheng, 2012) divide available techniques into two categories: statistical methods and machine learning methods. According to them, eigenvector spatial filtering, geographically and temporally weighted regression, space-time autoregressive integrated moving average (STARIMA), space-time geostatistical models, and spatial panel data models can be categorized as statistical methods, while support vector machines, artificial neural networks, and non-parametric regression techniques are categorized as machine learning techniques. Although interest of the researchers has been drawn towards statistical methods, in this research we use Artificial Neural Networks (ANNs) to forecast future dengue outbreaks in the Philippines since their ability to model complex nonlinear relationships in spatiotemporal data.

A detailed overview about historical dengue incidences in Philippines and arrangement of the ANNs to forecast the future dengue incidences are given in *Materials and Design of Experiment* section. Results obtained at the above section are summarized and discussed in *Results and Discussion* section. The conclusion of the research is given at the end.

2. Materials and Design of Experiment

Dengue Incidences in the Philippines

The dataset, which has been gathered by the Department of Health of the Philippines, contains monthly dengue incidences per 100,000 population of all regions in the Philippines from 2008 to 2016. The region map of the Philippines, which has 17 regions, is given in Figure 1.

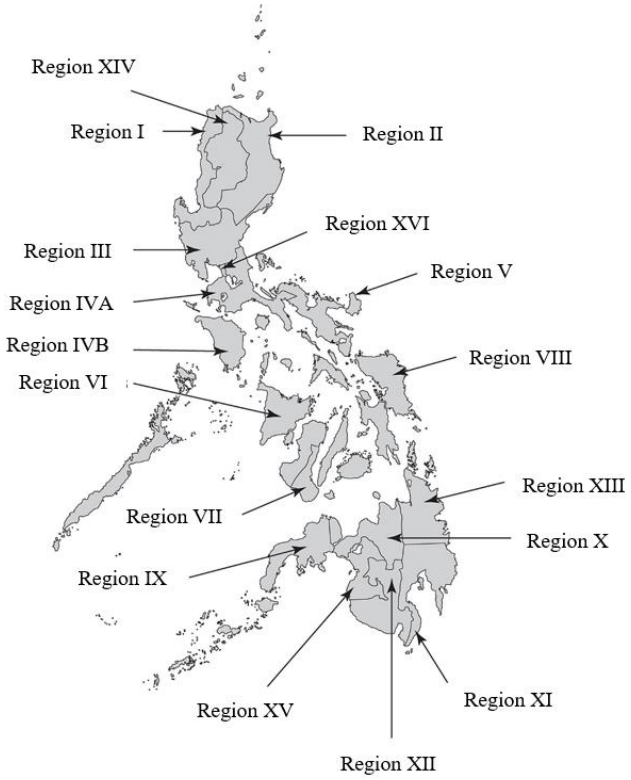

Figure 1 Administrative 17 regions of the Philippines

According to the gathered data, the highest number of incidences are recorded in 2013 (4919.09 per 100,000 population). Total incidences for the years from 2008 to 2016 show that regions XIV, VII, and VI have the highest number of dengue patients. These details are illustrated by Figure 2, where Figure 2.a shows the total number of dengue incidences of each year from 2008 to 2016 and Figure 2.b shows the total dengue incidences of each region from the region I to region XVI.

Sum of the monthly dengue incidences of all regions for each month shows that there is a peak after every two months from January. In other words, the number of dengue incidences increase from January and reaches to its first peak in March. Likewise, there are four peaks throughout the year which are in March, June, September, and December as illustrated in Figure 3.

