A statistical analysis of relationships between Koala occupancy and tree habitat in the Bermagui / Murrah area, NSW

Alan Welsh, Ross Cunningham and Christine Donnelly, ANU
May 2010

Objective

In the Bermagui / Murrah area, the optimal habitat for Koalas is not well known. The purpose of this report is firstly to present results of a detailed statistical analysis of Koala habitat selection based on the plot data provided, and secondly, to provide brief comment on other statistical analyses of these data (see Appendix).

Available data

Sample Design

Plot data were collected by Chris Allen and others (DECCW 2010). The survey teams conducted searches for evidence of Koalas in a pre-determined grid pattern across the study area. Initially the searches were conducted at sites located at either 350 m or 500 m intervals[1]. After 14 months of fieldwork the survey data was analysed. It was concluded that the larger areas of Koala activity would still be detected at sites at 1,000 m intervals. This coarser sampling interval was adopted for the remainder of the survey in order to increase the geographic coverage.

For sites where Koala activity was detected (detection sites), neighbouring sites within 350m of the detection sites were also surveyed to provide finer resolution in the delineation of the boundaries of the active areas.

The sample design and detected Koala activity are shown in Figures 1 and 2.

The sampling design incorporates systematic variation in the sampling intensity over the study region so the sample design is *non-uniform*. Initially, the sampling intensity did not depend on the distribution of the Koalas so the design is non-informative or *non–preferential*. However, the later fieldwork involved some preferential selection in that the sampling intensity was increased when evidence of Koala activity was found. This can have implications for subsequent inference: Diggle and Ribeiro (2007, “Model-based Geostatistics”) state that provided the sampling process is non-preferential the choice of the design does not impact on the statistical model for the data, but it does affect the precision of inferences which can be made from the data. The effects of a preferential design depend on the nature and purpose of the study. For example, estimation of population size is strongly affected by preferential sampling. In our analysis of this study, we convert the data to presence/absence of Koalas for each site and analyse occupancy. The effect of the preferential sampling is to increase the number of presences in the sample and make standard errors for the probability of presence smaller than they should be. However, the preferential sampling was used to delineate the regions of presence more precisely so the direct impact on the relationship of interest, the relationship between presence and habitat, is probably small.

Measurement

At each site, the base of the closest 30 live trees over 150mm diameter at breast height (dbh) were searched for Koala faecal pellets out to a metre from the trunks. The species and dbh of each of the 30 trees were recorded. The radius of each of the sites was only recorded in the later half of the survey. Of those sampled (n=202), the variation was not substantial (range 8-40m, mean = 20.1 ± 0.4 m), with 70% of these sites having radii between 15-25m. Covariates are recorded at the tree level but for convenience and simplicity of modeling we have derived site level covariates in a similar manner to McDougall and Saxon.

[1]

