



## **unmarked:** An R package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance

**Ian J. Fiske**

North Carolina State University

**Richard Chandler**

USGS Patuxent Wildlife Research Center

---

### **Abstract**

Ecological research uses data collection techniques that are prone to substantial and unique types of measurement error to address scientific questions about species abundance and distribution. These data collection schemes include a number of survey methods in which unmarked individuals are counted, or determined to be present, at spatially-referenced sites. Examples include site occupancy sampling, repeated counts, distance sampling, removal sampling, and double observer sampling. To appropriately analyze these data, hierarchical models have been developed to separately model explanatory variables of both a latent abundance or occurrence process and a conditional detection process. Because these models have a straightforward interpretation paralleling mechanisms under which the data arose, they have recently gained immense popularity. The common hierarchical structure of these models is well-suited for a unified modeling interface. The R package **unmarked** provides such a unified modeling framework, including tools for data exploration, model fitting, model criticism, post-hoc analysis, and model comparison.

*Keywords:* ecological, wildlife, hierarchical, occupancy, occurrence, distance, point count.

---

## **1. Introduction**

### **1.1. Imperfect detection and data collection in ecological research**

A fundamental goal of ecological research is to understand how environmental variables influence spatial or temporal variation in species abundance or occurrence. Addressing these research questions is complicated by imperfect detection of individuals or species. Individuals may go undetected when present for a variety of reasons including their proximity to the ob-

server, cryptic behavior, or camouflage. Thus, imperfect detection can introduce substantial measurement error and obscure underlying ecological relationships if ignored.

To accommodate imperfect detection of individuals or species, ecologists have developed specialized methods to survey wildlife populations such as site occupancy sampling, repeated counts, distance sampling, removal sampling, and double observer sampling (see Section 2 and (Williams *et al.* 2002) for definitions). This myriad of sampling methods are all unified by a common repeated-measures type of sampling design in which  $J$  observations are made at  $M$  spatial sample units during each of  $T$  seasons. As an example, Table 1.1 shows a typical repeated count dataset.

	Season 1			Season 2		
	Visit 1	Visit 2	Visit 3	Visit 1	Visit 2	Visit 3
Site 1	3	3	2	0	0	0
Site 2	0	0	0	0	0	0
Site 3	4	6	5	5	5	3
Site 4	0	0	0	0	0	2

Table 1: An example of the data structure required by **unmarked**’s fitting functions. Counts of organisms were made at  $M = 4$  sites during  $T = 2$  seasons with  $J = 3$  visits per season.

In addition to the repeated-measures structure, there are two important features to note about the data shown in Table 1.1. First, this design is often referred to as a ‘metapopulation design’ because the population may be regarded as an aggregation of subpopulations and because the spatial sampling is an explicit component of the problem. This is important because it provides a basis for modeling variation among sites as a function of site-specific covariates, and because metapopulation parameters such as local colonization and extinction can be directly estimated. Second, the counts are of individuals that may not be uniquely recognized, so although it may be possible to determine if three individuals were detected on two different visits (as was the case at Site 1 in Table 1.1), it is not always possible to determine if they were the *same* three individuals.

Both the metapopulation design and absence of individual recognition distinguish these data from data collected using another large suite of ecological sampling methods known as ‘capture-recapture’ methods. A vast number of models have been developed for capture-recapture data, and many of them are implemented in the software program *MARK* (White and Burnham 1999). The origins of **unmarked** (the package and its name) arose to fill the void of models for data from studies of unmarked individuals involving explicit spatial sampling.

## 1.2. Hierarchical Models for Metapopulation Designs

While many ecological studies produce data consistent with a metapopulation design, the development and implementation of models for inference under individual sampling protocols has proceeded in a piecemeal fashion, without a coherent analysis platform. Recently, however, a broad class of hierarchical models (see Royle and Dorazio (2008) for a general treatment) has been developed that offer a unified framework for analysis by formally recognizing that observations are generated by a combination of (1) a *state* process determining abundance or species occurrence at each site and (2) a *detection* process that yields observations conditional on the state process. The model for the state process describes abundance or occurrence at

each site, but due to imperfect detection, these quantities cannot be observed directly and are regarded latent variables.

This paper introduces **unmarked**, an R package that provides a unified approach for fitting this broad class of hierarchical models developed for sampling biological populations. Inference in **unmarked** is based on the integrated likelihood wherein the latent state variable is marginalized out of the conditional likelihood.

### 1.3. Scope and features of **unmarked**

**unmarked** provides tools to assist researchers with every step of the analysis process, including data manipulation and exploration, model fitting, post-hoc analysis, model criticism, and model selection. **unmarked** provides a growing list of model-fitting functions designed for specific sampling methods. The fitting functions each find the maximum likelihood estimates of parameters from a particular model (Section 3.3) and return an object that can be easily manipulated. Methods exist for performing numerous post-hoc analyses such as requesting linear combinations of parameters, back-transforming parameters to constrained scales, determining confidence intervals, and evaluating goodness of fit. The model specification syntax of the fitting functions was designed to resemble the syntax of R's common fitting functions such as **lm** for fitting linear models.

Although there is existing software for fitting some of these models (e.g. [Hines and MacKenzie \(2002\)](#)), there are a number of advantages to a unified framework within R. Many researchers are already familiar with R and use its powerful data manipulation and plotting capabilities. Sometimes many species are analyzed in tandem, so that a common method of aggregating and post-processing of results is needed, a task easily accomplished in R. Unlike other available software, **unmarked** makes it possible to map habitat-specific abundance and species distributions when combined with R's GIS capabilities. Another important advantage of **unmarked**'s approach is that researchers can simulate and analyze data within the same computational environment. This work flow permits simulation studies for power analysis calculations or the effectiveness of future sampling designs. All of this is made much simpler by analyzing the data within R, and using a single environment to complete all phases of the analysis is much less error-prone than switching between applications.

In this paper, Section 2 gives a brief summary of many of the models **unmarked** is capable of fitting. Section 3 describes general **unmarked** usage aided by a running data example.

## 2. Models implemented in **unmarked**

The list of models implemented in **unmarked** continues to grow as new models are developed. Table 2 shows the models available as of version 0.9-0. Rather than describe each fitting function in detail, this section provides a summary of several of the most common sampling techniques and how **unmarked** can be used to model the resulting data.

