The Factors are Using Correlation and Regression Relationship Analysis for Dengue Hemorrhagic Fever Outbreak in Northeast of Thailand

1Saisunee Jabjone, 2Benjapuk Jongmuenwai, 3Sudajai Lowanichchay
1 Information Technology Curriculum, Nakhonratchasima Rajabhat University
340 Suranaree road., Muang District, Nakhonratchasima, 30000, Thailand
Email: benjapuk.j@nrru.ac.th
2 Management Information System Curriculum, Nakhonratchasima Rajabhat University
Email: sudajai@hotmail.com
3 Information Technology Curriculum, Nakhonratchasima Rajabhat University
Email: a1102923@hotmail.com

Abstract
The factors are correlation and regression relationship for dengue hemorrhagic fever (DHF) outbreak in Northeast of Thailand. Which factors can use of analysis for climate change by using data in past of between 2007-2016 total of ten years. Data analysis objective were (1) to factor analyzed climate change for DHF outbreak in Northeast of Thailand and (2) to created model optimization for predictors by correlation statistic is called a correlation coefficient and regression analysis. Furthermore, the factor is complex six variables include of type of dengue, sex, area, rainfall, temperature and humidity for DHF: This correlation input of data for analysis can use data correlation between factor relationship (such as sex with area, sex with rainfall or area with rainfall). The regression analyze is used to calculate linear function for prediction of factors. Program are working in process and analyze using SPSS, R Studio and Excel. Therefore, data after for analyze with are compared different between actual and prediction value. This model correlation and regression results is accurate to 91.3 percent.

Keywords: Dengue Hemorrhagic Fever (DHF); Correlation; Regression; Relationship; Factor Analysis.

1.Introduction
Dengue fever occurs from RNA virus which can divide group four serotypes including of DEN-1, DEN-2, DEN-3 and DEN-4. The four serotypes found Antigen integrate cross reaction and cross protection of period short time [1]. In DHF and DSS, there is an acute increase in vascular permeability that leads to leakage of plasma into the extravascular compartments, resulting in haemoconcentration and hypotension. [2]. Dengue is one of the main problems of public health in the world. It is estimated that about 2.5 billion people are now at risk of dengue. [3]. In recent years, dengue fever are six diseases of Thailand and other physical countries as they are both transmitted by Aedes mosquitoes, which are in area Thailand [4]. Therefore, we have used to factors analysis of DHF outbreak in Northeast of Thailand and can use statistic the correlation and regression analysis. The correlation addresses the relationship between two different factors variables. The statistic is called a correlation coefficient. A correlation coefficient can be calculated when there are two or more sets of scores for the same individuals or matched groups.

This research is collection data of secondary data and DHF outbreak in northeast of Thailand, which of the data include of climate change rainfall, air temperature humidity, sex, area, type of dengue fever, time of period in the part of year 2007-2016. Therefore, we have studied factors analysis which epidemiologic of social science for find tool of predictors DHF in northeast of Thailand. Moreover, this study is contribution for surveillance DHF system development and reduce occurs crucial outbreak. The calculate of statistical using frequency percentage and correlation coefficient of Pearson and is analyzed by multiple regression analysis. This is a study of the relationship between dependent variable with independent variable (i.e. sex with age, sex with area, sex with type of dengue fever, sex with year, sex with rainfall, sex with air temperature, sex with humidity) and predictors of disease outbreaks, which will be useful in planning, controlling and preventing future outbreaks.
2. Related Works

The correlation coefficient is a measure that determines the degree to which two variables movements are associated. The range of values for the correlation coefficient is -1.0 to 1.0. If a calculated correlation is greater than 1.0 or less than -1.0, a mistake has been made. A correlation of -1.0 indicates a perfect negative correlation, while a correlation of 1.0 indicates a perfect positive correlation. Show equation as (1).

\[
p_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}
\]

(1)

While the correlation coefficient measures a degree to which two variables are related, it only measures the linear relationship between the variables. Nonlinear relationships between two variables cannot be captured or expressed by the correlation coefficient. A value of exactly 1.0 means there is a perfect positive relationship between the two variables. For a positive increase in one variables, there is also positive increase in the second variable. A value of exactly -1.0 means there is a perfect negative relationship between the two variables. This shows the variables move in opposite directions; for a positive increase in one variable, there is a decrease in the second variable. If the correlation is 0, this simply means there is no relationship between the two variables. The strength of the relationship varies in degree based on the value of the correlation coefficient. For example, a value of 0.2 indicates there is a positive relationship between the two variables, but it is weak.

