

GEEs and GLMMs for modelling RSFs

Dr. Nicola Koper
Natural Resources Institute
University of Manitoba

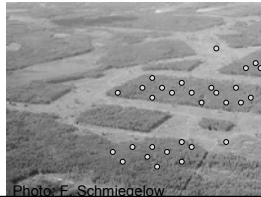
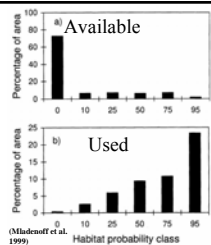
Photo by M. Manseau

Outline

- Background & Rationale
- Correlation structures
- Mixed Models
- GEEs
- Predictive capacity of models
- SAS:
 - GLMM
 - GEE
 - K-fold cross validation

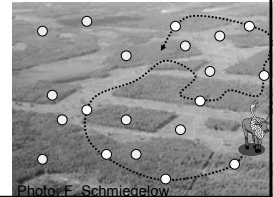
What are Resource Selection Functions?

- Models used to compare amount of used habitat with amount of available habitat
- Do animals use habitat in proportion to its availability, or do they select certain kinds of habitat over other kinds?



Resource Selection Functions with telemetry data

- Compare real locations collected by observing animals, e.g. using satellite collars, with random locations sampled from all across the landscape
- We assume that random points are “unused”, while we know real points were used
- If lots of real points are found in a habitat that only covers a small portion of the landscape, we assume the species is *selecting for* that habitat
- If a habitat covers a lot of area but we find few points there, we assume the species is *avoiding* that habitat



How are RSFs and RSPFs different?

- Resource Selection Probability Functions
- RSPF defines the actual probability of a resource being used
- RSF defines the relative probability of each habitat being used compared to other habitats
 - If Habitat_1 has a probability of use of 0.1 and Habitat_2 has a probability of use of 0.5, then Habitat_2 is 5 times as likely to be used
 - If Habitat_1 has a probability of use of 0.001 and Habitat_2 has a probability of use of 0.005, then Habitat_2 is 5 times as likely to be used
 - The probability of use has changed, but the relative probability of use hasn't changed

Lele and Keim 2006

How are RSFs and RSPFs different?

- RSPFs are much more sensitive than RSFs to the assumption that random points really are unused
 - However, random points could be “contaminated” with used points
 - This means that maybe caribou DID use some of the random points, we just never observed them there so assumed those points were unused
- Because we cannot be certain random points were unused, we restrict the rest of the presentation to RSFs

RSFs with satellite telemetry data

- Tons of data!
 - However, correlated
 - Cannot use traditional analysis method that assumes independence
 - Destructive sampling or long periods between locations does not solve the problem
 - Loses data, reducing accuracy of models
 - May still be correlated as long as a month apart (Cushman et al. 2005)
 - Our data retained some correlation at 3.33 days apart (dropping 95% of the data)

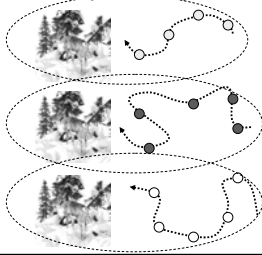


Other published attempts to deal with the problem

- Information theory (e.g., AIC)
 - Incorrect application:
 - SE assume independence
 - Likelihood, & therefore AIC, assumes independence
- Conditional logistic regression
 - Assumes equal correlation among points within groups; not met with telemetry data collected over time
 - Fortin (2005) adapted for “step scale” (how do they decide where to go given where they are starting?), but not broader spatial scales

GLMM

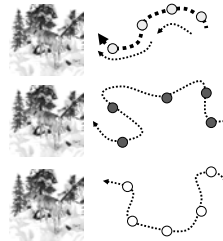
- Gillies et al. (2006) recommended use of models with fixed & random variables (including GLMM) to control for recording multiple samples from 1 animal



Including a random variable in a model (e.g. random = caribou) controls for this kind of correlation

GLMM

- Gillies et al. (2006) recommended use of models with fixed & random variables (including GLMM) to control for recording multiple samples from 1 animal



But not this kind

Gillies et al. mis-specified the correlation structure, because they did not take this important source of correlation into account

They assumed that all points from one animal are equally correlated, but in fact points that are closer in time are more correlated

2 potential problems with GLMM

1. GLMM are complex
 - Satellite telemetry data has 1000s of points within just a few animals; very complex matrix algebra involved
 - Can take forever for models to run (days), or may never converge because too complex
2. GLMM are *not robust* to mis-specifying the correlation structure
 - If the correlation structure is incorrect, the resulting parameter estimates are not trustworthy
 - The methods outlined by Gillies et al. are incorrect... but can we correct them and still use GLMM?



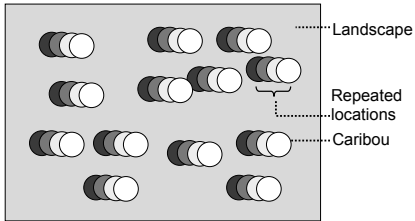
Can we correctly specify the correlation structure?

