# University of Alberta

## An Early Warning System for Dengue Disease in Yogyakarta, Indonesia

by

Christina Anne Haines

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Biostatistics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta
Spring 2006

*Your file Votre référence*
*ISBN: 0-494-13819-X*
*Our file Notre référence*
*ISBN: 0-494-13819-X*

# ABSTRACT

An early warning system for dengue disease in the province of Yogyakarta, Indonesia was developed. Using Poisson general linear models, accurate predictions of the number of recorded dengue cases between June 1999 and August 2004 were obtained up to four months in advance based on past dengue incidence and sea surface temperatures. The methodology and models developed can be easily implemented in future because the procedures are relatively simple and the required data are readily available. Further, reliable predictions can be obtained with an adequate amount of lead time that will successfully allow for proper planning and implementation of effective dengue control programs.

# ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Subhash Lele, for his years of guidance, suggestions and encouragement throughout my studies and in preparation of this thesis. His support truly helped me become the scholar I now am.

Special thanks to Dr. Dana A. Focks for providing me with the opportunity to work on the dengue disease early warning system in Yogyakarta, Indonesia. Without his resources and knowledge, this thesis would not have been possible.

Thank you to Dr. N.G. Narasimha Prasad and Dr. Christine Newburn-Cook for their time spent in reading my thesis and participating on my defense committee.

My appreciation extends to the Department of Mathematical and Statistical Sciences faculty members who taught me courses during my studies, and to the departmental secretaries for their assistance and warm smiles.

My sincere gratitude goes to my fiancé, Glenn Alloway, for his patience, understanding and commitment throughout the past two years of my studies and the preparation of this thesis.

Finally, I thank my parents, Barbara and David Gray, for their countless words of advice and encouragement.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1: INTRODUCTION

Dengue is considered the most important vector-borne disease that affects the human population (National Research Council, 2001). A global pandemic of dengue is occurring throughout the tropics and sub-tropics with an estimated fifty-million new cases of dengue infection occur yearly (WHO, 1998). This shows a substantial increase in the location and frequency of epidemic cycles and corresponding incidences over the last two decades. Inevitably, this has a large economic impact on both the government and affected families, making control efforts for dengue imperative.

## Chapter 1.1: The Spread of Dengue Disease

Dengue fever, originally discovered in the Philippines in 1953 (WHO, 1997), is a vector-borne virus transmitted to humans via the mosquito *Aedes aegypti*. While feeding on the blood of an infected individual, the female mosquito acquires one of the four virus serotypes responsible for dengue fever. The virus next replicates within the mosquito, rendering her capable of passing the virus on to all subsequent humans she feeds on, in addition to her offspring. With the spread of

dengue relying on mosquito populations, climate factors such as temperature, precipitation and humidity play an important role in its transmission and incidence, through both direct and indirect effects.

The *Aedes aegypti* can survive between 5 °C and 42 °C, however temperatures below 20 °C prevent offspring from hatching (Focks et al., 1995). Further, higher temperatures lead to an increase in the speed of the mosquito's metabolic rate and egg production, resulting in increased feeding requirements (Bradley, 1993). As such, warm temperatures directly promote an increase in the mosquito population, and hence in the dengue virus spread. In addition to temperature, humidity and rainfall also influence mosquito populations via an increase in the abundance of their available breeding sites. The *Aedes aegypti* rely on pools of standing water to act as breeding sites, and items such as drum barrels and discarded tires are prime locations as they easily collect water from rainfall (Sheppard et al., 1969). It is believed that heavy rainfall discourages breeding as containers collecting water begin to overflow, whereas drought conditions promote breeding due to elevated use of water storage containers by people (Moore et al., 1978). This is one example of an indirect effect that climate has on dengue. Ecologically, cities with high human densities promote the *Aedes aegypti* as the number of feeding opportunities is large. After having contracted dengue from one of the four virus serotypes, one is temporarily immune to that particular serotype. Unfortunately, this immunity is short-lived and as such herd immunity within the community is

also only short-lived (WHO, 2000). Further, if inadequate water or waste management is present within these cities, the number of available breeding sites increases, again promoting mosquito populations.

Besides climate, transportation and migration, urbanization, and poverty can also influence the incidence of dengue disease. With transportation and migration increasing, dengue has also shown an increase in its geographical span. Southeast Asia's dengue endemic in recent years is likely due to rapid urbanization providing plenty of feeding and breeding opportunities for the mosquito and the virus it carries to spread (National Research Council, 2001). Malnutrition and poor sanitation due to poverty also compromise a person's immune system thereby making them more susceptible to the dengue virus and its potentially lethal symptoms (WHO, 2004).

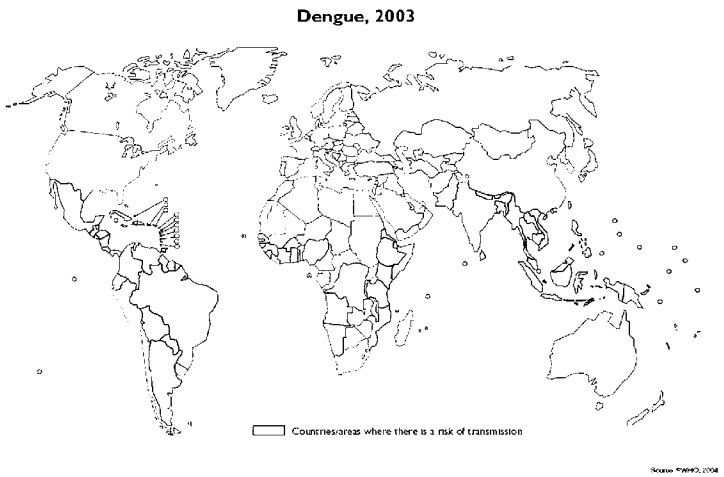


Figure 1.1.1: Geographical range of dengue as of 2003 (WHO, 2005)

## Chapter 1.2: Social and Economic Impacts of Dengue Disease

A major complication of dengue is in its diagnosis. Its symptoms include fever, severe headache, and aching muscles and bones - all of which are common for a wide variety of other diseases. As such, true dengue diagnosis is dependent on the results from expensive lab tests based on blood samples (WHO, 2000). This confirmed diagnosis is not always feasible due to its high cost, resulting in different diagnostic standards between countries. This problem makes disease management difficult and likely yields underestimated incidence counts (National Research Council, 2001).

Dengue is of high concern to policymakers in Southeast Asia due to its difficulty in diagnosis and management. In addition, it imposes a heavy burden on local hospitals, as dengue can occur throughout the entire year. A severe complication of this virus is dengue hemorrhagic fever (DHF), which is characterized by a long term high fever, vascular leakage of plasma, low platelet counts and circulatory failure (WHO, 2000). DHF has a case fatality greater than twenty percent, but this rate is reduced to less than one percent if proper treatment is received (WHO, 2000). Unfortunately the cost of proper medical treatment can be financially devastating to many families. Without accounting for time missed from work, the estimated cost for a DHF case is reported to be forty-three percent of the average

monthly family income. If symptoms become severe enough to require intensive care, the costs are estimated to be 188% of the average monthly family income (DeRoeck et al., 2003). In Jakarta, Indonesia, DHF has been reported to be the second highest cause of pediatric admission into public hospitals (DeRoeck et al, 2003), demonstrating how dengue persistently overwhelms hospital facilities, staff and finances in Southeast Asia. Politically, media concern regarding dengue is heightened as it equally affects all individuals, regardless of social economic status, due to the fact that it's vector borne, therefore forcing greater political recognition (DeRoeck et al., 2003).

## Chapter 1.3: Early Warning System for Dengue Disease

Clearly, dengue is a disease with great negative impacts on a country, both socially and economically. It results in overwhelmed medical facilities, heavy financial burdens on families, and lost work time. Further, it is projected that every 1 °C increase in future temperatures due to global warming will result in at least a thirty-one percent increase in incidence rates (Patz et al., 1998). Currently, mosquito control is the main preventative strategy for dengue disease (DeRoeck et al., 2003). Such measures however are difficult to sustain due to its high cost, short term impacts, and the high amounts of effort required by human resources.

The high costs associated with mosquito control therefore emphasize the need for an early warning system for dengue disease.

Early warning systems can provide a very helpful tool in effective mosquito control programs. Their use can help determine when mosquito control efforts should be increased based on their projected number of dengue incidences. Implementing mosquito control efforts only when an early warning system predicts an epidemic can reduce the need for human and other resources. Potentially the cost of mosquito control could also be reduced by focusing such efforts only in times projected to be of concern. Focusing mosquito control efforts could also improve its efficacy, resulting in a decrease in both morbidity and mortality due to dengue disease (WHO, 2004).

Dengue, being a vector borne virus, is climate sensitive. Thus, one can use climate factors to develop an early warning system. Specifically, characteristics of the El Nino/Southern Oscillation (ENSO) cycle nearing the equator in the Pacific Ocean can be employed, as it is known to be the cause of much of the world's climate variability (National Research Council, 2001). In Indonesia, El Nino is known to be associated with drought conditions. This could potentially increase the number of available breeding sites for mosquitoes through increased use of water storage containers. In addition to using climatic factors, increased

predictive power could result by including other factors related to population vulnerability, such as previous number of cases (WHO, 2004).

Currently, the use of early warning systems is not widespread. Reasons for this include the limited affordability and accessibility of accurate data. An early warning system must be capable of predicting incidences months in advance, in order to provide officials with adequate time to organize control efforts. Further, the predictive accuracy of such a system must be evaluated, for which there currently is no generally accepted criterion (WHO, 2004). For example, in Puerto Rico a dengue early warning system was developed based on daily temperature, precipitation and water values. This provided incidence predictions for two week intervals, three weeks in advance (Schreiber, 2001). Unfortunately, a system providing only three weeks notice may not be beneficial due to time constraints in organizing control efforts. Another early warning system for dengue used climate factors in conjunction with an early warning system for its closely related disease, Malaria (Cullen et al., 1984). This however was only capable of identifying past epidemics, and is not capable of predicting future dengue epidemics (Myers, M.F., 2000).

The advantage of using early warning systems to help control infectious diseases, such as dengue, is that the entire process of the disease does not need be known for future predictions. The goal of early warning systems is to provide reliable

future predictions about the number of projected incidences and to identify periods of time that are likely to be epidemic. Explanatory information concerning the relation between predictors and dengue epidemics is not a priority. If such predictions can be provided, much of the efforts and costs involved in mosquito control programs could be reduced, while increasing its efficacy. By improving mosquito control, the social and economic impacts of dengue disease could be dramatically decreased.

# Chapter 2: METHODOLOGY

It is essential that the early warning system created can provide reliable predictability several months in advanced, based on readily available predictors. These qualities are necessary to ensure that the system is both easy to implement and effective. In the province of Yogyakarta, Indonesia, the goal of an early warning system for dengue disease is to accurately indicate the level of concern authorities should have for a dengue epidemic. Further, being able to identify potentially epidemic months several months in advance, and then track any changes to the classification as the given month approaches, would also be useful for planning resources. Authorities involved in dengue control programs have indicated that a minimum of one month lead time for such predictions is essential for complete preparation and implementation of such programs. Two to three months lead time with less accuracy would also be valuable as "watch" indicators for the control program (Focks, 2003). This therefore requires a sequence of predictive models for the number of dengue cases. Based on the number of dengue cases predicted by each model, the level of concern for an epidemic within a given month can then be determined.

