CHAPTER 5

DOWNSCALING ABUNDANCE FROM THE DISTRIBUTION OF SPECIES:

Occupancy Theory and Applications

FANGLIANG HE AND WILLIAM REED

5.1 INTRODUCTION

One of the most important contributions to our understanding of how and why species distribute in landscapes is to document the significant correlation between abundance and distribution of species across a broad range of scales (Brown 1984, 1995, Gaston and Blackburn 2000). The correlation suggests that there is a general tendency that locally abundant species are more widely distributed in space than rare species, which forms a positive distribution-abundance (or occupancy-abundance) relationship. While the observed relationship of this macroecological pattern begs for ultimate biological accounts (Brown 1984, Hanski et al. 1993, Gaston 1994, Kolasa and Drake 1998, Gaston and Blackburn 2000), the mathematical forms of the relationship derived from physical, statistical and geometrical considerations have greatly advanced the study on the topics and have indeed provided a solid ground for fermenting biological explanation further (Maurer 1990, Wright 1991, Hanski et al. 1993, Leitner and Rosenzweig 1997, Hartley 1998, Kunin 1998, He and Gaston 2000, Kunin et al. 2000, Harte et al. 2001, He et al. 2002; see Holt et al. 2002 for a review). An important implication of the distribution-abundance correlation is to allow for the derivation of species abundance from information on species distribution, a downscaling process (Wu and Li, Chapters 1 and 2). Here we will follow this premise to derive abundance by examining the spatial distribution of species in landscapes based on the combinatorial theory of occupancy.

The combinatorial theory of occupancy can date back as far as Pierre Laplace (Barton and David 1962) and has a long application in physics (Feller 1967). Laplace's classical example of occupancy considers the following birth game. Assume that there are N births taking place within a year and that each birth has the same chance to occur in any of the 365 days. What Laplace wanted to know was how many days out of the 365 would have no births, i.e., the number of empty days.

J. Wu, K.B. Jones, H. Li, and O.L. Loucks (eds.),
Scaling and Uncertainty Analysis in Ecology: Methods and Applications, 89–108.

Similarly, in statistical mechanics physicists are interested in knowing how N particles occupy a space composed of M small cells. The most well-known models that describe the number of empty cells (without particles) include Maxwell-Botltzmann and Bose-Einstein models. In this chapter, however, we wanted to know the reverse: not how many cells are empty, but how many particles are there provided that the number of empty cells is known. Specifically, let's consider a real example illustrated in Figure 5.1a in which a 50 ha plot in a rain forest of Malaysia is evenly divided into 800 cells of 25 × 25 m each (Figure 5.1b). The distribution (or

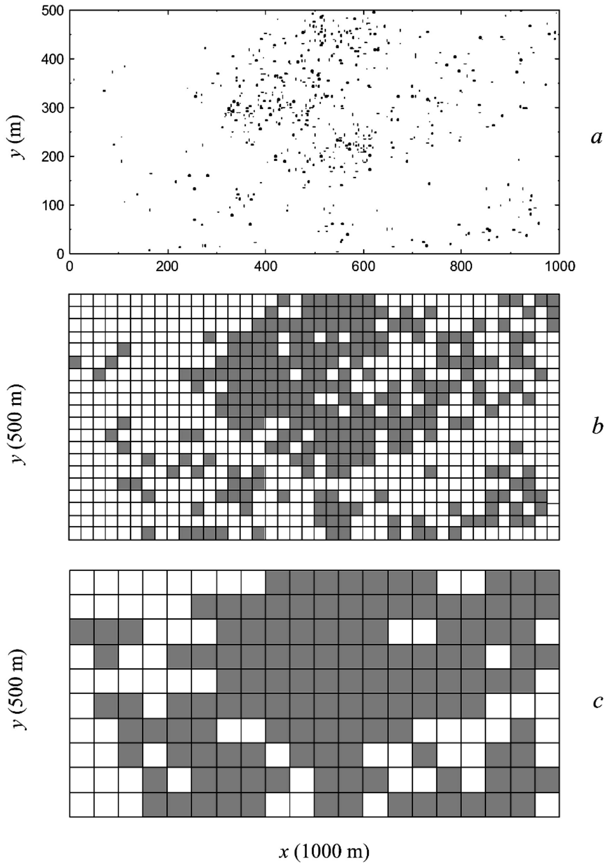


***Figure 5.1**. Example of distribution of canopy tree species Dacryodes rubiginosa in a 500 × 1000 m tropical rain forest plot of Malaysia. (a) The actual distribution of 591 stems of the species in the plot. (b) The lattice representation of the species distribution with a map resolution of 25 × 25 m. The area of occupancy A_{a1} by the species is 171875 m^2. (c) The coarse-scale lattice map generated by aggregating four adjacent cells in (b) with a map resolution of 50 ×50m. The area of occupancy A_{a2} by the species is 325000 m^2.*

occurrence map, binary map, or atlas) is so generated that a cell is grey if the species is present and white if it is absent. Thus, a grey cell has at least one tree, but can have many more. Given such a map, we want to find out how many trees there are; of course, for Figure 5.1 we already know the number of trees and their locations in the plot. Note that real distribution maps are usually not as regularly bordered as Figure 5.1, but, for simplicity, statistical derivations dealt with in this study will be based on a map with assumed regular borders. It will become clear later that the models so derived are equally applicable to irregular maps.

In its mathematical form, a distribution can be defined as

$$\boldsymbol{x} = (x_1, x_2, ..., x_M), \tag{5.1}$$

where the subscript $(1, 2, \ldots, M)$ is a (spatial) location index for the M cells, x_i is represented by either 0 or 1, depending on the absence or occurrence of the species in the cell. The vector $\boldsymbol{x}$ can be a random or systematic sample from a study area, or an exhaustive sample (census) that covers entire area of interest as illustrated in Figure 5.1b. Although random or systematic sampling is important, this study concentrates on exhaustive sampling.

