CHAPTER

1 DISTRIBUTION, ABUNDANCE, AND SPECIES RICHNESS IN ECOLOGY

CHAPTER OUTLINE

1.1 POINT PROCESSES, DISTRIBUTION, ABUNDANCE, AND SPECIES RICHNESS

Distribution and abundance are the two fundamental state variables in ecology (Begon et al., 1986; Krebs, 2009) and species richness is the most widely used measure for biodversity (Purvis and Hector, 2000; Balmford et al., 2003). All three are the focus of a preponderance of both theoretical ecological studies and especially of studies focused on specific management or conservation problems involving rare or endangered species, game animals, and invasive species. Interestingly, though, all three are only derived quantities, i.e., summaries of a more fundamental quantity: *point patterns*. Point patterns are the outcome of stochastic processes known as point processes, and, not surprisingly, statistical models describing them are called point process models (PPMs; Illian et al., 2008; Wiegand and Moloney, 2014). PPMs treat both the number *and* the locations of discrete points as random quantities governed by an underlying, continuous intensity field. The intensity is the expected number of points (e.g., animals or plants) per unit area in some study area and is the modeled parameter.

Both distribution and abundance are simple areal summaries of spatial point patterns for a single animal or plant species, that is, aggregations of a point pattern over some area. To develop a basic understanding of the relationships between a point pattern and abundance and occurrence, we jump right in and run our first simple data simulation in program R. Thus, consistent with how we often approach the understanding of a new model in the rest of this book, *we here use simulation to explain and to understand* a model, such as a PPM. Function `sim.fn` lets you experiment with the relationship between a point pattern and abundance and occurrence as a function of the intensity of the pattern (which is something that you cannot control and is the result of the biology you're interested in) and of the grid size, or more specifically, the size of the cells making up that grid; this is something that you

Applied Hierarchical Modeling in Ecology. http://dx.doi.org/10.1016/B978-0-12-801378-6.00001-1

can control—or somebody else has done it for you (for instance, the people who designed the monitoring program that produces the data you are analyzing). The default settings of the function are:

```
sim.fn(quad.size = 10, cell.size = 1, intensity = 1)
```

The function simulates animal or plant locations in a grid of cells forming a quadrat with total length (in arbitrary units) equal to `quad.size`, according to a Poisson process where individuals are randomly distributed in space. This process is characterized by a constant `intensity`, which is the average number of animals or plants ("points") per unit area. The resulting point pattern is then discretized by overlaying a grid with quadratic cells of length `cell.size`. It is only this discretization of space that lets one define abundance in the first place and then presence/absence, or occurrence, in a second step. Species richness, the third crucial quantity in the title of this chapter, is the sum of the species occurring at a site, hence, a summary of the point patterns not for a single species but for all species (or for some set of species) that occur at a site.

As usual for the data simulation functions in this book, execution of the function produces both numerical output (data that you can save and do things with after) and informative plots that visualize the simulated process and the resulting data set (Figure 1.1). We will use many such functions in this book; also note that we have a whole chapter on the simulation of data in R (Chapter 4).

To appreciate the randomness inherent in the stochastic process defined by this function, we encourage you to call the function repeatedly without random number seeds, or with different seeds, and with changed function arguments. There is nothing like data simulation to help you realize what stochasticity really means—that lack of exact reproducibility of a process, which can therefore only be predicted in some average sense. Therefore, we urge you to play! Play with this data simulation function and also with all other data simulation functions in this book. You will see that this book gives you much to play with. "To play" means that you vary the function arguments and observe the changes in the output from the process represented by the data simulation function. We are convinced that this can be a huge help for your understanding of the process represented by the function. In addition, our hierarchical models directly represent the processes underlying the observed data, hence, if you understand the data simulation process that serves to *assemble* a data set, you will also understand the model that serves to *disassemble* the data set in the analysis, where disassembly means to "break the data apart" into coefficients of covariates, random effects variances etc. (Kéry, 2010).

For now, we execute the function once, with a specific random number seed, so you get the same results as we do. Afterwards, you can do `str(tmp)` to see the objects created by the function and saved in the object `tmp`, but we simply focus on the graphical output for now. This is all that we need to make our point about the one-way deterministic relationship between a point pattern, abundance and distribution (remember that "distribution" is simply a certain spatial pattern of presence/absence).

```
set.seed(82)
tmp <- sim.fn(quad.size = 16, cell.size = 2, intensity = 0.5)
```

This relationship among the three quantities is visualized in the first three panels of Figure 1.1. Without spatial discretization, neither abundance nor occurrence (or presence/absence) is defined; both necessarily require discretization of continuous space into what you can think of as one or more "sites." In this simulation, a "site" is represented by one cell in the entire grid. You can perhaps think of the entire grid as a region wherein your study takes place. Only once we have established that discretization of space is abundance (which we like to denote as N) or occurrence (presence/absence,

