Chapter 2

Methodology

Data Source

The data were collected retrospectively during the period 1978- 1997 in order to study the climatic factors related to the incidence of DHF, and to compare the climatic factors between the West and East coasts of Southern Thailand.

The variables in this study are the monthly incidence of DHF, temperature, humidity, rainfall and rain days in the Southern Thailand. These are taken from two provinces in the West coast and two provinces in the East coast of the Southern Thailand during 1978 - 1997.

The selection criteria was as follows:

- The South of Thailand is divided into two sides the West and East coasts.
- The two provinces having the highest incidence of DHF in each side was selected.

Data Collection and Data Management

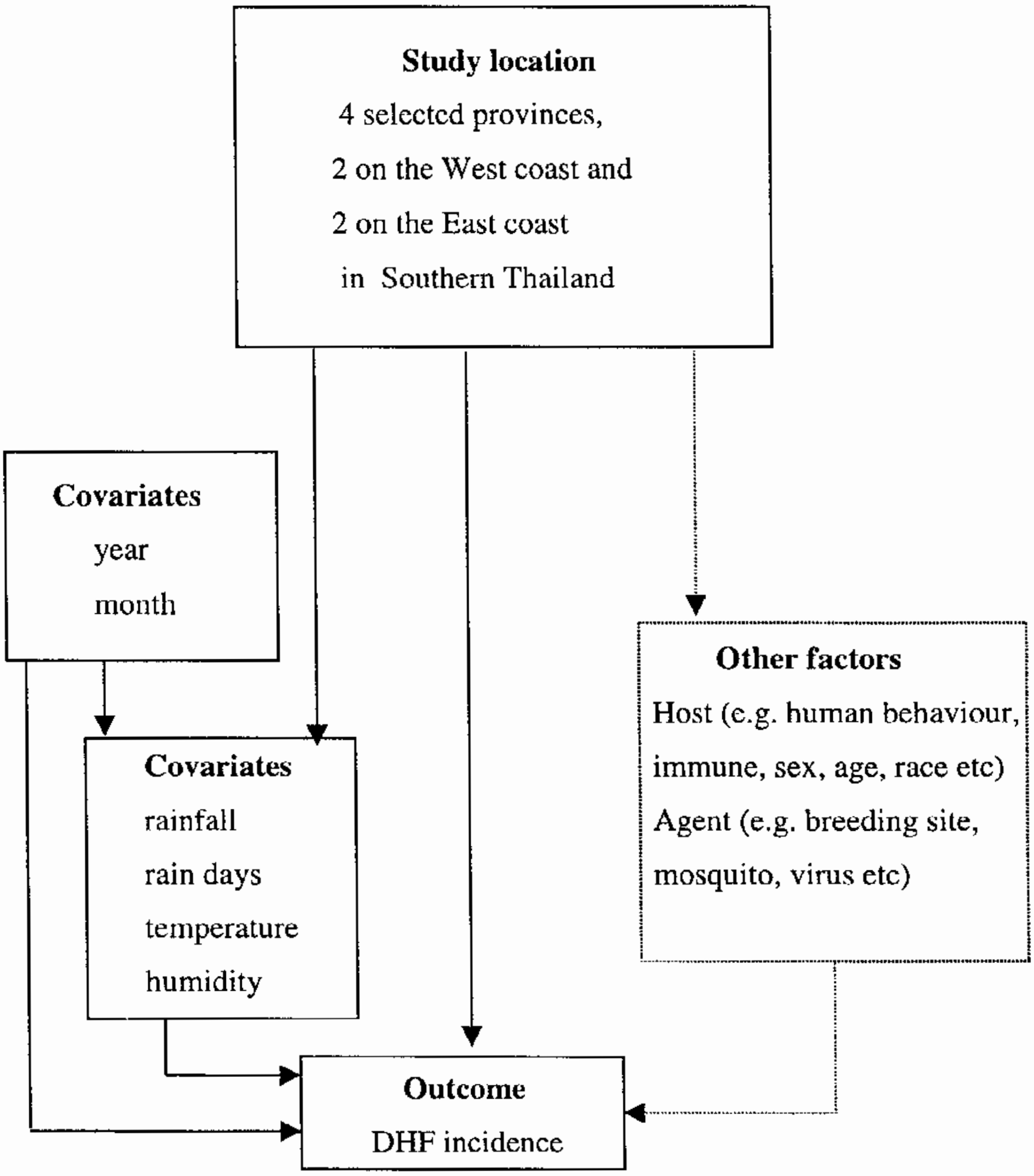
The data for the incidence of DHF were collected over the 20-year period 1978 to 1997 from Epidemiological reports at the Division of Epidemiology, Ministry of Public Health. The climatic factors are collected at the same period from the Meteorological Department, Ministry of Transport and Communications.

