Chapter 2

Methodology

In this chapter we described the method used in the study. The methodology comprises the following components.

(a) Study area

(b) Data sources

(c) Work flow diagram

(d) Data management

(e) Data analysis

2.1 Study area

Na Thap sub-district is located in Chana district of Songkhla province, southern part of Thailand. The total area of Na Thap sub-district covers 31.32 square kilometers.

The Land Development Department categorized types of land use into three groups, and each group was further divided into sub-groups as follows:

Natural: swamp forest, swamp forest-paddy field, lake, mangrove, river/canal, beach-forest, scrub/grass, wetland and casuarinas.

Farm: paddy field, rubber plant, coconut, shrimp farm, watermelon, cashew and mixed orchard.

Developed land: allocated project, abandoned paddy field, low land village sandpit and institutional land.

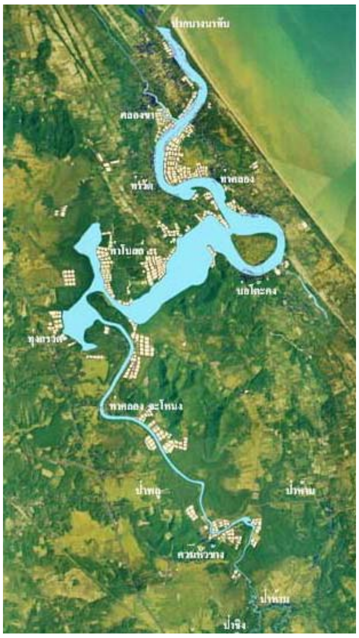


Figure 2.1: Na Thap sub-district

2.2 Data sources

This data applied secondary data from Thailand’s Land Development Department, namely the Land Development Department, region 12, Songkhla Province that had been based on surveys of land use in 1982 and 2000. Sample size is 381 regions of land use in Na Thap sub-district. Regions that had an area smaller than five hectare were combined. So, it was reduced from 381 regions to 174 regions. R program was used to arrange the map and to examine the change of land use.

2.3 Work flow diagram

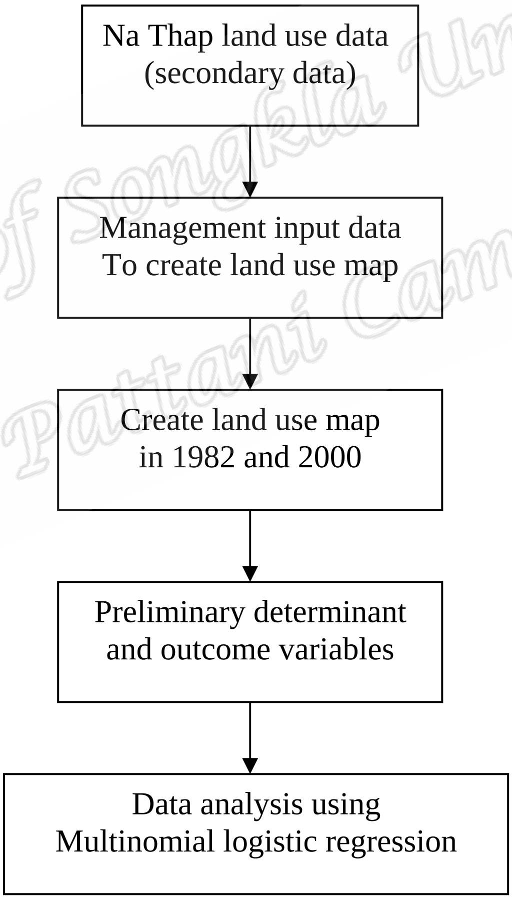

Na Thap land use data (secondary data)
Management input data To create land use map
Create land use map in 1982 and 2000
Preliminary determinant and outcome variables
Data analysis using Multinomial logistic regression

Figure 2.2 Work flow diagrams for determining the relationship between land use change and location

2.4 Data management

Software

The following computer programs were used for data analysis and thesis preparation. The open source statistical package R arranged land use map and to develop a model for described the land use change in each location.

