

9th National Convention on Statistics (NCS)
EDSA Shangri-La Hotel
October 4-5, 2004

Distance Sampling Simulated For Density Estimation

by
Rebecca S. Galela and Brigida A. Roscom

For additional information, please contact:

Author's name:	Rebecca S. Galela/Brigida A. Roscom
Designation:	Instructor 1/ Dean
Agency:	College of Education/Graduate School Mindanao State University – Iligan Institute of Technology
Address:	Andres Bonifacio Avenue, Iligan City
Telefax:	(06363) 223-2351/(06363)-223-2345
E-mail:	beckysgalela2002@yahoo.com / csm- bar@sulat.msuiit.edu.ph

Distance Sampling Simulated For Density Estimation

by

Rebecca S. Galela and Brigida A. Roscom¹

ABSTRACT

Line transects distance sampling for estimating density was simulated in three projects using matchsticks and rice grains 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. In all cases, selection of the chosen model was based on lowest Akaike's Information Criterion (AIC).

The experiences from the simulations were used to estimate the density of coconut trees in a 24-hectare farm land of Mabuhay, Liloy, Zamboanga del Norte.

Results of the simulation using the software package DISTANCE 3.5 showed that estimates by distance sampling was very close to the true density. Future research directions with mobile or clustered objects of interests may be considered.

Keywords and phrases: line transect, distance sampling, density estimation, detection function.

I. 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.

II. Objective

The study aims at illustrating the principles and methodology of distance sampling, construction of the detection function, the modeling process and model evaluation using simulated data. Matchsticks, rice grains and coconut trees were used for the illustration.

¹ Professor and Dean, Graduate School, Mindanao State University – Iligan Institute of Technology

III. 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 lengths of each transect lines 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.

A. 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 (Buckland 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.

B. 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 χ^2 goodness of fit test based on grouping the data.

C. 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 usually be at least 60 (Buckland et al, 1992).

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.

D. Analysis Guidelines

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

1. **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.
2. **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.
3. **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.

E. Simulation by Match Sticks (First Trial)

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.

F. Simulation by Match Sticks (Second Trial)

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 quadrants 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 that 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 similar to trial 1.

G. 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.

H. 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 quadrants 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.

IV. Results and Discussion

As shown in Table 1 the model with the lowest AIC is the uniform key + 1 cosine series (*). It yielded a density estimate of 8.3524 matchsticks / m² 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 Matchsticks Project (trial 1)

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 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 below refers to trial 2 using matchsticks.

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.

Table 2. Analysis of Matchsticks Project (Trial 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

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

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 3 below exhibits the 6 final combinations.

Table 3. Modeling the detection function

Model (key+adj)	W = 14		
	# 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 (%)	χ^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, Sp = simple polynomial, Hp = Hermite polynomial

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%), 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 χ^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 (MSDP₂), 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 $b = A(2) = 3.5$.

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.

A. Rice Grains Distance Project (RGDP)

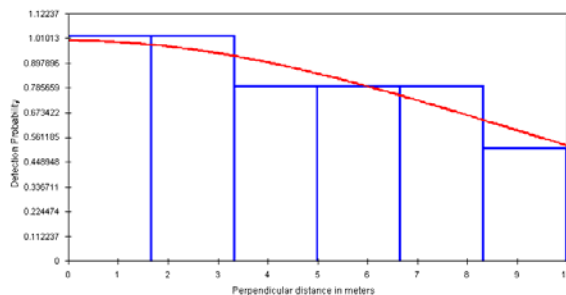
Line transects were traversed and distances between detected rice grains and transects were measured and recorded. We have the following information: Total effort = 2.87 m and n = 102 distances.

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/m².
N' = 982 grains

B. 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 points.
Half-normal key alone**

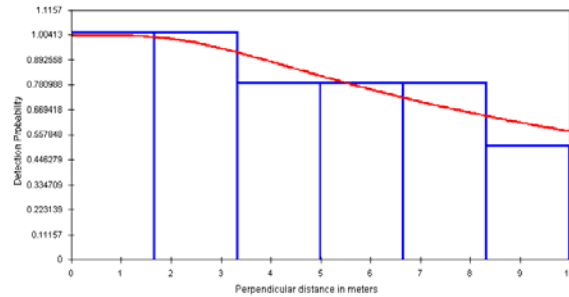


Figure 2. Hazard-rate key alone

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, U = Uniform, C = Cosine, Sp = Simple polynomial, Hp = Hermite polynomial, Hr = Hazard-rate

Lets look at the χ^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.

Table 6. Summary of Results

Project	Actual		Distanc es Measur ed	% Non detection	Error %	Estimated	
	Density	Total				Density	Total
MSDP ₁	8.4	504	112	6.15	0.57	8.352	501
MSDP ₂	8.4	504	115	20	3.33	8.12	487
Rice grains	448	1000	102	25	1.8	439.92	982
Coconut	29.42	706	87	-	0.146	29.46	707

BIBLIOGRAPHY

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

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 Kiado, Budapest, Hungary, pp. 267-81. 1973.

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/>