The data were stored in Microsoft Excel, and recorded, where necessary using, PFE (programmers file editor). Matlab version 5 (Hanselman and Littlefield et al, 1997) and Asp (McNeil et al, 1998) were used for graphical presentation and statistical analysis.

Schematic Diagram

Using the conceptual framework shown in Figure 2.1, we can investigate the relation between climatic factors and thc incidence of DHF of the southern part of Thailand. The climatic factors used in this study include rainfall, rain days, humidity and temperature from the selected provinces. The variables listed as "other factors" are relevant, but not includcd in this study.

Figure 2.1: Conceptual framework representation of DHF incidence and factors related to DHF incidence



Graphical Methods

The following graphical methods were used in this study.

1. Histograms and statistics for variables before and after transformation using logarithms and cube roots.
2. Comparison between the outcome and determinants by location.
3. Scatterplot matrix showing relationships between the outcome and determinants by location.
4. Time series plots of the variables (univariate).
5. Bivariate time series plot displaying association between DHF incidence and rainfall.

Statistical Methods

The statistical methods used for the data analysis may be described as follows.

1. Correlation analysis

The correlation is used to measure the linear association between an independent variable (x) and a dependent variable (y). The correlation coefficient is a measure of the strength of the linear, or straight-line, relationship between the variables. The correlation coefficient is defined as (McNeil et al, 1998, page 181)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \qquad (2.1)$$

It has been shown that r ranges from a minimum of -1 to a maximum value of 1. A correlation coefficient equal to 0 indicates no linear relationship between the two variables. In the analysis, the correlation between variables recorded at different times will be investigated.

2. Periodogram analysis

A time series is stationary if its statistical properties do not change with time. It is unlikely that a stationary time series will repeat itself exactly, but the series is repeatable in a probabilistic sense. The only aspect that changes is the sampling error, which does not contain useful information. These sampling fluctuations could be relatively large compared to the persistent characteristic.

These ideas lead to the sinusoid (the simplest function that repeats itself) and to the idea of measuring the amount of periodicity or repeatability in a time series by finding its covariance or correlation with a sine wave having a given period. A sinusoid is characterised by the property that taking a linear transformation of its argument only shifts its frequency and its phase or position relative to some origin. The cosine function is just a sine function whose argument is shifted by $\pi/2$, that is

$$\cos(x) = \sin(x + \pi/2) \qquad (2.2)$$

Since sinusoidal functions are periodic it is natural to use them as a basis for approximating a stationary time series. This basis comprises sinusoidal waves with different frequencies each defined on the time interval spanned by the data. The first component appears exactly once on this time interval, the second comprises two repeated sinusoids, the third three sinusoids, and so on. These components are also called *harmonics*. The functional form for the j^{th} harmonic is a cosine wave with some phase ϕ, that is, $\cos\{2\pi j(t-1)/n+\phi\}$, $t = 1,2, \ldots, n$.

Using the mathematical theory of Fourier analysis, any function defined at n equispaced points on a finite interval may be represented exactly by a constant plus $n-1$ harmonics. The number of different frequencies in these components, m, is $(n-1)/2$ or $n/2$ (depending on whether n is odd or even) since there is a sine and a cosine harmonic at each frequency. If n is even, this Fourier representation takes the form

$$y(t) = a_0 + \Sigma[a_j \cos\{2\pi j(t-1)/n\}+b_j \sin\{2\pi j(t-1)/n\}] + a_m \cos\{\pi(t-1)\} \qquad (2.3)$$

where the summation is from $j=1$ to $j=m-1$. (Since $\sin\{\pi(t-1)\}$ is 0 for all integers t, in this case there is no sine harmonic at the highest frequency). A similar result applies if n is odd. Using the fact that a linear combination of a sine function and a cosine function at the same frequency may be expressed as a single sinusoid with some phase ϕ, an alternative formula for the Fourier representation is

$$y(t) = a_0 + \Sigma A_j \cos\{2\pi j(t-1)/n+\phi_j\ \} \qquad (2.4)$$

where the amplitude $A_j = \surd\left(a_j^2 + b_j^2\right)$ and the summation is from 1 to m.

