Model Construction and Evaluation of Line Transect Distance Data for Density Estimation

REBECCA S. GALELA
BRIGIDA A. ROSCOM

Abstract

Line transect distance sampling for estimating density was simulated in two projects using match sticks as objects of observation. Possible detection functions were constructed and modeled. The selected model in each case was used to estimate the density and compared with the true density. For line transects distance sampling on match sticks $MSDP_1$, the best model selected was the uniform key function plus one cosine series. It yielded a density estimate of 8.352 matchsticks per square meter over the true density of 8.4. For line transects distance sampling on match sticks $MSDP_2$, the best model selected to estimate the density was the hazard rate key. It gave an estimate of 8.12 matchsticks per square meter over the true density of 8.4. In both cases selection of the chosen model was based on lowest Akaike's Information Criterion (AIC).

The experience from the simulations were used to estimate the density of coconut trees in a 24 hectare farmland of Liloy, Zamboanga del Norte and the density of mangrove trees in Langamon, Anibongan, Gigaquit, Surigao del Norte. The best model selected for the Coconut Distance Project (CDP) was the uniform key function+ one simple or Hermite polynomial, which gave a density estimate of 28.98 coconuts per hectare over the true density of 29.42. The hazard function without adjustment terms was the best model selected for the Mangrove Distance Project (MDP) . It yielded an estimate of 116.52 mangroves per hectare.

REBECCA GALELA is a faculty member of the Department of Professional Education, College of Education, MSU-Iligan Institute of Technology. She obtained her MS-Statistics at the Graduate School of MSU-Iligan Institute of Technology. BRIGIDA A. ROSCOM, Ph.D., is a full professor of MSU-Iligan Institute of Technology and currently the Dean of the Graduate School of MSU-IIT.

Results of the study, using the software package DISTANCE 3.5, showed the goodness of the estimate by distance sampling compared to the conventional strip transect full census methodology. For future research directions, a similar study might be conducted where objects of interest are mobile or clustered.

Keywords: Model Construction, Line Transect, Density Estimation

Introduction

Methods of estimating population abundance have been developed and the most recent is the distance method, which is based on object distances detected from points or lines. In distance sampling one traverses a randomly chosen path called line transects and measures the perpendicular distance from the path to the object detected. Estimates of density and abundance proved to be efficient even if some of the objects went undetected.

In the study a comprehensive computer software package called DISTANCE, version 3.5, was utilized. This software facilitated all the computations and plotting needed in the analysis.

At the heart of this analysis is a detection function, which has to be constructed and modeled. This function is produced in two stages. The first stage is a key function which captures the gross of the detection function and can usually be determined from visual inspection of a histogram of detection distances. The second stage is the addition of adjustment terms or expansion series of one or two parameters to the key function. So, linear combinations of key functions and adjustment terms are constructed and they are called models for the detection function. Among these models, the best model is selected based on some criteria. Using this best combination, statisticians and biologists or environmentalists in the field are able to estimate the density or abundance of a certain population with accuracy.

Objective

The study aims at illustrating the principles and methodology of distance sampling and the construction of the detection function. It also aims to compare line transect distance sampling to square plot sampling and strip transect sampling approaches in terms of density estimates and actual density with simulated data using match sticks here referred to as Match Stick Distance Project$_1$ ($MSDP_1$). The study also wants to illustrate the modeling process, model evaluation and finally density estimation using the simulated data on matchsticks here referred to as Match Stick Distance Project$_2$ ($MSDP_2$) and the primary data from Coconut Distance Project (CDP) and Mangrove Distance Project (MDP) collected in Zamboanga del Norte and Surigao del Norte, respectively.

Basic Concepts

The Detection Function

Central to the concept of line transect distance sampling is the detection function g (y), the probability of detecting an object, given that it is at distance y from the random line, or

$$g(y) = \text{prob}\{\text{detection} / \text{distance } y\}.$$

The distance y refers to the perpendicular distance from the centerline to the object of interest. The area occupied by the population of interest, the number of lines surveyed, the length of each transect line and the width of the area searched on each side of the transect line are known. The number of objects per unit area (D), and the population size (N), are the unknown parameters.

The Theory Underlying Line Transect Sampling

In strip transect sampling, if strips of width 2w and total length L are surveyed, an area of size $a = 2wL$ is censused. All n objects within the strips are enumerated, and estimated density is the expected number of objects per unit area:

$$\check{D} = n / 2wL$$

In line transect under distance sampling, however, density estimation uses probability detection function as outlined below.

Let D be the number of objects per unit area;
a be the area surveyed within distance w of the line;
x be a measured vertical distance from the transect line to the detected object;
P_a be the probability of detection for an object within this area unconditional on its position;
E(n) be the expected number of objects detected within distance w,

$E(n) = D \cdot a \cdot P_a;$

$D = E(n) / a \cdot P_a;$

$\check{D} = n / a \cdot P_a$; and,

$\check{D} = n / 2wLP_a$

P_a is conceptually to be derived using the distance data obtained. Suppose w is finite and let

f(x) = probability density function of perpendicular distances x and
g(x) = probability of detection given distance x.