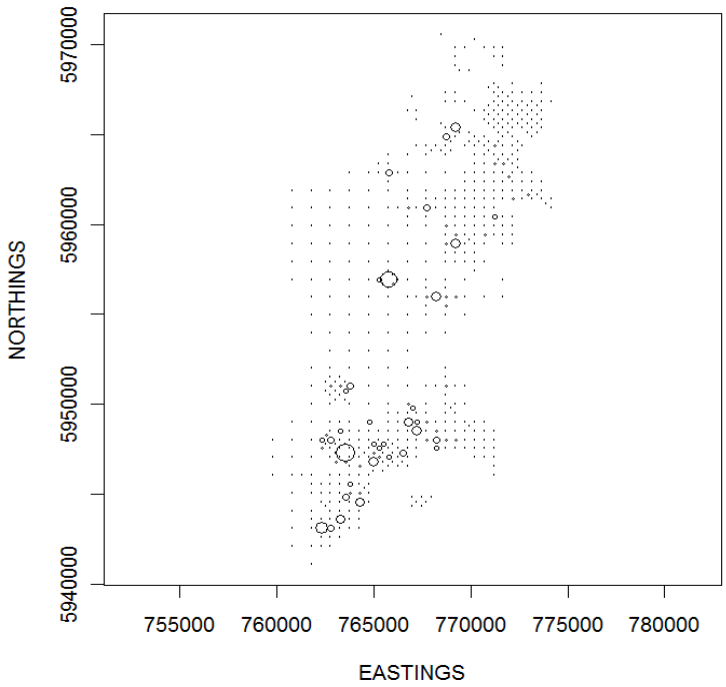

Figure 1. Sample design and koala activity. Each site is represented by a circle proportional in size to the number of trees where koala faecal pellets were found. The largest circle corresponds to a site with 9 trees and the smallest circles correspond to sites with no koala activity.

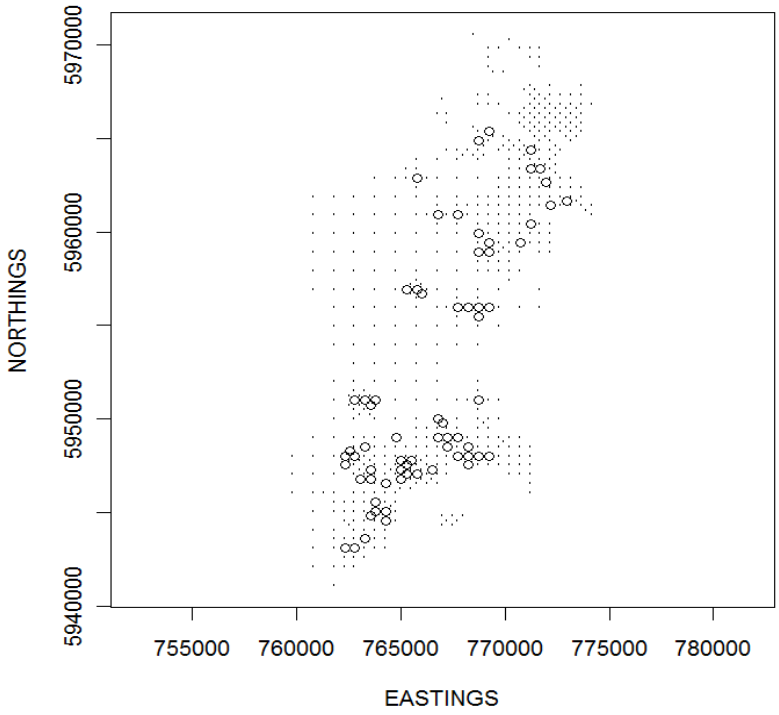

Figure 2. Sample design and koala activity. Each site is marked; large circles denote sites (68) where Koala faecal pellets were found.

Statistical Modeling

In this study, valid inferences about the covariates effects are of greater interest than the estimation of the spatial variation. The primary role of modeling the latent spatial process is to account for the spatial dependence in the observations and so guard against inferring spuriously significant covariate effects. (Ignoring the spatial correlation in the residual data usually results in standard errors being too small.) In this case however, spatial effects are of some interest in their own right, because the model provides a basis for constructing a map of smoothed predictions thus identifying areas of unusually high or low occupancy of Koalas. This might point to other, as yet unidentified, explanatory factors.

Our choice of modeling framework is discussed in detail by Diggle and Ribeiro (2007). Our approach was to try to find a plausible statistical model, compatible with the Koala data, both in terms of an appropriate model for the mean component (trend and covariate effects) and for the spatial covariance structure.

We considered modeling both site abundance (# of trees with scats) using a log-linear Poisson model and site occupancy (ie presence/absence binary data) using a logistic Bernoulli model. As the data are zero-inflated (many more zeros than would be expected for Poisson data), we believe that inferences based on a binary model for occupancy are more secure than inferences based on a Poisson model for abundance. For comparison we present results for pellet counts.

