Web-based Supplementary Materials for

Estimating the encounter rate variance in distance sampling,

by R. M. Fewster, S. T. Buckland, K. P. Burnham,

D. L. Borchers, P. E. Jupp, J. L. Laake, L. Thomas

Web Appendix A: Apparent Contagion Model

The Poisson-Gamma apparent contagion model is

$$E(n_i) = \beta l_i\,; \qquad \mathrm{var}(n_i) = \beta l_i + \gamma l_i^2\,; \qquad n_1, \ldots, n_k \text{ uncorrelated.} \tag{18}$$

Define $L = \sum_{i=1}^k l_i$ and $S = \sum_{i=1}^k l_i^2$. The line lengths $l_1, \ldots, l_k$ are regarded as fixed predictor variates. Under model (18), the true variance is $\mathrm{var}\,(n/L) = \beta/L + \gamma S/L^2$. We propose an estimator of the form

$$\widehat{\mathrm{var}}_{R4}\left(\frac{n}{L}\right) = \alpha \sum_{i=1}^k l_i^\phi \left(\frac{n_i}{l_i} - \frac{n}{L}\right)^2, \tag{19}$$

for values (ϕ, α) to be found. Expanding and taking model-expectations of the right-hand side of (19) gives (after some algebra):

$$E\left\{\alpha \sum_{i=1}^k l_i^\phi \left(\frac{n_i}{l_i} - \frac{n}{L}\right)^2\right\} = \beta\left[\alpha \sum_{i=1}^k l_i^\phi \left\{\frac{1}{l_i} - \frac{1}{L}\right\}\right] + \gamma\left[\alpha \sum_{i=1}^k l_i^\phi \left\{1 - \frac{2l_i}{L} + \frac{S}{L^2}\right\}\right]. \tag{20}$$

To gain a model-unbiased estimator of $\mathrm{var}(n/L)$, we equate the coefficients of β and γ in (20) with those of $\mathrm{var}\,(n/L) = \beta/L + \gamma S/L^2$. This yields two equations to be solved

simultaneously for ϕ and α. Manipulation provides the following expressions for ϕ and α:

$$v(k,\phi) = \sum_{i=1}^{k} l_i^{\phi} \left\{ \frac{1}{l_i} - \frac{2}{L} + \frac{2l_i}{S} - \frac{L}{S} \right\} = 0\,; \tag{21}$$

$$\alpha = \left\{ \sum_{i=1}^{k} l_i^{\phi} \left(\frac{L}{l_i} - 1 \right) \right\}^{-1}. \tag{22}$$

Here, $v(k,\phi)$ is defined as the left-hand side of (21). The problem reduces to solving equation (21) for ϕ. Although (21) can be solved numerically, solutions do not always exist for small k (e.g. $k < 10$). However, we can show that as $k \to \infty$, $v(k,\phi) > 0$ for $\phi < 2$ and $v(k,\phi) < 0$ for $\phi > 2$. Thus, for large k, at least one solution to (21) exists, and any solution converges to $\phi = 2$. This enables us to derive a closed-form approximation to $\widehat{\text{var}}_{R4}(n/L)$. Writing $\phi = 2 + \epsilon$, and constructing the first-order Taylor expansion of the equation $v(k,\phi) = 0$ about $\phi = 2$, gives the result for ϕ in the text.

Web Appendix B: Point Transect Sampling

Suppose there are k randomly placed points, each of which is visited t times. The design- and model-derived analogues of $\widehat{\text{var}}_{R2}$ and $\widehat{\text{var}}_{R3}$ are both:

$$\widehat{\text{var}}_{P1}\left(\frac{n}{kt}\right) = \widehat{\text{var}}\left(\frac{\overline{n}}{t}\right) = \frac{1}{t^2k(k-1)}\sum_{i=1}^{k}(n_i-\overline{n})^2, \tag{23}$$

where $n_i = \sum_{j=1}^{t} n_{ij}$, $n = \sum_{i=1}^{k} n_i$, $\overline{n} = n/k$, and n_{ij} is the number of objects detected at point i on visit j. This method assumes independence of $n_1, \ldots, n_k$, but does not assume that different counts at the same point are independent.

