Chapter 21
GLMM Applied on the Spatial Distribution of Koalas in a Fragmented Landscape

J.R. Rhodes, C.A. McAlpine, A.F. Zuur, G.M. Smith, and E.N. Ieno

21.1 Introduction

Predicting the spatial distribution of wildlife populations is an important component of the development of management strategies for their conservation. Landscape structure and composition are important determinants of where species occur and the viability of their populations. In particular, the amount of suitable habitat and its level of fragmentation (i.e. how broken apart it is) in a landscape can be important determinants of the distribution and abundance of biological populations (Hanski, 1998; Fahrig, 2003). In addition to the role of habitat, anthropogenic impacts, such as wildlife mortality on roads or direct wildlife-human conflict, can also have large impacts on the distribution and abundance of a species (Fahrig et al., 1995; Woodroffe and Ginsberg, 1998; Naves et al., 2003). Therefore, if we are to manage landscapes to successfully conserve wildlife, it is important that we understand the role of these landscape processes in determining their distributions.

In this chapter, we will model the impact of landscape pattern on the distribution of koalas (*Phascolarctos cinereus*, Fig. 21.1) in a landscape in eastern Australia. Koalas are folivorous arboreal marsupials restricted to the eucalypt forests of eastern and southeastern Australia. Across their geographic range, they feed on a wide range of tree species from the genus *Eucalyptus*, but mostly prefer only a few species in any particular area (Hindell and Lee, 1987; Phillips and Callaghan, 2000; Phillips et al., 2000). Koala habitat generally consists of forest associations containing their preferred tree species, although other factors, such as tree size, water availability, and nutrient status, can also be important determinants of habitat quality (Moore et al., 2004; Matthews et al., 2007). Since European settlement, koalas have suffered declines in their abundance and distribution due to clearing and degradation of eucalypt forests, together with historical hunting, disease, bushfire, drought, and urbanisation (ANZECC, 1998; Melzer et al., 2000; Phillips, 2000).

J.R. Rhodes (✉)
The University of Queensland, School of Geography, Planning and Architecture, Brisbane, QLD 4072, Australia

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_21,

Fig. 21.1 Young koala (photo by Dick Marks, Australian Koala Foundation. www.savethekoala.com)

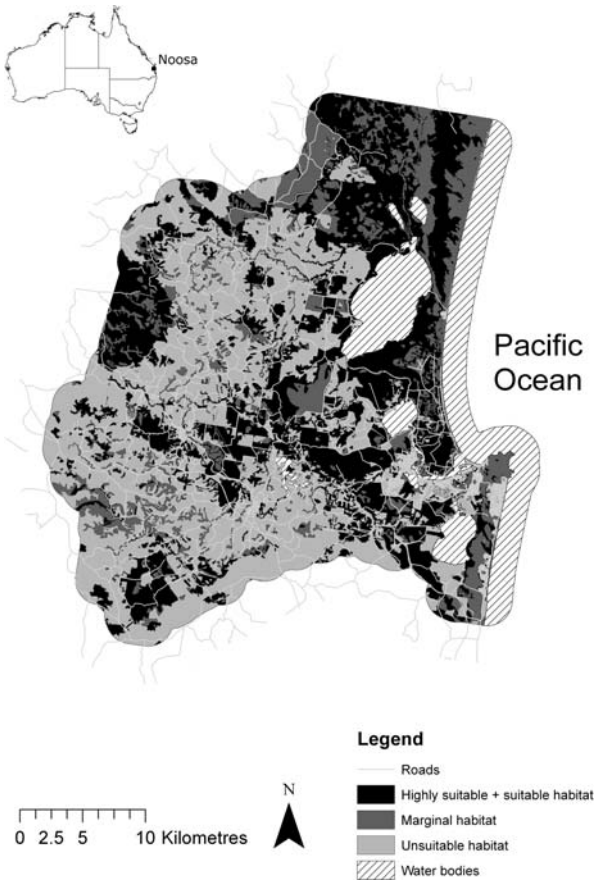

Fig. 21.2 Map of the study area (Noosa Local Government Area) showing the distribution of koala habitat and the location of roads (Australian Koala Foundation unpublished data)

The study area we consider for this chapter is the Noosa Local Government Area (LGA) in southeast Queensland, Australia (Fig. 21.2). Noosa has a subtropical coastal climate with native vegetation ranging from coastal heath to wet and dry eucalypt forests and subtropical rainforests. Over 50% of the original eucalypt forests have been cleared for farming and urban development (Seabrook et al., 2003). Koalas are, therefore, threatened by the loss and fragmentation of their habitat and by threats associated with urbanisation, such as cars and domestic dogs. To allow successful management strategies to be developed, it is important for conservation planners to be able to quantify the impact of these threats on koala distributions in the area.

We will use generalised linear mixed effects models (GLMM) to model the distribution of koalas using data on their presence and absence at sites located across the study area. We also take a multi-scale approach in the sense that our explanatory variables will be landscape characteristics measured at different landscape extents. The different landscape extents will be chosen to represent those scales thought to be most relevant for koala population dynamics, and hence their distributions. The chapter concentrates on dealing with collinearity and spatial auto-correlation for these types of landscape models. In addition, we present an information-theoretic approach to model selection, which allows us to assess both model and parameter uncertainty. We finish with a discussion on the implications of the results for koala conservation and what should be included in a scientific paper.

21.2 The Data

The data presented are based on surveys that were conducted to determine koala presence or absence at 300 locations in Noosa. This formed part of a larger study investigating the role of landscape change on koala distributions across eastern Australia (McAlpine et al., 2006; Rhodes et al., 2006). Using a form of stratified random sampling (McKay et al., 1979; Thompson, 1992) 100 sites were first located across the Noosa LGA. Then, within each site, three subsites were located 100 m apart. At each subsite, the presence or absence of koalas was then determined using standardised searches for koala faecal pellets around the bases of trees (as in Phillips and Callaghan, 2000). Previous work has identified the koala's preferred tree species in Noosa and these have been classified into primary, secondary, and supplementary species (Australian Koala Foundation (AKF) unpublished data). At each subsite, the percentage of trees that were primary and secondary species was recorded. Finally, the distribution of koala habitat (classified into highly suitable, suitable, marginal, and unsuitable habitat) and the location of paved roads were mapped in a geographical information system (AKF unpublished data, Fig. 21.2). The data on the presence/absence of koalas will be the response variable for our analysis, while the data on preferred tree species, habitat and roads will form the basis of the explanatory variables.

The data set can be accessed in R using the following code:

```
> library (AED); data (Koalas)
```

The resulting data frame contains a row for each subsite. The first two columns are the site and subsite ID numbers, the next two columns are the eastings and northings of the location of each subsite (in AMG ADG 1966 coordinates), the fifth column indicates whether koala pellets were found at the subsite (= 1) or not found at the subsite (= 0), and the remaining columns are the explanatory variables associated with each subsite (Table 21.1).