Then,

$$f(x)dx = \text{pr}\{\text{object is in } (x,x+dx) \text{ / objects is detected}\}$$

$$= \frac{\text{pr}\{\text{objects is in } (x,x+dx) \text{ and object is detected}\}}{\text{pr}\{\text{objects is detected}\}}$$

$$= \frac{\text{pr}\{\text{object is detected / object is in } (x,x+dx)\} \cdot \text{pr}\{\text{object is in } (x, x+dx)\}}{P_a}$$

$$f(x)dx = \frac{g(x) \cdot (dx \cdot L) / w \cdot L)}{P_a}$$

Hence,

$$f(x) = \frac{g(x)}{w \cdot P_a}.$$

Let $\mu = w \cdot P_a$ so that $f(x) = g(x) /$

By integration and f, a pdf,

$1 = \int_0^w f(x)dx = 1 / \mu \int_0^w g(x)dx$ or

$\mu = \int_0^w g(x)dx$.

Hence, P_a , the unconditional probability of detecting an object within w distance of the transect line, is μ area under the curve of the detection function divided by 1.0 w or

$$f(x) = g(x) / wP_a$$
$$f(x) = g(x) / \int_0^w g(x)dx .$$

From the assumption g(0) = 1, the probability density function, evaluated at zero distance, is

$$f(0) = g(0) / \int_0^w g(x)dx$$
$$f(0) = 1 / \int_0^w g(x)dx .$$
$$f(0) = 1 / \mu$$

The parameter $\mu = \int_0^w g(x)dx$ is a function of the measured distances. Therefore, we will often write the general estimator of density for line transect distance sampling simply as

$$\check{D} = n \cdot f(0) / 2L$$
$$\check{D} = n / 2L\mu$$

$\check{D}$ is valid whether w is bounded or unbounded (infinite) and when the data are grouped or ungrouped.

The density estimator expressed in terms of an estimated probability density function (pdf), evaluated at zero, is convenient, as a large statistical literature exists on the subject of estimating a pdf.

Hence, the estimator of density for line transect is

$$\check{D} = n / 2L\mu$$

where

$$\mu = \int_0^w g(x)dx$$

and μ is estimated with the modeling of g(x). Here, g(x) is the model of the detection function.

Note that x or y refers to the vertical distances measured from the line transect to the detected object.

Assumptions

Although many of the objects of interest may go undetected, unbiased estimates of density can still be made if the following conditions and assumptions are met (Anderson et. al., 1993).

- It is assumed that a population comprises objects of interest that are distributed in the area to be sampled according to some stochastic process. It is critical that the transect lines are placed randomly with respect to the distribution of objects.
- The observer must be able to recognize and correctly identify the objects of interest. The distances from the line to the identified objects must be measured without bias. Objects directly on the line are always detected with certainty. Objects are detected at their initial location.

Modeling of the Detection Function

The true detection function g (y) is not known. A flexible or 'robust' model for g (y) is essential. The strategy used here is to select a few models for g (y) that have desirable properties. Four properties desired for a model for g (y) are in order of importance (Anderson et. al., 1993):

- Model Robustness -The model is a general, flexible function that can take the variety of shapes that are likely for the true detection function.
- Shape Criterion -The detection function should have a 'shoulder' near the line, which means that detection remains nearly certain at small distances from the line.
- Efficiency-It is desirable to select a model that provides estimates that are relatively precise (i.e. have small variance).
- Model Fit- Test available of the fit of the model for g (y) to the distance data is the X^2 goodness of fit test based on grouping the data.

Transect Layout

A systematic design using parallel transects with a random start is a favored and practical layout. Multiple transects, usually of unequal lengths, are to be extended from boundary to boundary across the study area. Transects are placed sufficiently far apart to avoid an object from being detected on two neighboring transects. As a practical minimum, the sample size n should be 60 (Anderson et al, 1993).

In the program DISTANCE 3.5 (Laake et al, 1993), the candidate key functions offered are the following distributions: Uniform, Half normal, Hazard-rate and Negative exponential while the candidate series expansions are Cosine, Simple and Hermite polynomials.

Analysis Guidelines

Generally, three analysis phases were considered; the exploratory, model selection, and final inference and interpretation.

Exploratory phase: This phase involves the preparation of histograms of the distance data under several groupings to assess presence of heaping, evasive movement, outliers and the occasional gross error. The program, DISTANCE 3.5, allows exploratory option like grouping or truncation of data prior to further analysis. Truncation of the distance data is nearly always suggested, which is 5 – 10 % of the largest observations, even if no obvious outlier is noticed.

