ESTIMATING SPECIES ABUNDANCE

G.J. Melville
New South Wales Department of Primary Industries, Australia

A.H. Welsh
Mathematical Sciences Institute, Australian National University, Australia

Keywords: Adaptive Sampling, Calibration Study, Capture-recapture, Edge Bias, Expansion Estimator, Line Transect Sampling, Nearest Neighbor Distance, Quadrat Sampling, Removal Method, Transect Line, Trapping Web, T-square Sampling, Unstocked Quadrats, Wandering Quarter Method.

Contents

Summary

Estimating the size of a population of a species is of interest and importance to the study and/or monitoring of biological and other populations. The size of a population is also called the abundance and estimating the abundance is mathematically equivalent to estimating the number of plants or animals per unit area which is referred to as the density or the intensity of the population. This chapter reviews the rich variety of approaches to this difficult problem, including quadrat sampling, adaptive cluster sampling, line transect or distance sampling, nearest neighbor distance methods and capture-recapture methods.

1. Introduction

The estimation of species abundance or population size is often a fundamental aspect of the study and/or monitoring of biological and other populations. The different approaches to this problem can be distinguished by the different types of data which can be collected (including counts, distances etc), the different ways in which the data can be collected (quadrat sampling, adaptive sampling etc) and the different ways in which we can estimate abundance from these data.

Suppose we are interested in estimating the unknown number N of members of the species (which we refer to as objects) in a region R of area A. We treat R as homogeneous for simplicity. (If R has a known inhomogeneous structure, we can stratify R and treat each stratum separately; in this case, R is synonymous with a stratum of R.) A useful abstract version of the problem is obtained by treating the objects as points distributed in R according to some point process. The problem is to estimate N or, equivalently, N/A which is referred to as the object density or intensity. We will use intensity to avoid confusion with probability density functions and express all estimators as estimators of the abundance N rather than the intensity N/A.

Treating a biological population as a realization of a spatial point process is most useful for populations of stationary objects whose size is negligible compared to the area A of the region. Highly mobile species pose particular problems, rare and difficult to detect species introduce further complications. It is not surprising that there is a substantial literature on estimating species abundance for particular species. This usually begins by applying and adapting general methods to the specific species but may ultimately lead to the development of specialized techniques. It is impossible to review this literature in any generally useful way so we will restrict our focus to general methods applied in the idealized, abstract context.

As in all sampling, there are two different approaches to introducing randomness and these lead to different ways of thinking about estimation and inference. In the design-based framework, we condition on the number and location of the objects (so they are treated as fixed) and the randomness is induced by the sampling scheme used to collect the data. On the other hand, in the model-based framework, we condition on the sample and randomness is induced from the underlying stochastic process which determines the number and locations of the objects. The model-based approach does not preclude random sampling (which can be useful in providing information about the appropriateness of the model) but does not require it because it conditions on the observed sample. One difficulty in the literature on estimating species abundance is that the framework is not always explicitly specified and some methods incorporate aspects of both approaches.

The choice of inferential framework affects the way standard errors are obtained and the way we interpret the resulting inferences. The design-based approach is nonparametric in the sense that it depends only on probabilities which are under our control and known. (Standard error estimation requires joint inclusion probabilities which are not always straightforward to compute.) Since the method is intended to apply to any population, it does not necessarily do particularly well in any specific population and this is reflected in inefficient estimates. There are also a number of well-known difficulties stemming from the unconditional nature of the approach. Model-based methods condition on the observed sample and, provided the model holds, should perform better than design-based methods. The difficulty is that there is usually some uncertainty about the appropriateness of the model (i.e. the assumptions) so we have to deal with robustness issues and these can be quite complicated in sampling problems.

The class of models for the underlying spatial point process is limited by the complication of working with many of the models. The simplest and most tractable model is the homogeneous Poisson process. A homogeneous Poisson process with intensity λ (representing the mean number of objects per unit area) has N ~ $\text{Poisson}(\lambda A)$ and, conditional on *N*, the points distributed uniformly in *R*. For this reason, the homogeneous Poisson process is sometimes described as complete spatial randomness (CSR). This model is often too simple to account for observed heterogeneity such as clustering in the data. More useful models include the inhomogeneous Poisson process in which the intensity varies spatially over the region. Alternatively we can think of two-stage or parent-daughter processes with each stage following a separate stochastic model. In practice, we cannot distinguish between a two-stage (parent-daughter) Poisson process and heterogeneous Poisson process so we use whichever model is the more convenient.

