CHAPTER

20

Nonstandard GLMMs 1: Site-Occupancy Species Distribution Model

OUTLINE

20.1 INTRODUCTION

We have now seen a wide range of random-effects (also called mixed or hierarchical) models, including the normal, Poisson, and binomial generalized linear models (GLMs) with random effects. This means that some parameters are themselves represented as realizations from a random process. With the exception of the zero-inflated models in chapter 14, we have used a normal (or a multivariate normal) distribution as our sole description of this additional random component. However, nothing constrains us to use the normal distribution only and sometimes, other distributions will be appropriate for some parameters. The next two chapters illustrate two cases with discrete random effects that are assumed to be drawn from a Bernoulli or a Poisson distribution. Importantly, these effects have a precise biological meaning in these models: they correspond to the true, but imperfectly observed, state of occurrence (Chapter 20) or of abundance (Chapter 21).

Introduction to WinBUGS for Ecologists
DOI: 10.1016/B978-0-12-378605-0.00020-X

© 2010 Elsevier Inc. All rights reserved.

The modeling of animal and plant distributions is an important and active area of ecological research and applications. The most frequently applied method is a binomial GLM, or logistic regression (see Chapters 17–19), where the probability that an organism is found is modeled from what typically, and misleadingly, are called "presence–absence" data. These are binary indicators for whether a species was found (1) or not (0) in a spatial sample unit, and the effect of covariates is modeled through the logit link function. There are many variants of this approach, but the basic principle is often the same: so-called "presence–absence" data are directly modeled as coming from a Bernoulli distribution and the Bernoulli parameter is interpreted as the probability of occurrence.

However, a fundamental and extremely widely overlooked issue in almost all species distribution models is that detectability (p) of most species is imperfect—typically, a species will not always be detected where it occurs (MacKenzie and Kendall, 2002; Pellet and Schmidt, 2005; Kéry and Schmidt, 2008; Kéry et al., 2010b). In other words, detection probability is typically less than one ($p < 1$). This basic fact is very well known to field naturalists and reasonably understood for animals, and it even applies to populations of immobile organisms such as plants (Kéry, 2004; Kéry et al., 2006). However, it seems to have been overlooked by most professional ecologists dealing with distributional data (Araujo and Guisan, 2006; Elith et al., 2006).

As a consequence, virtually no study generated by current species distribution modelers actually models the true occurrence of a species as pretended or believed. Rather, the product of the two probabilities of occurrence and detection of a species is modeled. For noteworthy exceptions, see Gelfand et al. (2005), Royle et al. (2005), Latimer et al. (2006), and Altwegg et al. (2008), also see Royle et al. (2007), and Webster et al. (2008).

The widespread confusion about what is actually being modeled in most species distribution models has three main consequences (MacKenzie et al., 2002; Tyre et al., 2003; Gu and Swihart, 2004; MacKenzie, 2006; MacKenzie et al., 2006; Kéry and Schmidt, 2008; Royle and Dorazio, 2008; Kéry et al., 2010b):

1. species distributions will be underestimated whenever $p < 1$,
2. estimates of covariate relationships will be biased toward zero whenever $p < 1$, and
3. the factors that affect the difficulty with which a species is found may end up in predictive models of species occurrence.

The first is easy to understand, but the second has not been widely recognized, although Tyre et al. (2003), Gu and Swihart (2004), and MacKenzie et al. (2006, pp. 34–35) described this effect already a few years ago. Interestingly, and in contrast to a naïve analysis of count data in the presence of imperfect detection (cf. simple Poisson regression examples in Chapters 13–16), in distribution modeling, even a constant $p < 1$ will

bias low the slope estimate of the relationship between occurrence probability and a covariate. Presumably, this includes also a time covariate, i.e., situations where changes in distribution over time are modeled. As an example for the third effect, if a species has a higher probability to be found near roads, perhaps, because near roads, more people are likely to stumble upon it, then obviously roads or habitat types associated with roads will show up as important for that species, unless detection probability is accounted for. As an extreme example, when a species distribution map is constructed from road-kill records, then no matter how much roads might actually be avoided by that species in reality, the resulting distribution map will emphasize the great positive effect of roads on the distribution of the species!