The explanatory variables were chosen to represent characteristics of the landscape considered likely to be important determinants of the distribution of koalas. The variables can be split into those characterising habitat at the site-scale, and those characterising habitat and human impacts at broader landscape-scales (i.e., within 1, 2.5, or 5 km buffers around each subsite). The site-scale habitat variables (`pprim_ssite` and `psec_ssite`) measure the percentage of primary and secondary tree species at each subsite and reflect resource availability at this scale. Two of the landscape-scale variables (`pss` and `pm`) measure the percentage of the landscape, within each buffer, that is highly suitable plus suitable habitat and marginal

Table 21.1 Description of the explanatory variables

Variable name	Description	Detail description
`pprim_ssite`	Resources available at site-scale	Percentage of trees in each subsite that are primary tree species
`psec_ssite`	Resources available at site-scale	Percentage of trees in each subsite that are secondary tree species

`phss_1km`
`phss_2.5km`
`phss_5km`
	Habitat available at landscape-scale	Percentage of the landscape within 1, 2.5, and 5km, respectively, of each subsite that is highly suitable plus suitable habitat

`pm_1km`
`pm_2.5km`
`pm_5km`
	Habitat available at landscape-scale	Percentage of the landscape within 1, 2.5, and 5 km, respectively, of each subsite that is marginal habitat

`pdens_1km`
`pdens_2.5km`
`pdens_5km`
	Landscape fragmentation	Density (patches/100 ha) of habitat patches, consisting of highly suitable plus suitable plus marginal habitat, in the landscape within 1, 2.5, and 5 km, respectively, of each subsite

`edens_1km`
`edens_2.5km`
`edens_5km`
	Landscape fragmentation	Density (m/ha) of habitat patch edges, consisting of highly suitable plus suitable plus marginal habitat, in the landscape within 1, 2.5, and 5 km, respectively, of each subsite

`rdens_1km`
`rdens_2.5km`
`rdens_5km`
	Human impact at landscape-scale	Density (m/ha) of paved roads within 1, 2.5, and 5 km, respectively, of each subsite

habitat, respectively. These variables represent the amount of habitat resources available at the landscape-scale. Two of the landscape-scale variables (`pdens` and `edens`) measure the density of habitat patches and the density of habitat edges within each buffer, respectively. These variables represent the level of landscape fragmentation; patch density and edge density both tend to increase as habitat becomes more fragmented. Finally, one of the landscape-scale variables (`rdens`) measures the density of roads in each buffer and represents the level of human impact due to koala mortality of roads and general urbanisation.

21.3 Data Exploration and Preliminary Analysis

Two important issues to consider before building regression models of species' distributions are whether there is high collinearity between the explanatory variables and whether spatial auto-correlation between data points is likely to be an important factor. High collinearity can result in coefficient estimates that are difficult to interpret as independent effects and/or have high standard errors (Neter et al., 1990; Graham, 2003). Positive spatial auto-correlation violates the usual assumption of independence between data points and leads to the underestimation of standard errors, and elevated type I errors, if not accounted for (Legendre, 1993). Collinearity between explanatory variables and spatial auto-correlation are commonly encountered when using observational data to construct regression models of species' distributions. For both these issues, we examine whether they are likely to be a problem for the analysis of our dataset and then discuss how they can be addressed.

21.3.1 Collinearity

A simple first step for identifying collinearity is to look at the pairwise correlations between explanatory variables. We can generate a matrix of pairwise correlations between the explanatory variables in our dataset using the following code:

```
> cor(Koalas[, 6:22], method = "spearman")
```

This outputs a matrix of the Spearman rank correlations (results are not given here as it is too large). We have used the Spearman rank correlation coefficient, rather than the Pearson correlation coefficient because the Spearman rank correlation makes no assumptions about linearity in the relationship between the two variables (Zar, 1996). One could also use the `pairs` command to view pairwise plots of the variables. Booth et al. (1994) suggest that correlations between pairs of variables with magnitudes greater than ± 0.5 indicate high collinearity, and we use this rough rule-of-thumb here.

The first thing you notice from the correlation matrix is that the landscape variables measuring the same characteristic at different landscape extents tend to be highly positively correlated. For example, `phss_5km`, `phss_2.5km`, and

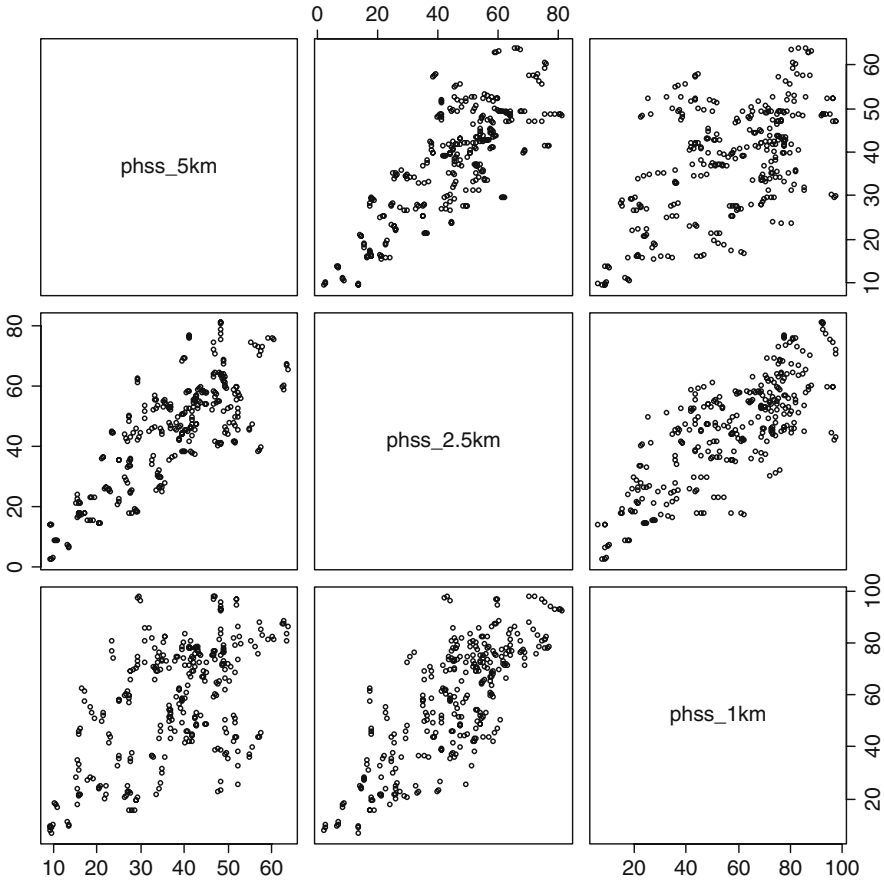

Fig. 21.3 Pairplot of the `phss_5km`, `phss_2.5km`, and `phss_1km` explanatory variables

`phss_1km` show high correlations with each other (Fig. 21.3). These variables measure the amount of highly suitable plus suitable habitat within distances of 5, 2.5, and 1 km of each subsite, respectively, and so they are spatially nested within each other (Fig. 21.4). The collinearity therefore arises because the variables calculated at the smaller landscape extents partly measure the same landscape characteristics as the variables calculated at the larger landscape extents.

You will also notice that the two landscape variables measuring habitat fragmentation (`pdens` and `edens`) are also highly positively correlated with each other. Areas with high patch densities tend to contain habitat patches that are smaller than those found in areas with low patch densities. Since small patches have more edge than large patches, this means that areas with high patch densities also tend to have high edge densities and vice versa, hence the high positive correlation. Finally, some of the patch density (`pdens`) and edge density (`edens`) variables tend to be somewhat negatively correlated with some of the habitat amount variables (`phss`

