cirad

ONE WEEK 2018

A week around Systemic Approaches in Health

Introduction to Distance Sampling

Dr Mathieu BOURGAREL

Kasetsart University

Workshop « Reservoir Hosts »

21- 23 November 2018

Abundance estimation - background

- Goal – estimate abundance over some defined area (A)
- *N* = true number of a particular species within the area (population size or abundance)
- Might count individuals seen or heard in a sampled area (a)
- *n* = Nb of animals counted

Abundance & Study Area

- Implicit or explicit in estimating abundance is the associated "area"…
- So even when we measure abundance or population size, *N*, there is an implied measure of density (D) where

$$\hat{D} = \frac{n}{a}$$

D = number of individuals (*n*) per unit area (*a*)

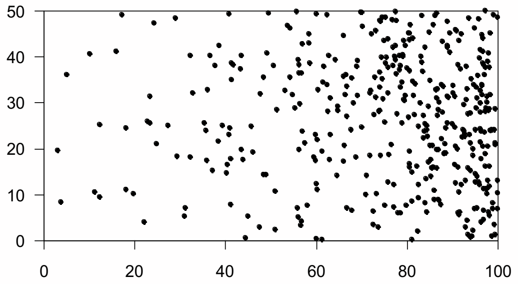
Abundance estimation: probability of detection

We have to estimate *N.*

*S*everal methods are available:

- variable-distance methods (includes line-transect and point-transect surveys)
- mark-recapture or mark-resight techniques (including removal methods)
- quadrat sampling with additional surveys to estimate detectability.

Estimation of the population size in the study area



Let

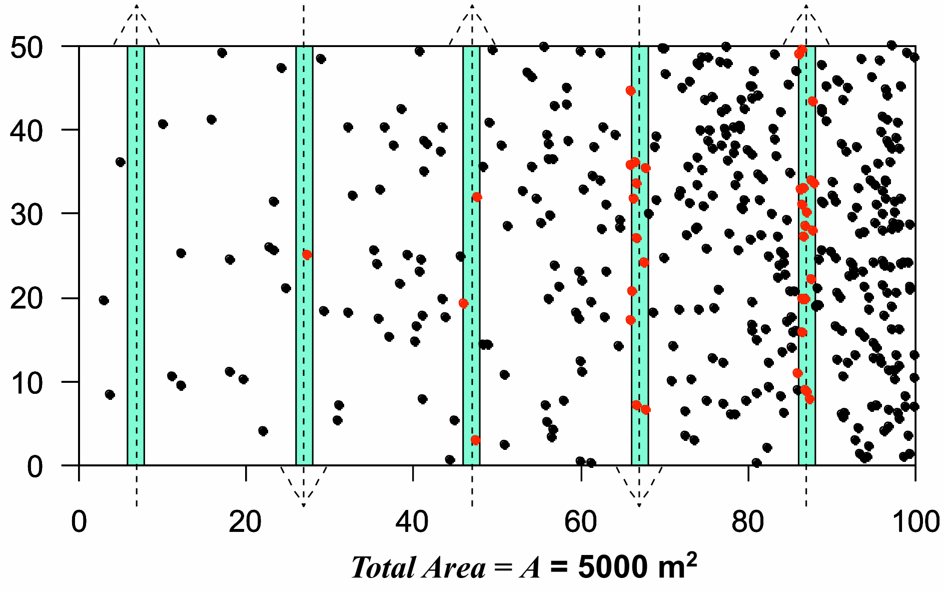
N = population size (abundance)
A = size of study region = 5000 m^2 or 0.5 hectares
D = Animal density = N/A

One method: count everything, i.e., a census (total count)

N = 412
D = 412/5000 = 0.0824 AI/m^2 or 824 AI/ha

Rarely possible in practice!

Strip Transect



Strip Transect

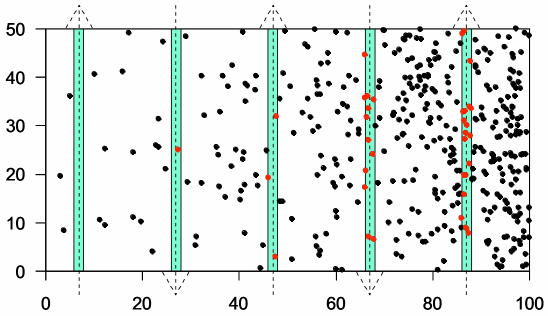
- Let

k = number of strips = 5

L = total line length = 50 x 5 = 250 m

w = the strip half-width = 1 m

Some sort of “random” placement of lines.



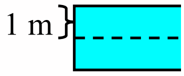
Strip Transect

If k = number of strips = 5,

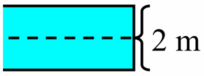
L = total line length = 50 x 5 = 250 m, and

w = the strip half-width = 1 m

Hypothesis : all animals within the strip are detected with a probability of 100 %



Impossible



=> biaised

then

a = area of region

n = number of animals counted/censused within the 5 strips = 36

Then extrapolate to area of interest :

$$\hat{N} = \frac{n}{a/A} = \frac{nA}{a} = \frac{36 \times 5000}{500} = 360$$

Distance Sampling Methods

- Distance sampling theory extends **the finite population sampling approach**
  - Given the *detection* of *n* objects, how many objects are estimated to occur within the sampled area?
- Implementation: In practice, a set of **randomly placed lines** or points are established and perpendicular distances are measured to those objects detected when traveling the line or surveying the points.
- The theory allows for the fact that some of the objects will be ***undetected*** and that there is a tendency for detectability to decrease with increasing distance from the transect line or point.

