CHAPTER 2

Methodology

This chapter describes the methodology including data source and data management for preliminary analysis, and statistical methods for analyzing the data. The graphical and statistical analyses were carried out using R program (R Development Core Team 2009).

This chapter presents graphical and statistical methods applied to abundance of resident birds. Graphical methods contain mosaic plots, bubble charts and box plots. The statistical methods include factor analysis and log-linear model.

2.1 Data source and Data management

Data used in this study were provided by the responsible agency of the Thale Noi non-hunting area, from Department of National Parks, Wildlife and Plant Conservation, Ministry of Natural Resources and Environment, Thailand for the period 2004 to 2007. Data were collected monthly by officers in seven stations of the Thale Noi non-hunting area for the four years (2004-2007). The surveys were designed to be conducted on the same day each month, simultaneously at each site, from 8 am to noon and from 1 pm to 4 pm, along the routes designed for counting birds. Species were identified using binoculars and field guides (Lekagul and Round 2005).

Therefore collected data included species names, date, month, year, site, numbers of bird sighted in a day (7 hours). These data were obtained in excel format which was modified and entered into suitable computer text files for managing and analyzing data as shown in Figure 2.1.

The term resident bird used in this study was defined using three criteria: (a) resident species status as categorized according to Lekagul and Round (2005), (b) the species was seen in the Thale Noi non-hunting area in each year of the data collection period, and (c) they had median incidence rate per day greater than zero. Using criteria (a) and (b), the bird species used thus contained 49 common resident birds among the 117 species observed for investigating bird distribution (study I). Moreover using criteria (a), (b) and (c) the bird species used thus contained 23 common resident birds among the 117 species observed for both investigating bird assemblages (study II) and examining methods for identifying incidence rates of bird species (study III). The study periods were defined as January-March, April-June, July-September, and October-December, giving 16 quarterly periods over four years, called ‘seasons’ for this study. Numbers of birds sighted in a day (7 hours) were converted to daily incidence rates. The incidence rates for each species were thus classified by 4 seasons, 7 sites, and 4 years.