- Problem: telemetry locations collected from one animal are correlated over time.

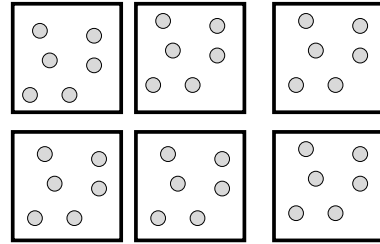


Modelling correlation structures

- Although Gillies et al. did not do so, it is possible to model correlation within groups, in this case, caribou

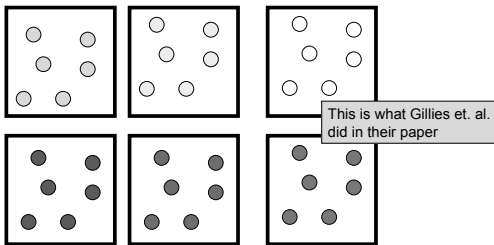


Modelling correlation structures



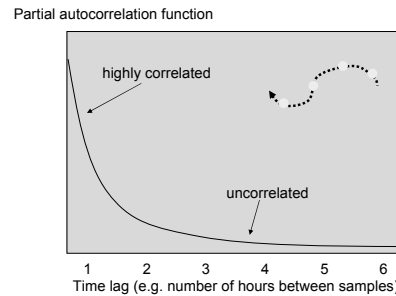
Independent: all samples are independent of one another

Modelling correlation structures



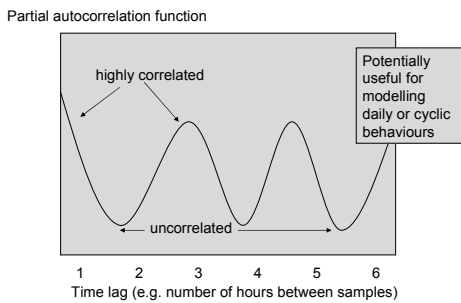
Compound symmetric / exchangeable: all plots are correlated with plots within the same group (caribou), but groups are independent of one another

Modelling correlation structures



Autoregressive: samples are more correlated with samples taken close to them in time

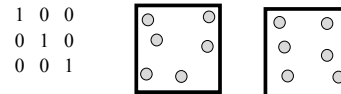
Modelling correlation structures



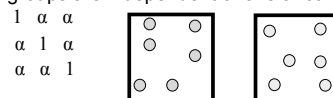
Banded: samples are equally correlated if (for example) taken 2 days apart or 4 days apart, but correlation doesn't necessarily go down over time

Correlation matrices

- Independent: assumes no correlation




- Compound symmetric / exchangeable: all plots are correlated with plots within the same group (caribou), but groups are independent of one another



Correlation matrices

- Autoregressive: assumes that correlation declines exponentially with time lag, but never declines to zero

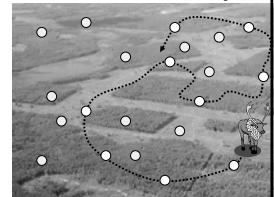
$$\begin{matrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{matrix}$$


- Banded: assumes data are correlated with all other data points, and that correlation *varies* with time lag, but may not *decline* as time lags increase. Can allow correlation to decline to zero.

$$\begin{matrix} 1 & \alpha_i & \alpha_{ii} \\ \alpha_i & 1 & \alpha_i \\ \alpha_{ii} & \alpha_i & 1 \end{matrix}$$

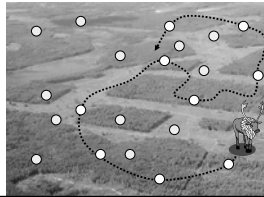

Can we correctly specify the correlation structure?

- Real locations
- Random locations
- Real and random locations have different correlation structures – cannot be correctly specified in GLMM



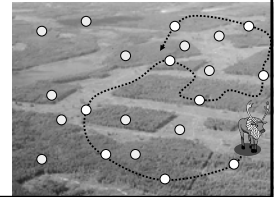
Can we correctly specify the correlation structure?

- No – not within the typical study design many of us want to use, to address the question: what habitat do caribou choose to use, relative to the habitat available to them?
- Can we model habitat use without correctly specifying the correlation structure?



Can we model habitat use without correctly specify the correlation structure?

- Yes – using both GLMM and GEE in combination with robust / empirical / sandwich standard errors
- What are GLMM and GEE, how can they be used for RSF, and how do they differ?