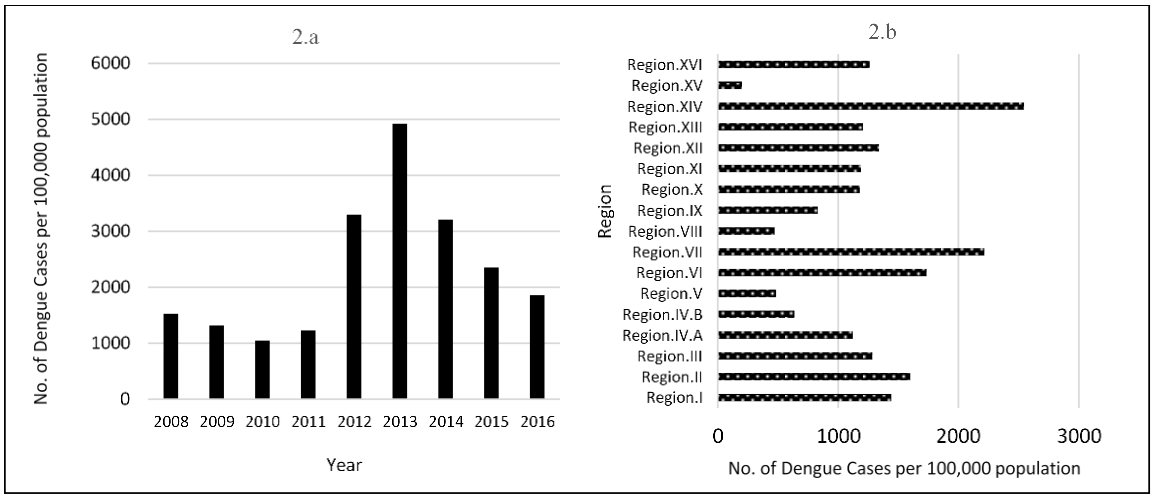

Figure 2 Total dengue incidences of each year from 2008 to 2016 and each region from I to XVI

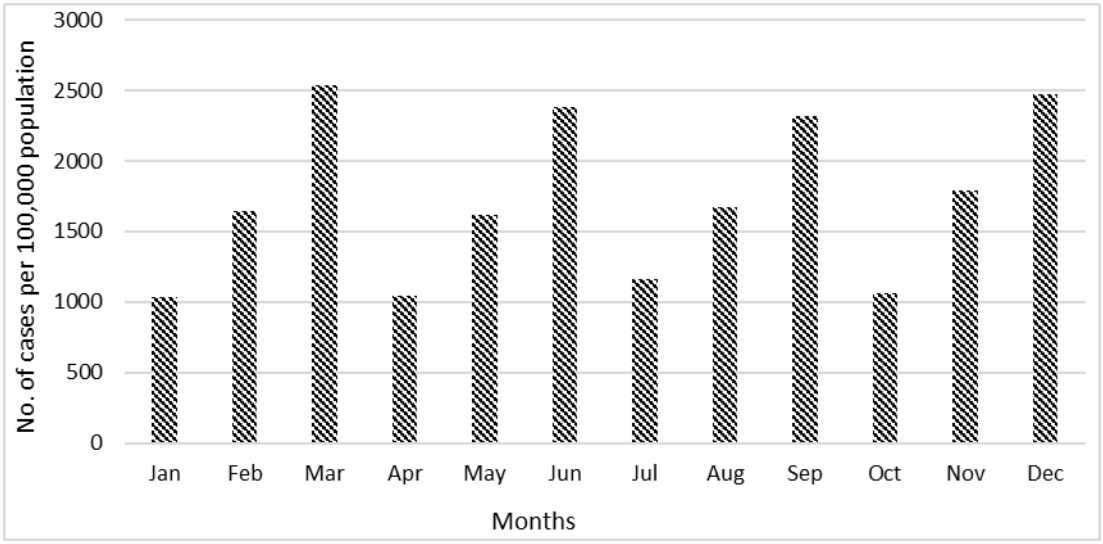

Figure 3 Sum of the monthly dengue incidences of all regions for each month

Even though researchers had used weather variables as inputs to forecast possible dengue outbreaks in the future (Descloux et al., 2012, Gharbi et al., 2011, Hii et al., 2012), rainfall and temperature data gathered by the World Bank Group for the same period have no visible or numerical correlation to the above monthly variations of dengue incidences in Philippines. These monthly variations of average rainfall and temperature series are presented against the monthly variation of total dengue incidences in the Philippines and given in Figure 4.a and 4.b, respectively.