Standard deviation is a measure of the dispersion of data from its average. Covariance is a measure of how two variables change together, but its magnitude is unbounded so it is difficult to interpret. By dividing covariance by the product of the two standard deviations, a normalized version of the statistic is calculated. This is the correlation coefficient. [5].

Then, for the regression model can used the residues from the adjusted model for each series in a study in which the dependent variable was the residue of dengue cases and variables waste were independent of the models adjusted for maximum, mean and minimum temperature, humidity. For statistical analysis of data the program Gretl 1.9.2 was used. Gretl (Gnu Regression acronym, Econometrics and Time-series Library) is a free software that collects and interprets econometric data, this software enables you to do a full analysis of temporal series, from graphical analysis of the data to forecasts. [3].

Dengue prevalence data (/1,000,000 population) for 2004 in a central region of Thailand were derived from the reported registry data on dengue of Ministry of Public Health (epid.moph.go.th). Rainfall data (average inches) in the studied area were derived from the Royal Irrigation Department, Thailand. The correlation between the rainfall and the prevalence of dengue was assessed by regression analysis. The least square equation plot prevalence (Y) versus rainfall (X) and the correlation coefficient (r) was calculated. All statistical analysis was performed using SPSS 10.0 for Windows. The detail of dengue prevalence in 22 provinces of a central region of Thailand and the average rainfall in each province is presented the least square equation plot prevalence (Y) versus rainfall (X) is Y = 17.5X + 24.2 (r = 0.88, p < 0.05). The main aim of this retrospective study was to investigate the correlation between the rainfall and the prevalence of dengue. The work described the effect of rainfall on dengue prevalence, which is important because of the need to develop tools to forecast variations in disease incidence and the risk related to the impact of change in climate [6].

The rainfall data (2002-03) collected from Ministry of Public Health were utilised for transformation of the infection and the rainfall data were derived from Royal Irrigation Department, Thailand. The correlation between the rainfall and the prevalence of dengue was assessed by regression analysis. The least square equation plot prevalence (Y) versus rainfall(x) is y = 3.0x + 4.6 (r=0.78, p<0.05) (r = 0.68, p<0.05). The study indicated that the prevalence of dengue infection in Thailand may depend on rainfall. Therefore, the surveillance and control of mosquito should be intensified during the period with high rainfall is recommended. However, the other confounding factors like ambient temperature and humidity which also determine the transmission of dengue should be looked into, before concluding that the increased prevalence is a result of rainfall alone. Further, similar studies to access the correlation between the rainfall and prevalence of infection in the other countries are required to confirm these observations [7].

Patient with DHF is a probable case of DF, having hemorrhagic tendency along with thrombocytopenia (platelets 100,000/cu. mm or less) and evidence of plasma leakage. Fifty normal
healthy age and sex matched individuals, having no history of febrile or other illness in the last three months were included as negative controls. The statistical analysis was performed with Epi Info version 3.3.2 (updated 2005, Center for Disease Control and Prevention, USA). Proportions of patients with abnormal clinical, hematological and biochemical findings between the two groups were compared using Chi \((x^2)\)/Fischer exact test. Mean values between four serotype groups were compared using analysis of variance (ANOVA). The results of correlation of dengue serotypes with clinical and hemorrhagic manifestations. Among the wide spectrum of mild and severe clinical and hemorrhagic manifestations analyzed, the incidence of only two severe manifestations, namely, abdominal pain \((p<0.05)\) and hepatomegaly \((p<0.05)\), was significantly different between four serotype groups. While frequency of clinical manifestations was most predominant in Den-2 infected cases, severe hemorrhagic manifestations were predominant in Den-4 infected group. Individual analysis of each serotype group demonstrated hepatomegaly to be significantly higher in Den-2 as well as Den-3 infected patients [8].

