CHAPTER 2

METHODOLOGY

Following are the steps and statistics used for the analysis of electricity usage in Pattani.

Data

The data for this analysis were obtained from the Pattani Electricity Authority Substation. This substation is situated in Puyud Subdistrict, Muang Pattani District, Pattani Province and controls electrical current for all of Pattani Province and some part of Songkhla Province. It is administratively divided into six feeders. The following data are daily meter readings for each feeder in 1996, giving 366 values for each, recorded in kilowatt.hour units. However, after the 338th day, the meter for feeder 8 was out of order. Therefore the substation used the meter for feeder6 to record usage for feeder 8 as well, causing a large increase in the reading for feeder6 after 338 days.

Methodology

The analysis are present in the following steps.

1. Graphs of daily consumption for each feeder and feeder combined.

2. Summary of the numerical analysis of the daily consumption.

3. Comparison of the means of electricity usage between feeders combined and between days.

4. Correlation analysis between feeder combined.

5. Trend analysis of daily consumption.

6. Comparison of the electricity usage between days.

7. Development of a model of electricity usage by time series.

Software

The following software was used in the analysis.

1. Microsoft Access Vcrsion2.0

Microsoft Access was used to construct the database file. The datafiles exported from in Microsoft Access. *Feeddif.num* is the file containing the electricity usage per day for each feeder and the total for Pattani Province. This data is shown in Table 7 of the appendix.

2. Matlab Version 4 (Hanselman & Littlefield,1995)

This software was used in the analysis to do lincar rcgrcssion, plot scatterplot matriccs and do time series analysis.

3. ASP (Asia Statistical Package)

Asp runs under Matlab Version 4 (Hanselman & Littlefield,1995). This program was developed by Dr. Don McNeil from Macquarie University and a team in the Department of Applied Mathematics and Computer Science in the Faculty of Science and Technology at Pattani Campus of Prince of Songkhla University (PSU) in Thailand. It was used to do linear regression, scatterplot matrices and time series.

4. SPIDA Version 6.08

SPIDA was used in the analysis to do Two-way Anova variable models to compare the mean of each feeder and was used in the correlation between feeders.

5. Microsoft word Version 6.0a

This was used to write the report.

Statistics used for the analysis.

1. Descriptive Statistics

1.1 Mean

Calculated from the formula

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

1.2 Standard Deviation (S.D.)

Calculated from the formula

$$S.D. \;=\; \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N}}$$

2. Two-way Analysis of Variances

Two way Analysis of Variances was used to compare the mean of the usage per day between each feeder, and to compare the mean usage per feeder for each day of the year. If the data array has r rows and no missing observation (giving n = r x c observations altogether), a correct p - value is based on an F - statistic defined as (McNeil, 1995: 73)

$$\mathrm{F} \;=\; \frac{(s_1 - s_{12})/(c-1)}{s_{12}/(n-c-r+1)}$$

where

$$S_1 \;=\; \sum_{j=1}^{c}\sum_{i=1}^{r}\,(y_{ij} - \bar{y}_i)^2$$

$$S_{12} \;=\; \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$$

and

$$\bar{y}_i \;=\; \frac{1}{c}\sum_{j=1}^{c} y_{ij}$$

$$\bar{y}_j \;=\; \frac{1}{r}\sum_{i=1}^{r} y_{ij}$$

$$\bar{y} \;=\; \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} y_{ij}$$

3. Coefficient of Correlation

The co-efficient of correlation was found between each feeder pair for the amount of electricity used each day. The researcher was able to calculate the coefficient from the following equation for the Pearson's Product-Moment Correlation Coefficient (Wert, Neidt and Ahmann, 1954:83).

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

When r = Pearson-Product-Moment Correlation Coefficient

n = Number of data

$\sum XY$ = the sum of X multiplied by Y

$\sum X$ = the sum of X

$\sum Y$ = the sum of Y

$\sum X^2$ = the sum of X squared

$\sum Y^2$ = the sum of Y squared

4. Time Series

A time series is a set of numerical data measured sequentially in time. The Measurements are often equispaced in time or nearly so. Time series data arise in economics and marketing (company sales or profits in successive months, stock market prices, currency relative values, etc.), in the physical sciences (barometric pressure in successive hours, snowfall in successive years, maximum and minimum daily air temperatures), in engineering processes (quality control charts, intervals between equipment failures), in biology and demography (sizes of animal population in successive seasons, birth and death rates in successive years, university enrollment, ect.), and in many other applications areas.

Four important objectives arise in time series analysis. These are

1. Forecasting future values of a series.
2. Estimating the trend or overall character of a time series.
3. Modeling the dynamic relations between two or more time series.