Variable

Conceptual frame work

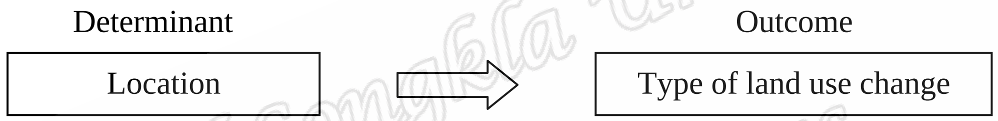


Figure 2.3 Conceptual frame work

The determinant was categorized into three groups: South, North and River. The outcome was type of land use change divided into nine groups: natural remaining natural, natural to farm, natural to developed, farm to natural, farm remaining farm, farm to developed, developed to natural, developed to farm and developed to developed.

2.5 Data analysis

Graphical Methods

First we use Universal Transverse Mercator (UTM) coordinate of Songkhla province, then selected data only Na Thap sub-district for construct example NaThap map. The input data file for create the land use map in R program.

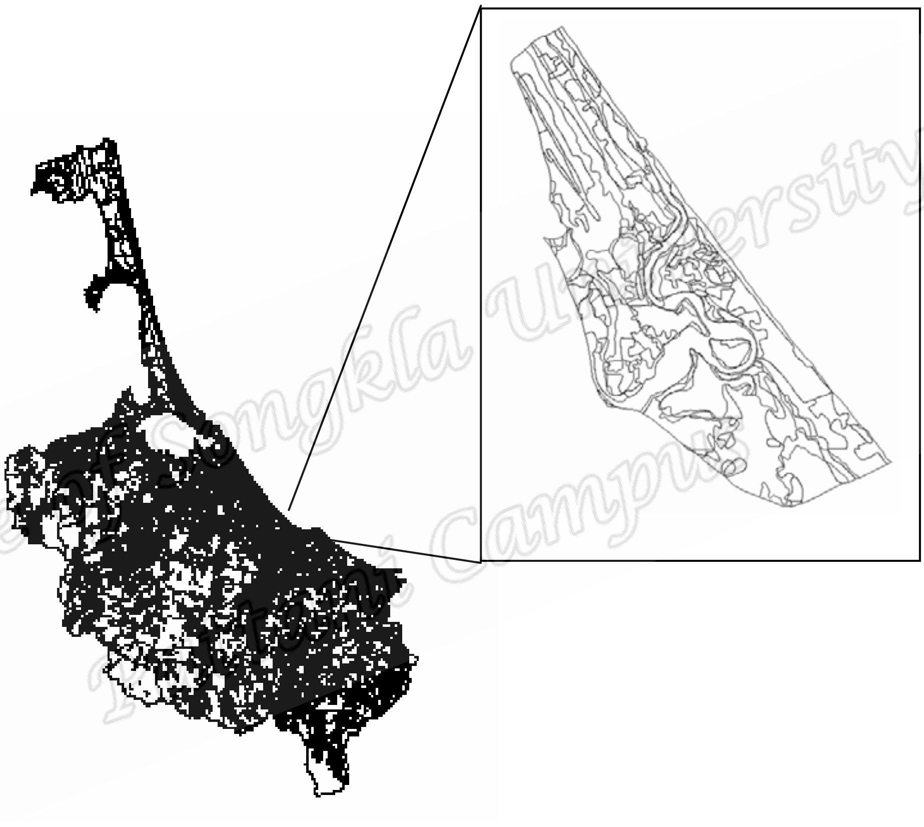


Figure 2.4: Songkhla province and Na Thap sub-district map

R program was used to create graphical display a Land use map. In this study used command *createmapNaThap.Rcm* to create Land use maps for 1982 and 2000.