Line Transect

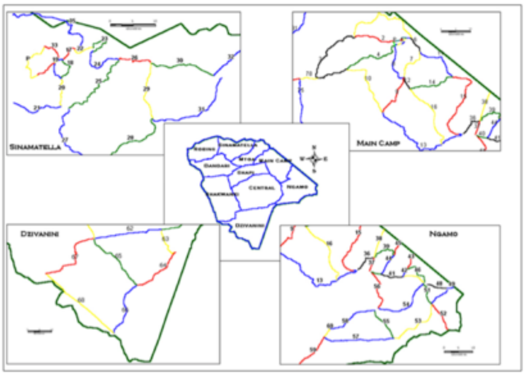
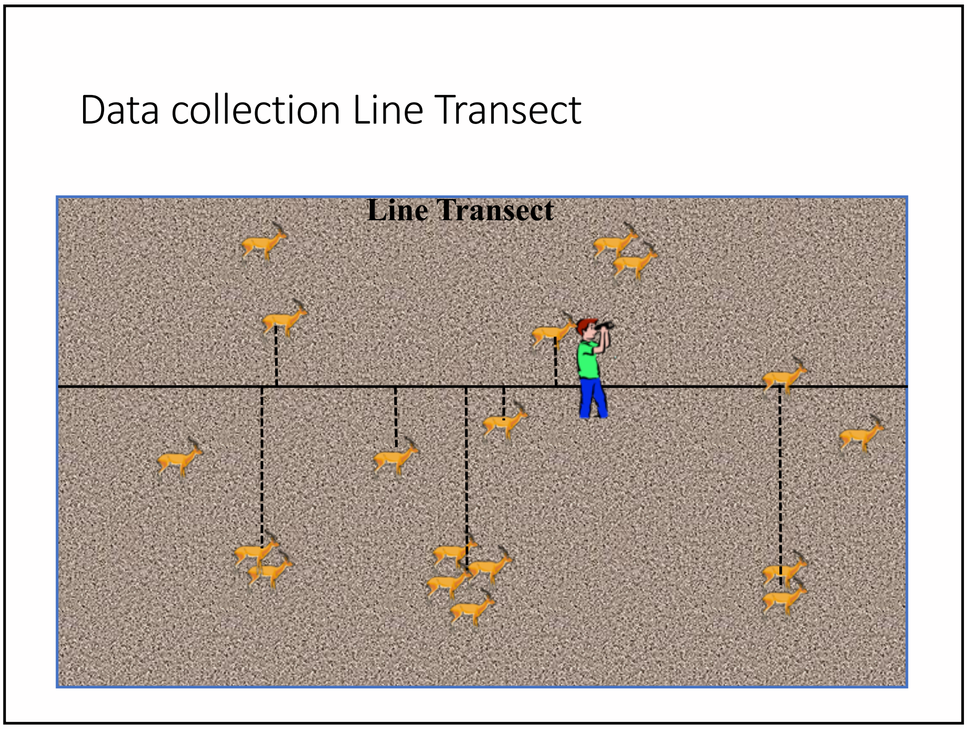
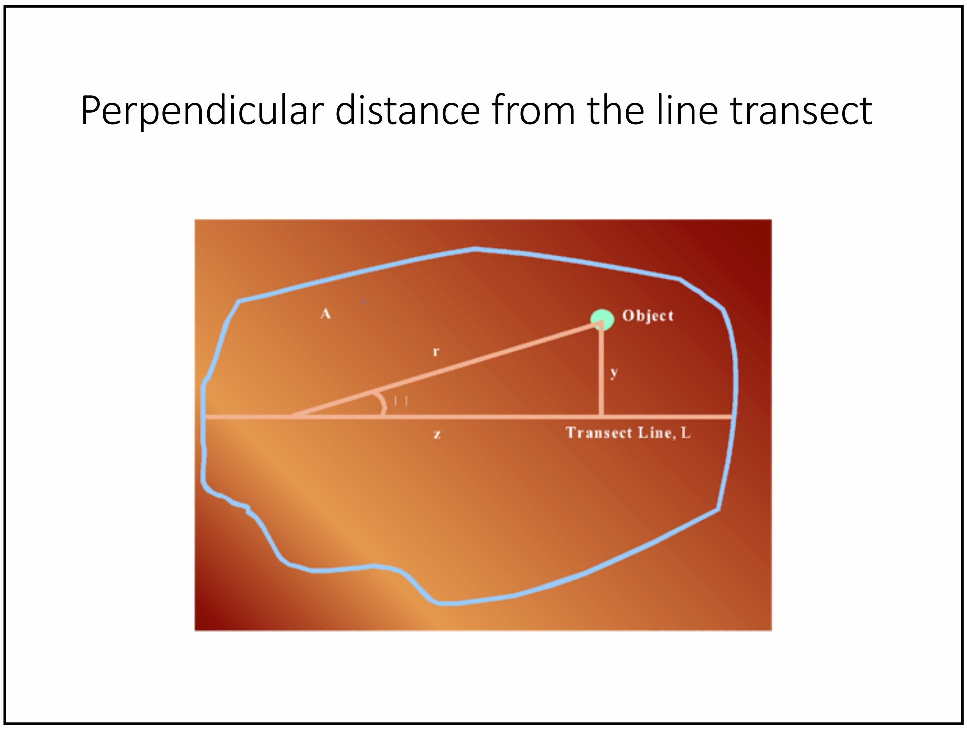


Figure 15 : Réseau routier dans 4 des blocks de Hwange et position des transects identifiés pour les comptages en voiture *Line transect* (chaque trait de couleur correspond à un transect).

