The 13th North American Caribou Workshop
Winnipeg, Manitoba, Canada
25–28 October, 2010

A guide to developing resource selection functions from telemetry data using generalized estimating equations and generalized linear mixed models

Nicola Koper[1] **& Micheline Manseau**[1, 2]

[1] Natural Resources Institute, University of Manitoba, 70 Dysart Rd., Winnipeg, MB, Canada, R3T 2N2 (koper@cc.umanitoba.ca).
[2] Western and Northern Service Centre, Parks Canada, 145 McDermot Ave, Winnipeg, MB, Canada, R3B 0R9.

Abstract: Resource selection functions (RSF) are often developed using satellite (ARGOS) or Global Positioning System (GPS) telemetry datasets, which provide a large amount of highly correlated data. We discuss and compare the use of generalized linear mixed-effects models (GLMM) and generalized estimating equations (GEE) for using this type of data to develop RSFs. GLMMs directly model differences among caribou, while GEEs depend on an adjustment of the standard error to compensate for correlation of data points within individuals. Empirical standard errors, rather than model-based standard errors, must be used with either GLMMs or GEEs when developing RSFs. There are several important differences between these approaches; in particular, GLMMs are best for producing parameter estimates that predict how management might influence individuals, while GEEs are best for predicting how management might influence populations. As the interpretation, value, and statistical significance of both types of parameter estimates differ, it is important that users select the appropriate analytical method. We also outline the use of *k*-fold cross validation to assess fit of these models. Both GLMMs and GEEs hold promise for developing RSFs as long as they are used appropriately.

Key words: autocorrelation; conditional models; empirical standard errors; GEE; generalized estimating equations; generalized linear mixed-effects models; GLMM; *k*-fold cross validation; marginal models.

Rangifer, Special Issue No. 20: 195–203

Introduction and rationale

This document provides a practical guide for developing resource selection functions from telemetry data, using generalized estimating equations and generalized linear mixed models, and outlines how to validate these models using *k*-fold cross validation. For more detailed explanations and to better understand the theory and mathematics behind these methods, readers should refer to Koper & Manseau (2009), in which we cover most of the topics within the present manuscript in more detail; Gillies *et al.* (2006) and Bolker *et al*. (2009) regarding GLMMs; and Boyce *et al*. (2002) regarding *k*-fold cross validation, as well as numerous excellent sources and textbooks referred to in those works. Fieberg *et al.* (2010) provides a useful and detailed comparison among various approaches to analyzing habitat selection, including GEEs and GLMMs. We also note that this paper discusses the development of resource selection functions (RSF), which estimate the *relative* probability of use of different habitat types (suitable vegetation), rather than resource selection probability functions (RSPF), which estimate the *actual* probability of a habitat being used; for more information on the additional assumptions and issues associated with RSPFs, see Lele & Keim (2006).

To facilitate the use of this paper as a guide, we outline a number of components important to RSF development below, and in most cases divide each section into *What*, *Why* and *How* subsections. Statistical codes are provided to conduct GEEs and GLMMs in SAS and to conduct GEEs in R. GLMM code is not provided for R because at the moment, there are no GLMM libraries that allow the user to request empirical standard errors.

Resource selection functions

What are resource selection functions?

Resource selection functions are models used to compare the amount of used habitat with the amount of available habitat (Manly *et al.*, 2002). If a habitat type, such as jack pine stand, is used by animals more than expected relative to the proportion of that habitat across the landscape, the habitat is assumed to be selected; if it is used less often than expected relative to the proportion of that habitat across the landscape, the habitat type is assumed to be avoided. For example, if 10% of the landscape is made up of jack pine stands, but the animals spend 25% in jack pine stands, then the assumption is that jack pine stands are selected.

Why are resource selection functions important?

Resource selection functions are used to quantify the relative importance of different vegetation or habitat types, or different components of the landscape, given the availability of those habitat types on the landscape. This helps define the realized niche of caribou or the species of interest.

How are resource selection functions developed?

When using telemetry data, there are different ways to estimate the resource use including a determination of the resource type associated with each telemetry point, the amount of different resources within a buffer around each telemetry point, the distance to different resource types from each point or the spatial characteristics of resource patches associated with each point. To quantify this, animal locations are imported from the satellite (ARGOS) or GPS telemetry data into a geographic information system (GIS) that also includes a land cover layer and then, the attributes are derived for each landscape parameters of interest and for each location point. To quantify availability, points are randomly generated within the individual's or herd's home range or within a certain distance of the telemetry points, and similarly the attributes are derived for the landscape parameters of interest. The total number of randomly generated points varies with each study; usually, the same number of telemetry points and random points are used, or the number of random points is a multiple (usually between 2 and 10) of the number of telemetry points.