A plausible model for site level binary data (observed presence/absence of Koala) is an extension of the linear logistic regression model which incorporates random effects to account for residual spatial effects. Let p_i denote the probability that koalas occupy the ith site. Then let

$$\log\{p_i/(1-p_i)\} = \alpha + \beta' z_i + S(x_i),$$

where $S(x_i)$ is a zero-mean stationary Gaussian process with isotropic covariance function

$$\gamma(u) = \sigma^2[\exp\{-(|u|/\phi)\} + v^2 I(u=0)], \quad \sigma^2 > 0, \quad \phi > 0, \quad v^2 > 0,$$

where I is the indicator function.

Model fitting

The sample semi-variogram of Koala pellet data (Figure 3) showed evidence of spatial dependence, consistent with the findings of Penman and Kavanagh, This is not surprising given the strong clustering evident in the observed data (Figures 1 and 2). When the data are expressed as binary data, the evidence for spatial dependence is not as strong (Figure 4). Figures 3 and 4 need to be interpreted with caution as the high proportion of zeros makes them relatively uninformative.

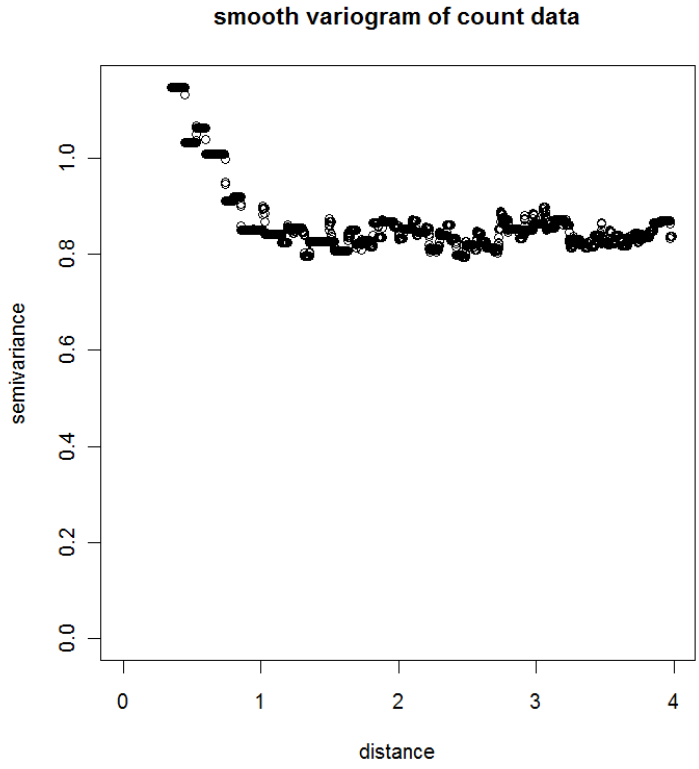

Figure 3. Variogram of koala pellet data. (The eastings and northings have been scaled by dividing by 1000.)

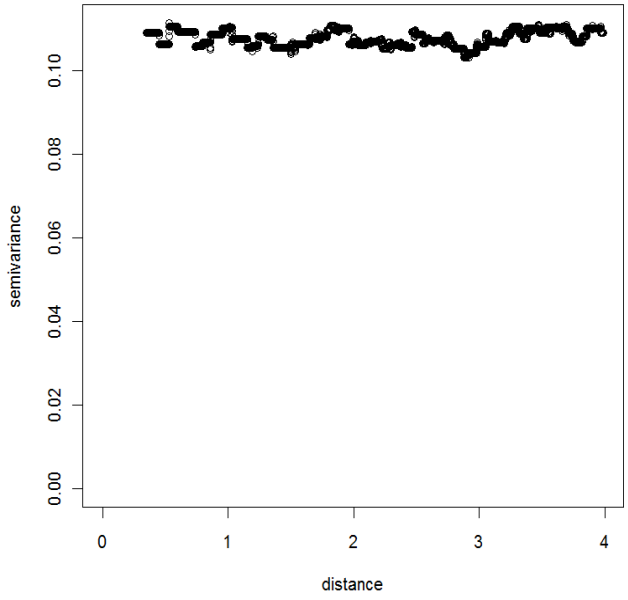

Figure 4. Variogram of binary data, when koala activity is expressed as presence or absence. (The eastings and northings have been scaled by dividing by 1000.)