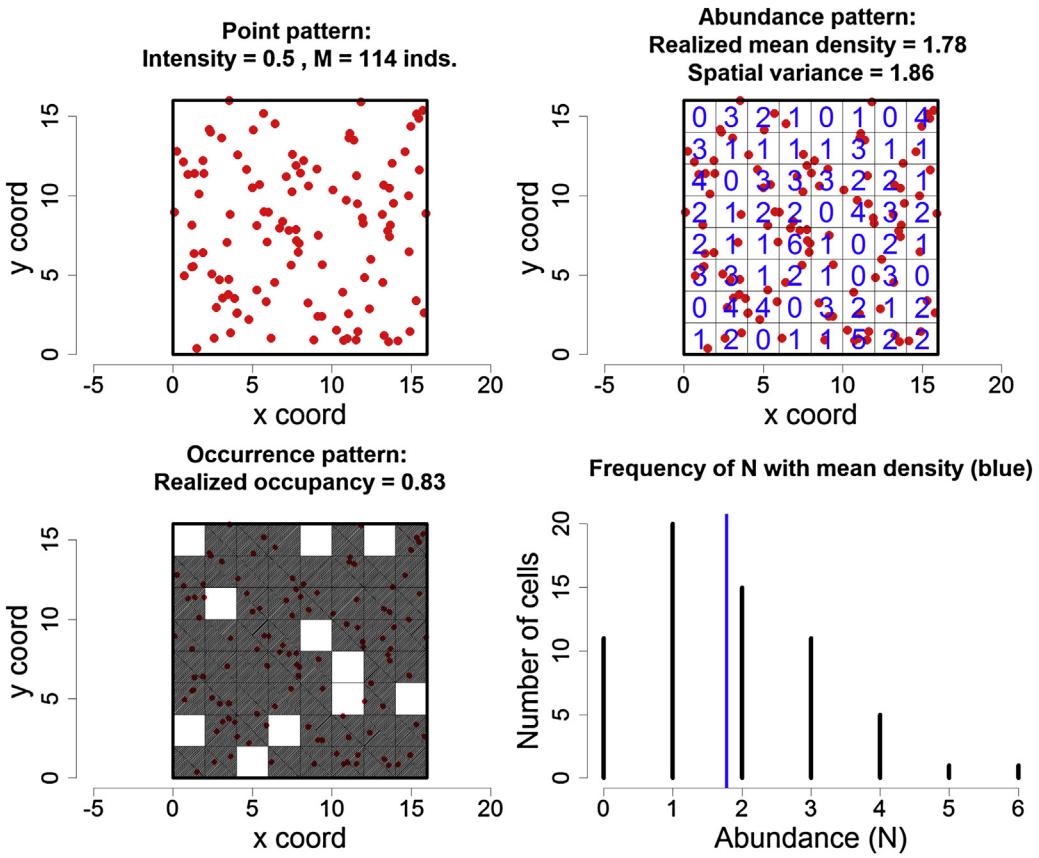

FIGURE 1.1

Relationship among three fundamental quantities in ecology: a *point pattern* of individual animals or plants (top left), a map of *abundance* with the local abundance (N) in each cell shown in blue (top right), and a species distribution map showing binary *presence/absence or occurrence* (bottom left), with occupied cells shown in gray and unoccupied cells in white. At the bottom right is the distribution of cell-based abundance (which is Poisson in this simulation), along with the mean shown in blue (which estimates the Poisson mean lambda). This figure is the graphical output from running function `sim.fn`.

which we like to denote as z) defined. Then, abundance N is simply the number of points falling into each "site" (i.e., cell)—if there is no point in a cell, abundance is zero; if there is one point, abundance is one; and so on. Furthermore, presence/absence (z) simply distinguishes the two cases where there is either no point in a cell (i.e., $N = 0$, this is an absence or nonoccurrence) or there is one or any number greater than one point in the cell (i.e., $N > 0$, this is a presence or occurrence). Thus, we can say that abundance is a first step of aggregating an underlying point pattern within some spatial discretization scheme, and occurrence is a second step in this aggregation over the spatial units. Alternatively, we can say that occurrence is a simple information-poor summary of abundance or "the poor man's abundance," where we only keep track of two abundance classes, one being zero (= "absence") and the

other greater than zero (= "presence"). Thus, the relationships between a point pattern, abundance and occurrence are deterministic in one way only—if you know the full pattern and are given some spatial discretization scheme, you have full knowledge also about abundance; and if you know the spatial pattern of abundance, you also perfectly know the spatial pattern of occurrence. In contrast, things are not so straightforward the other way round, e.g., from knowing a presence/absence pattern you cannot perfectly infer the underlying abundance distribution, although you can make explicit statistical inferences about abundance from simple occupancy data (He and Gaston, 2000; Royle and Nichols, 2003; Royle et al., 2005; Ramsey et al., 2015).

We can describe the spatial abundance pattern that emerges from this underlying spatial point pattern by discretization of space and summarizing the mean and the variance of the individual values of *N* in each cell. The way that this simulation works (i.e., with a uniform intensity over the entire field), the resulting numbers *N* will follow a Poisson distribution with mean *lambda*, where *lambda* is estimated by the mean abundance (or density) over the 256 cells. In turn, the spatial presence/absence pattern will follow a Bernoulli distribution with a "success parameter" that we will later call "occupancy probability," and which corresponds to the expected proportion of occupied cells (that is, cells with nonzero abundance).

This is perhaps the simplest possible manner to explain by simulation the relationship between a point pattern, abundance, and distribution—we use a so-called homogenous Poisson process, which is one with a constant intensity. When modeling the point pattern aggregate of abundance, this is equivalent to adopting a Poisson generalized linear model with an intercept only for the cell values of abundance. In real life, homogenous intensity fields arguably never exist, instead intensity is patterned due to environmental heterogeneity, which can be described by spatially indexed covariates or spatially correlated random site effects. Much of ecological modeling in space, also in this book, is aimed at identifying the nature and strength of such covariate relationships. When modeling distribution or abundance from real data, we very often find that there are too many zeros. That is, a species is absent from more sites than what we would expect under our model. Some authors therefore make a clear distinction between "distribution," which is something like a potential distribution area where a species can occur in principle, and "abundance," which describes the number of individuals only at sites that belong to that distribution area. Such authors then typically adopt zero-inflated Poisson or related zero-inflated models to describe what they perceive of as two distinct processes, distribution and abundance.

This is very different from the way in which we look at the two concepts of distribution and abundance. As just explained, in our view, "distribution" naturally follows from any given spatial distribution of abundance. We think that it rarely ever makes sense to conceive of two distinct mechanisms underlying a realized abundance distribution in space. Instead, we think that in almost all cases where there are too many zeroes in a data set, this is simply due to a failure to include in our model all adequate covariates to model these zeroes through the Poisson (or negative binomial, etc.) mean. We think that it is not very interesting to try and attribute much biology to what in our view is merely a deficiency of our abundance model and which manifests itself by a too high frequency of zeros.