The telemetry points from animal locations ("used" locations) are then compared with these randomly located points ("available" locations), to determine whether there is more or less use of each habitat type than expected given how much of each habitat type is available (Manly *et al.*, 2002). Selection can only be evaluated if availability can also be quantified. For example, if 60% of the caribou locations are in treed muskeg, but that habitat type makes up 75% of the landscape (or 75% of the random points), the results would indicate an avoidance of treed muskeg, even though more than half of the locations are in treed muskeg, because the proportion of telemetry points in treed muskeg is less than the proportion of random points in that habitat on the landscape.

Autocorrelation in telemetry data

What is autocorrelation in telemetry data?

Locations from satellite or GPS collars have provided us with a large amount of data which can be used to infer how animals use the landscape. In particular, once animals are collared, thousands or tens of thousands of locations for that animal are recorded, and by overlaying those locations on a land cover map using a GIS, the way that animal uses its landscape can be determined to a high degree of precision and accuracy.

These data points are, however, not independent of one another. There are two important sources of correlation in telemetry data. The first is that there are many data points from just a few caribou. Data points from a single individual are not independent of one another, and as such do not each provide us with a unique piece of information. For a comprehensive review of this issue, see Gillies *et al.* (2006).

The second source of correlation arises from the fact that animal locations are recorded sequentially. Locations can be recorded as often as every half hour, or less than once a day, as desired. Determining the optimal length of time between locations can be an important question, as more frequent locations result in shorter battery life. If locations are recorded too frequently, each location provides little new information about resource use; presumably, the current location of the animal is highly influenced by the previous location of the animal, or even a number of previous locations. Such data are serially autocorrelated. On the other hand, if locations are too infrequent, there may be insufficient data to evaluate habitat use relative to habitat availability, particularly when estimating the use of uncommon habitat types (see discussion in Fortin *et al.*, 2005).

While locations taken few minutes apart are probably highly correlated, and locations taken 5 days apart are much less correlated, the interval at which points become uncorrelated is not known. Indeed, Cushman *et al.* (2005) argued that locations may be correlated at intervals of a month apart. As such, there is no interval between data locations at which

telemetry locations are known to be independent of one another.

Why care about serial autocorrelation in telemetry data?
Autocorrelation between data points might be of interest to the researcher (Boyce *et al*., 2010). For example, this might help the researcher understand how likely an animal is to stay in a particular habitat type if it is already there. However, in some cases this correlation among data points is not of interest and becomes a statistical nuisance. While generally serial autocorrelation has relatively little effect on the parameter estimates that are derived from statistical models, they can affect any associated statistical comparisons, or any analysis that uses standard errors or confidence intervals. For example, if this correlation is ignored, it might be possible to estimate how much more or less animals use different habitat types in relation to what is available to them, but it would not be possible to determine whether this is a *statistically significant* habitat selection or avoidance.

This is because a key component of calculating statistical significance is knowing how much information is available to go towards comparisons of resource use. More information provides the user with more confidence that estimates of habitat selection or avoidance are trustworthy. However, when there is a lot of information from just a few animals, it can be hard to quantify how much information there really is. If the amount of information available is overestimated, the likelihood of making Type I errors (assuming that there is a statistically significant effect of a variable when in fact there is no effect; Clifford *et al*., 1989) is increased. To avoid this, we must take correlation among data points within animals into account.

How can serial autocorrelation in telemetry data be controlled for?
Gillies *et al*. (2006) recommended that random variables (also referred to as random effects in the literature) be included in RSFs to account for the fact that data points come from different animals, and that data points from individual animals are not independent from each other. In these models, a variable that represents the individual animal becomes the random variable (see section on *how GLMM work* for more information on random variables). One example of models that include random effects is generalized linear mixed models (GLMM). The "generalized" term refers to the fact that the error term associated with response variables need not follow a normal distribution; as resource selection functions compare used habitats (represented by "1" in the response variable) with available habitats (represented by "0" in the response variable), the response variables follow a binary (binomial) rather than a normal distribution. The "mixed" term in GLMM refers to the fact that both random effects and fixed variables (independent variables such as habitat type) are included in the model.