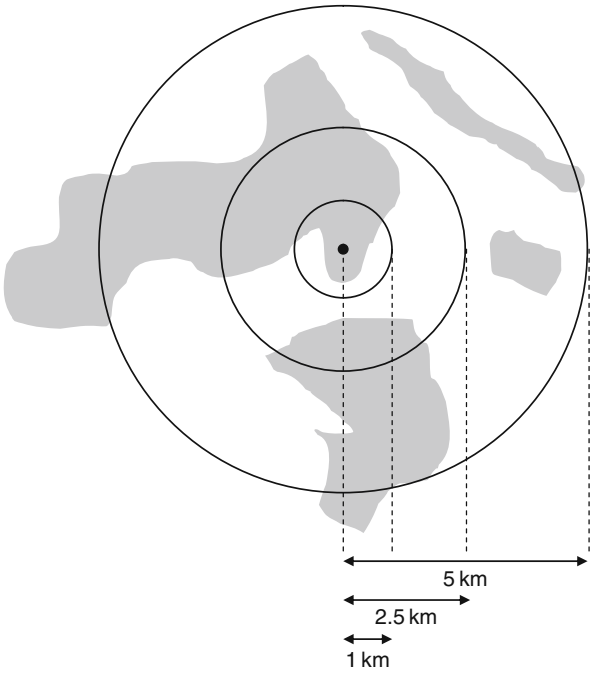

Fig. 21.4 Illustration of the nested landscape extents within which the landscape variables were calculated. The point in the centre represents a hypothetical subsite and the shaded areas represent hypothetical koala habitat

and pm). This occurs because the same processes that lead to habitat loss also tend to lead to a breaking apart of that habitat (i.e. fragmentation), resulting in greater numbers of patches with more edges. Therefore, landscape variables that measure fragmentation are often found to be correlated with those that measure habitat amount (Fahrig, 2003). However, in our data set, these correlations are only marginally more negative than –0.5 and are not considered a major concern at this stage.

There are several strategies that we could use to deal with the high collinearity found between the explanatory variables. These include (i) simply removing one or more variables so that the remaining variables are not highly correlated (Neter et al., 1990; Booth et al., 1994), (ii) using linear combinations of the variables rather than the variables directly in the model (Chatterjee and Price, 1991; Trzcinski et al., 1999; Villard et al., 1999), or (iii) using biased estimation procedures such as principal components regression or ridge regression (Neter et al., 1990; Chatterjee and Price, 1991). Here, we use the first two of these approaches to deal with collinearity because they are relatively straightforward to implement and appear adequate for our purposes.

We calculated the landscape variables at different landscape extents, because we were interested in the impact of landscape characteristics measured at different scales on koala presence at a site. We, therefore, ideally want to retain the nested structure, but reduce collinearity between the variables so that the coefficients in the model can be estimated precisely. To do this we recast each variable as linear

combination of the other variables. Suppose $\mathbf{X}_5$, $\mathbf{X}_{2.5}$, and $\mathbf{X}_1$ are landscape variables measured at the 5, 2.5, and 1 km landscape extents respectively. We can then create a new set of variables $\mathbf{Z}_5$, $\mathbf{Z}_{2.5}$, and $\mathbf{Z}_1$ such that:

$$\begin{aligned} \mathbf{Z}_5 &= \mathbf{X}_5 \\ \mathbf{Z}_{2.5} &= \mathbf{X}_{2.5} - \mathbf{X}_5. \\ \mathbf{Z}_1 &= \mathbf{X}_1 - \mathbf{X}_{2.5} \end{aligned} \tag{21.1}$$

Here the variable measured at the 5 km extent has remained the same, while the variables measured at the 2.5 and 1 km extents have been recalculated as the difference between the original variable and the one that it is nested within. We would expect the variables $\mathbf{Z}_5$, $\mathbf{Z}_{2.5}$, and $\mathbf{Z}_1$ to be less correlated with each other than $\mathbf{X}_5$, $\mathbf{X}_{2.5}$, and $\mathbf{X}_1$. This is because the new variables represent the value of the original variables relative to those they are nested within, rather than their absolute values. Now, if we use the variables $\mathbf{Z}_5$, $\mathbf{Z}_{2.5}$, and $\mathbf{Z}_1$, instead of $\mathbf{X}_5$, $\mathbf{X}_{2.5}$, and $\mathbf{X}_1$, in our regression model, the collinearity problem should be reduced and our coefficient estimates will be more precise.

To demonstrate the reduction in collinearity, consider the percentage of highly suitable plus suitable habitat variable (`phss`). First we need to create the new variables:

```
> Koalas$phss_2.5km_new <- Koalas[, "phss_2.5km"] -
                           Koalas[, "phss_5km"]
> Koalas$phss_1km_new <- Koalas[, "phss_1km"] -
                         Koalas[, "phss_2.5km"]
```

Note that we do not need to create a new variable for `phss_5km`; this variable always remains the same. We will also need to create new variables, called `pm_2.5km_new`, `pm_1km_new`, `pdens_2.5km_new`, `pdens_1km_new`, `edens_2.5km_new`, `edens_1km_new`, `rdens_2.5km_new`, and `rdens_1km_new` for each of the other landscape variables in a similar way (the code is on the book website). The reduction in collinearity for the percentage of highly suitable plus suitable habitat variables can be seen by looking at the correlation matrix for the new variables using the code:

```
> cor(Koalas[, c("phss_5km", "phss_2.5km_new",
              "phss_1km_new")], method = "spearman")
```

which shows substantially lower correlation between the variables (results are not given here). This reduced collinearity can also be seen by looking at pair plots for the new variables (Fig. 21.5) compared to the pair plots for the original variables (Fig. 21.3). The same reduction in collinearity is also seen in the other landscape variables.

In using this approach, it is important to note that the regression coefficients for the new variables will have different interpretations to those for the original

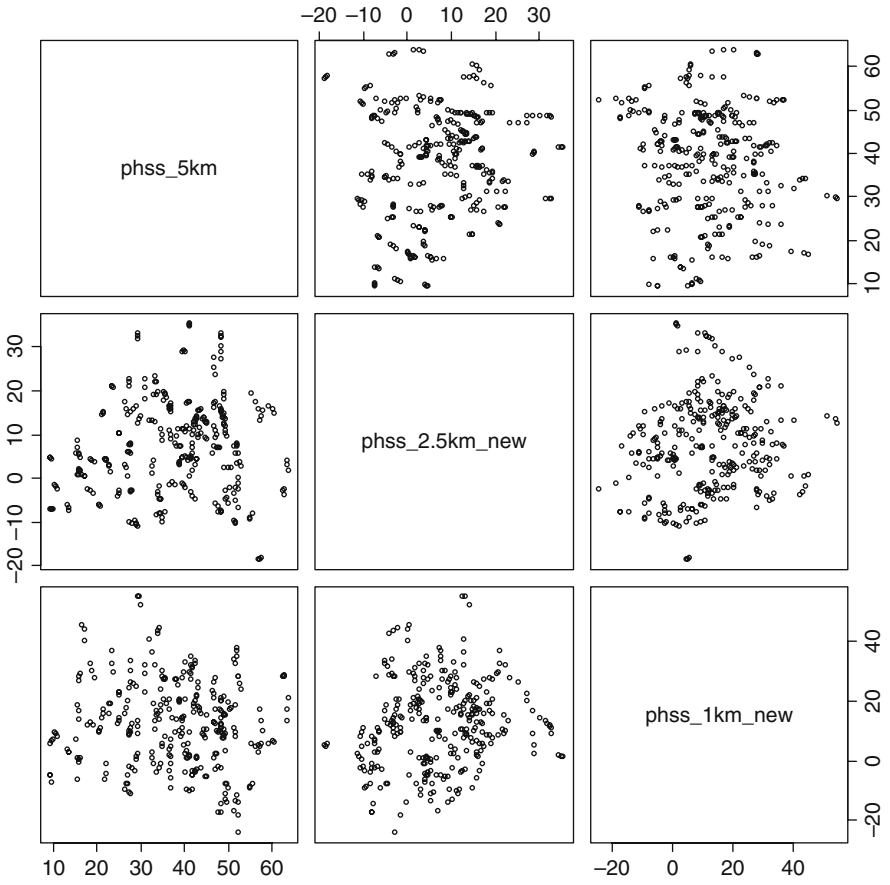

Fig. 21.5 Pairplot of the variables `phss_5km`, `phss_2.5km_new`, and `phss_1km_new`

variables. Fortunately, the coefficients for our new variables have a useful interpretation in terms of understanding the impact of landscape characteristics on koala presence. The interpretation of the coefficients for variables measured at the largest landscape extents remains the same. These coefficients quantify the broad-scale landscape effects on koala presence. However, the coefficients for variables measured at smaller landscape extents now represent landscape effects relative to the broader scale landscape context. This is a useful interpretation because it incorporates the dependence between fine-scale and broad-scale landscape effects on species distributions (O'Neil, 1989). Here, careful choice of the linear combinations of variables has resulted in new variables that are not highly correlated and have a useful interpretation. However new variables constructed from linear combinations of variables are not always so easily interpreted. Chatterjee and Price (1991) provide a good discussion on how to choose appropriate combinations of variables.