In contrast, a novel class of models with the rather peculiar name "site-occupancy models" (MacKenzie et al., 2002, 2003, 2006; Tyre et al. 2003) is able to estimate the true distribution of animals or plants free of any distorting effects of the difficulty with which they are found. This chapter deals with these models and shows how to fit them using WinBUGS.

As a motivating example, we consider an inventory of the beautiful Chiltern gentian (Fig. 20.1) conducted in 150 calcareous grasslands in

FIGURE 20.1 Chiltern gentian (*Gentianella germanica*), Slovenia, 2007. (*Photo: M. Vogrin*)

the Jura mountains. Our aim is to estimate the proportion or number of occupied sites and to identify environmental factors related to the occurrence of the gentian. Interestingly, *Gentianella germanica* is a typical plant of nutrient-poor sites, which are thus *a priori* often rather dry. However, within the class of nutrient-poor grasslands, it preferentially occurs on wetter sites. However, these sites often have a higher and denser vegetation cover, so the rather small gentian (5–40 cm height) may more frequently be overlooked at these better sites (we ignore here the fact that better sites may hold larger populations). This effect could mask its preference for wetter sites. None of the currently widespread methods for distribution modeling such as GLM, generalized additive models (GAM), or boosted regression trees (Elith et al., 2006) are able to tease apart the effects of a covariate that influences both the occurrence of a species and the ease with which it is found (i.e. detection probability).

We will use site-occupancy models (MacKenzie et al., 2002, 2003, 2006) to separately estimate the gentian's probability of occurrence (called occupancy or species distribution) and its probability to be detected at occupied sites (detection probability), along with covariate effects on either occurrence or detection. It may be claimed that site-occupancy models are currently the only genuine distribution models available. All other widespread distribution modeling approaches confound occurrence and detection and only estimate the *apparent occurrence*, or more explicitly, the *combination of the probability of occurrence and the probability of detection, given the occurrence*.

The price to be paid for this improved inference is a sort of repeated-measures design, i.e., at least some sites need to be visited twice or preferably more frequently. This field protocol may be called a metapopulation design because the same quantity (occurrence) is assessed at many spatial replicates (Royle, 2004b; Royle and Dorazio, 2006). It is from the pattern of detection or nondetection at multiply visited sites that we obtain the information about detection probability, separate from occurrence probability. See MacKenzie and Royle (2005), MacKenzie et al. (2006), and Bailey et al. (2007) for design considerations relevant to this model.

A balanced design, i.e., an equal number of visits to all sites, is not essential for site-occupancy models; it simply makes things easier to simulate and present. (See Chapter 21 for an alternative, "vertical" data format, which is more convenient for the analysis of unbalanced metapopulation data.) Therefore, in our inventory of *G. germanica*, we assume that each site is visited three times by an independent botanist, and every time she notes whether at least one plant of *G. germanica* is detected or not. The result of these surveys may be summarized in a binary string, such as 010 for a site, where a gentian is detected during the second, but not during the first or third surveys. Generally, for a species surveyed T times at each of R sites, survey results are summarized in an *R-by-T* matrix containing a 1 when the species is detected at site i during survey j and a 0 when it is not.

The genesis, and therefore the analysis, of detection/nondetection observation y_{ij} at site i during survey j is naturally described by a hierarchical, or state-space, model that contains one submodel for the only partially observed true state (occurrence, the result of the biological process), and another submodel for the actual observations. The actual observations result from both the particular realization of the biological process and of the observation process.

$z_i \sim \text{Bernoulli}(\psi)$ Biological process yields true state
$y_{ij} \sim \text{Bernoulli}(z_i \times p_{ij})$ Observation process yields observations

Hence, true occurrence z_i of *G. germanica* at site i is a Bernoulli random variable governed by the parameter ψ (occurrence probability), which is exactly the parameter that most distribution modelers wish they were modeling. The actual gentian observation y_{ij}, detection or not at site i during survey j (or "presence–absence" datum y_{ij}), is another Bernoulli random variable with a success rate that is the product of the actual occurrence of *G. germanica* at that site, z_i, and detection probability p_{ij} at site i during survey j. Hence, at a site where the gentian doesn't occur, $z = 0$, and y must be 0. Conversely, at an occupied site, we have $z = 1$, and *G. germanica* is detected with probability p_{ij}. That is, in the site-occupancy model, the detection probability is expressed *conditional on occurrence*, and the two parameters ψ and p are separately estimable if replicate visits are available.