Now suppose that point i is visited t_i times, where t_i can differ between points. This is similar to the line transect case with varying line lengths l_i, but with the important difference that t_i is chosen by the practitioner; it is not an observed piece of auxiliary information dictated by the point's position, as is the case with l_i. Lines with large l_i sample a large part of the survey region, so design-derived estimators allot them higher weight than short lines. By contrast, in point transect sampling, a design-derived estimator should allot all points equal weight, regardless of t_i. The design-derived encounter rate estimator ($\widehat{\text{er}}$) is therefore the sample mean of the response n_i/t_i. This means that $\widehat{\text{er}} = 1/k\sum_{i=1}^{k} n_i/t_i$, and the design-derived variance estimator is simply the sample variance:

$$\widehat{\text{var}}_{P2}\left(\frac{1}{k}\sum_{r=1}^{k}\frac{n_r}{t_r}\right) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\frac{n_i}{t_i} - \frac{1}{k}\sum_{r=1}^{k}\frac{n_r}{t_r}\right)^2. \tag{24}$$

Alternatively, we can use a model-based framework with the model $E(n_i) = \beta t_i$ and $\text{var}(n_i) = \sigma^2 t_i$. This model assumes that encounter rate β is constant throughout the region, so points with large t_i contribute more information about β than points with small t_i. For this model, the BLUE for encounter rate is $\widehat{\text{er}} = \widehat{\beta} = n/T$, where $T = \sum_{i=1}^{k} t_i$. The BLUE-based general formula of Thompson (2002:81) provides a model-unbiased variance estimator:

$$\widehat{\text{var}}_{P3}\left(\frac{n}{T}\right) = \frac{1}{T(k-1)}\sum_{i=1}^{k} t_i\left(\frac{n_i}{t_i} - \frac{n}{T}\right)^2. \tag{25}$$

This is equivalent to equation (3.79) of Buckland et al. (2001:79). Both $\widehat{\text{var}}_{P2}$ and $\widehat{\text{var}}_{P3}$ reduce to $\widehat{\text{var}}_{P1}$ when $t_i = t$ for all i. It is rare in practice for the number of visits per point to vary.

When the point transect survey design is systematic, a post-stratification scheme with non-overlapping strata can reduce bias in the encounter rate variance in the same way as for line transects. However, post-stratification using overlapping strata is harder to implement for point transect surveys. With a 2-dimensional arrangement of points there might be no obvious way of choosing which strata should overlap, especially if the region is irregularly shaped or point spacing is only systematic in one direction. An alternative is due to D'Orazio (2003), who noted that Wolter's 1-dimensional estimator for overlapping strata, v_2, is equal to the naive simple random sampling estimator (Wolter's v_1) multiplied by half the Durbin-Watson statistic ($D/2$), where D measures serial correlation in one dimension. Using Geary's spatial autocorrelation index c as the the 2-dimensional analogue of $D/2$, D'Orazio suggested the systematic variance estimator cv_1, which for equal-effort point transect encounter rate is:

$$\widehat{\text{var}}_{P4}\left(\frac{\overline{n}}{t}\right) = \frac{cv_1}{t^2} = \frac{\sum_{i=1}^{k}\sum_{j\neq i}^{k}(n_i - n_j)^2\delta_{ij}}{2kt^2\sum_{i=1}^{k}\sum_{j\neq i}^{k}\delta_{ij}},$$

where $\delta_{ij} = 1$ if points i and j are adjacent, and 0 otherwise. Note that the method can also be applied to line transect surveys in which lines are short and of equal length, and positioned on a regular grid. Stevens and Olsen (2003) developed a local neighborhood variance estimator that might also perform well in these cases.

To illustrate the point transect estimators, we use a survey of Scottish songbirds (Buckland, 2006), in which different point transect methodologies are compared. The survey involves four species: European chaffinch (*Fringilla coelebs*), great tit (*Parus major*), European robin (*Erithacus rubecula*) and winter wren (*Troglodytes troglodytes*). The survey

region was covered by 32 points spaced systematically along lines (Web Figure 1). Each point was covered twice during the survey period. We use data from the 'snapshot' point transect method (Buckland et al., 2001:173), pooled across the two visits, to illustrate the different encounter rate variance estimates with different degrees of post-stratification.