We believe that the recommendations by Gillies *et al.* (2006) initiated important progress in the trend towards using advanced statistical techniques for developing resource selection functions. However, they made an error by implicitly assuming that individual data points within animals were independent from one another. This is not correct, and GLMM are not robust to deviations from this assumption (Overall & Tonidandel, 2004, and see empirical analysis in Koper & Manseau, 2009); this means that statistical inferences made from models that ignore correlations among data points are likely to be incorrect, leading to increased rates of Type I errors. However, the correlation among telemetry points can be compensated by using empirical, rather than model-based standard errors (e.g., Hardin & Hilbe, 2003).

Empirical standard errors

What are empirical standard errors?
Empirical standard errors are also sometimes called robust standard errors, as they are robust to the lack of independence among data points (i.e., lack of independence among data points does not lead to incorrect empirical standard errors), or Huber-White sandwich standard errors (as applied by Gillies *et al.* 2006). The empirical standard error is generally larger than the model-based standard error, and the closer the modeled correlation structure to the true correlation structure, the closer together the model-based and empirical standard errors will be (Bishop *et al*., 2000). As such, the correlation should be modeled to reduce the size of standard errors and therefore, increase the power of the analyses. However, this is not possible when telemetry data are compared with random data points. The empirical standard errors are therefore required to correct for the correlation among data points.

Why should empirical standard errors be used for RSFs developed from telemetry data?
It is critical to use empirical standard errors if these are appropriate and necessary. There is often a very large difference between empirical and model-based standard errors, and this directly leads to differences in statistical inference. We found that model-based standard errors could be 1/10 the size of empiri-

cal standard errors (Koper & Manseau, 2009); not surprisingly, this has a dramatic effect on the apparent significance of independent variables. Empirical standard errors must be used to evaluate statistical significance of habitat selection behaviours when resource selection functions are developed using telemetry data.

How are empirical standard errors included in RSFs?
The variance function differs between empirical and model-based standard errors. This is accounted for by the selected statistical computer program when empirical standard errors are selected by the user.

In this paper, we cover two statistical approaches that can both be used with empirical standard errors: generalized linear mixed models (GLMM), and generalized linear models with generalized estimating equations (GEE). There are important practical and conceptual differences between these approaches that must be considered in determining which approach is appropriate. Below, we introduce GLMM and GEE, and follow with a comparison between the two. We then address validation of each type of model using *k*-fold cross validation (Boyce *et al*., 2002).

Generalized linear mixed-effects models (GLMM)

What are GLMM?
GLMM are sometimes also called generalized linear mixed models (GLME) or hierarchical models, and are referred to as longitudinal, clustered, latent-variable, or multilevel models. They are parametric, and are estimated using maximum likelihood theory or associated methods (see Quinn & Keough, 2002 for a clear explanation of maximum likelihood estimation). Mixed models include fixed and random independent variables. Fixed variables are differentiated from random variables in two ways: all levels of interest for the factor are included in the design, and inference is restricted to these levels. Random variables, in contrast, include randomly selected levels, and allow one to generalize inference over all possible levels of the random variable.

Why are GLMM useful?
Usually a random sample of caribou is monitored to allow the manager to infer habitat selection of all (or at least, other) caribou in a defined population. Differences in habitat selection among individual caribou should therefore be modeled using a random variable, because not all levels of interest are included in the design (i.e., all caribou in the population of interest), and the inference should be relevant to all possible levels of the random variable (i.e., all caribou in the population of interest).

Another benefit of mixed models is that they can be used to analyze hierarchical study designs. This means that one can use a single model to evaluate effects of local-scale variables nested within broad-scale variables. For example, there might be interest in evaluating effects of vegetation structure (e.g., canopy cover), which will be different at every location recorded, nested within caribou-scale variables (e.g., animal age), which will be the same for every data point within an animal. If the hierarchical nature of this design is ignored, there may be two unintended consequences: (1) if the caribou is considered the unit of replication, all local scale variables would have to be collapsed into a single value per caribou, thus losing an enormous amount of data and, therefore, statistical power; and (2) if the local data points are considered as the unit of replication, the degrees of freedom would be artificially increased at the caribou scale, introducing pseudoreplication into the design, and increasing the likelihood of making Type I errors. Mixed models allow us to analyze variables at both of these scales by using the random effect to indicate that local-scale variables are not completely independent of one another, because they are clustered within the broad-scale variables.