The clinical data were collected on predesigned questionnaire which include gender, clinical findings, complications and laboratory test done. Statistical analysis was done using statistical package for social science (SPSS) programme. Chi square test was used to compare categorical variables and Fischer exact test were applicable. The results of 633 confirmed dengue patients were included in this study. 248 were male and 319 were female (male to female ratio was approximately 3:4). In this study, 13.2% of the patients had no increase in the transaminases level \((\text{grade 0})\), 63.8% presented mild alterations in the liver enzymes levels \((\text{grade 1})\), 17.9% presented grade 2 liver involvement, 3.9% of the patients had progressed to acute hepatitis \((\text{grade 3})\) and 1.1% had severe liver damage with fulminant hepatic failure. In 86% of the patients there was elevation of the liver enzymes. All of them \((549 \text{ patients})\) had increase in the ALT level. The change in the ALT was seen in 82% of the patients [9].

Dengue cases have been diagnosed in Florida. The purpose of this study was to find an association between potential risk factors and the expansion of dengue fever in the United States. Guided by the eco-bio-social framework, which offers a broad assessment of risk factors for the illness, a retrospective design was used with archival data to correlate changes in climatic variables and imported dengue cases with autochthonous dengue cases in Southeast Florida from 1980 to 2013. A spearman correlation indicated weak correlations between temperature and autochthonous dengue cases \((r_s = 0.999, p = 0.000)\) and imported dengue cases with autochthonous dengue cases \((r_s = 0.162, p = 0.000)\). A negative binomial multivariate regression was used to analyze the expansion of dengue to each monthly unit of temperature, rainfall and imported dengue cases over 34 years. The results indicated that temperature \((\text{IRR} = 2.198; 95\% \ CI[1.903, 2.538])\) and precipitation \((\text{IRR} = 0.991; 95\% \ CI[0.988, 0.994])\) were predictors for the geographic expansion of dengue fever in Southeast Florida [10].

The total number of 258 patients was obtained from Siriraj Hospital, Bangkok, Thailand. The data set consists of 128 DF, 65 DHF I, 52 DHF II and 13 DHF III. The set of attributes consists of clinical attributes and hematological attributes. There are totally 48 attributes \((26 \text{ numerical attributes, 21 categorical attributes and one class attribute. During the treatment period, nurses and physicians followed the symptoms. The decision tree algorithm is applied for the feature selection process and it is found that the plasma leakage, the shock occurrence, the bleeding, the number of platelet, the level of ALT, the number of white blood cell, lymphadenopathy are the potential feature sets that can categorize the dengue patients. After the feature selection, the fuzzy logic approach is tried to see the classification performances the experimental result shows that Fuzzy logic outperforms Decision tree with the 97.94% of accuracy [11].

From 2007 to 2012, Ae. Aegypti mosquitoes were collected from three provinces in the central region of Thailand, including Nakhon Pathom, Ratchaburi and Samut Sakhon. These areas were selected primarily for three reasons: high mosquito density, minor differences in climatic factors, and a high DHF morbidity rate as reported in Thailand health information system. The forecasted cases were compared with the actual dengue cases reported by NTCAESI. The dataset in this experiment includes all variables from the three provinces. The best model \((\text{yielding the lowest MAPE, AIC and BIC})\) was selected in subsequent experiments. Four MPR models were constructed, To account for climate effect on dengue cases, the categorical variable of season was included into the model fitting. The first model \((\text{Model-1})\) deployed all four main predictors, whereas the second model \((\text{Model-2})\) and third model \((\text{Model-3})\) excluded insignificant terms. AgeRate and Mosquito were highly correlated with Mosquito \((r = 0.61, p < 0.001; \text{and } r = 0.57, p < 0.001)\), implying that each conveyed a similar
relationship to dengue cases as that of Mosquito; both variables were thus removed from Model-3. Subsequently, the interaction of two newly identified main factors (Season and Mosquito) was added into the fourth model (Model-4), according to the model selection process [12]. From details related work of above. This is conclusion research for correlation all of total in table 1.

Table 1 Conclusion and Comparation methods of research for DHF

<table>
<thead>
<tr>
<th>Author/Year</th>
<th>Study data of dengue patient</th>
<th>Regression and Multiple Analysis</th>
<th>Test Hypothesis (P value)</th>
<th>ANOVA</th>
<th>Data Mining and Fuzzy Logic</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viroj Wiwanitkit. (2005).</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Viroj Wiwanitkit. (2006).</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Rajni Kumaria. (2009).</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Ali K. Ageep and Abu elgasim S. (2012).</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Lugo Brunilda (2015).</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Nuanwan</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Soonthornphisaj. (2016).</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Siriyasatien et al. (2016).</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Benjakul Jongmuenvai, Sudajai Lowanichchai, and Saisunee Jabjone (2016).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

3. Proposed work

Pearson’s use signal $r$ or $r_{xy}$. The dataset has to interval two groups and more than. This is dataset and analyze by as SPSS or R studio program show in table 2.

