

A note on the detection function estimation in Line Transect Sampling: the case of variable sizes

Una nota sulla stima della detection function nel Campionamento per Transetti Lineari: il caso delle dimensioni variabili

Tonino Sclocco

DMQTE

G. d'Annunzio University - Pescara

sclocco@dmqte.unich.it

Marco Di Marzio

DMQTE

G. d'Annunzio University - Pescara

dimarzio@dmqte.unich.it

Riassunto: Nel campionamento per transetti lineari la stima della detection function $g(x)$ riveste primaria importanza. $g(x)$ fornisce la probabilità di avvistamento di un oggetto posto a una distanza perpendicolare x dal transetto. Se l'oggetto ha dimensioni variabili si dovrebbe stimare una funzione bivariata del tipo $g(x, s)$, operazione spesso complicata. In questo lavoro si affronta il problema della stima della funzione $g(x|S = s)$, proponendo una procedura inferenziale non parametrica che tiene conto dell'effetto dimensionale ma che non comporta problemi di stima su più dimensioni.

Keywords: Line Transects Sampling, Detection Function, Hazard Function, Local Regression.

1. Introduction

Many studies on biological populations require estimates of population density D on a region. Line transects sampling (LTS) is a methodology for estimating $D = \frac{N}{A}$, where A denotes the area of the region and N the population size. According to its simplest version, an observer travels along a randomly positioned line l recording the distances from the line to each object detected. Once accomplished the survey, the observer has sighted $n \leq N$ objects. For the i -th object sampled, $z_i \in [0, +\infty)$ denotes the distance between the observer and the orthogonal projection of i on l ; $r_i \in [0, +\infty)$ the 'radial' distance, i.e. the distance between the observer and i ; $x_i \in [0, +\infty)$ the perpendicular distance between i and l ; finally, θ_i the angle between r_i and z_i . The information resulting from this sampling scheme is indifferently formalized by one of these trivariate samples: $\{(X_i, Z_i, S_i)\}_{i=1, \dots, n}$ or $\{(R_i, \Theta_i, S_i)\}_{i=1, \dots, n}$.

'Distance bias', i. e. the tendency to detect fewer objects at greater distances from the line, constitutes a peculiar source of randomness proper of LTS as a distance sampling method. The detection function (df) $g(\cdot) : \mathbb{R}^+ \rightarrow [0, 1]$, giving the sighting probability at distance x , provides an adequate description of the effects of 'distance bias' on the survey. The same informative contents is proper of the hazard function (hf), i. e. the function giving the probability of sighting in correspondence of the z 's conditioned to a value of X (recently, Skaug and Schweder, 1999), here indicated as $h(z; x)$.

In view to formulate df as a function of hf , we introduce a particular version of hf , depending on r and θ rather than z and x . Given the random variables R and Θ , we define the hf of R for $\Theta = \theta$ as

$$h(r|\Theta = \theta) = \lim_{\Delta r \downarrow 0} \frac{\Pr[r \leq R < r + \Delta r | R > r, \Theta = \theta]}{\Delta r} = \frac{f(r|\Theta = \theta)}{1 - F(r|\Theta = \theta)}.$$

To simplify mathematics we assume the observer'sighting skill is the same for all directions, i. e.: $h(r|\Theta = \theta) = h(r)$. Consequently df takes the form (see Appendix):

$$g(x) = 1 - \exp\left(-\int_0^x h(r) dr\right). \quad (1)$$

2. Variable sizes

If objects' sizes, say s , are not constant, it is often the case that $g(\cdot)$ depends proportionally on them, for a given x . In this case a more carefully definition of df is needed: $g(x, s) : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow [0, 1]$. A direct estimation of $g(x, s)$ – monotonically not increasing on x and monotonically not decreasing on s – might be not simple owing to the bidimensionality of the domain and the small size of real samples.

Drummer and McDonald (1987) firstly studied this matter; they consider a class of parametric families among which to select a model for $g(x, s)$. In the current literature a nonparametric approach is prevalent: Chen (1996, 1999) applies to conditioned multi-dimensional kernels; Mack and Quang (1998) estimate characteristic values of bivariate density functions. In these nonparametric papers reasonable results are obtained for broad sample sizes.