## Chapter 2.1: The Data

To develop predictive models for an early warning system for dengue disease in Yogyakarta, Indonesia, a combination of dengue case data and climate data were utilized. The response variable was the number of monthly recorded dengue cases, using past dengue case and death counts along with sea surface temperatures as predictor variables. The dengue case data contain monthly records of dengue cases and deaths occurring in Yogyakarta, Indonesia from June 1985 to August 2004. Monthly anomalies, indicating the raw difference between the monthly mean and the monthly observation, for both cases and deaths were also calculated. These values were used to indicate the degree of departure a given months observations were from the norm. The monthly averages used to determine the anomalies were only based on the data between January 1985 and December 1999, and are summarized in table 2.1.1. This was done so that the 2000 to 2004 data could be used as validation data, with the pre-determined anomalies being independent of the validation set. As an example, the recorded number of cases in June 1985 was 54, yielding a monthly anomaly of 23.0 (77.0 - 54.0). The case and death data are important in the early warning system as a measure of the vulnerability of the population to dengue.

Table 2.1.1: Monthly averages of dengue cases and deaths (1985 and 1999)

| Month | Average Cases | Average Deaths |
|---|---|---|
| January | 131.6 | 3.3 |
| February | 107.2 | 3.1 |
| March | 123.5 | 3.5 |
| April | 148.5 | 3.9 |
| May | 139.7 | 2.9 |
| June | 77.0 | 1.5 |
| July | 52.3 | 1.1 |
| August | 47.9 | 1.8 |
| September | 64.3 | 2.5 |
| October | 71.8 | 1.9 |
| November | 71.5 | 1.9 |
| December | 72.0 | 2.7 |

The climate data were comprised of sea surface temperatures from two different sources. The first source was from the Japan Meteorological Agency (JMA), where monthly sea surface temperature (SST) anomalies are provided. These values are measured for the region between 4 °N to 4 °S and 90 °W to 150 °W, using a 2° grid over the ocean. The normal values used to calculate the monthly anomalies provided were determined based on recorded observations between 1961 and 1990, and are summarized in table 2.1.2 (JMA, 1991). These data are maintained to determine El Nino/Southern Oscillation (ENSO) events and are readily available and up to date at www.coaps.fsu.edu/products/jma_index.php. Further, with Yogyakarta being located at 7.5 °S and 110 °E (figure 2.1.1) it lies in near proximity to the JMA region resulting in high correlation with future air temperature (Focks, 2003). With these values influencing the air temperatures,

there is a subsequent effect on the mosquito population and transmission rates, as discussed in Chapter 1.

In addition to the JMA data, sea surface anomaly temperatures were also provided by the National Oceanic and Atmospheric Administration (NOAA), which are calculated using a different algorithm than the JMA temperatures. Specifically, the monthly SST anomalies were obtained for the regions covering 5 °N to 5 °S, 150 °W to 90 °W (titled "Nino3") and 5 °N to 5 °S, 170 °W to 120 °W (titled "Nino3.4"). Anomalies for these two regions were also calculated based on monthly averages between 1961 and 1990, and are summarized in table 2.1.2 as well. Figure 2.1.2 illustrates an example of the anomalies provided by NOAA based on February 2006 temperatures. Comparing the location (latitude and longitude) of these anomalies to the location of Yogyakarta in figure 2.1.1 (coloured green), the proximity of these values to Yogyakarta can be seen. Also notice that the JMA anomalies are also captured within the location of the anomalies in figure 2.1.2.

Table 2.1.2: Monthly averages of JMA and NOAA SSTs (1961 and 1990)

| Month | JMA Average SST | Nino3 Average SST | Nino3.4 Average SST |
|---|---|---|---|
| January | 25.4 | 25.61 | 26.51 |
| February | 26.2 | 26.35 | 26.69 |
| March | 26.9 | 27.08 | 27.14 |
| April | 27.1 | 27.40 | 27.69 |
| May | 26.6 | 27.06 | 27.76 |
| June | 26.1 | 26.37 | 27.49 |
| July | 25.2 | 25.58 | 27.07 |
| August | 24.6 | 24.95 | 26.70 |
| September | 24.6 | 24.82 | 26.64 |
| October | 24.6 | 24.89 | 26.60 |
| November | 24.6 | 24.95 | 26.51 |
| December | 24.9 | 25.08 | 26.48 |

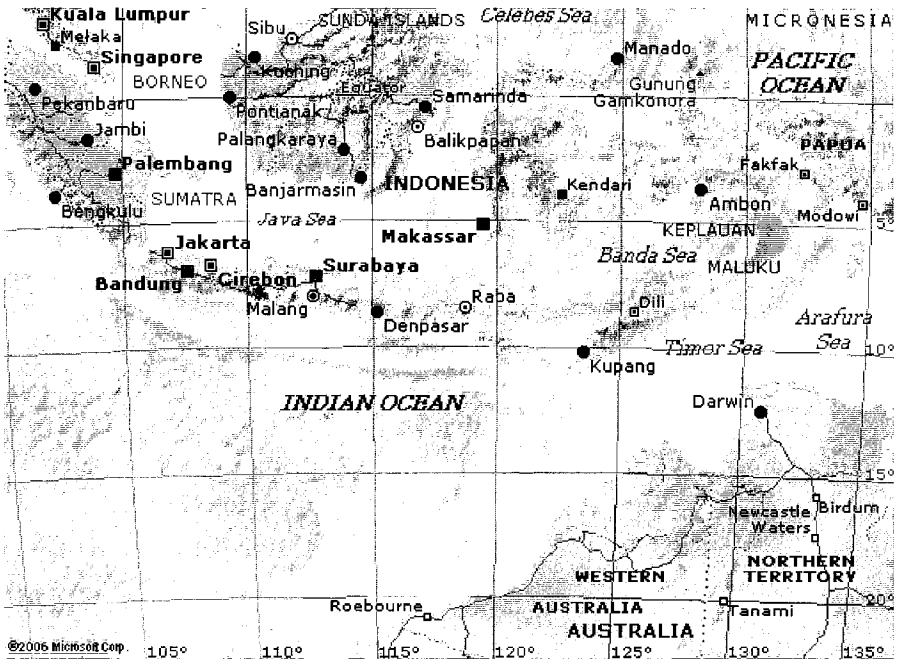


Figure 2.1.1: Map of Indonesia with green shaded area showing Yogyakarta (Borrowed from www.maps.msn.com)

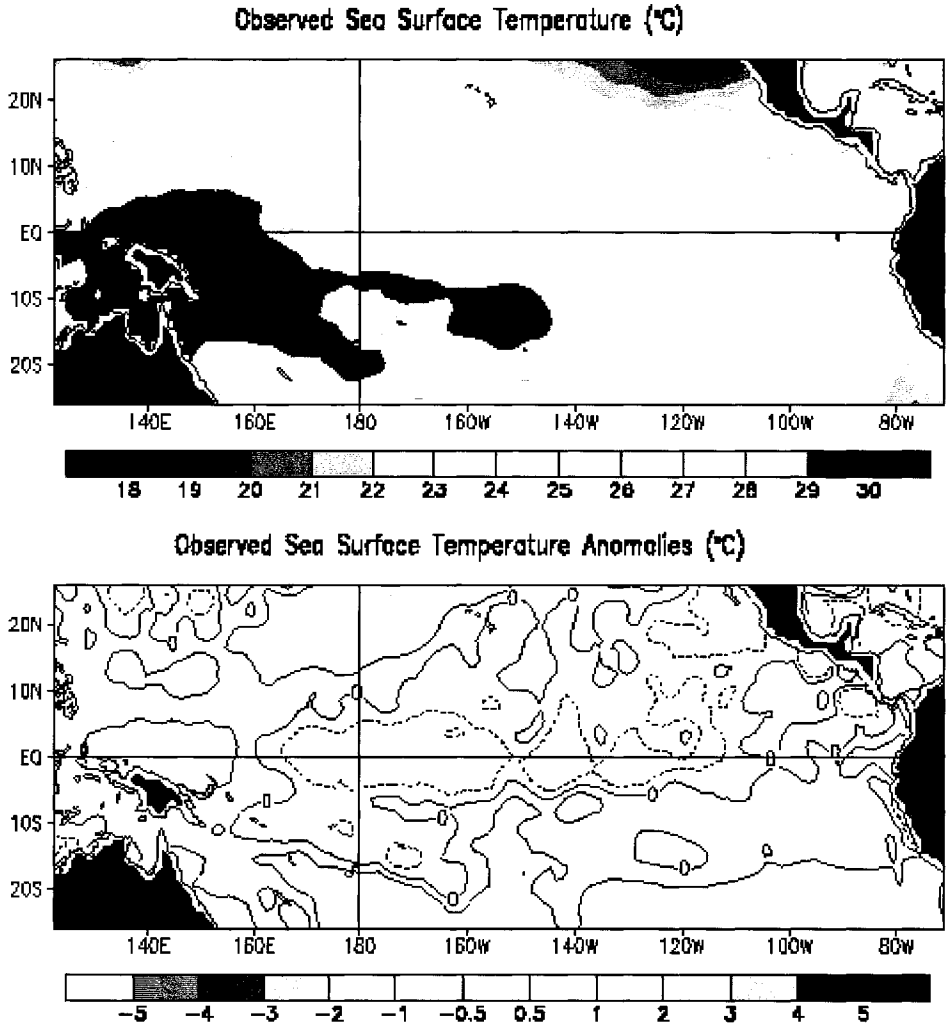


Figure 2.1.2: Sea surface temperature anomalies for February 2006, based on NOAA calculations (Borrowed from http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/enso_update/sstweek_c.gif)

Considering both the raw observations (cases, deaths, JMA SST, Nino3 SST, and Nino3.4 SST) and their anomaly observations, ten predictor variables were available. However, in order to develop predictive models based on past dengue cases and mortalities and climate, these variables needed to be time lagged. As

such, the predictor variables were lagged one to five months. In doing so, a predictive model five months in advance could be developed by using only observations recorded five months prior to the current month being predicted. A four month model could also be developed using only observation from five and four months earlier, and so on, up to one month in advance. By creating five to one month in advance models, the classification of a given month's epidemic concern could be tracked over its proceeding five months. Using time lagged variables resulted in increasing the number of predictor variables to fifty, for one month in advance models only. Of course, five month in advance models could only have the original ten variables to use in the models, as it is inappropriate to use four, three, two and one month lagged data. As such, the number of possible variables increased by ten for each model in the sequence. In addition, first order interactions between each of the appropriate predictor variables were also considered to be included in the model selection process, again substantially increasing the number of possible variables used in the early warning system models.

## Chapter 2.2: Classifying the Concern for a Dengue Epidemic

The epidemic severity of a given month was classified into one of five groups, with group 0 indicating very little concern for a dengue epidemic, and group 4 indicating very high concern for a dengue epidemic. In order to classify the predicted epidemic severity of a given month, monthly specific cutoff values to be used on the number of predicted cases were required. These selected cutoffs were primarily based on the monthly twenty-percent quantiles of dengue cases observed between June 1985 and June 1999. For some months, the cutoff ranges were slightly extended to avoid intervals that wee too restrictive.

For example, the observed number of dengue cases in June (1985 to 1999) were as follows:

```
  11  33  36  49  53  53  54
54  60  64  67  74 128 153 266
```

This yielded the following quantiles:

| 0% | 20% | 40% | 60% | 80% | 100% |
|---|---|---|---|---|---|
| 11.0 | 46.4 | 53.6 | 61.6 | 84.6 | 266.0 |

As such, the range for group 1 would only be [47, 54), which is too narrow. Therefore, this range was extended to [40, 54) without affecting the group distribution since no cases between forty and forty-seven were observed in the

June 1985 to June 1999 data. Table 2.2.1 summarizes the monthly cutoff values used for predicted epidemic severity classification.