When the average precipitation reaches its highest in July, average temperature reaches to its highest in May. Although average precipitation riches to its second peak in September, there is no clear second peak for average temperature as shown by Figure 4.b. Likewise, both average rainfall and precipitation series follow their own patterns and total dengue records has no correlation for those variations. However, the data series of the adjacent regions show higher correlations with each other. Therefore, our goal is to identify how these data from neighbouring regions affect the forecast of future dengue incidences. Data preparation and the model arrangement are discussed in the next section.

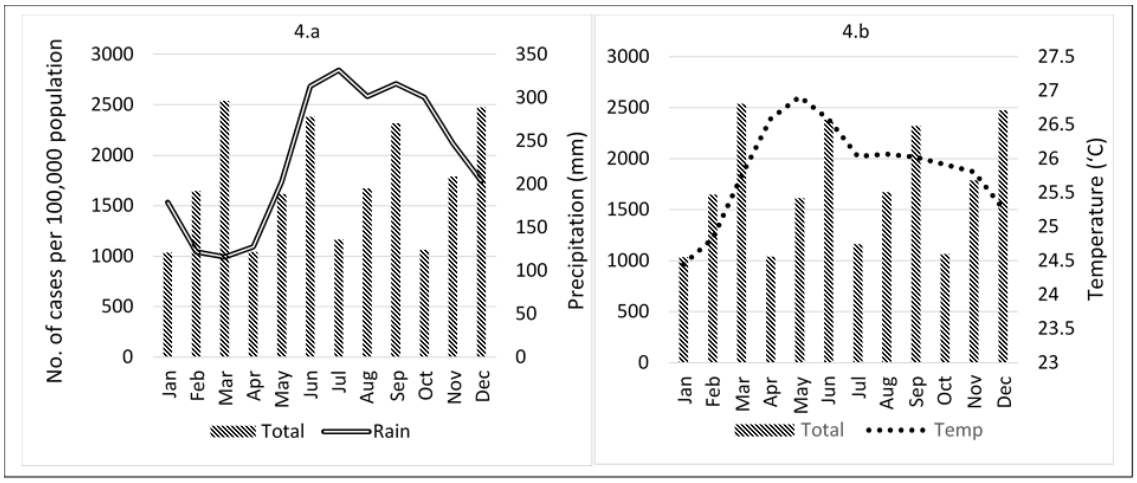

Figure 4 Average rainfall and temperature variations against monthly dengue incidences in the Philippines

Design of Experiment

Forecasting Dengue Incidences using the Historical Data of the Same Region

This is the ordinary way of forecasting the future dengue outbreaks. Historical data of the same region is used to forecast the future dengue incidences. Although the researchers have used time series techniques for the above purpose, we arrange an ANN to forecast the future dengue outbreaks using two inputs from the historical data of the same region. To forecast the possible dengue incidences of month *M*, data from month *M-1* and *M-12* are used as they have the most relevant information about the month *M*. An example case for the above arrangement can be identified as, to forecast the number of dengue incidences in November, 2016 for the region *i*, (*D*(*M*)*i*), October, 2016 data of the same region (*D(M*-1)*i*) and November, 2015 data of the same region (*D(M*-12)*i*) are used as inputs. Separate data series are arranged to forecast each region, separately. A sample data arrangement is given by Table 1 where it is used to train the ANN to forecast all the months in 2016 of region *i*, separately. However, to be fair for all months of the testing set, the ANN is trained with the same amount of data to forecast each month in the testing data set. As an example, even though the target series of the training data set starts from January 2009 to forecast dengue incidences in January 2016, target series starts from February 2009 to forecast dengue incidences in February data in 2016.

