American Journal of Theoretical and Applied Statistics
2023; 12(2): 18-31
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20231202.11
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Abundance Determination of Large Mammals with Distance Sampling Perspective, the Case of Elephants of the Mole National Park (MNP) of Ghana

Katara Salifu

Department of Statistics, Faculty of Physical Sciences, University for Development Studies, Tamale, Ghana

Email address:
skatara@uds.edu.gh

To cite this article:
Katara Salifu. Abundance Determination of Large Mammals with Distance Sampling Perspective, the Case of Elephants of the Mole National Park (MNP) of Ghana. *American Journal of Theoretical and Applied Statistics*. Vol. 12, No. 2, 2023, pp. 18-31.
doi: 10.11648/j.ajtas.20231202.11

Received: November 30, 2022; **Accepted**: January 3, 2023; **Published**: May 18, 2023

Abstract: Distance sampling with line transect method has been applied by many researchers to monitor and observe varied animals and plants with the aim of determining the population density and or abundance of animals. The application of this method has not received the needed attention in Ghana, in particular to monitor, observe, and estimate the densities and abundance of animals and plants in the game reserves of the Mole National Park (MNP) is not without exception and the statistics of these are always reported based on guesses and without any scientific proof. This study has seen the application of line transect methodology in the MNP in which the abundance estimates are statistically determined with both the classical and Bayesian philosophies of statistical approaches. An alternative means of detectability estimation using the total probability concept has been established to enhance the probability of detection of a rare and elusive population of large mammals. In performing statistical investigations on rare and elusive population, it appears insufficient to model from the classical perspective, the use of PRIOR knowledge as seen in the Bayesian context cannot be underestimated. This study proposed that the concept of Total Probability with prior knowledge of animals and plants in line transect surveys must be well embraced, Periodic censuss must be conducted regularly to help in establishing the rate of extinction of units of interest in wildlife and Distance sampling data with line transect sampling methodology need not be analysed using only the classical reasoning. Attention must be given to the existence and availability of prior knowledge of the units under study.

Keywords: Distance Sampling, Detectability, Line Transect, Density, Abundance, Data Augmentation, Mole National Park

1. Introduction

1.1. Background

The population size of the units under investigation supports much statistical analysis and investigation of which wildlife ecology and environmental biology form apart. In ecology or environmental biology, a complete study of the species concerned appears very challenging and density or abundance estimations of wildlife populations are based on sampling methods generally known as distance sampling. Distance sampling is a widely used group of closely related methods for estimating the density and or abundance of biological populations [18].

Line transect sampling is seen or has been observed to be one of the most widely used techniques for wildlife population size estimation and has been used for many types of populations, including bird, mammal, and plant species as well as other objects for which detectability depends on location relative to the observer [7]. The authors further stressed that under the transect methodology, observers typically survey the area of interest by traversing several spatially replicated lines to detect units of interest either in clusters of many or individuals while measuring or recording the distances of objects on either side of the line and modelling of detection with the observed distance data. When direct observations are not possible, counting animal signs provides a relative measure or "index", a measure often used to study secretive species [22, 26].

The most accurate estimates are obtained by methods

relying on the count of animals, or their signs, where the estimate of individuals per unit area allows the calculation of the total population [8, 11, 23, 24]. Partial counts enable a population subset to be obtained, allowing the whole population to be measured without the constraint of counting all individuals [22].

Assume $x_1, x_2, \ldots, x_n$ are perpendicular distances recorded by an observer from a transect line believed to follow a specific Probability Density Function (pdf), [5] suggested that the detectability function $g(x)$ is related with the pdf $f(x)$ by

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

And

$$f(0) = \left[\int_0^w g(x)dx\right]^{-1} \quad (2)$$

Where w is a truncated distance and the population density D is expressed as

$$D = \frac{E(n)f(0)}{2L} = \frac{n*f(0)}{2L} \quad (3)$$

Where L and E(n) length and expected number of units observed.

1.2. Problem Statement

Ghana has had an increase in population over the years with an estimated growth rate of about 2.07%. With such an increase in human population, human activities undoubtedly appear to increase, poaching for wild animals and plants in the MNP is not unlikely to occur. The occurrence of poaching may affect species numbers but by how much may remain unanswered for several decades.

The effective management and conservation of wild species of plants and animals appear dependent on methods with accurate estimates of density and abundance [11]. In Ghana and MNP in particular, the methods of density and abundance estimates appear unknown and estimates of mammals are based on guesses without scientific proof or establishment. This paper is thus aimed at establishing density estimates with a variety of principles.

1.3. Research Questions

1) Do the statistics of mammals in the MNP actual represent the figures of those units of interest?
2) Can abundance estimation with distance sampling technique be applied to Bayesian reasoning?

1.4. Study Objective

This study is meant to find density and abundance estimates of units of interest under study based on line transect laid down procedures and applications using both the classical and Bayesian reasoning with application of the principle of total probability concept.

2. Methods

2.1. The Study Area

The study area is located in the West Gonja District of the Savannah region of Ghana as seen in Figure 1 and about 184km from the Northern Regional capital of Tamale. It covers a total area of 4,755 square kilometers of the West Gonja District which inhabits several plant and animal species whose densities are of concern to management and conservation.

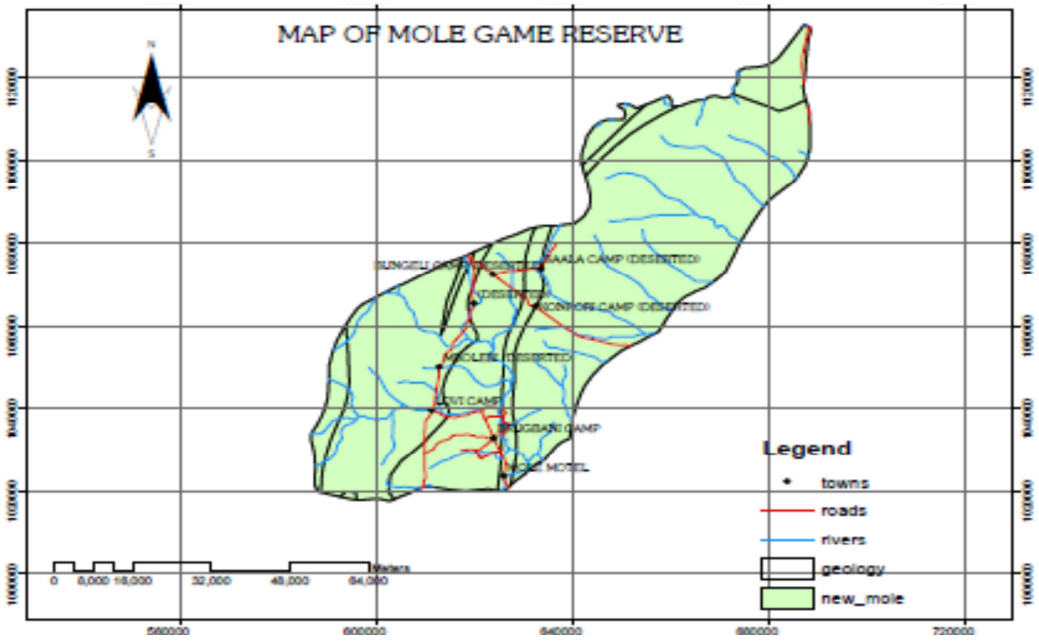


Figure 1. *Map of Ghana with Location of the MNP [16].*

Briggs, K. T., Tyler, W. B. and Lewis, D. B. revealed that the MNP represents Ghana's largest wildlife refuge which is located northwest Ghana on grassland savannah and riparian ecosystems at an elevation of 150m, with a sharp escarpment forming the southern boundary of the park [4]. The park's entrance is reached through the nearby town of Larabanga. This area of Ghana receives over 1000 mm per year of rainfall.

2.2. Pilot Survey

This provides a platform for a preliminary study to be carried out on a small scale in the study area with a purpose that includes the determination of the total length of the transect for reliable and precise estimates in the Mole National Park.

