Article:

Takashina, N, Kusumoto, B, Beger, M et al. (2018) Spatially explicit approach to estimation of total population abundance in field surveys. Journal of Theoretical Biology, 453. pp. 88-95. ISSN: 0022-5193

https://doi.org/10.1016/j.jtbi.2018.05.013

Spatially explicit approach to estimation of total population abundance in field surveys

Nao Takashina[1]*, Buntarou Kusumoto[2], Maria Beger[3,4] Suren Rathnayake[5], Hugh P. Possingham[3,6]

[1] *Tropical Biosphere Research Center, University of the Ryukyus,*
3422 Sesoko Motobu, Okinawa 905-0227, Japan

[2] *Center for Strategic Research Project, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan*

[3] *ARC Centre of Excellence for Environmental Decisions, School of Biological Sciences, The University of Queensland, St Lucia, QLD 4072, Australia*

[4] *School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK*

[5] *School of Biological Sciences, The University of Queensland, St Lucia, QLD 4072, Australia*

[6] *The Nature Conservancy, 4245 North Fairfax Drive Suite 100 Arlington, VA 22203-1606, USA*

Keywords: Field survey; population estimation; random sampling; spatial point process

---

*Corresponding author

Current address: Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Okinawa 904-0495, Japan

Email: `nao.takashina@gmail.com`

Abstract

The population abundance is fundamental in ecology and conservation biology, which provides essential information for predicting population dynamics and implementing conservation actions. While a range of approaches have been proposed to estimate population abundance based on existing data, data deficiency is ubiquitous. Where there is no information available, a population estimation will rely on labor intensive field surveys. Typically, time is one of the critical constraints in conservation and, often, management decisions must be made quickly under a data deficient situation. Hence, it is important to acquire a theoretical justification for survey method to meet a required estimation precision. However, there is no such a theory available in a spatially explicit context, while spatial considerations are critical to any field survey. Here, we develop a spatially explicit theory for population estimation, which allows us to examine the estimation precision under different survey designs and individual distribution patterns (e.g. random/clustered sampling and individual distribution). We demonstrate that the clustered sampling decreases the estimation precision when individuals form clusters, while sampling designs do not affect the estimation accuracy when individuals are distributed randomly. Regardless of the individual distribution, the estimation precision becomes higher with increasing total population abundance and the sampled fraction. Obtained insights provide theoretical bases for efficient field survey designs in information deficiency situations.

Introduction

Estimating the abundance of populations is important for ecological studies and conservation biology [1–7], as is the role of ecosystem monitoring to observe changes in ecosystems [8–10]. In conservation, such knowledge helps one to estimate the risk of extinction of threatened species [11,12], and to implement effective conservation actions [13].

While methods for statistically inferring population abundance with existing spatial data are well developed [4–6, 14, 15], information on the abundance of threatened or rare species is often rather limited and biased given budgetary constraints and access to remote sites [16,17], requiring further data collection or correction of data biases. For example, Reddy and Davalos [16] examined an extensive data set of 1068 passerine birds in sub-Saharan Africa, and they found that data on even well-known taxa are significantly biased to areas near cities and along rivers. Typically, time is one of the critical constraints in conservation areas facing ongoing habitat loss and environmental degradations [18]. In such cases, management decisions must be made quickly despite often having only limited knowledge of a system [13, 19, 20]. On the other hand, for many ecological studies and ecosystem monitoring programs, data must be accurate enough to be able to detect ecological change [9]. Hence, given time and budgetary constraints and required precision of data, it is desirable to set up an effective survey design to reduce time and effort of sampling.

Ultimately, we face trade-offs between data accuracy, time, and money. To tackle this trade-off and provide generic insights to people designing a population survey, we need to handle different sampling methods, choice of sampling unit scale, and data availability. However, most previous approaches are spatially implicit (e.g., [5, 6, 14, 15, 21]), and it is therefore not straightforward to compare the effect of different survey designs within a single theoretical framework applied. For example, the negative binomial distribution (NBD) is frequently used to describe the underlying individual distribution of a species. In the NBD, the parameter characterizing the degree of spatial aggregation is scale dependent, and needs to be calibrated for each sampling unit scale. However, this procedure is not intuitive and makes consistent comparison between survey designs difficult, as the parameter characterizing aggregation is usually inferred from observed data rather than biological mechanisms [14].

To develop generic insight into field survey performance under data deficient situations, we develop a spatially explicit theory for population abundance estimation, which allows us to consistently examine the estimation precision under various data collection schemes and different sampling scales. Specifically, we examine simple random sampling and cluster sampling [22, 23] as population sampling schemes. Cluster sampling reflects existing geographically biased sampling to some extent, and hence, it is expected to give a general insight into prevalent field survey designs. These sampling schemes are combined with spatial point processes (SPPs), a spatially explicit stochastic model, to reveal effects of different survey designs as well as different individual distribution patterns on the performance of population estimate. SPPs are widely applied in ecological studies due to their flexibility, applicability to many ecological distribution, and availability of biological interpretations [24–30]. Many examples come from studies of plant communities [24–26, 28, 29], but others include studies of coral communities [31], and avian habitat selection to examine distributions of bird nests [32]. Although individual distributions often show clustering patterns in plant and coral communities [25, 33–35], Bayard and Elphick [32] showed no statistical evidence of non-random distributions in avian habitat selection at two salt marshes. Therefore, we examine both clustering and random individual distribution patterns as example. By combining with sampling strategies, we provide the general properties of ”random/clustering sampling + random/clustering individual distributions” without information on target species. Therefore, facing to a data deficient situation, the best one can do is that merely assume if the species is randomly distributed or forming clusters in space to develop sampling designs.

However, the method developed is general enough and suitable for any sampling of organism or location used by an organism (e.g., nest and lek site) that is sedentary in space on a time scale of the field survey where its spatial distributions can be described by SPPs. Hence, the results of general sampling situation discussed may provide generic perspective of sampling designs.

Methods

In this analysis, we consider a situation where there is no prior spatial data available to infer the distribution and abundance of a target species. We assume that our estimate of population size is based only on field surveys where a fraction of sampled units α of the region of concern, W, is surveyed using a sampling unit size, S (Fig. 1: Note we also use the notation R to represent region in general. S is used when we specifically discuss the sampling unit.). We focus on a case where no measurement error occurs in each sampling unit, suggesting that sampling units should be chosen to ensure only trivial sampling errors in practice. It may vary for sampling in different systems. For example, such an area may be larger for counting plant species compared to counting coral species due to different visibility and accessibility of field surveys.