Table 2.2.1: Level of epidemic concern classification

| Month | Group 0 | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|---|
| Jan | < 71 | [71, 101) | [101, 151) | [151, 200] | > 200 |
| Feb | < 50 | [50, 71) | [71, 111) | [111, 160] | > 160 |
| Mar | < 45 | [45, 73) | [73, 96) | [96, 130] | > 130 |
| Apr | < 51 | [51, 71) | [71, 91) | [91, 116] | > 116 |
| May | < 65 | [65, 71) | [71, 96) | [96, 130] | > 130 |
| Jun | < 40 | [40, 54) | [54, 63) | [63, 84] | > 84 |
| Jul | < 30 | [30, 46) | [46, 65) | [65, 74] | > 74 |
| Aug | < 20 | [20, 33) | [33, 56) | [56, 70] | > 70 |
| Sep | < 30 | [30, 57) | [57, 65) | [65, 80] | > 80 |
| Oct | < 20 | [20, 46) | [46, 76) | [76, 94] | > 94 |
| Nov | < 25 | [25, 50) | [50, 71) | [71, 130] | > 130 |
| Dec | < 35 | [35, 66) | [66, 81) | [81, 95] | > 95 |

## Chapter 2.3: Poisson Generalized Linear Modeling

In developing models to predict the number of dengue cases, based on past observed cases, deaths and climate, general linear models (GLM) were employed. Moreover, since the counts of monthly dengue cases were being modeled, a Poisson distribution was used (McCullagh and Nelder, 1989). Using this distribution was most appropriate for the dengue case counts as the number of cases is strongly skewed to the right, with the majority of monthly observations having small counts (figure 2.3.1). In addition, this distribution assures that all predictions will be positive, which is essential since negative values of predicted cases are illogical.

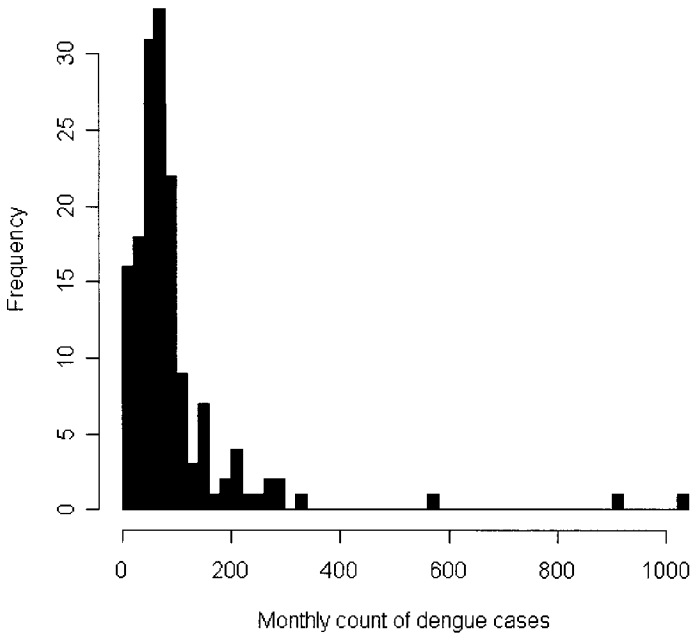


Figure 2.3.1: Monthly dengue cases (June 1985 to May 1999)

By using GLMs with a Poisson distribution, it is assumed that

$$\mathbf{Y} \mid \mathbf{X} \sim Poisson(\boldsymbol{\mu}).$$

The predictive models have the form

$$E[\mathbf{y} \mid \mathbf{x}] = \boldsymbol{\mu} = \exp\{\beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_p \mathbf{x}_p\}$$

where $\boldsymbol{y}$ is the vector of monthly observed dengue cases, $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ is the vectors of $p$ predictor variables, and $(\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the vector of estimated coefficients.

## Chapter 2.4: AICc Model Selection

To develop the predictive Poisson GLMs, stepwise model selection based on AICc was used to identify the model with the least number of predictors that best approximated the true, but unknown, model (Buckland et al. 1997). As such, the starting model contained all the appropriately time lagged predictor variables and their interactions. The model was then subsequently reduced by dropping the variable that resulted in the lowest AICc value for the remaining model. The variables previously dropped from the model were then separately reconsidered in the newly reduced model. The addition of the variable again resulting in the lowest AICc value for the model only occurred if it resulted in a lower AICc value than without the variable at hand. This sequence of dropping and adding variables to obtain the model with the lowest AICc value for the data at hand was repeated until no changes in the model resulted in a lower AICc. Unlike traditional modeling, if the two main factors in an interaction term were excluded from the model, the interaction term was not necessarily excluded. For example, if either anomaly cases or anomaly deaths were not in the best model, the interaction between these two variables was still considered. This was done because the main goal is predictability, not explanatory (Breiman, 2001).

The original Akaike information criterion, AIC, (Akaike. 1973) is calculated as

$$AIC = -2\log L + 2(m+1)$$

where $m$ represents the number of variables in the model, and log$L$ represents the log-likelihood of the model. However when the sample size is large or there are many predictor variables, AIC model selection can be too liberal and therefore may produce models that include unnecessary predictors (George, 2000). AICc provides a bias correction to the traditional AIC that is useful when the sample size is small or the number of parameters is large (Hurvich and Tsai, 1989). As mentioned previously, the number of parameters in the possible predictive models is very large, with one month in advance models having fifty potential variables, without considering their interactions. AICc is calculated as

$$AICc = AIC + \frac{2(m+1)(m+2)}{n-m-2},$$

where $n$ represents the number of observations used to develop the model, and $m$ and $AIC$ are as defined above.

With the anomaly variables having been derived directly from the raw variables by simply adding a constant, it is inappropriate to have both the raw variables and their anomaly counterparts in the models simultaneously. Therefore, it was required to determine which of the raw or anomaly variables should be used in order to obtain the best available model. For each of the five advance models

separately, this was determined by comparing the AICc values for the best stepwise model including only the following subset of variables:

1. Raw case/death data with raw SST data
2. Raw case/death data with anomaly SST data
3. Anomaly case/death data with raw SST data
4. Anomaly case/death data with anomaly SST data

The subset that yielded the stepwise model with the smallest AICc value was then deemed to be the most appropriate for the data at hand. With prediction accuracy being the goal, as opposed to understanding the mechanisms, the subset being used was free to change for each of the five predictive models. As an example, the five month in advance model using both raw case/death and SST data may be the most appropriate, while anomaly case/death data with raw SST data may be the most appropriate for the four month in advance model. This technique of choosing the most appropriate subset of the variables helped to reduce the number of possible variables and interactions in any given model, as well as eliminated many of the strong correlations existing between predictor variables. Figure 2.4.1 illustrates the correlations remaining between the raw variables. The SST variables do show strong correlations ($0.845 \leq r \leq 0.902$), indicating that multicollinearity was a concern. However, with prediction again being the main concern and not the mechanisms behind dengue incidence, the goal of model development is simply to find a function of the variables at hand that yields a good predictor of future dengue cases (Breiman, 2001). Therefore, all three of these SST variables were retained in the model selection process.

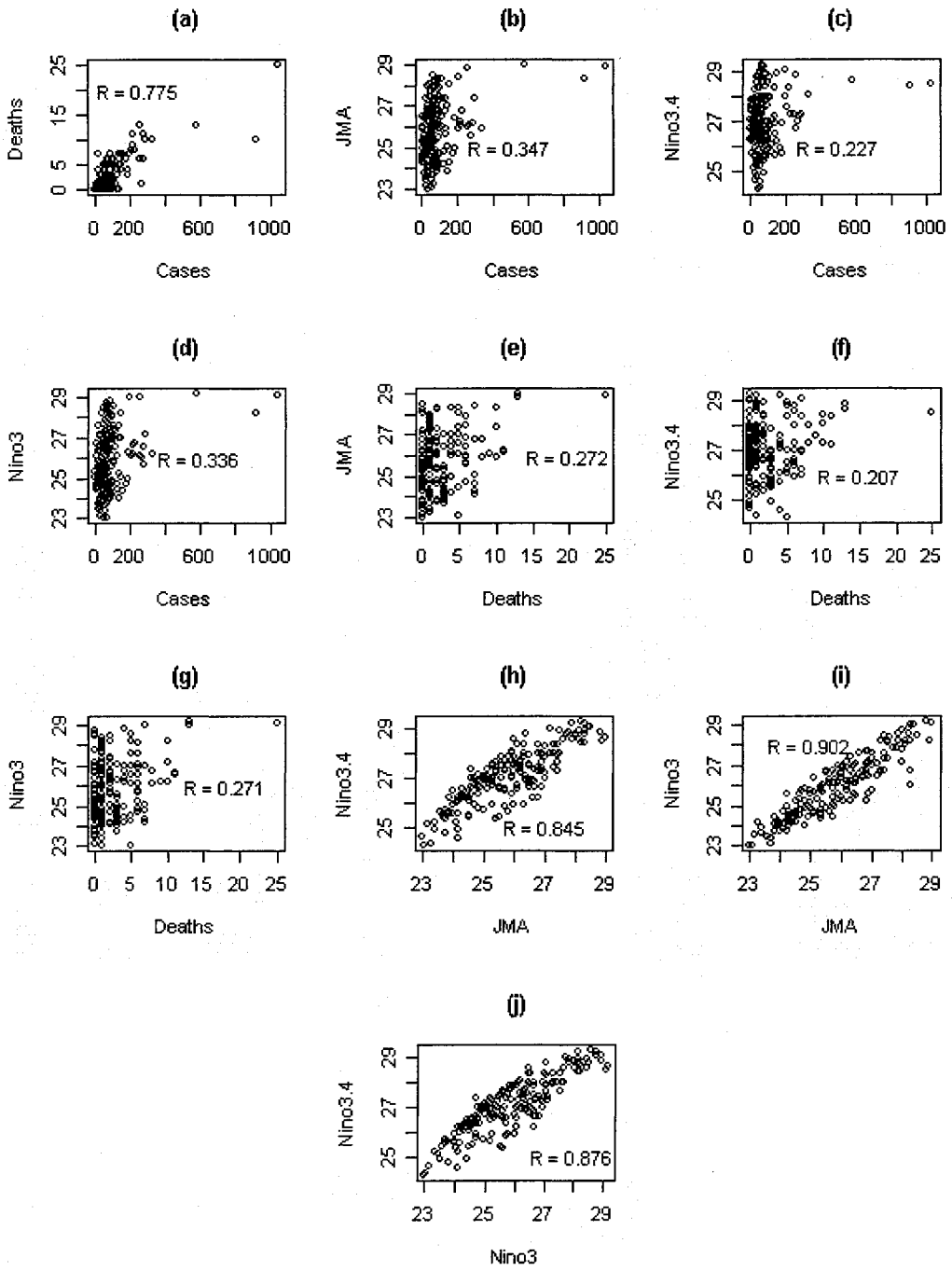


Figure 2.4.1: Scatterplots of the explanatory variables (June 1985 to May 1999)

Two sets of models for each of the four, three, two, and one month in advance predictions were developed. The first set merely contained the time lagged variables specific to the current month only. For example, the four month model would only contain the four month lagged data. These models were developed by choosing the best subset of raw/anomaly variables and the corresponding AICc stepwise model, as described above. With regards to the second set of models, the previous month(s) lags were also considered as appropriate variables. These models were developed by using the model from the first set as a base, or starting, model. The previous month(s) lags were then considered, using AICc stepwise model selection with the original model being the starting model. No interactions between lagged variables and their previous month's lags were permitted into the models. As an example, the four months in advance model was first developed using only the four month lagged data. Once the best model based only on these variables and their interactions was determined, the five month lagged data and their interactions were next considered to enter the original four month model, using AICc stepwise model selection. The interactions between the four and five month lagged variables were not permitted in the model selection.