Results are shown in Web Figure 2. Confidence intervals assume a log-Normal distribution for encounter rate, replacing the Normal z_α by the Student t_{df} quantile. The 95% confidence intervals are $(\widehat{\text{er}}/C,\ \widehat{\text{er}} \times C)$ (Buckland et al., 2001:77), where $\widehat{\text{er}}$ is the encounter rate estimate, $C = \exp\left[t_{df}\sqrt{\log\left\{1 + \widehat{\text{se}}(\widehat{\text{er}})^2/\widehat{\text{er}}^2\right\}}\right]$, and t_{df} is the upper 0.025 point of the Student's t-distribution with d.f.$= \sum_{h=1}^{H}(k_h - 1)$. Here, H is the number of post-strata, and k_h is the number of points in post-stratum h. For unstratified estimators, $H = 1$.

Web Figure 2 shows that the point transect results are not strongly affected by whether, and to what degree, stratification is used. Estimator $P4$, which uses the correction of D'Orazio (2003), usually yields some improvement in precision over the unstratified $P1$. However, we have not examined its suitability in any theoretical sense. Increasing stratification brings little change for inference from these data. Nonetheless, in view of the conclusive results for line transect estimators, we recommend that the post-stratified estimators are used for systematic point transect surveys where possible.

Web Appendix C: Simulation study

Here we describe in detail the simulation procedures and results from Section 6, for the line transect variance estimators derived in Sections 3 and 4. The simulations use a set of heavily trended populations. We primarily use a triangular region, in which there is substantial variability in line lengths. Results are also presented for a circular region and a rectangular region. Trends in object density can be highly correlated with line length, to provide a stringent test of the estimators in extreme conditions.

For the triangular region, all simulated object populations occupy half of the unit square, corresponding to the region $y \leq x$. Object x and y coordinates are generated from independent beta distributions, discarding any objects for which $y > x$ until a decided number of N object positions are obtained. Similarly, object positions are generated for the circular region, where the circle is centered on point $x = y = 0.5$ with radius 0.5, and for the rectangular region. We use six different sets of beta parameters, for each of which we generate 10,000 population realizations of N objects. The number of objects is variously $N = 1000$ and $N = 10{,}000$. This is the repeated survey framework, mimicking a population of mobile animals in which the population size is fixed but the spatial locations change according to an underlying spatial density.

Figure 1 shows a single population realization from each beta parameterization for the triangle region. For each realization, we generate a single survey of $k = 20$ vertical transects. All objects within horizontal distance w of a transect line are available for detection from that line, including those above the upper end of the line. The line length is measured strictly inside the region. To simulate detections, we use a half-normal detection function, with scale parameter σ and strip half-width w. Objects in the search strip have average detection probability $P_a = w^{-1} \int_0^w \exp\{-r^2/(2\sigma^2)\}\, dr$. We can make $P_a = 1$ by using large enough σ. The sampling fraction is $E(2wL/A) \simeq 40w$ for each region.

Figures 1 to 4 show results for encounter rate variance and overall density variance, $\widehat{\text{var}}(n/L)$ and $\widehat{\text{var}}(\hat{D})$, for a wide range of settings. Confidence interval coverage is also reported, rounded to the nearest percent. Confidence intervals assume a log-Normal distribution for n/L or $\hat{D}$ respectively, replacing the Normal z_α by the appropriate Student t_{df} quantile. For encounter rate n/L, the 95% confidence intervals are $(nL^{-1}/C,\ nL^{-1}\times C)$ (Buckland et al., 2001:77), where $C = \exp\left[t_{df}\sqrt{\log\left\{1+\widehat{\text{se}}(nL^{-1})^2/(nL^{-1})^2\right\}}\right]$. Here, t_{df} is the upper 0.025 point of the Student's t-distribution with degrees of freedom $\text{df}_{er} = \sum_{h=1}^{H}(k_h-1)$, with H the number of post-strata and k_h the number of lines in post-stratum h. For unstratified estimators, $H = 1$ and $k_h = 20$ lines. Under the post-stratification scheme with non-overlapping strata, $H = k/2 = 10$ and $k_h = 2$ lines per post-stratum. With overlapping strata, $H = k - 1 = 19$ and $k_h = 2$ for all post-strata.

For density $\hat{D}$, the 95% log-Normal confidence intervals are $(\hat{D}/C,\ \hat{D}\times C)$, where $C = \exp\left[t_{df}\sqrt{\log\left\{1+\widehat{\text{se}}(\hat{D})^2/\hat{D}^2\right\}}\right]$. Here, $\widehat{\text{se}}(\hat{D})^2$ is obtained through (1) using the selected estimator for $\text{cv}(n/L)$, and using the information matrix of the conditional likelihood of detection distances for $\text{cv}(\hat{P}_a)$; and t_{df} is the upper 0.025 point of the Student's t-distribution with degrees of freedom given by the Satterthwaite approximation (Buckland et al., 2001:78):

$$\text{df}_D = \frac{\text{cv}(\hat{D})^4}{\text{cv}(nL^{-1})^4/\text{df}_{er} + \text{cv}(\hat{P}_a^{-1})^4/(n-1)}\,,$$

where df_{er} is the encounter rate degrees of freedom given above, and n is the total number of observations.