First, we introduce an estimator of population abundance, its expected value and variance, which explicitly accounts for the effect of sampling unit size. These relevance to specific sampling schemes and individuals distribution patterns are the main concern of this paper. Next, we explain some basic properties of spatial point processes (SPPs), and models to describe spatial distribution patterns of individuals. Using this framework, we test our analytical results formula for population estimation.

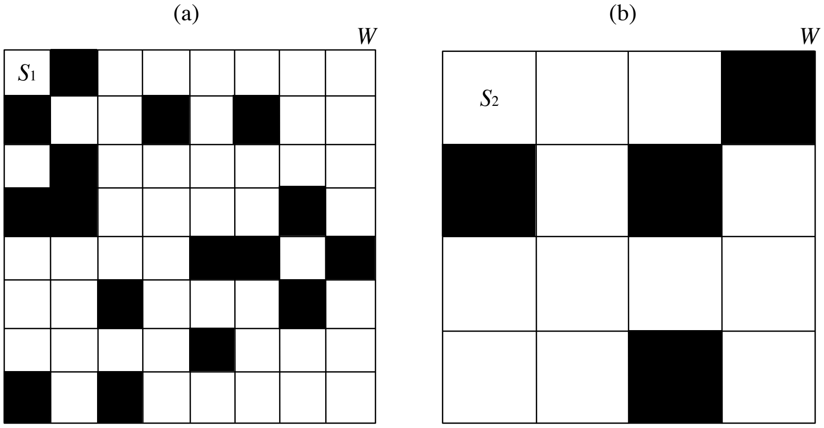

Figure 1: Example of simple random sampling with (a) smaller, and (b) larger sampling unit size, labeled S_1 and S_2, respectively. The whole region of concern W is divided into sampling units with equal size, and a certain fraction α is randomly sampled (shaded unit) without replacement, where all sampling units have the equal probability of being chosen. Essentially, applying larger sampling units corresponds to a cluster sampling. The examples show the case of $\alpha = 0.25$.

Survey design

Given parameters specifying the survey design noted above, a simple random sampling (SRS) without replacement [23] is conducted for collecting count data (Fig 1). In the SRS without replacement, all the sampling units have an equal probability of being chosen. The number of sampling units, N_t, and the sampled units, N_s, change with a sampling unit size, S. We assume all the sampling units have an equal size. With larger sampling units, the degree of the geographical sampling bias increases especially when the fraction of a sampled region is small (Fig 1). This design corresponds to one-stage cluster sampling [23], where either all or none of the area within the larger sampling units are in the sample. It is worth noting, however, that the degree of cluster sampling is relative: any SRS can be considered to be cluster sampling if it is compared to SRS with a smaller sampling unit size. In this article, we simply use these terms to imply that we are using relatively *small* and *large* sampling units.

Population estimator

Following the data collection, we apply the unbiased linear estimator of the population abundance in the region of concern W, $n(W)$ [22,23],

$$\begin{aligned}\hat{n} \mid S &= \frac{N_t}{N_s}\sum_{i}^{N_s} y_i, \\ &= \frac{N_t}{N_s}\sum_{k}^{\infty} n_k k,\end{aligned} \tag{1}$$

where, $\hat{n} \mid S$ is the estimated population abundance given sampling unit size S, y_i is the number of sampled individuals at the ith sampling trial, and n_k is the frequency of the sampled units holding k individuals ($n_k = 0$ for large k because the number of individuals within each sampling unit is finite). Note y_i and n_k change depending on the sampling unit size and underlying spatial point patterns. In the SRS without replacement with the number of sampled units N_s, the frequency n_k is only the random variable, following a multivariate hypergeometric distribution $p(n_k \mid S, N_s)$

with the mean $N_s p(k \mid S)$. Hence, the average population estimation $\hat{n}$ is

$$\begin{aligned} \mathrm{E}[\hat{n} \mid S] &= \frac{N_t}{N_s} \sum_{k}^{\infty} \mathrm{E}[n_k \mid S] k, \\ &= N_t \mathrm{E}[k \mid S]. \end{aligned} \tag{2}$$

The variance of the population estimate under the SRS without replacement is obtained by multiplying the finite population correction (fpc) $:= (N_t - N_s)/(N_t - 1)$ [22] by the variance under the SRS with replacement:

$$\begin{aligned} \mathrm{Var}[\hat{n} \mid S] &= (\mathrm{fpc}) \left(\frac{N_t}{N_s} \right)^2 \Big(\sum_{k}^{\infty} \mathrm{Var}[n_k \mid S] k^2 + \sum_{\substack{k,k' \\ k \neq k'}}^{\infty} \mathrm{Cov}[n_k n_{k'} \mid S] k k' \Big), \\ &= \frac{N_t^2}{N_s} \left(\frac{N_t - N_s}{N_t - 1} \right) \mathrm{Var}[k \mid S], \end{aligned} \tag{3}$$

where, the fact that the probability $p(n_k|S, N_s)$ follows a multinomial distribution with $\mathrm{Var}[n_k|S] = N_s p(k|S, N_s)(1 - p(k|S, N_s))$ and $\mathrm{Cov}[n_k n_{k'}|S] = -N_s p(k|S, N_s) p(k'|S, N_s)$ $(k \neq k')$ [36] are used. Therefore, the variance of the abundance estimate is determined by a constant multiplied by variance of individual numbers in the sampling unit.

Spatial distribution of individuals

To account for explicit spatial distributions of individuals, we use spatial point processes (SPPs) [24, 29]. The underlying models used in our analysis are the homogeneous Poisson process and Thomas process, generating random and cluster distribution patterns of individuals, respectively. Properties of these processes are found in the literature (e.g., [24, 29, 37]) and, hence, we only introduce the properties relevant to our questions.