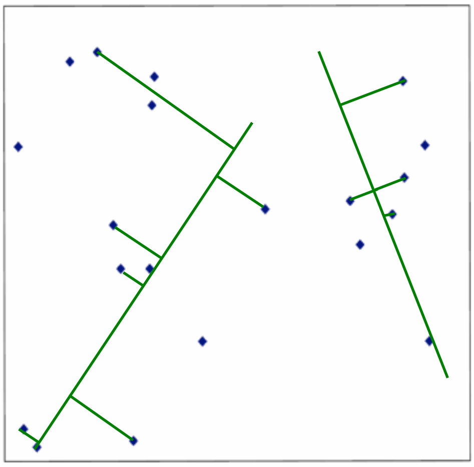


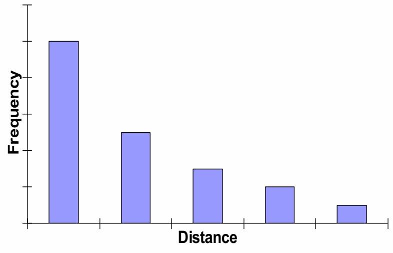
Data collection Line Transect



Perpendicular distance from the line transect

Distance Sampling



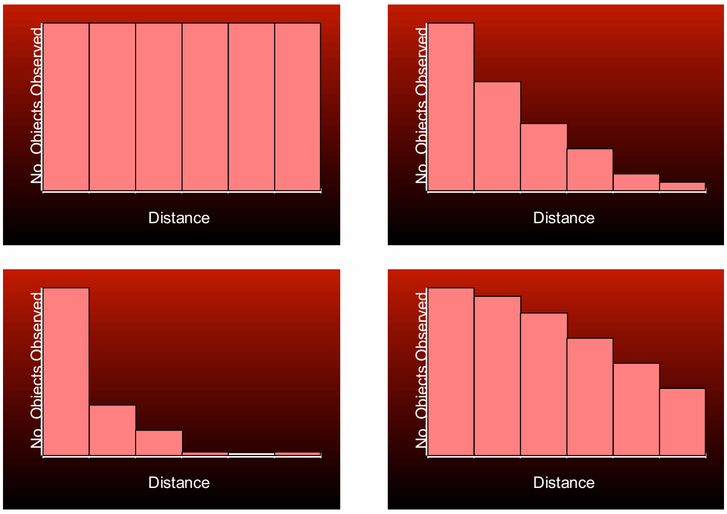


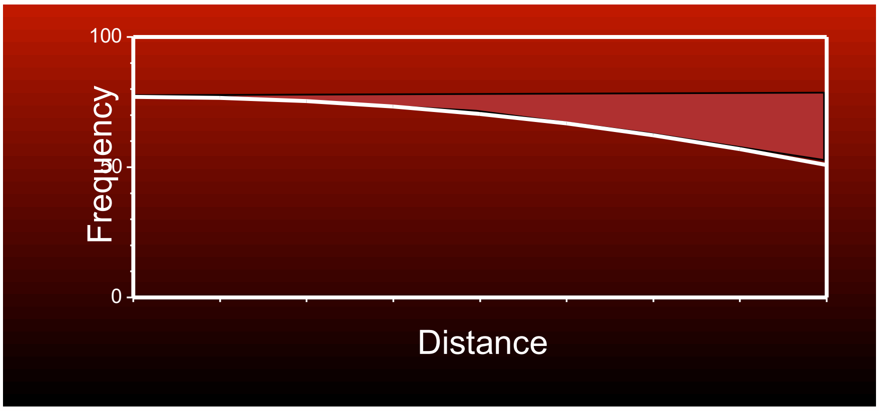
Distance Sampling

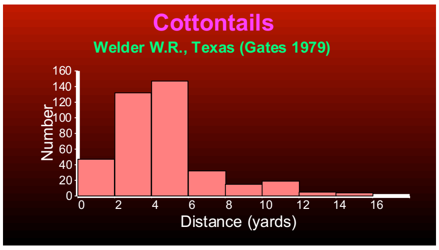
- General form of our density and abundance estimator

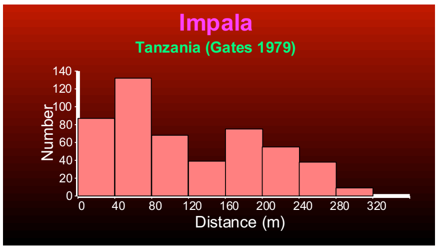
$$\hat{D} = \frac{n}{a\hat{P}_a} \qquad \hat{N} = \hat{D}A$$

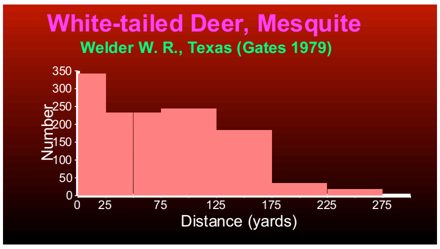
- Where:
  - *N* = abundance, n = count
  - P_a = detection probability in area ‘a’
  - *a* = area surveyed
  - A = total area
  - D = density

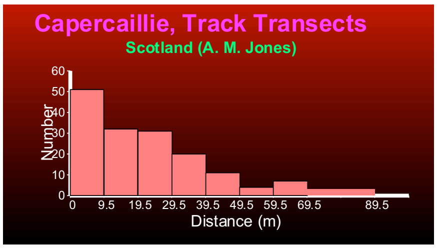

No. Objects Observed
Distance
No. Objects Observed
Distance
No. Objects Observed
Distance
No. Objects Observed
Distance


100
50
0
Frequency
Distance







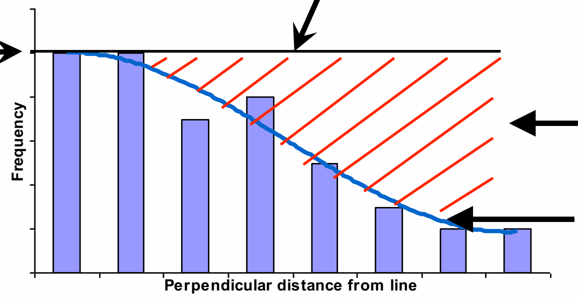


How is distance sampling data used?

- Fit a curve (i.e., a model!) to the data
- Intuitive estimate of probability of detection in the area, P_a , is the area under the curve (integration)

$$P_a = \frac{\text{area under curve}}{\text{area of rectangle}}$$

if you saw everything at all distances, on average the histogram bars should be here



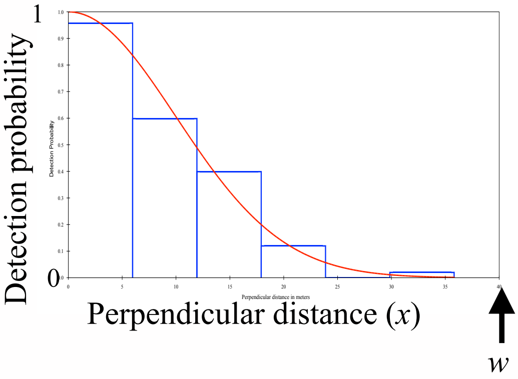
assume you see everything at zero distance

these animals are missed

Some function g(x)

Distance Sampling

The probability of detecting an object in the strip of area = 2*wL* is:



$$\hat{P}_a = \frac{\int_0^w \hat{g}(x)dx}{w}$$

Distance Sampling: density estimator

Substituting P_a into the density equation gives:

$$\hat{D} = \frac{n}{a\hat{P}_a} = \frac{n}{2Lw\hat{P}_a} = \frac{n}{2L\int_0^w \hat{g}(x)dx}$$

$$\hat{P}_a = \frac{\int_0^w \hat{g}(x)dx}{w}$$

D = density (number per unit area)

N = population size in the study area

n = number of objects or animals detected

P_a = probability of detection for an object within area, ***a***, regardless of its location.

w = width of area searched on each side of transect, or radius searched around point transect, or truncation point beyond which data not used in the analysis

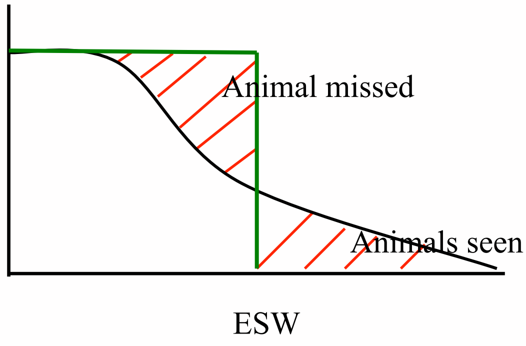
L = total line length,

g(0) = probability of detection on the line or point, usually assumed to be 1.

f(0) = the probability density function of detected distances from the line, evaluated at zero distance

Effective strip width (ESW)

The distance at which the number of animals missed within that distance is equal to the number of animals detected beyond that point



Modeling detection probability

Now, we need to model (estimate) an animals' detection probability:

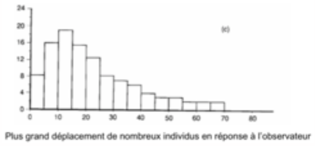
g(x) = Pr(animal detected/distance x)

What type of models might we use?

Crucial Assumptions

Unbiased estimates of density can occur from these distance data if certain assumptions are met. It is critical that these assumptions are met in establishment of field protocol:

1. Objects on the line or point are detected with probability one, p = 1.
2. Objects are detected at their initial location, prior to any movement.
3. Distances and angles are measured accurately.
4. Objects are spatially distributed in the area to be sampled according to some random process.
5. Randomly placed lines or points are surveyed.



