Distance sampling: Estimating densities of wildlife populations by modelling detectability

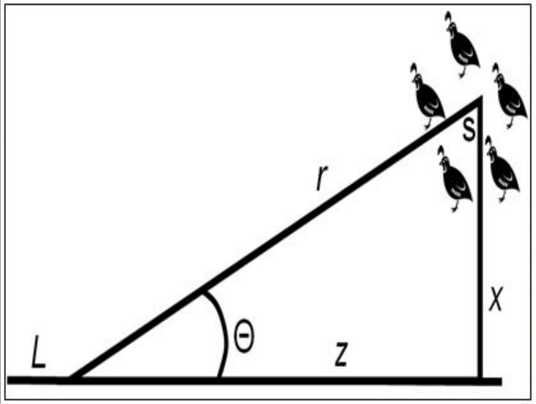
David Maphisa & Florian Weller

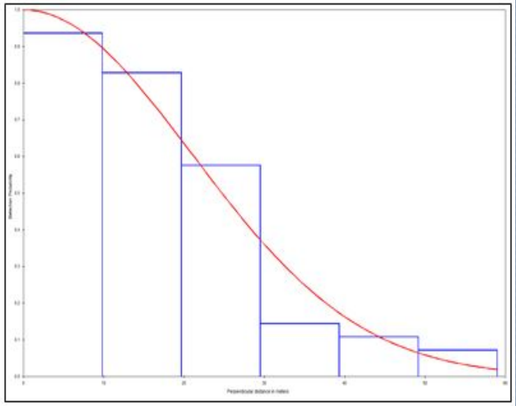
University of St Andrews

SEEC - Statistics in Ecology, Environment and Conservation

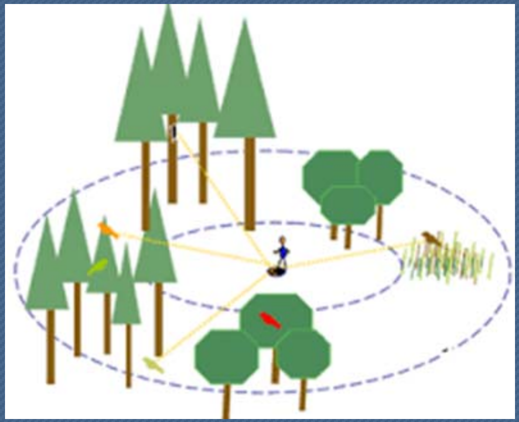
University of Cape Town · Universiteit van Kaapstad · iYunivesithi yaseKapa

Distance sampling is applicable to many species groups (birds, plants, mammals, whales etc.) Surveys can be done on foot, plane, boat etc.

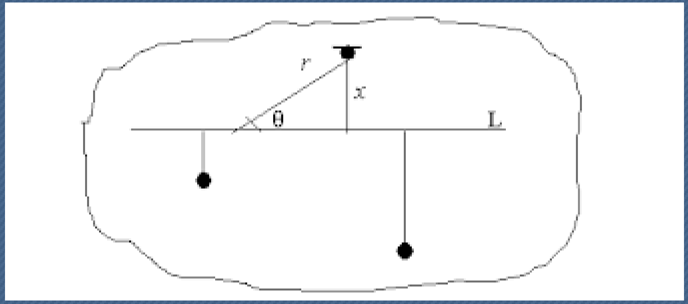


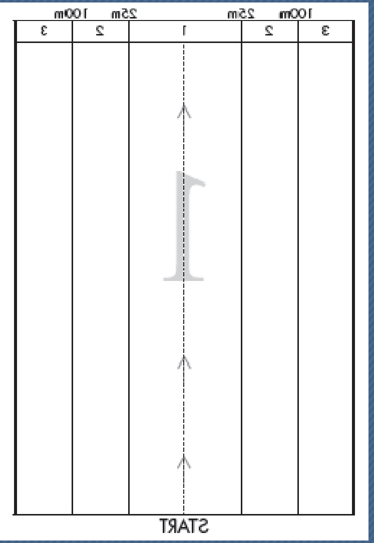


1. Point counts



2. Transects counts





The key is to account for observation process

Many factors can determine your ability to observe the species accurately
eg. vegetation (grassland birds)

Time of the day (or temp)

Cloud cover
etc……………………………

How many animals can you see on the 2nd pic below ?

We present two approaches to analysing your distance data

1. Package *unmarked* (function disamp) in R (Chandler et al. 2015)

I demonstrate this by using bird and vegetation data from my Phd thesis

My study site (FS-KZN boundary)

2. Florian will demonstrate Program DISTANCE (standalone) and talk more about assumptions and biases

SEEC - Statistics in Ecology, Environment and Conservation

The way package *unmarked* is implemented in R

Hierarchical distance sampling

Observation process/detection process
Allows inclusion of factors that may affect your ability to see or not see animals etc

Biological process/density (ha)
include factors that affect density

Data requirements

1. Bird distance sampling data (in my case fixed distance bands)

2. Habitat data - in my case detctn & density are affected by

 (i) fire – whether each transect was burned or not burned

(ii) Grazing – in 3 categories (none, light or heavily grazed)

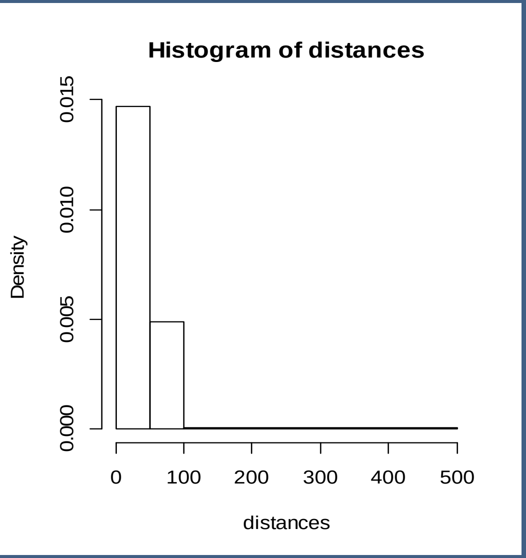
(iii) My main focus was on grass height (avh) and grass cover (cover)

all these datasets (birds and vegtn must be loaded in R) – merged into one object

Once this is done actual analysis is carried out (first detection and then density)

Out of eight species – work on one spcs at a time

```
data.PA <- bldata[bldata$Species==“African Pipit",]
data.PA$Season <- factor(data.PA$Season)

summary(data.PA)

ltUMF <- with(data.PA, {
 unmarkedFrameDS(y = cbind(X50m, X100m, X100m.1),
 siteCovs = data.frame(Year, Season, Area),
dist.breaks = c(0, 50, 100, 500)
tlength = rep(500,dim(data.PA)[1]), survey = "line", unitsIn = "m")})
ltUMF
summary(ltUMF)
hist(ltUMF)
```

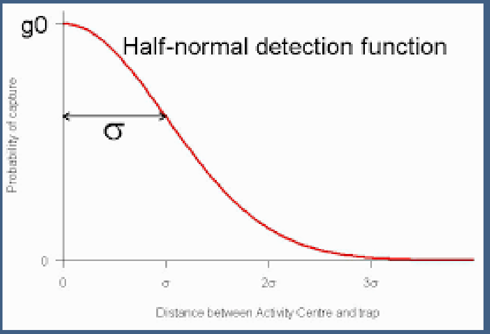


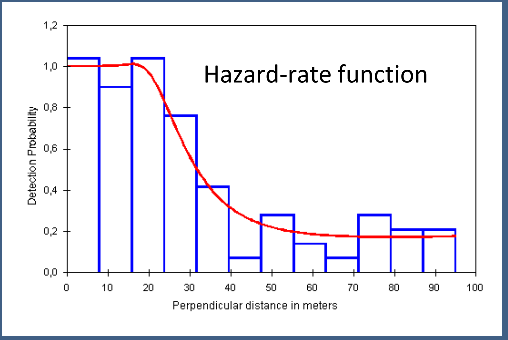
Model fitting – testing which model best fit your data
Several functions are available

(LcCm1_default <- distsamp(~ 1 ~ 1, ltUMF)) # default same as below

(LcCm1_halfNorm <- distsamp(~ 1 ~ 1, ltUMF, keyfun = "halfnorm", output = "density", unitsOut = "ha")) #

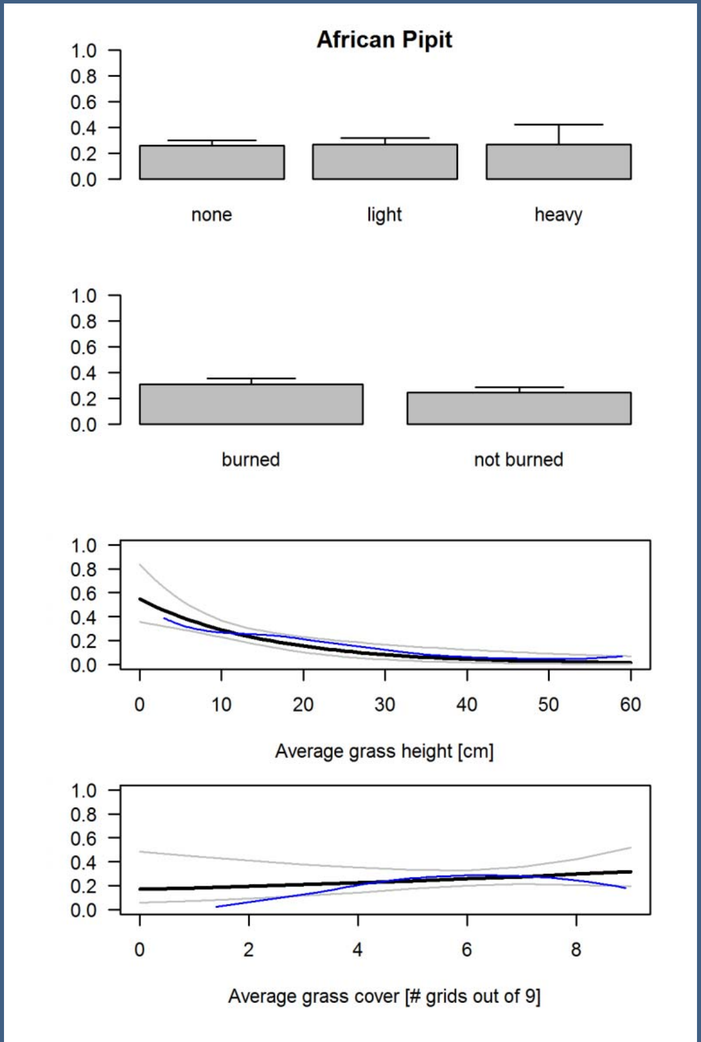
(LcCm1_hzrd <- distsamp(~ 1 ~ 1, ltUMF, keyfun = "hazard", output = "density", unitsOut = "ha")) #
Upon running the above codes /models - AIC values are produced to decide on best fit functn





For demonstration purposes this is what I did for African Pipit

```
PAmdetcn <- distsamp(~ (avh + cover) ~ 1, ltUMF))

(PAm8 <- distsamp(~ (avh + cover) ~ Grazing, ltUMF))
(PAm9 <- distsamp(~ (avh + cover) ~ Burning))
(PAm10 <- distsamp(~ (avh + cover) ~ avh, ltUMF))
(PAm11 <- distsamp(~ (avh + cover) ~ cover, ltUMF))
```