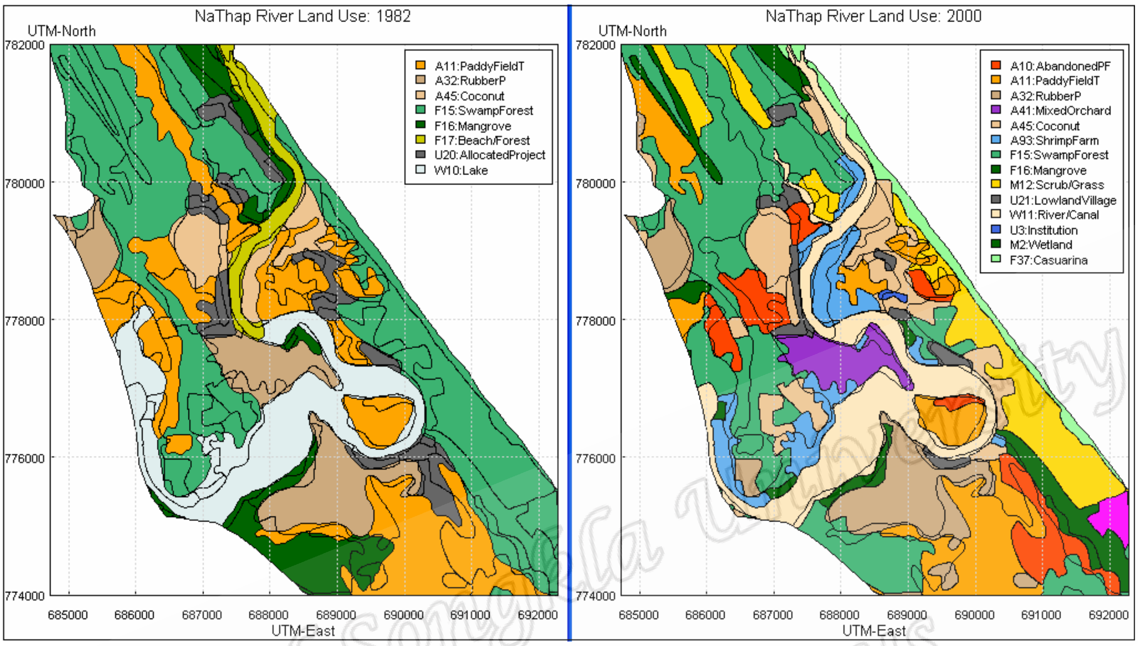


Figure 2.5: Land use map in 1982 and 2000 using R program

Figure 2.5, shows location and general area for specific land use in 1982 and 2000 of 381 regions. Sub-groups of land use were described with colors for example: in 1982 many areas were orange and light green that mean in 1982 Na Thap had a lot of paddy field and swamp forest, and some gray for allocated project etc.

The number of land use areas was reduced from 381 regions to 174 regions because some regions were combined. The R command file was use to remove borders and combine areas which were less than five hectares (Eso, 2010).