We found some evidence that spatial dependence was stronger in the north- south direction than in the east -west direction ie. spatial covariation is not isotropic. As Koala occurrence is relatively rare in our data, it is difficult to incorporate this feature into the model and we have chosen not to pursue it in detail.

We fitted the model using an MCMC algorithm within a Bayesian framework. The algorithm in R is written by Ribeiro and described in Diggle and Ribeiro (2007). To use the algorithm, we need to specify prior distributions for all the unknown parameters in the model so we imposed proper but vague priors on the parameters. For the parameters in the mean, α and β, we imposed independent normal priors with mean zero and variance 100. For σ^2, we imposed the inverse chi distribution with scale parameter 50 and degrees of freedom 5. These distributions have reasonably flat densities which should not have a strong impact on the analysis. We also tried to fit the model with prior distributions for the remaining parameters, ϕ and ν^2 (called phi and tausq.rel in the software), but we found that there was not enough information in the data to identify these parameters. Similar difficulties are reported for the examples presented in Diggle and Ribeiro (2007). As a result, we adopted the pragmatic strategy of fixing these parameters at specific values and then exploring the sensitivity of the conclusions to changing the fixed values. We explored the effect of setting ϕ = 1,2,3,5,10 and ν^2 = 0, 0.5, 1, 2. In the final model, we chose ϕ = 1 and ν^2 = 1. These choices correspond to a short range spatial correlation (so sites which are spatially well-separated are treated as uncorrelated) and a nugget effect at zero.

Spatial dependence

The scales of the estimated semi-variograms of the residuals have decreased slightly from the corresponding plots for the raw data but shapes are very similar, showing that the spatial models shrink the variation but do not have much impact on the correlation structure. This makes sense in the binary analysis as it confirms the initial finding that the correlation is weak.
Predictions from our spatial models are shown in Figures 5 and 6.

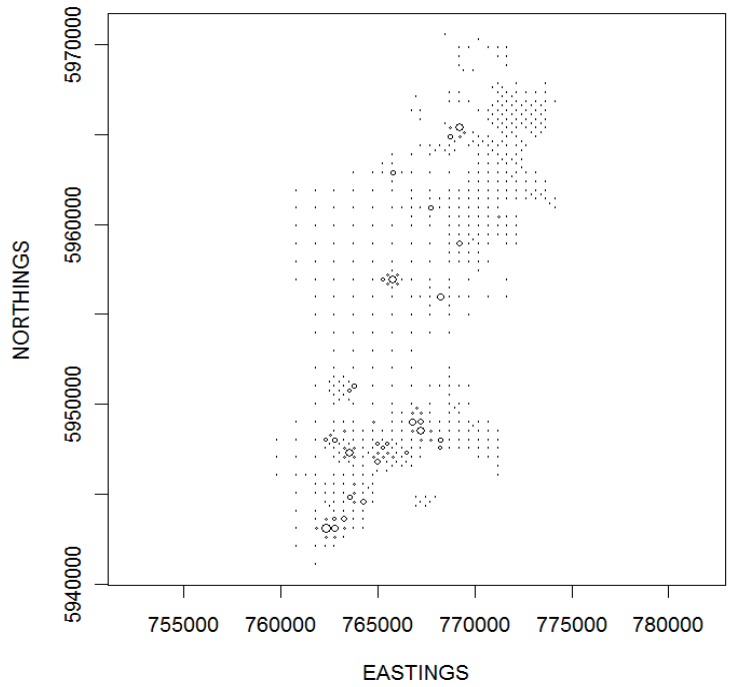

Figure 5. Plot of fitted values from spatial model of koala faecal pellet data. Each site is represented by a circle proportional in size to the number of trees where koala faecal pellets were found. The largest circles correspond to a predicted value of 2 and the smallest circles correspond to sites with no koala activity.