To deal with the collinearity between patch density (pdens) and edge density (edens) we could construct new variables based on linear combinations of the original variables. However, in this case, there are no obvious linear combinations that would result in easily interpreted coefficients. Many applications of species' distribution models require explanation to planners and the general public. Therefore, the ease of interpretation of the model is an important model building consideration, and rather than developing composite measures of patch density and edge density, we will simply retain only one of the variables as a measure of habitat fragmentation. The variable we retain is patch density because this is a straightforward and easily interpreted measure of fragmentation.

Having taken the steps described above, we now look at the variance inflation factors (VIFs) of the variables to assess the extent of any remaining collinearity. To do this, we first fit a generalised linear model with binomial response and logit link function (i.e. a logistic regression model), containing all explanatory variables, to the presence/absence data (McCullagh and Nelder, 1989; Hosmer and Lemeshow, 2000) and then calculate the VIFs for each variable from the resulting model. We use the vif function in the package Design to calculate the VIFs. The code to do this is as follows:

```
> Glm_5km <- glm(presence ~ pprim_ssite + psec_ssite +
      phss_5km + phss_2.5km_new + phss_1km_new +
      pm_5km + pm_2.5km_new + pm_1km_new + pdens_5km +
      pdens_2.5km_new + pdens_1km_new + rdens_5km +
      rdens_2.5km_new + rdens_1km_new,
      data = Koalas, family = binomial)
> library(Design)
> vif(Glm_5km)
```

and the output is:

Variable	VIF	Variable	VIF
pprim_ssite	1.121	psec_ssite	1.099
phss_5km	3.196	phss_2.5km_new	1.584
phss_1km_new	1.495	pm_5km	1.931
pm_2.5km_new	1.575	pm_1km_new	1.973
pdens_5km	2.474	pdens_2.5km_new	1.600
pdens_1km_new	1.273	rdens_5km	2.130
rdens_2.5km_new	1.368	rdens_1km_new	1.095

You can see that all the VIFs are well below 10, suggesting that collinearity is no longer a major issue (Neter et al., 1990; Chatterjee and Price, 1991). However, some authors do suggest a more stringent cut-off than this. For example, Booth et al. (1994) suggest that VIFs should ideally be less than 1.5. Later in this chapter, we consider alternative regression models where the largest landscape extent is only

2.5 or 1 km, rather than 5 km. In these cases, the variables measured at the largest landscape extent remain as the original variables, and new variables are only constructed for those variables nested within the largest landscape extent. Therefore, we also need to check the VIFs for the variables included in these models because the variable set is slightly different. This can done using the code

```
> Glm_2.5km  <- glm(presence ~ pprim_ssite +
      psec_ssite + phss_2.5km + phss_1km_new +
      pm_2.5km + pm_1km_new + pdens_2.5km +
      pdens_1km_new + rdens_2.5rdens_1km_new,
      data = Koalas, family = binomial)
> vif(Glm_2.5km)
```

for the 2.5 km maximum extent and the code

```
> Glm_1km <- glm(presence ~ pprim_ssite + psec_ssite +
      phss_1km + pm_1km + pdens_1km + rdens_1km,
      data = Koalas, family = binomial)
> vif(Glm_1km)
```

for the 1 km maximum extent. Note that for the 1 km maximum landscape extent, there are no new variables because there is no nesting within the 1 km extent. The VIFs for all variables are considerably less than 10 in both these cases. Therefore, the measures we have taken seem to have successfully reduced collinearity to acceptable levels.

21.3.2 Spatial Auto-correlation

There are two reasons for expecting spatial auto-correlation in the presence/absence data. First, spatial auto-correlation at the site-scale may occur because the distances between the subsites within individual sites are small relative to the size of koala home ranges. Average koala home range sizes in similar east coast habitats have been estimated at between 10–25 ha for females and 20–90 ha for males (AKF unpublished data, J. R. Rhodes unpublished data). Therefore, the occurrences of koalas at subsites within an individual site will tend to be correlated because they would often have been located within the same koala's home range. Second, spatial auto-correlation at broader scales may occur due to spatially constrained dispersal of koalas from their natal home ranges. Koala dispersal distances in nearby regions have been recorded to be around 3–4 km, but can be as high as 10 km (Dique et al., 2003). So, dispersal distances are substantially smaller than the spatial extent of the study area, and this could also lead to spatial auto-correlation between sites. We could also see spatial auto-correlation in the presence/absence data if the underlying spatial pattern of habitat is spatially auto-correlated. However, we would expect our explanatory variables to account for most of the spatial auto-correlation from this

source once the regression model is fitted to the data and is therefore considered to be of less concern.

One way to assess the extent of spatial auto-correlation is to look at correlograms of the data (Cliff and Ord, 1981; Bjørnstad and Falck, 2001). Correlograms are graphical representations of the spatial correlation between locations at a range of lag distances. Positive spatial correlation indicates that spatial auto-correlation between data points may be a problem. Negative spatial correlation may also indicate a problem, but this is fairly unusual in this kind of data; so we are mainly concerned with positive correlations. We use a spline correlogram to investigate auto-correlation in the presence/absence data. The spline correlogram that we use is essentially a correlogram that is smoothed using a spline function (Bjørnstad and Falck, 2001). To produce the correlograms, we need the `ncf` package (http://asi23.ent.psu.edu/onb1/software.html). A spline correlogram of the presence/absence data can be plotted using the code

```
> library(ncf)
> Correlog <- spline.correlog(x = Koalas[,  "easting"],
                              y = Koalas[,  "northing"],
                              z = Koalas[,  "presence"], xmax  = 10000)
> plot.spline.correlog(Correlog)
```

which produces Fig. 21.6A; a spline correlogram with 95% pointwise bootstrap confidence intervals and maximum lag distance of 10 km (note that it may take several minutes for this to run). You can see from the correlogram that significant positive spatial auto-correlation is present, but only at short lag distances of less than around 1 km. This suggests that spatial auto-correlation may be an issue for subsites located close to each other.