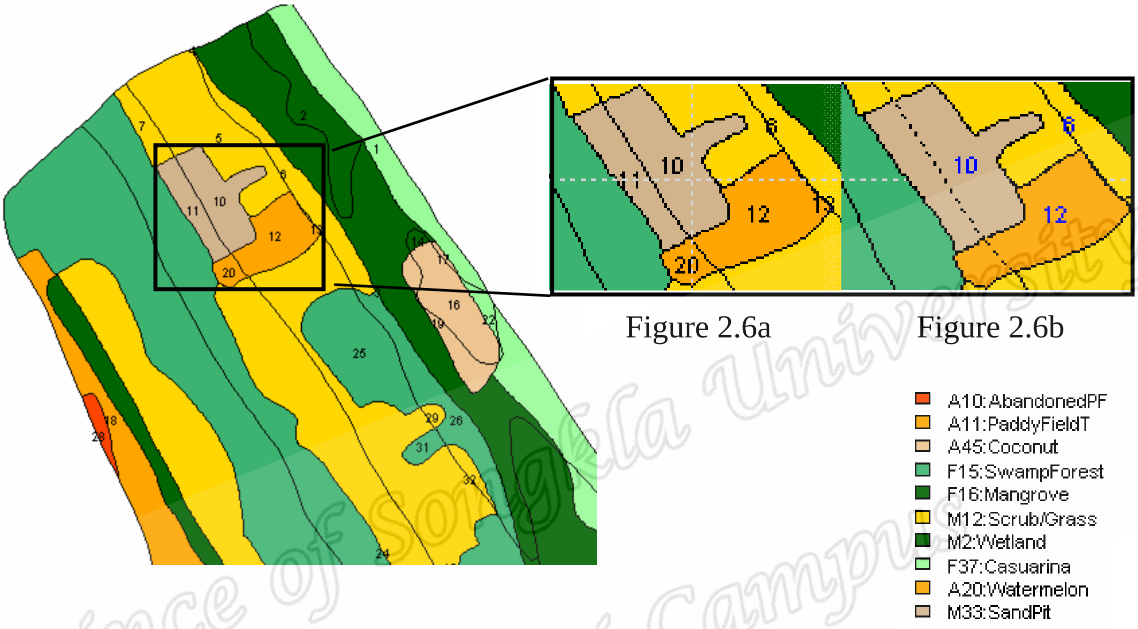


Figure 2.6a

Figure 2.6b

Figure 2.6 Combining area

For example in the north of Na Thap sub-district combine Sandpit areas (10 and 11) and Paddy Field (12 and 20) by changing solid border in figure 2.6a to deleted border in figure 2.6b. After combing regions, the code for main plot of Sandpit is 10 and main plot of Paddy Field is 12.

