CHAPTER 4

RESEARCH METHOD

4.1 Data collection

4.1.1 Secondary data collection

The data and information on history of pine caterpillar outbreaks in the province was collected from the Provincial departments including Department of Forest Development and Department of Forest Protection of Thua Thien Hue, from the technical sections of State Forest Enterprises and district forestry protection stations. The information collection consisted of area infected, density of larva, time of pine caterpillar outbreaks, and age of infected pine forest. The data collected were recorded during a period of 17 years (from 1987 to 2003).

Data on weather conditions were collected from several provincial offices, including the technical section of the State Forest Enterprises, and district offices. Data from three hydro-meteorological stations that represent whole province have been collected namely; Hue station (Central province, include Hue city and Huong Thuy district), A Luoi station (Western and Northern of province, include A Luoi, Huong Tra and Phong Dien district), and Nam Dong station (Southern of province, include Nam Dong and Phu Loc district). In the place where pine caterpillar outbreaks have occurred the weather data were collected during the epidemic period for 17 years (1987- 2003).

The weather data were collected on the monthly basis. The data included average air temperature, highest air temperature, lowest air temperature, total rainfall, mean relative humidity, number of sunshine hours, and number of rainfall days. Secondary data (historical pine caterpillar outbreak and weather data) were used to develop statistical models.

The study covered three districts: Phong Dien, Huong Thuy, and Phu Loc districts that were good representative of all locations where have pine forests (Figure 4.1). These districts also represent all weather condition in Thua Thien Hue province.

Besides, the weather data and information about history of pine caterpillar outbreaks in Huong Tra district (between Phong Dien and Huong Thuy district), and Quang Tri province (Thua Thien Hue neighbor province) were collected to validate the model (Figure 4.1).

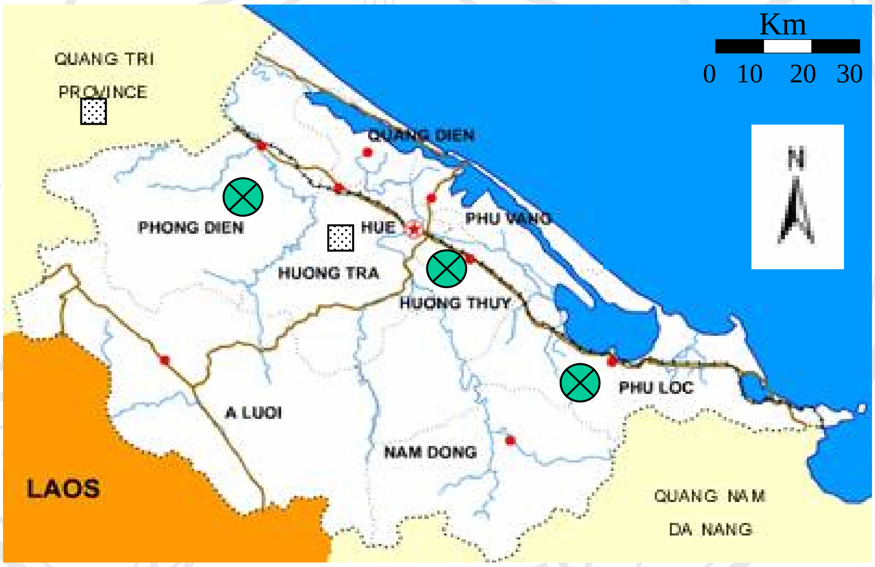

Figure 4.1: Map of study site for data collection.

(Source: www.thuathienhue.gov.vn/gioithieu/bando/).

Note: ⊗ *Locations where data were collected to develop models.*

▨ *Locations where data were collected to validate models.*

4.1.2 Primary data collection

Field survey was conducted in four districts (Phong Dien, Huong Tra, Huong Thuy, and Phu Loc) where pine forests are concentrated, and pine caterpillar epidemic usually occur. In each district, four forests at different forest age groups (younger than five years old, between six and 10 years old, from 11 to 15 years old, and between 16 and 20 years old) were selected for surveys. The forest, which was older than 20 years old, was not selected because the outbreaks do not often occur in the forest. In each forest, five plots were set up and five trees per plot were selected to collect the data. Field surveys were carried out in December when the epidemic often occurs. In field surveys, the data collection included the density of forest, the age of forest, the area damaged by pine caterpillar, and the average density of larvae in each area (forest age). The data from field surveys were used to validate the statistical models.