Table 1 Data arrangement to forecast dengue incidences of region *i* in each month of 2016 using same region's data

		Training			Testing		
Target	(*D*(*M*)*i*)	Jan 2009	……	Dec 2015	Jan 2016	……...	Dec 2016
Inputs	(*D(M*-1)*i*)	Dec 2008	……	Nov 2015	Dec 2015	……...	Nov 2016
	(*D(M*-12)*i*)	Jan 2008	……	Dec 2014	Jan 2015	……...	Dec 2015

Forecasting Dengue Incidences using Data from Neighbouring Regions

In this experiment, dengue incidences of each month of 2016 in all regions are forecasted using more than two inputs. In addition to the inputs identified in Table 1, data from regions which show higher correlation values is used as inputs to forecast future dengue outbarks. As an example, to forecast the number of dengue incidences in November, 2016 for the region *i*, (*D*(*M*)*i)*, October, 2016 data of the same region (*D(M*-

1)*i*), November, 2015 data of the same region (*D(M*-12)*i*), and October, 2016 data from neighbouring regions *j,* (*D*(*M*-1)*j*; where *j* can be 1,2,3,…..17 but $j \neq i$ and *j* and *i* are highly correlated) are used as inputs. Separate data series are arranged to forecast each region, separately. A sample data arrangement is given by Table 2 where it is used to train the ANN to forecast each month in 2016 of region *i*, separately.

Table 2 Data arrangement to forecast dengue incidences of region *i* in each month of 2016 using different regions' data

		Training			Testing		
Target	(*D*(*M*)*i*)	Jan 2009	.....	Dec 2015	Jan 2016	…….....	Dec 2016
Inputs	(*D(M*-1)*i*)	Dec 2008	.....	Nov 2015	Dec 2015	…….....	Nov 2016
	(*D(M*-12)*i*)	Jan 2008	.....	Dec 2014	Jan 2015	…….....	Dec 2015
	(*D(M*-1)*j*)	Dec 2008	.....	Nov 2015	Dec 2015	…….....	Nov 2016

There can be more than one data series in addition to the first two series which were created by the data from the same region. In other words, more than one region can be identified which show a higher correlation to the selected region. However, this number is not equal for all regions. Table 3 presents the regions which give their data to create inputs to forecast dengue incidences of their neighbouring regions. The total dengue incidence of all regions is also considered while checking the correlation and it is also used as an input if it shows a higher correlation to the data series of the selected region.

Table 3 Regions that support their neighbouring regions by providing their data as inputs

		Additional inputs																	
	Regions	I	II	III	IVA	IVB	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	Total
Targets	I		■	■	■											■			■
	II	■		■	■											■			■
	III	■	■		■													■	
	IVA			■		■	■											■	■
	IVB		■		■			■											■
	V				■		■	■											■
	VI					■	■		■					■					■
	VII				■		■	■						■					■
	VIII						■	■											
	IX											■	■	■					
	X										■		■	■	■		■		
	XI										■	■		■			■		
	XII										■	■	■				■		■
	XIII										■	■		■			■		
	XIV	■	■	■	■	■													■
	XV										■	■	■	■					
	XVI	■		■	■		■												

Designing the Artificial Neural Network

Once the data is ready, they are fed into the designed ANN. A network with four hidden layers (L_1-L_4) are used to forecast each region separately at both occasions. A preliminary test reveals that the number of neurons in the first hidden layer should be at least equal to the number of inputs to the network. Also, for the considered small dataset,

we try to keep the network simple. Considering these two concepts and the number of outputs, which is always one for all the cases, we gradually decrease the number of neurons of the hidden layers and make it to one at the end. Therefore, the number of neurons of the second (L_2), third (L_3), and fourth (L_4) hidden layers are kept the same for all the cases which are 4, 2, and 1, respectively. The number of neurons of the first hidden layer (L_1) is always equal to the number of inputs (L) to the network. As an example, to forecast the month M of the region I with the aid of data from neighbouring regions, a network with 7 inputs ($L = 7$), four hidden layers with neurons equal to 7, 4, 2, 1, and one output is used. Inputs for the selected example are ($D(M\text{-}1)_I$), ($D(M\text{-}12)_I$), ($D(M\text{-}1)_{II}$), ($D(M\text{-}1)_{III}$), ($D(M\text{-}1)_{IVA}$), ($D(M\text{-}1)_{XIV}$), ($D(M\text{-}1)_{ToT}$), and they can be arranged with the information given in Table 2 and 3. During the training process, backpropagation algorithm is used to training the network and training process stops when it completes 1000 training cycles or when the training error (Mean Squared Error) reaches zero. Figure 5 shows the network which can be adjusted to forecast each region with different data sets discussed in the *Design of Experiment* Section.