There may be rare exceptions, of course, where there are indeed two entirely distinct stochastic processes governing the abundance distribution in space. For instance, imagine the abundance of some terrestrial species in an archipelago. Clearly, any abundance greater than zero requires the colonization of an island beforehand and that is a stochastic process with binary outcome—either the island is colonized or it is not colonized. This may have nothing to do with the factors that determine abundance on that island once it is colonized, and, therefore in this example, it makes sense to imagine two separate mechanisms underlying the spatial variability of abundance as in a zero-inflated abundance model.

But in the vast majority of cases we think of such zero-inflated models simply as a modeling trick to make up for our lack of perfect knowledge of the covariates that really govern abundance. Therefore,

we are happy to adopt zero-inflated models to account for the resulting lack of fit (see, e.g., Chapter 6), but we would not usually claim that there was much biology in the zero-inflation part of the model. Especially, we would not adopt complicated covariate models in the zero-inflation part and we would *never* use the same covariates in both the zero-inflation part and in the abundance part of the model (the resulting model is probably near-unidentifiable; see also Ghosh et al., 2012).

After this brief discussion of the meaning of zero-inflated models, let's now look further at the actual numbers in Figure 1.1. The intensity of the field underlying the point pattern is 0.5, hence we would expect to have a total of $M = 16^2 * 0.5 = 128$ individuals in the entire quadrat, which has an area of 256 units. However, due to the randomness in the number of points inherent in a point pattern model, we only have 114 individuals in this realization of the process. At the chosen grain size (i.e., with `cell.size = 2`), the abundance in the 256 cells varies from 0 to 6 individuals and the mean *realized* abundance is 1.78, while we would have expected $\lambda = 2^2 * 0.5 = 2$ (the difference is sampling variability). In addition, the variance of local abundance (N) is 1.86, while we would expect 2 under a Poisson distribution with expected value of 2. Finally, the *realized* proportion of occupied cells (occupancy) is 0.83, where under a Poisson process with constant intensity we would expect $\psi = 1 - \exp(-\lambda) \approx 0.86$ (i.e., 1 minus the expected proportion of zero abundance). As always with simulated data sets, we are neatly confronted with the difference between the *expected* value of the output from a stochastic process, i.e., the average over an infinite number of realizations, and the actual value in one particular realization of the process. The difference represents sampling variability.

In addition, you can use such simulation functions to learn something about the simulated process in a more general and fundamental way. For instance, `sim.fn` lets you study the relationships among the intensity of a field in a homogenous Poisson point process (`intensity`) and the grain (`cell.size`) of the measurement of distribution or abundance on one hand, and the resulting numerical values of abundance (N) and occurrence (z) on the other. The following sets of commands let you study some of the relationships in a more qualitative manner. To be able to average in your mind over the randomness of the process, you should execute every line multiple times.

```
# Effect of grain size of study on abundance and occupancy (intensity constant)
tmp <- sim.fn(quad.size = 10, cell.size = 1, intensity = 0.5)
tmp <- sim.fn(quad.size = 10, cell.size = 2, intensity = 0.5)
tmp <- sim.fn(quad.size = 10, cell.size = 5, intensity = 0.5)
tmp <- sim.fn(quad.size = 10, cell.size = 10, intensity = 0.5)
```

Although the underlying point pattern is identical on average, you see how both the mean abundance N (and the variance of abundance N) and the proportion of the occupied cells (ψ) increase with increasing grain size, provided that the quadrat size remains constant. When the cell size is equal to the quadrat size, we always observe 100% occupancy for a species that occurs at all.

```
# Effect of intensity of point pattern (intensity) on abundance and occupancy
tmp <- sim.fn(intensity = 0.1)  # chose default quad.size = 10, cell.size = 1
tmp <- sim.fn(intensity = 1)
tmp <- sim.fn(intensity = 5)
tmp <- sim.fn(intensity = 10)
```

Now, you will observe that when a species is very rare (intensity is low), the occurrence and the abundance patterns will be essentially identical, since rarely will a cell be inhabited by more than a single individual; see also Figure 1.2. However, the greater the intensity, the less informative will the spatial pattern of occurrence be about the spatial variation in population density.

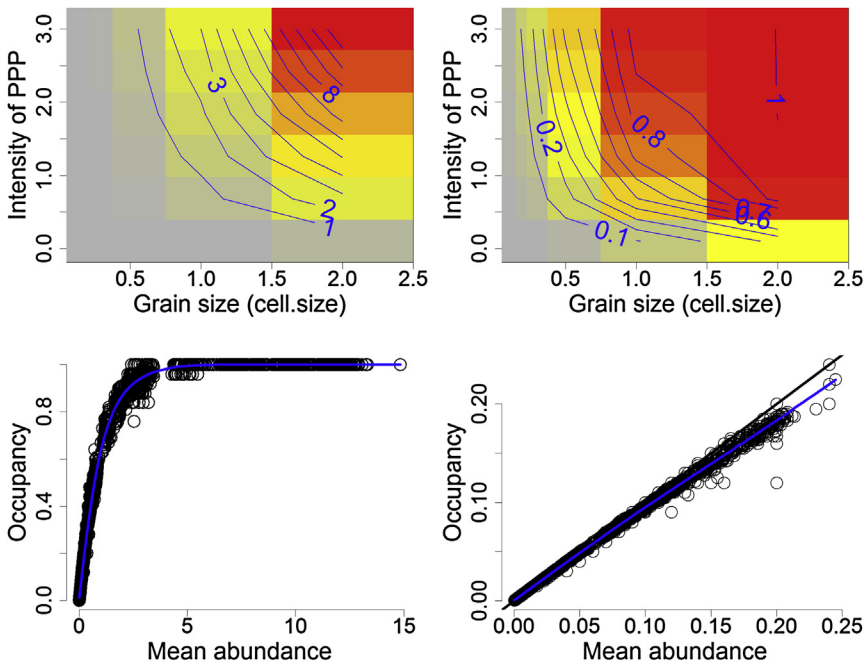

FIGURE 1.2

Relationships among intensity of the underlying Poisson point process and grain size and mean abundance per cell (top left), mean proportion of occupied cells (top right), and the relationship between mean occupancy and mean abundance for the full range of abundance created in the simulation (bottom left), and for a restricted range comprising only very small abundance values < 0.25 (bottom right); 1:1 line is added. Blue lines in bottom panels are smoothing splines with 4 d.f.