Homogeneous Poisson process

One of the simplest class of SPPs is the homogeneous Poisson process where the points (i.e. individuals) are placed randomly within the region of concern and the number of points given in the region R, $n(R)$, comes from a Poisson distribution with an average μ_R:

$$\text{Prob}(n(R) = k) = \frac{\mu_R^k}{k!} e^{-\mu_R}, \qquad (k = 0, 1, \dots) \tag{4}$$

where, μ_R is known as the intensity measure [24, 29] defined by

$$\mu_R = \lambda \nu(R), \tag{5}$$

where, $\lambda \coloneqq n(W)/\nu(W)$ is the intensity of individuals in the whole region W [29], and $\nu(R)$ is the area of region R.

Thomas process

The Thomas process, characterizing the clustering pattern of individuals, belongs to the family of Neyman-Scott processes [24, 29]. The Thomas process provides more general framework to address spatial ecological patterns since most species are clumped in nature rather than random [38]. Even though the model assumptions are minimal and does not assume a heterogeneous environment, it creates patterns consistent with species that live in heterogeneous environment (e.g., [25, 28]). The Thomas process is also amenable to an analytical approach, and therefore it is suitable to develop mathematical understanding by minimizing model complexity [24, 25, 28–30]. The Thomas process is obtained by the following three steps:

1. Parents are randomly placed according to the homogeneous Poisson process with a parent intensity λ_p.

2. Each parent produces a random discrete number c of daughters, realized independently and identically.

3. Daughters are scattered around their parents independently with an isotropic bivariate Gaussian distribution with variance σ^2, and all the parents are removed in the realized point pattern.

The intensity of individuals for the Thomas process is [29]

$$\lambda_{th} = \bar{c}\lambda_p, \tag{6}$$

where, $\bar{c}$ is the average number of daughters per parent. To allow population estimate comparisons between the two SPPs, we chose the intensity of the Thomas process so as to have the same average number of individuals within the region of concern W. Namely, the parameters λ_p and $\bar{c}$ satisfy

$$\lambda_{th} = \bar{c}\lambda_p = \lambda. \tag{7}$$

We also assume that the number of daughters per parents c follows the Poisson distribution with the average number $\bar{c}$.

Results

The total number of sampling units and sampled units are $N_t = \nu(W)/\nu(S)$ and $N_s = \lfloor \alpha N_t \rfloor$ respectively, where $\lfloor x \rfloor$ is the greatest integer not larger than x, and α is the fraction of sampled units $(0 \leq \alpha \leq 1)$. We are here interested in how the population estimates deviate from the true value. Therefore, one of the quantities to show these effect may be

$$\frac{\mathrm{E}[\hat{n} \mid S] \pm \mathrm{SE}[\hat{n} \mid S]}{\mathrm{E}[n(W)]}. \tag{8}$$

Note in the analysis below, we use $\lfloor \alpha N_t \rfloor = \alpha N_t$ for simplicity, but this approximation becomes negligible when αN_t is sufficiently large.

Population estimation under the homogeneous Poisson distribution

For the homogeneous Poisson process, $\mathrm{Var}[k|S]$ is equivalent to the variance of the Poisson process with average $\lambda\nu(S)$. Therefore, by substituting this expression into Eq. (3) and with some algebra, we obtain the SE of the population estimate of the homogeneous Poisson process

$$\mathrm{SE}_{po}[\hat{n} \mid S] = \sqrt{n(W)\left(\frac{1}{\alpha}-1\right)\frac{N_t}{N_t-1}}. \tag{9}$$

When the total number of sampling units is sufficiently large ($N_t \gg 1$), we obtain the simpler form

$$\mathrm{SE}_{po}[\hat{n} \mid S] \simeq \sqrt{n(W)\left(\frac{1}{\alpha}-1\right)}. \tag{10}$$

Under such circumstances, the standard error of the abundance estimation is only the function of the expected population total existing in the concerned region $n(W)$ and the sampling fraction α; and does not depend on the sampling unit size. Therefore, we can write $\mathrm{SE}_{po}[\hat{n} \mid S] = \mathrm{SE}_{po}[\hat{n}]$. Due to the term $n(W)^{1/2}$ in $\mathrm{SE}_{po}[\hat{n} \mid S]$, the relative variation from its average decreases with the factor $(1/\alpha-1)^{1/2}n(W)^{-1/2}$. These results were confirmed by numerical simulations, and they show good agreement with analytical results (Fig. 2).

Population estimation under the Thomas process

For the Thomas process, deriving a theoretical form of the variance of individuals given across sampling scales, $\mathrm{Var}[k|S]$, is challenging, although the probability generating functional of the Thomas process is known, e.g., [29]. Instead, we apply an approximated pdf of the Thomas process to obtain an explicit form of $\mathrm{Var}[k|S]$. By assuming that each daughter location has no correlation to its sisters locations, we derive the approximated pdf of the Thomas process (see Appendix for the detailed derivations):

$$p(n \mid S) \quad = \quad \sum_k \mathrm{Po}(k, \lambda_p \nu(S'))\mathrm{Po}(n, k\bar{c}p_d(S)). \tag{11}$$

where, $\mathrm{Po}(k,\lambda)$ is the Poisson distribution with the intensity λ, and $p_d(S)$ is the probability that an individual daughter produced by a parent situated in the region, $S + S_{out}$, falls in S. S_{out} is the surrounding region of S where parents can potentially supply daughters to the region S (See Appendix for the detailed definition of S_{out}). This probability is determined by the dispersal kernel

