Steven T. Hoekman. 2021. Multi-observer methods for estimating uncertain species identification. *Ecosphere* 12(9): e03648.

Metadata S2: Input files for simulation study of multi-observer methods for estimating uncertain identification

---

Author

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

File list (found within folder 'DataS2-Input_Files_for_Simulations\Input_Files')

Simulation input files

```
covariate_m_psi_22.csv
covariate_m_psi_32.csv
covariate_m_psi_33.csv
covariate_m_theta_22.csv
covariate_m_theta_32.csv
covariate_m_theta_33.csv
covariate_m_theta_os_22.csv
covariate_m_theta_psi_22.csv
covariate_s_psi_22.csv
covariate_s_psi_32.csv
covariate_s_psi_33.csv
covariate_s_theta_psi_22.csv
covariate_s_theta_22.csv
covariate_s_theta_32.csv
covariate_s_theta_33.csv
distinct_m_22.csv
distinct_m_32.csv
distinct_m_33.csv
omnibus_m_22.csv
omnibus_m_32.csv
omnibus_m_33.csv
omnibus_m_44.csv
omnibus_s_22.csv
omnibus_s_32.csv
omnibus_s_33.csv
omnibus_s_44.csv
```

R data files (found within folder 'DataS2-Input_Files_for_Simulations\Likelihood_Equations')

```
likelihood_equations_a2b2g25.RData
likelihood_equations_a3b2g25.RData
likelihood_equations_a3b3g18.RData
likelihood_equations_a4b4g10.RData
```

Description

The .csv format simulation input files define distinct models used in simulation analyses. These files can be used with R computer code in DataS3-R_Code_for_Simulations to reproduce simulation analyses in the companion article or they can be altered to produce user-specified simulation analyses.

The `likelihood_equations` .RData format files contain R list objects '`likelihood.equations`' required to provide pre-computed values for likelihood computations. See DataS3-R_Code_for_Simulations for additional details.

Naming conventions for .csv files

The first word of the name of each file corresponds to the simulations defined in the companion article (i.e., "omnibus", "covariate", and "distinct observer" simulations). The remaining parts of the file names define the specific model structure: "m" and "s" denote multi-observation method (MOM) and single-observation method (SOM) models, names of Greek letters specify covariates predicting true species probabilities ψ (psi) and classification probabilities θ (theta), and the final integers specify the number of observation states (*A*) and true species states (*B*). Input file "covariate_m_theta_os_22.csv" defined analyses assessing effects of un-modeled heterogeneity in classification probabilities for secondary observers (Fig. 2 in companion article).

Naming conventions for .RData format files

The prefix "likelihood_equations_" is followed by integers specifying the number of observation states (*A*), true species states (*B*), and the maximum group size (*g*).

User-specified inputs defining model structure and true parameter values

Each .csv format file defines distinct models that share a common model structure. Rows 5-6 specify features common to all models: model type, number of simulation replicates, and other details of model structure and optimization (Table 1). The optional field 'seed' specifies the value of the random number seed for generating simulation data, allowing replication of simulation results.

Columns on row 10 specify inputs for target sample sizes and true parameter values (naming conventions defined in Table 2; see companion article and supplemental Appendices S1-S3 for definitions of parameters). Inputs for true parameter values (Table 3) can vary among distinct models defined on separate rows below row 10. For distinct models with mean group size of 1 and/or with heterogeneous group parameters of 0, these parameters are omitted from data generation and estimation models.

Order of columns should follow examples found in .csv files. In general, column order from left to right is:

1) target samples sizes for individuals classified by primary and secondary observers (n),
2) heterogeneous group parameters (π or ρ),
3) intercept and slope coefficients for multinomial logistic regression predicting true species probabilities (beta coefficients in $\mathbf{B}_\psi$ in eqs. 3, 4 in the companion article),
4) intercept and slope coefficients for multinomial logistic regression predicting classification probabilities (beta coefficients in $\mathbf{B}_\theta$ in eqs. 3, 5 in the companion article),
5) classification probability parameters (θ),
6) true species probability parameters (ψ),
7) and mean group size parameters ($\overline{\mathbf{g}}_{.b}$).

Parameters for true species probabilities are ordered from true species states 1 to $B - 1$. Parameters for classification probabilities are ordered first by observation states 1 to A, and then by true species states 1 to B within each observation state. Parameters for correct classification are not specified, as these are complementary to other classification probabilities or comprise the baseline category in multinomial logistic regression models.

Table 1. User-specified inputs on rows 5-6 of .csv files define model type, model structure, and details of model optimization.