You can use a function such as this one for a formal simulation, to study the relationships among several quantities at a time. For instance, here is a little simulation to investigate the relationship between intensity and grain (`cell.size`) and the resulting mean density and occupancy proportion (ψ). We use the default quadrat size of 10 and vary both cell size and intensity in six steps each and record the mean abundance per cell and the realized proportion of occupied cells. We repeat this for a total of 100 data sets for each of the 36 combinations of the two factors `grain` and `int(ensity)`. When you switch off the plotting in the function, you generate $36 * 100$ data sets in barely four seconds!

```
simrep <- 100                    # Run 50 simulation reps
grain <- c(0.1,0.2,0.25,0.5,1,2) # values will be fed into 'cell.size' argument
int <- seq(0.1, 3,,6)            # values will be fed into 'lambda' argument
n.levels <- length(grain)        # number of factor levels in simulation
results <- array(NA, dim = c(n.levels, n.levels, 2, simrep)) # 4-D array !
for(i in 1:n.levels){            # Loop over levels of factor grain
  for(j in 1:n.levels){          # Loop over levels of factor intensity
    for(k in 1:simrep){
```

```
      cat("\nDim 1:",i, ", Dim 2:", j, ", Simrep", k)
      tmp <- sim.fn(cell.size = grain[i], intensity = int[j], show.plot = F)
      results[i,j,1:2,k] <- c(mean(tmp$N), tmp$psi)
    }
  }
}
```

We visualize the results in two image plots that show the average abundance (over the 100 simulated data sets) as a function of the six levels of each simulation factor (Figure 1.2, left) and the same for the mean realized proportion of occupied quadrats (Figure 1.2, right; code not shown).

We learn three things from Figure 1.2. First, we see that both abundance and occurrence do contain some information about the intensity of the underlying point process. Second, both abundance and occupancy are scale dependent (Figure 1.2, top), and, hence, you don't need to be a genius to recognize that neither abundance nor occupancy make sense when you don't know the spatial scale (here, `cell.size`) at which it is expressed (Fithian and Hastie, 2013). And third, there is a strong positive relationship between the mean abundance in a grid and the proportion of occupied cells (occupancy). At very small mean abundance, occupancy is exactly identical to average abundance (there is a slope of 1), while with increasing density, the slope of the relationship becomes shallower and eventually even zero, when all cells are occupied. Then, occupancy is no longer informative at all about either the underlying abundance or about the fundamental point pattern.

We have said that we can use R code for data simulation to explain a model, but of course the reverse is true also—that any data simulation implies a specific statistical model. Clearly, this data simulation process represents one particular model with many specific assumptions; for instance, we assume a homogenous Poisson process, which involves three things: that the spatial variability in abundance follows a certain pattern (that of a Poisson distribution), that there is no spatial heterogeneity in the suitability of the habitat and finally, that individuals are occurring independently of each other. All three are idealizations that will strictly never be true in real life. For instance, individuals may occur more aggregated (with larger spatial variance) or more evenly (with smaller spatial variance) than stipulated under the Poisson, the environment will be heterogeneous and so will be the intensity of the process and there may be repulsion (from territoriality) or aggregation (e.g., from social attraction) among individuals, all of which will again be manifest in the variance of abundance N. Also, we simulated a certain geometry, a square grid with an integer number of nested and contiguous cells, and this may not be adequate for some things that you might perhaps want to learn from such a simulation. As always with models, you need to use abstraction wisely—leave out only the things that are unimportant and keep those that are important; the same applies for simulation models.

In summary, the important insights that we wanted to gain from this simple simulation exercise are therefore: (1) that at the base of all abundance and distribution data resides a spatial point pattern (and that species richness is a summary of a whole collection of such species-specific point patterns), (2) that to assign a value of "abundance" or "occurrence," one must have a spatial scale and this is *only* possible when you discretize space, and (3) there is a one-way deterministic relationship in the relationship among the three scales of aggregation {point pattern, abundance, occurrence}, where knowledge of the one on the left gives perfect knowledge about the quantity to the right, while in the other direction, there is some—sometimes considerable—loss of information and therefore no simple relationship. Hierarchical models are extremely suited for describing processes with multiple scales, including combinations of two or more scales in the triple: point pattern, abundance, occurrence (Begon et al., 1986).

1.2 META-POPULATION DESIGNS

Interestingly, without even knowing about the relationship between point processes and their areal summaries of abundance and occurrence, people have always liked to discretize their entire study area into smaller subunits, or, put in another way, to replicate their study areas in space. This gives rise to what we call a "meta-population design" (Royle, 2004a; Kéry and Royle, 2010). We are a tad shy about this term because we do not mean to imply that the animals living in such discrete spatial units necessarily behave according to a formal metapopulation (Hanski, 1998; Sutherland et al., 2012, 2014). Rather, we could not come up with a better and more concise term for the extremely common case where distribution or abundance is studied at a collection of spatially replicated sites or where a whole study area is subdivided into smaller subunits, which we typically call a "site." This is a "meta-population design" to us, and to avoid annoying metapopulation ecologists, we sometimes put the term in quotes and add a hyphen. Nevertheless, we emphasize that the general sampling situation does include the formal metapopulation situation, and any model we discuss in this book can apply to classical metapopulations. Especially the dynamic models for occurrence (in Chapters 16 and 22 in *AHM* volume 2) are exactly metapopulation models for colonization/extinction dynamics in a presence/absence pattern.