The key to dispelling the multidimensional estimation problem involved by the form $g(x, s)$ could be a suitably modeling of R ; this approach is motivated by the consideration that follows. High dimensions increase the sighting probability, i. e.: 'percentually diminishes the distance between the observer and the object'; reduced dimensions diminish the sighting probability, i. e.: 'percentually increases the distance between the observer and the object'.

In the present notes we propose a model of r containing a transformation of s , say $\phi(\cdot)$, as scale factor. Moreover, on the basis of this scale model, we establish a nonparametric framework to estimate $\phi(\cdot)$. Through the observed values r_i , at first standardized by their estimated scales $\hat{\phi}(s_i)$ and after rescaled by $\hat{\phi}(s)$, we estimate $\hat{h}_s(\cdot)$. Finally, $\hat{g}(x|S = s)$ is obtained replacing $h(\cdot)$ by $\hat{h}_s(\cdot)$ in the formula (1).

3. The model

Consider a monotonic function $\phi(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, let $\phi(s_i)$ be the scale of the density function of the random variable R_i , $f_{R_i}(r; \phi(s_i))$. The following relation, in compliance with the parametric rule $t(\cdot; \phi(s_i)) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, holds

$$R_i = t\{R_i^*; \phi(s_i)\} = \phi(s_i)R_i^*. \quad \forall i \in n \quad (2)$$

The random variable R_i^* is immediately interpreted as the sighting distance free of S_i . For a value s of S , the following steps generate an estimation of the function $g(x|S = s)$:

- i) estimate $\phi(\cdot)$;
- ii) remove the size effect from data calculating $\hat{r}_i^* = r_i / \hat{\phi}(s_i)$;
- iii) rescale the omogeneous values $\hat{r}_i^*$ by $\hat{\phi}(s)$: $\hat{r}_{s,i} = \hat{r}_i^* \hat{\phi}(s)$;
- iv) estimate $h(\cdot)$ by $\hat{r}_{s,i}$ obtaining $\hat{h}_s(\cdot)$;

v) evaluate $\widehat{g}(x|S = s)$ through the rule

$$\widehat{g}(x|S = s) = 1 - \exp \left(- \int_0^x \widehat{h}_s(r) dr \right). \quad (3)$$

A formulation as (2), suitable for a nonparametric estimation, seems necessary owing to the practical impossibility to justify particular parametric models. For example, two simply parametric specifications, very different each others are: 1) $\phi(s_i) = s_i^\gamma$ $\gamma = 1, 2, \dots$; 2) $\phi(s_i) = 1 + \frac{s_i}{\gamma}$ $\gamma > 0$.

A scale model for X was proposed to estimate the parametric bidimensional detection function $g(x, s; \gamma) = g(x/s^\gamma)$ by Drummer and McDonald (1987), without detailed motivations about both the scale model and the choice of a parametric family for $g(x/s^\gamma)$.

The estimator (3) consists of two estimation steps: the estimation of $\phi(\cdot)$ and the estimation of $h(\cdot)$. This procedure might seem complex, but results necessary since two bias sources, distance and dimensions, exist. A single estimation step involves a bidimensional estimation task. Across steps i)-v), instead, a problem at time is approached, resolving two unidimensional problems.

4. Estimation

In order to estimate $\phi(\cdot)$, in the sequel we derive a model suitable for local regression.

A log version of (2) will be: $\log R_i^* = -\log \phi(S_i) + \log R_i$. Setting: $Y_i = \log R_i$, $\mu(\cdot) = \log \phi(\cdot)$, $\epsilon_i = \log R_i^*$, and $v(A)$ as $\text{Var}(A)$, we obtain

$$Y_i = \mu(S_i) + v^{1/2}(S_i) \epsilon_i \quad (4)$$

this model consists of a stochastic design variable S_i , a term of stochastic error ϵ_i , an unknown function $\mu(\cdot)$ and a response variable Y_i . The following procedure of conditional inference could be formulated.