In Equation 5.1, when $x_i = 1$, we know for sure that there is at least one individual occurring in that cell. Therefore, for an observation $\boldsymbol{x}$, we know that there are at least $\sum x_i$ individuals occurring in the M cells. But how many are actually there? This chapter was designed to answer this question. The reminder of the chapter is organized into three sections:

(1) We start from a classical occupancy model to derive an abundance estimate by assuming that the individuals of a species are randomly and independently placed in space. The classical occupancy estimate was evaluated by simulations and real data from a tropical rain forest of Malaysia.

(2) Following the same approach for the classical occupancy model, we derive an abundance estimate by assuming contagious distribution of the individuals. The estimate was also evaluated for the same data from the tropical rain forest.

(3) We show the connection of the estimates to species detectability in population sampling and derived a variance estimate to quantify uncertainty in detectability.

5.2 OCCUPANCY MODELS OF RANDOM PLACEMENT

5.2.1 The Classical Occupancy Model

The individuals of a species in an area can be distributed in many ways, which range from aggregated to regular patterns. Different spatial distributions result in different occurrence maps (Equation 5.1), even though the number of occupied cells may remain the same. In total, there are $\binom{M}{m}$ or $\binom{M}{u}$ possible maps for $\sum_{i=1}^{M} x_i = m$ and $u = M - m$, where m is the number of occupied cells, u is the number of empty cells. Figure 5.2 shows an example for $M = 4$, $m = 2$. Amongst these maps, the simplest case arises when all N individuals of a species are randomly distributed in a study area, A. This is equivalent to the situation of N individuals randomly placed into the M cells that comprise the area. Several models can be used to describe this random placement. In the statistical literature they are known as the "classical occupancy model" (Barton and David 1962, Kolchin et al. 1978).

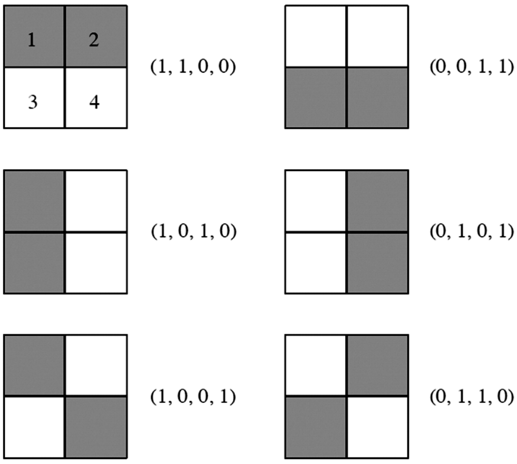


Figure 5.2. *Six possible maps of distributions for M = 4, m = 2 occurrences. Location index s = (1, 2, 3, 4) is shown in the upper left map.*

In the classical model, a species with N individuals is assumed to be randomly and independently distributed among the total number of M cells. The cell size is denoted as a, which defines the resolution of a map, or is called scale or grain in landscape ecology. It is clear that the probability that an individual falls in a given cell is simply *1/M* or *a/A* and the number of organisms, n, in a given cell follows a binomial distribution, i.e.,

$$p(n) = \binom{N}{n} \left(\frac{1}{M}\right)^n \left(1 - \frac{1}{M}\right)^{N-n}, \; n = 0, 1, 2, ..., N. \tag{5.2}$$

This model can be equally written in terms of areas as

$$p(n) = \binom{N}{n} \left(\frac{a}{A}\right)^n \left(1 - \frac{a}{A}\right)^{N-n}, \; n = 0, 1, 2, ..., N. \tag{5.3}$$

A realization of Equation 5.2 or Equation 5.3 produces an occurrence map as given in Equation 5.1. What we are interested in here is to estimate N given the occupancy in the map. The problem can be thought of as equivalent to placing N balls randomly and independently into M cells. Some cells will end up with no balls, some will have one ball, and others have several balls.

It can be shown that the random placement process will lead to the moment estimate of abundance N as (see the Appendix)

$$\hat{N} = \frac{\ln(1 - m/M)}{\ln(1 - 1/M)} \tag{5.4}$$

with approximate variance given as

$$V(\hat{N}) = \frac{V(u)}{[M\,(1 - 1/M)^N \ln(1 - 1/M)]^2} \tag{5.5}$$

where $V(u)$ is given by Equation A5 in the Appendix.

Equation 5.4 relates abundance N to the number of occupied cells m and the total number of cells M. A more desirable expression that explicitly links N to mapping scale a can be readily obtained as

$$\hat{N} = \frac{\ln(1 - A_a/A)}{\ln(1 - a/A)} \tag{5.6}$$

where A_a is the total occupied area ($= a \times m$), A is the total study area ($= a \times M$).

While Equation 5.4 was derived from regularly shaped maps, Equation 5.6 is suitable for both regular and irregular maps because the data on areas are used. Equation 5.6 was obtained previously by He and Gaston (2000) by a different approach and the derivation here is more rigorous. The variance given by He and Gaston (2000) is incorrect although it differs from Equation 5.5 by a small term.

The estimate given in Equation 5.6 can be further simplified for abundant species distributed in a large study area. It is easy to show that, when the study area $A \to \infty$ (or the total number of cells $M \to \infty$), Equation 5.6 becomes

$$\hat{N} = -\frac{A}{a}\ln(1 - \frac{A_a}{A}). \tag{5.7}$$

Its variance can be similarly derived from Equation 5.5 when $N >> M$ as

$$V(\hat{N}) = \frac{A}{a}\exp(\frac{Na}{A}). \tag{5.8}$$

5.2.2 Simulation Test and Applications

5.2.2.1 Simulation

The performance of Equation 5.4 or 5.6 was evaluated by generating a random distribution of a known number of "trees" (points) in an area. We simulated the distribution of "species" in a study plot of 500 × 1000 m. Three "species" were generated. The first one had 500 "trees" that were randomly located within the study plot. The plot was then divided into a lattice with scale a = 50 × 50 m to create a distribution map. Based on this map, the number of trees was estimated by using Equation 5.4. The simulation was repeated 100 times. The estimates are shown in Figure 5.3 (Species 1), along with the upper and lower bounds of the 95% confidence defined by $N \pm 1.96\sqrt{V(\hat{N})}$, where $V(\hat{N})$ is given by Equation 5.5.