4.2 Data analysis

4.2.1 The conceptual framework for developing and validating statistical model

The statistical models were developed to establish the relationships between independent variables and dependent variables. In this study, the independent variables including weather factors (continuous variables) and forest age (dummy variables). The dependent variables were larvae density in December (quantitative variable) and epidemic occurrence in December (qualitative variable).

Because larvae density in December, a dependent variable, is quantitative variable so the multiple linear regression analysis was used to establish the relationships between larvae density in December with weather factors and forest age. However, the binary logistic regression analysis was used to establish the relationships between epidemic occurrence in December with weather factors and forest age because the dependent variable, epidemic occurrence in December, is qualitative variable.

From the developed models, some models were selected to validate based on the value of R-square and standard error of estimate. The most appropriate models were chosen by using root mean square error (RMSE), paired samples t-test, Wilcoxon signed-rank test (for multiple linear regression) or by using classification table and symmetric test (Phi test) for binary logistic regression. It can be seen clearly in Figure 4.2.

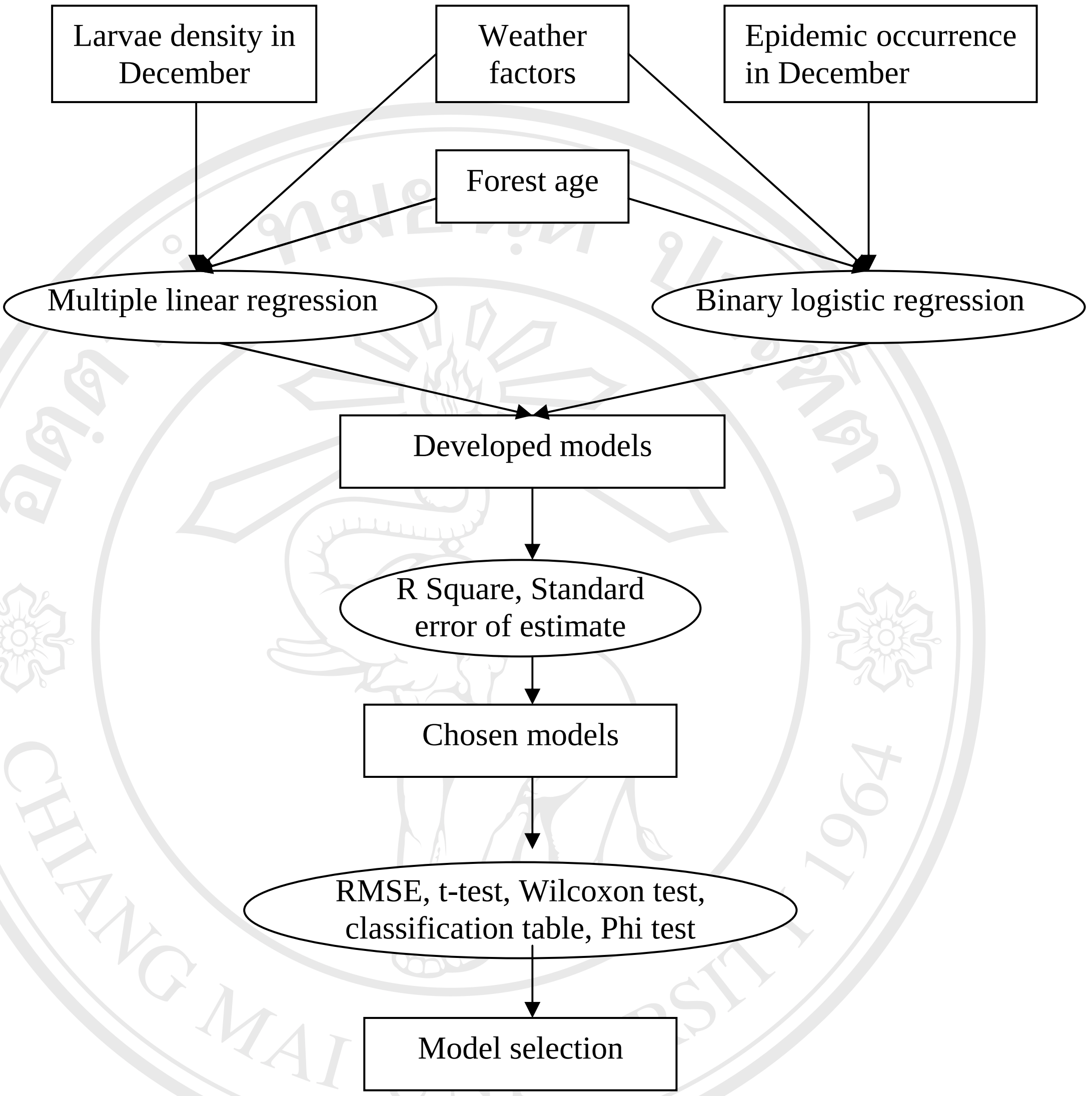
Larvae density in December
Weather factors
Epidemic occurrence in December
Forest age
Multiple linear regression
Binary logistic regression
Developed models
R Square, Standard error of estimate
Chosen models
RMSE, t-test, Wilcoxon test, classification table, Phi test
Model selection