4. Summarizing characteristic features of a time series.

Because time series methods are based on linear models, it is frequently necessary to transform the data. A logarithm transformation is usually needed for rates and financial data, whereas square roots are often better for transforming counts. The need for a transformation is usually apparent from an inspection of graph of the data. Many time series, including the transformed data have a linear trends. In these situations it may be useful to fit a straight line and use the residuals as a basis for further statistical analysis. This approach is not the only way of modelling a trend in a time series. and other methods will be considered, and shown to be superior in certain situations. A straight line may be fitted to time series data simply by least squares regression. If the data are denoted by y_t (t = 1, 2, ..., n), the least squares fitted line is given by the equation

$$\hat{y}_t = \hat{a} + \hat{b}t \qquad (1)$$

where

$$\hat{b} = \frac{\sum (y_t - y)(t - \bar{t})}{\sum (t - \bar{t})^2}$$

$$\bar{a} = \bar{y} - \hat{b}\,\bar{t}$$

A residual time series z_t is now obtained by subtracting the fitted lined from the data, giving

$$z_t = y_t - (\hat{a} + \hat{b}\,t) \qquad (2)$$

If only a short term forecast is needed, an intuitively appealing method is to fit a straight line model to the most recent values of the series and extrapolate this line. This procedure works well for forecasting series with slowly changing, since the forecast values can adapt quietly whenever a series changes its overall direction. If one were to fit a straight line to the two most recent values of a series and then use this line to forecast the observation, the forecast formula would be as follows.

$$\hat{y}_{t+1} = 2y_t - y_{t-1} \qquad (3)$$

This forecast formula may be generalized to take the form

This forecast formula may be generalized to take the form

$$\hat{y}_{t+1} = b_t y_t + b_2 y_{t-1} \tag{4}$$

where, to keep the forecasts on track with the data, the sum of the coefficients b_1+b_2 is assumed to be 1. If the most recent observations in the time series contain most of the information about where the next value is going to be, Equation (3) should provide a good forecast. On the other hand if there is very little information in the most recent data it would be more reasonable to regard the time series as a sample of independent observation, from which the best estimate of the next value is just the sample mean

$$\hat{y}_{t+1} = \frac{1}{n}\sum_{j=1}^{m-1} y_{t-j} = \bar{y} \tag{5}$$

We could also choose a weighted linear combination of the two forecast formulas, that is

$$\hat{y}_{t+1} = (1 - b_1 - b_2)\,\bar{y} + b_1 y_t + b_2 y_{t-1} \tag{6}$$

where $b_1 + b_2$ is now between 0 and 1. This formula may be expressed alternatively as

$$\hat{y}_{t+1} - \bar{y} = b_1(y_t - \bar{y}) + b_2(y_{t-1} - \bar{y}) \tag{7}$$

More generally still, the forecast may involve any number (p) of recent observations, talking the form

$$\hat{y}_{t+1} - \bar{y} = \sum_{k=1}^{p} b_k (y_{t-k+1} - \bar{y}) \tag{8}$$

or, more generally (for data that are not mean-corrected, as would be the case for a nonstationary series),

$$\hat{y}_{t+1} = b_0 + \sum_{k=1}^{p} b_k y_{t-k+1} \tag{9}$$

Forecasting the next value in time series thus involves choosing a set of coefficients b_k for $k = 1, 2, \ldots, p$ (or $k = 0, 1, 2, \ldots, p$, if the data are not mean-corrected) which is optimal in some sense, such as minimizing the squared error. Note that Equation (8) and (9) may be used to forecast more than one time unit ahead, simply by repeatedly using the forecasts as if they were observed data.

A residual time series z_t may be obtained by subtracting the forecast from each observation y_t. If the forecasts are obtained from Equation (9), the residual series is given by

$$z_t = y_t - b_0 - \sum_{k=1}^{p} b_k y_{t-k} \tag{10}$$

with a similar representation if Equation (8) is used. Ideally, there will no information present in these residuals since if information were present it could be used to improve the forecast formula. Consequently the residuals should resemble a white noise series. The process of using a formula such as Equation (8) (or (9)) to obtain a residual series is called *filtering* a time series, and the set of coefficients b_k (k= 1, 2,..., p) is called a *linear filter* with *span p*. If the span is known the filter coefficients may be estimated from the data using regression analysis, where the response variable comprises the time series (y_t , y_{t-1} ,.......,y_{p+1}) and there are p predictor variables, (y_{t-1}, y_{t-2}, ..., y_p), (y_{t-2} , y_{t-3} , ..., y_{p-1}), ...,(y_{t-p}, y_{t-p-1}, ...y_1). Note that the number of observations used to fit this regression model is n-p.