## Chapter 2.5: Model Diagnostics

To assess the quality of the models in terms of goodness-of-fit, the raw residuals were examined. Plots of standardized residuals versus fitted values were examined for any non-uniform patterns. A uniform band of points centering around the zero standardized residual value is ideal, indicating that constant variability in the residuals has been satisfied. Plots of observed cases versus fitted/predicted cases were also examined to visually assess the accuracy. A perfect 45° line in this plot is ideal, suggesting that the observed and fitted values have a perfect match. The log-transformed version of this plot was also examined for clarity, as the majority of cases are small, with a few outliers. Q-Q plots of the residuals were also examined to assess the normality of the residuals assumptions. Again, a perfect 45° line in this plot means that the residuals follow a perfect normal distribution. Unfortunately, goodness-of-fit tests for models with higher dimensions, or many explanatory variables, are less reliable and may have little power to detect non-conformances to the assumptions of the GLM. However, cross-validation techniques can help to alleviate this problem (Breiman, 2001).

To improve the goodness-of-fit assessment for the models, leave-one-out cross-validation was also performed. This was completed by building a model based on all observations, except for the first observation. The resulting model, which is

completely independent of the first observation by design, was then used to predict the first observation. The difference between the observed and the predicted number of cases for this first data point was then recorded, as well as the observed and predicted epidemic severity group classification. This process was next completed by building a model based on all observations, except for the second observation. The discrepancy between the observed and predicted number of cases, as well as the observed and predicted epidemic severity group classification, for this second data point were then recorded. This process was repeated over the entire dataset. The leave-one-out cross-validation prediction error (CVPE) was then calculated as follows:

$$CVPE = \frac{\sum_{i=1}^{n} (Observed_i - Fitted_i)^2}{n}.$$

Clearly, the smaller the CVPE, the better the model is in terms of predictive accuracy (Stone, 1974). A 5x5 contingency table representing the agreement between observed and predicted epidemic severity group classification was also tabulated, along with the cumulative percentages of correct classification, within one group, two groups, three groups, and four groups. Ideally, a very high percentage of the cases would be correctly classified and no observations would be classified three or four groups away from its observed classification.

The empirical distribution of the absolute difference between the observed and predicted cases was also calculated and plotted to provide a visual of the predictive accuracy. This graph displays the cumulative percentages of observations with a specified absolute difference, or smaller. Graphs with steeper slopes indicate better prediction accuracies. Similarly, the empirical distribution of the absolute difference between the observed and predicted epidemic severity group classification was also calculated and plotted. Again, graphs with steeper "steps" indicate better classification accuracies.

## Chapter 2.6: Principal Component Analysis

Even after having reduced the number of variables in each of the predictive models by using AICc stepwise model selection and avoiding the use of both raw variables and their anomaly counterparts, the potential of having a large number of parameters in the model is still problematic. As a worst case scenario, if no variables were removed from the one month in advance model, it could still potentially have seventy-five variables in the model, each of which requires a parameter estimate for its inclusion in the Poisson GLM. Unfortunately, increasing the number of estimated parameters has the impact of decreasing the confidence in the resulting predicted values. In addition, the three SST variables

are highly correlated (figure 2.4.1) posing a multicollinearity problem, as mentioned in Chapter 2.4. To overcome these problems, principal component analysis (PCA) was performed on the variables remaining in the models obtained from AICc stepwise selection. Using principal components can help improve parameter estimates that are often unstable in the presence of multicollinearity (Marx and Smith, 1990), and can also serve to reduce the number of estimated parameters.

To perform PCA, the correlation matrix between the variables, including any potential interactions in the best model was utilized. By using PCA, the original variables wee transformed in terms of the patterns revealed between the variables by the resulting components (Jolliffe, 1986). Specifically, the eigenvectors and eigenvalues of the correlation matrix were obtained. The eigenvectors, or components, represent perpendicular transformations of the original variable. Their corresponding eigenvalues indicated the amount of variability observed in the original variables that was captured by each of the components. A larger eigenvalue therefore indicates that its component is more important, while a smaller eigenvalue suggests that the component is less important in capturing patterns within the data. Performing vector multiplication of the original variables (including their interactions) with the components/eigenvectors produces a set of new transformed variables (scores) that were perpendicular, or uncorrelated. As a result, multicollinearity was no longer an issue. In addition,

by selecting only components that accounted for a large proportion of the observed variability in the original variables, the number of variables used in the Poisson GLM could be reduced, resulting in fewer estimated parameters.

The newly transformed variables, or scores, were then used sequentially to estimate predictive Poisson GLMs. For example, the first model only contained one variable, corresponding to the scores from the first (and most important) component. The second model contained two variables, corresponding to the scores from the first two (most important two) components. The third model contained three variables, and so on. The full model based on all principal components produced the exact same results as the Poisson GLM prior to PCA (Chapter 2.4 models). Similar to the model diagnostics described in Chapter 2.5, the residual plots, leave-one-out CVPEs, agreement contingency tables between the observed and predicted groups, and maximum difference between observed and predicted cases were determined for each of the sequential models. The model providing the smallest CVPE was taken to be the best model. Further, the CVPE values for two sets of models pertaining to the four, three, two, and one month in advance models were compared at this point. Again, of the two models being compared, the model producing the lowest CVPE was chosen as the final and best available model for the data at hand.

## Chapter 2.7: Model Validation

The true accuracy for each of the final five, four, three, two and one month in advance predictive models was determined by comparing the observed values the models were intended to predict, and the predicted values the models yielded. The predictive models and coefficients were determined based on June 1985 to May 1999 data only, with the intention of predicting observations between June 2000 and May 2001. As such, the true accuracy of the models was determined by first obtaining the new scores for the principal components using the appropriate variables observed in June 2000 to May 2001. These new scores were then used as the values of the explanatory variables in the final predictive models. The predicted values therefore represent the actual values that would be projected if the early warning system was in place. These results were then compared to the true observations the models were intended to predict, to obtain the true accuracy. The prediction error of each of the five models was calculated as

$$PE = \frac{\sum_{i=1}^{n} (Observed_i - Fitted_i)^2}{n},$$

which is the identical calculation as the CVPE. The maximum difference between the observed and predicted number of cases was also reported, along with the contingency tables summarizing the agreement between the observed and predicted groups.

In addition to calculating the prediction errors and contingency tables for group agreement, each of the months between June 2000 and May 2001 were randomly assigned to one of the five groups. The cumulative percentages of correct classification, within one, within two, within three and within four groups were then calculated and compared to the values obtained by the five predictive models developed. The values obtained by random selection served as a guideline to indicate if the five predictive models performed classification better than mere random allocation.

## Chapter 2.8: Modeling over the Years

The above process of developing a sequence of predictive models and evaluating their accuracies was completed five times in total. The first process described in detail above utilized the June 1985 to May 2000 as "training" data to build models for June 2000 to May 2001 prediction. However, because validation data was also available from June 2001 to May 2002, the process was repeated to develop predictive models for this particular time range as well. These models were developed using the June 1985 to May 2001 data however, meaning that an extra twelve observations were included in the dataset. Similarly, to predict June 2002

to May 2003 the models were developed using the June 1985 to May 2002 data. Predicting June 2003 to May 2004, models were based on June 1985 to May 2003 data, and finally predicting June 2004 to August 2004, models were based on June 1985 to May 2004 data. Table 2.8.1 summarizes the training datasets used to create each of the models, and the validation data they were intended to predict.

Table 2.8.1: Training and validation datasets for the early warning system models

| Model | Training data | | Validation data | |
|---|---|---|---|---|
| | Years | N | Years | N |
| 1 | June 1985 - May 2000 | 180 | June 2000 - May 2001 | 12 |
| 2 | June 1985 - May 2001 | 192 | June 2001 - May 2002 | 12 |
| 3 | June 1985 - May 2002 | 204 | June 2002 - May 2003 | 12 |
| 4 | June 1985 - May 2003 | 216 | June 2003 - May 2004 | 12 |
| 5 | June 1985 - May 2004 | 228 | June 2004 - Aug. 2004 | 3 |

This sequencing of models was developed to ensure that the most recent, and likely the most related data for a specific time of prediction would also be included. Using this approach in practice is highly feasible, and ensures that the predictive models are being updated on a yearly basis to accommodate any changes in the relation between current dengue disease incidence, and past cases and climate.

# Chapter 3: RESULTS

The results for June 2003 to May 2004 prediction are discussed in detail throughout Chapter 3.1, being the final year for which all twelve months in the validation dataset were available. Although the results for the remaining validation years are not identical, they are not immensely different either, as shown in Chapter 3.2. Recall that the methodology, fully described in Chapter 2, used to develop all of the early warning system models were identical, regardless of the year. For complete model details and estimates not provided in efforts to eliminate redundancy, please contact the author.

## Chapter 3.1: June 2003 through May 2004 Models

The reported Dengue cases between June 2003 and May 2004 were best predicted by the four, two and one month models and as such only these three models will be described in detail in efforts to avoid redundancy. Even though the actual prediction error for the five month in advance model was not the lowest of all the models, it failed to reflect any of the lows or highs in dengue cases observed for this year, as it essentially only provided average monthly counts of dengue

disease. As such, this model was deemed to be of little help, perhaps because the case, death and SST data were too early to accurately capture the current incidence of dengue and the future climate in Yogyakarta. Similarly, the three month in advance model also only provided approximately average counts, again resulting in too smooth of a yearly trend that failed to capture any of the observed lows or highs. However, the predictions were not as smooth as for the five month model, likely due to current population vulnerability being slightly better captured by the case and death data.

### Chapter 3.1.1: Four Month in Advance Model

Using the June 1985 through May 2003 four month lagged anomaly case, death and SST data with the raw five month lagged case, death and SST data, provided the model with the lowest AICc. The original GLM contained nineteen variables, and had a CPVE of 3068.9. Using PCA reduced the model dimensionality and eliminated any multicollinearity issues, by using only the first sixteen principal components. The first sixteen principal components accounted for 99.997% of the variability in the original explanatory variables. Based on CVPE, the GLM using these sixteen components was further improved by excluding the $2^{nd}$, $7^{th}$, $9^{th}$, and $14^{th}$ components (Appendix, table A.1), which had a lower CVPE of

2789.3, compared to 3006.5 with all sixteen components. Residual diagnostics for this model (Appendix, figure A.1) indicated less accuracy as predictions became smaller. This is likely due to the fact that there were many more observations with few dengue cases, and as such the variability is naturally larger for the smaller dengue cases. Of course, it would be inappropriate to eliminate the few large recorded dengue cases as outliers, because these are the months of particular interest and importance. The spread of positive residuals was also wider than negative residuals, indicating more severe underestimations, compared to overestimations, since the residuals are defined as "Observed-Fitted."

The four month in advance model for June 2003 to May 2004 observations was successful in that it was able to correctly predict the classification of three months, with five months being classified within one group of its true observation, eleven months within two groups, and all twelve months within three groups (table 3.1.1.1). Unfortunately, two of the months observed as very high epidemic concern (group 4) were only predicted as medium concern (group 2).