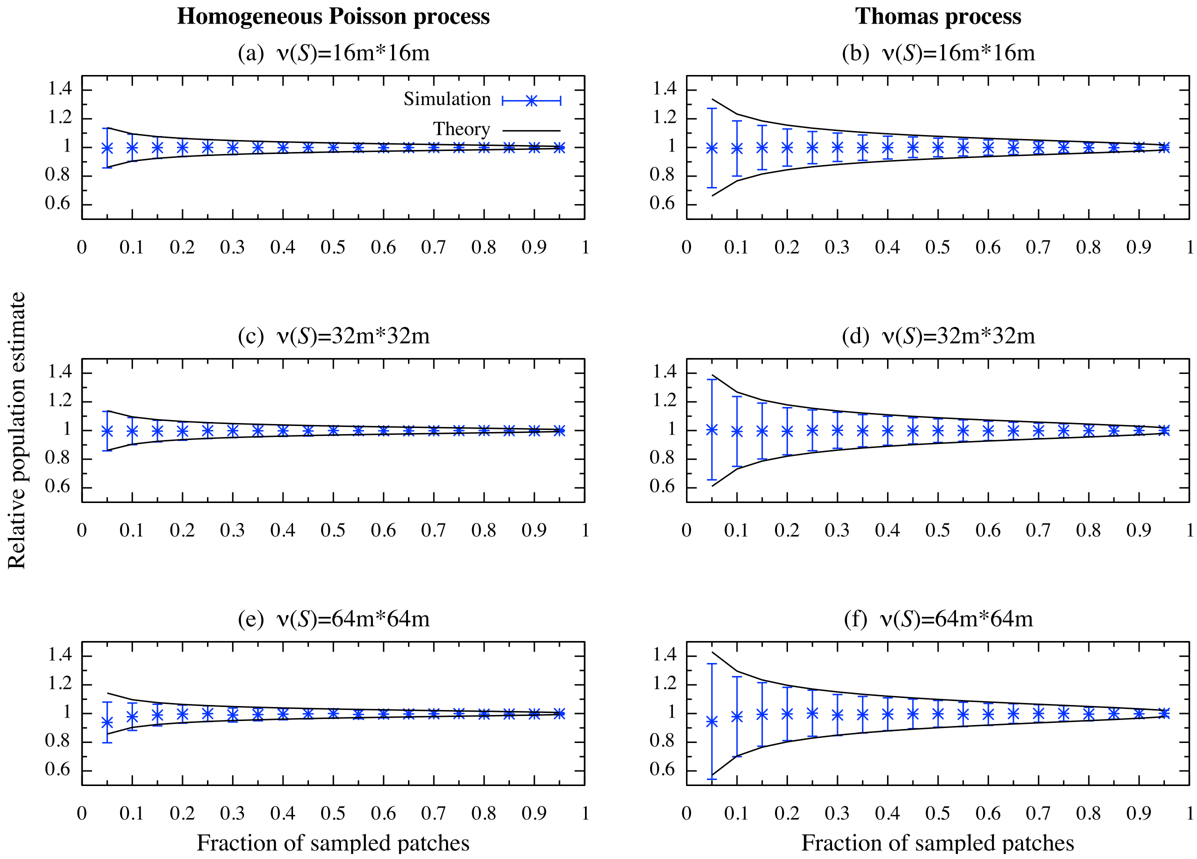

Figure 2: Relative value of the population estimate with the average individuals $\mathrm{E}[n(W)] = 10^3$ under the three sampling scales. Larger sampling area implies more cluster sampling. Each panel shows relative average estimate ± relative standard error (Eq. (8)) of simulation and theoretical results. Relative average estimate for theoretical results is omitted since it is an unbiased estimator. The parameter values used are $\bar{c} = 10$, $\sigma = 10$, and $\nu(W) = 2^{20}\mathrm{m}^2$ (1024m×1024m).

(See Eq. (A.3) in Appendix), and therefore, closely related to dispersal distance of the species. Thomas [39] refers to the form of Eq. (11) as the *double Poisson distribution*, in derivations of her original Thomas model, in which spatial effects are implicitly described. On the other hand, Eq. (11) explicitly handles spatial effect, such as the size of sampling unit S and the effect of dispersal $p_d(S)$. Eq. (11) enables us to derive an approximated form of $\mathrm{SE}_{th}[\hat{n}|S]$ (see Appendix for detailed derivation):

$$\mathrm{SE}_{th}[\hat{n}|S] = \mathrm{SE}_{po}[\hat{n}|S]\sqrt{\frac{\nu(S')}{\nu(S)}p_d(S)\left(1+\bar{c}p_d(S)\right)}. \tag{12}$$

This equation suggests that the standard error of the Thomas process, $\mathrm{SE}_{th}[\hat{n}|S]$, is described by the multiplication of $\mathrm{SE}_{po}[\hat{n}|S]$ and a term characterizing the degree of cluster of the Thomas process. Therefore, the similar discussions made for $\mathrm{SE}_{po}[\hat{n}|S]$ can also be applied to $\mathrm{SE}_{th}[\hat{n}|S]$. Especially, the effect of the expected population abundance $n(W)$ on the relative variation holds true in this situation. Eq. (12) suggest that increasing the average number of daughters, $\bar{c}$, increases the standard error. In addition, by definition of $p_d(S)$ Eq. (A.3), a smaller value of σ increases $p_d(S)$. Roughly speaking, a species with a large expected number of daughters, $\bar{c}$, and smaller dispersal distance of daughters, σ, form a high degree of clusters in individual distributions, and it increases the standard error of the population estimate $\mathrm{SE}_{th}[\hat{n}|S]$. The approximated $\mathrm{SE}_{th}[\hat{n}|S]$, Eq. (12), shows good agreement with the values obtained by the numerical simulations across sampling areas, although it shows slight deviations from the numerical values when the fraction of sampling patches is small (α is around 0.05-0.1; Fig. 2). Typically, increasing the sampling unit size (i.e., more clustered sampling) in population estimations increases the standard error, but it decreases with the fraction of sampled patches. We also confirmed the similar agreement between Eq. (12) and numerical simulations with different parameters (Fig. A.2).

Discussion

We examined a method for population estimation combined with spatial point processes (SPPs), spatially explicit model, to reveal effects of different survey regimes as well as individual distribution patterns on the precision of population estimates. By assuming the random and clustering placements of individuals as underlying distribution patterns, we analytically show that the individual distributions and sampling schemes, such as random sampling and cluster sampling, change significantly the standard error of the abundance estimate. In our sampling framework, increasing the sampling unit size corresponds to an increase of geographical bias of the sampling (i.e., cluster sampling; see Survey design). Typically, we find that the standard error of the abundance estimate is insensitive to the sampling unit size applied when the underlying individual distribution is the homogeneous Poisson process. On the other hand, the Thomas process analysis suggests that population estimate will result in less precise population estimates. Typically, under clustered individual distributions, the standard error increases as the degree of clustering sampling increases. We also show that the standard error of the population estimate increases with the parameter characterizing the degree of clustering of individual distributions. In addition, although for both individual distribution patterns, our results show that the absolute value of the standard error increases with the number of individuals, the relative standard error decreases with the factor proportional to $n(W)^{-1/2}$.