One way to look at this model in terms of the GLM framework, that features so prominently in this book, is as a hierarchical, coupled logistic regression. One logistic regression describes true occurrence, and the other describes detection, given that the species does occur. Another description of the model is as a nonstandard binomial GLMM with a binary distribution for the random effects—occupied or not occupied—instead of the normal distributions that we assumed in the "standard" binomial GLMM in Chapter 19 as well as most other mixed models in previous chapters.

The above equations describe just the simplest kind of a site-occupancy model; they can readily be extended to more complex cases. First, the ability to model covariate effects is crucial; indeed, covariates can easily be modeled into both the occurrence (ψ) and the detection (p) part of the model in a simple GLM fashion. That is, we can add to the model a statement like this:

$$\text{logit}(\psi_i) = \alpha + \beta^* x_i,$$

where x_i is the value of some occurrence-relevant covariate measured at site i, and α and β are parameters. The same can be done for the observation model also, i.e., the logit transformation of the detection parameter can be modeled by either a site covariate (x_i) or a survey covariate (x_{ij}).

We could also model the effects of many explanatory variables, of polynomial terms or even of splines.

Second, the model, as hitherto described, assumes a so-called closed population, i.e., a site is assumed to be either occupied or not occupied over the entire survey period. For plant surveys conducted during a single season, this may be a sensible assumption. But as soon as surveys are extended over several growing seasons, or the framework is applied to more short-lived or mobile organisms, such as most animals, occupancy status may change within the survey period. Indeed, colonization/extinction dynamics may be a research focus as in metapopulation studies (Hanski, 1998). The solution then is to collect data according to the robust design, where two or more surveys are conducted within a short time period (called a "season"), when the population can be assumed closed (Williams et al., 2002). This is repeated over multiple seasons between which the population may change. For such data, there is a dynamic formulation of the simple site-occupancy model described here to analyze occupancy, colonization, and extinction rates corrected for imperfect detection and to estimate covariate effects on each of these parameters, see MacKenzie et al. (2003, 2006), and Royle and Kéry (2007). In our case, however, we will simulate and analyze data only from a single season and assume a closed population.

20.2 DATA GENERATION

We now simulate the data for our inventory of *G. germanica*. We assume that the 150 sites visited form a sample of a larger number of calcareous grasslands in the Jura. The study objective is to use these data to learn about all calcareous grassland sites in the Jura and to see whether site humidity affects the distribution of *G. germanica*.

```
n.site <- 150          # 150 sites visited
```

We create an arbitrary continuous index for soil humidity, where −1 means dry and 1 means wet, and sort the data for convenient presentation of the results.

```
humidity <- sort(runif(n = n.site, min = -1, max =1))
```

Next, we create the positive true relationship between occurrence probability of *G. germanica* and soil humidity (Fig. 20.2). We do this by a logit-linear regression as customary for binomial responses. We choose the intercept and the slope for this relationship so that about 50% of all sites end up being occupied by the gentian.

```
alpha.occ <- 0        # Logit-linear intercept for humidity on occurrence
beta.occ <- 3         # Logit-linear slope for humidity on occurrence
```