Table 2. Pearson Correlation of DHF

<table>
<thead>
<tr>
<th>Interval</th>
<th>Type</th>
<th>Sex</th>
<th>Area</th>
<th>Rainfall</th>
<th>Temperature</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>1.000</td>
<td>0.000</td>
<td>-0.220</td>
<td>-0.022</td>
<td>0.104</td>
<td>0.014</td>
</tr>
<tr>
<td>Sex</td>
<td>0.000</td>
<td>1.000</td>
<td>-0.013</td>
<td>-0.011</td>
<td>-0.006</td>
<td>-0.008</td>
</tr>
<tr>
<td>Area</td>
<td>-0.220</td>
<td>-0.013</td>
<td>1.000</td>
<td>0.049</td>
<td>0.003</td>
<td>-0.033</td>
</tr>
<tr>
<td>Rainfall</td>
<td>-0.022</td>
<td>-0.011</td>
<td>0.049</td>
<td>1.000</td>
<td>-0.041</td>
<td>-0.027</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.104</td>
<td>-0.006</td>
<td>0.003</td>
<td>-0.041</td>
<td>1.000</td>
<td>0.019</td>
</tr>
<tr>
<td>Humidity</td>
<td>0.014</td>
<td>-0.008</td>
<td>-0.033</td>
<td>-0.027</td>
<td>0.019</td>
<td>1.000</td>
</tr>
</tbody>
</table>

From table 2 show values of test for association between paired samples, using one of Pearson’s product moment correlation coefficient for dengue hemorrhagic fever are total six factors. The correlation coefficient for predictors ($R^2$) equal to 91.3 percent. This is a study of the relationship between dependent variable with independent variable (i.e. sex with area, sex with type of dengue fever, sex with rainfall, sex with air temperature, sex with humidity) and predictors of disease outbreaks. Which is equation predictors dengue hemorrhagic fever as $y = 0.669 + -0.220 area + -0.022 rainfall + 0.104 temperature + 0.014 humidity$. When, we have conclusions effectiveness between correlation and regression include factor relationship of patient values of correlation coefficient equal to 0.955 and regression prediction equal to 91.3 percent and should produce regression estimates of 0.00 and 1.0 for $x$. Deviation from these values is evidence of bias and inefficiency in the predictors.
The F-test can be used to test the joint hypothesis that error regression equal to ±0.699. The correlation computes the variance of x and the correlation of x and y if these are vectors. If x and y are matrices then the correlations between the columns of x and the columns of y are computed. In additional, the correlation scales a covariance matrix into the corresponding correlation matrix efficiently.

The score or scale was found r value which include of smallest to r equal to .10 to .29 moderate equal to .30 to .49 and largest r equal to .50 to 1.0. Then, meaning of r value number 1 to 5 steps as:

1. r value is minus show variable X and variable Y to relationship opposite direction, if variable X addition is effect to variable Y reduce but if variable X reduce is effect to addition.
2. r value is plus show variable X and variable Y to relationship same way direction, if variable X addition is effect to increase but if variable X reduce is effect to reduce, too.
3. If r value is imminent 1, meaning of variable X and variable Y to relationship same way direction and to much relationship.
4. If r value is imminent -1, meaning of variable X and variable Y to relationship opposite direction and to much relationship.
5. If r value equal to 0 is show variable X and variable Y none of relationship.

We have use of descriptive interpretation of the data generated by the correlation analysis. The details can be written as follows.

1. sex with area equal to -0.220 is relationship least level and opposite direction, if number of area increase. Then, area to reduce but if sex to reduce can be area increase.
2. sex with rainfall equal to -0.011 is relationship least level and opposite direction, if number of rainfall increase. Then, rainfall to reduce but if sex to reduce can be rainfall increase.
3. sex with temperature equal to -0.006 is relationship least level and opposite direction, if number of temperature increase. Then, temperature to reduce but if sex to reduce can be temperature increase.
4. sex with temperature equal to -0.006 is relationship least level and opposite direction, if number of temperature increase. Then, temperature to reduce but if sex to reduce can be temperature increase.
5. sex with humidity equal to -0.008 is relationship least level and opposite direction, if number of humidity increase. Then, humidity to reduce but if sex to reduce can be humidity increase.