This Fourier representation is similar to linear regression analysis, where the sinusoidal components play the role of determinants or predictor variables. Since the number of parameters is exactly equal to number of data values, there is no residual error: the regression model provides a perfect fit to the data. Moreover it may be shown that the sum of products of sine and/or cosine harmonics over the range of frequencies is zero, which means that these harmonics are statistically uncorrelated with each other. Consequently each Fourier coefficient (a_j or b_j) is the regression coefficient of the time series y_t on the corresponding harmonic. The formulas for these coefficients (for n even) are as follows.

$$a_0 = \sum y(t)/n$$

$$a_m = \sum (-1)^{t-1} y(t)/n$$

$$a_j = (2/n)\sum y(t)\cos\{2\pi j(t-1)/n\}$$

$$b_j = (2/n)\sum y(t)\sin\{2\pi j(t-1)/n\}$$

We can see from these formulas that each Fourier coefficient may be interpreted as a covariance between the data and a sinusoid at the given frequency.

The *periodogram* of a time series $\{I_j, j = 1, 2, \ldots, m\}$is defined in terms of the amplitudes of the harmonics in the Fourier representation as

$$I_j = (n/2)(a_j^2 + b_j^2) \qquad (2.5)$$

The multiplier n/2 ensures that the jth periodogram value is equal to the component of the variance in the data accounted for by a sinusoidal function with frequency j/n. Since the sinusoidal terms are uncorrelated with each other, it follows that

$$\sum\{y(t) - \sum y(t)/n\}^2 = \sum I_j \qquad (2.6)$$

This useful formula is known as *Parseval's theorem.*

This relation is just an analysis of variance for a time series. So the sum of the periodogram ordinates is equal to the total squared error of the data, and consequently the periodogram shows how much of thc squared error of the data is accounted for by the various harmonics. For this reason it is useful to graph the *scaled* periodogram, obtained by dividing the periodogram by its sum. The scaled periodogram thus shows what *proportion* of the squared error is associated with each harmonic.

3. Autoregressive models

Anothcr useful graphical tool is the *correlogram*, or sample *autocorrelation* function, which comprises the set of estimated correlation cofficients between thc series and itself at various spacing. Thus the (auto) corrclation coefficient at spacing (or lag) *s* may be cstimated from the following formula

$$\mathbf{r}_s = \sum_{t=1}^{n-s} y\ t\ - \bar{y} \quad y\ t+s\ - \bar{y}\} \Big/ \sum_{t=1}^{n} \{y(t) - \bar{y}\}^2 \qquad (2.7)$$

and the correlogram is a graph of the series (r_s, s = 1, 2, ..., S) against the spacing s. Since the number of terms used to calculate the correlation coefficient at lag s is n-s where n is the length of the time series, the maximum spacing S should be substantially less than n.

According to statistical theory, when the sample sizc n is large the standard error of a correlation coefficicnt is approximately normally distributed with standard deviation 1/√n, which tcnds to 0 as n gets large. This means that as the length of an obscrvcd time series increases, the sample autocorrelation function of a stationary time series stabilises, approaching a smooth curve.

For a white noise process the theoretical correlation between observations at different spacing s is zero, so we would expect the graph of its sample autocorrelation function to approach the horizontal axis r = 0 as n gcts large. Based on the normal distribution which has 95% of its probability within 1.96 standard deviation of its mean, a 95% confidence interval for the autocorrelation at lag s ranges from −1.96/√(n-s) to 1.96/ √(n-s). In contrast, the periodogram values of white noise process, being exponentially distributed with constant standard dcviation, become more denscly packed as the length of the series increases.