Model selection: Model selection cannot proceed until proper truncation and grouping have been tentatively addressed. This phase begins once a data set has been properly prepared. Several robust models should be considered. The criteria that models for the detection function should satisfy (Akaike, 1973) conditions such as on robustness of the function (shape, estimator efficiency). The likelihood ratio tests are employed for each addition of adjustment term. The model with the lowest (AIC) is taken.

Final analysis and inference: The analyst selects a model believed to be the best for the data set under consideration. There may be several competing models that seem equally good. In most cases, there will be a subset of models that can be excluded from final consideration because they perform poorly relative to other models. Often, if two or three models seem to fit equally well to a data set, estimation of density under these models will be quite similar.

Methodology

1. Simulation by Match Stick Distance Project$_1$ (MSDP$_1$)

This project was purposed to compare density estimates of line transect distance sampling technique with the traditional square plot and strip transect sampling approaches.

Five hundred four black headed match sticks were stochastically distributed over a ten by six square meter field. The field was divided into sixty 1x1 square meter grids. The actual density was 8.4 match sticks per square meter.

i. Square plot sampling approach.

Forty percent of the total number of grids were taken as sample for the square plot sampling approach that is equal to 24 grids. Match sticks were counted in each grid and the total number n of match sticks in all 24 grids was taken. The total area a covered by the 24 grids was also computed so that estimate of match stick density $\check{D}$ was computed as $\check{D} = n / a$ while estimate of total population was computed as $\check{N} = \check{D}\ A$ where A is the actual area of the field of study. The estimates were then compared to the actual density and population.

ii. Strip transect sampling approach

In strip transect sampling approach, the first strip was established with a random start at grid 53 to grid 21 with length = 4.33 m and $2w$ = 0.333 m where w is half the width. Three more strips of the same width were established parallel to the first strip. They have the same lengths equal to 7.4 m each. All match sticks n within the strips were counted and recorded. The total area a covered by the strips was computed as a = $(L_t)(2w)$ where L_t is the total length of the four strips and w is half the width of the strips. The estimated density was then computed as $\check{D} = n / a$

while estimated population was computed as $\check{N} = \check{D}$ A. The resulting estimates were then compared to the actual density and population.

iii. Line transect distance sampling approach

A sixty-two meter serpentine was established over the field with a random start at grid 41 ending up at grid 50. This was done to facilitate the marking of matchsticks, which are supposed to be undetected. At every count of two every second match stick very near to or on the serpentine were marked red indicating that they were not to be detected in the line transect distance survey.

Five parallel line transects were established over the field with a random start at grid 50 and at 1 meter distance between transects. The lines were traversed and vertical distances between transects and matchsticks detected within width = 20 cm were recorded. Marked matchsticks within the stated width were considered undetected. The ungrouped vertical distances were then entered into DISTANCE 3.5 and analyzed. The data was truncated at the largest distance measured. Nine combinations of key functions and adjustment terms were constructed and modeled. The model with the lowest AIC was drawn out as the final model for the detection function together with its corresponding estimates of density and abundance. The final estimates of density and abundance were then compared to the actual density and total population.

All density estimates from the three sampling techniques were compared and the best technique was pointed out based on which gave the more truthful, if not accurate results.

2. Simulation by Match Stick Distance Project$_2$(MSDP$_2$)

Five hundred four matchsticks were prepared for distribution in a 6 by 10 square meter lot (60 m^2). Three hundred six of which were distributed in the first 18 square meter area, 144 in the next 18 square meter area, and 54 in the third 18 square meter area and none in the last 6 square meter area. This was done to establish a gradient of population. The whole area was divided into 60 square quadrats measuring 1 square meter each and the 22nd quadrate was randomly taken. A line transect was established starting at the 22nd quadrate parallel to the population gradient. Three line transects were also established to the right of and parallel to the first transect and two more to the left, a total of six transects with a total length of 60 meters. The transects were then traversed and distances between detected match sticks and line transects were recorded and for every five match sticks within the specified width the fifth is left as undetected. This fifth matchstick was randomly taken so as to facilitate a twenty percent possibility that not all matchsticks within the vicinity of observation are detected.

The ungrouped data were then entered into DISTANCE 3.5 and analyzed. The data was truncated at the largest vertical distance measured. To illustrate the process of modeling in detail twelve combinations of key functions and adjustment terms were constructed and the model with the lowest AIC was taken as the final model of the detection function. The resulting estimates of density and abundance were then compared to the actual.

3. Simulation by Rice Grains

One thousand rice grains were distributed over an area equal to 2.32 m^2. Line transects of total length = 2.87 meters were established over the area and vertical distances between detected rice grains and the line transects were recorded. Every fourth rice grain detected is not recorded. These observations were keyed in to DISTANCE 3.5. Using the existing four key functions and three expansion series 'built in' in the program, the researcher analyzed the data exhausting all 12 combinations. The one with the lowest AIC was taken as the best combination for the detection function.