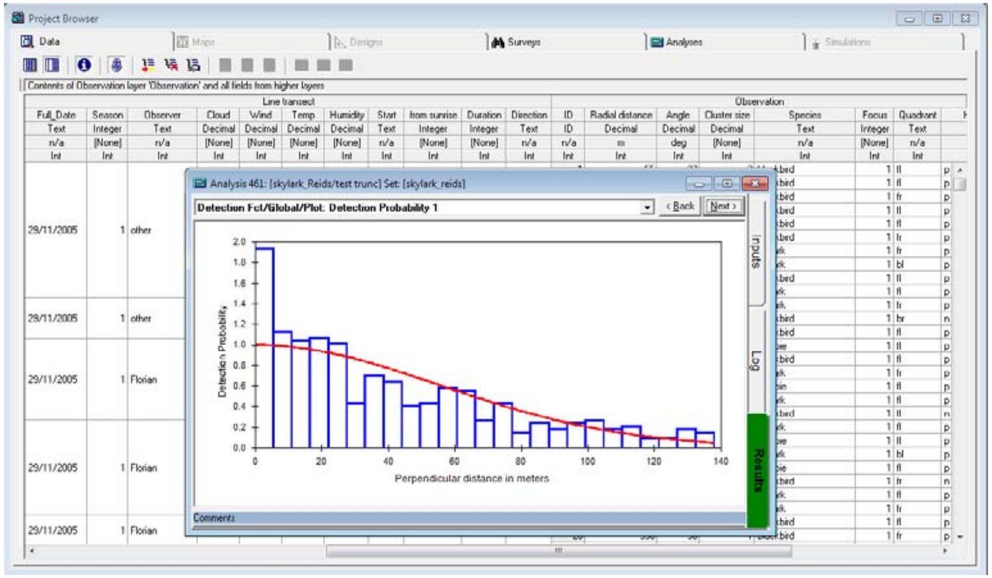
SEEC - Statistics in Ecology, Environment and Conservation

- **Multiple covariate distance analysis with program DISTANCE**
    - example 1: data in intervals
    - example 2: exact data
- **More about the basic assumptions of distance sampling**
    - how to deal with possible violations
- **General considerations about the modeling process**

Distance analysis with program DISTANCE

DISTANCE is a standalone Windows program for designing and analysing DS surveys . It is developed and maintained by a group based at the Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews.

- the same group maintains a number of DS R packages (primarily package *Distance*)
  - (note, this is a different implementation than *unmarked*)
- compatible with R on various levels; several of its components are in fact R modules
  - package *readdst* can port data and models directly from DISTANCE to R



Distance analysis with program DISTANCE

DISTANCE is a standalone Windows program for designing and analysing DS surveys . It is developed and maintained by a group based at the Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews.

DISTANCE vs R

Pro:

- easier entry and learning curve (learning R and DS at the same time can be daunting)
- better documentation (IMO)
- convenient output, analysis and error checking options
- survey design functionality (using spatial data)

Con:

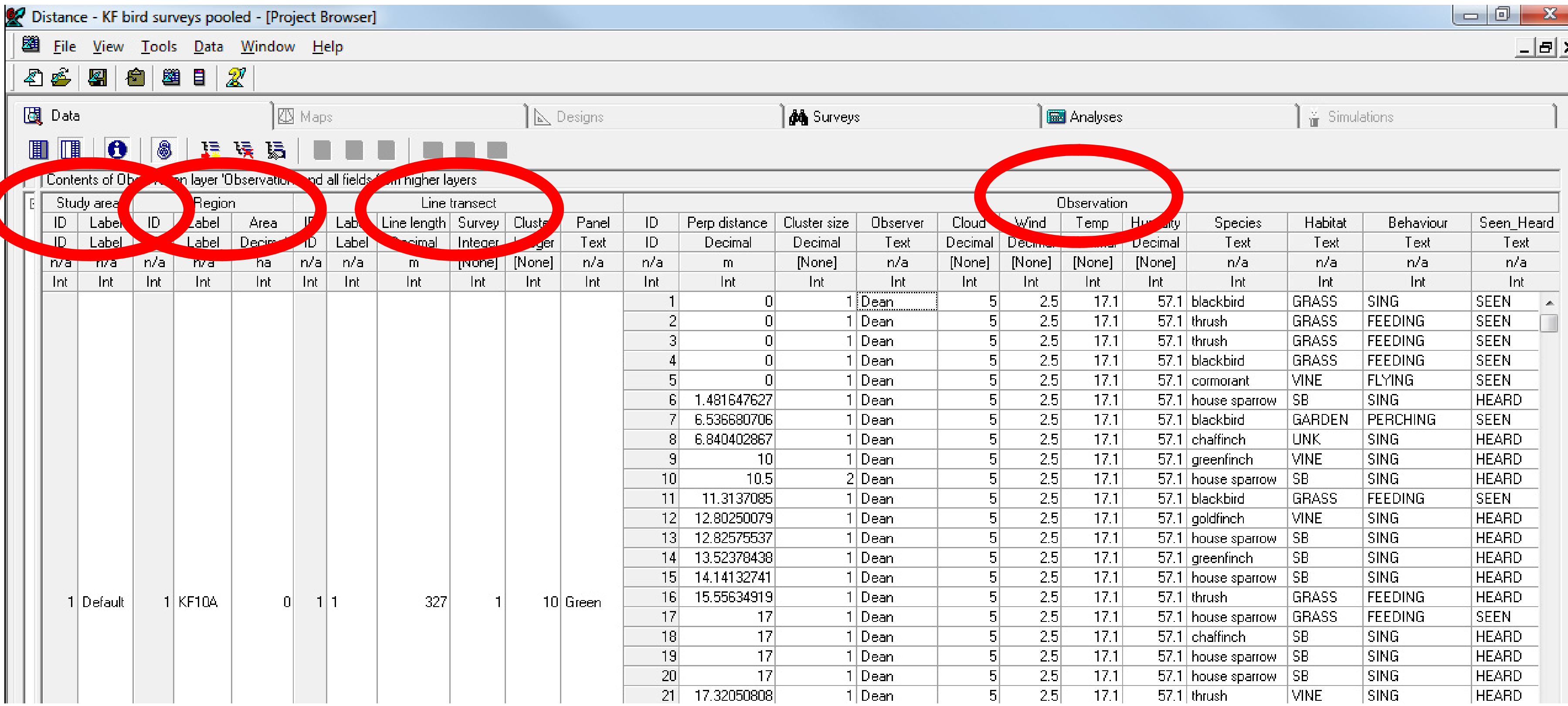
- R allows greater customization and better data handling
- development happens primarily with R packages and is then ported, thus DISTANCE can be expected to lag a little in development

Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

- **import data**
  - store data in external file and import (using import wizard), rather than entering it directly in DISTANCE
- set up **data filter** and **model definition**
- run **analyses** as combinations of filters and definitions

Example analysis:

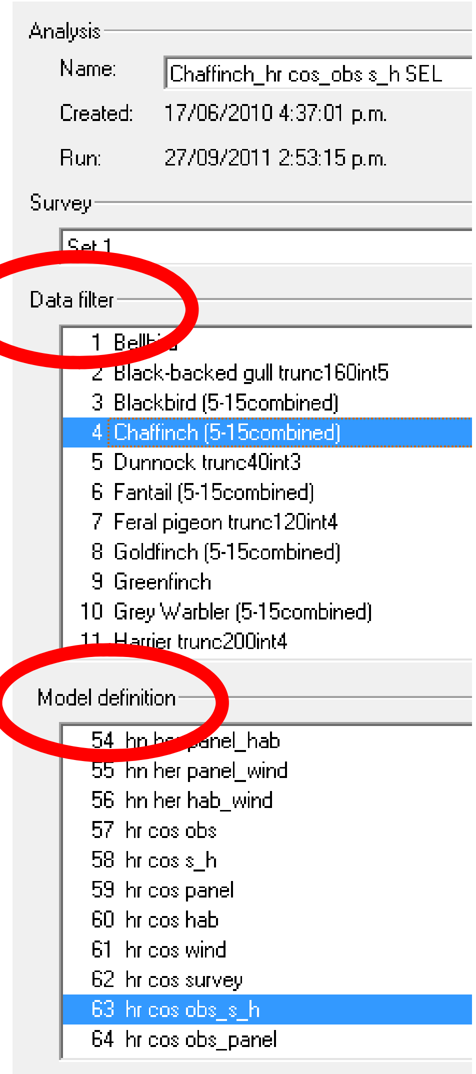
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand



Study area		Region			Line transect						Observation											
ID	Label	ID	Label	Area	ID	Label	Line length	Survey	Cluster	Panel	ID	Perp distance	Cluster size	Observer	Cloud	Wind	Temp	Humidity	Species	Habitat	Behaviour	Seen_Heard
ID	Label	ID	Label	Decimal	ID	Label	Decimal	Integer	Integer	Text	ID	Decimal	Decimal	Text	Decimal	Decimal	Decimal	Decimal	Text	Text	Text	Text
n/a	n/a	n/a	n/a	ha	n/a	n/a	m	[None]	[None]	n/a	n/a	m	[None]	n/a	[None]	[None]	[None]	[None]	n/a	n/a	n/a	n/a
Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
1	Default	1	KF10A	0	1	1	327	1	10	Green	1	0	1	Dean	5	2.5	17.1	57.1	blackbird	GRASS	SING	SEEN
											2	0	1	Dean	5	2.5	17.1	57.1	thrush	GRASS	FEEDING	SEEN
											3	0	1	Dean	5	2.5	17.1	57.1	thrush	GRASS	FEEDING	SEEN
											4	0	1	Dean	5	2.5	17.1	57.1	blackbird	GRASS	FEEDING	SEEN
											5	0	1	Dean	5	2.5	17.1	57.1	cormorant	VINE	FLYING	SEEN
											6	1.481647627	1	Dean	5	2.5	17.1	57.1	house sparrow	SB	SING	HEARD
											7	6.536680706	1	Dean	5	2.5	17.1	57.1	blackbird	GARDEN	PERCHING	SEEN
											8	6.840402867	1	Dean	5	2.5	17.1	57.1	chaffinch	UNK	SING	HEARD
											9	10	1	Dean	5	2.5	17.1	57.1	greenfinch	VINE	SING	HEARD
											10	10.5	2	Dean	5	2.5	17.1	57.1	house sparrow	SB	SING	HEARD
											11	11.3137085	1	Dean	5	2.5	17.1	57.1	blackbird	GRASS	FEEDING	SEEN
											12	12.80250079	1	Dean	5	2.5	17.1	57.1	goldfinch	VINE	SING	HEARD
											13	12.82575537	1	Dean	5	2.5	17.1	57.1	house sparrow	SB	SING	HEARD
											14	13.52378438	1	Dean	5	2.5	17.1	57.1	greenfinch	SB	SING	HEARD
											15	14.14132741	1	Dean	5	2.5	17.1	57.1	house sparrow	SB	SING	HEARD
											16	15.55634919	1	Dean	5	2.5	17.1	57.1	thrush	GRASS	FEEDING	HEARD
											17	17	1	Dean	5	2.5	17.1	57.1	house sparrow	GRASS	FEEDING	SEEN
											18	17	1	Dean	5	2.5	17.1	57.1	chaffinch	SB	SING	HEARD
											19	17	1	Dean	5	2.5	17.1	57.1	house sparrow	SB	SING	HEARD
											20	17	1	Dean	5	2.5	17.1	57.1	house sparrow	SB	SING	HEARD
											21	17.32050808	1	Dean	5	2.5	17.1	57.1	thrush	VINE	SING	HEARD