Such meta-population designs, or designs with spatially subdivided populations, are extremely common in ecology and all related sciences. In addition, they are adopted virtually everywhere in biological monitoring, where it is clear that you can't characterize the state of the environment from measurements taken only at a single site. Meta-population designs come in a large variety, and the number, size, and shape of cells (subunits) may all vary. Sometimes there is heterogeneity even within a single design, e.g., study sites in a collection differ in area and shape and also in their spatial configuration (e.g., intersite distance). Sites may be naturally defined by a habitat boundary and thus represent "habitat islands," such as ponds when you are studying fish or pond-breeding amphibians or mountaintops when you're interested in alpine plant life. This may then be the typical setting for formal metapopulations. Alternatively, sites may be defined arbitrarily, e.g., by laying a grid over a map and then calling a grid cell a site. Sites may come in two dimensions or they may be one-dimensional and follow linear structures such as rivers, coastlines, roads, or footpaths. Finally, one typically has some larger region that one wants to characterize in terms of the abundance or occurrence of some species, and the sampling fraction of a meta-population design may then differ between anything from almost zero to one, corresponding to the cases where only a small minority of the possible sites are surveyed on the one hand and the complete coverage of that region on the other.

Figure 1.3 shows just four examples among a myriad of possible "meta-population designs." In the top row we contrast coverage, with perfect regional coverage (all cells in the region of interest surveyed; left) and regional coverage of about 25% (right). In the bottom row, we contrast a systematic versus a random placement of the spatial replicates, with the actual spatial sample of 267 sites in the "meta-population design" of the Swiss breeding bird survey MHB (left; see Sections 6.9, 7.9, 10.9 and 11.3 for more information about that survey), while right is a hypothetical variant of that design where 267 sites are chosen randomly. In terms of the sampling fraction, the MHB has only about 0.64% coverage (267/42,000).

In addition, spatial subsampling is surprisingly common in meta-population designs, wherein each site (unit) is further subdivided into smaller spatial subunits, which may again cover the entire site or they may only cover part of the entire area of a site; see Sections 6.14 and 10.10, with Figure 10.13.

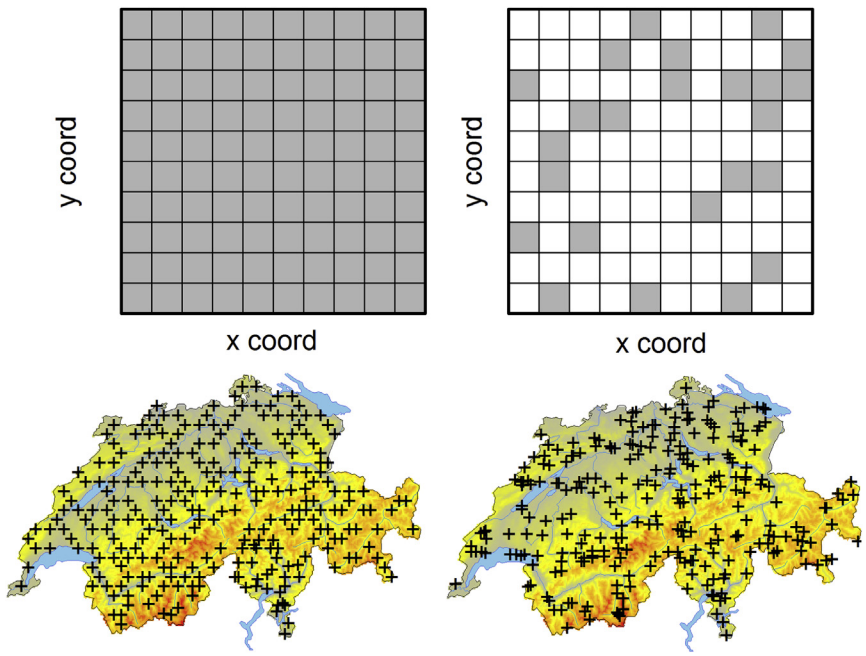

FIGURE 1.3

Four examples of meta-population designs: Top left: all cells (= sites) in a grid (= region) are surveyed; top right: only 25% of all sites in the region are surveyed; bottom left: the actual MHB meta-population design, with 267 1-km^2 quadrats laid out in an almost regular fashion that are surveyed from a region containing about 42,000 cells (= Switzerland); bottom right: a hypothetical variant of the MHB design where the same number of cells are placed randomly. In the bottom row increasing red means increasing elevation (the range is 200–4600 m).

And, finally, very often (especially in biodiversity monitoring) there is not only a spatial dimension in the study of distribution and abundance but also a temporal dimension, because there is an interest to study the dynamics of distribution or abundance over time. Thus, measurements of distribution and abundance are replicated temporally over longer time periods.

A final distinction may be made between the types of measurements that are taken at the collection of sites making up a whole "meta-population design." While an identical measurement protocol across units may perhaps be the most common approach, there are nevertheless many features of these measurements that may not be standardized, most of all the weather and other environmental conditions during which measurements are taken and commonly also the observer (i.e., it is rare that a single observer surveys all sites in a meta-population design). However, it is quite common also to measure several different things that are informative about the same underlying quantity. For instance, some sites may receive transect counts and others point counts (see Chapter 8). Or, there may be a combination of some sites where counts are conducted and others where only detection/nondetection observations (typically called "presence/absence data") are recorded. Of course, a very frequent source of such heterogeneity in sampling design is the need or the wish to combine the information from multiple schemes, where each scheme may be more homogenous by itself, but where there are

systematic differences between schemes (Solymos et al., 2013). Thus, in meta-population designs we are commonly faced with not only spatial variability in the biological quantity of interest (e.g., spatial variation in abundance) but almost always also with spatial variability in the measurement protocol with which we assess that quantity. In addition, there may be temporal variation in both aspects, too.

In summary, the "meta-population design" must be one of the most common study designs in all of ecology including applied fields such as conservation biology, wildlife biology, and especially in biodiversity monitoring. The design is so ubiquitous that perhaps some may question the need for a special term at all. Nevertheless, we felt we needed a label for it and so this is our label. The *AHM* book deals with the modeling of demographic data from meta-population designs on the distribution, abundance, and species richness of animals and plants, and of the parameters that underlie the changes of these variables over time (see next section). Hierarchical models are ideally suited to describe a quantity such as distribution or abundance that is available from a collection of sites and possibly repeatedly over time.