4. Primary Data from Coconut Distance Project (CDP)

A map of a 24-hectare farmland located at Mabuhay, Liloy Zamboanga del Norte was secured. A census of coconut trees was conducted. There were 706 coconut trees. Square quadrats were drawn over the map and quadrat 41 was randomly taken where the first line transect was established. Four parallel line transects with a total length of 1800 meters and a perpendicular distance of 60 m between each were traversed and distances between detected coconut trees and line transects were measured. The data were then entered into DISTANCE 3.5 for analysis.

5. **Primary Data From Mangrove Distance Project (MDP)**

This project was conducted to apply the theory of line transect distance sampling technique to actual field survey.

A good map of Langamon, Anibongan, Gigaquit, Surigao del Norte was secured and a reconnaissance from the opposite beach was made. The survey team looked for a possible gradient of mangrove trees. It was observed that mangrove trees were so dense at one side of the place where the beach is not facing the open sea but Alambeque river.

Quadrats were drawn over the map and the 15th quadrat was taken where the first line transect was established at S 65° W bearing, parallel to the gradient of mangrove population based on the team's observation. Four parallel line transects were placed across the area with a fifty meter perpendicular distance in between. Distances between mangroves detected and the line transects were measured and recorded. Truncated at the largest distance measured, the data were entered into DISTANCE 3.5 and analyzed.

Results and Discussions

1. **Match Stick Distance Project$_1$ (MSDP$_1$)**

Square plot sampling (SPS) approach

Twenty four square plots were censused and a total of 155 match sticks were counted. The total area covered by the plots was 24 square meters. The estimated density was 6.46 match sticks per square meter while the actual density was 8.4 a difference of 1.94 ms/ m^2.

Strip Transect Sampling (STS) approach

Eighty three match sticks were censused from the four strip transects with a total length of 26.53 meters. The width of the transects was 2w = 0.333 meters and the total area covered by the strips was 8.83 m^2. The estimated density was 9.4 ms/ m^2 and the estimated abundance was 564 match sticks.

Line transect distance sampling (LTDS)

Table 1 shows the details in the selection of the model detection function for LTDS. Nine combinations of key functions and adjustment terms were constructed The model with the lowest AIC is the uniform key + 1 cosine series (*). It yielded a density estimate of 8.3524 matchsticks / m^2 and an abundance estimate of 501 match sticks. The actual density was 8.4 and the actual number of matchsticks was 504, a difference of 0.048 in density and 3 matchsticks in abundance.

Table 1. Analysis of Match Stick Distance $Project_1$ data ($MSDP_1$: LTDS)

Model	AIC	D	N
Half-normal + Cosine	657.11	8.051	483
Half-normal + simple polynomial	657.11	8.051	483
Half-normal+Hermite polynomial	657.11	8.051	483
Uniform + Cosine *	656.21 *	8.352 *	501 *
Uniform + simple polynomial	658.24	8.302	498
Uniform I Hermite polynomial	658.24	8.302	498
Hazard-rate + Cosine	657.72	8.927	536
Hazard-rate + simple polynomial	657.72	8.927	536
Hazard-rate+ Hermite polynomial	657.72	8.927	536

Table 2. gives the summarized results of Match Stick Distance Project$_1$ (MSDP$_1$).

Table 2. Results of the three sampling approaches (MSDP$_1$)

	Density			Abundance		
	Actual	Estimated	Difference	Actual	Estimated	Difference
SPS	8.4	6.46	1.940	504	387.6	116.4
STS	8.4	9.40	1.000	504	564	60
LTDS	8.4	8.35	0.048	504	501	3

Where: SPS is square plot sampling
STS is strip transect sampling
LTDS is line transect distance sampling

Comparing the results of the three sampling approaches, the estimates yielded by line transect distance sampling (LTDS) are closest to the actual values.

2. Match Stick Distance Project$_2$ (MSDP$_2$)

Twelve combinations of detection functions were constructed out of the MSDP$_2$ data. Table 3 gives the summary.

The three combinations of the hazard-rate key with the three types of expansion series yield the same AICs. The number of adjustment terms is zero. Hence, the hazard-rate key here needs no expansion series. It means that the key function alone is sufficient to model the detection function

The half-normal key function alone is also sufficient, but with one simple or Hermite polynomial adjustment term, the precision is increased as implied in the decrease of the AIC from –456.48 to –457.47 or –457.05 respectively.

The uniform key function plus one simple or Hermite polynomial gives a lower AIC than the uniform key + three cosine series.