Therefore, we have presented the relationship between factor area with rainfall and temperature with type as follows.

1. area with rainfall equal to 0.049 is relationship least level and same way direction, if number of rainfall to increase. Then, area to increase, too, but if area to reduce can be rainfall to reduce, too.
2. temperature with type of dengue equal to 0.104 is relationship least level and same way direction, if number of temperature to increase. Then, type of dengue to increase, too, but if temperature to reduce can be type of dengue to reduce, too.

Moreover, we have described of regression analysis model to compute statistic by function linear and multiple regression used to found prediction values for DHF outbreak in Northeast of Thailand.
Figure 1. Shown in regression factor analysis period of data between 2007-2016 for DHF.

From figure 1. This graph is show regression factor analysis period of data between 2007-2016 for DHF outbreak of factor using function linear and multiple regression for prediction of model. We have compared actual and predictors value of analysis six factors. The dataset is a collection of data or a single statistical data where every factor of data represents variable and each factor has its own description. For prediction of dengue disease, we used dengue data set for prediction contains 15,000 records and detailed analysis of scoring using regression given by accuracy prediction for disease.

For experiment result showed that scoring accuracy of sex equal to 1.48 can describe trend is period data of 2017 the most is of female. The area equal to 1234,60 can told group the Mukdahan hospital Mukdahan district. The most people were Thai national and status being of treatment and died. Moreover, age of patient between 23-36 year and the epidemic is increasing every year. Forecasting is rainfall data between of period 26-30 every month. The humidity is level of 0.00.

4. Conclusion
This research is collection data of secondary data and DHF outbreak in northeast of Thailand, which of the data include of climate change rainfall, air temperature humidity, sex, area, type of dengue fever, time of period in the part of year 2007-2016. Therefore, we have presented data of DHF which of the regression ana l yze is used to calculate linear function for prediction of factors. Program are working in process and analyze using SPSS, R Studio and Excel. Therefore, they considered that calculate was compared different between actual and prediction value. This performance was found correlated and regression relationship analysis to results accurate 91.3 percent. The experiment results can have described trend was periodized data of 2017 the most is of female. The area equal to 1234,60 can told group the Mukdahan hospital Mukdahan district. The most people were found of data Thai national and being status of treatment until to died. Moreover, age of patient between 23-36 year and the recent epidemic is increasing every year. In addition, rainfall is forecasting data for day between of period 26-30 every month and result is 16 millimeters. The temperature and humidity are forecasting data level of 0.00.

References
[3] Bicalho Cristina Carolina, Safadi Thelma, Charret Cunha da Iraziet, 2014. The Influence of Climatic Factors on Dengue Epidemics in the Cities Cuiaba (Mato Grosso State) and Lavras (Minas

Authors Biography

Dr. Saisunee Jabjone is an Assistant Professor in Informatics Program at Nakhon Ratchasima Rajabhat University, Thailand. She received Doctor of Technology in Science from University of Technology, Sydney in 2005. She is having around 20 years of Professional experienced which includes teaching and researching experience. Her areas of research interests include Data Mining, Forecasting modeling and simulation, Computer applications in knowledge-based and expert systems.

Miss. Benjapuk Jongmuenwai is an lecture in management information system of Informatics Program at Nakhon Ratchasima Rajabhat University, Thailand. She received Master of Degree in Management Information System from Suranaree University of Technology in 2009. She is having around 6 years of experienced which includes teaching and researching experience. Her areas of research interests include Data Mining, Forecasting modeling and simulation, Computer applications in knowledge-based and expert systems, Artificial Intelligence.

Dr. Sudajai Lowanichchhai is an Assistant Professor in Informatics Program at Nakhon Ratchasima Rajabhat University, Thailand. She received Doctor of Phylosophy in civil Engineering from Thammasat University in 2009. She is having around 32 years of Professional experienced which includes teaching and researching experience. Her areas of research interests include Data Mining, Forecasting modeling and simulation, Computer applications in knowledge-based and expert systems, Artificial Intelligence.