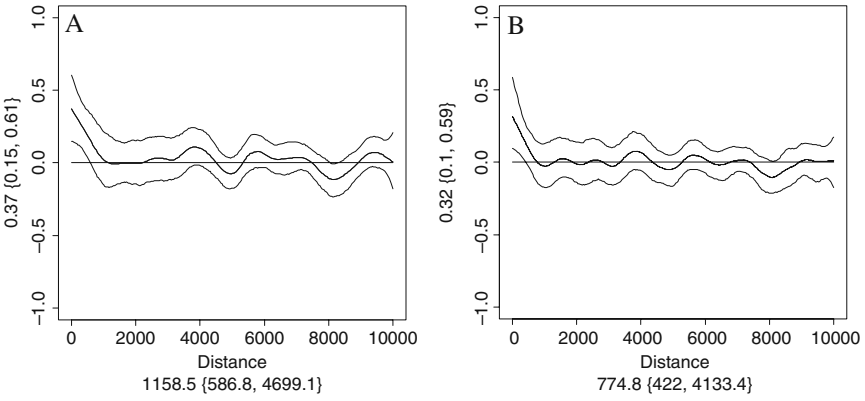

Fig. 21.6 Spline correlograms, with 95% pointwise bootstrap confidence intervals, of (**A**) the raw presence/absence data and (**B**) the Pearson residuals from a logistic regression model, including all the explanatory variables, fitted to the data

However, although spatial auto-correlation in the raw data is of interest, we are predominantly interested in whether there is any spatial auto-correlation in model residuals once any spatial auto-correlation explained by the explanatory variables has been accounted for. Therefore, we also look at the spatial auto-correlation in the Pearson residuals of the logistic regression model, containing all explanatory variables, that we fitted to the presence/absence data earlier in this chapter (`Glm_5km`). The following code will plot a spline correlogram of the Pearson residuals of this model:

```
> Correlog_Glm_5km  <-
       spline.correlog(x = Koalas[,  "easting"],
       y = Koalas[,  "northing"], xmax = 10000,
       z = residuals(Glm_5km, type = "pearson"))
> plot.spline.correlog(Correlog_Glm_5km)
```

and it produces Fig. 21.6B. Although there seems to be some overall reduction in spatial auto-correlation, compared to the raw data, significant positive spatial auto-correlation at short lag distances still remains. As significant positive auto-correlation only exists at short lag distances, it is probably the result of correlation between subsites within sites, rather than correlation between sites. Since the data are nested and the spatial scale of nesting coincides with the spatial scale of auto-correlation, one reasonably straightforward way to deal with this problem is to use GLMM (McCulloch and Searle, 2001). This approach would take account of dependencies within sites and we discuss the approach in more detail in the next section. However, if the data were not nested or the spatial scale of auto-correlation and the spatial scale of nesting did not coincide (e.g. if the dependencies occurred between sites, rather than within sites), then mixed effects models are likely to be less useful and alternative approaches are likely to be required. Alternatives include a broad range of autoregressive and auto-correlation models that explicitly incorporate the spatial dependence between locations (Keitt et al., 2002; Lichstein et al., 2002; Miller et al., 2007). A full discussion of these methods is beyond the scope of the chapter, but they are worth being aware of as alternatives for dealing with spatial auto-correlation.

21.4 Generalised Linear Mixed Effects Modelling

GLMMs are useful when data are hierarchically structured in some way. They account for dependencies within hierarchical groups through the introduction of random-effects (Pinheiro and Bates, 2000; McCulloch and Searle, 2001). In this study, the data are hierarchically structured in the sense that subsites are nested within sites, and we want to use mixed effects models to account for the spatial dependencies within sites. A suitable mixed effects model for these purposes can be constructed by introducing a random-effect for site into the standard logistic regression model. The resulting mixed effects model looks like this:

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{\beta}' \times \mathbf{X}_{ij} + b_i, \tag{21.2}$$

where p_{ij} is the probability of koala presence at subsite j in site i; $\boldsymbol{\beta}$ is a vector of model coefficients; $\mathbf{X}_{ij}$ is a vector of explanatory variables for subsite j in site i; and b_i is the random-effect for site i. Here, the b_i are drawn from a random variable B, that we will assume is normally distributed with a mean of zero and variance of σ^2, i.e., $B \sim \text{Normal}(0, \sigma^2)$. However, other random distributions can be assumed.

This provides an appropriate framework for modelling the distribution of koalas in our study area, but before progressing, we should first check that it will adequately account for the spatial auto-correlation that is present. To do this, we fit a logistic GLMM, including all the explanatory variables, to the data and once again look at a spline correlogram of the Pearson residuals. To fit the model, we will use the `glmmML` function in the package `glmmML`. Later in this chapter we compare alternative models using Akaike's information criteria (AIC) that require the calculation of the maximum log-likelihood of each model (Akaike, 1973; Burnham and Anderson, 2002). We use the `glmmML` function here because it estimates the model parameters by maximum likelihood and allows AICs to be calculated. An alternative would be to use the `lmer` function in the package `lme4` with the `Lapacian` or `adaptive Gauss-Hermite` methods. However, reliable AIC values cannot be calculated using some other mixed effects model functions such as `glmmPQL` in the package `MASS` because it maximises a penalised quasi-likelihood, rather than the full likelihood. The code to fit the mixed effects model is as follows:

```
> library(glmmML)
> Glmm_5km <- glmmML(presence ~ pprim_ssite +
          psec_ssite + phss_5km + phss_2.5km_new +
          phss_1km_new + pm_5km + pm_2.5km_new +
          pm_1km_new + pdens_5km + pdens_2.5km_new +
          pdens_1km_new + rdens_5km + rdens_2.5km_new +
          rdens_1km_new, cluster = site, data = Koalas,
          family = binomial)
```

The `cluster` argument indicates the grouping level for the random-effect. A spline correlogram of the Pearson residuals can then be generated using the code:

```
> Correlog.Glmm_5km <- spline.correlog(
         x = Koalas[, "easting"],
         y = Koalas[, "northing"],
         z = pres.glmmML(model = Glmm_5km,
         data = Koalas), xmax = 10000)
> plot.spline.correlog(Correlog.Glmm_5km)
```

which produces Fig. 21.7. Here the call to the function `pres.glmmML` (which can be found at the book website) calculates the Pearson residuals for the model. You

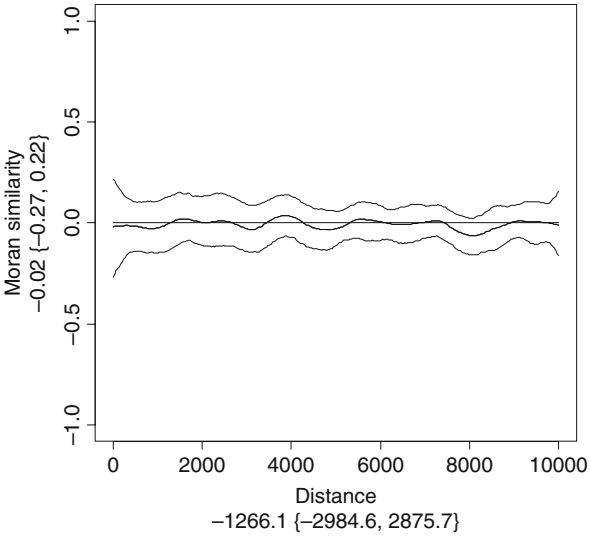

Fig. 21.7 Spline correlogram, with 95% pointwise bootstrap confidence intervals, of the Pearson residuals from a mixed effects logistic regression model, including all the explanatory variables, fitted to the data

can now see that there is no longer any obvious increase in spatial correlation at short lag distances. This suggests that the mixed effects model successfully accommodates the spatial auto-correlation within sites. This also helps to confirm that the main source of spatial auto-correlation at short lag distances is indeed the dependency between subsites within sites. In the following sections, we therefore use mixed effects logistic regression to model koala distributions in Noosa.