In practice, simple random sampling with a fine sampling unit may not easily be conducted due to time and budgetary constraints, and different accessibility to sites [16,23,40]. However, this sampling scheme enables us to obtain more reliable data since extensive sampling in inaccessible region may also lead to new discoveries [16]. Hence, this sampling scheme may be suitable for many ecological studies and ecosystem monitoring projects which require estimations to capture spatial and/or temporal patterns of the population. Alternatively, cluster sampling, which causes a geographical sampling bias, is often the favored survey design practically since it is less expensive and easy to implement [16,23]. Therefore, this survey design may be applied to managements where a target species require quick conservation action at a cost of precision of data. Most importantly, in line with the discussion of Takashina et al. [30], insights developed in the paper should be applied, by clearly setting a feasible goal of population estimate with time and economic constraints, before survey designs are developed.

Here we investigate population estimation with assuming that no data is available and with general ecological and sampling assumptions. However, our results provide generic insights into ecological survey design such as how the sampling unit size used and individual distribution patterns affect the precision of population estimation. Typically, it suggests that more clustered samplings and/or more clustered individual distributions cause less precise population estimations, but the precision improves with the fraction of sampled patches. For both ecological and conservation applications in mind, our sampling framework is kept as general as possible. Therefore, it allows one to further extend the framework to handle more complex situations where, for example, the concerned region holds multiple sampling unit sizes or a budgetary constraint is explicitly taken into consideration. Also, SPPs is not a only choice in our framework, but one can also use any spatially explicit models as long as the model allows to calculate Eq. (3). Especially, for analytical tractability, we focused on how individual distributions and sampling strategies affect the accuracy of population estimate by assuming no or sufficiently small measurement error. Although many empirical studies have adopted this assumption [41], imperfect detection is also frequently observed even in sessile organisms such as plants (e.g. [42, 43]). Also, if searching time is fixed, chance of imperfect detection would increase with survey area [44]. This indicates that the sampling unit size should be chosen while taking the scale-dependency of the imperfect detection into account. Further studies about how imperfect detection changes our predictions is highly beneficial for developing robust survey designs.

Acknowledgements

We would like to thank T. Fung and B. Stewart-Koster for their thoughtful comments. NT and BK were funded by the Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers of the Japan Society for the Promotion of Science, and they acknowledge the support for coordinating the research program from Dr Yasuhiro Kubota and Dr James D. Reimer.

Literature Cited

[1] E. C. Pielou, An introduction to mathematical ecology, Wiley-Interscience, New York, 1969.

[2] E. C. Pielou, Ecological Diversity, John Wiley and Sons Inc, New York, 1975.

[3] D. L. Otis, K. P. Burnham, G. C. White, D. R. Anderson, Statistical Inference from Capture Data on Closed Animal Populations, Source Wildl. Monogr. (62) (1978) 3–135. `doi:10.2307/2287873`.

[4] W. E. Kunin, Extrapolating species abundance across spatial scales, Science 281 (5382) (1998) 1513–1315. `doi:10.1126/science.281.5382.1513`.

[5] F. He, K. J. Gaston, Estimating Species Abundance from Occurrence, Am. Nat. 156 (5) (2000) 553–559. `doi:10.1086/303403`.

[6] F. L. He, K. J. Gaston, Occupancy-abundance relationships and sampling scales, Ecography (Cop.). 23 (4) (2000) 503–511. `doi:10.1111/j.1600-0587.2000.tb00306.x`.

[7] K. H. Pollock, J. D. Nichols, T. R. Simons, G. L. Farnsworth, L. L. Bailey, J. R. Sauer, Large scale wildlife monitoring studies: Statistical methods for design and analysis, Environmetrics 13 (2) (2002) 105–119. `doi:10.1002/env.514`.

[8] B. Goldsmith, Monitoring for Conservation and Ecology, Vol. 3, Chapman & Hall, 1991. `doi:10.1016/0305-1978(91)90074-A`.

[9] D. B. Lindenmayer, G. E. Likens, Adaptive monitoring: a new paradigm for long-term research and monitoring, Trends Ecol. Evol. 24 (9) (2009) 482–486. `doi:10.1016/j.tree.2009.03.005`.

[10] S. H. M. Butchart, M. Walpole, B. Collen, A. van Strien, J. P. W. Scharlemann, R. E. A. Almond, J. E. M. Baillie, B. Bomhard, C. Brown, J. Bruno, K. E. Carpenter, G. M. Carr, J. Chanson, A. M. Chenery, J. Csirke, N. C. Davidson, F. Dentener, M. Foster, A. Galli, J. N. Galloway, P. Genovesi, R. D. Gregory, M. Hockings, V. Kapos, J.-F. Lamarque, F. Leverington,

J. Loh, M. A. McGeoch, L. McRae, A. Minasyan, M. H. Morcillo, T. E. E. Oldfield, D. Pauly, S. Quader, C. Revenga, J. R. Sauer, B. Skolnik, D. Spear, D. Stanwell-Smith, S. N. Stuart, A. Symes, M. Tierney, T. D. Tyrrell, J.-C. Vie, R. Watson, Global Biodiversity: Indicators of Recent Declines, Science 328 (5982) (2010) 1164–1168. doi:10.1126/science.1187512.

[11] D. I. MacKenzie, J. D. Nichols, J. E. Hines, M. G. Knutson, A. B. Franklin, Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly, Ecology 84 (8) (2003) 2200–2207. doi:10.1890/02-3090.

[12] M. A. McCarthy, S. J. Andelman, H. P. Possingham, Reliability of Relative Predictions in Population Viability Analysis, Conserv. Biol. 17 (4) (2003) 982–989. doi:10.1046/j.1523-1739.2003.01570.x.

[13] H. S. Grantham, K. A. Wilson, A. Moilanen, T. Rebelo, H. P. Possingham, Delaying conservation actions for improved knowledge: How long should we wait?, Ecol. Lett. 12 (4) (2009) 293–301. doi:10.1111/j.1461-0248.2009.01287.x.

[14] E. Kuno, Evaluation of statistical precision and design of efficient sampling for the population estimation based on frequency of occurrence, Res. Popul. Ecol. (Kyoto). 28 (1986) 305–319.

[15] J. A. Royle, J. D. Nichols, Estimating abundance from repeated presenceabsence data or point counts, Ecology 84 (3) (2003) 777–790. doi:10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2.