Assumption 1: Objects on the line or point are detected with certainty

- Most Important assumption!
- Density estimate is biased low if g(0) < 1
- Examples of methods to ensure g(0) = 1
  - More effort on the line (slow consistent movements along line)
  - Video cameras (e.g., in aerial surveys)
  - One observer dedicated to observe animals on the line
- Advanced techniques when g(0) < 1
  - Multiple, independent observers to estimate objects missed on the line and adjust g(0).

Assumption 2: Objects are detected at their initial location

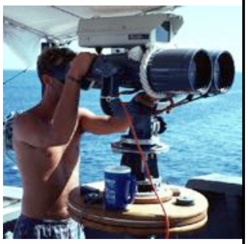
- Movement independent of observer is not a problem, but avoid counting objects multiple times on the same line or point.
- If animal reacts to the presence of the observer, distance must be measured from the position the animal was before reaction (when possible)
  - "Random" movement of animals that are detected at a later time is not a problem

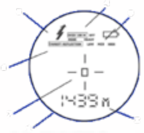
Assumption 2: Objects are detected at their initial location

- Movement away from or towards the line/point due to the observer creates bias
  - Animals that move, or flush, due to observer will cause a negative bias in density
  - Attraction to the line/point will cause a positive bias in density (e.g., mobbing behavior of birds)
  - Note: difficult to assess bias.

Assumption 3: Measurements are exact

- Exact measurements are obviously very important.
  - Use of measurement aids are highly recommended to reduce bias in approximations of observers.
    - Tape measures; Laser range finders; Compass for angles
- “Grouping” can reduce issues of measurement error, but obviously easier to begin with accurate measurements.





0
30
30
60
60
90
90

Assumption 3: Measurements are exact

- Established groupings can be used when exact measurements are difficult
  - 5-7 intervals
- Heaping – can result from “estimated” distances because of tendency of humans to choose distances at 0, 5, 10, etc. units.
- If angles and distances are measured, do not group prior to estimating perpendicular distances.
  - Put another way: if you are going to group measurements that include angles, do so during the analysis using the perpendicular distances.

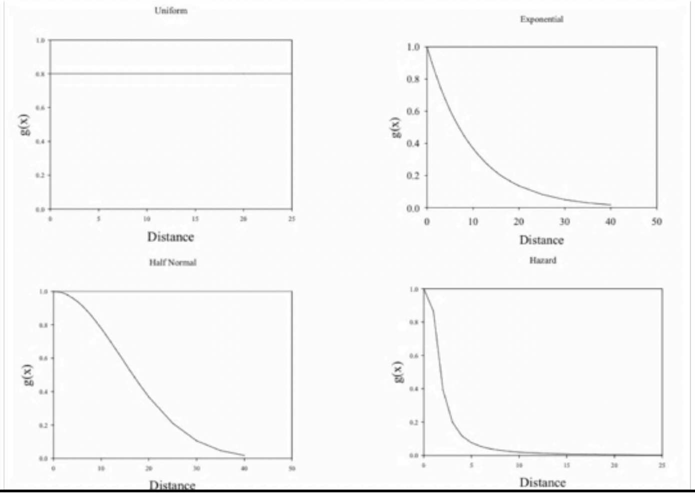
Assumption 3: Measurements are exact

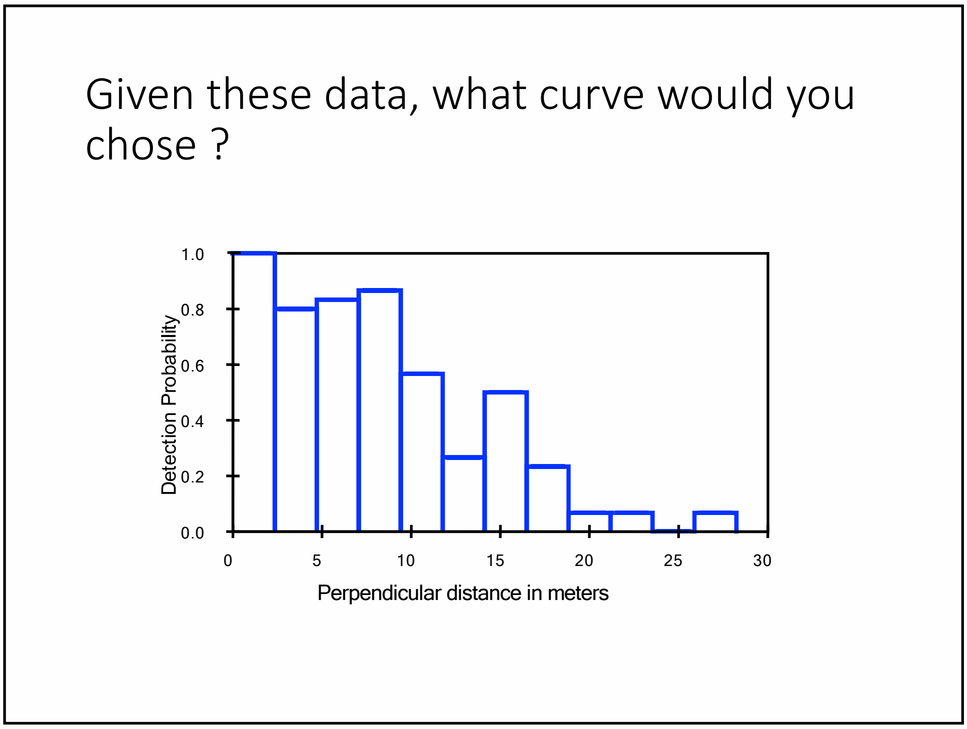
- Outliers: A few extreme outliers have little effect on the density estimate and should be truncated.
    - Including outliers can increase variance estimates of density, var(*D*), because detection function may overfit and add additional parameters.

Modeling the Detection Function

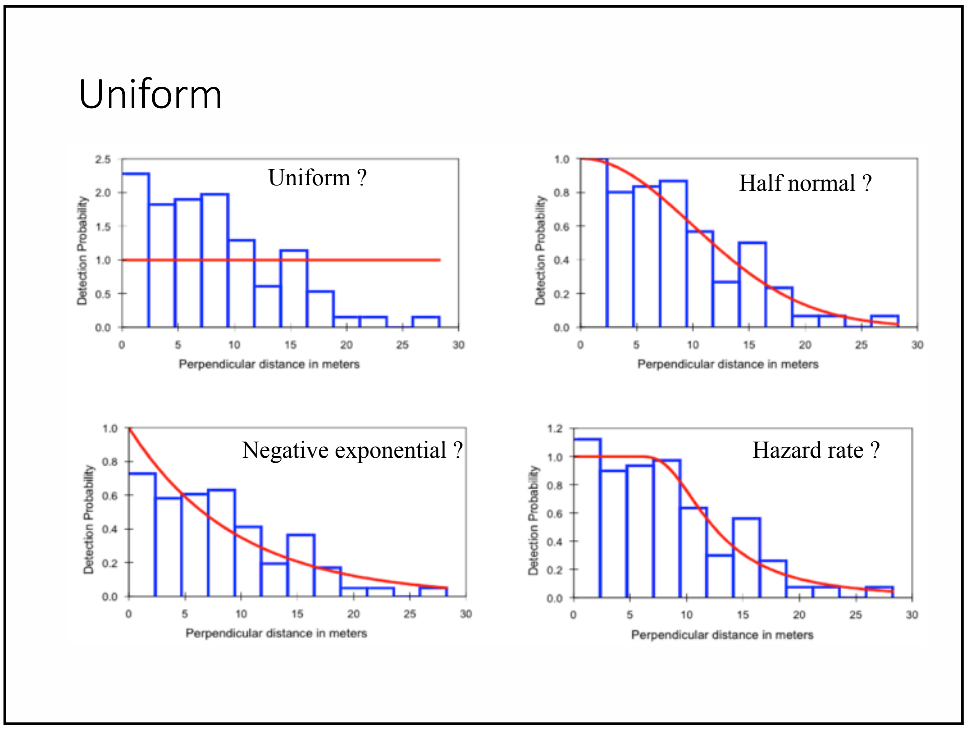
Several distribution curves available :