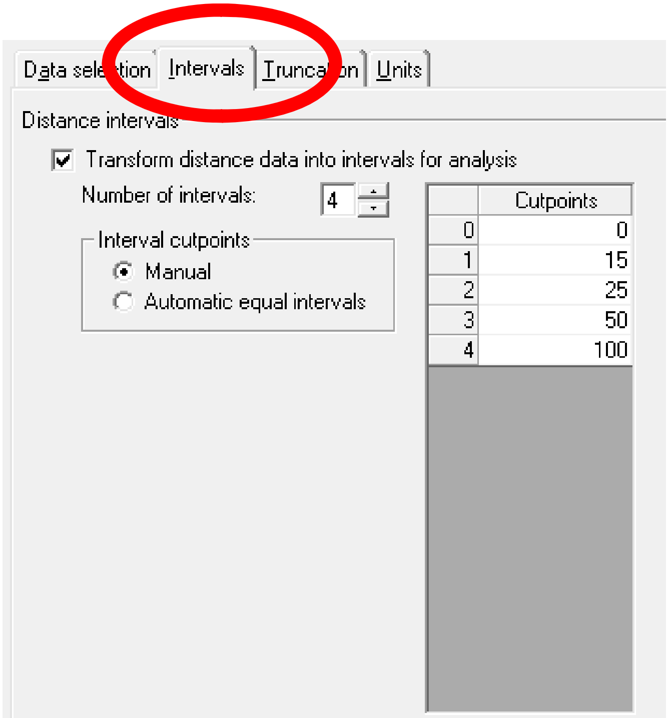
Example analysis:

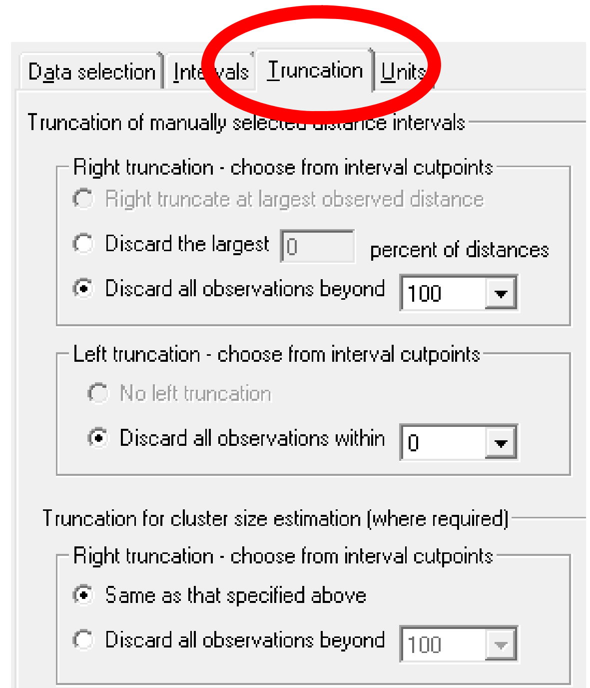
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand



Example analysis:

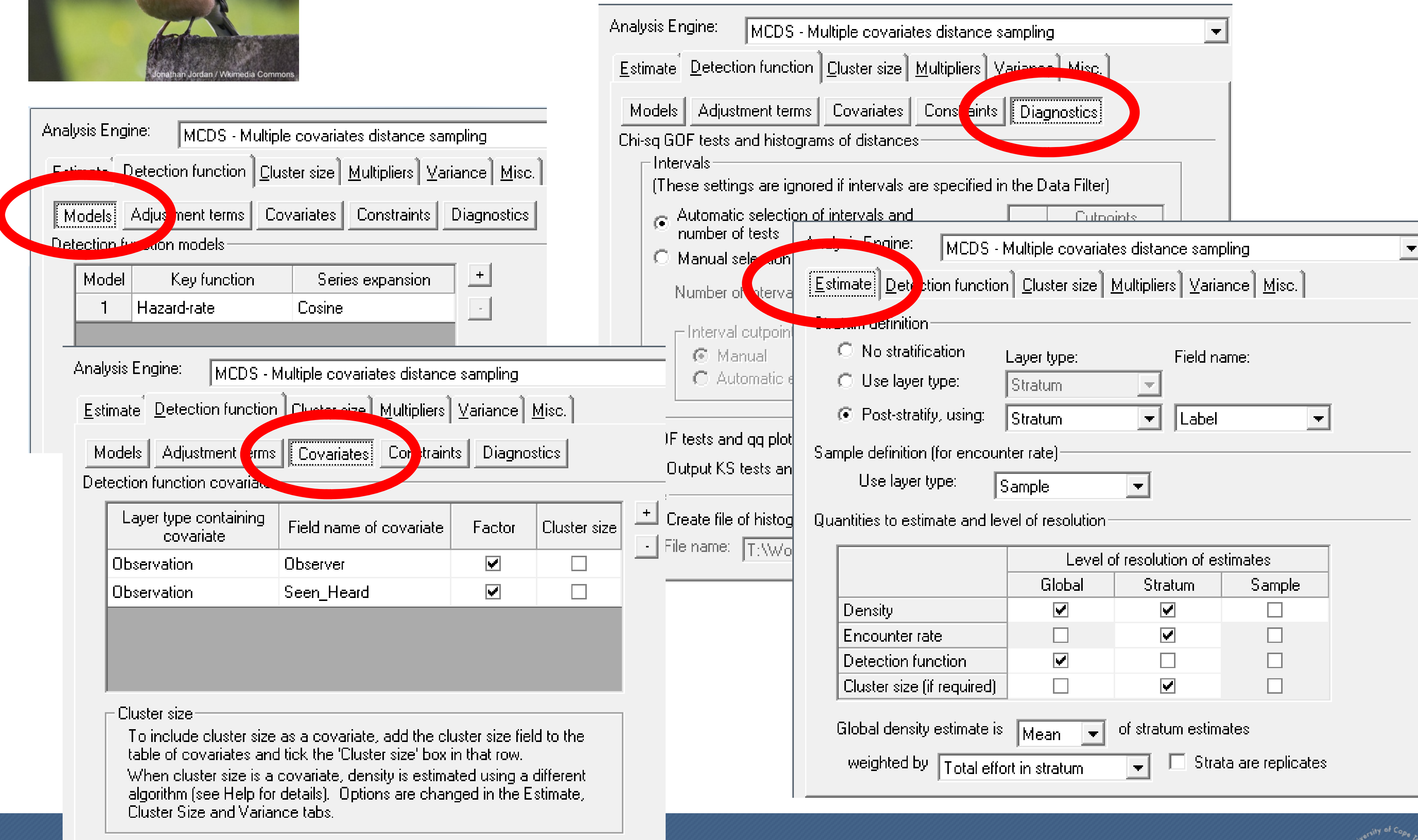
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand





Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

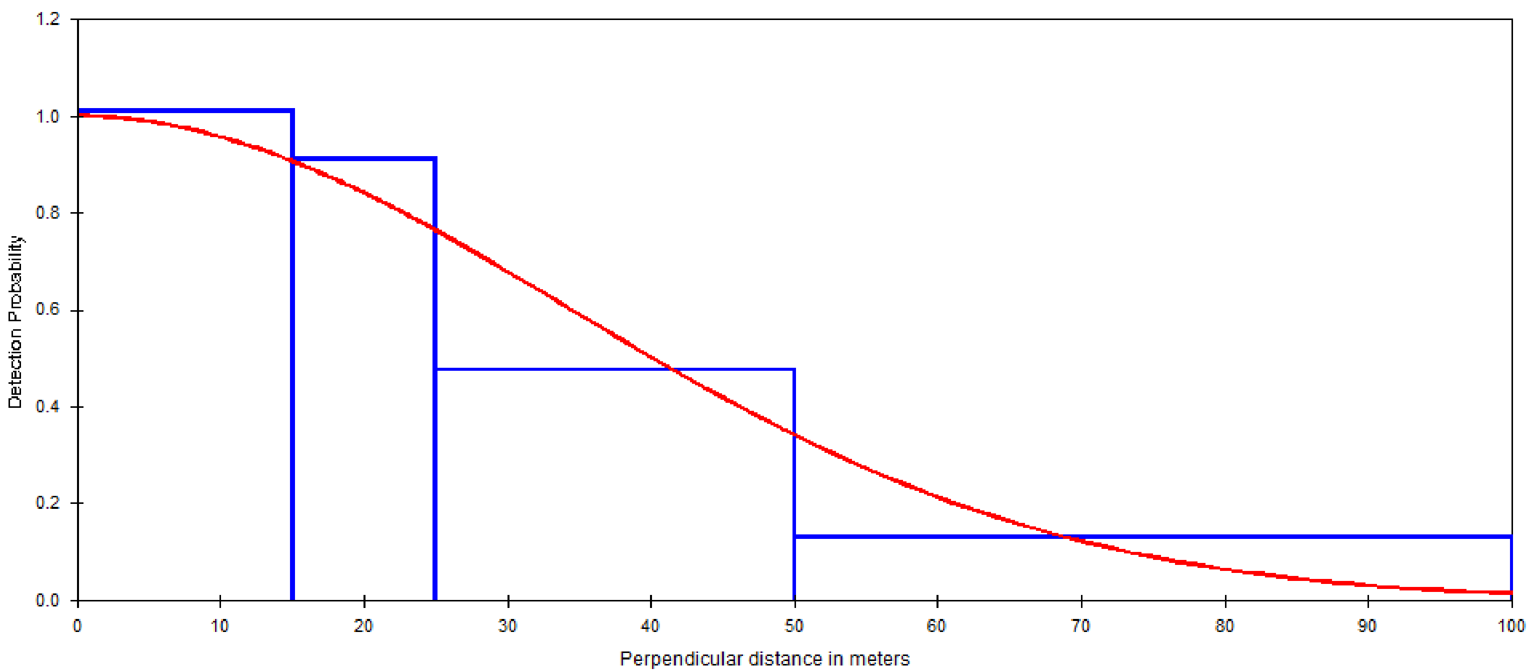


Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Basic model (halfnormal function, no covariates)

- detections are in 4 distance bands



Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Distance - KF bird surveys pooled - [Analysis 1541: [Chaffinch_test] Set: [Chaffinch]]