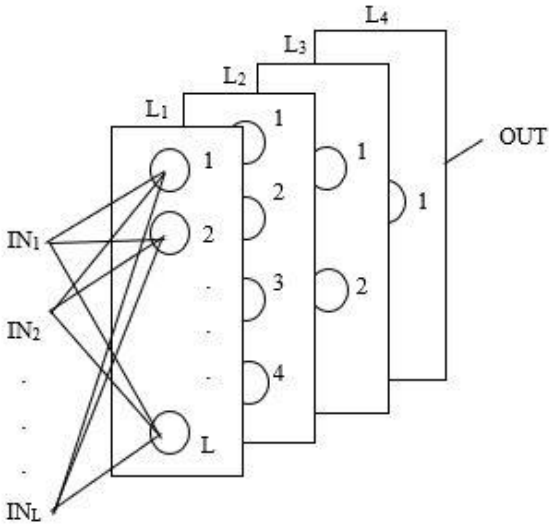

Figure 5 The proposed Neural Network which is adjustable to forecast each region, separately

However, to show that there is no bias on any of the experiments from the parameters of the ANN, the number of neurons in the first hidden layer is changed to 2 and 7 in the above example to forecast dengue incidences of the region I with data from the same region as given in Table 1. Therefore, the ANN trains with two different parameter sets when it is used to forecast future dengue incidences using data from the same region. Consequently, the same region is forecasted three times with different inputs and different ANN parameter arrangements: Experiment-1: Forecast each region with the aid of data from neighbouring regions and the number of neurons in the first hidden layer equal to the number of inputs. Experiment-2: Forecast each region with data from the same region and number of neurons in the first hidden layer equal to two. Experiment-3: Forecast each region with data from the same region and number of neurons in the first hidden layer equal to the number of neurons used during the first experiment. Obtained results are presented and discussed in the next section.

3. Results and Discussion

Dengue incidences of each month in 2016 of all regions are forecasted under the experiments discussed in the previous section. The error between the actual data series

and forecasted series of each region for the testing year 2016 is calculated in terms of the Mean Absolute Error (MAE) as shown in Equation 1.

$$MAE_i^E = \frac{1}{12} \times \sum_{m=1}^{12} |(D_A(M)i)_m - (D_F(M)i)_m| \quad 1$$

Both actual ($(D_A(M)i)$ and forecasted $(D_F(M)i)$ series consist of 12 elements (m = 1,2,3,….12) and it is equal to the number of months of a year. Therefore, sum of the absolute error divide by the number of months to get the Mean Absolute Error of region i (i = I, II,….XVI), of experiment E (E = 1,2,3). The MAE_i^E of each region obtained at different experiments are summerized in Table 5.

Table 5. Mean Absolute Errors of all regions at different experiments

Region	MAE^1	MAE^2	MAE^3	Region	MAE^1	MAE^2	MAE^3
I	2.42	4.90	5.63	**IX**	1.95	3.80	4.18
II	3.51	5.64	6.27	**X**	2.31	6.22	5.98
III	2.31	7.29	6.27	**XI**	1.60	3.09	2.81
IVA	1.65	4.70	4.70	**XII**	1.95	4.21	4.66
IVB	1.28	4.88	4.68	**XIII**	3.72	5.67	5.04
V	0.98	2.67	2.81	**XIV**	4.11	11.08	10.27
VI	1.71	5.53	4.82	**XV**	0.67	1.13	0.92
VII	7.40	11.80	11.43	**XVI**	2.16	4.94	4.53
VIII	0.83	2.45	2.24				