A number of other point processes have been put forward as methods for describing or simulating ecological data. Inhibition models attempt to describe the competitive and territorial behavior which is observed in ecological populations. Matern's static process is a modification of a homogeneous Poisson process in which all pairs of points which are less than *r* units apart are eliminated. This type of model is supposed to account for the competitive behavior which is observed in many species of plants and animals. Matern's sequential process is a dynamic variation, where a homogeneous Poisson process is generated one point at a time and a point is discarded if it lies within *r* units of any previously generated point. This process is more suited to animals, particularly those which exhibit territorial behavior. Diggle's sequential process is a modification of Matern's sequential process in which a point is discarded only if it is within *r* units of a retained point. All of the above processes are inflexible in the sense that the distance *r* is fixed and assumed to be the same for all pairs of points. The Strauss process allows *n* neighboring points within a distance *r*. The model has joint density proportional to $\alpha^N \beta^n$, where α and β are parameters and *N* is the number of points in the process. This process has been used to model the distribution of pine trees in a forest and the distribution of magnetic crystals in a rock.

Specifying the underlying stochastic process is only the first step. If we collect counts or distances, we need to derive their distributions from the underlying stochastic process. This is simplest for the homogeneous Poisson process and more complicated for other processes. The derived distributions do not necessarily distinguish between underlying processes, showing how difficult it can be to learn about the underlying process. The negative binomial distribution, for example, can arise in at least five different ways:-

- inverse binomial sampling - number of trials to *k*th success
- heterogeneous Poisson process - compound Poisson and gamma

- two-stage Poisson process - generalized Poisson and logarithmic
- constant birth-death-immigration process
- true contagion - mutual attraction of objects.

Hence establishing that a negative binomial distribution is appropriate does not reveal anything definite concerning the genesis of the population and therefore attempts to classify spatial patterns, including the use of statistical tests of fit to rule out competing models, are of doubtful benefit in determining the genesis of the pattern.

In highlighting various difficulties, it is not our intention to be negative. The problem of estimating species abundance is important and difficult; the difficulties pose challenges. It is remarkable what can and has been achieved and we are optimistic about what may be achieved in future.

The rest of this chapter consists of brief reviews of the main methods of estimating species abundance. We first discuss quadrat (Section 2) sampling which is based on counting the number of objects in selected areas called quadrats. We then consider adaptive sampling (Section 3) which is an extension of quadrat sampling to increase efficiency when trying to estimate the abundance of rare, clustered populations and line and point transect sampling (Section 4) which is an extension of quadrat sampling to allow for the fact that not all the objects in a quadrat are observed. We then discuss nearest neighbor distance sampling (Section 5) which is based on measuring distances without specifying quadrats. Finally, in Section 6, we discuss capture-recapture sampling which also avoids the need to specify quadrats explicitly.

2. Quadrat Sampling

As the name suggests, quadrats were originally square subsets of the study region which were sampled by throwing a wooden frame over the shoulder backwards. To avoid bias, both the location of the throwing point and the direction of throw need to be randomized and one has to allow for the possibility of overlap. A quadrat now more generally refers to a study area of fixed size and shape with square, rectangular and circular quadrats in most common use, and may be used as a synonym for sampling units (paddocks, trees, leaves, rocks, logs, rock pools, fishing nets etc). Quadrats may be permanently marked, for example by using wire or pegs; alternatively, in sparse populations the actual physical quadrat may not be needed at all - one only needs to locate quadrats accurately. Obviously, decision rules are needed to include or exclude objects on the boundary.

Suppose that the study region R is exhaustively partitioned into M non-overlapping quadrats of area a_i, each of which contains an unknown number of objects $Y_i,\ i=1,\ldots,M$. We collect data by selecting a sample s of m quadrats and counting the number of objects in each of the selected quadrats. We usually sample without replacement (so that no quadrats are chosen more than once). Different sized quadrats may be convenient when there is information on the distribution of the objects so that, for example, larger quadrats could be used in areas where the objects are sparse to save time in enumeration.