Randomly placed transect lines

Estimators $R1$ to $R4$, from equations (2), (3), (5), and (7), are assessed with random placement of transects (Figure 1). For each simulated survey, transects are generated independently with x-coordinates from Uniform$(w, 1-w)$. Lines are positioned between w and $1-w$ to ensure that search strips do not extend beyond the edge of the region. This

may cause slight bias in the estimates of encounter rate itself, due to a lower sampling probability for objects in the x-intervals $[0, w)$ and $(1 - w,\ 1]$. However, it ensures that the variance of the encounter rate is not influenced by search strips that extend beyond the region.

The boxplots in Figure 1 show summary results for $\widehat{\text{se}}(n/L) = \sqrt{\widehat{\text{var}}(n/L)}$ for each of the estimators. The sampling fraction is a realistic 0.01. We use perfect detection ($P_a = 1$) so the encounter rate variance is not complicated by detection variance. This corresponds to plot sampling. The thin horizontal lines across the boxes show the mean $\widehat{\text{se}}(n/L)$. The thick horizontal line is the sample value of $\text{sd}(n/L)$ from the 10,000 simulations. Good performance of the variance estimator is seen when the boxplot mean coincides with the sample $\text{sd}(n/L)$. The boxplots also show the variability of the standard error estimators.

Estimator $R1$, which does not take account of different line lengths, performs poorly and can be biased in either direction. The model-derived $R3$, based on the overdispersed Poisson or true contagion model, performs well in Population 1, where the uniformly distributed objects exactly satisfy the model underlying $R3$. However, it is not robust to other object distributions. Variance is overestimated when short lines carry the highest object densities, and underestimated when short lines carry the lowest object densities.

The best-performing estimators are the design-derived $R2$, and the Poisson-Gamma or apparent contagion model-derived $R4$. The good performance of $R2$ is expected, because the repeated survey framework is primarily design-based. Our new model-derived estimator, $R4$, exhibits a very slight improvement over $R2$ in terms of bias, although at the expense of a very slight deterioration in precision. Estimator $R4$ was consistently the best for all values of w and σ we investigated. In view of the simplicity of $R2$ and the tiny gain from using $R4$, we favor estimator $R2$ as the recommended estimator. However, it is useful to note that the Poisson-Gamma model (6) produces an estimator $R4$ that is

robust to the extreme trends exhibited here, and this model might be useful for future work.

Systematically placed transect lines

The remaining simulations use a systematic survey design instead of random placement of transects. The x-coordinate of the first transect line is generated as Uniform$(w, w+d)$, and subsequent lines are placed at fixed spacing d, where $d = (1-2w)/k$. Although the data are systematic, we wish to assess the random-based estimators $R2$, $R3$, and $R4$, as well as the new stratified estimators $S1$, $S2$, $O1$, $O2$, and $O3$. The traditional method of estimating var(n/L) from systematic designs has been to use $R3$, so we include $R2$, $R3$, and $R4$ to study the impact of this choice in the heavily trended simulation setting presented. Estimator $R1$ is omitted because it performs poorly and lacks justification.

Figure 2 shows results for the same settings as Figure 1, with perfect detection ($P_a = 1$) and small sampling fraction (0.01). In Figure 3, the effects of imperfect detection ($P_a = 0.60$) and high sampling fraction (0.40) are examined for different region shapes. Figure 4 further shows the effect of estimating overall density variance, var$(\hat{D})$, using equation (1), in which $\hat{P}_a = w^{-1} \int_0^w \exp\{-r^2/(2\hat{\sigma}^2)\}\, dr$, and $\hat{\sigma}$ is the maximum likelihood estimate. The sampling fraction ranges from moderate (0.1) to very high (0.8).