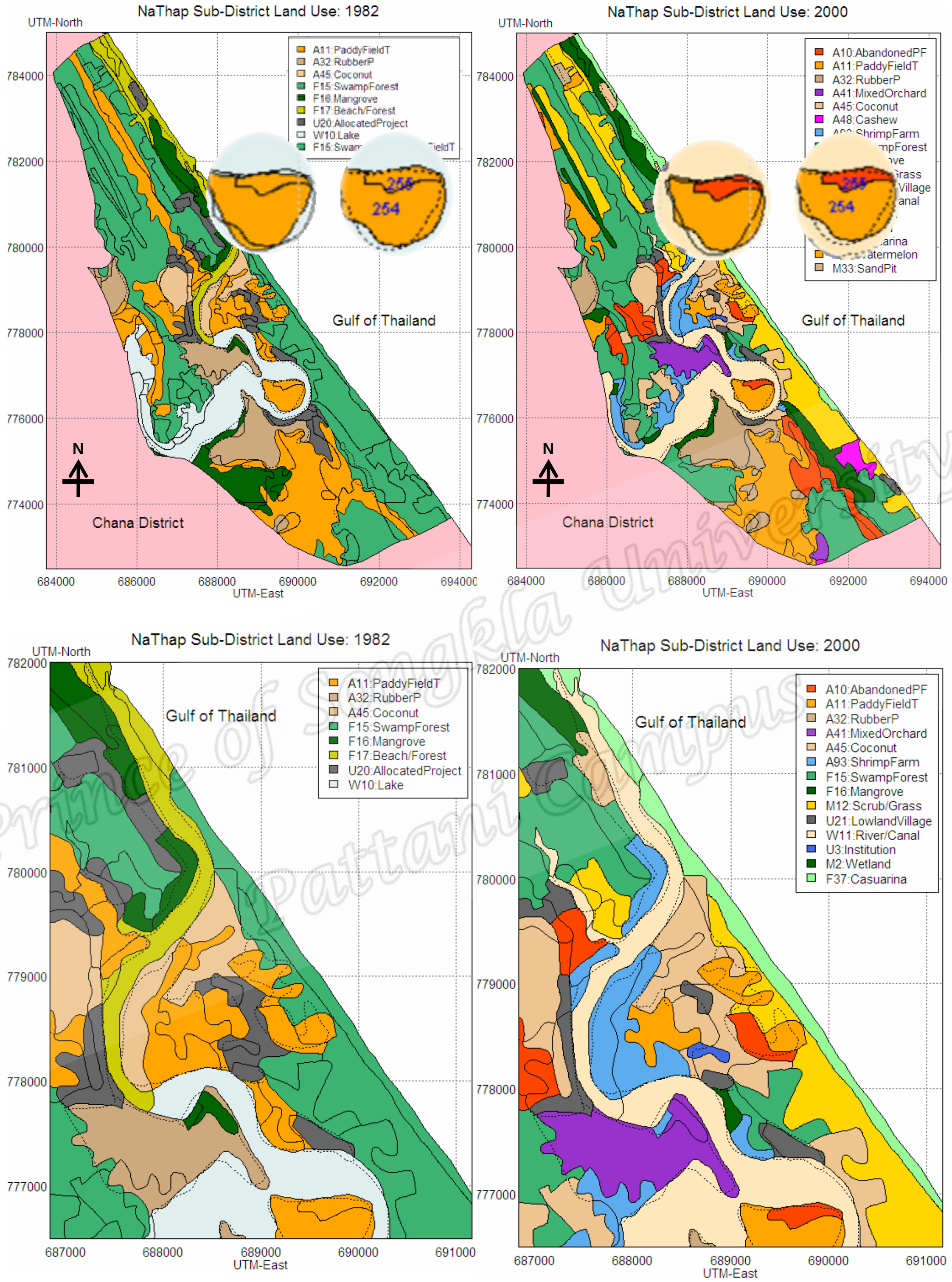


Figure 2.7: NaThap land use map after combine area

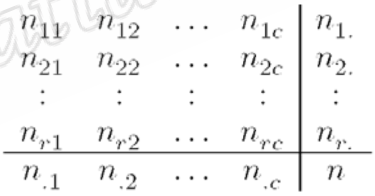
Figure 2.7 shows land use map with 174 regions in Na Thap after combined area. Regions that have an area smaller five hectare were combined.

Statistical methods

Statistics for descriptive analysis included percentages for location groups and types of land use, and odd ratios for land use change. Pearson’s chi-squared test was used to assess the association between categorical variables. Multinomial logistic regression was used to model the association between type of land use change (outcome) and the determinant.

Contingency Tables

If there is no response variable then to investigate the association between discrete variables a contingency table can be computed and a suitable test performed on the table. The simplest case is the two-way table formed when considering two discrete variables. For a data set of *n* observations classified by the two variables with *r* and *c* levels respectively, a two-way table of frequencies or counts with *r* rows and *c* columns can be computed.



$$\begin{array}{cccc|c} n_{11} & n_{12} & \cdots & n_{1c} & n_{1.} \\ n_{21} & n_{22} & \cdots & n_{2c} & n_{2.} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r.} \\ \hline n_{.1} & n_{.2} & \cdots & n_{.c} & n \end{array}$$

Figure 2.9: Contingency Table

If p_{ij} is the probability of an observation in cell *ij* then the model which assumes no association between the two variables is the model

$$p_{ij} = p_{i.}p_{.j} \tag{2.1}$$

where $p_{i.}$ is the marginal probability for the row variable and $p_{.j}$ is the marginal probability for the column variable, the marginal probability being the probability of observing a particular value of the variable ignoring all other variables. The appropriateness of this model can be assessed by two commonly used statistics: the Pearson chi-squared statistic

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-f_{ij})^{2}}{f_{ij}} \tag{2.2}$$

and the likelihood ratio test statistic

$$2\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}\times\log(n_{ij}/f_{ij}) \tag{2.3}$$

The f_{ij} is the fitted values from the model; these values are the expected cell frequencies and are given by

$$f_{ij}=n\hat{p}_{ij}=n\hat{p}_{i.}\hat{p}_{.j}=n(n_{i.}/n)(n_{.j}/n)=n_{i.}n_{.j}/n \tag{2.4}$$

Under the hypothesis of no association between the two classification variables, both these statistics have, approximately, a chi-squared distribution with $(c-1)(r-1)$ degrees of freedom. This distribution is arrived at under the assumption that the expected cell frequencies, f_{ij} , are not too small.