Outline

- Background & Rationale
- Correlation structures
- Mixed Models
- GEEs
- Model selection
- SAS:
 - GLMM
 - GEE
 - K-fold partitioning

Mixed models

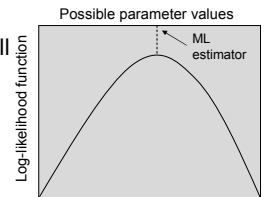
- Models that include both random and fixed variables:
 - Mixed models, mixed effects models (LME)
 - Hierarchical models
 - Correlated models
 - Clustered models
 - Longitudinal models
 - Multilevel models
 - GLME
 - GLMM
 - NLME

Generalized Linear Models

- GLMM are derived from Generalized Linear Models (GLM)
- “Generalized” means that the distribution of the response variable (e.g. probability of occurrence of caribou) does not have to be normal
 - Can use Poisson, binomial, negative binomial, etc.
 - In RSF = binomial
- Logistic regression is a special case of GLM where the distribution of the response variable is binary (1 or 0)
- Linear mixed models are a special case where distribution of the response variable is normal
- With telemetry data, must use *generalized* linear mixed models, to allow response data to follow a binomial distribution

Estimation procedures

- A different paradigm than OLS
- Maximum likelihood (ML)
 - Principle: Given a sample of observations, find estimates of parameters that maximize the likelihood of observing those data
 - Gives the likelihood of observing the data for all possible values of μ (pop. mean)



Maximum likelihood

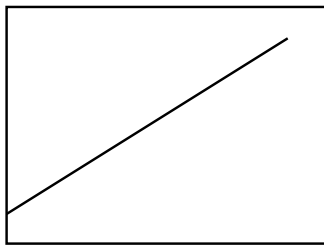
- Easier to work with log-likelihood than likelihood:
 - $L(\theta) = \sum \ln(f(y_i; \theta))$
 - Sum of the natural logs of the joint probability distributions of y over the possible values of θ
 - $f(y_i; \theta) \sim$ distribution (e.g. Gaussian, Poisson; binomial for RSF)
- Same log-likelihood as used in other procedures we are familiar with, e.g. AIC
- ML is robust to data missing at random

Mixed models: so-called because they contain both fixed and random variables

- Fixed:
 - All levels of interest for the factor included in design
 - Inference restricted to these levels
- Random:
 - Randomly selected levels
 - Generalize inference over all levels of random variable
- Example: Caribou (e.g., collars 1,2,3)
 - Restrict inference to just those caribou we studied, or to other caribou in the population?
 - Presumably all caribou; so “caribou” is treated as a random or grouping variable
 - Telemetry points are *grouped* within caribou
 - Controls for the fact telemetry points within caribou are not independent of one another, because they are from the same animal

“let’s pretend” linear mixed models

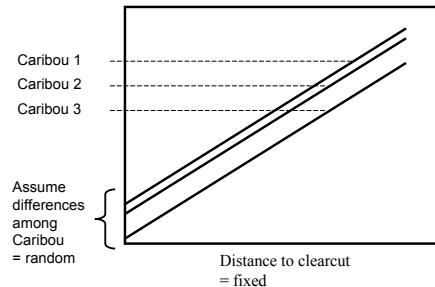
Probability of occurrence



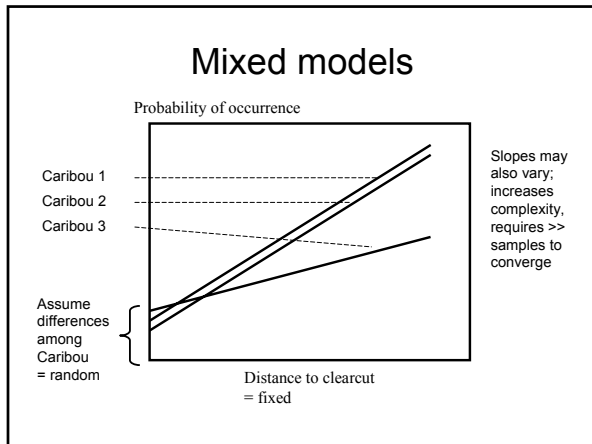
Distance to clearcut

Mixed models

Probability of occurrence



Distance to clearcut
= fixed

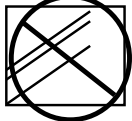


- ### BUT...
- Using random variables generally increases standard errors across all variables in the model, both random and fixed:
 - Expected Mean Square for the FIXED factor includes the variance component for the INTERACTION between the fixed and random factors ... if the variance for the interaction is high, power to detect effects of the fixed factors is lowered.
 - This makes sense... uncertainty generalizing results across *all possible* levels of random variables is higher than uncertainty of results across *specific* levels of variables that we measured
 - Power therefore lower than if all fixed parameters ... but analysis is more appropriate

- ### Estimation procedures for GLMM
- ML, or many other options, including:
 - Restricted maximum likelihood (REML)
 - Uses a *restricted* part of the likelihood
 - Penalized quasi-likelihood (PQL)
 - REML, others perform better than ML
 - ML tends to produce variances that are too small
 - However, only ML produces a true log-likelihood; all other likelihoods are approximation
 - ∴ only ML can be used if using log-likelihood for other purposes, e.g. information-theoretic model selection (AIC)
 - ML may not be possible for some GLMM designs with >1 random variable

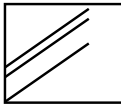
- ### Outline
- Background & Rationale
 - Correlation structures
 - Mixed Models
 - GEEs
 - Model selection
 - SAS:
 - GLMM
 - GEE
 - K-fold partitioning