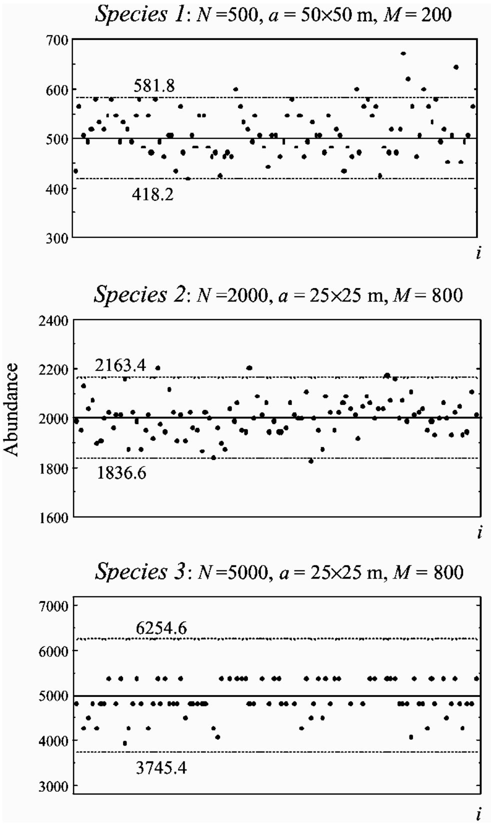
The second "species" had 2000 "trees", but this time the distribution was converted into a map with scale a = 25 × 25 m (as illustrated in Figure 5.1b). The third "species" had 5000 "trees" in a map at the same scale as for the second species. The results in Figure 5.3 show that Equation 5.4 estimates the abundances reasonably well for the randomly and independently distributed species. It appears that with the increase in N the approximate 95% confidence intervals constructed using the asymptotic variance of Equation 5.5 are too liberal when $N >> M$.

There was no estimation in the second simulation for species 3 (see the last table in Figure 5.3). This happens when a species fills up the entire area of a study. In this case, $m = M$, there is no solution to Equation 5.4.

5.2.2.2 *Applications*

We now apply Equation 5.4 to estimate the abundances of tree species in a lowland rainforest of Malaysia. The forest is located in the Pasoh Forest Reserve of Malaysia (2°55' N, 102°18' W). A 50 ha rectangular plot (500 × 1000 m) was initially established in 1987 and the census was repeated in 1990 and 1995 (Manokaran et al. 1999). The data from the 1995 census were used in this study. In each census, all free-standing trees and shrubs with diameter at breast height ≥ 1 cm were located by geographical coordinates on a reference map, and identified to species. In the 1995 survey, there were a total of 378224 trees belonging to 824 species. The most abundant species had 10470 individuals. Figure 5.1a is the distribution for one of the 824 species. The spatial patterns of the species surveyed in1990 were analyzed by He et al. (1997) which showed that about 80% of the species were aggregated, 20% had random distributions, and only one displayed a regular distribution. Because the abundance of each species was known, these census data allowed us to test the models that we developed.

Thirty-five of the 824 species were selected for analysis to represent the abundance range and spatial distribution patterns of the forest. The observed (true) abundances for the 35 species are listed in Table 5.1 together with the areas of occupancy at four scales (a = 10 × 10, 12.5 × 12.5, 25 × 25 and 50 × 50 m). Note that the total area of the study A is 500000 m^2 and that the area of occupancy A_a can be read from Table 5.1 for a given scale a. Substitute these three values into Equation 5.6, the abundance for each species could be estimated, and its corresponding variance can also be obtained from Equation 5.5. The results are shown in Table 5.2. The results for the simplified Equation 5.7 at a = 25 × 25 m are also presented in Table 5.2. It is clear that the simplified Equation 5.7 differs very little from Equation 5.6 even for rare species.



i	A_a	m	$\hat{N}$
1	442500	177	431.48
2	470000	188	561.27
3	460000	184	503.88
4	457500	183	491.79
5	462500	185	516.76
....	......		

i	A_a	m	$\hat{N}$
1	458125	733	1982.65
2	456250	730	1947.63
3	465000	744	2126.03
4	460625	737	2031.86
5	462500	740	2070.87
....	......		

i	A_a	m	$\hat{N}$
1	498750	798	4790.07
2	500000	800	--
3	497500	796	4235.91
4	498125	797	4465.90
5	498750	798	4790.07
....	......		

Figure 5.3. *Estimation of abundance for three simulated "species" in a 500 × 1000 m plot. The figures on the left-hand column are the outputs of 100 simulations for each species. The dashed lines are* $N \pm 1.96\sqrt{V(\hat{N})}$, *where* $V(\hat{N})$ *is given by Equation 5.5. The tables on the right-hand column are the outputs of the first five simulations for each species.*

Table 5.1. *Observed (true) abundance for 35 of 824 species in the Pasoh plot, and their area (m^2) of occupancy at four scales: 10 × 10, 12.5 × 12.5, 25 × 25 and 50 × 50 m.*