[16] S. Reddy, L. M. Dávalos, Geographical sampling bias and its implications for conservation priorities in Africa, J. Biogeogr. 30 (11) (2003) 1719–1727. doi:10.1046/j.1365-2699.2003.00946.x.

[17] T. A. Gardner, J. Barlow, I. S. Araujo, T. C. Ávila-Pires, A. B. Bonaldo, J. E. Costa, M. C. Esposito, L. V. Ferreira, J. Hawes, M. I. M. Hernandez, M. S. Hoogmoed, R. N. Leite, N. F. Lo-Man-Hung, J. R. Malcolm, M. B. Martins, L. A. M. Mestre, R. Miranda-Santos, W. L. Overal, L. Parry, S. L. Peters, M. A. Ribeiro, M. N. F. Da Silva, C. Da Silva Motta, C. A.

Peres, The cost-effectiveness of biodiversity surveys in tropical forests, Ecol. Lett. 11 (2) (2008) 139–150. doi:10.1111/j.1461-0248.2007.01133.x.

[18] A. E. Camaclang, M. Maron, T. G. Martin, H. P. Possingham, Current practices in the identification of critical habitat for threatened species, Conserv. Biol. 29 (2) (2015) 482–492. doi:10.1111/cobi.12428.

[19] M. Bode, K. A. Wilson, T. M. Brooks, W. R. Turner, R. A. Mittermeier, M. F. McBride, E. C. Underwood, H. P. Possingham, Cost-effective global conservation spending is robust to taxonomic group., Proc. Natl. Acad. Sci. U. S. A. 105 (17) (2008) 6498–6501. doi:10.1073/pnas.0710705105.

[20] A. Hastings, Timescales and the management of ecological systems, Proc. Natl. Acad. Sci.doi:10.1073/pnas.1604974113.

[21] J. L. Green, J. B. Plotkin, A statistical theory for sampling species abundances, Ecol. Lett. 10 (11) (2007) 1037–1045. doi:10.1111/j.1461-0248.2007.01101.x.

[22] J. Rice, Mathematical Statistics and Data Analysis, 3rd Edition, Vol. 72, Thomson Higher Education, Belmont, 2007. doi:10.2307/3619963.

[23] S. Lohr, Sampling: design and analysis, Nelson Education, 2009.

[24] N. A. C. Cressie, Statistics for Spatial Data, John Wiley & Sons, New York, 1993.

[25] J. B. Plotkin, M. D. Potts, N. Leslie, N. Manokaran, J. Lafrankie, P. S. Ashton, Species-area curves, spatial aggregation, and habitat specialization in tropical forests., J. Theor. Biol. 207 (1) (2000) 81–99. doi:10.1006/jtbi.2000.2158.

[26] K. Shimatani, Y. Kubota, Spatial analysis for continuously changing point patterns along a gradient and its application to an Abies sachalinensis population, Ecol. Modell. 180 (2-3) (2004) 359–369. doi:10.1016/j.ecolmodel.2004.04.036.

[27] N. Picard, C. Favier, A Point-Process Model for Variance-Occupancy-Abundance Relationships, Am. Nat. 178 (3) (2012) 383–396. doi:10.1086/661249.

[28] S. Azaele, S. J. Cornell, W. E. Kunin, Downscaling species occupancy from coarse spatial scales, Ecol. Appl. 22 (3) (2012) 1004–1014. doi:10.1890/11-0536.1.

[29] S. N. Chiu, D. Stoyan, W. S. Kendall, J. Mecke, Stochastic Geometry and Its Applications, John Wiley & Sons, New York, 2013.

[30] N. Takashina, M. Beger, B. Kusumoto, S. Rathnayake, H. Possingham, A theory for ecological survey methods to map individual distributions, Theor. Ecol.doi:https://doi.org/10.1007/s12080-017-0359-7.

[31] S. Muko, K. Shimatani, Y. Nozawa, Spatial analyses for nonoverlapping objects with size variations and their application to coral communities (2014).

[32] T. S. Bayard, C. S. Elphick, Using Spatial Point-Pattern Assessment to Understand the Social and Environmental Mechanisms that Drive Avian Habitat Selection, Auk 127 (3) (2010) 485–494. doi:10.1525/auk.2010.09089.

[33] L. R. Taylor, I. P. Woiwod, J. N. Perry, The Density-Dependence of Spatial Behaviour and the Rarity of Randomness, J. Anim. Ecol. 47 (2) (1978) pp. 383–406. doi:10.2307/3790.

[34] R. Condit, P. S. Ashton, P. Baker, S. Bunyavejchewin, C. V. S. Gunatilleke, I. A. U. N. Gunatilleke, S. P. Hubbell, R. B. Foster, A. Itoh, J. V. LaFrankie, H. S. Lee, E. Losos, N. Manokaran, R. Sukumar, T. Yamakura, Spatial patterns in the distribution of tropical tree species., Science 288 (1982) (2000) 1414–1418. doi:10.1126/science.288.5470.1414.

[35] M. Beger, G. P. Jones, P. L. Munday, Conservation of coral reef biodiversity: A comparison of reserve selection procedures for corals and fishes, Biol. Conserv. 111 (1) (2003) 53–62. doi:10.1016/S0006-3207(02)00249-5.

[36] Multinomial distribution. Encyclopedia of Mathematics. URL: http://www.encyclopediaofmath.org/index.php?title=Multinomial_distribution&oldid=28544 Data Accessed: Mar 18 2018.

[37] J. Illian, A. Penttinen, H. Stoyan, D. Stoyan, Statistical Analysis and Modelling of Spatial Point Patterns, Vol. 76, John Wiley & Sons, Chichester, 2008. doi:10.1002/9780470725160.

[38] R. Condit, P. S. Ashton, P. Baker, S. Bunyavejchewin, S. Gunatilleke, N. Gunatilleke, S. P. Hubbell, R. B. Foster, A. Itoh, J. V. LaFrankie, H. S. Lee, E. Losos, N. Manokaran, R. Sukumar, T. Yamakura, Spatial patterns in the distribution of tropical tree species, Science 288 (1982) (2000) 1414–1418. doi:10.1126/science.288.5470.1414.

[39] M. Thomas, A generalization of Poisson's binomial limit for use in ecology, Biometrika 36(1/2) (1949) 18–25.