- ### Generalized Estimating Equations
- Generalized estimating equations within a GLM (GEE) are a semi-parametric alternative to mixed models
 - Semi-parametric because β are estimated parametrically, while variance component is estimated non-parametrically
 - Introduces second-order variance component to describe the correlation within clusters
 - This variance component uses one of the correlation structures we introduced at the beginning of the presentation (e.g., independent, compound symmetric, autoregressive, banded, or others)

- ### Generalized Estimating Equations
- GEEs do NOT explicitly model differences among groups
 - The modification of the variance component adjusts for the correlation within groups
 - Compound symmetric (CS) or exchangeable (same) correlation structure within a GEE is equivalent to adding 1 random variable in mixed model
 - LME and GEE w/ CS give SAME results
 - I.e., if response variable is normal
 - **GLMM and GEE w/ CS give DIFFERENT results**
 - I.e., if response variable is NOT normal
- 

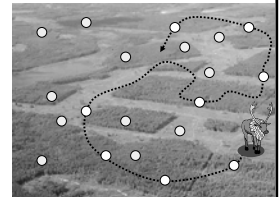
GLMM versus GEE

- GLMM accounts for correlations of observations within individuals by modeling differences between animals
 - Parametric, complex
- GEE accounts for differences among individuals by adjusting the standard error to account for the lack of independence of observations within individuals
 - Semi-parametric, simpler



Generalized estimating equations

- Benefits of GEEs:
 - Less analytically complex than GLMM; more likely to converge, and will take less time
 - Parameter estimates and empirical standard errors robust to misspecification of the correlation structure
 - While Gillies et al. (2006) argue robust standard errors alone did not perform as well as GLMM, robust standard errors with GEE might perform better.
- Disadvantages of GEEs:
 - Sensitive to selection of the link function



Background on GEEs: 3 components

1. Relationship between the response variable and the independent variables via a link function
2. Conditional variance of the data given the independent variables, including a scale parameter

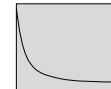
Background on GEEs: 3 components

3. Substitutes variance component of GLM with:

$$\mathbf{V}(\mu_i) = [D(\mathbf{V}(\mu_i))^{1/2} \mathbf{R}(\boldsymbol{\alpha})_{(ni \times ni)} D(\mathbf{V}(\mu_i))^{1/2}]_{ni \times ni}$$

- $\mathbf{V}(\mu_i)$ = variance of the marginal mean
- D = diagonal matrix
- $\mathbf{R}(\boldsymbol{\alpha})_{(ni \times ni)}$ = working correlation matrix
 - The correlation structure we are assuming is correct
 - $\boldsymbol{\alpha}$ can be 1 value (e.g. compound symmetric), or a matrix
 - E.g., Autoregressive:

$$\begin{matrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{matrix}$$



Model-based variance versus empirical variance

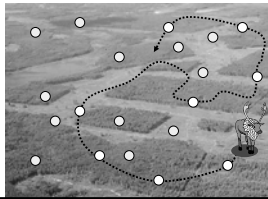
- This is used to derive the model-based variance
 - Variance based on assuming the correlation structure in the model is the real correlation structure
 - If we are wrong about the correlation structure, then this variance will also be wrong
- Often replaced by empirical, robust, or “sandwich” variance
 - Robust even if you’re using the wrong correlation structure, or if the correlation structure you’re using doesn’t correctly describe the correlation in the data

Model-based variance versus empirical variance

- Empirical variance is a theoretical variance, rather than the true variance, because it cannot be known for sure
- Can use empirical variance in GLMM OR GEE
- Empirical variance estimator often conservative
 - Larger than the model variance estimator
- The closer you are to using the correct correlation structure, the closer the empirical and model-based variances will be, and the smaller the empirical variance will be
 - Therefore, more powerful test if closer to correct correlation structure

Model-based variance versus empirical variance: RSFs

- Because the correlation structure in the real data is different from the correlation structure in the random data, we cannot model the correlation structure correctly
- Therefore, we *must* use the empirical variance estimator in RSF models derived using telemetry data



Example: model-based versus empirical standard errors

- Distance to hardwood-mixedwood stands
- Parameter estimate
Occurrence declines as we get further from HW-MW stands
- Does NOT change if change from model-based to empirical variance

	Empirical variance	Model-based variance
β	-0.19	-0.19
SE	0.069	0.007
p	0.782	0.008

Example: model-based versus empirical standard errors

- Distance to hardwood-mixedwood stands
- Standard error
Estimate of how variable the parameter estimate is
- Increases by factor of 10 when go from Model-based to Empirical variance

	Empirical variance	Model-based variance
β	-0.19	-0.19
SE	0.069	0.007
p	0.782	0.008

Example: model-based versus empirical standard errors

- Distance to hardwood-mixedwood stands
- p -value
By convention, generally treat values < 0.05 as indicating a significant effect of the variable
- Dramatic increase in p -value when go from model-based to empirical variance changes from significant to insignificant effect

	Empirical variance	Model-based variance
β	-0.19	-0.19
SE	0.069	0.007
p	0.782	0.008

Model-based variance versus empirical variance

- Message # 1 from this presentation:
- **ALWAYS USE THE EMPIRICAL STANDARD ERROR** when using GLMM or GEE with satellite telemetry data

Can we improve fit by trying to model the correlation structure?