21.4.1 Model Selection

We have now identified a suitable set of explanatory variables and an appropriate modelling framework. The next step is to identify which of the variables are important determinants of koala distributions and to identify a suitable and parsimonious approximating model that we can use to make predictions. Rather than using traditional null-hypothesis testing procedures for variable selection to achieve these aims, we will use an information-theoretic approach (Burnham and Anderson, 2002). Information-theoretic approaches provide a framework that allows multiple model comparisons to be made and the most parsimonious of these models to be identified. The process of identifying a parsimonious model involves trading off model bias against model precision and information-theoretic approaches achieve this by using appropriately constructed criteria to compare models (Burnham and Anderson, 2002). The criteria we use here is AIC, which is defined as

$$\mathrm{AIC} = -2L + 2K, \tag{21.3}$$

where L is the maximum log-likelihood of the model and K is the number of parameters in the model (Akaike, 1973). A model with a low AIC is more parsimonious

than a model with a high AIC. Note, however, that it is only the relative differences in AIC values between models that are important and that the absolute value of a model's AIC is meaningless (Burnham and Anderson, 2002). Information-theoretic approaches have certain advantages over traditional null-hypothesis testing approaches (Johnson, 1999; Anderson et al., 2000; Burnham and Anderson, 2001; Lukacs et al., 2007). These advantages include the ability to (i) evaluate multiple non-nested models relative to each other, (ii) quantify the relative support for multiple models simultaneously, and (iii) derive predictions that account for model uncertainty using model averaging; but see critiques by Guthery et al. (2005) and Stephens et al. (2005).

To implement this approach, we first develop a series of alternative mixed effects models that include different combinations of the explanatory variables. These alternative models can be thought of as different 'hypotheses' about the relationships between koala presence/absence and the explanatory variables. We then examine the support from the data for each of these models using AIC (sensu Hilborn and Mangel, 1997). This will be achieved by fitting each model to the data and ranking them by their AIC values. We will also calculate the relative probability of each model being the best model by calculating their Akaike weights, w_i. The Akaike weight for model i is defined as

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{j=1}^{R} \exp\left(-\frac{1}{2}\Delta_j\right)}, \tag{21.4}$$

where Δ_i is the difference between the AIC for model i and the model with the lowest AIC and the sum is over all the alternative models in the set $j = 1, \ldots, R$. Akaike weights are useful because they can be used to identify a 95% confidence set of models, and ratios of Akaike weights (evidence ratios) provide quantitative information about the support for one model relative to another (Burnham and Anderson, 2002). A 95% confidence set of models can be constructed by starting with the model with the highest Akaike weight and repeatedly adding the model with the next highest weight to the set until the cumulative Akaike weight exceeds 0.95. Akaike weights can also be used to calculate the relative importance of a variable by summing the Akaike weights of all the models that include that variable (Burnham and Anderson, 2002). We will therefore also calculate the 95% confidence set of models and the relative importance of the landscape-scale habitat amount, fragmentation, and road density variables.

In constructing the alternative models, we group the explanatory variables into four functional groups (1) site-scale habitat (`pprim_ssite` and `psec_ssite`); (2) landscape-scale habitat amount (`phss` and `pm`); (3) landscape-scale habitat fragmentation (`pdens`); and (4) landscape-scale road density (`rdens`). There is good evidence from other studies that site-scale habitat characteristics are a key determinant of the use of a site by koalas (Phillips and Callaghan, 2000; Phillips et al., 2000). Therefore, we include site-scale habitat in all the models and for

each landscape extent (1, 2.5, and 5 km), construct a model for all combinations of the landscape-scale habitat amount, landscape-scale habitat fragmentation, and landscape-scale road density variables. This leads to a total of 22 alternative models. However, we also construct a 'null' model that includes no explanatory variables as a check of our assumption of the importance of the site-scale variables. Note that for each landscape extent, the variables spatially nested within that spatial extent are also included in the model.

Before fitting each of these models to the data, the explanatory variables should be standardised so that they each have a mean of zero and standard deviation of one. This helps to improve convergence of the fitting algorithm and puts the estimated coefficients on the same scale, allowing effect sizes to be more easily compared. We can standardise the explanatory variables using the code

```
> Koalas_St <- cbind(Koalas[, 1:5],
      apply(X = Koalas[, 6:ncol(Koalas)], MARGIN = 2,
      FUN = function(x){(x - mean(x)) / sd(x)}))
```

which creates a new data frame, called `Koala_St`, of the standardised variables. We use these standardised variables as the explanatory variables in fitting the alternative models.

Rather than showing the code to fit each of the alternative models, we show the code to fit one of the models as an example. The code to fit the model including the site-scale habitat variables and the landscape-scale habitat amount variables at the 1 km extent is

```
> glmmML(presence ~ pprim_ssite + psec_ssite +
       phss_1km + pm_1km, cluster = site,
       data = Koalas_St, family = binomial)
```

which gives the following output:

```
                        coef se(coef)      z Pr(>|z|)
(Intercept)          -0.7427   0.2314 -3.210 0.001330
pprim_ssite           0.8576   0.2244  3.822 0.000132
psec_ssite            0.2319   0.1938  1.196 0.232000
phss_1km              0.2765   0.2479  1.115 0.265000
pm_1km                0.5573   0.2524  2.208 0.027200

Standard deviation in mixing distribution:  1.561
Std. Error:                                 0.3005
Residual deviance: 354.5 on 294 degrees of freedom
AIC: 366.5
```

This shows that the probability of koala presence increases with the percentage of preferred tree species at a subsite and the percentage of habitat in the surrounding landscape. The standard deviation of the random-effect is 1.56 and the model's AIC is 366.5.

The AICs, Akaike weights, and model rankings for all the models in the 95% confidence set are shown in Table 21.2. This table also shows the relative importance of landscape-scale habitat amount, fragmentation, and road density variables. The first thing to note is the large number of models in the 95% confidence set of models (14), indicating there is considerable model uncertainty. The Akaike weights confirm this with no models much more likely to be the best model than the other models. The best model includes the site-scale habitat and landscape-scale habitat amount variables at the 1 km extent. However, this model is only 1.7 times more likely to be the best model than the next best model, which also includes landscape-scale road density (evidence ratio = 0.174/0.101). In general the models at the 1 km landscape extent performed better than the models at the 2.5 and 5 km landscape extents. This suggests there is little gain in predictive performance from adding additional variables representing the landscape at extents broader than 1 km. The relative variable importances suggests that landscape-scale habitat amount and landscape-scale road density are more important determinants of koala distributions than landscape-scale fragmentation. However, due to the high model uncertainty, the differences in relative importance are not particularly large. Finally, the null

Table 21.2 The 95% confidence set of models

Rank	Site-scale habitat	Landscape-scale habitat amount	Landscape-scale habitat fragmentation	Landscape-scale road density	Landscape extent (km)	AIC	*w*
1	√	√			1	366.5	0.174
2	√	√		√	1	367.6	0.101
3	√				–	367.7	0.097
4	√			√	5	367.8	0.092
5	√	√			5	367.9	0.087
6	√	√	√		1	368.1	0.082
7	√			√	1	368.2	0.075
8	√	√	√	√	1	369.1	0.048
9	√	√			2.5	369.3	0.043
10	√		√		1	369.7	0.036
11	√			√	2.5	369.9	0.032
12	√		√	√	1	370.2	0.028
13	√		√		5	370.7	0.021
14	√	√		√	5	370.8	0.021
Relative importance	**–**	**0.590**	**0.261**	**0.431**			

AIC = Akaike's information criteria; *w* = Akaike weights; site-scale habitat = `pprim_ssite` + `psec_ssite`; landscape-scale habitat amount = `phss_1km` + `pm_1km` (1km extent), `phss_2.5km` + `phss_1km_new` + `pm_2.5km` + `pm_1km_new` (2.5km extent), `phss_5km` + `phss_2.5km_new` + `phss_1km_new` + `pm_5km` + `pm_2.5km_new` + `pm_1km_new` (5km extent); landscape-scale habitat fragmentation = `pdens_1km` (1km extent), `pdens_2.5km` + `pdens_1km_new` (2.5km extent), `pdens_5km` + `pdens_2.5km_new` + `pdens_1km_new` (5km extent); landscape-scale road density = `rdens_1km` (1km extent), `rdens_2.5km` + `rdens_1km_new` (2.5km extent), `rdens_5km` + `rdens_2.5km_new` + `rdens_1km_new` (5km extent).

model has an AIC of 382.2 and relative to the model only containing the site-scale habitat variables (AIC = 367.7), has an evidence ratio of almost zero. This indicates very strong support for our assumption that site-scale habitat variables are important determinants of koala presence or absence at a site.