2.2.1. Sample Size Selection

The number of line transects can constitute a component of a sample size in which a minimum of 10 - 20 replicated lines are recommended to allow for a reliable estimation of parameters of interest with at least 60 - 80 detection's of animals or a cluster of animals for reliable estimation and modeling of the detection function [6].

The variance estimate according [10] is:

$$\hat{v}ar(\hat{D}) = (\hat{D})^2\left\{[\hat{c}v(n)]^2 + \left[\hat{c}v(\hat{f}(0))\right]^2\right\} \quad (4)$$

$$\text{Where } [\hat{c}v(n)]^2 = var(n)/(E(n))^2 \quad (5)$$

$$\left[\hat{c}v(\hat{f}(0))\right]^2 = var(\hat{f}(0))/(E(\hat{f}(0)))^2 \quad (6)$$

Implies,

$$\left[\hat{c}v(\hat{D})\right]^2 = \frac{\hat{v}ar(\hat{D})}{(E(\hat{D}))^2} = \left[\frac{(\hat{D})^2\left\{[\hat{c}v(n)]^2+\left[\hat{c}v\left(\hat{f}(0)\right)\right]^2\right\}}{(\hat{D})^2}\right] = [\hat{c}v(n)]^2 + \left[\hat{c}v\left(\hat{f}(0)\right)\right]^2 = \left[\frac{var(n)}{(E(n))^2} + \frac{var(\hat{f}(0))}{(E(\hat{f}(0)))^2}\right] \quad (7)$$

Rewriting (4) to in the form

$$\left[\hat{c}v(\hat{D})\right]^2 = \frac{1}{n}[a_1 + a_2] = \frac{b}{n} \quad (8)$$

To determine the coefficient of variation to be tolerated on a pilot base where $var(n) = a_1 n$ and $var\left(\hat{f}(0)\right) = (f(0))^2\, a_2/n$ respectively, [12] provided evidence that the constant b may typically be between 2 and 4; however, *b=4* arise with less efficient estimators and a value of *b=2.5* is tenable with a risk of underestimating the needed line length of *b=1.5*.

If L_1 line length is covered in a pilot study to observe $n_1$units of interest, then the proportion $\frac{L}{n} = \frac{L_1}{n_1}$ holds without any loss of information where L and n respectively denote the total transects length covered in the study region and total number of units of interest observed (sampled) for the entire study period. Using equation (8) and $\frac{L}{n} = \frac{L_1}{n_1}$, the total length expected to cover can be expressed as:

$$L = \frac{b}{[\hat{c}v(\hat{D})]^2}\left(\frac{L_1}{n_1}\right) \quad (9)$$

2.2.2. Sampling Effort Determination

A pilot study revealed that on traversing a total of $L_1 = 2$ kilometres by the researcher, a total of $n_1 = 29$ elephants were detected. With the constraints on the part of the researcher, it was resolved to use a value of $b = 2.5$ as suggested by Eberhardt, L. L and tolerating a coefficient of variation of 5% ($\hat{c}v(\hat{D}) = 0.05$) [12]. The transect length obtained using equation (9) yields 69 km. This indicates that, at least a sampling effort of 69 km is expected to be covered.

2.3. Design for Selecting Transects

Katara S., S. K Amponsah and Bashiru I. I. S. stated that the sampling design in a line transect study is the procedure by which the transect locations are selected [17]. Desired properties of unbiasedness of estimators will be based as much as possible on the design rather than on assumptions about the population [26].

To select the appropriate transects for the study, a combination of both probability and non-probability sampling designs were employed for improve data quality and analysis. These designs included stratification, convenient, and systematic sampling, respectively. Due to the timing of the research and the nature of the study area, it was decided to stratify the study area into two major strata (a stratum with and without water bodies) within which separate designs are employed for transect placement. Areas, where water bodies are located, are believed to have more concentration of units of interest and hence more likely to have higher detectability. With this prior knowledge, more effort was allocated to this stratum to improve precision.

Conveniently identifying a random start line in each stratum, a combination of both continuous systematic design and discrete parallel transect lines were placed, respectively, in the strata to avoid discontinuation in detection from one transect to the other while ensuring an even spatial distribution of lines in the survey region.

2.3.1. Number of Transect Lines Used

Steven K. Thompson indicates that estimates from *n transects* are more preferred to those based on single transect [25], a recommendation emphasized by a number of authors [5, 12, 21, 25]. As recommended that a minimum of 10 - 20 replicated lines is capable of producing a reliable estimation and modeling [6], available resources could only permitted traversing through 10 replicated lines in the study area.

2.3.2. Methods of Observation

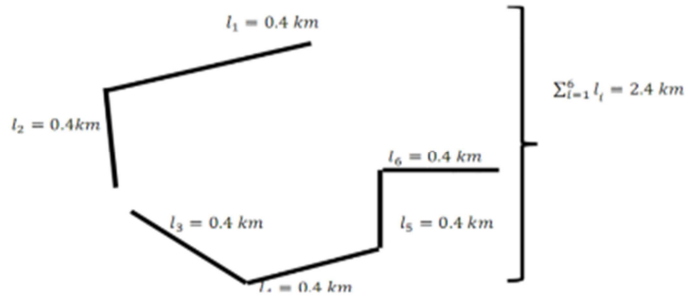
The study employs both direct and indirect methods of observation along transect lines placed in the study area. Transect lines are identified by employing simple and systematic sampling techniques. Direct methods are based on actual observation of the species in question (large mammals), while indirect methods rely

on interpreting the signs of animal presence. Visibility and detectability can often pose a problem when surveying terrestrial species, therefore, surveys relying on signs are ideal for estimating mammal abundance and habitat use [16, 22, 24]. The indirect method (the use of dung) was successful in confirming the present of the specific species, more particularly with the elephant species. To select the appropriate transects for the study, a combination of both probability and non-probability sampling designs were employed to improve data quality and analysis. These designs include stratification, convenient, and systematic, respectively. Due to the timing of the research and the nature of the study area, it was decided to stratify the study area into two major strata (a stratum with and without water bodies) within which separate designs are employed in transect placement. Areas, where water bodies are located, are believed to have more concentration of units of interest and hence more likely to have higher detectability. With this prior knowledge, more effort is allocated to this stratum to improve precision.

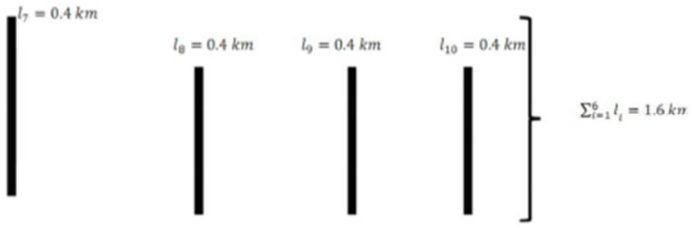
2.3.3. Transect Selection Procedure

To maximize the detectability of units of interest while satisfying the general requirements of line transect applications, the researcher must decide how many transect lines need to be placed in the coverage area and at what length each line must be stretched.

Given the nature of the study area and resource limitations, a total effort of 4 km per day was assumed to be covered in the study area during 20 days spread over 10 months, with at least 2 days spent on each visit to travel along a sample of ten (10) as seen from Figure 2. The first transects were conveniently chosen perpendicular to the baseline in each stratum and were thought of as random start samples. From there, the remaining eight non-overlapping and widely spaced lines were systematically positioned to cover and include specific locations with prior knowledge of a high concentration of the elephant population in the study area [17].



A: Continuous Zigzag lines (not to scale)



B: Systematically parallel lines (not to scale

Figure 2. *Traversed transect lines in the study area [17].*

2.4. Data Type

Two types of data were taken into account. An investigator and some supporting staff members collected the primary data in the study area. Ten conveniently and methodically set transect lines within two stratified areas were used to collect data on perpendicular distances as well as the quantity of elephants, hartebeests, waterbugs, and warthogs found at different points within the study zone. Regarding the elephant data in particular, footprints of units of interest and droppings or other ways were also taken into consideration as indicators of the presence of the unit of interest. The secondary data were previously gathered information regarding the study area's units of interest.