### 2.1. Site occupancy models

#### *Single season site occupancy model*

An important state variable in ecological research is species occurrence (or 'site occupancy'),

Model	Fitting function	Data	Citation
Occupancy from detection/non-detection	<code>occu</code>	<code>unmarkedFrameOccu</code>	MacKenzie <i>et al.</i> 2002
Abundance from detection/non-detection	<code>occuRN</code>	<code>unmarkedFrameOccu</code>	Royle and Nichols 2003
Abundance from repeated counts	<code>pcount</code>	<code>unmarkedFramePCount</code>	Royle 2004b
Abundance from distance sampling	<code>distsamp</code>	<code>unmarkedFrameDS</code>	Royle <i>et al.</i> 2004
Abundance from multinomial counts	<code>multinomPois</code>	<code>unmarkedFrameMPois</code>	Royle 2004a
Repeated multinomial counts	<code>gmultmix</code>	<code>unmarkedFrameGMM</code>	Chandler <i>et al.</i> In press
Multi-season occupancy	<code>colext</code>	<code>unmarkedMultFrame</code>	MacKenzie <i>et al.</i> 2003
Multi-season abundance	<code>pcountOpen</code>	<code>unmarkedFramePCO</code>	Dail and Madsen 2011

Table 2: Models currently handled by *unmarked* along with their associated fitting functions (Section 2) and data type (Section 3.1).

say  $Z_i$ , a binary state variable such that  $Z_i = 1$  if site  $i$  is occupied by a species and  $Z_i = 0$  otherwise. Much interest is focused on estimating functions of species occurrence (e.g., number of occupied sites) or to identify factors that are associated with changes in the probability of a site being occupied, i.e.,  $\psi_i = \Pr(Z_i = 1)$ . To estimate these parameters, researchers employ a sampling design, whereby surveyors visit a sample of  $M$  sites and record the binary response  $Y_{ij}$  of species detection ( $Y = 1$ ) or non-detection ( $Y = 0$ ) during  $j = 1, \dots, J_i$  visits to the  $i$ th site during a ‘season’ (MacKenzie *et al.* 2002). A key feature of species occurrence surveys is that false absences are possible, so that a species might go undetected ( $Y_{ij} = 0$ ) even if it is present ( $Z_i = 1$ ). A set of parameters  $p_{ij}$  accounts for this observation process, where  $p_{ij}$  is the probability of detecting the species at site  $i$  during visit  $j$  given that site  $i$  is truly occupied. That is,  $p_{ij} = \Pr(Y_{ij} = 1 | Z_i = 1)$ .

The key assumptions made when modeling these data are that the occupancy state at a site remains constant throughout the season and repeated visits at a site are independent. The first assumption is commonly referred to as the ‘population closure’ assumption, meaning that no births, deaths, emigrations, or immigrations occur during the season. Season, therefore, will generally refer to a very short time frame, such as a few months during a breeding season, or even a few minutes if repeated visits are made in quick succession. Replicate samples (i.e.,  $J > 1$ ) are necessary to obtain information about the detection rate separate from the occupancy rate.

The following hierarchical model describes the joint distribution of the observations conditional on the latent occupancy state, and the marginal distribution of the latent occupancy state variable:

$$\begin{aligned} Z_i &\sim \text{Bernoulli}(\psi_i) \quad \text{for } i = 1, 2, \dots, M \\ Y_{ij}|Z_i &\sim \text{Bernoulli}(Z_i p_{ij}) \quad \text{for } j = 1, 2, \dots, J_i \end{aligned}$$

Removing the latent  $Z$  variables by marginalization yields the likelihood:

$$L(\psi, p | Y_{ij}) = \prod_i^M \left\{ \prod_j^J (p^{Y_{ij}} (1-p)^{1-Y_{ij}}) \psi + I(Y_{i.} = 0)(1-\psi) \right\}, \quad (1)$$

where  $I$  is the indicator function taking the value 1 if its argument is true, and 0 otherwise. The model and likelihood are easily generalizable to accommodate covariates on detection probability or occupancy. Variables that are related to the occupancy state are modeled as

$$\text{logit}(\psi_i) = \mathbf{x}_i' \beta,$$

where  $\mathbf{x}_i$  is a vector of site-level covariates and  $\beta$  is a vector of their corresponding effect parameters. Similarly, the probability of detection can be modeled with

$$\text{logit}(p_{ij}) = \mathbf{v}_{ij}' \alpha,$$

where  $\mathbf{v}_{ij}$  is a vector of observation-level covariates and  $\alpha$  is a vector of their corresponding effect parameters. Examples of site-level covariates include habitat characteristics such as vegetation height. Observation-level covariates could include time of day, date, wind speed, or other factors that might affect detection probability. The function `occu` implements this model.

#### *Multi-season site occupancy model*

Sometimes the study objective is to understand the dynamics of the occupancy state over time. To obtain such information researchers conduct repeated occupancy studies (see Section 2.1.1) at the same sample of sites over consecutive seasons (MacKenzie *et al.* 2003) and seek to estimate probabilities of colonization ( $\gamma_{it}$ ) and extinction ( $\epsilon_{it}$ ), where colonization is the change of an unoccupied site to occupied and extinction is the change of an occupied site to unoccupied. If the occupancy status is assumed to evolve according to a Markov process, then a 2-state finite hidden Markov model describes these data. Let  $Y_{itj}$  denote the observed species occurrence status at visit  $j$  during season  $t$  to site  $i$ . Then

$$\begin{aligned} Z_{i1} &\sim \text{Bernoulli}(\psi_i) \\ Z_{it} &\sim \begin{cases} \text{Bernoulli}(\gamma_{i(t-1)}) & \text{if } Z_{i(t-1)} = 0 \\ \text{Bernoulli}(1 - \epsilon_{i(t-1)}) & \text{if } Z_{i(t-1)} = 1 \end{cases}, \\ &\text{for } t = 2, 3, \dots, T \\ Y_{itj} | Z_{it} &\sim \text{Bernoulli}(Z_{it} p_{itj}) \end{aligned}$$

This model generalizes the single season site occupancy model by relaxing the population closure assumption. To define the likelihood of this model, let  $\phi_0 = (\psi, 1 - \psi)$  and

$$\theta_{\mathbf{Y}_{it}} = \left( \frac{\prod_j^J p I(Y_{ijt} = 0)(1 - p_{ijt})}{I(\sum_j^J Y_{ijt} = 0)} \right)$$

Furthermore, let the matrix of one-step Markov transition probabilities be

$$\Phi_t = \begin{pmatrix} 1 - \epsilon & \epsilon \\ 1 - \gamma & \gamma \end{pmatrix}$$

The likelihood is then

$$L(\psi, \epsilon, \gamma, p | Y_{ijt}) = \prod_i^M \left\{ \phi_0 \left\{ \prod_t^{T-1} \Phi_t \theta_{Y_{it}} \right\} \theta_{Y_{iT}} \right\}, \quad (2)$$

which can be maximized by the `coltext` function in **unmarked**. Again, each of the four model parameters  $\psi$ ,  $\gamma$ ,  $\epsilon$ , and  $p$  can be modeled as functions of covariates on the logit scale.

## 2.2. Abundance models

### *Repeated count data*

Occupancy is a crude summary of population structure and dynamics and, in practice, considerable effort is focused on obtaining information about abundance (Dorazio 2007). A common sampling design for obtaining abundance information, analogous to the basic occupancy design described above, is to repeatedly visit a sample of  $M$  sites  $J$  times and record the number of unique individuals observed at each site. Similar assumptions are made as with occurrence data: (1) abundance at a site remains constant during a season and (2) counts at a site are independent. Royle (2004b) presented the following hierarchical model for repeated count data. Let  $N_i$  be the unobserved total number of individuals using a site and define  $Y_{ij}$  as the number of individuals observed during the  $j$ th visit. Then,

$$\begin{aligned} N_i &\sim f(\lambda_i, \theta) \quad \text{for } i = 1, 2, \dots, M \\ Y_{ij}|N_i &\sim \text{Binomial}(N_i, p_{ij}) \quad \text{for } j = 1, 2, \dots, J_i, \end{aligned}$$

where  $\lambda_i$  is the abundance rate at site  $i$  and  $p_{ij}$  is the detection probability during the  $j$ th visit to site  $i$ .  $f$  is a discrete distribution with support restricted to  $N_i \geq 0$  and  $\theta$  are extra parameters of  $f$  other than the location parameter,  $\lambda_i$ . **unmarked** currently supports  $f$  as Poisson or negative binomial. In the Poisson case, there is no  $\theta$ . In the negative binomial case,  $\theta$  is a dispersion parameter, which is useful when overdispersion is suspected.