- Independent (no correlation) versus compound symmetric (CS; grouped by caribou)
- Little improvement in fit in OUR study ($SE_E/SE_M = 8.65$ for Independent, 8.15 for CS)
- Implies that accounting for differences among animals removes only a small proportion of the correlation in the data
- However, in other data sets, both correlation structures should be compared ... perhaps it is more important in some areas than others

Interpreting parameter estimates: conditional versus marginal

- Marginal
 - interpretation of parameter estimates is on a *population-specific* basis
 - models effects of the independent variables on the population, independent of the correlation structure
 - GEE
 - Can be generated from GLMM, but marginal parameter estimates from GLMM are biased (downwards); interpretation not straightforward; relationships among covariates difficult to interpret
- Conditional
 - Interpretation of parameter estimates is *subject-specific*
 - coefficients model how individual responses change with respect to independent variables, not effects of independent variables on a population
 - GLMM

Interpreting parameter estimates: conditional versus marginal

- Marginal
 - “How does use of a particular drug change cancer rates across the population?”
- Conditional
 - “If we give a typical person this drug, how likely are they to recover?”
- Marginal
 - “What is the difference in habitat use of caribou between landscapes with lots of pine vs. little?”
- Conditional
 - “How would a typical caribou change it’s habitat use if their habitat changed from having a lot of pine to having very little pine?”

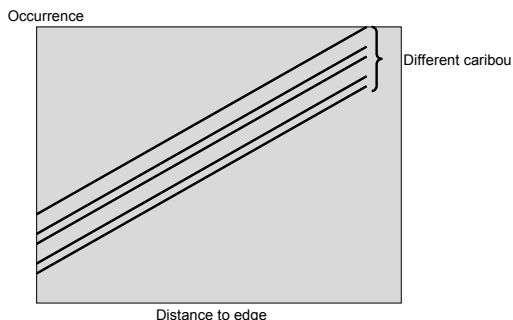
Management & Ecological implications of conditional versus marginal parameters

- Marginal response is appropriate if management is intended to influence the population
 - E.g., landscape-level plans, population recovery plans
 - Are you studying a subset of the population so you can understand how the whole population would respond?
 - = GEE
- Conditional response is appropriate if management is intended to influence particular individuals
 - E.g., conservation of highly endangered species or populations
 - Predicting changes to individuals if the habitat changes
 - Are you studying animals so you can manage those individuals, or future changes to typical individuals?
 - = GLMM

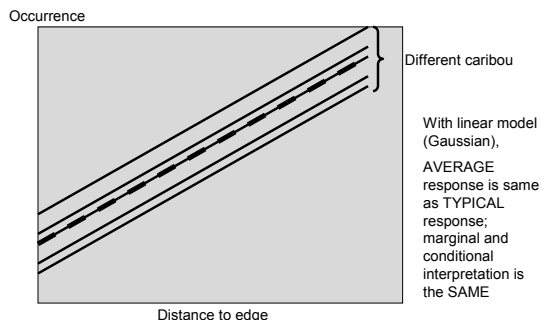
Interpreting parameter estimates: conditional versus marginal

