Steven T. Hoekman. 2021. Multi-observer methods for estimating uncertain species identification. *Ecosphere* 12(7): e03648.

Metadata S4: Input files and R v. 4.10 computer code for analyses of Glacier Bay murrelet survey data

---

Author

Steven T. Hoekman
Wild Ginger Consulting
P. O. Box 182
Langley, WA, USA 98260
steven.hoekman@protonmail.com

File list (found within folder 'DataS4-Analyses_of_Murrelet_Surveys')

R files

`murrelet_r_code.R`	R code for statistical analyses
`murrelet_data.RData`	R workspace with data and functions
`likelihood_equations_a3b2g25.RData`	R list with tables of pre-computed values

Dependencies (required)

R packages (available at <https://cran.r-project.org/web/packages/>)

`dplyr`	Data frame manipulation

Files description

murrelet_data.RData

An R workspace containing data, required R objects and functions, and statistical output for analyses of line transect surveys for two murrelet species. R objects in this workspace include:

'murrelet_data'
Data frame with survey data (Table 1). Individuals in each observed group were classified to one of three observation states: 1 = Kittlitz's murrelet, 2 = marbled murrelet, 3 = genus *Brachyramphus* murrelet (partial identification) by one of two primary observers. A sample of groups were photographed and also classified by each of three independent secondary observers from these images. Other fields include values of group-level covariates used to predict classification probabilities and true species probabilities.

'model_output'
List with pre-computed statistical output from the top model presented in Appendix S5 of the companion article. List elements include: parameter estimates ('par'), the negative log(likelihood) + any penalty term(s) ('value'), and the Hessian matrix for estimation of the variance-covariance matrix ('hessian').

The workspace also includes: the R function 'murrelet.model.f' used to conduct statistical analyses and other supplemental functions provided in the file murrelet_r_code.R, the R list 'likelihood.equations' provided in the likelihood_equations_a2b2g25.RData, and other R objects created by the R computer code (details below).

murrelet_r_code.R

R computer code for estimating uncertain identification using multi-observer method models for two species of murrelets during line transect surveys in Glacier Bay, Alaska, USA during July 2014. Code version 1.1.1 is updated to utilize the R statistical computing environment version 4.1 (R Development Core Team 2021). For code used to conduct analyses in the companion article, see release version 1.0.0 of the simulation engine (archived at Zenodo https://doi.org/10.5281/zenodo.4738002). The statistical model and its parameters (Table 2) are described in Appendix S5 of the companion article.

The R code loads required R packages, data, and functions; specifies initial values and constraints for model parameters; and conducts model optimization.

Analyses are conducted by using the R 'optim' function to call function 'murrelet.model.f' to produce the statistical output in the R list 'model'. The 'optim' function minimizes the negative log(likelihood) using the L-BFGS-B optimization method of Byrd et al. (1995), which allows box constraints placing lower and upper bounds on individual parameters (specified in vectors 'constraint_low' and 'constraint_up').

The estimated heterogeneous group probability ($\pi_{1,2}$) is also constrained using a penalty function so that it does not take inadmissible values. A penalty function within 'murrelet.model.f' defines a penalty term taking value 0 if heterogeneous group probability does not exceed the total availability of groups for either species, but otherwise taking a value >0 that scales with the magnitude of violation of this constraint. Adding the penalty term to the negative log(likelihood) enforces that optimization avoids violating the constraint.

The function 'murrelet.model.f' calls functions provided in section 'Required supplemental functions'. Most of these are simplified versions of those used in simulation analyses (DataS3-R_Code_for_Simulations). Detailed comments in code describe their inputs, outputs, and functioning.

likelihood_equations_a2b2g25.RData

An R file containing the list object 'likelihood.equations' required to provide pre-computed values for likelihood computations. See MetadataS3.pdf for additional details.

Workflow for statistical analyses

To execute statistical analyses for the murrelet survey data:
1) Place the .R and .RData format files within folder DataS4 into the R working directory.
2) In an R editor or integrated R development environment (e.g., *RStudio* by RStudio, Inc., Boston, Massachusetts, USA, <https://rstudio.com>), open file murrelet_r_code.R
3) Install and load required R packages in 'Load R Packages' section.
4) Load the R workspace murrelet_data.RData
      • The workspace includes data, objects, functions, and output from analyses.
5) Navigate to the 'Code for executing statistical analyses' section and follow instructions in the comments to generate a vector 'parameters_ini' of initial parameter values and vectors 'constraint_low' and 'constraint_up' of lower and upper box constraints.
6) Execute the 'optim' function code to produce list 'model' containing statistical output.
      • Optimization may require 10+ minutes.

Literature Cited

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 16:1190–1208.
R Development Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org

Table 1. Naming conventions and descriptions of fields in the data frame 'murrelet_data' containing data from line transect surveys. Classifications of survey observers are in fields with names starting with 'y'; suffixes (in brackets) separated by underscores denote observation states a and observer identities.

Field name	Field values	Description
y_[observer]_[a]	Integer	Counts of murrelets classified by survey observers to each observation state. Text suffix [observer] denotes primary observers 'p' or individual secondary observers 's1', 's2', and 's3', and the integer suffix [a] denotes observation state a.
group_size	Integer	Count of murrelets in a group
perpendicular_distance	Numeric	Estimated perpendicular distance of a group from the transect centerline in decameters (100's of m)
precipitation	Integer	Indicator variable taking value of 1 for groups located when rain, mist, or fog were present and value of 0 otherwise
density_area	Integer	Indicator variable taking value of 1 for groups on in areas with high expected density of Kittlitz's murrelets of value of 0 otherwise

Table 2. Naming conventions for model parameters. The base name (bold) for each type of parameter may have additional prefixes and suffixes (in brackets), separated by an underscore. Suffix(es) may denote observation states, true species states, and observer identity. For multinomial logistic regression predicting classification probabilities, prefixes denote intercept (b0) versus slope (b1) regression coefficients. Parameters are described in Appendix S5 of the companion article.

Parameter type (notation)	Name	Prefix	Suffix(es)
Heterogeneous groups π	**pi**_[*bb*]		Two integers denote true species states b in heterogeneous groups
Classification probabilities θ	[beta]_**distance**_[observer]_[*ab*]	Text 'b0' or 'b1' denotes intercept or slope regression coefficients for multinomial logistic regression with predictor *perpendicular_distance*	1) [observer] Text 'p' or 's' denotes primary or secondary observers 2) [*ab*] Two integers denote classification state a and true species states b
	[beta]_**precipitation**_[observer]	Text 'b1' denotes slope regression coefficients for multinomial logistic regression with predictor *precipitation*	[observer] Text 'p' or 's' denotes primary or secondary observers
True species probabilities ψ	**psi**_[*b*]_[density area]		1) [*b*] Integer denotes true species state b 2) [density area] Text denotes areas of 'low' or 'high' expected densities of Kittlitz's murrelets
Weight for hurdle model ω	**omega**_[*b*]		Integer denotes true species state b
Bernoulli probability τ	**tau**_[*b*]		Integer denotes true species state b
Geometric distribution parameter q	**q**		