Figure 4.2: The conceptual framework for developing and validating the statistical model to predict the outbreak of pine caterpillar.

4.2.2 Description of pine caterpillar distribution and other relative factors

4.2.2.1 Descriptive statistics

Descriptive statistics were used to modify the distribution of pine caterpillar outbreak (larvae density and outbreak frequency) and relative factors (forest age, forest density, and weather factors).

4.2.2.2 Cross tabulation (contingency table)

Cross tabulation (contingency table) was used to study the relationship between the epidemic occur and age of forest, or density of forest. χ^2 (chi-square) test was used to test the relationship between variables. The χ^2 test can be carried out by;

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] \qquad [4.1]$$

Where;

O and E = observed and expected frequencies, respectively (Gilbert, 1981).

The expected frequency is calculated by equation [4.2].

$$E = \frac{(Row\,Total)(Column\,Total)}{Grand\,Total} \qquad [4.2]$$

The hypothesis, there is no relationship between variables, can be tested basing on the χ^2 distribution with the degree of freedom is calculated by;

$$df = (R-1)(C-1) \qquad [4.3]$$

Where;

df = the degree of freedom

R = the number of rows in contingency table

C = the number of column in contingency table

4.2.3 Developing statistical models

4.2.3.1 Multiple linear regression analysis

The Multiple linear regression was used to develop the relationship between dependent variables, ***density of larvae in December***, and some independent variables.

$$\text{Density of larvae} = f(\text{weather factors, forest age}) \qquad [4.4]$$

The regression equation of this model can be shown as follows;

$$y = \alpha + \Sigma\beta_i X_i + \delta D + \varepsilon \qquad [4.5]$$

Where;

y = larvae density (number of larvae/tree in December) (LARDEN)

X_i = weather factors that include the highest temperature in the month (^{0}C) (TMAX), the lowest temperature in the month (^{0}C) (TMIN), average temperature in the month (^{0}C) (TAVER), average humidity in the month (%) (HUMID), total rainfall in the month (mm) (RAIN), number of rainfall days in the month (days) (RAIND), and number of sunshine hours in the month (hours) (SUNH).

D = dummy variable for forest age (FORAGE). Age of pine forests were classified into five categories: younger than five years old, between six and 10, between 11 and 15, between 16 and 20, and older than 20 years old.

β_i and δ = regression coefficients, α = intercept, and ε = error term.

According to the literature review, the outbreaks of pine caterpillar in Thua Thien Hue province usually occur in December when the fourth generation of this pest is growing. This study is an attempt to develop statistical models based on the weather factors during June to November when the third and fourth generation of this pest occurs.

Firstly, models were set up for each month with the same criteria of weather factors to find out which month can be used to predict the outbreak of this pest. Then in order to determine which weather factors have influence on the pine caterpillar outbreak, every weather factor has been used to develop models separately throughout the period. Finally, the significant independent variables from those models were used to develop another model. The result should be shown 14 models were established including six models for each month, seven models for each weather factor, and one model for selected independent variables.

The independent variables were entered in or removed from the model by using stepwise method supported by SPSS software. The significance of the relationship between individual partial independent variable and the response variable can be formally tested by using the t statistic, which is calculated by dividing the regression coefficient by the standard error of this independent variable, with (N-k-1) degree of freedom to evaluate the hypothesis that regression coefficient is equal to zero. If the regression coefficient is significantly different from zero, it means that there is a linear relationship between the explanatory and the response variable, and this explanatory variable should be entered to the model. Otherwise, the independent variable should be removed from the model (Hutcheson and Sofroniou, 1999).