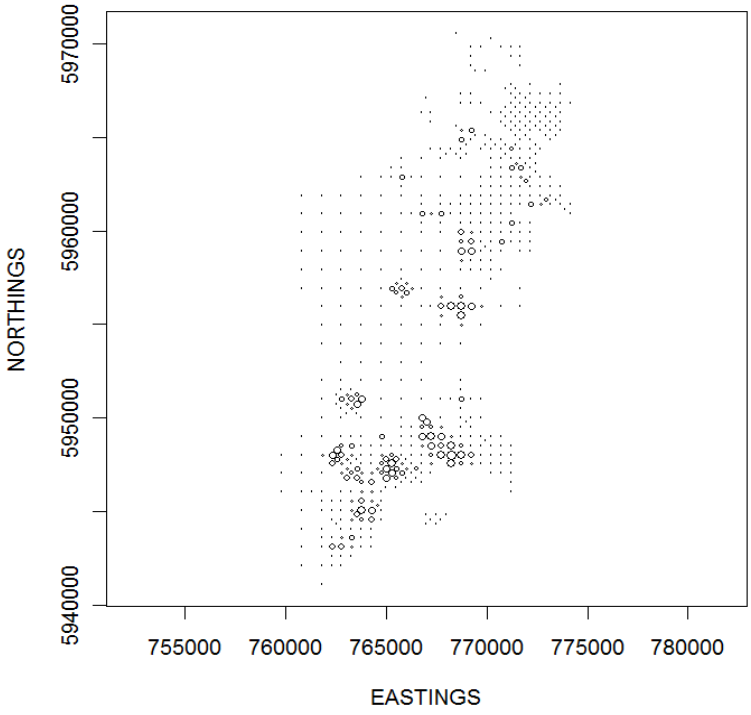

Figure 6. Plot of fitted values from spatial model of binary data. Each site is represented by a circle proportional in size to the predicted probability of koala activity. The largest circles correspond to a predicted value of 0.61 and the smallest circles correspond to a predicted value of 0.27.

Covariate effects

Using the models for the spatial effects described above, we examined many candidate covariate effects in an attempt to identify habitat covariates as predictors of Koala occupancy. We constructed covariates for the key tree species with particular emphasis on species found to be important by other analysts. These included number of stems, aggregate basal area and median values for dbh etc. Other derived variable have been suggested by other analysts but we felt that, given our results, there was little value in exploring these.

Sensitivity analysis

Just to reinforce confidence in our results we undertook a sensitivity analysis which involved sub-sampling our data by applying a spatial filter. This filter involved selecting data points that were at least 1km apart. For each subsample we fitted linear logistic models in an attempt to identify 'significant' explanatory habitat effects; apart from one or two weak effects no strong effects were found. Our general conclusion was that, after accounting for spatial effects, the models were considered to be of little practical use for prediction; this was consistent with our earlier findings**.**

Summary

Even though the Koala data are of high quality, two key features are that Koalas are highly clustered spatially and there is a relatively low incidence of occupancy. These two features have important consequences for statistical modeling. The high level of clustering means that for the purposes of identifying important covariates the effective sample size is much less than 589 and the true incidence of koala occupancy is small. A consequence of this is that if the strong clustering structure is not appropriately modeled, then spurious significant results will be found.