As with the occupancy model, covariates may be included at either the state (here, abundance) or detection levels, but abundance is modeled through a log link to enforce its positivity constraint.

$$\log(\lambda_i) = \mathbf{x}_i' \beta,$$

where  $\mathbf{x}_i$  is a vector of site-level covariates and  $\beta$  is a vector of their corresponding effect parameters. Similarly, the probability of detection can be modeled with

$$\text{logit}(p_{ij}) = \mathbf{v}_{ij}' \alpha,$$

where  $\mathbf{v}_{ij}$  is a vector of observation-level covariates and  $\alpha$  is a vector of their corresponding effect parameters.

The  $N_i$  variables are latent and so analysis is carried out based on the integrated likelihood obtained by marginalizing each  $N_i$  from the conditional likelihood (Royle *et al.* 2004):

$$L(\lambda, p|Y_{ij}) = \prod_i^M \left\{ \sum_{N_i=\max(\mathbf{Y}_i)}^{\infty} \left( \prod_j^J \frac{N_i!}{(N_i - Y_{ij})!} p^{Y_{ij}} (1-p)^{N_i - Y_{ij}} \right) \frac{e^{-\lambda} \lambda^{N_i}}{N_i!} \right\}. \quad (3)$$

This likelihood can be maximized using the `pcount` function.

The function `pcountOpen` implements a generalized form of this model designed for ‘open’ populations with temporal dynamics governed by apparent survival and recruitment parameters (Dail and Madsen 2011). Another related model was described by Royle and Nichols

(2003) and can be fit with the `occuRN` function. This model estimates abundance from site occupancy data by exploiting the link between abundance and detection probability.

### General multinomial-Poisson mixture model

Here we discuss a more general class of models that can be customized by the user for a variety of sampling methods, the multinomial-Poisson model (Royle 2004a). The general form of this model is

$$\begin{aligned} N_i &\sim \text{Poisson}(\lambda_i) \quad \text{for } i = 1, 2, \dots, M \\ \mathbf{Y}_i | N_i &\sim \text{Multinomial}(N_i, \boldsymbol{\pi}_i) \end{aligned} \quad (4)$$

where  $N_i$  is the latent abundance at site  $i$  as with the repeated count model, and  $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iJ})'$  is the vector of cell probabilities of corresponding to the vector of counts  $\mathbf{Y}_i$ . In general,  $\boldsymbol{\pi}_i$  is determined by the specific sampling method and  $\sum_j \pi_{ij} \leq 1$  because detection is imperfect and  $\pi_{iJ+1} = 1 - \sum_j \pi_{ij}$  is the probability of the species escaping detection at site  $i$ .

Note that the replication here is of a different nature than in previously described models – not over time necessarily – but still effectively replication as far as the design goes. That is, there are repeated measurements at each site, but with a multinomial protocol, the replicate counts are *dependent* instead of independent.

To illustrate the likelihood under a multinomial observation model, suppose that a sampling method produces a multinomial observation with 3 observable frequencies at each site. As before,  $N_i$  are latent variables, and so inference is based on the integrated likelihood of  $\mathbf{Y}_i$  which is:

$$L(\lambda, p | \mathbf{Y}_i) = \prod_{i=1}^N \left\{ \sum_{N_i = \sum_j (Y_{ij})}^{\infty} \left( \frac{N_i!}{Y_{i1}! Y_{i2}! Y_{i3}! Y_{i0}!} \pi_1^{Y_{i1}} \pi_2^{Y_{i2}} \pi_3^{Y_{i3}} \pi^{*N_i - Y_i} \right) \frac{e^{-\lambda} \lambda^{N_i}}{N_i!} \right\} \quad (5)$$

For this specific case of a multinomial-Poisson mixture, the likelihood simplifies analytically (to a product-Poisson likelihood). The function `multinomPois` can be used to maximize this likelihood.

For multinomial-Poisson sampling methods, the actual observations are an underlying categorical detection variable with  $M \leq J$  levels so that the  $J$ -dimensional  $\mathbf{Y}_i$  is derived from the  $M$ -dimensional raw counts in some sampling method-specific manner. Thus, it is necessary to model the detection at the raw observation level, denoted  $p_{ik}$  for  $k = 1, 2, \dots, M$  at site  $i$ . Then we derive the multinomial cell probabilities  $\boldsymbol{\pi}_i$  through the sampling technique-specific relationship  $\pi_{ij} = g(p_{ik})$  where  $p_{ik}$  is the underlying probability of detection and  $g$  is some sampling method-specific function.

Thus, the only two requirements to adapt **unmarked**'s general multinomial Poisson modeling to a new sampling method is to specify  $g$  and a binary 0-1 matrix that describes the mapping of elements of  $\mathbf{p}_i = (p_{i1}, \dots, p_{iR})'$  to elements of  $\boldsymbol{\pi}_i$ . This mapping matrix, referred to in **unmarked** as `obsToY`, is necessary to consistently clean missing values from the data and relate observation-level covariates with the responses. The  $(j, k)$ th element of `obsToY` is 1 only if  $p_{ik}$  is involved in the computation of  $\pi_{ij}$ . The detection function  $g$  is called `piFun` in **unmarked**.

Covariates may be included in either the state (here, abundance) or detection models, through  $\mathbf{p}_i$  (not  $\boldsymbol{\pi}_i$ ).

$$\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta},$$

where  $\mathbf{x}_i$  is a vector of site-level covariates and  $\boldsymbol{\beta}$  is a vector of their corresponding effect parameters. Similarly, the probability of detection can be modeled with

$$\text{logit}(p_{ij}) = \mathbf{v}_{ij}' \boldsymbol{\alpha},$$

where  $\mathbf{v}_{ij}$  is a vector of observation-level covariates and  $\boldsymbol{\alpha}$  is a vector of their corresponding effect parameters.

We now describe two common sampling methods that can be modeled with the multinomial-Poisson model: removal sampling and double observer sampling. These two methods are included in **unmarked**, but additional methods may easily be specified by defining a user-specified `piFun` function and `obsToY` matrix.