Species	*Abundance*	*Cell Size (m^2)*			
		10 × 10	*12.5 × 12.5*	*25 × 25*	*50 × 50*
1	1	100	156.25	625	2500
2	10	900	1562.50	5000	15000
3	13	1300	2031.25	8125	32500
4	22	1900	2968.75	11250	45000
5	27	2600	4062.50	15625	55000
6	30	2900	4531.25	14375	32500
7	50	5000	7500	27500	90000
8	98	9300	14375	51250	155000
9	115	9700	15156.25	45625	130000
10	122	11700	18437.50	65625	202500
11	155	14600	22031.25	79375	235000
12	157	14700	22812.5	83750	255000
13	177	16700	25312.50	82500	245000
14	207	19700	30468.75	105625	290000
15	302	27600	42031.25	140625	322500
16	325	30800	46406.25	158125	385000
17	333	31100	47968.75	159375	362500
18	384	33200	50312.50	162500	377500
19	405	36100	55156.25	175625	395000
20	490	44700	64687.50	195000	390000
21	520	43900	63750	161875	302500
22	522	45000	68593.75	199375	407500
23	537	47300	70625	203125	357500
24	742	63900	92968.75	262500	445000
25	874	72900	109062	290000	477500
26	891	74600	110469	286875	462500
27	1371	111400	156875	353750	482500
28	1419	115200	166562	376250	497500
29	2190	168000	231094	428750	492500
30	2793	166200	222812	410625	490000
31	3181	190300	246094	421250	492500
32	6031	308200	371562	478125	495000
33	7202	173400	200000	287500	392500
34	8571	186400	207656	275000	340000
35	10470	383300	433750	496875	500000

Table 5.2. *Estimated abundances for the 35 species in Table 5.1 using random placement Equation 5.6 and its simplified Equation 5.7 at four scales. The last row measures the "goodness-of-estimation" Δ of Equation 5.9.*

Cell Size (m^2)		*10 × 10*	*12.5 × 12.5*	*25 × 25*		*50 × 50*
Species	*True*	*Equation 5.6*	*Equation 5.6*	*Equation 5.6*	*Equation 5.7*	*Equation 5.6*
1	1	1	1	1	1	1
2	10	9.0	10.0	8.0	8.0	6.1
3	13	13.0	12.0	12.1	13.1	13.4
4	22	19.0	19.1	18.2	18.2	18.8
5	27	26.1	26.1	25.4	25.4	23.3
6	30	29.1	29.1	23.3	23.3	13.4
7	50	50.3	48.4	45.2	45.3	39.6
8	98	93.9	93.3	86.5	86.5	74.0
9	115	97.9	98.5	76.5	76.6	60.1
10	122	118.4	120.2	112.5	112.6	103.6
11	155	148.2	144.2	138.2	138.3	126.7
12	157	149.2	149.4	146.6	146.7	142.3
13	177	169.8	166.2	144.2	144.3	134.3
14	207	201.0	201.2	189.7	189.8	173.1
15	302	283.9	280.9	264.0	264.2	206.6
16	325	317.9	311.7	303.9	304.1	293.2
17	333	321.1	322.7	306.9	307.1	257.6
18	384	343.5	339.3	314.2	314.4	280.6
19	405	374.7	374.0	346.0	346.2	311.4
20	490	468.2	443.3	395.2	395.4	302.1
21	520	459.4	436.4	312.8	312.8	185.3
22	522	471.5	472.1	406.7	407.0	336.6
23	537	496.8	487.2	416.8	417.0	250.4
24	742	683.6	658.2	595.2	595.6	440.4
25	874	787.9	787.3	693.6	694.0	618.7
26	891	807.8	798.8	681.8	682.2	516.8
27	1371	1260.2	1204.7	982.8	983.4	668.8
28	1419	1309.3	1296.3	1116.4	1117.1	1057.0
29	2190	2047.2	1984.5	1557.8	1558.7	837.8
30	2793	2020.1	1887.4	1376.6	1377.4	780.5
31	3181	2394.8	2168.1	1477.7	1478.7	837.8
32	6031	4790.3	4348.6	2501.9	2503.4	918.7
33	7202	2129.2	1634.4	684.1	684.5	306.7
34	8571	2332.2	1717.1	638.4	638.8	227.3
35	10470	7274.3	6466.7	4057.6	4060.1	–
Δ		1.206	1.368	1.996	1.994	2.649

To measure the “goodness-of-estimation”, we define

$$\Delta = \sqrt{\sum b^2(\hat{N}_i)} \tag{5.9}$$

where $b(\hat{N}_i) = \dfrac{\hat{N}_i - N_i}{N_i}$ for species i.

The results in Table 5.2 show that except for those rare species there is considerable underestimation and that the underestimation becomes stronger with the increase of scale as evident from the measurement of Δ of Equation 5.9 (see the last row of Table 5.2). This is expected because few species in nature would present a truly random and independent distribution. Except at very low abundance, individuals of most species are typically aggregated. The underestimation of the random placement Equation 5.6 is largely due to the aggregation of a species, (overestimation would be more common if a species is actually at regular distribution). In other words, if the individuals of a species are not randomly and independently distributed, Equation 5.6 is biased. To reduce the bias, we need a method to take account of species aggregation.

5.3 OCCUPANCY MODELS OF CONTAGIOUS PLACEMENT

5.3.1 The Contagious Occupancy Model

Aggregated pattern arises when the distribution of individuals of a species among cells is produced by contagious processes. In this situation, the assumption of randomness and independence no longer holds; instead, a cell that already has an individual would be more likely to contain more individuals, and an occupied cell would be more likely to be adjacent to another occupied cell (and vice versa for empty cells). Barton and David (1959) show that contagious processes can either be modeled by a negative hypergeometric distribution arising from a Polyà urn model (they termed this model as pseudo-contagious process) or by a model of true contagion in which cells to be occupied are first selected at random and the number of individuals in the selected cells are then determined as realizations of logarithmically distributed random variables. It is well known that such a process generates the negative binomial distribution. For this latter model it can be shown that the moment estimate of abundance N of a species is (see Appendix B) given by

$$\hat{N} = Mk\left[\left(1 - \frac{m}{M}\right)^{-1/k} - 1\right] \tag{5.10}$$

where k is the aggregation parameter of the negative binomial distribution that takes positive values. Aggregated species have small k while random species have large k. Equation 5.10 can also be expressed in terms of areas as

$$\hat{N} = \frac{Ak}{a}\left[\left(1 - \frac{A_a}{A}\right)^{-1/k} - 1\right] \tag{5.11}$$

with approximate variance (see Equation B5)

$$V(\hat{N}) = \left(1 + \frac{N}{Mk}\right)^{2k+2} V(u) \tag{5.12}$$

where *u* is the number of empty cells and *V*(*u*) is given by Equation B4 in the Appendix.

