Chapter 2

Methodology

In this chapter we describe the source of data and summarize the statistical methods used to analyze these data. These methods include those used both for preliminary data analysis and for modeling. All of the graphical and statistical models were carried out using R program (Venables and Ripley, 2002)

2.1 Data source and variables

Data were obtained from the Climate Research Unit (CRU), the United Kingdom Meteorological Office (CRU, 2009). CRU provides monthly temperature averages for 5° by 5° latitude-longitude grid boxes based on data collected from weather stations, ships, and more recently satellites. These are available as separate files. The temperatures in South East Asia over the last 36 years from 1973 to 2008 were selected for this study. The study area lies between a circle latitude of 0^{o} N to 25^{o} N and longitude 85^{o} E to 110^{o} E and the global monthly surface temperatures cover both land and sea. The land area covers Thailand, Malaysia, Cambodia, Vietnam, Burma, and some parts of Indonesia. The sea covers Andaman Sea, Gulf of Thailand and some areas of South China Sea.

Flowchart of the data

The data obtained was anomaly temperature data on a 5 degree grid. The temperature anomaly is raw monthly temperature subtracted by monthly average temperature from 1961 to 1990. The selected data of Southeast Asia regions from 1973-2008 was converted to be raw temperature by adding back the monthly average temperature from 1961-1990 in each grid boxes to temperature anomalies data. The seasonally affected were adjusted and removed an autocorrelation. The last step statistical models were fitted to the temperature data.

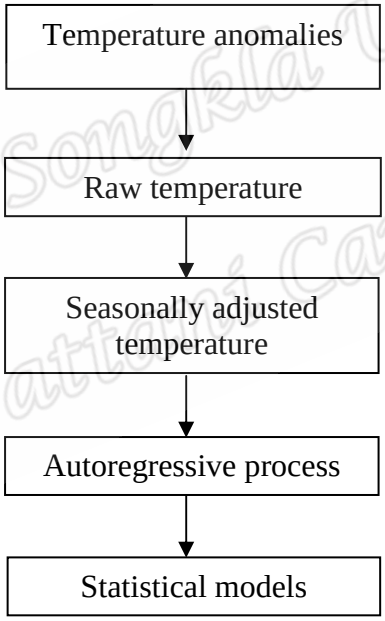

Figure 2.1 Flowchart of the data

Variables

The variables of interest in this study are monthly temperature averages, which is the outcome variable, time (month) and region, which are the determinant variables. There were 25 regions and 432 monthly temperatures for each region.

2.2 Statistical methods

Descriptive statistics

Since the outcome variable is continuous, the descriptive statistics are described by numerical summaries including mean, standard deviation, minimum and maximum.

Simple linear regression models

In this study the data were seasonally adjusted to remove the variation in the average monthly temperatures by subtracting the monthly average and then adding back the overall mean. A simple linear regression model was fitted for each grid-box taking the form

$$X_{it} = b_{0i} + b_{1i} d_t .$$

where X_{it} denotes the seasonally-adjusted temperature in grid-box *i* for month *t*, and d_t denotes the time elapsed in decade since 1973, b_{0i} is the average temperature in the grid-box over the period, b_{1i} is the estimated rate of increase in temperature per decade.

Autoregressive process

Autoregressive (AR) models were also used to account for auto-correlation among residuals from the fitted linear models. A second AR model is fit to the residual, which take form

$$R_{it} = a_{1i} R_{i(t-1)} + a_{2i} R_{i(t-2)} .$$

Where R_{it} are the residuals, $R_{i(t-1)}$ are lag 1 residuals, $R_{i(t-2)}$ are lag 2 residuals, a_{1i} and a_{2i} are the estimated parameters of the model (AR coefficients).