[40] H. Possingham, I. Ball, S. Andelman, Mathematical methods for identifying representative reserve networks, in: Quant. Methods Conserv. Biol., Springer New York, 2000, pp. 291–306. doi:10.1007/0-387-22648-6_17.

[41] K. F. Kellner, R. K. Swihart, Accounting for imperfect detection in ecology: A quantitative review (2014). doi:10.1371/journal.pone.0111436.

[42] G. Chen, M. Kéry, J. Zhang, K. Ma, Factors affecting detection probability in plant distribution studies, J. Ecol. 97 (6) (2009) 1383–1389. doi:10.1111/j.1365-2745.2009.01560.x.

[43] G. Chen, M. Kéry, M. Plattner, K. Ma, B. Gardner, Imperfect detection is the rule rather than the exception in plant distribution studies, J. Ecol. 101 (1) (2013) 183–191. doi:10.1111/1365-2745.12021.

[44] K. D. Clarke, M. Lewis, R. Brandle, B. Ostendorf, Non-detection errors in a survey of persistent, highly-detectable vegetation species, Environ. Monit. Assess. 184 (2) (2012) 625–635. doi:10.1007/s10661-011-1991-0.

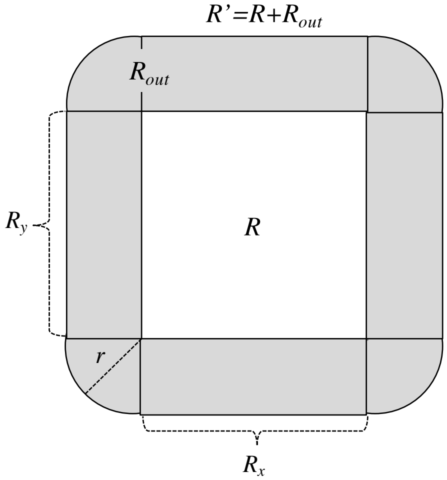

Figure A.1: R is the concerned region with area $R_x \times R_y$. Parents outside R with a distance less than r from the edges of R (parents in R_{out}) may also contribute to the number of daughters in the concerned region R. The whole region where parents can supply daughters to R is $R' = R + R_{out}$.

Appendix

Derivations of an approximated pdf of the Thomas process

Here, we derive an approximated form of the probability distribution function (pdf) of the Thomas process. For this purpose, we firstly introduce two regions R' and R_{out}. Let R' be the region where a parent potentially supples the daughters to the region R. Then R' is decomposed into two regions $R' = R + R_{out}$, where R_{out} is the surrounding region of R and satisfies with $R' \setminus R$ (Fig. A.1). Here, we approximate the probability that n individuals fall in the region R with k' individuals produced by parents in R' by the binomial distribution, though sisters (i.e., daughters share a same parent) locations depend on its parent location. Under this assumption, the probability that n individuals found in region R is described

$$\begin{aligned} p(n|R) &= \underbrace{\sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')}}_{\text{no. parents in } R'} \underbrace{\sum_{k'} \begin{pmatrix} k' \\ n \end{pmatrix} p_d(R)^n (1-p_d(R))^{k'-n} \sum_{k' \in \mathcal{K}} \prod_i^k \frac{\bar{c}^{k'_i}}{k'_i!} e^{-\bar{c}}}_{\text{Prob}(n \text{ daughters fall in } R \text{ provided } k' \text{ daughters produced by parents in } R')} \qquad (A.1) \\ &= \sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')} \sum_{k'} \begin{pmatrix} k' \\ n \end{pmatrix} p_d(R)^n (1-p_d(R))^{-n} e^{-\bar{c}k} \sum_{k' \in \mathcal{K}} \prod_i^k \frac{\{\bar{c}(1-p_d(R))\}^{k'_i}}{k'_i!}, \end{aligned}$$

where, $k' = k'_1 + \cdots + k'_k$ and k'_i is the number of daughters produced by parent i. $\sum_{k' \in \mathcal{K}}$ runs all the combinations of k' satisfies $\sum_i k'_i = k'$. As one can easily see $\sum_{k' \in \mathcal{K}} k'! \prod_i^k \{\bar{c}(1-p)\}^{k'_i}/k'_i!$ is the coefficient of expansion of $(\lambda_1 + \cdots + \lambda_k)^{k'_1 + \cdots + k'_k}$, where we set $\lambda_1 = \cdots = \lambda_k = \bar{c}(1-p_d(R))$. Therefore, Eq. (A.1) becomes

$$\begin{aligned} p(n|R) &= \sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')} \sum_{k'}^{\infty} \frac{1}{(k'-n)!n!} p_d(R)^n (1-p_d(R))^{-n} e^{-\bar{c}k} \sum_{k' \in \mathcal{K}} k'! \prod_i^k \frac{\{\bar{c}(1-p_d(R))\}^{k'_i}}{k'_i!}, \\ &= \sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')} \frac{1}{n!} p_d(R)^n (1-p_d(R))^{-n} e^{-\bar{c}k} \sum_{k'}^{\infty} \frac{(\bar{c}k(1-p))^{k'}}{(k'-n)!}, \\ &= \sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')} \frac{(\bar{c}k(1-p_d(R)))^n}{n!} p_d(R)^n (1-p)^{-n} e^{-\bar{c}k} \sum_{k'}^{\infty} \frac{(\bar{c}k(1-p))^{k'-n}}{(k'-n)!}, \\ &= \sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')} \frac{(\bar{c}k)^n}{n!} p_d(R)^n e^{-\bar{c}k} e^{\bar{c}k(1-p_d(R))}, \\ &= \sum_k \frac{(\lambda_p \nu(R'))^k}{k!} e^{-\lambda_p \nu(R')} \frac{(\bar{c}k p_d(R))^n}{n!} e^{-\bar{c}k p_d(R)}, \\ &= \sum_k \mathrm{Po}(k, \lambda_p \nu(R')) \mathrm{Po}(n, k\bar{c} p_d(R)). \qquad (A.2) \end{aligned}$$

where, $\mathrm{Po}(k, \lambda)$ is the poisson distribution with the intensity λ and $p_d(R)$ is the probability that an individual daughter produced by a parent within R' falls in R. Since a parent location is randomly chosen in R', we calculate $p_d(R)$ as follows