1.3 STATE AND RATE PARAMETERS

Abundance, distribution, and species richness may perhaps be the single most widely studied quantities in all of ecology and related fields. However, very frequently there is an interest not only in these state variables but also in the parameters that govern the rate of change of these quantities, i.e., that drive population dynamics (rate variables). For instance, the single most important quantity in biodiversity monitoring seems to be the "trend," that is, some sustained rate of change over time of some quantity such as abundance or occurrence. And, a trend is only the simplest description of population dynamics; the most detailed description of the dynamics of an animal population is achieved by the four vital rates: birth rate, immigration rate, death rate, and emigration rate, and these may be stratified by age, sex, or possibly other classes in a study population. One vital rate that has received particular interest in animal ecology and other fields such as life-history evolution (Stearns, 1992) is survival probability. There is a very well worked out theory for estimating survival from temporally replicated samples of marked, wild animals, e.g., in the celebrated Cormack-Jolly-Seber model (Cormack, 1964; Jolly, 1965; Seber, 1965; Pollock et al., 1990), in its variant for ring-recovery data (Brownie et al., 1985) and in a multitude of generalizations including multistate (Arnason, 1972; Hestbeck et al., 1991; Brownie et al., 1993; Arnason and Schwarz, 1999) and related models (Barker, 1997; Pradel et al., 1997; Kendall et al., 2003; Pradel, 2005; Bonner and Schwarz, 2006).

In animal ecology there is fairly strong divide between models and methods (and interestingly also a little in the people applying these models) that target state variables (e.g., estimate abundance) or that aim at rate variables (e.g., estimate survival). This divide is mostly artificial and to a large part due to limitations of past and current models and the associated software to fit these models. Hence, up to very recently, there was a huge divide in ecological statistics between models for "closed populations" (which essentially meant abundance estimation) and models for "open populations" (which first and foremost meant survival estimation). (Indeed, you can still see this divide in the way in which we split up the two volumes of our book.) However, nowadays this distinction becomes increasingly blurred and especially in the context of hierarchical models you will see how easily we can bridge what may have been thought of as a deep divide perhaps only 10 years ago. In addition, the ease with which we can fit "hybrid open/closed" models to our data is largely due to the power of the Bayesian model fitting machinery and in practice, to BUGS software (Schofield et al., 2009). This is especially interesting also for population

dynamics modeling such as population viability analysis (Beissinger, 2002) and related population modeling (Buckland et al., 2004b, 2007; Schofield and Barker, 2008; Tavecchia et al., 2009; Newman et al., 2006, 2014), including matrix projection models (Caswell, 2001; Link et al., 2003). Finally, another type of population model where the divide between open and closed populations is completely dropped is the fascinating area of *integrated population modeling* (IPMs; Baillie, 1991; Besbeas et al., 2002; Brooks et al., 2004; Schaub et al., 2007; Abadi et al., 2010a,b; Schaub and Abadi, 2011).

1.4 MEASUREMENT ERROR MODELS IN ECOLOGY

The error in a measurement is the difference between the measured value and the true value of some quantity. Probably by far the best-known type of measurement error in ecology is that associated with the measurement of continuous quantities such as body size, body mass, or the content of some pollutant in the air or the water. Their measurement is likely to be affected by a large number of small causes that act additively, giving rise to measurement errors that typically behave according to a normal distribution. An important consequence of this is that the measurements are unbiased, i.e., on average right on target—positive and negative errors simply cancel out in the average over repeated measurements. This type of measurement error seems to be the one that people have universally in mind when they think about this topic in ecology. For instance, such a type of measurement error is typically accommodated in the residual of a regression model.

However, things are very different for counts of discrete variables such as abundance, and this includes the binary variable "presence/absence," i.e., when you deal with the aggregation of data from an underlying point process. For them, you can undercount and overcount, and the mechanisms leading to the two types of errors are not the same, but can be very different. Thus, there is one set of mechanisms that lead to false-negative errors, when an individual is overlooked or a species is missed at a site where it occurs. This type of error cannot be reasonably described by a normal distribution but is typically described by a binomial or a Bernoulli distribution—given that there are N individuals out there and there is some probability p to detect any single one of them, the number of individuals detected (C) will be binomial:

$$C \sim Binomial(N, p) \quad \# \text{ False-negative measurement error model for counts}$$

Here, p is the detection or encounter probability of an individual and it represents the complement of the false-negative error rate, i.e., the associated error rate is $1 - p$. Similarly, for the presence/absence state of a species at a site, z, where $z = 1$ denotes presence and $z = 0$ denotes absence, we can specify the following measurement error model for the presence/absence measurement or detection/nondetection datum y at an occupied site:

$$y \sim Bernoulli(p) \quad \# \text{ False-negative measurement error model for det./nondet. obs.}$$

In either case, and unlike the normal model for measurement errors with continuous variables, the average of repeated measurements will *not* be unbiased with respect to the target quantities N (abundance) or z (presence/absence). Rather, the mean will be equal to Np for counts and equal to p for presence/absence measurements (detection/nondetection observations) at an occupied site. In contrast, the maximum among a series of n measurements will increasingly approach the true values N or z, when the number of replicate measurements n is increased. How quickly the maximum approaches N or z again depends on detection probability p (see Exercise 3).

False-negative detection error is "the" detection error that is addressed in the vast majority of capture-recapture and related methods that you have probably ever heard of (e.g., in Otis et al., 1978; Seber, 1982; Buckland et al., 2001, 2004a; Borchers et al., 2002; Williams et al., 2002; Amstrup et al., 2005; MacKenzie et al., 2006; Royle and Dorazio, 2008; King et al., 2009; Link and Barker, 2010; Kéry and Schaub, 2012; McCrea and Morgan, 2014; Royle et al., 2014). At the base of virtually all of these methods is the binomial or Bernoulli measurement error model from above. Arguably, false negative errors occur in almost all data sets of distribution and abundance, regardless whether they are for animals or for plants (Kéry and Gregg, 2003, 2004; Kéry, 2004; Kéry et al., 2005a, 2006; Chen et al., 2009, 2013).