File View Tools Analysis - Results Window Help

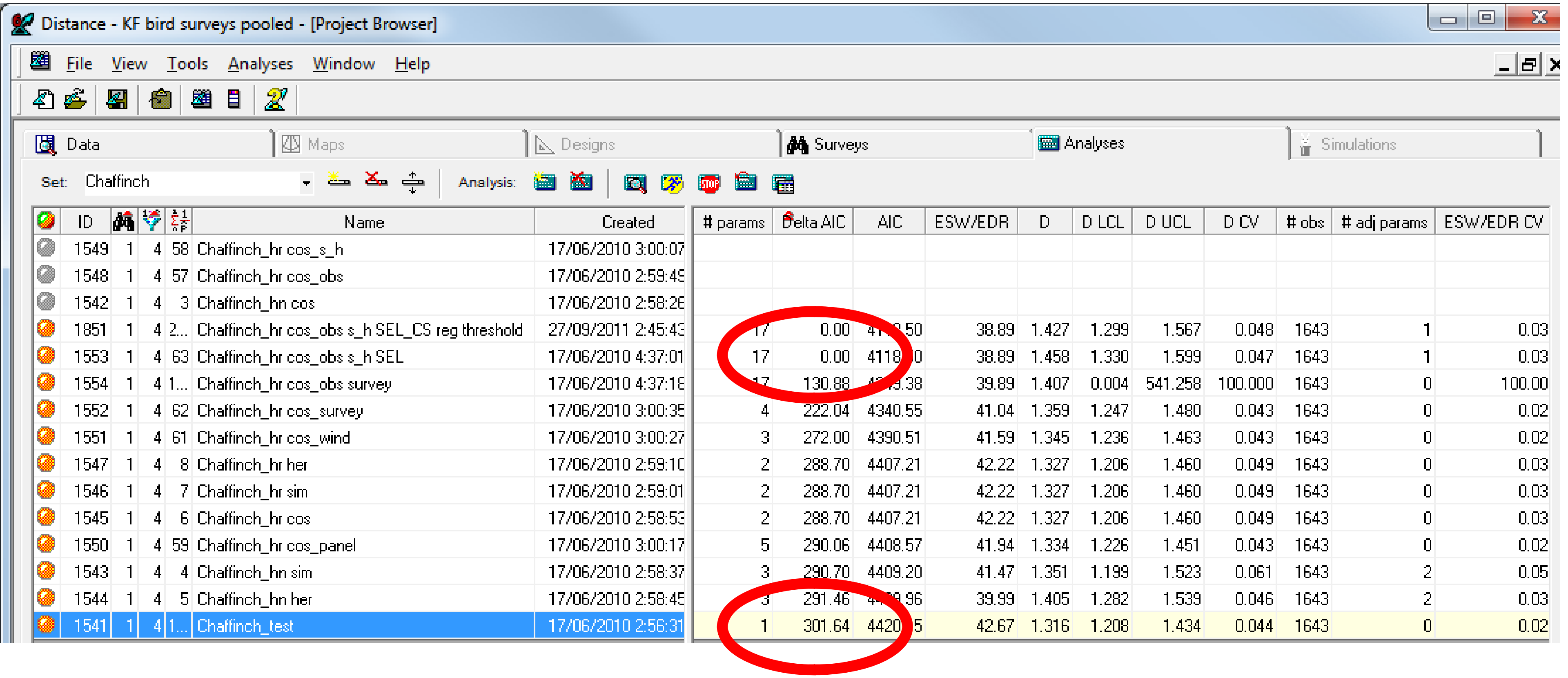
Detection Fct/Global/Model Fitting

```
Effort        :    160666.3
# samples     :   827
Width         :    100.0000
Left          :       0.0000000
# observations:  1643

Model
   Half-normal key, k(y) = Exp(-y**2/(2*A(1)**2))
      A( 1) bounds = ( 1.0000      , 0.10000E+07 )

       Iter   LN(likelihood)     Parameter Values
      -------------------------------------------------------
         1   -2209.90           35.1205
         2   -2209.90           35.1205
         3   -2209.84           35.0849
         4   -2209.58           34.9130
         5   -2209.08           34.2043
         6   -2209.07           34.1632
         7   -2209.07           34.1657
      -------------------------------------------------------
      Results:
      Convergence was achieved with    7 function evaluations.
      Final Ln(likelihood) value =  -2209.0746
      Akaike information criterion =    4420.1494
      Bayesian information criterion =    4425.5537
      AICc =    4420.1519
```

```
                                Estimate      %CV     df     95% Confidence Interval
                               ------------------------------------------------------
Stratum: KF10A
Half-normal/Cosine
                        DS      1.6602       23.32    10.16 0.99532        2.7692
                        D       1.6841       23.64    10.72  1.0064        2.8181
Stratum: KF10B
Half-normal/Cosine
                        DS      1.3024       21.19    27.51 0.84746        2.0016
                        D       1.3418       21.33    28.21 0.87117        2.0666
Stratum: KF10C
Half-normal/Cosine
                        DS      1.4464       35.64    23.15 0.70757        2.9568
                        D       1.5680       35.70    23.31 0.76637        3.2083
Stratum: KF11A
Half-normal/Cosine
                        DS      1.4658       20.58    19.38 0.95756        2.2439
                        D       1.6454       21.22    21.86  1.0646        2.5433
Stratum: KF11B
Half-normal/Cosine
                        DS      1.1422       27.38    13.15 0.63933        2.0405
                        D       1.5448       29.16    16.78 0.84492        2.8244
Stratum: KF11C
Half-normal/Cosine
                        DS      1.5196       18.95    29.69  1.0353        2.2306
                        D       1.5568       18.98    29.86  1.0601        2.2861
Stratum: KF12A
Half-normal/Cosine
                        DS     0.36954       29.60    29.28 0.20432       0.66835
                        D      0.39287       29.81    30.09 0.21652       0.71286
```

Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand



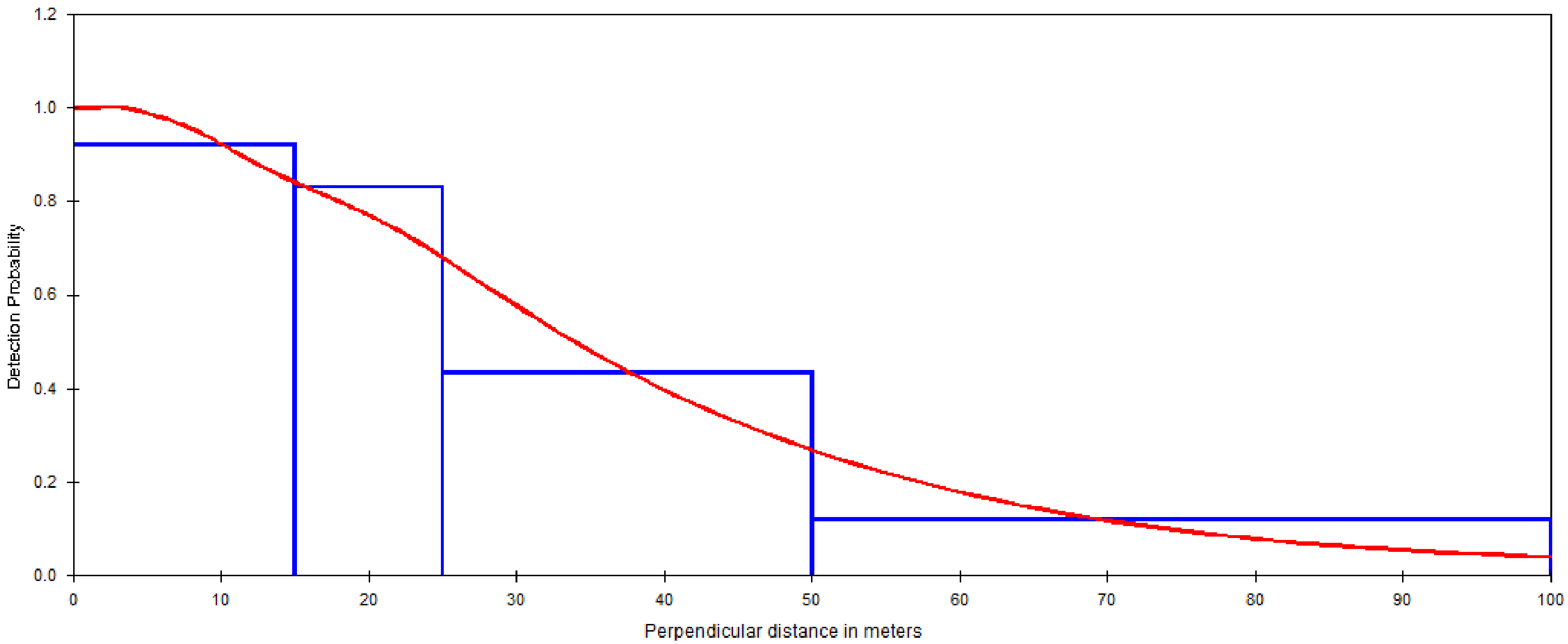
	ID				Name	Created
	1549	1	4	58	Chaffinch_hr cos_s_h	17/06/2010 3:00:07
	1548	1	4	57	Chaffinch_hr cos_obs	17/06/2010 2:59:49
	1542	1	4	3	Chaffinch_hn cos	17/06/2010 2:58:26
	1851	1	4	2...	Chaffinch_hr cos_obs s_h SEL_CS reg threshold	27/09/2011 2:45:43
	1553	1	4	63	Chaffinch_hr cos_obs s_h SEL	17/06/2010 4:37:01
	1554	1	4	1...	Chaffinch_hr cos_obs survey	17/06/2010 4:37:18
	1552	1	4	62	Chaffinch_hr cos_survey	17/06/2010 3:00:35
	1551	1	4	61	Chaffinch_hr cos_wind	17/06/2010 3:00:27
	1547	1	4	8	Chaffinch_hr her	17/06/2010 2:59:10
	1546	1	4	7	Chaffinch_hr sim	17/06/2010 2:59:01
	1545	1	4	6	Chaffinch_hr cos	17/06/2010 2:58:53
	1550	1	4	59	Chaffinch_hr cos_panel	17/06/2010 3:00:17
	1543	1	4	4	Chaffinch_hn sim	17/06/2010 2:58:37
	1544	1	4	5	Chaffinch_hn her	17/06/2010 2:58:45
	1541	1	4	1...	Chaffinch_test	17/06/2010 2:56:31

# params	Delta AIC	AIC	ESW/EDR	D	D LCL	D UCL	D CV	# obs	# adj params	ESW/EDR CV
17	0.00	[illegible]	38.89	1.427	1.299	1.567	0.048	1643	1	0.03
17	0.00	[illegible]	38.89	1.458	1.330	1.599	0.047	1643	1	0.03
17	130.88	[illegible]	39.89	1.407	0.004	541.258	100.000	1643	0	100.00
4	222.04	4340.55	41.04	1.359	1.247	1.480	0.043	1643	0	0.02
3	272.00	4390.51	41.59	1.345	1.236	1.463	0.043	1643	0	0.02
2	288.70	4407.21	42.22	1.327	1.206	1.460	0.049	1643	0	0.03
2	288.70	4407.21	42.22	1.327	1.206	1.460	0.049	1643	0	0.03
2	288.70	4407.21	42.22	1.327	1.206	1.460	0.049	1643	0	0.03
5	290.06	4408.57	41.94	1.334	1.226	1.451	0.043	1643	0	0.02
3	290.70	4409.20	41.47	1.351	1.199	1.523	0.061	1643	2	0.05
3	291.46	[illegible]	39.99	1.405	1.282	1.539	0.046	1643	2	0.03
1	301.64	[illegible]	42.67	1.316	1.208	1.434	0.044	1643	0	0.02

Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

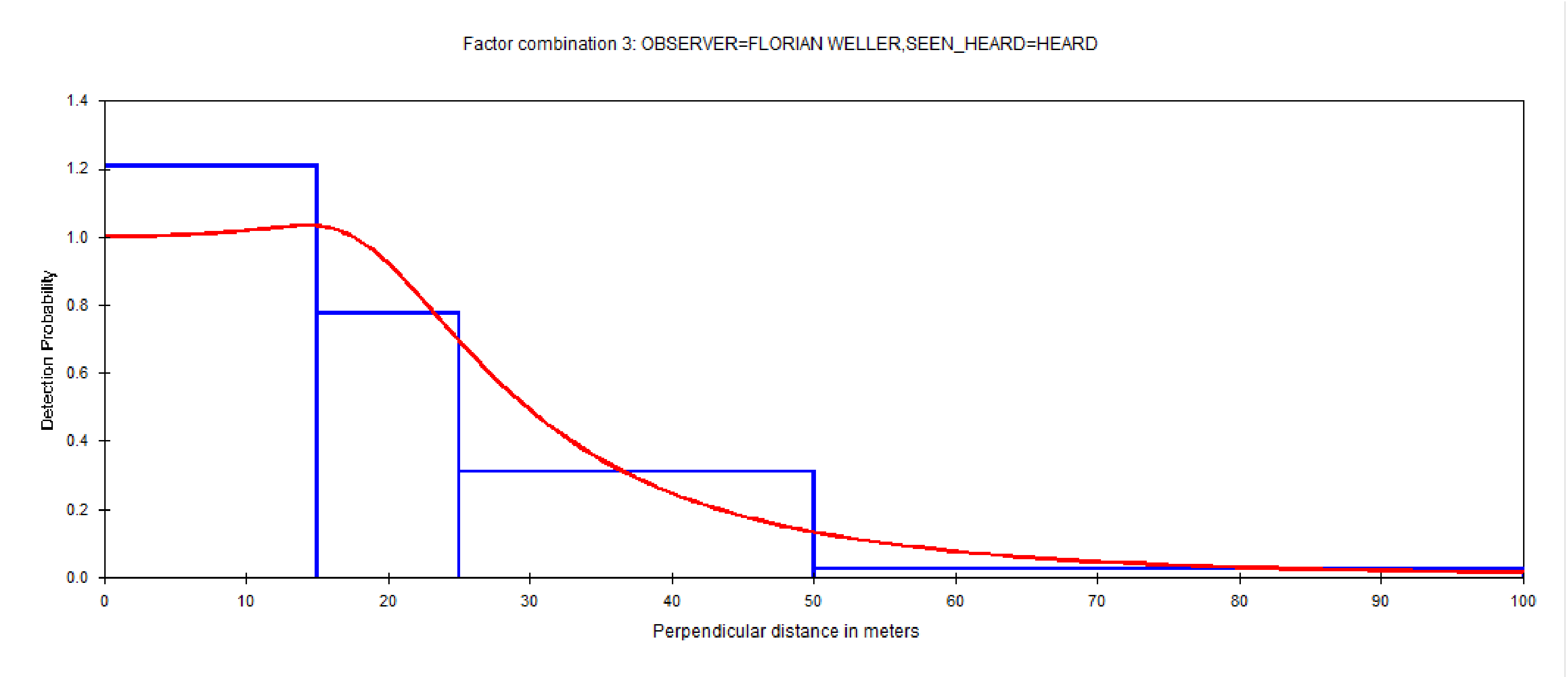
Hazard rate function, covariate: observer *(average of covariate levels)*



Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

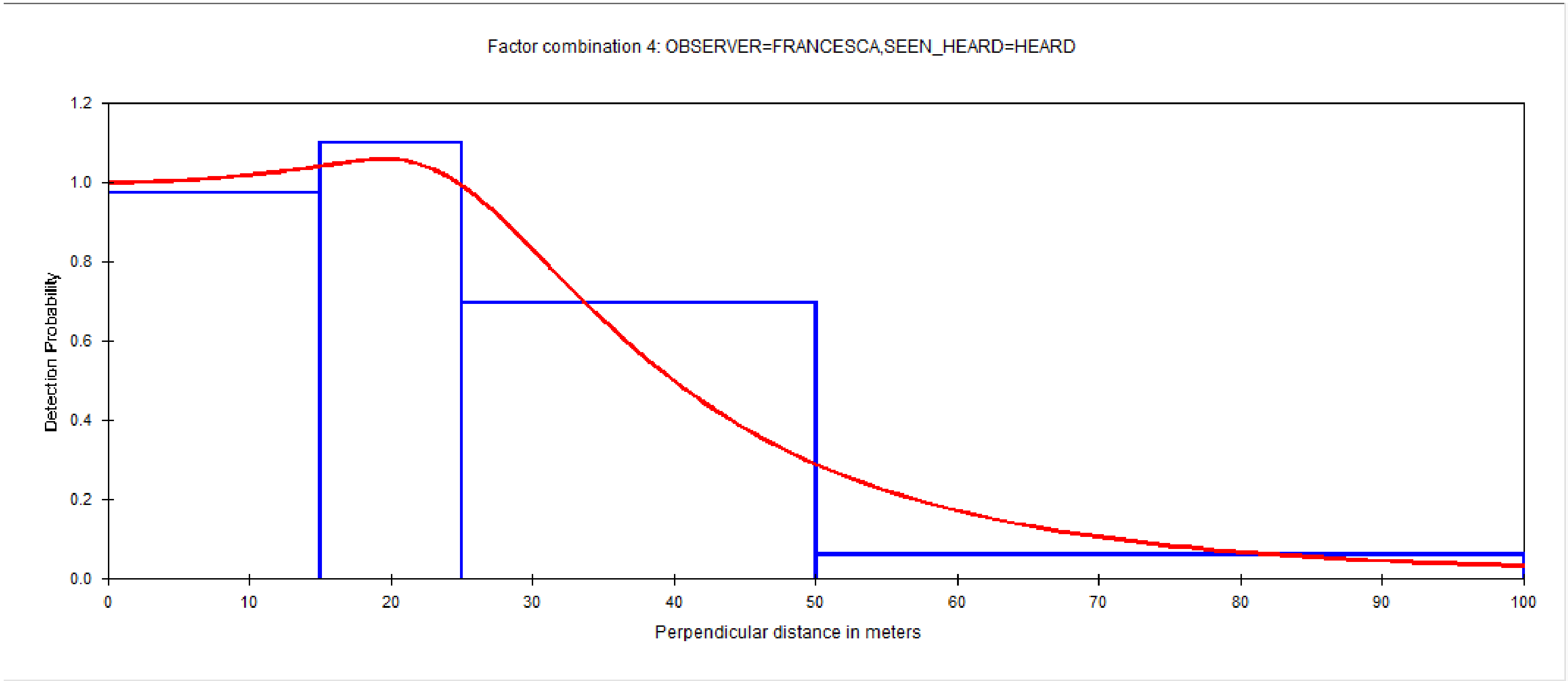
Hazard rate function, covariate: observer *(covariate level 1)*



Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

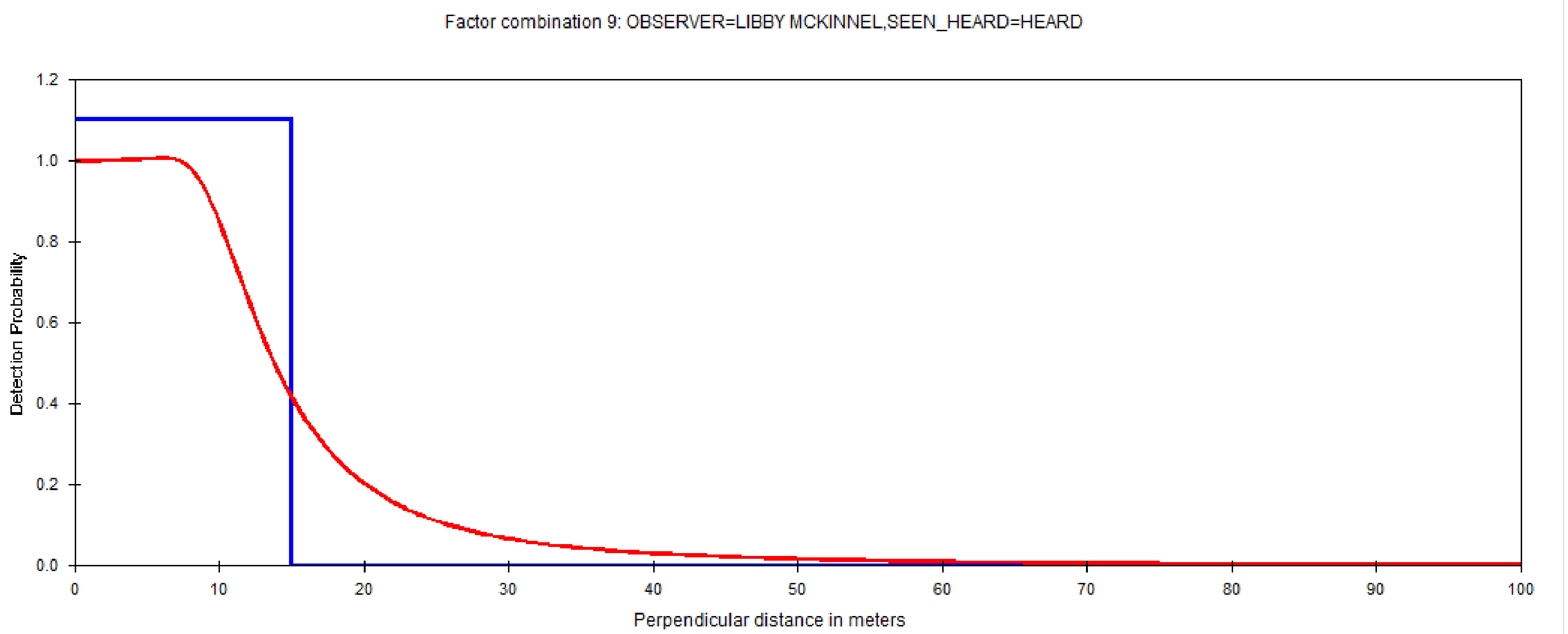
Hazard rate function, covariate: observer *(covariate level 2)*



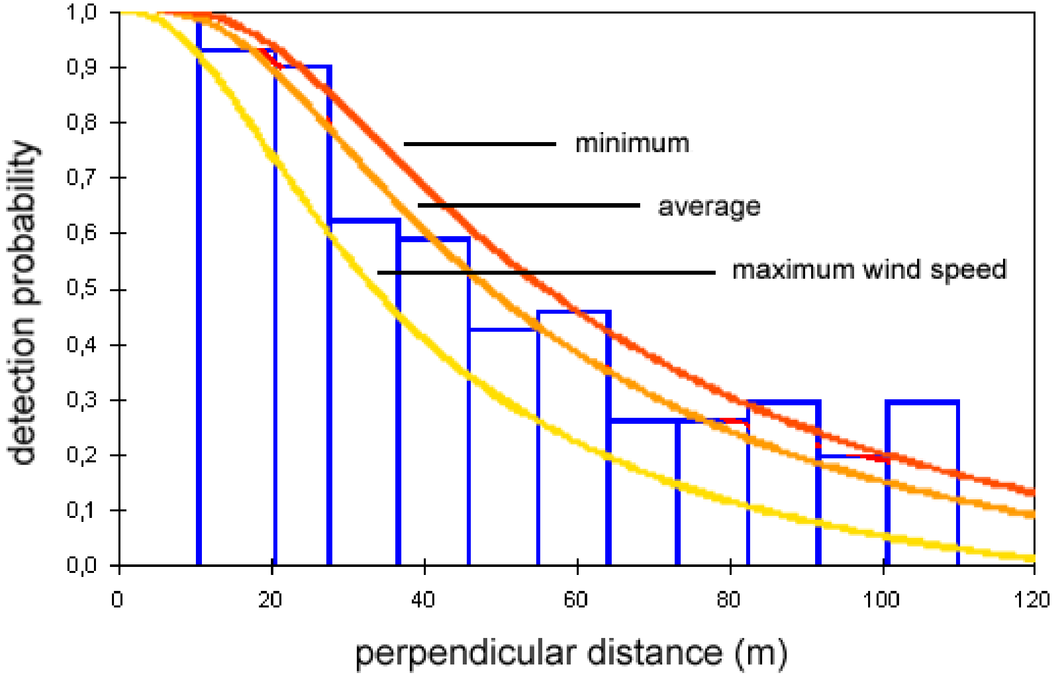
Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Hazard rate function, covariate: observer *(covariate level 3)*



Covariates may be factorial (e.g. observer) or continuous (e.g. wind speed)



Covariates may be factorial (e.g. observer) or continuous (e.g. wind speed)

<u>Keep in mind</u>:
all covariates that enter into this model fitting process are covariates of *detectability*, not *abundance*. Here we are only concerned with effects on the detection process, to derive a largely unbiased survey result.