Table 3.1.1.1: Four months in advance agreement between predicted and observed groups (June 2003 to May 2004)[1]

| | true.0 | true.1 | true.2 | true.3 | true.4 |
|---|---|---|---|---|---|
| pred.0 | 0 | 0 | 0 | 0 | 0 |
| pred.1 | 0 | 2 | 0 | 1 | 0 |
| pred.2 | 3 | 2 | 1 | 0 | 2 |
| pred.3 | 1 | 0 | 0 | 0 | 0 |
| pred.4 | 0 | 0 | 0 | 0 | 0 |

| | Count | Percentage | Cumulative |
|---|---|---|---|
| correct | 3 | 25.000 | 25.000 |
| within1 | 5 | 16.667 | 41.667 |
| within2 | 11 | 50.000 | 91.667 |
| within3 | 12 | 8.333 | 100.000 |
| within4 | 0 | 0.000 | 100.000 |

**Max.diff**
177.55

**Prediction.error**
4985.9

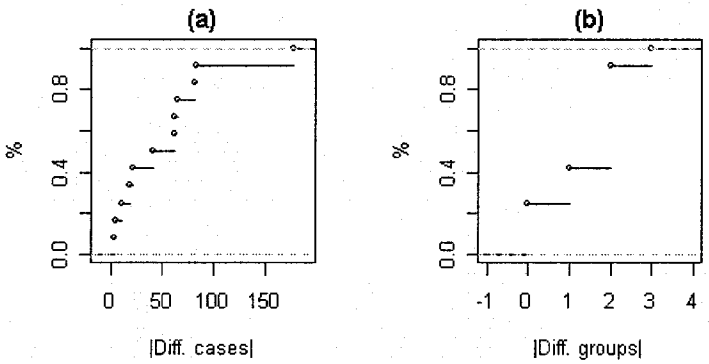


Figure 3.1.1.1: Four month in advance prediction accuracy (June 2003 to May 2004)

[1] The column names "true.0", "true.1", etc. indicate the observed group classification for each month. The row names "pred.0", "pred.1", etc. indicate the predicted group classification for each month. The rows "correct", "within1", etc. indicate if the predicted classifications were correctly classified, classified within 1 group of the observed classifications, etc. The subsequent columns "count", "percentage" and "cumulative" represent the number of months, percentage of months and cumulative percentage of months for each row.

## Chapter 3.1.2: Two Month in Advance Model

The original two month in advance GLM contained thirty-five variables, and used the combination of two through five month lagged raw case/death data with two, three and five month lagged raw SST, and four month anomaly SST data. However PCA substantially reduced the model dimensionality and eliminated multicollinearity, by using only the first fifteen principal components, accounting for 99.5% of the variability observed in the original thirty-five variables. A model using only the scores from the first fifteen components had a CVPE of 2859.9 was much lower than 14411.0 CVPE from the original model. Removing the $9^{th}$ component further improved the model, with an even lower CVPE of 2753.1. For details on the estimated coefficients for each component in the GLM, see table A.2 in the Appendix. The residual diagnostics again indicated less accuracy as predictions became smaller, with more severe underestimates due to wider spread of the positive residuals (Appendix, figure A.2).

Based on prediction error, the two month in advance model for June 2003 to May 2004 observations was not as successful as the four month model, with a much higher error of 6420.6, compared to 4985.9. However, the group classifications were improved with one of the high epidemic concern months (group 4) now being accurately classified, while the other is predicted as little concern (group 1).

Two months are overestimated, resulting in high epidemic concern classification when their true observations were very low concern. With early warning systems however, it is better to overestimate than underestimate as underestimations will result in unexpected, and unplanned for, dengue epidemics. Thirty-three percent of the months are now correctly classified, which is an improvement compared to the twenty-five percent from the four month model (table 3.1.2.1).

Table 3.1.2.1: Two months in advance agreement between predicted and observed groups (June 2000 to May 2001)

| | true.0 | true.1 | true.2 | true.3 | true.4 |
|---|---|---|---|---|---|
| pred.0 | 0 | 0 | 0 | 1 | 0 |
| pred.1 | 1 | 3 | 1 | 0 | 1 |
| pred.2 | 1 | 1 | 0 | 0 | 0 |
| pred.3 | 0 | 0 | 0 | 0 | 0 |
| pred.4 | 2 | 0 | 0 | 0 | 1 |

| | Count | Percentage | Cumulative |
|---|---|---|---|
| correct | 4 | 33.333 | 33.333 |
| within1 | 7 | 25.000 | 58.333 |
| within2 | 8 | 8.333 | 66.666 |
| within3 | 10 | 16.667 | 83.333 |
| within4 | 12 | 16.667 | 100.000 |

Max.diff
165.88

Prediction.error
6420.6

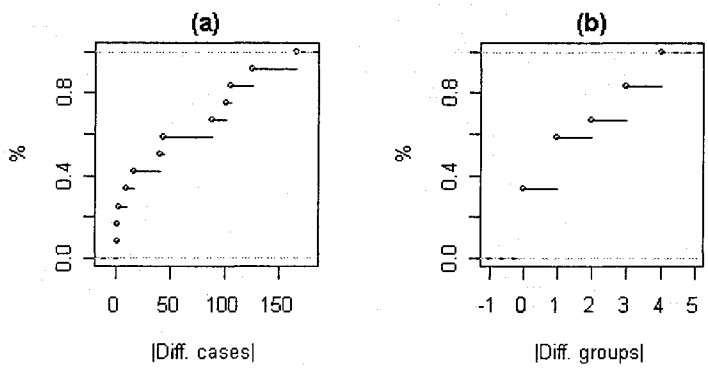


Figure 3.1.2.1: Two months in advance prediction accuracy (June 2003 to May 2004)

Recall that the cutoffs used for classification were determined from a statistical point of view and may not represent the best choice from an epidemiologist's perspective. In contrast, prediction error is not subjective and is therefore a better indicator of model performance in comparison to the cumulative agreement percentages. With this in mind, the four month model appears to better predict June 2003 to May 2004. This is likely due to climate still being an important driving factor for dengue disease, which four month in advance SST would better capture than two month in advance SST values.

### Chapter 3.1.3: One Month in Advance Model

Originally, forty-two variables were used in the GLM, consisting of one, two, four and five month lagged raw case, death, and SST data, along with three month anomaly case, death, and SST data. PCA was unable to reduce the dimensionality of the model, as the lowest CVPE was 2356.0 when all components were used, although multicollinearity was still resolved. Once the $19^{th}$, $25^{th}$, $26^{th}$, $27^{th}$, $30^{th}$, $31^{st}$, and $40^{th}$ components were excluded, the CVPE lowered to 1407.7. Table A.3 in the appendix shows the estimated one month in advance GLM model, while figure A.3 in the appendix still indicated less accuracy as predictions became smaller, except that the residuals now show similar spread in both the

positive and negative direction. This suggests that underestimates are no longer more severe than overestimates, unlike the four and two months in advance models.

The prediction error for the one month in advance model was 11848.0, which is very large compared to the four and two month models (4985.9 and 6420.6, respectively). Closer examination of the one month in advance predictions indicated that this model predicted April 2004 to have 330.2 Dengue cases, when it only had eleven recorded cases. Thus, April was overestimated by 319.2, which contributed to almost seventy-two percent of the prediction error. Excluding this one month's prediction, the prediction error greatly reduced to only 3367.9, which is lower than both the four and two month models. Again, overestimations are preferred to underestimations for an early warning system, as its better to be safe than sorry with regards to tracking potential epidemics. Considering the two months observed as very high epidemic concern, one month was again correctly classified, while the other month was classified as medium concern (group 2), which offered an improvement to both the four and two month models. Thirty-three percent of the months were still correctly classified (table 3.1.3.1), as with the two month model.

Table 3.1.3.1: One month in advance agreement between predicted and observed groups (June 2003 to May 2004)

| | true.0 | true.1 | true.2 | true.3 | true.4 |
|---|---|---|---|---|---|
| pred.0 | 0 | 0 | 0 | 1 | 0 |
| pred.1 | 2 | 3 | 1 | 0 | 0 |
| pred.2 | 0 | 1 | 0 | 0 | 1 |
| pred.3 | 1 | 0 | 0 | 0 | 0 |
| pred.4 | 1 | 0 | 0 | 0 | 1 |

| | Count | Percentage | Cumulative |
|---|---|---|---|
| correct | 4 | 33.333 | 33.333 |
| within1 | 8 | 33.333 | 66.666 |
| within2 | 9 | 8.333 | 74.999 |
| within3 | 11 | 16.667 | 91.666 |
| within4 | 12 | 8.333 | 100.000 |

Max.diff
319.02

Prediction.error
11848

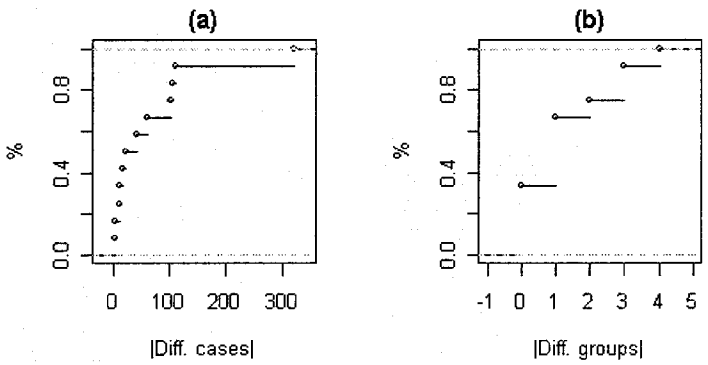


Figure 3.1.3.1: One month in advance prediction accuracy (June 2003 to May 2004)

## Chapter 3.1.4: Model Comparison

The reported dengue cases between June 2003 and May 2004 were best predicted by the one, two, and four month in advance models. Based on actual prediction

errors, the one month performed the best if the overestimation for April was unaccounted for, followed by the four, and then two month model. This is not extremely surprising, given that the one month model would be expected to provide the most accurate predictions, with current states of dengue incidence being the main driving factor at this point in time. It may seem slightly surprising that the four month model outperformed the two month model with respect to prediction error, but as previously mentioned, this may be contributed to the fact that the four month model better captured the future climate conditions, for which the two month model is too late to capture. Further, four month in advance climate may be more important than the current dengue incidence two month in advance.

Figure 3.1.4.1 compares the predicted trend for each of the one, two, and four month in advance models to the actual recorded dengue cases for June 2003 to May 2004. Examining this figure clearly shows that the one month model best captured the yearly trend of dengue cases, while the two month in advance model seems to capture the trends second best, as it actually shows an increase in dengue cases for March. This is extremely important for an effective early warning system. Although the four month in advance model does depict increasing predictions as the observed cases increases, the observed spike in March is not clearly captured by this model. This therefore indicates that the two month in advance model does indeed provide a better warning than the four month model,

which contradicts the results based on prediction error alone. This clearly demonstrates the need to examine the general trend any given model is producing with regards to predicted dengue cases, and not just the prediction error alone. Also notice that the one month in advance model greatly overestimated April, as mentioned earlier, missing the observed peak by one month. Fortunately however, the one month in advance model still managed to flag March as high epidemic concern, especially compared to the lower predictions for the previous months.

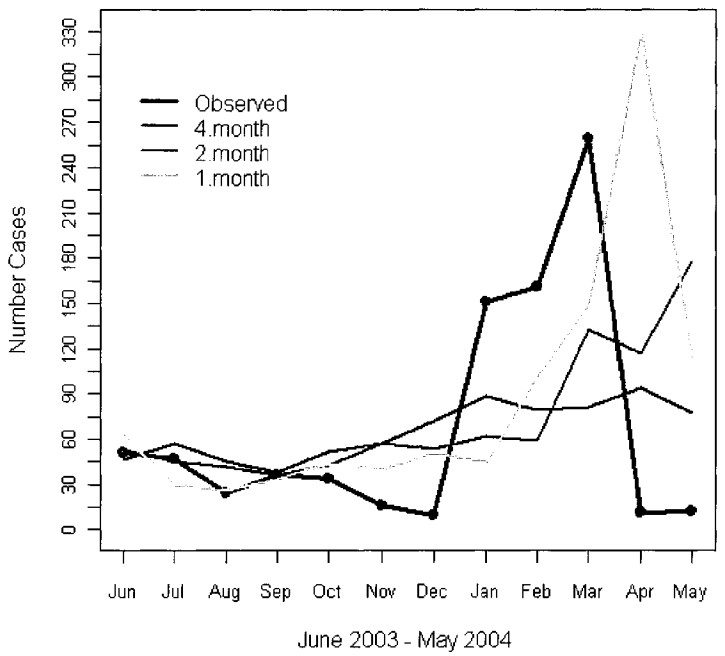


Figure 3.1.4.1: Observed and predicted number of cases from the three predictive models (June 2003 to May 2004)

With regards to the classification of epidemic severity resulting from model prediction, the one and two month in advance models both correctly classified 33.3% of the observations. For the remaining cumulative percentages, the one month in advance model out performed the two month in advance model, however the four months in advance model appears to have performed better than the one month model, based on the cumulative percentages of agreement alone. All three of the models have larger cumulative percentages of agreement than random allocation alone (table 3.1.4.1), demonstrating that the predictive models do indeed perform better than random classification alone, as hoped. Once again, it must be emphasized that the cutoffs used for classification of epidemic severity were determined using a statistical approach, as opposed to an epidemiology point of view. Therefore the groupings used may not be the most appropriate.