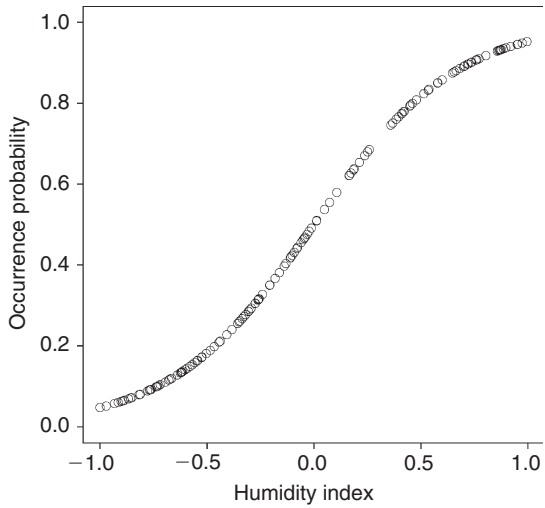

FIGURE 20.2 The unobserved true relationship between humidity and occurrence probability of the Chiltern gentian.

```
occ.prob <- exp(alpha.occ + beta.occ * humidity) / (1 + exp(alpha.occ + beta.occ
* humidity))

plot(humidity, occ.prob, ylim = c(0,1), xlab = "Humidity index", ylab =
"Occurrence probability", main = "", las = 1)

true.presence <- rbinom(n = n.site, size = 1, prob = occ.prob)
true.presence          # Look at the true occupancy state of each site
sum(true.presence)     # Among the 150 visited sites

> true.presence <- rbinom(n = n.site, size = 1, prob = occ.prob)
> true.presence # Look at the true occupancy state of each site
  [1] 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
 [36] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0
 [71] 0 0 0 1 1 0 0 1 1 1 1 0 1 1 0 1 1 1 1 1 1 0 1 0 0 1 0 0 1 1 1 1 1 1 1 0 1
[106] 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[141] 0 1 1 1 1 1 1 1 0 1
> sum(true.presence)# Among the 150 visited sites
[1] 75
```

This is the *true state* of the gentian system we are studying, i.e., the realization of the stochastic biological process we're interested in. This state is only imperfectly observable in nature, even for plant populations (Kéry et al., 2006). However, it is what we would like to observe and what we would like to relate to habitat variables such as humidity in our distribution model for *G. germanica*.

Unfortunately, we only observe a degraded image of that true state of nature, where the degradation is because of the fact that *G. germanica* may be overlooked and particularly so at the wetter sites with higher vegetation. So, next we simulate this effect to obtain our actual "presence–absence" observations. After simulating the biological process (which resulted in 150 realizations of true occurrence z_i), we now model the observation process that resides between the true biological process ("truth") and our observations (Fig. 20.3).

```
alpha.p <- 0              # Logit-linear intercept for humidity on detection
beta.p <- -5              # Logit-linear slope for humidity on detection
det.prob <- exp(alpha.p + beta.p * humidity) / (1 + exp(alpha.p + beta.p *
humidity))

plot(humidity, det.prob, ylim = c(0,1), main = "", xlab = "Humidity index", ylab
= "Detection probability", las = 1)
```

Assuming no false-positive errors, i.e., that no other species is erroneously identified as *G. germanica*, the Chiltern gentian can only be detected at sites where it occurs. Hence, the effective detection probability is the product of true occurrence (z_i) and this detection probability (`det.prob`):

```
eff.det.prob <- true.presence * det.prob
```

Importantly, this effective detection probability or apparent occurrence probability is precisely the quantity modeled by conventional species distribution models (e.g., Elith et al., 2006)! Its expectation is the product of occurrence probability and detection probability, i.e., $\psi \times p$.

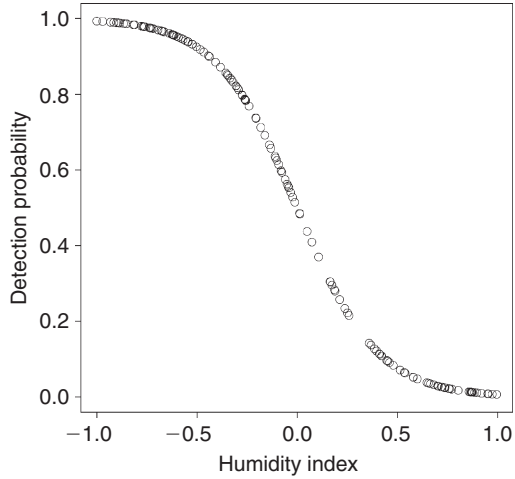

FIGURE 20.3 Relationship between site humidity and detection probability in the Chiltern gentian.