Many covariates can of course affect *both* detectability and abundance (e.g. site identity, vegetation type, time of year,…). However, modelling their effects on density/abundance itself is a separate (later) job, and not part of DS proper.

(*unmarked* combines these steps sequentially, but they are still different processes)

perpendicular distance (m)

Another example:
Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

While the chaffinch data was collected in specific **distance bands**, the skylark detection distances (collected with a range finder) were analysed as **exact measurements**.

Pro:

- biases in detection process are easier to identify in high resolution data
- better model fit can be achieved

Con:

- requires a large number of detections to be feasible
- prone to overfitting
- you will likely end up partitioning the data into intervals anyway, to deal with biases (see below)

→ **some partitioning** into intervals / detection bands is generally a good idea; if not in the field then during the analysis stage

Allowing the detection function to fit more closely to the data

Detection function	Form
Uniform	$1/w$
Half-normal	$\exp(-y^2/2\sigma^2)$
Hazard-rate	$1 - \exp(-(y/\sigma)^{-b})$
Negative exponential	$\exp(-ay)$

where y is distance and w is truncation distance

Extended form: detection function = key function **+ series expansion**

$g(y)$ = key (y) [1 + series (y)]

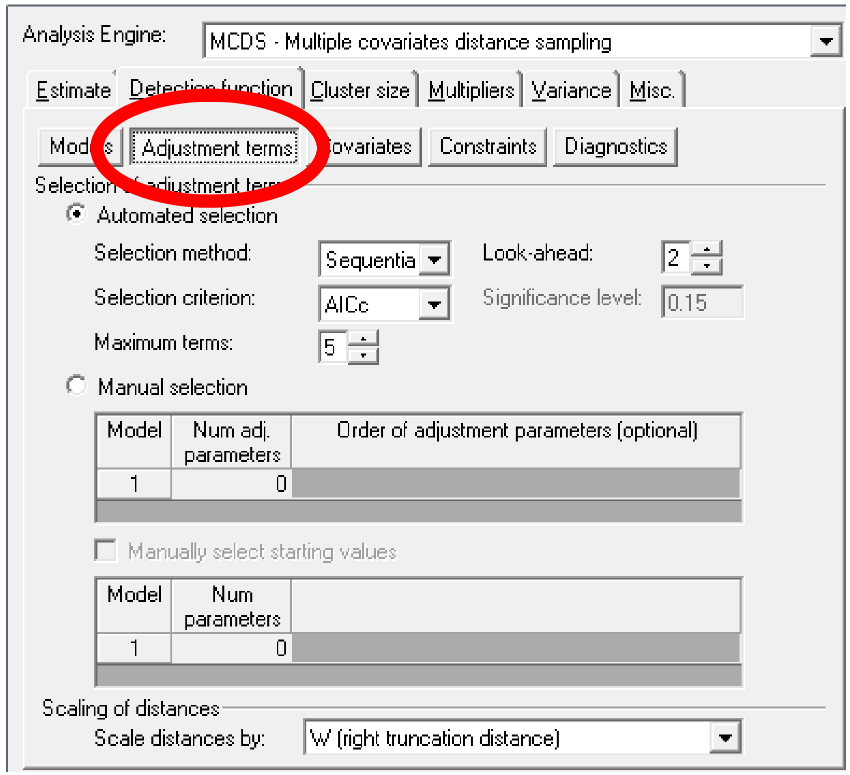
Series expansions add **flexibility** to the shape of the key function

Cosine $$\sum_{j=1}^{m} \alpha_j \cos(j\pi y / w)$$

Simple polynomial $$\sum_{j=1}^{m} \alpha_j (y / w)^{2j}$$

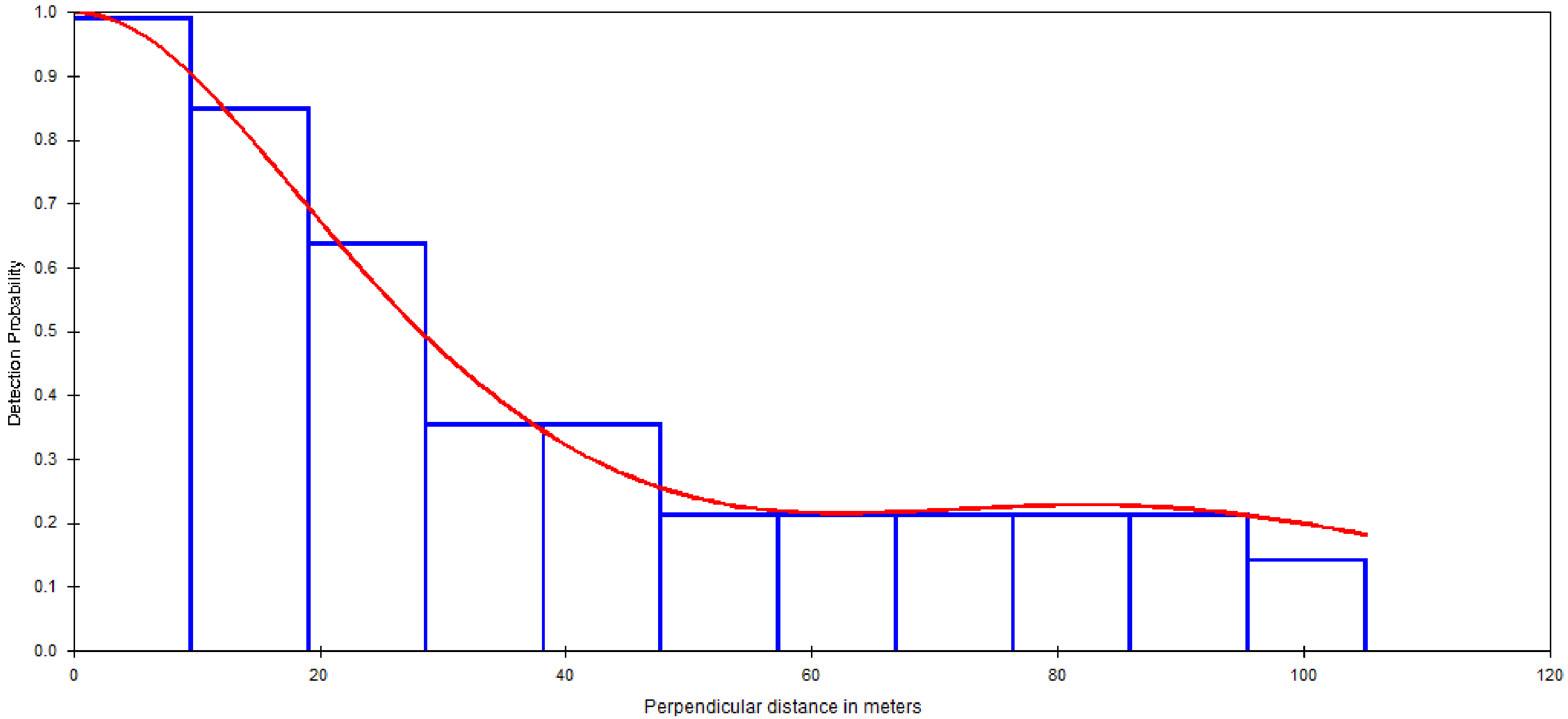
Hermite polynomial $$\sum_{j=2}^{m} \alpha_j H_{2j}(y / \sigma)$$

Allowing the detection function to fit more closely to the data



Song thrush data

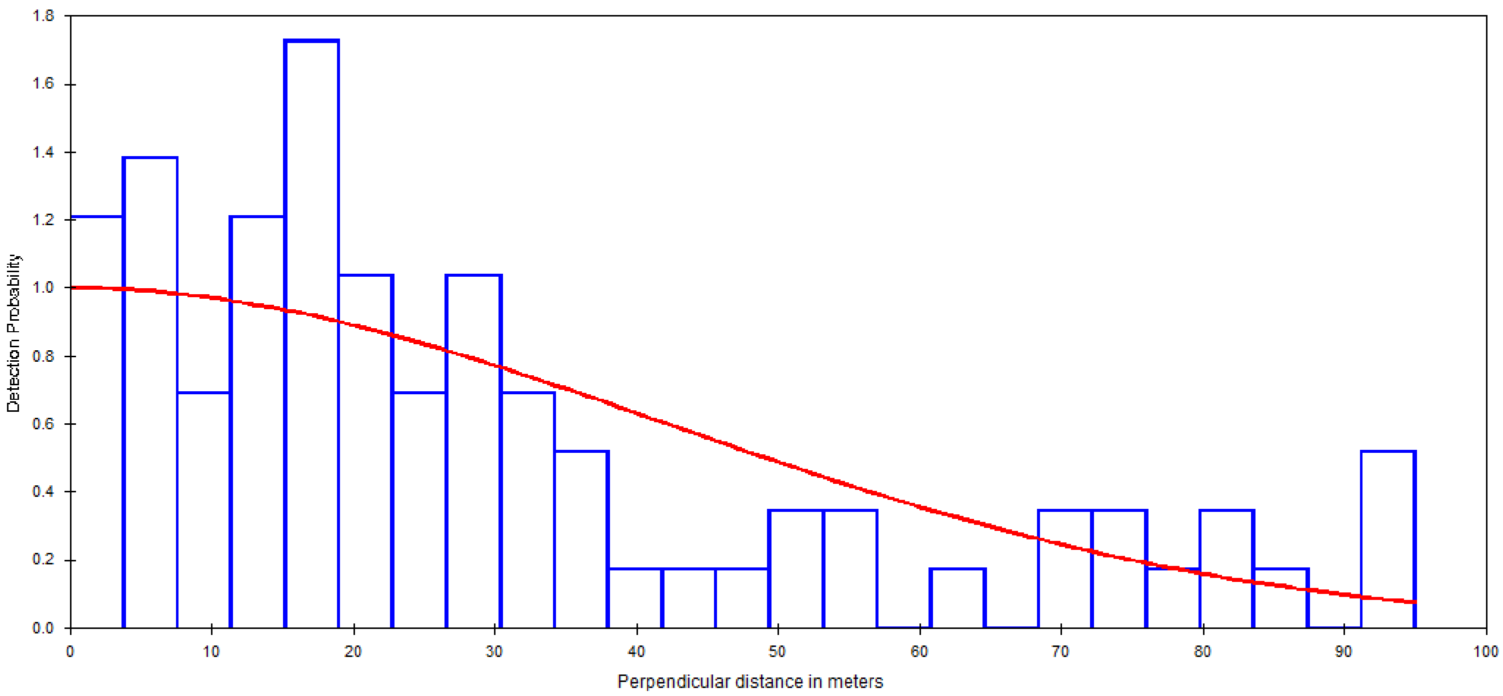
Basic model (halfnormal function, no covariates) **+ 2 series expansion terms**