To select the most appropriate model, all developed models need to validate. The models were selected to validate in next step were measured by some appropriate indicators those are R^2 statistic and standard error of estimate. The coefficient of determination (R^2) is calculated by;

$$R^2 = \frac{\sum\left(\hat{y} - \bar{y}\right)^2}{\sum\left(y - \bar{y}\right)^2} \qquad [4.6]$$

Where;

y = the observed value of y.

$\hat{y}$ = the value of y predicted from the model.

$\bar{y}$ = the mean value of y.

$\sum\left(\hat{y}-\bar{y}\right)^2$ = explained variation

$\sum\left(y-\bar{y}\right)^2$ = total variation

R^2 is commonly interpreted as the percentage of the variability in y (dependent variable) that is explained by x (independent variables) when it is used to predict y. It provides an indication of how well the model fits the data.

Whilst R^2 provides an indication of explanatory power of model, it does not indicate the level of significance. To do this need to test the hypothesis that the regression coefficient (β) equals zero. A test of this hypothesis is provided by the F test, which can be calculated using R^2 statistic (Hutcheson and Sofroniou, 1999).

$$F=\frac{R^2/k}{(1-R^2)/(N-k-1)} \qquad [4.7]$$

Where;

R^2 = the coefficient of determination

N = the number of cases used to construct the model (sample size)

k = the number of terms in the model (not including the constant)

If the value of F is not significant, the null hypothesis stated that there is a non-linear relationship between x and y, is accepted. If, on the other hand, the value of F is significant the null hypothesis is rejected and the alternative hypothesis stated that there is a significant linear relationship between x and y, is accepted (Hutcheson and Sofroniou, 1999).

The standard error of estimate (SE) is a measure of the errors that can be calculated by;

$$SE = \sqrt{\frac{\sum (y - \hat{y})^2}{N - k - 1}} \qquad [4.8]$$

Where;

SE = standard error of estimate

$\sum\left(y - \hat{y}\right)^2$ = error variation that can be calculated by

$$\sum\left(y - \hat{y}\right)^2 = \sum\left(y - \bar{y}\right)^2 - \sum\left(\hat{y} - \bar{y}\right)^2 \qquad [4.9]$$

Error variation should be as small as possible because the lower the standard error of estimate is the better model.

4.2.3.2 Binary logistic regression analysis

The binary logistic regression was used to assess the factors affecting on the outbreak of pine caterpillar.

Epidemic occur = f(weather factors, forest age) [4.10]

The equation for the relationship between dependent variable and the independent variables can be transformed into;

$$\text{Logit}(Y) = \alpha + \Sigma\beta_i X_i + \delta D \qquad [4.11]$$

Where;

Logit(Y) = the natural logarithm of the odds.

$$Logit(Y) = Ln\left(\frac{P_i}{1-P_i}\right) \qquad [4.12]$$

In which;

P_i = probability of epidemic occur in December. (It was called epidemic occurrence only if the average density of larvae was greater than five larvae per tree, the proportion of infected tree was greater than 50%, and the infected area was greater than one hectare).

$(1-P_i)$ = the probability of epidemic not occur.

$\frac{P_i}{1-P_i}$ = odds of epidemic occur.

The probability of epidemic occur can be calculated from the model by;

$$P_i = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i+\delta D)}} \qquad [4.13]$$

X_i are weather factors that included the highest temperature in the month (^{0}C) (TMAX), the lowest temperature in the month (^{0}C) (TMIN), average temperature in the month (^{0}C) (TAVER), average humidity in the month (%) (HUMID), total rainfall in the month (mm) (RAIN), number of rainfall days in the month (days) (RAIND), and number of sunshine hours in the month (hours) (SUNH) during the period between June and November.

D is dummy variable for forest age (FORAGE). Ages of pine forest were classified into five categories: younger than five years old, between six and 10, between 11 and 15, between 16 and 20, and greater than 20 years old.

$\beta_{i,}$ δ = regression coefficient and α = constant.

The same multiple regression was used for density of larvae. The models were developed by each month with all factors, individual weather factors in all months, and with selected independent variables from those kinds of models. The result should be shown 14 models were established including six models by each month, seven models by each weather factor, and one model by selected independent variables.

The independent variables were entered in or removed from the model by using stepwise method supported by SPSS software. The significance of the relationship between particular independent variable and the response variable can be tested by using the Wald statistic, which has a chi-square distribution when squared. Wald statistic is the ratio of the coefficient to its standard error. The significance level for the Wald statistic based on the chi-square distribution with the degree of freedom is one for single variable, but for the dummy variable, the degree of freedom equal to one less than the number of categories (Norušis, 2003).