Autocorrelation functions

The autocorrelation function is an important tool for describing the properties of a stationary process (Chatfield, 1996). Sample autocorrelation coefficients measure the correlation between observations at different distances apart. Regarding the first observation in each pair as one variable, and the second observation as a second variable, the correlation coefficient between x_t and x_{t+1} is given by

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x}_1)(x_{t+1} - \bar{x}_2)}{\sqrt{[\sum_{t=1}^{N-1}(x_t - \bar{x}_1)^2 \sum_{t=1}^{N-1}(x_{t+1} - \bar{x}_2)^2]}}$$

where $\bar{x}_1$ is the mean of the first observations and $\bar{x}_2$ is the mean of the second observations.

We can find the correlation between observations a distance *k* apart, which is given by

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}$$

This is called the autocorrelation coefficient at lag *k*. Where $\bar{x}$ is the mean of *N-k* observations.

The correlogram

A useful aid in interpreting a set of autocorrelation coefficients is a graph called a correlogram where the sample autocorrelations, r_k are plotted against the lags, *k*. If a time series is completely random, 5% of the values of r_k would be expected to lie between $\pm 2/\sqrt{N}$ as shown in Figure 2.1.

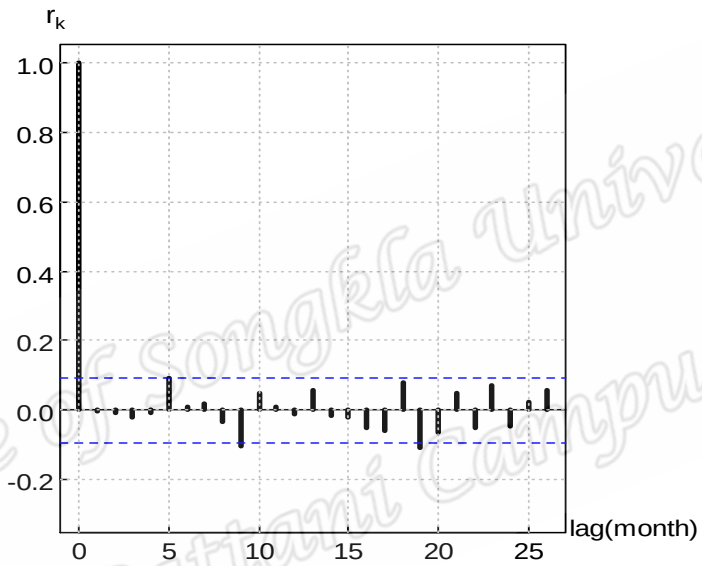

Figure 2.2 Example of a correlogram, the dotted line indicates $\pm 2/\sqrt{N}$

Linear spline model

Linear regression is an analytic approach commonly used to examine the relationship between a dependent and independent variables. Predictor variables may be separated into logical categories or we may add additional terms that are functions of existing predictors such as spline modeling, which may provide a better fit. Splines are continuous lines or curves. The join points that mark one transition to the next are referred to as knots. Knots give the curve freedom to change direction and more closely follow the data. A linear spline model is given by

$$s(t) = c_0 + c_1 t + \sum_{j=1}^{m} c_{j+1}\left(t - T_j\right)_+$$

where c_0 is a constant

c_1 is parameter for the first of time t

c_{j+1} are parameters for time t when knots are defined

T_j are knots

j is number of knots, j=1,2,3,…,m

t_+ is t if $t > 0$

A linear spline defined on an interval (0, T) takes the form where t_+ is t if $t > 0$, otherwise. The constants T_j are called knots. A linear spline is a piecewise continuous linear function with discontinuities in its derivative at each knot (McNeil et al, 2011). Suppose that (y_i, t_i) are data comprising measurements of average temperatures y_i, say, at successive months t_i. To fit a linear spline by linear regression, we simply take y as the outcome variable and the components t, $(t - T_1)_+$, $(t - T_2)_+$ … as predictors. For the global temperature data, there are 36 × 12 = 432 months in a 36-year period, and we assume that the fitted splines have knots after 12 and 24 years (after 144 and 288 months), so the model contains 4 parameters, c_0, c_1, c_2 and c_3. These constants are the value at the first month, the slope over the first 12 years, the increase in slope from the first to the second 12-year period, and the further increase in slope from the second to the third 12-year period, respectively.