**Removal sampling** Popular in fisheries, removal sampling is commonly implemented by visiting a sample of  $M$  sites  $J$  times each and trapping or otherwise capturing and then removing individuals at each visit. Thus,  $Y_{ij}$  is the number of individuals captured at the  $j$ th visit for  $j = 1, 2, \dots, J$ . We can specify  $g$  for removal sampling as follows. The probability of an individual at site  $i$  being captured on the first visit is  $\pi_{i1} = p_{i1}$ . The probability of capture on the  $j$ th visit is

$$\pi_{ij} = \prod_{k=1}^{j-1} (1 - p_{ik}) p_{ij},$$

for  $j = 2, \dots, J$  and the probability of not being sampled is

$$\pi_{i,J+1} = 1 - \sum_{j=1}^J (\pi_{ij})$$

Thus, the mapping matrix is an  $J \times J$  matrix with ones in the upper triangle,

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

**Double observer sampling** Double observer sampling involves collecting data by a team of two surveyors simultaneously visiting a site. Each observer records a list of detected organisms, and at the end of the survey the two observers attempt to reconcile their counts. If individuals are not uniquely marked, this may be a difficult task in practice; however, assuming that individuals can be distinguished, the data at each site are a vector of length three ( $\mathbf{Y}_i$ ), corresponding to the numbers of individuals seen by only observer one, only observer two, and both observers. Thus, for double observer sampling,  $g$  is defined as follows.

$$\pi_i = \begin{pmatrix} p_{i1}(1 - p_{i2}) \\ (1 - p_{i1})p_{i2} \\ p_{i1}p_{i2} \\ (1 - p_{i1})(1 - p_{i2}) \end{pmatrix}$$



The `obsToY` mapping matrix for double observer sampling is the following  $3 \times 2$  matrix.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

**Distance sampling** One of the most widely used sampling methods in animal ecology is distance sampling (Buckland *et al.* 2001), which involves recording the distance to each individual detected at  $M$  sites (often referred to as ‘transects’) on a single occasion. Detection probability is modeled as a function of distance ( $d$ ) to the observer, for example using the half-normal detection function  $p = \exp(-d^2/(2*\sigma^2))$  where  $\sigma$  is the half-normal shape parameter. In practice, the distance measurements are often binned into  $J$  distance intervals, which allows them to be modeled as multinomial outcomes with cell probabilities  $\pi_i$  computed as the product of the probability of detection and the probability of occurrence in each distance interval (Royle *et al.* 2004).

As currently implemented, the general form of the distance sampling model is identical to the multinomial-Poisson mixture described above, with the sole difference being that instead of

$$\text{logit}(p_{ij}) = \mathbf{v}'_{ij}\alpha, \quad (6)$$

we have

$$\log(\sigma_i) = \mathbf{v}'_i\alpha, \quad (7)$$

Here, the log link is required because  $\sigma$  is a positive shape parameter of the detection function. In addition, distance sampling data are associated with many unique attributes such as distance interval cut-points and survey method (line-transect vs point count); therefore, we created the specialized function `distsamp` to fit the multinomial-Poisson model to distance sampling data.

**Generalization** The flexibility of the `multinomPois` function is extended even further in the `gmmultmix` function, which allows for a negative binomial mixture (Dorazio *et al.* 2005) and relaxes the population closure assumption (Chandler *et al.* In press).

### 3. unmarked usage

**unmarked** provides data structures, fitting syntax, and post-processing that form a cohesive framework for the analysis of ecological data collected using a metapopulation design. In order to achieve these goals, **unmarked** uses the S4 class system (Chambers 2008). As R’s most modern system of class-based programming, S4 allows customization of functions, referred to as methods, to specific object classes and superclasses. For example, when the generic `predict` method is called with any **unmarked** model fit object as an argument, the actual `predict` implementation depends on the specific model that was fit. Use of class-based programming can provide more reliable and maintainable software while also making the program more user-friendly (Chambers 2008).

#### 3.1. Preparing data

**unmarked** uses a custom S4 data structure called the **unmarkedFrame** to store all data and metadata related to a sampling design. Although this at first appears to add an extra layer of work for the user, there are several reasons for this design choice. The multilevel structure of the models means that standard rectangular data structures such as **data.frames** or matrices are not suitable for storing the data. For example, covariates might have been measured separately at the site level and at the visit level. Furthermore, the length of the response vector  $\mathbf{Y}_i$  at site  $i$  might differ from the number of observations at the site as in the multinomial Poisson model. In some cases, metadata of arbitrary dimensions may need to be associated with the data. For example, in distance sampling it is necessary to store the units of measure and the survey design type. Aside from these technical reasons, [Gentleman \(2009\)](#) pointed out that the use of such portable custom data objects can simplify future reference to previous analyses, an often neglected aspect of research. Repeated fitting calls using the same set of data require less code repetition if all data are contained in a single object. Finally, calls to fitting functions have a cleaner appearance with a more obvious purpose when the call is not buried in data arguments. The parent data class is called an **unmarkedFrame** and each **unmarked** fitting function has its own data type that extends the **unmarkedFrame**. To ease data importing and conversion, **unmarked** includes several helper functions to automatically convert data into an **unmarkedFrame**: **csvToUMF** which imports data directly from a comma-separated value text file, **formatWide** and **formatLong** which convert data from data frames, and the family of **unmarkedFrame** constructor functions.

An **unmarkedFrame** object contains components, referred to as slots, which hold the data and metadata. All **unmarkedFrame** objects contain a slot for the observation matrix **y**, a **data.frame** of site-level covariates **siteCovs**, and a **data.frame** of observation-level covariates **obsCovs**. The **y** matrix is the only required slot. Each row of **y** contains either the observed counts or presence-absence data at each of the  $M$  sites. **siteCovs** is an  $M$ -row **data.frame** with a column for each site-level covariate. **obsCovs** is an  $MJ$ -row **data.frame** with a column for each observation-level covariate. Thus each row of **obsCovs** corresponds to a particular observation, with the order corresponding to site varying slower and observation within site varying faster. Both **siteCovs** and **obsCovs** can contain NA values corresponding to unbalanced or missing data. If any terms in the model specification are missing for that observation, **unmarked** automatically removes observations. **unmarked** provides constructor functions to make creating **unmarkedFrames** straightforward. For each specific data type, specific types of **unmarkedFrames** extend the basic **unmarkedFrame** to handle model-specific nuances.

### *Importing repeated count data*

Here is an example of creating an **unmarkedFrame** for repeated count data (Section 2.2.1). First, load Mallard (*Anas platyrhynchos*) point count dataset described in [Kéry et al. \(2005\)](#).

```
> library(unmarked)
> data(mallard)
```

Loading the mallard data makes three objects available within the R workspace. The matrix **mallard.y** contains the number of mallards counted at each of  $M = 239$  sites (rows) on  $J = 3$  visits (columns). The site-level covariates are columns of the **mallard.site** **data.frame**, which has  $M$  rows. The observation-level covariates are a **list** with separate  $M \times J$  matrices

for each observation-level covariate. The `unmarkedFrame` constructors can accept `obsCovs` in this `list` format or as a `data.frame` in the format described in Section 3.1.