Analogous the F test used in multiple linear regression, the Maximum Log-Likelihood ratio (Log(LL)) was used to test the hypothesis that all regression coefficients (β_i) equal zero for the logistic regression, but it is 2Log(LL), which is used in testing the hypothesis. The 2Log(LL) is calculated by equation [4.14].

$$2\text{Log(LL)} = 2[\text{LL (M)} - \text{LL (0)}] \qquad [4.14]$$

Where;

LL (0) = log likelihood of model with only constant variable

LL (M) = log likelihood of model with independent variables.

The 2Log(LL) has an approximated Chi-squared distribution with (n-1) degree of freedom. If χ^2 (model) is statistically significant ($p<0.05$) then we reject the null hypothesis and conclude that information about the independent variables allows us to make better prediction than we could make without the independent variables (Menard, 1995).

The goodness-of-fit of binary logistic models were measured by some important indicators such as Pearson χ^2, R^2_N, and in some case Hosmer-Lemeshow statistic (chi-square). In this work, the Hosmer-Lemeshow statistic was used.

The Cox and Snell R^2 (R^2_{CS}) and the Nagelkerke R^2 (R^2_N) statistics has the same purpose as R^2 in a linear regression model, although the variation in a logistic regression model must be defined differently. However, the Cox and Snell R^2 cannot achieve the maximum value of 1, so the Nagelkerke R^2 is a modification of Cox and Snell R^2 in order to achieve the value of 1. They are calculated by equation [4.15] and [4.16] (Norušis, 2003).

$$R^2_{CS} = 1 - \sqrt[N]{\left[\frac{L(0)}{L(M)}\right]^2} \qquad [4.15]$$

$$R^2_N = \frac{R^2_{CS}}{R^2_{MAX}} \qquad [4.16]$$

Where;

R^2_{CS} = the Cox and Snell R^2.

R^2_N = Nagelkerke R^2.

L(0) = the likelihood for the model with only a constant.

L(M) = the likelihood for the model included independent variables.

N = the sample size (number of case to construct model).

R^2_{MAX} is calculated by;

$$R^2_{MAX} = 1 - \sqrt[N]{\left[L(0)\right]^2} \qquad [4.17]$$

The Nagelkerke R^2 which shows the percentage of the variation in the outcome variable is explained by the logistic regression model.

The Hosmer-Lemeshow chi-square test was used to test the goodness-of-fit of the observed and predicted number of pine caterpillar outbreak occurs. The data were divided into 10 approximately equal groups based on the estimated probability of the epidemic occur and see how the observed and predicted number of the epidemic occur and non-occur. The chi-square value is calculated by;

$$\chi^2 = \sum\left[\frac{(O - E)^2}{E}\right] \qquad [4.18]$$

Where;

O and E = observed and predicted number of cases in each of the cells in table.

The degree of freedom is calculated as the number of groups minus 2. The observed significance level for the chi-square value should determine to accept or reject the null hypothesis, there is no different between the observed and predicted values (Norušis, 2003). All steps and procedures were processed by SPSS software.

4.2.4 Model selection and validation

First of all, six models (three models for multiple linear regression and three models for logistic) were selected. One model was set up by using all weather factors in each month, another one model developed by each weather factor in all months, and one model set up by selected independent variable. The models were selected based on the highest value of coefficient of determination (R^2), and the lowest value of standard error of estimate (SE) in multiple linear regression, and the highest of coefficient of determination (Nagelkerke R^2) and the lowest Hosmer-Lemeshow chi-square value for the logistic regression.

Using historical data from another province (Quang Tri), and district (Huong Tra), and from observation (field surveys) to validate the selected models, choosing the most appropriate model. For the binary logistic model, the most appropriate model should have the highest percentage of correctness in classification table and should be significant in symmetric test. However, for the multiple regression, the most appropriate model should have the lowest root mean square error (RMSE) value, and highest coefficient of determination (R^2) value. In addition should be significant by using paired sample t-test as well as Wilcoxon signed-ranks test that were supported by SPSS procedures.

The RMSE value is calculated as equation [4.19]

$$RMSE = \sqrt{\frac{1}{N}\sum\left(O - E\right)^2} \qquad [4.19]$$

Where;

O and E = observed and predicted value from the model, respectively.

N = the number of cases were used to test model.