Further it is known that when occupancy probabilities of a species are very low, the ability to be able to establish the statistical significance of an effect tends to be low. That is, there is low statistical power. Figure 7 (Nicholls and Cunningham 1995) shows the effect of increasing rarity (i.e. a low probability of detection) on the standard error of log (relative risk) while maintaining a constant sample size. Here relative risk is a measure of the extent to which a site having a particular attribute is more (or less) likely to have a species present than a site without the attribute. The effect is that for data on rare species, a change in a factor of interest (e.g., a covariate) may result in a large relative change in odds of a given species being present, but this may not translate into a statistically significant effect. This is particularly evident as occurrence falls below 5%. Nicholls and Cunningham (1995) provide an example in the context of predicting the distribution of the Koala and, given occupancy by these species is very low, the ability to establish the statistical significance of an effect tends to be low.

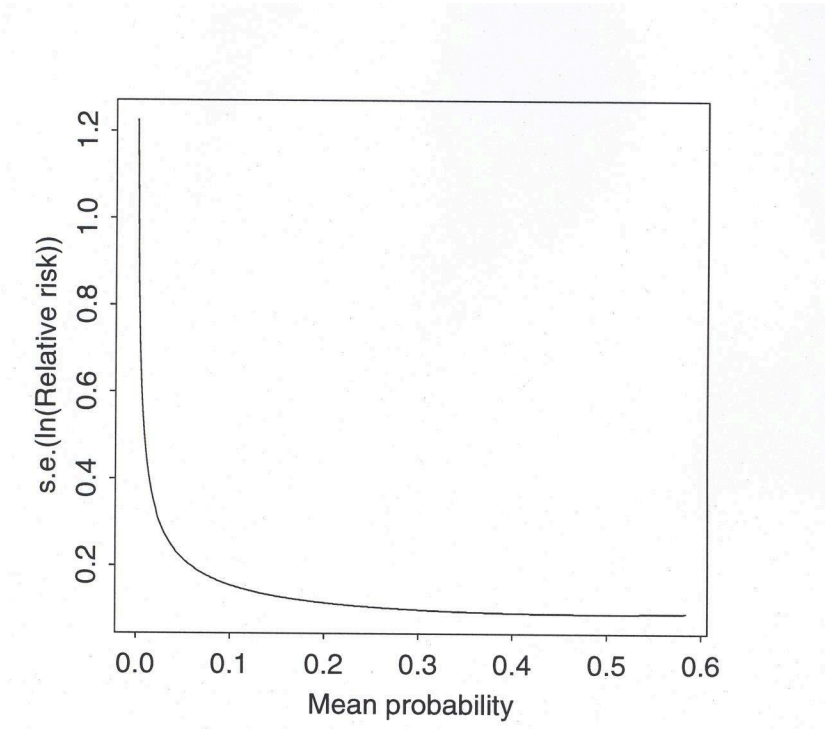

Figure 7. Relationship between the standard error of ln(relative risk) and rarity.

References

Diggle, Peter J,and Ribeiro Jr, Paulo J (2007) Model-based Geostatistics Springer Series in Statistics

Nicholls, A. O., and R. B. Cunningham. 1995. Strategies for evaluating wildlife management. Koala Conservation in the south-east forests. *In* S. Cork, S. Feary, and C. Mackowski, editors. Proceedings of an expert workshop. National Parks and Wildlife Service, Sydney, NSW, Australia.

APPENDIX

Comments on previous analyses

a) Keith McDougall, Chris Allen and Michael Saxon (February, 2010) used bivariate spatial analysis (Rowlingson and Diggle 1993) based on K functions to compare the distribution of two point patterns (the point patterns being the presence/absence of Koalas and selected tree data at a range of scales).
In terms of the departure of the bivariate K function from the upper probability limit (Fig. 5), the authors found that the following attributes were most strongely correlated with the presence of Koala pellets in this study area:

1. *Eucalyptus bosistoana*; any number or DBH,
2. *Eucalyptus muelleriana*; > 4 trees / plot,
3. *Eucalyptus longifolia*; > 276 mm DBH (> 30 percentile),
4. *Eucalyptus globoidea*; > 392 mm DBH (> 70 percentile),
5. *Eucalyptus tricarpa*; > 351 mm DBH (> 40 percentile.