The first important feature is the bias seen in estimators $R2$, $R3$, and $R4$, for the trended populations 2 to 6. The bias is high for a small sampling fraction (Figure 2), and extreme for large sampling fractions (Figure 4). This is because systematic designs gain greatest efficiency over random designs when the sampling fraction is large, because systematic designs do not allow overlapping transects. At large sampling fractions, random designs will have substantial overlap (unless they randomly sample non-overlapping strips, as in Buckland et al. 2001:235), so they are effectively sampling a smaller fraction of the region. However, even when this effect is negligible at sampling fraction 0.01, a comparison of the

true variances on Figures 1 and 2 shows that systematic designs are still more efficient than random designs for highly trended populations. Estimators $R2$, $R3$, and $R4$ cannot be assumed to have good properties under systematic designs.

The post-stratification approach greatly reduces the bias in variance estimation. At small sampling fractions (Figure 2) the bias is almost completely eliminated. For larger sampling fractions (Figures 3 and 4), some bias remains, usually positive. This is expected at large sampling fractions because there is an effect of finite population sampling. However, other simulations (not reported here) show that incorporating a finite population correction leads to substantial underestimates of variance in the repeated survey framework, which is important to avoid. In most real survey situations, the sampling fraction is less than 0.1 and this issue does not arise. Figures 3 and 4 show that the post-stratified estimators nonetheless provide a dramatic improvement over the random-based estimators when the sampling fraction is high, and the remaining bias is not of great concern.

Estimators $S1$ and $O1$, and estimators $S2$ and $O2$, show the effects of switching from non-overlapping strata to overlapping strata. In each case, the increased d.f. of the overlapping scheme produces an improvement in precision, but little if any change in bias. The model-derived estimator, $O3$, is not recommended, and exactly mirrors the direction of bias in the similar estimator $R3$. There is very little difference in performance between estimators $S1$ and $S2$, or between estimators $O1$ and $O2$.

Our simulation results, including many not shown, indicate that trended populations with disparate line lengths are problematic for estimator $R3$, but the problem is overcome by using $R2$ or $R4$. For systematic designs, the random-based estimators are not suitable if the population is highly trended. The post-stratification scheme effectively deals with the problem, especially for the most realistic case of small sampling fractions. At high sampling fractions, we do not have theory that acknowledges finite population effects of

spatial sampling while still allowing for variance in detectability and object positions; however, the post-stratification scheme still represents an important improvement, and high sampling fractions are rare in practice. Explicit spatial modeling would be a promising approach to variance estimation when the sampling fraction is high. Detectability (P_a), region shape, and effect of estimating $\text{var}(\hat{D})$ rather than $\text{var}(n/L)$ have little or no impact on the results. Overall, taking account both of estimator performance and simplicity, we recommend estimator $R2$ for random survey designs. For systematic designs, we recommend $O2$ when the design can accommodate overlapping strata in whole or part, and $S2$ otherwise. Estimator $R2$ is used for the within-stratum variances in $S2$ and $O2$ (eqns (13) and (16)).

Web References

Buckland, S. T. (2006). Point transect surveys for songbirds: robust methodologies. *The Auk* **123**, 345–357.

D'Orazio, M. (2003). Estimating the variance of the sample mean in two-dimensional systematic sampling. *Journal of Agricultural, Biological, and Environmental Statistics* **8,** 280–295.