We store the results of each survey, 1 (gentian detected) or 0 (no gentian detected), in an n.site-by-3 matrix and fill it by simulating coin-flips (i.e., drawing Bernoulli trials) with the detection probabilities just computed. We note that this is the first time in the book that we use a two-dimensional array for our data. As a consequence, we will see a double `for` loop in the BUGS code to analyse these data.

```
R <- n.site
T <- 3
y <- array(dim = c(R, T))

# Simulate results of first through last surveys
for(i in 1:T){
    y[,i] <- rbinom(n = n.site, size = 1, prob = eff.det.prob)
}
```

Hence, `y` now contains the results of our simulated surveys to find *G. germanica* at 150 sites.

```
y    # Look at the data
sum(apply(y, 1, sum) > 0) # Apparent distribution among 150 visited sites
> y# Look at the detection/nondetection data
       [,1]  [,2]  [,3]
  [1,]    0     0     0
  [2,]    0     0     0
[ ... ]
[149,]    0     0     0
[150,]    0     0     0
> sum(apply(y, 1, sum) > 0)# Apparent distribution
[1] 31
```

On average (if we simulate this stochastic system many times) our parameter values yield about 42% detected gentian populations. Let's see what a naïve analysis of these observations would tell us about the relationship between humidity and the occurrence of *G. germanica*. (I call this analysis naïve because it omits an important system component, the observation process. The simplest way to analyze this relationship is by a logistic regression of an indicator for "ever detected" (here called `obs`) on humidity.

```
obs <- as.numeric(apply(y, 1, sum) > 0)
naive.analysis <- glm(obs ~ humidity, family = binomial)
summary(naive.analysis)
lin.pred <- naive.analysis$coefficients[1] + naive.analysis$coefficients[2] *
humidity
plot(humidity, exp(lin.pred) / (1 + exp(lin.pred)), ylim = c(0,1), main = "",
xlab = "Humidity index", ylab = "Predicted probability of occurrence", las = 1)
```

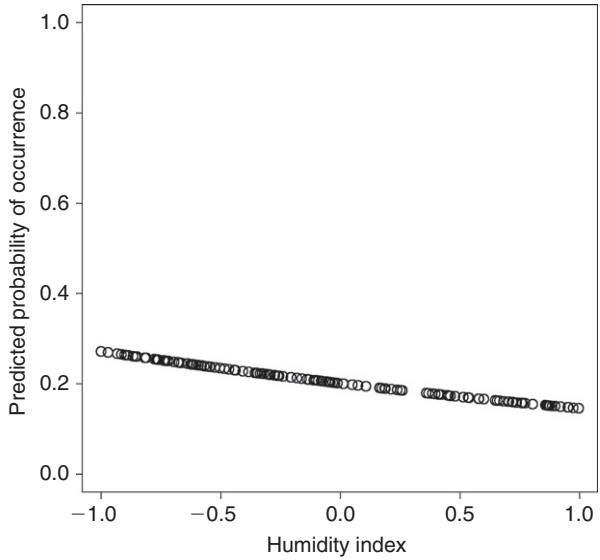

FIGURE 20.4 A naïve analysis of the apparent relationship between occurrence of the Chiltern gentian and humidity. This conventional species distribution model ignores detection probability.

We see that in a naïve analysis, the gentian's preference for wetter sites is totally masked (Fig. 20.4; although in different realizations of the data-generating process, you might also find significant positive or even negative slopes for humidity). Let's see whether a site-occupancy model can do better.

20.3 ANALYSIS USING WinBUGS

A variety of site-occupancy models (e.g., with covariates, for single or multiple seasons, single or multiple species) may be fitted using maximum likelihood in the free Windows-based programs MARK (see *http://welcome.warnercnr.colostate.edu/~gwhite/mark/mark.htm*) and PRESENCE (see *http://www.mbr-pwrc.usgs.gov/software/doc/presence/presence.html*). *R code for obtaining maximum likelihood estimates (MLEs) can be found in the Web appendix of the book by Royle and Dorazio (2008). Furthermore, there is a new R package called unmarked which allows to fit these models using maximum likelihood (Fiske and Chandler, 2010). Hence, here we will directly use WinBUGS to fit the site-occupancy model in a Bayesian mode of inference.*