2.5. Detectability

In the realm of the basic distance sampling framework, the variable of interest is assumed to have been recorded without error for each unit in the sample. In a survey of elusive events such as birds, mammals, or the homeless, some units of interest may remain undetected. The probability that an object in a selected region is observed, whether seen, heard, caught, or detected by some other means represents its detectability. The process of detection can either be passive or active form depending on how the units of interest or objects are observed or detected.

2.5.1. Constant Detectability Within a Given Area A

Assume the probability of detection for a given unit of interest in a given region is a constant p throughout the region of area A. Denote y as the number of units observed in the region, while the actual (exact) number in the region is τ; i.e., the population total. Assume also that the detection of units of interest are statistically independent- i.e., the detection of one unit is not being influenced by another. We note that a unit in a region is

either detected or not with probabilities p and q, respectively. This follows a binomial distribution with the expectation.

$$E(y) = \tau p \tag{10}$$

And variance

$$var(y) = \tau pq \tag{11}$$

Since p is assumed known or constant, an estimator of the population total is expressed as

$$\tau = \frac{y}{p} \tag{12}$$

with variance expressed as

$$Var(\widehat{\tau}) = \frac{1}{p^2} var(y). \tag{13}$$

The density estimate with known detectability is now

$$\widehat{D} = \frac{y}{Ap} \tag{14}$$

with variance;

$$\widehat{Var(\widehat{D})} = \frac{1}{(Ap)^2} var(y). \tag{15}$$

2.5.2. Unknown Detectability over a Region of Area A

In general terms, the detectability of objects in the study region appears unknown and can be estimated by alternative methods including mark – recapture, ratio estimation with the use of auxiliary information, the number of units observed by observers in the air and on the ground, to mention but a few. For this study, I decided to employ the following two cases for reasons of convenience.

Using Coverage Region

This considers the detectability as a ratio of the area covered in the study region and the total area of the entire study region. Consider a line transect in which a region at a distance w from the line is searched which has an area a. If the area of the entire study region is A, the fraction

$$p_{c_i} = \frac{a}{A} \tag{16}$$

for all i's.

Suppose the transect lines are located randomly and independently of the units of interest, then on average, any covered object in the coverage region with probability p_{c_i} referred to as the "coverage probability". Now assume y objects are observed in the covered region with some units remaining unsighted.

Denote

$$z_j = \begin{cases} 1, if\ objects\ are\ sighted\ in\ the\ coverage\ area \\ 0, otherwise \end{cases}. \tag{17}$$

We can reasonably consider

$$p = \sum_{j=1} p_{c_i} * z_j \text{ for } j = 1, 2, \dots y \tag{18}$$

with abundance estimated as

$$\hat{\tau} = \frac{y}{p} \tag{19}$$

2.5.3. Density Estimation

Density is the number of objects (animals) per unit area A. It is mathematically expressed as;

$$D = \frac{\tau}{A} \tag{20}$$

Implying that

$$\widehat{D} = \frac{y}{Ap} \tag{21}$$

Equation (21) represents an estimated parameter based on sampled y animals or units of interest with p detectectability or probability of detection. As indicated in [10], the density estimator is solely dependent on the Probability Distribution Function (pdf) at distance zero ($f(x = 0)$). However, with Bayesian philosophy which depends on the principle of total probability, where detections are made on selected transect lines with the number of detections conditional on the probability of detections on each transect i where $i = 1, 2, 3, \dots, n$ with which the density is seen to be expressed as

$$\widehat{D} = \frac{y}{Ap} = \frac{y}{2wl \sum_{i=1}^{n} P(T_i) * P(A/T_i)} \tag{22}$$

This equation appears to look independent of the probability distribution function and can easily be applied in both cases of known and unknown distributions of detection and detectabilities of units of interest.

3. Analysis and Discussion

3.1. The Case of Detected/Observed Elephants in the MNP

The study was structured to collect data in two locations (coded as 1&2); Two Sessions (coded as 1&2); and Two seasons (coded as 1&2) throughout the study period. In all, a total of 111 elephants were detected in 69 observations. The location, session, season, and month are considered as covariates in further analysis to identify their possible impact on the detection of units of interest in the study area.

Table 1. *Descriptive Statistics of Elephants in the MNP.*

Variable	Total Count	N	N*	Mean	v	StDev	Variance	CoefVar
Cluster size	69	69	0	1.6087	0.0951	0.7900	0.6240	49.11
Distance	69	69	0	22.05	2.99	24.87	618.59	112.81
N for Variable	Minimum	Q1	Median	Q3	Maximum	IQR	Mode	Mode
Cluster size	1.0000	1.0000	1.0000	2.0000	4.0000	1.0000	1	38

Distance	0.10	5.51	12.05	26.58	95.65	21.07	*	0
Variable	Skewness	Kurtosis						
Cluster size	1.20	0.88						
Distance	1.62	1.65						

Of the 69 observations made during the research period, a mean cluster size of about two elephants was observed at an average distance of about 22 meters from the centre of the transect line. As Table 1 reveals, all observed measurements turned to differ from the mean values of the distance and cluster sizes by + or - 24.87 and 0.7900, respectively. 75% of the observation is also at distances less than 26.58 meters.

3.2. Estimates of Rate of Detection of Elephants

Given an estimated population size of elephants of at least 450 and with an average observation time of 9 hours per day, it is expected that an average of at least 50 elephants is observed per day. The 50 units or above is thus set as a benchmark to test the rate of occurrence. On the assumption of the event (Detection of elephants per observation) as either a binomial or Poisson process that describes the occurrence of an event in a given amount of time (day), area, or other observation space, one sample Poisson rate is employed to compare the rate to a target value and to estimate the rate of occurrence. The 1-sample Poisson rate procedure calculates a confidence interval and conducts a hypothesis test for the rate of occurrence in a one-sample Poisson model. In this application, the investigation will be made to detect whether the number of detection exceeds a certain (benchmark value) per day/visit.

Table 2. *Test and CI for One-Sample Cluster size Poisson Rate.*

Test of rate = 50 vs rate not = 50						
Total		Rate of				
Variable	Occurrences	N	Occurrence	95% CI	Z-Value	P-Value
Cluster size	111	69	0.053623	(0.043648, 0.063599)	-321.37	0.000
	"Length" of	Mean				
Variable	Observation	Occurrence	95% CI	Z-Value		
Cluster size	30	1.60870	(1.30943, 1.90796)	-321.37		

From Table 2, a total of 111 observation units of interest were detected in 69 observation times within the 30 days length of the observation period of the survey. This implies that at the 95% confident level, the average detection rate per day equals $30 * 0.053623 = 1.60869 = 2$. This estimate is believed to fall between 39.2829 and 57.2388, which is a range of values that is likely to contain the rate.

3.3. Modelling Detectability Function of Elephants in the MNP

The modelling is carried out using DISTANCE. Modelling by DISTANCE is an iterative procedure [8] that first applies a key function to approximate the underlying data distribution and then adjusts the fit of the basic model by manipulating additional series parameters. This was performed in two stages.

The first stage is done allowing the detection of units of interest to be conditional on perpendicular distances only with the use of all possible key functions embedded in DISTANCE with all available adjustment series expansion combinations. The second phase involved making room for detectability to be dependent on additionally identified factors as covariates. This allows incorporating other factors likely to influence the ability to detect units of interest extends the key function and the adjustment terms making room for the covariates into either the scale or the shape or both [19]. The study observed that apart from distance, the model performance can be influenced by weather conditions, time of survey to mention but a few referred to as covariates. These covariates have some possibility and ability to affect the likelihood of detecting units of interest under study in the Mole National Park.