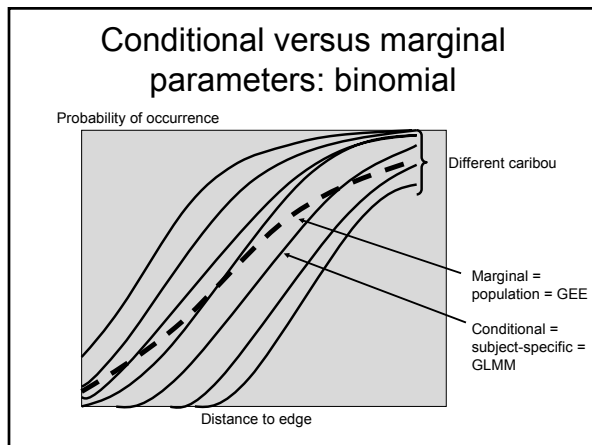
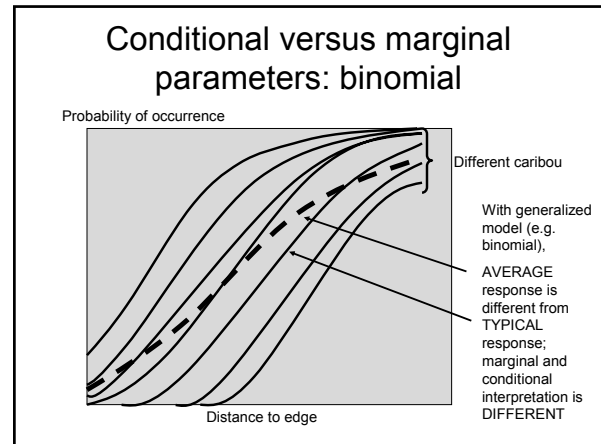
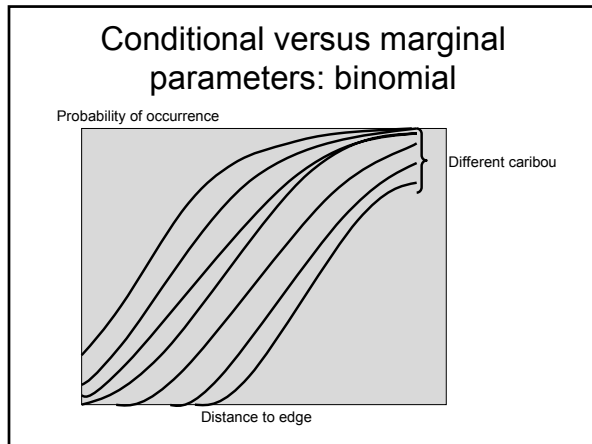
- In LME (any mixed model with a normal distribution), no difference between conditional and marginal interpretation / estimates
- In generalized linear models (not a normal distribution; e.g., binomial distribution), conditional versus marginal interpretation has a **strong** effect on parameter estimates, standard errors, significance
 - Marginal effects are usually smaller
- **Critical** to:
 - Use appropriate parameter estimates
 - Interpret parameter estimates correctly

Conditional versus marginal parameters: Gaussian



Conditional versus marginal parameters: Gaussian





Marginal (GEE) versus Conditional (GLMM) parameter estimates

	Marginal (GEE)	Conditional (GLMM)
Treed muskeg	1.991 $p < 0.001$	0.920 $p = 0.003$
Jack pine	0.168 $p = 0.008$	0.704 $p = 0.080$
Spruce	1.223 $p < 0.001$	0.575 $p = 0.062$

- ### Conditional versus marginal parameters
- Message # 2 from this presentation:
 - **Selecting a conditional versus marginal parameter estimate is very important!**
 - It will change your results
 - It is critical that once you select which approach to use, that your interpretation reflects that analysis.

- ### Advantages of GEE
- Simpler analysis than GLMM (because not parametric)
 - fewer convergence problems
 - usually takes less time to converge
 - Some additional useful correlation structures are possible
 - Estimation is on marginal (population) basis
 - Parameter estimates, and sandwich or empirical variance estimates, are robust to misspecification of the correlation structure

Disadvantages of GEE

- Cannot use in conjunction with AIC because no log-likelihood, adjusted for the correlation, is produced
 - Our simulation studies (Barnett et al. 2010) show problems with the quasi-likelihood equivalent to AIC: QIC
 - Highly biased towards selecting more complex models
 - Corrected QIC is being developed currently

GLMM versus GEE

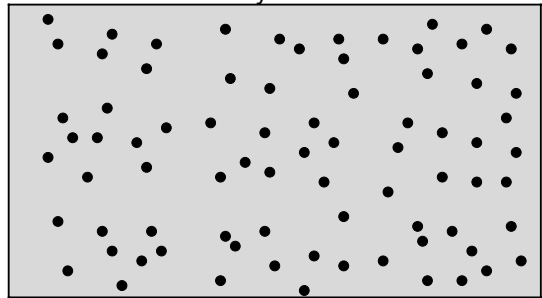
- GLMM better if:
 - Hierarchical (nested) design
 - Sample sizes within groups approx. equal
 - It converges
 - You want Conditional estimates
- GEE better if:
 - Convergence problems with GLMM
 - Sample sizes within groups quite variable
 - Clustering is a nuisance, not the focus of study
 - You want Marginal estimates

Outline

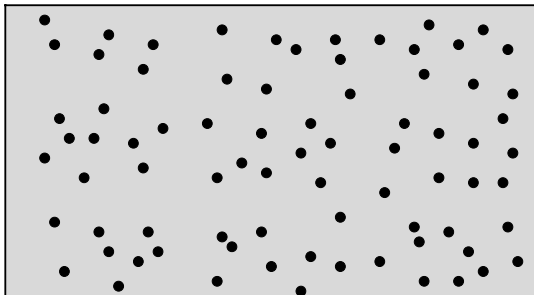
- Background & Rationale
- Correlation structures
- Mixed Models
- GEEs
- Predictive capacity of models
- SAS:
 - GLMM
 - GEE
 - K-fold partitioning

How well does my subsample of data predict behaviour of the population?