As an added feature, we will perform a posterior predictive check based on the sum of absolute residuals. The resulting Bayesian *p*-value allows us to judge whether the assumed model is appropriate for our data set (based on the selected discrepancy measure).

```
# Define model
sink("model.txt")
cat("
model {

# Priors
 alpha.occ ~ dunif(-10, 10)      # Set A of priors
 beta.occ ~ dunif(-10, 10)
 alpha.p ~ dunif(-10, 10)
 beta.p ~ dunif(-10, 10)
# alpha.occ ~ dnorm(0, 0.01)  # Set B of priors
# beta.occ ~ dnorm(0, 0.01)
# alpha.p ~ dnorm(0, 0.01)
# beta.p ~ dnorm(0, 0.01)

# Likelihood
 for (i in 1:R) {  #start initial loop over the R sites
 # True state model for the partially observed true state
    z[i] ~ dbern(psi[i])             # True occupancy z at site i
    logit(psi[i]) <- alpha.occ + beta.occ * humidity[i]

    for (j in 1:T) {  # start a second loop over the T replicates
       # Observation model for the actual observations
       y[i,j] ~ dbern(eff.p[i,j])   # Detection-nondetection at i and j
       eff.p[i,j] <- z[i] * p[i,j]
       logit(p[i,j]) <- alpha.p + beta.p * humidity[i]

       # Computation of fit statistic (for Bayesian p-value)
       Presi[i,j] <- abs(y[i,j]-p[i,j])   # Absolute residual
       y.new[i,j]~dbern(eff.p[i,j])
       Presi.new[i,j] <- abs(y.new[i,j]-p[i,j])
    }
 }

fit <- sum(Presi[,])# Discrepancy for actual data set
fit.new <- sum(Presi.new[,])   # Discrepancy for replicate data set

# Derived quantities
  occ.fs <- sum(z[])        # Number of occupied sites among 150
}
",fill=TRUE)
sink()

# Bundle data
win.data <- list(y = y, humidity = humidity, R = dim(y)[1], T = dim(y)[2])

# Inits function
zst <- apply(y, 1, max)        #Good starting values for latent states essential !
inits <- function(){list(z = zst, alpha.occ=runif(1, -5, 5), beta.occ = runif(1,
-5, 5), alpha.p = runif(1, -5, 5), beta.p = runif(1, -5, 5))}
```

```
# Parameters to estimate
params <- c("alpha.occ","beta.occ", "alpha.p", "beta.p", "occ.fs", "fit",
"fit.new")

# MCMC settings
nc <- 3
nb <- 2000
ni <- 12000
nt <- 5

# Start Gibbs sampler
out <- bugs(win.data, inits, params, "model.txt", n.chains=nc, n.iter=ni, n.burn
= nb, n.thin=nt, debug = TRUE)
```

Before inspecting the parameter estimates, we first check the adequacy of the model for our data set using a posterior predictive check (Fig. 20.5).

```
plot(out$sims.list$fit, out$sims.list$fit.new, main = "", xlab = "Discrepancy for
actual data set", ylab = "Discrepancy for perfect data sets", las = 1)
abline(0,1, lwd = 2)
```

We then compute a Bayesian *p*-value based on the posterior predictive distributions of our discrepancy measures. In our graph, this corresponds to the proportion of points above the line and for an adequate model, it should be around 0.5 and not approach too much either toward 0 or 1.

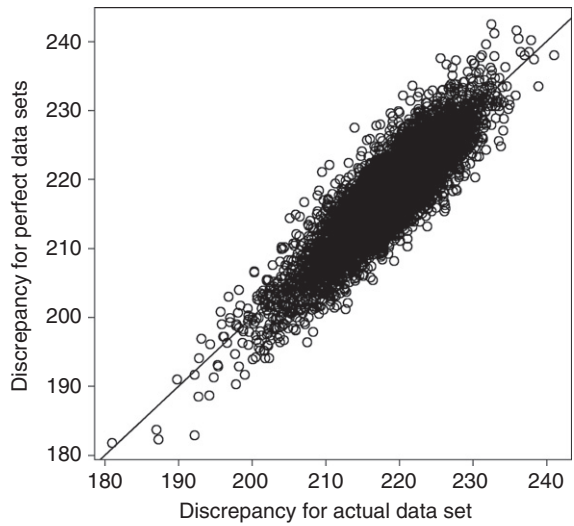

FIGURE 20.5 Posterior predictive check of the adequacy of the site-occupancy model for the gentian data based on the sum of absolute residuals. A well-fitting model has an even number of circles on either side of the 1:1 line.

```
mean(out$sims.list$fit.new > out$sims.list$fit)
> mean(out$sims.list$fit.new > out$sims.list$fit)
[1] 0.3688333
```