In the case of the 2 × 2 table, i.e., $c = 2$ and $r = 2$, the chi-squared approximation can be improved by using Yates's continuity correction factor. This decreases the absolute value of $(n_{ij}-f_{ij})$ by 1⁄2. For 2 × 2 tables with a small value of n the exact probabilities can be computed; this is known as Fisher's exact test.

Chi-squared and odds ratio

Pearson's chi-squared test and 95% confidence intervals for odds ratios are used to assess the associations between the determinant variables and the outcome of this study. The formulas based on contingency tables (McNeil, 1998) are as follows (X is a determinant of interest, Y is the outcome).

2×2 table

X is the determinant and Y is the outcome. The outcome is binary (0 or 1). The odds ratio is a measure of the strength of an association between two binary variables, that is, both the outcome and the determinant are dichotomous (McNeil, 1998).

The ratio of these odds is referred to as the odds ratio. Therefore, the estimate the odds ratio is

$$OR = \frac{a \times d}{b \times c} \tag{2.5}$$

One method of testing the null hypothesis of no association between the determinant and the outcome is to use the z-statistic, $z = \ln(OR)/SE$, where SE is the standard error of the natural logarithm of the odds ratio. An asymptotic formula for this standard error is given by

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{2.6}$$

A 95% confidence interval for the population odds ratio is thus

$$95\%\ \text{CI} = \text{OR} \times exp\ (\pm 1.96\ SE\ [\ln OR]) \tag{2.7}$$

Pearson’s chi-square statistic is defined as

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \tag{2.8}$$

The p-value is the probability that a chi-squared distribution with 1 degree of freedom exceeds this statistic.

Multinomial Logistic Regression

Logistic regression provides a method for modeling the association between a nominal outcome and multiple determinants. It is similar in many ways to linear regression. This model is easily extended to handle multiple determinants. For *m* determinants,

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{j=1}^{m} \beta_j x_j \tag{2.9}$$

The only other statistical assumption is that the outcomes are mutually independent. The model is known as simple logistic regression.

The logistic regression model described by Equation 2.9 can be further extended to situations in which the outcome variable is nominal with more than two categories. If these outcome categories are coded as 0, 1, 2, …, *c* and p_k is the probability that an outcome has the value *k*, the model takes the form, for $0 < k \le c$,

$$p_k = \frac{\exp\left(\alpha_k + \sum_{j=1}^{m} \beta_{jk} x_j\right)}{1 + \sum_{k=1}^{c} \exp\left(\alpha_k + \sum_{j=1}^{m} \beta_{jk} x_j\right)} \qquad (2.10)$$

This model is known as polytomous logistic regression or multinomial logistic regression (Hosmer and Lemeshow, 1989) is a type of logistic regression that deals with dependent variables that are nominal that is, there are multiple response levels and they have no specific order. In this study, here there are four outcome categories, Natural remaining Natural, Natural to Farm/Developed, Farm/Developed to Natural and Farm/Developed to Farm/Developed). The determinant (Location) is also nominal with three categories, North South and river. In the analysis to follow, a reference group has to be chosen for comparison, the appropriate group would be the Natural remaining Natural, was chosen to be a reference group. The main problem with multinomial logistic regression is the enormous amount of output it generates; but there are ways to organize that output, both in tables and in graphs, that can make interpretation easier. (McNeil, 1998).

The preliminary analysis about our variables was going to present in chapter 3.