Table 3. *Half Normal Key with Varied Values of Truncation and Associated AIC Values.*

Filtration	Eff	Number Observed	Total Para	Specific parameters		ESW	D	N	P	f(0)	Log L	AIC Value	GoF Chi
				Key Para	Adj par								
Untrun	7.20	111	2	1	1	31.80	24.237	194	0.27	0.03	-452.58	909.16	0.000
40m	7.20	93	1	1	0	17.66	36.575	293	0.44	0.06	-311.57	625.15	0.443
	7.20	93	2	1	1	15.42	41.891	335	0.39	0.06	-162.49	328.98	0.368
60m	7.20	98	2	1	1	17.85	38.118	305	0.30	0.06	-348.34	700.68	0.091
	7.20	98	2	1	1	17.31	39.317	315	0.29	0.06	-149.63	303.25	0.263
80m	7.20	107	2	1	1	22.93	32.405	259	0.29	0.04	-407.46	818.91	0.004
	7.20	107	2	1	1	23.30	31.897	255	0.29	0.04	-164.91	333.82	0.036
100m	7.20	109	2	1	1	26.68	28.375	227	0.27	0.04	-428.11	860.23	0.000
	7.20	109	2	1	1	26.65	28.401	227	0.27	0.04	-154.30	312.61	0.000
120m	7.20	111	2	1	1	31.87	24.183	193	0.27	0.03	-452.70	909.40	0.000
	7.20	111	2	1	1	32.02	24.077	193	0.27	0.03	-153.82	311.65	0.000

Table 4. *Hazard Rate Key to Varied Values of Truncation and Associated AIC Values.*

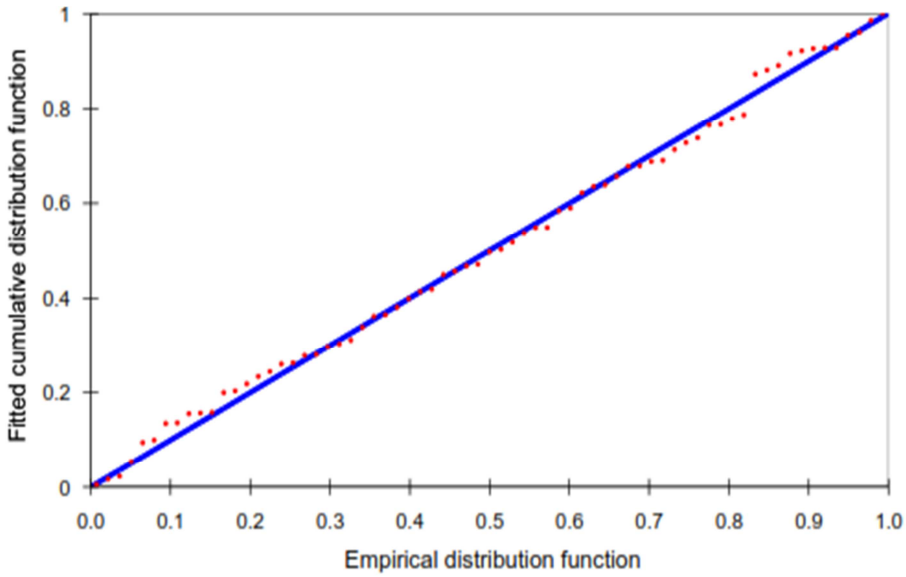
Filtration	Eff	Number Observed	Total Para	Specific parameters		ESW	D	N	P	f(0)	Log L	AIC Value	GoF Chi
				Key Para	Adj par								
Untrun	7.20	111	2	2	0	17.86	43.157	345	0.15	0.06	-439.10	882.20	0.000
40m	7.20	93	3	2	1	14.71	43.919	351	0.37	0.07	-309.23	624.45	0.326
	7.20	93	3	2	1	14.61	44.190	354	0.37	0.07	-161.58	329.17	0.146
60m	7.20	98	2	2	0	16.44	41.396	331	0.27	0.06	-347.14	698.29	0.236
	7.20	98	2	2	0	16.23	41.920	335	0.27	0.06	-148.76	301.52	0.434
80m	7.20	107	3	2	1	17.64	42.121	337	0.22	0.06	-404.71	815.42	0.000
	7.20	107	3	2	1	18.65	39.836	319	0.23	0.05	-163.78	333.56	0.006
100m	7.20	109	2	2	0	17.52	43.202	346	0.18	0.06	-423.05	850.10	0.000
	7.20	109	2	2	0	18.35	41.243	330	0.18	0.05	-151.37	306.75	0.001
120m	7.20	111	2	2	0	17.87	43.141	345	0.15	0.06	-439.13	882.27	0.000
	7.20	111	2	2	0	14.61	52.751	422	0.12	0.07	-141.76	287.51	0.033

With the application of the various keys at varied filter points as revealed from Table 3 and Table 4, the HN key produces a minimum AIC value of 303.25 at the 60m truncation with 8 histogram bin plots. The HR key, NE key, and the UN key functions produce a respective minimum AIC value of 287.51, 296.34, and 331.01 at various filter points as indicated in the tables in Appendix. In view of the respective AIC values, the HR key functions appear to produce a somewhat good fit compared to the other key functions. The question we pose here is how does the data support this? Exploring further to adjustment of terms and visual inspection on Qq plots with an additional test on Kolmogorov –Smirnov and Cramer –von Mises goodness of fit, we observed that the Qq plots reveal a visual representation of how well the individual key functions well fit the data observed. It is evident from the Qq plots that the elephant data is best described by the hazard key where in most points appears on or close to the blue line with little systematic departure.

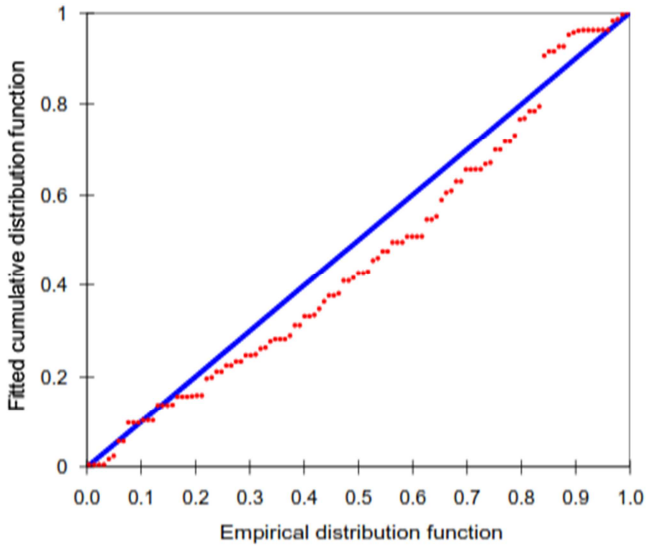
3.4. Kolmogorov-Smirnov and Cramer-von Mises family Test

(i). Kolmogorov-Smirnov test of the HR key
D_n= 0.0460
p = 0.9731

Cramer-von Mises family test of the HR key
W-sq (uniform weighting) = 0.0297
0.900 < p <= 1.000
Relevant critical values:
W-sq crit (alpha=0.900) = 0.0000
C-sq (cosine weighting) = 0.0150
0.900 < p <= 1.000
Relevant critical values:
C-sq crit (alpha=0.900) = 0.0000
(ii). Kolmogorov-Smirnov test of the NE key
D_n = 0.1132
p = 0.1162
Cramer-von Mises family test of the NE key
W-sq (uniform weighting) = 0.3556
0.050 < p <= 0.100
Relevant critical values:
W-sq crit (alpha=0.100) = 0.3472
W-sq crit (alpha=0.050) = 0.4621
C-sq (cosine weighting) = 0.2361
0.050 < p <= 0.100
Relevant critical values:
C-sq crit (alpha=0.100) = 0.2353
C-sq crit (alpha=0.050) = 0.3153



A: Hazard Rate (HR) function



B: Negative exponential (NE) function

Figure 3. *Global Qq plots of HR and NE keys.*

The Kolmogorov-Smirnov and Cramer-von Mises family test looks at the Qq plot of Figure 3 and transforms the location of the points into a test statistics with an associated p-value testing hypothesis that the model fits the data. In each case of the p values, we have no statistical evidence to reject the claim that the HR performs better than the NE key.