The model seems to fit well, so we compare the known true values from the data-generating process with what the site-occupancy analysis has recovered.

```
cat("\n *** Known truth ***\n\n")
alpha.occ ; beta.occ ; alpha.p ; beta.p
sum(true.presence) # True number of occupied sites, to be compared with occ.fs
sum(apply(y, 1, sum) > 0)        # Apparent number of occupied sites
cat("\n *** Our estimate of truth *** \n\n")
print(out, dig = 3)

> cat("\n *** Known truth ***\n\n")

  *** Known truth ***

> alpha.occ ; beta.occ ; alpha.p ; beta.p
[1] 0
[1] 3
[1] 0
[1] -5

> sum(true.presence) # True number of occupied sites, to be compared with occ.fs
[1] 75

> sum(apply(y, 1, sum) > 0) # Apparent number of occupied sites
[1] 31

> cat("\n *** Our estimate of truth *** \n\n")

  *** Our estimate of truth ***

> print(out, dig = 3)
Inference for Bugs model at "model.txt", fit using WinBUGS,
  3 chains, each with 12000 iterations (first 2000 discarded), n.thin = 5
  n.sims = 6000 iterations saved
             mean     sd    2.5%     25%     50%     75%   97.5%  Rhat n.eff
alpha.occ   0.445  0.466  -0.417   0.127   0.427   0.744   1.394 1.001  3900
beta.occ    3.737  1.094   1.769   2.971   3.673   4.444   6.043 1.002  2600
alpha.p    -0.255  0.259  -0.767  -0.432  -0.256  -0.084   0.256 1.001  6000
beta.p     -5.514  0.817  -7.237  -6.030  -5.484  -4.955  -3.988 1.001  6000
occ.fs     76.752  5.812  62.000  74.000  78.000  81.000  85.000 1.001  5800
fit       218.370  6.399 204.297 214.500 218.800 222.600 229.900 1.001  6000
fit.new   217.376  7.246 201.900 212.900 217.700 222.200 230.600 1.001  6000
[ ... ]
```

Thus, the site-occupancy species distribution model succeeds well in recovering the true relationships between humidity and occurrence and detection probability, respectively, which we built into the data. Particularly impressive is its ability to estimate the true number of occupied sites: gentians were only detected at 31 of the known 75 sites where they actually occurred, and the site-occupancy model estimates this number at 76.7, with a 95% credible interval of 62–85. (Note that when analyzing my simulated data set from the book website you will get slightly different results because of Monte Carlo sampling error.)

Finally, we will graphically compare the results from the naïve and the site-occupancy analysis with truth by plotting the true and estimated relationships between occurrence probability of the Chiltern gentian and site humidity (Fig. 20.6). This plot shows the mean predictions under the models; credible intervals could be added if we wished so.

```
plot(humidity, exp(lin.pred) / (1 + exp(lin.pred)), ylim = c(0,1), main = "",
ylab = "Occurrence probability", xlab = "Humidity index", type = "l", lwd = 2,
col = "red", las = 1)
points(humidity, occ.prob, ylim = c(0,1), type = "l", lwd = 2, col = "black")
lin.pred2 <- out$mean$alpha.occ + out$mean$beta.occ * humidity
points(humidity, exp(lin.pred2) / (1 + exp(lin.pred2)), ylim = c(0,1), type =
"l", lwd = 2, col = "blue")
```

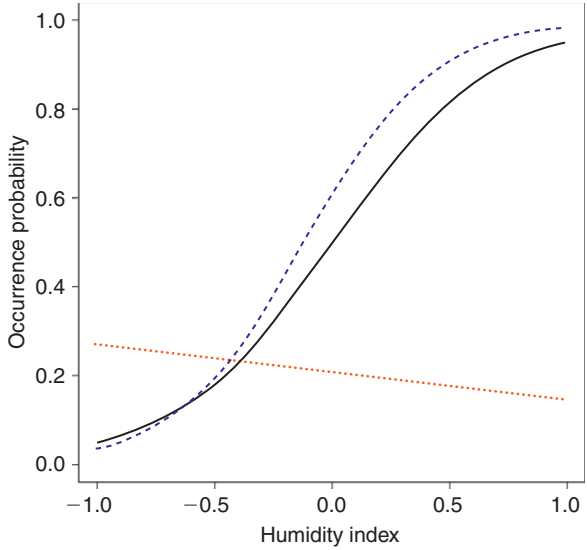

FIGURE 20.6 Comparison of true and estimated relationship between occurrence probability and humidity in the Chiltern gentian (*G. germanica*) under a site-occupancy model (dashed line) and under the naïve approach that ignores detection probability (dotted line). Truth is shown in solid line.