Table 3.1.4.1: Cumulative group agreement percentages[1]

| | 4.month | 2.month | 1.month | Random |
|---|---|---|---|---|
| **Correct** | 25.0 | 33.3 | 33.3 | 0.0 |
| **within1** | 41.7 | 58.3 | 66.7 | 16.7 |
| **within2** | 91.7 | 66.7 | 75.0 | 50.0 |
| **within3** | 100.0 | 83.3 | 91.7 | 83.3 |
| **within4** | 100.0 | 100.0 | 100.0 | 100.0 |

[1] The columns "4.month", "2.month", "1.month" and "Random" represent the predicted classifications from the 4, 2 and 1 month in advance model, followed by random classification respectively. The rows "correct", "within1", etc. indicate the cumulative percentages of correctly predicted classifications, predicted classifications within 1 group of the observed classifications, etc.

## Chapter 3.2: Model Overview for the Remaining Years

### Chapter 3.2.1: June 2000 through May 2001 Models

For June 2000 to May 2001 prediction, the prediction errors for the five through one month in advance models were 2186.3, 2812.7, 2461.4, 2576.5, and 1536.2, respectively. Based on these values, the one month in advance model was again the best, next followed by the five month, then three month in advance models. However the trends presented in figure 3.2.1.1 indicate that the five month model again just simply provided average predictions and failed to capture the observed trend. In contrast, the four month in advance model does appear to somewhat capture the trends, although it also seems to be predicting highs and lows two to three months too early, and fails to spike in May. The three month in advance model again appears too smooth and unable to capture the overall trend, while the two month model also seems inadequate. Finally, the one month in advance model does parallel the observed trend nicely, with the exception of the drop in prediction in April and too low of a prediction for May.

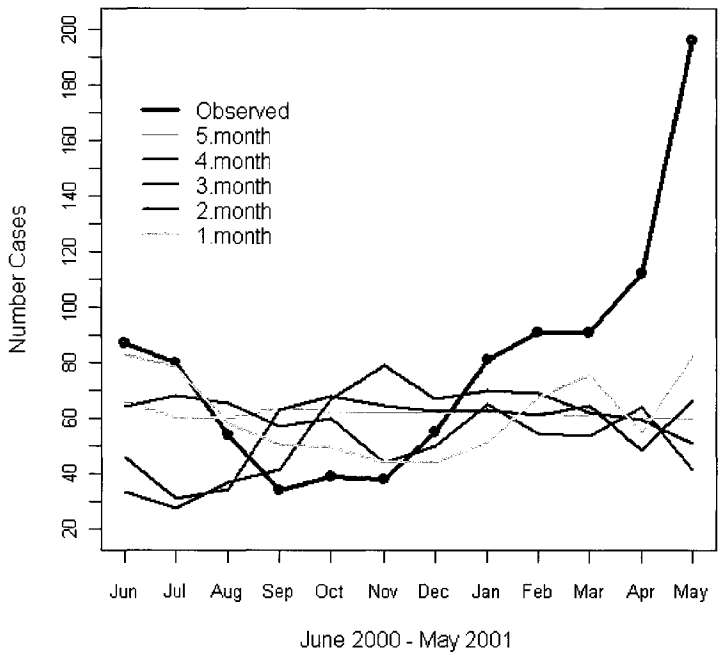


Figure 3.2.1.1: Observed and predicted number of cases from the five predictive models (June 2000 to May 2001)

## Chapter 3.2.2: June 2001 through May 2002 Models

The prediction errors for the five through one month in advance models were 391.2, 1384.3, 979.7, 692.5, and 194.2 respectively. Accordingly, the one month in advance model was again the best, next followed by the five, then two month in advance models. In addition, the June 2001 to May 2002 predictions yielded the smallest prediction errors of all the years predicted. This however is likely due to the fact that little variability was actually observed (range of 32 to 94) in this

year's dengue cases, and there was no evidence of a dengue epidemic. As usual, the five month in advance model offers no real assistance in an early warning system, with predictions being too moderate. Both the three and four month models seem to completely miss the actual observed trend in that the opposite appears to be predicted. Both the one and two month in advance predictions do roughly parallel the yearly trend, with the one month model being slightly more accurate. Again, no real epidemic concerns are present in this data, as indicated by all the predictive models.

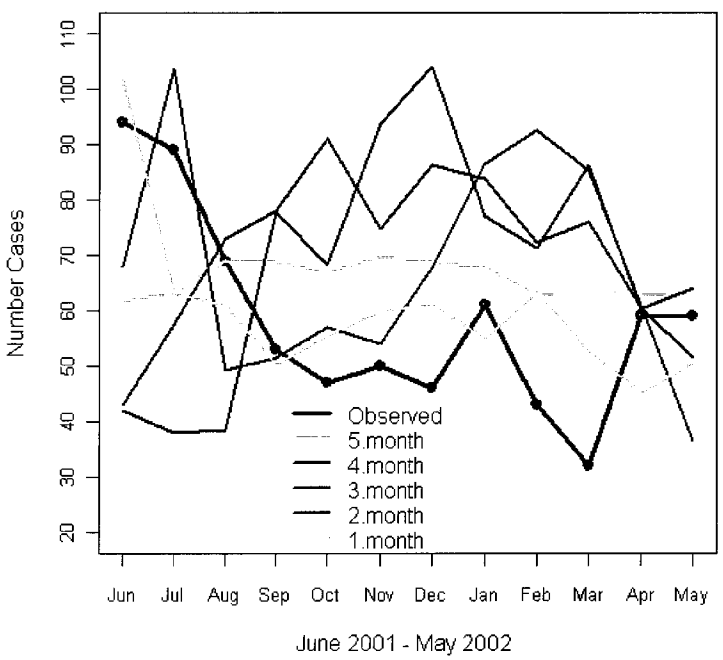


Figure 3.2.2.1: Observed and predicted number of cases from the five predictive models (June 2001 to May 2002)

## Chapter 3.2.3: June 2002 through May 2003 Models

The prediction errors for the five through one month in advance models were 3843.6, 866.1, 2257.9, 954.3, and 535.0, respectively. As with all the previous models, the one month in advance model again offered the best predictions, followed by the four and then two month in advance models. These results closely parallel the results for June 2003 to May 2004, discussed in detail. Also notice that the range of dengue cases for this year is again fairly small (3, 78), with no real concern of an epidemic. Figure 3.2.3.1 shows that all the models generated overestimates and were therefore more conservative than the actual records. The one, two and four month predicted trends parallel the observed trend better than the five and three month in advance predictions. Further, the four month in advance model again appears to be making predictions a couple of months too early.

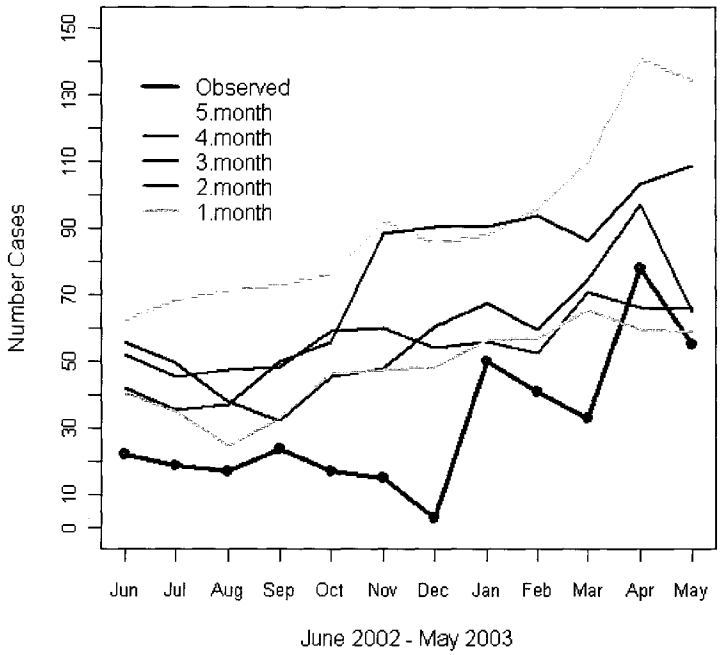


Figure 3.2.3.1: Observed and predicted number of cases from the five predictive models (June 2002 to May 2003)

## Chapter 3.2.4: June 2004 through August 2004 Models

With only three months available for 2004 validation, it is difficult to truly assess the accuracy of these predictive models. Regardless, the prediction errors for the five through one month in advance models were calculated as 4968.8, 2834.5, 1815.9, 1045.4, and 1135.5, respectively. For these three months, it therefore appears as though the two month in advance model provided the most accurate predictions, closely followed by the usual one month in advance model. Neither

of the three, four or five month models appear to be adequate. These results are supported by the three month trends depicted in figure 3.2.4.1 below.

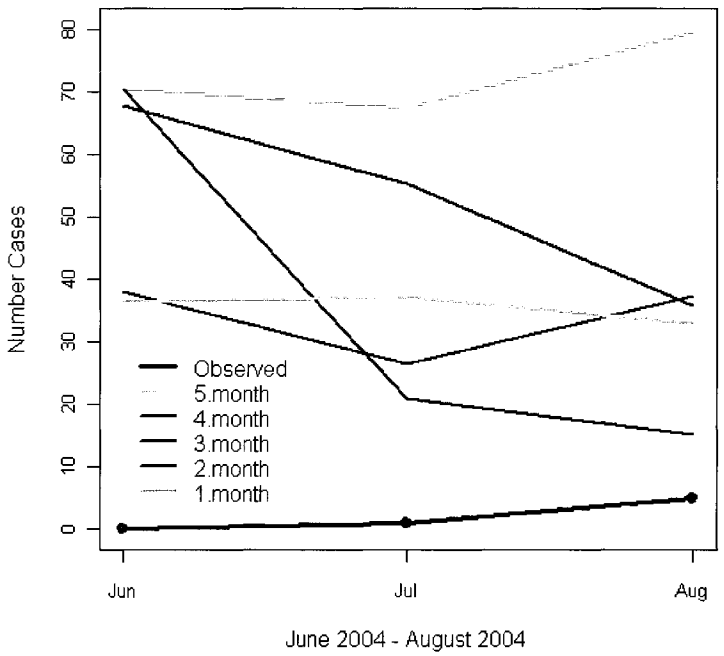


Figure 3.2.4.1: Observed and predicted number of cases from the five predictive models (June 2004 to August 2004)

## Chapter 3.3: Overview of the Models

The actual prediction errors for each of the five models over the five predicted years ranged from a minimum of 194.2 (one month in advance model for June 2001 to May 2001 prediction) to a maximum of 11848.4 (one month in advance

model for June 2003 to May 2004 prediction). However almost seventy-two percent of the latter prediction error can be contributed to the model overestimating the number of dengue cases in April 2004 by 319.2 cases. Therefore, without this one month, the prediction error reduces to a much more reasonable 3357.7.