Table 3. Analysis of Match Stick Distance $Project_2$ data ($MSDP_2$)

Model detection function	Model selected	Number of parameters	Number of adjustment terms	AIC
Half-normal + C	1	1	None	-456.48
Half-normal + Sp	2	2	1	-457.47
Half-normal + Hp	2	2	1	-457.05
Uniform + C	4	3	3	-457.28
Uniform + Sp	2	1	1	-457.79
Uniform + Hp	2	1	1	-457.79
Hazard-rate + C	1	2	None	-458.80*
Hazard-rate + Sp	1	2	None	-458.80*
Hazard-rate + Hp	1	2	None	-458.80*
Negative exp + C	1	1	None	-453.27
Negative exp + Sp	2	2	2	-455.79
Negative exp + Hp	2	2	2	-455.79

Where: Hp is hermite polynomial, C is cosine,
Sp is Simple polynomial, exp is exponential

The negative exponential plus two simple or two Hermite polynomial terms also gives a better fit than just the key alone. The corresponding AICs of these combinations are higher compared to other combinations. Since good fit yields minimum AICs, the negative exponential key combinations are eliminated.

All three combinations of the half-normal key are taken. So, from here, six combinations are singled out to model the detection function. Table 4 exhibits the 6 final combinations.

Table 4. Modeling the detection function

Model (key+adjustment terms)	W = 14		
	Number of parameters		AIC
	Key	adj	
Half-normal key	1	0	-456.48
Half-normal + Simple Polynomial	2	1	-457.47*
Half-normal + Hermite Polynomial	2	1	-457.05
Uniform + Cosine	3	3	-457.28
Uniform+Simple/Hermite Polynomial	1	1	-457.79*
Hazard-rate key	2	0	-458.80*

The number of combinations should further be reduced for final comparison. Out of the six combinations three were taken based on minimum AIC: the half-normal key + one simple polynomial, the uniform key + one Hermite or simple polynomial, and the hazard-rate key (alone). These three contending combinations, as shown in Table 4, were further analyzed for the best to approximate the detection function.

Table 4. Best model fit

Model (key+adj)	W = 14							
	# of parameters		AIC	D	N	CV (%)	X^2 – P-value	P
	Key	Adj						
Hn + Sp	2	1	-457.47	8.08	485	7.89	0.486	0.846
U+Sp/Hp	1	1	-457.79	8.86	532	9.97	0.347	0.772
Hr	2	0	-458.80*	8.12	487	10.00*	0.503*	0.843*

where: Hn = half-normal, U = uniform, Hr = hazard-rate

In terms of AIC values, all three combinations are considered good models already. There is not much to choose between the models since they have similar AICs. All the three are very close but the lowest AIC corresponds to the hazard-rate key. Lowest AIC is an estimate of the best approximating model. Assuming no problems with the data, Len Thomas-author of Distance Book and DISTANCE 3.5, say she would go for the one with the lowest AIC).

Hence, since the AIC value of the hazard-rate key is the lowest it is chosen as the best model for the detection function to estimate population density and abundance of match sticks in this particular study. To assess its adequacy, the coefficient of variation (cv in %), Chi Square Goodness of fit (χ^2 GOF) test and probability of detectability (P) are to be considered. A 10 to 20 percent cv is usually good. All 3 models have good cv_s. With respect to model fit, higher χ^2 - p value is preferred because it implies better fit. The $\chi 2$ - p value of the hazard-rate key (0.503) is the highest, meaning it has the best fit. With respect to P which is probability of higher object detection in the area, the hazard-rate key is only a little bit lower than the half-normal + one simple polynomial by 0.003.

Hence, for this line transect sampling data ($MSDP_2$), the hazard-rate key with no adjustment terms is the best detection function model. The chosen model is:

$$g(y) = 1 - \exp(-(y/\sigma)^{-b})$$

where the parameters are $\sigma = A(1) = 0.1097$ and $b = A\ (2) = 3.5$ (Appendix I).

Hence, $$g(y) = 1 - \exp[-y/0.1097)^{-(3.5)}].$$

The estimated density of matchstick per square meter is 8.12 while the actual density is 8.4, a difference of 0.28. The estimated abundance is 487 while the actual is 504 matchsticks, a difference of 17 matchsticks. This estimate is closest to 504 matchsticks.

3. Rice Grain Distance Project (RGDP)

Line transects were traversed and distances between detected rice grains and transects were measured and recorded. We have the following information: Total effort = 2.87 m and n = 102 distances.

The data were keyed in to Distance 3.5 for analysis. Analysis of twelve combinations of key functions and adjustment terms were made.

The combination with the lowest AIC was taken. The following are the results.

Model: Uniform + 1 Cosine term
AIC = -611.86 (lowest AIC)
D' = 439.92 grains/m^2.
N' = 982 grains

4. Coconut Distance Project (CDP)

From the twenty-four hectares there are a total of 706 trees or 29.417 trees in one hectare. The analysis is based on the largest distance, which is 9.98 meters. The total number of detected objects is 87 trees. The total number of line transects is 4, which sums up to a total length of 1800 meters. The following are some histograms of the data at 6 cut points with different key functions.