It is evident from Fig. 20.6 that nonaccounting for detection in species distribution models may lead one spectacularly astray. However, some might argue that we have simulated a pathological case, and that one would rarely find such a situation in nature. This may be true. But, we don't know until we have conducted the right analysis. And, we know that even a constant detection probability <1 not only biases the apparent distribution in conventional models but also biases low the strength of covariate relationships. Also, given suitable data, the site-occupancy distribution model can correct for that. This should make it a serious candidate for species distribution modeling when replicate observations are available from at least some sites.

In practice, this very positive conclusion about the model and its implementation in WinBUGS needs to be moderated somewhat. Performance of the model will be inferior with smaller samples (e.g., with fewer sites, a smaller proportion of occupied sites or lower detection probability, or fewer replicate visits) and presumably also in the presence of important unmeasured covariates. On the application side, it must be said that WinBUGS can be painful when fitting these slightly more complex models. For instance, it is essential to provide adequate starting values, in particular, for the latent state (occupancy, the code bit `z = zst`). There may also be prior sensitivity. For instance, altering the set B prior precision from 0.01 to 0.001 may cause WinBUGS to issue one of its dreaded trap messages (e.g., `TRAP 66 (postcondition violated)`). Sometimes, they are produced even with uniform or `normal(0, 0.01)` priors. Many other seemingly innocent modeling choices may influence success or failure when fitting a particular model to a given data set.

Hence, for suitable data, site-occupancy models in WinBUGS are great, but a comprehensive analysis of a more complex model may have to be accompanied by a few simulations to check the quality of the inference for the particular case. In addition, sometimes you must be prepared to do quite some amount of painstaking trial and error until the code works. But then, to a large part, this applies quite generally for more complex models, not just site-occupancy models and not just to models fitted using WinBUGS.

20.4 SUMMARY

The site-occupancy model is an extended logistic regression that can estimate true occurrence probability (ψ) and the factors affecting it, while correcting for imperfect detection. The extension is represented by the model component for detection probability (p): conventional logistic regression is a special case, when $p = 1$. Site-occupancy models are the only current framework for inference about species distributions that model true rather than apparent distributions; the latter is the product

of occurrence and detection probability (i.e., $\psi \times p$). Our example shows that not accounting for detection probability may lead to spectacularly wrong inferences about the distribution of a species under a conventional, naïve modeling approach. In contrast, the site-occupancy model applied to replicated "presence–absence" data was better able to estimate the true system state (site occupancy and covariate relationship).

EXERCISES

1. *Prior sensitivity*: Conduct a simple prior sensitivity analysis. Compare the inference under the normal and the uniform sets of priors. Plot a histogram of the posterior draws for each of the four primary model parameters to see whether the uniform (−10, 10) priors were not too restrictive to be uninformative.
2. *Site and survey covariates*: We have fitted a site covariate, i.e., one that varied among sites but not among surveys. Incorporate a survey covariate into simulation and analysis code, i.e., one that varies by individual survey. An example might be an inventory conducted by different people with differing and known experience. Experience could be rated on a continuous scale from 0 to 1 and more than one person would be sent to each site. Hint: A sampling covariate has the same 2-D-format as the observed detection/nondetection data.
3. *Swiss hare data*: Collapse the hare counts of one particular year (e.g., 2000) to binary data where a 1 indicates the observation of a "large population" (say, a count ≥10). Estimate the proportion of sites inhabited by a large population, i.e., one capable of producing a count ≥10. A site-occupancy model will correct for the fact that a "large population" may appear small, i.e., produce a count ≥10, or is missed altogether. You may fit a site covariate such as elevation on the occurrence probability of "large" populations (i.e., on ψ).
4. *Simulation study*: Extend the simulation in this chapter to see under which conditions the site-occupancy model is superior in performance to a binomial GLM. Things to vary might be the number of sites (e.g., 20, 50, 150, 500), number of replicate visits (e.g., 2, 3, 5, 10), true average occupancy, and true average detection probability. This is a larger project that could be part of a thesis.