With the inclusion of adjustment terms to adequately describe the model, it is necessary to investigate to determine how many adjustment terms are needed to provide an optimum solution which adequately improves the maximum description with a balance between bias and variance in the parsimony principle. In this principle, an increase in the number of parameters results in a decrease in bias and an increase in variance.

The analysis of the models suggested a hazard rate key function with no adjustment term as the best model because it had the lowest AIC value. The detectability function.

$g(x) \propto key(x)[1 + series(x_s)]$ is found to be of the form $1 - e^{[-(x/\sigma)^{-b}]}$ with the scale parameter $\sigma = A(1) = 6.3894963 \approx 6.4$ and the shape parameter $b = A(2) = 1.454576 \approx 1.5$ respectively. Thus, the general form of the model is expressed as

$$g(x) = 1 - e^{[-(x/6.4)^{-1.5}]} \quad (23)$$

3.5. The Detection Function Model Conditional on Distance and Other Covariates

During the data collection period, the investigator identified session, season, and stratum as some variables which are considered to have a potential power to influence the detection function [19]. That is, the ability to observe units of interest may appear to depend on these factors other than distance alone. These factors are termed covariates considered as factors with two levels each. Incorporating these into the detection function where the covariates assume a variable z, then the model is considered to take the form $g(x,z) = key(x,z)[1 + series\ (x)]$ as suggested by [9] where $g(x,z)$ is the probability of detecting an object of interest at a distance x and covariates z.

Table 5. *MCDS with Hazard Rate Key with Varied Truncation and AIC Values.*

Filtration	Eff	Number Observed	Total Para	Specific parameters		ESW	D	N	P	f(0)	Log L	AIC Value	GoF Chi
				Key Para	Adj par								
Untrun	3.60	111	5	2	0	17.82	86.531	346	0.15	0.06	-435.73	888.46	0.000
40m	3.60	93	5	2	0	13.4	96.371	385	0.34	0.07	-307.68	625.35	0.066
	3.60	93	5	2	0	16.76	77.054	308	0.42	0.06	-159.37	328.74	0.028
60m	3.60	98	5	2	0	17.97	75.740	303	0.30	0.06	-344.09	698.19	0.114
	3.60	98	5	2	0	16.59	82.047	328	0.28	0.06	-144.76	299.53	0.021
80m	3.60	107	5	2	0	19.81	75.013	300	0.25	0.05	-404.27	818.55	0.000
	3.60	107	5	2	0	21.26	69.892	280	0.27	0.05	-166.10	342.19	0.000
100m	3.60	109	5	2	0	17.99	84.164	337	0.18	0.06	-420.58	851.16	0.000
	3.60	109	5	2	0	20.53	73.739	295	0.21	0.05	-148.47	306.94	0.000
120m	3.60	111	5	2	0	17.82	86.523	346	0.15	0.06	-435.76	881.51	0.000
	3.60	111	5	2	0	29.18	52.827	211	0.24	0.03	0.00	10.00	0.000

As indicated in Table 5, various filtration points were considered for both the HN and HR with observer group, season, and session as covariates at 2 levels with distances scaled by the maximum distance at which detection is considered possible in this study. According to the AIC values, the HR key appears more appropriate in fitting the

detection function with the covariates inclusion at 60m truncation with 8bins specified. Conditioning distance and other identified covariates yields a minimum AIC = 299 with $g(x)$ taking the form $g(x) = 1 - \exp(-(x/16.33957)^{-2.507})$ in which the shape parameter $= A(2) = 2.507$ and the scale parameter $= s = A(1) * Exp((A(3)) + (A(4)) + (A(5)))$ with estimated values $A(1) = 13.13, A(3) = 0.6966, A(4) = -0.4176$ and $A(5) = -0.06031$. Hence $s = 16.33957$.

We can observe that the covariates included in the hazard model have increased both the scale and shape parameters from 6.4 to about 16.33957 and 1.5 to about 2.507 respectively.

3.6. Enhancement of the Detectability Through Bayesian Approach

The Bayesian inference is a process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as prediction for new observations [1]. This approach is an alternative to the classical form of statistical analysis whose inference is done on the posterior which dependents on both the likelihood and the prior in which the parameters are seen to be random with a known probability distribution. The posterior is an updated form of the prior after the observation.

The decision to perform this analysis is based on the fact that both the variables of interest and the parameters to be estimated are considered as random variables believed to have or follow a particular probability distribution. This contrasts the classical or frequentist approach wherein parameters are considered fixed with unknown values which are found by maximizing the likelihood function. With the application of the Bayes theorem, the posterior distribution of the parameters given the data has a density proportional to the product of the likelihood of the data given the parameters and the prior distribution of the parameters.

In fitting the Bayesian model to estimate abundance and other parameters of interest, I adopt a simulation method using the Markov Chain Monte Carlo principle [14, 15] with data augmentation, which makes it possible to sample from very high dimensional joint posterior densities while ensuring that units undetected but present in the study area are catered for in the modelling process for both observed and unobserved units of interest.

3.6.1. Bayesian Models of Elephant Detectability

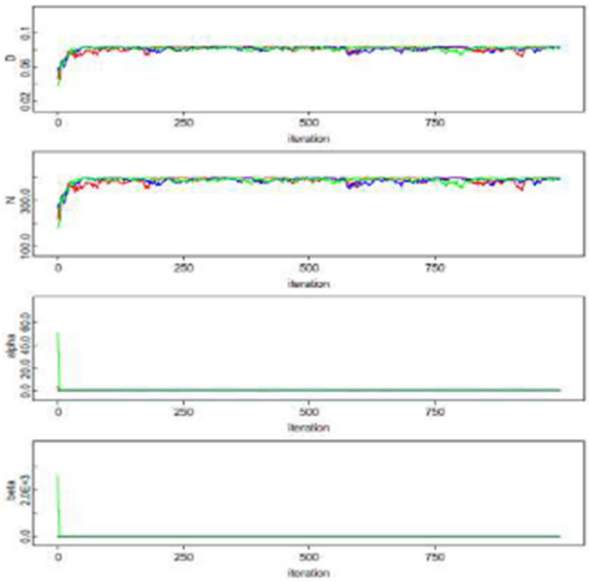
Recall that the posterior distribution is proportional to the product of the PRIOR and the LIKELIHOOD and on the assumption of a unit either to be detected or not believed to be characterized binomially and distributed as $P(x/\theta) = \theta^x(1-\theta)^{n-x}$. If the distribution of elephants in the study region is seen to be described by the Weibull distribution as seen earlier, which I considered to represent my PRIOR believe, then the posterior distribution associated with this BINOMIAL WEIBULL combination is defined as

$$p(\theta/x) \propto p(\theta) * p(x/\theta) = \{\theta x^{\theta-1} exp(-x^{\theta})\} * \{\theta^x(1-\theta)^{n-x}\} = x^{\theta-1}\,\theta^{1+x}(1-\theta)^{n-x} e^{(-x^{\theta})}\, x > 0 \quad (24)$$

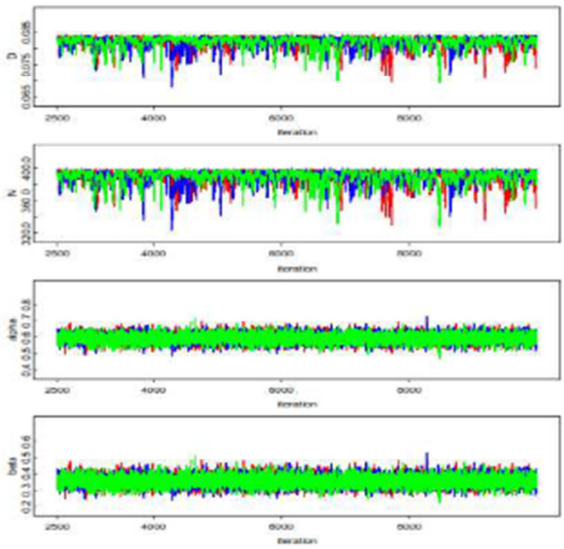
The above equation exhibits non-conjugacy property with the Weibull-Binomial combination, I have adapted the approach of [3] to model the relationship between distances of detection and cluster sizes of units observed through Bayesian approach using Gibbs sampling.