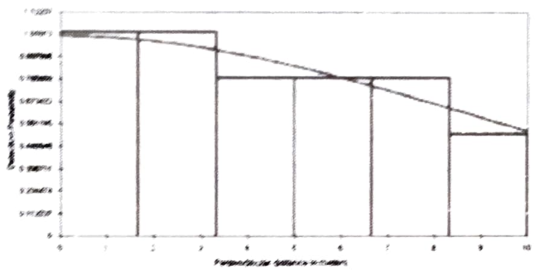

Figure 1. Detection Probability plots at 6 cut point Half-normal key alone

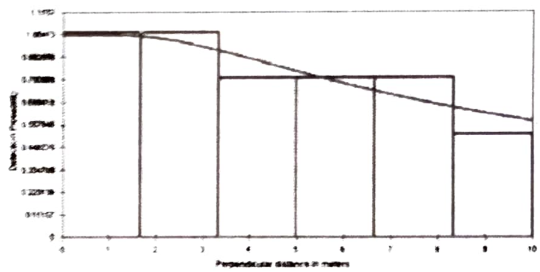

Figure 2. Hazard-rate key alone

As what can be observed in the histograms, all four models showed a broad shoulder so any of them can be used to model the data. But we have to choose the best. In Table 5 below, the smallest AIC is 399.41, which corresponds to the half-normal key alone and the uniform key plus one-simple/Hermite polynomial. How to judge between these two?

Table 5. Distance 3.5 Analyses on Coco Distance Project

Model Key + adj	W = 10 M						
	AIC	D	N	Cv(%)	X^2 value	X^2 p value	P
Hn	399.41*	29.46	707	15.28	0.679	0.954*	.406*
U + C	399.62	30.33	728	16.15	0.919	0.922	.378
U+ Sp / Hp	399.41*	28.98	696	13.98	0.648	0.958*	.411*
Hr	401.41	29.64	711	21.60	0.791	0.852	.321

Where: Hn = Half-normal, C = Cosine, Hp = Hermite polynomial, U = Uniform, Sp = Simple polynomial, Hr = Hazard-rate

Lets look at the X^2 goodness of fit test. The uniform key plus one simple or Hermite polynomial has better fit than the half-normal key alone. In terms of detection probability value it has a higher P-value.

Hence, the uniform key plus one simple or Hermit polynomial model is the best model for the detection function of the coconut distance data.

5. Mangrove Distance Project (MDP)

Mangrove trees are detected whether they are just seedlings, sapling or matured. Vertical distances between detected mangrove trees and the line transect are measured and recorded. The number of trees detected are $n_1 = 33$, $n_2 = 36$, $n_3 = 11$, and $n_4 = 0$, hence, a total of 80 vertical distances. The total transect length $L_t = 960$ meters.

With truncation at largest distance, the researcher entered the data into DISTANCE 3.5 for analysis. Histograms at different cut points (Figures 3-6) were plotted to explore the data and to see its characteristic. In here, the histogram with five cut points shall be used for illustration and each of the candidate model will be fitted into it.