2.1. Design-Based Quadrat Sampling

All the classical finite population sampling designs (including simple random sampling, stratified, cluster, probability proportional to size (pps), random systematic designs etc) and their associated estimators can be used to construct design-based estimators of abundance. If we select quadrats by simple random sampling without replacement, we have the expansion estimator

$$\hat{N}_{\mathrm{E}} = \frac{M}{m}\sum_{i\in s} Y_i. \tag{1}$$

If, as is often the case, the counts in the *i*th quadrat are proportional to the area of the quadrat, we can use the ratio estimator

$$\hat{N}_{\mathrm{R}} = \frac{A}{a}\sum_{i\in s} Y_i. \tag{2}$$

On the other hand, if we use probability proportional to size sampling, the selection probability for the *i*th quadrat is a_i/A, and the Horvitz-Thompson estimator of the total will be

$$\hat{N}_{\mathrm{HT}} = A\sum_{i=1}^{m} \frac{Y_i}{a_i}. \tag{3}$$

These are all classical estimates and expressions for estimating their variances are well known. For example, the variance of the expansion estimator under simple random sampling without replacement is estimated by the familiar

$$\hat{V}(\hat{N}_{\mathrm{E}}) = \frac{M^2}{m(m-1)}\left(1-\frac{m}{M}\right)\sum_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2. \tag{4}$$

2.2. Model-based Quadrat Sampling

Most models for quadrat sampling treat the quadrat counts $Y_1,\ldots,Y_M$ as independent random variables with some count distribution. For example, the Poisson model assumes that the quadrat counts are independent Poisson variables with mean $a_i\lambda$, where λ is the mean intensity. We can estimate λ by $\hat{\lambda} = a^{-1}\sum_{i\in s} Y_i$ and then estimate abundance by

$$\hat{N} = A\hat{\lambda} = \frac{A}{a}\sum_{i\in s} Y_i, \tag{5}$$

which is the same as the ratio estimator.

Under the binomial model the quadrat counts are independent binomial (ra_i ,p) variables (treat ra_i as an integer for simplicity). The need to estimate r (the largest possible count in a quadrat) as well as p (the probability that an object is present in a quadrat) makes this an unusual model. The parameter r is usually estimated by the largest observed count $\hat{r}$ and p is estimated by $\hat{p} = (a\hat{r})^{-1}\sum_{i\in s} Y_i$. The estimate of population size is given by $A\hat{r}\hat{p}$ which is the ratio estimator. Similarly, the negative binomial model has parameters p, where $p/(p+1)$ is the probability of an object being present in a quadrat, and $a_i k$, which is related to the degree of clustering. The estimate of abundance is again the ratio estimator. The variance of the estimator depends on the model. The general form of the estimated prediction variance is

$$\hat{V}_{\mathrm{P}}(\hat{N}) = \frac{A^2}{a}\left(1-\frac{a}{A}\right)V\hat{[}Y_i] \qquad (6)$$

where $V\hat{[}Y_i] = \hat{\lambda}$ under the Poisson model, $V\hat{[}Y_i] = \hat{r}\hat{p}(1-\hat{p})$ under the binomial model and $V\hat{[}Y_i] = \hat{k}\hat{p}(1+\hat{p})$ under the negative binomial model.

The Poisson distribution has been used to model the distribution of spiders under boards. However, most animals are clustered rather than randomly distributed. Although much discussed, the binomial model has only rarely been used in practice. The negative binomial model has been widely used, particularly in modeling insect counts. In an interesting study, Shiyomi and Nakamura distributed aphids on barley plants both randomly and uniformly. Within a few days the number of aphids per plant was Poisson distributed, presumably due to random deaths and plant to plant movement, but after a short period of reproduction (1 week), the number of aphids per plant had a negative binomial distribution.

Other models for count data are also sometimes used. Compound and generalized models provide enormous scope for describing actual populations. However several authors have pointed out that no precise biological meaning can be attached to the parameters of the various distributions. In addition it has been noted that several distributions can often fit the same data equally well. For example micro-arthropods and rice leaf grasshoppers have been modeled using the Neyman type A, Polya-Aeppli, negative binomial and discrete log-normal models.