Unfortunately, the 196 dengue cases observed in May 2001 was not captured by any of the five models for that particular year's prediction, with its largest prediction being 82.3 cases by the one month in advance model. The less extreme observation of 112 cases in April 2002 was also not well captured by any of the models, since its largest prediction was only 64.2 by the two month in advance model. June 2001 through May 2003 did not present any extreme observations in terms of the number of dengue cases, with the largest number of cases being 78 in April 2003. With no observed epidemic concerns here, the models were generally conservative, yielding larger estimates than actually observed. The peak number of cases for the June 2003 to May 2004 year occurred in March, with 259 cases. This severity was somewhat captured by the two and one month models, having predictions of 132.5 and 149.0 cases respectively. As previously mentioned, figure 3.1.4.1 suggests that the one-month in advance model was off by one month for this prediction, since the following month (April) was predicted to have 330.2 cases. Finally, with regards to the June 2004 to August 2004 observations, which were all very low (only 0 to 5 cases), the models all produced

overestimates, as one might expect. In addition, the models were indeed capable of producing classifications of the dengue epidemic severity that were better than mere random classification alone, as shown in table 3.1.4.1.

The one and two month in advance models captured the overall "up and down" trends in dengue cases occurring each year. To a lesser extent, so did the four months in advance models, except that they tended to offer predictions that were between one to three months too early. The three and five month in advance models typically provided similar estimations to each other, and produced yearly trends that were too smooth to be of much help in an early warning system. For implementation purposes, it would be most reasonable to visually examine the trends (such as those presented in figures 3.1.4.1 through 3.2.4.1) in the predicted dengue cases each model is offering, and be more prepared to increase mosquito control efforts when the models produce peaks or gradual increases in their predictions, paying closer attention to the one, two, and four month in advance models. For the complete datasets, R programs used to develop the predictive models, and resulting models, please contact the author.

# Chapter 4: DISCUSSION

## Chapter 4.1: Other Options for Predictive Models

In addition to considering first order interactions between each of the basic five variables used in the models, various transformations were also considered. Using the log-transformed cases as opposed to the original form, was considered as the response variable, however this did not improve the linear relationship between the number of cases and the exponential of the explanatory variables. Similarly, using functions such as the square and inverse of the explanatory variables did not improve the model fits.

Although Poisson GLMs were used, the observed variances in the number of dengue cases were much larger than the observed means. To account for this violation of the Poisson distribution, the negative binomial distribution could have instead been used to develop the GLMs. By doing so, the models accounted for extra variation in the observed number of cases. Although these models produced lower AICc values due to improvements in the log-likelihood, the trend in the residuals observed with the simpler Poisson GLMs were still present with these more complex models. Further, the CVPE values were actually *higher* for these

models, compared to the Poisson GLMs. With the main goal of the early warning system models being predictive accuracy, the increased CVPE values indicated that the negative binomial GLMs were not as accurate as the Poisson GLMs, with regards to prediction capabilities.

As the first step to model reduction, stepwise model selection was used based on AICc values. This method was chosen as a fast way to reduce the number of possible variables and their interactions in each of the predictive models. However this method of model selection is not fool proof and may result in a model far away from the most optimal model, given the data at hand (Hawkins, 1973). For this reason, both backward and forward model selection alone were also explored, again based on AICc, and compared to the model obtained using stepwise selection. Usually the three models obtained were identical, but on the rare occasion they were slightly different, due to one or two changes in the interaction(s) included. However, comparing these three models developed, the AICc values for the stepwise models were always the smallest.

An alternative to AICc based stepwise model selection would have been to choose the models based on CVPEs for each of the possible subset of models, and chose the model resulting in the lowest CVPE (Stone, 1977). However, this would have been extremely time consuming and would limit the ease of actually implementing this procedure to develop predictive models for an early warning

system in practice. The current choice of methodology for developing the predictive models ensures a reasonably fast and easy method that can be used in practice for early warning system, while producing reliable results. This is an extremely important concept, as using complex methodology would limit the practicality of such models and would thereby render the early warning system useless.

## Chapter 4.2: Other Options for Explanatory Variables

The previous case and death records along with sea surface temperatures were capable of producing fairly accurate results, even though they did have difficulty in capturing some of the extreme number of cases, which somewhat jeopardizes the reliability of the models. The advantage of this data is that it is easily available and updated on a monthly basis, which allows for quick and timely development of the models for use in an early warning system. Unfortunately however, they may not be the best predictors for dengue cases, as shown by the large residuals when fewer monthly cases occur and the less than perfect predictions they yielded for the validation datasets. These results could potentially be improved by gaining access to other climate and vulnerability indicators. For example, rainfall, humidity, and temperature could be used as

alternate climate variables. A count of the number of water storage containers in the region could provide an indicator of population vulnerability, based on the number of breeding sites available for the *Aedes aegypti.* The availability and frequency of updates for these variables would first need to be considered before their inclusion would be of much help though, regardless of how well they improve the predictive accuracy and reliability of the models. If the data are not readily available and continuously up-to-date, they will obviously be of little help in a real-time early warning system.

## Chapter 4.3: Classification Cutoffs Review

Once again, it must be highly emphasized that the cutoffs used to determine each month's level of epidemic concern was purely based from a statistical standpoint. As such, it is highly doubtful that these cutoffs are the most appropriate for defining and classifying epidemic concern. Is seventy-one to one-hundred (cutoffs for January, group 1) dengue cases for January really not considered to be of great concern? It is difficult to answer this question without considering the social and economic impact this level of cases has on society, or the future impacts on spread of dengue this would have. For this reason, it is difficult to assess the performance of the models based on the cumulative percentages of

agreements presented, as it is unknown how reasonable and representative the cutoffs used actually are. One possible modification that could be used for the cutoffs is to consider the mean number of cases for each month, along with their standard deviations. This would therefore be more representative of the monthly norms and variability. Again, this is from a statistical view however, and may not be truly representative of dengue concern from an epidemiology point of view.

## Chapter 4.4: Final Comments

These models would be extremely easy and practical to actually implement in a real-time early warning system for dengue disease in Yogyakarta, Indonesia. The data used to develop the models are updated on a monthly basis, and readily available. As well, the methodology created to develop each of the predictive models was not overly complex, nor are they overly time consuming. The models developed for June 2000 to August 2004 do not have a perfect track record, in the sense that they did not capture every single month with higher than usual incidences of dengue. They were however capable of suggesting general increasing and decreasing trends in the number of cases that would be appropriate to use as early warnings. Further, many of the models erred on the side of caution by providing overestimates of the number of dengue cases. For obvious reasons,

this is clearly much more useful than models that consistently underestimated and provided no indication of months considered to be high epidemic concern. The goal was to provide models that could be used in an early warning system – models that could help raise awareness and provide advance warnings with regards to elevated number of dengue cases. Overall, the models developed here were successfully capable of providing such warnings up to four months in advance. If implemented in practice, this early warning system could help to substantially reduce the number of dengue cases and epidemics in Yogyakarta, Indonesia. Closely watching the predictions and trends offered by these models could provide valuable signals indicating when early preparation for dengue control efforts would be most beneficial.

# BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory. Ed. B.N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.

Bradley, D.J. (1993). Human tropical diseases in a changing environment, pp. 146-162. In Environmental Change and Human Health. Lake, J., Brock, G. and Ackrill, K. eds. Ciba Foundation Symposium, London UK: CIBA Foundation.

Breiman, L. (2001). Statistical modeling: the two cultures. Statistical Science (16:3), 199-231.

Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997). Model selection: an integral part of inference. Biometrics (53), 603-618.

Cullen, J.R., Chitprarop, U., Doberstyn, E.B., and Sombatwattanangkul. (1984). An epidemiological early warning system for malaria control in northern Thailand. Bulletin of the World Health Organization (62), 107-114.

DeRoeck, D., Deen, J., and Clemens, J.D. (2003). Policymakers' views on dengue fever/dengue hemorrhagic fever and the need for dengue vaccines in four southeast Asian countries. Vaccine (22), 121-129.

Focks, D.A., Brenner, R.A., Daniels, E., and Keesling, J.E. (1995). A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation and samples of simulation results. American Journal of Tropical Medicine and Hygiene (53), 489-506.

Focks, D.A. (2003). A review of entomological sampling methods and indicators for dengue vectors. Geneva, TDR, 2003 (TDR/IDE/Den/03.1).

George, E.I. (2000). The variable selection problem. Journal of American Statistical Association (95), 1304-1308.

Hawkins, D.M. (1973). On the investigation of alternative regressions by principal component analysis. Applied Statistics (22), 275-286.

Hurvich, C.M. and Tsai, C.L. Regression and time series model selection in small samples. Biometrika (76), 297-307.

JMA. (1991). Climatic charts of sea surface temperatures of the western north pacific and the global ocean. pp. 51. Tokyo: Japan Meteorological Agency.

Jolliffe, I.T. (1986). Principal component analysis. New York: Springer-Verlag.

Marx, B.D. and Smith, E.P. (1990). Principal component estimation for generalized linear regression. Biometrika (77), 23-31.

McCullagh, P. and Nelder, J.A. (1989). Generalized linear models, 2nd ed. London: Chapman and Hall.

Moore, C.G., Cline, B.L., Ruiz-Tibén, E., Lee, L., Romney-Joseph, H., and Rivera-Correa, E. (1978). *Aedes aegypti* in Puerto Rico: Environmental determinants of larval abundance and relation to dengue virus transmission. American Journal of Tropical Medicine and Hygiene (27), 1225-1231.

Myers, M.F. (2000). Forecasting disease risk for increased epidemic preparedness in public health. Advances in Parasitology (47), 309-330.

National Research Council. (2001). Under the weather: climate, ecosystems, and infectious disease. Washington: National Academy Press.

Patz, J.A., Martens, W.J.M, Focks, D.A., and Jetten, T.H. (1998). Dengue fever epidemic potential as projected by general circulation models of global climate change. Environmental Health Prospectives (106), 147-152.

Schreiber, K.V. (2001). An investigation of relationships between climate and dengue using a water budgeting technique. International Journal of Biometeorology (45), 81-89.

Sheppard, P.M., Macdonald, W.W., Tonn, R.J., and Grabs, B. (1969). The dynamics of an adult population of *Aedes aegypti* in relation to dengue hemorrhagic fever in Bangkok. Journal of Animal Ecology (38), 661-702.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, B (36), 111-147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, B (39), 44-47.