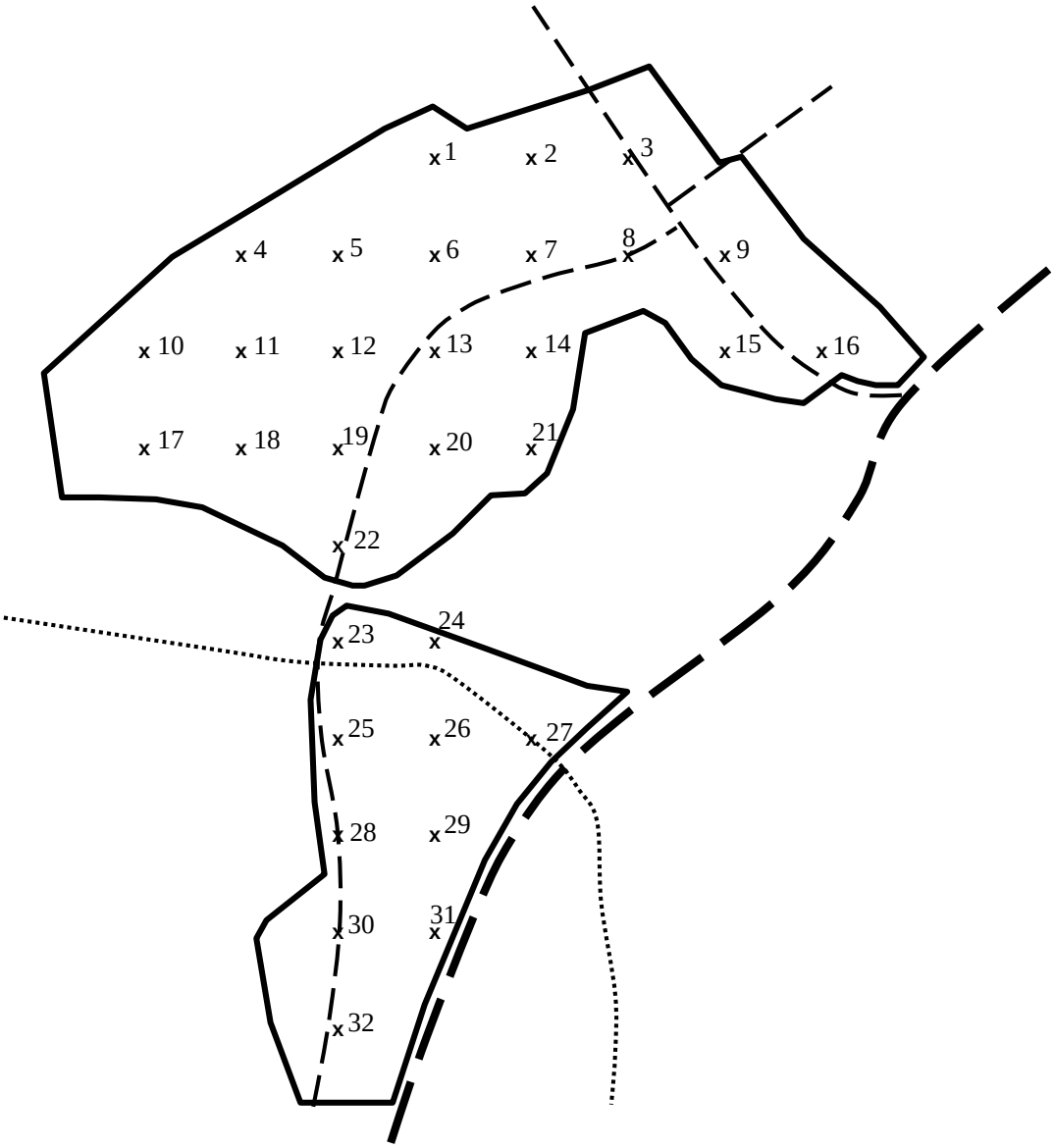
Stevens, D. L. Jr and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* **14,** 593–610.