-**Uniform**	**- exponential**
-***half-normal***	**- hazard rate**

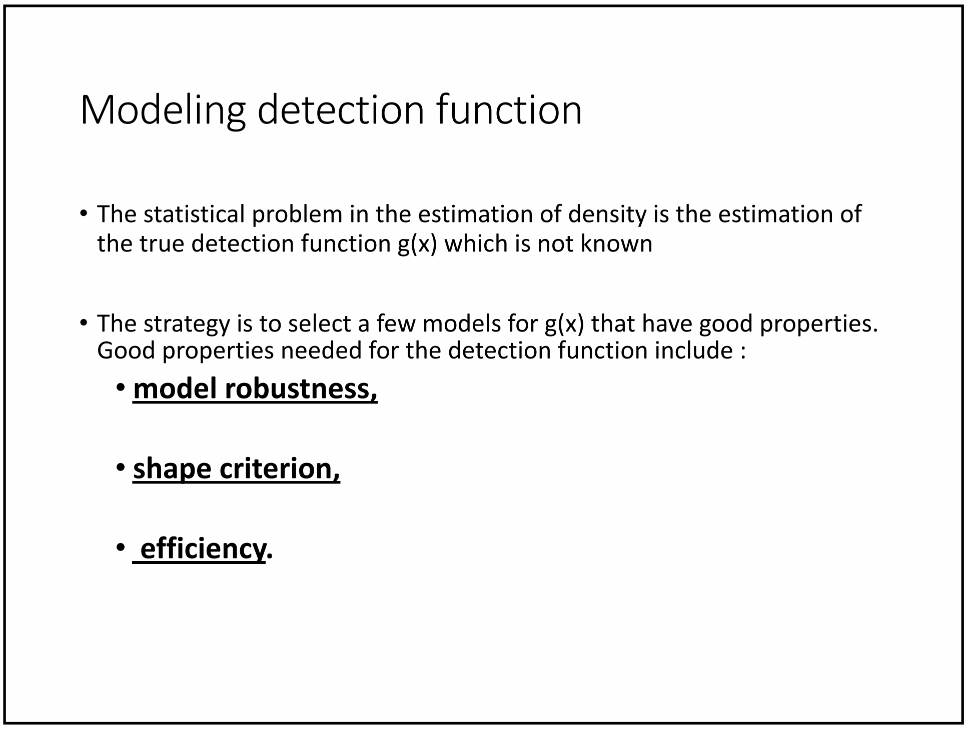




Given these data, what curve would you chose ?

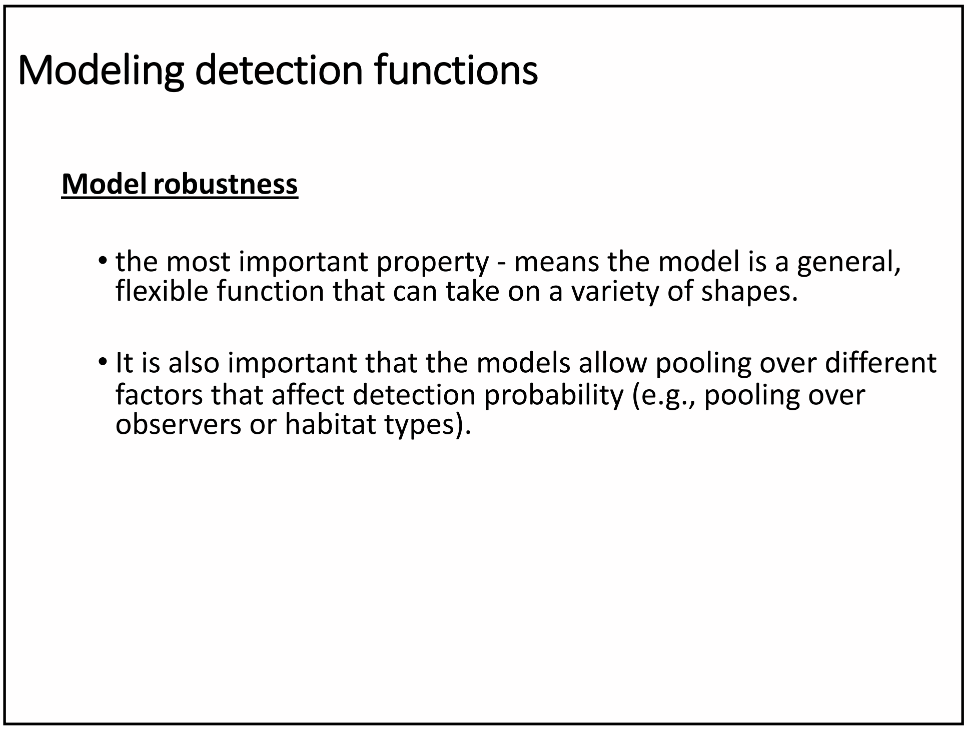


Uniform



Modeling detection function

- The statistical problem in the estimation of density is the estimation of the true detection function g(x) which is not known
- The strategy is to select a few models for g(x) that have good properties. Good properties needed for the detection function include :
  - **model robustness,**
  - **shape criterion,**
  - **efficiency.**



Modeling detection functions

Model robustness

- the most important property - means the model is a general, flexible function that can take on a variety of shapes.
- It is also important that the models allow pooling over different factors that affect detection probability (e.g., pooling over observers or habitat types).

Modeling detection functions

Efficiency

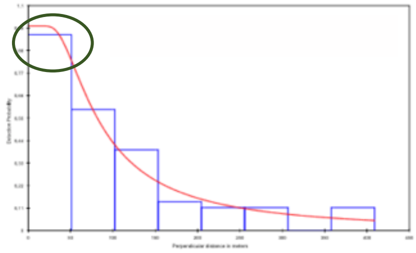
- the model provides estimates with small variances (that are relatively precise).