5.3.2 Test for the Contagious Abundance Estimate

Given a distribution map, there are two unknown parameters *N* and *k* to be estimated in Equation 5.10 or 5.11. To use the method of moments one would normally equate the observed first and second moments with their theoretical mean and variance (see Equation B3 and B4 in the Appendix). In this case, however, a map only has one single observation on *u*, and its variance *V*(*u*) is not available. Splitting the map to create more observations will not work, because there will be a new unknown parameter for each part of the map (the number of organisms in that part). A possible alternative is to group cells to produce a coarser scale map (Kunin 1998, He and Gaston 2000). This will lead to two equations of Equation 5.10 or 5.11 for two unknown variables *N* and *k* with the assumption that the aggregation parameter *k* remains the same at both scales. We realize that this assumption does not necessarily hold in reality (Pielou 1957, Taylor et al. 1978). However, if the difference in scale for the two maps is not large, this assumption may be plausible.

From the fine-scale map we can read the area of occupancy A_{a1} at scale a_1. The second map can be produced as follows. If any of the adjacent cells at the fine-scale map are occupied, then the aggregated cell at coarse-scale is occupied; otherwise, it is left empty. The second map has a coarse scale a_2 and an area of occupancy A_{a2} (see Figure 5.1c for an example).

With the two maps so generated, *N* and *k* in Equation 5.10 can be evaluated numerically using, e.g., Newton-Raphson method. The estimated abundances for 35 of the 824 species are shown in Table 5.3. The first one is calculated in terms of two maps: the fine-scale map with $a_1 = 12.5 \times 12.5^2\text{m}$ and the coarse-scale map with $a_2 = 25 \times 25$ m^2. The second map pair is the two maps with scales at $a_1 = 25 \times 25$ m^2 and $a_2 = 50 \times 50$ m^2. The results show that Equation 5.10 works fairly well. Compared with the random placement Equation 5.4 in the previous section, the estimation is substantially improved (compared the *Δ*'s in the last rows of Tables 5.2 and 5.3). Estimation for rare species (e.g., abundance ≤ 2000) appears to work particularly well, which is indeed the strength of the method (Equation 5.10) because the abundance information on rare species is the major concern of conservation.

Table 5.3. *Estimated abundance for the 35 species in Table 5.1 using the contagious occupancy Equation 5.11 in terms of two map pairs: 12.5 × 12.5 − 25 × 25 m, and 25 × 25 − 50 × 50 m. The last row measures the "goodness-of-estimation" Δ of Equation 5.9.*

Species	*True*	*12.5 × 12.5 − 25 × 25*	*25 × 25 − 50 × 50*
1	1	1	1
2	10	11.0	9.1
3	13	13.0	13.0
4	22	19.4	18.0
5	27	26.4	26.2
6	30	32.1	34.0
7	50	49.5	47.6
8	98	96.0	91.7
9	115	110.3	85.1
10	122	123.1	115.9
11	155	146.3	142.7
12	157	150.4	148.1
13	177	175.8	147.9
14	207	205.4	196.3
15	302	287.2	294.3
16	325	314.3	307.7
17	333	328.5	329.5
18	384	348.9	328.0
19	405	384.7	359.9
20	490	463.0	446.8
21	520	515.1	439.5
22	522	500.8	439.8
23	537	518.5	576.3
24	742	683.4	687.0
25	874	826.6	724.2
26	891	851.0	774.8
27	1371	1312.8	1206.3
28	1419	1375.2	1138.1
29	2190	2209.1	2492.7
30	2793	2200.8	2043.2
31	3181	2659.5	2193.3
32	6031	6341.9	14253.4
33	7202	4803.7	1603.1
34	8571	9078.3	4603.2
35	10470	8548.8	3528.3
Δ		0.542	1.920

It is apparent that a considerable degree of underestimation still remains for those very abundant species. This underestimation is also observed for abundant insects (Warren et al. 2003). However, in that study they did not estimate abundances for rare species which are actually very simple to compute, their conclusions are thus unfortunately biased.

Similar to Equation 5.4, the accuracy of the estimation of Equation 5.10 also depends on the scale of observation. The results as measured by Δ of Equation 5.9 (the last row of Table 5.3) show that the estimation becomes progressively poorer with the increase in the scale from the map pair of $12.5 \times 12.5 - 25 \times 25$ m^2 to that of $25 \times 25 - 50 \times 50$ m^2. In addition to the effect of scale on the accuracy of abundance estimates, the precision (i.e., the variances of Equation 5.5 and 5.12) of the estimates also varies with scale. Figure 5.4 shows the effect of spatial aggregation on the variance-scale relationship for Equations 5.5 and 5.12. At random distribution (Equation 5.5), variance in abundance monotonically increases with scale (the dashed line), while the variance of Equation 5.12 can be hump-shaped for aggregated species. The practical implications of the variance-scale relationship are that sampling scale for randomly distributed species should be as small as possible if high precision is to be achieved, and that, for aggregated species, the model scales that lead to high variance should be avoided in order to achieve high precision.