The following call to `unmarkedFramePCount` organizes the observations and covariates into an object that can be passed to the `data` argument of the fitting function `pcount`.

```
> mallardUMF <- unmarkedFramePCount(y = mallard.y, siteCovs = mallard.site,
+   obsCovs = mallard.obs)
```

Printing `unmarkedFrames` shows them as `data.frames` to make it easier to verify that the data were converted correctly. Here only the first five sites (rows) and the first two visits are shown.

```
> mallardUMF[1:5, 1:2]
```

Data frame representation of `unmarkedFrame` object.

	y.1	y.2	elev	length	forest	ivel.1	ivel.2	date.1	date.2
1	0	0	-1.173	0.801	-1.156	-0.506	-0.506	-1.761	0.310
2	0	0	-1.127	0.115	-0.501	-0.934	-0.991	-2.904	-1.047
3	3	2	-0.198	-0.479	-0.101	-1.136	-1.339	-1.690	-0.476
4	0	0	-0.105	0.315	0.008	-0.819	-0.927	-2.190	-0.690
5	3	0	-1.034	-1.102	-1.193	0.638	0.880	-1.833	0.167

The first two columns contain the count data. The site-level covariates are elevation (`elev`), transect length (`length`), and the proportion of forest covering the site (`forest`). The two observation-level covariates are a measure of survey effort (`ivel`) and the date of the survey (`date`). Note that these variables are standardized to mean of 0 and unit variance. We can also check data contents with a quick summary.

```
> summary(mallardUMF)
```

unmarkedFrame Object

239 sites

Maximum number of observations per site: 3

Mean number of observations per site: 2.76

Sites with at least one detection: 40

Tabulation of y observations:

0	1	2	3	4	7	10	12	<NA>
576	54	11	9	6	1	1	1	58

Site-level covariates:

	elev	length	forest
Min.	:-1.436e+00	Min. :-4.945e+00	Min. :-1.265e+00
1st Qu.	:-9.565e-01	1st Qu.:-5.630e-01	1st Qu.:-9.560e-01
Median	:-1.980e-01	Median : 4.500e-02	Median :-6.500e-02

Mean	:-4.603e-05	Mean	:-2.929e-05	Mean	: 6.695e-05
3rd Qu.:	9.940e-01	3rd Qu.:	6.260e-01	3rd Qu.:	7.900e-01
Max.	: 2.434e+00	Max.	: 2.255e+00	Max.	: 2.299e+00

Observation-level covariates:

	ivel		date
Min.	:-1.753e+00	Min.	:-2.904e+00
1st Qu.:	-6.660e-01	1st Qu.:	-1.119e+00
Median	:-1.390e-01	Median	:-1.190e-01
Mean	: 1.504e-05	Mean	: 7.259e-05
3rd Qu.:	5.490e-01	3rd Qu.:	1.310e+00
Max.	: 5.980e+00	Max.	: 3.810e+00
NA's	: 5.200e+01	NA's	: 4.200e+01

The summary reveals that only 40 sites have at least one detection. The tabulation of y observations provides additional evidence of sparse counts, with no mallards being detected during 576 of the surveys. The 58 NA values correspond to missing data.

### *Importing removal sampling data*

To illustrate the slightly different syntax for removal sampling data example, we will import data from a removal survey of ovenbirds (*Seiurus aurocapillus*) described by [Royle \(2004a\)](#). The data consist of a list named `ovendata.list` with a matrix named `data` containing the removal counts for 4 visits and a `data.frame` called `covariates` containing site-level covariates information.

```
> data(ovendata)
```

This loads a list named `ovendata.list` with components for the observed count data and covariates. The only additional specification required when creating an `unmarkedFrame` for the multinomial-Poisson model is to specify the particular type of data as either `removal`, `double`, or `userDefined`.

```
> ovenFrame <- unmarkedFrameMPois(ovendata.list$data, siteCovs = ovendata.list$covariates,
+   type = "removal")
> head(ovenFrame, 5)
```

Data frame representation of `unmarkedFrame` object.

	y.1	y.2	y.3	y.4	site	ufp	trba
1	0	0	0	0	CAT003	23.43	62.5
2	1	0	0	0	CAT004	15.62	70.0
3	0	0	0	0	CAT011	35.54	90.0
4	0	0	0	0	CAT013	25.78	60.0
5	0	0	0	0	CAT017	25.00	122.5

### *Preparing multi-season occupancy data*

The data structures are more complex when surveys were conducted over more than one

season. The following toy examples demonstrates how to prepare data for the `colex` fitting function if there were only four sites, two visits per season, and three seasons.

```
> M <- 4
> J <- 2
> T <- 3
> y <- matrix(0:1, nrow = M, ncol = J * T)
> rownames(y) <- paste("site", 1:M, sep = "")
> cn <- paste("season", rep(1:2, each = 3), sep = "")
> cn <- paste(cn, paste("v", rep(1:3, 2), sep = ""), sep = "")
> colnames(y) <- cn
> y
```

	season1v1	season1v2	season1v3	season2v1	season2v2	season2v3
site1	0	0	0	0	0	0
site2	1	1	1	1	1	1
site3	0	0	0	0	0	0
site4	1	1	1	1	1	1

```
> umf <- unmarkedMultFrame(y = y, numPrimary = T)
```

The function `unmarkedMultFrame` accepts `siteCovs` and `obsCovs` like all other `unmarkedFrame` classes, and it also has an argument `yearlySiteCovs` that can accept a list of  $M \times T$  `data.frames` containing season-level covariates. The `numPrimary` argument specifies the number of seasons.

### 3.2. Manipulating unmarkedFrames

The various components of `unmarkedFrames` can be extracted and subsetted in a manner similar to the methods used to manipulate standard R objects. Subsetting can be accomplished using the ‘bracket’ notation. For example, the first five rows of data and all three removal occasions can be extracted from the ovenbird removal study using

```
> ovenFrame[1:5, 1:3]
```

Data frame representation of `unmarkedFrame` object.

	y.1	y.2	y.3	site	ufp	trba
1	0	0	0	CAT003	23.43	62.5
2	1	0	0	CAT004	15.62	70.0
3	0	0	0	CAT011	35.54	90.0
4	0	0	0	CAT013	25.78	60.0
5	0	0	0	CAT017	25.00	122.5

In some cases, the covariate data may need to be manipulated after the `unmarkedFrame` has been created. The following code demonstrates how to extract the site-level covariates, standardize those that are continuous (columns 2 and 3), and reinsert them back into the `unmarkedFrame`.

```
> sc <- siteCovs(ovenFrame)
> sc[, 2:3] <- scale(sc[, 2:3])
> siteCovs(ovenFrame) <- sc
```

### 3.3. Fitting models

As introduced in Section 2, each type of survey design has a corresponding fitting function. For example, to fit a repeated count model, we call `pcount`. Table 2 provides a summary of all models that **unmarked** currently fits. With the exception of the ‘open’ population models (`colext`, `gmultmix`, and `pcountOpen`), all fitting functions use a double right-hand sided formula syntax representing the hierarchical model structure. Specifically, covariates affecting the detection process are specified following the first tilde, and covariates of the state process follow the second tilde. No left-hand side of the formula is specified because the `unmarkedFrame` defines the response variable uniquely as the `y` slot.