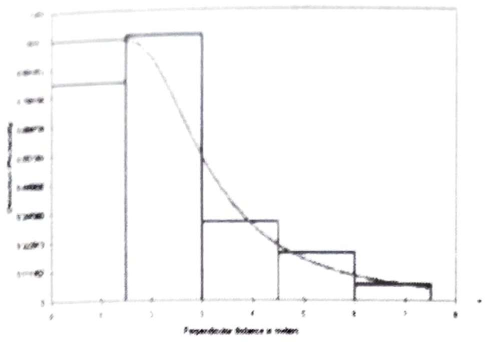

Figure 3. MDP data at 5 cut points

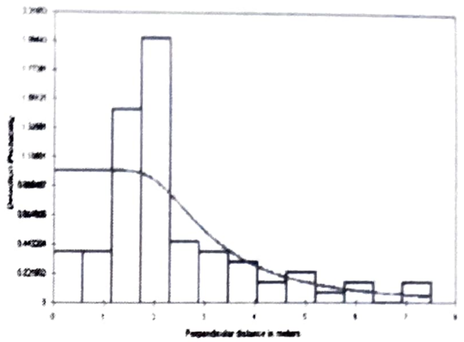

Figure 4. MDP data at 13 cut points

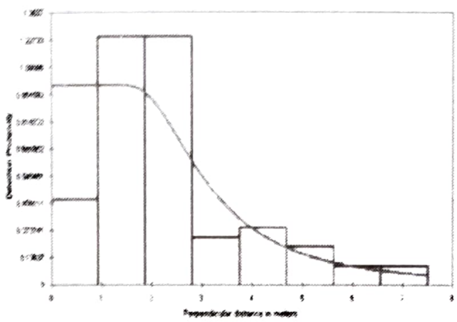

Figure 5. MDP data at 8 cut points

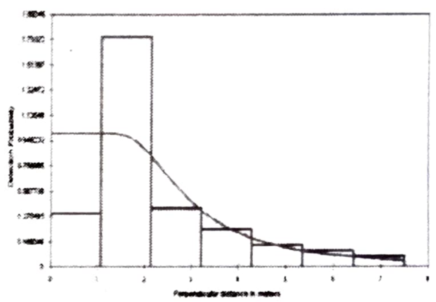

Figure 6. MDP data at 7 cut points

To start the analysis on model construction and selection the researcher chooses one data filter and constructs twelve combinations of model definitions and names it analysis set 1. At the data filter the data is truncated at largest distance, which is 7.5 meters. According to Anderson et. al.(1993), one can discard 5 –10 % of the data from the largest distances, or one can discard all observations beyond a specific distance from the transect line.

Table 6-9 show these combinations together with their fit into the histogram of data (Figures 7-10). Table 6 is about the hazard rate key combinations.

Table 6. The hazard-rate key combinations

Model combination	# of models fitted	Model selected	# of parameters	Adjustment terms		AIC
				#	Order	
Hr + C	2	1^{st}	2	0	None	273.91
Hr + Sp	2	1^{st}	2	0	None	273.91
Hr + Hp	2	1st	2	0	none	273.91

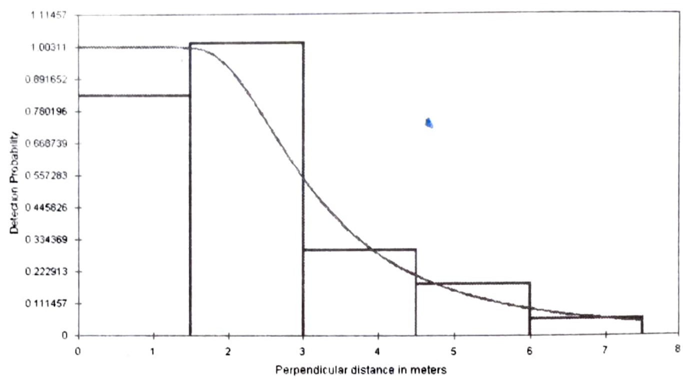

Figure 7. The hazard rate key (alone), p = 0.224

Each of the hazard rate key combinations yields two models. The first model is just the key function alone while the second model is with the addition of the adjustment terms with their corresponding order. The addition of adjustment terms (i.e. cosine term, simple or hermite polynomials) to the key function has not improved the fit of the data as implied by a higher AIC in model 2 (Appendix A: Model fitting) so that DISTANCE 3.5 selected the 1st model as the final model for the hazard rate key combinations based on lowest AIC. Take notice of the AICs, no wonder, all three combinations have the same AIC. Table 7 is all about the uniform key combinations.

Table 7. The Uniform key combinations

Model comb'n	# of models fitted	Model selected	# of parameters	Adjustment terms		AIC
				#	Order	
U + C	4	3rd	2	2	1 , 2	280.15
U + Sp	4	3rd	2	2	2 , 4	281.49
U + Hp	5	4th	3	3	2, 4, 6	283.05

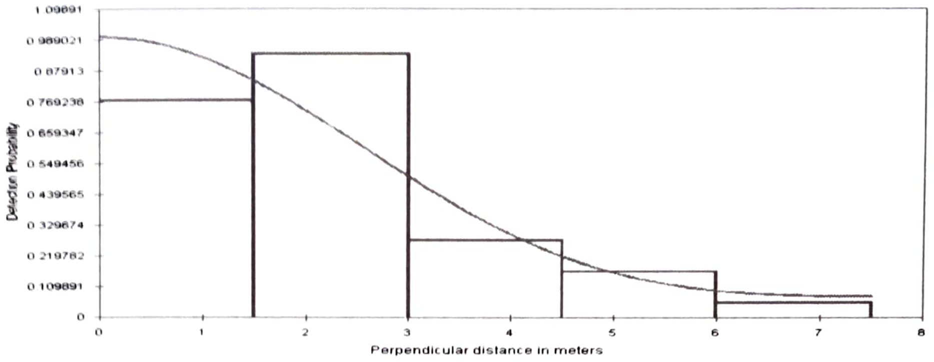

Figure 6. The Uniform key + two cosine series of orders 1 and 2

There are also three uniform key combinations. The uniform key + cosine combination has four models fitted into the data and the third model was selected based on minimum AIC. The uniform key + simple polynomial combination has four models fitted into the data and the third model was selected again based on minimum AIC. The uniform key + Hermite polynomial combination has five models fitted into the data and the fourth model was selected .

For the uniform key combinations, the model with the lowest AIC is the uniform key function plus two cosine series of orders 1 and 2.

There are also three half-normal key combinations which yield two models each (Table 8).

Table 8. The half-normal key combinations.