Comment

The K-function methodology was developed for analysis of completely observed spatial point pattern data. Observing the processes at a selected set of points is different from completely observing the processes over the whole. Data here were collected using a multi-level spatial grid framework - 350m, 500m and 1000 m apart - so the observation of the true spatial point pattern is far from complete. Imposing a further level of sampling to deal with the identified spatial dependence problems adds to the incompleteness of the process. We do not know how the incompleteness of the underlying process affects inferences pertaining to associations between Koala and key habitat variables.

b) 'Comments on the analysis of koala data in south-eastern NSW undertaken by Keith McDougall, Chris Allen and Michael Saxon' by Trent Penman and Rod Kavanagh, 18 February 2010

In a model of the probability of occurrence (hereafter "p(occ)") the authors first consider the spatial correlation in the data. Haining (2003) recommend the use of a spatially lagged response variable (hereafter SLRV). They used a SLRV to account for the abundance of Koalas in the neighbouring plots while weighting them according to the inverse distance between plots. All 589 plots were included in the calculation of the SLRV and the subsequent analysis of relationships.

Equation 1: Calculation of SLRV for point i, where n is the total number of sites other than i, w_{ij} is the weight given of site j from site i. In this case, the weight is the inverse of the distance between site i and j and x_j is the response at site j, i.e. the number of trees with confirmed Koala pellets.

$$SLRV_i = \left(\sum_{j=1}^{n} w_{ij} \times x_j \right) / \sum_{j=1}^{n} w_{ij}$$

They then used a generalized additive modeling (GAM) framework to assess the probability of occurrence of a Koala at a site. GAM's were chosen as they allow for non-linear relationships which might be expected for the relationships examined. The response variable was a binary response being either the presence or absence of a Koala at a plot. Models did not account for the abundance or activity of a koala at the plot. Independent variables in the model were SLRV, median and sum of the DBH of trees at the plot and the counts of tree

species which occurred on more than 5% of plots in the study area. Tree species were only considered individually in a model in an attempt to test the findings of the bivariate analysis

For all models, the authors found a significant positive linear relationship with the presence of Koala's and the SLRV ($p<0.0001$**). SLRV appeared to be having the greatest influence on the p(occ) in all models.** Varying relationships occurred with DBH. Mean plot DBH was not significantly related to occurrence of koalas, whereas the median plot DBH was significantly positively related to probability of occurrence in a linear manner. It is important to note that these relationships are not always consistent with the recommendations of the authors, namely *Eucalyptus bosistoana* which showed no relationship in our models and *E. muelleriana* which showed an initial increase in p(occ) but after 4 trees per plot the p(occ) declined. The deviance explained by the models was relatively low ranging from 14 to 17%.

Comment

These authors use a traditional auto-logistic linear regression approach to model the presence /absence of Koala and to identify potential habitat variables as predictors of the estimated probability of occupancy of a site by Koala. The method attempts to account for spatial dependence by including a constructed neighborhood variable (SLRV) in the mean function of the model. It is unclear whether the SLRV variable adequately accounts for the spatial dependence in these data. If this variable fails to model the dependence structure then standard errors associated with estimated parameters of habitat variables will be too small and so there are potential errors of falsely claiming statistical significance of covariates.

General Comments

Clearly there can be interplay between the mean model and the covariance structure for data. Analyst have to make a judgment as to which terms to include in the mean model (as trend) as well as whether trends should be removed before estimating spatial correlation. There is no clear answer to these questions and the choices have a lot to do with the primary objective of the analysis.

These analyses show some weak evidence of relationships between the probability of occupancy and several habitat variables. However the strength of relationship is such that they are not that useful for prediction.

By modeling spatial dependence, prediction can be based on random spatial effects but these predictions are of little practical use since the results cannot be extrapolated to other regions. Such models yield a parametrically smoothed map showing the spatial distribution of the target response in the study region, but do not produce similar maps for new regions.