#### *Fitting a repeated count model*

Continuing the Mallard example, the following call to `pcount` fits a binomial-Poisson mixture model (Section 2.2.1). The following code specifies that detection probability  $p$  should be modeled by day of year, including a quadratic term. We also wish to model abundance using elevation and proportion of area forested. As described in Section 2.2.1, covariates of detection are on the logit-scale and covariates of abundance are on the log scale for the repeated count model.

```
> fm.mallard.1 <- pcount(~date + I(date^2) ~ elev + forest,
+   data = mallardUMF, K = 50)
> fm.mallard.1
```

Call:

```
pcount(formula = ~date + I(date^2) ~ elev + forest, data = mallardUMF,
       K = 50)
```

Abundance:

	Estimate	SE	z	P(> z )
(Intercept)	-1.853	0.236	-7.87	3.62e-15
elev	-1.232	0.234	-5.27	1.36e-07
forest	-0.756	0.165	-4.60	4.32e-06

Detection:

	Estimate	SE	z	P(> z )
(Intercept)	0.3071	0.2149	1.4290	0.15301
date	-0.4091	0.1482	-2.7612	0.00576
I(date^2)	-0.0077	0.0853	-0.0903	0.92801

AIC: 521.2001

This initial fit suggests that Mallard abundance decreases with increasing elevation and forest. It also looks like a linear model might suffice for the detection model, so we subsequently fit the linear detection model as follows:

```
> fm.mallard.2 <- pcount(~date ~ elev + forest, data = mallardUMF,
+      K = 50)
> fm.mallard.2
```

Call:

```
pcount(formula = ~date ~ elev + forest, data = mallardUMF, K = 50)
```

Abundance:

	Estimate	SE	z	P(> z )
(Intercept)	-1.857	0.232	-7.99	1.37e-15
elev	-1.236	0.230	-5.38	7.59e-08
forest	-0.756	0.164	-4.60	4.23e-06

Detection:

	Estimate	SE	z	P(> z )
(Intercept)	0.298	0.188	1.58	0.113983
date	-0.401	0.114	-3.51	0.000449

AIC: 519.2081

This seems to be a better model according to both the Wald p-value and AIC. The result suggests that detection probability decreases during the course of a year.

### *Fitting a multinomial-Poisson model*

Here we demonstrate fitting a multinomial-Poisson mixture model to removal sampling data. The Ovenbird data has no observation-level covariates, so detection probability is assumed constant across visits. It is not necessary to specify that removal sampling was used when fitting the model because this information is already stored in the `ovenFrame` data. We model abundance as a function of understory forest coverage (`ufp`) and average basal tree area (`trba`).

```
> fm.oven.1 <- multinomPois(~1 ~ ufp + trba, ovenFrame)
> fm.oven.1
```

Call:

```
multinomPois(formula = ~1 ~ ufp + trba, data = ovenFrame)
```

Abundance:

	Estimate	SE	z	P(> z )
(Intercept)	0.102	0.119	0.864	0.388
ufp	0.100	0.126	0.794	0.427
trba	-0.171	0.135	-1.262	0.207

Detection:

Estimate	SE	z	P(> z )
0.288	0.233	1.24	0.217

AIC: 326.1387

### 3.4. Examining fitted models

Objects returned by **unmarked**'s fitting functions also make use of the S4 class system. All fitted model objects belong to the **unmarkedFit** parent class. Thus, common operations such as extracting coefficient estimates, covariance matrices for estimates, and confidence intervals have been adapted to behave similar to R's base functions.

For example, we can extract estimated coefficients either from the entire model, or from the state or detection levels by specifying the **type** argument.

```
> coef(fm.mallard.2)

      lam(Int)   lam(elev) lam(forest)      p(Int)      p(date)
-1.8565048  -1.2355497  -0.7555178   0.2977558  -0.4006066

> coef(fm.mallard.2, type = "state")

      lam(Int)   lam(elev) lam(forest)
-1.8565048  -1.2355497  -0.7555178
```

To check which types are available for a model, use the **names** method.

```
> names(fm.mallard.2)

[1] "state" "det"
```

Similarly, the **vcov** function extracts the covariance matrix of the estimates, using the observed Fisher information by default. **vcov** also accepts a **type** argument, as does the convenience method **SE**, which returns standard errors from the square root of the diagonal of the covariance matrix.

```
> vcov(fm.mallard.2, type = "det")

      p(Int)      p(date)
p(Int) 0.03549024 0.00344853
p(date) 0.00344853 0.01302865
```

Extracting confidence intervals proceeds in a similar fashion. By default, the asymptotic normal approximation is used.



```
> confint(fm.oven.1, type = "state", level = 0.95)
```

```

              0.025      0.975
lambda(Int) -0.1299722 0.33470834
lambda(ufp)  -0.1471741 0.34776195
lambda(trba) -0.4361311 0.09440937
```

Profile confidence intervals are also available upon request. This can take some time, however, because for each parameter, a nested optimization within a root-finding algorithm is being used to find the profile limit.

```
> ci <- confint(fm.oven.1, type = "state", level = 0.95,
+   method = "profile")
```

```

Profiling parameter 1 of 3 ... done.
Profiling parameter 2 of 3 ... done.
Profiling parameter 3 of 3 ... done.
```

```
> ci
```

```

              0.025      0.975
lambda(Int) -0.1390786 0.32676614
lambda(ufp)  -0.1477724 0.34811770
lambda(trba) -0.4368444 0.09469605
```

The profile confidence intervals and normal approximations are quite similar here.

### *The nonparametric bootstrap*

Nonparametric bootstrapping can also be used to estimate the covariance matrix. **unmarked** implements a two-stage bootstrap in which the sites are first drawn with replacement, and then within each site, the observations are drawn with replacement. First, bootstrap draws must be taken using the **nonparboot** function, which returns a new version of the **unmarkedFit** object with additional bootstrap sampling information. Thus, this new fit must be stored, either in a new fit object or the same one, and then subsequently queried for bootstrap summaries. To illustrate, we use the removal sampling data instead of the repeated count data because computations are much faster; however, bootstrapping is available for any of the models in **unmarked**.

```
> set.seed(1234)
> fm.oven.1 <- nonparboot(fm.oven.1, B = 100)

> SE(fm.oven.1, type = "state")
```

```

lambda(Int)  lambda(ufp)  lambda(trba)
  0.1185431    0.1262615    0.1353444
```

```
> SE(fm.oven.1, type = "state", method = "nonparboot")
```

```
lambda(Int) lambda(ufp) lambda(trba)
0.1256949    0.1156153    0.1197142
```

The bootstrapping and asymptotic standard errors are similar. Additional bootstrap samples can be drawn by calling `nonparboot` again.

```
> fm.oven.1 <- nonparboot(fm.oven.1, B = 100)
```

### *Linear combinations of estimates*

Often, meaningful hypotheses can be addressed by estimating linear combinations of coefficient estimates. Linear combinations of coefficient estimates can be requested with `linearComb`. Continuing the Ovenbird example, the following code estimates the log-abundance rate for a site with `ufp = 0.5` and `trba = 0`.

```
> (lc <- linearComb(fm.oven.1, type = "state", coefficients = c(1,
+ 0.5, 0)))
```

Linear combination(s) of Abundance estimate(s)

```
Estimate    SE (Intercept) ufp trba
0.153 0.130          1 0.5    0
```

Multiple sets of coefficients may be supplied as a design matrix. The following code requests the estimated log-abundance for sites with `ufp = 0.5` and `trba = 1`.

```
> (lc <- linearComb(fm.oven.1, type = "state", coefficients = matrix(c(1,
+ 0.5, 0, 1, 1, 0), 2, 3, byrow = TRUE)))
```

Linear combination(s) of Abundance estimate(s)

```
Estimate    SE (Intercept) ufp trba
1    0.153 0.130          1 0.5    0
2    0.203 0.166          1 1.0    0
```

Standard errors and confidence intervals are also available for linear combinations of parameters. By requesting nonparametric bootstrapped standard errors, **unmarked** uses the samples that were drawn earlier.

```
> SE(lc, method = "nonparboot")
```

```
[1] 0.1354092 0.1659877
```