Differencing

If a time series has a linear trend it is reasonable to fit a straight line and to use this line as a first step for forecasting the series. Subtracting a line may produce residuals that look stationary, and these residuals in turn may be further modeled using time series techniques. Another way of removing a linear trend from a time series is to *difference* it. The series of first differences is simply

$$Dy_t = y_t - y_{t-1} \tag{11}$$

where D denotes the difference operator. More generally a series may be differenced at any lag s, giving

$$D_s y_t = y_t - y_{t-s} \tag{12}$$

If a series has a seasonal component at spacing s, differencing at lag s may remove this component. Differencing is thus the special case of filtering that arises when all of the coefficients in the filter are 0 except b_s, which is 1. Differencing may be repeated. For example differencing twice at lag 1 gives

$$D^2 y_t = D(y_t - y_{t-1}) \ = y_t - 2y_{t-1} + y_{t-2} \tag{13}$$

(secondary differencing), while, more generally, differencing at lags 1 and s gives

$$DD_s y_t = D(y_t - y_{t-s}) = y_t - y_{t-1} - y_{t-s} + y_{t-s-1} = D_s D y_t \qquad (14)$$

Differencing is widely used by econometricians as a technique for producing a series of residuals that looks stationary and is thus more amenable to statistical analysis. The method has also been suggested by statisticians including Box & Jenkins (1976).

Spectrum 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. Another way of looking at this is to say that the character of the series persists as you move forward or backward in time, and the only aspect that changes is the sampling error, which does not contain useful information. Of course 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 give period. A sinusoid is characterized 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

$$\cos(x) = \sin(x + \pi/2) \qquad (15)$$

Since sinusoidal function are periodic it is natural to use them as a basis for approximating a stationary time series. This basic comprises sine waves with different frequencies each defended 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 cosine wave with some phase $\varnothing$, that is , $\cos\{2\pi j(t-1)/n+\varnothing\}$, t=1, 2, ..., 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

$$y_t = a_0+\sum[a_j\cos\{2\pi j(t-1)/n\}+b_j\sin\{2\pi j(t-1)/n\}]+a_m\cos\{\pi(t-1)\} \quad (16)$$

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 formula 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 ∅, an alternative formula for the Fourier representation is

$$y_t = a_0 + \sum A_j\cos\{2\pi j(t-1)/n\}+\varnothing_j \} \quad (17)$$

where the amplitude $A_j = \sqrt{(a_j^2 + b_j^2)}$ 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 perfection for the data. Moreover it may be show that the sum of products or 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, .., m} is defined in terms of the amplitudes of the harmonics in the Fourier representation as

$$I_j = (n/2)(a_j^2 + b_j^2) \quad (18)$$

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

$$\Sigma(y_t-\Sigma y_t/n)^2 = \Sigma(I_j) \qquad (19)$$

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 the squared error of the data is accounted for by each 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.Note that the frequency j/n is expressed in terms of the number of cycles per unit time. Since the values of j are 1, 2, ..., m, the lowest frequency is 1/n, corresponding to a period equal to the whole range of the data. and the highest frequency is close to 0.5 (exactly 0.5 if n is even), corresponding to cycles of length 2 with the data oscillating from one value to the next. A function *tsplot* may be used to show a periodogram of a time series.

Autoregressive Models

We saw how the periodogram and its logarithm may be used to investigate the character of a time series. Another useful graphical tool is the *correlogram,* or sample *autocorrelation function*, which comprises the set of estimated correlation coefficients between the series and itself at various spacing. Thus the (auto)correlation coefficient at spacing (or lag) s may be estimated from the formula

$$r_s = \frac{\sum_{t=1}^{n-s}(y_t-\bar{y})(y_{t+s}-\bar{y})}{\sum_{t=1}^{n}(y_t-\bar{y})^2} \qquad (20)$$

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 size n is large the standard error of a correlation coefficient is approximately normally distributed with standard deviation 1/√n, which tends to 0 as n gets large. This means that as the length of an observed time series increases, the sample autocorrelation function of a stationary time series stabilizes, approaching a smooth curve. For a white noise process the theoretical correlation between observations at different spacing is zero, so one would expect the graph of its sample autocorrelation function to approach the horizontal axis r = 0 as n gets large. Based on the normal distribution which has 95% of its probability within 1.96 standard deviations 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 a white noise process, being exponentially distributed with constant standard deviation, do not settle down as the length of the series increases. Instead they become more densely packed, as we saw in the preceding section. Ljung & Box (1978) suggested using the statistic

$$\mathrm{Q} = n(n+2)\sum_{s=1}^{m}\frac{r_s^{\,2}}{n-s} \qquad (21)$$

where m is a specified integer substantially less than the series length n, to 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.