How do GLMM work?
Differences among caribou, represented by the random variable, are modeled by allowing the *intercept* of the relationship between each independent fixed variable and the dependent variable to be different for each caribou. It is, in fact, possible to assume that both the intercept and the slopes of the relationships vary among caribou (e.g., see Gillies *et al*., 2006), but this is more complex than allowing only the intercept to vary among caribou (and is usually unnecessary and results in an overparameterized model; L. Lix, University of Saskatchewan, pers. comm.) and will not be discussed here. Allowing the intercept to vary among caribou recognizes that animals differ from one another, but means that habitat or landscape structures are assumed to have a similar effect on different animals.

While mixed models are an important tool for dealing with clustered sampling designs, they are not a panacea. The addition of a random effect tends to increase standard errors of all the fixed variables in the model (Hox, 2002, Quinn & Keough, 2002). The consequence of including a random effect is a reduction of analytical power, but the inference is then correct.

Generalized estimating equations (GEE)

What are GEEs?

Generalized estimating equations are a semi-parametric alternative to GLMMs. They are semi-parametric because the parameter estimates are estimated parametrically and the variances are estimated non-parametrically.

Usually, part of the process of defining the variance structure is to define the correlation structure of the data points within individual caribou; for example, data collected sequentially over time could be modeled differently from data that were clustered spatially, say across a number of different isolated islands. Correlation structures can include, among others, an independent correlation structure (in SAS, corr=ind), which assumes no correlation among data points; a compound symmetric or exchangeable correlation structure (in SAS, corr=CS), which assumes that data from a single animal is correlated within that animal, but all data points within animals are equally correlated; and an autoregressive correlation structure (in SAS, corr=AR(1)), which assumes that data points within animals that are closer together in time are more correlated than data points that are farther away. We remind the reader that the latter structure is a reasonable assumption for the used data points, but not for the random points (Koper & Manseau, 2009).

Why are GEEs useful?

When sample sizes (number of caribou) are sufficiently high, GEEs with empirical standard errors have the enticing property of producing both parameter estimates and standard errors that are trustworthy even when the correct correlation structure cannot be known (Fitzmaurice *et al*., 2004). This is important when developing RSFs, because the correlation structure between telemetry points and random points cannot be modeled.

How do GEEs work?

GEEs deal with the correlation caused by collecting numerous samples from each individual (e.g., numerous locations from one caribou) by adjusting the standard error to compensate for the lack of independence among samples. This involves using empirical standard errors, rather than model-based standard errors, as discussed above (Hardin & Hilbe, 2003).

Because tests are more powerful if the covariance structure can be modeled, users should still compare model fit between models with different covariance structures, and use the model that fits the data the best. Covariance structures can be compared by taking the ratio of the empirical standard error to the model-based standard error (SE_E/SE_M), and the model with the ratio that is closest to 1 is the model that fits the data the best (Bishop *et al*., 2000). Although the non-parametric alternative to AIC, the quasi-likelihood under the independence model information criterion, QIC (Pan, 2001), is also theoretically capable of this comparison, our research has demonstrated that this criterion is biased (Barnett *et al*., 2010). Therefore, we recommend that QIC should not be used for comparisons among models until it is redeveloped, a process that is in progress (J. Hilbe, 2010, pers. comm.). Because the correlation structure among the used data points differs from the correlation structure among random points, this correlation structure cannot be modeled correctly, and there will be some dependence on the fact that empirical standard errors are robust to misspecification of this structure.

Choosing between GLMM and GEE for developing RSFs

What are the main differences between GLMM and GEE?

There are a number of practical and conceptual differences between GLMM and GEE, and these must be considered before determining which method is appropriate for analyzing any data set. These differences are summarized in Table 1, and are discussed in more detail below.

Parametric and semi-parametric modelling

Because GLMMs are parametric, while GEEs are semi-parametric, the analytical process for generating GLMM is more complex, takes longer, and is more likely to fail to converge (Agresti, 2002). Nonetheless, in our experience GLMM can generally be used successfully for developing resource selection functions using telemetry data.

Hierarchical versus non-hierarchical models

GLMMs model differences among animals directly, and this allows for a hierarchical data analysis that directly models effects of independent variables at different spatial or temporal scales. This hierarchical analysis is not possible with GEEs. GEEs do not directly model differences among animals, but instead account for the lack of independence among samples within animals by adjusting the standard error via an altered variance estimate.