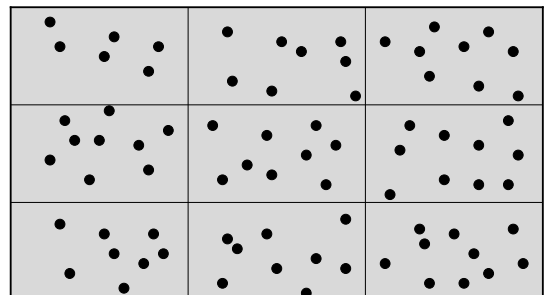
How trustworthy is the inference?



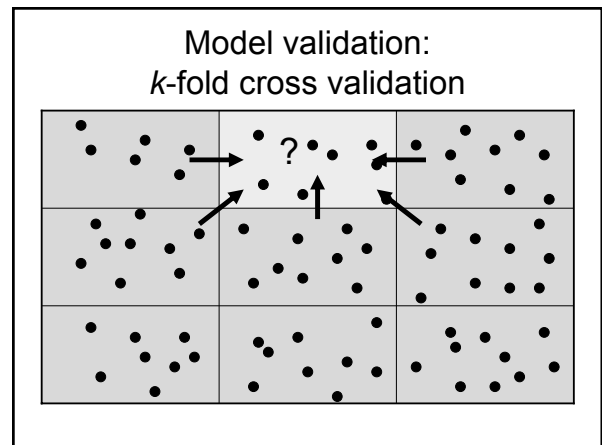
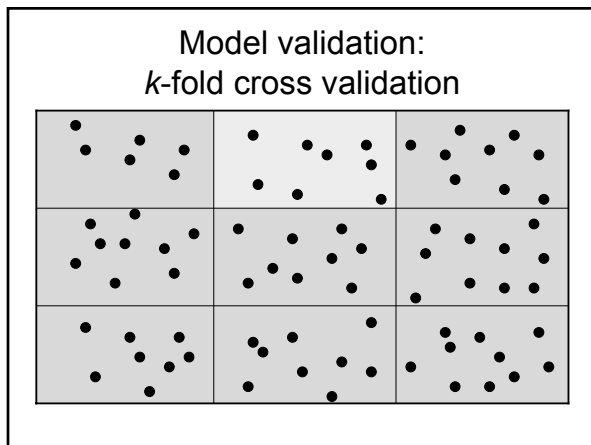
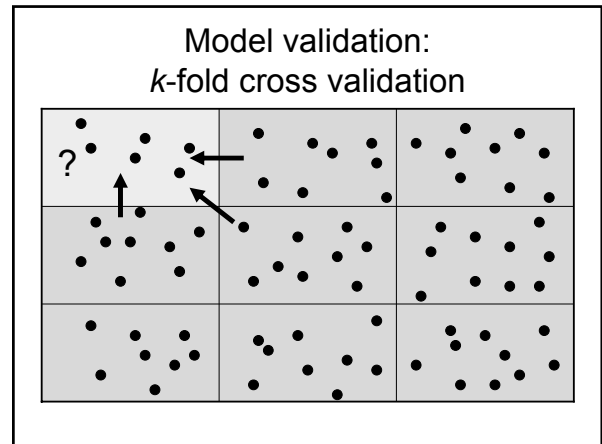
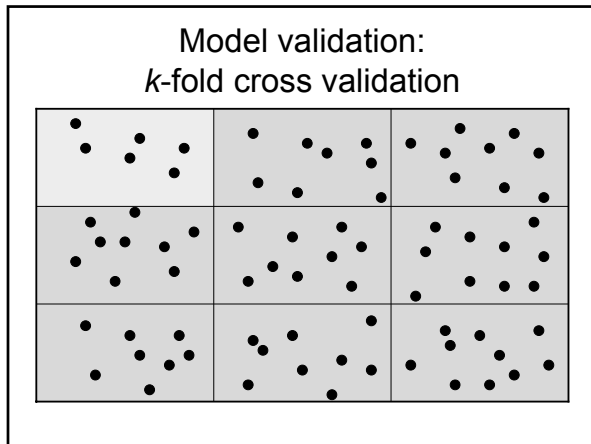
Model validation: *k*-fold cross validation



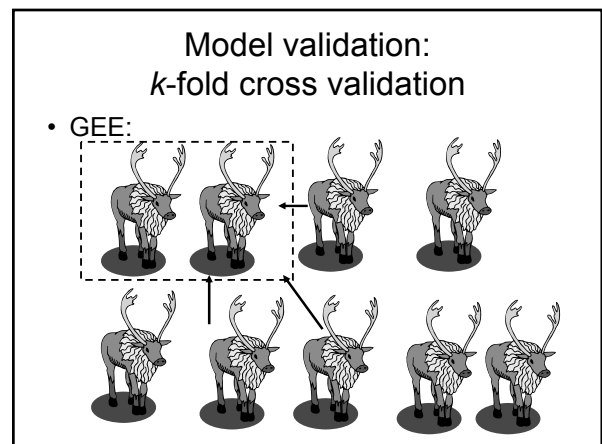
Model validation: *k*-fold cross validation