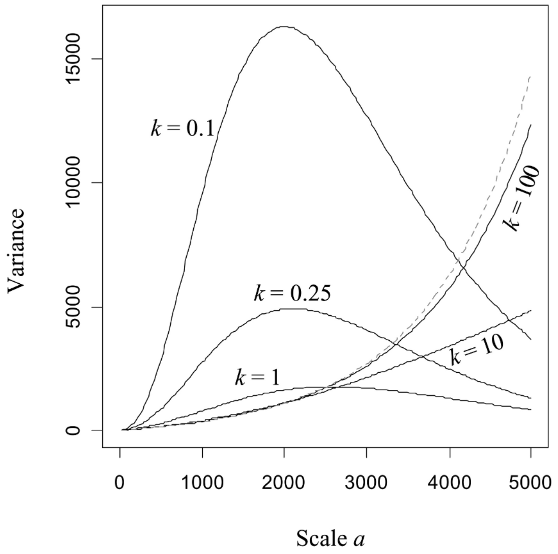


Figure 5.4. Variance-scale relationships. The dashed curve is the variance for the classical random placement estimate (Equation 5.5). The solid curves are the variances of Equation 5.12 with the aggregation parameter k varying from 0.1 to 100. The plot is produced by setting N=500 and a from 0 to 5000. Note Equation 5.12 approaches Equation 5.5 at large k.

5.4 UNCERTAINTY IN OCCUPANCY AND SPECIES DETECTABILITY

Equation 5.11 is an important occupancy-abundance model in ecology (Wright 1991, Hanski et al. 1993, Hartley 1998, He and Gaston 2000), which is typically written in the standard form as

$$p = 1 - \left(1 + \frac{Na}{Ak}\right)^{-k} \tag{5.13}$$

where p is the proportion of occupied area, or A_a/A.

More than any other occupancy-abundance models in the literature (see Holt et al. 2002 for a review), Equation 5.13 unifies occupancy (p), species abundance (N), the spatial pattern of the species (k), mapping scale (a), and the extent of study area (A) into a single mathematical form. In addition, Equation 5.13 is mathematically very flexible in that many other occupancy-abundance models are its special cases (He and Gaston 2000, He et al. 2002). For example, it is easy to show that the random placement Equation 5.6 is a special case of Equation 5.11 at $k = -N$. Equation 5.13 also provides a basic tool for investigating other biodiversity patterns, such as species-area curves (He and Legendre 2002, He et al. 2002) and beta diversity patterns (Plotkin and Muller-Landau 2002).

When used for sampling populations, occupancy p is often referred to as species detectability. The occupancy-abundance model (Equation 5.13) suggests that the detectability depends not only on the abundance of the species but also on its spatial distribution and the size of sampling unit. This finding is useful in sampling design. For instance, for a given abundance we know from Equation 5.13 that strong aggregation (i.e., small positive k) leads to small detectability and randomly distributed species (large k) have large detectability. So, in order to retain a high level of detectability for an aggregated species, it is necessary to use a large sample area (i.e., large a). Similarly, Equation 5.13 would help us calculate the size of sample areas for rare and common species for a predetermined detectability p.

Another important sampling issue is that the presence of a species in a site may or may not be observed in the field, i.e., there is always an uncertainty associated with detectability. The nondetection may mean that the species is truly absent or that it is missed because of insufficient survey efforts or sampling errors (MacKenzie et al. 2002). The latter scenario will inevitably lead to underestimation of occupancy rates. This uncertainty in occupancy p can be quantified according to the theory of occupancy in the Appendix. Because $p = m/M = 1 - u/M$, where the number of empty cells u is a random variable, it is easy to show that p has a variance of $V(p) = V(u)/M^2$, where the variance $V(u)$ is given either by Equation A5 or B4, depending on whether the random placement model or the contagious model is used. $V(p)$ provides information on the uncertainty in the detectability.

5.5 CONCLUSIONS

Based on the discussions in previous sections, we make the following conclusions:

(1) The widely recognized distribution-abundance macroecological pattern suggests that distribution and abundance of species are closely correlated so that we may infer about one from the other. This premise provides an essential basis for deriving information on abundance in terms of the distribution of species. In this chapter, we approached this problem by modeling the distribution of species in landscapes with the occupation process that N balls are placed into M cells following the theory of combinatorial occupancy.

(2) Two abundance estimates were derived from the theory of occupancy. The first one, as given by Equation 5.6, was derived under the assumption that the N unknown balls are randomly and independently placed into M cells. The second estimate (Equation 5.11) was derived from the contagious process that generates aggregated distribution of species (the negative binomial distribution). The random placement Equation 5.6 is a special case of Equation 5.11 at $k = -N$.

(3) While Equation 5.6 predicted very accurately the abundance of randomly placed species, it underestimated the abundance of aggregated species. Equation 5.11 greatly improved the accuracy of the estimation for real species because it accounts for aggregation with the addition of parameter k. Nevertheless, both simulated and observed data showed that the accuracy of the estimates consistently decreases with scale (i.e., the mapping resolution). The underestimation is particularly serious for abundant species as evidenced in Tables 5.2 and 5.3. Similarly, high intensity of aggregation would result in a poor estimation. In the extreme case, if a species is so highly aggregated that all of the individuals are clustered in a single cell, none of the methods could differentiate this species from the one that has only one individual and occurs also in a single cell.

(4) Equation 5.13 is a fundamental occupancy-abundance model that unifies occupancy (p), species abundance (N), the spatial pattern of the species (k), mapping scale (a) and the extent of study area (A) into a single mathematical form. The model suggests that occupancy (or species detectability) depends not only on the abundance of the species but also on its spatial pattern and the size of sample unit. This finding would help us understand the factors that may influence the detectability of species in a field survey and, thus, design the survey in order to maintain a desirable level of detectability (e.g., to calculate the size of sample unit). The derived variance for the occupancy p can be used to quantify the uncertainty in the detectability.