Modeling detection functions

Shape criterion

- the detection function should be monotonically decreasing and have a "**shoulder**" near the transect line or point. That is, detection remains nearly certain out to some distance from the transect line. Program DISTANCE includes a shape criterion that forces it to fit functions with a shoulder.

- In order to m ɜn grouped into distance rvations removed fror



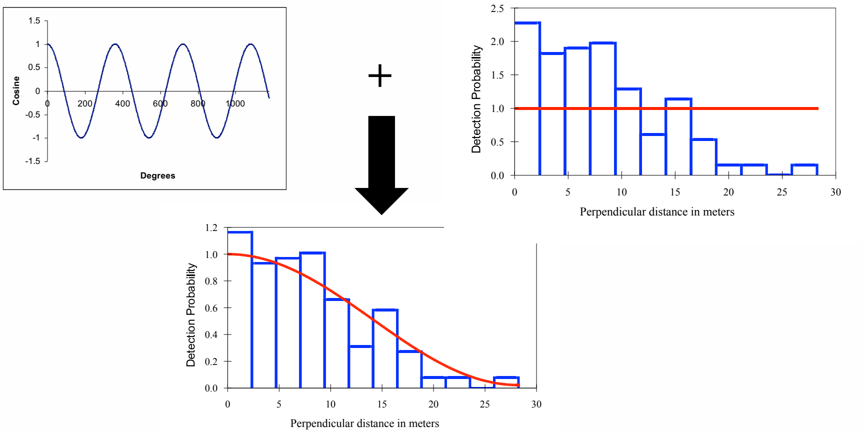
Detection function models

- Take the form of a ‘key’ function
  - uniform,
  - half-normal,
  - hazard-rate
  - Negative exponantial

- plus a ‘series’ expansion
  - cosine,
  - simple polynomial,
  - hermite polynomial) => adjustment

Series expansions for adjusting key functions

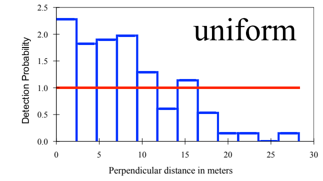
- Uniform with a single cosine adjustment term
- Adds parameters



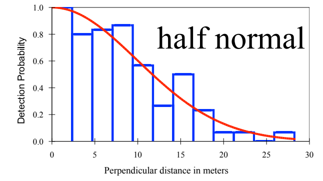
RECOMMANDED MODELS for density estimation
(Thomas et al. 2010)

- Uniform + Cosine
- Half Normal + Cosine
- Half Normal + Hermite
- Hazard Rate + Polyniomal

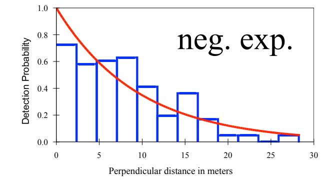
How do we selected among models? And among number of adjustment terms?



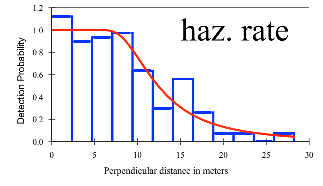
$g(x) = 1/w$



$g(x) = e^{-x^2/2\sigma^2}$

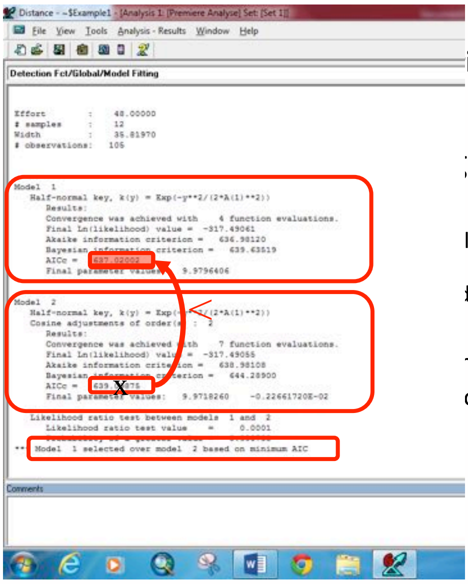


$g(x) = e^{-x/\lambda}$



$g(x) = 1 - e^{-(x/\sigma)^{-b}}$

Model selection


Distance - ~$Example1 - [Analysis 1: [Premiere Analyse] Set: [Set 1]]
File View Tools Analysis - Results Window Help
Detection Fct/Global/Model Fitting
Comments

```
Effort        :    48.00000
# samples     :    12
Width         :    35.81970
# observations:   105

Model  1
   Half-normal key, k(y) = Exp(-y**2/(2*A(1)**2))
      Results:
      Convergence was achieved with    4 function evaluations.
      Final Ln(likelihood) value =  -317.49061
      Akaike information criterion =   636.98120
      Bayesian information criterion =   639.63519
      AICc =    637.02002
      Final parameter values:   9.9796406

Model  2
   Half-normal key, k(y) = Exp(-y**2/(2*A(1)**2))
   Cosine adjustments of order(s) :  2
      Results:
      Convergence was achieved with    7 function evaluations.
      Final Ln(likelihood) value =  -317.49055
      Akaike information criterion =   638.98108
      Bayesian information criterion =   644.28900
      AICc =    639.09875
      Final parameter values:   9.9718260    -0.22661720E-02

   Likelihood ratio test between models  1 and  2
      Likelihood ratio test value    =     0.0001
      [illegible]
*** Model  1 selected over model  2 based on minimum AIC
```

X

iterion: AIC

$_e$(L) + 2q

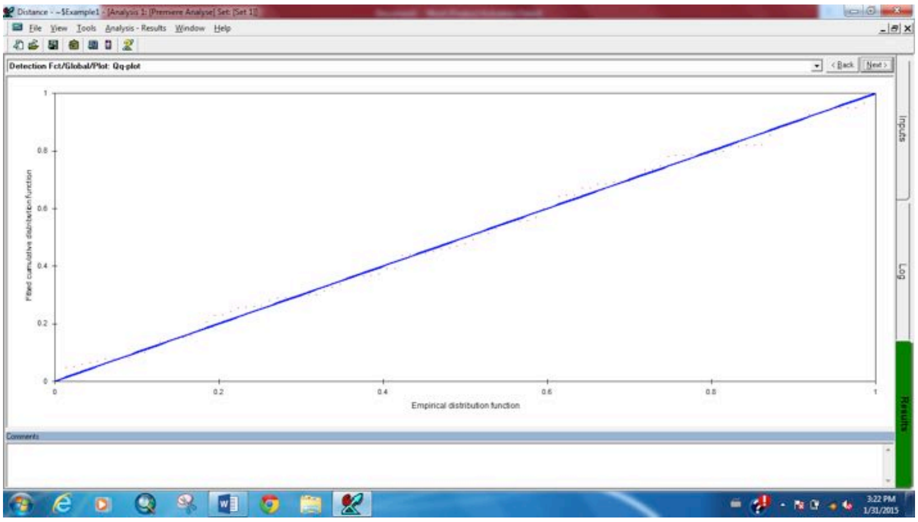
luated at the maximum parameters) an q is the del