Given there is no single model that is clearly the best, a sensible approach is to acknowledge this model uncertainty and make inferences based on model averaging (Burnham and Anderson, 2002). Model averaging allows coefficients to be estimated and model predictions to be made that account for the inherent model uncertainty in addition to parameter uncertainty. In essence, these approaches derive weighted average predictions, where the weights are the relative model probabilities. When model uncertainty is present, this has considerable advantages over more traditional step-wise and null-hypothesis approaches to model selection, where you only end up with a single best model. Model averaged predictions are likely to be more robust than those derived from a single best model. Burnham and Anderson (2002) provide useful guidelines for conducting model averaging using AIC, and see McAlpine et al. (2006) and Rhodes et al. (2006) for examples of model averaging applied to predicting koala distributions.

21.4.2 Model Adequacy

So far, we have examined the relative support from the data for each model. However, this tells us little about how well the models fit the data or whether there are any departures from model assumptions. Traditionally, the fit of logistic regression models have been assessed using global goodness-of-fit tests based on the deviance or Pearson χ^2 statistics. However, the distributional properties of these statistics are not well understood, making the tests somewhat difficult to apply in practice (Hosmer and Lemeshow, 2000). Further, despite the convenience of global goodness-of-fit tests, it is unclear to what extent it is sensible to condense model fit into a single number or test (Landwehr et al., 1984). An alternative to global goodness-of-fit tests is to use a range of graphical methods to assess how well a model fits the data. Here, we concentrate on quantile-quantile plots and partial residual plots (Landwehr et al., 1984). Logistic regression quantile-quantile plots are useful for assessing whether the error distribution of the data is modelled correctly and to detect more general departures from model assumptions. Partial residual plots are useful for assessing systematic departures from model assumptions, such as linearity. We will apply these diagnostic procedures to the most parsimonious model, although they can equally be applied to model averages if model averaged predictions are to be made.

A quantile-quantile plot consists of a graph of quantiles of residuals assuming the fitted model is the true model, against the actual quantiles of the residuals from the fitted model. If there are no major deviations from the model assumptions, then these points should lie close to the 1:1 line. Since the distribution of the residuals in logistic regression is not well understood, Landwehr et al. (1984) propose a simulation approach for constructing a logistic regression quantile-quantile plot. Their basic approach is as follows:

1. From the fitted model, calculate the residuals r_i.
2. Order the r_i, giving $r_{(i)}$.
3. Simulate M data sets from the fitted model.
4. Fit the model to the M data sets.
5. Compute the residuals r_i^* for the models fitted to the M data sets and order them to get $r_{(i)}^*$.
6. Calculate the medians of the ordered residuals from the M replicates. (Landwehr et al. (1984) use a slight modification here where they interpolate within the distribution of the simulated residuals to avoid plotting negative against positive residuals.)
7. Plot the median simulated ordered (interpolated) residuals against the ordered residuals from the original model fit.
8. Calculate confidence intervals for the simulated ordered (interpolated) residuals from the M replicates.
9. Plot the median simulated ordered (interpolated) residuals against the upper and lower confidence intervals.

If we apply this approach to the most parsimonious model, with $M = 1000$, we get the plot shown in Fig. 21.8. The code for creating this plot and the required functions `res.glmmML` and `fitted.glmmML` can be found at the book website. You will see that the points lie quite close to the 1:1 line and within the simulated 95% point-wise confidence interval. This suggests there are no major departures from the model assumptions.

The partial residual plot for a particular covariate consists of a graph of the values of the covariate against its partial residuals. Partial residuals (r_{par}) are defined as

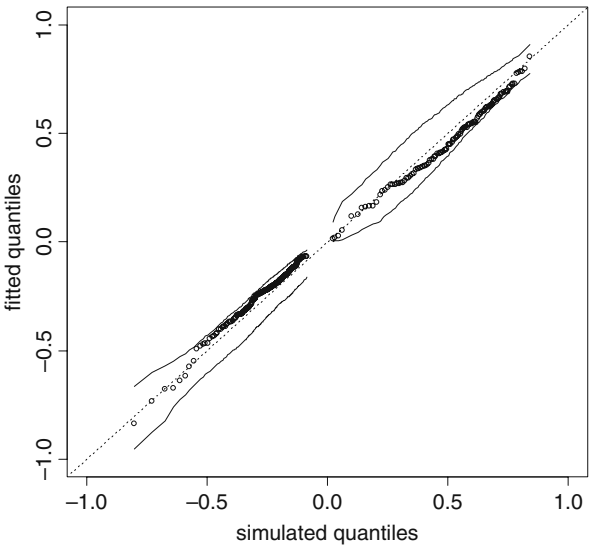

Fig. 21.8 Quantile-quantile plot with 95% pointwise confidence bounds

$$r_{\text{par}} = \frac{y - \hat{p}}{\hat{p} \times (1 - \hat{p})} + X \times \hat{\beta}_X \tag{21.5}$$

where y is the observed data (1 or 0), $\hat{p}$ is the estimated probability for the fitted model, X is the covariate value, and $\hat{\beta}_X$ is the estimated coefficient for the covariate X for the fitted model (Landwehr et al., 1984). If a partial residual plot is linear, then a linear assumption for this covariate is appropriate. However, if a partial residual plot is non-linear, this indicates that a linear assumption may not be appropriate, and in that case, the shape of the curve can suggest an appropriate functional form for the covariate. Due to the dichotomous nature of binomial data, partial residual plots for logistic regression show two groups of points; one for the 1 observations and one for the 0 observations. Therefore, it is necessary to fit a smoothed curve to the points to assess whether it is linear or non-linear. The partial residual plots for the four covariates in the most parsimonious model with smoothed curves fitted using the `loess` function are shown in Fig. 21.9. The code for creating these plots and the required functions `res.glmmML` and `fitted.glmmML` can be found at the book website. All of the curves are moderately non-linear, but especially so for

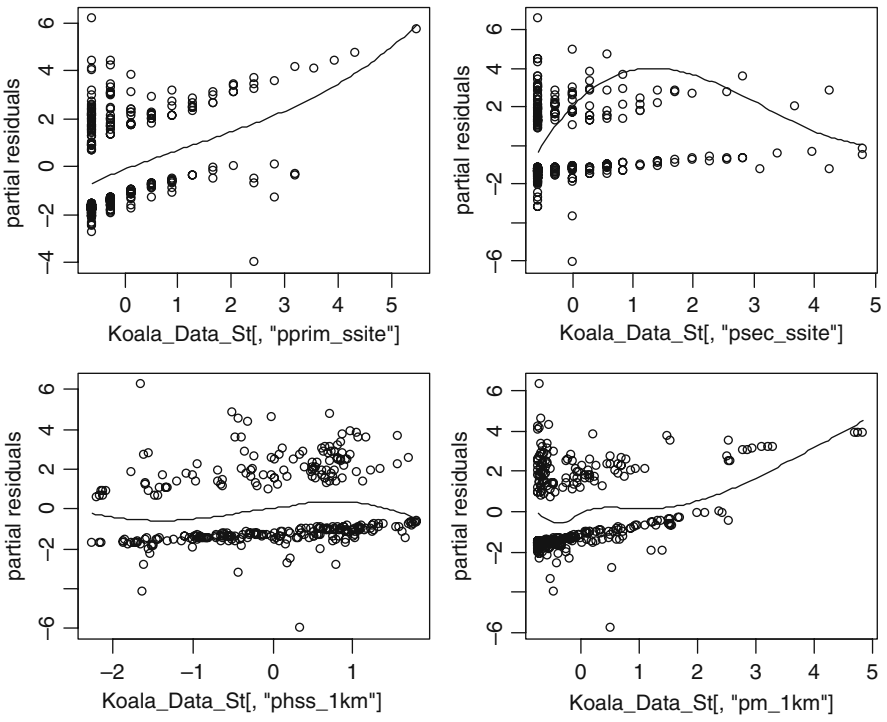

Fig. 21.9 Partial residual plots for `pprim_ssite`, `psec_ssite`, `phss_1km`, and `pm_1km` for the highest ranked model

the psec_ssite curve. The shape of the psec_ssite curve suggests that the inclusion of a quadratic term for this covariate might be appropriate.