Model combination	Number of models fitted	Model selected	Number of parameters	Adjustment terms		AIC
				Number	Order	
Hn + C	2	1st	1	0	None	279.87
Hn + Sp	2	1st	1	0	None	279.87
Hn + Hp	2	1st	1	0	none	279.87

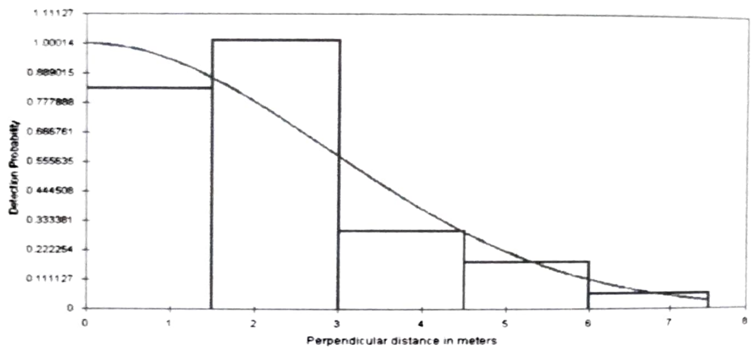

Figure 7. The half-normal key (alone) p= 0.13710

The addition of adjustment terms to the half-normal key does not improve the fit of the key function so that the program chose the first model for each of the three combinations based on minimum AIC. So, the half-normal key combinations is represented only by the key function with no adjustment terms. The AIC for this model is 279.870.

The p-value (Figure 7) is 0.13710 which is not significant. Hence, the half-normal key alone fits well into the data.

The three negative exponential key combinations yield four models each (Table 9). Based on minimum AIC, the third model in all three combinations is selected. All three combinations have three parameters and two adjustment terms. The cosine series are of orders 1 and 2 while the simple and Hermite polynomials are of orders 2 and 4. Table 9 shows that the AIC values of the negative exponential combinations are quite large as compared to the other combinations so that this set of combinations is eliminated first in the selection of the model detection function.

Table 9. The negative exponential combinations

Model combination	Number of models fitted	Model selected	#of parameters	Adjustment terms		AIC
				#	Order	
Ne + C	4	3rd	3	2	1, 2	282.148
Ne + Sp	4	3rd	3	2	2, 4	283.352
Ne + Hp	4	3rd	3	2	2, 4	283.352

Table 10 gives the results of mangrove distance data analysis where five combinations are finally considered. Three (*) models are contending for the best, the Uniform key plus two cosine terms, the half-normal key alone and the hazard-rate key alone. To choose the best model, the lowest AIC is considered.

Table 10

Truncated at largest distance. Width = 7.5 m							
Model	AIC	D	N	cv%	X^2 val	$\chi 2$ - p value	P
U + C	280.15*	125.75	503	15.16	5.689	0.058	0.442
U + Sp	281.49	105.91	424	13.27	6.251	0.044	0.525
U + Hp	283.05	108.09	432	13.34	5.930	0.015	0.514
Hn	279.87*	116.86	467	13.78	5.530	0.137	0.475
Hr	273.91*	116.51	466	16.48	2.992	0.224	0.477

The hazard-rate key has the lowest AIC which is 273.91. To assess its adequacy, the fit is considered by looking at the χ^2 – p value. It has the best fit which is 0.224. Hence, the Hazard-rate key, alone, is the best model for this mangrove distance sampling data.

Hence, the model for the detection function that best fits the data of Mangroves is the following:

$$g(y) = 1 - \exp[-(y/\sigma)^{-b}]$$

where σ = A(1) =2.78 and b = A(2) = 3.015 (Appendix A: Parameter estimates).

Hence $$g(y) = 1 - \exp[-(y/2.78)^{-3.015}].$$

With this detection function model, the density estimate is 116.51 mangroves per hectare and the abundance estimate is 466 mangroves.

Table 11. Summary of Results

Project	Actual		Distances Measured	% Non detection	Error %	Estimated	
	Density	Total				Density	Total
$MSDP_1$	8.4	504	112	11.81	0.57	8.352	501
$MSDP_2$	8.4	504	115	20	3.33	8.12	487
Rice grains	448	1000	102	25	1.80	439.92	982
Coconut	29.42	706	87	-	1.42	28.98	696
Mangrove	-	-	-	-	-	116.51	466

Bibliography

Akaike, H. Information Theory and an extension of the maximum likelihood principle, in International Symposium of Information Theory, 2nd ed (B. N. Petran and F. Csaaki), Akadeemiai Kiadi, Budapest, Hungary, pp. 267-81. 1973.

Anderson, D. R. et al. Distance Sampling: Estimating Abundance of Biological Populations. Chapman and Hall, London, 1993.

Thomas, L. et al. Distance Research Unit for Wildlife Population Assessment. University of St. Andrews, UK. 1998.

http://www.ruwpa.st-and.aac.uk/distance/