(5) Two obvious questions need to be answered: How can we further incorporate the information on scale (i.e., mapping resolution) and aggregation to improve the estimation? What scale should be used for mapping a distribution to ensure a certain level of accuracy? The questions about scale appear to be more challenging. Answers necessarily depend on the life history properties of organisms. For example, for insects with small body sizes and highly aggregated distribution, a small mapping scale compatible with the size of the insects should be used (e.g., in centimeters or a few meters), while for large body trees, a relatively large mapping

scale may be used (e.g., in 10 or 100 meters). The underestimation caused by strong aggregation may be solved by some ad hoc methods. The aggregated mapping method used in this chapter (Figure 5.1b, c) assumed that k in Equation 5.11 was constant. This assumption may be relaxed by correcting the k by comparing the observed map with its random counterpart.

(6) This study deals only with exhaustive survey of a distribution map as defined by Equation 5.1. It will also be interesting and useful to consider Equation 5.1 as a random sample from a distribution map. If one knows that a species is randomly and independently distributed, the classical occupancy estimate Equation 5.6 can be applied to estimating abundance in this sampling scheme. But in other situations where the random and independent assumption does not hold, the estimation of abundance is a challenging task. This is certainly an interesting problem deserving further investigation.

APPENDIX: DERIVATION OF OCCUPANCY-ABUNDANCE MODELS

A. The Random Placement Occupancy Model

Assume a distribution map of m occupied cells out of M total number of cells. Let u ($= M - m$) be the number of empty cells, and let E_i be the event that the i^{th} cell is empty and $\overline{E}_i$ be the event complementary to E_i. Then the probability that *one particular*, say the first, cell is empty is $p(E_1) = (1 - \frac{1}{M})^N$ which is equivalent to $n = 0$ in the binomial distribution Equation 5.2, or obtained by replacing A in Equation 5.3 by $a \times M$.

The probability that *two particular*, say the first two, cells are empty is $p(E_1 E_2) = (1 - \frac{2}{M})^N$. This probability can again be derived from Equation 5.2 with $n = 0$ by replacing 1 by 2 since there are *two* empty cells, or by replacing a by $2 \times a$ in Equation 5.3. Similarly, the probability that u *particular* cells are empty is $p(E_1 ... E_u) = (1 - \frac{u}{M})^N$.

Here we shall not be interested in a *particular* set of cells but the number of u empty cells given N balls being placed into M cells. From Figure 5.2, we know there are $\binom{M}{u}$ possible combinations for u (out of M) empty cells. Thus the probability that there are u empty cells is $p(u) = \binom{M}{u} p(E_{i_1} ... E_{i_u} \overline{E}_{j_u} ... \overline{E}_{j_{M-u}})$. It is equivalent to

$$p(u) = \binom{M}{u} p(E_1 ... E_u \overline{E}_{u+1} ... \overline{E}_M). \tag{A1}$$

Because E_1, …, E_u, $\overline{E}_{u+1}$, …, $\overline{E}_M$ are independent events, by some probability operations, we arrived at

$$\begin{aligned} p(E_1 \ldots E_u \overline{E}_{u+1} \ldots \overline{E}_M) &= \left(1-\frac{u}{M}\right)^N \sum_{i=0}^{M-u} (-1)^i \binom{M-u}{i} \left(1-\frac{i}{M-u}\right)^N \\ &= \sum_{i=0}^{M-u} (-1)^i \binom{M-u}{i} \left(1-\frac{u+i}{M}\right)^N . \end{aligned} \tag{A2}$$

Finally, the probability that there are u empty cells given N balls randomly and independently placed into M cells is derived by substituting Equation A2 into Equation A1, i.e.,

$$p(u) = \binom{M}{u} \sum_{i=0}^{M-u} (-1)^i \binom{M-u}{i} \left(1-\frac{u+i}{M}\right)^N , \text{ for } u = 0, 1, \ldots, M. \tag{A3}$$

The factorial moment of the number of u empty cells of the probability mass function equation A3 is known to be (Johnson et al. 1993, p. 415)

$$\lambda_{[r]} = \frac{M!}{(M-r)!}(1-\frac{r}{M})^N ,$$

where $\lambda_{[r]} = E(\frac{u!}{(u-r)!})$. Thus the expectation and variance of the number of empty cells are

$$E(u) = \lambda_{[1]} = M(1-\frac{1}{M})^N \tag{A4}$$

$$\begin{aligned} V(u) &= \lambda_{[2]} + E(u) - E^2(u) \\ &= M(M-1)(1-\tfrac{2}{M})^N + M(1-\tfrac{1}{M})^N - M^2(1-\tfrac{1}{M})^{2N}. \end{aligned} \tag{A5}$$

The variance for the number of occupied cells m is the same as Equation A5 for a given map with fixed M because $V(u) = V(M-m) = V(m)$.

Given an occurrence map, it is obvious that the moment estimate of $E(u)$ is simply M - m. Hence, the estimate of N can be solved from Equation A4 as

$$\hat{N} = \frac{\ln(u/M)}{\ln(1-1/M)} = \frac{\ln(1-m/M)}{\ln(1-1/M)} . \tag{A6}$$

The approximate variance of the abundance estimate $\hat{N}$ in Equation A6 can be easily obtained by applying the delta method to Equation A6, i.e.,

$$V(\hat{N}) = \left[\frac{\partial(\hat{N}(u))}{\partial u}\right]^2_{u=E(u)} V(u), \tag{A7}$$

where $V(u)$ is as Equation A5 and $\hat{N}$ is as Equation A6. The derivative is evaluated at $E(u)$ of Equation A4. The variance so obtained is

$$V(\hat{N}) = \frac{V(u)}{\left[M(1-1/M)^N \ln(1-1/M)\right]^2}. \tag{A8}$$

B. The Contagious Occupancy Model

For the contagious process that generates the negative binomial distribution, Barton and David (1959) show that the distribution of the number of empty cells u has probability mass function:

$$p(u) = \frac{\binom{M}{u}}{(kM+N-1)^{(N)}} \sum_{i=0}^{M-u} (-1)^i \binom{M-u}{i} \left[kM - k(u+i) + N - 1\right]^{(N)}, \tag{B1}$$

where $i^{(j)} = \dfrac{i!}{(i-j)!}$, k is the aggregation parameter of the negative binomial distribution, N is the (unknown) number of organisms of a species distributed in a defined area with size A, M is the total number of cells dividing A.