Parameter name	Values[†]	Description
Model[‡]	'M'	Model without covariates
	'M.psi'	Model with covariate predicting true species probabilities ψ (psi)
	'M.theta'	Model with covariate predicting classification probabilities θ (theta)
	'M.theta.p'	Model with covariate predicting classification probabilities θ (theta) only for primary observers
	'M.theta.psi'	Model with independent covariates predicting classification probabilities θ (theta) and true species probabilities ψ (psi)
	'M.theta+psi'	Model with one covariate predicting both classification probabilities θ (theta) and true species probabilities ψ (psi)
reps	Integer > 1	Number of simulation replicates for each distinct model
A	1 ≤ Integer ≤ 4	Number of observation states. If A = B + 1, observation state A is partial identification.
B	1 ≤ Integer ≤ 4	Number of true species states
O_p	0 ≤ Integer ≤ 1	Number of primary observers
O_s	1 ≤ Integer ≤ 4	Number of secondary observers
mx_model	'constant'	Model for heterogeneous group probabilities as described in eq. 7 in the companion article.
	'encounter'	Model for heterogeneous group affinities (Appendix S3: eqs. S3, S4), where group-level heterogeneous group probabilities vary relative to group-level true species probabilities
n_bins	3 ≤ Integer ≤ 50	Number of equal-interval bins (15-20 usually work well) for values of covariates in multinomial logistic regression models. For all other values, covariate values are continuous.
n_bootstrap	Integer > 2	Number of non-parametric bootstrap resamples applied to each simulation replicate to estimate standard errors and confidence intervals for parameters derived from estimated regression coefficients of multinomial logistic regression models. Reasonably precise SEs typically require 30+ resamples.
t_limit	0	No upper constraints for estimated classification probabilities θ (theta)
	0 < Numeric < 1	Upper constraint for estimated classification probabilities θ (theta) is fixed probability 't_limit'
	'theta + 0.1'	Upper constraints for estimated classification probabilities θ (theta) are defined relative to true values by an expression. Here, constraints are each true value + 0.1. Other expressions may be used (e.g., 'theta * 1.5').
seed	integer > 1	Value of random number seed for generating simulation data (optional)

[†]Quotes denote character strings. All other values are integer or numeric within specified range.
[‡]The same model specification values are used for MOM and SOM models. SOM models are specified by omitting classification probability parameters for primary observers (Table 2).

Table 2. Naming conventions for parameters defined in row 10 of .csv format files. Parameters have a base name (bold) specifying the type of parameter (column 1). Suffix(es) are appended to denote observation state, true species state(s), and observer identities. For multinomial logistic regression coefficients predicting true species probabilities or classification probabilities, a prefix denotes intercept (b0) versus slope (b1) coefficients. Suffixes and prefixes are separated from the base name and each other by an underscore. Parameters are described in the companion article and supplemental Appendices S1-S3.

Parameter type (notation)	Name	Prefix	Suffix
Target sample size $n^{\dagger}$	**n**_[observer]		Text denoting primary 'p' or secondary 's' observers
Heterogeneous groups π or ρ	**mix**_[*bb*]		Two integers denoting true species states *b*
Classification probabilities $\theta^{\ddagger}$	[beta]_**theta**_[observer]_[*ab*]	Text 'b0' or 'b1' denotes intercept or slope regression coefficients for multinomial logistic regression predicting classification probabilities	1) [observer] Text denotes primary 'p' or secondary observers 's1', 's2', . . . 2) [*ab*] Two integers denote classification state *a* and true species state *b*
True species probabilities ψ	[beta]_**psi**_[*b*]	Text 'b0' or 'b1' denotes intercept or slope regression coefficients for multinomial logistic regression predicting true species probabilities	Integer denoting true species state *b*
Mean group size $\bar{g}$	**g**_[*b*]		Integer denoting true species state *b*

†Target sample size for secondary observers must be ≤size for primary observers.
‡If only one secondary observer is specified, secondary observers are assumed identical (i.e., to have the same classification probabilities). If no primary observer is specified, data generation and analyses are for SOM models rather than MOM models.

Table 3. User-specified inputs on rows 11+ of .csv files that define true parameter values.

Parameter (notation)	Parameter column name	Values	Comments
Heterogeneous group probability π_{wx}	mix_[*bb*]	0 < Numeric < 1	For each species, heterogeneous group probabilities should not exceed the proportion of their groups among all groups (δ_b in eq. 7 in the companion article)
Heterogeneous group affinity ρ	mix_[*bb*]	0 < Numeric < 1	Used for models including a covariate predicting true species probabilities. See Appendix S3: eqs. S3 and S4.
Classification probabilities θ	theta_[observer]_[*ab*]	0 < Numeric < 1[†]	For each observer *o* and true species state *b*, summed classification probabilities across *A* observation states should not exceed one ($\sum_a \theta_o^{a\|b} \leq 1$).
Classification regression coefficients $\beta^{v\|b}$	b0_theta_[observer]_[*ab*]	Numeric	
	b1_theta_[observer]_[*ab*]	Numeric	
True species probabilities ψ	psi_[*b*]	0 < Numeric < 1	Summed true species probabilities should not exceed one ($\sum_b \psi_b \leq 1$).
True species regression coefficients β_v	b0_psi_[*b*]	Numeric	
	b1_psi_[*b*]	Numeric	
Group size $\bar{g}$	g_[*b*]	Numeric ≥ 1	Maximum sizes of individual groups are limited by maximum group size supported by `likelihood_equations` .RData files.

[†]To specify a model with certain identification by a secondary observer, include a single secondary observer with a value of 0 for all classification probabilities. Supported for models ‘M’, ‘M.psi’, and ‘M.theta.p’ (see Appendix S2: Table 1) without un-modeled heterogeneity in parameters.