WHO. (1997). Dengue hemorrhagic fever: diagnosis, treatment, prevention and control, 2nd ed. Geneva: World Health Organization. Available online at: *http://www.who.int/csr/resources/publications/dengue/Denguepublication/en/*

WHO. (1998). Dengue and dengue hemorrhagic fever. Fact sheet no. 117. Available online at: *http://www.who.int/mediacentre/factsheets/fs117/en/index.html*

WHO. (2000). WHO report on global surveillance of epidemic-prone infectious diseases: Dengue and dengue hemorrhagic fever. Available online at: *http://www.who.int/csr/resources/publications/dengue/CSR_ISR_2000_1/en/print.html*

WHO. (2004). Using climate to predict disease outbreaks: a review. Available online at: *http://www.who.int/globalchange/publications/oeh0401/en/print.html*

WHO. (2005). International travel and health: situation as on 1 January 2005, pp. 76. Available online at: *http://whqlibdoc.who.int/publications/2005/9241580364_chap5.pdf*

# APPENDIX

Table A.1: Four months in advance GLM (June 2003 to May 2004 prediction)[1]

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 4.26390 | 0.00841 | 506.99 | < 2e-16 | *** |
| Comp.1 | -0.07168 | 0.00230 | -31.20 | < 2e-16 | *** |
| Comp.3 | 0.17749 | 0.00393 | 45.17 | < 2e-16 | *** |
| Comp.4 | 0.07698 | 0.00689 | 11.17 | < 2e-16 | *** |
| Comp.5 | -0.25097 | 0.01071 | -23.43 | < 2e-16 | *** |
| Comp.6 | 0.20672 | 0.01081 | 19.12 | < 2e-16 | *** |
| Comp.8 | 0.19520 | 0.02019 | 9.67 | < 2e-16 | *** |
| Comp.10 | -0.54090 | 0.02987 | -18.11 | < 2e-16 | *** |
| Comp.11 | 0.53176 | 0.04621 | 11.51 | < 2e-16 | *** |
| Comp.12 | -0.83561 | 0.07689 | -10.87 | < 2e-16 | *** |
| Comp.13 | 0.42218 | 0.07615 | 5.54 | 3.0e-08 | *** |
| Comp.15 | 2.52589 | 0.10033 | 25.17 | < 2e-16 | *** |
| Comp.16 | -1.15959 | 0.10474 | -11.07 | < 2e-16 | *** |

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

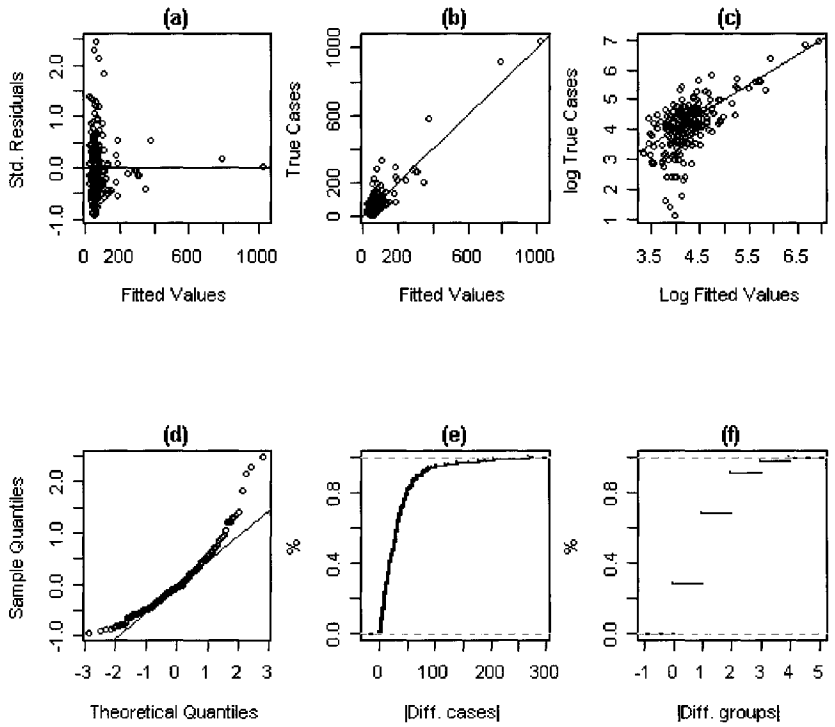


Figure A.1: Four month in advance GLM diagnostics

[1] The row names "Comp.1", "Comp.3", etc. indicate component 1, 3, etc. from principal component analysis.

Table A.2: Two months in advance GLM (June 2003 to May 2004 prediction)

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 4.24850 | 0.00845 | 502.79 | < 2e-16 | *** |
| Comp.1 | -0.06397 | 0.00161 | -39.75 | < 2e-16 | *** |
| Comp.2 | 0.06220 | 0.00257 | 24.23 | < 2e-16 | *** |
| Comp.3 | -0.07452 | 0.00424 | -17.59 | < 2e-16 | *** |
| Comp.4 | -0.11384 | 0.00496 | -22.94 | < 2e-16 | *** |
| Comp.5 | 0.09558 | 0.00725 | 13.18 | < 2e-16 | *** |
| Comp.6 | 0.23487 | 0.00609 | 38.60 | < 2e-16 | *** |
| Comp.7 | -0.17150 | 0.00847 | -20.24 | < 2e-16 | *** |
| Comp.8 | -0.10003 | 0.01021 | -9.80 | < 2e-16 | *** |
| Comp.10 | 0.00518 | 0.01343 | 0.39 | 0.700 | |
| Comp.11 | -0.13227 | 0.01765 | -7.49 | 6.7e-14 | *** |
| Comp.12 | -0.05915 | 0.01914 | -3.09 | 0.002 | ** |
| Comp.13 | -0.37641 | 0.02027 | -18.57 | < 2e-16 | *** |
| Comp.14 | -0.19240 | 0.02420 | -7.95 | 1.8e-15 | *** |
| Comp.15 | 0.43811 | 0.02077 | 21.10 | < 2e-16 | *** |

---

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

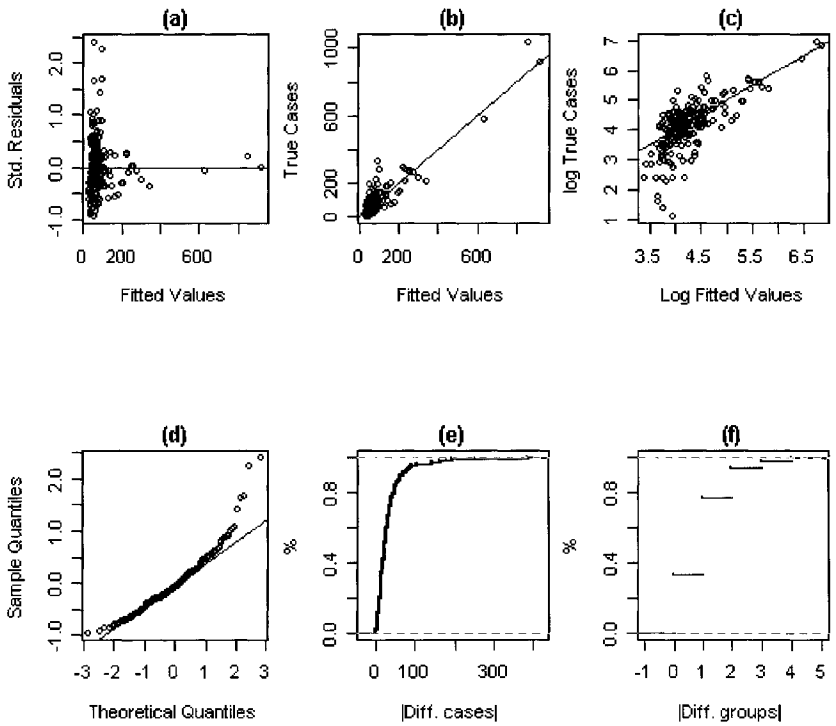


Figure A.2: Two months in advance GLM diagnostics

Table A.3: One month in advance GLM (June 2003 to May 2004 prediction)

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 4.19548 | 0.00891 | 470.72 | < 2e-16 | *** |
| Comp.1 | -0.07663 | 0.00158 | -48.65 | < 2e-16 | *** |
| Comp.2 | 0.06269 | 0.00265 | 23.67 | < 2e-16 | *** |
| Comp.3 | 0.12114 | 0.00288 | 42.01 | < 2e-16 | *** |
| Comp.4 | 0.01695 | 0.00408 | 4.15 | 3.3e-05 | *** |
| Comp.5 | 0.06374 | 0.00515 | 12.38 | < 2e-16 | *** |
| Comp.6 | 0.13278 | 0.00627 | 21.16 | < 2e-16 | *** |
| Comp.7 | -0.10815 | 0.01030 | -10.50 | < 2e-16 | *** |
| Comp.8 | -0.06046 | 0.00949 | -6.37 | 1.9e-10 | *** |
| Comp.9 | 0.06153 | 0.01056 | 5.83 | 5.6e-09 | *** |
| Comp.10 | 0.03858 | 0.01226 | 3.15 | 0.00165 | ** |
| Comp.11 | -0.07221 | 0.01694 | -4.26 | 2.0e-05 | *** |
| Comp.12 | -0.28476 | 0.01828 | -15.57 | < 2e-16 | *** |
| Comp.13 | -0.16927 | 0.01937 | -8.74 | < 2e-16 | *** |
| Comp.14 | 0.57123 | 0.02006 | 28.47 | < 2e-16 | *** |
| Comp.15 | 0.12454 | 0.02853 | 4.37 | 1.3e-05 | *** |
| Comp.16 | -0.14195 | 0.03244 | -4.38 | 1.2e-05 | *** |
| Comp.17 | 0.05302 | 0.02982 | 1.78 | 0.07543 | . |
| Comp.18 | -0.43597 | 0.03658 | -11.92 | < 2e-16 | *** |
| Comp.20 | 0.79538 | 0.07970 | 9.98 | < 2e-16 | *** |
| Comp.21 | 0.39911 | 0.09732 | 4.10 | 4.1e-05 | *** |
| Comp.22 | -0.35935 | 0.11472 | -3.13 | 0.00173 | ** |
| Comp.23 | 1.25066 | 0.11639 | 10.75 | < 2e-16 | *** |
| Comp.24 | -0.32015 | 0.15214 | -2.10 | 0.03535 | * |
| Comp.28 | -0.40769 | 0.23008 | -1.77 | 0.07640 | . |
| Comp.29 | -2.23986 | 0.45237 | -4.95 | 7.4e-07 | *** |
| Comp.32 | -5.58444 | 0.52407 | -10.66 | < 2e-16 | *** |
| Comp.33 | 2.27249 | 0.94175 | 2.41 | 0.01582 | * |
| Comp.34 | -2.70312 | 0.98369 | -2.75 | 0.00600 | ** |
| Comp.35 | -6.88218 | 1.10160 | -6.25 | 4.2e-10 | *** |
| Comp.36 | -13.07012 | 1.45024 | -9.01 | < 2e-16 | *** |
| Comp.37 | -15.15099 | 1.52915 | -9.91 | < 2e-16 | *** |
| Comp.38 | 24.94054 | 2.02371 | 12.32 | < 2e-16 | *** |
| Comp.39 | -8.70275 | 2.26402 | -3.84 | 0.00012 | *** |
| Comp.41 | -29.39826 | 6.00936 | -4.89 | 1.0e-06 | *** |
| Comp.42 | -33.04340 | 7.64107 | -4.32 | 1.5e-05 | *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

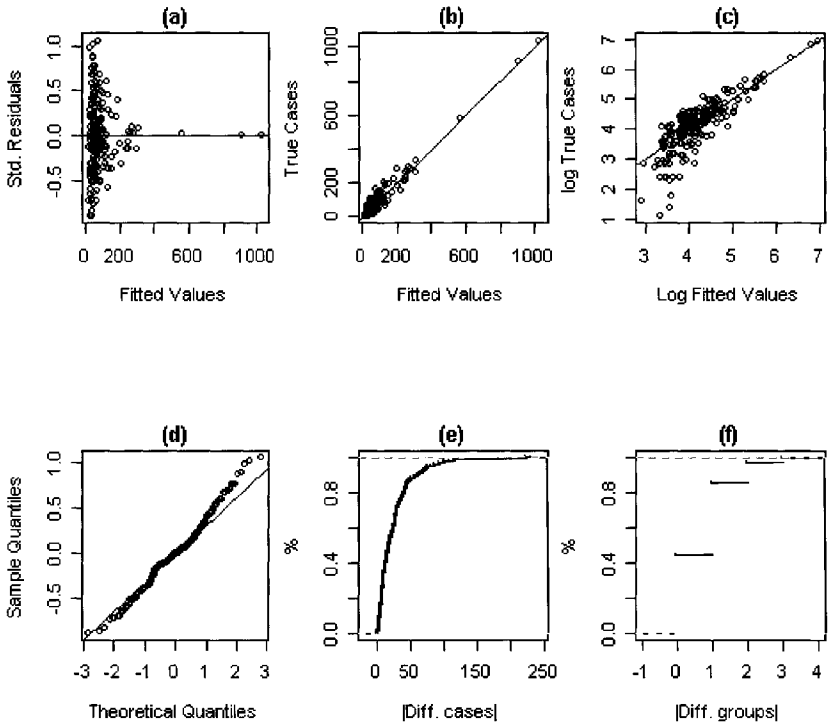


Figure A.3: One month in advance GLM diagnostics