3.6.2. Abundance Estimation of Hazard Rate Detectability with Data Augmentation

The study organized the observed data according to a 4 X 5 factorial layout of mammal types by covariates wherein the units detected are augmented with unobserved units in the study region. By augmentation, I have observed that the study region contains no more than 398 elephants.



A: 1000 iterations



C: 10000 updates with discards

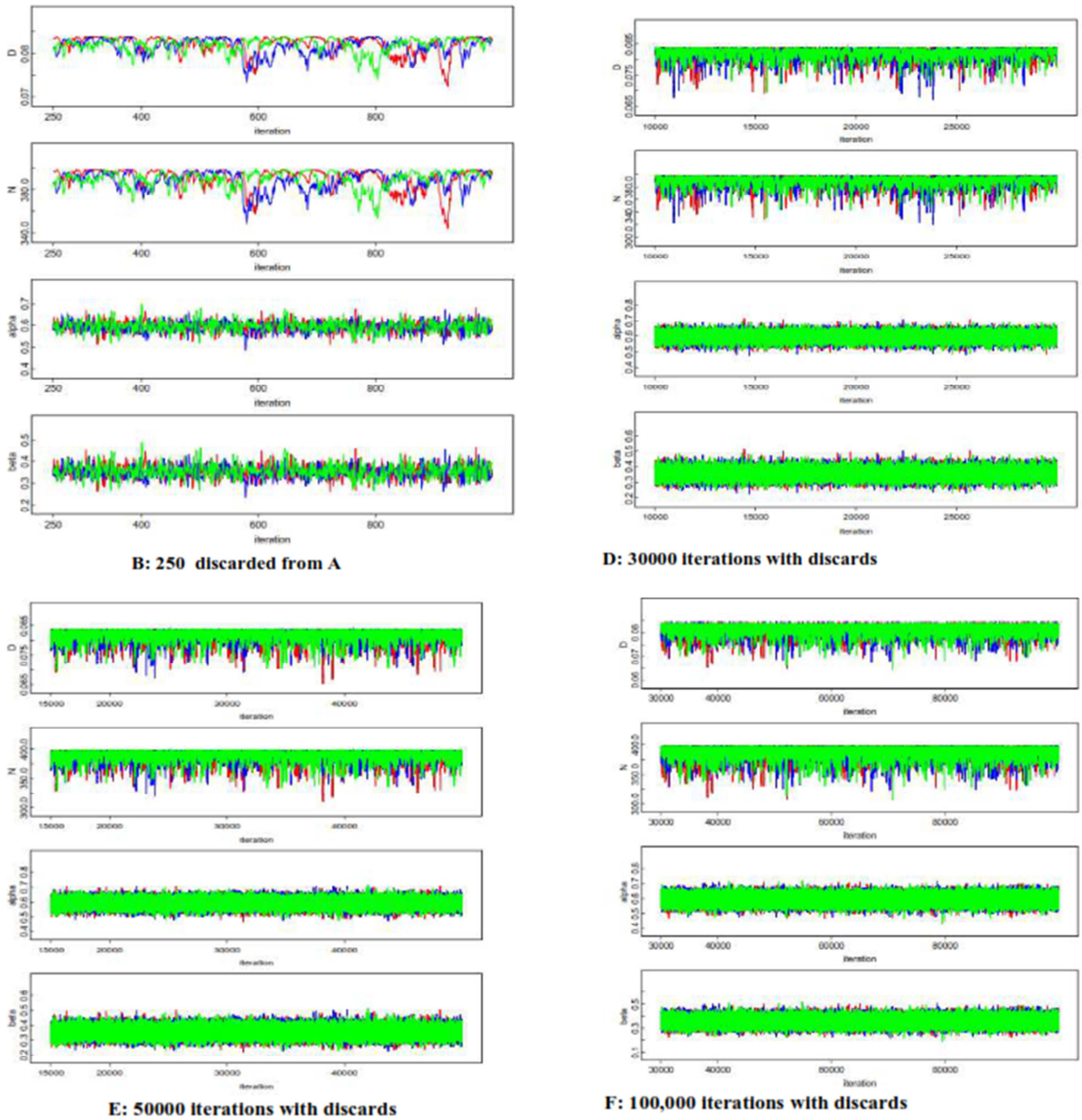


Figure 4. *History plots of different levels of iterations of three chains.*

Figure 4 presents the history plots for the first 1,000 iterations of all three chains through 100000 updates for the specified augmented model. The figure reveals that within the first few iterations of all three chains, the initially sampled samples from the target distribution contribute less significantly in parameter estimation, which eventually influences the mixing ability of the chains. Discarding the first 250 samples as burn-in and visually inspecting all labels confirm better mixing after discarding the initial samples even though convergence has not been properly achieved. There is a slow mixing of the chains and high autocorrelation and requires many more iterations to reach a stable equilibrium distribution (i.e., converged). With further updates up to 30000 through to 100000 as indicated in the above, I have realized that the convergence improves as the number of iterations increases with a relatively perfect and stable convergence at 50000 iterations and beyond.

By perusing the BGR plots as seen in Figure 5, at the varied level of iterations and burn-ins, we can observe the stability within and between chains occurring beyond the 20000 updates, wherein we can conclude to have obtained complete mixing with complete convergence among all three chains sampled from the same posterior distribution. For all BGR's, after several iterates, the red line representing the ratio begins

near 1 and remains there with the blue and green lines coming together to become horizontal beyond the 20000 iterates reflecting the initial samples that must be discarded as burn-in in the MCMC iteration. Values around 1 indicate convergence, with 1.1 considered as an acceptable limit by [13].

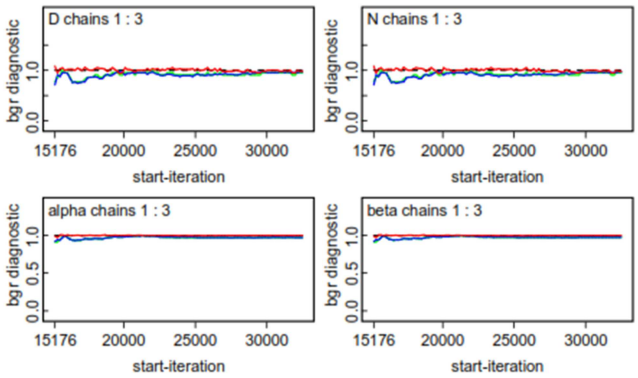


Figure 5. *BGR diagnostic plot from 50000 iterations and 15000 iterates discarded.*

3.6.3. Autocorrelation Plots of Burn-ins of the Hazard Rate Detectability

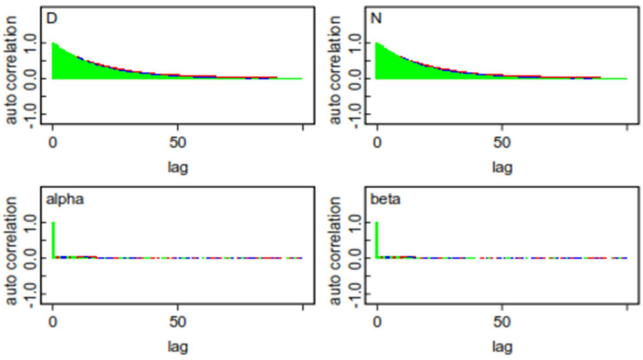


Figure 6. *Autocorrelation from 100000 iterations to 30000 burn-in.*

The autocorrelation plots represent a plot of lags on the x-axis and the magnitude of the autocorrelation (between -1 and +1) on the y-axis with the highest magnitude (+1) at lag zero (0) at which the remaining magnitudes at the various lags are compared and decreases steadily as the lags increases until it reaches a threshold lag beyond which it is essentially zero. Usually sampled iterates are realized conditional on the other until a realization of the target distribution where convergence between chains are achieved and autocorrelation is a quantitative measure of this dependence. Whereas correlation is a measure of association between two quantitative variables measured on the same object, autocorrelation refers to correlation with self. i.e., correlation within the same chain. Lag k autocorrelation in MCMC output is the correlation between samples drawn k iterations apart.