The best *MAE* of each region comes with the Experiment-1, where data from the neighbouring regions is used to calculate the future dengue outbreaks. The best *MAE* given by Experiment-1 belongs to region XV and it is equal to 0.67. These value by Experiment-2 and 3 are 1.13 and 0.92 and belong to the same region. The average *MAEs* are calculated for each experiment to decide the best way of forecasting the future dengue occurrences in the Philippines out of these three methods using Equation 2.

$$Average\ MAE^E = \frac{1}{17} \times \sum_{i=I}^{XVI} MAE_i^E \quad 2$$

The $Average\ MAE^{(1,2,3)}$ are 2.39, 5.29, and 5.13, respectively. Therefore, the percentage decrease of *MAE* when the forecasting model use data from neighbouring regions compared to Experiment-2 and 3 are 54.92% and 53.52%, respectively. These is a significant improvement of results. However, the change of ANNs' parameters from Experiment-2 to Experiment-3 does not make a significant difference. The change decreases the $Average\ MAE$ from 5.29 to 5.13. The graph given in Figure 6 shows the variation of average regional *MAE* throughout the testing year.

The best monthly *MAE* obtained at experiment 1, 2, and 3 are 1.38 (September), 2.94 (July), 2.78 (July). However, monthly variations of $MAE^{(1,2,3)}$ show that the steps identified at Experiment-1 are the best to forecast monthly dengue incidences in the Philippines as it obtains the lowest *MAE* for all months.

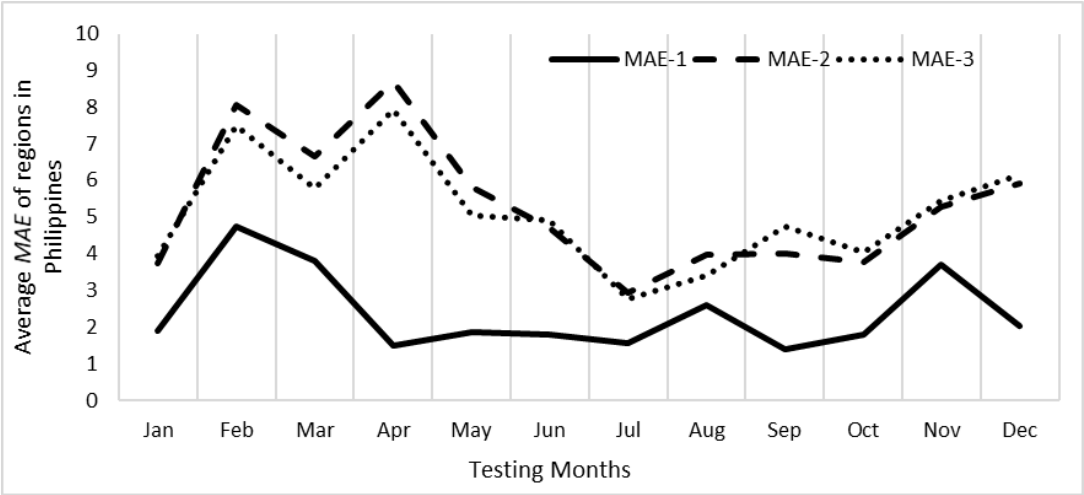

Figure 6 Average regional *MAE* throughout the testing year

4. Conclusion

According to the results discussed in the previous section, steps used at the Experiment-1 are identified as the best way to forecast all regions in the Philippines. Adding data as inputs from the neighbouring regions improves the forecasting performances significantly. To eliminate adding noises to the training data, a data analysis and preprocessing step is recommended.

However, this research focuses on finding the effect of adding data from neighbouring regions to forecast future dengue incidents and does not focus on applying multiple strategies to reduce the forecasting error. The dataset used in this research is quite small and parameters used while arranging the ANNs might not be the optimum. Therefore, the research can be further extended to reduce the forecasting error by focusing more on these limitations and more other strategies.