*Back-transforming linear combinations of coefficients*

Estimates of linear combinations back-transformed to the native scale are likely to be more interesting than the direct linear combinations. For example, the logistic transformation is applied to estimates of detection rates, resulting in a probability bound between 0 and 1. This is accomplished with the `backTransform`. Standard errors of back-transformed estimates are estimated using the delta method. Confidence intervals are estimated by back-transforming the confidence interval of the original linear combination.

```
> (btlc <- backTransform(lc))
```

```
Backtransformed linear combination(s) of Abundance estimate(s)
```

	Estimate	SE	LinComb	(Intercept)	ufp	trba
1	1.16	0.151	0.153		1 0.5	0
2	1.22	0.203	0.203		1 1.0	0

```
Transformation: exp
```

```
> SE(btlc)
```

```
[1] 0.1510141 0.2032272
```

```
> confint(btlc)
```

	0.025	0.975
1	0.9033915	1.501747
2	0.8846296	1.695386

**3.5. Model selection**

**unmarked** performs AIC-based model selection for structured lists of `unmarkedFit` objects. To demonstrate, we fit a few more models to the Ovenbird removal data, including an interaction model, two models with single predictors, and a model with no predictors.

```
> fm.oven.2 <- update(fm.oven.1, formula = ~1 ~ ufp * trba)
> fm.oven.3 <- update(fm.oven.1, formula = ~1 ~ ufp)
> fm.oven.4 <- update(fm.oven.1, formula = ~1 ~ trba)
> fm.oven.5 <- update(fm.oven.1, formula = ~1 ~ 1)
```

Next, we organize the fitted models with the `fitList` function and then use the `modSel` method to rank the models by AIC.

```
> fmList <- fitList(`lam(ufp+trba)p(.)` = fm.oven.1, `lam(ufp*trba)p(.)` = fm.oven.2,
+   `lam(ufp)p(.)` = fm.oven.3, `lam(trba)p(.)` = fm.oven.4,
+   `lam(.)p(.)` = fm.oven.5)
> modSel(fmList)
```

	nPars	AIC	delta	AICwt	cumltvWt
lam(trba)p(.)	3	324.77	0.00	0.35	0.35
lam(ufp)p(.)	3	325.73	0.96	0.21	0.56
lam(ufp+trba)p(.)	4	326.14	1.37	0.17	0.73
lam(.)p(.)	2	326.28	1.51	0.16	0.90
lam(ufp*trba)p(.)	5	327.17	2.40	0.10	1.00

It looks like the best model includes only tree basal area as a predictor of abundance. We can examine this relationship using the `predict` method (Figure 1).

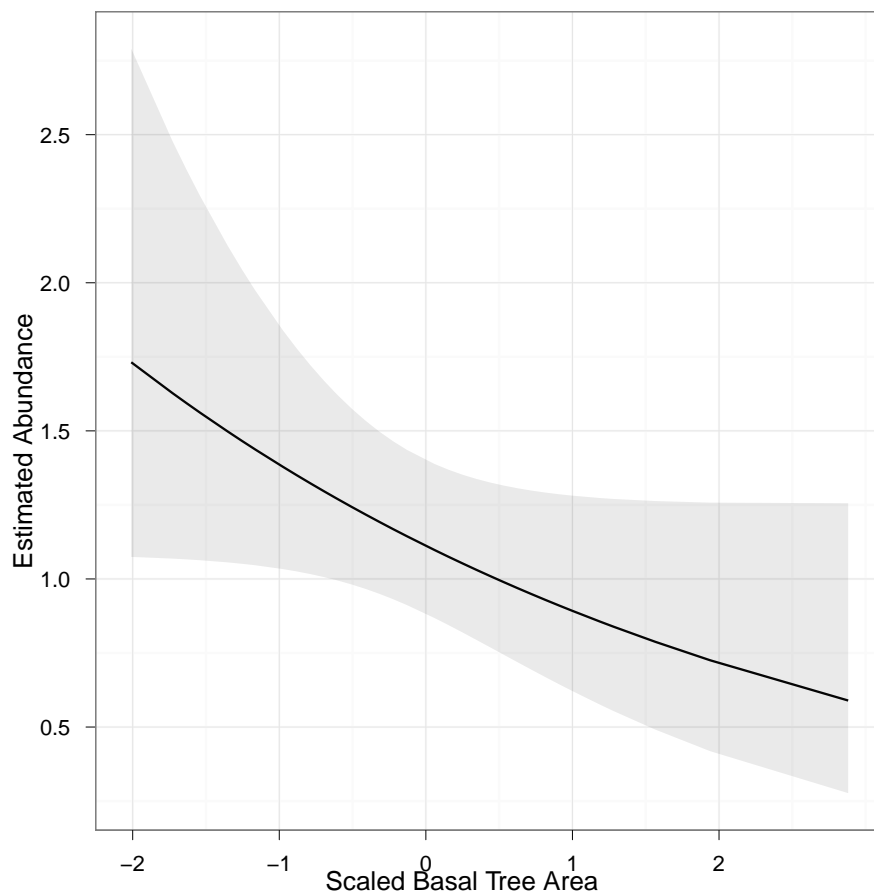


Figure 1: Examine estimated abundance for Ovenbird removal data. Band is 95% confidence interval.

```
> preddata <- predict(fm.oven.4, type = "state", appendData = TRUE)
> library(ggplot2)
> qplot(trba, Predicted, data = preddata, geom = "line",
+       xlab = "Scaled Basal Tree Area", ylab = "Estimated Abundance") +
+   geom_ribbon(aes(x = trba, ymin = lower, ymax = upper),
+             alpha = 0.1) + theme_bw()
```

`predict` functions much like `linearComb` except that new data can be passed to it as a `data.frame` rather than a design matrix. When the first argument given to `predict` is a list of models created by `fitList`, `predict` computes model-averaged predictions, which may be useful in the presence of high model selection uncertainty.

### 3.6. Goodness of fit and the parametric bootstrap

To conduct goodness of fit tests, **unmarked** provides a generic parametric bootstrapping function `parboot`. It simulates data from the fitted model and applies a user-defined function that returns a fit-statistic such as the Pearson's  $\chi^2$ .