Re-fitting the most parsimonious linear model with a quadratic term for psec_ssite gives the model

```
                          coef se(coef)      z Pr(>|z|)
(Intercept)            -0.4809   0.2576 -1.866 0.062000
pprim_ssite             0.8908   0.2292  3.887 0.000101
psec_ssite              0.9161   0.3718  2.464 0.013700
I(psec_ssite^2)        -0.2820   0.1360 -2.074 0.038100
phss_1km                0.3095   0.2522  1.227 0.220000
pm_1km                  0.5972   0.2581  2.314 0.020700

Standard deviation in mixing distribution:  1.581
Std. Error:                                 0.3065
Residual deviance: 349.3 on 293 degrees of freedom
AIC: 363.3
```

which confirms the improvement in the model with a reduction in AIC of 3.2 units. Since this is a more parsimonious model than the linear model, the preference would be to use this to make predictions, rather than the linear model, or alternatively to include models with a quadratic term for psec_ssite in the model set for making model-averaged predictions.

In considering the adequacy of our models, we have only compared model predictions against the data that they were fitted to. However, we often want to use species' distribution models to make predictions for a new area or for a new site. In this case, simply comparing predictions to the data used to fit the models will tend to overestimate the predictive performance of the models. One way to overcome this is to fit the models to one data set and then compare model predictions to an independent data set (Pearce and Ferrier, 2000). This is known as cross-validation. However, we rarely have the luxury of a completely independent data set; so simulation-based cross-validation using random samples from the data used to fit the models is often used instead (Stone, 1974; Efron and Tibshirani, 1997). We do not consider these approaches in detail here, but they are important aspects of model validation and it is important to be aware of them. For specific discussion on the validation of wildlife distribution models, see Pearce and Ferrier (2000) and Vaughan and Ormerod (2005).

21.5 Discussion

In this chapter, we have demonstrated the use of GLMM for modelling species distributions. The use of GLMM was an effective way of dealing with spatial autocorrelation in the data, but this may not always be the case, such as if spatial autocorrelation existed between sites. However, other approaches, such as autoregres-

sive models, do exist that could be used to deal with between-site auto-correlation (e.g., Miller et al., 2007). We also found that constructing simple linear combinations of nested landscape variables was useful for reducing collinearity, while still maintaining an easily interpreted model. This approach is particularly useful for landscape-scale studies such as this, where landscape effects are often conceptualised as occurring at a range of nested spatial extents. We also demonstrated an information-theoretic approach (using AIC) to model selection and the identification of the most parsimonious models. The information-theoretic approach allowed us to quantify the level of model uncertainty and provided the potential to calculate model-averaged predictions. Model-averaged predictions are useful in contexts such as the one presented here, where there is reasonably high model uncertainty, because predictions are not conditional on a single model (Burnham and Anderson, 2002). The information-theoretic framework was also found to be useful for ranking the landscape-scale covariates in terms of their importance. Identifying the importance of each covariate in this way has an important practical application for prioritising management actions for the conservation of koalas.

One of the primary aims of this chapter was to model koala distributions to help understand the key landscape- and site-scale factors determining the presence of koalas. We found strong evidence that the percentage of preferred tree species at the site-scale was positively related to koala occupancy. This is consistent with other studies indicating that koalas often select certain preferred tree species (Phillips and Callaghan, 2000, Phillips et al., 2000) or select habitats containing high proportions of preferred tree species (Rhodes et al., 2005). We also found that koala occupancy was positively related to the amount of habitat at the landscape-scale, which was more important than the density of roads, which in turn was more important than habitat fragmentation. It is generally accepted that the amount of habitat tends to be more important than habitat fragmentation for the viability of wildlife populations (Fahrig, 2003). Our analyses suggest this is the case for the koala in Noosa and that the conservation priority should be habitat protection, rather than just seeking particular landscape configurations that minimise fragmentation. However, fragmentation effects may become more important as habitat is lost (Flather and Bevers, 2002). It is interesting to note that road density was almost as important as habitat amount. Increasing road density decreases the chance of finding koalas and this may simply reflect the general effects of urbanisation and associated threatening processes. It is generally accepted that areas around habitat patches, known as the habitat matrix, can have important implications for the viability of species (Ricketts, 2001). This may be what is happening here with factors associated with urban development, such as vehicle collision mortality and dog attacks, negatively impacting koala populations. Mitigation of these factors would therefore also seem to be an important conservation priority for koalas in Noosa.

It is interesting to note that the landscape-scale variables measured at the 1 km scale tended to be the best descriptors of koala presence (Table 21.2). We would expect the scale at which the landscape affects the presence of koalas to be related to the scale of koala movements such as natal dispersal and movements within individual home ranges. Koalas have average dispersal distances of several kilometres

(Dique et al., 2003), and so the scale of the landscape effects is at the shorter end of the distribution of koala dispersal distances. This suggests that the spatial dynamics of koala populations in Noosa are influenced predominantly by koalas dispersing over short distances and by movements of individuals within their home ranges, rather than by less common long distance dispersal movements.

21.6 What to Write in a Paper

When writing a scientific paper you need to be selective about what you include, while still ensuring that the methods are sufficiently detailed to allow readers to repeat your study and that the research findings are clearly explained. We have presented a great deal more information in this chapter than would be required for a scientific paper. Although there is no single recipe for what to include and what not to include in a paper, based on the analysis presented in this chapter, we give a broad outline of what we think should be included.

In the introduction section, we would aim to give a clear statement of the biological and wildlife management issues addressed by the research. The last paragraph of this section should explicitly state the specific questions that the research addresses, and very briefly, outline what was done. In the methods section, we would have a description of the study site and the data collection methods. Then we would briefly describe the exploratory analysis we conducted in relation to collinearity and spatial auto-correlation. Although the description of these steps should be brief, it would be important to describe the transformations of the explanatory variables and perhaps include the graphs showing the reduction in collinearity (e.g. Figs. 21.3 and 21.5). The remainder of the methods section should then describe the alternative models we fitted to the data, the use of AIC in comparing the models, and the methods used to assess model adequacy. The results section should include a description of the key findings of the statistical analyses and the assessment of model adequacy. It is not necessary to describe every single aspect of these results, but sufficient details should be included to give the reader a clear picture of the key findings. Other things to include here would be a table showing the model rankings with AICs, coefficient estimates, and standard errors for at least the best model(s) and graphical demonstration of model adequacy (e.g. Fig. 21.8). A useful additional figure that we do not show here would be a map of predictions and their associated standard errors based on the best, or model-averaged, model (see, e.g. Rhodes et al. (2006)). Finally, the discussion section should indicate the implications of the results in terms of the issues raised in the introduction and highlight the applied or theoretical advances the study has made. A key component of the discussion should be identifying any limitations of the work and suggesting future research directions.

Acknowledgments This work was funded by the Australian Research Council, the Australian Koala Foundation, and The University of Queensland.