The above models can be extended to incorporate covariate information such as the size of the quadrat, physical aspects (e.g. soil type), type of vegetation, time of day or year, distance from water and so on, provided these covariates or benchmark variables are available for the whole population of quadrats. The parameters of the model are estimated from the quadrat sample, predicted values are then obtained for quadrats which were not sampled and the prediction variance of the final estimate is derived from the model. This approach has been used with two-part Poisson and negative binomial models to estimate the abundance of seabird nests.

A simpler but less efficient approach is based on using the number of quadrats with zero counts $n_0 = \sum_{i \in s} I(Y_i = 0)$ (the unstocked quadrats) to construct an estimator. For example, if all quadrats have equal area *a*, under the Poisson model $P(Y_i = 0) = \exp(-\lambda a)$ so $\lambda = -a^{-1} \log P(Y_i = 0)$ and we can estimate *N* as

$$\hat{N} = -\frac{A}{a}\log\left(n_0/m\right), \tag{7}$$

with estimated variance

$$\hat{V}(\hat{N}) = -\frac{A^2}{na^2}\left(\frac{m}{n_0} - 1\right). \tag{8}$$

2.3. Quadrat Size

Quadrat sampling is sensitive to quadrat size and this choice is somewhat arbitrary. Particularly if the intention is to use a model-based analysis, the quadrat size can determine the appropriateness of the model. There are several rules of thumb in the literature for choosing quadrat size. One possibility is to choose the quadrat size which has a mean count of 1.6 - for the Poisson model this results in 20% zero quadrats. Alternative suggestions include obtaining a mean count of 1.0 (40% zero quadrats) and 4.0 (2% zero quadrats).

3. Adaptive Cluster Sampling

Adaptive cluster sampling is an extension of quadrat sampling which is intended to improve efficiency when the objects of interest are clustered and possibly rare. The basic idea is to take an initial probability sample (such as a simple random sample without replacement) of quadrats and then sample additional quadrats around those quadrats in which at least one object is detected. This second adaptive stage of sampling continues until the adjacent quadrats contain no objects.

Although adaptive sampling schemes have been of interest for some time, adaptive cluster sampling as described here derives from the work of Thompson and Seber.

As in ordinary quadrat sampling, we enumerate the quadrats by $i = 1, \ldots, M$ and let Y_i denote the number of objects in the *i*th quadrat. The quadrats (which are usually taken to be square and of equal size) are the sampling units.

Suppose that we select an initial sample s_1 of m_1 quadrats using a probability design such as simple random sampling without replacement. Suppose that the *i*th quadrat is in the initial sample so $i \in s_1$. If the *i*th quadrat does not contain any objects $(Y_i = 0)$, no further sampling is carried out around it. If the *i*th quadrat contains at least one of the objects of interest $(Y_i > 0)$, we add the neighborhood $Nbhd_i$ of the *i*th unit to the sample and count Y_j, $j \in Nbhd_i$, the number of objects in these additional quadrats. The neighborhood $Nbhd_i$ of the *i*th quadrat is defined to be the *i*th quadrat and the set of four quadrats with an edge in common with this quadrat. If any of the additional quadrats in $Nbhd_i$ contain the object of interest, we add the neighborhoods of these quadrats to the sample and count the number of objects in them. We proceed in this way until we obtain neighborhoods in which none of the quadrats contain any objects and then stop. Thus we obtain a set of quadrats with at least one common boundary which contain the objects of interest surrounded by a set of empty quadrats (called edge quadrats) which have at least one common boundary with the object containing quadrats. This set of quadrats that are observed as a result of including the *i*th quadrat in the initial sample (including the edge quadrats) is called the *i*th cluster. Thus the final sample *s* consists of isolated quadrats containing no objects (clusters of size one) and clusters of quadrats in which the inner quadrats contain objects and which include an outer ring of empty edge quadrats.

Clusters of quadrats arise naturally from the sampling process but are not convenient to work with. The difficulties are caused by the edge units because selecting edge quadrats does not result in the cluster being selected and clusters can overlap in the sense of having edge quadrats in common. One way to get around these difficulties is to omit the edge units from the clusters of size greater than one. The resulting object is called a network. The network Net_i of the *i*th quadrat is the cluster generated by the *i*th quadrat with the (empty) edge units removed. Empty units are defined to be networks of size one so that both clusters of size one and edge units are networks of size one. Networks cannot overlap and form an exhaustive partition of the region of interest, making them simpler to use than clusters.