With all autocorrelation plots at different iterated values as revealed in Figure 6 with varied discarded initial samples, we can observe that all three superimposed chains obtained a zero autocorrelation beyond lag 50.

Table 6. *Summary statistics of 100000 iterations with 30000 burn-in.*

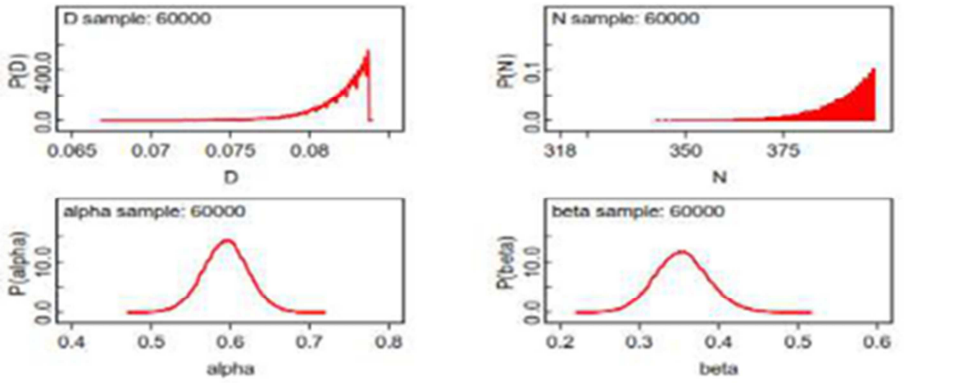
	Mean	sd	MC_error	val2.5pc	median	val97.5pc	Start	Sample
N	318.0	9.191	0.1198	314.0	392.0	398.0	30001	210000
Alpha	0.5938	0.0285	1.081E-4	0.5378	0.5938	0.6498	30001	210000
Beta	0.3535	0.03386	1.272E-4	0.2892	0.3526	0.4223	30001	210000

Table 6 revealed that the Bayesian estimate of the abundance of elephants in the MNP is estimated at 318, believed to lie within a credible range of 364 and 398 for all varied iterated levels of the hazard rate function with an error not exceeding 0.1927 and a standard deviation of about 9.191. As a rule of thumb, $MC\ error < 1 - 5\%\ of\ the\ posterior\ SD$ [2]. Following this rule, $MC\ error = 0.1198 < 1 - 5\%\ of\ the\ posterior\ SD = 1 - 0.05 * 9.191 = 0.54045$ which indeed is satisfied making the inference and estimates of the parameter reliable and valid. With 100000 iterations and 30000 burn–in, the Half Normal detectability with the Bayesian approach produced an abundance estimate of the elephant as 333.2 falling within a credible range of 333.0 and 391.0 with an MC error and SD of 0.5667 and 31.05 respectively.

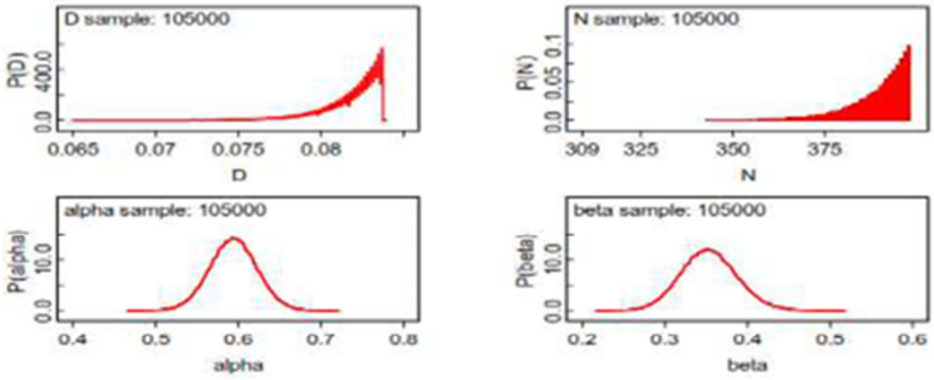
Similarly, $MC\ error = 0.5667 < 1 - 5\%\ of$ the

posterior $\text{SD} = 1 - 0.05 * 31.05 = -0.5525 > MC\ error$, an indication that model is not appropriate for the observed data. In the case of the negative exponential detectability, the abundance is estimated at 385.3, falling within a credible range of 352.0 and 352.0 with an MC error and SD of 0.1764 and 12.36 respectively. But $MC\ error = 0.1764 < 1 - 5\%\ of$ the posterior $\text{SD} = 1 - 0.05 * 12.36 = 0.382 > MC\ error$ which confirms the Hazard rate as best as were found by the classical approach.

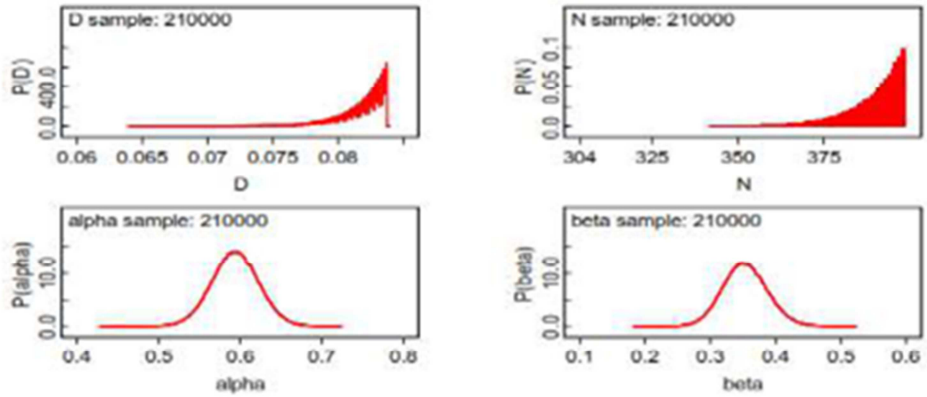
3.6.4. Posterior Probability Density Plots for Varied Iterates and Burn-ins



a: Posterior probability distribution plots after 30000 iterations



b: Posterior plots from 50000 iterations with 15000 burn-in



c: Posterior probability from 100000 iterations with 30000 burn-in

Figure 7. *Posterior probability at different iterations and burn-ins.*

With the convergence achieved at 500000 iterations and beyond as indicated in Figure 7, a kernel smoothed posterior distribution of all monitored parameters with the hazard rate, the Half Normal, and the Negative Exponential detectability functions, respectively, revealed that the kernel smoothed histogram of the DENSITY and ABUNDANCE parameters of both the hazard rate and the Negative Exponential functions appear negatively skewed with a symmetric shape in the ALPHA and BETA parameters with much smoother as the number of updates increases.

Table 7. *Bayesian and classical estimates of the abundance of elephants in the MNP.*

Function	Abundance_Classical	Abundance_Bayesian	Error of Estimation
Hazard Rate	328	318	0.01198
Half Normal	315	333	0.5667
Negative Exponential	439	386	0.01764
Weibull	N/A	307	0.01077

In Table 7, various estimates have been realized for the elephant species with a function specific for both the Bayesian and the classical approaches. Even though there exist differences in the estimation procedures, the findings, however, are similar.