Marginal versus conditional parameter estimates

When response variables are binary (or otherwise non-normal), there is an important difference in the meaning of the parameter estimates gener-

Table 1. Comparison between the use of generalized estimating equations (GEEs) and generalized linear mixed models (GLMMs) for developing resource selection functions using telemetry data.

	GEE	GLMM
Analysis	Semi-parametric	Parametric
Method of dealing with correlation	Adjusts standard error to account for correlation of data points within groups (animals)	Models differences among animals directly, usually by allowing intercept to vary among animals
Complexity	Simpler	More complex
Convergence	More likely	Slightly less likely
Robustness	Parameter estimates and standard errors robust to misspecification of the correlation structure when using empirical standard errors	Standard errors robust to misspecification of the correlation structure when using empirical standard errors
Interpretation	Marginal	Conditional
Information theory	Use with QIC is not recommended	Can use with AIC
Treatment of hierarchical data	All nested levels treated equally – better if clustering is a nuisance, not the focus of the study	Explicitly models hierarchical or nested sampling design
Sensitivity to sample sizes	More robust to differences in sample sizes within groups	Sample sizes within groups should be approximately equal

ated between GEEs and GLMMs (Fitzmaurice *et al*., 2004: 364), and this can result in large differences in parameter estimates and standard errors between the two approaches (Fitzmaurice *et al*., 2004; Koper & Manseau, 2009). This is primarily because GLMMs produce conditional (subject-specific) parameter estimates (see Agresti, 2002 for reasons why marginal estimates derived from GLMMs should be avoided), while GEEs produce marginal (population-specific) parameter estimates. Conditional parameter estimates model how a typical *individual* might respond to independent variables. Marginal parameter estimates evaluate effects of independent variables on the *population*.

Two examples may help clarify the difference in interpreting marginal and conditional parameter estimates. First, we will consider an example derived from epidemiology. A marginal question might be, "what is the effect of this drug on cancer rates across a population?" This type of study would be designed to compare how many people got cancer in populations that received the drug, and how many people got cancer in populations that did not receive the drug. This is a population-specific approach because it addresses how the independent variable, use of a drug, affects a whole population.

An equivalent conditional question would be, "what is the likelihood of a typical patient recovering if we give them this drug?" This type of study would be designed to compare whether people who received the drug were more likely to get cancer than people who did not receive the drug. This is a subject-specific approach because it addresses how the independent variable, use of a drug, affects the likelihood of a typical individual getting cancer.

The difference between these two approaches may seem like semantics until one reflects on the position of a patient. Most individuals will care much more about what the effect of the drug might have on their own probability of getting cancer, compared with the effect of the drug on cancer rates across a population. This results in a very real difference in the interpretation of marginal and conditional population estimates.

The difference is also important from a wildlife management perspective. An example of a marginal question might be, "what is the difference in habitat use of caribou between landscapes with high or low jack pine cover?" This type of study might be designed to compare whether populations of caribou that lived in landscapes with high jack pine cover demonstrated different habitat selection from popula-

tions of caribou that lived in landscapes with low jack pine cover. This is a marginal or population-specific approach because it addresses how the independent variable, jack pine cover, affects habitat selection across a population.

An equivalent conditional question would be, "how would a typical caribou change its habitat use if its environment changed from having high jack pine cover to relatively little jack pine cover?" This type of study might be designed to compare whether individuals changed their habitat selection if their landscape changed from one of high jack pine cover to low jack pine cover through forestry activities. This is a conditional or subject-specific approach because it addresses how the independent variable, jack pine cover, affects habitat selection of a typical individual.

Again, these questions are different and address different management issues. The marginal approach might be more appropriate for trying to understand effects of habitat on the population of interest; for example, for the development of population recovery plans. In such cases, the interest is on how an entire population will respond to management. The conditional approach might be more appropriate if there is interest in how future changes in an environment might affect a typical caribou; for example, if evaluating the potential impact of future forestry activities on individuals of a population. Regardless, we emphasize that this decision is important because it will change the interpretation of the parameter estimates, will change the actual parameter estimates, and will change their apparent significance. For further discussion about differences between marginal and conditional parameter estimates, and a useful graphical explanation, see Fitzmaurice *et al*. (2004).