In the design-based analysis, the quadrats are the basic sampling units but the inclusion probabilities (the probability of including a quadrat in the sample) are unknown. Thompson defined the partial inclusion probability π_i' to be the probability that the initial sample intersects Net_i and then constructed Horvitz-Thompson estimators based on the networks

$$\hat{N}_{\text{AS}} = \sum_{i \in s} \frac{Y_i}{\pi_i'}. \tag{9}$$

After re-expressing the estimator in terms of networks, the variance can be estimated in a standard way.

The estimator $\hat{N}_{\text{AS}}$ is designed unbiased so can be improved by applying the Rao-Blackwell theorem. However, whether this rather complicated step is worth doing in practice is unclear.

Adaptive cluster sampling can be extended to stratified populations of quadrats. This is straightforward unless networks are allowed to straddle stratum boundaries.

-
-
-

Bibliography

Barry, S. and Welsh, A.H. (2001) Distance sampling methodology. *Journal of the Royal Statistical Society B*, **63**, 31-53. [An examination of the foundations and theory of distance sampling.]

Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. (1993) *Distance Sampling*. Chapman & Hall, London. [An exposition of classical distance sampling methodology.]

Byth, K. and Ripley, B.D. (1980) On sampling spatial patterns by distance methods. *Biometrics*, **36**, 279-284. [An examination of nearest neighbor and nearest object distance methods.]

Cormack, R.M. (1979) Models for capture-recapture. In *Sampling Biological Populations*, R.M. [A review of capture-recapture methodology.]

Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press Inc. (London) Ltd, London. [Includes nearest neighbor and nearest object distance methods.]

Hall, P., Melville, G. and welsh, A.H. (2001) Bias correction and bootstrap methods for a spatial sampling scheme. Bernoulli, 7(6), 829-846. [Statistical presentation of the “wandering quarter” sampling method.]

Melville, G.J. and Welsh, A.H. (2001) Line transect sampling in small regions. *Biometric*s, **57**, 1130-1137. [Presents the calibration approach to distance sampling.]

Ripley, B.D. (1981) *Spatial statistics*. John Wiley & Sons, New York. [A general text which includes nearest neighbor and nearest object distance methods.]

Seber, G.A.F. (1973) *The estimation of Animal Abundance*. Griffin: London. Second Edition (1982). [A classic work in the field.]

Seber, G.A.F. (1986) A review of estimating animal abundance. *Biometrics*, **42**, 267-292. [A recent review of the field.]

Thompson, S.K. and Seber, G.A.F. (1996) *Adaptive Sampling*. John Wiley & Sons, New York. [A clear presentation of the adaptive sampling methodology.]

Biographical Sketches

Gavin Melville was born in Parkes, NSW and grew up on a grazing property near Mt Hope, NSW. He obtained a Bachelor of Science degree from Macquarie University in 1984 and also has a MLitt in cosmology (New England University, 1987), a Graduate Diploma in statistics (Australian National University, 1990) and a PhD in ecological sampling (Australian National University, 1999). Gavin spent eight years in Canberra, being employed by the Australian Bureau of Statistics, Commonwealth Department of Health and Australian Institute of Health. Since 1992 he has worked as a biometrician at the Trangie Agricultural Research Centre, NSW and is an author and co-author of several scientific publications. Current research interests are centered on various sampling problems which arise in agricultural and natural resource situations. Gavin is a member of the Statistical Society of Australia.

A.H. Welsh is Professor of Statistics at the Australian National University. He completed a BSc (Hons) at the University of Sydney and a PhD at the Australian National University before becoming an Assistant Professor at the University of Chicago in 1984. He returned to the Australian National University as a lecturer in 1987 where he remained for 14 years. He took up a chair in Statistics at the University of Southampton, United Kingdom, in 2001. He returned to the Australian National University as E.J. Hannan Professor of Statistics in the Mathematical Sciences Institute in 2004. He has been a fellow of the Institute of Mathematical Statistics since 1990 and was awarded the Moran Medal by the Australian Academy of Science in 1990. His main research interests include statistical inference, statistical modeling, robustness, nonparametric methods, analysis of sample surveys and ecological monitoring.