Another example:

Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

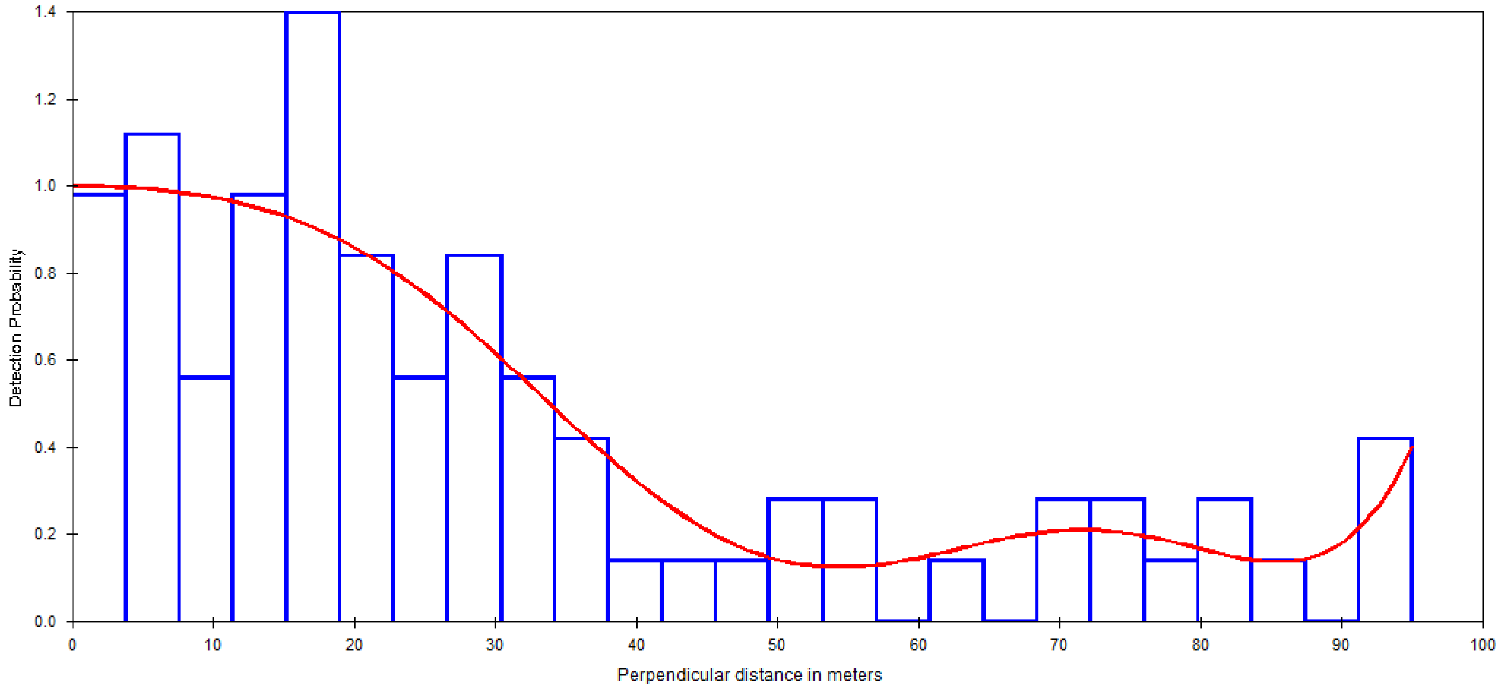
Basic model (halfnormal function, no covariates)



Another example:

Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

Basic model (halfnormal function, no covariates) **+ 5 series expansion terms -> overfitting**

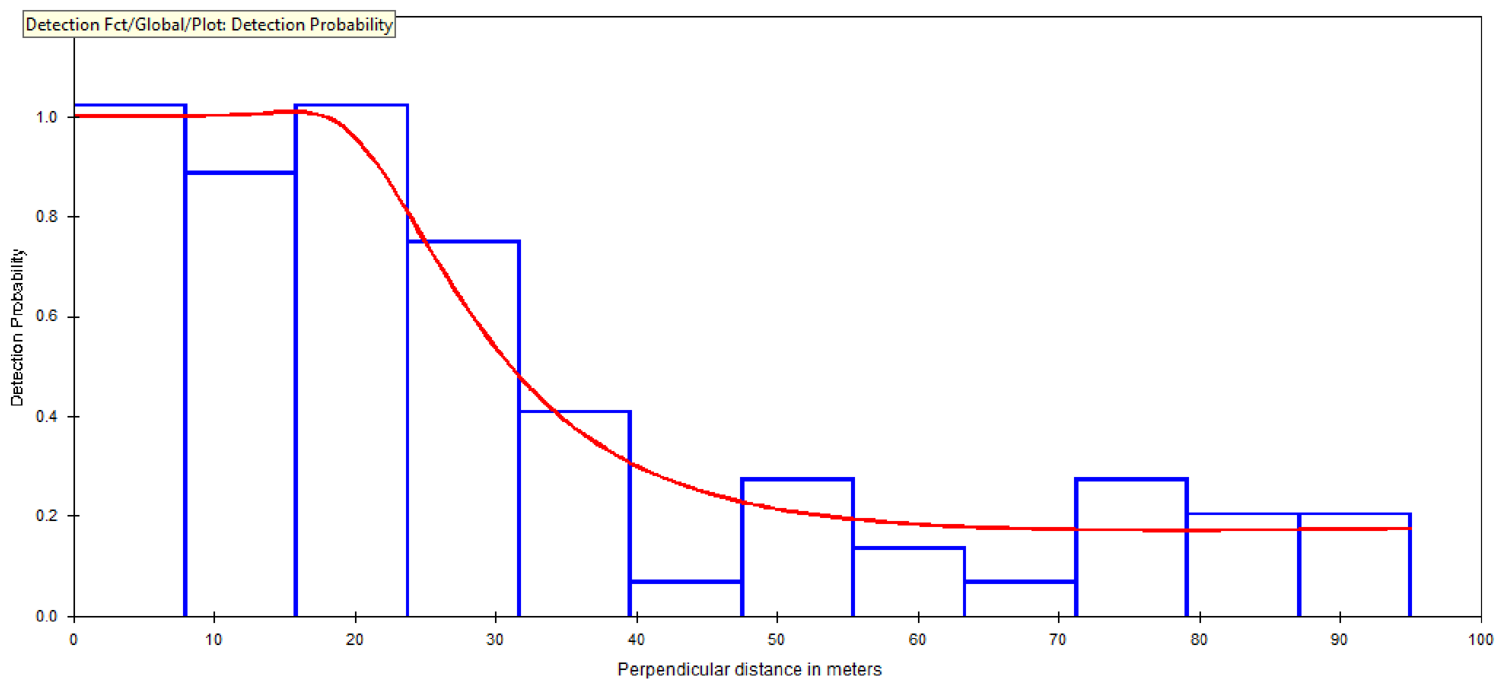


Another example:

Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

Hazard rate function, no covariates

- **partitioned into 8 m intervals**



- **Basic assumptions of distance sampling**
- **General considerations about model choice**

Assumption #1:

All animals on the line are detected (i.e., detection probability at distance 0 is 1)

- depends on study subject; often an issue in shipboard surveys because ship may block field of view, and/or animals may dive
- in terrestrial surveys, can generally be assumed
- animals may **move away** from the line before observer passes that point; however, this does not result in a non-detection but in **movement before detection** (see below)

dealing with violations:

- this is tricky, because g(0) = 1 is a fundamental assumption of the fitting process
- can use a 2nd observer to “guard the centerline” (may result in **over**estimation)
- there are procedural fixes but they aren’t pretty
- try **not** to violate this one

Assumption #2:

Animals are randomly and independently distributed

- three sources of bias:
  a) populations are **clustered** (flocks etc.) but individual detections are treated as independent
     - artificially shrinks the confidence interval of the detection estimate
  b) transects are not placed independently of **gradients of density** (e.g. roads)
     - can lead to strongly biased detection at specific distances
  c) transects are **too close together**

dealing with violations:

a) record not individuals but **clusters + cluster size**, then incorporate cluster size into the detection funtion
b) place transects either **randomly**, or **across gradients of density**
c) make sure that **maximum detection ranges** do not overlap between transect

Recording clusters instead of individuals

Analysis Engine: MCDS - Multiple covariates distance sampling

Estimate | Detection function | Cluster size | Multipliers | Variance | Misc.

Cluster size estimation method

(These options are ignored unless objects are clusters.)

- (•) Use size-bias regression method
- ( ) Use mean of observed clusters
- ( ) Use size bias regression method if regression significant at an alpha-level of [0.15] ; use mean of observers if not significant

Size-bias regression method

- (•) Regress ln(cluster size) against estimated g(x)
- ( ) Regress cluster size against estimated g(x)
- ( ) Regress ln(cluster size) against distance x
- ( ) Regress cluster size against distance x

```
Effort          :    34965
# samples       :    71
Width           :    80.00000
# observations:    157

Model
   Half-normal key, k(y) = Exp(-y**2/(2*A(1)**2))
      A( 1) bounds = (0.80000      , 0.10000E+07 )

       Iter    LN(likelihood)     Parameter Values
      -------------------------------------------------------------------
         1    -649.744            30.6225
         2    -649.744            30.6225
         3    -649.635            30.9778
         4    -649.496            31.9478
         5    -649.494            32.0888
         6    -649.494            32.1064
         7    -649.494            32.1067
      -------------------------------------------------------------------
      Results:
      Convergence was achieved with    7 function evaluations.
      Final Ln(likelihood) value =  -649.49381
      Akaike information criterion =   1300.9877
      Bayesian information criterion =   1304.0438
      AICc =   1301.0134
```

```
                        Estimate      %CV     df      95% Confidence Interval
                                 ------------------------------------------------
 Average cluster size
                        1.1911        4.62   156.00   1.0872        1.3049
 Half-normal/Cosine
                r       [illegible]E-01
                r-p     0.13957
                E(S)    1.1365        2.55   155.00   1.0808        1.1952
```

Assumption #3:

Animals do not move before detection

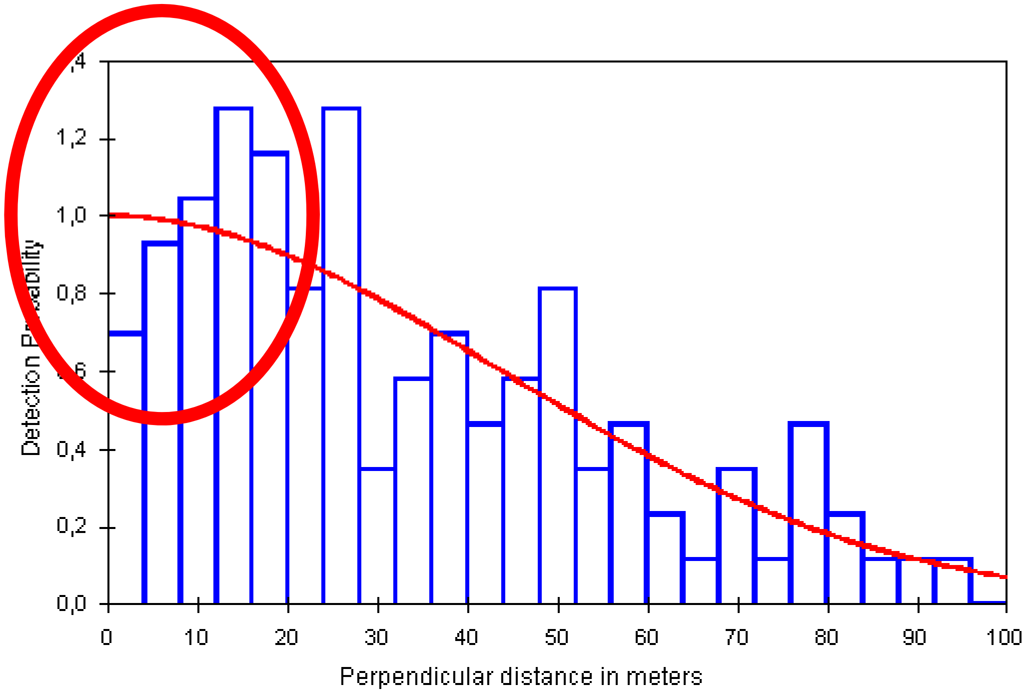
- bias resulting from movement is **negligible if movement is random**
- movement in response to the observer can result in negative (**avoiding** observer) or positive (**attracted** to observer) bias in detectability
- particularly avoidance behaviour is quite common

dealing with violations:

- average out & bridge dips/spikes in the distribution by
  - partitioning data into **intervals**
  - using **models with a shoulder** (e.g. hazard-rate)

Skylark data

Indications of **observer avoidance** in the data



Assumption #4:

Measurements are exact (angles and distances)

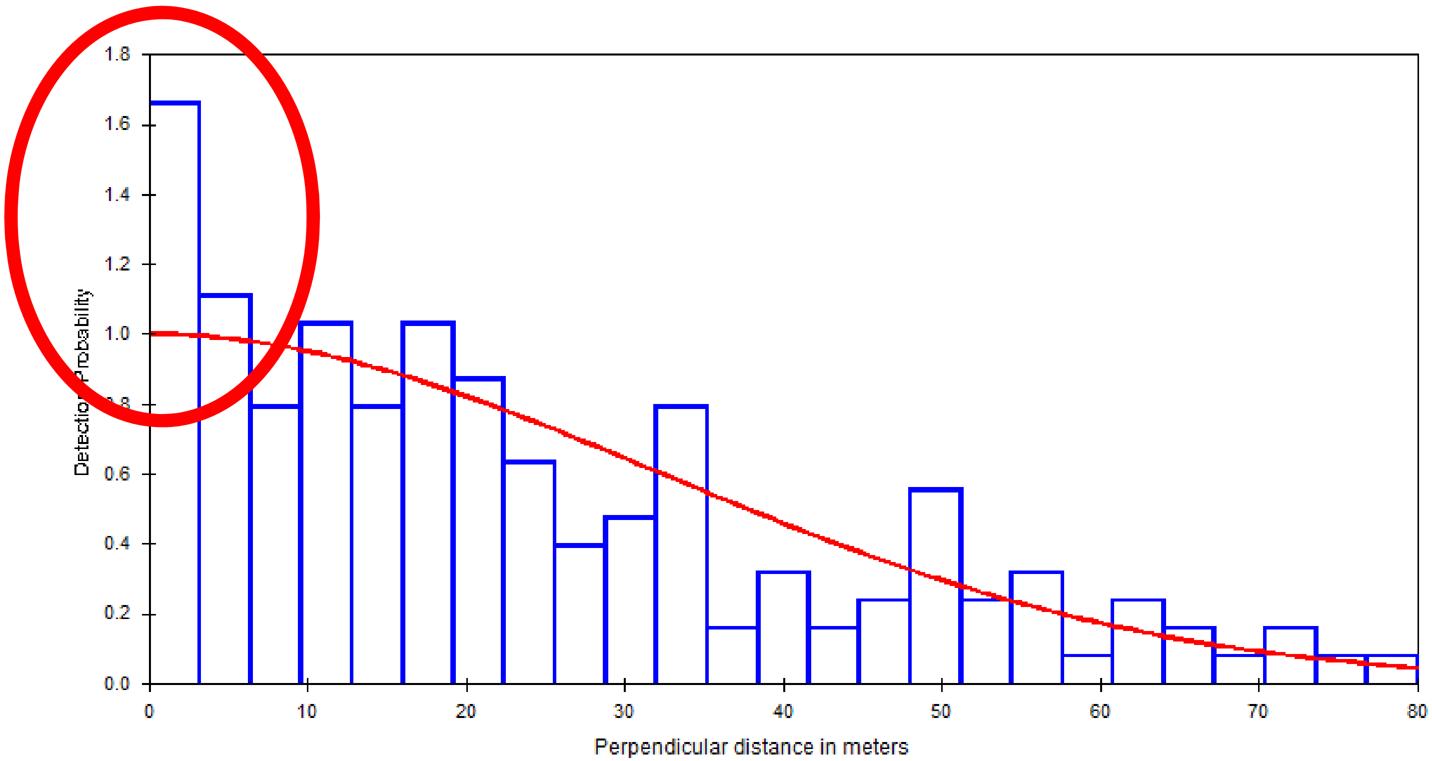
- the effect of **random** errors is negligible, but **systematic** errors can introduce bias; this often happens with **rounding**
- rounding angles or distances to preferred values can result in **heaping**
- **angle rounding to zero** is particularly common

dealing with violations:

- **avoid dead reckoning** in the field by using **tools**: range finders, angle boards, etc.
- smoothen & average instances of rounding in the data by **partitioning into intervals**

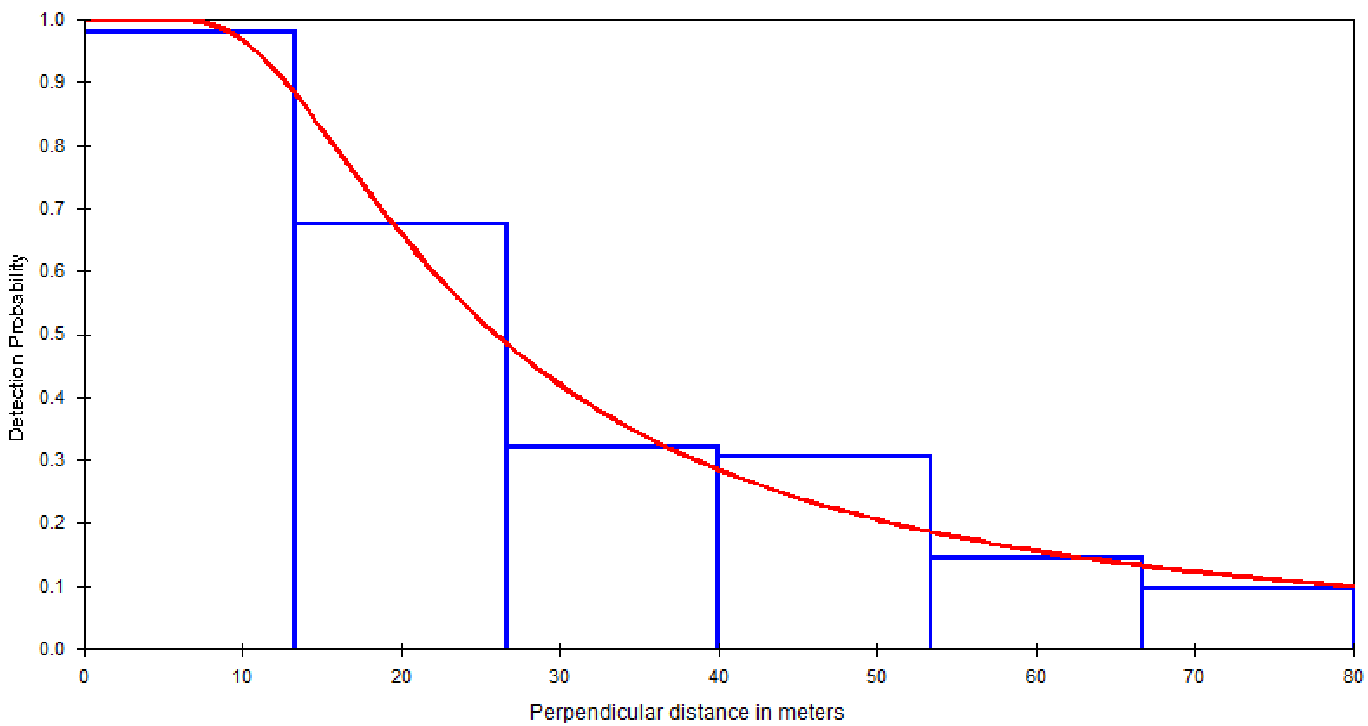
Blackbird data

Indications of angle **rounding to zero** in the data



Blackbird data

Hazard rate (shoulder), partitioning into intervals



When to use distance sampling?

available number of samples / accuracy of abundance estimate

Census
(complete count)

Double count
(rapid large-scale counts, calibrated by a few small-scale censuses)

Distance sampling
(modelled detectability)

- \+ large improvement in accuracy by modelling detectability
- \+ much lower effort than census
- – requires reasonably high number of records (>60)

Index count
(count not corrected for detectability, may be sufficient for tracking relative changes over time)

Presence/absence
(e.g. Occupancy)

When to use ***multiple covariate*** distance sampling? (as opposed to basic DS)

-> this is a question of **level of detection function** vs **level of abundance/density estimate**

Assume e.g. that one species was surveyed at several sites, and you want an estimate for each site. There are three basic options:

a) Fit a **separate** detection function to each site dataset

→ most accurate outcome, best adaptation to different site parameters; BUT requires sufficient sample sizes at **each** site

b) Fit a **global** detection function using all pooled detection, then use this function on **site datasets** to get site-level estimates

→ may save your bacon if sample sizes at individual sites don't support fitting a function; BUT assumes that detection does **not vary** (substantially) between sites

c) Use **multiple covariate distance sampling**: fit a global detection function, specifying site as a **covariate** (i.e., allow the global function to vary for data subsets defined by covariate levels)

→ this is halfway between the other two options in requirements and result accuracy

The same choice applies to any covariate (canopy cover, season, sex, etc.)

→ this is essentially a **model selection question**; if the choice isn't clear, AIC can be used to help choose between approaches (as all three use the same data set)

Useful resources

- The Distance project website http://distancesampling.org/
    - software downloads (R packages & program Distance)
    - Distance mailing list / Google group
    - huge distance sampling article bibliography

- Books (by Steve Buckland *et al*.):
    - Introduction to Distance Sampling: Estimating Abundance of Biological Populations. (2001) Oxford University Press. *(in Main Library)*
    - Advanced Distance Sampling. (2004) Oxford University Press.
    - Distance Sampling: Methods and Applications. (2015) Springer.

- Program Distance documentation (very useful)

- Method reference: Thomas, L., S.T. Buckland, E.A. Rexstad, J. L. Laake, S. Strindberg, S. L. Hedley, J. R.B. Bishop, T. A. Marques, and K. P. Burnham. 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. Journal of Applied Ecology 47: 5-14. DOI: 10.1111/j.1365-2664.2009.01737.x