nallest AIC

of fit

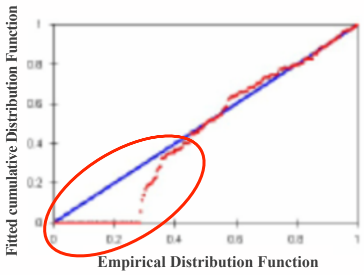
Goodness of Fit

- 1st test : q-q plot

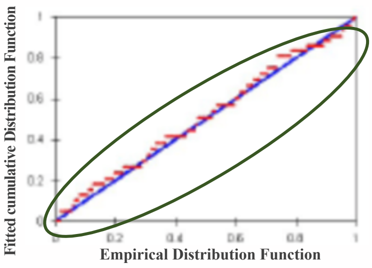

Distance - ~$Example1 - [Analysis 1: [Premiere Analyse] Set: [Set 1]]
File View Tools Analysis - Results Window Help
Detection Fct/Global/Plot: Qq-plot
< Back
Next >
Fitted cumulative distribution function
Empirical distribution function
0
0.2
0.4
0.6
0.8
1
Inputs
Log
Results
Comments
3:22 PM
1/31/2015

Goodness of Fit

- 1st test : q-q plot



The model doesn't fit

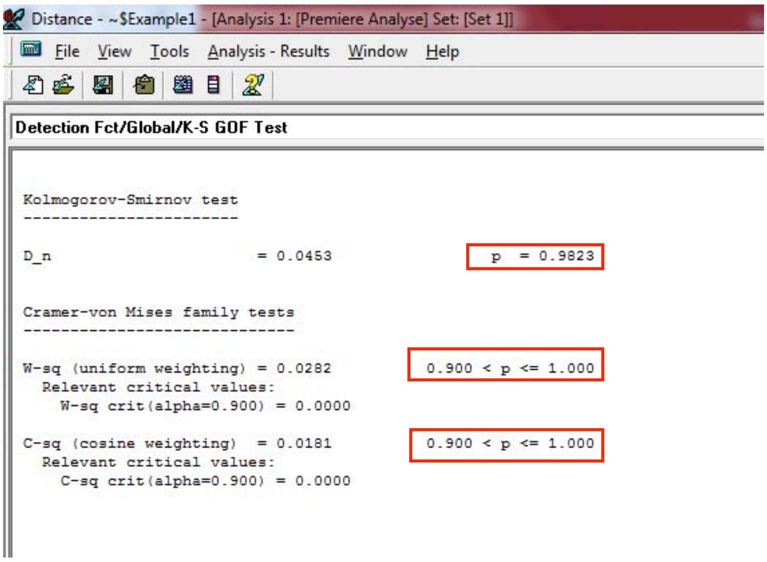


The model fits

Goodness of Fit

- **2nd Test : Test Kolmogorov Smirnov**
  - Non parametric test used to compare 2 distributions (observed data and estimated date produced by the model)
- **3rd Test : Test Cramés-von Mises:**
  - Compare 2 distributions (observed data and estimated date produced by the model) .
  - Possibility to give a weight to the data (especially near 0)
- 2 test must be non significative (P> 0,05): no difference between the 2 distributions (observed and estimated)

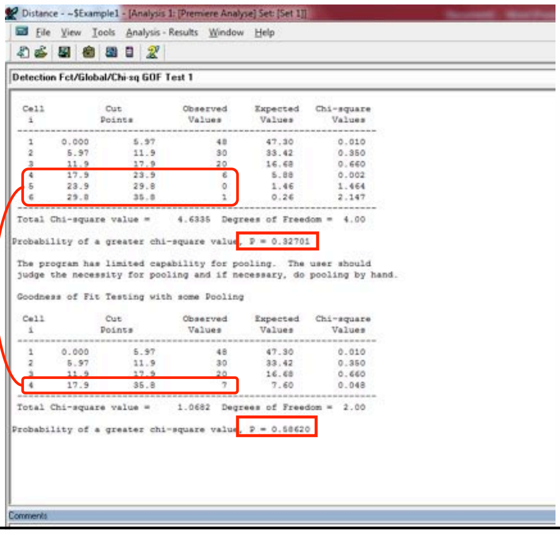
Goodness of Fit


Distance - ~$Example1 - [Analysis 1: [Premiere Analyse] Set: [Set 1]]
File View Tools Analysis - Results Window Help
Detection Fct/Global/K-S GOF Test

```
Kolmogorov-Smirnov test
-----------------------

D_n                      = 0.0453              p  = 0.9823


Cramer-von Mises family tests
-----------------------------

W-sq (uniform weighting) = 0.0282         0.900 < p <= 1.000
  Relevant critical values:
    W-sq crit(alpha=0.900) = 0.0000

C-sq (cosine weighting)  = 0.0181         0.900 < p <= 1.000
  Relevant critical values:
    C-sq crit(alpha=0.900) = 0.0000
```

Goodness of Fit

- 4th test : Chi-Square test


Distance - ~$Example1 - [Analysis 1: [Premiere Analyse] Set: [Set 1]]
File View Tools Analysis - Results Window Help
Detection Fct/Global/Chi-sq GOF Test 1
Comments

```
 Cell          Cut           Observed    Expected   Chi-square
  i           Points          Values      Values      Values
 ----------------------------------------------------------------
  1     0.000       5.97          48       47.30       0.010
  2      5.97       11.9          30       33.42       0.350
  3      11.9       17.9          20       16.68       0.660
  4      17.9       23.9           6        5.88       0.002
  5      23.9       29.8           0        1.46       1.464
  6      29.8       35.8           1        0.26       2.147
 ----------------------------------------------------------------
 Total Chi-square value =    4.6335  Degrees of Freedom =  4.00

Probability of a greater chi-square value, P = 0.32701

 The program has limited capability for pooling.  The user should
 judge the necessity for pooling and if necessary, do pooling by hand.

 Goodness of Fit Testing with some Pooling

 Cell          Cut           Observed    Expected   Chi-square
  i           Points          Values      Values      Values
 ----------------------------------------------------------------
  1     0.000       5.97          48       47.30       0.010
  2      5.97       11.9          30       33.42       0.350
  3      11.9       17.9          20       16.68       0.660
  4      17.9       35.8           7        7.60       0.048
 ----------------------------------------------------------------
 Total Chi-square value =    1.0682  Degrees of Freedom =  2.00

Probability of a greater chi-square value, P = 0.58620
```