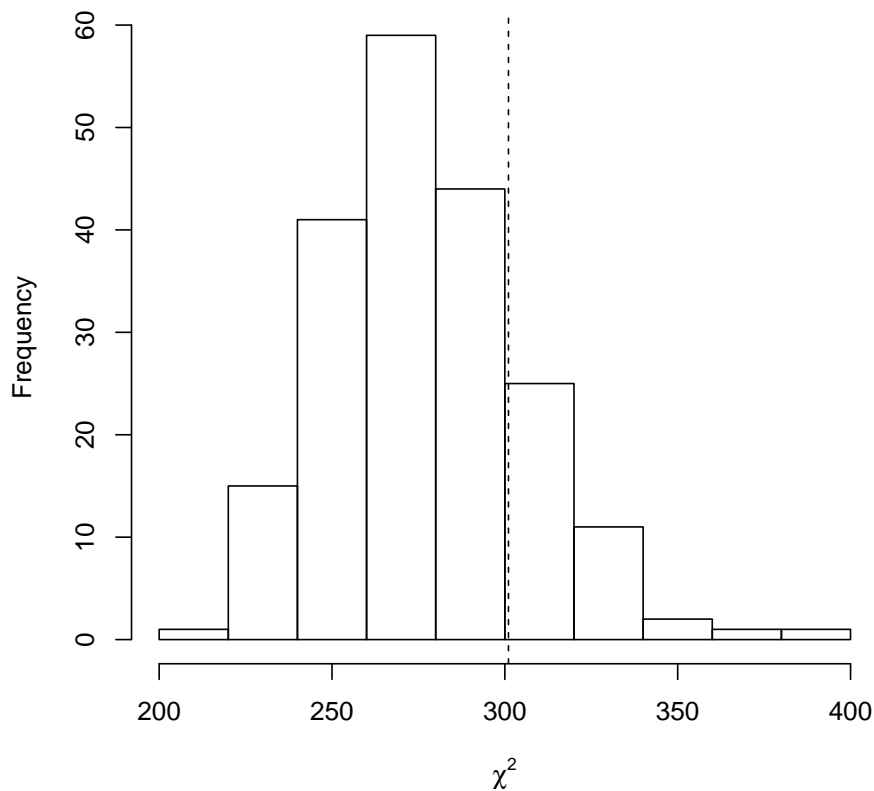


Figure 2: Graphical assessment of model fit by parametric bootstrapping. The dashed line is the observed chi-squared statistic. The histogram approximates the expected sampling distribution

```
> set.seed(1234)
> chisq <- function(fm) {
+   observed <- getY(fm@data)
+   expected <- fitted(fm)
+   sum((observed - expected)^2/expected)
```

```
+ }
> pb <- parboot(fm.oven.1, statistic = chisq, nsim = 200)

> plot(pb, main = "")
```

The above call to `plot` with a parametric bootstrap object as the argument produces a useful graphic for assessing goodness of fit (Figure 2). The plot suggests that the model adequately explains these data.

Beyond serving as a tool to evaluate goodness of fit, `parboot` can be used to characterize uncertainty in any derived quantity of interest.

## 4. Future directions for unmarked development

**unmarked** has become a stable and useful platform for the analysis of ecological data, but several areas of development could improve its utility. First, new models need to be added to cover the range of sampling techniques and population dynamics commonly encountered. Table /Reftab:modspace illustrates some of the gaps that need to be filled. In most cases, models to fill these gaps have not been developed so more research is needed.

Sampling method	Population Dynamics		
	Closed	Open to temporary emigration	Open to demographic processes
Occurrence sampling	<code>occu</code>	—	<code>colext</code>
Repeated counts	<code>pcount</code>	—	<code>pcountOpen</code>
Removal sampling, double observer sampling, and other multinomial designs	<code>multinomPois</code>	<code>gmultmix</code>	—
Distance sampling	<code>distsamp</code>	—	—

Table 3: Model fitting functions classified by sampling method and population dynamics. Missing cells indicate models that have not been developed but are likely to be investigated in the future.

Second, each of the models in **unmarked** assumes independence among sites. However, ecologists often use sampling methods such as cluster sampling that induce spatial dependence. Typically, this is done for logistical convenience, but because few methods are available to account for spatial correlation and imperfect detection probability, the spatial dependence is often ignored. Rather than this being a weakness of the sampling design, we envision that this dependence can be used as information regarding the spatial distribution of individuals. Third, many of the likelihoods are written in pure R, which can be slow for large problems. We are currently translating many of these functions into C++ with the help of the R package **Rcpp** (Eddelbuettel *et al.* 2011).

Finally, Markov chain Monte Carlo methods could be implemented for all of these models allowing for Bayesian inference. An important advantage of Bayesian analysis over classical methods is that the latent abundance or occurrence variables can be treated as formal parameters. Thus posterior distributions could easily be calculated for derived parameters such

as the proportion of sites occupied. Bayesian analysis also would provide a natural framework for incorporating additional sources of random variation. For example, one could model heterogeneity among sites not accounted for by covariates alone.

## Acknowledgements

The authors thank Andy Royle of the United States Geological Survey's Patuxent Wildlife Research Center, who provided initial funding and inspiration for this work.

## References

- Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001). *Introduction to Distance Sampling*. Oxford University Press, Inc., New York, USA.
- Chambers JM (2008). *Software for data analysis: Programming with R*. Springer Verlag, New York, USA.
- Chandler RB, Royle JA, King DI (In press). "Inference about density and temporary emigration in unmarked populations." *Ecology*.
- Dail D, Madsen L (2011). "Models for Estimating Abundance from Repeated Counts of an Open Metapopulation." *Biometrics*.
- Dorazio RM (2007). "On the choice of statistical models for estimating occurrence and extinction from animal surveys." *Ecology*, **88**(11), 2773–2782.
- Dorazio RM, Jelks HL, Jordan F (2005). "Improving removal-based Estimates of Abundance by Sampling a Population of Spatially Distinct Subpopulations." *Biometrics*, **61**(4), 1093–1101.
- Eddelbuettel D, Francois R, with contributions by Douglas Bates, Chambers J (2011). *Rcpp: Seamless R and C++ Integration*. R package version 0.9.1, URL <http://CRAN.R-project.org/package=Rcpp>.
- Gentleman R (2009). *R programming for bioinformatics*. Chapman & Hall/CRC, New York, USA.
- Hines JE, MacKenzie DI (2002). "PRESENCE." URL <http://www.mbr-pwrc.usgs.gov/software/presence.html>.
- Kéry M, Royle JA, Schmid H (2005). "Modeling avian abundance from replicated counts using binomial mixture models." *Ecological Applications*, **15**(4), 1450–1461. URL <http://www.esajournals.org/doi/abs/10.1890/04-1120>.
- MacKenzie DI, Nichols JD, Hines JE, Knutson MG, Franklin AB (2003). "Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly." *Ecology*, **84**(8), 2200–2207.

- MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002). “Estimating site occupancy rates when detection probabilities are less than one.” *Ecology*, **83**(8), 2248–2255.
- Royle JA (2004a). “Generalized estimators of avian abundance from count survey data.” *Animal Biodiversity and Conservation*, **27**(1), 375–386.
- Royle JA (2004b). “N-mixture models for estimating population size from spatially replicated counts.” *Biometrics*, **60**(1), 108–115.
- Royle JA, Dawson DK, Bates S (2004). “Modeling abundance effects in distance sampling.” *Ecology*, **85**(6), 1591–1597.
- Royle JA, Dorazio RM (2008). *Hierarchical Modeling and Inference in Ecology*. Academic Press, London.
- Royle JA, Nichols JD (2003). “Estimating abundance from repeated presence-absence data or point counts.” *Ecology*, **84**(3), 777–790.
- White GC, Burnham KP (1999). “Program MARK: survival estimation from populations of marked animals.” *Bird Study*, (46 Supplement), 120–138.
- Williams BK, Nichols JD, Conroy MJ (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego, USA.

### Affiliation:

Ian Fiske  
 Department of Statistics  
 North Carolina State University  
 2311 Stinson Drive  
 Campus Box 8203  
 Raleigh, NC 27695-8203  
 E-mail: [ijfiske@ncsu.edu](mailto:ijfiske@ncsu.edu)  
 Richard Chandler  
 USGS Patuxent Wildlife Research Center  
 Gabrielson Lab, Room 226  
 12100 Beech Forest Rd.  
 Laurel, MD 20708  
 E-mail: [rchandler@usgs.gov](mailto:rchandler@usgs.gov)