3.6.5. Using the Principle of Total Probability

Suppose A represents an event that the unit of interest (elephants) is detected and T_i represents the event that the detection occur in transect i with partitions represented as follows:

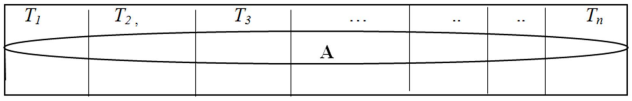


Figure 8. *Total Probability representation.*

By the principle of the total probability, the probability of observing event A which appears as a subset of Ti's as Figure 8 illustrates is estimated as; $P(A) = P(T_1) * P(A/T_1) + P(T_2) * P(A/T_2) + \cdots + P(T_n) * P(A/T_n)$

Table 8. *Total probability concept of detectability estimation with prior knowledge.*

	Function	Design	Detectability _ From Literature	Posterior Estimates
Observed	Hazard Rate	CDS	0.12	N/A
		MCDS	0.28	N/A
	Weibull	CDS	N/A	0.5736
		MCDS	N/A	0.5979
Simulated	Weibull	CDS	N/A	0.5149
		MCDS	N/A	0.5777

The restriction of four specific functions consisting of half normal, hazard rate, negative, exponential, and uniform for analyzing species with varied characteristics has proved to be insufficient in explaining the observational differences of plants and animals. The biological differences in plants and animals can be thought of to be reflected in functions that may appear to describe their detection and detectabilities. The use of prior knowledge in ecological studies subject to density and abundance estimation of rare events in the Bayesian ideology appears critical and necessary. As seen in Table 8, the Weibull function as determined by best fit detection and detectability of the elephant species in the MNP, appears to perform better in abundance in the Bayesian form than the classical method of statistical investigation of the functions specified in the distance software. This confirms the need for increase of the functional space of the application software for distance sampling with line transect applications to cater for all differences in biological characteristics of the units of interest under study.

In Table 6, we can observe that the posterior estimate of the updated Weibull prior on the binomial likelihood appear to increase the detectability of the elephant species in the MNP. This is seen to suggest that the use of prior information in the detection of rare species in ecological studies with the line transect methodology of distance sampling cannot be downplayed.

4. Conclusions and Recommendations

In conclusion, the classical or frequentist approach determines the abundance and detectability estimates of the elephant in the MNP with the hazard rate detectability function based on a minimum AIC of 287.51 at 422 and 0.12 respectively. Using the total probability concept, the detectability appears to be enhanced as seen in Table 6.

With the Bayesian principle using the Gibbs sampling approach, the research has established that, based on minimum error and runtime, the hazard rate function determines the abundance estimate of elephants at 318, believed to be within a credible range of 314 and 398. This has made the researchers to believe that the actual count of the elephant population in the MNP appears far less than what management of the park wants the public to believe. Moreover, the observation of units of interest conditional on distance has been modelled in a more general perspective with additional categorical predictors wherein all regression equations are provided at various factor level combinations with about 99.98% of explained variation in the model by the data augmentation process and less than 20% of the explained variability for non-augmented case.

Furthermore, in performing statistical investigations on rare and elusive population, it appears insufficient to model from the classical perspective, the use of PRIOR knowledge as seen in the Bayesian context cannot be underestimated. In addition, data recording and presentation in this type of study requires a more generalization in the form described in this study as "augmentation" process. Based on the findings of this study, the researcher wishes to propose the following recommendations based on the line transect application of the distance sampling methodology:

1) The concept of Total Probability with prior knowledge of animals and plants in line transect surveys must be well embraced.
2) Application of data augmentation process in data representation with Line Transect Methodology involving observation of more than one species at a time.
3) Periodic census must be conducted regularly to help in establishing the rate of extinction of units of interest in wildlife.
4) That Distance sampling data with line transect sampling methodology need not be analyse using only the classical reasoning. Attention must be given to the existence and availability of prior knowledge of the units under study.
5) That every effort should be made by researchers in ecological and biological science to make maximum use of prior knowledge of units of interest.

Conflict of Interest

The content of this paper contains no conflict of interest from any source.

References

[1] Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin (2009). Bayesian Data Analysis; 2nd ed., Chapman and Hall /CRC.

[2] Andrew Thomas (2014), Open BUGS Developer Manual, Dept. of Mathematics & Statistics, University of St Andrews, Scotland.

[3] Bradley, P. C. and Alan, E. G. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. Statistics and Computing. 1, 119 - 128. https://doi.org/10.1007/BF01880086

[4] Briggs, K. T., Tyler, W. B. and Lewis, D. B. (1985). Aerial surveys for seabirds: methodological experiments. Journal of Wildlife Management, 49.

[5] Burnham, K. P. and Anderson, D. R. (1976). Mathematical Models for non-parametric inferences from line transect data. Biometrics, 32, 325 – 36.

[6] Buckland, S. T., Rexstad, E. A., Marques, T. A., Oedekoven, C. S. (2015). Modelling Detection Functions. In: Distance Sampling: Methods and Applications. Methods in Statistical Ecology. Springer, Cham. https://doi.org/10.1007/978-3-319-19219-2_5

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

[8] Buckland S. T., Anderson D. R., Burnham K. P. & Laake J. L. (1993). Distance sampling Estimating abundance of biological populations. Chapman & Hall, London.

[9] Buckland S. T., Goudie I. B. J. & Borchers D. L. (2000). Wildlife population assessment: past developments and future directions. Biometrics 56, 1-12.

[10] Burnham, K. P., Anderson D. R. & Laake, J. L. (1980). Estimation of density from line transects sampling of biological populations. Wildlife Monographs 72: 1-202.

[11] Cassey P. & McArdle B. H. (1999). An assessment of distance sampling techniques for estimating animal abundance. Environmetrics 10, 261-78.

[12] Eberhardt, L. L (1978b). Appraising variability in population studies. Journal of Wildlife Management, 42, 207 – 38.

[13] Gelman, A., & Hill, J. (2007). Data analysis using regression and multi- level/hierarchical models. New York: Cambridge University Press.

[14] Gilks W (1992) Derivative-free adaptive rejection sampling for Gibbs sampling. In Bayesian Statistics 4, (J M Bernardo, J O Berger, A P Dawid, and A F M Smith, eds), Oxford University Press, UK, pp. 641-665.

[15] Gilks W R, Richardson S and Spiegelhalter D J (Eds.) (1996) Markov chain Monte Carlo in Practice. Chapman and Hall, London, UK.

[16] Hayward M. W., de Tores P. J., Dillon M. J., Fox B. J. & Banks P. B. (2005). Using faecal pellet counts along transects to estimate quokka (Setonix brachyurus) population density. Wildlife Research 32, 503-7.

[17] Katara S., S. K Amponsah and Bashiru I. I. S. (2018). Distributional analysis with line transect methodology of the distance sampling techniques: Case of large mammals of the Mole National Park (MNP) of Ghana. African Journal of Mathematics and Computer Science Research. 11 (1), 1-13.

[18] Len Thomas, Stephen T. Buckland, Kenneth P. Burnham, David R. Anderson, Jeffrey L. Laake, David L. Borchers & Samantha Strindberg (2002). Distance Sampling, Volume 1, pp 544–552 John Wiley & Sons, Ltd, Chichester, in Encyclopaedia of Environmetrics (ISBN 0471 899976).

[19] Marques, T. A., Thomas, L., Fancy, S. G., and S. T. Buckland (2007). Improving estimates of bird density using multiple covariate distance sampling. The Auk. 124: 1229 – 1243.

[20] Marshall A. R., Lovett J. C. & White P. C. L. (2008). Selection of line-transect methods for estimating the density of group-living animals: lessons from the primates. *American Journal of Primatology* **70**, 452-62.

[21] Overton, W. S., and D. E. Davis (1969). Estimating the number of animals in wildlife populations. Pp. 405 – 455. Wildlife Management Techniques. The Wildlife Society, Washington, D. C.

[22] Sadlier L. M. J., Webbon C. C., Baker P. J. & Harris S. (2004). Methods of monitoring red foxes Vulpes vulpes and badgers Meles meles: are field signs the answer? Mammal Review 34, 75-98.

[23] Schwarz C. J. & Seber G. A. F. (1999) Estimating animal abundance: review III. Statistical Science 14, 427-56.

[24] Seber, G. A. F. (1982). The estimation of animal abundance and related parameters, Macmillan, New York.

[25] Steven K. Thompson (2012). Sampling Third Edition, Willey Series in Probability and Statistics.

[26] Sutherland W. J. (2006). *Ecological census techniques: A handbook*. Cambridge University Press, Cambridge.