The r^{th} factorial moment of u is

$$\lambda_{[r]} = M^{(r)} \frac{(kM-1)!(kM-kr+N-1)!}{(kM+N-1)!(kM-kr-1)!} \tag{B2}$$

from which the expectation and variance of u can be found. However, Barton and David (1959) show that, even for relatively small M and N, the pmf given by Equation B1 can be well approximated by a normal distribution with mean and variance:

$$E(u) = M\left(1 + \frac{N}{Mk}\right)^{-k} \tag{B3}$$

$$V(u) = Me^{-2\mu}\left(e^{\mu} - 1 - \mu\right) \tag{B4}$$

where $\mu = \dfrac{N}{M}$.

Given a binary map, it is straightforward that the observed first moment estimate of $E(u)$ is $M - m$. Therefore, from Equation B3 the moment estimate of N, Equation 5.10, is resulted. The variance of the estimate $\hat{N}$ of Equation 5.10 can be derived from Equation B3 and Equation B4 using the delta method following Equation A7:

$$V(\hat{N}) = \frac{1}{M^2}\left(1 + \frac{N}{Mk}\right)^{2k+2} V(u) \tag{B5}$$

where $V(u)$ is given by Equation B4.

REFERENCES

Barton, D. E., and F. N. David. 1959. Contagious occupancy. Journal of the Royal Statistical Society, Series B. 21:120-133.

Barton, D. E., and F. N. David. 1962. Combinatorial Chance. Charles Griffin, London, UK.

Brown, J. H. 1984. On the relationship between abundance and distribution of species. American Naturalist 124:255-279.

Brown, J. H. 1995. Macroecology. University of Chicago Press, Chicago.

Feller, W. 1967. An Introduction to Probability Theory and Its Applications. Wiley, New York.

Gaston, K. J. 1994. Rarity. Chapman & Hall, London, UK.

Gaston, K. J., and T. M. Blackburn. 2000. Pattern and Process in Macroecology. Blackwell, London, UK.

Hanski, I., J. Kouki, and A. Halkka. 1993. Three explanations of the positive relationship between distribution and abundance of species. Pages 108-116 *in* R. E. Ricklefs and D. Schluter, editors. Species diversity in ecological communities: historical and geographical perspectives. University of Chicago Press, Chicago.

Harte, J., T. Blackburn, and A. Ostling. 2001. Self-similarity and the relationship between abundance and range size. American Naturalist 157:374-386.

Hartley, S. 1998. A positive relationship between local abundance and regional occupancy is almost inevitable (but not all positive relationships are the same). Journal of Animal Ecology 67:992-994.

He, F., and K. J. Gaston. 2000. Estimating species abundance from occurrence. American Naturalist 156:553-559.

He, F., and P. Legendre. 2002. Species diversity patterns derived from species-area models. Ecology 83:1185-1198.

He, F., K. J. Gaston, and J. Wu. 2002. On species occupancy-abundance models. Ecoscience 23:503-511.

He, F., P. Legendre, and J. V. LaFrankie. 1997. Distribution patterns of tree species in a Malaysian tropical rain forest. Journal of Vegetation Science 8:105-114.

Holt, A. R., K. J. Gaston, and F. He. 2002. Occupancy-abundance relationships and spatial distribution: a review. Basic & Applied Ecology 3:1-13.

Johnson, N. L., S. Kotz, and A. W. Kemp. 1993. Univariate Discrete Distributions. Wiley, New York.

Kolasa, J. 1989. Ecological systems in hierarchical perspective: breaks in community structure and other consequences. Ecology 70:36-47.

Kolchin, V. F., B. A. Sevast′yanov, and V. P. Chistyakov. 1978. Random Allocations. Winstons & Sons, Washington, D.C.

Kunin, W. E. 1998. Extrapolating species abundance across spatial scales. Science 281:1513-1515.

Kunin, W. E., S. Hartley, and J. L. Lennon. 2000. Scaling down: on the challenging of estimating abundance from occurrence patterns. American Naturalist 156:560-566.

Leitner, W. A., and M. L. Rosenzweig. 1997. Nested species-area curves and stochastic sampling: a new theory. Oikos 79:503-512.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

Manokaran, N., J. V. LaFrankie, K. M. Kochummen, E. S. Quah, J. E. Klahn, P. S. Ashton, and S. P. Hubbell. 1999. The Pasoh 50-ha Forest Dynamics Plot: 1999 CD-ROM Version. Forest Research Institute of Malaysia, Kepong, Malaysia.

Maurer, B. A. 1990. The relationship between distribution and abundance in a patchy environment. Oikos 58:181-189.

Pielou, E. C. 1957. The effect of quadrat size on the estimation of the parameters of Neyman's and Thomas' distributions. Journal of Ecology 45:31-47.

Plotkin, J. B., and H. C. Muller-Landau. 2002. Sampling the species composition of a landscape. Ecology 83:3344-3356.

Taylor, L. R., I. P. Woiwod, and J. N. Perry. 1978. The density dependence of spatial behaviour and the rarity of randomness. Journal of Animal Ecology 47:383-406.

Warren, M., M. A. McGeoch, and S. L. Chown. 2003. Predicting abundance from occupancy: a test for an aggregated insect assemblage. Journal of Animal Ecology 72:468-477.

Wright, D. H. 1991. Correlations between incidence and abundance are expected by chance. Journal of Biogeography 18:463-466.