References

ALTHOUSE, B. M., NG, Y. Y. & CUMMINGS, D. A. 2011. Prediction of dengue incidence using search query surveillance. *PLoS neglected tropical diseases,* 5**,** e1258.

BARBAZAN, P., YOKSAN, S. & GONZALEZ, J.-P. 2002. Dengue hemorrhagic fever epidemiology in Thailand: description and forecasting of epidemics. *Microbes and infection,* 4**,** 699-705.

CHOUDHURY, Z. M., BANU, S. & ISLAM, A. M. 2008. Forecasting dengue incidence in Dhaka, Bangladesh: A time series analysis.

DESCLOUX, E., MANGEAS, M., MENKES, C. E., LENGAIGNE, M., LEROY, A., TEHEI, T., GUILLAUMOT, L., TEURLAI, M., GOURINAT, A.-C. & BENZLER, J. 2012. Climate-based models for understanding and forecasting dengue epidemics. *PLoS neglected tropical diseases,* 6**,** e1470.

DRAKE, J. M. 2005. Limits to forecasting precision for outbreaks of directly transmitted diseases. *PLoS medicine,* 3**,** e3.

GHARBI, M., QUENEL, P., GUSTAVE, J., CASSADOU, S., LA RUCHE, G., GIRDARY, L. & MARRAMA, L. 2011. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC infectious diseases,* 11**,** 166.

HAWORTH, J. & CHENG, T. 2012. Non-parametric regression for space–time forecasting under missing data. *Computers, Environment and Urban Systems,* 36**,** 538-550.

HII, Y. L., ZHU, H., NG, N., NG, L. C. & ROCKLÖV, J. 2012. Forecast of dengue incidence using temperature and rainfall. *PLoS neglected tropical diseases,* 6**,** e1908.

LI, Z., DUNHAM, M. H. & XIAO, Y. Stiff: A forecasting framework for spatiotemporal data. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2002. Springer, 183-198.

LINTHICUM, K. J., ANYAMBA, A., TUCKER, C. J., KELLEY, P. W., MYERS, M. F. & PETERS, C. J. 1999. Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya. *Science,* 285**,** 397-400.

LUZ, P. M., MENDES, B. V., CODEÇO, C. T., STRUCHINER, C. J. & GALVANI, A. P. 2008. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *The American journal of tropical medicine and hygiene,* 79**,** 933-939.

MATTAR, S., MORALES, V., CASSAB, A. & RODRÍGUEZ-MORALES, A. J. 2013. Effect of climate variables on dengue incidence in a tropical Caribbean municipality of Colombia, Cerete, 2003–2008. *International Journal of Infectious Diseases,* 17**,** e358-e359.

MYERS, M. F., ROGERS, D., COX, J., FLAHAULT, A. & HAY, S. 2000. Forecasting disease risk for increased epidemic preparedness in public health. *Advances in Parasitology*. Elsevier.

NSOESIE, E. O., BROWNSTEIN, J. S., RAMAKRISHNAN, N. & MARATHE, M. V. 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses,* 8**,** 309-316.

PHUNG, D., HUANG, C., RUTHERFORD, S., CHU, C., WANG, X., NGUYEN, M., NGUYEN, N. H. & DO MANH, C. 2015. Identification of the prediction model for dengue incidence in Can Tho city, a Mekong Delta area in Vietnam. *Acta tropica,* 141**,** 88-96.

POKRAJAC, D. & OBRADOVIC, Z. Improved spatial-temporal forecasting through modelling of spatial residuals in recent history. Proceedings of the 2001 SIAM International Conference on Data Mining, 2001. SIAM, 1-17.

PROMPROU, S., JAROENSUTASINEE, M. & JAROENSUTASINEE, K. 2006. Forecasting Dengue Haemorrhagic Fever Cases in Southern Thailand using ARIMA Models.

SILAWAN, T., SINGHASIVANON, P., KAEWKUNGWAL, J., NIMMANITYA, S. & SUWONKERD, W. 2008. Temporal patterns and forecast of dengue infection in Northeastern Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health,* 39**,** 90.