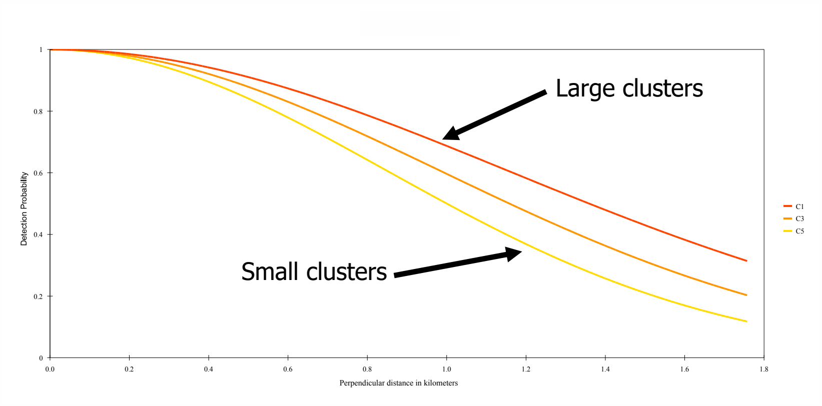
Clusters

- When objects are naturally clustered, it is better to consider the "object of interest" as the cluster versus the individuals within the cluster.
  - Herds of antelope

Cluster

- Measurements are from the point or line to the center of the cluster.
- Sample size, *n*, is the number of clusters versus to the number of individual objects.
- Size bias: larger clusters have a tendency to be more detectable,
  - result is an overestimate of cluster size as small-sized clusters are missed.
- How can size bias detected and accounted for?

Scale parameter a function of cluster size, *s*

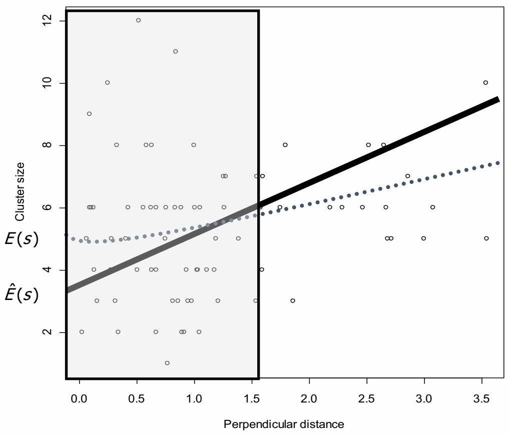


Large clusters

Small clusters

Regression methods

Linear regression of *s* on *x*



Distance Sampling:
Line transect density estimator

- We know *g(0)* = 1 (the key assumption for distance sampling), so using the assumption that g(0) = 1, the pdf can be evaluated at zero distance so that f(0), the pdf, can also be evaluated at distance 0 :

$$f(0) = \frac{g(0)}{\int_0^w g(x)} = \frac{1}{\int_0^w g(x)dx}$$

- When *f(x)* is evaluated at 0, (the fitted pdf for perpendicular distances evaluated at zero distance), it is a function of the measured distances.

Now the general estimator of density for line transect sampling, based on the measured distances, is:

$$\hat{D} = \frac{n}{2LwP_a} = \frac{n}{2L\int_0^w \hat{g}(x)dx} \rightarrow\rightarrow \hat{D} = \frac{n \cdot \hat{\mathrm{f}}(0)}{2\mathrm{L}}$$

Note: *D* and f(0) are estimated!

Estimate of Precision

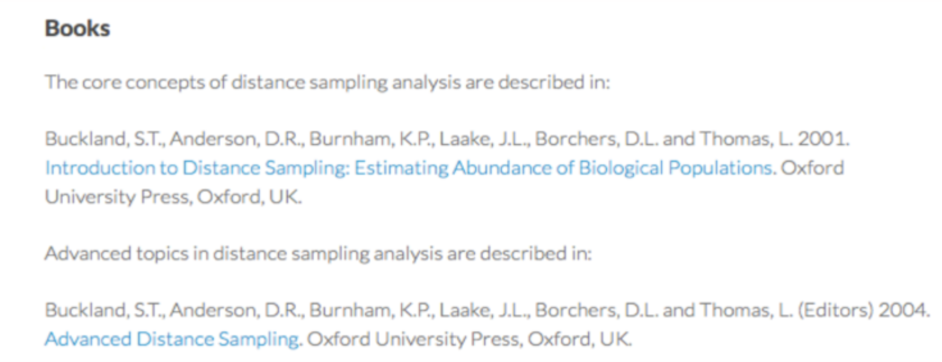
- Having produced a point estimate of density, we must also produce an estimate of the precision. This variance formula consists of two pieces; uncertainty in *f(0)* and uncertainty in the number of individuals we detect:

$$\hat{\mathrm{var}}(\hat{D}) = \hat{D}^2 \left[\frac{\hat{\mathrm{var}}(n)}{n^2} + \frac{\hat{\mathrm{var}}(\hat{f}(0))}{\hat{f}(0)^2} \right]$$

- An empirical estimate of variance using bootstrapping over the lines or points is often a better approach and is an option in Program DISTANCE

Recommanded literature

- Software :
  - http://distancesampling.org/Distance/distance72download.html



Books

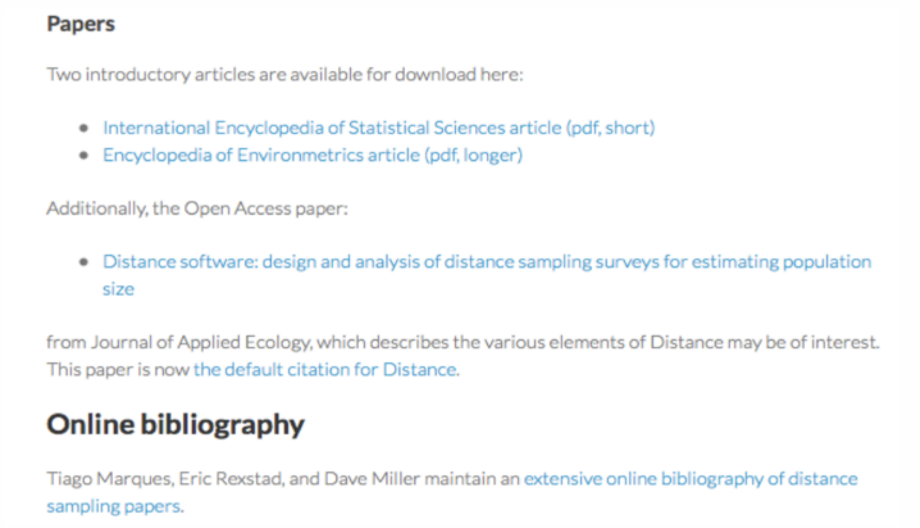
The core concepts of distance sampling analysis are described in:

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. 2001. Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press, Oxford, UK.

Advanced topics in distance sampling analysis are described in:

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. (Editors) 2004. Advanced Distance Sampling. Oxford University Press, Oxford, UK.

Recommanded literature



Papers

Two introductory articles are available for download here:

- International Encyclopedia of Statistical Sciences article (pdf, short)
- Encyclopedia of Environmetrics article (pdf, longer)

Additionally, the Open Access paper:

- Distance software: design and analysis of distance sampling surveys for estimating population size

from Journal of Applied Ecology, which describes the various elements of Distance may be of interest. This paper is now the default citation for Distance.

Online bibliography

Tiago Marques, Eric Rexstad, and Dave Miller maintain an extensive online bibliography of distance sampling papers.