How are GEE and GLMM run on statistics programs?
An example of code that can be used for conducting a GLMM using Proc GLIMMIX in SAS is given in Appendix I. An example of code that can be used for conducting a GEE using Proc GENMOD in SAS is given in Appendix II. An example of code that can be used for conducting a GEE using the library geepack in R is given in Appendix III. Koper & Manseau (2009) provides a case study using GLMM and GEE on woodland caribou GPS relocation data.

Model validation

What is model validation?
Model validation allows us to determine how well a dataset, which is collected from a sample of the population of interest, predicts habitat selection by the population from which the sample is drawn. A common approach for validating resource selection functions is to use *k*-fold cross validation (Boyce *et al*., 2002). An important benefit of this approach is that it may be used with any resource selection function, regardless of the statistical approach used to develop that function. Therefore, it can be applied to models developed using both GLMMs and GEEs.

Why is model validation important?
It allows us to determine the trustworthiness of models.

How are models validated using k*-fold cross validation?*
k-fold cross validation starts by separating the data set into bins (a number of different groups, say k = 10 for this example). A model is developed using all of the data except data from a single bin. Then the fit of the data from the withheld bin is evaluated to the model developed from the other data. This comparison produces a correlation coefficient, r.

This process is repeated, withholding data from one bin at a time, until each bin has been withheld once. This produces k correlation coefficients which are then averaged. The idea behind this approach is that it gives us the opportunity to evaluate the fit of each model using data that are independent of the data used to develop the model.

How are bins selected for the k*-fold cross validation?*
There are several ways in which data can be separated into bins for these comparisons, each of which produces different results. With marginal population estimates, the interest is in predicting habitat selection of other animals in the population, using data from just a few individuals. Therefore, to evaluate predictive capacity of marginal models, models should be developed by withholding all of the data from one or two individuals at a time, and then evaluating how the models developed using the remaining animals to predict habitat selection by those animals (Koper & Manseau, 2009).

At this point, however, we diverge slightly from the recommendations we provided in Koper & Manseau (2009). Previously, we argued that for conditional models, we should withhold a portion of the data from each animal, develop models using all of the remaining data, and then test model fit using the withheld data (Koper & Manseau, 2009). This is still appropriate if the interest is in predicting habitat selection by the specific individuals surveyed, for example, if managers are interested in predicting effects of future management on these animals. However, conditional models can also be used to predict habitat selection of a *typical* animal; in that case, the

interest is still in generalizing results to other animals in the population of interest. We note that this is not the same as predicting the effects of habitat on a population overall, but instead still focuses on the habitat selection of individuals; however, that might include the habitat selection of individuals from outside of the study sample. If that is the purpose of the conditional model, then we recommend that the user should follow the process recommended for marginal models; bins should be developed by withholding all the data from one or two animals, and then evaluating how well the models predict the habitat selection of those animals.

Like any statistical model, *k*-fold cross validation has some drawbacks. It is often misused, most commonly by withholding data from each individual in the data set, instead of withholding all the data from certain individuals. Further, the comparison between model predictions and the binned data gives only a coarse estimate of model fit. Apparent fit can change with number of bins, which is determined arbitrarily. Finally, there are no guidelines to indicate what threshold of *r* represents a "good" fit of the model (Pearce & Boyce, 2006). While *k*-fold cross validation remains an important tool in model validation, improvements are likely to continue with time.

Summary: using GLMMs and GEEs to develop RSFs using telemetry data

Both generalized estimating equations and generalized linear mixed models can be used to develop resource selection functions that are robust to the lack of independence among numerous locations collected from individual animals, if they are used in conjunction with empirical standard errors. The decision of which approach to apply should depend on whether a marginal or conditional approach should be taken, which in turn depends on the research or management goals. Following the development of the RSF, *k*-fold cross validation can be useful for model validation; usually this should be conducted by withholding all data collected from individual animals, developing RSFs with the remaining animals, and then comparing these models against the data from the withheld animals.

Acknowledgements

We thank all organizers and volunteers of the 2010 North American Caribou Workshop. Funding was provided by the University of Manitoba and the Natural Sciences and Engineering Research Council.

References

Agresti, A. 2002. *Categorical Data Analysis*. John Wiley & Sons, Hoboken.

Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F. K. A., & Manseau, M. 2010. Using information criteria to select the correct variance-covariance structure for longitudinal data in Ecology. – *Methods in Ecology and Evolution* 1: 15-24.

Bishop, J., Die, D., & Wang, Y.-G. 2000. A generalized estimating equations approach for analysis of the impact of new technology on a trawl fishery. – *Australian and New Zealand Journal of Statistics* 42: 159-177.