Suppose n couples (s_i, y_i) are available; if: i) $E(\epsilon) = 0$; ii) $\text{Var}(\epsilon) = 1$; iii) $v^{1/2}(S_i) = \sigma \in \mathbb{R}^+$, we estimate the function $\mu(s) = E(Y|S = s)$ by $v(s) = \text{Var}(Y|S = s)$.

An interpretation of the model (4) is provided by

Proposition 1 *The (4) implies to assume an hazard model independent on S_i and rescaled by $\phi(S_i)$ (see Appendix):*

$$\Pr \{R_i > r|S_i\} = \phi(S_i) h_{R_i^*}(r\phi(S_i)) \quad (5)$$

The model (4) is appropriate for a nonparametric regression algorithm. From the theory above we can estimate $g(x|S = s)$ in a double manner: conforming to the steps iii) – v) or to the step suggested by

Proposition 2 *If (1) holds, proposition 1 implies (see Appendix):*

$$g(x|S = s) = g(x\phi(S_i)).$$

Consequently, steps iii) – v) of section 3 can be replaced by

iii') Obtain $\widehat{g}(x|S = s)$ estimating $g(\cdot)$ on the basis of the values $x_i\phi(s)$.

For $h_s(\cdot)$ estimation we use the kernel method. To estimate $g(\cdot)$ is preferable a local regression algorithm for the optimal boundary properties of this class of estimators.

Appendix

Proof of formula (1).

Since $S(r) = 1 - F(r)$, we have:

$$g(x) = 1 - S(x); \quad h(r) = \frac{f(r)}{S(r)} = -\frac{d \log S(r)}{dr}.$$

Then $S(r) = \exp\left(-\int_0^r h(u) du\right)$ and finally:

$$g(x) = 1 - \exp\left(-\int_0^x h(u) du\right).$$

Proof of Proposition 1.

Consider the hf of the random variable Y_i , with $i \in [1, \dots, N]$. We have:

$$\begin{aligned} \Pr\{R_i > r | s_i\} &= \Pr\{Y_i > \ln r | s_i\} \\ &= \Pr\{v^{1/2}(s_i) \epsilon_i > \ln r - \mu(s_i) | s_i\} \\ &= \Pr\{\exp(v^{1/2}(s_i) \epsilon_i) > r \exp(-\mu(s_i)) | s_i\} \\ &> r \exp(-\mu(s_i)) | s_i\} = 1 - F_{R_i^*}(r \exp(-\mu(s_i))). \end{aligned}$$

But:

$$\begin{aligned} h_{R_i} &= \frac{-d \ln \Pr\{R_i > r | s_i\}}{dr} \\ &= \frac{\exp(-\mu(s_i)) f_{R_i^*}(r \exp(-\mu(s_i)))}{1 - F_{R_i^*}(r \exp(-\mu(s_i)))} \\ &= \exp(-\mu(s_i)) h_{R_i^*}(r \exp(-\mu(s_i))) = \phi(s_i) h_{R_i^*}(r \phi(s_i)). \end{aligned}$$

Proof of Proposition 2.

If formula (1) holds, proposition 1 involves that

$$g(x|S=s) = 1 - \exp\left(-\int_0^x ah(ra) dr\right)$$

with $a = \phi(S_i)$. If $y = ra$ we have:

$$g(x|S=s) = 1 - \exp\left(-\int_0^{ax} h(y) dy\right) = g(ax).$$

References

- Chen, S. X. (1996). Studying school size effects in line transect sampling using the kernel method, *Biometrics* **52**, 1283–1294.
- Chen, S. X. (1999). Estimation in independent observer line transect surveys for clustered populations, *Biometrics* **55**, 754–759.
- Drummer, T. D. and McDonald, L. (1987). Size bias in line transect sampling, *Biometrics* **43**, 13–21.
- Mack, Y. P. and Quang, P. X. (1998). Kernel methods in line and point transect sampling, *Biometrics* **54**, 606–619.
- Skaug, H. J. and Schweder, T. (1999). Hazard models for line transect surveys with independent observers, *Biometrics* **55**, 29–36.