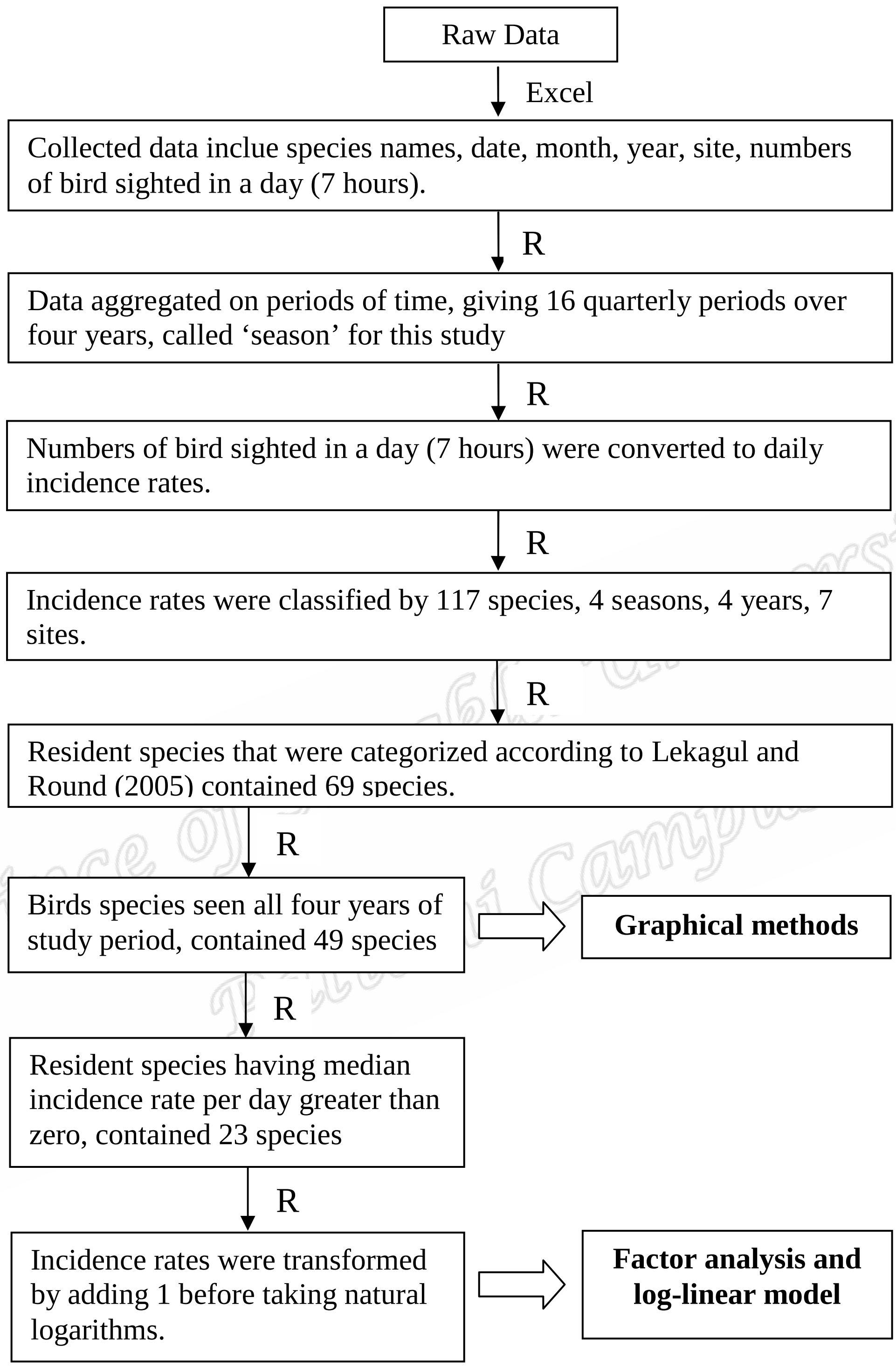

Figure 2.1 Path diagram for data management

2.2 Graphical methods

Mosaic plots (Hofmann 2003) were used to display the associations between overall incidence rates and the site and season factors. These graphics represent values such as counts or incidence rates in a contingency table by rectangular tiles with areas proportional to the values. Since mosaic plots become difficult to decipher when one of the factors in the contingency table has a large number of levels, we used bubble charts to depict the associations between the species-specific incidence rates and the site and seasonal factors. Bubble charts are structurally identical to mosaic plots, but use circles rather than rectangles.

Box plots were used to display incidence rates of each species in preliminary analysis before data were analysed by using factor analysis and log-linear model. These graphs indicate the degree of dispersion and skewness of data, and identify outliers (Tukey 1977).

2.3 Statistical methods

Appropriate statistical methods were used to identify groups of species with respect to incidence rates by season and location. The statistical methods used in this study with log-transformed data, after preliminary analysis by using box plots indicating the degree of dispersion and skewness of data, and identify outliers (Tukey 1977), are the factor analysis (Lattin et al. 2003) and log-linear model.

The factor analysis

The factor analysis (Lattin et al. 2003) used is essentially the same as that conventionally used in ecological studies (see, for example, Schnell et al. 1977, Pearson 1993, Sampantarak et al. 2011). The aim of the factor analysis in this study was to allocate the 23 bird species into a smaller number of interpretable groups that tended to appear on the same occasions. The incidence rates of the 23 species thus comprise the variables of interest. Each variable has 112 observations corresponding to combinations of season, year and site. The initial step is the determination of the matrix of correlation coefficients between these 23 variables. The second step is the estimation of factors (groups of species) from the correlation matrix. Ideally each factor (which must contain at least two species to contribute to the factor analysis) contains species that have large correlations with each other and small correlations with species in other groups. To achieve this, species that are not correlated with any other species are said to have high "uniquenesses" and are conventionally omitted from the factor analysis. The factors comprise weighted linear combinations of the species and may be rotated to maximize the weights within each group and minimize the weights outside each group. The resulting weights are called "loadings". Species are assigned to factors based on their loadings. "Promax" rotation was used in preference to "varimax", which requires the rotations to be orthogonal (Browne 2001).

Log-linear model

Since bird data used are both count and rate, Poisson and negative binomial were used for count data and log- linear model was used for incidence rates. The notations are described as follows, Supposed that n_{ist} is the number of observed birds in cells defined by species *i*, site *s*, and season *t* and t_{ist} is the corresponding period of time for observation (7-hour day). Denoting the corresponding mean incidence rate by λ_{ist} we consider negative binomial GLM of the form

$$\ln(\lambda_{ist}) \; = \; t_{ist} + \mu + \alpha_i + \beta_{st} \qquad (1)$$

The terms α_i and β_{st} represent species and season-site effects, respectively, and are centred at 0, so μ is a constant encapsulating the overall incidence rate. The generalized linear model based on Poisson distribution with mean λ_{ist} where t_{ist} is an offset term (Venables and Ripley 2002). The Poisson model is special case when variance equal to mean. The negative binomial distribution has mean λ and variance λ(1+λ/θ), so its variance-mean ratio is greater than 1. The smaller values of θ correspond to greater dispersion. The model fit is assessed by comparing deviance residual with normal quantiles, and it is also informative to plot observed counts against corresponding fitted values based on the model. The model also gives adjusted incidence rates for each factor of interest.

The model with sum contrasts is used rather than treatment contrast (Venables and Ripley 2002, Tongkumchum and McNeil 2009) for constructing confidence intervals to compare the incidence rates for each level of factor with the overall mean incidence rate. An advantage of these confidence intervals is that they can be classified into three groups according to their location entirely above, around, or below, the overall mean.

A log-linear model can be used for the rate. The additive log-linear model for the incidence rates with normally distributed error is

$$\ln\left(n^*_{ist} / t_{ist}\right) \; = \; y_{ist} \; = \; \mu + \alpha_i + \beta_s + \gamma_t \qquad (2)$$

In this model μ is a constant and α_i, β_s and γ_t are effects of species, sites and periods respectively. In this case n^*_{ist} is a simple modification of the number of observed birds and t_{ist} is the corresponding period of time for observation (7-hour day). To ensure that each number of observed birds is positive before taking log transformation, to each is added a constant 1. This method is a common practice since these transformed rates are finite and remain zero when the incidence rate is zero (Clark and Warwick 1994). The model fit is assessed by plotting deviance residuals against normal quantiles for the negative binomial model and similarly by plotting standardized residuals against normal quantiles for log-linear regression model. The r-squared is used to see how much of the variation in the data is accounted for by the model.