Bolker, B. B., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. – *Trends in Ecology and Evolution* 24: 127-135.

Boyce, M. A., Pitt, J., Northrup, J. M., Morehouse, A. T., Knopff, K. H., Cristescu, B., & Stenhouse, G. B. 2010. Temporal autocorrelation functions for movement rates from global positioning system radiotelemetry data. – *Philosophical Transactions of the Royal Society B*. 365: 2213-2219.

Boyce, M. A., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. 2002. Evaluating resource selection functions. – *Ecological Modelling* 157: 281-300.

Clifford, P., Richardson, S., & Hémon, D. 1989. Assessing the significance of the correlation between two spatial processes. – *Biometrics* 45: 123-134.

Cushman, S. A., Chase, M., & Griffin, C. 2005. Elephants in space and time. – *Oikos* 109: 331-341.

Dobson, A. J. & Barnett, A. G. 2008. *An Introduction to Generalized Linear Models.* 3rd edn. Chapman & Hall, Boca Raton, Florida.

Fieberg, J., Matthiopoulos, J., Hebblewhite, M., Boyce, M. S., & Frair, J. L. 2010. Correlation and studies of habitat selection, red herring or opportunity? – *Philosophical Transactions of the Royal Society B*. 365: 2233-2244.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. 2004. *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken.

Fortin, D., Beyer, H. L., Boyce, M. S., Smith, D. W., Duchesne, T., & Mao, J. S. 2005. Wolves influence elk movements: behavior shapes a trophic cascade in Yellowstone National Park. – *Ecology* 86: 1320-1330.

Gillies, C. S., Hebblewhite, M., Nielsen, S. E., Krawchuk, M. A., Aldridge, C. L., Frair, J. L., Saher, D. J., Stevens, C. E., & Jerde, C. L. 2006. Application of random effects to the study of resource selection by animals. – *Journal of Animal Ecology* 75: 887-898.

Hardin, J. W. & Hilbe, J. M. 2003. *Generalized Estimating Equations*. Chapman and Hall, New York.

Hox, J. J. 2002. *Multilevel analysis: Techniques and applications.* Lawrence Erlbaum Publishers, Mahwah, N.J.

Koper, N. & Manseau, M. 2009. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. – *Journal of Applied Ecology*. 46: 590-599.

Lele, S. R. & Keim, J. L. 2006. Weighted distributions and estimation of resource selection probability functions. – *Ecology* 87: 3021-3028.

Manly, B. F., McDonald, L. L., Thomas, D. L., McDonald, T. L., & Erickson, W. P. 2002. *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. 2nd edn. Kluwer, New York, New York, USA.

Overall, J. E. & Tonidandel, S. 2004. Robustness of generalized estimating equation (GEE) tests of significance against misspecification of the error structure model. – *Biometrical Journal* 46: 203-213.

Pan, W. 2001. Akaike's information criterion in generalized estimating equations. – *Biometrics* 57: 120-125.

Pearce, J. L. & Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. – *Journal of Applied Ecology* 43: 405-412.

Quinn G. P. & Keough M. J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, New York.

Appendices

Appendix I

SAS script for generalized linear mixed-effects model, annotated. The SAS script uses the procedure, "GLIMMIX".

```
proc sort data = YOURNAME;
by GROUPINGVARIABLE;
run;

* It is necessary for mixed model data to be ordered first by the grouping variable. As such, it is good practices to include a proc sort script prior to any GLMM, to be sure that data are sorted by group prior to the analysis. For RSFs, grouping variable would usually be caribou ID;

TITLE1 'GLIMMIX model';
proc glimmix data = YOURNAME empirical;

* This ensures that the standard errors provided are empirical standard errors;

class GROUPINGVARIABLE INDEPENDENT1 INDEPENDENT2 INDEPENDENT3;

*Above includes all categorical variables;

model RESPONSE = INDEPENDENT1 INDEPENDENT2 INDEPENDENT3 /solution ddfm = betwithin dist = binomial link = logit CL;

* RESPONSE is the name of the column with the response variables (1s and 0s), other variables are the independent variables of interest. ddfm = changes the way that degrees of freedom are calculated. Betwithin stands for Between – Within, the most intuitive method of calculating standard errors. An alternative sometimes preferred by statisticians is Satterthwaite, ddfm = SATTERTH.
random intercept /subject = GROUPINGVARIABLE TYPE = vc;
nloptions tech = nrridg;

*uses newton-raphson with ridging optimization technique, previous line may not be necessary for many data sets;
Title 'Glimmix model';
output out = Glimmixconditional pred = p Pearson = PEARSRESID UCL = UPPER LCL = LOWER;

*creates an output file with residuals, which can be analyzed in SAS or exported to Excel for further examination;
Run;
```