In addition to false-negative errors, we may have false-positive errors, i.e., for abundance, we may overcount, most typically because we either count the same individual multiple times or because we mistake one species for another. For occurrence it means that we *think* we detected a species at a site where either it does not occur at all or we think we detected it at a site where it does occur, but what we saw was not the target species; see Chambert et al. (2015) for more on this subtle distinction. Methods that accommodate false-positive measurement errors in ecological models for abundance, occurrence and the the associated vital rates are still in their infancy and have essentially been developed in only two fields. For abundance, their development seems to have been restricted to genetic capture-recapture so far (e.g., Lukacs and Burnham, 2005; Wright et al., 2009; Yoshizaki et al., 2009; Link et al., 2010). For occurrence, the seminal paper by Royle and Link (2006) introduced occupancy models with both types of errors (i.e., false-negative and false-positive), but the main thrust of the development that led to more practically useful models came later with Miller et al. (2011, 2013b), Sutherland et al. (2013), and Chambert et al. (2015). However, at the time of writing the development of models with false-positive errors is both an active but also a difficult field, especially when (which is usually the case) we simultaneously have to confront both types of errors.

When we deal with areal-summary data from discretized "simple" point patterns, i.e., abundance or occurrence data, these are the two fundamental types of measurement errors. However, in addition, there are state classification errors when you classify individuals by age, size, or other states as in multistate models (Arnason, 1972; Hestbeck et al., 1991; Brownie et al., 1993) and when you distinguish different types of "occurrence" such as the occurrence of nonreproductive individuals versus that of reproductive individuals as in multistate occupancy models (Royle and Link, 2005; Nichols et al., 2007, see Chapter 18 in volume 2). (This means that you aggregate a *marked* point pattern, i.e., one where there are different types of points; Illian et al., 2008.) In these settings, the errors are simply a generalization of the two fundamental types of measurement error for discrete variables to multiple states.

When we deal directly with the underlying point pattern data, there is a third fundamental type of error: location error. That is, a difference between the true coordinates at which an individual is when you detect it and the coordinates that you record and feed into your model. Location error needs to be addressed in such spatially explicit models for abundance or density because otherwise biased estimators result. The simplest spatially explicit modeling framework for abundance, conventional distance sampling (see Chapter 8 in this book and Buckland et al., 2001), assumes away the problem by requiring zero location error as one of its main assumptions. Interestingly, once we "forget" the invididual's locations by aggregating a point pattern to become abundance or occurrence data, location error "vanishes" and is translated into either false-positive or false-negative error. If location error is such that an individual is erroneously recorded in a neighboring cell, then that record becomes a false

positive in that cell and will correspond to a false negative in the cell where the individual really is located. If location error does not lead to the recording of the individual in a different cell from the cell where the individual really is, it remains without a consequence in the modeling of abundance or occurrence.

And to finish our brief exegesis on measurement error in ecology, there is yet another type of measurement error: that in the covariables. This is quite different from the other measurement error types in this section, which are all associated with the response in a model, not with a covariate in the model. The issue of errors in covariables in models for distribution and abundance is exactly analogous to that in any other (regression) model; essentially, measurement errors in (continuous) covariables attenuates the slope estimate, that is, erroneously pulls the estimate towards zero. There is a pretty large body of research in statistics on this type of measurement error (Stefanski, 2000), and there are few if any novel considerations in the context of statistical ecology, and therefore we don't give it special attention.

Hence, it is typically *not* enough for an ecological model of distribution and abundance to simply describe the spatial variability of a process and possibly also the temporal dynamics in abundance or occurrence. Rather, to achieve unbiased inferences about the demographics of distribution and abundance it will be necessary to explicitly model the measurement error processes that underlie your data at hand. Studies employing meta-population designs typically face two sequential inferential steps (see Figure 1.4). The first is from the sample of surveyed sites to some larger, statistical "population" of sites in which we are interested (or the "region" we talked about in Section 1.2). We need a statistical model to describe the variability among these sites and the sampling of the surveyed sites to infer quantities in the entire region. And second, we need a second statistical model to describe the randomness in the measurement process, typically to estimate and therefore correct for false-negative and false-positive error rates. This two-step sampling procedure is ubiquitous in ecology and especially in biodiversity monitoring. It has been presented in a particularly lucid way in the seminal paper by Yoccoz et al. (2001), where they denote the two steps as "spatial variability and survey error" and "detection error."

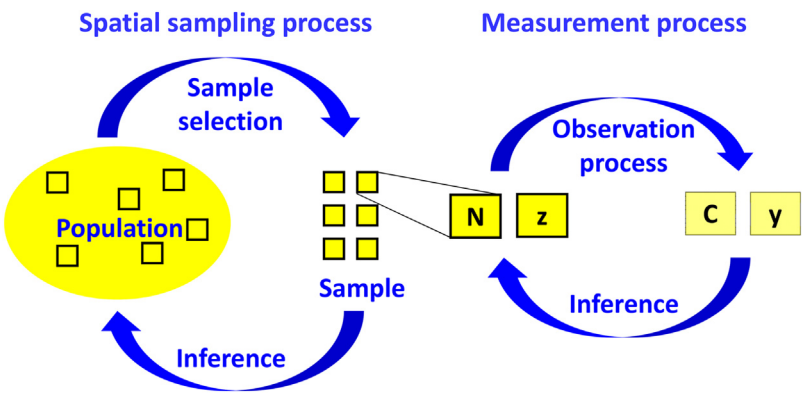

FIGURE 1.4

The two sampling processes in ecology that typically underlie the measurement of abundance or occurrence: first, spatial sampling and then the measurement of the desired quantity. *N* and *z* denote the typical quantities of interest (abundance and presence/absence) and *C* and *y* denote their measurements, a count or a detection/nondetection measurement, respectively (see also Yoccoz et al., 2001).

In summary, all of this calls for ecological models for distribution, abundance, and related demographic quantities that have at least two components: one submodel for the spatial and possibly also temporal variability in the focal quantity of interest and another submodel for the measurement process. Hierarchical models in this book achieve this aim in an admirable fashion.