$$p_d(R) = \frac{1}{\nu(R')} \int_{R'} \int_R \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) d\mathbf{x} d\mathbf{y}, \qquad (A.3)$$

where, $\mathbf{x}$ and $\mathbf{y}$ are location in R and R', respectively. By referring Fig. A.1, $\nu(R')$ is calculated as

$$\nu(R') = (2r + R_x)(2r + R_y) - r^2(4 - \pi), \tag{A.4}$$

where, r is the distance that on average a fraction u of daughters scattered by the parent (placed center) are covered. r is calculated by converting the expression of the isotropic bivariate gaussian on cartesian coordinates, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy 1/(2\pi\sigma^2)\exp\{-(x^2+y^2)/(2\sigma^2)\}$, to the one on the polar coordinates, and solving about r

$$r = \sqrt{-2\sigma^2 \log(1-u)}, \tag{A.5}$$

where, in the analysis, we set $u = 0.99$ (i.e., 99% of daughters fall within this distance).

Standard error of the Thomas process

Using Eq. (A.2), we calculate the first moment and the second moment of the point number k in region R

$$\begin{aligned} \mathrm{E}[n(R)] &= \lambda_p \bar{c} p_d(R)\nu(R') = \lambda_p \bar{c}\nu(R), && \text{(A.6)} \\ \mathrm{E}[n(R)^2] &= \lambda_p \bar{c} p_d(R)\nu(R')(1 + \bar{c} p_d(R) + \lambda_p \bar{c} p_d(R)\nu(R')). && \text{(A.7)} \end{aligned}$$

Using Eqs (3), (9), (A.6), and (A.7) and the fact $\lambda_p \bar{c} = \lambda = n(W)/\nu(W)$, $N_t = \nu(W)/\nu(S)$, and $N_s = \alpha N_t$, we calculate Eq. (12) as follows:

$$\begin{aligned} \mathrm{SE}_{th}[\hat{X}|S] &= \sqrt{\lambda_p \bar{c} p_d(S)\nu(S')(1 + \bar{c} p_d(S))\frac{N_t^2}{N_s}\left(\frac{N_t - N_s}{N_t - 1}\right)}, \\ &= \sqrt{n(W)\left(\frac{1}{\alpha} - 1\right)\frac{N_t}{N_t - 1}\frac{\nu(S')}{\nu(S)} p_d(S)\,(1 + \bar{c} p_d(S))}, \\ &= \mathrm{SE}_{po}[\hat{X}|S]\sqrt{\frac{\nu(S')}{\nu(S)} p_d(S)\,(1 + \bar{c} p_d(S))}. \end{aligned} \tag{A.8}$$

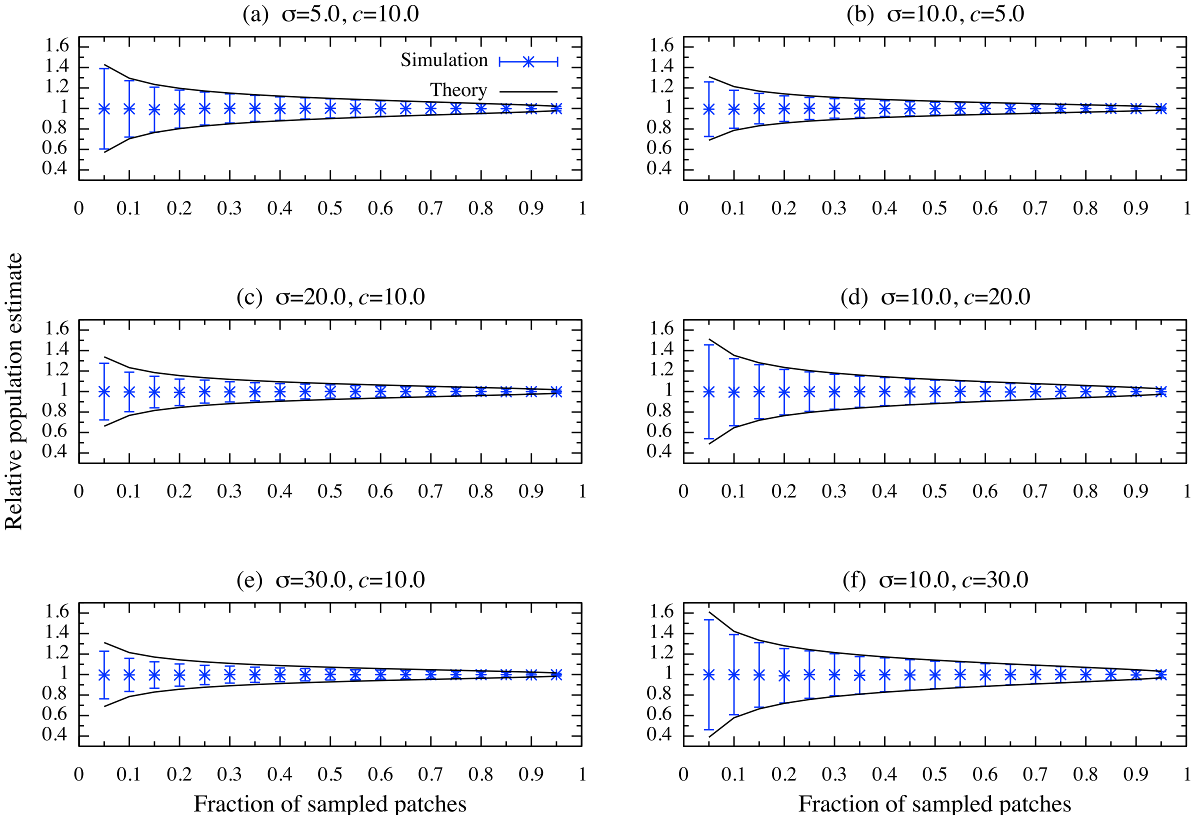

Figure A.2: Relative value of the population estimate with the average individuals $\mathrm{E}[n(W)] = 10^3$ with different parameters. Sampling area is 32m×32m. Each panel shows relative average estimate ± relative standard error (Eq. (8)) of simulation and theoretical results. Relative average estimate for theoretical results is omitted since it is an unbiased estimator. Total area is $\nu(W) = 2^{20}\mathrm{m}^2$ (1024m×1024m).