Ljung & Box (1978) suggested using the statistic

$$Q = n(n+2)\sum_{s=1}^{m} \frac{r_s^2}{n-s} \qquad (2.8)$$

where m is a specified integer substantially less than the series length n, to test the hypothesis that a time series is a sample from a white noise process. If it is necessary to fit a linear model involving p parameters to transform the series to a white noise process, where these parameters are estimated from the data, then Q is distributed approximately as a chi-squared distribution with m-p degrees of freedom.

4. Forecasting

Forecasting a time series is easy if the noise component is white noise. If the noise is modelled using an autoregressive process, the forecasting procedure is more complicated. From the model it can be seen that the s-step forecast (that is, the forecast at s time units in the future) of a noise series z(t) may be obtained recursively. For a first-order autoregressive process – called an ar(1) process for short – the forecasting formula is quite simple. If T is the time index of the last value observed, the next value of the time series is

$$z(T+1) = a\,z(T) + w(T+1) \qquad (2.9)$$

Since w(t) is white noise, the forecast of w(T+1) is 0, so the forecast of z(T+1) is just

$$z_f(T+1) = a\,z(T).$$

Similarly, the forecast at t = T+2 is

$$\begin{aligned} z_f(T+2) &= a\,z_f(T+1) + w_f(T+2) \\ &= a^2\,z(T). \end{aligned}$$

Continuing in this way, the general formula for forecasting the value at lead s is

$$z_f(T+s) = a^s\,z(T).$$

The forecasting formula for the ar(2) process is rather more complicated. However numerical values for the forecasts at any lead s for an autoregressive process of any order p, can be obtained by using computer program. So if estimates of the autoregressive coefficients $a_1, a_2, \ldots, a_p$ are available, we can forecast ahead as far as we wish using the observed values of the time series.

5. Moving average process

Another type of model for a time series called a *moving average* model is used for smoothing a time series when it fluctuates a lot.

If the original series (of elevations) is denoted by z(t), t = 1, 2, …, a smoothed series could take the simple form

$$x(t) = \{z(t-1) + z(t) + z(t+1)\}/3 \qquad (2.10)$$

This is called a *three-term equally-weighted moving average*. It involves replacing the series by the average of each observation and its two adjacent observations.This formula is natural if the direction is unimportant, that is, if there is no difference between increasing and decreasing values of t. But in time series analysis it is more natural for a smoothing formula to involve only present and past values of the series (since the future is usually not yet observed). So the formula for the 3-term equally weighted moving average becomes

$$x(t) = \{z(t) + z(t-1) + z(t-2)\}/3 \qquad (2.11)$$

6. Bivariate time series analysis

Bivariate time series analysis is used to show the association between two stationary time series processes. Bivariate time series have two functions the *squared coherence* and the *crosscorrelation* obtained by using the function *tsbiplot* function in ASP. The squared coherence, like the periodogram of a univariate time series, is a function of frequency. If two series are independent, the squared coherence should be constant on average limit is 0 and which contains each squared coherence with probability 0.95, assuming the series are independent.

Like periodogram values, the squared coherences are mutually independent. So we would expect 5% of them to be outside a 95% confidence interval, indicated by dotted lines on the graph, when the null hypothesis of independence is true.

The cross-correlation function is similar to the autocorrelation function of a univariate time series. Dotted lines corresponding to the limits of intervals centered at 0 containing each crosscorrelation with probability 0.95, are used to assess the independence hypotheses.

7. Time series analysis with predictor variable using regression analysis

The model of time series analysis can be used to predict the dependent variable. And this model is similar the model of multiple regression analysis. It can describe the relationship between dependent and independent variables in each the time period. For example, a forecast of the incidence of DHF may be based on a relationship with epidemiologic factors such as environmental, host and agent factors. These variables are called predictor or independent variables and the incidence of DHF is referred to as the predicted or dependent variable.

Steps in establishing a model of time series regression analysis are as follows:

1. Starting with the univariate model, the incidence of DHF over time for each province is assessed.

2. Upon completion of the univariate analysis, we enter each predictor variable of interest in the model obtained. Then dividing it by its standard error assesses the regression coefficient obtained from each model. If the resulting value is greater than 2, this indicates that the variable of interest is significant (p-value = 0.10).

3. For a model having more than one significant variable, all significant variables should be included in the model.