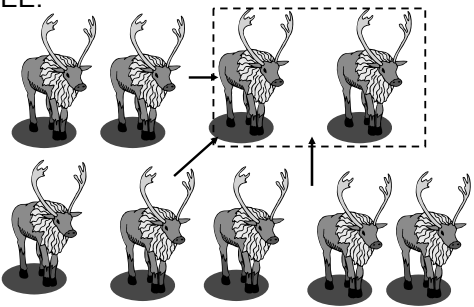
Separate into "bins"



- Model validation:
k-fold cross validation
- GEE:
 - Marginal (population-specific) estimates
 - Predict habitat selection of all animals in the population from a subset of that population
 - Therefore, withhold 15% of the individuals
 - Develop models with remaining 85% of the animals
 - Compare fit with the withheld individuals



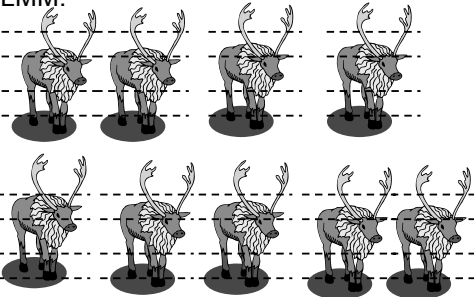
Model validation: k-fold cross validation

- GEE:
 

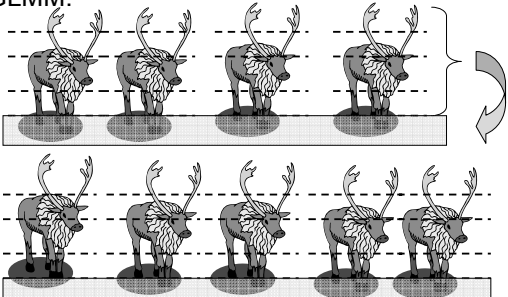
Model validation: k-fold cross validation

- GLMM:
 - Conditional (subject-specific) estimates
 - Predict habitat selection *by the same animals*
 - Therefore, withhold 15% of the points from each animal
 - Develop models with remaining 85% of the points
 - Compare fit with the withheld points from the same animals

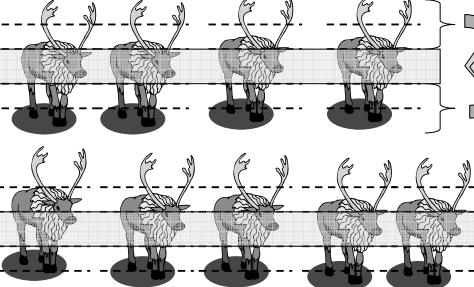
Model validation: k-fold cross validation

- GLMM:
 

Model validation: k-fold cross validation

- GLMM:
 

Model validation: k-fold cross validation

- GLMM:
 

Model validation: k-fold cross validation

- Obviously, fit will appear better when compare against data from the same animal ($r = 0.962$), relative to the fit compared with data from other animals ($r = 0.739$)
 - Higher r implies better fit
- Just shows that we are better at predicting habitat selection of the animals used to develop the models, compared with predicting habitat selection of other animals from the same herd
- Therefore, this approach cannot be used to compare models developed from GEE and GLMM
 - ie, CANNOT answer the question, "Which model is better for my data, GEE or GLMM?"
- This is appropriate because the meaning of the GEE parameter estimates is not the same as the meaning of the GLMM parameter estimates

Is k -fold cross validation the answer?

- Advantages:
 - Can be used with any model that produces parameter estimates
 - Conceptually straightforward
 - Becoming common in the literature
- Disadvantages:
 - Misused in the literature (e.g., withholding points when should be withholding animals)
 - Bins can at best give only coarse estimate of whether model fits the data well
 - Apparent fit can change if change number of bins
 - No guidelines for thresholds for r : what is a high enough value for a model to “fit well”?
- Best available, but still room for improvement!

Model validation: k -fold cross validation

- Message # 3 from this presentation:
- Inference of marginal and conditional models differs
- **Therefore, the appropriate way to test model fit differs**
- Test fit against withheld **ANIMALS** for marginal model
- Test fit against withheld **POINTS** for conditional models

Summary

- We need to account for correlation over time in telemetry data
- This cannot be accurately modeled because correlation in real data differs from correlation in random data
- Both GLMM and GEE can be used IF we use empirical standard errors instead of model-based standard errors



Summary

- Generally, if we want marginal parameter estimates, we should use GEE
- If we want conditional parameter estimates, we should use GLMM
- Because inference differs for marginal and conditional parameter estimates, the way to test for model fit also differs



Acknowledgements

- Support:
J. Hilbe, L. Lix, S. Keobouasone, L. Fitch, M. Manseau
- Field work:
F. Moreland, D. Frandsen, A. Arsenault, T. Trottier, T. Tokaruk
- Funding and in-kind:
NSERC, Parks Canada, University of Manitoba, Saskatchewan Environment and Resource Management, Weyerhaeuser, Prince Albert Model Forest

Photo: L. Neufeld