Web Figure Captions

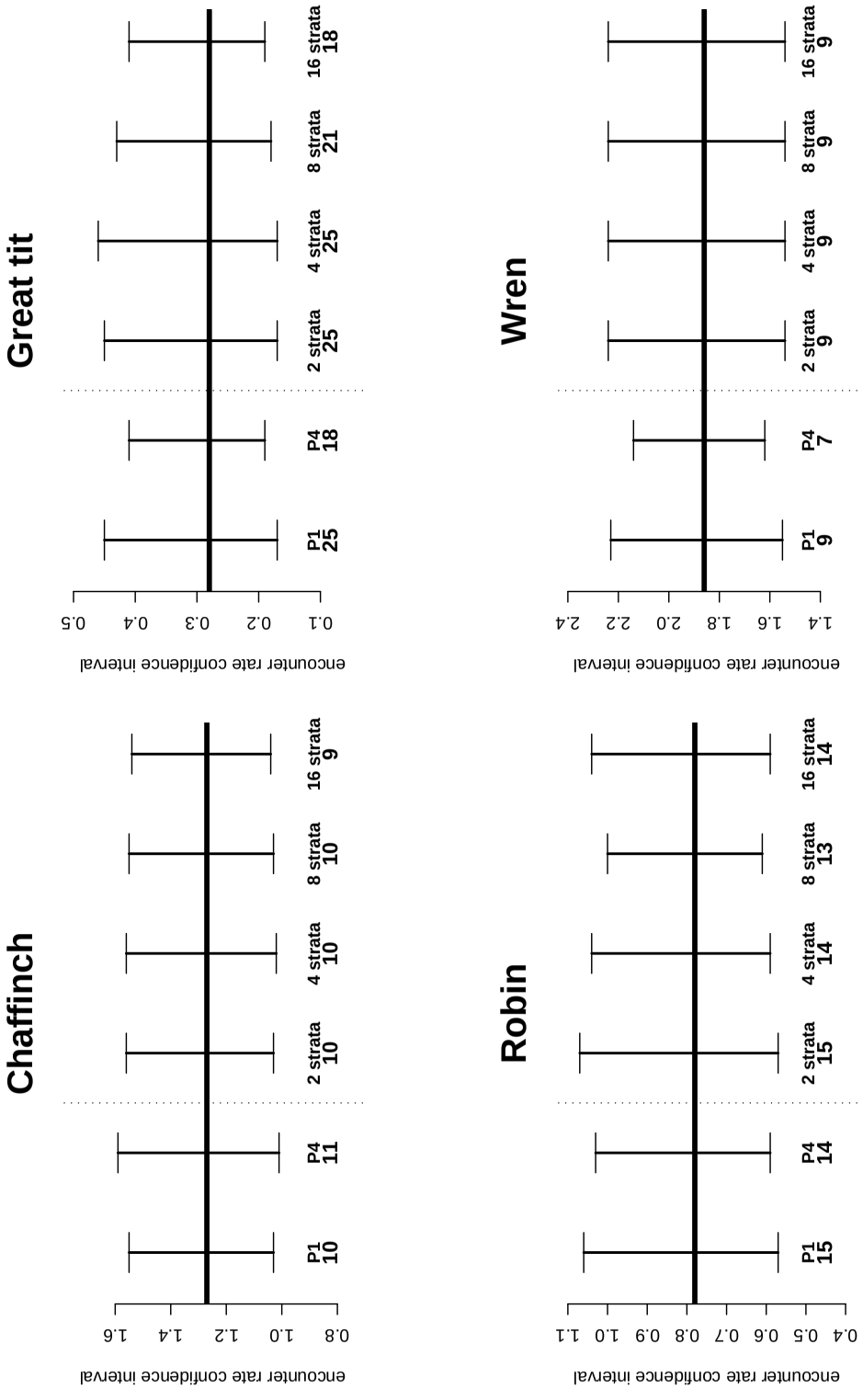
Web Figure 1. The study site at Montrave in Fife, Scotland. The area comprises parkland and mixed woodland. The dotted line is a small stream, the thin dashed lines are tracks, and the thick dashed line is a main road. The points are laid out on a systematic grid with 100m separation, shown by crosses and labeled $1, 2, \ldots, 32$.

Web Figure 2. Estimated encounter rate (thick horizontal line), 95% confidence intervals (vertical bars), and %CVs for point transect surveys for chaffinch, great tit, robin, and wren. Each point was covered twice ($t = 2$). Estimator $P1$ is applied without stratification. Estimator $P4$ is the unstratified point transect estimator corrected by Geary's autocorrelation index, intended to mimic a scheme of overlapping strata in two dimensions. The stratified estimators use $P1$ within each stratum and sum across strata for the final result. Results are shown for two strata ({1-22}, {23-32}); four strata ({1,4-6,10,11,17}, {2,3,7-9,15,16}, {12-14,18-22}, {23-32}); eight strata ({1-3,7}, {4-6}, {8,9,15,16}, {10,11,17,18}, {12-14,21}, {19,20,22}, {23-27}, {28-32}); and 16 strata ({1,6}, {2,3}, {4,5}, {7,14}, {8,9}, {10,17}, {11,12}, {13,20}, {15,16}, {18,19}, {21,22}, {23,24}, {25,28}, {26,27}, {29,31}, {30,32}). Strata were formed as far as possible by combining neighboring points within similar habitats.

Web Figure 1


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Web Figure 2


Great tit
encounter rate confidence interval
0.1 0.2 0.3 0.4 0.5
P1 25
P4 18
2 strata 25
4 strata 25
8 strata 21
16 strata 18
Wren
encounter rate confidence interval
1.4 1.6 1.8 2.0 2.2 2.4
P1 9
P4 7
2 strata 9
4 strata 9
8 strata 9
16 strata 9
Chaffinch
encounter rate confidence interval
0.8 1.0 1.2 1.4 1.6
P1 10
P4 11
2 strata 10
4 strata 10
8 strata 10
16 strata 9
Robin
encounter rate confidence interval
0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
P1 15
P4 14
2 strata 15
4 strata 14
8 strata 13
16 strata 14