1.5 HIERARCHICAL MODELS FOR DISTRIBUTION, ABUNDANCE, AND SPECIES RICHNESS

Hierarchical models are a sequence of probability models that are ordered by their conditional probability structure in the sense that they describe conditionally dependent random variables (see Chapter 2). In the context of models described and analyzed in this book, we use hierarchical models to describe both the true state of nature that is not observable (or only partly so) and also to describe the measurement error. Typically, our hierarchical models have one submodel for the true state of interest and another for measurement error although in some cases we might have more than one submodel for either. In the usual case where our model accommodates false-negative detection error only, the "bottom" level of the hierarchical model (where the data are) is a binomial (or Bernoulli) distribution where the "success probability" has the interpretation of the detection probability. In a sense, this turns most models in this book into some fancy sort of a logistic regression, but with possibly very complicated random effects structures. Other features of the spatiotemporal pattern of occurrence or abundance, or how we observe those, may be represented by additional levels in the model, especially groupings by site, species, etc.

Hierarchical models are our favorite framework for inference about distribution, abundance, species richness and related demographic quantities in populations, meta-populations, communities, and metacommunities. They are ideally suited to accommodate, in a single model, multiple data sets, multiple sources of variability such as spatial, temporal, and spatiotemporal variability, and multiple scales of measurement, while at the same time they rigorously propagate the combined uncertainty into every estimand from the model (Clark, 2007; Cressie et al., 2009; Cressie and Wikle, 2011; Hobbs and Hooten, 2015). Especially their Bayesian implementation with Markov chain Monte Carlo (MCMC) methods is almost limitless in its power to be applied to real data. In addition, hierarchical models represent a natural "compartmentalization" of a big, complex system into a sequence of smaller and usually far less complex subsystems. (In fact, 'sequential models' might be an equally fitting term for hierarchical models.) This is an ideal framework for describing jointly the true state and dynamics of an underlying system of interest, such as an animal or plant population or meta-population, and the potentially complex measurement processes with various and possibly heterogeneous types of error.

Understanding distribution and abundance and developing models that describe these things can be unified under a common framework of point process models (PPMs). Various models that we cover in this book either involve an explicit PPM (Chapters 8 and 9) or else we develop models for quantities that can only be sensibly understood as aggregations of an underlying point process (Chapters 6, 7, 10, and 11). Hence, we might call these two classes of models explicit and implicit PPMs.

1.6 SUMMARY AND OUTLOOK

In this chapter, we have clarified the meaning of three things: (1) of what distribution, abundance, and species richness are, (2) of what we call meta-population designs, and (3) of the types of measurement

errors associated with studies of distribution, abundance, and related demographic quantities. First, we have seen that in spite of their foundational role for all of ecology, distribution, abundance and species richness are only derived quantities that can moreover only be defined if we discretize space. The first two are the result of aggregating an underlying point pattern over a study area or its subdivisions, and the last one is the result of aggregating not a single point pattern (for one species) but the point patterns of *all* occurring species at some site (or of some defined group of species, such as Red-list species.). We have derived this one-way deterministic relationship between a point pattern, abundance, and occurrence by way of a trivial little simulation in R. This simulation also emphasized that occurrence is nothing but an information-reduced summary of abundance, wherein instead of the full abundance distribution we simply keep track of two abundance states, $N = 0$ and $N > 0$. Finally, in this simulation we illustrated the relationship between occupancy probability and mean abundance, which is almost exactly linear at very low values of mean abundance and becomes increasingly shallow with higher mean abundance. Once mean abundance is so high that 100% of the spatial units are occupied, the slope is zero.

Second, the discretization of space required to even define abundance and occurrence based on an underlying point pattern is very often made at a collection of spatial replicates, leading to the study of what may be called one spatially subdivided population or of a collection of several spatially replicated populations. We called this design a "meta-population design," though it may or may not be inhabited by a true metapopulation in the technical sense of that term (e.g., Hanski, 1998; Sutherland et al., 2012, 2014).

Third, we have described the two fundamental types of errors for discrete model responses such as abundance and occurrence: false-negative and false-positive errors. We have discussed some of the typical mechanisms that give rise to these errors and have seen that location error in the underlying point pattern translates into one of these two types of errors for the aggregated data.

Finally, we have introduced hierarchical models, which are our modeling framework of choice to describe distribution, abundance, species richness, and other demographic quantities such as population trends or vital rates, including survival probability. Hierarchical models are perfectly suited to accommodate the spatial replication inherent in a meta-population design as well as all kinds of measurement error and idiosyncrasies of the measurement protocol. Thus, hierarchical models provide us with a tremendous power to describe and understand how populations, meta-populations, communities, or metacommunities vary over space and time. Add to this the ease with which you can specify these models in the BUGS language and you see golden times approach for ecological modelers.

Finally, the segmenting of a single big, complicated process into a sequence of linked, smaller, and simpler subprocesses, which is a hallmark of hierarchical modeling (or "sequential modeling"), has the potential to radically change the way in which you approach modeling and inference of ecological systems. It is our experience that hierarchical modeling not only invites you, but actually almost *forces* you, to adopt a much more mechanistic way of thinking about your study systems. Thus, hierarchical modeling may change, and improve, the very way in which you approach science. Hierarchical models are the subject of the next 10 chapters in this first volume of *Applied Hierarchical Modeling for Ecologists* and then of another 14 or so chapters in volume 2 of the book.

EXERCISES

1. Use that first R simulation function (`sim.fn`) to improve your intuition about the nature of distribution and abundance. NOTE: This function lacks one important feature that is always present in real-world abundance and abundance data: it does NOT model measurement error.

2. Function `sim.fn` is most appropriate for plants. Why? Because there is no movement. Can you think of ways in which the results from our simulations (e.g., Figure 1.2) would change if there was also movement? For volume 2, we envision developing a version of the function that also includes movement, so then you can try out whether your hunches were correct.
3. In Section 1.4, we said that the maximum among a series of n measurements will come increasingly close to the true value N or z with increasing number of measurements. Devise a little simulation in R to show the relationship between the maximum of a collection of counts as a function of both true abundance N and detection probability p.