SAS model for generalized linear mixed-effects model, without annotation

```
proc sort data = YOURNAME;
by GROUPINGVARIABLE;
run;

TITLE1 'GLIMMIX model';
proc glimmix data = YOURNAME empirical;
class GROUPINGVARIABLE INDEPENDENT1 INDEPENDENT2 INDEPENDENT3;
model RESPONSE = INDEPENDENT1 INDEPENDENT2 INDEPENDENT3 /solution ddfm = betwithin dist = binomial link = logit CL;
random intercept /subject = GROUPINGVARIABLE TYPE = vc;
Title 'Glimmix model';
output out = Glimmixconditional pred = p Pearson = PEARSRESID UCL = UPPER LCL = LOWER;
Run;
```

Appendix II

SAS script for developing generalized linear model with generalized estimating equation, annotated. The SAS code uses the procedure, "GENMOD"

```
proc sort data = YOURNAME;

by GROUPINGVARIABLE;
run;

* It is necessary for mixed model data to be ordered first by the grouping variable. As such, it is good practices to include a proc sort script prior to any GLMM, to be sure that data are sorted by group prior to the analysis. For RSFs, grouping variable would usually be caribou ID;

TITLE1 'GEE model';
proc genmod data = YOURNAME descending;

*descending command ensures that used habitat is compared with available habitat, rather than the reverse. By including "descending", this ensures that positive parameter estimates indicate that habitat is selected, while negative parameter estimates indicate that habitat is avoided;
class GROUPINGVARIABLE INDEPENDENT1 INDEPENDENT2 INDEPENDENT3;
*Above includes all categorical variables;

model RESPONSE = INDEPENDENT1 INDEPENDENT2 INDEPENDENT3 / dist = binomial corrb;

* RESPONSE is the name of the column with the response variables (1s and 0s), other variables are the independent variables of interest;
repeated subject = GROUPINGVARIABLE / corr = CS modelse;

*corr = indicates the correlation structured desired (Independent = IND, Compound Symmetric = CS, Autoregressive = AR(1). Model SE will produce both model and empirical standard errors, so that the ratio of SE_E to SE_M can be compared to evaluate model fit;
output out = RESIDS predicted = inverselogit reschi = pearsresid stdreschi = stpearsresid STDXBETA = stdxbeta xbeta = logit;
Run;
```

SAS script for developing generalized linear model with generalized estimating equation, without annotation

```
proc sort data = YOURNAME;
by GROUPINGVARIABLE;
run;

TITLE1 'GEE model';
proc genmod data = YOURNAME descending;
class GROUPINGVARIABLE INDEPENDENT1 INDEPENDENT2 INDEPENDENT3;
model RESPONSE = INDEPENDENT1 INDEPENDENT2 INDEPENDENT3 INDEPENDENT4 / dist = binomial corrb; [ ]
repeated subject = GROUPINGVARIABLE / corr = CS modelse; [...]
output out = RESIDS predicted = inverselogit reschi = pearsresid stdreschi = stpearsresid STDXBETA = stdxbeta xbeta = logit; [...]
Run;
```

Appendix III

R script for developing generalized linear model with generalized estimating equation, annotated. The R script uses the library "geepack" (R code from Dobson & Barnett 2008).

```
>geeind<-geeglm (RESPONSE ~ INDEPENDENT1 INDEPENDENT2 INDEPENDENT3,

# RESPONSE is the name of the column with the response variables (1s and 0s), other variables are the independent variables of interest;

family = binomial, data = YOURNAME, id = GROUPINGVARIABLE, wave = time, corst ="independence")

#corst = indicates the correlation structured desired (Independent = independence, Compound Symmetric = exchangeable, Autoregressive = AR1)
```

R script for developing generalized linear model with generalized estimating equation, withouth annotation.

```
>geeind<-geeglm (RESPONSE ~ INDEPENDENT1 INDEPENDENT2 INDEPENDENT3, family = binomial,

data = YOURNAME, id = GROUPINGVARIABLE, wave = time, corst ="independence")
```